ALGUNOS ESTUDIOS CUANTITATIVOS APLICADOS EN PRESENCIA DE GRANDES MASAS DE DATOS EN LAS CIENCIAS SOCIALES

Carlos N. Bouza Herrera* , Byron Oviedo** y Sira Allende*
*Universidad de La Habana
**Universidad Técnica Estatal de Quevedo

ABSTRACT
Engineers and managers of firms are involved in the "numeric revolution", which is identified with the so-called BIG DATA problem. When dealing with Big Data is needed to use particular methods of low computational complexity .
The aims of this paper are related with promoting the interest of taking into account the problems posed by dealing with Big Data, when developing social science studies. The use of quantitative methods is discussed considering the use of clustering techniques . An illustration is presented by analyzing a real case.

KEYWORDS: Data Mining, Big data, Clúster, Data Analysis, Social Sciences

MSC: 91F99, 62P12, 62H30

RESUMEN
Los ingenieros y ejecutivos de las empresas, se encuentran inmersos en la "revolución numérica" que identificamos la confrontación con BIG DATA. Al tratar con Big Data se abre la necesidad de desarrollar métodos particulares de baja complejidad computacional dado que la complejidad computacional. En esta contribución intentamos promover el interés en considerar el reto que pone ante los estudios en las ciencias sociales la existencia de Grandes masas de datos. Solo el uso de métodos cuantitativos puede permitir establecer las características esenciales de la data bajo estudio
Se discute el uso de métodos de aglomeración. Se ilustra como estos fueron usados en un estudio concreto.

PALABRAS CLAVE: Data Mining, BigData, Clústeres, análisis de datos, ciencias sociales

1. ALGUNOS ELEMENTOS

Los ingenieros y ejecutivos de las empresas, ya sean pequeñas o grandes se encuentran inmerso en la revolución numérica que identificamos la confrontación con BIG DATA, o datos de alta complejidad VVV (Volumen, Velocidad, Veracidad). Cabe a la estadística un papel crucial dada la existencia de métodos y modelos y algoritmos para procesar los datos simplificar la información relevante. En fin, es capaz de resumir la información que bien dan los datos al determinar prototipos o representantes, particionar la población al construir clústeres o segmentar los datos, desarrollar la presentación de los datos mediante representaciones gráficas con sentido y que sean interpretarles fácilmente. Al tratar con Big Data se abre la necesidad de desarrollar métodos particulares de baja complejidad computacional dado que la complejidad computacional que se aparece es inconmensurablemente grande. El interés en el tema ha dado lugar a libros especializados como Scholz (2017), amén de un gran número de papers.
Si pensamos en la data y la perspectiva contemporáneas notamos que han emergido una gran cantidad de datos generados por el hombre y que son de uso en la investigación social. Cada individuo se asocia a información sobre los textos que genera, la auto-representación del mismo, su intercambio de mensajes, sus datos identificativos (fecha de nacimiento, sexo, estudios etc.) y otros como gastos etc. Esta data en general está no estructurada y posee una riqueza semántica y nos lleva a pensar que cada individuo es representado por una larga sucesión de bites de información (¿miles?) y por tanto cuando representamos una población podemos tener una matriz con miles de columnas y millones de filas. Esto representa que ante cada estudio vemos que los métodos convencionales se enfrentan a la llamada "Social Data Explosion". Esta ha generado la necesidad de elaborar nuevas teorías y estudios sobre la llamada Big Social Data (BSD). Profundas discusiones han dado lugar a un cuerpo de ideas como las discutidas por Dasgupta, (2013), Einav-Levin (2014, Gandomi- Haider (2015), Belsey . (2005), Chen (2010), por ejemplo.
Nos enfrentamos a un fenómeno de naturaleza numérica muy complejo pues ahora tenemos la posibilidad de recolectar muchos datos provenientes de diversas fuentes, como las transacciones en el comercio, los censos, los datos meteorológicos, y en general los provenientes de las mediciones hechas a tiempo real como son los

capturaos por cámaras de seguridad, cajas de supermercados, tráfico en las WEB. Además, estos son de muy diversa índole, pues tenemos acceso al mismo tiempo de textos, variables medidas, datos categóricos, imágenes etc. Un análisis de estos aspectos puede obtenerse en Marwick (2015), King (2011) y Kum-Ahalt-Carsey (2011).

2. BIG DATA

Como Big Data se identifican los conjuntos de datos cuyo volumen hace que no podemos manejarles con los comunes sistemas de cómputo. En particular se identifican como tales datas que tienen un volumen que se mueve entre of 100 terabytes hasta peta bytes, pueden estar estructurados o no y que se actualizan constantemente,
BD plantea un reto pues en ella aparecen nuevas formas de la data que requieren de redefinir como tratar con lo cualitativo junto con lo cuantitativo. Esto es un reto para las ciencias de los datos y ha dado lugar a una nueva rama de especialización: Data Analytics.
Podemos citar los aspectos más importantes que caracterizan la problemática actual del mundo del BD:

- El Volumen: este se incrementa dado el uso proliferante de imágenes de satélites, los que cubren el planeta; la expansión de sensores, la telefonía celular y el uso de las cámaras para hacer de fotos y videos y subirles al mundo digital
- La Velocidad: dado que las imágenes se mueven al a velocidad de la luz se plantea que el procesamiento necesariamente deba acercarse a ella cada vez más.
- La Veracidad: se plantea si los atributos que se obtienen garantizan una calidad adecuada por pasar algunas pruebas de control. Métodos para hacer cruzamientos con información real, identificar observaciones extremas etc., deben ser desarrollados e incluidos en las herramientas.
- El Valor: al trabajar en tiempo real las decisiones deben tomar en cuenta la dinámica de cambio del fenómeno estudiado. Técnicas analíticas particulares deben ser consideradas. Hay una necesidad de su desarrollo y evaluación

Hasta donde llegará el volumen del BD, no es predecible. Hoy se suben más videos en un día que los que se hacían en los primeros 50 años de la televisión. Un cálculo conservador es que en la nube de datos hay cerca de 2.8 trillones de gigabytes. Consideremos por ejemplo un estudio de las parejas casadas hace dos años en un país. Si nos interesan los problemas sociales de ellas podemos medir estabilidad matrimonial, gustos, inversiones personales y de pareja, proyección de la familia etc. Analizaremos los records de los individuos casados hace dos años en la nube de datos, en tiempo real. Estos cambian diariamente. Entonces de cada pareja podemos considerar datos de ella, como la foto de la boda, y multiplicaremos por dos el número de datos en la información sobre el uso de los celulares, los datos socio demográficos, la información colgada en las redes sociales como Facebook y Twitter etc. Como obtener la información, como manejarla y el cómo filtrar esos datos, llegando a un número manejable de valores, es un problema del tipo BD en estudios sociales.
Nos enfrentamos al leer literatura especializada con nuevos y variados conceptos en las investigaciones que propenden a ayudar a entender el ambiente digital y sus efectos en la sociedad. BD plantea un reto pues en ella aparecen nuevas formas de la data que requieren de redefinir como tratar con lo cualitativo junto con lo cuantitativo. Esto es un reto para las ciencias de los datos y ha dado lugar a una nueva rama de especialización: Data Analytics.
La literatura sobre BSD, Mandiberg (2012), Varian (2013), Einav- Levin (2014) , parte de diversos puntos de vista. Lo cierto es que BSD es usada generalmente para obtener una visión, de lo que reflejan los datos generados por los grupos sociales y las interacciones entre los individuos, para poder describir o predecir con el propósito de influir en las decisiones de los grupos estudiados. Lo más común es que las investigaciones apunten hacia el desarrollo de estudios analíticos y que no se preste mucha atención al reto conceptual que plantea el uso de BSD. Sin conceptualizar, es difícil (¿imposible?) entender los fenómenos reflejados por la data. El manejo de una conceptualización por parte de investigadores plantea retos que se deben enfrentar. Un enfoque se basa en considerar que BSD es una ciencia que trata con las relaciones entre el mundo físico y el de los datos existentes en la nube de información. Como todo proviene de lo que explicita la masa de individuos (mundo real) y que esto se refleja en la data que les identifica en la nube. Al enfocar desde esta posición, deben resolverse los problemas que plantea el volumen y la variedad de los datos, así como la

velocidad de acceso y procesamiento computacional. Los que usan este enfoque tratan de clarificar los conceptos de como resolver problemas de la semántica y de lo difuso, presentes en frases, imágenes etc., y en cómo se hacen las interconexiones entre ellas. Otra forma de ver estos problemas es considerar que se trata con la relación entre lo físico, que le da un carácter cuantitativo para el estudio social, y las conexiones dentro lo subjetivo (ideas, preferencias, intercambios entre los elementos sociales). Quiere inferirse sobre el comportamiento de los agentes humanos. En tal caso, se considera que la data generada por el mundo físico revela como los agentes se comportan. Entonces se trata primero de los problemas de ubicar la información en la nube que permite describir, eficientemente, la subjetividad del mundo físico. Se enfocarán los investigadores en ver cómo identificar los datos más importantes en la nube, dada lo dificultad de identificar que data es la adecuada, y como esta varia en el tiempo. Vea Gandomi-Haider (2015), King (2011).
El mundo de la informática considera que los datos son primarios para la BSD y que solo ellos pueden caracterizar el panorama social. No hace falta más que manejar adecuadamente los datos en la nube y determinar un cubo. Los estudios se enfocan hacia la pérdida de información, el efecto del tamaño de los vectores en las dinámicas que les hacen variar con el tiempo y en tratar eficientemente con la estructura información para determinar el cubo. Ver Bello-Orgaz-Camacho (2014), Dasgupta, A. (2013) por ejemplo

3. BIG DATA EN LAS CIENCIAS SOCIALES

Puede decirse que viniendo de las Ciencias Sociales lo usual es que se considere que la BSD caracteriza la sociedad. Burges and Burns son de esta línea de pensamiento. En él se considera que una BSD no es sino una expresión de la sociedad en su conjunto y que deben reverse los problemas de accesibilidad a los datos. En este caso se concentran las investigaciones en ver los aspectos de autenticidad de los datos obtenidos, la accesibilidad a ellos y a evaluar la potencialidad de una BSD para eliminar investigaciones particulares para generar datos. Entonces se prioriza el ver como a partir de estudiar una BSD se minimice (¿elimine?) la necesidad de interactuar con el mundo físico
Si seguimos una línea de pensamiento como la de Bello-Orgaz et al (2014) BSD debe ser visto como un método novedoso para que las compañías puedan extrae, r información útil para ellas. Así que el interés se centra en mejorar la velocidad de acceso, caracterizar utilitariamente la variedad, valor y veracidad de la data. Absolutizar uno de esos enfoques es un error. Debemos ver que en el concepto mismo de BSD aparecen problemas comunes. Al investigar, los especialistas darán prioridad a los aspectos que sus conceptos identifican mejor. Lo cierto es que las ciencias que hagan uso de BDS plantean problemas nuevos al novedoso campo de la analítica de los datos (data analytics). Las universidades están en general no enfrentando la formación de especialistas que posean las capacidades necesarias para ser analistas de BSD con la velocidad con que es necesaria.
Cuando hablamos de CS identificamos investigaciones y aplicaciones que integran a las Ciencias Sociales las modernas herramientas de la computación. Sus fundamentos van a requerir de la Psico-sociología, Sociología, Análisis de Redes Sociales, Antropología etc. Otras herramientas ya están bien establecidas en áreas como las de Organización, Comunicación, Computación. Esto fija la existencia de una importante interacción entre las nuevas tecnologías de la comunicación y la sociedad. Esta interacción favorece el desarrollo de ambas. La CS permite modelar usando la moderna tecnología computacional para medir la auto-representación, la inter-comunicación y también soporta el desarrollo y mantenimiento de las relaciones entre individuos a través de lo digital. Así, las infraestructuras de comunicación incrementan su importancia y desarrollan nuevas herramientas (Webs, Bases De Datos, Multimedia, Tecnologías sin cables etc.). Por ello el entramado de compañías y negocios, gobiernos y partidos políticos requieren de especialistas de nuevo tipo. Probeles de est tipo son abordado por Katz (2008), Kum-Ahalt- Carsey (2011).
Cuando hablamos de la ciencia de los Grandes Volúmenes de Datos (Big Data Science) nos referimos a que ella trata del procesamiento y manejo de grandes volúmenes de datos, a gran velocidad y que trata de resolver el conflicto creado por la una gran variedad de los datos. Ella trata de extraer una valoración confiable y valiosa del fenómeno bajo estudio que generó la data. Así que, al hablarse del papel de los estudios de "Big Data", se sabe que se busca dar un servicio a aplicaciones que tratan con un problema de una escala enorme de datos. Este tipo de trabajo es necesario para las corporaciones para manejar en sus sistemas.
Los especialistas que laboran pegados a los negocios centran su interés en la tecnología de BD, dada por la posibilidad de estudiar grandes masas de personas, inmiscuyéndose en su vida privada a través de sus datos generados por las redes sociales, el uso de la telefonía celular tarjetas de crédito etc. Esto abre un campo a críticas sobre el efecto de esto. Al tener acceso información privada debe conceptualizarse los elementos éticos y la validez de la información captada de la nube de datos con fines que pueden afectar al individuo.

Cuando hablamos de Data Analytics consideramos que a través de su uso vamos a poder configurar el panorama que describe la existente masa de datos accesibles. Esta permite hacer una descripción de la data, permite explorarla para descubrir correlaciones desconocidas, hacer predicciones para predecir eventos, ciclos y tendencias y además para prescribir que acciones tomar. Esto da a los métodos de la estadística un rol que nunca antes fue concebido y a la informática un campo muy amplio de desarrollo. El aprendizaje sobre los problemas estudiados es uno de los aspectos más importantes y existe un amplio arsenal de métodos. Muchos de ellos aparecen en Hastie-Tibshirani- Friedman (2009):
Es claro que BD plantea un reto para los teóricos, pues el uso adecuado de la BSD permite guiar a elaborar teorías y hacer aplicaciones en diversas áreas de la sociedad. Este es un campo multidisciplinario que abre la necesidad de elaborar teorías fundamentadas sobre como estudiar los fenómenos sociales. Este es un campo donde deben laborar mancomunadamente científicos especializados en lo social, lo del comportamiento humano, lo del aprendizaje automático, lo de la computación así como, matemáticos y físicos.
Las herramientas cuantitativas más usadas en BD son:

- Clasificación
- Clustering
- Regresión
- Simulación
- Detección de Anomalías
- Predicción Numérica
- Optimización

Estas son usadas en forma combinada. No existe una teoría sobre cómo hacer la analítica de los problemas donde se trata con BD. En particular la problemática planteada a la informática lleva al desarrollo de softwares que sean especialmente eficientes.

4. UNAS APLICACIONES

En esta sección veremos algunas aplicaciones a problemas reales en las que se ilustra el uso de algunas de estas herramientas en estudios de las Ciencias Sociales.

Aplicación 1.
Geboers et al. (2014), Monash (2010) plantearon como la finalidad última de estudiar el comportamiento de los estudiantes es proveer de herramientas para mejorar el desempeño de la educación. Esto ha dado pie a innumerables discusiones académicas. Las investigaciones empíricas por su parte, dan una visión de las problemáticas particulares.
Un problema que es de interés por las ciencias sociales estudiar el comportamiento del rendimiento estudiantil. Algunos problemas de BigData en tales estudios de este tipo aparecen en Van Dewerfhorst-Mijs (2010), Freire (2015), Geboers et al. (2014), Chetty-Friedman-Rockoff (2017). Nosotros hemos analizado este problema en Allende et al (2017). Este estudio lo catalogamos como un problema de Big Data. Se debía analizar 40 782 datos provenientes de cinco preuniversitarios y 178 variables para cada entrada. El objetivo era explicar un modelo para evaluar el comportamiento de los estudiantes en la enseñanza universitaria a partir de los resultados en la enseñanza pre-universitaria.
Aplicamos el Análisis de Componentes Principales (PCA) para reducir el número de variables en el modelo. Este es un procedimiento matemáticos que se basa en el uso de una transformación ortogonal que convierte el conjunto de observaciones valores de las variables en unas no correlacionadas. Esta se denominan componentes principales. El número de ellos es menor o igual al de las variables originales. En el primer componente se agrupan las que explican la mayor cantidad de la varianza, en el Segundo estarán aquellas que explica la mayor parte de la varianza residual. Esto se repite. Los componentes son ortogonales a los componentes.
En el ejemplo que consideramos solo las variables con peso estimado del factor mayor que 0,100. Bajo esta restricción obtenemos

Tabla . Los componentes son ortogonales a los componentes.

	Primer Componente	**Segundo Componente**	**Tercer Componente**
Matemática 1	0,752	0,127	0,116

Matemática 2	0,650	0,111	0,107
Matemática 3	0,788	0,273	0,127
Física 1	0,512	0,771	0,164
Español 2	0,539	0,261	0,107
Inglés	0,178	0,690	0,142
Educación física	0,106	-0,205	0,536
Porciento acumulado del total de la varianza explicada	0,685	0,803	0,884

Las dos primeras componentes juntas explican más del 80% de la varianza. Por ello tiene sentido considerarles como las importantes para explicar el comportamiento en la universidad. La primera componente podemos llamarla "componente matemática", partiendo del hecho de que las notas en matemática son las variables más importantes y podemos evaluarlas como importantes en la promoción en la universidad las calificaciones en idioma español 2. La segunda podemos llamarla "complementos" donde son importantes la nota de Física 1 y el conocimiento del Inglés.
Nosotros fijamos la búsqueda de 5 clústeres y utilizamos la primera componente primero el segundo después. Evaluamos cuan similares fueron las particiones computando, para cada estudiante usado para determinar la partición,

$$I(u,j) = \begin{cases} 1 \; si \; u \; y \; j \; están \; en \; el \; mismo \; clúster \; en \; las \; dos \; particiones \\ 0 \; en \; otro \; caso \end{cases}$$

Evaluamos la similitud al computar

$$S(clustering) = \frac{1}{M(M-1)} \sum_{j=1}^{M} \quad \sum_{u \neq j} I(u,j)$$

En este caso

$$S(clustering) = 0{,}661$$

Usando la muestra de validación, de tamaño M*, cada estudiante fue clasificado utilizando el criterio

$$u \in C(t) \; si \; \left\| X(u) - p_{C(t)} \right\| = Min_h \left\| X(u) - p_{C(h)} \right\|$$

El clúster C(t) es representado por el centroide $\boldsymbol{p_{C(t)}}$ y utilizamos como norma la Euclideana. La eficiencia de la identificación fue evaluada al computar el número de estudiantes que fueron mal clasificados. Es decir, se calculó

$$P(mal \; clasificado) = número \; de \; mal \; clasificados \; / M^*$$

En el ejemplo se obtuvo

Componente	$\boldsymbol{P(mal\ clasificado)}$
Primer Componente	*0,481*
Segundo Componente	*0,640*

Por tanto, el uso del componente matemático posee una mejor predicción del éxito de un estudiante de pre universitario.
No se tomó en cuenta las carreras. En un estudio más profundo esto debería llevarse a cabo.

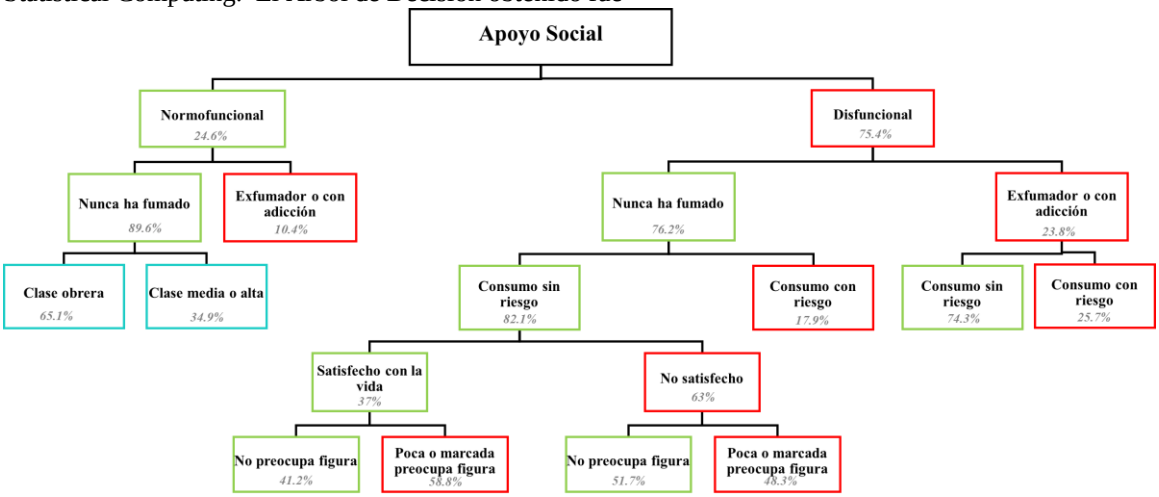
Aplicación 2.
Andrade-Sánchez et al. (2019) realizaron un estudio transversal, de una población de 1267 adolescentes de doce escuelas del estado de Colima, México; 589 (46.5%) hombres y 678 (53.5%) mujeres, con una media de 16.5 años de edad (±0.8). obtuvieron información sobre aspectos clasificados como Apoyo Social Funcional, Clase Social, Tipo de Familia , Patrón de Alimentación, Actividad Física, Preocupación por la figura, Adicción al Tabaco Consumo de Alcohol, Satisfacción con la vida. Realizaron un análisis de clases latentes en la búsqueda del mejor predictor a través del coeficiente de predictividad definido por:

$$\tau_{i/j} = \frac{\left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{f_{ij}^2}{f_{..} f_{.j}} - \sum_{i}^{I} (f_{i.}/f_{..})^2 \right)}{1 - \sum_{i=1}^{I} (f_{i.}/f_{..})^2} = \frac{\Phi^2}{1 - \sum_{i=1}^{I} (f_{i.}/f_{..})^2}$$

y usando el el índice de Catanova calculado a través de $C = (I-1)(J-1)\,\tau_{i/j}$. I y J representan, respectivamente, el número de categorías de la respuesta y el predictor. Desarrollaron un Análisis de Correspondencias No Simétrico (ACNS) definiendo las categorías fuertes y débiles como sigue

$\varphi_{j1} \geq 1$ Categorías fuertes por la derecha.
$|\varphi_{j1}| < 1$ Categorías débiles.
$\varphi_{j1} \leq -1$ Categorías fuertes por la izquierda.

donde φ_{j1} representa la correspondiente coordenada sobre el primer eje factorial del ACNS conformado con la variable respuesta y el correspondiente predictor. Los cálculos fueron realizados a través de R Project for Statistical Computing. El Árbol de Decisión obtenido fue



Asique detectaron nueve perfiles resultantes

Nodo	Descripción	% de la población
1:	Familia normofuncional y nunca han fumado.	22.05%
2:	Familia normofuncional y exfumador o con algún tipo de dependencia al tabaco.	2.56%
3:	Familia disfuncional; nunca han fumado; bebedores sociales; satisfechos con la vida; no les preocupa su figura.	7.18%
4:	Familia disfuncional; nunca han fumado; bebedores sociales; satisfechos con la vida; de poca a marcada preocupación por su figura.	10.26%
5:	Familia disfuncional; nunca han fumado; bebedores sociales; no satisfechos con la vida; no les preocupa su figura.	15.38%
6:	Familia disfuncional; nunca han fumado; bebedores sociales; no satisfechos con la vida; de poca a marcada preocupación por su figura.	14.36%
7:	Familia disfuncional; nunca han fumado; consumo de riesgo de bebidas alcohólicas.	10.26%
8:	Familia disfuncional; exfumador o con algún tipo de dependencia al tabaco; bebedores sociales.	13.33%
9:	Familia disfuncional; exfumador o con algún tipo de dependencia al tabaco; con consumo de riesgo de bebidas alcohólicas.	4.62%

Los caracterizados por el perfil 1 tienen niveles normales en casi todas las variables; en el perfil 3 se agrupan quienes tienen algún tipo de disfunción (leve o grave) en su familia; y en el perfil 5 se detectó que los jóvenes no están satisfechos con la vida. Ellos conforman un grupo

En un segundo grupo están estudiantes sin hábitos nocivos de salud y además tener una familia con algún tipo de disfuncionalidad. Estos se preocupan por su figura más de lo recomendado (perfil 4) y no tienen niveles adecuados de satisfacción con la vida.

El tercer grupo está conformado por el resto de los perfiles (2, 7, 8 y 9) se caracterizan por hábitos nocivos, como son el consumo de tabaco con algún tipo de adicción y el consumo de riesgo de bebidas alcohólicas.

Aplicación 3.

Machado-Daza (2918) utilizó la base de datos de Scopus para caracterizar el comportmsiento del tema "Valor percibido por el cliente" en el periodo 1963-2017. Se consideraron como filtros en la investigación Business, Management and Accounting; Social Sciences y Econometric, Economic and Finance. Una vez filtrada la base se usó VOSviewer v.1.6.5 para construir diagramas bibliográficos y visualizar la co-citación de revistas y co-ocurrencia del corpus de texto extraído del título y resumen de cada uno de los artículos. Para el análisis de co-ocurrencia, se utilizó la función de minería de texto del VOSviewer.

En el anlisis se hizo uso de Principales Componentes del concepto de VPC el que está compuesto por: calidad del producto, calidad del servicio y precio el que midió mediante

$$\text{Valor Percibido Para El Cliente} = \frac{\text{Calidad el producto } + \text{Calidad del servicio}}{\text{Precio}}$$

Los documentos aparecieron en 96 vea los 7 más importantes en la tabla siguiente

Tabla 5. Ranking de Publicaciones académicas

Documento	n	SJR	Cuartil	índice H
Industrial Marketing Management	14	1,413	Q1	90
International Journal Of Contemporary Hospitality Management	5	1,329	Q1	35
Journal Of Retailing And Consumer Services	5	0,669	Q2	42
Service Industries Journal	5	0,471	Q2	42
Actual Problems Of Economics	4	0,124	Q4	6
International Journal Of Production Research	4	1,445	Q1	91
International Journal Of Service Industry Management	4	0,781	Q1	69

La Figura 1 muestra los resultados de co-citación de las revistas que reciben no menos de 50 co-citaciones. El tamaño del círculo refleja el número de citas que la revista ha recibido, mientras que la distancia entre dos revistas indica la fuerza de relación entre ellas.

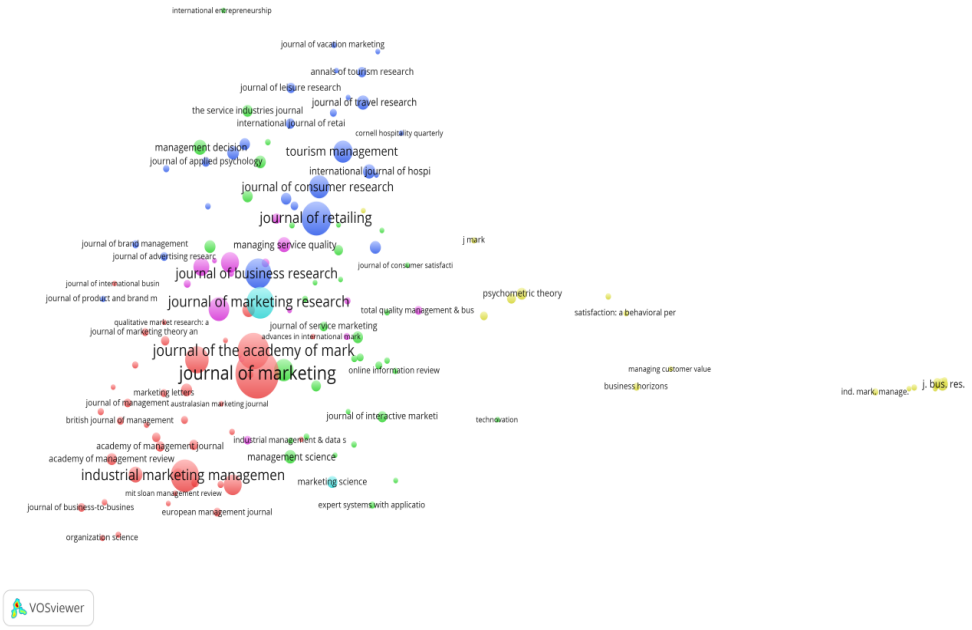


Figura 1. Visualización de la red de co-citación

Se identificaron 6 grupos. El grupo 1 (rojo) consta de 40 revistas que publican sobre temas de administración, mercadeo, estrategia y psicología aplicada; el grupo 2 (verde) consta de 35 revistas que publican sobre temas como servicio, mercado minorista, producción y aspectos virtuales como e-commerce, información on-line etc.; el grupo 3 (azul) con 27 revistas que publican sobre mercadeo, gestión, logística y turismo; el grupo 4 (amarillo) con 19 revistas que publican sobre marketing, valor del consumidor, satisfacción; el grupo 5 (violeta) consta de 15 revistas que publican sobre marketing de servicios, gerencia de la calidad del servicio y temas bancarios; y el grupo 6 (azul claro) con 2 revistas que publican sobre investigación de mercados y ciencia del marketing.
En la figura 2 se muestra la densidad de la red de co-citación. Las revistas con más disposición a este tipo de conceptos están en las zonas de color rojo, tal como Journal of Marketing y Journal of the Academy of Marketing y las que menos interés tienen son las que están en las zonas de color azul. El color verde denota interés medio en el tema.

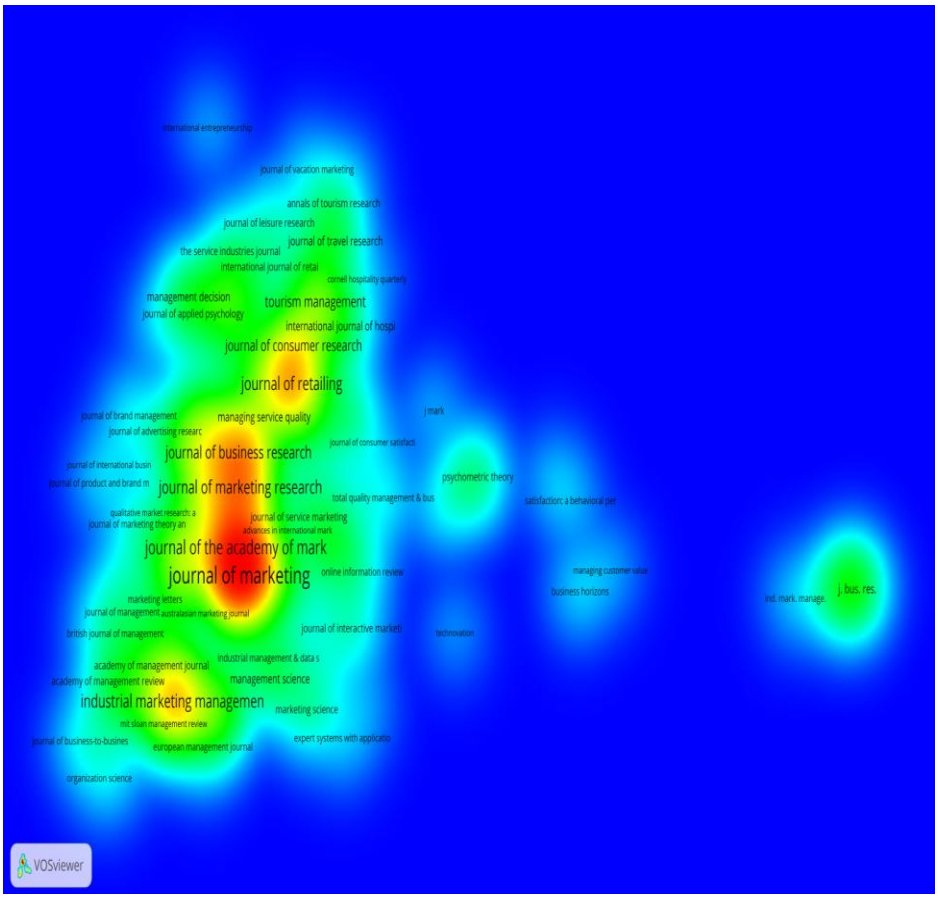


Figura 2. Densidad de la red de co-citación

En la figura 3 se muestra la red de co-ocurrencia de términos que se utilizan en el título y en el resumen de los documentos. La figura 4 muestra la densidad de la red de co-ocurrencia, en la que se observa los componentes del concepto principal más investigados de acuerdo con las líneas de investigación que se desprenden de los segmentos anteriormente analizados. Los componentes con mayor profundización son lealtad del consumidor, lealtad y satisfacción del consumidor. Un segundo grupo, son el mercado y CPV, y el tercer grupo, beneficios, perspectivas y teoría.

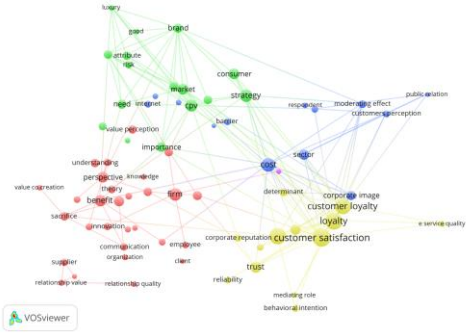


Figura 3. Visualización de la red de co-ocurrencias

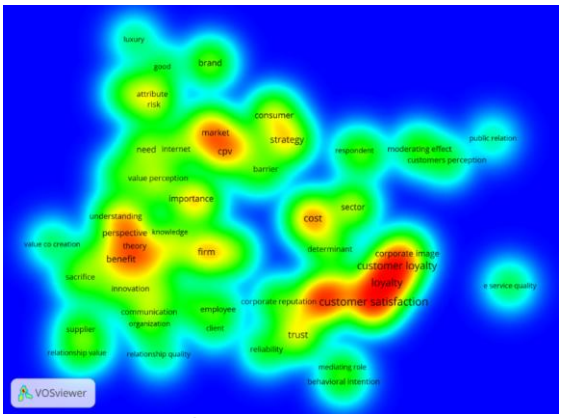


Figura 4. Visualización de densidad de la red de co-ocurrencia

En la figura 5, se observa la densidad de cada uno de los segmentos analizados y en la 6 la evolución histórica que han presentado los documentos, mostrándose en azul los componentes del concepto anteriores a 2008 y en rojo los más recientes. Se puede observar que las investigaciones están orientándose hacia conceptos como relaciones públicas y el rol mediador del valor percibido por el cliente

.

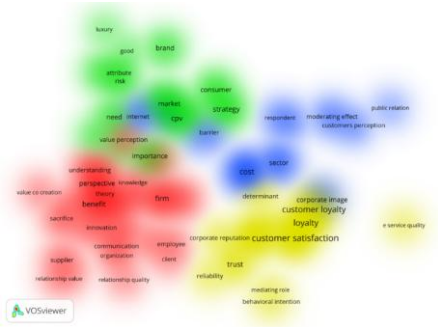


Figura 5. Densidad por segmentos de co-ocurrencia

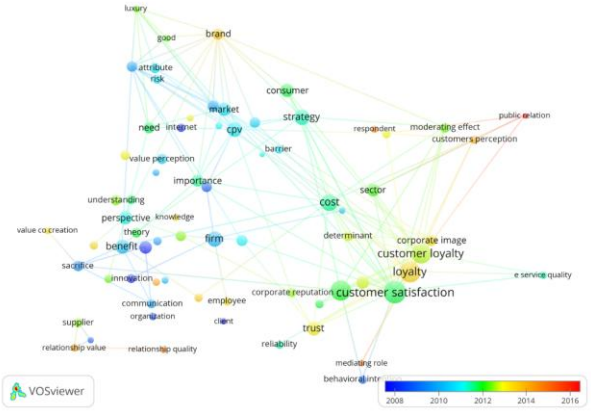


Figura 6. Red de visualización superpuesta por año de publicación

Aplicación 4.

Vázquez et al (2018) utilizaron la base de datos obtenidas por las encuestas de satisfacción realizadas a los turistas bajo los auspicios del Ministerio de la Industria Turística (MINTUR) de Cuba. La misma está compuesta por una tabla principal y cuatro tablas de referencia. La tabla principal contiene la información relevante de la opinión del cliente en sí (Id del cliente, las preferencias por la comida y la bebida, el trato del personal del hotel hacia los mismos entre otros indicadores). Las tablas de referencia describen la recopilación de determinados campos de la tabla de Opinión del Cliente. Como gestor de bases de datos se utilizó *PostgreSQL*, el cual permite realizar subconsultas SQL. Para efectuar la predicción se recurrió al algoritmo J48 utilizando como variable en estudio el indicador que se desea analizar . Las reglas usadas para hacer la clasificación fueron

Regla 1:Si la variedad del producto es Excelente, entonces la satisfacción general de los alimentos es Excelente (1)

Regla 2: Si la variedad del producto es Bueno, entonces la satisfacción general de los alimentos es Bueno (11/1)

Regla 3: Si la variedad del producto es Malo y la nacionalidad es cubano, entonces la satisfacción general de los alimentos es Regular (2)

Regla 4: Si la variedad del producto es Malo y la nacionalidad es americano, entonces la satisfacción general de los alimentos es Malo (2)

Regla 5: Si la variedad del producto es Pésimo, entonces la satisfacción general de los alimentos es Malo (2)

Regla 6: Si la variedad del producto es Regular, entonces la satisfacción general de los alimentos es Malo (1)

Para realizar el proceso de minado de datos se empleó la librería *WEKA* (B. R. R. *et al*., 2013), y se seleccionaron tres algoritmos. (*Clustering*, Clasificación y Asociación).

Al aplicar el *Clustering* se empleó el método particional *K-means*, con el objetivo de intentar minimizar la varianza total intra-grupo o la función de error cuadrático. En este caso, su uso tributó a la creación de una herramienta de verificación y testeo de la calidad de los servicios de alimentos y bebidas; en particular, del vector de características seleccionado.

Para emplear la Clasificación se recurrió al algoritmo J48, cuya función es crear una descripción eficiente de un conjunto de datos mediante la utilización de un árbol de decisión. El árbol resultante describe el conjunto de entrada a la perfección. Además, el árbol puede ser utilizado para predecir nuevos valores, asumiendo siempre que el conjunto de entrenamiento sobre el cual se trabaja es representativo respecto a la fuente de información original.

La Asociación se puso en práctica con la herramienta *Apriori*, cuya finalidad es descubrir reglas sustentadas probabilísticamente con un intervalo de confianza. Su aplicación posibilitó descubrir patrones significativos dentro de un conjunto de datos, que proporcionan una mejor comprensión del comportamiento general de los indicadores analizados.

El árbol de decisión obtenido fue

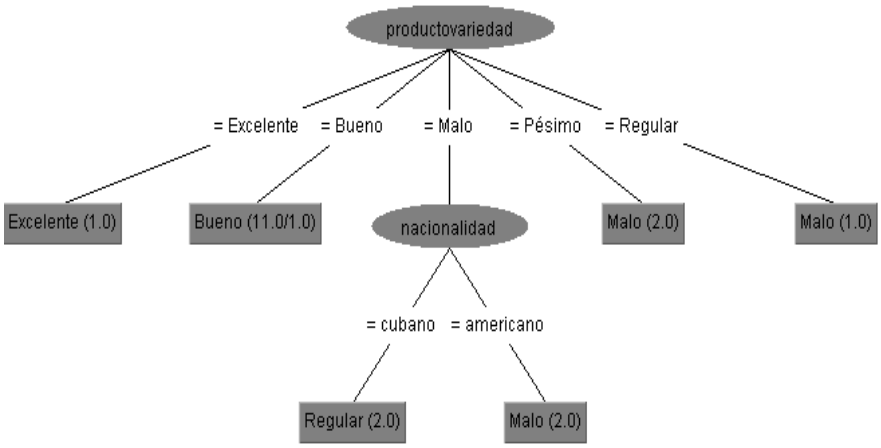


Figura 7. Árbol de Decisión de encuestas de satisfacción realizadas a los turistas.

La próxima figura ofrece una representación gráfica de la agrupación de las distintas opiniones de los clientes y la salida que efectuó la herramienta *WEKA* utilizando los mismos parámetros que se manejaron para la aplicación.

La integración de los algoritmos matemáticos, con su implementación computacional constituyen actualmente una herramienta para el análisis de la información en las instalaciones del MINTUR.

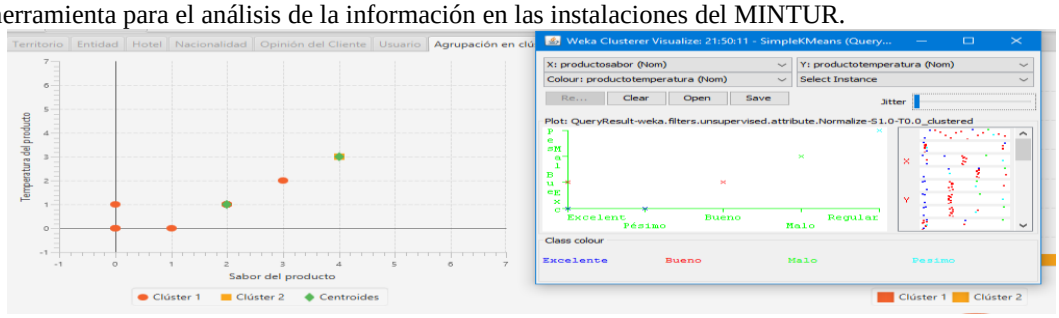


Figura 8. Agrupación de las distintas opiniones de los clientes y la salida de la herramienta

Acknowledgements: The autors would like to sincerely thank to the editors and two anonymous referees for their constructive comments.

RECEIVED: JULY, 2019.
REVISED: OCTBER, 2019.

REFERENCIAS

[1] ANDRADE-SÁNCHEZ, A. I., GALINDO-VILLARDÓN, MA. P.; SALAZAR-C., and C. M. HERNÁNDEZ-GONZÁLEZ, S. (2019): Caracterización de los Hábitos de Vida, Percepción y Apoyo Social en Estudiantes Preuniversitarios Mediante el Algoritmo Taid. **Inv. Operacional**, 40, 192-200

[2] BELLO-ORGAZ, G. and D CAMACHO (2014): Evolutionary clustering algorithm for community detection using graph-based information. Presented in **Evolutionary Computation (CEC), 2014 IEEE Congress** on, 930-937

[3] BELSEY, B. (2005): Cyberbullying: an emerging threat to the "always on" generation. Recuperado el 5, 5 2010; 5. **http://www.cyberbullying.ca/pdf/Cyberbullying_Article_by_Bill_Belsey.pdf.**

[4] CHETTY, R., J. FRIEDMAN, and J. ROCKOFF (2017): Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. **Am. Econ. Rev**.104, 2593–2632 .

[5] CHEN, W. (2010): How to tame big bad data. 2010. Recuperado el 15, 10, 2015. http://blog.magnitudesoftware.com/2010/08/25/tame-big-bad-data/.

[6] DASGUPTA, A. (2013): Big data: The future is in analytics. Geo Spatial World. https://www.geospatialworld.net/. Consulted Mayo, 5, 2017.

[7] EINAV, L. , and J. LEVIN (2014): The data revolution and economic analysis. En " **Innovation Policy and the Economy**", J. Lerner, S. Stern, Eds. Univ. of Chicago Press, Chicago, 14, 1–24.

[8] FREIRE F. C. (2015): Online digital social tools for professional self-promotion. A state of the art review. **Revista Latina de Comunicación Social**, 70, 288–99.

[9] GANDOMI A. and HAIDER M. (2015): Beyond the hype: big data concepts, methods, and analytics. **Int. J. Inf. Manag**. 35, 137–44.

[10] GEBOERS, E., F. GEIJSEL, W. ADMIRAAL and G. TEN DAM (2014): Typology of Student Citizenship. **European Journal of Education**, 1, 35-40.

[11] HASTIE, T., R. TIBSHIRANI and J. FRIEDMAN (2009): **The Elements of Statistical Learning: Data Mining, Inference and Prediction.** Springer, New York.

[12] KING, G. (2011): .Ensuring the data-rich future of the social sciences. **Science**, 331, 719–721 .

[13] KATZ J. E. (2008): **Handbook of mobile communication studies**. The MIT Press, London:

[14] KUM, . H. C. , S. AHALT, and T. M. CARSEY (2011): Dealing with data: Governments records. **Science**332, 1263 .

[15] MACHADO-DAZA, A. (2018): Valor Percibido por el Cliente: Cartografia Bibliometrica. **Investigación Operacional** , 39,549-561.

[16] MANDIBERG M. (2012): **The social media reader**. NYU Press, New York University.

[17] MARWICK A. E. (2015): **Status update: celebrity, publicity, and branding in the social media age**. Yale University Press, New Haven.

[18] MONASH, C. (2010): Three broad categories of data. Recuperado el 6, 7, 2016. http://www.dbms2.com/2010/01/17/three-broad-categories-of-data/. .

[19] SCHOLZ, T. M. (2017): **Big Data in Organizations and the Role of Human Resource Management A Complex Systems Theory-Based Conceptualization**. Peter Lang, New York.

[20] STOFFEL, K., and BELKONIENE, A. (1999): Parallel k/h-Means Clustering for large

[21] data sets, in: **P. AMESTOYet al. (Eds), Euro-Par'99, LNCS** 1685, 1451–1454. (Berlin, Heidelberg, Springer-Verlag).

[22] VAN DEWERFHORST, H.G., and MIJS, J. J. B. (2010): Achievement inequality and the institutional structure of educational systems: a comparative perspective. **Annual Review of Sociology**, 36, 407–428.

[23] VARIAN, H. (2013): Beyond big data. presented at the National Associate for Business Economics Annual Meeting, San Francisco, CA, 7 to 10 September 2013; http://people.ischool.berkeley.edu/~hal/Papers/2013/BeyondBigDataPaperFINAL.pdf.

[24] VÁZQUEZ ALFONSO, Y., M. YNFANTE MARTÍNEZ, A. DÍAZ VASALLO, L. and E. VELASTEGUÍ LÓPEZ (2018): Sistema informático de apoyo a la toma de decisiones en los servicios de restauración de la red hotelera en Cuba. **13th ICOR**, La Habana, Cuba.