Universidad de Costa Rica
Facultad de Educación
Instituto de Investigación en Educación

REVISTA ACTUALIDADES INVESTIGATIVAS EN EDUCACIÓN

http://revista.inie.ucr.ac.cr/ ISSN 1409-4703

Las evaluaciones internas del sistema nacional de investigadores de México a través de un análisis clúster

The internal evaluations of the national system of researchers of Mexico through a cluster analysis

Volumen 18, Número 1
Enero-Abril
pp. 1-32

Este número se publica el 1° de enero de 2018
DOI: https://doi.org/10.15517/aie.v18i1.31408

Gerardo Reyes Ruiz

Revista indizada en **REDALYC**, **SCIELO**

Revista distribuida en las bases de datos:

LATINDEX, **DOAJ**, **REDIB**, **IRESIE**, **CLASE**, **DIALNET**, **SHERPA/ROMEO**, **QUALIS-CAPES**, **MIAR**

Revista registrada en los directorios:

ULRICH'S, **REDIE**, **RINACE**, **OEI**, **MAESTROTECA**, **PREAL**, **CLACSO**

Las evaluaciones internas del sistema nacional de investigadores de México a través de un análisis clúster

The internal evaluations of the national system of researchers of Mexico through a cluster analysis

Gerardo Reyes Ruiz[1]

***Resumen**: El Sistema Nacional de Investigadores de México (SNI) evalúa, selecciona y reconoce, mediante un estímulo económico, el capital humano nacional que realiza investigación de calidad. Esta logística puede ser considerada como una selección de proyectos, la cual conlleva, obligatoriamente, a la elección de capital humano especializado. En este artículo se utiliza la técnica de análisis y agrupamiento de datos conocida como clustering (k Means) para profundizar sobre los criterios seguidos por el SNI en cuanto a dicha elección de investigadores. Una vez que se conoce el perfil productivo de cada nombramiento definido por el SNI, y a través de la distancia de Hamming, se realiza un análisis comparativo entre los datos estimados y reales asociados a cada nombramiento. Las estimaciones permitieron concluir que no se justifica la actual clasificación en cuatro agrupaciones (nombramientos), tal vez ello se deba a que los evaluadores del SNI utilizan información no recolectada en las variables reportadas por las solicitudes. Además, se demuestra la necesidad de mejorar la información estadística utilizada como base de datos para la evaluación; se señalan las diferencias en las clasificaciones estimadas para las siete áreas del conocimiento definidas por el SNI y se recomiendan algunos de los resultados para complementar las evaluaciones por pares, realizadas actualmente, siempre que se mejore la cantidad y calidad de la información disponible. Sin duda, ello debe de servir para hacer más eficiente la futura selección de proyectos de investigación y desarrollo concernientes a un programa de la política pública de investigación en México.*

***Palabras clave**: método de evaluación, estadísticas científicas; análisis comparativo, investigador*

***Abstract**: The National System of Researchers of Mexico (SNI) evaluates, selects, and recognized by an economic stimulus to national human capital that makes quality research. This logistics can be considered as a selection of projects, which leads, inevitably, to the choice of specialized human capital. This article uses the technique of analysis and clustering of data known as clustering (k Means) to deepen on the criteria followed by the NSR with regard to the choice of researchers. Once the productive profile of each appointment defined by SNI, and through the Hamming distance is known, is a comparison between the actual and estimated data associated with each appointment. Estimates allowed to conclude that it is not justified the current classification into four groups (appointments), perhaps this is due to that the evaluators of the SNI used information not collected on variables reported by requests. In addition, demonstrates the need for improved statistical information used as the database for the evaluation; the differences that exist in the ratings for the seven knowledge areas defined by the SNI and recommended some of the results to supplement assessments by peers today, provided that improvements are designated the quantity and quality of available information. Certainly, this should serve to streamline the future selection of projects of research and development concerning a programme of public policy research in Mexico.*

***Keywords**: evaluation methods, scientific statistics; comparative analysis, research workers*

[1] *Investigador en la Universidad Autónoma del Estado de México (UAEM), México.*

Dirección electrónica: greyesru@uaemex.mx

***Artículo recibido**: 2 de mayo, 2017*
***Enviado a corrección**: 31 de agosto, 2017*
***Aprobado**: 13 de noviembre, 2017*

1. Introducción[2]

Los modernos métodos automatizados de medición, recolección, recopilación y análisis de datos en todos los ámbitos de la ciencia, la industria y la economía proporcionan más y más datos con un aumento gradual en la complejidad de su estructura (Washio y Motoda, 2003). Esta creciente complejidad se justifica en gran medida por la necesidad de una rica y cada vez más precisa descripción de los fenómenos del mundo real y también debido al rápido progreso de la medición y el análisis de técnicas versátiles que facilitan la exploración de dichos fenómenos (Blum y Mitchell, 1998; Dietterich, Lathrop y Lozano-Perez, 1997; Gärtner, Flach, Kowalczyk y Smola, 2002; Goethals, Hoekx y Van den Bussche, 2005; Kailing, Kriegel, Pryakhin y Schubert, 2004). Por ello, y con el fin de gestionar el enorme volumen de datos tan complejos, se emplean sistemas de bases de datos (Kriegel et al, 2007). Con el arribo de la experimentación de alto rendimiento y tecnologías de conexión a internet cada vez más veloces, la generación y transmisión de grandes volúmenes de datos han visto enormes cambios de automatización en las últimas décadas. Como resultado, la ciencia, la industria e incluso los individuos tienen que afrontar el reto de hacer frente a enormes conjuntos de datos que en ciertas ocasiones son demasiado grandes para el análisis manual (Kriegel et al, 2007).

El *Data Mining* (DM) o Minería de Datos (MD), a menudo también denominada Descubrimiento del Conocimiento en Bases de Datos[3] (*Knowledge Discovery in Databases-KDD*), es una subdisciplina relativamente joven de la informática, con miras a la interpretación automática de grandes conjuntos de datos (Han y Kamber, 2006). Esta nueva rama de la ciencia considera varias técnicas de análisis como el aprendizaje de ordenadores, el reconocimiento de patrones, los sistemas de bases de datos, la inteligencia artificial y la estadística, por mencionar tan solo algunos, y entre sus múltiples objetivos se encuentra el análisis de grandes volúmenes de datos (Fayyad, Piatetsky-Shapiro y Smyth, 1996; Han y Kamber, 2006; Shian-Chang, En-Chi y Hsin-Hung, 2009; Tan, Steinbach y Kumar, 2006). Actualmente existen múltiples algoritmos de MD que son adaptados a diversos campos de aplicación para realizar diferentes tareas sobre el análisis de datos (Kittler, Hatef, Duin y

[2] Parte de este trabajo de investigación fue presentado en el *XVII Congreso Internacional de Contaduría, Administración e Informática* celebrado los días 3, 4 y 5 de octubre de 2012 en la Facultad de Contaduría y Administración, Ciudad Universitaria-UNAM, México.

[3] La definición clásica del descubrimiento de conocimiento en bases de datos es la que se describe en Fayyad, Piatetsky-Shapiro y Smyth (1996) como un proceso no trivial de identificación válida, novedosa, potencialmente útil, comprensible y, en definitiva, de patrones en los datos. Además, a la minería de datos la interpretan como un paso en el proceso de KDD, el cual consiste, grosso modo, en la aplicación de análisis de datos y algoritmos de descubrimiento.

Matas, 1998; Kriegel, Kröger, Pryakhin, y Schubert, 2004; Kriegel, Pryakhin y Schubert, 2005; Weidmann, Eibe y Bernhard, 2003; Wu *et al*, 2008). La MD suele abordar ciertos enfoques para algunos subtipos de datos, es el tema fijado para las cadenas de datos especializadas o listas de valores posibles (Yarowsky, 1995). Muchos enfoques de clasificación o agrupamiento necesitan tan solo de datos numéricos -algoritmo *K-means*-, mientras que otros lo hacen exclusivamente para datos categóricos -algoritmo *k-modes*- pero, a menudo, los distintos enfoques se combinan para obtener resultados más apropiados -algoritmo *k Prototypes, algoritmo Harmony K-means*. (Huang, 1998; Mahdavi y Abolhassani, 2009).

Se podría pensar que trabajar únicamente con cualquiera de estas dos categorías de datos limita las técnicas comunes de agrupamiento, y en consecuencia, se prohíbe la agrupación de datos del mundo real (Huang, 1998). No obstante, es de suma importancia mencionar que encontrar un modelo adecuado para representar el fenómeno de estudio, incluso en un corto periodo de tiempo, no es un asunto trivial; alcanzar con facilidad el uso, e inclusive, reducir la parametrización es un objetivo de mayor importancia, incluso si los datos de entrada no son muy complejos.

En la actualidad, la selección de proyectos es una de las principales estrategias para cualquier organización, principalmente porque en su evaluación se hacen cada vez más enfáticos los factores medioambientales y sociales que, aunados a otros factores (de mercado, técnicos y financieros), hacen que la viabilidad del proyecto se oriente principalmente hacia los productos o servicios de la vida cotidiana. Todos estos factores requieren del manejo de una gran cantidad de información y de hacer suposiciones inteligentes para realizar la mejor selección de proyectos (Prasanta, 2006). Por otra parte, y como bien se comenta en Kan y Zhou (2007), una selección de inversiones necesariamente implicará la toma de las decisiones acerca de qué proyectos deberá apoyar una organización dentro de sus márgenes de capital y de que estén lo más apegados a la contribución de su objetivo general: maximizar el valor actual neto de la empresa o la riqueza de los accionistas. En la práctica, este objetivo se basa en un determinado número de métodos y criterios de selección cuyo uso depende, en primera instancia, del entorno de decisión y, posteriormente, de las características directas para las inversiones consideradas.

En México, y más aún en el Sistema Nacional de Investigadores (SNI), se realiza año tras año precisamente la selección de capital humano especializado. Esta selección se hace tomando en cuenta, principalmente, la producción científica realizada por un investigador, al

menos durante sus últimos tres años (CONACYT, 2017). El SNI (inversor) finalmente decidirá, mediante la valoración de pares, cuáles solicitudes aceptará (portafolio) para con ello apoyarlas económicamente (inversión) durante cierto periodo de tiempo, el cual al concluir dará pauta a valorar la permanencia de dicho investigador, si este así lo desea, tomando en cuenta nuevamente su última producción científica (rendimiento). Desde esta perspectiva, la selección de un proyecto de investigación (en nuestro caso será la solicitud presentada por un investigador mexicano al SNI) bien puede verse como una inversión (asumiendo que dicho proyecto de investigación sea aprobado por el SNI), y la elección simultánea de varios de ellos como la integración de un portafolio de inversión (por supuesto, dicho portafolio será integrado por todas las solicitudes aprobadas por el SNI en un determinado año). De esta manera, y como bien se desprende de Reguia (2014), las organizaciones innovan para mantener su composición competitiva y, con ello, hacerse de ventajas competitivas cada vez más innovadoras.

En este sentido, la selección de proyectos conlleva obligatoriamente a la selección de recursos humanos. Es cierto que un buen proyecto no necesariamente implica un buen capital humano. Sin embargo, es más probable que un buen capital humano implique un buen proyecto, el cual se espera propicie un impacto considerable en el desarrollo y el crecimiento económico de una determinada región. En este contexto Khurram, Kirsten y Phanindra (2007) demostraron que aquellos países que han destinado considerables inversiones económicas para el desarrollo humano lograron una producción más eficiente y, como resultado, mantuvieron tasas de crecimiento más altas. En relación con la creación y transferencia de nueva ciencia y tecnología la inversión en estos investigadores mexicanos está más que justificada, ya que como mencionan Schultz (1961) y Krueger y Ruttan (1989) estas no serían posibles si un país no posee un nivel de capital humano intelectual adecuado para derivar todo el beneficio del que es capaz. En contraparte, el riesgo de la inversión podría ser que el investigador no renueve su solicitud y con ello se pierda dicha inversión tanto en conocimiento como en lo económico.

La logística que asume el SNI anualmente para la selección de proyectos y/o capital humano intelectual conlleva una gran inversión de tiempo y de recursos tanto humanos como económicos. No obstante, y al contar con un soporte o apoyo técnico para llevar a cabo las mencionadas evaluaciones, se podrían optimizar todos esos recursos y hacer más eficiente su logística. En esta, la selección de una solicitud por parte del SNI indudablemente dependerá del factor humano. Por ello, se justifica la existencia de una Comisión Evaluadora

en cada una de las siete áreas del conocimiento definidas por dicho sistema de investigación. Sin embargo, ¿cómo ha sido esta selección de solicitudes? Es decir, ¿existe en realidad una correspondencia entre la información presentada al SNI por cada investigador aprobado y el nombramiento que le otorga dicho sistema de investigación? Y ¿los resultados (*outputs* de investigación) justifican las resoluciones adoptadas por dicho sistema de investigación? El presente trabajo pretende avanzar en este conocimiento con base en una técnica de análisis de datos conocida como *clustering*.

Tras esta introducción, el presente artículo se divide en cinco apartados: en el primero se comenta el objetivo de este estudio; en el segundo se describe la metodología así como los datos utilizados. En el tercero se presenta brevemente el SNI; en el cuarto, los resultados obtenidos y, por último, se muestran unos comentarios a modo de conclusiones.

2. Objetivo

El objetivo que persigue este artículo es mostrar que una técnica de análisis y agrupamiento de datos sirve como apoyo y soporte técnico para hacer más eficiente la selección de recursos humanos especializados que integran un programa concerniente a la política pública de investigación en México.

Al conocer la correspondencia que existe entre el nombramiento asignado a los investigadores aprobados por el SNI y su producción científica, mediante una técnica de agrupamiento de datos (*k Means*), se pueden detectar las características predominantes del SNI. Es decir, se puede apreciar el potencial intelectual y productivo de los investigadores mexicanos seleccionados por las siete Comisiones Evaluadoras definidas por este sistema. Este análisis definitivamente permitirá valorar si la asignación realizada por el SNI se basa en la producción científica reportada por cada investigador mexicano aceptado durante el periodo de 1996 a 2003.

Además, a través de este trabajo de investigación se manifiestan implícitamente, aunque con una perspectiva más de mediano plazo, las bases para automatizar la logística de la convocatoria emitida anualmente por el SNI. Es decir, se cimientan las bases para justificar el nombramiento asignado a cada investigador en relación con su producción científica reportada al SNI desde el punto de vista cuantitativo. Todo ello, con miras a servir de apoyo a los evaluadores del SNI y hacer sus valoraciones más representativas.

3. Metodología

El proceso de *clustering* consiste en dividir los datos en grupos de objetos similares (Bao, Han y Wu, 2006). Entonces esta técnica se puede usar para investigar la cercanía entre objetos y obtener la validación de una clasificación. En los métodos tradicionales de *cluster*, la función objetivo está basada en algoritmos de agrupamiento. Dicha función se hizo más popular al convertirse en un problema de optimización (Fisher, 1936). Es decir, el análisis de *clusters* es un problema focalizado en dividir un conjunto de datos $\{x_i\}_1^n$, de algún espacio ***X***, en una colección de grupos disjuntos pero similares entre ellos (MacQueen, 1967). En este contexto, el algoritmo *k means* surge como un método para la clasificación, y actualmente es considerado como un algoritmo exclusivo de agrupamiento no jerárquico; si un específico conjunto de datos pertenece a un grupo definido entonces no puede pertenecer a otro grupo simultáneamente. No obstante, uno de los principales problemas de este método es seleccionar el mejor valor de k, es decir, el número de clases o grupos. Por su parte, Kuo, Ho y Hu (2002) señalan que para estos métodos no jerárquicos se puede tener mayor precisión si el punto de partida y el número de las agrupaciones son preestablecidos. Es decir, *k means* es un algoritmo de aprendizaje no supervisado que resuelve eficientemente el problema de agrupamiento. Por tanto, la idea es definir los centroides k, uno para cada *cluster*. En otras palabras, estos centroides cambian su ubicación paso a paso (iteraciones) hasta que no se realicen más cambios, es entonces cuando se constituyen dichos centroides. En este sentido, y siguiendo el trabajo de Soto, Flores, y Vigo (2004), el algoritmo denominado *k means* proporciona k *clusters* $\{k_j\}_1^m$ cuando se minimiza la siguiente función objetivo:

$$J = \sum_{j=1}^{m} \sum_{i=1}^{n} \left\| x_i^{(j)} - k_j \right\|^2$$

Donde $\|\cdot\|$ es una distancia, previamente seleccionada, entre un conjunto de puntos $x_i^{(j)}$ y el centroide k_j del correspondiente *cluster*. Toda vez que el número de iteraciones ha concluido, un elemento pertenece tan solo a un *cluster* y no a varios simultáneamente[4]. En

[4] Una extensión del algoritmo de *k means* es, precisamente, el algoritmo de Fuzzy k Means (FKM). En este último un elemento sí puede pertenecer a varios grupos simultáneamente (véase Dunn, 1974; Bezdek, 1981; Dae-Won, Kwang y Doheon, 2004; Campello, Hruschka y Alves, 2009).

este trabajo se utiliza la herramienta para el análisis de datos conocida como *k Means* (Anderberg, 1973; Bock, 2008; MacQueen, 1967). Se considera este algoritmo de datos porque permite detectar tanto el nivel de asociación como la importancia de las variables involucradas. Además, este algoritmo está considerado entre los mejores diez algoritmos para la clasificación de datos (Wu et al, 2008). El insumo para esta técnica de análisis de datos es, en gran medida, la producción científica[5] reportada al SNI por cada investigador que solicitó el ingreso/permanencia a dicho sistema de investigación mexicano durante el periodo 1996-2003. Por otra parte, para detectar el total de artículos por investigador en el ISI[6], y reportados al SNI de 1996 a 2003, se utilizaron las bases de datos denominadas *Science Citation Index* (SCI) y *Social Science Citation Index* (SSCI), ambas ubicadas en el apartado ISI *Web of Knowledge*. Es decir, se hace uso de la información presentada por cada investigador al SNI para ser aceptado en dicho sistema, así como de la información del *Institute for Scientific Information* (ISI), la cual hace referencia a las publicaciones realizadas por al menos un investigador mexicano. Estas tres fuentes de información son consideradas con una periodicidad anual y para el periodo comprendido por los años de 1996 a 2003. A pesar de que la información del SNI estuvo acotada por el año 2003, y debido a que las estimaciones no involucran una variable cuantitativa temporal, es decir, que haga énfasis al tiempo o periodo alguno, los resultados de las estimaciones no se limitan a un periodo de estudio determinado. En consecuencia, y para los fines de este capítulo, se puede suponer que dicho periodo de estudio hace referencia a los últimos ocho años del SNI[7].

Con el algoritmo *k means* se pretende, sumado al factor humano, obtener unos dictámenes más robustos y eficientes por parte del SNI. Este análisis tiene sentido, ya que gran parte de las variables utilizadas por esta técnica de agrupamiento y análisis de datos son cuantitativas (Huang, 1998). Es decir, a través de esta técnica de agrupamiento de datos, se detectan las características predominantes de los investigadores mexicanos

[5] Por ejemplo, en la información integrada por el SNI se contempla el número de citas recibidas a los trabajos, al menos de su último nombramiento, realizados por cada investigador aprobado en dicho sistema de investigación. Es claro que el número de citas resulta ser uno de los instrumentos más habituales para valorar la calidad investigadora, aunque también está sujeto a problemas como acuerdos entre autores para realizar citaciones cruzadas o bien, diferencias entre áreas respecto a la práctica de proceder a citaciones, entre otras.

[6] Se reconocen indicadores más complejos como el factor H o el índice de Bauwens (1998). No obstante, estos indicativos servirían tan solo para "calificar" a los investigadores mexicanos que ya cuentan con una considerable trayectoria de publicar. Es decir, al involucrar indicadores compuestos, la evaluación del SNI dejaría a los investigadores jóvenes (Candidato) y que apenas cuentan, en la mayoría de los casos con cierta experiencia para publicar, aún más en desventaja respecto a los criterios de evaluación. Ello debido a que un investigador joven, al obtener un parámetro bajo, sesgaría la objetividad del evaluador y se desvirtuaría la calidad de su investigación.

[7] Se sabe que el periodo de estudio es limitado. Sin embargo, este no restringe los alcances del presente estudio, ya que hasta el día de hoy el SNI integra la misma información para emitir sus dictámenes. Por lo que bien puede suponerse un periodo de estudio más actual.

aceptados por el SNI, de 1996 a 2003. Este análisis muestra, por una parte, la correspondencia que existe entre las evaluaciones internas realizadas en el SNI y los diferentes perfiles de los investigadores aprobados por este sistema de investigación. Por otra parte, dicho análisis sirve para detectar el potencial de dichos investigadores que formaron parte del SNI durante algún periodo de tiempo. Toda vez que se estimaron los respectivos *clusters*, la información obtenida para los promedios reales mediante el algoritmo de *k means* y los promedios estimados de una solicitud aprobada por el SNI, de 1996 a 2003, permitió llevar a cabo un comparativo mediante la distancia de Hamming (Hamming, 1950). Se utilizó la distancia de Hamming porque los reactivos involucrados en el análisis bien pueden ser considerados como atributos de un perfil deseado. Con ello, se muestran finalmente los potenciales nombramientos que definieron, a su vez, a cada una de las áreas definidas por el SNI durante el periodo de 1996 a 2003.

Como consecuencia, y mediante estas técnicas de análisis y agrupamiento de datos, en la medida de que se disponga de mayor pero sobre todo mejor información por parte del SNI, entonces se obtendrán evaluaciones más robustas, las cuales conllevarán a tener un panorama más claro del potencial de los investigadores mexicanos que integran al SNI. Desde el punto de vista de la automatización, y en términos de una estimación (Greene, 2008), lo que se pretende es, para cada dato pronosticado, llevar el nivel subjetivo humano a niveles poco significativos. Es decir, obtener las estimaciones más robustas que finalmente serán ratificadas por los evaluadores que conforman las Comisiones Evaluadoras del SNI. Por lo tanto, en la medida que aumente la calidad de la información recabada por el SNI mejor será el soporte técnico proporcionado a dichos evaluadores para la asignación de un nombramiento.

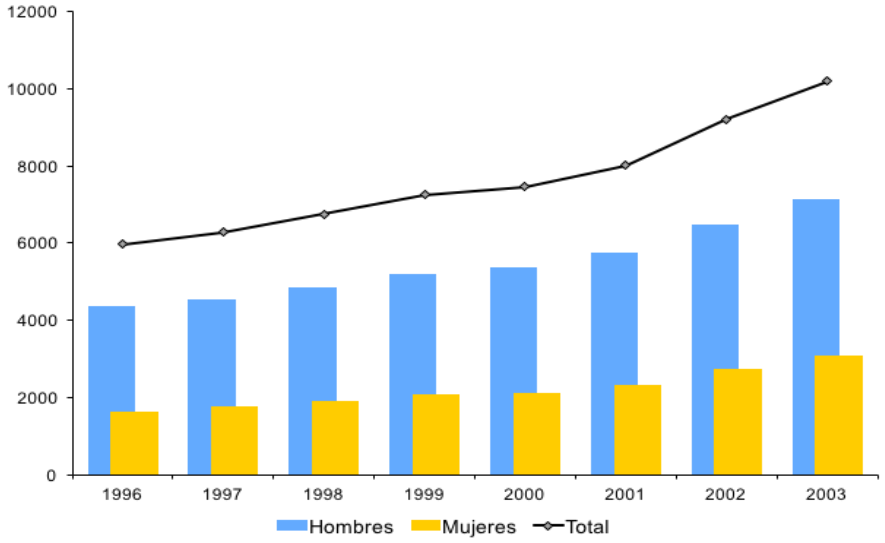
Como ya se mencionó con anterioridad, resaltar que tan solo se contó con la producción científica de aquellos investigadores mexicanos aprobados por el SNI durante el periodo de 1996 a 2003. Además, se desconoció la producción científica de las solicitudes no aprobadas por dicho sistema de investigación. Consecuentemente, se orientó el análisis a conocer si el SIN, en realidad, ha llevado a cabo una buena selección de solicitudes. No obstante, la experiencia de la información integrada para los investigadores aprobados por el SNI permitió responder el interrogante de cómo se han llevado a cabo las evaluaciones en dicho sistema. No se ha podido considerar un período temporal más amplio debido a que el SNI solo proporcionó la información hasta el año 2003. Este hecho, sin duda, es una limitación pero se entiende que no anula el interés del artículo, puesto que permite ver

igualmente las potencialidades de la técnica aplicada y, además, permite valorar la racionalidad de los criterios de evaluación[8] aplicados por el SNI, similares a los utilizados, incluso, en un periodo más reciente.

4. El Sistema Nacional de Investigadores (SNI)

Es un subprograma del Programa de Fomento a la Investigación Científica, establecido por el Gobierno Federal, cuya conducción y operación, así como el establecimiento de sus objetivos y funciones, organización y reglamentación interna, están a cargo del Consejo Nacional de Ciencia y Tecnología (CONACyT). El SNI de México, tiene por objeto promover y fortalecer, a través de una evaluación, la calidad de la investigación científica y tecnológica y la innovación que se produce en el país (CONACyT, 2017).

Figura 1. Total de investigadores vigentes en el SNI por año y por género, 1996-2003.

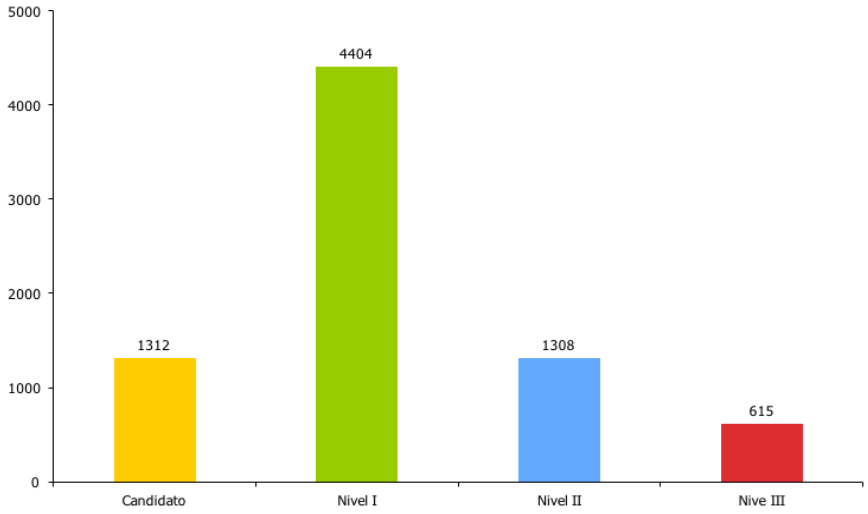


Fuente: Elaboración propia con información del SNI, 2012.

[8] Actualmente el SNI define siete áreas del conocimiento: I) Físico Matemáticas y Ciencias de la Tierra; II) Biología y Química; III) Medicina y Ciencias de la Salud; IV) Humanidades y Ciencias de la Conducta; V) Sociales; VI) Biotecnología y ciencias agropecuarias y; VI) Ingeniería y Tecnología. Todas ellas valoran la producción científica de un solicitante, la cual comprende: artículos, libros, libros traducidos, libros editados, capítulos de libros, tesis dirigidas, citas realizadas a sus trabajos de investigación, patentes, desarrollos tecnológicos, distinciones recibidas, grupos de investigación, estancias posdoctorales, reseñas, estancias de investigación y cursos académicos impartidos. Estos criterios asumen que un artículo tiene la misma ponderación que una patente o cualquier otro criterio evaluado, al menos desde el punto de vista cuantitativo. Por lo que no debe confundirse y asumir que una reseña, por ejemplo, es más importante que una cita realizada, ya que al no existir una ponderación en los Criterios Internos de Evaluación del SNI, definitivamente dicha valoración quedará sujeta al criterio subjetivo del evaluador.

El ingreso al SNI es voluntario y gratuito para el solicitante. Una vez analizada su solicitud por la correspondiente Comisión Evaluadora, en alguna de las siete áreas del conocimiento definidas por este sistema de investigación mexicano, al solicitante se le comunica su valoración positiva o negativa y, en el primer caso, un nombramiento (véase Figura 2) como miembro del SNI con la adscripción a un nivel (Candidato a Investigador, Investigador Nacional Nivel I, Investigador Nacional Nivel II o Investigador Nacional Nivel III), que además tiene asociada una compensación económica variable. Actualmente, pertenecer a dicho sistema de investigación supone un reconocimiento a la calidad y prestigio académico del investigador, resultado de una producción científica de considerable trascendencia a nivel nacional y, en algunos casos, en el ámbito internacional.

Figura 2. Promedio[9] de investigadores vigentes por categoría del SNI, 1996-2003.



Fuente: Elaboración propia con información del SNI, 2012.

La distribución por género (véase Figura 1), para el periodo de 1996 a 2003, de los investigadores vigentes mostró una tendencia creciente tanto para los hombres (5446 en promedio) como para las mujeres (2193 en promedio) en el SNI. Sin embargo, la razón promedio entre hombre-mujer para este periodo fue de 2.5 a 1. De esta manera, el promedio de los investigadores vigentes por cada área del conocimiento definida por el SNI y género se muestra en la Tabla 1.

[9] Se utiliza el promedio de investigadores debido a que una persona puede estar contabilizada varias veces en el periodo de estudio.

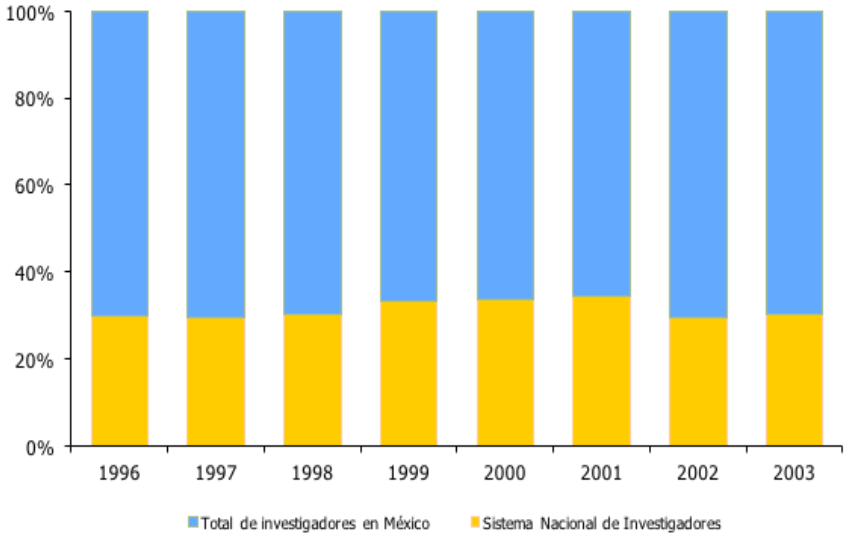
Tabla 1. Promedio de investigadores vigentes por área del conocimiento del SNI, 1996-2003.

Área	Promedio 1996-2003			% Total
	Hombres	Mujeres	Total	
Area I: Físico-Matemáticas y Ciencias de la Tierra	1,304	242	1,546	20.2
Area II: Biología y Química	994	544	1,538	20.1
Area III: Medicina y Ciencias de la Salud	494	292	786	10.3
Area IV: Humanidades y Ciencias de la Conducta	696	623	1,318	17.3
Area V: Sociales	589	258	848	11.1
Area VI: Biotecnología y Ciencias Agropecuarias	575	131	706	9.2
Area VII: Ingeniería y Tecnología	794	103	898	11.7
Promedio de todas las áreas	5,446	2,193	7,639	100.0

Fuente: Elaboración propia con información del SNI, 2012.

En relación con el total de solicitudes recibidas por este sistema de investigación, se puede mencionar que el índice de solicitudes aprobadas pasó del 72.2% en 1996 al 77.6% en el año 2003. Estos resultados implican que el índice de solicitudes no aprobadas pasó del 27.8% al 22.4% de 1996 a 2003 respectivamente. Sin duda, y durante el periodo de estudio, este sistema de investigación favoreció el ingreso/permanencia de los investigadores mexicanos. Otro resultado de suma importancia es la participación de los investigadores que pertenecen al SNI con respecto a los investigadores mexicanos que no tienen dicho registro, definidos estos últimos como de tiempo completo (véase Figura 3). Es interesante apreciar que durante cada año, de 1996 a 2003, la participación del SNI, en relación con los investigadores mexicanos de tiempo completo, fue de una tercera parte.

Figura 3. Participación del SNI respecto al total de investigadores mexicanos en equivalente de tiempo completo, 1996-2003.



Fuente: Elaboración propia con información del SNI, 2012.
OECD: Main Science and Technology Indicators, 2005-2.
RICYT: Principales Indicadores de Ciencia y Tecnología, 2004.

5. Resultados y su análisis

En este apartado se asume lo siguiente: 1) los dictámenes emitidos por cada una de las siete Comisiones Evaluadoras del SNI son dados, es decir, congruentes o no, son irrepetibles y, 2) las resoluciones, correctas o no, son perfectibles. Es por ello que existen Subcomisiones Evaluadoras, que tienen por objetivo evaluar las solicitudes de inconformidad, y pueden emitir un dictamen diferente al de la Comisión Evaluadora correspondiente. Asimismo, la información involucrada para este análisis *clustering* fue la mostrada en la Tabla 2. El número total de *clusters* definidos para este algoritmo fueron cuatro (C1, C2, C3 y C4)[10], a semejanza y/o similitud de los nombramientos del SNI. Con esta relación se pretende establecer una correspondencia entre los nombramientos otorgados por dicho sistema de investigación y los grupos pronosticados mediante la técnica de agrupamiento *k means*. La similitud entre cada grupo pronosticado y el correspondiente nombramiento del SNI estará en función de la producción científica reportada para dicho sistema por cada uno de estos cuatro nombramientos internos.

[10] El cálculo de los clústers C1, C2, C3 y C4 se llevó a cabo con el paquete estadístico SPSS en su versión 17.

Tabla 2. Descripción de las variables de agrupamiento para realizar el *clustering*.

1	SEXO (sexo del solicitante). Hombre=1 y Mujer=0
2	EDAD (edad del investigador al momento de presentar su solicitud al SNI).
3	GRADO (último grado académico reportado al SNI por el solicitante). Doctorado=3; Maestría=2; Especialidad=1 y Licenciatura=0
4	NA (nivel anterior del investigador, únicamente para reingresos vigentes). Nivel III=3; Nivel II=2; Nivel I=1 y Candidato=0
5	NIVEL (nivel asignado por el SNI al momento de ingreso). Nivel III=3; Nivel II=2; Nivel I=1 y Candidato=0
6	AREA (área de conocimiento definida por el SNI). Físico-Matemáticas y ciencias de la tierra=1; Biología y Química=2; Medicina y ciencias de la salud=3; Humanidades y Ciencias de la conducta=4; Sociales=5; Biotecnología y ciencias Agropecuarias=6; Ingeniería y Tecnología=7.
7	DISC_SNI (disciplina asociada al área de conocimiento definida por el SNI). Clave del SNI asignada a cada disciplina del conocimiento.
8	SIT (situación del solicitante; Reingreso Vigente, Nuevo Ingreso, Reingreso No Vigente). Reingreso Vigente=2; Nuevo Ingreso=1 y Reingreso No Vigente=0
9	INST (institución de adscripción en México del solicitante). Clave del SNI asignada a cada institución de adscripción.
10	UBIC_MEX (ubicación geográfica en México de la institución de adscripción del solicitante). Aguascalientes=1;…..Distrito Federal=9;……:Zacatecas=31 y NE=0
11	ART_SNI (artículos reportados al SNI por el solicitante ya sean estos publicados, aceptados ó enviados). Número de artículos=0,1,2,…..n
12	CAP_LIB (capítulos de libros reportados por el solicitante). Número de capítulos de libros=0,1,2,…..n
13	CITAS (citas recibidas a los trabajos del solicitante). Total de citas=0,1,2,…..n
14	DES_TEC (desarrollos tecnológicos realizados por el solicitante). Número de desarrollos tecnológicos=0,1,2,…..n
15	DISTIN (distinciones recibidas por el solicitante). Número de distinciones=0,1,2,…..n
16	DOCENCIA (total de cursos académicos impartidos por el solicitante). Número de cursos impartidos=0,1,2,…..n
17	EST_INV (estancias de investigación realizadas por el solicitante). Número de estancias de investigación=0,1,2,…..n
18	POSDOC (posdoctorados realizados por el solicitante). Número de posdoctorados=0,1,2,…..n
19	GRU_INV (grupos de investigación a los que pertenece el solicitante). Número de grupos de investigación=0,1,2,…..n
20	INVITA (invitaciones a congresos nacionales o internacionales). Número de congresos=0,1,2,…..n
21	LIBROS (libros reportados por el solicitante). Número de libros=0,1,2,…..n
22	LIBEDIT (libros editados reportados por el solicitante). Número de libros editados=0,1,2,…..n
23	LIB_TRAD (libros traducidos reportados por el solicitante). Numero de libros traducidos=0,1,2,…..n
24	MEMORIAS (memorias de congresos reportados por el solicitante). Número de memorias=0,1,2,…..n
25	PATENTES (patentes registradas reportadas por el solicitante). Número de patentes=0,1,2,…..n
26	RESENAS (reseñas reportadas por el solicitante). Número de reseñas=0,1,2,…..n
27	TESIS (tesis dirigidas reportadas por el solicitante). Número de tesis=0,1,2,…..n
28	PUBIC_ISI (total de publicaciones del investigador con registro SNI en el ISI). Número de publicaciones en el ISI=0,1,2,…..n

Fuente: Elaboración propia con información del SNI y el ISI, 2012.

Los promedios reales[11] para una solicitud aprobada por el SNI, de 1996 a 2003, por nivel y concepto se muestran en la Tabla 3, mientras que los promedios estimados se presentan en la Tabla 4. Como resultado de aplicar el algoritmo *k means*, el 87.1% del total de observaciones fueron clasificadas en el conglomerado C2 de la Tabla 4. Este conglomerado captó el mayor número de solicitudes aprobadas durante dicho periodo (véase Figura 4), ya que para los investigadores con un nombramiento de Candidato, el 91.7% fue clasificado en el mencionado conglomerado C2; para los investigadores Nivel I esta clasificación fue del 90.3%; para los investigadores Nivel II fue del 77.2% y para los investigadores Nivel III fue del 66.7%. Mencionar que el 3.5% del total de solicitudes aprobadas por el SNI no fueron clasificadas en ningún conglomerado.

Tabla 3. Promedios reales para una solicitud aprobada en el SNI, por concepto y nivel 1996-2003.

Concepto	Candidato	Investigador Nacional		
		Nivel I	Nivel II	Nivel III
Artículos	3.7	8.8	15.6	23.0
Publicaciones en el ISI	0.3	0.7	1.4	2.5
Capítulos de libros	0.6	1.8	3.2	5.3
Citas realizadas	2.1	14.1	46.0	84.6
Desarrollos tecnológicos	0.2	0.5	0.5	0.9
Distinciones recibidas	1.9	2.9	4.6	6.1
Cursos académicos impartidos	0.2	0.2	0.1	0.7
Estancias de investigación	0.2	0.1	0.1	0.1
Estancias posdoctorales	0.4	0.6	1.0	1.1
Grupos de investigación	0.2	0.6	1.0	1.2
Invitaciones a congresos	5.7	9.9	13.2	20.5
Libros	0.3	0.8	1.3	1.8
Libros editados	0.1	0.2	0.5	0.9
Libros traducidos	0.0	0.1	0.1	0.2
Memorias en congresos	1.6	3.0	4.2	5.4
Patentes	0.0	0.1	0.2	0.3
Reseñas	0.1	0.3	0.5	0.7
Tesis dirigidas	1.9	5.2	8.1	9.8

Fuente: Elaboración propia con información histórica del SNI, 2012.

[11] El promedio real hace referencia a la media aritmética obtenida en cada concepto evaluado (artículos, publicaciones en el ISI, capítulos de libros, etc.) en cada una de las Comisiones Evaluadoras del SNI.

Tabla 4. Promedios estimados[12] para una solicitud aprobada en el SNI por concepto, 1996-2003.

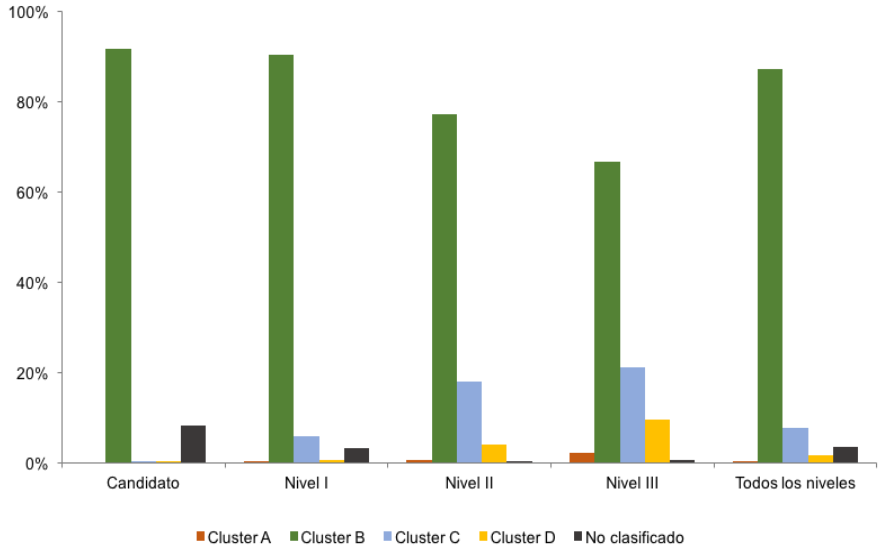
Concepto	Conglomerado			
	C1	C2	C3	C4
Artículos	63.6	8.0	24.3	40.2
Publicaciones en el ISI	4.4	0.7	2.2	3.3
Capítulos de libros	6.7	1.8	3.6	5.8
Citas realizadas	825.4	6.2	109.1	344.5
Desarrollos tecnológicos	0.1	0.5	0.7	0.5
Distinciones recibidas	11.2	2.9	6.0	8.5
Cursos académicos impartidos	0.3	0.2	0.3	0.4
Estancias de investigación	0.2	0.1	0.1	0.1
Estancias posdoctorales	1.5	0.6	1.3	1.8
Grupos de investigación	1.0	0.6	1.0	1.2
Invitaciones a congresos	40.1	9.2	19.0	30.0
Libros	1.9	0.8	1.1	1.3
Libros editados	0.9	0.3	0.5	0.6
Libros traducidos	1.2	0.1	0.1	0.1
Memorias en congresos	7.0	2.7	5.9	9.0
Patentes	0.1	0.1	0.3	0.3
Reseñas	0.2	0.3	0.3	0.9
Tesis dirigidas	15.1	4.9	10.0	11.7

Fuente: Elaboración propia con información histórica del SNI, 2012.

Por su parte, de la figura 4 se desprenden tres importantes comentarios: 1) los nombramientos asignados por el SNI, de 1996 a 2003, convergen hacia un solo conglomerado, y en particular al *cluster* C2 de este análisis; 2) los niveles del SNI denominados Candidato y Nivel I presentan una ubicación más cercana hacia el mencionado *cluster* C2 y, 3) los niveles superiores del SNI (Nivel II y Nivel III) lo hacen también, pero con una ubicación más lejana.

[12] Para los conceptos artículos, citas realizadas e invitaciones a congresos se obtuvieron estimaciones muy variables. Este resultado, sin duda, implica que existen investigadores del SNI que necesitan más, pero sobre todo mejor, información para ser definidos. Es decir, existen Candidatos a Investigador con producción científica que bien pudiesen ser clasificados en niveles superiores del SNI o bien, existen investigadores en los niveles superiores del SNI que no reportaron suficiente producción para justificar su nombramiento.

Figura 4. Distribución de los nombramientos del SNI en los conglomerados obtenidos mediante el algoritmo *k means*, 1996-2003.



Fuente: Elaboración propia con información histórica del SNI, 2012.

En este contexto, y a partir de la Figura 4, no se aprecia otro agrupamiento, además del *cluster* C2, que muestre una participación significativa. De estos últimos resultados se puede deducir que la producción científica reportada al SNI, por todos los investigadores mexicanos aprobados, no justifica la existencia de cuatro niveles, sino que solo se justifica uno o como máximo dos (en primera instancia tan solo un conglomerado definido por los Candidatos o bien, como segunda instancia, un primer conglomerado que agrupe a los investigadores con los nombramientos Candidato-Nivel I y otro agrupamiento que contenga a los investigadores con los nombramientos Nivel II-Nivel III).

La información obtenida para los promedios reales, mediante el algoritmo de *k means* (Tabla 3) y los promedios estimados (Tabla 4) de una solicitud aprobada por el SNI, de 1996 a 2003, permite realizar un comparativo mediante la distancia de Hamming[13]. Esta distancia se define de la siguiente manera:

$$\varphi[\delta_{\mu(x)},\delta_{\mu(y)}] = \frac{1}{n}\sum_{k=0}^{n} |x_k - y_k|$$

[13] Se hace uso de la distancia Hamming ya que los ítems considerados bien pueden ser considerados como atributos de un perfil deseado.

Donde:

$\mu(x)$ es el vector de los promedios reales en cada nivel del SNI

$\mu(y)$ es el vector de los promedios estimados en cada nivel del SNI

$\delta_{\mu(x)}$ define todos los atributos del conjunto $\mu(x)$

$\delta_{\mu(y)}$ define todos los atributos del conjunto $\mu(y)$

x_k es el k-ésimo atributo del conjunto $\mu(x)$

y_k es el k-ésimo atributo del conjunto $\mu(y)$

n es el total de atributos

Se utiliza esta distancia para detectar la similitud que existe entre los vectores reales de la producción asociada a cada nivel del SNI y los vectores estimados mediante el algoritmo de *k means* (véase Tabla 5). Es decir, si no existiera el criterio subjetivo en el proceso de evaluación del SNI, entonces los nombramientos definitivamente tendrían una distribución muy diferente a la observada en el periodo de estudio.

Tabla 5. Matriz de distancias Hamming para los promedios reales y los promedios estimados mediante el algoritmo *k means*.

Nivel / Cluster	C1	C2	C3	C4
Candidato	53.4	1.1	9.3	24.5
Nivel I	51.8	0.6	7.6	22.8
Nivel II	49.0	3.4	4.7	19.9
Nivel III	45.6	7.0	1.8	16.6

Fuente: Elaboración propia con información histórica del SNI, 2012.

Los resultados de la Tabla 5 permiten deducir lo siguiente: a) el total de solicitudes aprobadas por el SNI, durante el periodo de estudio, presentó un claro agrupamiento hacia dos conglomerados (C2 y C3); b) tres de los cuatro nombramientos definidos en el SNI (Candidato, Nivel I y II) mostraron una clara convergencia hacia un solo conglomerado, a decir, el *cluster* C2; c) los investigadores Nivel III convergen hacia el clúster C3 y; d) el *clúster* C2 presentó bastante similitud (menor distancia de Hamming) con el vector real referente a un investigador Nivel I. Este último resultado implica que casi un 90% de los investigadores mexicanos aprobados por el SNI, de 1996 a 2003, tuvieron el perfil productivo de un investigador Nivel I. Para validar la coherencia de los resultados obtenidos en la Tabla 5, se calculó la matriz de distancias Hamming para los promedios reales de los criterios evaluados a los investigadores aprobados por el SNI de 1996 a 2003 (véase Tabla 6).

Destacar que de la Tabla 6 se desprende un resultado de suma importancia: en la realidad del SNI el perfil productivo de un Candidato es muy similar al perfil productivo de un Nivel I.

Tabla 6. Matriz de distancias Hamming para los promedios reales de los criterios evaluados en el SNI, por nivel 1996-2003.

Nivel del SNI	Candidato	Nivel I	Nivel II	Nivel III
Candidato	0.0	1.7	4.6	8.1
Nivel I	1.7	0.0	2.9	6.4
Nivel II	4.6	2.9	0.0	3.5
Nivel III	8.1	6.4	3.5	0.0

Fuente: Elaboración propia con información histórica del SNI, 2012.

Estos últimos resultados obtenidos para el total de solicitudes aprobadas por el SNI, de 1996 a 2003, se obtuvieron también para la gran mayoría de las áreas del conocimiento definidas por dicho sistema de investigación (véase Tabla 7). Más aún, para el Área I se encontró que tres nombramientos del SNI pueden ser clasificados en uno solo (Candidato, Nivel I y Nivel II), mientras que los investigadores con más experiencia (Nivel III) definitivamente pueden ser considerados por separado. En el Área II se definieron dos conglomerados de investigadores; por una parte, los dos niveles inferiores del SNI (Candidato y Nivel I) y por otra parte, los dos niveles superiores (Nivel II y Nivel III). Así, en el Área III también se definieron 2 conglomerados, uno de ellos integra a los investigadores con un nombramiento de Candidato y Niveles I y el otro conglomerado a los dos niveles superiores del SNI (Nivel II y Nivel III). Para el Área IV se conformaron dos agrupamientos, en el primero de ellos se definieron a los niveles Candidato - Nivel I - Nivel II y en el otro clúster a los investigadores mexicanos del SNI con un nombramiento de Nivel III.

En el Área V se identificaron dos grupos de investigadores, aquellos con un nombramiento de Candidato - Nivel I - Nivel II y los investigadores con el nivel superior del SNI (Nivel III). De la misma manera, en el Área VII se identificaron dos conglomerados, el primero de ellos agrupó a los investigadores con un nombramiento de Candidato - Nivel I - Nivel II y el segundo agrupó a aquellos investigadores con un nombramiento de Nivel III. Finalmente, en el Área VII tan solo se detectó un conglomerado. Estos resultados muestran que en seis de las siete áreas del SNI se identificaron, a lo más, dos conglomerados (también se pueden apreciar los dendogramas generados en la Figura 5 que enfatizan este

resultado) en donde se concentraron una gran proporción de dichas solicitudes aprobadas, de 1996 a 2003.

Otro resultado que se deduce de la Tabla 7 es que, en todas las áreas definidas por el SNI, se utilizaron criterios internos de evaluación diferentes, ya que al diferir el conglomerado para cada nivel, entonces bien, se puede decir que cada área valoró criterios científicos diferentes al aprobar una solicitud durante el periodo de estudio. Además, sobresale el hecho de que en la gran mayoría de estas áreas de conocimiento el nivel superior del SNI (Investigador Nacional Nivel III) se diferencia claramente de los otros nombramientos. Ello implica que la información integrada por el SNI para definir a estos últimos investigadores debería ser diferente pero, sobre todo, debería primar la calidad de sus investigaciones (por ejemplo, artículos con calidad JCR, citas en revistas de alto impacto, desarrollos internacionales, etc.).

Tabla 7. Matriz de distancias Hamming para los promedios reales y los promedios estimados mediante *k means*, por área de conocimiento del SNI.

Área	Nivel/Clúster	C1	C2	C3	C4
AREA I: Físico-Matemáticas y Ciencias de la Tierra	Candidato	1.2	81.0	11.1	33.8
	Nivel I	0.4	79.5	9.6	32.3
	Nivel II	3.9	76.4	6.0	28.8
	Nivel III	7.6	73.4	2.4	25.0
AREA II: Biología y Química.	Candidato	50.7	1.0	8.7	23.3
	Nivel I	48.7	1.1	6.7	21.2
	Nivel II	44.7	5.1	2.6	17.2
	Nivel III	39.5	10.4	2.7	12.1
AREA III: Medicina y Ciencias de la Salud.	Candidato	49.0	9.5	24.1	1.1
	Nivel I	46.2	6.7	21.3	1.7
	Nivel II	40.4	0.9	15.5	7.6
	Nivel III	35.9	3.9	11.1	12.3
AREA IV: Humanidades y Ciencias de la conducta.	Candidato	6.2	0.5	35.6	15.7
	Nivel I	5.1	0.7	35.0	14.6
	Nivel II	4.1	1.7	34.7	13.6
	Nivel III	2.8	3.6	34.1	11.7
AREA V: Sociales.	Candidato	7.1	21.0	7.9	0.71
	Nivel I	5.7	19.6	6.5	0.68
	Nivel II	4.8	18.1	5.0	2.2
	Nivel III	4.6	15.8	3.2	4.5
AREA VI: Biotecnología y Ciencias Agropecuarias.	Candidato	42.9	0.8	21.1	7.9
	Nivel I	41.4	0.9	19.4	6.2
	Nivel II	39.5	2.9	17.4	4.2
	Nivel III	39.7	7.6	12.8	2.0
AREA VII: Ingeniería y Tecnología.	Candidato	52.2	27.1	9.4	1.1
	Nivel I	50.7	25.6	7.9	0.4
	Nivel II	47.2	22.0	4.4	3.9
	Nivel III	19.9	2.5	6.1	2.0

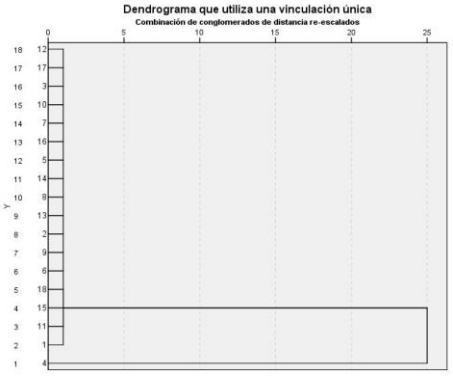
Fuente: Elaboración propia con información histórica del SNI, 2012.

Al continuar con el análisis de los resultados, en la Figura 5 se presentan los dendogramas, por área del conocimiento del SNI, correspondientes a los conglomerados (niveles) estimados y reales con base en los o*utputs* de investigación, considerados por dicho sistema de investigación de 1996 a 2003. Los casos se representan en las filas y las etapas de la fusión en las columnas. A través de estos dendogramas, sin lugar a dudas, se pueden visualizar gráficamente los comentarios vertidos durante este apartado. Es decir, los mencionados dendogramas describen y caracterizan visualmente el hecho de que los criterios definidos en cada una de las áreas definidas por el SNI, durante el periodo de 1996 a 2003, fueron diametralmente distintos y que el número de nombramientos debería de analizarse con más detalle, inclusive por área del conocimiento del SNI. En este contexto, y considerando que los conglomerados estimados no son los mismos o no deberían de ser los mismos en cada una de las áreas de conocimiento del SNI, entonces no puede hacerse un análisis respecto al conglomerado más representativo de todas las áreas.

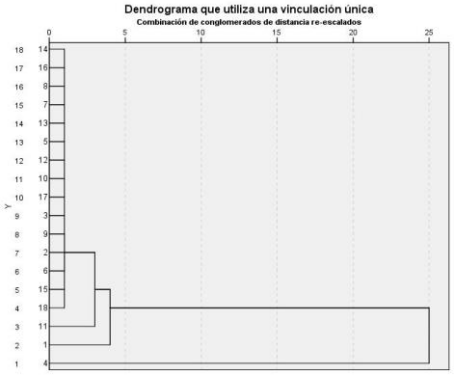
Figura 5. Dendogramas por área del conocimiento definida en el SNI, correspondientes a los conglomerados (niveles) estimados y reales.

AREA I: Físico-Matemáticas y Ciencias de la Tierra.

Promedios estimados

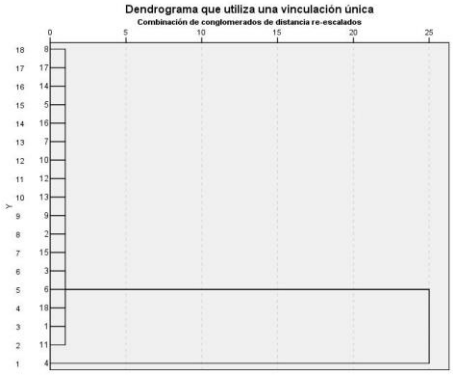


Promedios reales

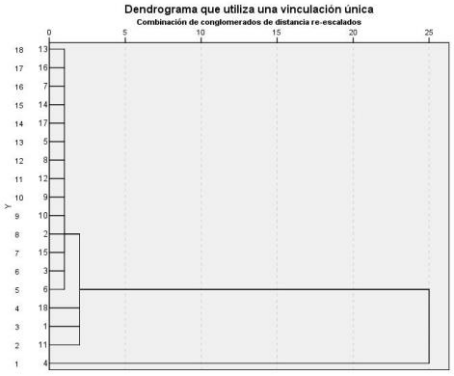


AREA II: Biología y Química.

Promedios estimados

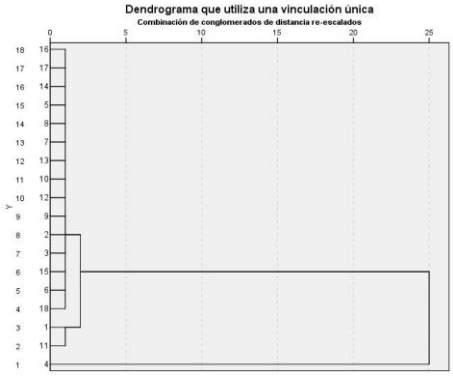


Promedios reales

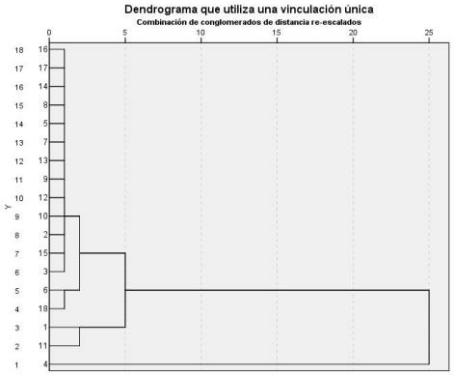


AREA III: Medicina y Ciencias de la Salud.

Promedios estimados

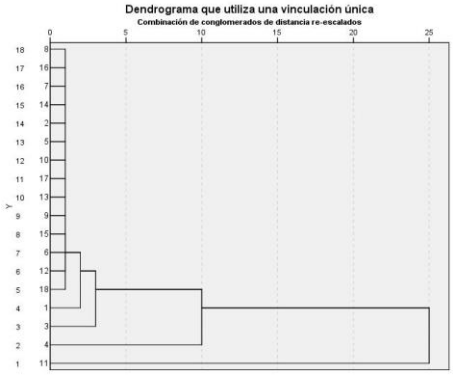


Promedios reales

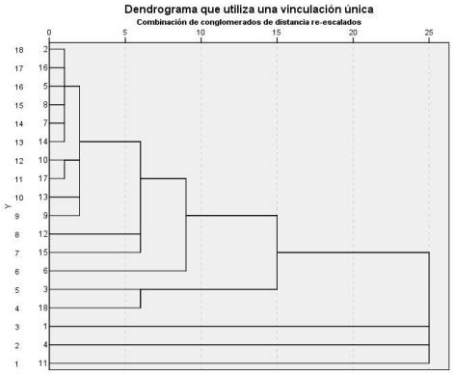


AREA IV: Humanidades y Ciencias de la Salud.

Promedios estimados

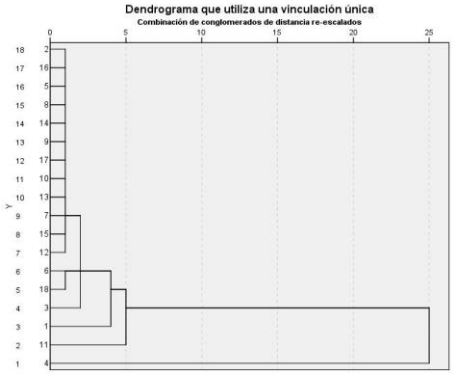


Promedios reales

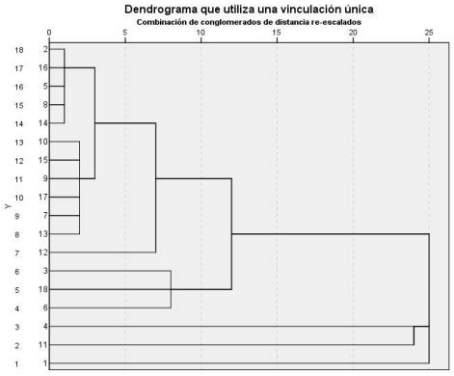


AREA V: Sociales.

Promedios estimados

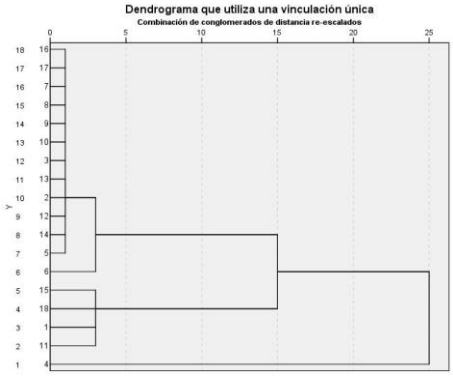


Promedios reales

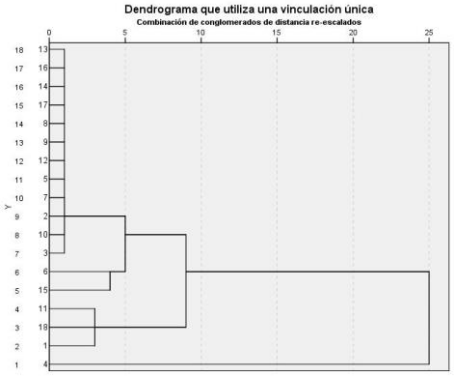


AREA VI: Biotecnología y Ciencias Agropecuarias.

Promedios estimados

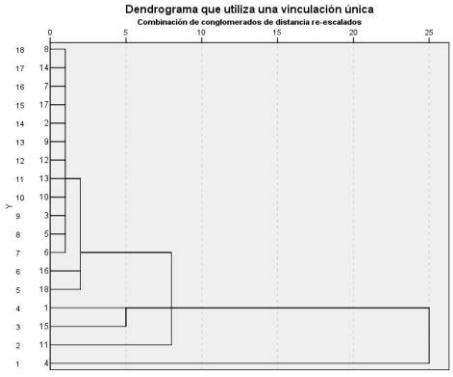


Promedios reales

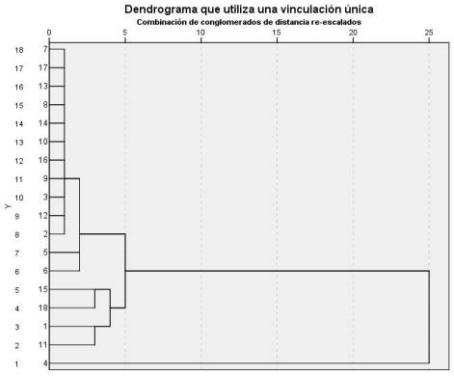


AREA VII: Ingeniería y Tecnología.

Promedios estimados



Promedios reales



Fuente: Elaboración propia con información del SNI, 2012.

En la Tabla 8 se muestra el agrupamiento de las solicitudes aprobadas en cada conglomerado estimado mediante el algoritmo *k means*, según el nivel otorgado durante el periodo de 1996 a 2003 en cada una de las siete áreas definidas por el SNI (el subíndice *i=1,2,…..,7,* en cada uno de los cuatro clúster estimados, hace referencia al área de conocimiento definida por el SNI).

Tabla 8. Distribución en los conglomerados obtenidos mediante el algoritmo *k means* y la participación relativa en cada uno de ellos por nivel y área del SNI, 1996-2003.

Área	Nivel del SNI	Clúster				NE	Total %
		$C1_i$	$C2_i$	$C3_i$	$C4_i$		
AREA I: Físico-Matemáticas y Ciencias de la Tierra.	Candidato	89.0	0.0	0.0	0.0	11.0	100.0
	Nivel I	93.0	0.0	4.4	0.0	2.6	100.0
	Nivel II	76.4	0.1	20.3	2.9	0.3	100.0
	Nivel III	65.1	0.7	22.8	10.7	0.7	100.0
	Total	86.8	0.1	8.3	1.4	3.5	100.0
AREA II: Biología y Química	Candidato	0.0	85.3	0.2	0.0	14.5	100.0
	Nivel I	0.0	85.5	10.0	1.1	3.4	100.0
	Nivel II	1.1	61.8	27.3	9.3	0.5	100.0
	Nivel III	6.2	55.1	23.6	15.2	0.0	100.0
	Total	0.4	80.7	11.0	2.6	5.3	100.0
AREA III: Medicina y Ciencias de la Salud.	Candidato	0.0	0.7	0.0	94.3	5.0	100.0
	Nivel I	0.2	13.5	2.2	81.0	3.1	100.0
	Nivel II	3.0	35.7	10.2	50.4	0.6	100.0
	Nivel III	4.0	26.3	22.2	47.5	0.0	100.0
	Total	0.8	14.4	3.8	78.0	3.0	100.0
AREA IV: Humanidades y Ciencias de la Salud.	Candidato	3.8	90.4	0.0	0.2	5.6	100.0
	Nivel I	11.5	85.7	0.0	0.7	2.1	100.0
	Nivel II	17.0	80.3	0.0	2.3	0.4	100.0
	Nivel III	23.0	70.0	0.5	6.6	0.0	100.0
	Total	12.4	84.2	0.0	1.3	2.0	100.0
AREA V: Sociales.	Candidato	1.4	0.0	0.0	95.4	3.3	100.0
	Nivel I	8.1	0.3	1.8	86.6	3.1	100.0
	Nivel II	14.3	0.9	8.3	76.5	0.0	100.0
	Nivel III	19.7	5.1	17.1	57.3	0.9	100.0
	Total	8.8	0.6	3.5	84.6	2.4	100.0
AREA VI: Biotecnología y Ciencias Agropecuarias.	Candidato	0.0	95.5	0.0	0.8	3.7	100.0
	Nivel I	0.0	86.1	0.6	9.4	3.9	100.0
	Nivel II	0.0	71.0	1.2	27.4	0.3	100.0
	Nivel III	1.5	51.5	17.6	27.9	1.5	100.0
	Total	0.0	84.9	1.1	10.7	3.2	100.0
AREA VII: Ingeniería y Tecnología.	Candidato	0.0	0.0	0.0	94.2	5.8	100.0
	Nivel I	0.0	0.0	3.3	92.1	4.6	100.0
	Nivel II	0.0	4.0	22.1	73.9	0.0	100.0
	Nivel III	0.9	3.5	40.9	53.9	0.9	100.0
	Total	0.0	0.6	6.4	88.7	4.1	100.0

Fuente: Elaboración propia con información histórica del SNI, 2012.

Los resultados de las Tablas 6 y 8 permiten hacer una valoración para los nombramientos otorgados en cada área del SNI durante el periodo de 1996 a 2003. Este análisis permite conocer cuál es el potencial de un investigador del SNI en cada una de sus áreas del conocimiento. Más aún, y a través de este último resultado, se aprecia la capacidad productiva de cualquier investigador mexicano que pertenezca a dicho sistema de investigación. Además, y debido a que las áreas de conocimiento del SNI utilizan diferentes criterios de evaluación, los nombramientos emitidos por cada una de ellas no tienen por qué

ser homogéneos. Es decir, el nombramiento de Candidato, desde el punto de vista cualitativo, no es el mismo en cada una de estas áreas del conocimiento del SNI.

De esta manera, y como se desprende de la Tabla 6, para el Área I: Físico-Matemáticas y Ciencias de la Tierra sobresalieron dos clústers (el C1 y el C3). En este sentido, para esta área del conocimiento se encontró que el 86.8% de las solicitudes aprobadas por el SNI, de 1996 a 2003, fueron agrupadas en el clúster $C1_1$ (Tabla 8). Por su parte, de la Tabla 6 se deduce que la menor distancia hacia el clúster C1 la mostró el nombramiento Nivel I. Con este mismo análisis se aprecia que el clúster $C3_1$ agrupó tan solo al 8.3% de dichas solicitudes aprobadas por el SNI (Tabla 8), mientras que, con base en la Tabla 7, se obtuvo que la menor distancia hacia el clúster C3 fue para el nombramiento Nivel III; en efecto, fue el único nombramiento que definía a este clúster. A partir de estos resultados, se puede afirmar que los evaluadores del Área I: Físico-Matemáticas y Ciencias de la Tierra se inclinan por otorgar el nombramiento de Nivel I, y no tanto porque dichos evaluadores se inclinen o prefieran otorgar el nombramiento Nivel I, más bien la producción académica presentada, durante los años de 1996 a 2003, por un investigador mexicano aprobado por el SIN, en esta área del conocimiento, los hacía merecedores de ese nombramiento.

Con el mismo razonamiento, es fácil mostrar que, para el Área V: Sociales, los dictaminadores se inclinaron más por otorgar el nombramiento de Nivel I que por el nombramiento de Nivel III. De hecho, del total de las evaluaciones realizadas, durante el periodo de 1996 a 2003, otorgaron a 9 de cada diez solicitudes aprobadas, el nombramiento de Nivel I (el 84.6%). En otras palabras, para esta área de conocimiento del SNI se puede decir que sus dictaminadores no apreciaron evidencias para otorgar otro nombramiento que no fuera el Nivel I. Por lo tanto, mediante este análisis se obtuvieron los resultados de la Tabla 9 para cada área definida por el SNI.

Tabla 9. Comparación de los nombramientos estimados por área de conocimiento del SNI.

Área	Comparación de niveles
AREA I: Físico-Matemáticas y Ciencias de la Tierra.	Nivel I ⟩ Nivel III
AREA II: Biología y Química.	Candidato ⟩ Nivel II
AREA III: Medicina y Ciencias de la Salud.	Candidato ⟩ Nivel II
AREA IV: Humanidades y Ciencias de la Salud.	Candidato ⟩ Nivel III
AREA V: Sociales.	Nivel I ⟩ Nivel III
AREA VI: Biotecnología y Ciencias Agropecuarias.	Candidato ⟩ Nivel III
AREA VII: Ingeniería y Tecnología.	Nivel I

Fuente: Elaboración propia con información del SNI, 2012.

Los resultados de la Tabla 9 son contundentes, ya que, como bien se ha mencionado con anterioridad, en el Área I: Físico-Matemáticas y Ciencias de la Tierra, la producción científica presentada por la gran mayoría de los investigadores mexicanos al SNI, de 1996 a 2003, los hizo meritorios de un nombramiento de Nivel I; para el AREA II: Biología y Química, este resultado fue de Candidato; para el AREA III: Medicina y Ciencias de la Salud también fue Candidato; en el AREA IV: Humanidades y Ciencias de la Salud, resultó ser Candidato; en el AREA V: Sociales, fue el Nivel I; para el AREA VI: Biotecnología y Ciencias Agropecuarias, fue Candidato y, para el AREA VII: Ingeniería y Tecnología resultó ser el Nivel I. Según estos resultados se puede dilucidar que en cuatro áreas, la producción científica presentada por sus investigadores mexicanos tuvo el potencial de un investigador con el nombramiento de Candidato. Es decir, de un investigador relativamente joven y que además cuenta con poca experiencia para realizar investigación de calidad como lo sería un investigador con un nombramiento de Nivel III. De manera complementaria, en tres áreas del SNI, la producción científica presentada por sus investigadores mostró un potencial de investigación de un Nivel I. Ahora bien, recuérdese que, con base en los resultados de la Tabla 6, el vector promedio asociado a un Candidato presentó la menor distancia de Hamming respecto al vector promedio asociado a un investigador con un nombramiento de Nivel I. Por lo tanto, y desde el punto de vista cuantitativo, la producción científica de un investigador mexicano aprobado por el SNI, con un nombramiento de Candidato, de 1996 a 2003, fue equiparable a la producción científica reportada por un investigador aceptado en el SNI con un nombramiento de Nivel I. De esta manera, se puede argumentar que el potencial productivo del SNI, visto este como un todo, está en correspondencia con el perfil productivo de un investigador mexicano con un nombramiento de Candidato a Investigador Nacional. Por último, y como conclusión de la aplicación de esta técnica de clasificación, debe señalarse que la producción científica, y más aún, la información solicitada por el SNI no estuvo, al menos durante el periodo de 1996 a 2003, en concordancia con el nombramiento otorgado, independientemente del área de conocimiento, puesto que se deberían agrupar a todos los investigadores, a lo mucho, en dos conglomerados.

5. Comentarios a modo de conclusiones

En este artículo se aplicó una técnica de agrupamiento de datos (*k means*), misma que ha permitido profundizar en la clasificación del conjunto de investigadores que fueron evaluados positivamente por el SNI, de 1996 a 2003. Sin duda, esta técnica de análisis de

datos muestra aspectos significativos y trascendentales de las solicitudes aprobadas por las Comisiones Evaluadoras de las siete áreas del conocimiento definidas por el SNI, durante el periodo antes mencionado. Es decir, esta técnica ha permitido obtener tanto clasificaciones alternativas, basadas en un algoritmo estadístico, como calibrar el nivel de coherencia interna de la realizada por el SNI, la cual se basa estrictamente en la información cuantitativa aportada. En otras palabras, con esta técnica, para el análisis de datos, se ha verificado la coherencia que existe entre la producción científica solicitada para cada investigador aprobado por el SNI y el nombramiento otorgado en dicho círculo de investigación, durante el periodo de 1996 a 2003, para cada una de las áreas del conocimiento definidas por dicho sistema. A su vez, este soporte de información coadyuvará para la toma de las decisiones de los evaluadores que integrarán las futuras Comisiones Evaluadoras del SNI, para que, con ello, la asignación de sus nombramientos sea más eficiente y dinámica.

Modelar el mundo o tratar de modelar al mundo, solo crea, obviamente, una representación simplificada de la realidad. Así, y tomando en cuenta por una parte, la experiencia de las evaluaciones realizadas por el SNI durante el periodo comprendido por los años de 1996 a 2003 y, por otra parte, la producción científica presentada por un investigador mexicano al SNI, con este trabajo se presenta un instrumento que sirva de apoyo y respaldo técnico para las Comisiones Evaluadoras del SNI, en cuanto a la asignación de un nombramiento se refiere. En otras palabras, los resultados de este artículo sirven como un sustento cuantitativo para los evaluadores del SNI y, con ello, tienen la finalidad de hacer más representativa la asignación del nombramiento de una solicitud aprobada por dicho sistema.

El algoritmo *k means* muestra que el SNI, visto como un todo, presenta un perfil productivo similar al de un investigador Nivel I, el cual, a su vez, tiene un perfil productivo equiparable al de un Candidato a Investigador Nacional. Este algoritmo permitió conocer la dinámica (o rigidez) de los nombramientos otorgados por los evaluadores que integran cada una de las Comisiones Evaluadoras definidas en el SNI. Conocer este potencial es saber de lo que son capaces de crear y desarrollar los investigadores que pertenecen a una élite de investigación en México. La diversidad de dichos nombramientos justifica, a su vez, los Criterios Internos de Evaluación que se utilizan hoy por hoy en cada una de las siete áreas del conocimiento definidas por el SNI. Por lo tanto, este resultado, sin duda, justifica el nivel de producción y penetración tanto en el contexto nacional como en el ámbito internacional de los investigadores mexicanos con un registro en el SNI durante el periodo de 1996 a 2003.

Por su parte, es importante mencionar que, a través de la técnica de agrupamiento de datos utilizada en este trabajo, se presentan evidencias estadísticas que bien podrían cimentar las bases para una nueva recapitulación de los Criterios Internos de Evaluación en cada una de las siete áreas del conocimiento definidas por el SNI. De este resultado se deduce que la producción científica reportada, de 1996 a 2003, por los investigadores mexicanos aprobados en este sistema de investigación no se corresponde con el nombramiento otorgado por los evaluadores que conformaron cada una de sus siete Comisiones Evaluadoras. Es decir, los evaluadores consideraron otros elementos, que no son identificados a través de la producción científica reportada, para otorgar un nombramiento, al menos durante el periodo comprendido por los años de 1996 a 2003. Como consecuencia, la información solicitada por el SNI debe ser diferente para cada nombramiento y, más aún, debe ser diferente en cada una de las siete áreas del conocimiento definidas por dicho círculo de investigación mexicano. Ello debido a que se ha mostrado que los Criterios Internos a evaluar en cada una de estas áreas son diferentes y poco homogéneos, al menos desde el punto de vista cuantitativo. Indudablemente, ello se debe a que cada área del conocimiento tiene distintos ámbitos para difundir y/o transferir el conocimiento.

Una consecuencia inmediata del algoritmo para detectar las características predominantes del SNI es que no se justifican nombramientos superiores al del Nivel I en el SNI, a partir del nivel de producción científica reportado a dicho sistema de investigación mexicano. Por lo tanto, no se justifica la existencia de cuatro niveles diferenciados de investigadores (Candidato, Nivel I, Nivel II y Nivel III), sino que como máximo se justifican dos niveles (como primer alternativa un conglomerado definido tan solo por los Candidatos o bien, como segunda alternativa, un primer conglomerado que agrupe a los investigadores con los nombramientos Candidato-Nivel I y un segundo agrupamiento que contenga a los investigadores con los nombramientos Nivel II-Nivel III). Este resultado confirmó que los evaluadores deben utilizar información adicional a la reportada en la base de datos del SNI, la cual debería ser integrada en una misma base de datos. Quizás, los dos bloques (nombramientos) que parecen ser los más claros son: a) los investigadores mexicanos con registro en el SNI, que están definiendo una línea propia de investigación y b) los investigadores mexicanos del SNI que ya cuentan con experiencia en investigación o tienen una línea de investigación consolidada, al menos en el ámbito nacional.

Con base en la producción científica reportada al SNI, por cada solicitud aprobada, el nombramiento definido como Nivel I bien podría ser el que separe a estos dos grupos de investigadores. No obstante, es claro que los investigadores con un nombramiento de Nivel III se diferencian de los otros investigadores del SNI. Por ello, es recomendable que para estos investigadores consolidados se incluyan indicadores sobre la calidad de su investigación (adicional a la cantidad). En este sentido, se sugiere la incorporación de variables como el número de artículos con calidad JCR, páginas y/o citas publicadas en revistas del ISI-JCR, libros y capítulos de libros en editoriales de prestigio internacional.

Del análisis efectuado en el presente trabajo se desprende la necesidad de utilizar información estadística de mejor calidad para proceder a la evaluación. En este sentido, la producción científica, reportada al SNI, debe dejar de ser un trámite más para el investigador que desee presentar su solicitud de ingreso a este sistema de investigación en México. Al ser presentada con mayor responsabilidad, por parte del solicitante, mejor será la información integrada por el SNI y, en consecuencia, serán más representativos los nombramientos emitidos por cada una de sus siete Comisiones Evaluadoras. Sin embargo, es claro que la valoración final de una solicitud seguirá dependiendo del criterio individual y subjetivo de cada evaluador que integra una Comisión Evaluadora del SNI. Por ello, y desde el punto de vista de una estimación, este trabajo avanza y cimienta las bases para que, sumado al factor humano, se obtengan dictámenes más robustos y representativos por parte del SNI. Este análisis tiene sentido, ya que gran parte de las variables suministradas a la técnica de agrupamiento, utilizada en este trabajo, son cuantitativas. Como consecuencia, para cada estimación se espera lo siguiente: en la medida en que se disponga de mayor y mejor información por parte del SNI, entonces se obtendrán evaluaciones más robustas, las cuales, a su vez, conllevarán a tener un panorama más claro del potencial de los investigadores mexicanos que integran al SNI.

Por último, en este artículo se presenta un algoritmo de agrupamiento de datos para hacer más eficiente el proceso de selección en el SNI con el objetivo de; primero, captar información de calidad y de primera mano; segundo, hacer más eficiente la dinámica (recepción) de la información; tercero, reducir el tiempo de respuesta por parte del SNI; cuarto, obtener resultados más representativos los cuales, a su vez, se trasformen en nombramientos más confiables; quinto y tal vez lo más importante, la técnica aplicada se ofrece como buen instrumento para complementar la evaluación del SNI por pares, siempre y cuando mejore la información cualitativa que ahora parecen utilizar los evaluadores, y que no

queda recogida en las variables actuales. Mencionar que los aportes del presente trabajo son relevantes en la medida en que el sistema de evaluación de investigadores en México es la base para asignar recursos de investigación, con lo cual todas las propuestas que ayuden a mejorar dicho sistema de evaluación coadyuvarán, sin lugar a dudas, a incrementar la eficiencia y transparencia en la asignación de recursos para la investigación.

Referencias

Anderberg, Michael R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Bauwens, Luc. (1998). *A New Method to Rank University Research in Economics in Belgium*, mimeo. CORE, Université Catholique de Louvain, Belgium

Bao, Zhiqiang, Bing, Han and Wu, Shunjun. (2006). A General Weighted Fuzzy Clustering Algorithm. En Aurélio Campilho and Mohamed Kamel (Eds), *Image Analysis and Recognition. ICIAR 2006. Lecture Notes in Computer Science*, (Vol. 4142, pp. 102-109). Springer, Berlin, Heidelberg. Recuperado de https://link.springer.com/chapter/10.1007/11867661_10

Bezdek, James C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Ed. Plenum Press.

Blum, Avrim y Mitchell, Tom. (julio, 1998). Combining labeled and unlabeled data with co-training*. Proceedings of the 11th annual conference on computational learning theory (COLT), Madison, USA*, 92-100.

Bock, Hans-Hermann. (2008). Origins and extensions of the k-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics*, *4*(2), 1-18. Recuperado de https://eudml.org/doc/130880

Campello, Ricardo, Hruschka, Eduardo R. y Alves, Vinícius S. (2009). On the efficiency of evolutionary fuzzy clustering. *Journal Heuristics*, 15, 43-75. Recuperado de https://link.springer.com/article/10.1007/s10732-007-9059-6

Consejo Nacional de Ciencia y Tecnología, CONACyT. (2017). Reglamento del Sistema Nacional de Investigadores. México. Recuperado de http://www.conacyt.gob.mx/index.php/el-conacyt/sistema-nacional-de-investigadores/marco-legal

Dae-Won, Kim, Kwang, H. Lee and Doheon, Lee. (2004). On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, *37*(10), 2009-2025. Recuperado de https://dl.acm.org/citation.cfm?id=2793552

Dietterich, Thomas G., Lathrop, Richard H. and Lozano-Perez, Tomás. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Inteligence*, *89*(1-2), 31-71. Recuperado de http://www.sciencedirect.com/science/article/pii/S0004370296000343

Dunn, Joseph. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well separated cluster. *Journal of Cybernetics*, *3*(3), 32-57. Recuperado de http://www.tandfonline.com/doi/abs/10.1080/01969727308546046

Fayyad, Usama, Piatetsky-Shapiro, Gregory y Smyth, Padhraic. (1996). Knowledge discovery and data mining: Towards a unifying framework. *Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD)*, *Portland, USA*, 82-88. Recuperado de https://dl.acm.org/citation.cfm?id=3001460&picked=prox

Fisher, Ronald Aylmer. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179-188.

Gärtner, Thomas, Flach, Peter A., Kowalczyk, Adam and Smola, Alex J. (july, 2002). Multi-instance kernels. *Proceedings of the 19th international conference on machine learning (ICML). Sydney, Australia*, 179-186. Recuperado de https://dl.acm.org/citation.cfm?id=656014

Goethals, Bart, Hoekx, Eveline y Van den Bussche, Jan. (2005). Mining tree queries in a graph*. The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, Illinois, USA*, 61-69. Recuperado de https://dl.acm.org/citation.cfm?id=1081870&picked=prox

Greene, William H. (2008). *Econometric Analysis* (6ª ed.). New York University: Prentice Hall.

Han, Jiawei, y Kamber, Micheline. (2006). *Data Mining: Concepts and Techniques* (2a. ed.). USA, Waltham: Elsevier.

Hamming, Richard Wesley. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*(2), 147-160.

Huang, Zhexue. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283-304. Recuperado de https://link.springer.com/article/10.1023/A:1009769707641

Kailing, Karin, Kriegel, Hans-Peter, Pryakhin, Alexey and Schubert, Matthias. (2004). Clustering multi-represented objects with noise*. Proceedings of the 8th Pacific-Asia conference on knowledge discovery and data mining (PAKDD). Sydney, Australia*, 394-403.

Kan, Raymond and Zhou, Guofu. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, *42*(3), 621-656. Recuperado de http://apps.olin.wustl.edu/faculty/zhou/KZ_JFQA_W07.pdf

Khurram, Jamali, Kirsten, Wandschneider y Phanindra, V. Wunnava. (2007). The effect of political regimes and technology on economic growth. *Applied Economics*, *39*(11), 1425-1432. Recuperado de https://econpapers.repec.org/article/tafapplec/v_3a39_3ay_3a2007_3ai_3a11_3ap_3a1425-1432.htm

Kittler, Josef, Hatef, Mohamad, Duin, Robert P.W. y Matas, Jiri. (1998). On combining classifiers. *IEEE Trans Pattern Analysis and Machine Intelligence*, *20*(3), 226-239.

Kriegel, Hans-Peter, Borgwardt, Karsten M., Kröger, Peer, Pryakhin, Alexey, Schubert, Matthias and Zimek, Arthur. (2007). Future trends in data mining. *Data Min Knowl Disc*, 15, 87-97.

Kriegel, Hans-Peter, Kröger, Peer, Pryakhin, Alexey and Schubert, Matthias. (April 2004). Using support vector machines for classifying large sets of multi-represented objects. *Proceedings of the 4th SIAM international conference on data mining (SDM). Florida, USA,* 102-113.

Kriegel, Hans-Peter, Pryakhin, Alexey y Schubert, Matthias (april, 2005). Multi-represented kNN-classification for large class sets. *Proceedings of the 10th international conference on database systems for advanced applications (DASFAA). Beijing, China*, 511-522.

Krueger, Anne and Ruttan Vernon. (1989). Development thougth and development assistance. In *Aid and Development* (pp. 13-28). Baltimore, USA: The Johns Hopkins University Press.

Kuo, Renjieh, Ho, L. M., and Hu, C. M. (2002). Integration of self-organizing feature map and k-means algorithm for market segmentation. *Computers and Operations Research*, *29*(11), 1475-1493.

MacQueen, James B. (1967). Some methods for classification and analysis of mulivariate observations. In L.M. LeCam, J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability Volume: 1 Statistics*. University of California Press, Berkely, 281-297. Recuperado de https://projecteuclid.org/euclid.bsmsp/1200512992

Mahdavi, Mehrdad y Abolhassani, Hassan. (2009). Harmony K-means algorithm for document clustering. *Data Min Knowl Disc*, *18*(3), 370-391.

Prasanta, Kumar Dey. (2006). Integrated project evaluation and selection using multiple-attribute decision-making technique. *International Journal Production Economics*, *103*(1), 90-103.

Reguia, Cherroun. (2014). Product innovation and the competitive advantage. *European Scientific Journal, 1*, 140-157.

Schultz, Theodore W. (1961). Investment in human capital. *American Economic Review*, *51*(1), 1-17.

Shian-Chang, Huang, En-Chi, Chang and Hsin-Hung, Wu. (2009). A case study of applying data mining techniques in an outfitter's customer value analysis. *Expert Systems with Applications*, *36*(3), 5909-5915.

Soto, Jesús A., Flores-Sintas, Antonio and Vigo, M. Isabel. (2004). Marco formal para una nueva función objetivo en agrupación difusa. *Revista Iberoamericana de Inteligencia Artificial*, *8*(23), 35-41.

Tan, Pang-Ning, Steinbach, Michael and Kumar, Vipin. (2006). *Introduction to Data Mining*. USA: Pearson Addison New York, Wesley.

Washio, Takashi and Motoda, Hiroshi. (2003). State of the art of graph-based data mining. *ACM SIGKDD Explorations Newsletter*, *5*(1), 59-68.

Weidmann, Nils, Eibe, Frank and Bernhard, Pfahringer. (September, 2003). A two-level learning method for generalized multinstance problems. *Proceedings of the 14th European conference on machine learning (ECML), Cavtat-Dubrovnik, Croatia*, 468-479. Recuperado de https://link.springer.com/chapter/10.1007/978-3-540-39857-8_42

Wu, Xindong, Kumar, Vipin, Quinlan, J. Ross, Ghosh, Joydeep, Yang, Qiang, Motoda, Hiroshi … Steinberg, Dan. (2008). Top 10 algorithms in data mining. *Knowl Inf Syst*, *14*(1), 1-37.

Yarowsky, David. (1995). Unsupervised word sense disambiguation rivaling supervised methods*. ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Stroudsburg, PA, USA*, 189-196.