

# Reseña del Estratificador INEGI

## Review of the INEGI Stratifier

Víctor Alfredo Bustos y de la Tijera y Ana Miriam Romo Anaya\*

\* INEGI, alfredo.bustos@inegi.org.mx y miriam.romo@inegi.org.mx, respectivamente.

Coloridos globos antes de su lanzamiento en el Parque Nacional de Göreme, Capadocia, Turquía/MarBom/iStock



El Instituto Nacional de Estadística y Geografía (INEGI) ha buscado desarrollar alternativas que le permitan cumplir simultáneamente con, por una parte, la necesidad de preservar la privacidad de sus informantes, tanto personas físicas como morales, y por la otra, la de difundir información de México que contribuya a la toma de decisiones. El *Estratificador INEGI* representa un esfuerzo en esa dirección e ilustra una estrategia a seguir: primero, desarrollar servicios para hacer disponibles herramientas de análisis estadístico para conjuntarlas con datos que no pueden ser difundidos; segundo, limitar sus resultados a los que no ponen la confidencialidad en riesgo. Ello debe permitir al usuario decidir el tipo de análisis, los datos a ser usados y el nivel de desagregación geográfica deseado. Así, obtiene los resultados que requiere, más allá de los productos censales tradicionales, y el Instituto logra que la información reciba usos adicionales. En el caso particular del *Estratificador INEGI*, un primer conjunto de herramientas se refiere a tres procedimientos para formar grupos de unidades similares en su interior, y diferentes entre grupos. Los datos consisten en valores porcentuales de variables censales referidos a distintas desagregaciones geográficas: estatal, municipal y manzanas dentro de localidades. Un segundo conjunto está formado por gráficas y criterios numéricos para evaluar cada resultado y permitir al usuario elegir entre ejercicios alternativos de análisis. El servicio le permite exportar los resultados de todas las estratificaciones realizadas en una misma sesión, o únicamente de las seleccionadas. Estos consisten solo del listado de identificadores de las áreas geográficas involucradas y de sus correspondientes etiquetas de estrato.

**Palabras clave:** censura; confidencialidad; estratificación; censos; servicios en línea.

## Introducción

Las oficinas nacionales de estadística (ONE) han incorporado herramientas de las tecnologías de la información y la comunicación (TIC) como complemento a la información que ya se distribuía en medios impresos. Su versatilidad permite empacar

The National Institute of Statistics and Geography (INEGI) has sought to develop alternatives that allow it to simultaneously comply with, on the one hand, the need to preserve the privacy of its informants, both individual and companies, and on the other, the need to disseminate information of Mexico that contributes to decision-making. The *INEGI Stratifier* represents an effort in this direction and illustrates a strategy to be followed: first, develop services to make statistical analysis tools available to combine them with data that cannot be disseminated; second, to limit its results to those that do not put confidentiality at risk. This should allow the user to decide the type of analysis, the data to be used and the desired level of geographic disaggregation. Thus, it obtains the results it requires, beyond traditional census products, and the Institute ensures that the information receives additional uses. In the particular case of the *INEGI Stratifier*, a first set of tools refers to three procedures to form groups of similar units within each of them, and different between groups. The data consist of percentage values of census variables referring to different geographical disaggregations: state, municipal and blocks within localities. A second set consists of graphs and numerical criteria to evaluate each result and allow the user to choose between alternative analysis exercises. The service allows you to export the results of all stratifications carried out in the same session, or only those selected. These consist only of the list of identifiers of the geographical areas involved and their corresponding stratum labels.

**Key words:** censored information; confidentiality; stratification; censuses; on-line services.

volúmenes importantes de esta para hacerla accesible a través de una ventanilla única, es decir, mediante una sola interfaz de usuario que permita combinar datos correspondientes a diferentes ejercicios de captación, periodos y niveles de desagregación geográfica. Ejemplo de lo anterior lo constituye el Sistema de Información Municipal en

Bases de Datos (SIMBAD) que el INEGI puso en funcionamiento a mediados de la década de los 90 del siglo pasado.<sup>1</sup>

Sin embargo, siguiendo en esa dirección, las ONE enfrentan de distintas maneras la disyuntiva entre hacer disponible información más desagregada para fines de análisis y de toma de decisiones, por un lado, y evitar que la que corresponde a algunos individuos pueda llevar a su identificación, violando así su privacidad y el requisito de preservar la confidencialidad, por el otro. Por esta razón, se desarrollaron técnicas para anonimizar bases de datos, englobadas en cinco tipos de operaciones: generalización, supresión, anatomización, permutación y perturbación.<sup>2</sup> No obstante, su eficiencia ha sido puesta en duda recientemente,<sup>3,4</sup> en particular, a partir del advenimiento de métodos de análisis desarrollados alrededor de los *Big Data* y de la accesibilidad a múltiples bases de datos.

La anterior afirmación nos empuja a buscar estrategias alternativas para proteger la privacidad de los individuos. Además de nuevos algoritmos de anonimización, es útil explorar otras avenidas que faciliten el acceso y la explotación de los datos en poder de las ONE en apoyo a una mejor toma de decisiones. En el caso mexicano, la *Ley del Sistema Nacional de Información Estadística y Geográfica (LSNIEG)*, en su artículo 4, establece que el sistema deberá "... promover el conocimiento y uso de la información..." a efecto de coadyuvar en el desarrollo nacional, apoyando su planeación, instrumentación y seguimiento.

El INEGI ha venido actualizando los productos que, en su momento, eran distribuidos a través de

discos compactos con estos fines, como el Sistema para la Consulta de Información Censal (SCINCE). Estos, por su naturaleza, incluían bases de datos con desagregación geográfica hasta el nivel que impida la identificación de los informantes. En aquellos casos en los que esta lo amerite, la base de datos se somete a una censura que excluye valores de variables para municipios, Áreas Geoestadísticas Básicas (AGEB) o manzanas con tres o menos personas, hogares o establecimientos. Las versiones actuales en línea han preservado estas estrategias, pero en general no han incorporado salvaguardas que tomen en cuenta la aparición de nuevas tecnologías, como las ya mencionadas. Adicionalmente, y en concordancia con experiencias internacionales, el INEGI ha hecho disponibles microdatos a nivel de registro. Conforme a las prácticas tradicionales, bases de datos de encuestas por muestreo, en particular en hogares, así como muestras aleatorias de datos censales, han sido distribuidas por diversos medios. En ambos casos, la identificación de los individuos y de las viviendas ha sido removida con fines de anonimización.

Más aún, el INEGI ha puesto en servicio el denominado Laboratorio de Microdatos. Mediante este mecanismo, bases de datos censales a nivel de registro individual se ponen a disposición de usuarios e investigadores dentro de un ambiente controlado, aunque solamente para algunos temas. Los servicios del Laboratorio<sup>5</sup> y de procesamiento remoto están disponibles para funcionarios de organismos internacionales y nacionales de estadística, investigadores de instituciones académicas y estudiantes de nivel superior. En el primer caso, es necesario acudir a las instalaciones del Instituto en la Ciudad de México.<sup>6</sup> Originalmente, se ofrecían SPSS, Excel, ArcGIS, R y Stata como herramientas para procesamiento de datos.

De forma similar y siguiendo la tendencia que va desde los productos impresos a los servicios estadísticos en línea, a partir del Censo de Población

1 INEGI. *Notas, Revista Trimestral de Información y Análisis del Instituto Nacional de Estadística, Geografía e Informática*. México, INEGI, 1998:2, pp. 55-57 (DE) <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825001564> consultado el 01/03/2022.

2 Eyupoglu C., M. A. Aydin, A. H. Zaim y A. Sertbas. *An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques*. *Entropy (Basel)*. 2018 May 17; 20(5):373. doi: 10.3390/e20050373. PMID: 33265463; PMCID: PMC7512893.

3 Rocher, L., J. M. Hendrickx, & Y. A. de Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models", en: *Nat Commun*. 10, 3069, 2019 (DE) <https://doi.org/10.1038/s41467-019-10933-3>

4 European Commission. *Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques* (DE) [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)

5 Reglas de operación del Laboratorio de Microdatos: [https://sc.inegi.org.mx/repositorioNormateca/Or\\_24Mar20.pdf](https://sc.inegi.org.mx/repositorioNormateca/Or_24Mar20.pdf)

6 Existe una segunda sede del Laboratorio en el Centro de Investigación y Docencia Económicas, A. C., en Santa Fe, Ciudad de México.

ción y Vivienda (CPV) 2010,<sup>7</sup> se instrumentó una estrategia paralela consistente en el desarrollo de rutinas de análisis estadístico usando el lenguaje R. La principal motivación de este ejercicio consiste en hacer disponibles los datos de manera indirecta a través de las mencionadas rutinas, evitando que el usuario tenga acceso a estos. El usuario deberá poder seleccionar los proyectos de generación de información estadística, las variables, así como las desagregaciones geográficas que satisfagan sus requerimientos. Del mismo modo, podrá elegir rutinas de análisis de entre la gama ofrecida y solicitar estadísticas que le permitirán evaluar la calidad de los resultados. Los obtenidos así consistirán en valores estadísticos de los que no sea posible discernir la información individual; por ejemplo, la estimación de los parámetros de un modelo de regresión o el entrenamiento de un algoritmo de inteligencia artificial. El primer ejemplo, y hasta ahora el único, en esta dirección está dado por el denominado *Estratificador INEGI*.

7 Programa estadístico del INEGI.

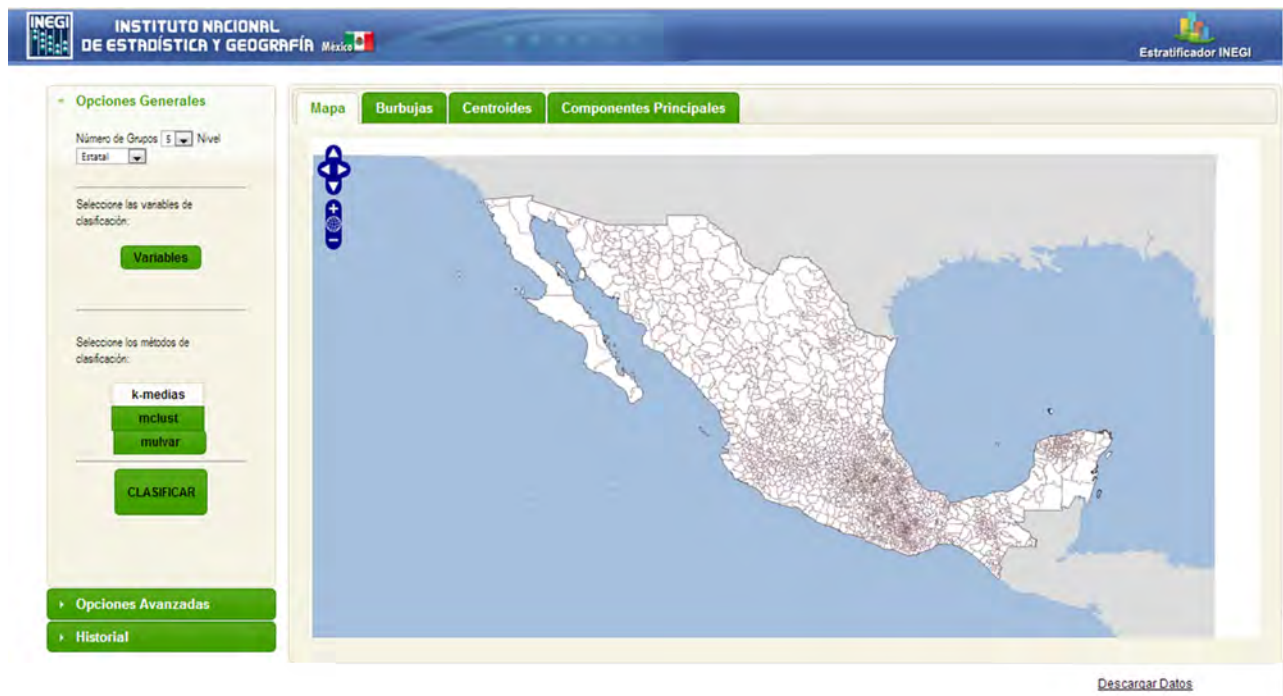
## La primera versión

A partir de la publicación de los resultados del CPV 2010 se inició el desarrollo del *Estratificador INEGI*. La pantalla de entrada al servicio publicado (ver figura 1) concedía acceso a la información censal con una desagregación geográfica hasta el nivel de municipio. El usuario podía fijar el número de grupos a formar y elegir las variables relevantes a su investigación. Seleccionaba el método de clasificación a usar de entre los tres disponibles. Los resultados se mostraban tanto en forma de mapas como de gráficas. Al finalizar la sesión, le era posible exportar archivos con su selección de resultados.

Por lo pronto, no profundizaremos sobre los anteriores temas, discusión que postergaremos al hablar de la versión actualizada. Baste mencionar el resultado del análisis de una bitácora de uso del servicio a lo largo del par de años que mediaron entre este y su actualización. A pesar de disponer solamente de una interfaz en español, se registra-

Figura 1

### Pantalla principal de la versión 1.0 del *Estratificador INEGI*



ron accesos desde más de 25 países en tres continentes (figura 2). Por supuesto, casi todos tuvieron lugar desde el territorio mexicano, con la salvedad de seis estados. Desde 22 entidades de los Estados Unidos de América (EE. UU.) se realizaron también consultas (figura 3).

El interés despertado por el servicio se reflejó en una serie de invitaciones para realizar presentaciones ante diversas instituciones gubernamentales y de educación superior, tanto públicas como privadas. A partir de lo anterior, se concluyó que la inclusión de utilerías adicionales beneficiaría a los

Figura 2

**Países desde los que tuvo lugar el uso de la versión 1.0 del *Estratificador INEGI***

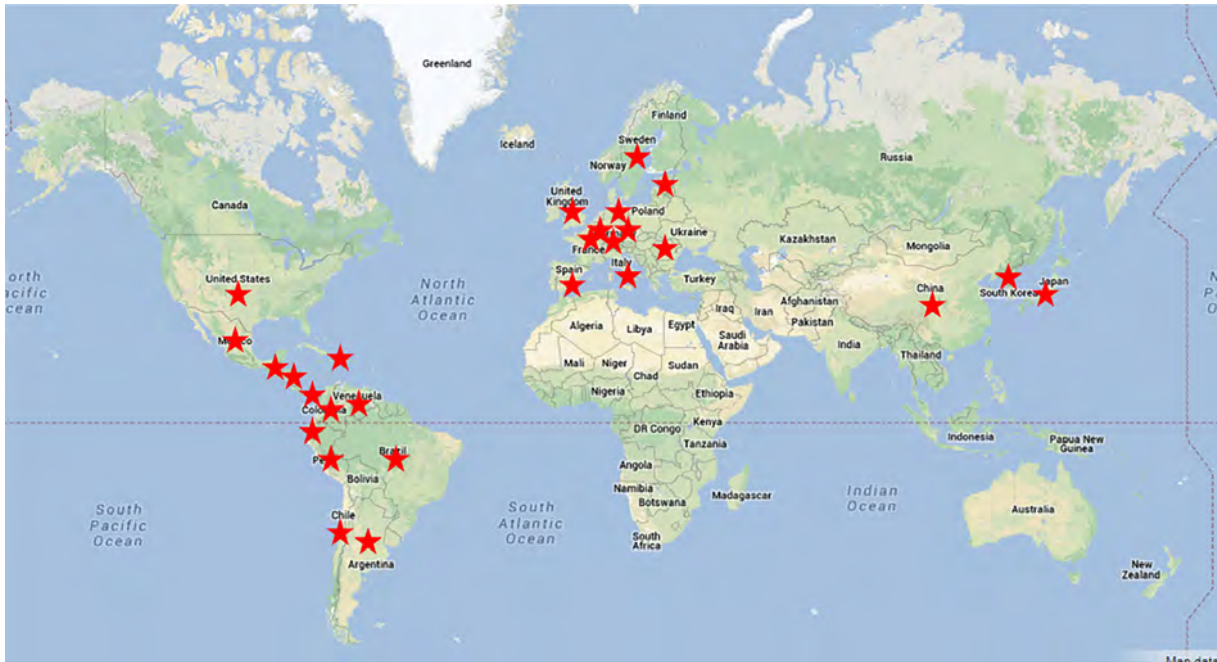


Figura 3

**Estados de los EE. UU. en los que se hicieron consultas a la versión 1.0 del *Estratificador INEGI***



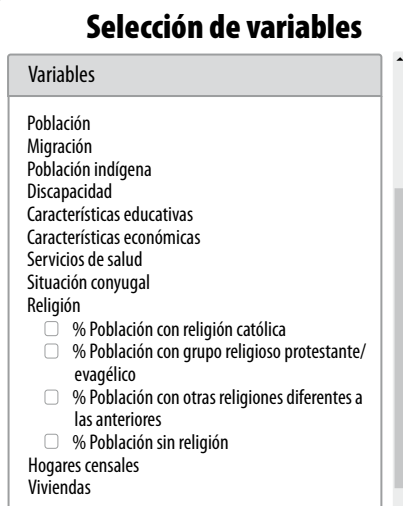
usuarios. Ello dio lugar a una versión actualizada del *Estratificador INEGI*.

## Estratificador INEGI, V. 1.1

Con base en los lineamientos anteriores, se desarrollaron diversas herramientas para cumplir con las etapas de elaboración de: consulta, análisis de la información, despliegue gráfico de resultados, cálculo de medidas de calidad y exportación de resultados.

La consulta consiste, primeramente, en la selección de las variables que entrarán en el análisis (ver figura 4). Se ofrece al usuario un listado de temas censales, cada uno de los cuales abre, a su vez, un menú de variables incluidas; todas están expresadas como porcentajes.

Figura 4

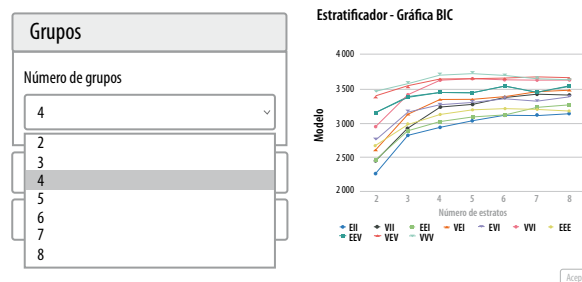


Se brinda, también, la opción de fijar de antemano el número de clases o estratos a formar. Incluido en uno de los procedimientos para realizar la agrupación de unidades, se agrega un criterio de información bayesiano que apoya la determinación de este número (ver figura 5).

En vista de que algunos ejercicios de estratificación hacen uso de transformaciones de los datos, se concede al usuario la posibilidad de aplicarlas con el fin de reproducirlos y compararlos. La primera de ellas, la estandarización, elimina las unidades de medida en los datos, lo que parece innecesario

Figura 5

## Selección del número de grupos a formar

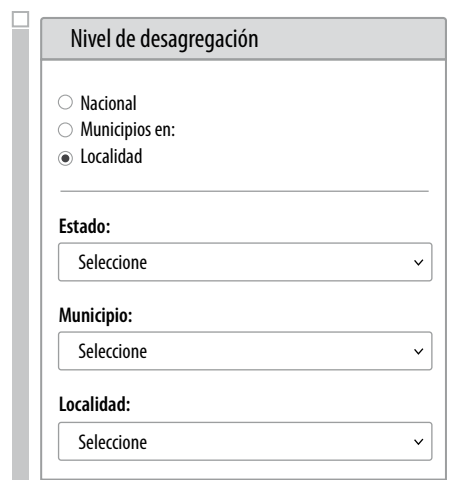


cuando estos son representados como porcentajes; sin embargo, el Índice de Marginación del Consejo Nacional de Población (CONAPO) recurrió a esta bajo las mismas circunstancias, por lo cual la hemos conservado. El Índice se basaba, además, en la primera componente principal de los datos estandarizados, por lo que el cálculo de este y otros indicadores compuestos se ha mantenido.

Acto seguido, el usuario selecciona el nivel de desagregación geográfica para estas variables. A saber, estos son: estatal, municipal nacional, municipal restringido a un conjunto de entidades seleccionadas y manzanas dentro de localidades urbanas<sup>8</sup> (ver figura 6).

Figura 6

## Selección del nivel de desagregación geográfica a trabajar



<sup>8</sup> En una primera instrumentación, los datos a nivel de manzana no estuvieron sujetos a censura por confidencialidad.

Para el paso final, el usuario tiene a su disposición tres procedimientos de estratificación que puede usar indistintamente (ver figura 7). El primero de ellos se incluye con fines de comparación, pues en este se basó el producto impreso denominado *Niveles de bienestar en México*,<sup>9</sup> que se publicó a partir de la información del XI Censo General de Población y Vivienda 1990. El segundo es el método multivariado clásico denominado de las *k-medias*.<sup>10</sup> El tercero, incluido en librerías de R, es el denominado *mclust*,<sup>11</sup> y que está basado en un modelo para el comportamiento de los segundos momentos de la información disponible, bajo un supuesto de *gaussianidad* local a corroborar. A partir de la prueba de todos los modelos realizada por este con base en la información disponible, se sugieren tanto el número de grupos a formar como el modelo de comportamiento para los segundos momentos a ser usado.

Al finalizar la construcción de la consulta, se muestra al usuario un resumen de sus selecciones (ver figura 8). Este texto podrá ser consultado en el historial de estratificaciones, en apoyo a los ejercicios de comparación de resultados, indicando las circunstancias en que cada uno de ellos fue llevado a cabo.

9 INEGI. *Niveles de bienestar en México*. México, INEGI, 1993 (DE) <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825450557> consultado el 01/03/2022.

10 Hartigan, J. A. *Clustering Algorithms (Probability & Mathematical Statistics)*. John Wiley & Sons Inc., 1975.

11 *Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation* (DE) <https://cran.r-project.org/web/packages/mclust/mclust.pdf> consultado el 01/03/2022.

Figura 7

### Elección del método de estratificación

**Métodos**

Estandarizar

Componentes principales

**Métodos:**

*k-medias*

*mulvar*

*mclust*

Figura 8

### Resumen de consulta construida mediante la versión 1.1 del *Estratificador INEGI*

**Variables:**

- % Pobladores en viviendas que disponen de televisor
- % Pobladores en viviendas que disponen de refrigerador
- % Pobladores en viviendas que disponen de automóvil o camioneta

**Grupos: 4**

**Nivel de desagregación:** Nacional por municipio

**Método:** *k-medias*

## Resultados

Una vez hechas las selecciones anteriores, el sistema permite visualizar los resultados tanto en forma de mapas como de gráficas hasta en cuatro dimensiones, como se muestra en la figura 9. En el primer caso partiendo de un código de color, que va del rojo, que representa las unidades que muestran condiciones más desfavorables, hasta el verde oscuro en el otro extremo; con este es posible obtener diversos *k* mapas (figuras 10 y 11).

Se obtiene así un mosaico que permite relacionar unidades que, aunque geográficamente distantes, muestran condiciones similares; en otras palabras, se desfigura la noción de regiones geográficas gruesas que parecieran ser homogéneas, para dar lugar a tantos *archipiélagos* como grupos se hayan formado, cuyos integrantes se encuentran repartidos a lo largo y ancho del territorio nacional, o de la región seleccionada.

Manteniendo el mencionado código de color, también es posible desplegar gráficas de globos en planos definidos por dos variables (ver figura 12). Adicionalmente, se pueden incorporar datos relativos al valor de una tercera variable, que dará un mayor o menor volumen al globo. Cada globo se referirá a una unidad geográfica y mostrará el color de la clase a la que esta pertenece, complementando la información de las variables hasta ahora consideradas con una cuarta fuente de información.

Figura 9

### Pantalla resumen de resultados de la versión 1.1 del *Estratificador INEGI*

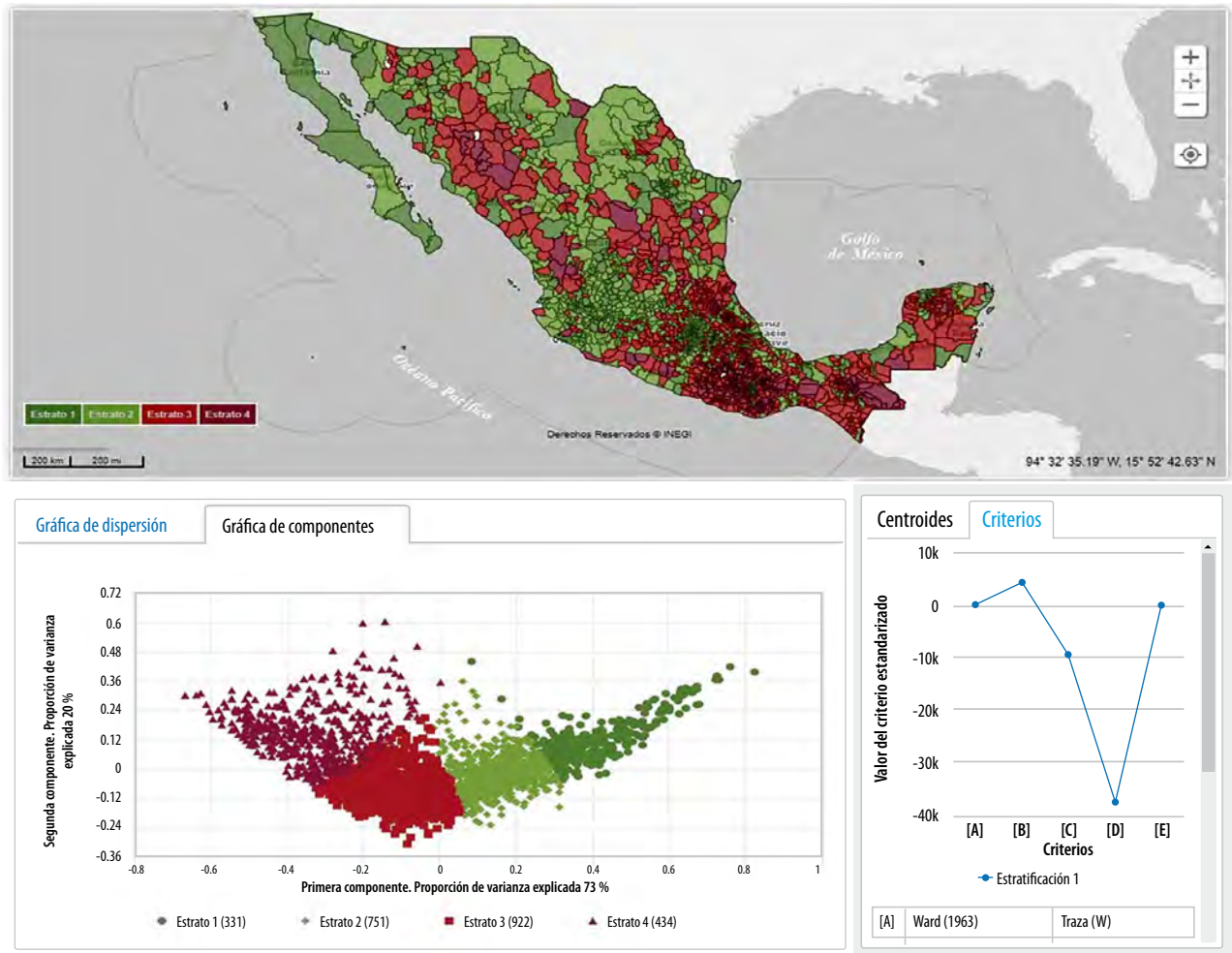


Figura 10

### Ejemplo de estratificación nacional por municipio

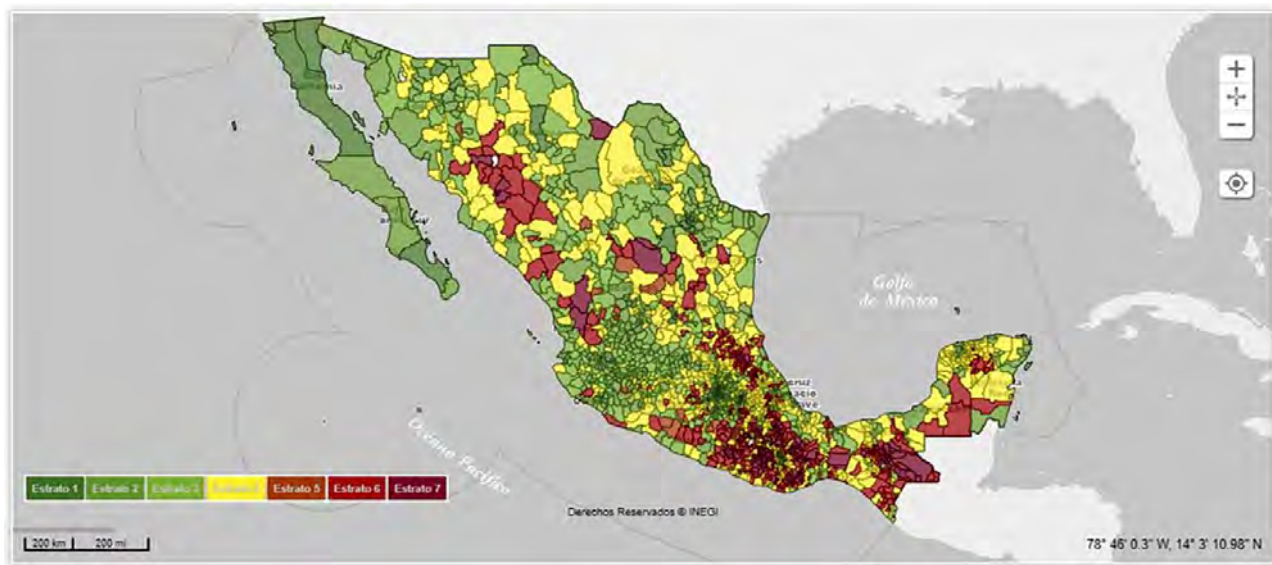




Figura 11

### Ejemplo de estratificación regional por municipio

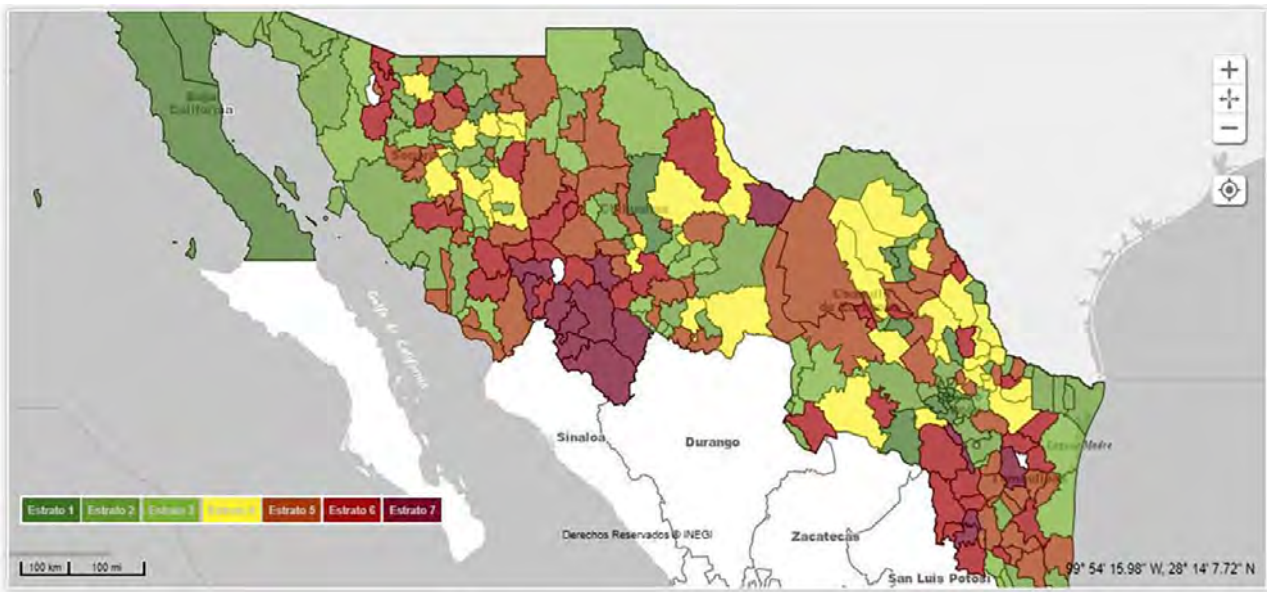
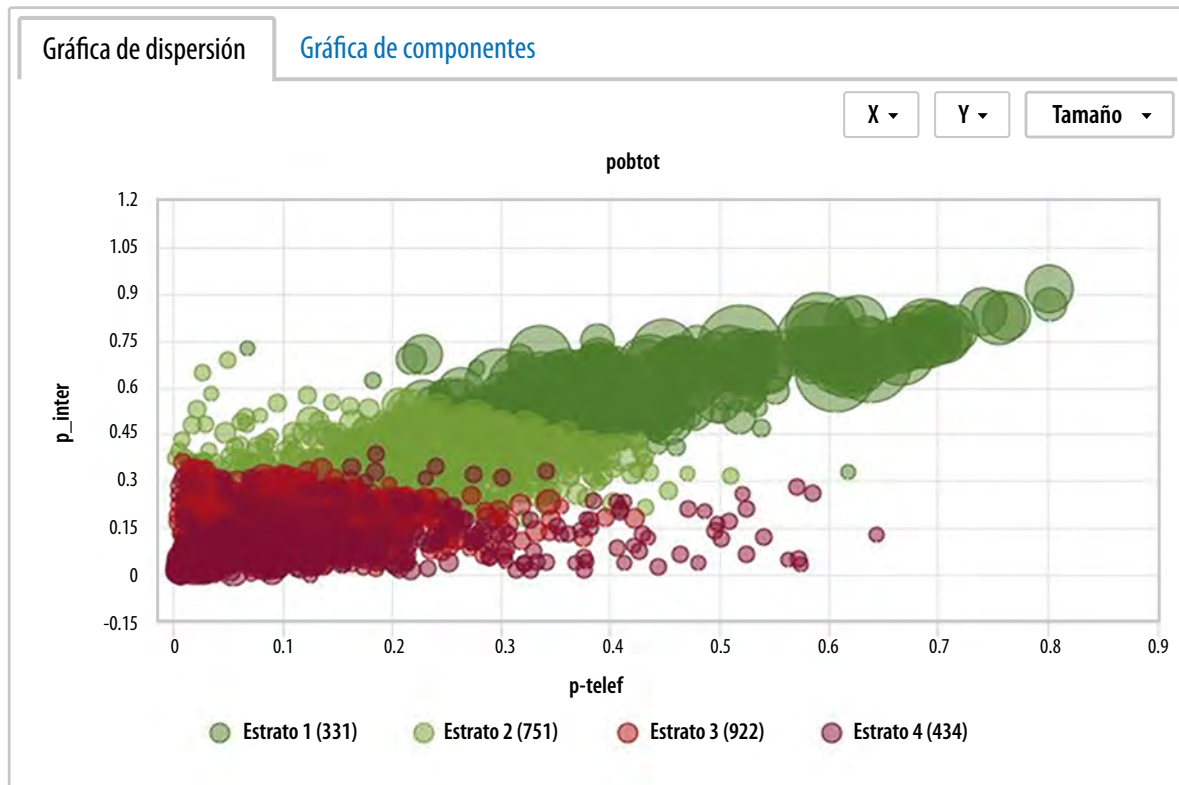


Figura 12

### Gráfica de dispersión



Cuando el número de variables consideradas para llevar a cabo una estratificación es grande, la cantidad de parejas utilizadas para definir los ejes lo es mucho más. Ello dificulta hacer una evaluación después de observar un conjunto numeroso de gráficas. Es por ello que hemos considerado útil incluir una representación gráfica para la cual los ejes representan a las primeras dos componentes principales (ver figura 13). Esto da lugar a una representación más canónica, auxiliar en la interpretación de los resultados.

Otra representación gráfica disponible es la que se refiere a los valores de los promedios, dentro de cada uno de los grupos formados, para cada variable considerada (ver figura 14). De esta manera, se aprecian discrepancias más o menos importantes entre los grupos; cuando las diferencias para una variable son muy pequeñas, sería posible concluir que esta importa poco en la conformación de los grupos resultantes, por lo que podría ser omitida en ejercicios subsecuentes. Cuando la línea asociada a uno de los grupos se

encuentra generalmente por arriba de la de otro, se puede concluir que el primero presenta condiciones más favorables.

Figura 14

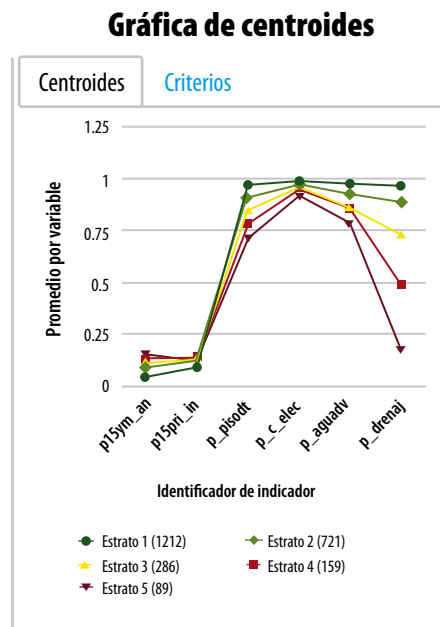
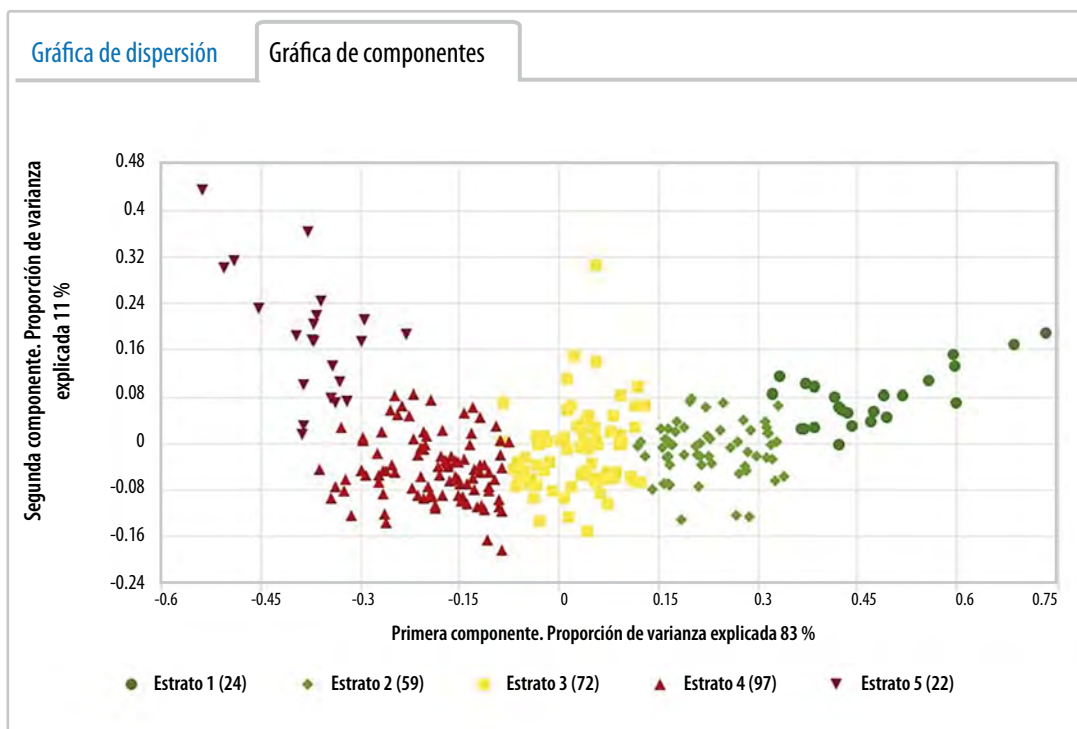


Figura 13

### Gráfica de burbujas con las dos primeras componentes principales como ejes



Se obtiene así una forma simple de lograr un ordenamiento entre los grupos a la manera del denominado Grado de Marginación, derivado del Índice de Marginación del CONAPO sin pasar por la innecesaria pérdida de información que tiene lugar con el uso de índices resumen.<sup>12</sup>

En efecto, es necesario enfatizar que los ordenamientos así obtenidos hacen un uso más eficiente de la información disponible que aquellos que parten de formar un resumen unidimensional y de aplicar a este un método univariado de estratificación, como es el caso del Índice del CONAPO (ver figura 15). De hecho, las herramientas proveídas por el servicio permiten reproducir la metodología para la construcción de este índice basada en la primera componente principal, como fue habitual hasta hace poco tiempo.

Cuando se recurre al uso de componentes principales, como hacía el CONAPO tiempo atrás, resulta evidente que ignorar la información contenida en la segunda de ellas conducirá a una estratificación menos que óptima. Algo similar ocurrirá al hacer uso de otro tipo de índices resumen, como ha sucedido recientemente.

12 Bustos, V. A. "Niveles de marginación: una estrategia multivariada de clasificación", en: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. Vol. 2, Núm.1. México, INEGI, 2011 (DE) <https://rde.inegi.org.mx/index.php/2011/01/03/niveles-de-marginacion-una-estrategia-multivariada-de-clasificacion/> consultado el 02/03/2022.

## Comparaciones

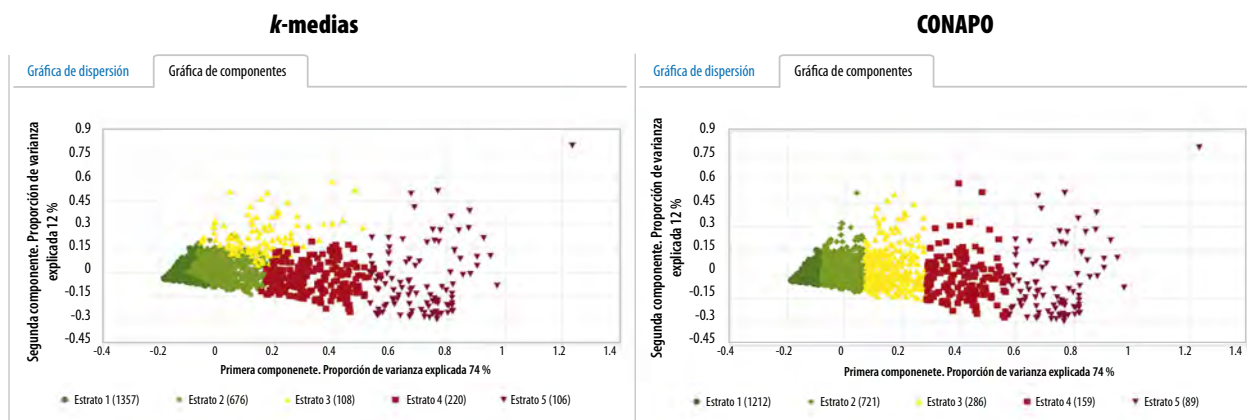
Durante una misma sesión, es posible realizar una serie de ejercicios a partir de la selección de distintas combinaciones de variables, unidades geográficas, número de grupos y método. El sistema los recuerda a todos para futura referencia (ver figura 16). De esta manera, se facilita la comparación entre resultados obtenidos bajo circunstancias diversas.

Entonces, desde las primeras gráficas es posible formarse una impresión de las ventajas relativas de algunos ejercicios. Así, los restantes podrían dejar de ser considerados desde esta etapa. Sin embargo, creemos útil tomar en cuenta toda la información que el sistema despliega antes de tomar una decisión.

Por ejemplo, para una misma localidad y utilizando las mismas variables, las manzanas fueron clasificadas en cinco, seis y siete grupos. El método de clasificación de las *k*-medias fue utilizado en los tres primeros casos; para el cuarto, se usó *mc*lust en su versión para los modelos más general. Estrictamente, solo la segunda y la cuarta estratificaciones son comparables. La representación mediante mapas muestra ya discrepancias importantes (ver figura 17).

Las gráficas de burbujas en el espacio de las dos primeras componentes principales también exhiben diferencias notables (ver figura 18). Las que hacen uso del método de las *k*-medias exhiben una

**Figura 15**  
**Estratificaciones nacionales de municipios en cinco grupos usando las mismas variables**



### Botones de recuperación de ejercicios realizados en la misma sesión

Historial de ejercicios X

- Estratificación 1
- Estratificación 2
- Estratificación 3
- Estratificación 4

**Estratificación 4**

**Variables:**

- % Población de 15 años y más analfabeta
- % Población de 15 años y más con primaria incompleta
- % Pobladores en viviendas con piso de material diferente de tierra
- % Pobladores en viviendas que disponen de luz eléctrica
- % Pobladores en viviendas que disponen de agua entubada en el ámbito de la vivienda
- % Pobladores en viviendas que disponen de drenaje

**Grupos: 5**  
 Nivel de desagregación: Nacional por municipio  
 Método: mulvar Componentes principales (1)

Ver

gradualidad a lo largo del eje asociado a la primera componente, excepto para el último grupo en el que, además, es aparente la influencia de los valores de la segunda. Destacamos que nada similar ocurre con el último ejercicio para el cual solo es apreciable el comportamiento descrito para el último grupo.

Las gráficas de centroides (ver figura 19) narran una historia semejante. Por ejemplo, las primeras cuatro variables consideradas en el estudio influyen poco o nada en la caracterización de los grupos. Entre ellas, nuevamente, la que se refiere al cuarto ejercicio es la que muestra una mayor discrepancia; sin embargo, vale la pena destacar que los grupos formados por mclust muestran entre sí diferencias pequeñas, contra lo que sería de esperarse.

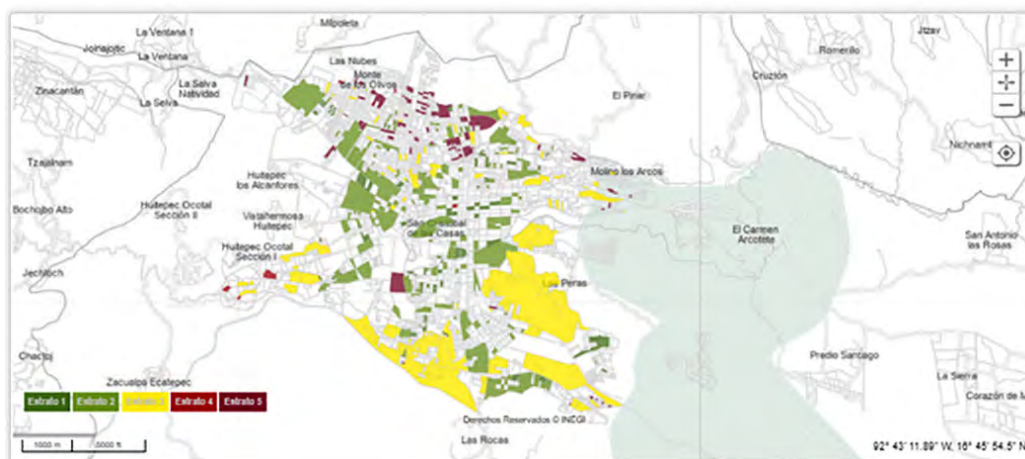
El *Estratificador INEGI* pone también a disposición del investigador un conjunto de hasta cinco criterios para fines de comparación entre los diversos ejercicios (ver cuadro 1). De este modo, se brinda un apoyo adicional para determinar el

ejercicio que muestra resultados más favorables. En todos los casos, valores pequeños para los criterios son deseables. En general, los que se refieren a determinantes de matrices indican grupos más homogéneos. En vista de la simetría de las matrices involucradas, lo mismo es cierto para los criterios basados en la traza. Sus valores han sido reescalados de modo que se encuentren siempre entre 0 y 1; se tiene así que el peor resultado, en términos de uno de ellos, toma el valor 1, en tanto que el más favorable tiene el de 0.

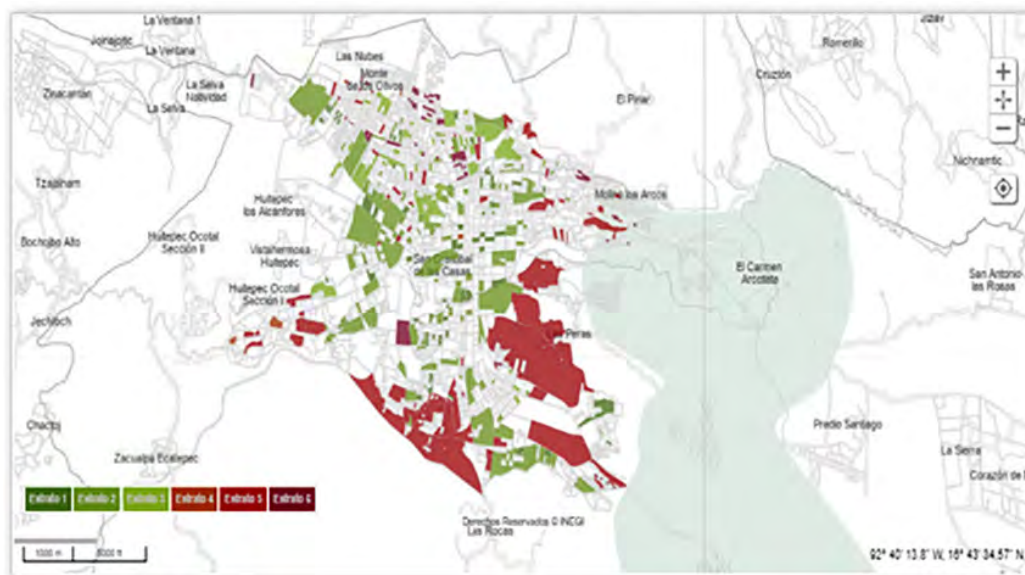
Para nuestros cuatro ejercicios, se tiene que los resultados que arrojan son mixtos (ver figura 20). Por ejemplo, el basado en mclust da los peores resultados, y con mucho, para los primeros cuatro criterios; lo opuesto ocurre para el quinto. Por su parte, de entre los otros tres, el que contempla la formación de siete grupos proporciona los mejores resultados cuando se consideran solamente los primeros tres criterios. El que se refiere a seis grupos es un cercano segundo lugar en los mismos términos, quedando el que tiene cinco siempre en

### Cuatro estratificaciones para una misma localidad

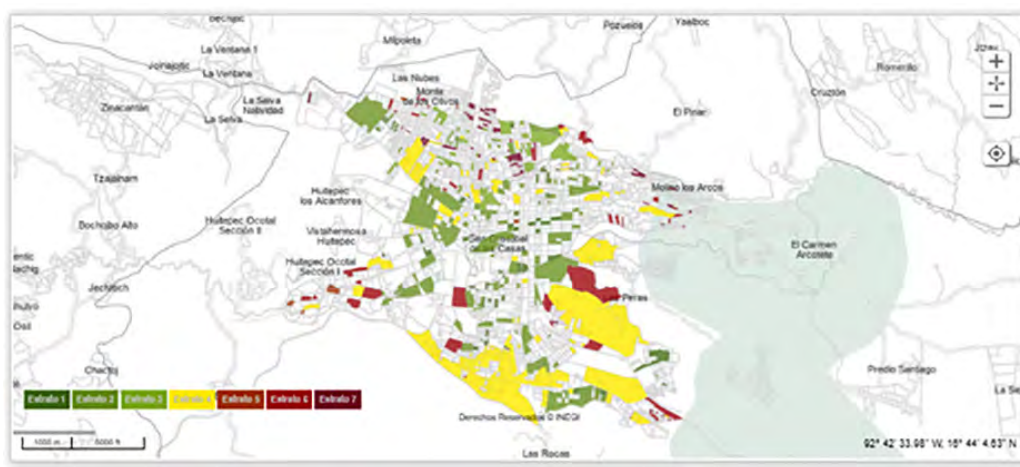
*k*-medias, cinco estratos



*k*-medias, seis estratos



*k*-medias, siete estratos



### Cuatro estratificaciones para una misma localidad

mclust(vv), seis estratos

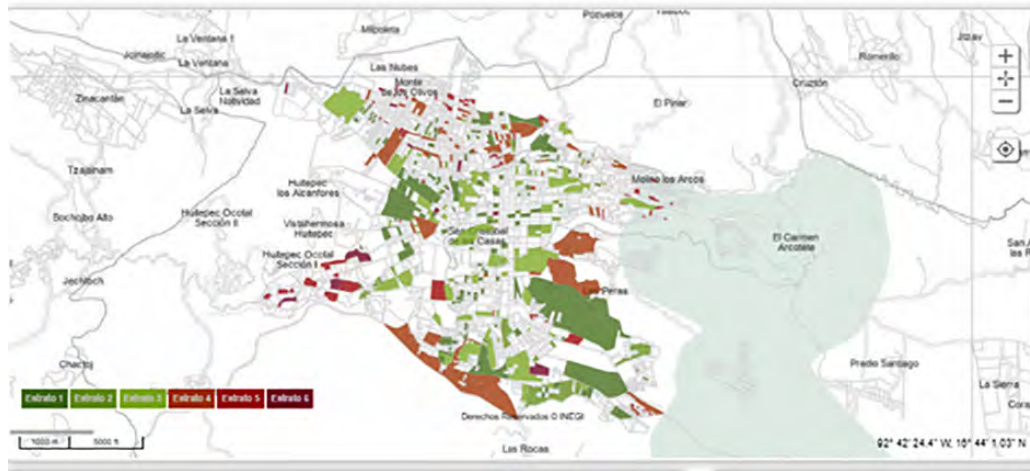


Figura 18

### Gráficas de burbujas para estratificaciones anteriores, respectivamente

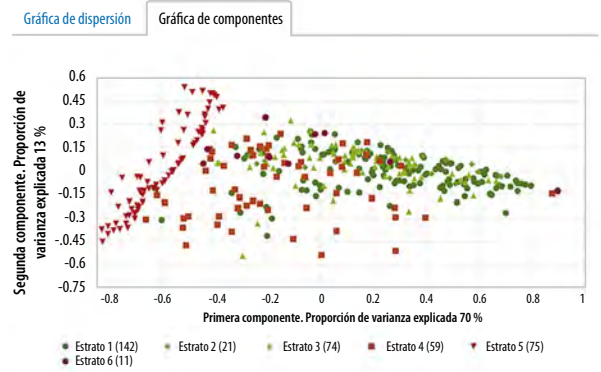
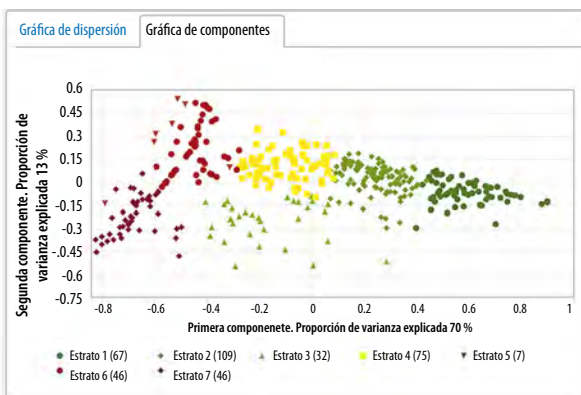
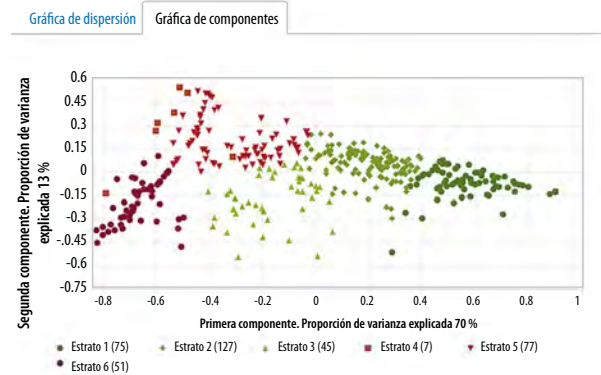
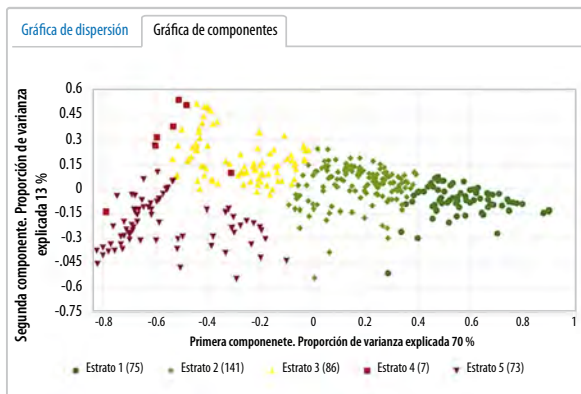
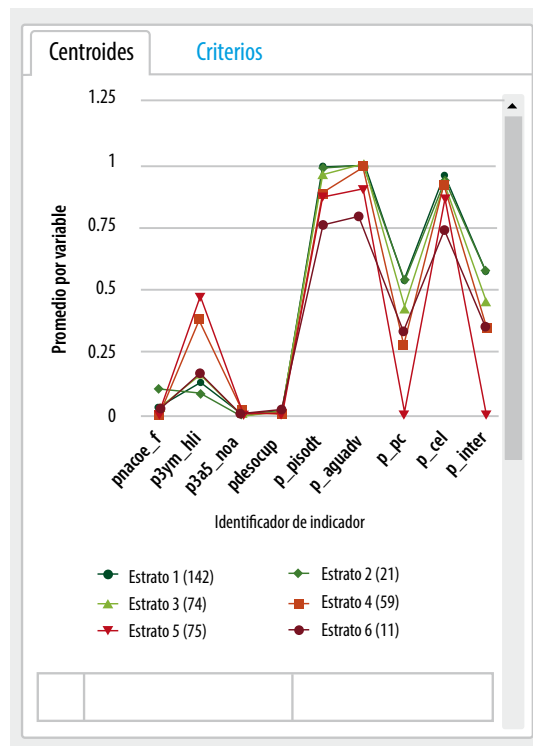
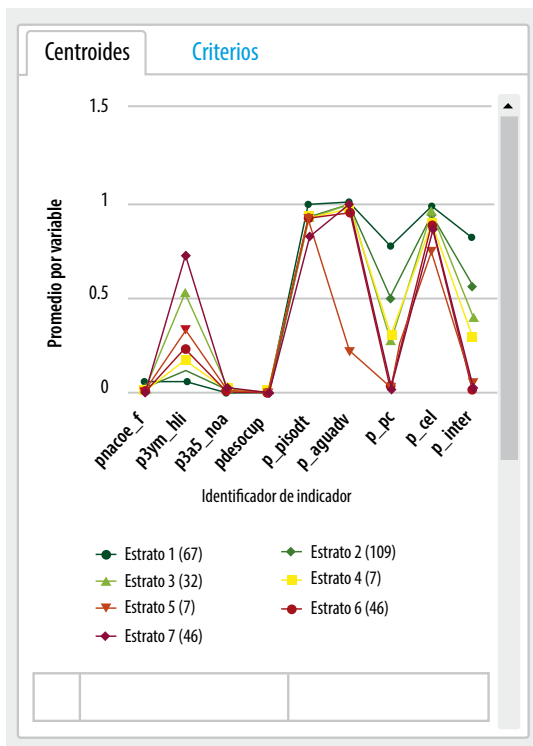
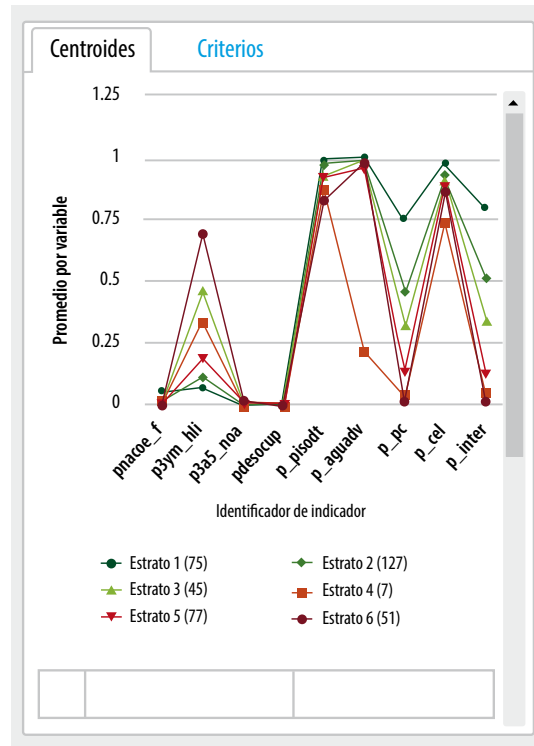
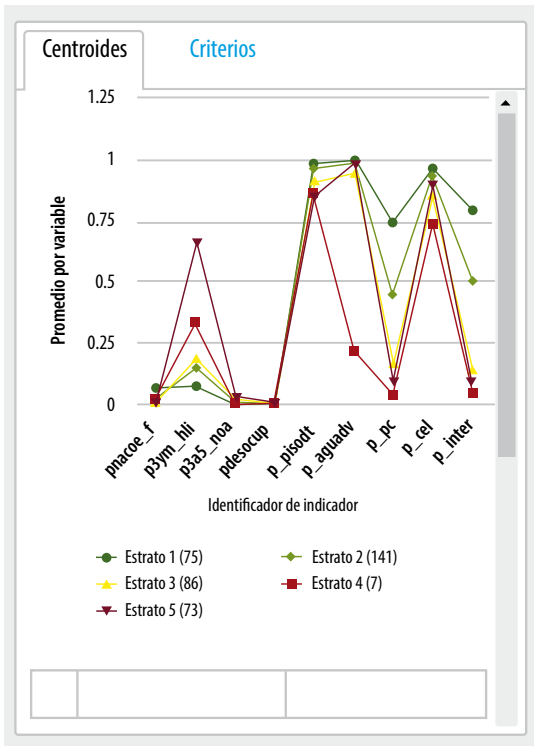


Figura 19

### Gráficas de centroides

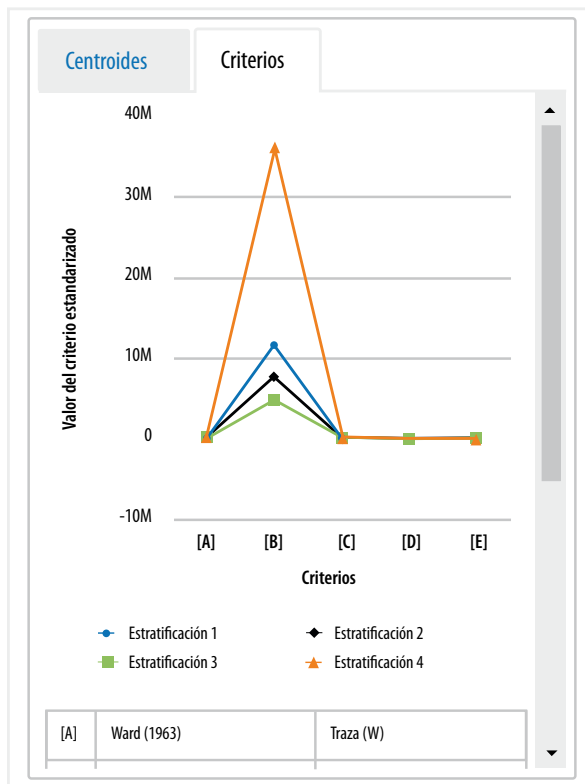


### Criterios útiles en la evaluación de resultados de ejercicios de estratificación

Criterios	
Determinante de la suma de las matrices locales de sumas de cuadrados y productos cruzados (Ward)	$=  W $
Suma de cuadrados de las distancias euclidianas de cada punto al centroide de su grupo (SC)	$= tr(W) = \sum_{i=1}^h w_{ii}$
Suma ponderada del logaritmo de trazas de matrices locales de covarianzas (SLT)	$= \sum_{k=1}^K n_k \log (tr(\Sigma_k))$
Suma ponderada del logaritmo de determinantes de matrices locales de covarianzas (SLD)	$= \sum_{k=1}^K n_k \log ( \Sigma_k ) = \log \left( \prod_{k=1}^K  \Sigma_k ^{n_k} \right)$
Promedio ponderado de las relaciones entre determinantes locales y globales (DEff)	$= \sum_{k=1}^K \left( \frac{n_k}{n} \right) \frac{ \Sigma_k }{ \Sigma }$

Figura 20

#### Gráfica de criterios según ejercicio de estratificación



tercer lugar. Lo anterior se invierte cuando se toma en cuenta al cuarto criterio. El último de los criterios proporcionados podría modificar las impresiones alcanzadas hasta ahora, pues los ejercicios con cinco y seis grupos se alejan del resultado óptimo, en tanto que el de siete alcanza un decoroso segundo lugar.

El investigador tiene ahora a su disposición diversas estrategias basadas en la información anterior. Por ejemplo, puede decidir que centrará su atención en solamente uno de los criterios, pues es aquel cuyo uso le parece más adecuado. Por otro lado, ante la incertidumbre, podrá optar por el que mostró un mejor comportamiento para el mayor número de casos. Como se indicó, el servicio no toma una decisión con base en las anteriores comparaciones, sino que permite al usuario actuar como lo considere más conveniente.

Una vez que el investigador decide cuáles resultados conservar, puede exportar un archivo en formato de valores separados por comas (CSV), que incluirá la identificación de cada una



de las unidades incluidas en el estudio, así como el indicador numérico del grupo al cual pertenecen estas; es decir, la base de datos para los resultados no incluye a ninguna de las variables utilizadas en el estudio, con lo que nos acercamos a preservar la confidencialidad de la información (cuadro 2).

## Trabajo futuro

Con la publicación de los resultados del CPV 2020 se tomó la decisión de enriquecer la base de datos

del *Estratificador INEGI*, a la vez que se actualizarán los estándares de programación para hacerlos coincidir con los que actualmente están vigentes en el sitio del Instituto. Para esta nueva versión, no se tiene contemplado añadir nuevas facilidades.

Será más adelante cuando se incorporen utilidades adicionales sugeridas en algún momento por el público usuario. Entre estas destaca, en particular, la facilidad de añadir bases de datos en poder de los usuarios. Múltiples estudios sobre inseguridad, salud o educación se verían beneficiados aprovechando sinergias entre fuentes de información.

Cuadro 2

Continúa

### Vista parcial de la tabla de resultados exportada

	A	B	C	D	E
1	Estado	Municipio	Clave geográfica	Estado	
2	Baja California ▶	Ensenada	2001	1	
3	Baja California ▶	Mexicali	2002	1	
4	Baja California ▶	Tecate	2003	1	
5	Baja California ▶	Tijuana	2004	1	
6	Baja California ▶	Playas de Rosarito ▶	2005	1	
7	Coahuila de Zaragoza ▶	Abasolo	5001	3	
8	Coahuila de Zaragoza ▶	Acuña	5002	2	
9	Coahuila de Zaragoza ▶	Allende	5003	1	
10	Coahuila de Zaragoza ▶	Arteaga	5004	3	
11	Coahuila de Zaragoza ▶	Candela	5005	3	
12	Coahuila de Zaragoza ▶	Castaños	5006	3	
13	Coahuila de Zaragoza ▶	Cuatro Ciénegas ▶	5007	3	
14	Coahuila de Zaragoza ▶	Escobedo	5008	3	
15	Coahuila de Zaragoza ▶	Francisco I. Madero ▶	5009	2	
16	Coahuila de Zaragoza ▶	Frontera	5010	1	
17	Coahuila de Zaragoza ▶	General Cepeda ▶	5011	4	
18	Coahuila de Zaragoza ▶	Guerrero	5012	2	
19	Coahuila de Zaragoza ▶	Hidalgo	5013	2	
20	Coahuila de Zaragoza ▶	Jiménez	5014	2	

## Vista parcial de la tabla de resultados exportada

	A	B	C	D	E
21	Coahuila de Zaragoza	▶ Juárez	5015	3	
22	Coahuila de Zaragoza	▶ Lamadrid	5016	3	
23	Coahuila de Zaragoza	▶ Matamoros	5017	3	
24	Coahuila de Zaragoza	▶ Monclova	5018	1	
25	Coahuila de Zaragoza	▶ Morelos	5019	3	
26	Coahuila de Zaragoza	▶ Múzquiz	5020	3	
27	Coahuila de Zaragoza	▶ Nadadores	5021	2	
28	Coahuila de Zaragoza	▶ Nava	5022	1	
29	Coahuila de Zaragoza	▶ Ocampo	5023	3	
30	Coahuila de Zaragoza	▶ Parras	5024	3	
31	Coahuila de Zaragoza	▶ Piedras Negras ▶	5025	1	
32	Coahuila de Zaragoza	▶ Progreso	5026	3	
33	Coahuila de Zaragoza	▶ Ramos Arzipe ▶	5027	1	
34	Coahuila de Zaragoza	▶ Sabinas	5028	1	
35	Coahuila de Zaragoza	▶ Sacramento	5029	2	
36	Coahuila de Zaragoza	▶ Saltillo	5030	1	
37	Coahuila de Zaragoza	▶ San Buenaventura ▶	5031	1	
38	Coahuila de Zaragoza	▶ San Juan de Sabinas ▶	5032	1	
39	Coahuila de Zaragoza	▶ San Pedro	5033	3	
40	Coahuila de Zaragoza	▶ Sierra Mojada	5034	3	
41	Coahuila de Zaragoza	▶ Torreón	5035	1	
42	Coahuila de Zaragoza	▶ Viesca	5036	4	
43	Coahuila de Zaragoza	▶ Villa Unión	5037	3	
44	Coahuila de Zaragoza	▶ Zaragoza	5038	3	
45	Chihuahua	Ahumada	8001	2	
46	Chihuahua	Aldama	8002	1	
47	Chihuahua	Allende	8003	3	