Caracterización geográfica de la vivienda en México: un enfoque de Ciencia de Datos

Jacobo Gerardo González León[*]
Miguel Félix Mata Rivera[*]

Recibido 16 de octubre de 2018, aceptado 09 de noviembre de 2018

Resumen

Una tendencia global en los primeros años del siglo XXI ha sido la generación de plataformas y datos abiertos por parte de entidades gubernamentales para la consulta pública. Estos datos permiten a los ciudadanos generar nuevas técnicas y mecanismos para su análisis con la intención de lograr mejoras en el sistema político, social y económico de un país. En esta investigación, proponemos una metodología de Ciencia de Datos centrada en la caracterización geográfica de la vivienda mexicana, utilizando datos abiertos del período 2014-2017 y aplicando técnicas de aprendizaje automatizado no supervisado. La aplicación de esta metodología permitió descubrir, que son 7 los tipos de viviendas en México, todos ellos distribuidos a lo largo y ancho del territorio mexicano. Los resultados son desplegados en mapas que podrán ser de gran utilidad en ciencia aplicada, para el estudio de fenómenos humanos y geográficos en diferentes dominios del conocimiento, tales como: seguridad, transporte, salud, economía, industria, agricultura, entre otras.

Palabras clave: *caracterización, vivienda, clusterización, análisis de datos.*

Resumo

Uma tendência global nos primeiros anos do século XXI tem sido a geração de plataformas e dados abertos por entidades governamentais visando a consulta pública. Esses dados permitem que os cidadãos criem novas técnicas e mecanismos para sua análise com a intenção de alcançar melhorias no sistema político, social e econômico de um país. Nesta pesquisa, propomos uma metodologia de Ciência de dados

[*] Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional (IPN), Unidad Avenida Instituto Politécnico Nacional núm. 2580, Col. Barrio La Laguna Ticomán, 07340 Gustavo A. Madero, Ciudad de México, México, correos electrónicos: jgonzalezl1007@alumno.ipn.mx, mmatar@ipn.mx

focada na caracterização geográfica de casas mexicanas, usando dados abertos do período 2014-2017 e aplicando técnicas automatizadas de aprendizado não supervisionado. A aplicação desta metodologia permitiu descobrir que existem sete tipos de casas no México; todos eles estão distribuídos em todo o território mexicano. Os resultados são exibidos em mapas que serão muito úteis em ciência aplicada, para o estudo de fenômenos humanos e geográficos em diferentes domínios do conhecimento, tais como segurança, transporte, saúde, economia, indústria, agricultura, entre outros.

Palavras-chave: *caracterização, habitação, agrupamento, análise de dados.*

Abstract

A global trend in the first years of the 21st century has been the generation of platforms and open data by government entities for public consultation. These data allow citizens to generate new techniques and mechanisms for their analysis with the intention of achieving improvements in the political, social and economic system of a country. In this research, we propose a Data Science methodology focused on the geographical characterization of Mexican houses, using open data from the period 2014-2017 and applying unsupervised automated learning techniques. The application of this methodology allowed discovering, that there are seven types of houses in Mexico; all of them are distributed throughout the Mexican territory. The results are displayed on maps that will be very useful in applied science, for the study of human and geographical phenomena in different domains of knowledge, such as security, transport, health, economy, industry, agriculture, among others.

Key words: *characterization, housing, clustering, data analysis*.

Introducción

Hoy en día, los Sistemas de Información recuperan y gestionan datos de muchas actividades humanas. Estos datos se han ido acumulando en grandes volúmenes a lo largo del tiempo. En estos almacenes de datos, no sólo se ha guardado información, también está contenida la historia, ya que los datos tienen la propiedad intrínseca de registrar consigo el paso del tiempo. Simultáneamente a esta acumulación de datos, también han surgido distintas maneras de procesarlos para obtener utilidad de ellos. Un ejemplo es el enfoque estadístico, que basándose en la aplicación de funciones de resumen ha permitido inferir comportamientos sobre los mismos, permitiendo examinar estos volúmenes de datos y encontrar información. En este enfoque resulta común buscar el ajuste de estas características (por ejemplo, considere una encuesta con la siguiente pregunta: ¿Cuántas habitaciones tiene su casa?, si alguien decide no responder esa pregunta, al final se promedia o se ajusta entre las que si se respondieron en la encuesta, para poder desplegar el resultado) a distribuciones

probabilísticas (Yang y Jargowsky, 2005). Sin embargo, hoy el enfoque ha cambiado, ya que no se busca encontrar un modelo estadístico que se ajuste a los datos, sino todo lo contrario, se busca analizar todo el conjunto de datos para poder conocer el modelo que mejor describe los mismos. Estas capacidades sin duda alguna, nos están llevando a una cuarta revolución industrial donde los datos son la materia prima (Oguro, 2016).

No obstante, los datos por sí solos carecen de significado, por lo que el problema continúa siendo ¿cómo tratar los datos? El objetivo entonces no ha cambiado y se centra en la extracción de información. En el contexto actual se habla de la Ciencia de Datos, entendida como un conjunto de técnicas que combina el procesamiento de datos, estadística aplicada, algoritmos de aprendizaje automatizado, arquitecturas distribuidas de servidores, análisis exploratorio, entre muchas otras más, y que su incorporación en cualquier área del conocimiento podría producir mejoras sustanciales (Gibert, Horsburgh, Athanasiadis y Holmes, 2018). Este enfoque, que últimamente está en *vogue*, no resulta tan nuevo como parece, puesto que tiene su primer registro en el año 1962, cuando el estadista John Tukey en su libro "*The Future of Data Analysis*" invita a la comunidad estadista a involucrarse en acelerar el desarrollo computacional, aplicar sus teoremas y generar estándares y medidas de validación, que sirvan como técnicas de interpretación, para entender y visualizar a los datos. De esta manera, según Tukey, se obtendrían resultados más precisos y certeros (Donoho, 2017).

La incorporación de la Ciencia de Datos en el estudio de la Geografía podría ser útil en el desarrollo futuro de esta disciplina, pues se podrían generar nuevas metodologías y datos para su incorporación en el análisis espacial de fenómenos y actividades humanas, tales como incidencia delictiva, comercio, movilidad, industrialización, turismo, y demás.

Considerando este contexto, en este trabajo se presenta la propuesta de una metodología para caracterizar geográficamente la vivienda mexicana, esto a través del agrupamiento de la similitud de sus características. Finalmente, se presentan los resultados del análisis de los tipos de viviendas obtenidos desplegándolos en un mapa, en donde se representan los componentes espaciales y geográficos para su posterior integración en otros estudios donde, por ejemplo, se podrían responder preguntas del tipo ¿las casas del Sur de México se parecen a las del Norte, o, por lo contrario, son más parecidas a las del centro del país?

La vivienda en México

En 1990 el Instituto Nacional de Estadística y Geografía (INEGI) realizó el XI Censo General de Población y Vivienda, donde surge una investigación enfocada a describir las características físicas y socio-espaciales de las viviendas. En esta investiga-

ción se regionaliza al país en 10 zonas ordenadas por la marginación de su población, originalmente, a través de la *Geografía de la Marginación* (1983) y, posteriormente, ajustada a través de los índices y grados de marginación publicados en *Indicadores Socioeconómicos e Índice de Marginación Municipal* por el Consejo Nacional de Población (CONAPO) en 1993 (Scheingart y Solís, 1995). A partir de la conformación de estas regiones y el tamaño de la localidad, se caracterizó a la población y las viviendas de México con una muestra del 1% del censo, y se generaron índices como el de "calidad de la vivienda", "calidad de los materiales empleados en la construcción de la vivienda", "distribución de las viviendas", etc.

Recientemente, los productos desarrollados por INEGI están basados en la *Encuesta Intercensal 2015* y arrojan sólo descripción estadística de los datos crudos asociadas a variables como "promedio de ocupantes en viviendas particulares habitadas", "porcentaje de viviendas particulares habitadas con drenaje", "porcentaje de viviendas particulares habitadas con electricidad".

Ante esta disponibilidad de datos asociados con múltiples variables que permiten caracterizar la vivienda en México, surge la posibilidad de realizar un análisis, utilizando técnicas de la Ciencia de Datos, que permita identificar el panorama actual de la vivienda en el país.

Metodología de Ciencia de Datos para la caracterización geográfica de la vivienda

El desarrollo de este trabajo se ha centrado en responder a la siguiente pregunta de investigación: ¿Cómo aplicar las técnicas de Ciencia de Datos para la caracterización de las viviendas mexicanas y el descubrimiento de conocimiento geográfico? A través de esta pregunta inicial se pretende descubrir las similitudes y diferencias entre las casas del norte y sur de México .Para contestar a esta pregunta se propone el procesamiento de conjuntos de datos relacionados con la temática de este trabajo a través de la metodología propuesta en la Figura 1.

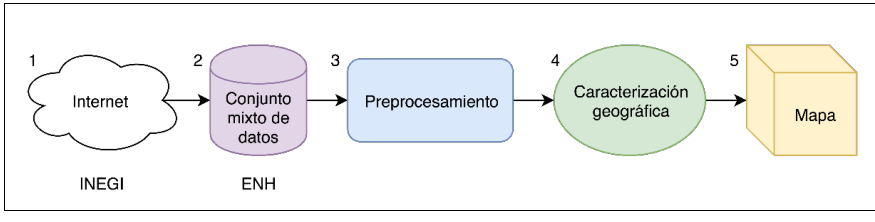

Figura 1. Metodología propuesta para la caracterización de un conjunto de datos mixto proveniente de encuestas.

La implementación de esta metodología permite que los datos estructurados sean desplegados geográficamente a través de las técnicas de aprendizaje no supervisado, lo que va a permitir, por ejemplo, identificar cuales casas de la República mexicana no tienen drenaje y conocer su distribución espacial.

La metodología consiste de 5 etapas: 1) extraer datos del portal del INEGI, 2) integrar los datos del INEGI con los datos de la Encuesta Nacional del hogar (ENH), 3) preprocesarlos para limpiarlos y adecuarlos para poder usarlos, 4) realizar una caracterización geográfica, lo que implica utilizar un algoritmo de clasificación no supervisada, ya que a priori no se conocen cuantos grupos o clases de viviendas hay en México. Finalmente, 5) visualizar los grupos identificados en un mapa.

Fuente de datos

En este trabajo se trataron datos obtenidos del repositorio de datos abiertos que publica el INEGI <www.inegi.com.mx> sobre la Encuesta Nacional del Hogar (ENH)[1] y comprenden el periodo 2014-2017. Esta encuesta está organizada en una base de datos por cada año y está compuesta por tres tablas con características de personas, hogares y viviendas, organizadas como sigue:

- Personas: contiene características sociodemográficas y ocupacionales de los integrantes del hogar, entre ellas, edad, sexo, parentesco, asistencia escolar, estado conyugal, tipo de trabajo, etc.
- Hogar: contiene características de los hogares que habitan los integrantes, entre ellas, número de integrantes del hogar, si se tiene trabajador doméstico, niñero o cuidador de ancianos, etc.
- Vivienda: contiene características de las viviendas que habitan los integrantes de los hogares encuestados, entre ellas, tipo de vivienda, material de paredes, disponibilidad de agua, etc.

Estas tablas se concatenaron en un solo conjunto de datos, el cual se denomina mixto (contiene datos de tipos numérico y de tipo categórico) donde el tamaño de la muestra es de 64 mil viviendas por cada año, siendo un total de 256 mil viviendas. La información recopilada de las encuestas abarca la siguiente cobertura temática: educación, salud, vivienda, ocupación y tecnologías de la información.

Conjunto mixto de datos obtenido de encuestas

El conjunto de datos provenientes de encuestas del INEGI suele ser de naturaleza mixto, es decir, contiene datos de tipo numérico, de tipo categórico, así como valores cuantitativos para expresar cantidades y valores cualitativos para expresar respuestas a las preguntas que hace el encuestador. Entre las preguntas que recogen

[1] <http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/enh/2014/default.html>.

estas encuestas se pueden encontrar, por ejemplo: si la vivienda dispone de tinaco en la azotea, si cuenta con cocina, o la forma en que se abastece de agua.

Como se mencionó con anterioridad, en este trabajo se realiza una concatenación de las tablas recopiladas, por medio de las claves de entidad, municipio y localidad. Esto permitió generar como resultado un conjunto de datos de 254,928 observaciones y 200 variables. A modo de ejemplo, en la Tabla 1 se muestran algunas de las diversas variables consideradas en este trabajo.

Tabla 1
Muestra de algunas variables del conjunto mixto de datos

Variable	*Etiqueta*	*Tipo*
mat_pared	Material de paredes	Categórico
antiguedad	Antigüedad de la vivienda	Numérico
num_cuarto	Número de cuartos	Numérico
focos_inca	Número de focos incandescente	Numérico
tinaco_azo	Dispone de tinaco	Categórico
…	…	…

Preprocesamiento

En la etapa de preprocesamiento se realizan 4 tareas previas a la aplicación de técnicas vinculadas a la Ciencia de Datos y que van a permitir obtener resultados de calidad durante el proceso de análisis. Estas tareas se centran en limpieza, rellenado, transformación y agrupación de datos.

Con respecto a la limpieza, esta consiste en corregir los errores presentes en los datos. En el rellenado se predicen los valores faltantes en función de las variables que presentan datos completos. En la transformación se convierten los datos categóricos en numéricos para su uso en posteriores etapas. Finalmente, en la agrupación de datos se resume la información en las localidades geográficas de interés para su caracterización.

Tras la aplicación de estas tareas, se obtuvo como resultado 5,640 localidades, producto de la combinación de entidades, municipios y localidades en esta etapa de preprocesamiento. Este dato pone de manifiesto que se trabaja con una muestra de datos, ya que según el Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades[2] se tienen 304, 221 localidades en el territorio mexicano, pero la ENH sólo se cuenta con la información de 5, 640 localidades, o sea, que, a partir de esta encuesta, sólo se podría caracterizar el 1.8% del total de localidades.

[2] <http://www.inegi.org.mx/geo/contenidos/geoestadistica/catalogoclaves.aspx>.

Caracterización geográfica

Para proceder a la caracterización geográfica de los tipos de viviendas que existen en toda una región se necesita encontrar alguna técnica para generar esta categorización. La técnica que se adopta en este trabajo es el Aprendizaje Automatizado No Supervisado a través de la Clusterización (*clustering,* en inglés). Esta técnica permite agrupar los datos en función de la similitud de sus variables (por ejemplo: qué tanto se parece la variable X a la variable Y). Para ello, se utilizó el método *k-means* (Hartigan & Wong, 1979), un algoritmo de partición que divide los datos en acumulaciones de puntos (clústeres) utilizando la distancia euclidiana. De tal modo, que las distancias entre observaciones dentro de cada clúster queden minimizadas, de tal forma que así se puede saber si un punto pertenece a un clúster o no. La dificultad estriba en que, además, debemos saber cuántos grupos de puntos se pueden formar (cuál es el número óptimo de clústeres para todos los datos que se están analizando).

Para encontrar el valor *k* clústeres (un parámetro que indica el número óptimo de clústeres, en el cual se debe dividir todo un conjunto de datos), se eligió calcularlo usando el máximo *índice de Dunn*, que se define como la razón entre la distancia más pequeña entre todas las observaciones y la distancia máxima en el clúster, también llamada el diámetro del clúster. La utilización de este índice será útil para hallar la mejor separación (agrupamiento) de los tipos de viviendas presentes en los datos, con lo cual se va a conocer si una vivienda pertenece a una cierta clase (grupo), las cuales están representadas en observaciones (datos de encuesta) (Desgraupes, 2017).

Tras ello, se ejecutó una heurística del *índice Dunn* estableciendo un rango de dos a 15 clústeres. Esto permite encontrar el valor máximo de agrupaciones, el cual se alcanza en este caso cuando existen siete clústeres, es decir, que conforme a todos los datos de viviendas considerados se detectan siete grupos con características comunes, por lo que se puede afirmar que, en términos simples, todas las viviendas analizadas se agrupan en siete tipos (véase Figura 2).

Por otra parte, para visualizar el conjunto de datos clusterizado en el espacio se requeriría de un sistema de coordenadas de 61 dimensiones, ya que ese es el número de variables que tienen las viviendas de acuerdo a la encuesta del INEGI. Sin embargo, nuestros actuales métodos de visualización no nos permitirán apreciar estas dimensiones, por lo cual, se hizo una reducción de dimensiones, aplicando Análisis de Componentes Principales (PCA, en sus siglas en inglés). Con este método (véase Figura 3) se redujo de 61 a 2 dimensiones, por medio de un cambio de coordenadas en función de los componentes no correlacionados con mayor varianza. Así, en este conjunto de datos los dos componentes que se buscan están situados en las dimensiones 1 y 2, que representan el 19.9% y el 5.6% de la varianza total.

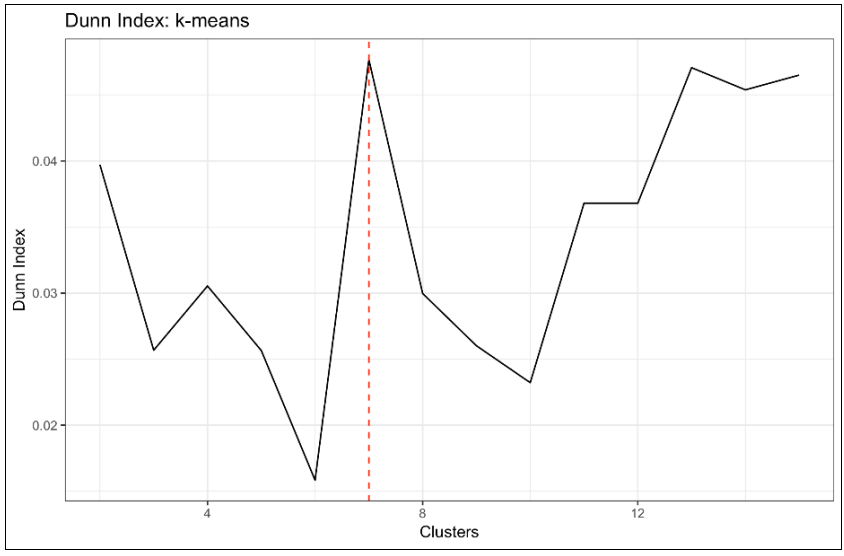

Figura 2. El índice Dunn es una medida para elegir la mejor partición en función de la cohesión de clústeres.

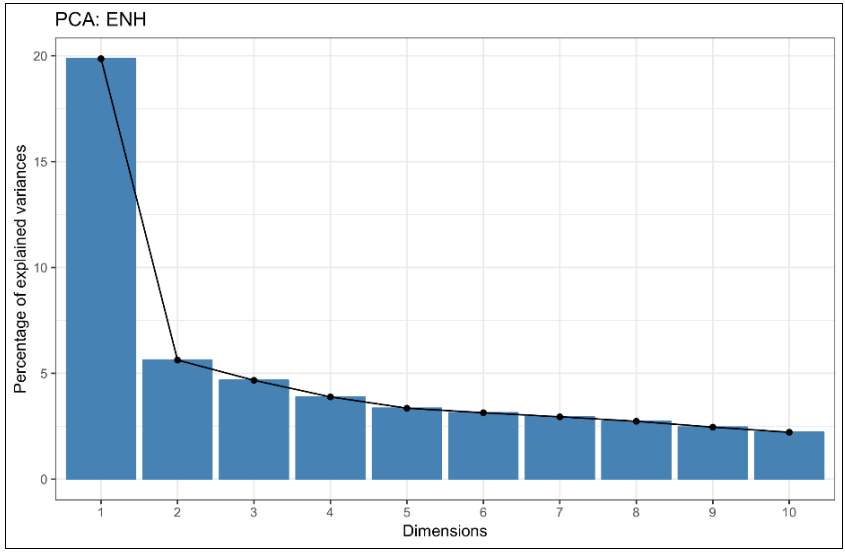

Figura 3. Análisis de Componentes Principales para el conjunto mixto de datos.

Al reducir el número de dimensiones, el resultado de la clusterización del algoritmo *k-means* con los siete clústeres encontrados con el *índice Dunn* permite generar la visualización recogida en la Figura 4, donde se pueden presentar las diferentes localidades, representadas por puntos, contenidas en estos siete clústeres deformas curvilíneas. Los clústeres son de naturaleza circular, por la *distancia euclidiana* que utiliza el método elegido. Esto se puede verificar en la forma de las agrupaciones, donde la mejor muestra está asociada con el clúster 3, que es prácticamente redondo (es decir, no es irregular). El objetivo de clusterizar los datos considerados es encontrar la estructura y organización de los datos, que a simple vista es difícil localizar, sobre todo, en altas dimensiones como es este caso. De esta manera, los clústeres conformados se interpretan como los tipos de viviendas encontrados a través del proceso de la caracterización en función de sus atributos.

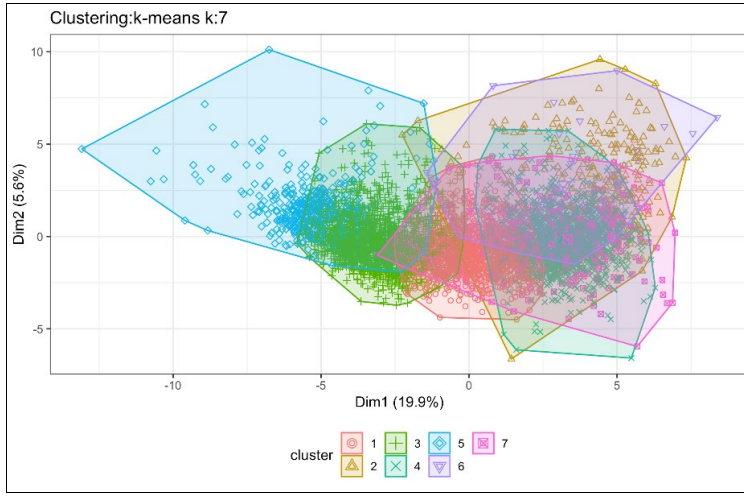

Figura 4. Los 7 clústeres encontrados por medio del algoritmo k-means.

Por lo tanto, al conjunto de datos asociados a la encuesta del INEGI, que contiene las localidades y sus características, se le puede agregar la columna del tipo de vivienda, resultado de la clusterización realizada en este trabajo. La distribución de estas localidades se encuentra en la Tabla 2, donde se pueden ver las 5, 640 localidades que poseen información y algunas variables a modo de ejemplo de la caracterización realizada. Con propósitos ilustrativos y para comprender qué características tienen en común cada clúster conformado se mencionan solo algunos de los atribu-

tos (por ejemplo, el tipo de vivienda 3 es un grupo que tiene: 4 habitaciones, drenaje de red pública, internet, entre otros aspectos).

Tabla 2
Distribución de las 5,640 localidades en los 7 tipos

Tipo	*Localidades*	*Antigüedad*	*Núm. de Habs.*	*Drenaje*	*Internet*	*Computadora*	*…*
3	1,664	23	4	Red pública	Sí	Sí	…
4	1,624	16	3	Tubería que va a dar a una barranca o grieta	Sí	Sí	
1	1,328	19	3	Red pública	Sí	Sí	…
5	441	20	5	Red pública	No	No	…
7	283	16	3	Tubería que va a dar a una barranca o grieta	Sí	Sí	…
2	246	18	3	Una tubería que va a dar a un río, lago o mar	Sí	Sí	…
6	54	15	2	Tubería que va a dar a una barranca o grieta	Sí	Sí	…

Visualización

Resaltando que el objetivo de esta investigación es encontrar la caracterización geográfica de las viviendas mexicanas (y agruparlas), se procede a la visualización de estos grupos o clústeres. Para ello, se realiza una visualización geográfica de la distribución de los tipos de vivienda en el país, presente en la Figura 5. En dicha representación se pueden visualizar algunos hallazgos interesantes. Un ejemplo de esto, es el hecho de que en el centro del país se observa cómo se concentran la mayoría de las casas del tipo 1 y 3, las cuales cuentan con alrededor de 20 años de antigüedad, mientras en la distribución del centro hacia el Sur del país predomina el tipo 7, donde hay el mayor número de cuartos en las viviendas, y el tipo 4, donde están presentes las viviendas que cuentan con drenaje a través de una tubería con destino a una barranca o grieta.

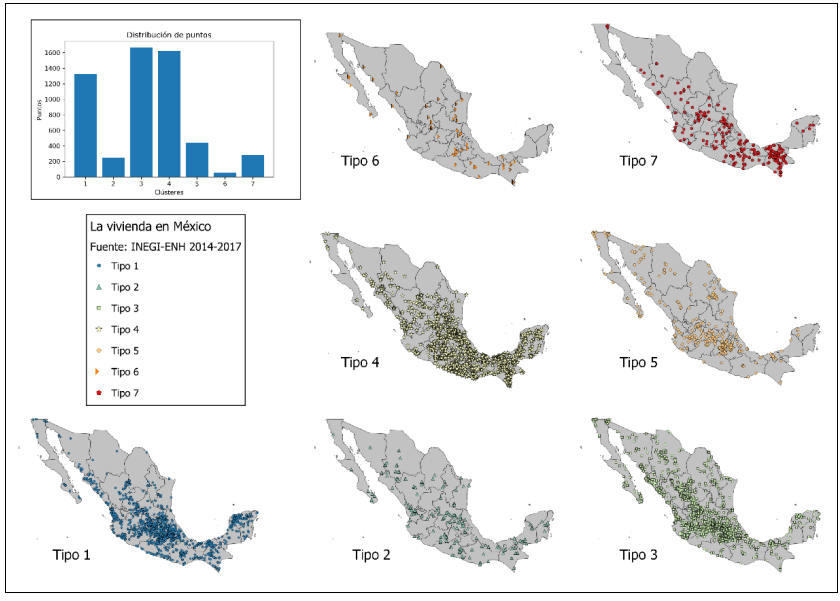

Figura 5. En este gráfico se muestran los 7 tipos de viviendas, así como su distribución de las localidades, encontrados a través de la clusterización de la ENH.

Conclusiones

La implementación de la metodología propuesta permitió encontrar la caracterización geográfica de las viviendas mexicanas, para el periodo 2014-2017, con base en la similitud de sus variables. A partir de este trabajo y la conformación de localidades agrupadas en clústeres se podría realizar un estudio más detallado del estado actual de las viviendas en México. Este estudio, por ejemplo, podría ir orientado a la detección de la brecha digital entre localidades, ya que al evaluar las viviendas del clúster 5, localizadas en la región Norte en frontera con USA, pese a que estas viviendas tienen características similares a las del centro del país, las mismas se caracterizan por no poseer computadora e Internet, por tanto, ¿serían estas las localidades más rezagadas tecnológicamente?

El aporte de este trabajo de investigación se centra en la implementación de técnicas de Ciencia de Datos mediante el preprocesamiento de los datos y Aprendizaje Automatizado No Supervisado, que permiten la agrupación de las localidades para la generación de información geográfica y su representación gráfica a través de mapas.

Con respecto al trabajo futuro, este consiste en caracterizar otras variables socioculturales y generar más capas de información para utilizar operaciones espacia-

les que permitan estudiar fenómenos humanos, tales como: el comercio, movilidad, delincuencia, entre otros.

Bibliografía

Desgraupes, B. (2017). *Clustering Indices.* Lab Modal'X. Paris, Francia: University Paris Ouest.

Donoho, D. (2017). "50 Years of Data Science", *Journal of Computational and Graphical Statistics*, 26, 749.

Gibert, K.; Horsburgh, J.S.; Athanasiadis, I.N. and Holmes, G. (2018). "Environmental Data Science", *Environmental Modelling & Software*, 106, 4-12.

Gold, J. (2009). "Behavioral Geography", en R. Kitchin and N. Thrift, *International Encyclopedia of Human Geography*, Elsevier Science, 282-293.

Hartigan, J. A. and Wong, M.A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm", *Journal of the Royal Statistical Society*, 28(1), 100-108.

Oguro, K. (2016). Big data–key to the 4th industrial revolution. *Japan SPOTLIGHT*, 24-27.

Scheingart, M. and Solís, M. (1995). *Vivienda y Familia en México: Un Enfoque Socio-Espacial Tomo VIII.* Aguascalientes, México, Instituto Nacional de Estadística Geografía e Informática.

Yang, R. and Jargowsky, P.A. (2005). "Descriptive and Inferential Statistics", en K. Kempf-Leonard, *Encyclopedia of social measurement, Volume 1*, San Diego, California: Elsevier Academic Press, 659-658.