UNIVERSIDAD DEL AZUAY

DEPARTAMENTO DE POSGRADOS

SEGMENTACIÓN DE USUARIOS POR CONSUMO DE AGUA POTABLE, MEDIANTE TÉCNICAS DE *CLUSTERING*

Trabajo de graduación previo a la obtención del título de:

Magister en Matemática Aplicada

Autor:

Estalin Fabián Mejía Hidalgo

Directora:

Ing. María Fernanda Samaniego Larriva Mgtr.

Cotutora:

Ing. María Verónica Rodas Ochoa Mgtr.

Cuenca – Ecuador

2024

AGRADECIMIENTOS

Agradezco a Dios por sus bendiciones durante mi trayectoria profesional. También quiero extender mi gratitud al grupo de investigación SWACH y a ETAPA EP por brindarme la oportunidad de llevar a cabo este estudio y facilitarme los datos necesarios para el mismo. A mis tutoras, la Ing. Fernanda Samaniego y la Ing. Verónica Rodas, por su aporte y acompañamiento durante el desarrollo de este trabajo.

ÍNDICE

ÍNDICE DE TABLAS IV

ÍNDICE DE FIGURAS IV

ÍNDICE DE ANEXOS IV

1. INTRODUCCIÓN 3

2. REVISIÓN BIBLIOGRÁFICA 4

3. METODOLOGÍA 8

3.1. Recolección de datos 8

3.2. Exploración y análisis de datos 9

3.3. Preparación de datos 10

3.4. Transformación e imputación de datos 10

3.5. Clusterización 11

4. RESULTADOS Y DISCUSIÓN 12

4.1. Preparación de datos 12

4.2. Depuración de datos 12

4.3. Usuarios totales de cada año considerados para el proceso de *clustering* 14

4.4. Selección del número de clústeres con *kmeans* 15

4.5. Distribución de puntos agrupados por *k-means* 16

4.6. Evolución *clustering* por años (Clúster 1) 18

4.7. Comparación Espacio-temporal (2009-2023) 18

5. CONCLUSIONES Y RECOMENDACIONES 19

6. BIBLIOGRAFÍA 22

7. ANEXOS 25

7.1. Análisis espacio-temporal de consumos de agua potable de la ciudad de Cuenca 25

7.2. Evolución de centroides de la clusterización por años 32

ÍNDICE DE TABLAS

Tabla 1: Algoritmos de *clustering* aplicados al consumo de agua potable. Elaboración Propia. 7

Tabla 2: Variables de la base de datos de ETAPA EP. Elaboración Propia 9

Tabla 3: Número de clústeres a partir del porcentaje de varianza explicada. Elaboración Propia 16

Tabla 4: Centroides de k-means con 6 clústeres. Elaboración Propia 17

Tabla 5: Rangos de consumo de agua potable para la ciudad de Cuenca. Elaboración Propia. 17

ÍNDICE DE FIGURAS

Figura 1: Esquema del proceso de segmentación de usuarios de consumo de agua potable. Elaboración Propia 8

Figura 2: Fragmento de la base de datos de consumo de agua potable (m3) preparada para el análisis de clustering. Elaboración Propia 12

Figura 3: Porcentaje de datos válidos y porcentaje de datos eliminados de cada año. Elaboración Propia 13

Figura 4: Porcentajes de datos eliminados. Elaboración Propia 13

Figura 5: Usuarios de agua Potable de la ciudad de Cuenca considerados aptos para el estudio. Elaboración Propia 14

Figura 6: Método del codo con k-means para evaluar el número adecuado de clústeres. Elaboración Propia. **Fuente:** (Torgo, 2011). 15

Figura 7: Distribución de los puntos asignados por clustering en 2D. Elaboración Propia 16

Figura 8: Evolución de centroides para el clúster 1 asociado al enfoque de clusterización por años. Elaboración Propia 18

Figura 9: Segmentos de mapas de la ciudad de Cuenca (2009-2023) 19

ÍNDICE DE ANEXOS

Anexo 1: Análisis Espacio-temporal de consumos de agua potable de la ciudad de Cuenca (2009-2023) 32

Anexo 2: Evolución de centroides de clustering por cada año. Elaboración Propia 34

SEGMENTACIÓN DE USUARIOS POR CONSUMO DE AGUA POTABLE, MEDIANTE TÉCNICAS DE *CLUSTERING*

Estalin Fabián Mejía Hidalgo emajiama12@es.uazuay.edu.ec
María Fernanda Samaniego Larriva mafersamaniego@azuay.edu.ec
María Verónica Rodas Ochoa vrodas@etapa.net.ec

RESUMEN

En los últimos años, la ciudad de Cuenca ha experimentado un aumento en la densidad poblacional, resaltando la necesidad de asegurar la sostenibilidad del agua. Por esta razón, la presente investigación tiene como propósito llevar a cabo la segmentación de los usuarios de agua potable de la empresa ETAPA EP en la ciudad de Cuenca, en función de sus perfiles de consumo. Para lograr este objetivo, se realizó una preparación y transformación de los registros mensuales de los clientes de agua potable desde el año 2009 al 2023. Mediante la aplicación del algoritmo *k-means*, se realizó una clusterización por cada año y una global. La primera aplicación, muestra el comportamiento individual de los grupos de consumo por año, mientras que la segunda define grupos globales para el periodo de estudio. Este enfoque permitió establecer rangos de consumo para observar la evolución espacio-temporal del uso de agua en la ciudad.

Palabras clave

Clusterización, Consumo de agua, K-means, Imputación de datos, Perfiles de consumo de agua.

ABSTRACT

In recent years, the city of Cuenca has experienced an increase in population density, highlighting the need to ensure water sustainability. For this reason, the purpose of this research is to carry out the segmentation of the drinking water users of ETAPA EP in the city of Cuenca, according to their consumption profiles. To achieve this objective, the monthly records of drinking water customers from 2009 to 2023 were prepared and transformed. By applying the k-means algorithm, a clustering was performed for each year and a global one. The first application shows the individual behavior of consumption groups by year, while the second defines global groups for the study period. This approach made it possible to establish consumption ranges to observe the spatial and temporal evolution of water use in the city.

Keywords

Clustering, Water consumption, K-means, Data imputation, Water consumption profiles.

1. INTRODUCCIÓN

En los últimos 40 años, a nivel mundial, el uso del agua ha experimentado un incremento de 1% anual, atribuible al progreso socioeconómico, aumento poblacional y a los hábitos de consumo, especialmente en países en desarrollo (UNESCO, 2023). A todos los factores mencionados anteriormente, se suma las alteraciones a las que han sido expuestos los recursos hídricos debido al cambio climático, pues este proceso afecta directamente a los procesos hidrometeorológicos e indirectamente al desarrollo productivo de la sociedad (UNESCO, 2020).

La ciudad de Cuenca, forma parte de la provincia del Azuay y es pionera en la gestión integral de recursos hídricos. Conforme al último censo realizado en el año 2022 por el Instituto Nacional Ecuatoriano de Censos (INEC), la ciudad registra una población total de 596.101 habitantes, 361.524 en el área urbana y 234.577 en el área rural (INEC, 2022). Esto supone un reto en la demanda de agua para la población a cargo de la Empresa Pública Municipal de Teléfonos, Agua Potable y Alcantarillado (ETAPA EP).

Debido a la situación actual, asegurar la cantidad y calidad de agua en el presente y futuro se ha vuelto un tema de vital importancia. Por esta razón, en el marco de cooperación interinstitucional entre universidades de Ecuador y Bélgica, la municipalidad de Cuenca y la empresa de agua potable ETAPA EP, nace el proyecto SWACH (Gestión sostenible del agua bajo cambio climático en el Sur de Ecuador). El cual se encuentra estudiando el manejo sostenible del agua para Cuenca, bajo escenarios de cambio climático (SWACH, 2023).

En esta línea, el presente estudio se enfoca en la segmentación de usuarios según sus características de consumo de agua potable, durante el periodo comprendido entre 2009-2023 mediante técnicas de *clustering*. Se analizará el comportamiento temporal y espacial de los grupos identificados en los años mencionados.

Durante el desarrollo, se llevará a cabo la exploración y transformación de datos por año, pues es necesario recalcar que los consumos han sido tomados en contadores mecánicos y registrados de forma manual, lo cual podría indicar fallas o inconsistencias en la información debido a errores mecánicos o humanos.

2. REVISIÓN BIBLIOGRÁFICA

El manejo del agua en ambientes urbanos es considerado una parte importante para lograr la adaptación al cambio climático y poder alcanzar ciudades resilientes (Herslund & Mguni, 2019). Cabe recalcar que la problemática actual y futura con relación a la cantidad y calidad del agua, no solo se debe a la variabilidad climática, si no al crecimiento demográfico y su comportamiento conductual y cultural (Retamal et al., 2011).

Según Adamala (2017), para los gestores el contar con grandes bases de datos ayuda a mejorar muchos aspectos como: la planificación de los sistemas hídricos, programas de mitigación, la contaminación ambiental, la predicción de modelos y detectar cambios en los ecosistemas. Para resaltar la importancia de estudiar patrones en los consumos de agua, el estudio de Waraga et al. (2021) indica que, explorar patrones en esta área objeto de estudio mejora la gestión eficiente del agua en entornos urbanos, satisfaciendo necesidades de consumo actual sin afectar las futuras.

Rahim et al. (2021) menciona que, una de las técnicas que ha ido adquiriendo relevancia en diferentes áreas como: educación, negocios, salud, biología y agua, es la de aprendizaje no supervisado. Este tipo de aprendizaje se emplea para extraer conocimiento de un conjunto de datos que carecen de información predefinida, siendo su principal objetivo identificar similitudes entre los datos sin una guía previa (Chander & Vijaya, 2021).

Los algoritmos de aprendizaje no supervisado, pueden ser empleados para comprender mejor los comportamientos de los usuarios de agua potable a través de modelos de agrupación, segmentando los patrones en subconjuntos en función de características semejantes entre ellos (Saxena et al., 2017). Ghamkhar et al. (2023) argumenta que en caso de no existir etiquetas en los datos, la única solución es usar aprendizaje no supervisado.

La Tabla 1 muestra información bibliográfica del uso de técnicas de *clustering* en diferentes partes del mundo, destacando temas de vital importancia como: la elección del algoritmo, la base de datos disponible y el objetivo de cada investigación.

#	Autores	Título	Zona de estudio	Algoritmo	Recolección de Datos	Descripción (datos, variables y/o características)
1	(Laspidou et al., 2015)	*Exploring patterns in water consumption by clustering*	Isla Skiathos, Grecia	Mapas de Kohonen	Contadores Mecánicos	Datos trimestrales de consumo para 2006-2013. En el artículo se estudió 168 contadores con 3 tipos de consumidores diferentes. Mapas de Kohonen, es un tipo de red neuronal que forma parte del aprendizaje no supervisado para agrupar y visualizar la información. Se usaron 5 variables a partir del consumo trimestral medio, máximo y la relación de consumos trimestrales de cada año con el consumo trimestral medio.
2	(Ioannou et al., 2017)	*Data mining for household water consumption analysis using selforganizing maps*	Sosnowiec, Polonia	Mapas de Kohonen	Contadores Inteligentes	Detección de patrones diarios en consumo doméstico, registrados cada 30 segundos por 445 días. 13 variables, que describen los consumos a lo largo del día como: consumo en la mañana, tarde, noche, medio día, madrugada, horas de trabajo, horas de descanso en oficinas, horas laborales en bancos, horas de descanso en bancos, horas de trabajo y descanso en tiendas, consumo total y desviación estándar del consumo diario
3	(Brentan et al., 2018)	*Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems*	El artículo evalúa 3 estudios de caso en ciudades brasileñas	*K-means* Mapas de Kohonen	No se especifica	El estudio considera 3 tipos de datos por separado: Calidad del agua, Red de distribución de agua y Transitorios hidráulicos. En el artículo se expresa que *K-means* puede ser más rápido que los mapas autoorganizados basados en redes neuronales, explica también que este método obtiene grupos demasiado grandes lo cual es un inconveniente en la toma de decisiones. Los mapas autoorganizados en el estudio se utilizaron como parte del preproceso.

#	Autores	Título	Zona de estudio	Algoritmo	Recolección de Datos	Descripción (datos, variables y/o características)
4	(Leitão et al., 2019)	*Detecting urban water consumption paterna: a time-series clustering approach*	Coimbra, Portugal	*K-means*	No especificado, se asume contador inteligente por frecuencia de 1h	Datos horarios de tipo residencial en un período de 1 año, tomados de la empresa gestora del agua. Los datos están sujetos a imprecisiones ya sea por lectura, comunicación o los instrumentos de medición.
5	(Rahim et al., 2021)	*A clustering solution for analyzing residential water consumption patterns*	Melbourne, Australia	*K-means* Agrupación Jerárquica	Contadores inteligentes	2 tipos de datos: a partir de ingeniería de funciones y datos de tiempo de uso y probabilidades ponderadas. Para el primer conjunto de datos se usó *K-means* con la técnica del codo y para el segundo se usó agrupación jerárquica. El estudio determina que *K-means* es la técnica con mejor rendimiento.
6	(Ioannou et al., 2021)	*Exploring the Effectiveness of Clustering Algorithms for Capturing Water Consumption Behavior at Household Level*	Ohio, EE.UU	Mapas de Kohonen *K-means* y Agrupación Jerárquica	Contadores Inteligentes (Caudalímetro Electromagnético)	El estudio se realizó en 21 hogares, durante 210 días en intervalos de media hora. En este estudio se utilizó algoritmos de mapas autoorganizados como parte del preprocesamiento, para reducir la dimensionalidad y coste computacional. El número de clústeres se eligió usando la suma de cuadrados dentro del clúster comúnmente conocido como (WCSS), mide la suma de las distancias al cuadrado entre cada punto de datos de un clúster y el centroide.
7	(Alkelani & Awad, 2021)	*K-Means Clustering Based Model for Fair Water Distribution of Urban Regions Depending on Consumption*	Jenin, Palestina	*K-means*	No se especifica	La ciudad se divide en 39 zonas, en el estudio se calculó el consumo total de cada área en un mes entre enero y noviembre del 2019. Se agruparon zonas según su consumo de agua, de esta manera se mejoró la distribución de agua en la ciudad de Jenin.

#	Autores	Título	Zona de estudio	Algoritmo	Recolección de Datos	Descripción (datos, variables y/o características)
8	(Olmedo et al., 2023)	Aplicación de algoritmos de *Machine Learning* para la segmentación del consumo de agua potable. Caso de estudio en Catamayo, Ecuador	Catamayo. Ecuador	*K-means* DBSCAN	Contadores mecánicos	Histórico de consumo mensual desde enero a noviembre de 2019, datos categorizados por tipo de consumo. Segmentación según su consumo de agua mensual de usuarios, correspondientes a la categoría de residencial. El estudio contó con 6.492 registros. Análisis de clústeres con el método del codo y de la silueta.
9	(Ghamkhar et al., 2023)	*An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity*	Irán, no se especifica las ciudades de estudio	DBSCAN	Contadores Mecánicos	Datos mensuales de consumo proporcionados por registros de la empresa de agua potable encargada. El objetivo de la investigación fue detectar usuarios finales con patrones de consumo anormal en sistemas de distribución de agua. El estudio no realizó el preprocesamiento de datos.

Tabla 1: Algoritmos de *clustering* aplicados al consumo de agua potable. Elaboración Propia.

3. METODOLOGÍA

Para segmentar usuarios de consumo de agua potable durante los años de estudio correspondientes al periodo 2009-2023, se implementaron varios procesos con el objetivo de preparar una base de datos adecuada para el análisis de *clustering* (Figura 1). Esto permitió agrupar a los usuarios de agua potable de acuerdo con sus patrones de consumo.

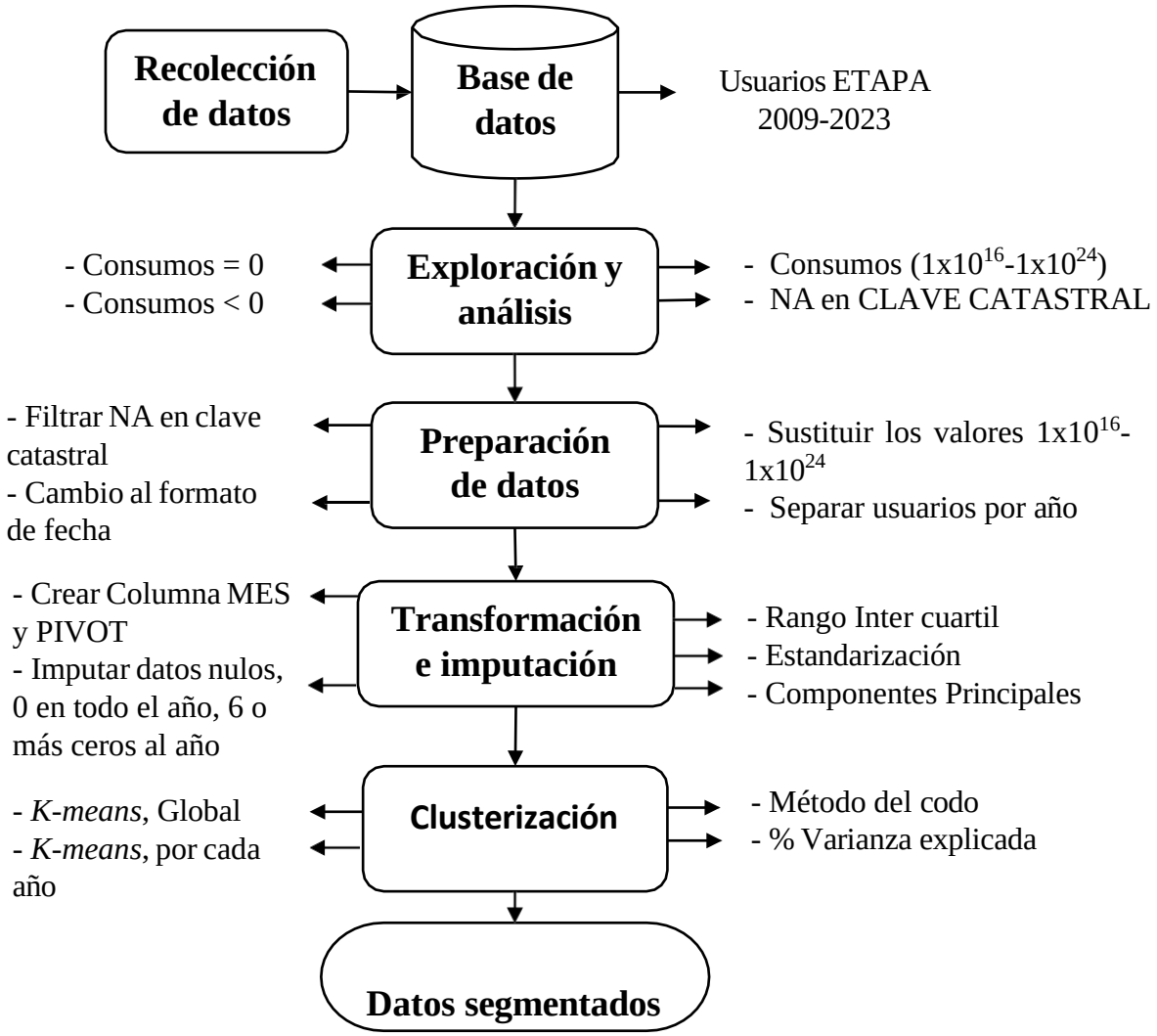


Figura 1: Esquema del proceso de segmentación de usuarios de consumo de agua potable. Elaboración Propia.

3.1. Recolección de datos

La información necesaria para el estudio fue aportada por la empresa ETAPA EP a través del convenio interinstitucional de entidades que forman parte del proyecto SWACH. Los datos corresponden a los consumos de agua potable en m^3 de los usuarios de ETAPA EP del histórico de consumos desde el año 2009 al 2023 (ETAPA EP, 2023a).

Las lecturas son tomadas mensualmente en los medidores mecánicos, ubicados en la ciudad de Cuenca de forma manual. Por tal motivo, la información es susceptible a presentar errores tanto mecánicos como humanos. El conjunto de datos aportado por ETAPA EP consta de 24.669.798 filas y 7 columnas (Tabla 2), las cuales se organizaron por fecha y meses de consumo para cada usuario para el análisis de *clustering.*

CODINS	Es el código de instalación del medidor.
CLAVE CATASTRAL	Identificador correspondiente al bien inmueble.
MEDIDOR	Es el número de instalación del medidor.
FECHA DE VALORACIÓN	Fecha en la cual se realiza la toma de datos.
LECTURA ANTERIOR	Valor de la lectura del mes anterior en el medidor (m^3).
LECTURA ACTUAL	Valor de lectura del mes actual en el medidor (m^3).
CONSUMO M3	Es la resta de la lectura actual y la lectura anterior, dando como resultado el consumo de agua potable en el mes (m^3).

Tabla 2: Variables de la base de datos de ETAPA EP. Elaboración Propia
Fuente: ETAPA EP (2023).

Para el estudio se contó con información georreferenciada de los predios de la ciudad de Cuenca. De esta forma, es posible enlazar la información de los usuarios de agua potable con el predio georreferenciado, mediante el código de la "CLAVE CATASTRAL " proporcionado por ETAPA EP en la base de datos.

3.2. Exploración y análisis de datos

Para el análisis y la exploración de información, se utilizó el software de libre acceso Jupyter Notebook (Kluyver et al., 2016). Durante el análisis de datos se localizó la presencia de valores nulos, negativos y consumos de 0 m^3. En la columna de "CLAVE CATASTRAL" se encontraron 104.654 filas de valores nulos y la columna de "FECHA DE VALORACIÓN" fue identificada en formato numérico.

En la base de datos, existía la presencia de usuarios con valores de $1x10^{16}$ a $1x10^{24}$ en sus consumos mensuales. Estos valores se asignan para indicar que en el mes se realizó un cambio de medidor, siendo necesario imputar estos valores para el análisis.

3.3. Preparación de datos

En la preparación de datos, inicialmente, se eliminaron 104.654 registros con valores faltantes de la columna “CLAVE CATASTRAL”. Este código es importante para enlazar los consumos de agua potable con la base georreferenciada de los predios de la ciudad. El formato de la columna “FECHA DE VALORACIÓN”, se cambió de numérico a fecha.

Para tratar los valores entre $1x10^{16}$ y $1x10^{24}$, debido a cambios en los medidores, se utilizó el valor de la columna “LECTURA ACTUAL”. Dado que los medidores registran el flujo de agua de manera continua, en estos casos, el valor de “LECTURA ANTERIOR” corresponde al medidor antiguo mientras que el valor de “LECTURA ACTUAL” refleja el nuevo conteo del medidor, lo que representa el consumo acumulado del mes. Posteriormente, los códigos de usuarios fueron separados por el año de estudio correspondiente.

3.4. Transformación e imputación de datos

Este paso implica una conversión, limpieza y refinación de los datos más profunda, los cuales ya han sido previamente preparados para este proceso. De esta forma, se garantiza una estructura y variables adecuadas para el análisis de *clustering*.

Para iniciar con esta fase, se creó una columna que identifica el mes de consumo de cada usuario. Se empleó esta columna para realizar una operación de pivoteo, según los meses, manteniendo la columna “CODINS” y “CLAVE CATASTRAL” para identificar los consumos mensuales con el código de instalación del medidor y el predio.

Los usuarios con presencia de 1 o 2 valores nulos fueron ajustados, sustituyendo los datos faltantes con la mediana de sus consumos normales en el año. Consecuentemente, se filtró códigos con más de 3 valores nulos, negativos y medidores con consumos de 0 m^3 en todos los meses del año de valoración.

Los medidores con consumos de cero en 6 o más meses del año fueron eliminados, pues podría tratarse de equipos nuevos o fuera de funcionamiento, lo cual no aporta información de los consumos del usuario en ese año.

Para los medidores con pocas cifras iguales a cero, se utilizó un criterio de relleno, reemplazando los ceros por la mediana de los consumos normales del usuario. Para abordar datos anómalos en los consumos mensuales de cada usuario, los cuales se originan tanto por

errores en los registros como por eventos inusuales, se empleó la técnica de rango inter cuartil (Vinutha et al., 2018). Con este proceso, se manejó los datos atípicos, sustituyéndolos por la mediana de los meses considerados normales para ese usuario.

Con la biblioteca scikit-learn (sklearn), se realizó un proceso de estandarización de los datos como paso previo para aplicar la técnica de componentes principales. Este procedimiento se desarrolló con la finalidad de reducir la dimensionalidad de los datos (Raschka et al., 2022).

3.5. Clusterización

Tras realizar una revisión bibliográfica de las diferentes técnicas de *clustering* en el consumo de agua en entornos urbanos (Tabla 1), se distinguió a *k-means* como uno de los algoritmos más utilizados en esta área, convirtiéndolo en una herramienta adecuada para el desarrollo de esta investigación. Su simplicidad y efectividad para agrupar los datos en conglomerados homogéneos como lo menciona Olmedo et al. (2023), justifican su aplicación en este estudio.

Para determinar el número adecuado de clústeres, se emplearon dos enfoques: uno analítico y otro gráfico. En el método gráfico, se aplicó el método del codo, mientras que, para el enfoque analítico, se utilizó el software de libre acceso RStudio para determinar el porcentaje de varianza entre los clústeres (Torgo, 2011). Este valor representa una medida de calidad que indica la proporción de variabilidad en los datos totales explicada por las diferencias entre los clústeres (1).

$$\%\, de\ varianza\ explicada\ = \frac{Suma\ de\ cuadrados\ intra\ cluster}{Total\ de\ suma\ cuadrados} * 100 \qquad (1)$$

Total de suma cuadrados = Suma de cuadrados Intra clúster + Suma de cuadrados inter clúster).

Se realizaron dos enfoques para la clusterización de los usuarios de agua potable, utilizando la técnica de *clustering* "*k-means"*. En el primer enfoque, se aplicó el algoritmo de *clustering* a cada año por separado; mientras que en el segundo enfoque, se clusterizó todos los años en conjunto. De esta manera, se puede obtener usuarios asignados a grupos de consumo homogéneos según los patrones de consumo de cada año, así como también grupos basados en los patrones generales de consumo a lo largo del tiempo.

4. RESULTADOS Y DISCUSIÓN

4.1. Preparación de datos

Para agrupar a los usuarios del servicio de agua potable de la ciudad de Cuenca en conglomerados homogéneos, según los perfiles de consumo; fue necesario convertir la base de datos en una estructura adecuada para clusterizar. Como se muestra en la Tabla 2, los datos proporcionada por la empresa pública ETAPA EP, no presentaban una estructura y variables idóneas para el proceso de *clustering.*

Con la ayuda de la librería pandas de Jupyter Notebook, se obtuvo una base de datos de cada año (Figura 2), apta para el análisis. En la base de datos mencionada, se identificó a los usuarios que corresponden a las filas por el código de instalación de medidor "CODINS" (Tabla 2), incluyendo sus consumos mensuales de agua en m^3 durante el año. Además, se conservó la "CLAVE CATASTRAL" (Tabla 2), para enlazar a los usuarios de agua potable con los predios georreferenciados de la ciudad de Cuenca.

Ene (m3)	Feb (m3)	Mar (m3)	Abr (m3)	May (m3)	Jun (m3)	Jul (m3)	Ago (m3)	Sep (m3)	Oct (m3)	Nov (m3)	Dic (m3)
26.00	25.00	34.00	28.27	27.00	25.00	30.00	26.00	25.00	34.00	31.00	28.00
29.00	19.45	27.00	14.00	11.00	14.00	20.00	19.00	21.00	22.00	20.00	17.00
31.00	25.00	24.00	17.00	20.00	26.64	28.00	29.00	28.00	36.00	28.00	27.00

Figura 2: Fragmento de la base de datos de consumo de agua potable (m3) preparada para el análisis de *clustering.* Elaboración Propia.
Fuente: ETAPA EP (2023).

4.2. Depuración de datos

A partir de la estructura de datos preparada para los usuarios de cada año (Figura 2), se realizó la depuración y transformación, eliminando usuarios anómalos que no aporten información para el análisis.

En la Figura 3 se observa el porcentaje de datos tomados como válidos para el análisis, los cuales representan alrededor del 80-85% del porcentaje total. Los usuarios eliminados para este análisis representan alrededor del 15 a 20% de los datos anuales totales, estos usuarios

fueron filtrados por presentar consumos anómalos a lo largo del año como: usuarios con datos nulos, códigos con consumo de cero en todo el año, 6 o más consumos con cero en todo el año y consumos negativos.

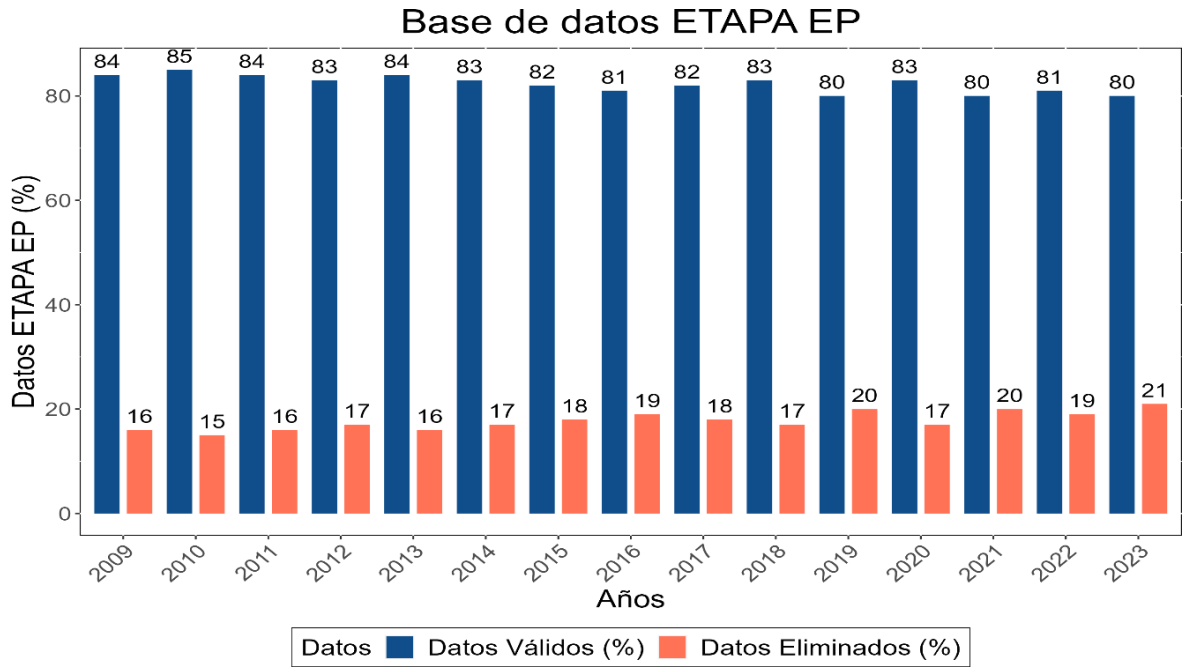


Figura 3: Porcentaje de datos válidos y porcentaje de datos eliminados de cada año. Elaboración Propia
Fuente: ETAPA EP (2023).

En la Figura 3 el porcentaje de datos eliminados en cada año ronda el 15 y 20%, estas cantidades resaltan la necesidad de conocer los diferentes aportes de valores anómalos, debido a su gran presencia en los registros a través de los años de estudio (Figura 4).

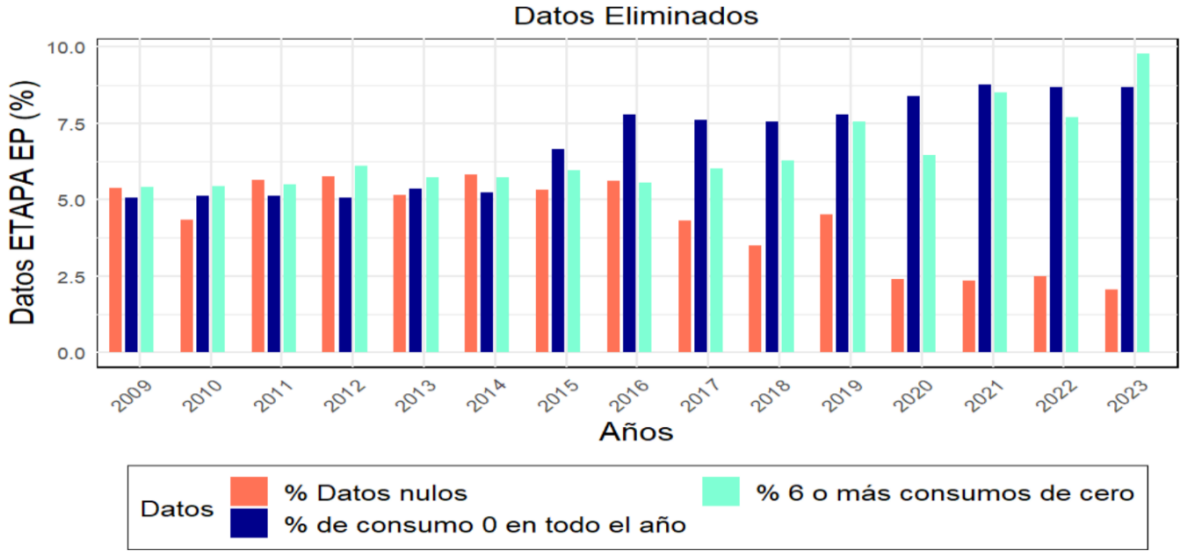


Figura 4: Porcentajes de datos eliminados. Elaboración Propia
Fuente: ETAPA EP (2023).

En la Figura 4, el porcentaje de datos más alto corresponde a códigos de usuarios con consumo de cero en todo el año a partir del año 2015. Por lo cual, se podría asumir que son medidores que han dejado de funcionar o que la vivienda esta deshabitada. Desde el año 2009 hasta el 2014 este valor representa más o menos el 5% en cada año y a partir del año 2015 al 2023 este valor oscila entre 7 al 9%.

Los usuarios con 6 o más consumos iguales a cero en el año (Figura 4), de igual forma presentan un porcentaje llamativo a través del tiempo. Con un valor entre el 5 al 10%, siendo el porcentaje más alto el año 2023. Estos valores, podrían indicar usuarios con un registro reciente, obteniendo consumos menores a 6 meses en sus consumos mensuales. De tal manera, esto impedirá que el usuario de agua potable aporte información relevante de sus patrones de consumo durante el año en cuestión.

A partir del año 2020, los valores nulos se sitúan alrededor del 2.5%, mostrando una disminución con respecto a los años anteriores. Esto podría indicar una mejora en el sistema de gestión en la toma de datos por parte de la empresa ETAPA EP en los últimos años.

4.3. Usuarios totales de cada año considerados para el proceso de *clustering*

Una vez finalizada la preparación y transformación de datos, los usuarios considerados válidos para el análisis de *clustering* rondan el 80-85% de códigos totales en cada año (Figura 3). La Figura 5 muestra el número y el aumento de medidores a lo largo del tiempo, pues la ciudad de Cuenca es una de las más grandes del país y con un crecimiento acelerado.

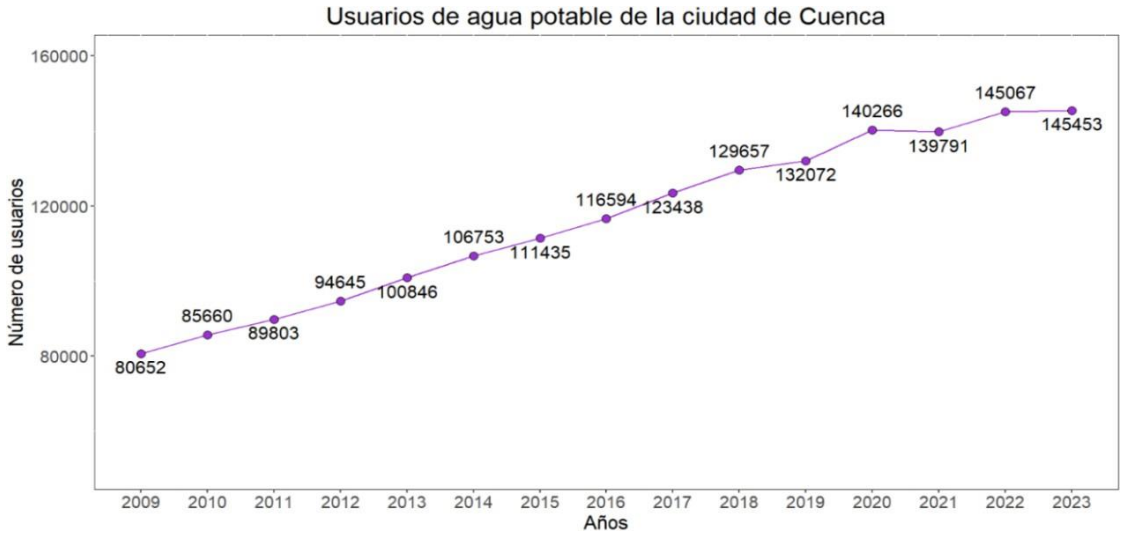


Figura 5: Usuarios de agua Potable de la ciudad de Cuenca considerados aptos para el estudio. Elaboración Propia.
Fuente: ETAPA EP (2023).

En la Figura 5 en el año 2009, el número de usuarios de agua potable considerados aptos para el estudio son 80.652 y en el año 2023 son 145.453 usuarios. Estas cantidades representan el 84% y 80% de datos válidos en los años mencionados de la Figura 3, mostrando un incremento de 64.801 usuarios a lo largo de los años de estudio.

La Figura 5 refleja una tendencia notable al crecimiento, especialmente desde el año 2009 al 2018, esto podría atribuirse a la expansión territorial que ha presentado la ciudad a lo largo del tiempo. Sin embargo, la tendencia a partir del año 2019, es menor con relación a los años anteriores, lo cual podría indicar que la ciudad sigue creciendo en usuarios de agua potable que han cambiado las características de sus instalaciones. Dicho de otra manera, los predios con un solo medidor podrían estar cambiando a más medidores al modificar su edificación.

4.4. Selección del número de clústeres con *kmeans*

Uno de los parámetros para la seleccionar del número de clústeres fue utilizar el método del codo (Figura 6), el cual muestra que el número adecuado de clústeres podrían estar entre 5 y 6, de acuerdo con la suma de cuadrados dentro del clúster.

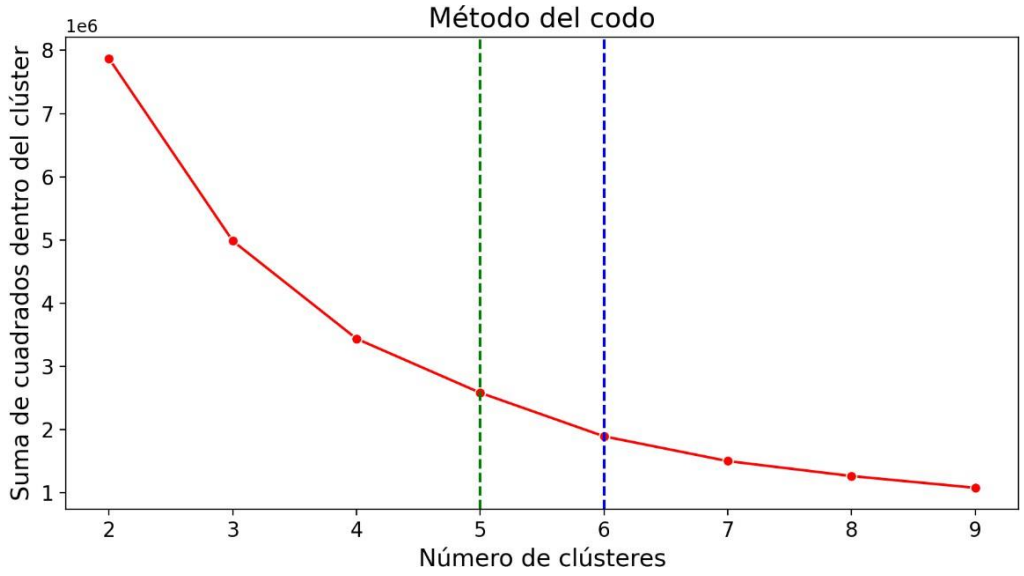


Figura 6: Método del codo con *k-means* para evaluar el número adecuado de clústeres. Elaboración Propia.
Fuente: (Torgo, 2011).

En la Tabla 3, se puede observar que el menor porcentaje de varianza se encuentra entre el clúster número 6 y 7, con un valor de 1% de diferencia. Por tal motivo, a partir del clúster número 6 el porcentaje de varianza no es representativo, indicando que no es necesario asignar más grupos a la clusterización.

Clúster 2	Clúster 3	Clúster 4	Clúster 5	Clúster 6	Clúster 7	Clúster 8
61%	75%	80%	87%	91%	92%	94%

Tabla 3: Número de clústeres a partir del porcentaje de varianza explicada. Elaboración Propia.
Fuente: (Torgo, 2011).

4.5. Distribución de puntos agrupados por *k-means*

Uno de los parámetros para determinar el número de clústeres es el método del codo, como se muestra en la Figura 6 y el porcentaje de varianza determinado en la Tabla 3. Se clusterizó la base de datos de usuarios de ETAPA EP ajustando el hiper parámetro a 6 clústeres en *k- means.* La Figura 7 presenta la asignación de los diferentes puntos en su grupo correspondiente. Mostrando así, la segmentación de los usuarios de agua potable, según sus perfiles de consumo.

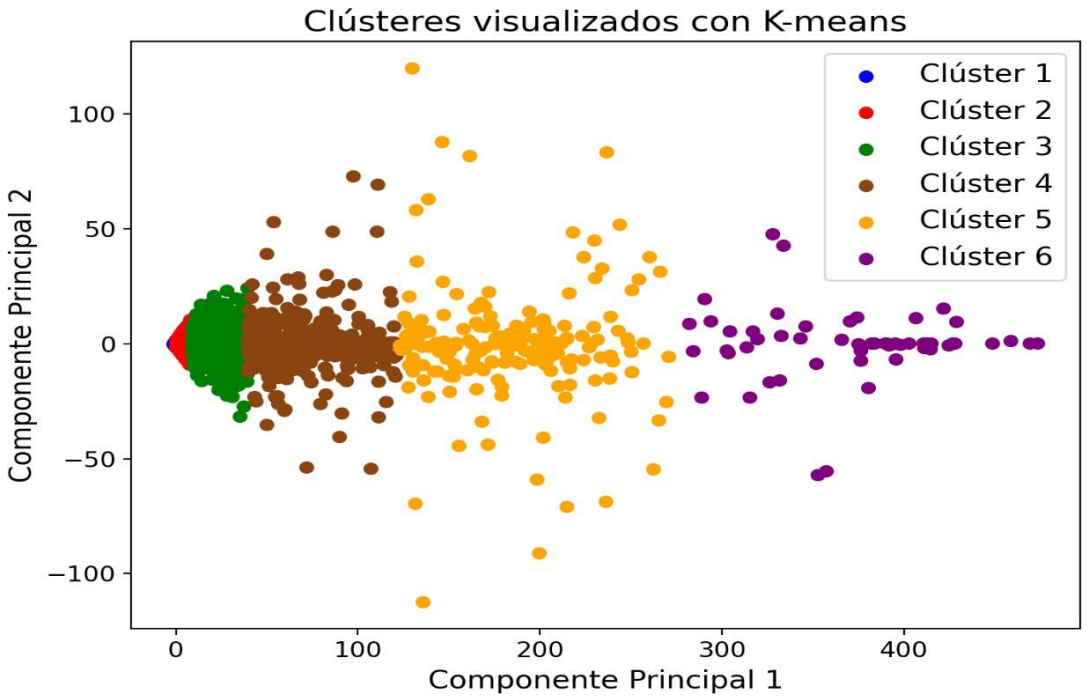


Figura 7: Distribución de los puntos asignados por *clustering* en 2D. Elaboración Propia.
Fuente: ETAPA EP (2023).

La Tabla 4 muestra los centroides obtenidos por *k-means* para los usuarios de agua potable, indicando que a los usuarios asignados en el clúster número 1 corresponden a una media de consumo de 13.90 m^3. Este grupo representa el consumo más bajo perteneciendo al tipo de consumo residencial. El segundo grupo presenta una media de 50.73 m^3, lo cual podría corresponder a usuarios de tipo residencial alto, según lo descrito por Olmedo et al. (2023) en su estudio de segmentación de usuarios de agua potable de tipo residencial. En dicho estudio, se obtuvo 2 grupos de consumo de tipo residencial en rangos de 25 a 29m^3 y 52 a 62 m^3.

Según el tarifario del servicio de agua potable y saneamiento de la empresa ETAPA EP, los consumos de agua potable mayores a 50 m^3 son considerados usuarios de tipo comercial o industrial (ETAPA EP, 2023b). De acuerdo con esta información el clúster número 3, que obtuvo una media de 380.05 m^3 se lo podría identificar como usuarios de tipo comercial. Por otro lado, el clúster número 4, 5 y 6 presentan medias de consumo mayores a 1300 m^3. Estos valores altos de consumo de agua podrían considerarse de tipo industrial.

Mes	Clúster 1 (m^3)	Clúster 2 (m^3)	Clúster 3 (m^3)	Clúster 4 (m^3)	Clúster 5 (m^3)	Clúster 6 (m^3)
1	13.78	50.81	379.70	1316.04	3854.95	7756.60
2	13.80	50.55	376.81	1305.39	3823.44	7692.85
3	14.14	51.32	381.35	1320.60	3867.63	7781.54
4	14.04	51.30	382.37	1325.10	3881.24	7809.50
5	14.16	51.93	388.53	1349.07	3952.12	7953.86
6	13.81	50.28	375.81	1305.78	3825.36	7699.43
7	13.82	50.59	379.85	1322.67	3875.60	7802.44
8	13.63	50.69	383.32	1337.49	3920.08	7893.64
9	13.59	49.84	375.72	1311.49	3843.63	7740.21
10	13.92	50.16	376.31	1313.50	3849.08	7751.39
11	14.16	50.84	380.72	1328.17	3891.79	7836.95
12	13.77	50.48	380.05	1325.53	3884.53	7821.88

Tabla 4: Centroides de *k-means* con 6 clústeres. Elaboración Propia.
Fuente: ETAPA EP (2023).

Con la ayuda de las medias de los grupos encontrados a partir de la Tabla 4, es posible obtener rangos de consumo de agua potable en la ciudad de Cuenca (Tabla 5). Para la obtención del tipo de consumo medio o residencial 2, se promedió el valor de la media de los centroides del clúster número 1 con la media de los centroides del clúster número 2.

Rango	Tipo de consumo
0-14	Residencial 1 (Consumo Bajo)
>14-32	Residencial 2 (Consumo Medio)
>32-51	Residencial 3 (Consumo Alto)
>51-380	Comercial 1
>380-1322	Comercial 2
>1322	Industrial

Tabla 5: Rangos de consumo de agua potable para la ciudad de Cuenca. Elaboración Propia.

4.6. Evolución *clustering* por años (Clúster 1)

La clusterización aplicada por separado a cada año permite observar las características de los patrones de consumo para cada año en particular (Anexo 2), siendo posible evidenciar la evolución de los clústeres a lo largo de los años, como muestra la Figura 8 para el clúster número 1.

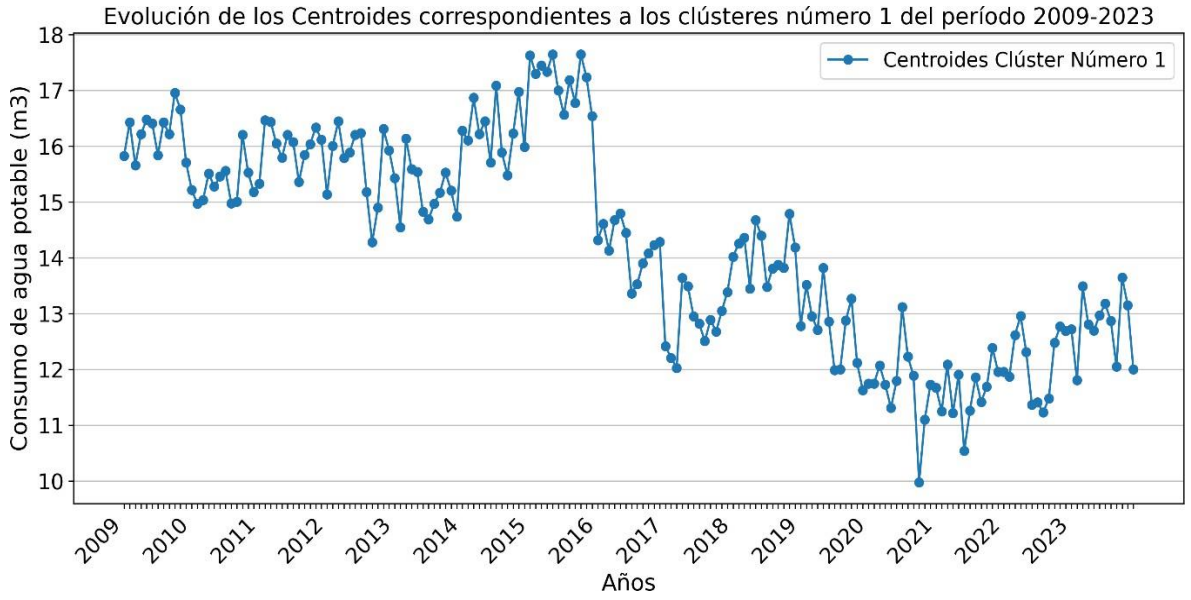


Figura 8: Evolución de centroides para el clúster 1 asociado al enfoque de clusterización por años. Elaboración Propia.
Fuente: ETAPA EP (2023).

La Figura 8 indica que en la ciudad de Cuenca el clúster número 1 pertenece a consumos bajos, durante los años 2009 a 2015 los rangos de consumo estuvieron entre 14 a 17 m^3 y a partir del año 2016 se destaca una disminución considerable en los consumos de la ciudad llegando a rangos entre 14 y 10 m^3. Esto podría atribuirse a las políticas económicas establecidas por la Empresa Municipal de Telecomunicaciones y Agua Potable (ETAPA EP) en el año 2015, modificando el tarifario con un aumento del 100% al metro cúbico de agua, pasando de USD 0.20 a 0.40 centavos de dólar (El Comercio, 2015).

4.7. Comparación Espacio-temporal (2009-2023)

El Anexo 1 muestra un aumento considerable de usuarios correspondientes al consumo residencial 1 y 2, siendo los grupos con mayor porcentaje de crecimiento en el análisis espacio-temporal desde el año 2009 al 2023. Gracias al análisis espacial, se muestra el cambio en los hábitos de consumo de los usuarios. Según la Figura 9, han cambiado sus rangos de

comercial 2 con un color verde oscuro a comercial 1 con un color verde claro y en algunos casos de Residencial 3 correspondiente al color azul a comercial 1 con un color verde claro.

Por otro lado, se puede observar que existe mayor cantidad de usuarios de color azul del año 2023 con relación al 2009, indicando que usuarios de residencial 2 han aumentado su promedio de consumo de agua.

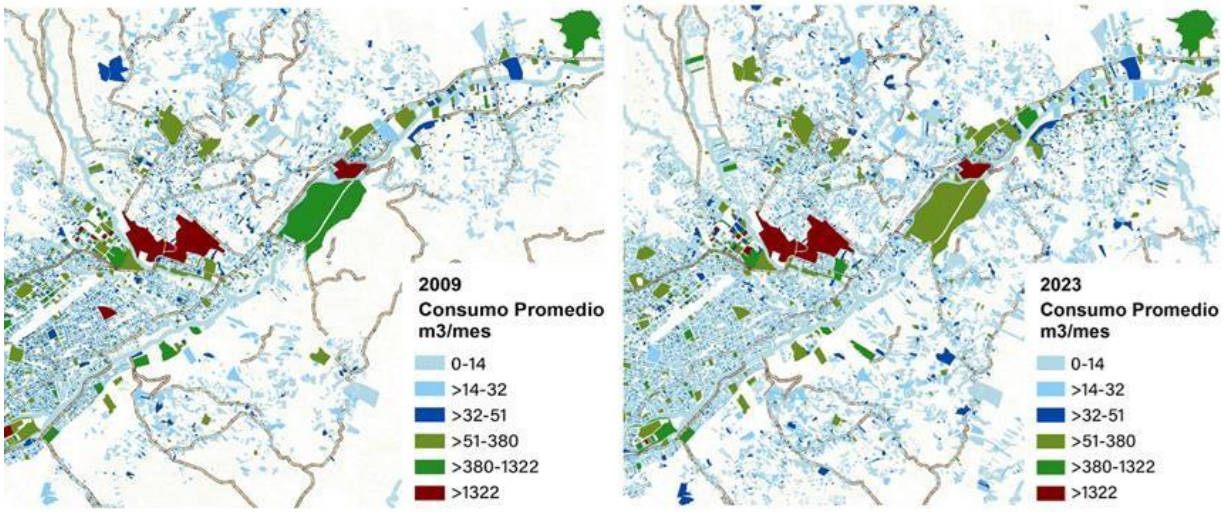


Figura 9: Segmentos de mapas de la ciudad de Cuenca (2009-2023)
Realizado por: Rodas Verónica, Mejía Estalin,2024

5. CONCLUSIONES Y RECOMENDACIONES

Para la aplicación de *clustering* al histórico de consumos de agua potable de la ciudad de Cuenca, se realizó una preparación y transformación de datos (Figura 1). Durante el proceso se pudo evidenciar la existencia de un alto porcentaje de datos anómalos en cada año en la Figura 3. Este porcentaje resaltaría uno de los inconvenientes más importantes de la gestión del agua en entornos urbanos. Por ello, el problema podría derivarse en gran medida del uso de medidores mecánicos y la recolección manual de datos, lo cual afecta directamente la calidad de la información ya sea por errores humanos o mecánicos; siendo uno de los principales obstáculos a la hora de plantear algoritmos para evaluar patrones de consumo de agua en una ciudad.

La gran cantidad de datos a manejar requirió la implementación de un algoritmo de aprendizaje no supervisado sencillo y que no demande altas capacidades computacionales para su desarrollo. Para la agrupación en base a su perfil de consumo en esta investigación, se utilizó *k-means*, el cual, junto con el método del codo en la Figura 6 y el porcentaje de varianza descrito en la Tabla 3, determinaron que el número de grupos en la ciudad de

Cuenca, según su perfil de consumo son 6 (Figura 7). Estando en una media de consumo de 13.90 m^3 para el primer grupo, 50.73 m^3 para el segundo, 380.05 m^3 para el tercer grupo, 1321.74 m^3 para el cuarto grupo, 3872.45 m^3 para el quinto y el último con una media de 7795.02 m^3. Estos valores se obtuvieron a partir de la información de centroides de la Tabla 4.

La clusterización por años presentada en el Anexo 2 podría ayudar a comprender el comportamiento anual de la población con relación a sus hábitos de consumo, los cuales pueden variar debido a influencias de factores sociales, económicos y ambientales. Sin embargo, este criterio podría segmentar a los predios de forma diferente en cada año, considerando su inviabilidad para el análisis de los cambios en el comportamiento de los usuarios de ETAPA EP a través del tiempo. Por esta razón, se clusterizó a todos los años en conjunto, obteniendo valores de consumo generales para todos los años, este enfoque ayudó a plantear rangos de consumo para observar la evolución espacio-temporal del uso de agua potable en la ciudad de Cuenca mostrada en el Anexo 1.

A partir de las medias de consumo de cada grupo obtenidas a partir de la Tabla 4, se puede sugerir rangos de consumo para la ciudad de Cuenca. Estos valores se los ha designado de la siguiente manera: Residencial 1 de consumo bajo (0-14), residencial 2 de consumo medio (>14-32), residencial 3 de consumo alto (>32-51), comercial 1 (>51-380), comercial 2 (>380-1322) y consumo industrial (>1322) (Tabla 5).

Con la ayuda de los rangos de consumo obtenidos en la Tabla 5, se realizó el análisis espacio-temporal del histórico de consumo de agua potable para el periodo 2009-2023 (Anexo 1), evidenciando el crecimiento que la ciudad ha experimentado en los últimos años. Los usuarios de tipo residencial 1 y 2 son los que han tenido un incremento mayor a lo largo de los años de estudio, mostrando un aumento de estos grupos en los alrededores del perímetro urbano. Esto destaca la necesidad de aplicar políticas de sostenibilidad dirigidas a los usuarios de tipo residencial.

Como futuras líneas de investigación, se recomienda analizar los comportamientos de consumo de una parte representativa de la ciudad a través de medidores inteligentes, lo cual aportara información mucho más concreta y en tiempo real de las conductas de consumo evaluadas.

La base de datos depurada de cada año servirá para futuros estudios concatenando los datos de consumos con información demográfica y de posesión de bienes de los usuarios. De esta

forma, se podría plantear algoritmos de aprendizaje no supervisado que evalúen grupos de consumo basándose en una mayor cantidad de variables como, por ejemplo: Número de habitantes, número de vehículos y condición social.

Por último, sería interesante identificar usuarios que no correspondan a los grupos asignados y sean considerados como atípicos. Pues una de las debilidades del algoritmo *k-means* es que asigna todos los puntos cercanos a un grupo, de esta forma no se podría conocer usuarios que no formen parte de ninguno de los grupos establecidos.

6. BIBLIOGRAFÍA

Adamala, S. (2017). Sirisha Adamala. An Overview of Big Data Applications in Water Resources Engineering. *Machine Learning Research*, *2*(1), 10–18. https://doi.org/10.11648/j.mlr.20170201.12

Alkelani, Z., & Awad, M. (2021). *K-Means Clustering Based Model for Fair Water Distribution of Urban Regions Depending on Consumption*.

Brentan, B., Meirelles, G., Luvizotto, E., & Izquierdo, J. (2018). Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems. *Environmental Modelling & Software*, *106*, 77–88. https://doi.org/https://doi.org/10.1016/j.envsoft.2018.02.013

Chander, S., & Vijaya, P. (2021). 3 - Unsupervised learning methods for data clustering. In D. Binu & B. R. Rajakumar (Eds.), *Artificial Intelligence in Data Mining* (pp. 41–64). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-820601-0.00002-1

El Comercio. (2015, May 4). Las tarifas de agua potable suben en Cuenca. *El Comercio* . https://www.elcomercio.com/actualidad/ecuador/tarifas-aguapotable-cuenca.html

ETAPA EP. (2023a). HIST_AGP_VR2.csv. In *Base de datos de consumos de agua potable de la ciudad de Cuenca período (2009-2023). Proporcionados por ETAPA EP para fines de investigación.*

ETAPA EP. (2023b). *Tarifario del servicio de Agua Potable y Saneamiento 2023*. Conozca El Tarifario Actualizado Del Servicio de Agua Potable y Saneamiento. https://www.etapa.net.ec/agua-potable-y-saneamiento/agua-potable/tarifario-agua-potable-2023/

Ghamkhar, H., Jalili Ghazizadeh, M., Mohajeri, S. H., Moslehi, I., & Yousefi-Khoshqalb, E. (2023). An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity. *Sustainable Cities and Society*, *94*, 104516. https://doi.org/https://doi.org/10.1016/j.scs.2023.104516

Herslund, L., & Mguni, P. (2019). Examining urban water management practices – Challenges and possibilities for transitions to sustainable urban water management in Sub-Saharan cities. *Sustainable Cities and Society*, *48*, 101573. https://doi.org/https://doi.org/10.1016/j.scs.2019.101573

INEC. (2022). *Censo Ecuador*. https://censoecuador.ecudatanalytics.com/

Ioannou, A. E., Creaco, E. F., & Laspidou, C. S. (2021). Exploring the effectiveness of clustering algorithms for capturing water consumption behavior at household level. *Sustainability*, *13*(5), 2603.

Ioannou, A. E., Kofinas, D., Spyropoulou, A., & Laspidou, C. (2017). Data mining for household water consumption analysis using self-organizing maps. In *European Water* (Vol. 58).

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., & Corlay, S. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. *Elpub*, *2016*, 87–90.

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A., & Tassiulas, L. (2015). Exploring Patterns in Water Consumption by Clustering. *Procedia Engineering*, *119*, 1439–1446. https://doi.org/https://doi.org/10.1016/j.proeng.2015.08.1004

Leitão, J., Simões, N., Marques, J. A. S., Gil, P., Ribeiro, B., & Cardoso, A. (2019). Detecting urban water consumption patterns: A time-series clustering approach. *Water Science and Technology: Water Supply*, *19*(8), 2323–2329. https://doi.org/10.2166/ws.2019.113

Olmedo, A. T., Castro, J. R., Reátegui, R., & Castillo, T. (2023). Aplicación de algoritmos de Machine Learning para la segmentación del consumo de agua potable. Caso de estudio en Catamayo, Ecuador. *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI58278.2023.10211900

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Ahmed, T., Giurco, D., & Blumenstein, M. (2021). A clustering solution for analyzing residential water consumption patterns. *Knowledge-Based Systems*, *233*, 107522. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107522

Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing Ltd.

Retamal, M. R., Rojas, J., & Parra, O. (2011). Percepción al cambio climático y a la gestión del agua: aportes de las estrategias metodológicas cualitativas para su comprensión. *Ambiente & Sociedade*, *14*(1), 175–194. https://doi.org/10.1590/S1414-753X2011000100010

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/https://doi.org/10.1016/j.neucom.2017.06.053

SWACH. (2023). *Sustainable Water Management Under Climate Change in Southern Ecuador*. https://swach.uazuay.edu.ec/

Torgo, L. (2011). *Data mining with R: learning with case studies*. chapman and hall/CRC.

UNESCO. (2020). Informe mundial de las Naciones Unidas sobre el desarrollo de los recursos hídricos 2020: agua y cambio climático. In *Informe de las Naciones Unidas sobre el desarrollo de los recursos hídricos en el mundo 2016: agua y empleo, p. 99-106, illus.* https://unesdoc.unesco.org/ark:/48223/pf0000373611.locale=es

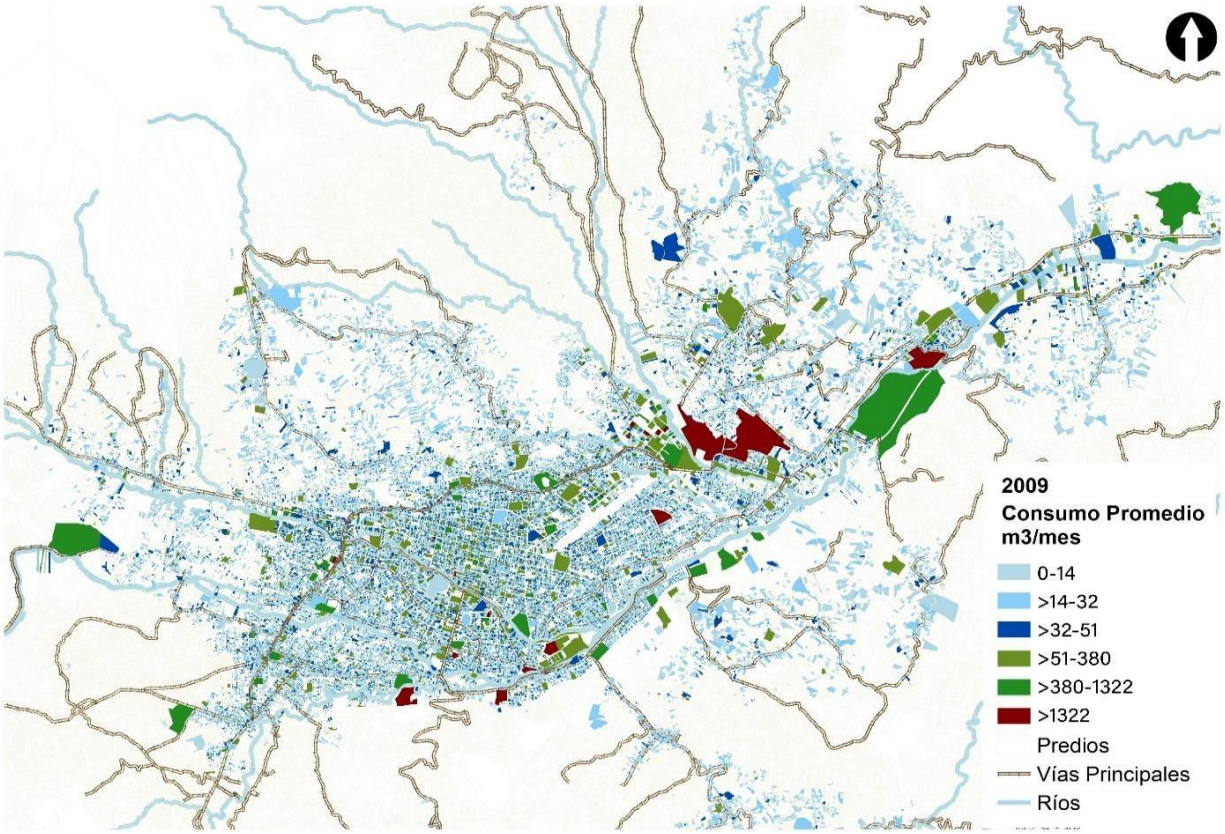
UNESCO. (2023). *Informe Mundial de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos 2023: alianzas y cooperación por el agua*. https://unesdoc.unesco.org/ark:/48223/pf0000386807

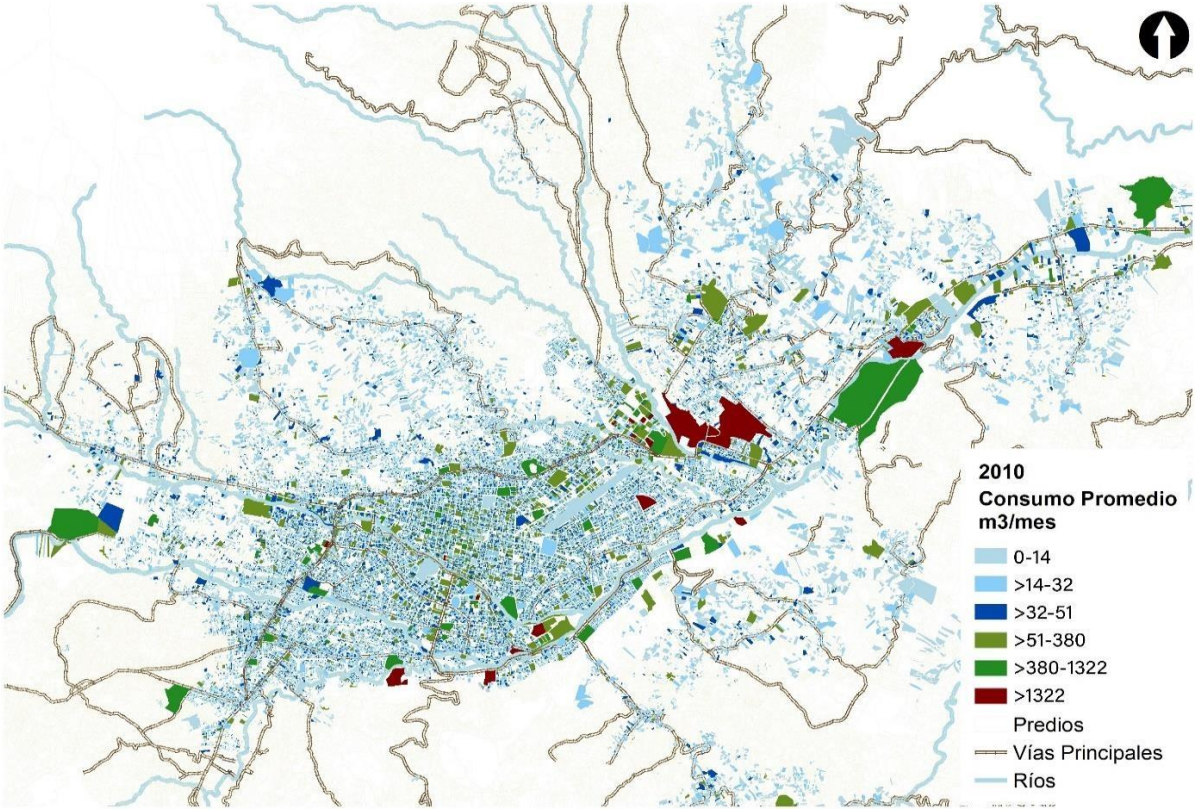
Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In S. C. Satapathy, J. M. R. S. Tavares, V. Bhateja, & J. R. Mohanty (Eds.), *Information and Decision Sciences* (pp. 511–518). Springer Singapore.

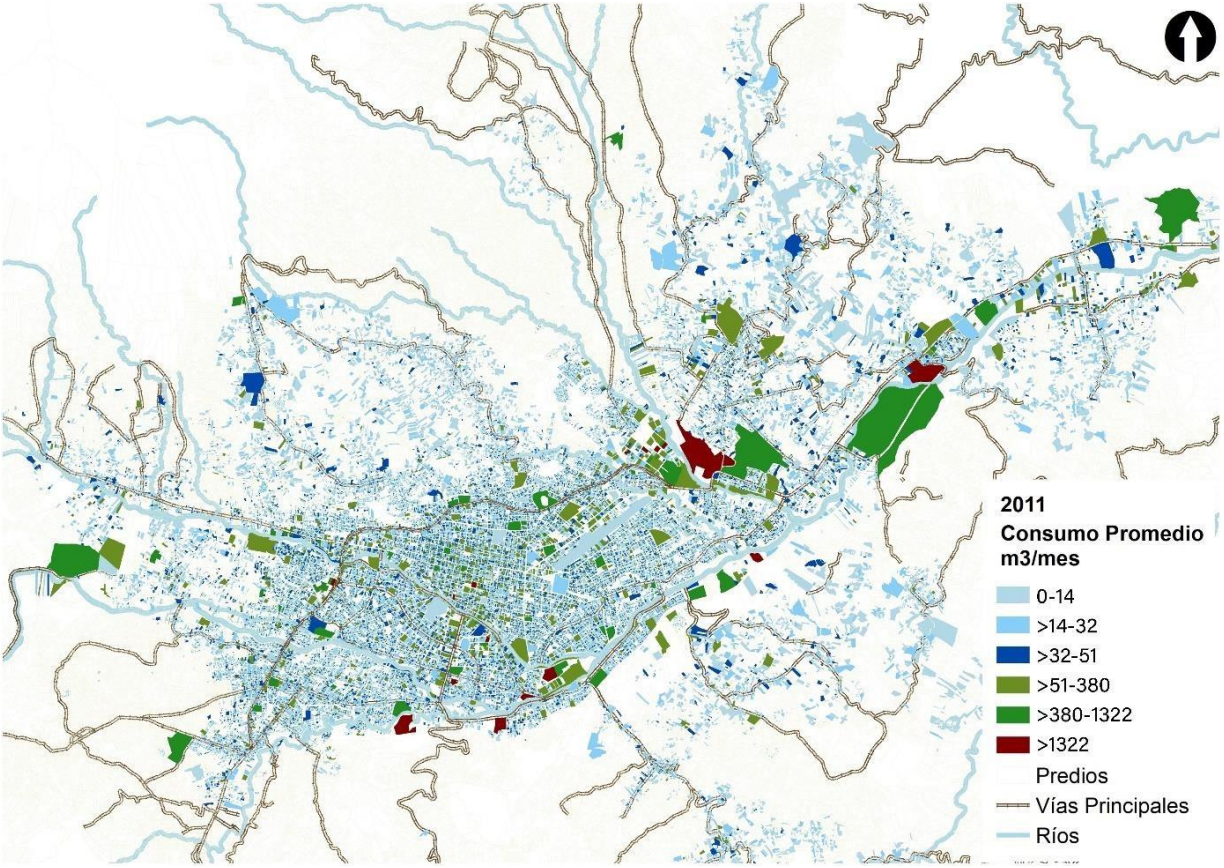
Waraga, O. A., Abdeljaber, A., Talib, M. A., & Abdallah, M. (2021). Investigating Water Consumption Patterns Through Time Series Clustering. *2021 14th International Conference on Developments in ESystems Engineering (DeSE)*, 44–49. https://doi.org/10.1109/DeSE54285.2021.9719367

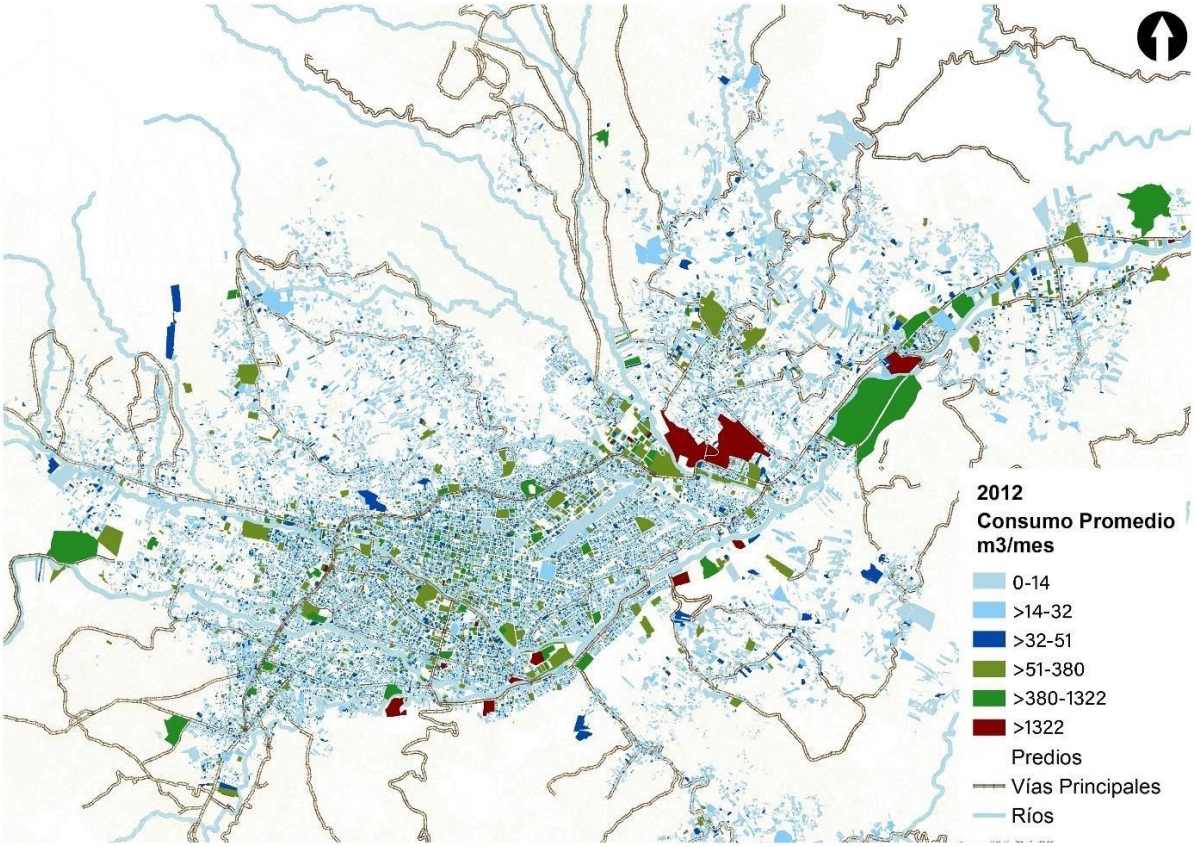
7. ANEXOS

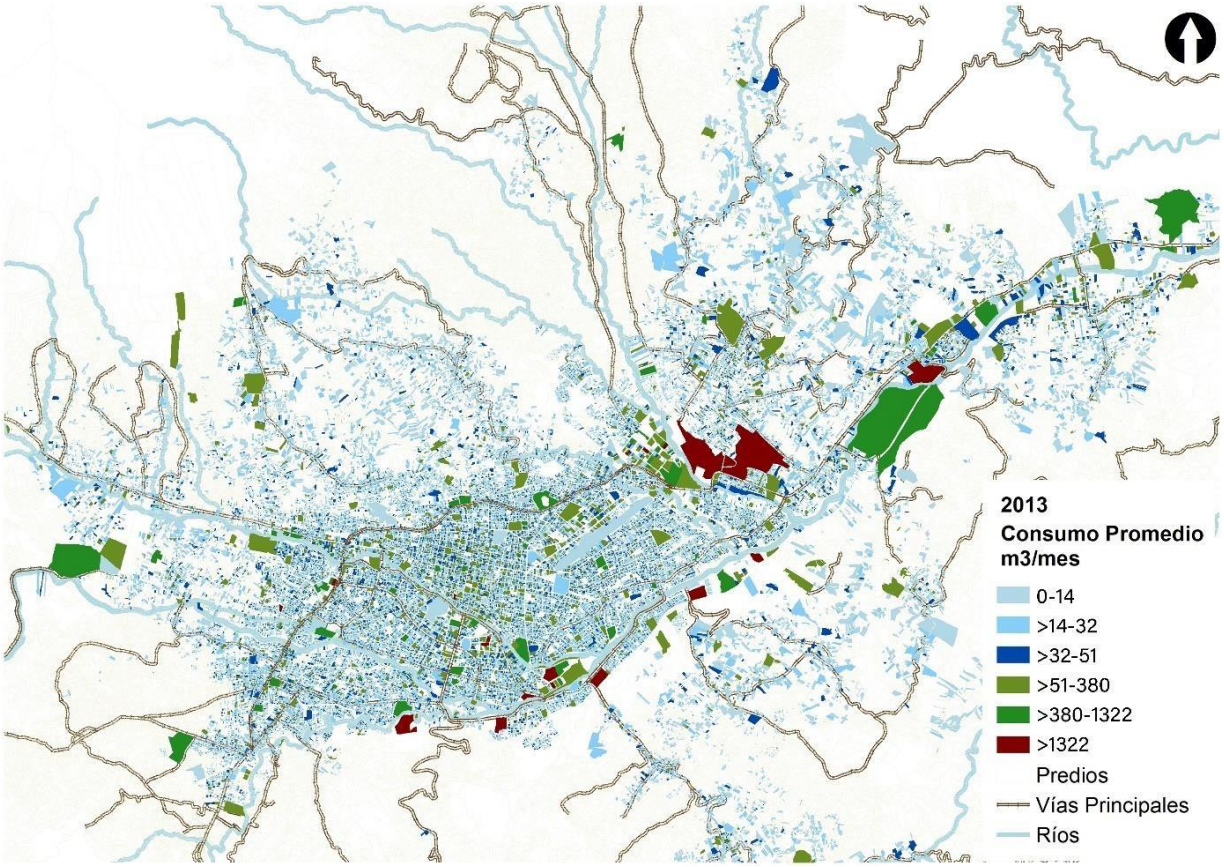
7.1. Análisis espacio-temporal de consumos de agua potable de la ciudad de Cuenca

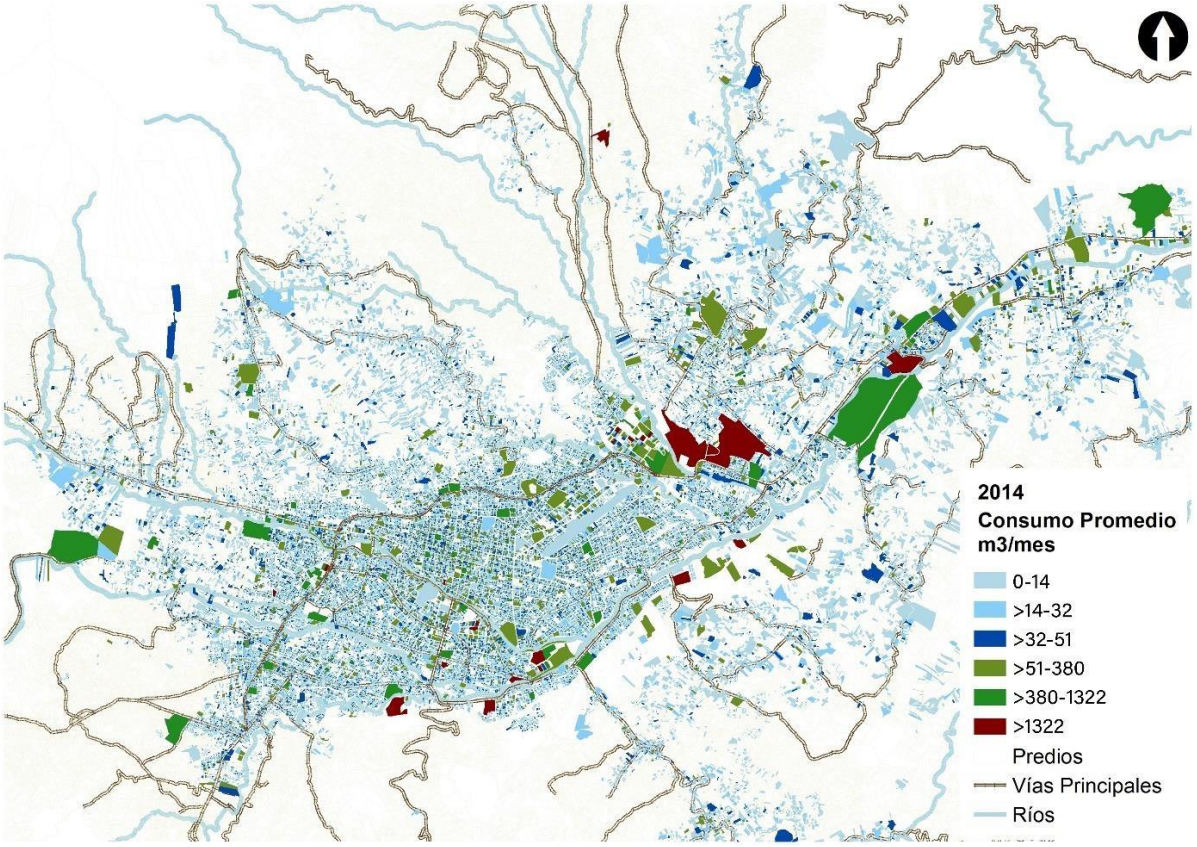


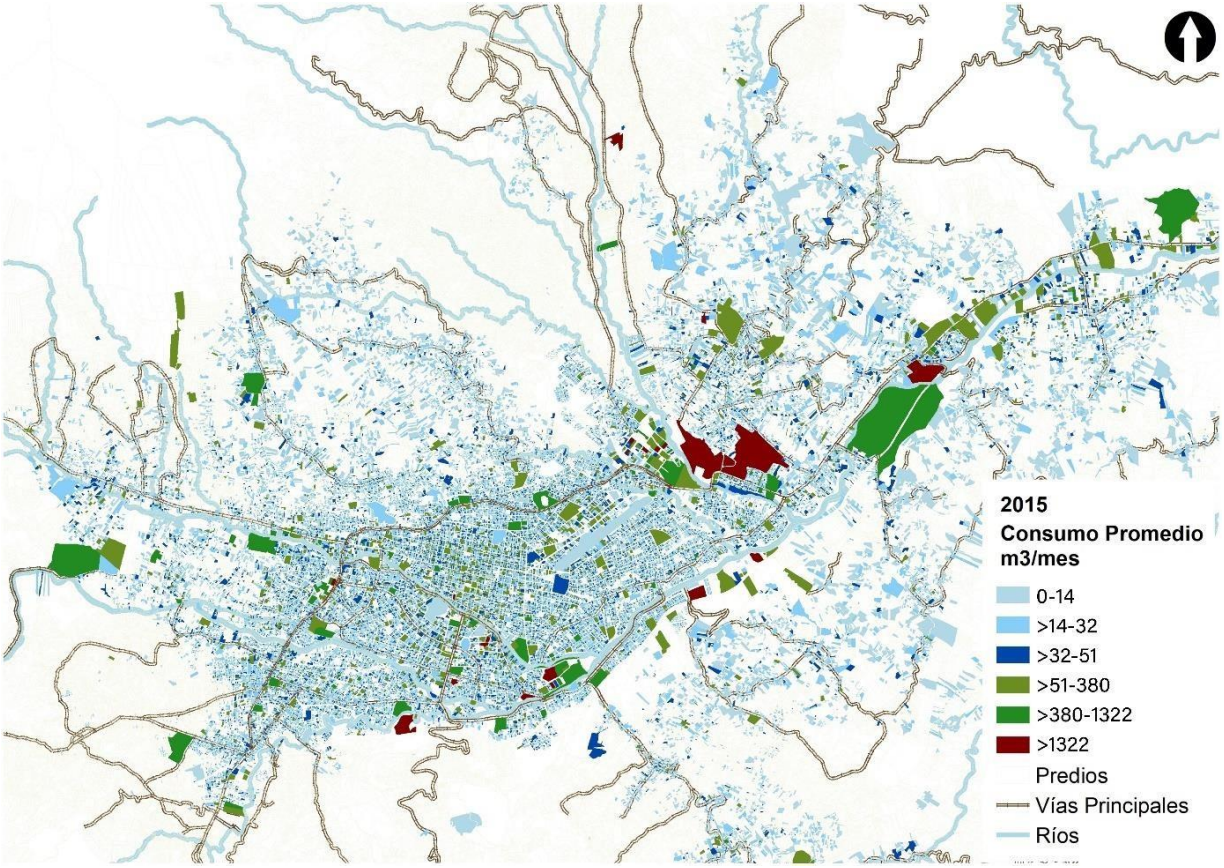


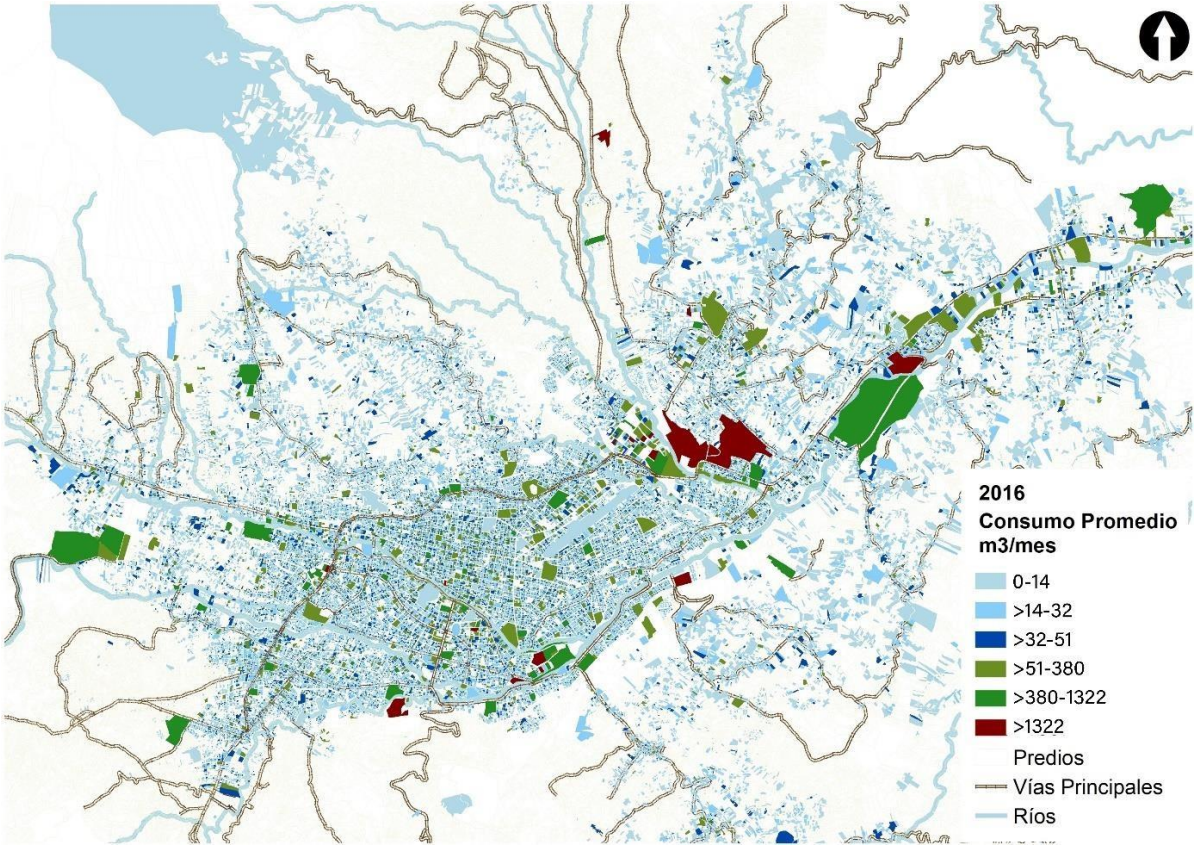

2011
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

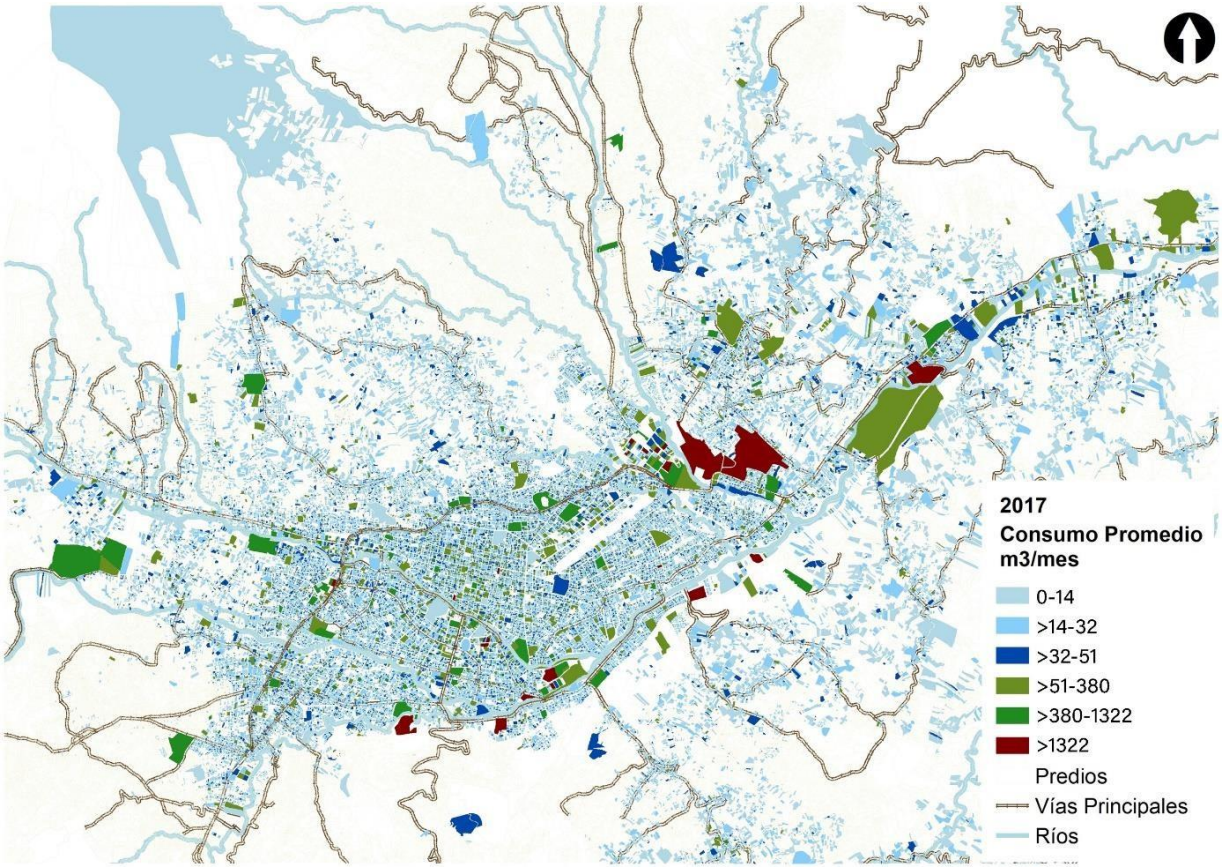

2012
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

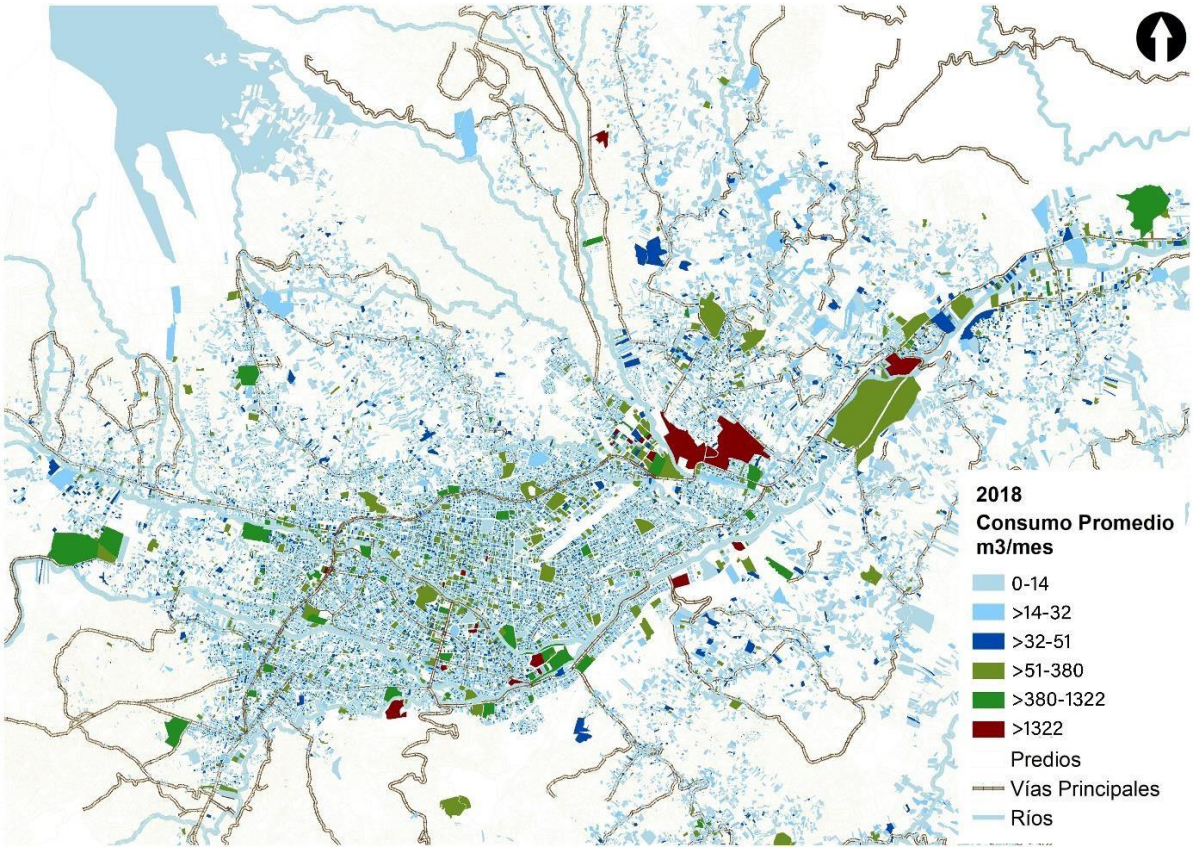

2013
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

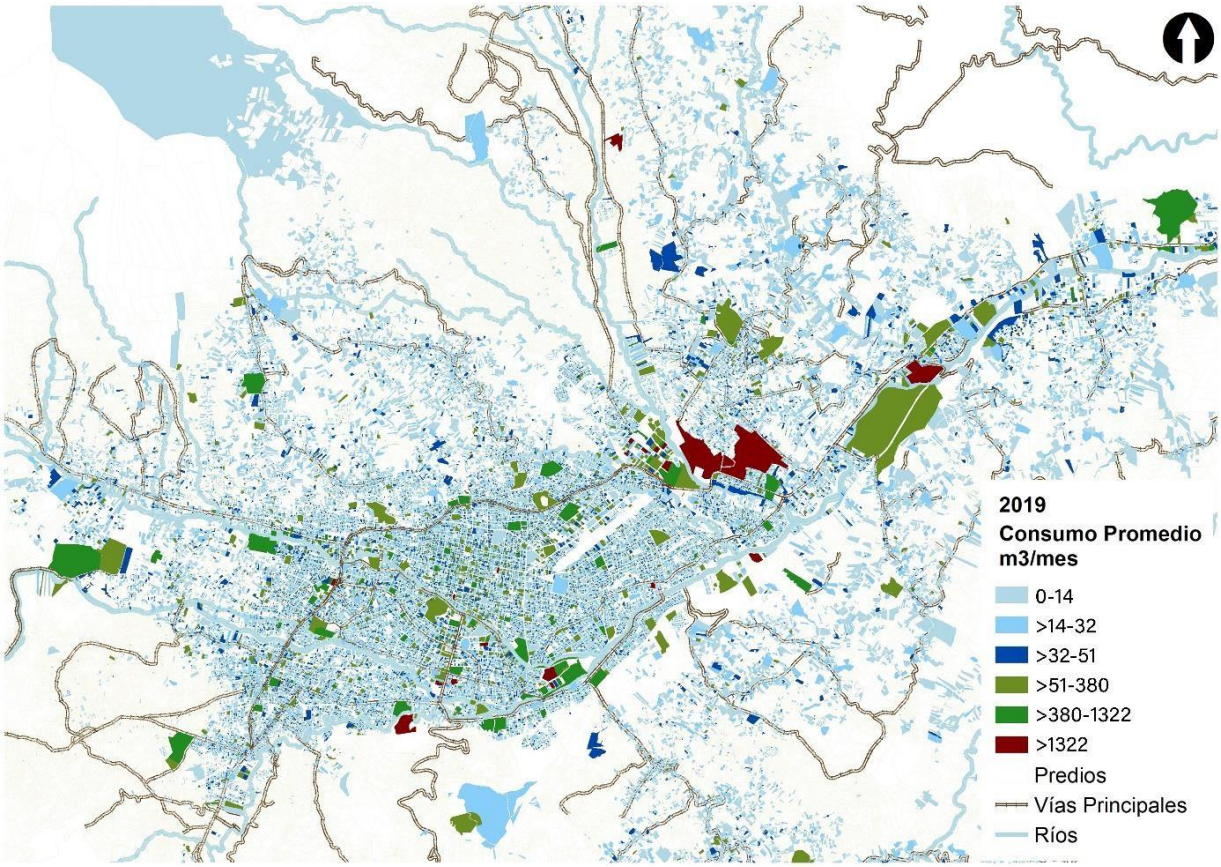

2014
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

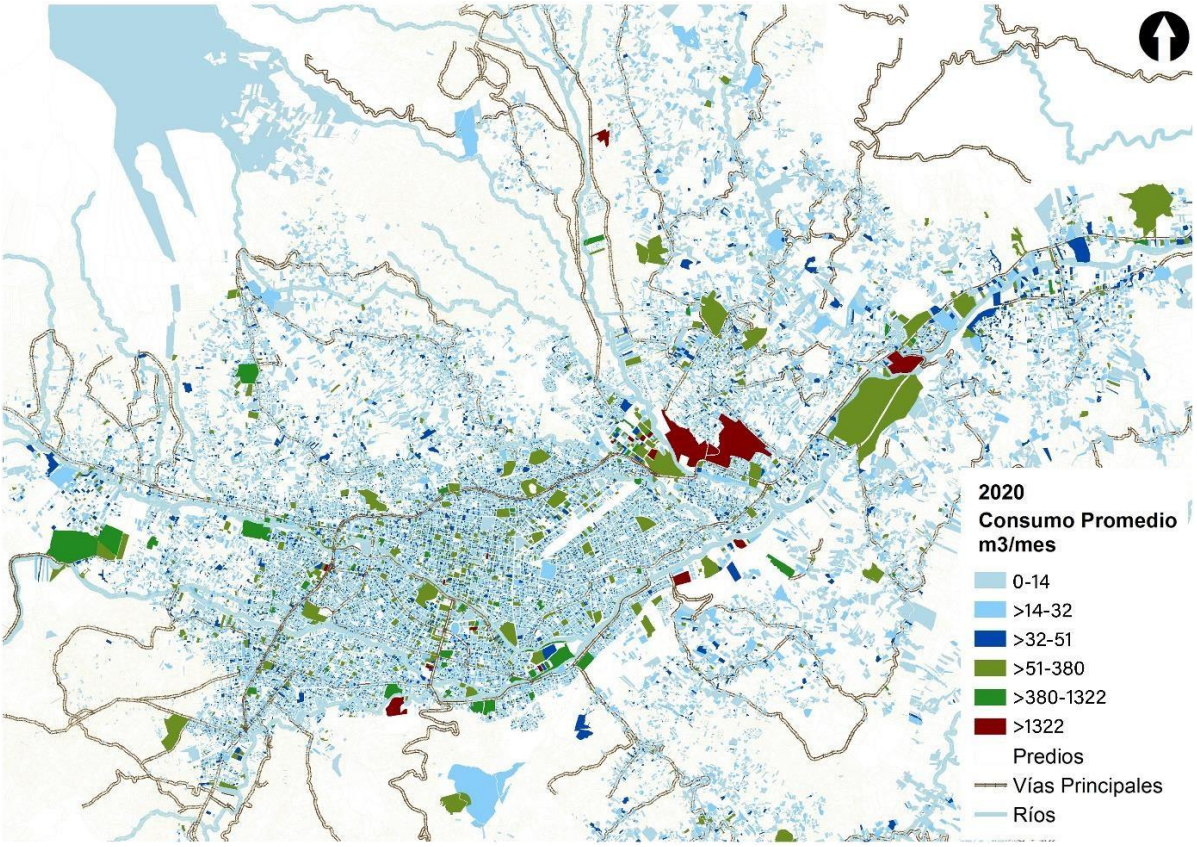

2015
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

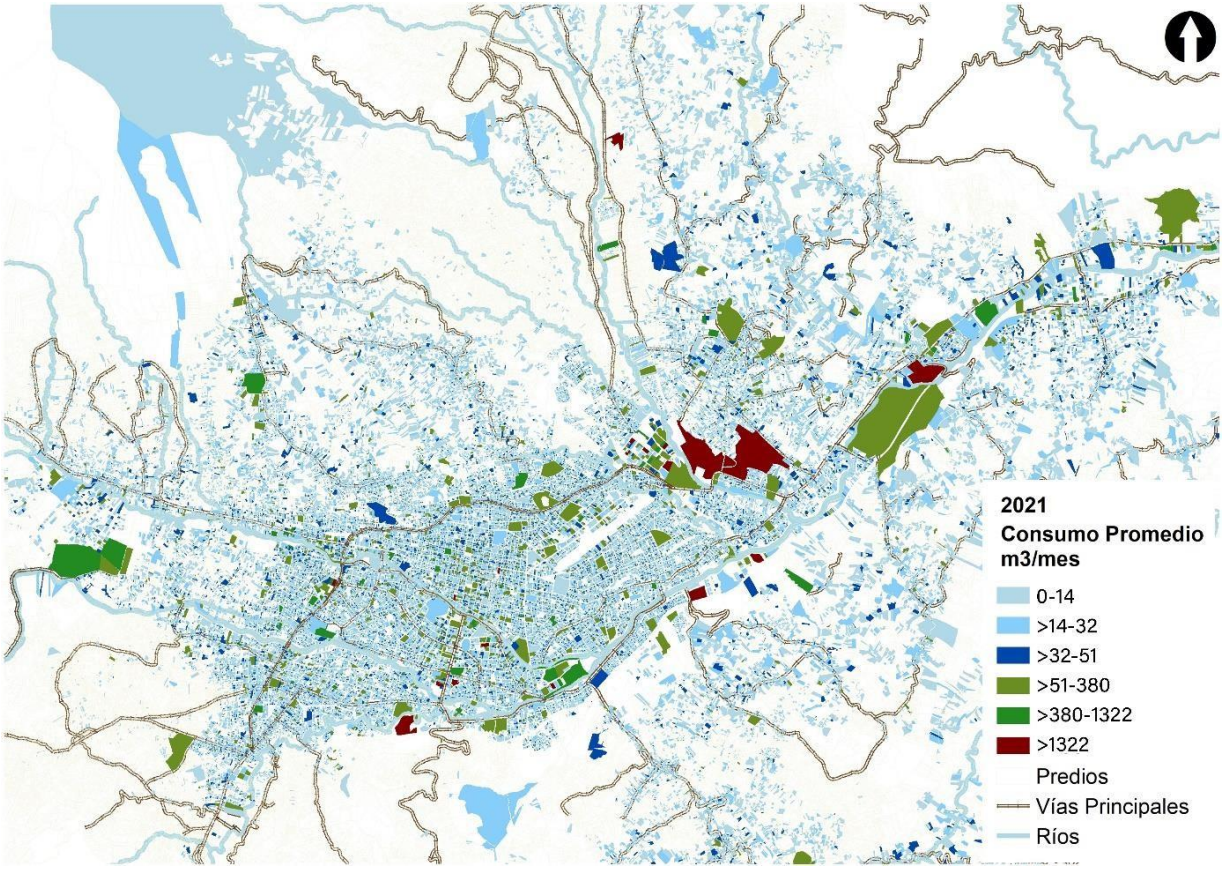

2016
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

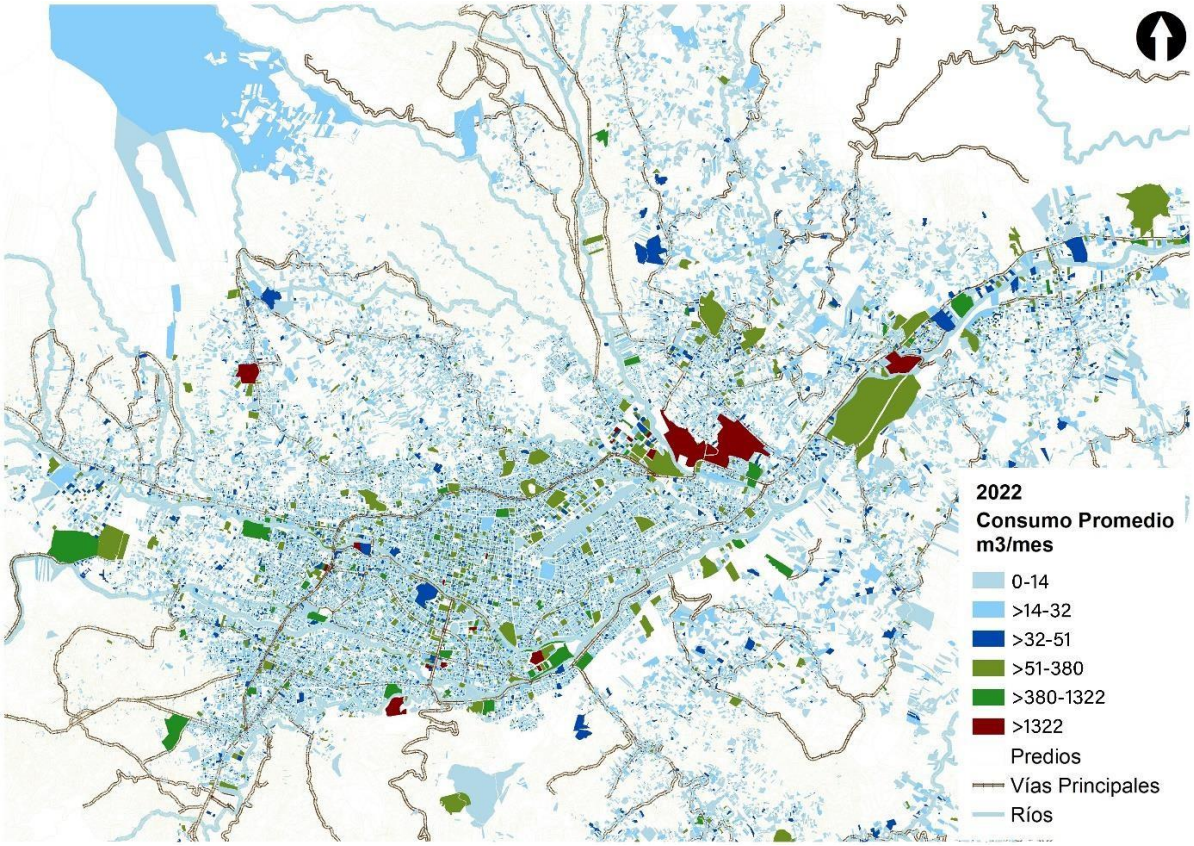

2017
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

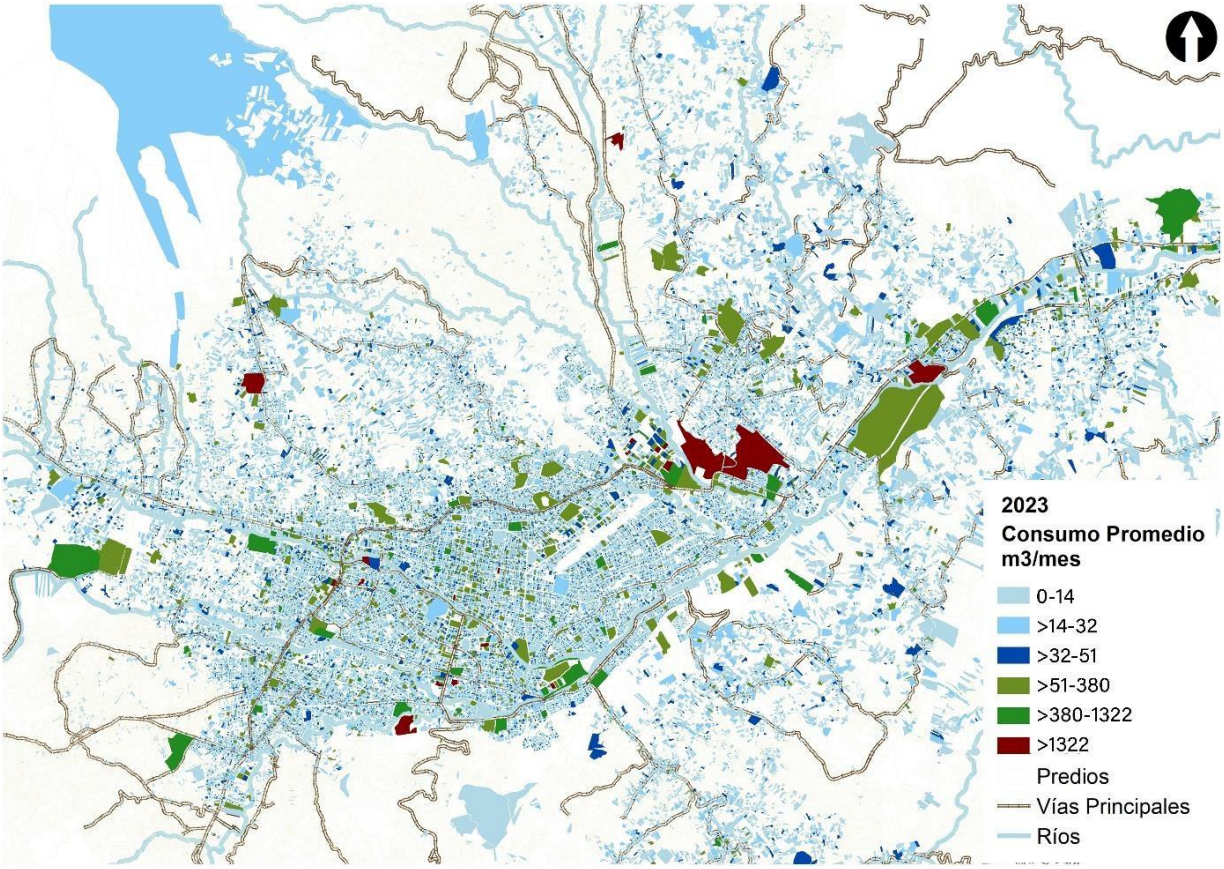

2018
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos


2019
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos


2020
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

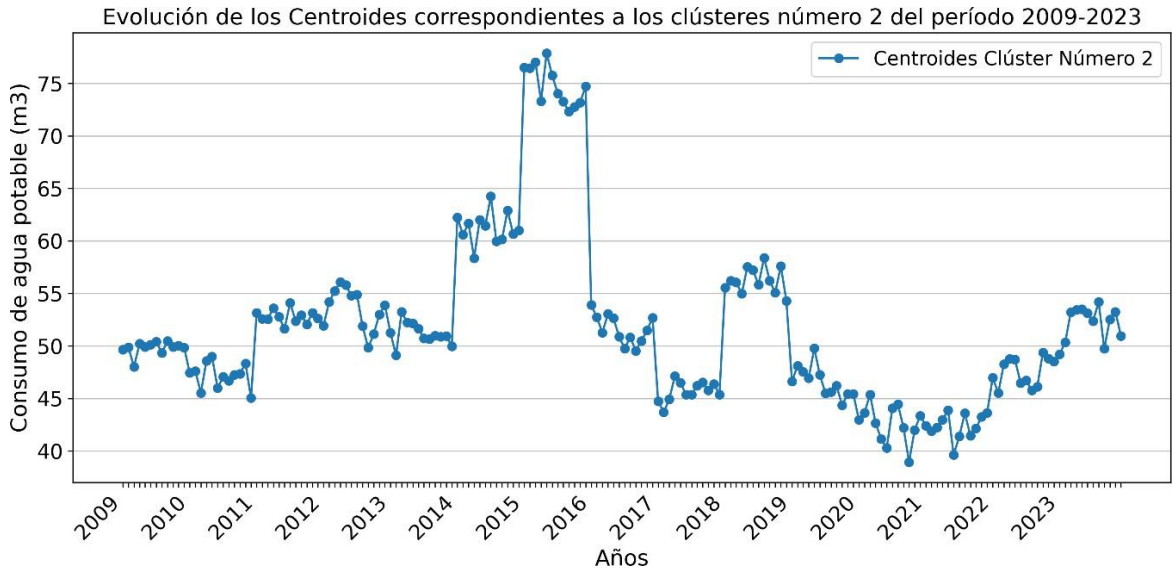

2021
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

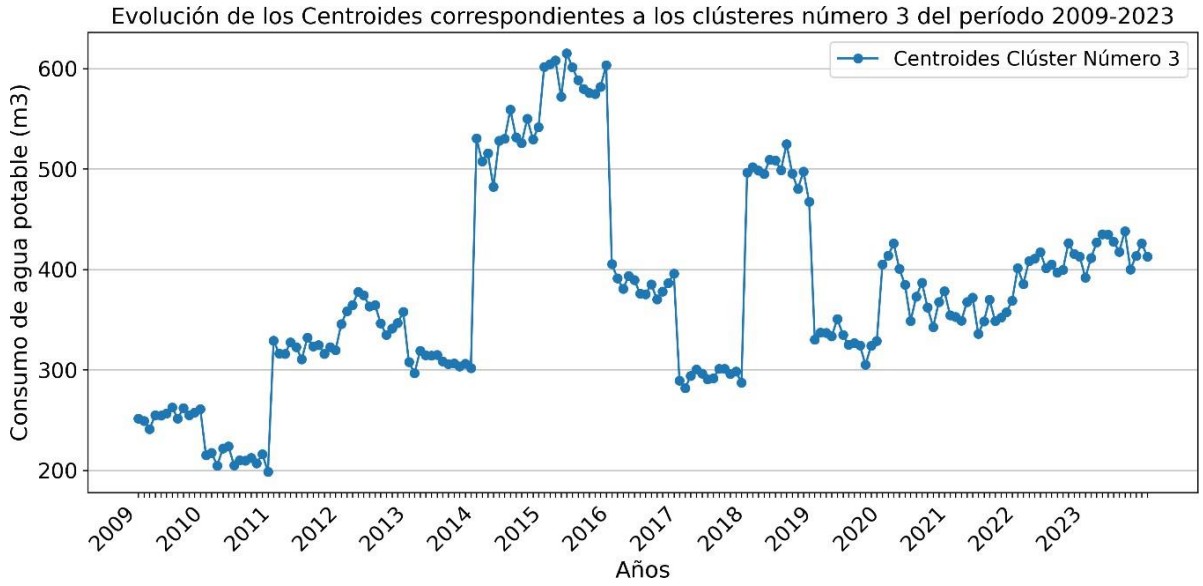

2022
Consumo Promedio m3/mes
0-14
>14-32
>32-51
>51-380
>380-1322
>1322
Predios
Vías Principales
Ríos

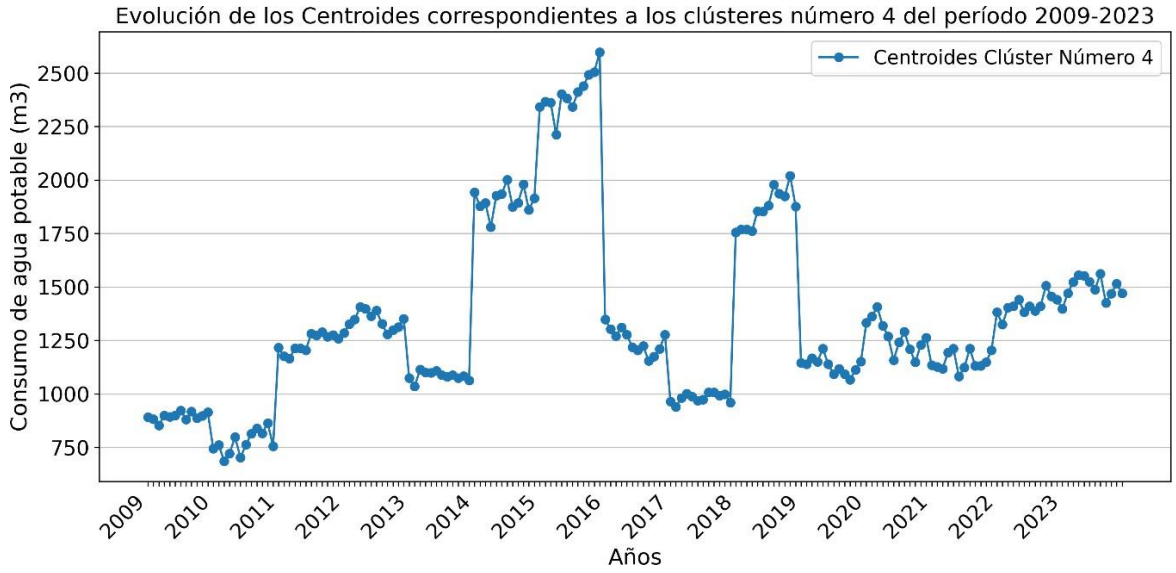


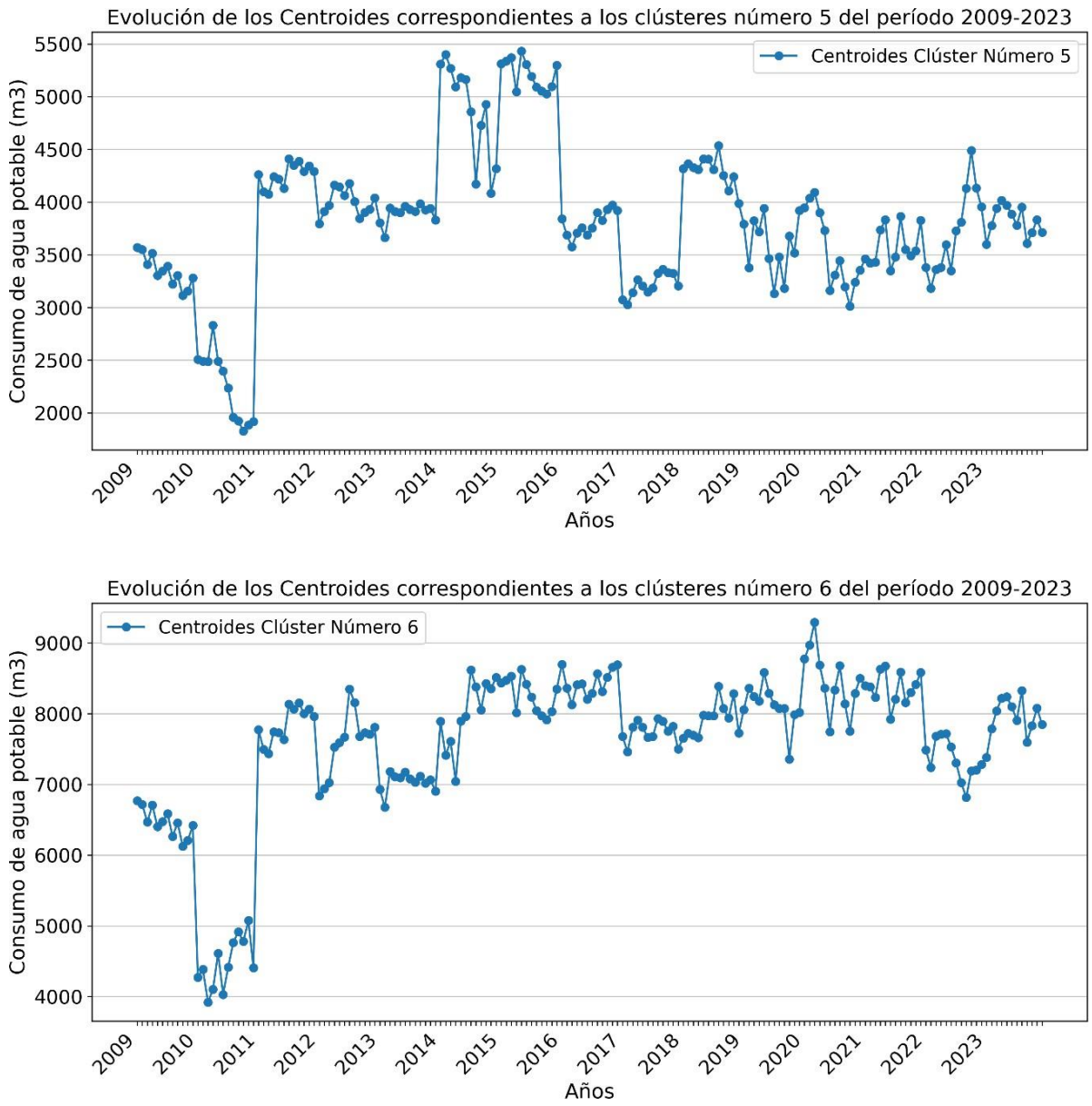
Anexo 1: Análisis Espacio-Temporal de consumos de agua potable de la ciudad de Cuenca (2009-2023).
Realizado por: Rodas Verónica, Mejía Estalin, 2024.

7.2. Evolución de centroides de la clusterización por años




Evolución de los Centroides correspondientes a los clústeres número 3 del período 2009-2023
Centroides Clúster Número 3
Consumo de agua potable (m3)
600
500
400
300
200
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Años


Evolución de los Centroides correspondientes a los clústeres número 4 del período 2009-2023
Centroides Clúster Número 4
Consumo de agua potable (m3)
2500
2250
2000
1750
1500
1250
1000
750
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Años



Anexo 2: Evolución de centroides de *clustering* por cada año. Elaboración Propia.
Fuente: ETAPA EP (2023).