Universitat de Girona

MÉTODO PARA DESCUBRIR PATRONES ESPACIO-TEMPORALES DE COMPORTAMIENTO DE USUARIOS ELÉCTRICOS UTILIZANDO HERRAMIENTAS DE APRENDIZAJE AUTOMÁTICO

Mario Alberto Flor Ambrosi

Universitat de Girona

TESIS DOCTORAL

Método para descubrir patrones espacio-temporales de comportamiento de usuarios eléctricos utilizando herramientas de aprendizaje automático

MARIO ALBERTO FLOR AMBROSI

2021

Universitat de Girona

TESIS DOCTORAL

Método para descubrir patrones espacio-temporales de comportamiento de usuarios eléctricos utilizando herramientas de aprendizaje automático

MARIO ALBERTO FLOR AMBROSI

2021

PROGRAMA DE DOCTORADO EN TECNOLOGÍA

Supervisado por: Dr. Sergio Herraiz Jaramillo y Dr. Iván Contreras Fernández-Dávila

Memoria presentada para optar al título de doctor por la Universidad de Girona

A mi padre Carlos †, a mi madre Marianita,

a mis hijos Leonardo, Mario y Sofía,

y a Mónica, con amor.

AGRADECIMIENTO

El autor fue premiado por la Universidad de Girona y SENESCYT-Ecuador, con una beca predoctoral de la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación, (SENESCYT) - Ecuador.

Mi agradecimiento por su apoyo al grupo de investigación eXiT (Enginyeria de control i sistemes intel·ligents) y de manera especial a mi amigo Iván Contreras.

Mario Alberto Flor Ambrosi

Girona, España

Julio 2021

LISTA DE PUBLICACIONES Y COMUNICACIONES

1. Flor, M., Herraiz, S., Contreras, I. (2021). Electricity load profiling using machine learning. *V Jornades d' Investigadors Predoctorals de la Universitat de Girona.* Spain, 2021.
2. Flor, M., Herraiz, S., Contreras, I. (2021). Definition of residential power load profiles clusters using machine learning and spatial analysis. *Energies (ISSN 1996-1073)* . Enviado.

LISTA DE ABREVIATURAS

Los acrónimos y abreviaturas que se muestran a continuación se pueden encontrar a lo largo de esta tesis.

Acrónimos y abreviaturas

AMI	Advanced metering infrastructure (Infraestructura de medición avanzada)
ANN	Artificial neural networks (Red neuronal artificial)
BAN	Building area network (Red de área de edificios)
CGIS	Canadian geographic information system (Sistema de información geográfico de Canadá)
DTW	Dynamic time warping (Deformación dinámica de tiempo)
ED	Empresa de distribución de electricidad
FAN	Field area network (Red de área de campo)
GPS	Global Positioning System (Sistema de posicionamiento global)
HAN	home area network (Red de hogar)
IAN	industrial area network (Red de área industrial)
ID	Identificador
LSTM	Long short-term memory (Red de memoria a corto y largo plazo)
MDM	Meter data management (Sistema de gestión de mediciones)
NAN	Neighborhood area network (Red de área de vecindario)
PMU	Phasor measurement unit (Unidad de medición de fasores)
RMSE	Root mean squared error (Raíz del error cuadrático medio)
RNN	Recurrent neural network (Red neuronal recurrente)
RTU	Remote terminal unit (Unidad terminal remota)
SBD	Shape based distance (Distancia basada en la forma)
SIG	Sistema de información Geográfico
SM	Smart meter (Medidor inteligente)
sMAPE	Symmetric mean absolute percentage error (Error porcentual absoluto medio simétrico)
WAN	Wide area network (Red de área amplia)

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUM

A questa investigació presenta un enfocament nou per definir zones geogràfiques amb perfils típics de consum elèctric a partir dels registres de mesuradors intel·ligents, utilitzant mètodes d'aprenentatge automàtic i anàlisi espacial.

Les empreses d'electricitat han de garantir la qualitat i fiabilitat del servei elèctric. Per aconseguir aquest objectiu, les empreses de distribució d'electricitat requereixen conèixer en detall i amb una periodicitat adequada els perfils de consum dels seus clients. Els moderns dispositius de telemesura, com els mesuradors intel·ligents, obren la porta a una immensa quantitat de dades i nous anàlisis, a causa de la major freqüència i precisió de la informació del consumidor, però els mètodes convencionals no poden abordar els voluminosos i ràpids conjunts de dades que generen aquests dispositius. L'objectiu d'aquesta investigació és utilitzar tècniques d'aprenentatge automàtic combinades amb l'anàlisi espacial per generar perfils de càrrega més eficients i precisos a les zones d'estudi.

L'estudi analitza una voluminosa base de dades amb mesuraments de 4 anys recollits per 1000 mesuradors intel·ligents georeferenciats a la ciutat de Guayaquil a l'Equador.

A l'estudi s'implementa una metodologia d'aprenentatge no supervisat per agrupar i classificar les sèries temporals de mesures d'energia, utilitzant la tècnica de deformació dinàmica de temps per descobrir perfils de càrrega típics d'acord amb el seu consum característic setmanal. A continuació, es va realitzar una anàlisi espai temporal restringit per definir zones geogràfiques amb comportament constant i predictible.

Per comprovar el benefici d'obtenir aquesta agrupació espacial, s'utilitza la informació dels seus membres per millorar el pronòstic de càrrega a curt termini utilitzant una xarxa neuronal recurrent amb memòria a curt i llarg termini.

Els resultats d'aquest estudi han demostrat que el patró de consum d'energia en àrees pròximes està relacionat i es pot utilitzar en models que aprofiten aquesta informació com a avantatge. L'anàlisi temporal de les mesures recollides pels mesuradors intel·ligents revela 2 perfils mensuals de consum significativament diferents. A més, es va demostrar que per al cas de l'Equador el perfil de càrrega no canvia significativament a causa de la variabilitat climàtica, sinó a efectes temporals com els festius llargs. L'anàlisi espacial va definir 21 zones geogràfiques on tots els mesuradors tenen el mateix comportament de consum en el mateix període de temps, informació que va ser utilitzada per millorar la previsió de consum energètic dels clients d'aquestes zones en un 2.46%.

Tal i com ha demostrat aquest estudi, el coneixement del perfil de càrrega de zones geogràfiques representa un actiu valuós per a la planificació i disseny de les xarxes de distribució així com per a una planificació d'activitats de manteniment i operació més eficient, i delimitats a zones geogràfiques específiques. Així, aquesta informació es converteix en una entrada important per definir estratègies millor delimitades, com ara la priorització de zones per a campanyes de sensibilització al consumidor, o estimacions dels factors de demanda futura en zones geogràfiques determinades per reduir la inversió en xarxes o centrals elèctriques.

RESUMEN

Esta investigación presenta un enfoque novedoso para definir zonas geográficas con perfiles típicos de consumo eléctrico a partir de los registros de medidores inteligentes, utilizando métodos de aprendizaje automático y análisis espacial.

Las empresas de electricidad deben garantizar la calidad y confiabilidad del servicio eléctrico. Para lograr este objetivo, las empresas de distribución de electricidad requieren conocer en detalle y con una periodicidad adecuada los perfiles de consumo de sus clientes. Los modernos dispositivos de telemedición, como los medidores inteligentes, abren la puerta a una inmensa cantidad de datos y nuevos análisis, debido a la mayor frecuencia y precisión de la información del consumidor. Sin embargo, los métodos convencionales no pueden abordar los voluminosos y rápidos conjuntos de datos que generan estos dispositivos. El objetivo de esta investigación es utilizar técnicas de aprendizaje automático combinadas con el análisis espacial para generar perfiles de carga más eficientes y precisos en las zonas de estudio.

El estudio analiza una voluminosa base de datos con mediciones de 4 años recogidos por 1000 medidores inteligentes georreferenciados en la ciudad de Guayaquil en Ecuador.

En el estudio se implementa una metodología de aprendizaje no supervisado para agrupar y clasificar las series temporales de mediciones de energía, utilizando la técnica de deformación dinámica de tiempo para descubrir perfiles de carga típicos de acuerdo con su consumo característico semanal. A continuación, realizamos un análisis espacio temporal restringido para definir zonas geográficas con comportamientos de consumo constantes y predecibles.

Para comprobar el beneficio de obtener esta agrupación espacial, se utiliza la información de sus miembros para mejorar el pronóstico de carga a corto plazo utilizando una red neuronal recurrente con memoria a corto y largo plazo.

Los resultados de este estudio han demostrado que el patrón de consumo de energía en áreas cercanas está relacionado y se puede utilizar en modelos que aprovechan esta información como ventaja. El análisis temporal de las medidas recogidas por los medidores inteligentes reveló 2 perfiles mensuales de consumo significativamente diferentes. Además, se demostró que para el caso de Ecuador el perfil de carga no cambia significativamente debido a la variabilidad climática, sino a efectos temporales como los feriados largos. El análisis espacial definió 21 zonas geográficas donde todos los medidores tienen el mismo comportamiento de consumo en el mismo período de tiempo, información que fue utilizada para mejorar la previsión de consumo energético de los clientes de estas zonas en un 2.46%.

Tal y como ha demostrado este estudio, el conocimiento del perfil de carga de zonas geográficas específicas representa un activo valioso para la planificación y diseño de las redes de distribución, así como para una planificación de actividades de mantenimiento y operación más eficiente. Esta información además se convierte en un insumo importante para definir estrategias mejor delimitadas, como por ejemplo, la priorización de zonas para campañas de sensibilización al consumidor, o estimaciones de los factores de demanda futura en zonas geográficas determinadas para reducir la inversión en redes o centrales eléctricas.

ABSTRACT

This research presents a novel approach to define geographic areas with typical electricity consumption profiles from smart meter records, using machine learning and spatial analysis methods.

Distribution system operators must guarantee the quality and reliability of the electric service. To achieve this objective, electricity distribution utilities need to know in detail and with an adequate periodicity the consumption profiles of their customers. Modern telemetering devices, such as smart meters, open the door to an immense amount of data and new analysis, due to a higher frequency and precision of consumer electrical consumption. However, conventional methods cannot deal with the voluminous and fast gathered data by smart meters. The objective of this research is to apply machine learning techniques combined with spatial analysis to generate more efficient and accurate load profiles in the areas of study.

The study analyzes a voluminous database of measurements gathered during 4 years by 1000 georeferenced smart meters located in the city of Guayaquil in Ecuador.

In the study an unsupervised learning methodology to group and classify the time series of energy measurements, using the dynamic time warping technique to discover typical load profiles according to their characteristic weekly consumption, is applied. Next, we perform a restricted space-time analysis to define geographic areas with constant and predictable behavior.

To test the benefit of obtaining this spatial grouping, the information of its members is used to improve the forecast of short-term load using a recurrent neural network with short and long-term memory.

The results of this study have shown that the pattern of energy consumption in nearby areas is related and can be used in models that take advantage of this information. The

temporal analysis of the measurements collected by the smart meters revealed two significantly different monthly consumption profiles. In addition, it was shown that in the case of Ecuador the load profile does not change significantly due to climatic variability, but to temporary effects such as long holidays. The spatial analysis defined 21 geographical areas where all meters have the same consumption behavior in the same period of time, information that was used to improve the forecast of energy consumption of customers in these areas by 2.46%.

As this study has shown, knowledge of the load profile of geographical areas represents a valuable asset for the planning and design of distribution networks, as well as for a more efficient maintenance and operation activities, and limited to specific geographical areas. This information also becomes an important input to define strategies, such as the prioritization of areas for consumer awareness campaigns, or the estimation of future demand factors in specific geographic areas to reduce investment in networks or power plants.

CAPÍTULO 1

1 INTRODUCCIÓN

La electricidad juega un papel clave en el desarrollo económico de la sociedad (Burke, Stern, & Bruns, 2018). La electricidad está relacionada con todos los aspectos del desarrollo, incluida la producción, salud, la educación y la seguridad. La creciente demanda de electricidad puede llegar a acelerarse aún más como resultado de la electrificación del transporte y la climatización, la creciente demanda de dispositivos digitales conectados y, en general, el aumento de los ingresos del consumidor final. El aumento de la demanda de electricidad fue una de las razones clave por las que las emisiones mundiales de CO2 del sector eléctrico alcanzaron un récord en 2018 (Harvey, 2018). Las empresas eléctricas necesitan estar a la vanguardia de los esfuerzos para combatir el cambio climático y la contaminación, gracias a la disponibilidad en el mercado de multitud de tecnologías que generan bajas emisiones de CO2. De igual manera, las energías limpias (solar, eólica, hidráulica), proporcionan una plataforma para reducir las emisiones de CO2 a través de combustibles basados en la electricidad, como el hidrógeno o los combustibles líquidos sintéticos.

Por lo tanto, los gobiernos deberían priorizar la planificación y la formulación de políticas energéticas limpias en las ciudades modernas para lograr un suministro constante y confiable de electricidad a la sociedad. Se estima que la demanda global de la electricidad crecerá a un ritmo mayor que el resto de las energías, más de un 2% anual, situándose con una cuota en respecto al total de energía consumida entre el 24% y 31% para 2040 (IEA, 2019). Actualmente, las potencias económicas mundiales, EEUU y China, se sitúan en las primeras posiciones en cuanto al consumo total de energía eléctrica (ver *Figura 1*) con un total de 3865 y 6510 TWh, respectivamente (Enerdata, 2020)**.** Aunque la mayor parte de la atención de la industria energética está centrada en la producción de energía, la reducción del consumo de energía se presenta como la gran apuesta para el futuro cercano y representa un muy posible cambio de paradigma respecto el constante aumento de la producción.

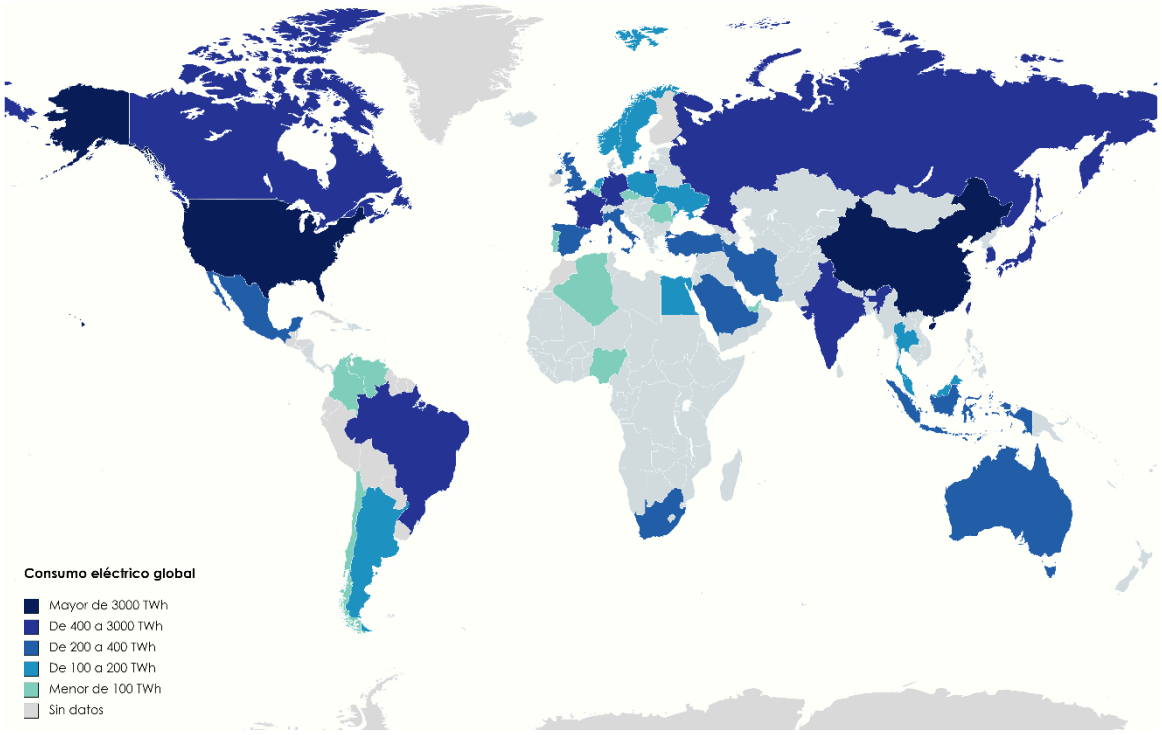

Figura 1. Representación por rangos del consumo de energía eléctrica (TWh) global. Los países en colores más oscuros muestran el mayor consumo **(Enerdata, 2020)**.

Junto con el consumo de electricidad, sus costes económicos han aumentado considerablemente a lo largo de los años debido al rápido incremento de la demanda y a que la electricidad no puede almacenarse ni conservarse de forma eficiente y sencilla. Por tanto, es necesario un análisis de datos a gran escala para identificar dónde se produce el consumo y cómo se puede abordar para reducirlo. Por ello, los métodos que permiten prever el consumo de energía, y sus aplicaciones, se han convertido en un problema activamente investigado. En este nuevo paradigma, donde prima más la eficiencia que la producción, es donde técnicas avanzadas y algoritmos inteligentes de análisis de datos pueden aportar soluciones satisfactorias a la industria. Con el fin de minimizar el consumo de energía, investigadores de diversos campos del conocimiento han juntado esfuerzos y han creado formas de mejorar la concordancia entre la producción y el consumo de energía. La gran cantidad de datos generados hoy en día, así como la evolución de la inteligencia artificial y del análisis espacial, hace necesaria la combinación de estas tecnologías para aportar soluciones en el ámbito.

1.1 Infraestructura de medición avanzada

Una de las tecnologías que permite a las empresas de electricidad avanzar hacia nuevas soluciones que lleven a aumentar la eficiencia energética, la planificación oportuna del servicio y la optimización de la operación, es la implementación de una infraestructura de medición avanzada (AMI).

La infraestructura AMI ayuda a las empresas de distribución de electricidad (ED) a mejorar las estrategias de operación y ahorro de energía por medio de una comunicación bidireccional con el consumidor del servicio eléctrico (usuario). Esto permite transformar al usuario en un agente activo al disponer de la información de su consumo para cambiar su patrón o programar sus dispositivos de control de carga inteligentes (termostatos, lavador o acondicionadores de aire) para que regulen su consumo en función de criterios y directrices predeterminadas (precio por franjas horarias). Desde las EDs, los sistemas AMI apoyan en la aplicación de estrategias de gestión del lado de la demanda para lograr aplanar los picos en la demanda de electricidad (Praveen & Rao, 2020) y reducir el consumo general de electricidad.

Además, AMI agilita la detección y diagnóstico de fallas, se pueden usar los datos AMI para detectar consumos inusuales o interrupciones de energía y agilitar la restauración provocando un ahorro en costos operativos y ayudando a las EDs a mejorar sus indicadores de calidad de servicio. Por otro lado, utilizando la información AMI de las EDs, y técnicas avanzadas de analítica de datos, las agencias de regulación y control de electricidad podrían optimizar los mecanismos de fijación de precios basando sus cálculos en patrones de consumo de energía más reales y dinámicos.

El AMI es una infraestructura que integra una serie de tecnologías que incluye medidores inteligentes, una infraestructura de comunicación y un sistema de gestión de mediciones, más conocido en su término anglosajón como *meter data management* (MDM).

1.1.1 Medidores Inteligentes

Un medidor inteligente, o más conocido en su término en inglés como *smart meter*, es un dispositivo electrónico que registra variables eléctricas (consumo, nivel de voltajes, factor de potencia, entre otros) y las trasmite periódicamente. Este tipo de dispositivos han ido sustituyendo a los clásicos medidores electromecánicos, por ejemplo, en Europa a partir de directivas como la 2009/72/EC para el mercado eléctrico o la 2012/27/EU para la eficiencia energética.

Un medidor inteligente típico consta principalmente de tres módulos: el módulo de medición para detectar las tasas de consumo de energía en tiempo real, ya sea según un cronograma o a petición; un módulo de comunicación para trasmitir datos almacenados y recibir comandos operativos, y un módulo de control que permite a la ED ejecutar los comandos de control como conexión o reconexión de energía y gestión de alarmas. Además, debe disponer de funciones como autonomía en el caso de un corte de energía, sincronización temporal con el resto de medidores, y facilitar al usuario la información para conocer su consumo en tiempo real y tarifa.

Las características principales de los medidores inteligentes de electricidad se resumen en (Mohassel, Fung, Mohammadi, & Raahemifar, 2014):

- Proporcionar datos de consumo para el usuario y la ED.
- Medición neta y tarifa horaria.
- Notificación de fallo e interrupciones.
- Operaciones remotas (activar / desactivar).
- Limitación de carga para fines de respuesta a la demanda.
- Monitoreo de la calidad de la energía que incluye: voltaje y corriente, potencia activa y reactiva, factor de potencia.
- Detección de robo de energía.
- Comunicación con otros dispositivos inteligentes.

Los medidores inteligentes pueden reducir drásticamente muchos de los costos tradicionales de operación, incluida la lectura de medidores, servicios al cliente, corte y reconexión de servicio, gestión de pérdidas no técnicas entre otras funciones.

1.1.2 Comunicación

La arquitectura de comunicación de la red inteligente está definida por el estándar IEEE Std 2030-2011 que es una norma que establece un modelo de referencia de interoperabilidad de redes inteligentes del sistema de energía eléctrica, y proporciona una base para la terminología, las características, el rendimiento funcional y los criterios de evaluación. En esta norma se detalla la infraestructura de comunicación en una disposición de red jerárquica (Khan & Khan, 2013) :

- La red de hogar, o normalmente conocida por los términos anglosajones *home area network* (HAN), es la primera red ubicada en la capa de usuario. Paralelamente a esta, tenemos la red de área industrial, conocida también como *industrial area network* (IAN) y la red de área de edificios conocida como *building area network* (BAN).

- La segunda red, ubicada en la capa de distribución, se denomina red de área amplia, o más conocida por los términos anglosajones wide area network (WAN). Esta red comprende la red de área de vecindario, o neighborhood area

network (NAN), y la red de área de campo, o field area network (FAN). Estas redes tienen un mejor ancho de banda y están equipadas con varios sistemas de control y monitoreo tales como unidades terminales remotas (RTU), medidores inteligentes y unidades de medición de fasores (PMU).

- La llamada red principal, usualmente conocida con los términos en inglés core network, es la que lleva toda la información hacia la empresa eléctrica. Esta red generalmente es privada y de muy alta velocidad, normalmente, implementada con fibra óptica.

La información se toma en los puntos de acceso (medidores inteligentes) y se envía a los puntos de agregación (en una subestación o torre de comunicación) a través de una red de WAN, y luego utilizando la red principal se transmiten los datos a la empresa eléctrica para almacenarlos y validarlos en un sistema de gestión de mediciones (MDM) y desde allí, por medio de interfaces, a las aplicaciones transaccionales de las empresas eléctricas. En la *Figura 2* se muestra un esquema general de la jerarquía de la red en una infraestructura AMI.

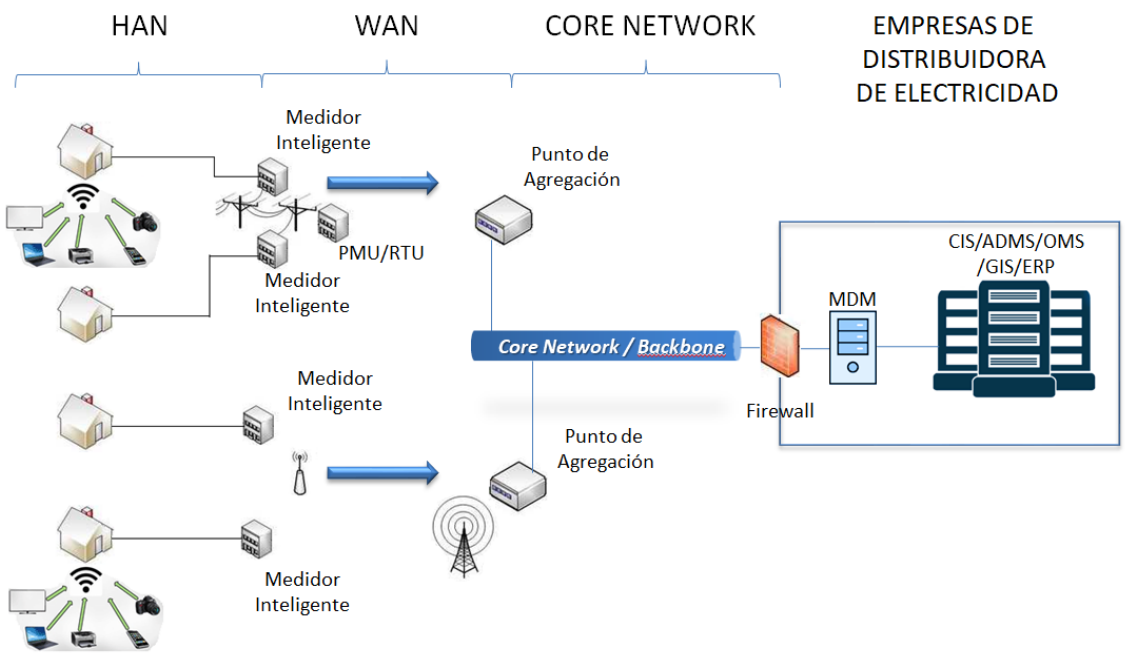

Figura 2. Jerarquía en red de comunicaciones AMI. Se diferencia una red HAN que conecta los dispositivos del hogar a los medidores inteligentes, una red WAN que conecta los medidores inteligentes y la red principal que trasmite esta información a la empresa eléctrica.

1.1.3 Sistema de gestión de mediciones

El sistema de gestión de mediciones, más conocido por las siglas en inglés de *meter data management* (MDM), almacena las mediciones, las valida, limpia y procesa; gestiona el despliegue y monitorea los activos de la red de comunicaciones; verifica la restauración de energía, hace lectura o corte y reconexión de suministro a petición.

Un MDM debe contar con interfaces de programación de aplicaciones a los múltiples sistemas que dependen de los datos del medidor, como por ejemplo los sistemas comerciales, los sistemas de manejo de interrupciones, los sistemas de administración de la distribución y los sistemas de información geográfica.

1.2 Análisis espacial

Un sistema de información geográfica (SIG) es un sistema diseñado para capturar, almacenar, manipular, combinar y analizar datos espaciales, esto significa que esta información esta referenciada con sus ubicaciones en la tierra, comúnmente también llamada información georreferenciada.

Un SIG revela conocimientos más profundos sobre los datos y eventos que suceden en un entorno de estudio utilizando análisis espacial. Con el análisis espacial se puede encontrar relaciones de superposición, proximidad o conectividad, lo que permite descubrir eventos, tendencias o patrones. Esto facilita entender los datos y predecir comportamientos, lo que apoya a la toma de decisiones y el diseño de estrategias. El término sistema en un SIG, hace referencia a que tiene varios componentes que interactúan entre ellos. Los componentes más evidentes de un SIG son el software, hardware y por supuesto los datos georreferenciados. Además, los usuarios y los procesos son componentes esenciales, ya que son los usuarios los que deben aplicar los procesos para actividades como la recopilación de información, el diseño del modelo de datos, la administración de los datos espaciales, el análisis y la comunicación de los resultados, entre otras actividades. Esta combinación de 5 componentes conforma un SIG y da lugar a una potente tecnología analítica con base científica y fiable.

El componente software en un SIG hace referencia a las aplicaciones tecnológicas que proporcionan las funciones y herramientas informáticas para agilitar, automatizar, analizar y mejorar la visualización y comunicación de los resultados. Estas aplicaciones pueden ser comerciales o de acceso libre, para su ejecución en servidores, computadores personales, equipos móviles o en la nube. Dependiendo de estos factores se dispone de diferentes funcionalidades y especialidades científicas.

El software SIG se ejecuta en una amplia gama de tipos de hardware, desde un servidor en la nube hasta un teléfono móvil. Además, se utilizan dispositivos de hardware para recoger la información en el terreno como sistema de posicionamiento global (GPS) y equipos móviles de medición. Existe además información que debe ser

digitalizada por medio de escáner o en la actualidad con métodos modernos de aprendizaje automático como el reconocimiento de imágenes o características. Finalmente, esta información se puede distribuir por medio de impresoras, plotters o servidores de publicación web, entre otros.

Por otro lado, cuando nos referimos a los datos, la mayoría de estos se pueden asociar con ubicaciones geográficas. Un SIG integrará datos espaciales con sus atributos. Los datos espaciales nos dicen dónde ocurre algo y los datos de atributos nos dicen lo que ocurre. La naturaleza de lo que se analiza influye en la forma de representarlo. En su forma más elemental, el mundo real se puede representar mediante dos tipos de datos: i) los datos discretos, almacenados con su ubicación geográfica exacta y que son conocidos en el área como tipo vector, y ii) los datos continuos, representados por cuadrículas regulares conocidos también como datos de tipo ráster. Los datos tipo vector se pueden representar por medio de polígonos para el caso de por ejemplo límites políticos, manzanas o lotes. Para elementos como carreteras, redes de transmisión de energía o ríos se los pueden representar mediante líneas, y por medio de puntos, localizaciones concretas, como la ubicación de un trasformador de energía o un poste eléctrico. Por otro lado, los datos tipo ráster sirven para representar datos no discretos, como por ejemplo la temperatura, elevación o precipitaciones.

Los datos espaciales, adicionalmente a su ubicación, contienen propiedades geométricas y topológicas, y de acuerdo al tipo de dato, las propiedades geométricas pueden incluir perímetro, áreas, distancia, dirección, o volumen. Las propiedades topológicas representan relaciones espaciales tales como la conectividad, la inclusión y la vecindad. Estas propiedades permiten realizar nuevos análisis espaciales, como por ejemplo conocer los clientes afectados aguas abajo de un equipo de corte abierto en una red de distribución de energía.

Los SIG modernos incluyen entre sus funcionalidades el análisis de la información temporal con los datos espaciales. Su objetivo es capturar los aspectos geográficos y temporales de los datos de entrada. Un evento en un conjunto de datos espacio-temporal describe un fenómeno espacial y temporal que existe en un determinado momento *t* y en la ubicación *u*, donde la propiedad espacial es la ubicación, y la

propiedad temporal es la marca de tiempo o intervalo de tiempo para el cual el objeto espacial es válido.

El componente usuarios en el SIG, se refiere a todos los individuos que utilizan el SIG, desde las personas que levantan información en campo hasta los especialistas técnicos que diseñan el modelo de datos espacial o administran el SIG, así como los usuarios que los utilizan para sus tareas diarias o toman decisiones tácticas o estratégicas con los análisis espaciales generados. Cualquiera que sea la aplicación, el usuario es la clave para un SIG poderoso, al igual que la capacitación del usuario y la claridad en las metas institucionales. Por último, un SIG eficiente y escalable se obtiene mediante un diseño y modelado basado en reglas de negocio propias de cada organización, por lo que los métodos y proceso de cada organización deben sustentarse en estándares de la industria y bases científicas.

1.2.1 Aplicaciones del sistema de información geográfica en el área de negocio de servicios eléctricos

Las empresas de electricidad modernas se han concientizado en que utilizar los SIG para tomar decisiones sobre ubicaciones es estratégico para el éxito de su organización, como por ejemplo el lugar más idóneo para ubicar una subestación o un almacén de materiales. Adicionalmente, los mapas de los SIG proporcionan un marco visual para comprender, diseñar y establecer acciones, transformándose en un lenguaje que mejora la comunicación entre departamentos, campos profesionales, organizaciones y público en general.

Las principales áreas de acción del SIG en empresas de distribución de energía eléctrica son la gestión de activos y la operación. Por un lado, los activos, los empleados y los clientes están todos ubicados en algún lugar y deben relacionarse con la ubicación del resto para determinar rutas idóneas, costos y riesgos. El análisis espacial facilita la planificación y gestión de estos activos. El análisis espacial permite responder a preguntas como: ¿dónde están expandiéndose sus redes de distribución de energía?, o ¿cuál es el diseño de la ruta más eficiente de una red de distribución de energía respetando las áreas ambientales protegidas?. Por otro lado, al incorporar el análisis

espacial a los datos provenientes de tecnologías de teledetección como sistemas de control y supervisión o sistemas de gestión de la distribución, se incrementa el conocimiento del comportamiento o la vulnerabilidad de la red, permitiendo una operación más eficiente y segura, además de disminuir los gastos operativos y tiempo de restauración de servicio. Con el análisis espacial y temporal se puede lograr una gestión eficiente de los grupos de operación, de los materiales y las rutas de los vehículos a los lugares de trabajo; en estos ejemplos la ubicación y la optimización de los tiempos son claves. Por otro lado, con un análisis espacial se pueden descubrir las condiciones de riesgo (vegetación, geografía, tráfico, clima, fauna) que amenazan a los trabajadores, instalaciones o medio ambiente para una mejor gestión. Al incorporar a los procesos de las empresas eléctricas el análisis espacial, se logra incrementar la satisfacción del cliente debido principalmente a la atención de incidentes con tiempos de respuesta más cortos, información más clara por medio de mapas fáciles de interpretar y el uso eficiente de recursos.

1.2.2 Análisis espacio-temporal

Cuando observamos un mapa, de forma intrínseca empezamos a analizar su contenido, ya sea hallando agrupaciones, buscando distancias o entendiendo patrones. Este proceso lo hace nuestra mente de forma natural y se denomina análisis espacial. El concepto fundamental del análisis espacial está en apilar capas con distinta información y compararla entre sí basándose en la ubicación geográfica.

El principal beneficio de un SIG consiste en su capacidad para realizar análisis espaciales, integrando las características de los elementos espaciales y sus relaciones con la información estadística de los atributos. De esta forma, la combinación de análisis espaciales como conectividad, vecindad o proximidad, con análisis estadísticos como correlación, medidas de dispersión, análisis de significancia o series temporales permite descubrir relaciones espaciales que de otro modo no serían evidentes.

Actualmente los análisis espaciales que se pueden realizar con un SIG permiten resolver problemas complejos orientados a ubicaciones como modelar eventos, interpretar patrones o tendencias, descubrir o predecir cambios en un contexto

espacial y temporal, lo que brinda nuevas perspectivas de conocimiento para la toma de decisiones.

Como ya hemos mencionado en la introducción, la demanda de la electricidad ha aumentado mundialmente, debido principalmente a la concientización de utilizar energías más limpias, los avances en la tecnología, el internet de las cosas, la introducción de los vehículos eléctricos y un mayor uso de electrodomésticos para el confort en el hogar. Este incremento en la demanda ha provocado que el análisis de la demanda de energía para la identificación de soluciones sostenibles sea esencial para los futuros sistemas de energía (Lund, Østergaard, Connolly, & Mathiesen, 2017). Por otro lado, la demanda total de energía se obtiene con la agregación de la demanda de cada usuario agrupada en un espacio (barrio, provincia o país) y tiempo (horas, meses o años). Por lo tanto, el análisis del consumo de energía se enriquece con información tanto espacial como temporal.

Uno de los factores clave de éxito es una gestión adecuada de los datos y un modelado espacio-temporal preciso. Por ejemplo, un aspecto importante a considerar es la forma en que se definen los datos, ya que puede tener un fuerte impacto en los resultados. Los análisis pueden dar respuestas diferentes dependiendo de si el espacio se evalúa por país, estados, o subestación eléctrica, así como si el tiempo se evalúa por año, mes, día de la semana u hora. Por lo tanto, de acuerdo a cómo se definan los datos se pueden obtener patrones diferentes, más suavizados o aplanados que dependerán de su uso o interpretación.

Una fortaleza que no se consigue con los análisis puramente espaciales, o de series temporales por separado, es la facilidad de descubrir las relaciones de los objetos espacio-temporales con sus vecinos, por su presencia o ausencia en los periodos de análisis. El descubrir estas relaciones permite estudiar tendencias, influencias o patrones en el tiempo.

La frecuente, detallada y elevada cantidad de mediciones recolectadas con tecnologías de medición inteligente y teledetección, juega un papel importante para el monitoreo y supervisión de los comportamientos de consumo de energía. Con esta basta información y la aplicación de técnicas de análisis de datos modernas (análisis espacio

temporal, aprendizaje automático) las empresas de electricidad pueden extraer nueva información útil para la toma de decisiones y desarrollar planes de eficiencia energética y programas de respuesta a las demandas más oportunos y eficientes.

CAPÍTULO 2

2 OBJETIVOS

La identificación de áreas geográficas con un comportamiento predecible de consumo de electricidad puede ser un insumo importante para las empresas eléctricas en actividades como la planificación, operación y mantenimiento, así como realizar análisis técnicos más detallados, como por ejemplo una comprensión más profunda de la demanda de electricidad como en (Cano, Groissböck, Moguerza, & Stadler, 2014) y (Esther & Kumar, 2016), analizar criterios para establecer mecanismos de apoyo al ajuste de la tarifa o tarifas dinámicas como en (Mahmoudi-Kohan, Moghaddam, & Sheikh-El-Eslami, 2010), o para el establecer mecanismos para la eficiencia energética y optimizar las demandas energéticas como en (Kwac, Flora, & Rajagopal, 2014) y (Beaudin & Zareipour, 2015), entre otros. Además, las empresas de distribución de energía eléctrica definen el comportamiento de consumo de sus usuarios con las curvas del perfil de carga y generalmente las empresas definen un único perfil de carga para todos los clientes residenciales y otro para los clientes comerciales. Por lo que poder disponer de un perfil de carga zonificado abriría la puerta a análisis técnicos más precisos a las empresas de distribución de energía.

En consecuencia, el objetivo general de esta tesis es entender si la componente espacial aporta información adicional en la previsión de consumo de energía de los usuarios residenciales.

Este objetivo general se desglosa en tres diferentes objetivos específicos:

- Encontrar nuevos perfiles de comportamiento de consumo para usuarios residenciales, por medio de técnicas de agrupamiento no supervisado de series temporales.
- Utilizar las ubicaciones de los medidores inteligentes y la clasificación obtenida del proceso de agrupamiento del objetivo anterior para encontrar zonas geográficas que muestren un comportamiento de consumo constante y predecible.
- Diseñar una técnica de previsión de consumo semanal de electricidad con granularidad horaria, utilizando redes neuronales recurrentes y la información espacial obtenida de las secciones anteriores.

2.1 Estructura

El resto de esta tesis está organizado de la siguiente manera:

- En el capítulo 3: Estado del Arte. Se presenta el estado actual de los sistemas de información geográfica y las técnicas actuales de inteligencia artificial aplicadas al análisis de la demanda de energía eléctrica.
- En el capítulo 4: Metodología. Se describe la metodología y técnicas utilizadas en este trabajo de investigación, desarrollando las técnicas utilizadas para:
    - El pre procesamiento de los datos.
    - La generación de series temporales.
    - El agrupamiento de series temporales.
    - El análisis espacial y temporal.
    - La implementación de una red neuronal recurrente del tipo LSTM.
- En el capítulo 5: Resultados. Se demuestra con mapas, tablas y gráficas los resultados de la aplicación de la metodología utilizada en la ciudad de Guayaquil (Ecuador) y se definieron zonas geográficas de comportamiento de consumo homogéneo.
- En el capítulo 6: Discusión. Está dedicado a discutir los hallazgos obtenidos en esta investigación. Se argumenta sobre los tipos de comportamiento de consumo característicos de usuarios residenciales de la ciudad de Guayaquil y los beneficios de la definición de sus zonas geográficas.

- En el capítulo 7: Conclusiones. Se señalan las contribuciones de este trabajo y se abordan las perspectivas de futuras investigaciones.

CAPÍTULO 3

3 ESTADO DEL ARTE

En esta sección se presenta el estado actual de los sistemas de información geográfica y la aplicación de técnicas de inteligencia artificial en el ámbito de la demanda de energía eléctrica.

3.1 Evolución del SIG

El SIG evolucionó desde 1960 de un concepto a una ciencia en la actualidad, apoyado por los avances tecnológicos en áreas como el diseño asistido por computadora, la teledetección, la geografía, la estadística, y las ciencias de la computación. En 1963 fue Roger Tomlinson quien diseñó el primer SIG (Foresman, 1998) al planificar y desarrollar un sistema de información geográfica para realizar un inventario de los recursos naturales del gobierno de Canadá al que llamó “Canadian geographic information system” (CGIS). Fue Tomlinson el primero en adoptar un sistema de enfoque de capas para el manejo de mapas y utilizó por primera vez el término “sistema de información geográfica” en su publicación (Tomlinson, 1969). En 1964, Howard Fisher formó en la Universidad de Harvard el laboratorio de computación gráfica y análisis espacial, en la Harvard graduate school of design (Chrisman, 2006), donde se desarrollaron y perfeccionaron nuevos conceptos y técnicas en el manejo y análisis de datos espaciales. Además, allí se desarrollaron aplicaciones de software que sirvieron de base para posteriores desarrollos en entes gubernamentales, centros de investigación y otras universidades.

En la década de los 80 emergieron los primeros sistemas computarizados para uso comercial, los cuales fueron evolucionando desde su funcionamiento en servidores centrales hasta computadores personales y en diferentes bases de datos. No fue hasta la década de los 90 que las aplicaciones SIG despegaron debido principalmente a la proliferación y abaratamiento de los computadores con mejor procesamiento y capacidad de almacenamiento en disco y en memoria, el lanzamiento de nuevos satélites, la integración de tecnología de teledetección y el creciente desarrollo de aplicaciones de software (Waters, 2016).

En el siglo XXI, creció sustancialmente la disponibilidad de datos georreferenciados en una variedad de formatos como: 2D, 3D, radar, láser con tecnología LiDAR (Dubayah & Drake, 2000), por otro lado también decrecieron los costos de equipos para levantamiento de información geoespacial como: drones, cámaras térmicas y GPS de precisión. Estos factores junto con el surgimiento de oferentes de estos datos y servicios, catapultaron la investigación en técnicas de análisis especial y la comercialización de estas (Goodchild, 2018). En la época actual, con el movimiento hacia la computación en la nube y disponibilidad de varias opciones de software libre, el SIG se ha convertido en una herramienta accesible e indispensable para el entendimiento y distribución de la información espacial.

3.2 La inteligencia artificial en la gestión del consumo eléctrico

Los medidores inteligentes proporcionan una granularidad y precisión en los datos que permite superar métodos clásicos de análisis como es el caso de la definición de una única curva de consumo típico para clientes residenciales o comerciales. Pero, esta abundante y constante fuente de información hace que los métodos convencionales de análisis no sean capaces de manejarlos de forma oportuna. Esta situación, junto con la creciente disponibilidad de mejores procesadores y técnicas de análisis más eficientes, ha promovido el uso de técnicas de inteligencia artificial para estudiar los patrones que pueden formar los datos en series temporales. La aplicación de técnicas basadas en datos, como el aprendizaje automático, a la información generada por las redes

inteligentes ha permitido abordar problemas de una manera más personalizada y buscar resultados más precisos.

La literatura muestra que se han aplicado diversas técnicas de aprendizaje automático para analizar datos provenientes de fuentes AMI y otros sensores de red para cumplir con los requisitos de las empresas eléctricas. En la investigación de (Nagi, Yap, Tiong, Ahmed, & Mohammad, 2008) se integra una máquina de vectores de soporte con algoritmos genéticos para detectar pérdidas no técnicas y preseleccionar automáticamente a los usuarios sospechosos de robo de energía para ser examinados en el sitio. También en la investigación de (Monedero, et al., 2012) se utilizan técnicas como el coeficiente de Pearson, redes bayesianas y árboles de decisión para la detección de pérdidas no técnicas. En el trabajo de (Wytock & Kolter, 2014), se desarrolló un método para la desagregación de energía de datos AMI, utilizando supervisión contextual para conocer el uso detallado de la energía de clientes residenciales.

Más específicamente, las técnicas de agrupamiento (Halkidi, Batistakis, & Vazirgiannis, 2001) se pueden utilizar en series temporales de demanda de electricidad diaria para agrupar perfiles similares en los mismos grupos y revelar los perfiles de carga más típicos como en (Chicco, Napoli, & Piglione, 2006), (Hsiao, 2014), (Lavin & Klabjan, 2015) y (Zhou, Yang, & Shen, 2017). Estas técnicas de agrupamiento pueden extraer patrones de consumo en diferentes períodos de tiempo, tales como períodos mensuales en (Gouveia & Seixas, 2016), (Viegas, Vieira, Melício, Mendes, & Sousa, 2016) y (Rhodes, Cole, Upshaw, Edgar, & Webber, 2014), estacionales (invierno, verano, etc.) o anuales en (Kwac, Flora, & Rajagopal, 2014) y (Abreu, Pereira, & Ferrão, 2012). Estas técnicas revelan información importante sobre los hábitos de consumo de los usuarios y su relación con las variables de tiempo.

Con respecto al componte espacial, los hábitos de consumo de energía pueden verse afectados por elementos como barreras geográficas (ríos, bosques), límites políticos (divisiones provinciales o estados), áreas comerciales (parques industriales, zonas de libre comercio o puertos), características del suelo u orientación de los edificios, entre

otros. Los sistemas de información geográfica (SIG) recopilan estos elementos espaciales y sus relaciones para realizar análisis espaciales. Los análisis espaciales han sido utilizados para pronosticar la demanda de energía, por ejemplo, en (Yarbrough, et al., 2015) se utilizó un mapa de calor basado en factores de coincidencia para identificar la demanda máxima. En (Janetzko, Stoffel, Mittelstädt, & Keim, 2014) se aplicó análisis espacial para detectar anomalías en el consumo de energía. En (Tascikaraoglu & Sanandaji, 2016) proporcionaron una metodología que facilita la explotación de las estructuras de baja dimensión que gobiernan las interacciones entre los usuarios de medidores inteligentes residenciales circundantes. Otro ejemplo es el trabajo presentado en (Melo, Carreno, & Padilha-Feltrin, 2012) donde los autores estudiaron cómo se distribuye la carga eléctrica en una ciudad utilizando agentes independientes del área y las relaciones entre diferentes áreas vecinas. También en (Melo, Padilha-Feltrin, & Carreno, 2015), se utiliza regresión espacial para determinar la probabilidad de que las regiones rurales se conviertan en áreas urbanas como parte de la expansión urbana mediante la relación espacial de la carga instalada y las variables socioeconómicas distribuidas en el área de estudio.

La incorporación del análisis de las variables de tiempo con el análisis espacial es lo que se conoce como análisis espacio-temporal y estas tecnologías están basadas en cubos espacio-temporales. El término cubo espacio-temporal se refiere a una representación geográfica donde el tiempo se trata como una tercera dimensión. Uno de los primeros usos fue el del geógrafo Hägerstand en (Hägerstrand, 1970) donde se introdujo el concepto del cubo espacio-temporal para analizar el comportamiento y las interacciones de las personas en el espacio y en el tiempo. Los cubos espacio-temporales se han empleado para varios análisis y visualización interactiva de esta información como en (Kraak, 2003) y (Thakur & Hanson, 2010), y se han presentado conceptos y taxonomías de operaciones estandarizadas como en (Bach, Dragicevic, Archambault, Hurter, & Carpendale, 2014). En (Xu, Yue, Katramatos, & Yoo, 2016), se propone un método autorregresivo vectorial basado en los medidores más cercanos (k-nearest) con entrada exógena, que modela la variación espacio-temporal del consumo de electricidad para la previsión de carga residencial individual. En (Zhang, Feng, & Jian,

2016), los autores aplicaron un análisis temporal y espacial para estudiar los patrones de evolución de energías alternativas y mejorar la planificación y construcción de sistemas energéticos mediante un modelo autómata celular. En (Niu, Wu, Liu, Huang, & Nielsen, 2021) se propone un enfoque de análisis visual de datos espacio-temporales, que integra las series temporales de consumo de energía de medidores inteligentes, en un diagrama de dispersión bidimensional para una exploración visual que permite comprender los comportamientos de la demanda de energía.

Si bien estos estudios han aplicado metodologías de agrupamiento, series temporales y análisis espacial, en nuestro conocimiento hasta el momento, no existe una investigación que incorpore en el mismo estudio la técnica de agrupación de series temporales utilizando deformación de tiempo dinámico (Cuturi & Blondel, 2017) integrada con un análisis espacio temporal para definir perfiles de carga en zonas geográficas específicas como es el caso de esta investigación. Adicionalmente, existen estudios que han demostrado que pronosticar cargas residenciales independientes es más desafiante que pronosticar cargas comerciales o agregadas, como en (Wijaya, Vasirani, Humeau, & Aberer, 2015) y (Sevlian & Rajagopal, 2014). La principal razón de esta mayor complejidad es un incremento en la variabilidad del perfil de carga, normalmente producido por el uso de electrodomésticos que generan fluctuaciones importantes en los patrones de consumo. Estas fluctuaciones son a menudo impredecibles debido a la naturaleza dinámica de los comportamientos de los residentes del hogar. Es por esta razón que existen estudios que demuestran que la hibridación de técnicas de agrupamiento con técnicas de redes neuronales artificiales (ANN) mejora el rendimiento de los estudios, como en (Shahzadeh, Khosravi, & Nahavandi, 2015). Las metodologías de la ANN varían desde implementaciones clásicas como en (Biswas, Robinson, & Fumo, 2016) hasta enfoques ANN más complejos como redes neuronales recurrentes (RNN) con una arquitectura *long short-term memory (LSTM)* como en (Marino, Amarasinghe, & Manic, 2016) o máquinas de Boltzmann restringidas como en (Mocanu, Nguyen, Gibescu, & Kling, 2016) y (Ryu, Noh, & Kim, 2017), entre otras. Sin embargo, estas técnicas híbridas no han incluido el análisis espacio temporal a la predicción con las redes neuronales recurrentes como es el caso de esta propuesta.

CAPÍTULO 4

4 METODOLOGÍA

En esta sección se explican los diferentes métodos y procedimientos que se han aplicado en este estudio. La metodología general propuesta se divide en cuatro fases. Primero, se realiza el pre procesamiento de los datos recogidos por la AMI a partir de los medidores inteligentes, de forma que se obtenga un conjunto de datos limpio para generar series de tiempo que representen el perfil de carga de cada usuario residencial. En segundo lugar, se clasifica a los usuarios por sus perfiles de carga mensuales aplicando una metodología de agrupamiento de series temporales, con el objetivo de encontrar conjuntos de usuarios con tipos de comportamiento de consumo de electricidad similares. Luego, usamos esta clasificación de comportamiento de los usuarios y su ubicación geográfica para realizar un análisis espacio temporal, y encontrar clústeres espaciales de usuarios ajustados a la realidad particular de cada zona geográfica. Finalmente, para comprobar el beneficio de obtener esta agrupación espacial, se utilizan las mediciones de los usuarios en estos clústeres espaciales con un método de aprendizaje automático, específicamente una red neuronal recurrente, para pronosticar el consumo energético de una semana. La *Figura 3*, esquematiza en términos generales las fases de este estudio.

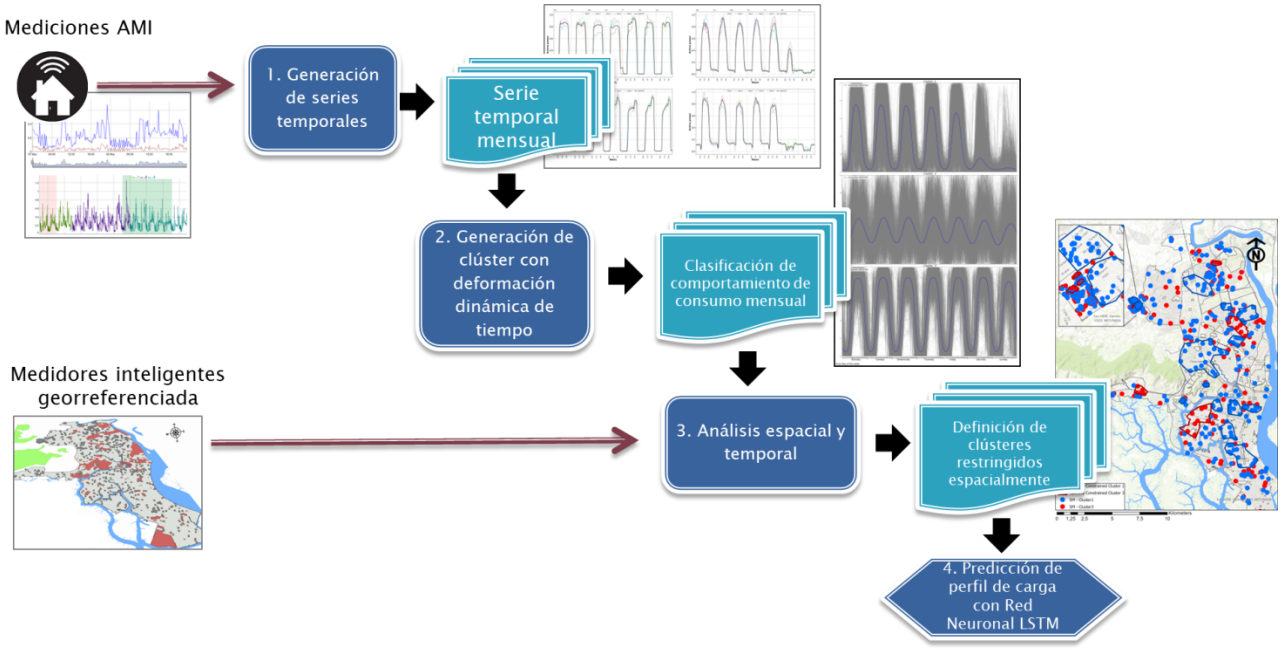

Figura 3. Diagrama general de la metodología empleada.

Fase 1 - Generación de series temporales, Fase 2 - Clasificación de comportamiento utilizando agrupaciones de series temporales, Fase 3 - Análisis espacio-temporal, Fase 4 - Predicción de consumo eléctrico utilizando una red neuronal recurrente.

4.1 Pre procesamiento y generación de series temporales

En esta sección se explican los diferentes procedimientos que se han llevado a cabo en la fase inicial de pre procesamiento de los datos de las mediciones originales. Primero se realiza un análisis exploratorio para definir si los datos fuente contienen valores faltantes o atípicos, y luego se realizan una serie de pasos para obtener un conjunto de datos limpio, completo y consistente (ver *Figura 4*) que servirá de insumo para generar series de tiempo que representen el perfil de carga mensual de cada usuario.

Los experimentos llevados a cabo a lo largo de esta tesis se han basado en una muestra de 1000 medidores inteligentes de usuarios residenciales en la ciudad de Guayaquil (Ecuador). La base de datos comprende un período total de 4 años de 2014 a 2017, con una periodicidad de las mediciones de 15 minutos. Los medidores están georreferenciados con un sistema de coordenadas WGS-1984 en la zona 17 Sur. El total

de registros de la base de datos estudiada representa aproximadamente 130 millones de registros.

Las medidas recogidas incluyen las siguientes variables:

• La posición geográfica de los medidores inteligentes en coordenadas X y Y en los campos:

Coord_x: campo numérico de tipo real.

Coord_y: campo numérico de tipo real.

• Marca de tiempo con la fecha y hora de la medición en el campo:

Fecha: campo de tipo date con formato (DD-MM-AAAA HH:00)

• Código para identificar al medidor inteligente. Respetando las políticas de confidencialidad de los consumidores se recodificó las mediciones para anonimizar la información.

Código: campo de tipo texto de longitud [15 caracteres].

• La potencia activa en kW medida cada 15 minutos:

Activa: campo numérico de tipo real.

Como requisitos imprescindibles en la adquisición de los datos para poder aplicar las técnicas propuestas en esta tesis son que los medidores inteligentes estén georreferenciados y tengan una precisión acorde con la densidad de zona de estudio. En el caso de nuestra zona de estudio se trabajó con una precisión de 5 metros aproximadamente. Por otro lado para garantizar que contamos con información significativa del comportamiento de consumo, que puede estar influenciada por componentes estacionales, es imprescindible contar como mínimo con 10 meses consecutivos de mediciones en cada año de estudio por cada medidor inteligente.

4.1.1 Pre procesamiento de los datos

Un análisis exploratorio señaló valores faltantes y atípicos en los registros de potencia activa en las mediciones de los medidores inteligentes. Se encontraron datos faltantes en períodos cortos de tiempo menores a 2 horas, probablemente debido a fallas o interrupciones en las comunicaciones entre los medidores inteligentes y los concentradores de datos. También se encontró información faltante en períodos más largos de tiempo (días o semanas). Esto se debe generalmente porque algunas residencias eran casas de alquiler, vacacionales o de campo, y sus usuarios acostumbran desconectar la energía cuando están deshabitadas. Para mejorar la información de mediciones de energía activa, se aplicaron los pasos descritos a continuación (ver *Figura 4*) como pasos de pre procesamiento a las mediciones de cada medidor inteligente en la base de datos de 4 años.

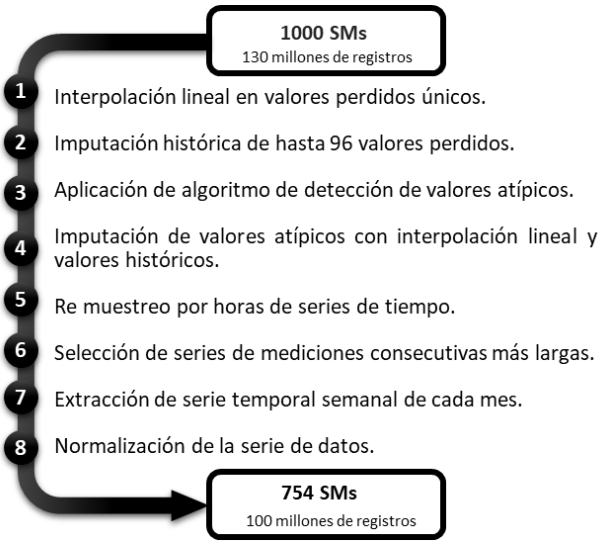

Figura 4. Flujo de pre procesamiento de los datos y generación de serie temporal.

El primer paso consiste en calcular y asignar un valor (imputación) a las mediciones que tengan un solo valor faltante (dato nulo) entre dos mediciones válidas. Esto se lo realiza por medio de una interpolación lineal entre las mediciones anterior y posterior a la faltante. Para los casos en los que no exista mediciones en más de un registro hasta un máximo de 96 registros (1 día) se realiza la imputación de esos valores con las mediciones de energía activa del mismo medidor inteligente a la misa hora y el mismo día de la semana inmediatamente anterior.

Para la detección de los valores atípicos, también conocidos como *outliers,* se definió un límite superior y otro inferior para los valores de energía activa en las mediciones de los medidores inteligentes. Aquellos valores que se encuentren sobre el límite superior o sean menores del límite inferior se consideran *outliers*. El límite superior se calculó con una media móvil centrada de 30 días. A esta media móvil se le sumó el error absoluto medio y un factor de 3 multiplicado por su desviación estándar. El factor de 3 es un resultado empírico ya que, de acuerdo con la varianza de los datos recogidos, los valores más allá de las 3 desviaciones estándar eran erróneos. Para el caso del límite inferior, se definió como límite cero ya que todas las mediciones deben ser positivas. De forma similar al proceso implementado con los valores faltantes, a los *outliers* detectados se les imputa valores haciendo una interpolación lineal entre los elementos que tenga un solo dato fuera de los límites establecidos. Para los casos en los que existan más de un *outlier* se realiza la imputación de esos valores con las mediciones de energía activa del cliente a la misma hora y el mismo día de la semana inmediatamente anterior. En la *Figura 5* se muestran ejemplos de *valores atípicos* detectados con los límites superior, inferior y el valor de media móvil.

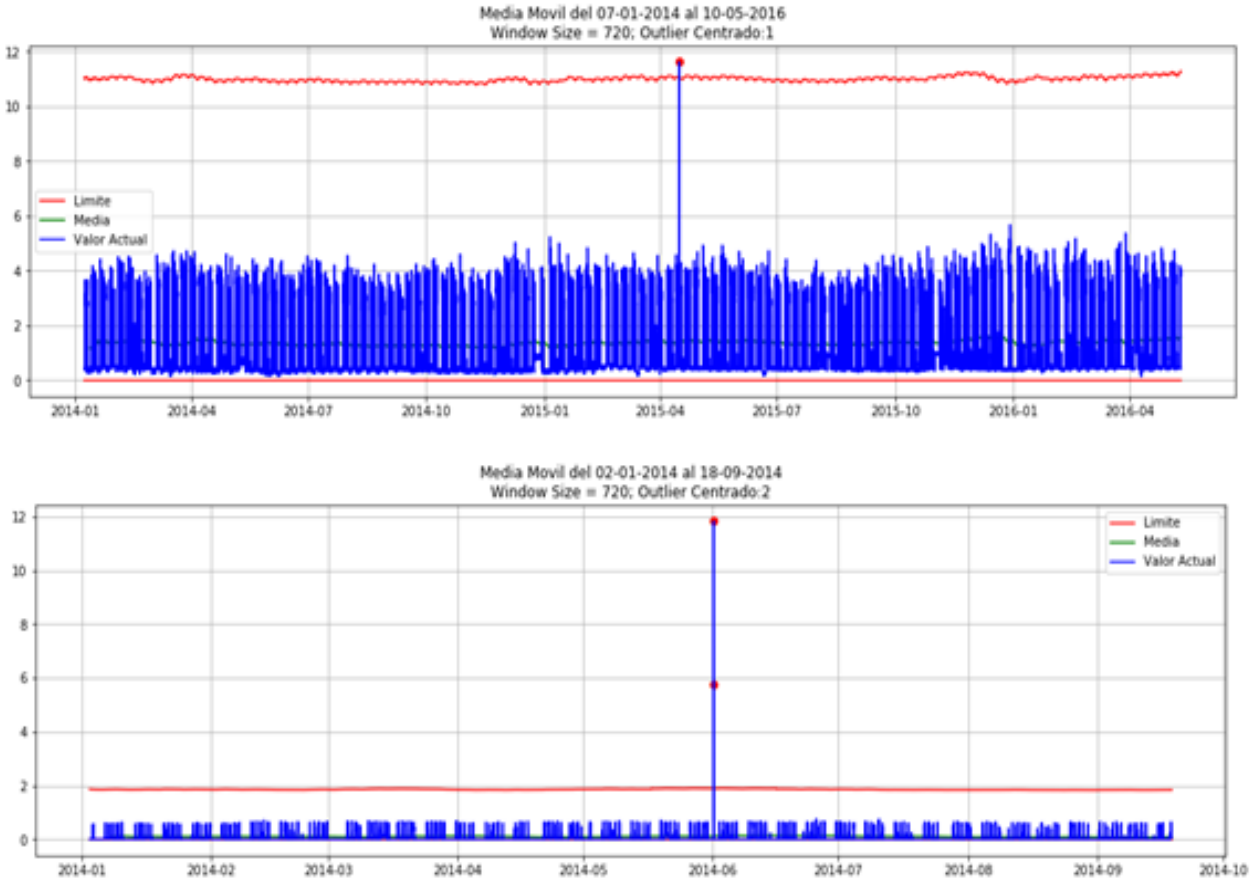

Figura 5. Gráfico de valores atípicos. Los límites superior e inferior se dibujan en color rojo para la detección de valores atípicos.

Debido a que el objetivo de este análisis es modelar el comportamiento de consumo mensual de los usuarios, se simplificaron las mediciones para que estas modelen el comportamiento de forma horaria. Para realizar esta tarea, las mediciones de cada medidor inteligente, recogidas cada 15 minutos, se remuestrearon para obtener mediciones horarias con el valor de potencia activa promedio en este período.

Para garantizar que el análisis de comportamiento se realiza siempre con mediciones correspondientes a un mismo usuario, solo se seleccionaron aquellos medidores inteligentes que, tras pasar por los procesos de limpieza de datos anteriormente descritos, proporcionaban como mínimo 10 meses de datos consecutivos. Es importante que las mediciones sean consecutivas debido a que largos períodos de falta de consumo eléctrico podrían representar un cambio de inquilino o propietario de la vivienda.

4.1.2 Generación de series temporales

Una serie temporal es una sucesión de observaciones de una variable realizadas a intervalos ordenados de tiempo. El análisis en las series temporales se centra en identificar patrones en los datos, es decir, entender lo que sucede en la evolución del tiempo.

Las series temporales están compuestas por la adición de tres componentes: la tendencia, la componente estacional y la componente residual. La tendencia es la componente general de la serie y representa el movimiento anual de una serie independientemente de los otros componentes, es decir modela el nivel subyacente y regular de la serie. La componente estacional son las fluctuaciones reconocibles en los valores de la variable en una periodicidad inferior a un año. Es decir, son oscilaciones en corto plazo que se repiten regularmente. Estas pueden presentarse mensualmente, como es el caso del aumento de la frecuencia de viajes en los meses de verano. También existen estacionalidades semanales o diarias en series horarias, como por ejemplo el cambio del consumo de energía durante los fines de semana. La última componente de la serie temporal es la componente residual, que es el comportamiento de la serie que se debe a pequeñas causas impredecibles que son transitorios e irregulares de la serie.

Entonces dada una serie temporal denotada por y_t, se puede modelar como:

$$\mathrm{y}_t = T_t + E_t + r_t \tag{1}$$

donde:

el subíndice t representa el instante de tiempo (horas, meses, años).

T es la tendencia.

E es la componente estacional.

r es la componente residual.

En este estudio se pretende modelar el comportamiento de consumo mensual de los usuarios por medio de semanas típicas de comportamiento en cada uno de los meses del año. Por lo tanto, a nuestras series de datos les debemos eliminar la tendencia. Un método general para eliminar la tendencia cuando esta evoluciona lentamente en el

tiempo, como es el caso de la electricidad en los consumos residenciales, es la diferenciación. En este método se considera que en el instante *t* la tendencia es similar a la tendencia en el instante $t - 1$. Por lo tanto, si restamos a cada valor de la serie el valor anterior, la nueva serie diferenciada estará sin tendencia. Con respecto a la componente residual esta fue suavizada en el proceso de remuestreo a datos horarios en el pre procesamiento de los datos y puede contener patrones ocultos que serán útiles en el proceso de predicción, por lo que se mantendrán en la serie temporal. Como nuestro modelo es univariante (potencia activa) y estamos modelando a corto plazo (un mes), la media modela suficientemente bien y más eficientemente que otros modelos más complejos, por lo tanto, para obtener un comportamiento típico, se calcula a partir de las mediciones de potencia activa disponibles de cada mes, calculando el valor promedio para la misma hora y el mismo día de la semana. Por ejemplo, la hora 12 de un lunes de marzo se calcula como el promedio de todas las mediciones disponibles para la hora 12 de todos los lunes de ese mes. Este cálculo se realiza de la misma manera para las 168 horas (24 horas x 7 días) de una semana en un mes. De esta manera se generaron una serie temporal que representa una semana característica de consumo eléctrico de un mes (para ver ejemplos de los meses de marzo y de agosto de 2017 ver *Figura 6* y *7*). Por último, y para poder hacer las series de tiempo comparables, se normalizaron las series temporales en el rango entre 0 y 1.

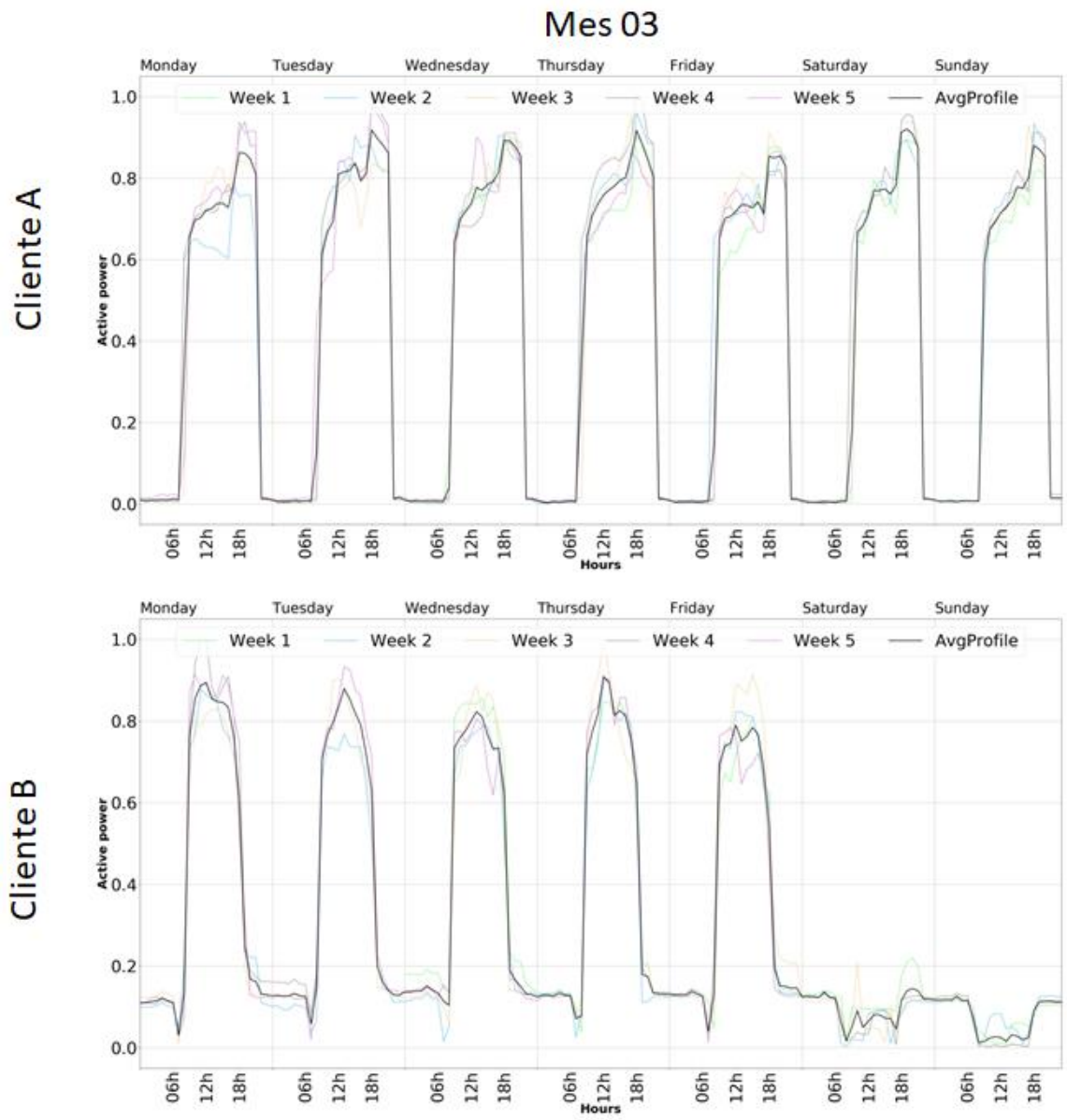

Figura 6. Ejemplo de serie temporal normalizada (a). Generada para el mes de marzo de 2017 para dos medidores inteligentes diferentes.

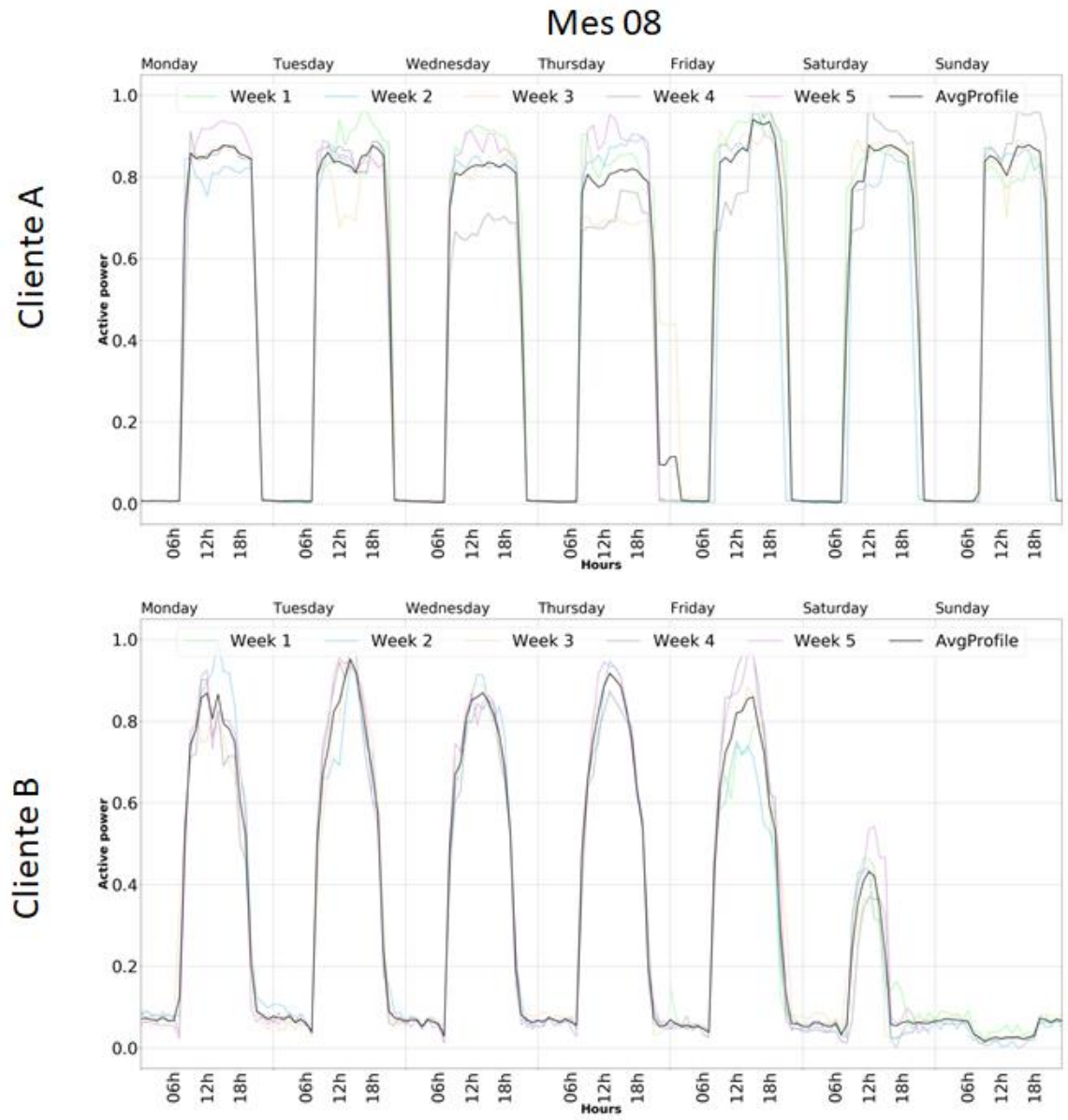

Figura 7. Ejemplo de serie temporal normalizada (b). Generada para el mes de agosto de 2017 para dos medidores inteligentes diferentes.

4.2 Generación de clústeres

En esta sección implementamos una metodología de clasificación no supervisada para agrupar las series temporales generadas en la sección anterior y descubrir los perfiles característicos de comportamiento de consumo de electricidad en usuarios residenciales.

Analizar y extraer información de datos de series temporales pertenecientes a mediciones de medidores inteligentes es una tarea compleja. Se utiliza un conjunto de técnicas de análisis de clústeres para evaluar su desempeño y comprender la estructura macroscópica y las relaciones entre las series analizadas.

En este trabajo se han evaluado cuatro metodologías de agrupamiento de series temporales. En primer lugar, se ha implementado un algoritmo clásico de k-medias basado en la distancia euclidiana como en (Lavin & Klabjan, 2015). El algoritmo de k-medias es una de las técnicas más utilizadas para medir similitudes entre perfiles de carga en clústeres (Chicco, 2012). Sin embargo, debido al uso de la distancia euclidiana puede generar errores considerables al calcular la distancia entre series de tiempo (Hino, Shen, Murata, Wakao, & Hayashi, 2013). Para hacer frente a este reto, se utilizó la técnica de deformación de tiempo dinámica llamada en inglés *dynamic time warping,* (DTW) (Sakoe & Chiba, 1971). DTW tiene un mayor desempeño para la agrupación de series temporales (Petitjean, Ketterlin, & Gançarski, 2011). La métrica DTW ha sido ampliamente aplicada para determinar la disimilitud entre series temporales como en (Aach & Church, 2001) y (Bar-Joseph, Gerber, Gifford, Jaakkola, & Simon, 2002). Esta métrica permite encontrar la distancia mínima entre dos series temporales desplazándose sobre el eje del tiempo, lo que permite agrupar perfiles con formas similares independientemente de su temporalidad. La DTW se basa en la distancia de Levenshtein, introducida por (Sakoe & Chiba, 1971). El objetivo de esta técnica es encontrar la mejor coincidencia entre dos series temporales alineando las coordenadas dentro de ambas secuencias. En la *Figura 8* se muestra un ejemplo de alineación DTW de dos series temporales. El cálculo de la distancia DTW entre dos series temporales estaría dada por:

$$D(A_i, B_j) = d(a_i, b_j) + \min \left\{ \begin{array}{c} D(A_{i-1}, B_{j-1}) \\ D(A_i, B_{j-1}) \\ D(A_{i-1}, B_j) \end{array} \right\} \quad (2)$$

donde:

A_i y B_j son dos subseries de las series temporales con los elementos $A = (a_1, a_{2,} \dots, a_T)$ y $B = (b_1, b_{2,} \dots, b_T)$.

d es la distancia entre los elementos de las secuencias.

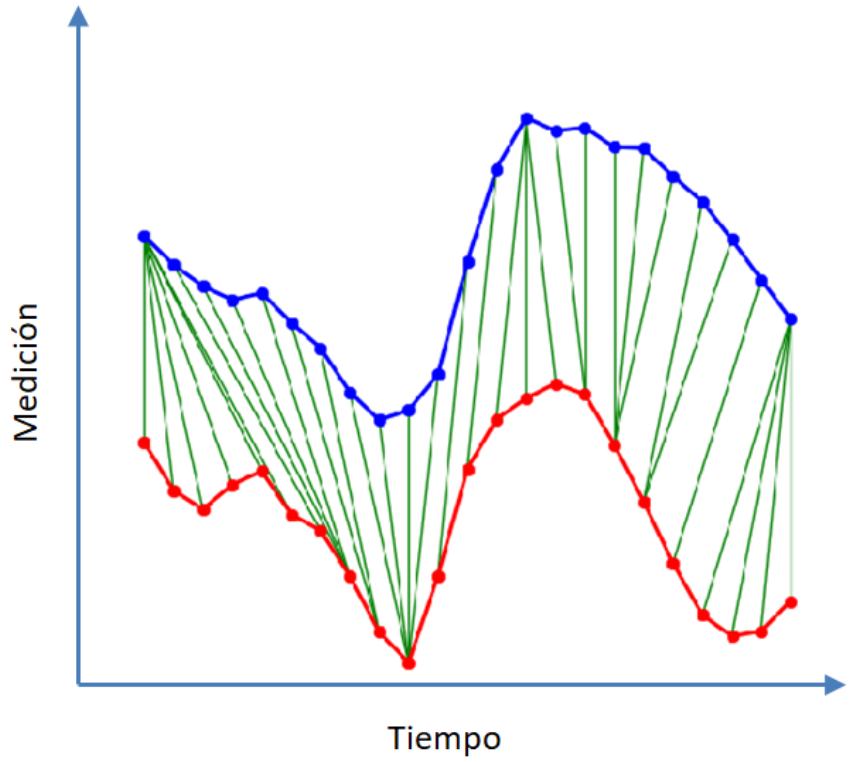

Figura 8. Alineación de dos series temporales utilizando la técnica DTW

Para mejorar el rendimiento de esta clasificación por medio del cálculo de su gradiente se debe definir una formula diferenciable. DTW no es diferenciable debido a la no diferenciabilidad de la operación de minimización *min* en su fórmula (2). *Soft*-DTW es una función de disimilitud diferenciable a la que se pueden calcular su gradiente, por lo tanto, es adecuado para la generación de clúster ya que tiene un rendimiento mejor en comparación con DTW como se demuestra en (Cuturi M. , 2011). El método *soft*-DTW (Cuturi & Blondel, 2017) propone reemplazar la operación de minimización en el cálculo de la métrica en DTW por una operación de minimización suavizada (*soft-min*). En la formula *soft-min* se reemplazan los costos por su exponencial negativa, y las

operaciones mínimo y suma por operaciones suma y multiplicación respectivamente. La formulación de *soft-min* es:

$$min^{\gamma}(a_1, a_{2,} \dots, a_T) = -\gamma \log \sum_i e^{-a_i/\gamma} \tag{3}$$

donde γ es un factor de suavizado.

Entonces, la formulación de *soft-DTW* sería la siguiente forma:

$$\text{soft-DTW}^{\gamma}(A, B) = \min_{\pi}^{\gamma} \sum_{(i,j)\in\pi} \|A_i, B_j\|^2 \tag{4}$$

Además, como se discutió en (Janati, Cuturi, & Gramfort, 2020), *soft*-DTW es sensible a las variaciones del tiempo. *Soft*-DTW, a diferencia de DTW, proporciona una puntuación de similitud promedio ponderada en todas las rutas de alineación en lugar de centrarse en la única mejor alineación. Además, una ventaja que hemos introducido con la implementación de *soft*-DTW es la generación de un efecto de eliminación del ruido cuando esta hace el suavizado de la serie temporal.

Un método de agrupamiento adicional que se evaluó fue la metodología de *k-Shape* (Paparrizos & Gravano, 2015). El algoritmo *k-Shape* es un procedimiento iterativo de refinamiento que tiene como objetivo encontrar clústeres y preservar las formas de las secuencias de las series temporales. Esta técnica aplica una medida estadística de correlación cruzada normalizada para encontrar el centroide de cada grupo y luego, actualiza los miembros de cada grupo. La correlación cruzada es una medida estadística con la que podemos determinar la similitud de dos secuencias $A = (a_1, a_{2,} \dots, a_T)$ y $B = (b_1, b_{2,} \dots, b_T)$, incluso sin estar totalmente alineadas. Esta correlación cruzada mantiene la serie temporal B estática, deslizando A sobre B para calcular su producto interno por cada desplazamiento s de A. Con todos los posibles desplazamientos A_s con $s \in [-T, T]$ se produce $CC_w(A, B) = (c_1, \dots c_w)$. Por lo tanto, la correlación cruzada de la secuencia con longitud $2T - 1$ se define como:

$$CC_w(A,B) = R_{w-T}(A,B), \qquad w \in \{1,2,\dots,2T-1\} \tag{5}$$

donde $R_{w-T}(A,B)$ se calcula como:

$$R_k(A,B) = \begin{cases} \sum_{l=1}^{T-k} a_{l+k} \cdot b_l, & k \geq 0 \\ R_{-k}(B,A), & k < 0 \end{cases} \tag{6}$$

Este algoritmo realiza en cada iteración un paso de asignación y un paso de refinamiento. En el paso de asignación se calcula la disimilitud de las series de tiempo con los centroides utilizando su propia distancia, llamada *shape based distance* (SBD). SBD es una versión normalizada de la medida de correlación cruzada y se utiliza para actualizar las pertenencias del clúster asignando cada serie de tiempo al grupo del centroide más cercano. La normalización consiste en hacer que la media del conjunto de datos de la serie temporal sea 0 y la desviación estándar del mismo sea 1. La distancia basada en la forma da resultados entre 0 y 2, donde 0 indica similitud perfecta para secuencias de series de tiempo y se la formula de la siguiente manera:

$$\text{SBD}(A,B) = 1 - \max_w \left(\frac{CC_w(A,B)}{\sqrt{R_0(A,A) \cdot R_0(B,B)}} \right) \tag{7}$$

En el paso de refinamiento, los centroides del grupo se actualizan para reflejar los cambios en la pertenencia al grupo en el paso anterior, y para determinar el centroide del grupo se selecciona la posición en la que se puede maximizar la correlación cruzada normalizada. El algoritmo repite estos dos pasos hasta que no se produzcan cambios en la pertenencia a los clústeres o se alcanza el número máximo de iteraciones permitidas.

Uno de los factores críticos de éxito para los procesos de agrupación con k-medias consiste en definir el número adecuado de clústeres que se van a generar, para lo cual se utilizó el método del codo (Thorndike, 1953) y el coeficiente de silueta (Rousseeuw, 1987). El método del codo compara las distancias intra cluster ($Inerica$) obtenidas para

diferente número de clústeres (k). La $Inerica$ es la suma de las distancias al cuadrado entre cada elemento (x_i) y el centroide (c) del mismo clúster.

$$Inerica = \sum_{i=0}^{n} |x_i - c|^2$$

Cuanto más grande es el número de clústeres (k), la $Inerica$ tiende a disminuir, lo que significa que los clústeres son más compactos. El método del codo consiste en calcular la $Inerica$ para diferentes valores de k, luego trazar un gráfico lineal que represente la $Inerica$ respecto al número de clústeres k, y por último se localiza el valor k que satisfaga el hecho de que su incremento no mejora sustancialmente su $Inerica$. Este punto se visualiza en el gráfico como un codo y se debe a que su distancia media intra-cluster comienza a disminuir de forma lineal.

Por otro lado, el coeficiente de silueta proporciona una representación de que tan similar es un punto a su propio grupo en comparación con otros grupos. El valor de Silueta para el elemento i está dado por $S(i)$:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

donde:

$a(i)$ = distancia media de i a todos los otros puntos en el mismo clúster.
$b(i)$ = distancia media de i a todos los puntos en el clúster más cercano.

El coeficiente de silueta considera como número óptimo de clústeres aquel que maximiza la media de $S(i)$ de todas las observaciones.

El coeficiente de silueta puede variar entre -1 y 1, los valores cercanos a 1 indican que los elementos están bien clasificados, los valores cercanos a -1 indican una clasificación errónea y los valores cercanos a 0 indican que los elementos están en la frontera de decisión entre dos clústeres.

El método del codo proporciona el candidato al valor de (k), mientras que el coeficiente de silueta mide la calidad de la agrupación, por lo que se utilizó ambos métodos para determinar el mejor valor de (k).

4.3 Análisis espacio temporal

En esta sección realizamos un análisis espacial y temporal de los clústeres descubiertos en la sección anterior, es decir, de los perfiles típicos de comportamiento en usuarios residenciales. El objetivo es determinar las relaciones espaciales y temporales, entendiendo estas como las asociaciones geométricas y temporales entre sus posiciones, momento de las mediciones y sus relaciones con otros eventos, en este caso, la influencia de estos componentes con los clústeres obtenidos.

Una vez agrupados cada comportamiento de consumo mensual en alguno de los clústeres encontrados con el proceso anterior (Sección 4.2), se etiquetan con su respectiva clases de comportamiento típico: [$Tipo_1$, $Tipo_2$, …, $Tipo_n$]. Tras el etiquetado, se realiza un análisis espacio temporal utilizando un sistema de información geográfico. El análisis espacio temporal tiene como objetivo determinar si las similitudes en el patrón de consumo de energía están relacionadas con su ubicación específica o con la proximidad espacial y temporal de otros medidores inteligentes. El análisis híbrido realizado en este estudio, que combina análisis de agrupamiento temporal y espacial, es una de las contribuciones destacadas de este estudio. Este análisis nos permite modelar geográficamente nuestro problema, permitiendo posteriormente explorar, interpretar y detectar patrones importantes ocultos en el conjunto de datos.

La tecnología de análisis espacio temporal, está basada en el uso de cubos espacio temporales y puede analizar de manera integral los patrones, modelando las características espaciales y temporales simultáneamente. Estos cubos permiten representar visualmente el cambio de sus atributos a través del tiempo por medio de mapas que pueden reflejar intuitivamente la distribución y tendencia de los datos en el espacio-tiempo. El término cubo de espacio-tiempo se refiere a una representación

geográfica donde el tiempo se trata como una tercera dimensión. Como se mencionó en la introducción, fue Hägerstand quien introdujo el modelo del cubo espacio temporal, y consiste en representar un elemento en 3 dimensiones (un cubo) que representa las características espaciales en el plano formado por los ejes x, y, mientras que el eje t (la altura del cubo) representa el tiempo. La *Figura 9* muestra un cubo espacio temporal donde cada celda, que recibe el nombre de bin, representa una medición en el espacio (x, y), que se sobreponen uno encima de otro de acuerdo al paso del tiempo.

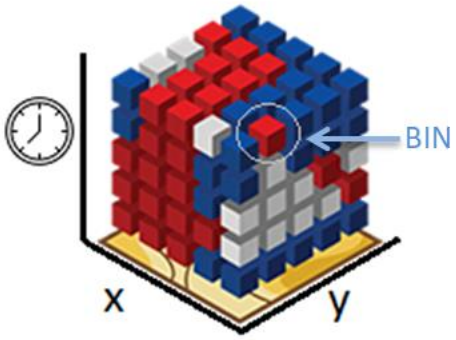

Figura 9. Representación de cubo espacio temporal. Cada bin representa una medición en el espacio x, y que se sobreponen uno encima de otro de acuerdo al paso del tiempo

En este estudio, los $bins$ del cubo espacio temporal que se posicionan en la misma ubicación, definida por (x, y), comparten el mismo código de medidor inteligente. Por otro lado, los $bins$ que abarcan el mismo plano temporal (t) comparten el mismo mes del año en el periodo de análisis. Los $bins$ que están asociados al mismo medidor inteligente representarán una serie temporal. Además, en nuestro caso, cada bin tiene como atributo el tipo de comportamiento encontrado con el proceso de agrupamiento de la sección anterior.

Como ejemplo, se graficó en la *Figura 10* el cubo espacio-tiempo implementado en nuestro estudio, mediante la utilización del software ESRI Arcgis Pro. En esta implementación, los colores de cada bin representan la clasificación del comportamiento encontrada. Los $bins$ representan los comportamientos mensuales en un año, donde los bins más antiguos (primero meses de año) se encuentran en la parte inferior del cubo y los más recientes (últimos meses del año) en la parte superior del

mismo. Se debe considerar que la agrupación visual de los símbolos puede ser un efecto de la perspectiva de la representación de los objetos tridimensional en la imagen bidimensional.

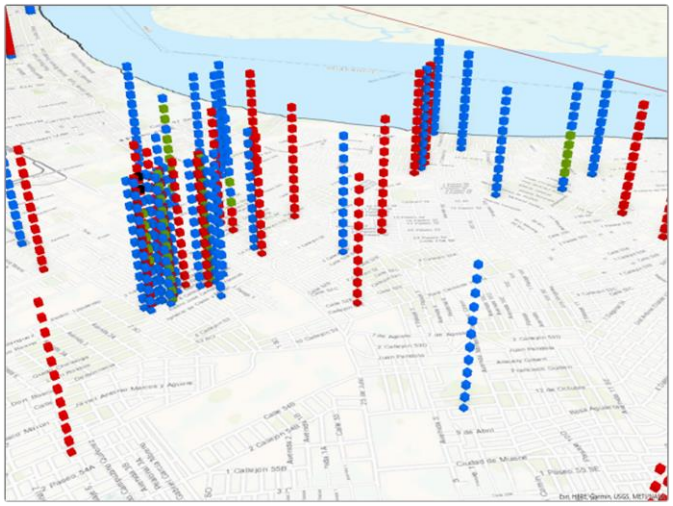

Figura 10. Ejemplo de cubo espacio-temporal. Los colores de cada $\boldsymbol{bin}$ *representan la clasificación del comportamiento. Cada bin representa los comportamientos en mensuales en un año, los* $\boldsymbol{bins}$ *más antiguos (primeros meses del año) se encuentran en la parte inferior del cubo y los más recientes (últimos meses del año) en la parte superior.*

4.3.1 Análisis temporal

Combinando el SIG y la técnica de los cubos espacio temporales se realiza un análisis de series temporales. El análisis temporal consiste en dividir la colección de los cubos generados en función de la similitud de sus características. De esta forma, las series temporales se parecerán si sus valores tienden a cambiar al mismo tiempo, es decir, sus valores se correlacionan a lo largo del tiempo (Montero & Vilar, 2014). A partir del análisis de estas similitudes (ESRI, 2019), se definen las ubicaciones geográficas con su comportamiento típico y se genera una tabla indicando las ubicaciones, es decir los medidores inteligentes, que tienen unos comportamientos constantes o variables a lo largo del período de análisis.

4.3.2 Análisis espacial

De la misma forma que en la Sección 4.2, donde necesitábamos una metodología y una métrica para la generación de clústeres, necesitamos definir una metodología y una distancia apropiada para analizar la formación de estos clústeres espaciales. Por lo que, el siguiente paso fue determinar la distancia donde los procesos espaciales que generan el agrupamiento son más marcados. Para determinar esta distancia realizamos un *análisis de autocorrelación espacial incremental* a las ubicaciones de los medidores inteligentes. El *análisis de la autocorrelación espacial incremental* ejecuta un *análisis de autocorrelación espacial* para una secuencia de distancias en aumento, midiendo la intensidad del agrupamiento espacial para cada distancia.

El *análisis de autocorrelación espacial* es una aplicación del índice global Moran's I, el cual fue inicialmente sugerido por (Moran, 1969) y popularizado por (Cliff & Ord, 1973) con un esquema de ponderación y estadísticas necesarias para identificar la autocorrelación espacial. El análisis de *autocorrelación espacial* permite evaluar si el patrón en las ubicaciones de los medidores inteligentes esta agrupado, disperso o es aleatorio mediante el cálculo de tres variables: i) un índice global Moran's I, ii) una puntuación llamada $z\text{-}score$ y iii) un valor p.

El análisis global Moran's I es una estadística de productos cruzados entre desviaciones de la media de dos ubicaciones (Anselin, 2019). El índice global Moran's I se formula:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{8}$$

donde:

i y j son ubicaciones.

$w_{i,j}$ son los pesos espaciales entre la observación i y j.

S_0 es la suma de todos los pesos espaciales $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} W_{i,j}$.

n es el número total de observaciones.

z es la desviación a la media de los atributos de las ubicaciones, se calcula como $z = x - \hat{x}$, donde x es el valor actual y $\hat{x}$ es la media de los valores.

La intensidad del agrupamiento espacial está determinada por el cálculo de un valor llamado $z\text{-}score$. A medida que aumenta la distancia, también aumenta el $z\text{-}score$, lo que indica la intensificación del agrupamiento. El valor de $z\text{-}score$ para la estadística está formulado como:

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{9}$$

donde*:*

I es el índice global Moran's I

$E[I] = \frac{-1}{n-1}$

$V[I] = E[I^2] - E[I]^2$

El análisis de auto correlación espacial global Moran's I se basa en una hipótesis nula de aleatoriedad espacial, es decir, que se considera que es igualmente probable que un valor ocurra en cualquier ubicación. El cálculo está basado en permutaciones, este calcula una distribución de referencia para la estadística bajo la hipótesis nula de aleatoriedad espacial, permutando aleatoriamente los valores observados sobre las ubicaciones. La estadística se calcula para cada uno de estos conjuntos de datos reorganizados aleatoriamente, lo que produce una distribución de referencia. La distribución de referencia se utiliza para calcular un valor p que sirve para evaluar la importancia del índice. Esto se formula como:

$$p = \frac{R + 1}{M + 1} \tag{10}$$

donde*:*

R es el número de veces que el índice calculado de datos espaciales aleatorios es igual o más extremo que el estadístico observado.

M es igual al número de permutaciones.

Los valores $z\text{-}score$ y p indican si la diferencia entre el dato aleatorio y el estudiado es estadísticamente significativa o no. Los valores del índice no se pueden interpretar directamente, solo pueden interpretarse dentro del contexto de la hipótesis nula. Cuando el valor p devuelto es estadísticamente significativo, se puede rechazar la hipótesis nula, es decir que los valores analizados están más agrupados o dispersos espacialmente de lo que se esperaría si los procesos espaciales subyacentes fueran aleatorios. Con este conocimiento se procede a interpretar el $z\text{-}score$, si este es positivo, es decir, mayor a la media, representa valores muy similares en ubicaciones vecinas, y si el $z\text{-}score$ es negativo, es decir menor a la media, representa valores muy diferentes en ubicaciones vecinas.

El análisis *de la autocorrelación espacial incremental* grafica en un plano cartesiano los valores de distancia en el eje de las abscisas y el $z\text{-}score$ en el eje de las ordenadas como se pude ver en la *Figura 11*. Cuando en el análisis se encuentra un $z\text{-}score$ que es estadísticamente significativo y forma un pico en el gráfico nos indica la distancia donde los procesos espaciales que generan el agrupamiento son más marcados. Un pico se forma cuando existe un aumento en el $z\text{-}score$ seguido luego por una caída del mismo.

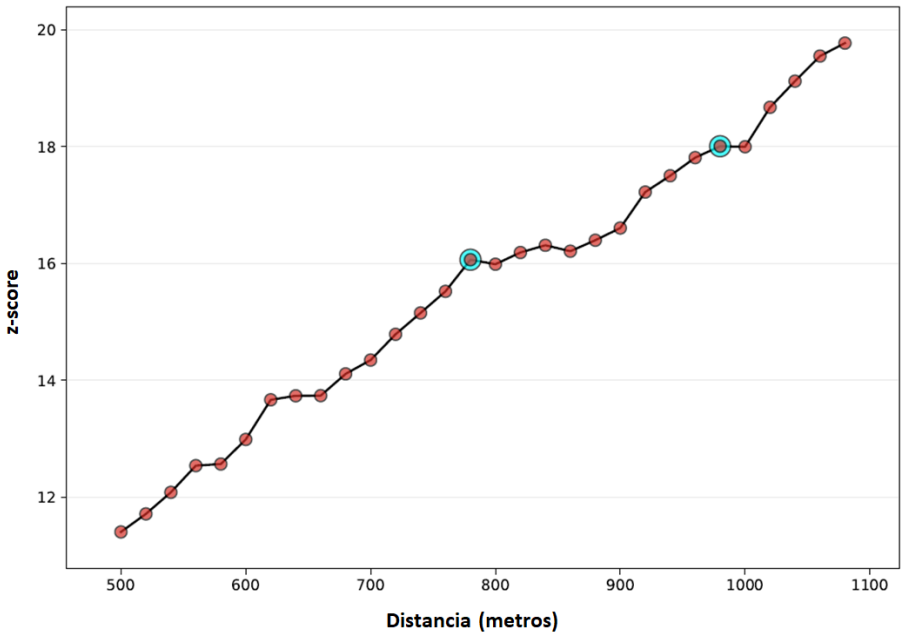

Figura 11. Análisis de la autocorrelación espacia. Los $\boldsymbol{z\text{-}score}$ picos reflejan distancias donde los procesos espaciales que promueven la agrupación son más pronunciados. La herramienta marca con un círculo celeste el primer pico y el máximo picos encontrados.

El *análisis de la autocorrelación espacial incremental* requiere una distancia inicial desde la que se empieza a calcular los incrementos, esta distancia inicial se definió calculando la *banda de distancia a partir del recuento de vecindades*, como se muestra en la *Figura 12*. El cálculo de la *banda de distancia* crea una lista de distancias entre cada entidad y el N-ésimo vecino más cercano. La lista de distancias contiene los valores: mínimo, máximo y promedio. El valor máximo es la distancia calculada para que todos los medidores inteligentes tengan al menos N vecinos. El valor mínimo es la distancia calculada para que al menos un medidor inteligente tenga N vecinos. El valor promedio es la distancia promedio que existe desde cada medidor inteligente a los N vecinos más cercanos. En nuestro análisis utilizamos el valor promedio como la distancia inicial para el *análisis de la autocorrelación espacial incremental*.

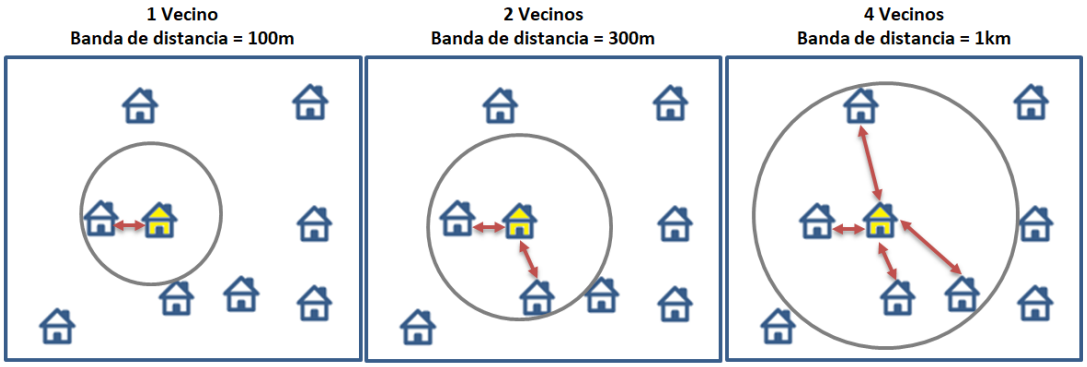

Figura 12. Banda de distancia al N-ésimo vecino.

4.3.3 Análisis de proximidad

Con el valor del $z\text{-}score\ pico$ definido como resultado del análisis de *autocorrelación espacial incremental* de la sección anterior, se utilizó el SIG para realizar un análisis espacial de proximidad. Se establece el valor del $z\text{-}score\ pico$ como la distancia máxima del análisis de proximidad, descartando los medidores inteligentes que están más allá de esta distancia y descartando también los medidores inteligentes que tienen menos de *N* vecinos dentro de esta distancia.

El valor de *N* se definió como 4, por considerar como mínimo los 4 lados de un espacio bidimensional. Por consiguiente, cada clúster tendrá como mínimo 5 medidores inteligentes, es decir el propio medidor inteligente a analizar más sus 4 vecinos.

4.3.4 Análisis de clústeres espacialmente restringidos

El siguiente paso consiste en agrupar espacialmente los medidores inteligentes que pertenece a un mismo tipo de comportamiento, sin ningún otro elemento espacialmente significativo entre ellos, como pueden ser ríos, zonas comerciales o medidores inteligentes pertenecientes a otro tipo de comportamiento. Para este análisis nos soportamos en los análisis espaciales del SIG para primero realizar un agrupamiento restringido espacialmente utilizando un árbol de expansión mínimo con un método llamado SKATER (Assunção, Neves, Câmara, & da Costa Freitas, 2006). SKATER es un método que agrupa elementos espacialmente contiguos que son homogéneos. Para esto primero se crea un grafo de conectividad que captura la relación de vecindad entre los objetos espaciales. El costo de cada borde en el grafo es la disimilitud entre los nodos a los que se une, la estructura del grafo es simplificada mediante un árbol de expansión mínimo, que luego es sucesivamente particionado eliminando las aristas que unen regiones con alta disimilitud, para así obtener los objetos espaciales contiguos que tienen la máxima homogeneidad. Un grafo de conectividad captura las relaciones de adyacencia entre objetos. Cada objeto está asociado con un vértice y vinculado por aristas a sus vecinos. Estas aristas tienen como costo la disimilitud de los atributos entre cada par de vecinos.

Un árbol de expansión es un grafo conectado sin circuitos. Un circuito es una ruta en la que el primer y el último nodo son iguales. Es decir, un árbol de expansión es uno donde dos nodos cualesquiera están conectados por una ruta única, contiene todos los nodos Q del grafo y el número de aristas del árbol es $Q - 1$ (ver *Figura 13*). Con referencia al árbol de expansión *mínimo*, es un árbol de expansión con un costo mínimo, donde el costo es la suma de las disimilitudes en todas las aristas del árbol. El método SKATER poda el árbol de expansión mínimo, produciendo dos árboles no conectados más

pequeños con posibles nuevas particiones y, cuyos bordes unen áreas similares que son candidatos a agrupaciones espaciales.

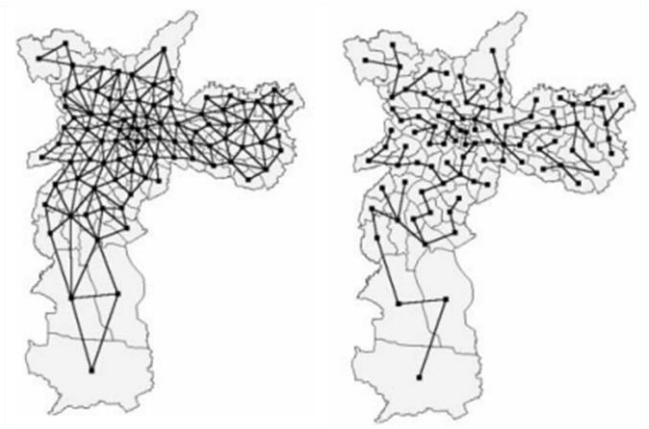

Figura 13. Grafo de conectividad y árbol de expansión mínimo.

A la izquierda un grafo de conectividad. A la derecha un árbol de expansión mínimo. Fuente: **(Assunção, Neves, Câmara, & da Costa Freitas, 2006)**.

Al final del análisis se obtienen los medidores inteligentes que son del mismo tipo y son contiguos en el espacio, etiquetándolos con un identificador (ID) de estos nuevos subgrupos espaciales que llamaremos clústeres espaciales en adelante. En esta nueva clasificación se consideran solo los clústeres espaciales generados que tienen por lo menos N vecinos, el resto de medidores inteligentes se reclasifica en un solo grupo como medidores inteligentes que no generaron un clúster espacial. En el siguiente análisis generaremos las zonas de los clústeres espaciales, restringiendo estas por elementos espaciales considerados significativos. Este análisis considera elementos significativos: las zonas no urbanizables (bosques, zonas comerciales, puertos, entre otros), los ríos y el límite del área de estudio. Primero se generan polígonos de Thiessen (Thiessen, 1911) con las ubicaciones de los medidores inteligente espacialmente clasificados. Los polígonos de Thiessen se crean por la división del plano euclidiano en n polígonos convexos de manera que cada polígono contenga exactamente un punto y cualquier punto dentro de un polígono dado esté más cerca de su punto generador que de cualquier otro (ver Figura 14), definiendo de esta manera su área de influencia. A los polígonos de Thiessen se les substrae las áreas que se sobreponen con los polígonos en

las capas de ríos y de zonas no urbanizable, y se perfila los polígonos substrayendo la zona externa al límite del área de estudio.

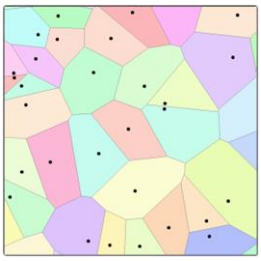

Figura 14. Polígonos de Thiessen. Cada polígono contiene exactamente un punto y el perímetro de estos polígonos es equidistante a los puntos vecinos.

A los polígonos resultantes se les realiza una agregación espacial por el código de clasificación de los clústeres espaciales, de forma que se generan nuevos polígonos más amplios determinado por su ID de clúster espacial. Se cuantifica cuantos medidores inteligentes se ubican dentro de las nuevas áreas agregadas, y se descartan las áreas que contenga menos de N medidores inteligentes dentro de sus áreas. Este proceso es iterativo y requiere adicionalmente la participación del conocimiento de las zonas geográficas para afinar las zonas que cubre los clústeres espaciales definitivos.

4.3.5 Definición de zonas de influencia

Para validar las áreas definidas por los clústeres espaciales, previamente se debe separar una muestra de medidores inteligentes etiquetada con su tipo de comportamiento. La validación se realiza comparando los tipos de comportamiento de los medidores inteligentes de muestra, con el tipo de comportamiento de las zonas de influencia dentro de la que se ubican geográficamente. La zona de influencia de los clústeres espaciales tiene la forma de un círculo con el centro ubicado en centro medio geométrico de sus miembros y un radio igual al doble de la distancia estándar entre esos miembros. La distancia estándar es una métrica de la distribución de entidades alrededor del centro, similar al modo en que una desviación estándar mide la distribución de los valores de datos alrededor del valor medio de la estadística.

En un mapa de dos dimensiones la distancia estándar se formula:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \hat{X}\right)^2}{n} + \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}\right)^2}{n}} \tag{11}$$

donde*:*

x_i, y_i son las coordenadas para cada medidor inteligente i.

$\hat{X}, \hat{Y}$ son el centro medio para los medidores inteligentes en el clúster espacial.

n es el número de medidores inteligentes en el clúster espacial.

El centro medio geométrico está definido en las coordenadas $\hat{X}, \hat{Y}$ y se calculan promediando las coordenadas x y y de todos los medidores inteligentes en el área de estudio. Se formula de la siguiente manera:

$$\hat{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \hat{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{12}$$

donde*:*

x_i, y_i son las coordenadas para cada medidor inteligente *i.*

n es el número de medidores inteligentes en el clúster espacial*.*

De esta manera se extraen subgrupos que se caracterizan por ser directamente colindantes y tener el mismo comportamiento en los mismos periodos de tiempo. Consecuentemente, el siguiente análisis evaluará la hipótesis de que la predicción del consumo de energía de los usuarios, dentro de estas áreas geográficas, estará mejor representado si incluimos en el análisis las mediciones de sus vecinos.

4.4 Predicción de consumo de energía con redes neuronales

En esta sección modelamos el conocimiento extraído de la sección anterior usando una red neuronal recurrente con arquitecta *long short-term memory* (LSTM) para pronosticar los consumos de energía por hora de una semana.

El desarrollo de redes neuronales se remonta a 1958 cuando Frank Rosenblatt diseñó su estructura básica, conocida como el perceptrón, con la que generó una pequeña red neuronal (Rosenblatt, 1958). Desde entonces se han propuesto varias estructuras de redes neuronales más complejas y para diferentes propósitos. En una red neuronal básica una entrada se procesa a través de una función de activación para producir una salida. Sin embargo, para datos correlacionados como las series temporales es necesario relacionar los datos de entrada entre sí. Las redes neuronales recurrentes o *recurrent neural networks* (RNN) tienen la capacidad de trasmitir información de su estado anterior, reflejando las características de una secuencia o de la temporalidad. En las redes neuronales recurrentes se incluye esta entrada adicional (*h)* que se la llama el estado oculto de la red. A estas redes se las llama recurrentes porque ejecutan el mismo proceso para cada elemento de la secuencia de datos de entrada, con una diferencia, la salida de los cálculos anteriores se trasmite a los siguientes pasos mediante el estado oculto. De esta forma, el proceso permite que la información se transmita cíclicamente a través de las neuronas de un paso de la red al siguiente.

En la *Figura 15* se representa la estructura de una RNN como múltiples copias de una misma red con su función de activación *A*, donde cada copia genera una salida y_t procesada para cada nueva entrada x_t en la secuencia, pero adicionalmente por cada paso de tiempo (t) la red emite un nuevo estado oculto h_t como una función de esa entrada y del estado oculto anterior h_{t-1}. Es decir, las RNNs son capaces de conectar información previa a la tarea actual.

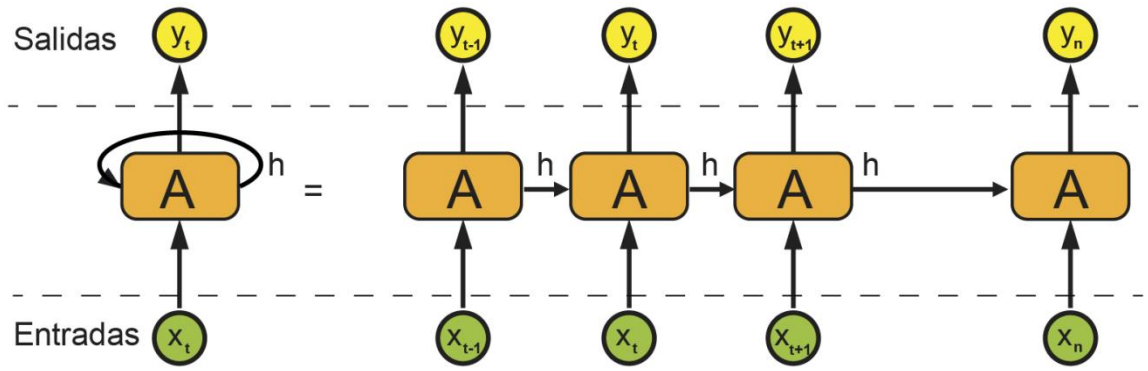

Figura 15. Diagrama de red Neuronal Recurrente. A la izquierda la representación de una RNN y a la derecha la RNN detallada.

A pesar que las RNN exhiben una capacidad superior para modelar secuencias, sufren del llamado desvanecimiento de gradiente durante el proceso de retro propagación explicado en (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001) y (Bengio, Simard, & Frasconi, 1994). Esta desventaja las hace incapaces de aprender dependencias a largo plazo, es decir, relaciones entre entidades que están separadas por varios pasos temporales. El algoritmo de retropropagación (Werbos, 1974) compara las salidas finales y_t de una red neuronal con una muestra preseleccionada de validación de los datos. Esta muestra es utilizada para retropropagar el error y ajustar los pesos desde las últimas capas hacia las primeras con el objetivo de minimizar el error. Hochreiter demuestra en su artículo (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001) que cuando las redes neuronales tienen múltiples pasos, el gradiente del error disminuirá exponencialmente con cada paso en el proceso de retropropagación, por lo que el entrenamiento de una red neuronal recurrente básica con una dependencia a largo plazo se vuelve muy lento y no es capaz de ajustar los pesos correctamente.

Para resolver este problema, Hochreiter y Schmidhuber diseñaron las LSTM (Hochreiter & Schmidhuber, 1997), un tipo especial de redes neuronales recurrentes. La principal ventaja de las redes LSTM es que puede evitar eficazmente el problema del desvanecimiento de gradiente (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001). Las redes LSTM, al igual que las RNN, tienen una estructura en forma de cadena, pero en lugar de tener una sola función de activación en su celda de memoria, las redes LSTM tienen tres estructuras llamadas compuertas (de olvido, de entrada y de salida), a través

de las cuales se puede eliminar o agregar información a un *estado de la celda* (c_t). El estado de la celda es como una banda transportadora que recorre directamente a lo largo de cada celda de memoria, con interacciones en cada una de ellas que no le afectan exponencialmente.

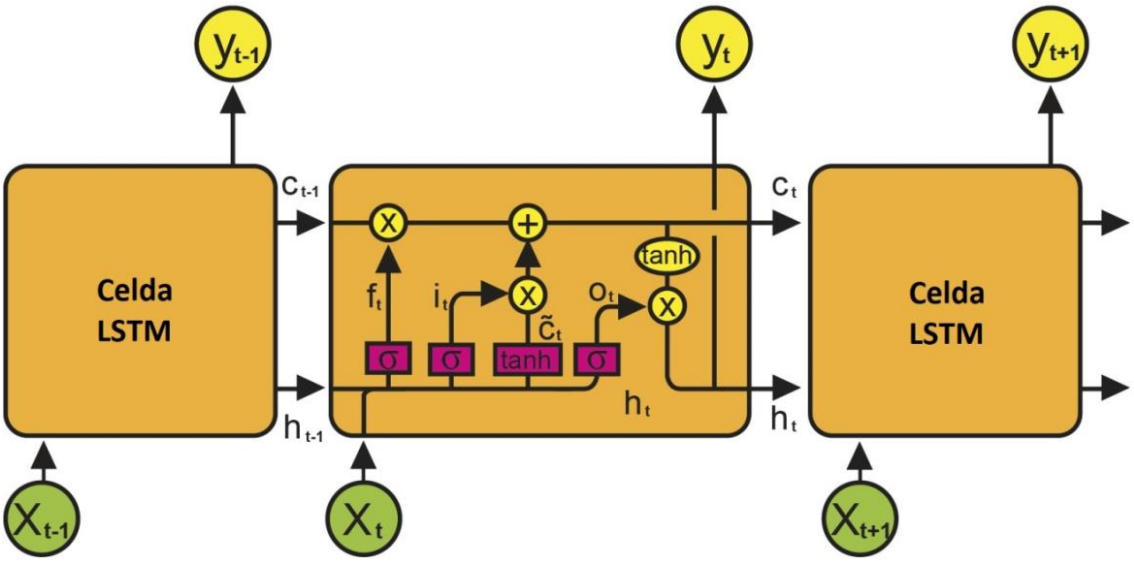

Figura 16. Estructura de una celda de memoria LSTM. La celda contiene una compuerta de olvido $(\boldsymbol{f_t})$, una compuerta de entrada $(\boldsymbol{i_t \cdot \tilde{c}_t})$, y una compuerta de salida$(\boldsymbol{o_t})$.

En la *Figura 16* se representa la estructura de una celda de memoria en una red LSTM donde las líneas representan el flujo de los datos, que pueden bifurcase representando una copia, o unirse representando una concatenación. Las figuras rectangulares de color rosado representan estructuras de redes neuronales con solo una función de activación. Las figuras circulares de color amarillo representan operaciones entre vectores. Las celdas LSTM tiene 3 entradas: el dato de entrada de la secuencia (x_t), el estado oculto de la celda anterior en las redes neuronales recurrentes (h_{t-1}), y el estado de la celda anterior de las redes LSTM (c_{t-1}).

La compuerta de olvido (f_t) define qué información se eliminará del estado de la celda. Esta compuerta toma (h_{t-1}) concatenado con (x_t) y, por medio de una capa sigmoidea, genera valores entre 0 y 1 que multiplicará por la entrada (c_{t-1}) que, recordemos, es el estado de la celda anterior. Las multiplicaciones por 0 eliminarán la información de

(c_{t-1}) y las multiplicaciones por 1 mantendrán las mismas. Esta primera compuerta se puede formular de la siguiente manera:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{13}$$

donde:

σ representa la función de activación sigmoidal.

W_f son los pesos.

b_f es el sesgo.

La compuerta de entrada define la nueva información que se agregará al estado de la celda. Para esto, primero, una capa de tangente hiperbólica $(\tilde{c}_t)$ crea un vector de nuevos valores candidatos entre -1 y 1, que serán agregados al estado de la celda. Estos valores se multiplican por una capa sigmoidea (i_t) que determina el nivel de actualización de los valores del estado de la celda de forma similar a como se explicó con la compuerta de olvido (f_t). Esto se formula de la siguiente manera:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{14}$$

$$\tilde{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{15}$$

donde:

σ representa la función de activación sigmoidal.

W_i son los pesos de capa sigmoidea (i_t).

W_c son los pesos de capa hiperbólica $(\tilde{c}_t)$

b_i y b_c son sus respectivos sesgos.

La compuerta de salida calcula el nuevo estado oculto (h_t) a ser trasmitido a la siguiente celda de memoria. La salida se basa en el estado de la celda al que se le aplica una operación $tanh$ y se multiplica por la salida de una capa sigmoidea (o_t) aplicado a la concatenación de (h_{t-1}) con (x_t).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{16}$$

$$h_t = o_t \cdot tanh(c_t) \tag{17}$$

donde:

σ representa la función de activación sigmoidal

W_o son los pesos

b_o es el sesgo.

El nuevo estado de celda (c_t) que saldrá de la celda de memoria LSTM será la multiplicación de f_t por el estado de celda anterior (c_{t-1}) , y sumado a la multiplicación de (i_t) con $(\tilde{c}_t)$.

$$c_t = f_t \cdot c_{t-1} + \ i_t \cdot \tilde{c}_t \tag{18}$$

La estructura de las celdas de la red LSTM permite el aprendizaje de dependencias a largo plazo, es decir, pueden reconocer el contexto de toda la secuencia, razón por la que se manifiesta que estas redes tienen una memoria a largo plazo. Esta característica ha popularizado el uso de las redes LSTM especialmente en el campo de reconocimiento de patrones, en el análisis de series temporales y el procesamiento del lenguaje.

En este último paso de nuestro análisis validamos la contribución de este estudio utilizando los resultados del análisis espacial y temporal como un componente de una red neuronal recurrente de tipo LSTM.

La red LSTM se aplica con dos enfoques diferentes para comparar las predicciones resultantes. En el primer enfoque, se codificó una red LSTM que utiliza solo las mediciones del mismo medidor inteligente en un análisis univariante, es decir que la red LSTM solo tiene como entrada de datos la secuencia de una variable. En el segundo enfoque, se implementó un análisis multivariante en la red LSTM. En este último caso, además de las propias mediciones del medidor objetivo, incluimos las mediciones de N medidores inteligentes que se encuentran dentro del mismo clúster espacial.

A medida que se agregan más capas ocultas LSTM, la red podrá inferir comportamientos más complejos en nuestra serie temporal y aumentar la precisión de la predicción, por lo que en nuestro modelo se utilizan dos capas ocultas LSTM asumiendo comportamientos diarios y semanales. Al final se incluye una capa de salida con 168 neuronas, que coincide con el número de pasos horarios que debemos predecir en una semana (24 horas multiplicado por 7 días). Además, se alterna una capa de abandono de neuronas (dropout) entre cada capa LSTM con el fin de agilizar el aprendizaje y evitar un sobre ajuste. Las capas dropout consisten en capas que solo actualizan un porcentaje de los pesos de las neuronas en las iteraciones de entrenamiento de la red, mientras el resto permanecen constantes. El entrenamiento de la red neuronal se establece con un máximo de 100 épocas con parada anticipada si el error cuadrático medio no mejora en 5 épocas. La red LSTM se esquematiza en la *Figura 17*.

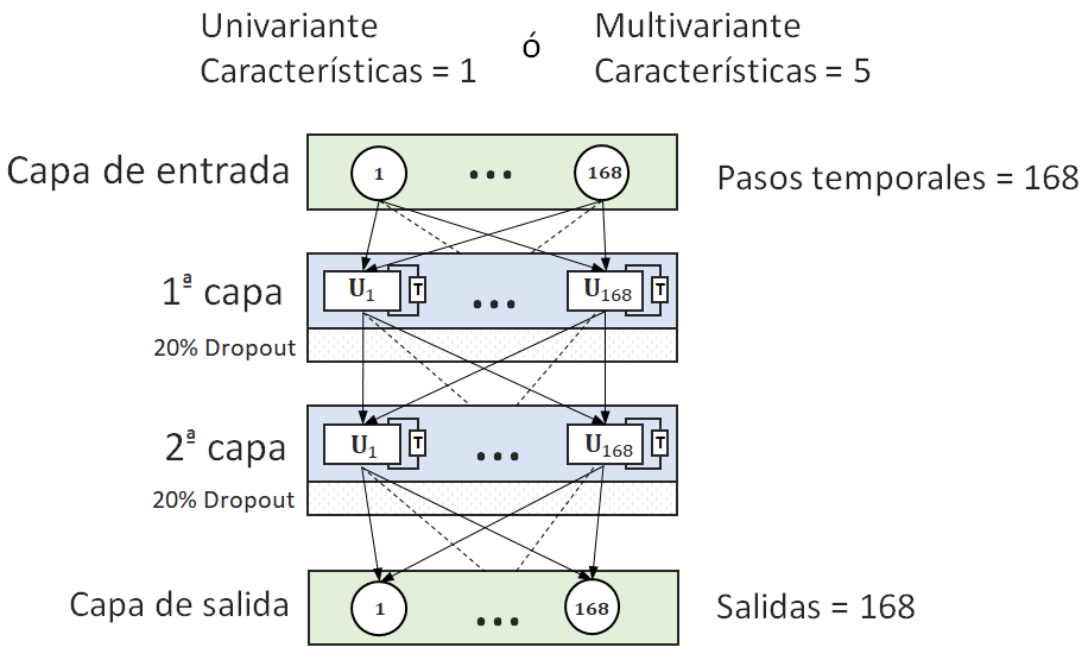

Figura 17. Red LSTM. Red neuronal recurrente con 2 capas LSTM y dropout de 20% por cada capa.

4.5 Métricas de rendimiento

El rendimiento del modelo de pronóstico se midió utilizando 2 métricas: la raíz del error cuadrático medio y el error porcentual absoluto medio simétrico llamadas en inglés respectivamente *root mean squared error* (RMSE) y *symmetric mean absolute percentage error* (sMAPE).

La métrica sMAPE se define como:

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{(|y_t| + |\hat{y}_t|)/2} \tag{19}$$

donde:

y_t es el valor real.

$\hat{y}_t$ es el valor pronosticado.

Se utilizó sMAPE como una métrica relativa para poder comparar las diferentes predicciones de los consumos eléctricos. Debido al hecho de que los valores de las series temporales son positivos, esta métrica tiene la ventaja de que tiene definidos el límite inferior (0%) y el límite superior (200%). Aunque se tiene conocimiento de la ligera asimetría que favorece a los modelos que sobreestiman el pronóstico con sMAPE, se asume este sesgo de pronóstico para este estudio. En esta métrica, para los casos en que la observación y la predicción suman cero, el valor es indeterminado y, por tal razón se ignora estos casos en el cálculo del sMAPE.

La validación de las zonas de los clústeres espaciales definidos se la realizó por medio de indicadores de clasificación de especificidad, precisión y exhaustividad, comparando los valores reales del tipo de comportamiento del medidor inteligente de muestra, con el tipo de comportamiento de la zona de influencia dentro de la que se ubica geográficamente. Las fórmulas se indican en la *Tabla 1*:

$$Especificidad = \frac{TN}{TN + FP}$$	(20)
$$Precision = \frac{TP}{TP + FP}$$	(21)
$$\text{Exhaustividad} = \frac{TP}{TP + FN}$$	(22)
$$Fscore = \frac{2 * \text{Exhaustividad} * Precision}{(\text{Exhaustividad} + Precision)}$$	(23)

Tabla 1. Métricas de rendimiento empleadas. Se abrevian los términos: verdadero positivo (TP), verdadero negativo (TN), falso positivo (FP) y falso negativo (FN).

CAPÍTULO 5

5 RESULTADOS

La metodología propuesta en la sección anterior se aplicó utilizando el software Python y la biblioteca TSLearn para la generación de clústeres de series temporales (Tavenard, et al., 2020) y la biblioteca Keras (Keras, 2021) para la aplicación de las redes neuronales. Para realizar los análisis espacio temporales se utilizó el Sistema de Información Geográfica ESRI®ArcGIS Pro.

Aunque nuestra base de datos inicialmente incluye las mediciones de 1000 medidores inteligentes (SM), el porcentaje de medidores inteligentes disponibles con respecto al número total de medidores en la ciudad de Guayaquil es significativamente bajo. La ciudad tiene una concentración muy alta de usuarios eléctricos, llegando a más de 700.000 clientes en un área urbanizada de aproximadamente 190 km2. Por tanto, la muestra está bastante dispersa, principalmente por la existencia de contadores electromecánicos que aún no han sido sustituidos. Por lo tanto, nuestro conjunto de datos inicial será analizado estrictamente a través de diferentes procesos para obtener un conjunto de datos final enfocado en validar los objetivos de este estudio. Después del pre procesamiento descrito en la sección 4.1.1 de la metodología, se obtuvo un total de

754 medidores inteligentes con al menos 10 meses de mediciones consecutivas al año, lo que representó una base de datos con aproximadamente 100 millones de registros.

Con el fin de mejorar la legibilidad y seguimiento de este estudio, la *Figura 18* muestra las fases de cada paso que se desarrollan en el resto del documento.

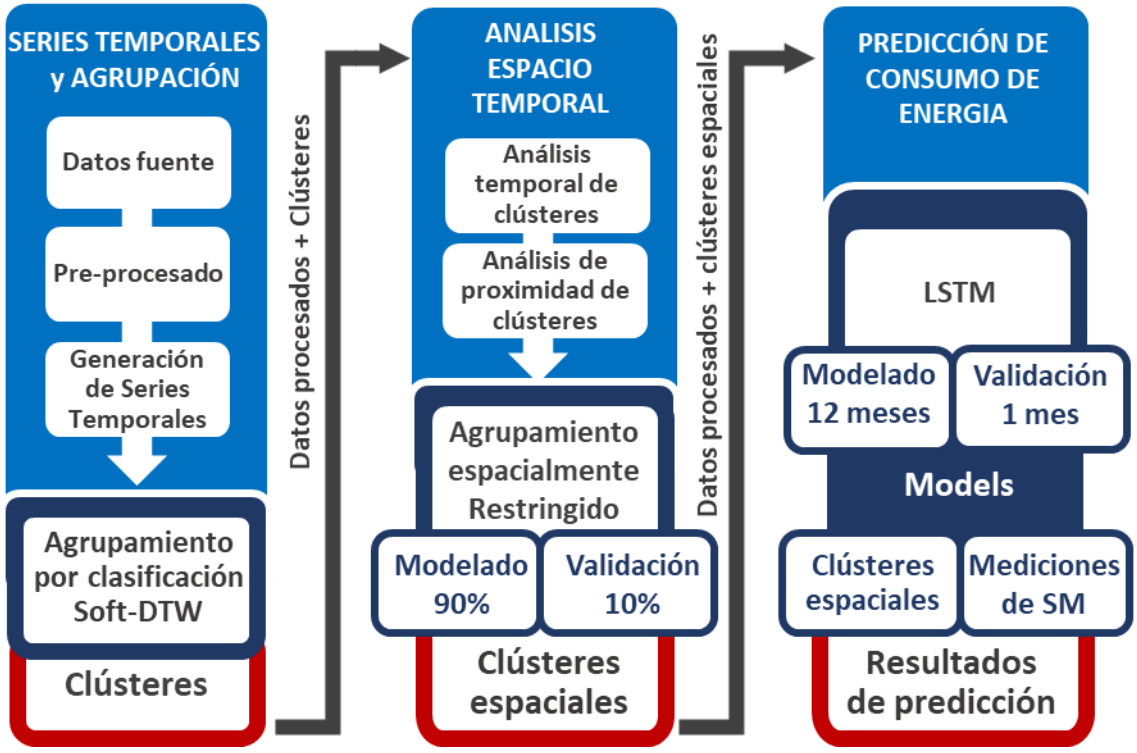

Figura 18. Fases de cada paso del análisis.

5.1 Clúster de series temporales

La generación de los clústeres de las series temporales se realizó utilizando los métodos K-Shape y el método k-means con las métricas indicadas en la Sección 4.2 (euclidiana, DTW y *soft*-DTW). Todos los métodos se ejecutaron con tres diferentes valores de clústeres objetivo (k = 3, 4, 5). Para el caso de *soft*-DTW se utilizaron 3 valores del hiperparametro gamma γ = [0.5, 1, 2]. Los clústeres generados se validaron con el método del codo y el coeficiente de silueta.

Utilizando el método del codo se determinó que el mejor agrupamiento es de 3 y se utilizó el coeficiente de silueta para obtener una representación de qué tan bien se encuentran los datos dentro de su grupo. En la *Tabla 2* se presentan los resultados de los métodos y número de clústeres analizados, ordenados de mejor a peor desempeño según el coeficiente de silueta:

Métodos de agrupamiento	**k**	**Coeficiente de silueta**
k-means (*Soft*-Dtw γ = 1)	3	0,5775
k-means (*Soft*-Dtw γ = 2)	3	0,5297
k-means (*Soft*-Dtw γ =2)	4	0,5280
k-means (DTW)	3	0,5234
K-Shape	3	0,5213
k-means (*Soft*-Dtw γ = 1)	4	0,5187
k-means (Euclidean)	3	0,5132
K-Shape	5	0,5119
k-means (*Soft*-Dtw γ = 0,5)	3	0,5083
K-Shape	4	0,4907
k-means (*Soft*-Dtw γ = 2)	5	0,4479
k-means (DTW)	4	0,4235
k-means (*Soft*-Dtw γ = 1)	5	0,3714
k-means (*Soft*-Dtw γ = 0,5)	5	0,3675
k-means (*Soft*-Dtw γ = 0,5)	4	0,3617
k-means (Euclidean)	4	0,3341
k-means (Euclidean)	5	0,3023
k-means (DTW)	5	0,2755

Tabla 2. Resultados de coeficiente de silueta en métodos de agrupamiento aplicados. El valor k lista la cantidad de clúster generados por cada método.

Como muestran los resultados de la tabla anterior, el método k-means con métrica *soft*-DTW con k=3 y γ=1 es el que mejor se desempeña en comparación con el resto de métodos y parametrizaciones analizadas. Por consiguiente, se clasificó las mediciones con método. La *Figura 19* muestra en color azul los baricentros obtenidos de la

aplicación de *soft*-DTW con k=3 y γ=1. Estos baricentros se utilizarán como lo perfiles de comportamiento característico de los tres clústeres encontrados.

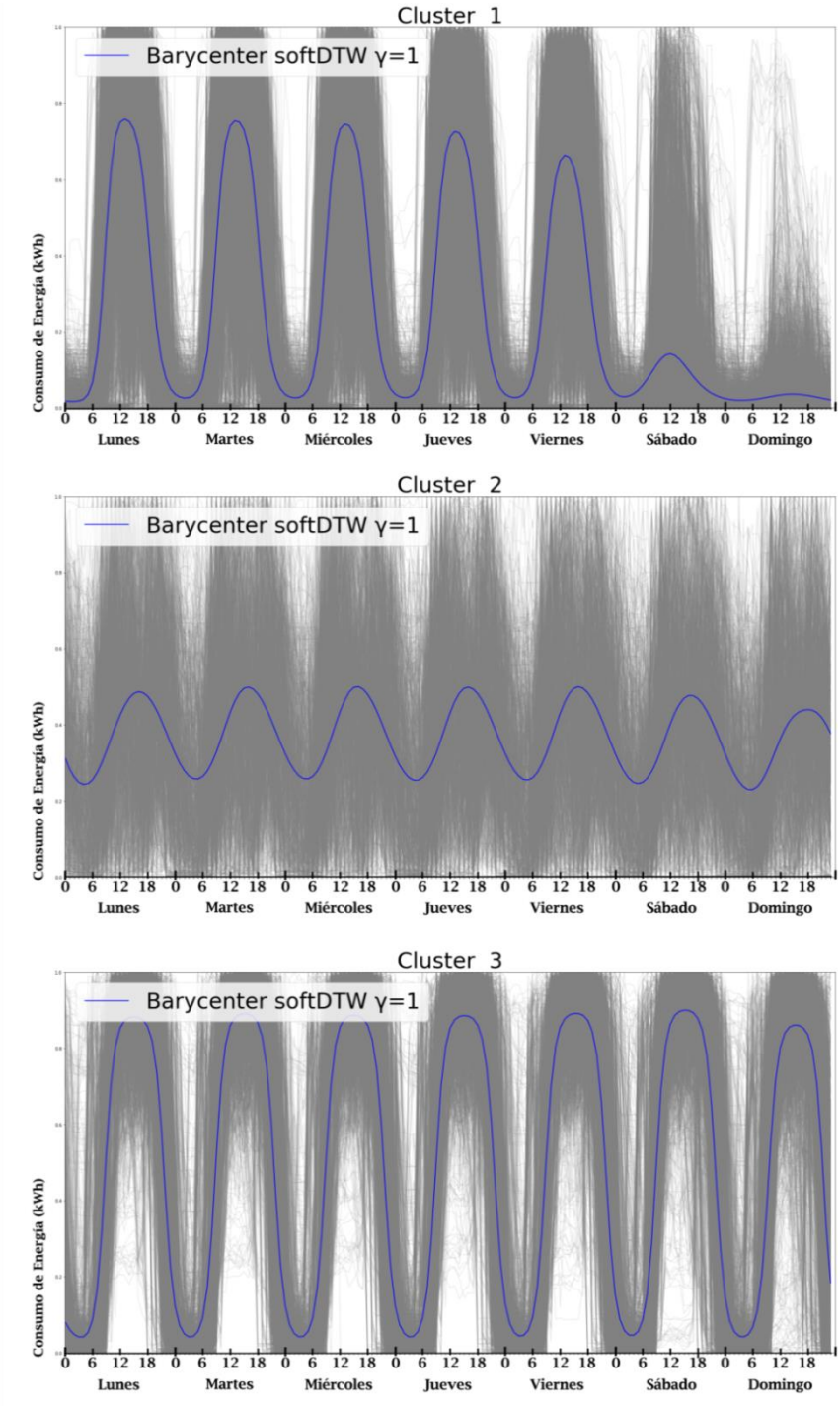

Figura 19. Series temporales agrupadas. Las series temporales son horarias y modelan una semana típica de consumo por cada mes de cada medidor inteligente. Las series temporales se representan en gris y en azul el baricentro de cada clúster.

5.2 Análisis espacio-temporal

Disponiendo de los medidores inteligentes georreferenciados y su comportamiento de consumo clasificado en los tres clústeres resultantes del análisis anterior, se utilizó el sistema de información geográfico para realizar un análisis espacio temporal y explorar el patrón de comportamiento de consumo de los medidores inteligentes.

En primer lugar, para cada medidor inteligente se analiza el comportamiento de consumo mensual definido por su clasificación en los clústeres temporales. Es decir, se determina si cambian de clúster durante los años o, por el contrario, permanece constante. Para este análisis, se representa mediante cubos espacio temporales superpuestos el comportamiento de consumo de cada medidor inteligente en un mapa. El período de tiempo del 2014 al 2017, representado en cubos mensuales es bastante extenso, y da como resultado una superposición significativa de símbolos en un mapa. Por esta razón, en la *Figura 20* solamente se visualiza la clasificación de los medidores inteligentes del año 2017. En la mencionada figura se puede observar que en la ubicación de cada medidor inteligente se apilan hasta 12 cubos que representan el comportamiento de consumo de cada medidor inteligente en los 12 meses del año, comenzando desde la parte inferior con enero y concluyendo con diciembre en la parte más alta. Como se puede notar la mayoría de los medidores inteligentes tiene un comportamiento constante, es decir, mantienen el mismo comportamiento de consumo durante todos los meses, y predominan los clientes de tipo de comportamiento 1 y 3, existiendo muchos menos medidores inteligentes que combinan el comportamiento con los de otros tipos, principalmente con el tipo 2.

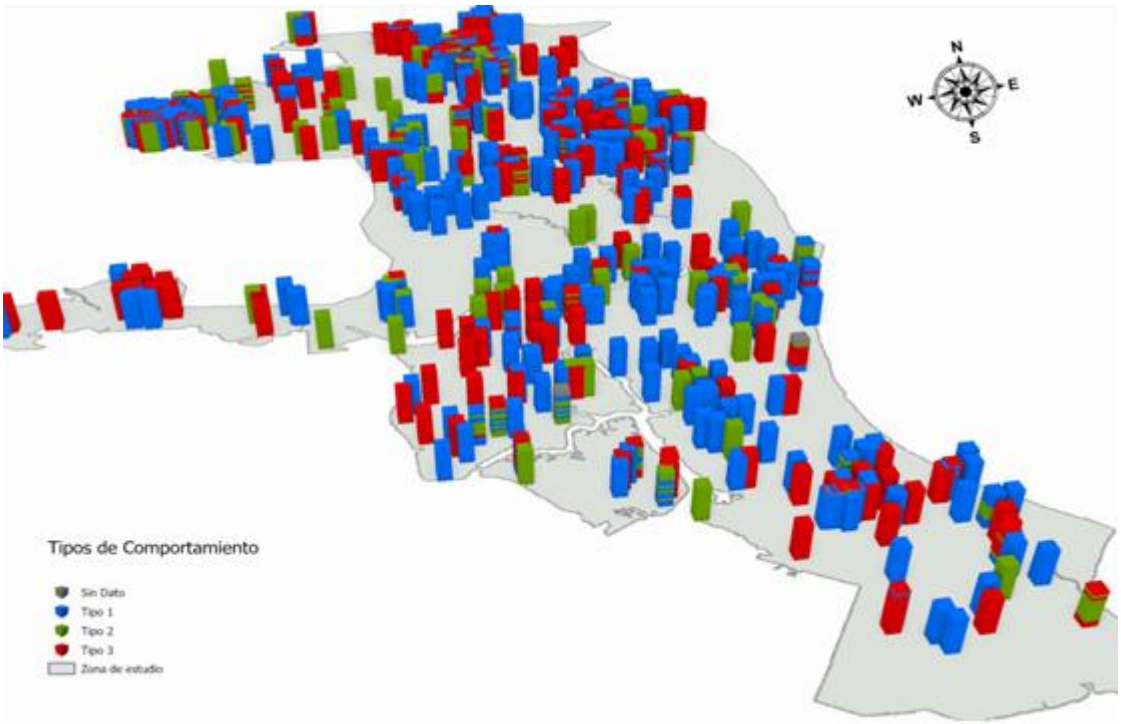

Figura 20. Mapa de cubos espacio temporales. Los cubos representan los tipos de comportamiento mensual en el año 2017, el color de los cubos representan los tipos de comportamientos y en gris la ausencia de datos.

En la *Tabla 3* se presenta un resumen de los resultados, se pudo observar que, de los 754 medidores considerados, 594 medidores inteligentes tienen información completa y un tipo de comportamiento constante todos los años, estos representan el 78,78% del total de medidores inteligentes. De manera similar, hay otros 82 medidores inteligentes, que representa el 10.88% del total, que tiene entre 10 a 11 meses de información completa por año y además mantienen un comportamiento constante todos los años. Los 78 medidores inteligentes restantes, que representan un 10.34% del total, cambian su tipo de comportamiento en diferentes ocasiones en los años.

Meses con información completa por año. (2014-2017)	Tipo de comportamiento	Cantidad de medidores inteligentes	Porcentaje
12 meses	Constante	594	78,78%
Entre 10 y 11 meses	Constante	82	10,88%
Entre 10 y 12 meses	Variable	78	10,34%
Total		754	100,00%

Tabla 3. Resultado del análisis de los cubos espacio-temporales. Tipo de comportamiento, indica si el SM es constante en su tipo de comportamiento (pertenecía al clúster) en el transcurso del período de análisis.

Para el siguiente paso se seleccionan solo los medidores inteligentes que contienen al menos 10 meses de información completa por año y mantienen un comportamiento constante. Comportamiento constante quiere decir que pertenece al mismo tipo de comportamiento todos los meses durante todo el período de análisis (2014-2017). Estos medidores representan el 89,66% de los medidores inteligentes analizados en esta sección, dando un total de 676 medidores inteligentes. En la *Figura 21* se puede diferenciar en el mapa los medidores constantes y variables en su comportamiento.

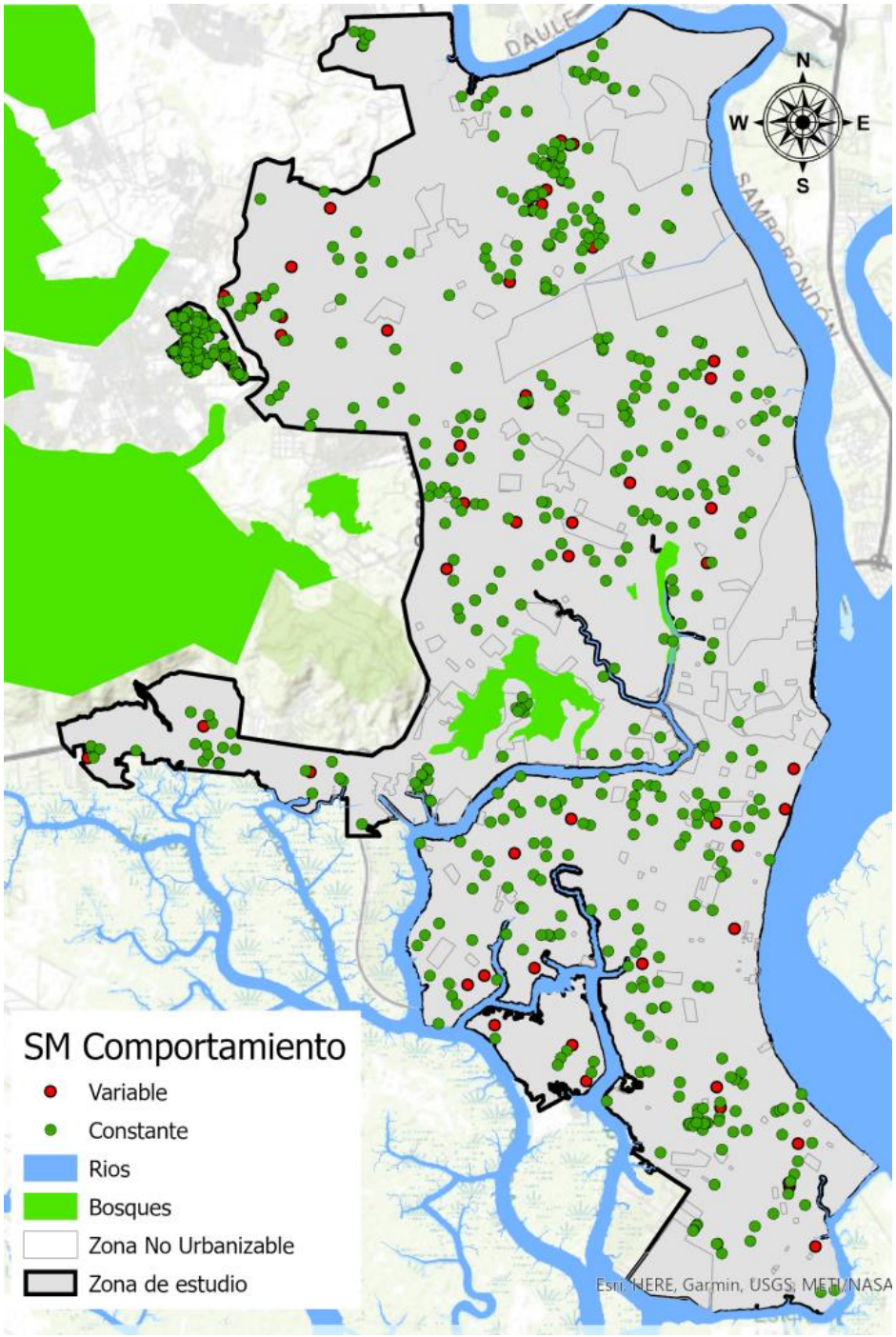

Figura 21. Mapa temático de ubicación de medidores inteligentes. El color diferencia el comportamiento constantes o variables en el periodo de análisis.

Una vez seleccionados los medidores inteligentes con comportamientos constantes, debemos constatar que tengan los suficientes vecinos para probar nuestra hipótesis. Para esto se realiza el análisis espacial de proximidad definido en la Sección 4.3.3 de la metodología. El cálculo del radio medio en el que un medidor inteligente tiene 4 vecinos se calculó en 980m, que se ha redondeado a 1 km. La *Tabla 4* muestra los resultados

obtenidos con este criterio. La mencionada tabla muestra que 20 medidores inteligentes quedan separados más de 1km de otros medidores (menos de 4 vecinos). Estos 20 medidores inteligentes, la mayoría clasificados como tipo 2, se descartan del análisis debido principalmente a que pertenecen a casas en el campo que están aisladas, de los cuales no podremos obtener datos suficientes de sus vecinos espaciales para realizar un agrupamiento restringido espacialmente, que es el siguiente paso de esta fase de análisis. La *Figura 22* muestra un mapa con los medidores inteligentes aislados que se descartaron.

Clúster	Total	> 1 km	< 1 km
Tipo 1	370	2	368
Tipo 2	74	16	58
Tipo 3	232	2	230
	676	20	656

Tabla 4. Resultado del análisis de proximidad. Un total de 20 medidores inteligentes están fuera de la distancia de análisis (1km).

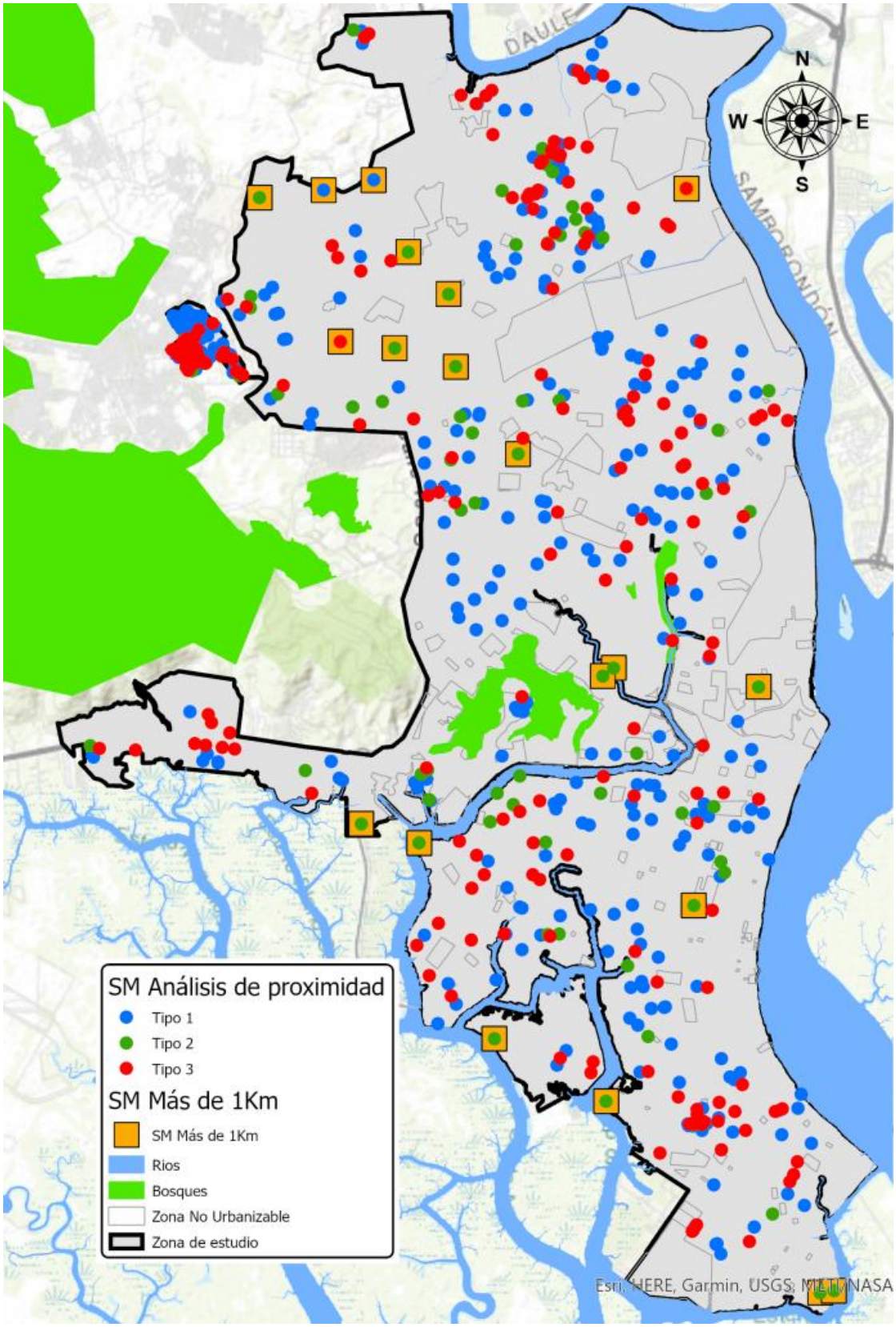

Figura 22. Medidores inteligentes aislados. Los medidores que tienen menos de 4 vecinos en una distancia de 1 km están representados con un cuadrado amarillo que los contiene.

Los resultados del análisis de proximidad de la *Tabla 4*, respecto a los medidores inteligentes aislados que están principalmente clasificados de tipo 2, incentivaron a profundizar más en el análisis de la consistencia de los elementos pertenecientes a cada clasificación. Por esta razón, se compararon las métricas RMSE y sMAPE entre la serie temporal de cada medidor inteligente y el baricentro del clúster en el que fueron clasificados cada mes. La *Tabla 5* muestra las métricas de comparar el baricentro de los clústeres con sus integrantes. El RMSE medio de los medidores inteligentes del tipo 2 es mayor que el doble del RMSE del tipo 1 y del tipo 3. De manera similar, el sMAPE es un 17,47% y 22,43% mayor que el sMAPE del tipo 1 y el tipo 3, respectivamente. Estos valores apuntan a que los miembros del tipo 2 tienen una alta variabilidad y representan aquellos medidores inteligentes que no fueron identificados completamente en los tipos 1 o 3. Por esta razón y dado que nuestro objetivo es enfocarnos en conjuntos de medidores inteligentes con ubicaciones contiguas y comportamiento similares, se excluyó del análisis a los medidores inteligentes clasificados en el tipo 2.

Clúster	RMSE	sMAPE	SMs	Proporción
Tipo 1	0,5541	45,78	368	56,10%
Tipo 2	1,5594	63,25	58	8,84%
Tipo 3	0,6078	40,82	230	35,06%
		Total:	656	100,00%

Tabla 5. Métricas RMSE y sMAPE. Análisis entre la serie temporal de cada medidor inteligente y el baricentro del clúster en el que fueron clasificados.

Luego de suprimir los medidores inteligentes del tipo 2, el total de medidores inteligentes seleccionados para continuar con el análisis de clústeres espacialmente restringidos es de 598. Para validar los resultados de la generación de clústeres restringidos espacialmente, se necesitó separar una muestra aleatoria del 10% de estos medidores inteligentes que representa 59 medidores. Por lo que la cantidad final de medidores inteligentes disponibles para realizar el análisis espacialmente restringido es de 539 medidores inteligentes.

$$SM(Tipo1) + SM(Tipo3) - SM(Muestra) = 539\,SM$$

En la *Figura 23* se grafica los medidores inteligentes seleccionados para el agrupamiento restringido espacialmente, puntos azules representan los medidores inteligentes con el comportamiento clasificado por el tipo 1 y en rojo los pertenecientes al tipo 3.

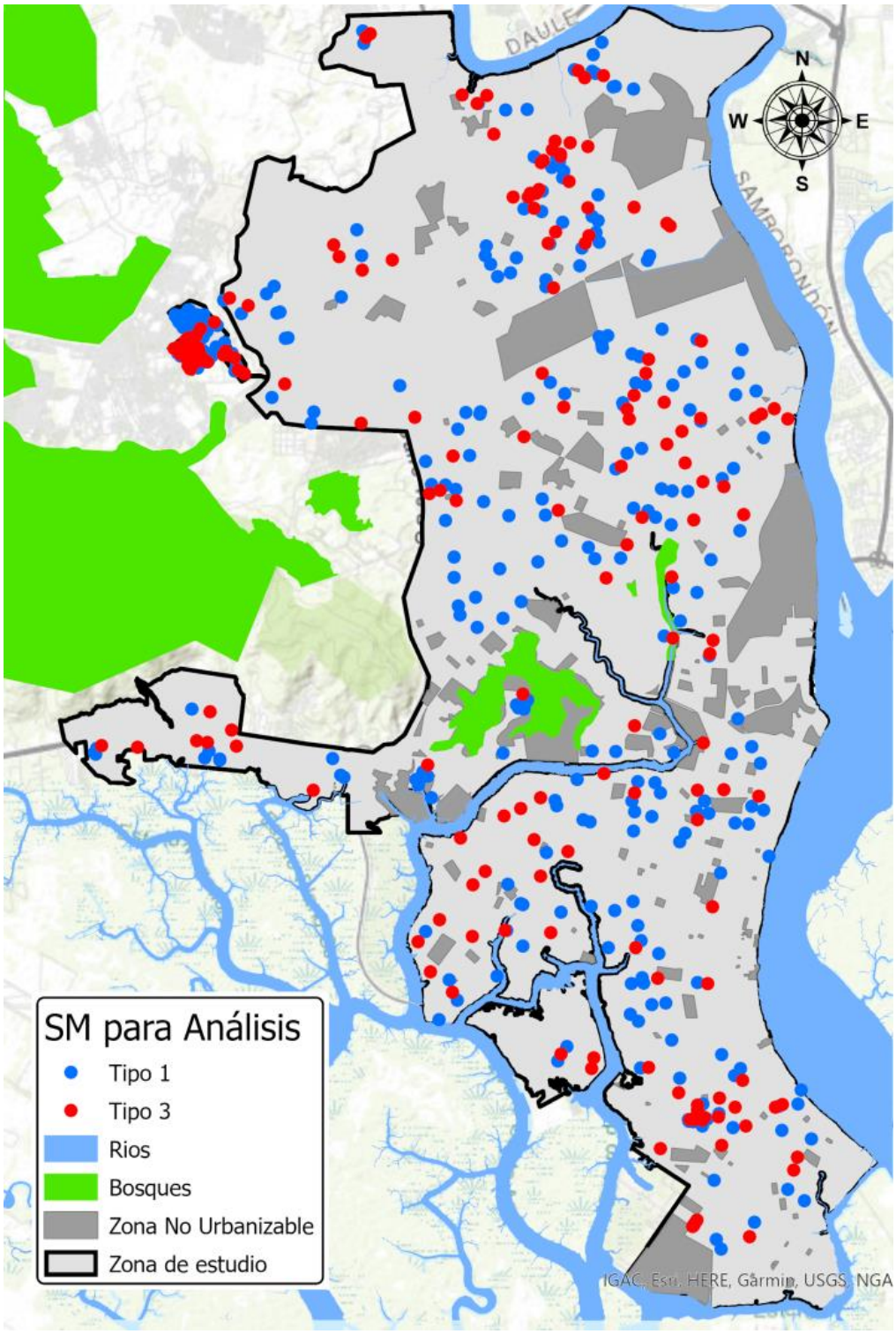

Figura 23. Medidores inteligentes seleccionados para agrupamiento restringido espacialmente.

Los resultados de la generación de clústeres espacialmente restringidos se presentan en la *Tabla 6*. Los resultados establecieron 21 subgrupos espaciales que en el resto del documento los llamaremos solamente clústeres espaciales. En el caso de los clústeres espaciales pertenecientes al tipo 1, se definen 14 zonas que contienen 143 medidores inteligentes. Por otro lado, para el caso de los clústeres espaciales pertenecientes al tipo 3, los resultados muestran 7 zonas con 46 medidores inteligentes. Este hecho indica que el comportamiento del tipo 1 es más común, frecuente y espacialmente estable que el comportamiento del tipo 3. En la *Figura 24* se presenta un mapa de los medidores inteligentes georreferenciados en la ciudad de Guayaquil, incluidas las zonas definidas para los clústeres espaciales. Los puntos azules representan los medidores inteligentes con el comportamiento clasificado por el tipo 1 y en rojo los pertenecientes al tipo 3. Los polígonos fueron generados para visualizar las zonas de los clústeres espaciales a las que pertenecen los medidores inteligentes, igualmente los polígonos cuyo borde es de color azul definen las zonas que incluyen los medidores inteligentes pertenecientes a la clasificación del tipo 1, y en color rojo para los medidores inteligentes pertenecientes al tipo 3.

Este conjunto final de clústeres espaciales tienen el mismo comportamiento de consumo, tienen al menos 5 miembros a menos de 1 km y son espacialmente contiguos, es decir, no están separados por un río, no están en una zona no urbanizable y no existe otros medidores inteligentes pertenecientes a otros tipos de comportamiento entre ellos.

Tipo de Comportamiento	SM	Cantidad de SM en clústeres espaciales	Cantidad de clústeres espaciales
Tipo 1	332	143	14
Tipo 3	207	46	7
	539	189	21

Tabla 6. Cantidad de clústeres espaciales generado.

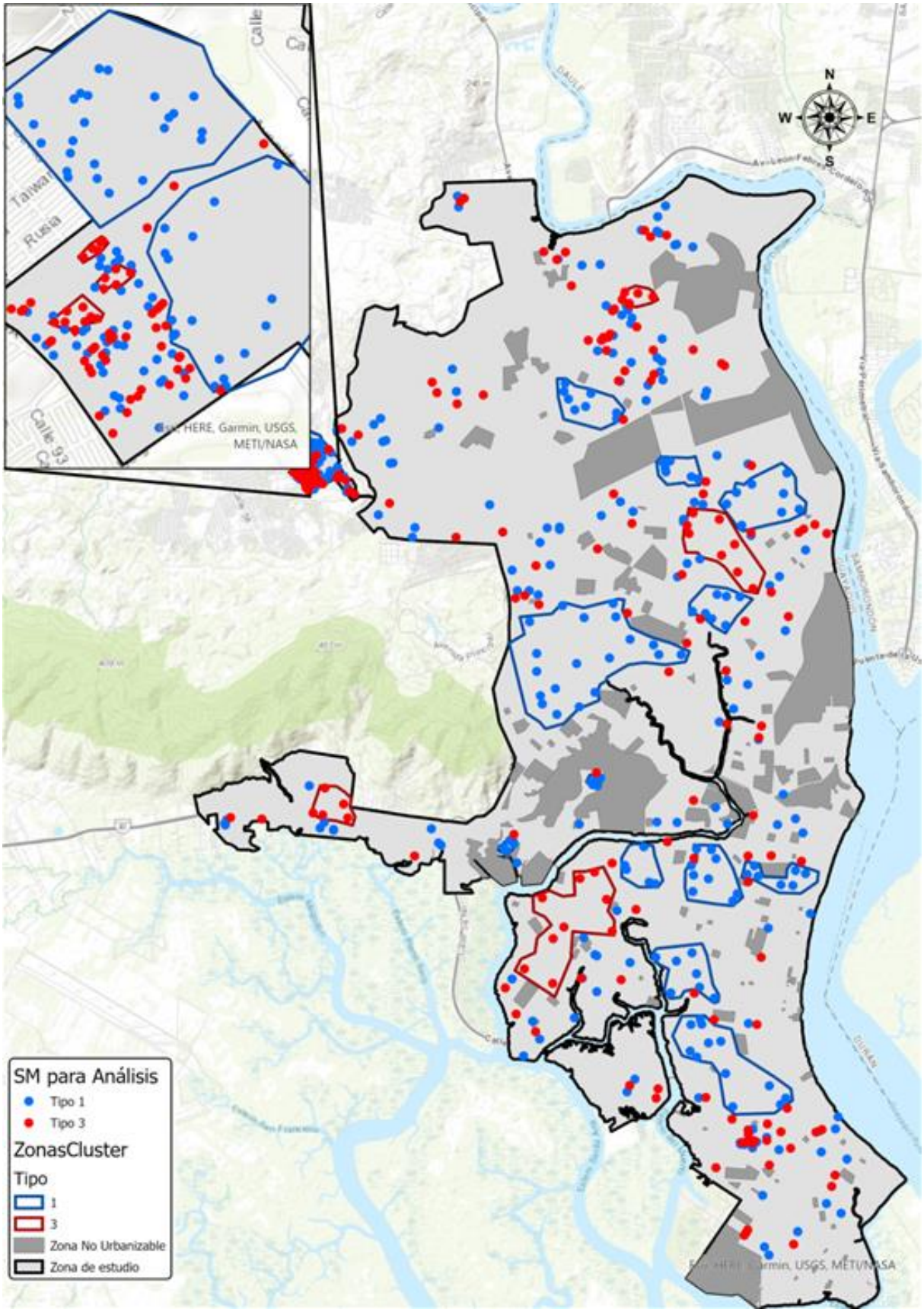

Figura 24. Zonas de clústeres espacialmente restringidos.

5.2.1 Validación

Se realizó una evaluación para determinar si un nuevo medidor inteligente ubicado en el área de influencia de los clústeres espaciales definidos tiene un comportamiento similar al de sus vecinos. Para esto se utilizaron los medidores inteligentes seleccionados como muestra de validación, y las zonas de influencia de los clústeres espaciales explicado en la metodología en la Sección 4.3.5. Se realizó una superposición entre la capa de las zonas de influencia y la capa de medidores inteligentes de validación, para determinar si el comportamiento de consumo de los medidores inteligentes de validación coincide con el comportamiento definido en las zonas de influencia de los clústeres espaciales. Si los medidores inteligentes se ubican dentro de una zona de influencia y coincide con su tipo de comportamiento se considera un valor correctamente clasificado, si no coincide se registra como incorrecto y si están ubicados fuera de la zona de influencia se descartan del análisis.

La *Tabla 7* muestra los resultados del proceso de validación. Los resultados fueron satisfactorios ya que todas las métricas se acercan al 90%, confirmando que las áreas encontradas tienen un comportamiento característico para los medidores inteligentes analizados.

Métricas	Total
Especificidad	90,32%
Precisión	86,96%
Exhaustividad	95,24%
F-Score	90,91%

Tabla 7. Métricas de clasificación entre las zonas de influencia y los medidores de muestra.

5.3 Predicción de consumo eléctrico utilizando redes neuronales LSTM

En esta sección se muestran los resultados de la predicción del consumo eléctrico de un usuario residencial mediante la aplicación de una red neuronal recurrente, utilizando la información de los medidores inteligentes agrupados en los clústeres espaciales obtenidos anteriormente. Como se indicó en la sección de metodología, se utiliza una red neuronal de tipo *long short-term memory* (LSTM).

Con el objetivo de comprobar el valor añadido que pueden aportar las agrupaciones espaciales obtenidas, se compara la predicción del consumo de energía utilizando dos redes neuronales LSTM con la misma configuración de capas. En la primera red se realizó un análisis univariante, utilizando solo la potencia activa horaria medida por el propio medidor inteligente del que se quiere obtener la predicción, mientras que en la segunda red se realizó un análisis multivariado, utilizando también como datos de entrada las medidas de potencia activa de los 4 medidores inteligentes más cercanos pertenecientes al mismo clúster espacial. Las redes se entrenan con el 80% de los datos de las series temporales dejando el 20% restante de la serie temporal para hacer las pruebas de validación.

Los resultados promedio de realizar por 30 ocasiones las predicciones de los medidores inteligentes de cada clúster espacial con ambos modelos se muestran en la *Tabla 8*. La tabla presenta valores del sMAPE para ambas aproximaciones y la diferencia porcentual entre ellas. Se puede observar que utilizar los datos de los vecinos pertenecientes al mismo clúster espacial mejora los resultados de la predicción en un 2,46%.

Tipo	sMAPE (%)		
	Univariante media	Multivariante media	Diferencia%
Tipo 1	19,72	16,84	2,88
Tipo 3	14,70	12,65	2,05
	17,21	14,75	2,46

Tabla 8. Resultados de medición promedio de sMAPE para redes LSTM univariante y multivariante.

CAPÍTULO 6

6 DISCUSIÓN

El análisis del conjunto inicial de mediciones de energía activa de los medidores inteligentes en Guayaquil (Ecuador) recopilada durante 4 años, dio como resultado 3 clústeres correspondientes a tipos de comportamiento diferentes. Un análisis más detallado de los clústeres reveló que solo 2 de ellos eran significativos. Para los medidores clasificados en cada uno de los clústeres, se calculó un perfil mensual típico del consumo durante una semana, el cual, reveló que el perfil de potencia no cambia significativamente debido a la poca variabilidad climática de la zona de Guayaquil ubicada en el Ecuador. Este resultado enfatiza el hecho de que estamos refiriéndonos al perfil de comportamiento de consumo, más no al consumo en sí. El consumo de electricidad en los edificios residenciales depende de las actividades de los ocupantes (Bin & Dowlatabadi, 2005), un ejemplo, y uno de los factores más representativos en el consumo de energía, está dado por los equipos acondicionadores de aire en los hogares. En una ciudad como Guayaquil las temperaturas y la humedad son muy altas, tal y como muestra la *Figura 25* y la *Figura 26* (weatherspark, 2018). Estas condiciones, por ejemplo, acostumbrarían a un hogar a encender el acondicionador de aire a la hora del almuerzo todos los días del año alrededor del mediodía. Ese sería un

comportamiento de consumo típico de esa residencia, pero la potencia consumida por el acondicionador de aire en los meses febrero, marzo y abril sería mayor.

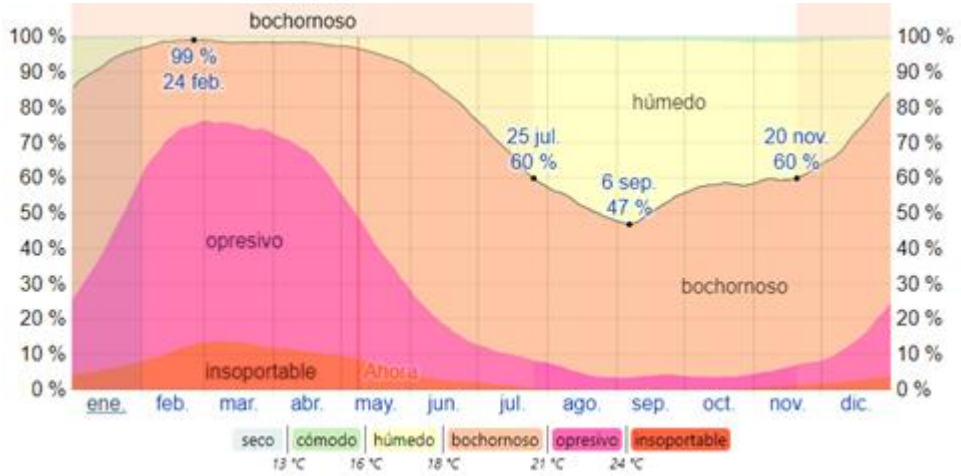

Figura 25. Porcentaje de humedad promedio en la ciudad de Guayaquil. Categorizados por el punto de rocío **(weatherspark, 2018)**.

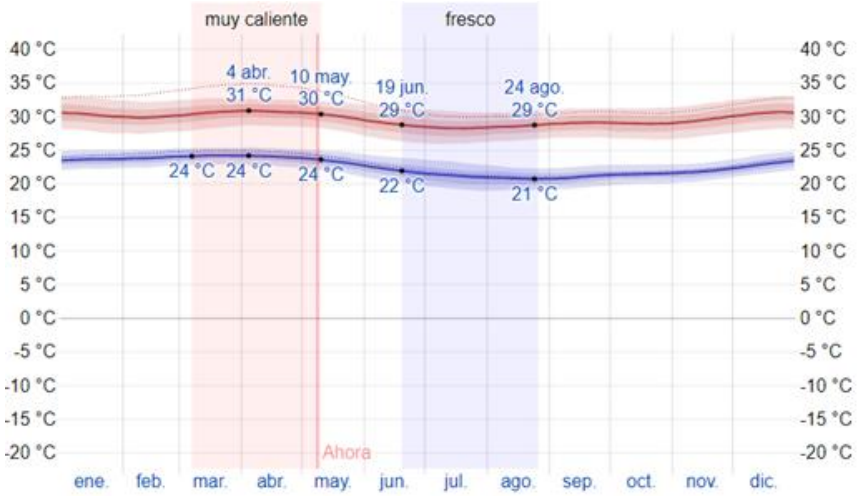

Figura 26. Temperatura promedio en la ciudad de Guayaquil. La línea roja representa la temperatura máxima, la línea azul la temperatura mínima **(weatherspark, 2018)**.

.

Los tipos de comportamiento encontrados, complementados con los clústeres espaciales definidos con el análisis espacial, son una información valiosa para las empresas eléctricas, pues les permite planificar mejor sus actividades de operación y mantenimiento por zonas geográficas, minimizando el impacto de estas actividades en

los clientes. Según los resultados se puede evidenciar que, aunque todos los clientes son residenciales, no se les puede generalizar con un único perfil de carga. Incluso es más difícil aún determinar zonas geográficas con un comportamiento único, como se puede ver en el resultado de la generación de los clústeres espaciales donde solo 189 de los 539, es decir un 35% de los clientes, se puede zonificar con un comportamiento típico de consumo. Sin embargo, la información de las zonas geográficas y los medidores inteligentes que sí se pudieron agrupar en clústeres espaciales con comportamiento típico, constituyen un insumo valioso para análisis más complejos y geográficamente definidos. Para probar esto, en el estudio se utilizó una red neuronal recurrente de tipo long short-term memory (LSTM) para predecir el consumo de energía de un consumidor, utilizando no solo las mediciones de su medidor inteligente, sino también los datos recopilados por los otros medidores inteligentes que pertenecen al mismo clúster espacial. Los resultados mostraron que la precisión de la previsión mejora en un 2,46% en promedio. Esta mejora en el pronóstico demuestra la valiosa y prometedora información que se puede obtener del análisis espacial. Las empresas eléctricas podrían realizar análisis cualitativos a los miembros de estos clústeres espaciales para determinar los factores comunes y descubrir variables relevantes a considerar en la clasificación de los comportamientos de los usuarios residenciales.

De igual manera el poder establecer zonas geográficas con comportamientos de consumo claramente definido y estable en el tiempo, permiten a los administradores de las empresas de distribución tomar estrategias de inversión o mantenimiento de las redes de forma zonificada y con información más precisa.

Para ejemplificar la valiosa información encontrada en los clústeres espaciales, se han graficado los perfiles de carga con sus valores reales (sin normalizar) y estratificados dentro de cada clúster para evitar que se suavicen al promediar usuarios con consumos mayores pero con igual comportamiento. La estratificación dentro de cada clúster se realizó de acuerdo a la energía mensual consumida: i) menor que 130 kWh/mes, ii) entre 130 y 500 kWh/mes, iii) entre 500 y 1000 kWh/mes y iv) más de 1000kWh/mes. Los dos primeros estratos (i, ii) son los que obtienen mayores subsidios económicos del gobierno y los siguientes van perdiendo los beneficios hasta el último segmento (iv) que no tiene

ningún subsidio. La *Figura 27* muestra los perfiles de carga promedio por hora de lunes a domingo para los tipos de comportamiento encontrados, y la *Figura 28* muestra igualmente los perfiles de carga promediados por día. Como se puede evidenciar en ambos gráficos los clientes del tipo 1 demandan una mayor cantidad de energía los días lunes en todos los estratos, mientras que los clientes del tipo 3 la demandan los días sábados igualmente en todos los estratos. Para los clientes del tipo 1 lo más adecuado sería planificar las actividades de mantenimiento durante los fines de semana, mientras que para los clientes del tipo 3 deberían evitarse hacerlo los sábados. Además, si las obras de mantenimiento se planificaran en días laborables, tendrían menos impacto a partir de las 17:00 horas, en el caso de los clientes del tipo 1, y después de las 19:00 horas para los clientes del tipo 3. Para ambos tipos, las mañanas hasta las 10:00 horas serían periodos críticos para realizar actividades de mantenimiento debido al rápido incremento de la demanda de energía en ese periodo. El poder contar con esta información de consumo precisa, actualizable y zonificada geográficamente, mejorará sustancialmente las tareas de mantenimiento y optimizará recursos.

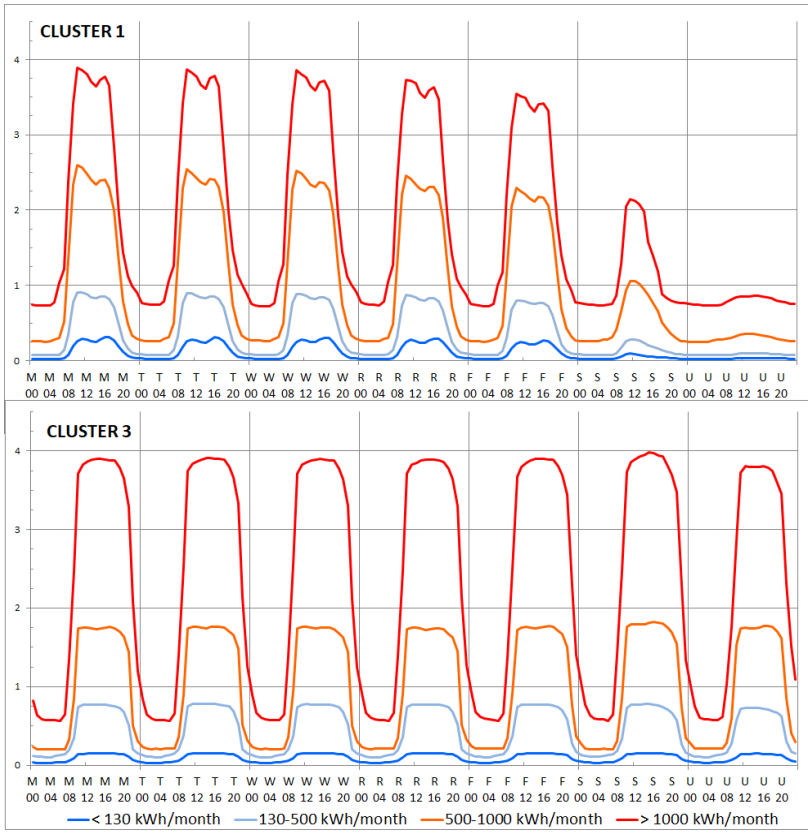

Figura 27. Perfiles de carga promedio por hora. Grafica los consumos no normalizados de lunes a domingo y estratificado en función de la energía consumida.

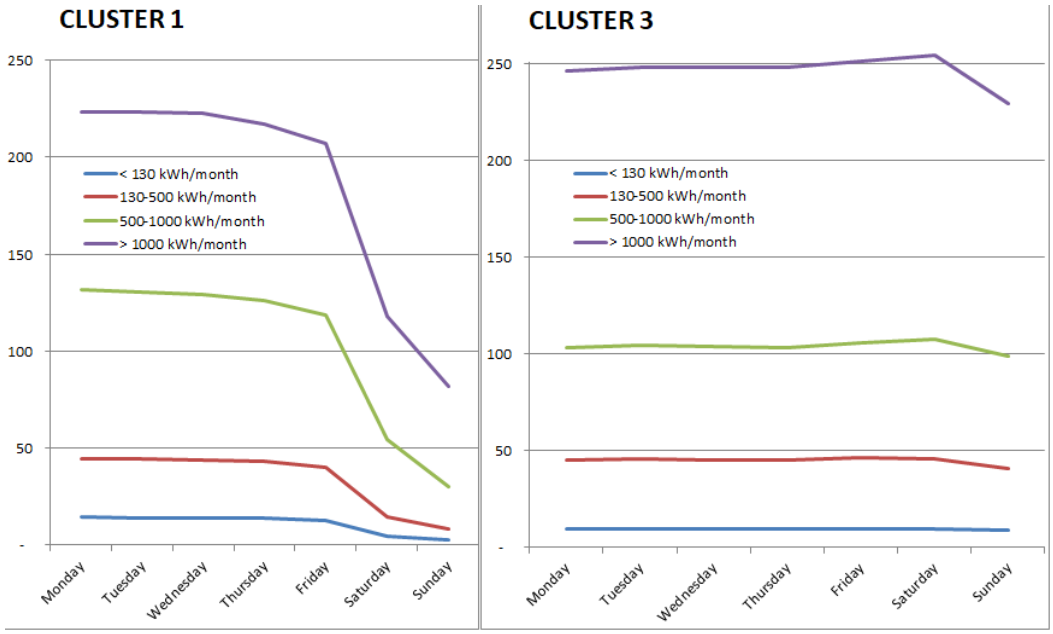

Figura 28. Perfiles de carga promedio por día. Grafica los consumos no normalizados de lunes a domingo y estratificado en función de la energía consumida.

Las posibles limitaciones del método propuesto desde el punto de vista del análisis espacio-temporal estarían dadas por la falta de respeto por parte de la ciudadanía de la ordenanzas municipales para uso de los espacios ya que se pueden encontrar zonas residenciales con un alto componente de comercios clandestinos (bares, tiendas, cyber café, talleres de microempresas), lo que hace necesario una validación en campo o el apoyo de personal que conoce las zonas para afinar las zonas definidas de los clústeres espaciales. Desde el punto de vista del análisis utilizando redes neuronales recurrentes la principal limitación es el tiempo de procesamiento para actualizar y reentrenar el modelo, que se debería realizar periódicamente (acorde con el crecimiento de la zona de estudio) y que pueden tomar varias horas o incluso días.

CAPÍTULO 7

7 CONCLUSIONES

Los datos disponibles en el sector eléctrico ecuatoriano han crecido enormemente por la implementación de nuevos sistemas informáticos homologados, los cuales tienen un inmenso potencial que puede ser explotado. Si incorporamos en los procesos y flujos de trabajo el enfoque combinado de aprendizaje automático y análisis espacial se podrán realizar análisis que permitirán encontrar nuevo conocimiento y de esta manera aportar en la planificación, mantenimiento y controlar la operación más eficientemente. En este estudio hemos presentado una metodología que permite demostrar que el patrón de consumo de energía en áreas cercanas está relacionado y se pueden utilizar modelos que aprovechan esta información como ventaja. El uso de herramientas de aprendizaje automático combinado con el análisis espacial, permite definir y descubrir nuevos perfiles de comportamiento de consumo de usuarios residenciales y determinar zonas geográficas cuyo comportamiento es más marcado y constante, lo que nos permite mejorar la previsión de consumo energético.

Los cambios en políticas económicas, energéticas o estados de excepción, como el caso de la pandemia por covid-19 que está viviendo el mundo actualmente, pueden conducir a un cambio en los factores que determinan la demanda de energía. Por lo tanto, estos factores deben identificarse y evaluarse de forma dinámica. Al emplear aprendizaje automático en combinación con análisis espacial, se pueden emitir pronósticos para

ajustar las predicciones de la carga de energía en toda una región o en áreas geográficas específicas, lo que puede servir de insumo tanto para la gestión técnica como comercial. Desde una perspectiva comercial, la metodología presentada permite estimar con mayor precisión la energía no suministrada a diferentes zonas de la ciudad durante apagones o priorizar zonas en campañas de concientización de consumo, mientras que desde un punto de vista técnico permite una mejor planificación de las actividades de mantenimiento, y una estimación más precisa de los factores de demanda futura, que son útiles para la planificación de la red y la reducción de inversiones en redes y/o centrales eléctricas.

7.1 Contribución

Las principales contribuciones de esta tesis se pueden resumir en los siguientes aspectos, los cuales manifiestan la consecución de los objetivos presentados en esta tesis:

- Se han analizado y comparado diferentes algoritmos de aprendizaje automático y agrupamiento de series temporales para mejorar la capacidad de definir perfiles de carga mensual de los usuarios residenciales, que es un insumo valioso para apoyar en la toma de decisiones de las empresas eléctricas.
- Se ha descubierto y analizado dos perfiles de carga significativos para usuarios residenciales, personalizados a la realidad de la ciudad de Guayaquil.
- Además, se han definido zonas geográficas en la ciudad de Guayaquil donde el comportamiento de consumo de los medidores inteligentes es constante y más predecible.
- En este estudio se han aplicado análisis espacio temporales a un gran conjunto de datos de medidores inteligentes residenciales, lo que demuestra que la información del vecindario espacial es una fuente importante de información que puede mejorar la toma de decisiones y las habilidades de predicción.
- Además, hemos apoyado nuestros experimentos con un nuevo marco de aprendizaje de máquina híbrido que aplica una técnica de agrupamiento de series temporales utilizando deformación dinámica de tiempo e información

espacial de sus vecinos junto con una red neuronal recurrente con una arquitectura LSTM.

Como aporte y transferencia de la investigación realizada se ha enviado la publicación de un artículo titulado “Definition of Residential Power Load Profiles Clusters Using Machine Learning and Spatial Analysis” donde se presentan las contribuciones realizadas en el campo de análisis espacio temporal y las redes LSTM.

7.2 Trabajos futuros

Como trabajos futuros en la línea de esta tesis, se pueden considerar varias mejoras a la metodología presentado en los anteriores capítulos:

- Ejecutar un análisis similar del año 2020 y 2021 cuando los comportamientos de consumo se vieron afectados por el incremento del teletrabajo y el confinamiento de las personas debido a la pandemia originada por el covid-19. Este nuevo análisis serviría para determinar cómo ha cambiado el comportamiento de consumo y si los comportamiento se mantienen en las zonas geográficas descubiertas o se han generado nuevas zonas.
- A pesar de los resultados prometedores de la metodología, se necesita un mayor conjunto de medidores inteligentes para validar plenamente el enfoque propuesto. Considerando que la empresa eléctrica de Guayaquil está en un continuo proceso de actualización de sus medidores de energía se podrían evaluar y experimentar con más medidores inteligentes y mediciones para mejorar los resultados obtenidos.
- En esta tesis se han proporcionado una serie de técnicas empíricas de geoprocesos para mejorar la restricción espacial, sería interesante profundizar en el plano teórico o empírico para derivar una estrategia más robusta, capaz de ajustar las zonas de los clústeres espaciales más rápidamente.
- Realizar un análisis similar pero con usuarios de tipo comerciales y comparar la precisión con la de los usuarios residenciales de esta investigación.

- Probar la precisión y el rendimiento utilizando otros tipos de redes neuronales que actualmente se están aplicando para análisis de series temporales como son las redes neuronales bidireccionales (Schuster & Paliwal, 1997) y redes neuronales convolucionales (Borovykh, Bohte, & Oosterlee, 2017) (Yang, Nguyen, San, Li, & Krishnaswamy, 2015).
- Finalmente, es importante considerar los criterios de seguridad. A medida que la cantidad de medidores inteligentes aumenta, también aumentan los problemas de seguridad asociados con los medidores inteligentes. La información detallada del consumo de los usuarios puede revelar su estilo de vida, por lo que sería importante estudiar mecanismos de seguridad que garanticen la confidencialidad y validez de las mediciones realizadas con los medidores inteligentes.

BIBLIOGRAFÍA

Aach, J., & Church, G. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics, 17*(6), 495-508.

Abreu, J. M., Pereira, F. C., & Ferrão, P. (2012). Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and buildings, 49*, 479–487.

Anselin, L. (2019, 09 10). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Routledge*, 111-126. Retrieved from https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fnref3

Assunção, R., Neves, M., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science, 20*(7), 797-811.

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., & Carpendale, S. (2014). A review of temporal data visualizations based on space-time cube operations. *Eurographics conference on visualization.*

Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., & Simon, I. (2002). A new approach to analyzing gene expression time series data. *Proceedings of the sixth annual international conference on Computational biology.*, (pp. 39-48). New York.

Beaudin, M., & Zareipour, H. (2015). A review of modelling and complexity. Renewable and sustainable energy reviews. *Renewable and sustainable energy reviews, 45*, 318-335.

Bengio, Y., Simard, P., & Frasconi, P. (1994, March). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5*(2), 157-166.

Bin, S., & Dowlatabadi, H. (2005). Consumer lifestyle approach to US energy use and the related CO2 emissions. *Energy policy, 33*(2), 197-208.

Biswas, M. R., Robinson, M. D., & Fumo, N. (2016). Prediction of residential building energy consumption: A neural network approach. *Energy, 117*, 84–92.

Borovykh, A., Bohte, S., & Oosterlee, C. (2017). Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*.

Burke, P., Stern, D., & Bruns, S. (2018, November). The impact of electricity on economic development: a macroeconomic perspective. *International Review of Environmental and Resource Economics, 12*(1), 85-127.

Cano, E., Groissböck, M., Moguerza, J., & Stadler, M. (2014). A strategic optimization model for energy systems planning. *Energy and buildings, 81*, 416-423.

Chicco, G. (2012). Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy, 42*(1), 68-80.

Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on power systems, 21*(2), 933-940.

Chrisman, N. (2006). *"Charting the unknown." How computer mapping at Harvard became GIS.* Redland: ESRI Press.

Cliff, A., & Ord, J. (1973). Spatial Autocorrelation. *Pion*.

Cuturi, M. (2011). Fast global alignment kernels. *Proceedings of the 28th international conference on machine learning (ICML-11)*, (pp. 929–936).

Cuturi, M., & Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. *International Conference on Machine Learning*, (pp. 894-903).

Dubayah, R., & Drake, J. (2000). Lidar remote sensing for forestry. *Journal of forestry, 98*(6), 44-46.

Enerdata. (2020). *Global Energy Statistical Yearbook*.

ESRI. (2019). *How Time Series Clustering works*. Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/learnmoretimeseriesclustering.htm

Esther, B., & Kumar, K. (2016). A survey on residential demand side management architecture, approaches, optimization models and methods. *Renewable and Sustainable Energy Reviews, 59*, 342-351.

Foresman, T. W. (1998). *The history of geographic information systems: perspectives from the pioneers* (Vol. 397). Prentice Hall PTR Upper Saddle River, NJ.

Goodchild, M. (2018). Reimagining the history of GIS. *Annals of GIS, 24*(1), 1-8.

Gouveia, J. P., & Seixas, J. (2016). Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings, 116*, 666–676.

Hägerstrand, T. (1970). What about people in regional. *Papers of the Regional Science Association XXIV*, (pp. 7-21).

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*, 107–145.

Harvey, C. (2018, December 6). *Scientific American*. Retrieved from News, E&E: https://www.scientificamerican.com/article/co2-emissions-reached-an-all-time-high-in-2018/

Hino, H., Shen, H., Murata, N., Wakao, S., & Hayashi, Y. (2013). A versatile clustering method for electricity consumption pattern analysis in households. *IEEE Transactions on Smart Grid, 4*(2), 1048-1057.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A field guide to dynamical recurrent neural networks. IEEE Press*.

Hsiao, Y.-H. (2014). Household electricity demand forecast based on context information and user daily schedule analysis from meter data. *IEEE Transactions on Industrial Informatics, 11*, 33–43.

IEA. (2019). *World Energy Outlook 2019*. Retrieved from IEA Paris: https://www.iea.org/reports/world-energy-outlook-2019

Janati, H., Cuturi, M., & Gramfort, A. (2020). Spatio-Temporal Alignments: Optimal transport through space and time. *International Conference on Artificial Intelligence and Statistics*, (pp. 1695-1704).

Janetzko, H., Stoffel, F., Mittelstädt, S., & Keim, D. (2014). Anomaly detection for visual analytics of power consumption data. *Computers & Graphics, 38*, 27-37.

Keras. (2021, 01). *Keras about*. Retrieved from https://keras.io/about/

Khan, R., & Khan, J. (2013). A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network. *Computer Networks, 57*(3), 825-845.

Kraak, M.-J. (2003). The space-time cube revisited from a geovisualization perspective. *Proc. 21st International Cartographic Conference*, (pp. 1988-1996).

Kwac, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid, 5*, 420–430.

Kwac, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid, 5*(1), 420-430.

Lavin, A., & Klabjan, D. (2015). Clustering time-series energy data from smart meters. *Energy efficiency, 8*, 681–689.

Lund, H., Østergaard, P., Connolly, D., & Mathiesen, B. (2017). Smart energy and smart energy systems. *Energy, 137*, 556-565.

Mahmoudi-Kohan, N., Moghaddam, M., & Sheikh-El-Eslami, M. (2010). An annual framework for clustering-based pricing for an electricity retailer. *Electric Power Systems Research, 80*(9), 1042-1048.

Marino, D. L., Amarasinghe, K., & Manic, M. (2016). Building energy load forecasting using deep neural networks. *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*, (pp. 7046–7051).

Melo, J. D., Carreno, E. M., & Padilha-Feltrin, A. (2012). Multi-agent simulation of urban social dynamics for spatial load forecasting. *IEEE Transactions on Power Systems, 27*, 1870–1878.

Melo, J. D., Padilha-Feltrin, A., & Carreno, E. M. (2015). Spatial pattern recognition of urban sprawl using a geographically weighted regression for spatial electric load forecasting. *2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*, (pp. 1–5).

Mocanu, E., Nguyen, P. H., Gibescu, M., & Kling, W. L. (2016). Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks, 6*, 91–99.

Mohassel, R., Fung, A., Mohammadi, F., & Raahemifar, K. (2014, May). A survey on advanced metering infrastructure. *International Journal of Electrical Power & Energy Systems, 63*, 473-484.

Monedero, I., Biscarri, F., León, C., Guerrero, J., Biscarri, J., & Millán, R. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. *International Journal of Electrical Power & Energy Systems, 34*(1), 90-98.

Montero, P., & Vilar, J. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software, 62*(1), 1-43.

Moran, P. (1969). Notes on Continuous Stochastic Phenomena. *Biometrika, 37*(1/2), 17–23.

Nagi, J., Yap, K., Tiong, S., Ahmed, S., & Mohammad, A. (2008). Detection of abnormalities and electricity theft using genetic support vector machines. *TENCON 2008-2008 IEEE Region 10 Conference* (pp. 1-6). IEEE.

Niu, Z., Wu, J., Liu, X., Huang, L., & Nielsen, P. (2021). Understanding Energy Demand Behaviors through Spatio-temporal Smart Meter Data Analysis. *Energy*, 120493.

Paparrizos, J., & Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data.*, (pp. 1855-1870).

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition, 44*(3), 678-693.

Praveen, M., & Rao, G. (2020). Ensuring the reduction in peak load demands based on load shifting DSM strategy for smart grid applications. *Procedia Computer Science, 167*, 2599-2605.

Rhodes, J., Cole, W., Upshaw, C., Edgar, T., & Webber, M. (2014). Clustering analysis of residential electricity demand profiles. *Applied Energy, 135*, 461-471.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Ryu, S., Noh, J., & Kim, H. (2017). Deep neural network based demand side short term load forecasting. *Energies, 10*, 3.

Sakoe, H., & Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. *3*, 65-69.

Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing., 45*(11), 2673-2681.

Sevlian, R., & Rajagopal, R. (2014). Short term electricity load forecasting on varying levels of aggregation. *arXiv preprint arXiv:1404.0058*.

Shahzadeh, A., Khosravi, A., & Nahavandi, S. (2015). Improving load forecast accuracy by clustering consumers using smart meter data. *2015 international joint conference on neural networks (IJCNN)*, (pp. 1–7).

Tascikaraoglu, A., & Sanandaji, B. M. (2016). Short-term residential electric load forecasting: A compressive spatio-temporal approach. *Energy and Buildings, 111*, 380–392.

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., . . . Woods, E. (2020). Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research, 21*(118), 1-6.

Thakur, S., & Hanson, A. (2010). A 3D visualization of multiple time series on maps. *201 14th International Conference Information Visualisation* (pp. 336-343). IEEE.

Thiessen, A. (1911). Precipitation averages for large areas. *Monthly Weather Review, 39*(7), 1082-1084.

Thorndike, R. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267-276.

Tomlinson, R. (1969). A geographic information system for regional planning. *Journal of Geography (Chigaku Zasshi), 78*(1), 45-48.

Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V. M., & Sousa, J. M. (2016). Classification of new electricity customers based on surveys and smart metering data. *Energy, 107*, 804–817.

Wang, X., Qin, Y., Wang, Y., Xiang, S., & Chen, H. (2019). ReLTanh: An activation function with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis. *Neurocomputing, 363*, 88-98.

Waters, N. (2016). GIS: history. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, 1-13.

weatherspark. (2018). *El clima promedio en Guayaquil*. Retrieved from weatherspark.com: https://es.weatherspark.com/y/19346/Clima-promedio-en-Guayaquil-Ecuador-durante-todo-el-a%C3%B1o

Werbos, P. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*.

Wijaya, T. K., Vasirani, M., Humeau, S., & Aberer, K. (2015). Cluster-based aggregate forecasting for residential electricity demand using smart meter data. *2015 IEEE international conference on Big data (Big data)*, (pp. 879–887).

Wytock, M., & Kolter, J. (2014). Contextually supervised source separation with application to energy disaggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28.*

Xu, J., Yue, M., Katramatos, D., & Yoo, S. (2016). Spatial-temporal load forecasting using AMI data. *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, (pp. 612–618).

Yang, J., Nguyen, M., San, P., Li, X., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. *Twenty-fourth international joint conference on artificial intelligence.*

Yarbrough, I., Sun, Q., Reeves, D., Hackman, K., Bennett, R., & Henshel, D. (2015). Visualizing building energy demand for building peak energy analysis. *Energy and Buildings, 91*, 10-15.

Zhang, L., Feng, J., & Jian, X. (2016). Model of energy alternative in spatial load forecasting. *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, (pp. 2106–2110).

Zhou, K., Yang, C., & Shen, J. (2017). Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China. *Utilities Policy, 44*, 73–84.


UdG