MINERÍA DE DATOS PARA LA SOSTENIBILIDAD URBANA

Tesis Doctoral

Programa:
DOCTORADO EN TECNOLOGÍAS INFORMÁTICAS

Escuela:
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Departamento:
LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN

Doctorando:
FRANCISCO RODRÍGUEZ GÓMEZ

Directores:
JOSÉ DEL CAMPO ÁVILA
LLANOS MORA LÓPEZ

Junio 2023

UNIVERSIDAD DE MÁLAGA

AUTOR: Francisco Rodríguez Gómez

https://orcid.org/0000-0002-7707-6762

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

UNIVERSIDAD DE MÁLAGA

ANDALUCÍA TECH
Campus de Excelencia Internacional

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D. FRANCISCO RODRÍGUEZ GÓMEZ
Estudiante del programa de doctorado TECNOLOGÍAS INFORMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: MINERÍA DE DATOS PARA LA SOSTENIBILIDAD URBANA

Realizada bajo la tutorización de la Dra. LLANOS MORA LÓPEZ y dirección del Dr. JOSÉ DEL CAMPO ÁVILA y la Dra. LLANOS MORA LÓPEZ DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 27 de marzo de 2023

Fdo.: Francisco Rodríguez Gómez	Fdo.: Tutor/a
Fdo.: Director/es de tesis	

Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10 **E-mail:** doctorado@uma.es

UNIVERSIDAD DE MÁLAGA

AUTORIZACIÓN PARA LA LECTURA E INFORME SOBRE LA TESIS DE:
D. FRANCISCO RODRÍGUEZ GÓMEZ

La Dra. Llanos Mora López, Catedrática del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de tutora y directora de la tesis doctoral de Francisco Rodríguez Gómez, titulada **Minería de datos para la sostenibilidad urbana**; y el Dr. José del Campo Ávila, Profesor Contratado Doctor del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de director de dicha tesis, AUTORIZAN su lectura.

Asimismo, INFORMAN que las publicaciones que avalan la tesis no han sido utilizadas en tesis anteriores.

Málaga, 27 de marzo de 2023.

Fdo.: Llanos Mora López

Fdo.: José del Campo Ávila

Agradecimientos

Me gustaría comenzar esta sección de agradecimientos dando las gracias a mis directores de tesis: Llanos y José. Esta tesis nunca hubiera sido posible sin su ayuda. Gracias por compartir conmigo vuestra sabiduría y vuestros consejos. Gracias por ser mis guías en mi nueva aventura en el mundo de la investigación y la docencia. Y gracias por vuestra confianza, apoyo, amistad, y tiempo.

También quiero mostrar mi gratitud especialmente a Luis, IP del proyecto URSUS-DM en el que se enmarca y del que nace esta tesis. Gracias por tu conocimiento, ayuda, y por contar siempre conmigo.

Por supuesto, también me gustaría dar las gracias al resto de mis compañeros del equipo de investigación del proyecto URSUS-DM, sin los que tampoco hubiera posible la realización de este trabajo de investigación. Gracias a Domingo y Marta.

También me gustaría mostrar mi agradecimiento al departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, por darme la oportunidad de trabajar como profesor en el área de conocimiento de Lenguajes y Sistemas Informáticos.

Gracias a mi madre y a mi hermana, por ayudarme tanto y apoyarme siempre, y por su incesante empeño en que me aventurase en el mundo de la investigación y de la docencia universitaria.

Gracias a todos por confiar en mí.

Índice general

Página

Índice de figuras

Índice de tablas

Resumen

En este trabajo de investigación se han propuesto nuevos modelos y procedimientos basados en técnicas de minería de datos que permiten abordar y resolver diferentes problemas relacionados con la sostenibilidad urbana. La sostenibilidad urbana consiste en el desarrollo e implantación de modelos urbanísticos que garanticen y mejoren el bienestar y la calidad de vida de los seres vivos de las ciudades, sin degradar el entorno, ni poner en compromiso a las futuras generaciones. Los modelos desarrollados han contribuido al descubrimiento de conocimiento utilizando datos heterogéneos. Además, a partir de los resultados obtenidos se han implementado varias herramientas de simulación y estimación para ayudar en la toma de decisiones.

Los dominios de aplicación del trabajo son dos dentro del ámbito de la sostenibilidad urbana. Uno de ellos ha sido la caracterización y predicción de la generación y consumo de energía en entornos urbanos. El otro ha sido el análisis y modelización del papel de la infraestructura verde en la mitigación del efecto isla de calor, teniendo en cuenta sus características y distribución geográfica. En ambos casos, se han propuesto metodologías para abordar estos problemas y se han implementado distintas herramientas de ayuda para la toma de decisiones en estos ámbitos.

En el contexto de la generación eléctrica, se han implementado modelos para la estimación del recurso energético utilizado por sistemas fotovoltaicos (radiación solar), modelos para la evaluación del potencial de generación fotovoltaica de un área, modelos para la predicción a corto plazo de la producción fotovoltaica de instalaciones y modelos para la caracterización de consumos de electricidad para usuarios domésticos. Se ha desarrollado una herramienta de ayuda a la toma de decisiones en la evaluación del potencial energético de un área urbana.

Se han producido los avances que se enumeran en el campo de generación de electricidad a partir de energías renovables y consumo eléctrico doméstico:

- Se ha implementado una metodología para realizar predicciones a corto plazo de la producción fotovoltaica (para el día siguiente) en un área determinada. Esta metodología utiliza modelos de predicción de radiación solar global horaria a corto plazo desarrollados previamente e incorpora el conocimiento de expertos que se considera como un aspecto fundamental para conseguir resultados de calidad. Las

predicciones obtenidas son necesarias, entre otras aplicaciones, para la toma de decisiones en el mercado energético con el fin de asegurar la correcta integración de los sistemas solares fotovoltaicos conectados a la red eléctrica. La metodología propuesta es más sencilla que las alternativas anteriores y consta, únicamente, de dos fases: una primera que es capaz de capturar las características más relevantes del terreno para crear nueva información (usando algoritmos de agrupamiento) y una segunda fase que usa dicha información para realizar estimaciones de la producción fotovoltaica horaria atendiendo a modelos que han sido generados considerando datos meteorológicos y de radiación.

- Se ha construido una herramienta de código abierto, URSUS-PV, para estimar la electricidad potencial que se puede generar en instalaciones fotovoltaicas a corto plazo (durante el día siguiente) y a largo plazo (promedio diario para todos los días del año) en un área urbana de interés (barrios, calles, edificios complejos). Podría ser potencialmente útil para múltiples tipos de usuarios, como pueden ser las administraciones públicas, las empresas del sector fotovoltaico, cooperativas o comunidades de vecinos. Uno de los beneficios más significativos de esta herramienta es la automatización de un proceso complejo. Inicialmente el proceso se realizó manualmente para obtener resultados globales en áreas urbanas de interés. Después de realizar un proceso de extracción de datos, dicho procesamiento manual se automatizó.
- Se han identificado perfiles de consumo eléctrico horario para usuarios domésticos incorporando en la selección de perfiles el conocimiento de expertos del dominio. Para ello, se ha propuesto un algoritmo que permite determinar automáticamente el número de perfiles de consumo (clústeres) más apropiado.

En el caso de la modelización de la infraestructura verde se han desarrollado dos aplicaciones informáticas a partir del análisis realizado. La primera tiene como objeto identificar las zonas más desfavorables de un área urbana en términos de altas temperaturas y escasez de espacios verdes. La segunda hace posible la predicción de la temperatura que se conseguiría en una zona si se instalaran infraestructuras verdes en localizaciones cercanas.

En el campo de la naturación urbana y urbanismo sostenible, se han producido distintos avances:

- Se ha desarrollado una metodología para la identificación automática de zonas urbanas con alta temperatura y escasez de vegetación creando una herramienta informática que implementa esta metodología. Aunque existían previamente estudios relacionados con la influencia de las zonas verdes en las temperaturas urbanas, no existía hasta el momento una metodología para determinar qué zonas tienen mayores necesidades de establecimiento de vegetación para bajar su temperatura.

- Se ha caracterizado qué tipología y tamaño de zonas verdes tienen más efecto sobre la bajada de temperatura, así como a qué distancias se observan las mayores influencias. Esta última cuestión representa la mayor innovación ya que estudios anteriores no tienen en cuenta las distancias de las zonas verdes a las áreas de afección.
- Se ha conseguido entrenar un modelo que permite realizar predicciones de cómo evolucionaría la temperatura en una ubicación determinada si se incorporara vegetación en zonas cercanas. Este modelo ha sido implementado en una herramienta que posibilita la realización de simulaciones para ver en qué localizaciones se maximizan los efectos de disponer nuevas zonas verdes y poder realizar inversiones públicas informadas que permitan optimizar los recursos y los resultados obtenidos con los mismos.

Desde el punto de vista de la utilidad de los trabajos desarrollados, se puede afirmar que, en al ámbito del enverdecimiento de las ciudades, pueden contribuir a la mitigación del calor extremo producido en las áreas urbanas, ya que las herramientas desarrolladas pueden ayudar en la toma de decisiones sobre la instalación de elementos de infraestructura verde (jardines verticales, techos verdes, parques ...) en las zonas que se detecten como más desfavorecidas. Fomentar dicha actividad en las ciudades repercute de forma positiva en la salud los ciudadanos y facilita un mayor ahorro energético en climatización, entre otros muchos beneficios.

En el ámbito del uso de energía solar en entornos urbanos, el sistema desarrollado puede contribuir a la evaluación de los mejores emplazamientos para estos sistemas y a la evaluación del potencial de generación de los mismos. Su implantación en las mejores ubicaciones, en detrimento del uso de fuentes de energía contaminantes, conducirá a modelos urbanos sostenibles y generará una serie de beneficios como pueden ser una mejor calidad del aire, mejor salud de los ciudadanos (menos problemas respiratorios, menos enfermedades relacionadas con las temperaturas extremas ...), menor consumo energético en climatización a partir de la electrificación con renovables, y por lo tanto, un ahorro en las facturas de los consumidores. Todas estas ventajas se obtienen como consecuencia de la reducción o eliminación de los gases de efecto invernadero al sustituir su uso por energías renovables.

La caracterización de los perfiles de consumo eléctrico de los hogares en las ciudades, puede tener un impacto positivo sobre la mejora de la sostenibilidad urbana, ya que permite ofrecer tarifas y servicios personalizados a los consumidores y realizar una gestión más eficiente de los consumos.

Palabras clave: ubicación de instalaciones fotovoltaicas, caracterización de perfiles de consumo eléctrico de los hogares, número óptimo de clústeres, efecto isla de calor urbano, predicción de temperatura de la superficie terrestre, naturación urbana

1

Introducción

En este capítulo se describe de forma breve el contexto de esta investigación. Ésta se ha enfocado en el desarrollo y uso de técnicas de minería de datos para su utilización en la mejora de la sostenibilidad urbana. También se presenta la motivación que la originó así como los objetivos que se plantearon y que se han alcanzado. Por último, se describe la metodología que se ha seguido y las principales aportaciones que se han generado.

1.1. Sostenibilidad urbana

El rápido crecimiento de las ciudades y las distintas actividades no sostenibles que en ellas se llevan a cabo, como el uso de energías contaminantes, edificaciones masivas, ausencia de actividades de reciclaje y el despilfarro de recursos energéticos, hacen que sea importante e imprescindible la necesidad de reconducir las ciudades hacia modelos más sostenibles.

Los modelos urbanos no sostenibles tienen un impacto muy negativo para el medio ambiente, los seres vivos, y el planeta, por lo que se hace indispensable una actuación inmediata. Algunas de las devastadoras consecuencias que provocan esta configuración de las áreas urbanas son enfermedades, golpes de calor, problemas respiratorios, muertes, calor extremo, sequías, extinción de especies o mala calidad del aire.

Para lograr modelos urbanos sostenibles, el problema puede abordarse desde diferentes

perspectivas como pueden ser, por citar algunas, el reciclaje, el uso de energías renovables, la naturación urbana, los desplazamientos con medios de transporte no contaminantes y el ahorro energético.

La sostenibilidad urbana se ha convertido en un reto al que hay que hacer frente a corto plazo, ya que las ciudades son los sistemas que mayor impacto tienen sobre el planeta y, por ello, se hace indispensable mejorar la organización y la gestión urbana tal y como se señala en el trabajo de Rueda (2012). Son un factor imprescindible en la lucha contra el cambio climático. En este sentido, los Objetivos de Desarrollo Sostenible (ODS) establecidos por la Agenda 2030 de las Naciones Unidas (ONU) (United Nations General Assembly, 2015) constituyen una agenda modelo global que debería servir como base para las acciones que se desarrollen en el ámbito urbano de cara a conseguir ciudades sostenibles.

Las tecnologías impulsadas por la inteligencia artificial se han propuesto como facilitadoras de cada uno de los ODS porque permiten la automatización, la trazabilidad y la optimización (Palomares y otros, 2021). En particular, la cantidad de datos recopilados en el dominio de la sostenibilidad urbana es tan alta que solo pueden explotarse desde una perspectiva de inteligencia artificial (Geyer y otros, 2021).

Las ciudades se han convertido en un factor determinante del cambio climático, ya que constituyen un lugar donde se consume gran parte de la energía (64 % del uso de energía primaria global) y se emiten altos niveles de gases de efecto invernadero (70 % del total global), debido al uso de combustibles fósiles como fuentes de energía (International Energy Agency, 2016).

Existen alternativas para que los ciudadanos reduzcan estas emisiones; una de ellas es el uso de energías renovables para satisfacer los consumos eléctricos. Así, sustituir las fuentes de energía contaminantes por energías renovables que respeten el medio ambiente es uno de los requisitos imprescindibles para conseguir ciudades energéticamente sostenibles y favorecer la lucha contra el cambio climático. Además, el cambio a fuentes de energía renovables en detrimento de las energías contaminantes mejora la salud y la calidad de vida por la no emisión de gases contaminantes, uno de los grandes problemas de las ciudades. Precisamente, uno de los objetivos propuestos en la Agenda 2030 para el Desarrollo Sostenible de la ONU es *hacer ciudades inclusivas, seguras, resilientes y sostenibles* (United Nations General Assembly, 2015).

La sostenibilidad urbana pretende implementar modelos orientados a garantizar la salud y la calidad de vida de los seres vivos que habitan en las ciudades, sin poner en compromiso a las futuras generaciones, respetando el planeta y el medio ambiente. Por lo tanto, redirigir las ciudades hacia modelos sostenibles es una prioridad para los organismos nacionales e internacionales. El desarrollo, la implantación y el uso de herramientas que faciliten la adopción de modelos urbanos sostenibles, se ha convertido en un reto para la comunidad

científica y para las administraciones públicas.

Los problemas a los que se intenta dar respuesta, desde el punto de vista de aplicación de la investigación desarrollada en este trabajo, dentro de la sostenibilidad urbana, son los siguientes:

- La selección automática del mejor emplazamiento para instalaciones fotovoltaicas en entornos urbanos y la evaluación del potencial energético fotovoltaico de áreas urbanas.
- La caracterización y modelización de los perfiles de consumo eléctrico de los hogares para contribuir a facilitar la gestión de la demanda energética y reducir el consumo eléctrico.
- La detección de zonas más desfavorables por el efecto isla de calor urbano (UHI). Este efecto es responsable de las temperaturas extremas que se producen en las zonas urbanas, alejadas del entorno rural, debido a la escasez de vegetación y a los materiales empleados para las construcciones masivas.
- La predicción de temperaturas analizando las características urbanas del entorno cercano en las zonas más desfavorables.
- La simulación de diferentes escenarios urbanos modificando las características de los tipos de vegetación del entorno cercano. Esta simulación permitirá realizar predicciones sobre las posibles mejoras de la temperatura de las zonas más desfavorables.

1.2. Minería de datos

En la actualidad se encuentran disponibles de forma gratuita una gran cantidad de datos de calidad que permiten analizar y modelar las ciudades, su situación, y su comportamiento (imágenes de satélites, imágenes láser, datos de estaciones meteorológicas, datos de consumo eléctrico, etc.). A partir de esos datos, es posible descubrir conocimiento oculto, novedoso, relevante y útil.

La dimensionalidad que alcanzan dichos datos (tamaño, heterogeneidad, cantidad ...), hacen prácticamente imposible su tratamiento de forma manual, por lo que es necesario desarrollar herramientas que sean capaces de llevar a cabo un proceso automático que permita descubrir conocimiento a partir de tal cantidad de datos.

Para llevar a cabo de forma automática dichas tareas de extracción e integración de datos, de procesamiento y limpieza, y de extracción de conocimiento oculto, útil, y relevante, se utilizan los procesos de minería de datos. Estos, a partir de la definición de objetivos,

permiten la generación de modelos que puedan ser utilizados desde un punto de vista aplicado en forma de herramientas.

Las técnicas de minería de datos ofrecen un enfoque adecuado para abordar los problemas que requieren el uso de fuentes de información heterogéneas y en las que las relaciones entre las diferentes variables involucradas no son triviales.

Una de las hipótesis de este trabajo de investigación es que el estudio, análisis y modelización de diferentes fenómenos que inciden en la sostenibilidad urbana puede abordarse mediante el uso de técnicas de minería de datos. Además, hay que tener en cuenta que es necesario un enfoque multidisciplinar para enfrentar los problemas relacionados con la sostenibilidad urbana. El desarrollo de herramientas que incidan en una mejora de ésta puede tener un impacto muy beneficioso en la planificación urbana, lo que supondrá una mejora del entorno.

1.3. Objetivos

El objetivo general que se planteó en este trabajo de investigación fue la generación de conocimiento y procedimientos basados en técnicas de inteligencia artificial y minería de datos que permitieran abordar y solucionar diferentes problemas que tienen una incidencia directa en la sostenibilidad urbana.

Este objetivo se concreta, por una parte, en el desarrollo de nuevos modelos basados en datos para caracterizar los problemas y las relaciones entre las distintas informaciones disponibles y en el desarrollo de nuevos métodos híbridos de minería de datos para descubrir conocimiento utilizando datos heterogéneos y, por otra parte, en la construcción de herramientas de simulación que implementen estos modelos y sirvan para apoyar la toma de decisiones.

Es decir, el objetivo general se desglosa en estos dos objetivos específicos:

1. **Construir modelos que extraigan información relevante a partir de elementos que influyen en la sostenibilidad urbana**.
2. **Utilizar esos modelos para desarrollar nuevos sistemas inteligentes que ayuden en la tarea de la toma de decisiones para alcanzar modelos urbanos sostenibles.**

Todo el conocimiento que se extrae como resultado de un proceso de minería de datos en forma de modelo es de gran utilidad para los expertos y puede ser utilizado por herramientas que ayuden en la toma de decisiones para alcanzar ciudades sostenibles.

1.4. Metodología

Existen diferentes metodologías que establecen las fases que debe seguir un proceso de minería de datos. La mayoría de ellas comparten múltiples características, como pueden ser las fases y tareas más determinantes, además de la iteratividad e interactividad del proceso. Por ser una de las más utilizadas en proyectos de minería de datos, en este trabajo de investigación se ha empleado la metodología CRISP-DM (Chapman y otros, 2000) para guiar todos los procesos de minería de datos llevados a cabo en los trabajos de investigación que se han abordado. En la Figura 1.1 se muestran las fases que se establecen en la metodología CRISP-DM, así como las actividades que se desarrollan en cada una de ellas.

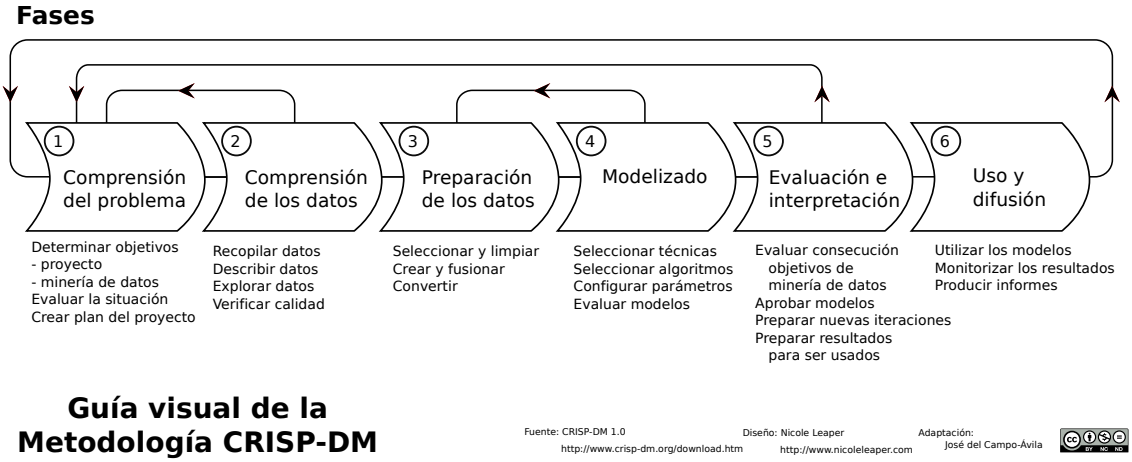


Figura 1.1: Fases y tareas de la metodología CRISP-DM.

La primera fase es la de comprensión del problema. Esta fase tiene como actividades principales: (a) la *determinación de objetivos*, que constituyen el núcleo del proyecto en sí; (b) la *evaluación de la situación*, que constituye un punto de partida realista, al tiempo que permite intuir la posibilidad de alcanzar los objetivos propuestos (se expone en el capítulo 2); y (c) la elaboración del *plan del proyecto*, que organiza todos los aspectos necesarios para avanzar en la consecución de los objetivos.

Las fases siguientes están dedicadas a la comprensión y preparación de los datos (fases 2 y 3 en el esquema de la Figura 1.1). Posteriormente se pueden aplicar algoritmos de aprendizaje automático sobre los datos para inducir modelos que representen el conocimiento extraído (fase 4). Una vez se hayan generado modelos que hayan podido ser validados, es necesario que sean evaluados por los expertos para comprobar su capacidad para ser utilizados (fase 5). La última fase de esta metodología (fase 6) se centra en la utilización y difusión de los resultados.

La Figura 1.2 muestra de forma esquemática las dimensiones de estudio de este trabajo. Puede observarse que los procesos de minería de datos aplicados en cada una de los dominios de investigación deben ser guiados y supervisados por el conocimiento de los expertos.

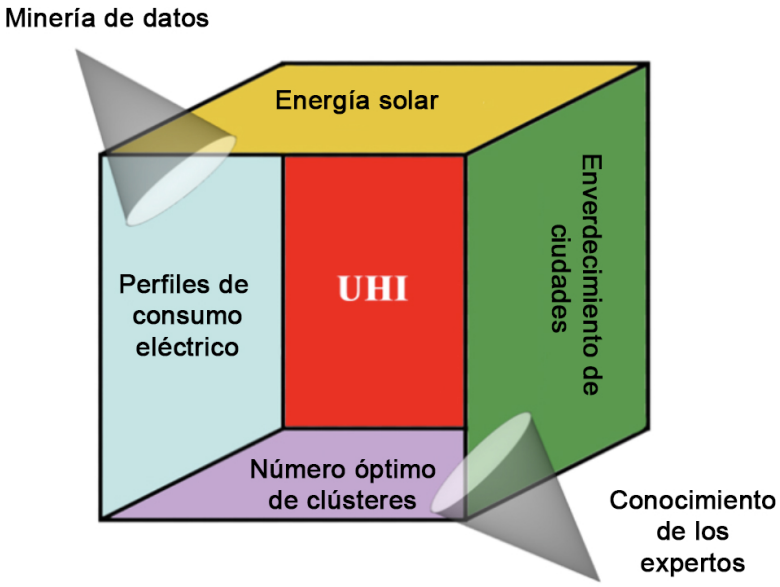


Figura 1.2: Dimensiones de la tesis. Elaboración propia

Una vez extraído el conocimiento en forma de modelos, y evaluado su potencial y validez, estos modelos se implementan en herramientas inteligentes. El objetivo principal de estas herramientas es servir de ayuda en el reto de alcanzar lo antes posible modelos urbanos sostenibles.

1.5. Aportaciones de la tesis

En esta sección, se describen las principales aportaciones que se han alcanzado como resultado de los trabajos de investigación realizados. En concreto, son las siguientes:

- Se han desarrollado modelos y herramientas que permiten determinar de forma automática los mejores emplazamientos en áreas urbanas en los que realizar instalaciones fotovoltaicas.
- Se ha desarrollado un sistema que permite la evaluación automática del potencial de generación fotovoltaica de un área urbana.

- Se ha propuesto un novedoso algoritmo para la determinación del número óptimo de clústeres en problemas de caracterización de perfiles de consumo de los hogares.
- Se ha propuesto una nueva metodología para caracterizar los perfiles de consumo eléctrico de los hogares de una ciudad.
- Se han analizado los elementos urbanos más influyentes sobre la temperatura de la superficie terrestre (LST)[1] analizando las características del entorno cercano.
- Se han desarrollado herramientas que permiten determinar las zonas más desfavorables por el efecto isla de calor urbano.
- Se han extraído modelos y generado herramientas que permiten predecir la temperatura de la superficie terrestre en cualquier área de cualquier ciudad, analizando las características urbanas del entorno.
- Se han inducido modelos y se han desarrollado herramientas que permiten realizar simulaciones de diferentes escenarios modificando la vegetación urbana (cantidad, tipo, y distancia), para estimar la LST que se obtendría si se aplicaran dichas configuraciones.

El desarrollo de herramientas informáticas basadas en los resultados obtenidos por procesos de minería de datos es una alternativa real para mejorar la sostenibilidad urbana, como se puede deducir de los antecedentes que se describirán en la Sección 2.3.

Estas herramientas pueden ser de utilidad a administraciones públicas, a planificadores urbanos y a particulares.

El conocimiento ofrecido por las herramientas inteligentes desarrolladas, se podrá tener en cuenta para determinar las posibles zonas óptimas en las que realizar instalaciones fotovoltaicas o de infraestructuras verdes urbanas, así como para priorizar las áreas en las que es más urgente actuar.

El uso combinado de las herramientas desarrolladas generará una serie beneficios: facilitar la integración de energías renovables en detrimento de las energías contaminantes, ayudar a conseguir un incremento de instalaciones de infraestructura verde urbana por la ayuda en la detección de las zonas donde sea más necesario intervenir para reducir las temperaturas, todo esto supondrá una mejor calidad del aire, menos enfermedades, menos calor, mejor calidad de vida ... y, por lo tanto, un importante avance hacia modelos urbanos sostenibles dirigidos a luchar contra el cambio climático.

[1]LST: Land Surface Temperature

1.6. Estructura de la tesis

Esta memoria está estructurada en 5 capítulos: en este primer capítulo se presenta una breve introducción al trabajo de investigación que se ha realizado. El segundo capítulo corresponde a los fundamentos necesarios para la realización del trabajo de investigación. El tercer y cuarto capítulo describen las propuestas metodológicas y resultados que se han obtenido en este trabajo de investigación relacionados con el desarrollo de nuevos modelos, algoritmos y herramientas que pueden contribuir a la sostenibilidad urbana, tanto en la modelización de los usos de la energía en la ciudad (Capítulo 3), como en el desarrollo de herramientas inteligentes para asesorar acerca del enverdecimiento de las ciudades (Capítulo 4). El último capítulo (Capítulo 5), corresponde a las conclusiones extraídas a partir del trabajo de investigación y a los posibles trabajos futuros. En el anexo se detallan los artículos publicados relacionados con las líneas de investigación de la tesis, así como los artículos que se encuentran en revisión.

2

Material y métodos. Fundamentos

En este capítulo se describen los materiales y fundamentos necesarios para los trabajos de investigación realizados.

Las principales fuentes de datos que se han utilizado para el entrenamiento de modelos y generación de conocimiento se detallan en la Sección 2.1. En la Sección 2.2 se describen los principales conceptos, fórmulas y fundamentos teóricos que han sido necesarios para poder alcanzar los objetivos propuestos. En la Sección 2.3 se realiza una introducción a la minería de datos, y se describen con más detalle las técnicas que se han utilizado en este trabajo, así como el software utilizado, los conceptos necesarios, y las aplicaciones de la minería de datos en el ámbito de la sostenibilidad urbana. También se presentan las métricas que se han utilizado para evaluar la calidad de los modelos.

2.1. Fuentes de datos

A continuación, se describen las principales fuentes de datos que han sido utilizadas para la consecución de los objetivos descritos en 1.5.

2.1.1. Imágenes satelitales: Landsat-8 y Sentinel-2

En la actualidad, las imágenes satelitales constituyen una de las herramientas más empleadas para modelar las ciudades, conocer su estado, y extraer una serie de características relevantes para la lucha contra el cambio climático. Existen multitud de satélites (MODIS, Landsat, Sentinel ...) que a diario toman imágenes de las ciudades, lo que favorece poder realizar estudios actualizados, profundos y detallados de las mismas. Los satélites utilizan sensores remotos para obtener información sobre diversas características de la superficie terrestre.

Para los objetivos establecidos en las fases iniciales de los procesos de minería de datos, se decidió centrar el estudio en dos de ellas: Sentinel-2 y Landsat-8, ya que según los expertos son las más adecuadas para alcanzar los objetivos establecidos.

Las imágenes satelitales tienen muchas aplicaciones prácticas en diversos sectores, como por ejemplo, clasificar el tipo de cobertura del suelo, detectar zonas verdes, y evaluar la salud de los cultivos.

Las imágenes satelitales Landsat-8 y Sentinel-2 son, en la actualidad, dos de las imágenes más utilizadas para la obtención y cálculo de variables determinantes para caracterizar y conocer el estado actual de las ciudades. Las imágenes Landsat-8 cubren una resolución de $30 \times 30 m/pixel$, mientras que las imágenes Sentinel-2 tienen una resolución de $10 \times 10 m/pixel$. Estas últimas tienen una utilidad especial para la caracterización de las zonas verdes de las ciudades con mayor detalle ya que permiten calcular el llamado índice de vegetación de diferencia normalizada (NDVI), como se describe en la Sección 2.2.2. Por su parte, las imágenes Landsat constituyen, en la actualidad, una de las tecnologías más empleadas para el cálculo de la temperatura de la superficie terrestre (LST) siguiendo los pasos descritos en la Subsección 2.2.3.

Las imágenes satelitales Landsat-8 y Sentinel-2 pueden obtenerse de forma gratuita desde Earth Explorer[1] (USGS, Department of the Interior, U.S.A.).

2.1.2. Nube de puntos LiDAR

La tecnología LiDAR consiste en la emisión de luz sobre los elementos de la superficie terrestre a través de un dispositivo láser por parte de vehículos aéreos como avionetas o drones. A partir de las características de la señal devuelta tras el rebote de la luz entre el dispositivo y los elementos urbanos de las ciudades, esta tecnología es capaz de generar nubes de puntos con las alturas de los elementos urbanos así como reconocer de qué elemento urbano se trata.

[1] `http://earthexplorer.usgs.gov`

En el trabajo de Sharma y otros (2021) se muestra todo el potencial de esta tecnología en términos de extracción de características urbanas. Una de las ventajas de usar imágenes LiDAR es la simplicidad que aporta a la fase de preparación de datos, ya que incluye una clasificación urbana para cada punto de la imagen (por ejemplo, edificio, vegetación o agua). La resolución es de $0,5$ puntos cada m^2.

Las imágenes LiDAR que se han utilizado en este trabajo se han obtenido del Centro Nacional de Información Geográfica (CNIG)[2] de forma gratuita para diferentes ciudades españolas.

2.1.3. Datos meteorológicos

La Agencia Estatal de Meteorología de España (AEMET)[3], ofrece información meteorológica de observaciones convencionales. De todas las que se registran por la AEMET, en este trabajo se han utilizado los registros históricos de las variables radiación global horaria y la temperatura ambiente. Además, la AEMET también ofrece información de tipo predictivo para varios parámetros meteorológicos; de éstos, en este trabajo se han utilizado las predicciones de temperatura, humedad, velocidad del viento, radiación solar y nubosidad. Todas estas variables, registros históricos y predicciones, son necesarias para poder evaluar el potencial de generación fotovoltaica en entornos urbanos.

Para acceder a los datos de observaciones convencionales, radiación y temperatura, y/o datos predictivos, AEMET ofrece una API que permite su descarga para las fechas que se especifiquen. En la Tabla 2.1 se describen las variables que se han recopilado, de forma diaria. Estas variables se han utilizado para los modelos de predicción a corto plazo de generación eléctrica de sistemas fotovoltaicos.

2.1.4. Datos de consumo eléctrico de los hogares

Los datos de consumo de electricidad de los hogares se recopilan, en general, por empresas privadas. Alguno de estos datos están disponibles públicamente gracias a organismos nacionales que apoyan la difusión de datos abiertos, como el descrito por Toussaint (2019).

Para poder desarrollar una metodología que caracterice los perfiles de consumo eléctrico de los hogares, se han utilizado dos fuentes de información distintas: la primera es un un conjunto de datos privado de consumo de electricidad en el sudeste de España, y la segunda fuente de información es un conjunto de datos público de Irlanda (Commission for Energy Regulation (CER), 2012).

[2] `https://centrodedescargas.cnig.es`
[3] `http://www.aemet.es`

Tabla 2.1: Parámetros meteorológicos ofrecidos por AEMET y su descripción

PARÁMETRO	**DESCRIPCIÓN**
gd_previo	Índice de radiación global observado el día anterior
kd_previo	Índice de claridad observado el día anterior
t9_12_diaprevio	Temperatura observada entre las 9:00 y las 12:00
t13_15_diaprevio	Temperatura observada entre las 13:00 y las 15:00
h8_diaprevio	Humedad observada a las 8:00
h14_diaprevio	Humedad observada a las 14:00
t_dia_diaprevio	Temperatura observada
precip_diaprevio	Índice de precipitaciones observado
Predicc_temp day_C	Predicción de temperatura
Predicc_relat_humidity_day_0_1	Predicción de humedad relativa diaria en el rango [0,1]
Predicc_cloudy_sky_day_0_1	Predicción de nubosidad diaria en el rango [0,1]
Predicc_cloudy 10_11_12	Predicción de nubes entre las 10:00 y las 12:00
Predicc_cloudy_13_14_15	Predicción de nubes entre las 13:00 y las 15:00
Predicc_temp_l 0_11_12	Predicción de la temperatura entre las 10:00 y las 12:00
Predicc_temp_l 3_14_15	Predicción de la temperatura entre las 13:00 y las 15:00
Predicc_R_hum 10_11_12	Predicción de la humedad entre las 10:00 y las 12:00
Predicc_R_hum 13_14_15	Predicción de la humedad entre las 13:00 y las 15:00

La selección de dos conjuntos de datos permitirá probar fácilmente la metodología que se ha desarrollado en este trabajo en dos regiones con climas diferentes (un aspecto relevante tal y como han indicado los expertos consultados). El clima del sudeste de España se caracteriza por un clima mediterráneo (veranos calurosos y secos e inviernos húmedos y frescos) y su temperatura media anual es de $18°C$. Irlanda, por su parte, tiene un clima oceánico templado (veranos cálidos, sin estación seca e inviernos frescos) y su temperatura media anual es de $10°C$.

Los valores de consumo registrados en Irlanda corresponden a periodos de consumo de 30 minutos, mientras que los de España son consumos horarios. Para poder trabajar con el mismo intervalo en ambos conjuntos de datos, los datos de Irlanda se han preparado de forma que los consumos sean horarios.

Tras esta preparación, en los dos conjuntos de datos, cada observación tiene 24 valores correspondientes a las 24 horas de consumo de un usuario durante un día.

Los conjuntos de datos finales a los que se les ha aplicado la metodología propuesta se describen a continuación:

- El conjunto de datos del sudeste de España se registró desde enero de 2020 hasta diciembre de 2021 para más de 3.000 usuarios. Se utilizaron un total de 2.396.741 observaciones después de filtrar los datos.
- El conjunto de datos de Irlanda se ha recopilado de aproximadamente 6.000 usuarios

durante dos años (2009-2010). Se utilizaron un total de 2.522.976 observaciones después de filtrar los datos.

El proceso de filtrado de datos que se ha aplicado a los 2 conjuntos de datos se detalla en la sección 3.3.1, en la que se describe la metodología propuesta.

2.2. Definiciones

En las siguientes secciones se presentan los conceptos y definiciones que se utilizan en los trabajos de investigación que se han realizado.

2.2.1. Cálculo del potencial de energía solar

Para estimar la energía que generará un sistema de energía solar fotovoltaica es necesario conocer, por una parte, información sobre la energía (radiación) que recibirá en función de la posición relativa sol-tierra y ubicación, y, por otra, información relativa a las características de la instalación, como área, tecnología, orientación (α) e inclinación (β) del sistema y factores de pérdidas.

Los datos de radiación solar que se registran por los servicios meteorológicos suelen ser valores medidos sobre superficie horizontal. A partir de estos datos y teniendo en cuenta la hora en que se registra la medida (posición relativa SOL-TIERRA) y la orientación e inclinación de la superficie que se va a evaluar se puede calcular la energía que se recibirá en esa superficie utilizando las expresiones propuestas por Iqbal (1983) y Coronas y Villarrubia (1983).

Primero es necesario calcular las componentes difusa (G_d) y directa (G_b) de radiación sobre superficie horizontal, a partir de la radiación global (G_g), para lo que se utilizan las siguientes expresiones:

$$G_d = \begin{cases} G_g \cdot (1 - 0{,}09 k_h) & 0 \leq k_h \leq 0{,}22 \\ G_g \cdot (0{,}9511 - 0{,}16 k_h + 4{,}388 k_h^2 - 16{,}638 k_h^3 + 12{,}336 k_h^4) & 0{,}22 < k_h \leq 0{,}8 \\ G_g \cdot 0{,}165 & 0{,}8 < k_h \end{cases} \tag{2.1}$$

$$G_b = G_g - G_d \tag{2.2}$$

donde k_h es el índice de transparencia horario:

$$k_h = \frac{G_g}{G_0} \tag{2.3}$$

con G_0 definido como la radiación solar extraterrestre horaria que se obtiene usando la expresión:

$$G_0 = I_{sc} E_0 \sin(\alpha) \tag{2.4}$$

donde I_{sc} es la constante solar , E_0 es el factor de excentricidad y α es la altura solar. El valor para la constante solar que se ha utilizado es 1367 W/m2.

A partir de estas componentes, se obtienen las correspondientes sobre superficie inclinada y/o orientada, $G_{d\beta\alpha}$ y $G_{b\beta\alpha}$ y la componente reflejada $G_{r\beta\alpha}$, utilizando las expresiones propuestas por Coronas y Villarrubia (1983). La radiación global sobre superficie inclinada y/o orientada es la suma de esta tres componentes:

$$G_{g\beta\alpha} = G_{b\beta\alpha} + G_{d\beta\alpha} + G_{r\beta} \tag{2.5}$$

Una vez conocida la radiación que recibirá el sistema fotovoltaico según la orientación e inclinación de la instalación y la hora, se puede calcular la energía que producirá utilizando el modelo propuesto por Osterwald (1986). Siguiendo este enfoque, para realizar estimaciones de la energía horaria producida por cada sistema, se utilizan también los siguientes datos meteorológicos: radiación solar horaria global sobre superficie de paneles de la instalación, y temperatura horaria. La potencia (P) generada por el sistema se estima mediante la expresión:

$$P = P_{STC} \frac{G_{\beta,\alpha}}{1000} (1 + \gamma(T_{mod} - 25)) GL \tag{2.6}$$

donde P_{STC} es la potencia del sistema en condiciones estándar, $G_{\beta,\alpha}$ es la irradiancia global en la superficie de los módulos, β es la inclinación de los módulos, α es la orientación, γ es el coeficiente de temperatura de la potencia máxima de los módulos, T_{mod} es la temperatura del módulo, y GL es el coeficiente global de pérdidas del sistema. La expresión incluye tanto las pérdidas producidas por temperatura como otro tipos de pérdida (ensuciamiento, pérdidas espectrales, etc.).

Por otro lado, en un escenario de estimación a corto plazo, la energía fotovoltaica producida se puede calcular con un distintos intervalos de anticipación. Utilizando el modelo propuesto por del Campo-Ávila y otros (2021) se calcula la predicción de la radiación solar horaria para el día siguiente. Este modelo utiliza como variables independientes la mayoría de las variables de entrada significativas seleccionadas en la

propuesta de Castangia y otros (2021), y toma como entrada los datos meteorológicos registrados durante el día en curso y determinadas previsiones meteorológicas para predecir el tipo de radiación esperada para el día siguiente (soleado, nublado, parcialmente soleado, etc.).

El modelo proporciona el índice de transparencia horaria esperado para cada hora entre las 9:00 y las 16:00 del día siguiente. Estos valores se utilizan para realizar la estimación de la radiación solar global por hora para el día siguiente lo que, junto con los datos de cada uno de los tejados del área de interés, permite realizar la estimación de energía horaria que se generará el día siguiente entre las 9:00 y las 16:00h.

2.2.2. Cálculo del índice de vegetación (NDVI) en imágenes satelitales

El Índice de Vegetación de Diferencia Normalizada (NDVI) es un índice numérico utilizado para evaluar la presencia de vegetación. Su rango oscila entre -1 y 1. La Tabla 2.2 ilustra la correspondencia entre los rangos de NDVI y varios elementos urbanos basada en el trabajo de Fusami y otros (2020).

Tabla 2.2: Elementos urbanos y rangos de NDVI

	Sin vegetación		Con Vegetación		
	Agua	Otros	Escasa	Moderada	Densa
Rango de NDVI	$[-1{,}0, 0{,}0]$	$(0{,}0, 0{,}2]$	$(0{,}2, 0{,}4]$	$(0{,}4, 0{,}6]$	$(0{,}6, 1{,}0]$

El valor NDVI para cada píxel de una imagen satelital Sentinel-2 se calcula usando las bandas roja ($b4$) e infrarroja cercana ($b8$) de acuerdo con la siguiente expresión:

$$\text{NDVI} = \frac{b8 - b4}{b8 + b4} \tag{2.7}$$

Además de la utilización de imágenes del satélite Sentinel-2, también se puede obtener el valor NDVI a partir de imágenes del satélite Landsat-8. El procedimiento es similar y sólo es preciso cambiar las bandas que se utilizan. Concretamente, la banda roja se identifica como $b4$ y la infrarroja cercana como $b5$. De tal forma que el cálculo se realiza de acuerdo con la siguiente expresión:

$$\text{NDVI} = \frac{b5 - b4}{b5 + b4} \tag{2.8}$$

Cuando se necesita mayor precisión para la clasificación de zonas urbanas en base al NDVI, se utilizan las imágenes Sentinel-2, ya que ofrecen una mayor resolución tal y como se describe en la sección 2.1.1.

2.2.3. Cálculo de la temperatura en la superficie terrestre

La temperatura de la superficie terrestre (LST) se utiliza a menudo como una métrica para identificar las áreas más afectadas por el efecto isla de calor urbano (UHI). Para calcular el valor de LST se emplean imágenes del satélite Landsat-8. Para generar una capa ráster con valores LST para cada píxel de una imagen seleccionada de una ciudad, el primer paso es calcular la radiación espectral de la parte superior de la atmósfera (TOA) (L_λ) utilizando los factores de reescalado proporcionados en el archivo de metadatos (se puede usar la Banda 10 o la Banda 11) (Mejbel Salih y otros, 2018):

$$\mathrm{L}_\lambda = M_L \cdot Q_{cal} + A_L \tag{2.9}$$

donde M_L y A_L representan los factores de cambio de escala multiplicativos y aditivos específicos de la banda de los metadatos, respectivamente; y Q_{cal} es el valor de píxel del número digital del producto estándar cuantificado y calibrado.

A continuación, la radiación espectral TOA (L_λ) se convierte en temperatura de brillo (BT) usando la siguiente fórmula:

$$\mathrm{BT} = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)} \tag{2.10}$$

donde K_1 y K_2 son constantes de conversión térmica específicas de banda de los metadatos, correspondientes a la misma banda utilizada para calcular L_λ.

Para obtener el valor de la temperatura LST son necesarios algunos cálculos intermedios (Mejbel Salih y otros, 2018). El término principal es la emisividad de la superficie terrestre (LSE o ε_λ), que está influenciada por ciertas constantes de emisividad y la proporción de vegetación (calculada a partir de valores NDVI) según:

$$\mathrm{LST} = \frac{BT}{1 + \frac{\lambda \cdot BT}{\rho} \cdot \ln \varepsilon_\lambda} \tag{2.11}$$

donde λ es la longitud de onda media de la banda utilizada y ρ es una constante (dependiente de constantes universales como las de Boltzmann o Planck).

2.2.4. Componentes conexas

El etiquetado de componentes conexas (Connected Components Labelling – CCL –), es una técnica de procesamiento de imágenes utilizada para asignar una etiqueta única a todos los píxeles que están conectados, siguiendo criterios de vecindad y de similitud (He

y otros, 2017). Es una técnica que se ha utilizado en varios de los trabajos de investigación realizados y expuestos en esta tesis.

El objetivo común en dichos trabajos ha sido la identificación de componentes conexas en imágenes ráster[4]. Por ejemplo, una imagen ráster del índice NDVI (presentado en la sección 2.2.2) con los valores discretizados según la Tabla 2.2, será una imagen con la clasificación de diferentes tipos de terreno urbano. Concretamente se usará 1 para los elementos urbanos que correspondan con agua, 2 para zonas sin vegetación, 3 para zonas con vegetación escasa, 4 para zonas con vegetación moderada, y 5 para zonas con vegetación densa.

En los algoritmos de etiquetado de componentes conexas, suele ser común seleccionar el nivel de detalle de vecindad. Normalmente 4 u 8 vecinos. En este trabajo, los algoritmos empleados buscan similitud en los píxeles de los 8 vecinos para formar componentes conexas.

En la Figura 2.1 se muestra un ejemplo en el que se seleccionan diferentes componentes conexas en función del valor de los píxeles. Se generan 4 componentes conexas: una para los píxeles con valor 1, otra para los píxeles con valor 2, otra para los píxeles con valor 3 y otra para los píxeles con valor 4.

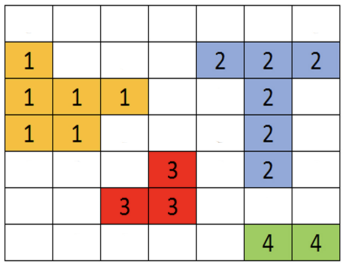


Figura 2.1: Ejemplo de generación de componentes conexas en imágenes. Elaboración propia.

2.2.5. Geometría básica de triángulos

El algoritmo propuesto para determinar el número óptimo de clústeres de forma automática, descrito en la sección 3.3.1.1, se basa en conceptos geométricos relativos a los triángulos. A continuación, se repasan los principales conceptos teóricos en los que se fundamenta la construcción del nuevo algoritmo.

Sean p, q y r tres puntos que definen un triángulo y p_x, p_y, q_x, q_y, r_x y r_y sus coordenadas

[4]Imágenes con un valor para cada píxel. Formato muy común en entornos de computación urbana.

para el eje X y el eje Y. La función que calcula el área de ese triángulo es:

$$area(p,q,r) = \left| \frac{1}{2} \cdot ((q_x \cdot p_y - p_x \cdot q_y) + (r_x \cdot q_y - q_x \cdot r_y) + (p_x \cdot r_y - r_x \cdot p_y)) \right| \quad (2.12)$$

Sean p y r los dos vértices más lejanos de un triángulo. La función que calcula la pendiente (en grados) es:

$$pendiente(p,r) = \arctan\left(\frac{r_y - p_y}{r_x - p_x}\right) \cdot 180°/\pi \quad (2.13)$$

2.3. Minería de datos para el desarrollo de herramientas inteligentes

El ritmo al que se generan grandes cantidades de datos relativos a cualquier dominio de la vida real supone una oportunidad y un reto para la ciencia. Un correcto procesamiento de los datos adecuados podría conducir al descubrimiento de conocimiento oculto, válido, y relevante, lo que podría suponer una serie de avances en cualquier ámbito o dominio: medio ambiente, salud, etc. Debido al enorme tamaño de los datos, a su heterogeneidad, y a su ritmo de crecimiento en la mayoría de los dominios se hace inviable el procesamiento manual de los mismos, por lo que parece imprescindible la aplicación de metodologías que permitan la automatización de los procesos que deben llevarse a cabo para la extracción de información. Algunas de esas metodologías, son las *metodologías de minería de datos* que, combinadas con el conocimiento de los expertos, permiten automatizar todos los procesos necesarios para la extracción de conocimiento a partir de los datos, siguiendo una serie de etapas bien definidas y organizadas.

Un proceso de minería de datos no es más que un proceso cuya finalidad es la extracción de conocimiento novedoso, útil y comprensible, a partir de la selección, extracción, y procesamiento de los datos (Chapman y otros, 2000).

Las metodologías de minería de datos definen una serie de etapas y actividades que permiten la extracción de dicho conocimiento. Las etapas y actividades son similares en todas las metodologías, compartiendo tres etapas (Freitas, 2002): el pre-procesamiento de los datos, la generación de modelos para la extracción y descubrimiento del conocimiento, y la evaluación de los modelos. En la Figura 2.2, se muestra el esquema con las tres etapas más relevantes.

Otra característica que comparten la mayoría de las metodologías es que el proceso que se sigue para lograr los objetivos es *iterativo* e *interactivo*. El proceso de extracción de conocimiento es *interactivo* porque requiere de la intervención de los expertos para evaluar

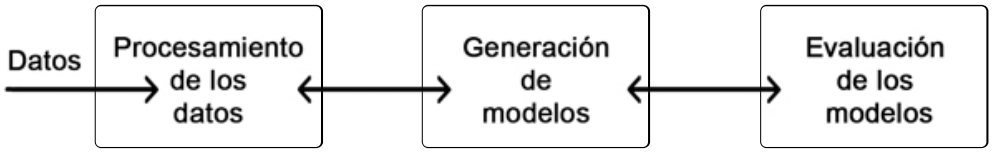


Figura 2.2: Esquema general de las etapas del proceso de minería de datos. Elaboración propia.

la validez y la utilidad de los modelos generados, y para reconducir las futuras iteraciones sobre las etapas de la metodología de minería de datos para guiar los modelos hacia los objetivos deseados. La salida de una etapa no tiene por qué ser el inicio de la siguiente, sino que puede reconducir y realimentar una etapa anterior. Es *iterativo* porque requiere de varias iteraciones no lineales en las que se puede ir saltando de unas etapas a otras para reconducir el proceso. La sucesión de iteraciones dará lugar a un proceso más refinado.

Las *metodologías de minería de datos* están diseñadas para descubrir conocimiento gracias a la generación de modelos. La fase de generación de modelos se sitúa como etapa clave en cualquier proceso de minería de datos. La entrada a esta etapa de modelizado, como se observa en la Figura 2.2, son datos pre-procesados de forma acorde a la técnica de modelizado que se vaya a utilizar.

Algunas de las principales disciplinas que dan soporte a las metodologías de minería de datos son: el aprendizaje automático, la estadística, el big data, la computación paralela, la recuperación de información, y el procesamiento de imágenes y señales. El *aprendizaje automático o computacional* (*machine learning* en inglés) suele ser la disciplina con más peso para realizar el análisis inteligente de los datos, ya que dispone de una gran cantidad de algoritmos que permiten generar modelos para la resolución de problemas de distintos tipos.

Según sea el tipo de las tareas que se tengan que resolver para tratar un problema, se utilizarán unos algoritmos de aprendizaje automático u otros. Los principales tipos de tareas son las tareas *descriptivas* y las tareas *predictivas* (Hernández Orallo y otros, 2004). Las tareas descriptivas están enfocadas a explicar o resumir los datos, a identificar propiedades. Se puede citar como ejemplo de este tipo de tareas las tareas de agrupamiento. Las tareas predictivas están enfocadas a predecir valores futuros o desconocidos de variables relevantes. Como ejemplo de este tipo de tareas se encuentran las tareas de clasificación o regresión.

Otra forma de clasificar el aprendizaje automático se basa en la información de la que se dispone. Si se conocen los valores de una variable sobre la que se quieren realizar

predicciones, se trata de un problema de *aprendizaje supervisado*. En estos casos, los algoritmos de aprendizaje automático pueden entrenar modelos ya que se conoce el valor de esa variable (clase). En el aprendizaje *no supervisado* no existe ninguna clase, por lo que los algoritmos no pueden entrenar modelos para realizar predicciones. En estos casos, los algoritmos suelen generar modelos que resuelven tareas descriptivas. Un ejemplo de aprendizaje no supervisado, son los problemas de agrupamiento.

El objetivo final de un proceso de minería de datos consiste en construir modelos que representen el conocimiento oculto en los datos. Estos modelos se utilizarán para la creación de herramientas inteligentes que permitan avanzar en el campo de aplicación del proceso. La alta dimensionalidad de los datos es el principal problema, sobre todo si el volumen de datos es alto, existen múltiples fuentes heterogéneas, y para poder abordar el problema hay que trabajar con múltiples variables.

Existen numerosos ejemplos de aplicación de la minería de datos en el campo de la sostenibilidad, como son los algoritmos predictivos de aprendizaje automático basados en los datos para estimar el uso de energía solar en los edificios de las ciudades (Kontokosta y Tull, 2017), la integración de diferentes fuentes de datos para extraer nuevo conocimiento (Dutta y otros, 2014) o la determinación de las posibles mejores localizaciones urbanas para realizar nuevas instalaciones (Doorga y otros, 2018).

En el proceso de minería de datos, los métodos y algoritmos que se utilizan en la fase de modelizado constituyen un aspecto clave. En el contexto de la sostenibilidad urbana se usa la misma variedad de técnicas que en cualquier otro dominio: aprendizaje supervisado, no supervisado, o una combinación de ambas técnicas (Fan y otros, 2008; del Campo-Ávila y otros, 2021), empleando algoritmos para estimar diferentes modelos como los árboles de decisión, regresiones lineales o redes neuronales (Moreno-Sáez y otros, 2013; Moreno-Sáez y Mora-López, 2014), que utilizan como entrada datos de múltiples fuentes heterogéneas (imágenes, datos meteorológicos, sensores, etc.) (Dutta y otros, 2014; Kong y otros, 2020).

Otras herramientas necesarias para mejorar la gestión y planificación de las futuras intervenciones en materia de urbanización sostenible son las que permiten simular diferentes escenarios y predecir el comportamiento esperado. Existen diferentes métodos de simulación, pero en el ámbito de la sostenibilidad se pueden destacar los siguientes (Moon, 2017): modelado y simulación de sistemas basados en agentes, en eventos discretos y en sistemas dinámicos. Entre estos, los sistemas inteligentes basados en agentes se han usado, por ejemplo, para la planificación del uso del terreno y el transporte (Motieyan y Mesgari, 2018).

2.3.1. Software para proyectos de minería de datos

En la actualidad *python* (Python Core Team, 2019) y *R* (R Core Team, 2019) son dos de los lenguajes más utilizados en proyectos de minería de datos y de inteligencia artificial. Su potencial, simplicidad, y amplio abanico de funciones, librerías, multidisciplinaridad, diversidad de aplicaciones, y acceso simple y gratuito a la información necesaria, los hacen idóneos para trabajar en proyectos de estas características y en proyectos de cualquier dominio.

Existen librerías muy completas con algoritmos de aprendizaje automático para resolver múltiples tareas como la inducción o entrenamiento de modelos, así como para evaluar su calidad. La presentación de resultados de procesos experimentales es imprescindible en proyectos de este tipo, siendo otro de los puntos fuertes de estos lenguajes. El desarrollo de herramientas interactivas basadas en datos, suele ser otro de los aspectos que ofrecen estos lenguajes, y que encaja con algunos de los objetivos de la tesis.

En la actualidad, para tareas de minería de datos espacial [5], se suelen utilizar sistemas de información geográfica GIS como Q-GIS o Arc-GIS. *R* dispone de una amplia gama de librerías específicas de reciente desarrollo que permiten extraer información relacionada con la sostenibilidad urbana a partir de imágenes satelitales o aéreas. Esta característica, unida a su versatilidad para poder crear herramientas web basadas en modelos conducidos por datos, hace idóneo su uso para algunos de los trabajos de investigación que se han realizado.

Por todo lo anterior, la selección de alguno de estos lenguajes, o una combinación de los mismos, parece idónea para los trabajos de investigación que se han realizado en esta tesis.

2.3.2. Algoritmos de agrupamiento

El agrupamiento (o *clustering*) es una técnica de aprendizaje automático no supervisado que se utiliza para separar y agrupar las observaciones de un conjunto de datos en diferentes subconjuntos o grupos. Las observaciones en cada subconjunto deben ser lo más similares posible entre sí y lo más diferentes posible con respecto a las observaciones de otros subconjuntos; todo ello considerando alguna función de distancia que permita evaluar esa similitud y separación.

En esta sección se describen los aspectos más relevantes a tener en cuenta para realizar una tarea de agrupamiento: distancia entre observaciones, algoritmos, indicadores para

[5] Extracción de conocimiento útil, novedoso, y relevante, a partir del procesamiento de imágenes aéreas o satelitales

evaluar la idoneidad de los modelos y determinación del número adecuado de clústeres.

2.3.2.1. Tipos de algoritmos de agrupamiento

Existe un gran cantidad de algoritmos de agrupamiento y estos se pueden clasificar según el modo en que realizan dicho agrupamiento (Xu y Tian, 2015): particional, jerárquico, en función de la densidad, y en función de la teoría de grafos. Dentro de toda esta variedad de algoritmos, algunos de ellos están mejor preparados para trabajar con conjuntos de datos masivos en diferentes dominios. Atendiendo a los requisitos de tiempo para la ejecución, se deberían potenciar los algoritmos con complejidades inferiores, aquellos de un orden $O(n)$ o, a lo sumo, $O(n \cdot log(n))$. A continuación se describen algunos de los algoritmos más utilizados:

- *Particionales.* Entre estos, k-means, propuesto por MacQueen (1967), es el más utilizado. Se empieza seleccionando k grupos, cada uno de los cuales consta de un único punto aleatorio (*centroide*). A partir de esos grupos iniciales, k-means asigna cada observación del conjunto de datos al grupo cuyo *centroide* está más cercano a esa observación, según la función distancia que se utilice. Un proceso de actualización continúa cambiando los *centroides* hasta que se cumplan algunos criterios de convergencia. Este procedimiento es de baja complejidad temporal, $O(n)$, siendo n el número de observaciones.
- *Jerárquicos.* De entre estos, Bisecting k-means es uno de los más utilizados (Steinbach y otros, 2000). Es un algoritmo de agrupamiento jerárquico divisivo que usa k-means para refinar divisiones sucesivas. En lugar de dividir el conjunto de datos en grupos de k desde el principio, la bisección de k-means divide un grupo en dos subgrupos en cada paso de bisección (mediante el uso de k-means) hasta que se obtienen k clústeres. Su complejidad temporal es, como k-means, lineal en el número de observaciones ($O(n)$). Si el número de clústeres es grande, bisecting k-means podría ser incluso más eficiente que el algoritmo clásico de k-means. Otro enfoque jerárquico con baja complejidad temporal ($O(n)$) es BIRCH (Zhang y otros, 1996).
- *Basados en densidad.* DBSCAN es el algoritmo de agrupamiento más conocido en esta categoría (Ester y otros, 1996). La idea básica es agrupar aquellas observaciones del espacio de datos que se encuentran en una región con alta densidad. OPTICS es una evolución de DBSCAN que incorpora dos parámetros (el radio del vecindario y el número mínimo de puntos en un vecindario) para mejorar la inducción del modelo (Ankerst y otros, 1999). Ambos algoritmos muestran una complejidad ligeramente mayor que la lineal: $O(n \cdot log(n))$.

- *Basados en la teoría de grafos.* Power Iteration Clustering (PIC) es una técnica de agrupación basada en la teoría de grafos (Lin y Cohen, 2010). Requiere una matriz de afinidad normalizada por filas y puede encontrar la partición de agrupamiento usando complejidad lineal en el número de observaciones ($O(n)$). Aunque su complejidad es baja, asume la existencia de una matriz de afinidad de forma previa a la ejecución del algoritmo PIC, lo que es un proceso costoso. Para cualquier otro método basado en la idea de particionar el grafo (como CLICK (Sharan y otros, 2003)), también se debe construir un grafo ponderado no dirigido, según Dhanapal y Perumal (2016). Cada observación en el conjunto de datos es un vértice y el valor de similitud entre dos observaciones es el peso del borde que conecta los dos vértices. La generación de esta matriz, en caso de no conocerse previamente, conlleva una complejidad temporal superior ($O(n^2)$).

2.3.2.2. Funciones de distancia

Se necesita una función de distancia para medir la similitud entre las observaciones, de forma que el algoritmo de agrupamiento pueda reunir las observaciones similares.

Una de las funciones más comunes es la distancia euclídea. En caso de considerar m variables (o atributos) para definir una observación se puede formular (en espacio euclidiano m-dimensional) como:

$$euclidea\_dist(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} \tag{2.14}$$

Al calcular el error de ajustar todo el conjunto de datos (definido por n observaciones) en un modelo (con k clústeres), la suma de las distancias euclídeas al cuadrado (SSE) entre cada observación (o_j donde $j \in [1, n]$) y el *centroide* ($centroide_i$ donde $i \in [1, k]$) del clúster correspondiente ($o_j \in C_i$) se usa para obtener la distorsión media (MD) en el modelo completo (Shi y otros, 2021):

$$SSE = \sum_{i=1}^{k} \sum_{o_j \in C_i} euclidea\_dist(o_j, centroide_i)^2 \quad ; \quad MD = \frac{SSE}{n} \tag{2.15}$$

Otros tipos de distancias tienen en cuenta la forma de las curvas, como la distancia Dynamic Time Warping (DTW) (Sakoe y Chiba, 1978) o el kernel distancia (Cuturi, 2011).

2.3.2.3. Índices para evaluar agrupamientos

Para evaluar la calidad de los modelos de agrupamiento, existen muchos índices, pero cuando no hay información sobre los grupos reales a los que pertenece cada observación solo es posible la evaluación interna. Los indicadores internos suelen evaluar qué tan similares son las observaciones de un mismo clúster (cohesión), y lo diferentes que son las observaciones de un clúster con respecto a las observaciones de otros clústeres (separación). Existe una gran diversidad, pero las métricas más comunes que se utilizan para evaluar la calidad de los modelos de agrupamiento son las siguientes:

- *Índice de Silhouette*, definido por Rousseeuw (1987), es una métrica que combina: (a) la disimilitud promedio de una observación con todos los demás objetos de su propio grupo (cohesión); y (b) la disimilitud promedio de una observación con todos los demás objetos de su grupo más cercano (separación). Esta combinación da como resultado el cálculo del índice para cada observación $s(o_j)$. Cuando este índice se agrega para cada observación en un grupo y luego se agrega para todos los grupos, se obtiene el ancho de silueta promedio general. Este valor general oscila en el rango [-1,1], donde los valores más altos indican un modelado adecuado de los grupos.
- *Índice de Davies-Bouldin*, descrito por Davies y Bouldin (1979), calcula para cada par de grupos o clústeres, la relación entre: (a) su dispersión intra-clúster (o cohesión); y (b) la distancia entre sus *centroides* (separación). A continuación promedia las proporciones máximas para cada grupo. El índice de Davies-Bouldin oscila en el rango [0,1], donde los valores más bajos significan mejores modelos de agrupamiento.

2.3.2.4. Determinación del número óptimo de clústeres

Encontrar cómo se deben asignar correctamente todos los datos a diferentes grupos es un problema fundamental, incluso cuando se conoce el número óptimo de grupos, pero este problema se agrava cuando se desconoce dicho número (Ezugwu y otros, 2021). Existen diferentes alternativas para buscar el número adecuado de clústeres o grupos. Los dos enfoques principales son: (a) creando modelos de agrupamiento con diferentes valores de k en un rango predefinido ($k \in [k_{min}, k_{max}]$) y posteriormente determinando los mejores valores; y (b) construyendo una sucesión de modelos de agrupamiento, con k tomando valores crecientes y no necesariamente consecutivos, hasta que la aceptación de una condición detenga la búsqueda.

Existe una gran diversidad de métodos y algoritmos siguiendo el primer enfoque. Uno de los más sencillos se basa en los valores del índice de Silhouette anteriomente mencionado,

ya que una forma de elegir el número de clústeres, k, es seleccionar el valor que resulte en un valor más alto del índice de Silhouette.

Otra idea es la conocida como método del codo (elbow), que considera como mejor número de clústeres aquel en el que ya no hay un cambio significativo entre el valor del índice para ese número de clústeres y el valor del índice para el siguiente número propuesto de clústeres. Esta identificación se realiza manualmente visualizando una curva, lo que puede provocar situaciones en las que los expertos no puedan identificar claramente el punto del codo (curvas suaves). Para superar esta dificultad en el trabajo de (Shi y otros, 2021) se propone la selección del valor de k donde se observa el ángulo mínimo entre valores consecutivos. En el paquete Nbclust (Charrad y otros, 2014) se calculan varias métricas (incluso decenas de métricas) y se selecciona el valor de k votado por la mayoría. Argumentan que es difícil llegar a una decisión unánime sobre el número óptimo de agrupamientos.

Con respecto al segundo enfoque, existen también múltiples alternativas. En la Sección 3.1.2 se señalará cómo dos algoritmos, G-Means (Hamerly y Elkan, 2003) o X-Means (Pelleg y Moore, 2000), se han utilizado en el contexto de los patrones de consumo eléctrico de los hogares. Sin restringir el alcance a este dominio, Ezugwu y otros (2021) hacen una revisión en la que se compilan otros algoritmos de agrupamiento dinámico (como *PLDC*, *PSOAC* o *IDisABC*). Independientemente del dominio, estos autores concluyen que existe un gran problema con los conjuntos de datos a gran escala, especialmente cuando se manejan problemas de agrupamiento del mundo real.

2.3.3. Algoritmos de aprendizaje supervisado

Los algoritmos de clasificación de aprendizaje automático implican la identificación de relaciones entre los datos de entrada y los datos de salida. Los datos de entrada pueden ser discretos o continuos, mientras que los datos de salida (clase) suelen ser una variable discreta (Liu y Wu, 2012). Los métodos de regresión se emplean cuando la variable de salida o clase es numérica. En las técnicas de aprendizaje supervisado, se conoce la clase a la que pertenecen las observaciones y los modelos aprenden a predecir la variable dependiente (clase) a partir de ejemplos etiquetados.

2.3.3.1. Tipos de algoritmos de aprendizaje supervisado

Existe una gran cantidad de algoritmos de clasificación y regresión; esta revisión se centra en los algoritmos de regresión que son los más adecuados para abordar los problemas que se tratan en este trabajo de investigación. Se enumeran aquellos que se consideran

relevantes para el trabajo.

Las redes neuronales artificiales ANN consisten en nodos interconectados que se organizan en capas de entrada, capas ocultas y capas de salida (Murtagh, 1991). Las conexiones entre neuronas (pesos) se ajustan para alterar la relación entre entradas y salidas y, con la ayuda de funciones de activación, filtrar valores de entrada y generar una salida deseada. Las ANN a menudo emplean la técnica de retropropagación (backpropagation) para ajustar los pesos de la red para ajustar la salida deseada a las variables de entrada de los ejemplos etiquetados.

Las máquinas de vectores de soporte SVM son un tipo de algoritmo de clasificación que utiliza clasificadores lineales para identificar el hiperplano que divide los datos en diferentes categorías (Cortes y otros, 1995). Los modelos específicos para regresión, como Support Vector Regression (SVR) (Smola y Schölkopf, 2004) presentados por Awad y Khanna (2015), son una variante de SVM diseñada específicamente para problemas en que la clase es numérica. El algoritmo busca la curva o línea (hiperplano) que mejor capture la tendencia de los datos. Genera dos bandas simétricas paralelas a la curva separadas por un margen específico (denominado ε).

Los árboles de decisión son algoritmos de aprendizaje automático supervisados que pueden generar modelos predictivos interpretables para problemas de clasificación o regresión. Cada rama del árbol representa una regla expresada como una declaración condicional que se puede entender sin conocimiento experto. C4.5 Quinlan y Ross (1993) es un algoritmo de árbol de decisión ampliamente utilizado. Cuando la variable objetivo a predecir es numérica, se utilizan árboles de regresión como pueden ser los generados por el algoritmo CART (Breiman y otros, 1984).

Otro enfoque comúnmente utilizado para mejorar la precisión de modelos aislados es el de los sistemas multiclasificadores. Entre ellos Random Forest (RF) (Breiman, 2001) es un algoritmo basado en árboles de decisión que ha demostrado un gran éxito (Wyner y otros, 2017). Su uso implica generar múltiples árboles de decisión usando diferentes subconjuntos de atributos. Si se emplean árboles de regresión en lugar de árboles de decisión, se crea un bosque de árboles de regresión.

Bagging (Breiman, 1996) es otro de esos métodos para crear sistemas multiclasificadores. En él se entrenan múltiples modelos débiles con diferentes conjuntos de observaciones seleccionadas al azar del conjunto de datos. Los datos de entrenamiento para cada modelo nuevo pueden incluir datos usados para entrenar otros modelos. La salida del modelo se forma combinando la salida de los modelos débiles.

XGBoost (XGB), o eXtreme Gradient Boosting, es un algoritmo predictivo de aprendizaje automático supervisado que utiliza el principio del boosting (otro tipo de

sistema multiclasificador) para generar secuencialmente múltiples modelos de predicción débil (Chen y Guestrin, 2016). Cada modelo se basa en los resultados del modelo anterior para crear un nuevo modelo con un poder predictivo mejorado y una mayor estabilidad en sus resultados. Se utiliza un algoritmo de optimización como el descenso de gradiente para combinar estos modelos débiles en un modelo más robusto.

2.4. Métricas para la evaluación de modelos

Para evaluar la precisión de los modelos inducidos por algoritmos de aprendizaje supervisado para problemas de regresión se utilizarán las siguientes métricas (Stroock, D. W, 2011): error absoluto medio (MAE), error cuadrático medio (RMSE) y el coeficiente de determinación (R^2).

El coeficiente de determinación, también conocido como R^2, es una medida estadística que proporciona información sobre el ajuste de un modelo de regresión. Específicamente, representa la proporción de la varianza en la variable dependiente (respuesta) que es predecible a partir de la variable independiente (predictor). La expresión para calcular el coeficiente de determinación R^2 es:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.16}$$

donde $\hat{y}_i$ representa el valor predicho de la variable dependiente para la observación i, $\bar{y}$ es el valor medio de la variable dependiente, y_i es el valor observado de la variable dependiente para observación i, y n es el número de observaciones. La fórmula también se puede expresar como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.17}$$

donde el numerador representa la suma de errores al cuadrado (SSE) y el denominador representa la suma total de cuadrados (TSS). Por lo tanto, R^2 puede interpretarse como la proporción de TSS que es explicada por SSE. El valor de R^2 varía de 0 a 1, donde 0 indica que el modelo no explica nada de la variabilidad de los datos de respuesta en torno a su media, y 1 indica que el modelo la explica perfectamente.

R^2 a menudo se usa como una métrica de bondad de ajuste para determinar qué tan bien se ajusta el modelo a los datos. Una ventaja de R^2 es que es fácil de interpretar, ya que proporciona una indicación simple de la proporción de variabilidad en la variable de respuesta que explica el modelo. Sin embargo, una desventaja es que R^2 no proporciona

información sobre la validez de los supuestos del modelo o la importancia estadística del modelo. Por lo tanto, debe usarse junto con otras métricas para evaluar la calidad general del modelo. En este trabajo de investigación se utilizan MAE y RMSE, definidos de la siguiente manera:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.18}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.19}$$

Tanto MAE como RSME ofrecen información útil sobre el rendimiento de un modelo de regresión, ya que miden el error en el que se incurre al aproximar y_i a $\hat{y}_i$, pero tienen diferentes interpretaciones y aplicaciones. MAE se usa a menudo para comparar los errores de predicción de diferentes modelos, mientras que RSME se usa a menudo para evaluar el ajuste general de un modelo a los datos. En general, un valor más bajo para cualquiera de las métricas indica un mejor ajuste del modelo a los datos.

3

Modelos para la generación y consumo de energía en entornos urbanos

En este capítulo se describen los trabajos realizados en el ámbito de la energía para contribuir a una mejor sostenibilidad urbana.

Por una parte, se ha elaborado una metodología que permite conocer cuál es el potencial de generación de electricidad con sistemas de energía solar fotovoltaica en un área urbana, tanto a largo como a corto plazo. Esta metodología permite también determinar cuáles son los mejores emplazamientos para este tipo de instalaciones. La generación de electricidad con este tipo de sistemas puede contribuir de manera significativa a la descarbonización de las ciudades.

Por otra parte, se analizan los consumos energéticos domésticos para generar modelos que permitan caracterizarlos y estudiar cómo ese conocimiento puede contribuir a una mejor demanda y gestión del consumo de energía en la ciudad.

Se presenta primero el estado del arte sobre los aspectos energéticos en la ciudad en cuanto a generación con energías renovables y consumo de electricidad. Después se describen los trabajos de investigación y resultados obtenidos en cada uno de estos dos ámbitos.

3.1. Estado del arte

En esta sección se describe el estado del arte sobre los modelos de generación y consumo de energía en entornos urbanos. Se ha realizado un estudio y análisis de las principales técnicas, tecnologías y trabajos que tienen relación con la energía en entornos urbanos.

Por una parte, respecto a la generación de electricidad con sistemas de energía solar fotovoltaica, se ha hecho una revisión de los trabajos que abordan cómo se puede conocer o predecir con antelación la energía solar que se generará a corto plazo por este tipo de sistemas. También se ha caracterizado la energía que podría ser generada a largo plazo por los mismos. En ambos casos, el objetivo es poder determinar los mejores emplazamientos urbanos en los que realizar instalaciones fotovoltaicas.

Respecto a la caracterización de los consumos energéticos domésticos, se ha hecho una revisión de las diferentes propuestas que se han realizado, con especial interés en aquellas que persiguen identificar los perfiles de consumo eléctrico más comunes en los hogares.

3.1.1. Emplazamiento de instalaciones fotovoltaicas en ciudades

Para analizar la posible generación de electricidad en las ciudades mediante sistemas fotovoltaicos es importante analizar la evolución que ha tenido en los últimos años este tipo de sistemas. Así, se puede afirmar que la energía solar ha experimentado un fuerte incremento en los mix energéticos respecto a otros sistemas de generación.

En la Figura 3.1 se muestra cómo ha sido esta evolución a nivel mundial desde el año 2016 al 2021 según el informe International Renewable Energy Agency, IRENA (2023). La participación de las energías renovables en el crecimiento de la capacidad total de generación de energía instalada alcanzó el 81 % en 2021, frente al 79 % en 2020. La cuota renovable de la generación total en 2021 fue de un 38,3 %.

Mucha de esta capacidad instalada corresponde a grandes infraestructuras fotovoltaicas que están alejadas de las ciudades; sin embargo, se ha producido un fuerte aumento de instalaciones distribuidas en edificios, industria y viviendas en Europa, Estados Unidos y Japón (Agency, 2020), lo que se conoce como instalaciones fotovoltaicas de autoconsumo. Esto es muy importante ya que la producción local aumenta la responsabilidad de los ciudadanos en el consumo y ahorro de energía, y contribuye a aumentar la autosuficiencia energética.

La disponibilidad de espacio para realizar instalaciones fotovoltaicas en las zonas urbanas es limitada. Se estima que los tejados representan entre el 20 % y el 25 % de la superficie urbana (Akbari y Rose, 2008). Son, por tanto, un excelente recurso a explotar mediante

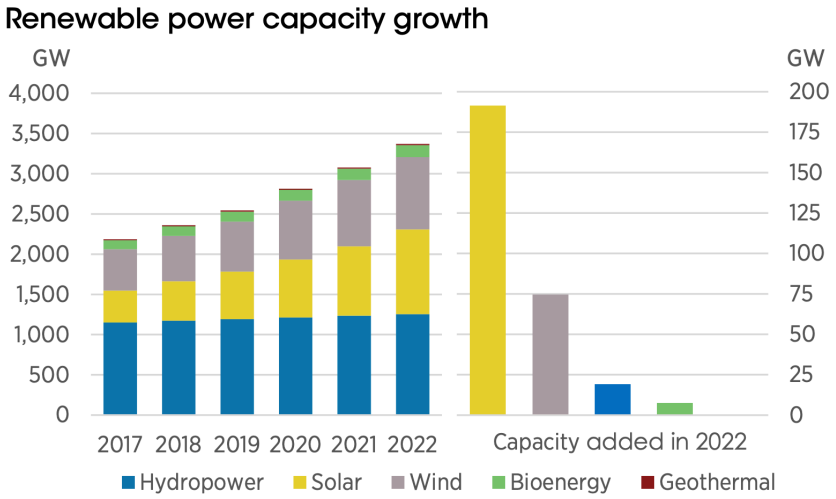


Figura 3.1: Evolución de la capacidad de generación de electricidad con energías renovables. Fuente: (International Renewable Energy Agency, IRENA, 2023).

la instalación de sistemas fotovoltaicos.

Disponer de herramientas que ayuden en la toma de decisiones para aumentar la producción energética de una zona (barrio o complejo de edificios) sería beneficioso para identificar los lugares más adecuados para tal intervención. Las predicciones sobre la producción energética esperada a largo plazo también son muy convenientes para evaluar la idoneidad de instalar estas infraestructuras de cara a poder conocer con antelación parámetros de rentabilidad económica. Lo mismo ocurre con las estimaciones a corto plazo de la producción fotovoltaica, puesto que pueden ayudar a los propietarios de instalaciones de autoconsumo a realizar una gestión más eficiente, y a las comercializadoras a mejorar la integración de este tipo de sistemas en la red eléctrica.

Aunque exista una tendencia al alza en la instalación de sistemas fotovoltaicos en las ciudades, aportar a las administraciones públicas herramientas que pudieran ayudar a a los planificadores urbanos a determinar las posibles áreas urbanas óptimas en las que realizar instalaciones fotovoltaicas, sería aún más beneficioso para la lucha contra el cambio climático y para guiar a las ciudades hacia modelos sostenibles, ya que podría favorecer y agilizar el proceso de sustitución de las energías contaminantes por energías renovables.

Para desarrollar herramientas que permitan localizar las posibles áreas óptimas urbanas en las que realizar instalaciones fotovoltaicas, es necesario que el sistema permita realizar estimaciones de la energía solar producida en diferentes áreas de interés. A continuación, se describen los principales modelos, tecnologías para la creación de estos sistemas, y las herramientas encontradas para realizar estimaciones de energía solar en las ciudades.

3.1.1.1. Segmentación semántica de los tejados

Entre los trabajos encontrados, se utilizan principalmente dos técnicas: aprendizaje supervisado con redes neuronales (Khoshboresh-Masouleh y otros, 2020; Pan y otros, 2020), y aprendizaje no supervisado con algoritmos de agrupamiento o *clustering* como k-means (Gavankar y Ghosh, 2019; El Joumani y otros, 2017).

Al emplear aprendizaje supervisado para entrenar modelos capaces de segmentar los tejados a partir de imágenes de las ciudades pueden surgir diversos problemas como el sobreentrenamiento o la dificultad para generalizar y extender el uso de dichos modelos a otras ciudades. Al entrenar los modelos con imágenes de las ciudades en las que se van a usar, son muy sensibles a las características urbanas de dichas ciudades, por lo que no funcionan bien para segmentar tejados en otras ciudades o países, ya que la tipología de tejados suele variar.

Para solventar el anterior problema, es posible utilizar la tecnología LiDAR, que consiste en una serie de imágenes en 3D de las ciudades, tomadas por drones o aviones que permiten obtener información sobre la altura y clasificación de cada elemento urbano (edificios, vegetación, etc.) (Awrangjeb y otros, 2013).

Una vez conocida la ubicación de los tejados, el siguiente paso consiste en la extracción de las características necesarias para realizar los cálculos de producción de energía fotovoltaica.

3.1.1.2. Extracción de las características de los tejados

Como se ha descrito anteriormente en los trabajos encontrados, para calcular las características de los tejados para poder realizar predicciones de energía fotovoltaica, es muy común utilizar procesamiento de imágenes (como hace, por ejemplo, Google Sun Roof), o aprendizaje automático para predecir las características como el área o la inclinación. El principal problema de estas técnicas puede ser un elevado tiempo de procesamiento, el sobreentrenamiento y la difícil extensión de la herramienta a otras ciudades.

A partir de los modelos LiDAR que ya ofrecen información de las alturas de los edificios puede generarse fácilmente un modelo de alturas normalizado, NDSM (Rao y otros, 2022)[1], que se obtiene eliminando la altura del terreno. En este modelo 2D, la altura de los objetos se mide al mismo nivel. Utilizando estas tecnologías, modelos, y transformaciones, el sistema puede determinar de forma automática las cubiertas disponibles y las alturas de los diferentes edificios detectados. A partir del NDSM, el cálculo de la inclinación,

[1]NDSM: Normalized Digital Surface Model

orientación, y área son tareas sencillas.

3.1.1.3. Predicción de energía solar en ciudades

En el trabajo realizado por Freitas y otros (2015) se llevó a cabo una revisión de la literatura de los modelos disponibles para realizar estimaciones de energía solar en las ciudades mediante el análisis de diferentes algoritmos. A esas herramientas se les han ido incorporando otras más modernas y, en la actualidad, las principales herramientas encontradas para predecir la energía solar generada en los tejados son las siguientes:

- *Deep photovoltaic nowcasting* (Zhang y otros, 2018). Herramienta que realiza predicciones de la energía producida en un sistema fotovoltaico a muy corto plazo (un minuto). La tecnología utilizada se basa en redes neuronales artificiales entrenadas con imágenes aéreas del cielo y valores energéticos asociados. El sistema funciona con los paneles fotovoltaicos y cámaras previamente instaladas que toman fotos del cielo. El aprendizaje se basa en la energía obtenida y las fotos del estado del cielo, por lo que no ayuda a establecer la ubicación óptima de las instalaciones, como es el objetivo de nuestro sistema.
- *DeepRoof* (Lee y otros, 2019). Esta herramienta ha sido desarrollada en la Universidad de Massachusetts y utiliza datos de seis ciudades diferentes de Estados Unidos, centrándose en una ciudad en Framingham (Massachusetts). Utiliza imágenes satelitales y datos inmobiliarios para estimar el tamaño y la geometría (orientación e inclinación) de los tejados. Estas imágenes deben ser etiquetadas por los expertos para crear el conjunto de entrenamiento. El sistema utiliza un enfoque de aprendizaje profundo para estimar el potencial solar del tejado. Los edificios se seleccionan individualmente a través de una interfaz web o se pueden enumerar para el procesamiento por lotes, pero no se pueden identificar automáticamente en una región de interés. El principal problema es que el acceso público actualmente no está disponible.
- *Google Sun Roof* (Google, 2021, May 7). Herramienta desarrollada por Google que permite estimar a largo plazo la producción solar media diaria que se conseguiría en la cubierta de un edificio concreto seleccionado manualmente en el mapa de una ciudad. Actualmente está disponible para ciudades de los Estados Unidos de América. El sistema utiliza imágenes de satélite de Google para calcular las características de los tejados e información meteorológica para realizar predicciones energéticas.
- *Huella Solar* (HuellaSolar, 2021, May 7). Permite seleccionar zonas de una ciudad y realiza la estimación energética mensual y anual. El cálculo se realiza para toda

el área en función de la radiación. Por lo tanto, la segmentación de edificios no es automática y debe realizarse manualmente si solo se desean tejados. Un punto positivo es que considera el sombreado y proporciona una herramienta para estimar la energía producida por las fachadas. Las ciudades que se ofrecen están limitadas en la versión gratuita y se pueden incluir nuevas en la versión registrada, pero se requiere un conocimiento avanzado sobre mapas e imágenes.

Tal y como se ha descrito, los tejados se han convertido en uno de los mejores lugares en los que realizar instalaciones fotovoltaicas en las ciudades. Para poder llevar a cabo, las tareas de predicción de energía solar en los tejados, el primer paso consiste en detectar de forma automática los tejados urbanos. Para realizar esta tarea, se llevó a cabo un estudio para revisar las últimas novedades y tecnologías que permitieran realizar una segmentación semántica de los tejados[2] automática de los tejados a partir de imágenes de ciudades.

3.1.2. Consumo eléctrico en hogares

Los datos de consumo de electricidad son un ejemplo real donde los medidores inteligentes registran grandes cantidades de datos, y algunas propuestas diferentes los transforman en información útil (Rajabi y otros, 2020). A partir de estos datos es posible obtener los patrones de consumo eléctrico de los consumidores domésticos desde una perspectiva global. Descubrir patrones a partir de los datos de consumo eléctrico es una tarea que puede agilizarse y automatizarse con un proceso de minería de datos que permitirá inducir modelos de suministro, usos energéticos, sostenibilidad y competitividad.

La caracterización de los perfiles de consumo eléctrico de los hogares permitirá a las empresas comercializadoras de energía eléctrica ofrecer mejores tarifas y servicios personalizados a sus clientes, e incluso hacerles recomendaciones para que cambien su consumo a momentos en los que los precios de la electricidad son más bajos. Otro de sus beneficios es que permite una mejor y más eficiente gestión de la red eléctrica. Por todo lo anterior, podemos concluir que disponer de modelos que permitan caracterizar los perfiles de consumo eléctrico de los hogares (TLP) [3], supone un avance en la lucha contra el cambio climático, y en la mejora de la sostenibilidad urbana.

El uso de métodos de aprendizaje automático para descubrir perfiles de consumidores es una realidad desde hace algunos años. Desde principios de siglo, ha habido trabajos académicos que han propuesto el uso de mapas autoorganizados de Kohonen (SOM) para encontrar clústeres que representen los perfiles de consumo típicos de los clientes (Figueiredo y otros, 2003; Verdu y otros, 2004).

[2]Consiste en asignar una etiqueta o categoría a cada píxel de una imagen
[3]TLP: Typical Load Profiles

El uso creciente de medidores inteligentes está acelerando la aparición de grandes conjuntos de datos que registran el consumo de energía por hora (millones de registros), pero los requisitos computacionales de los algoritmos de agrupamiento limitan la cantidad de registros y variables que se pueden procesar usando cantidades estándar de tiempo y memoria.

Propuestas más recientes, como la presentada por Räsänen y otros (2010), continúan utilizando el mismo enfoque, pero considerando miles de consumidores en lugar de los cientos utilizados en trabajos anteriores. En ese trabajo utilizaron datos de consumo de electricidad horaria durante un año de 3.989 pequeños clientes ubicados en Northern-Savo, Finlandia. Los autores agregaron otros métodos de agrupamiento (como k-means y agrupamiento jerárquico) a los mapas autoorganizados (SOM), pero incluso con mejoras en hardware, las limitaciones computacionales requerían una reducción del conjunto de datos. Así, lo primero que hicieron fue procesar los datos y hacer una selección de los mismos. Para decidir el número de clústeres y así poder configurar el algoritmo k-means, utilizaron el índice de Davies-Boulding y el promedio dentro de la varianza. Como conclusión de su trabajo propusieron 19 perfiles de consumidores diferentes.

En otro trabajo, se proponen varias clases de perfil de carga (PC) para caracterizar el consumo de electricidad de los consumidores de los hogares (McLoughlin y otros, 2015). Se utilizan métodos de aprendizaje no supervisados (k-means, k-medoids y SOM) para caracterizar patrones diurnos, intradiarios y estacionales de uso de electricidad. Los datos se tomaron de un conjunto de datos en el que participaron 5.000 hogares y empresas irlandesas (Archivo Irlandés de Datos de Ciencias Sociales – ISDDA – (Commission for Energy Regulation (CER), 2012)). En concreto, utilizaron datos de un periodo de seis meses para un total de 3.941 consumidores. También emplearon el índice de Davies-Bouldin para identificar el número más apropiado de clústeres, y encontraron 10 patrones diferentes de consumo de electricidad (PC).

En la revisión de Cembranel y otros (2019), se presenta un resumen de los métodos utilizados para agrupar perfiles de consumo eléctrico. Concluyen que el algoritmo k-means presenta los mejores resultados, aunque señalaron el problema de decidir el número correcto de clústeres. Para evitar este problema, los autores sugieren utilizar algoritmos automáticos, como G-means (Hamerly y Elkan, 2003) o X-means (Pelleg y Moore, 2000). A pesar de sus ventajas, estos algoritmos presentan otras limitaciones, por ejemplo, G-means se ha utilizado en el ámbito de la caracterización del consumo de electricidad (Mets y otros, 2016) pero los datos originales tenían que aproximarse mediante una representación de tipo *wavelet*. La reducción de la dimensionalidad es necesaria porque G-means requiere datos no tan altamente dimensionales. En cuanto a X-means, es lento para grandes cantidades de datos.

Rajabi y otros (2020) compara varias técnicas de agrupamiento (k-means, fuzzy C-means, agrupamiento jerárquico y SOM) utilizadas en el contexto de los patrones de consumo eléctrico. Los índices para evaluar la calidad de los modelos de agrupamiento (CVI) utilizados son el error cuadrático medio, Silhouette, Davies-Bouldin, y Dunn. La principal fuente de datos es nuevamente el conjunto de datos irlandés (ISDDA) donde se aplican algunos filtros y transformaciones para reducir la dimensionalidad. Después de comparar el rendimiento de los algoritmos de agrupamiento aplicados mediante los CVI anteriormente descritos, proponen un número diferente de clústeres según el tamaño y la naturaleza de los conjuntos de datos construidos. En el caso de usar cientos de curvas de consumo eléctrico diarias (en particular, 356 curvas de un cliente residencial), concluyen que el número óptimo de clústeres se encuentra en el rango de 8 a 10. Por otro lado, al utilizar miles de clientes proponen 32 clústeres (16 para días laborables y 16 para fines de semana).

Toussaint y Moodley (2020) enfatizan que la selección de un conjunto útil de clústeres para identificar los patrones de consumo de electricidad típicos en los hogares requiere una amplia experimentación y conocimiento del dominio. En este experimento, se han utilizado los CVI de la Silhouette y Davies Bouldin para determinar el número óptimo de clústeres. El conjunto de datos que han utilizado se recopiló durante 20 años en Sudáfrica y acumula más de 3 millones de consumos eléctrico diarios. Identifican un total de 59 perfiles de consumo. Además de su propuesta, también incluyen un análisis de los algoritmos de agrupamiento más utilizados para esta tarea, concluyendo que k-means es el algoritmo más utilizado y el que mejores resultados obtiene.

En algunos trabajos como en el de Bourdeau y otros (2021), se usan diferentes distancias en el proceso de agrupamiento: además de la distancia euclídea clásica, se incluye la distancia Dynamic Time Warping (DTW).

Algunas características pueden destacarse resumiendo los trabajos relacionados antes mencionados:

- La dimensionalidad del conjunto de datos, ya sea por el número de observaciones o por el número de variables que las definen, es una de las principales limitaciones cuando se ejecutan algoritmos de agrupamiento para inducir modelos. Es relativamente fácil aprender a partir de miles de perfiles de consumo eléctrico diarios, pero la tarea se vuelve mucho más complicada cuando hay millones involucrados. Es habitual muestrear el conjunto de datos para utilizar aproximadamente 100.000 observaciones de consumo diario como máximo. Considerar 24 variables, una por hora en un día, también es muy común.
- Existen muchas configuraciones para realizar la tarea de agrupamiento (algoritmos, distancias, CVIs). K-means parece ser el algoritmo más utilizado, quizás por su baja

complejidad (Xu y Tian, 2015), y el que logra mejores resultados en este contexto.

- Determinar un único número óptimo de clústeres (k) es una tarea que se aborda desde dos perspectivas diferentes: considerando múltiples valores para k y obteniendo el mejor valor *a posteriori* (comprobando las medidas de CVI). Los tamaños propuestos en diferentes trabajos de investigación son muy heterogéneos: desde unos pocos clústeres (menos de 10) a muchos de ellos (centenas) pasando por algunas decenas. Uno de los principales factores para evitar o reducir esta variabilidad puede pasar por la inclusión del conocimiento de los expertos.

3.2. Selección automática de emplazamientos urbanos para instalaciones fotovoltaicas

Se presenta en esta sección la metodología que se ha propuesto en este trabajo de investigación para la organización de las fuentes de datos, el tratamiento de estos y la obtención de los modelos de predicción a corto y largo plazo de la energía generada por sistemas fotovoltaicos. Además, se describe la implementación de estos modelos y los resultados obtenidos.

3.2.1. Propuesta metodológica

La Figura 3.2 muestra un esquema del proceso de minería de datos utilizado para el desarrollo de una herramienta inteligente de ayuda a la determinación de los posibles emplazamientos urbanos adecuados para realizar instalaciones fotovoltaicas. A continuación se describen las fases involucradas en dicho esquema.

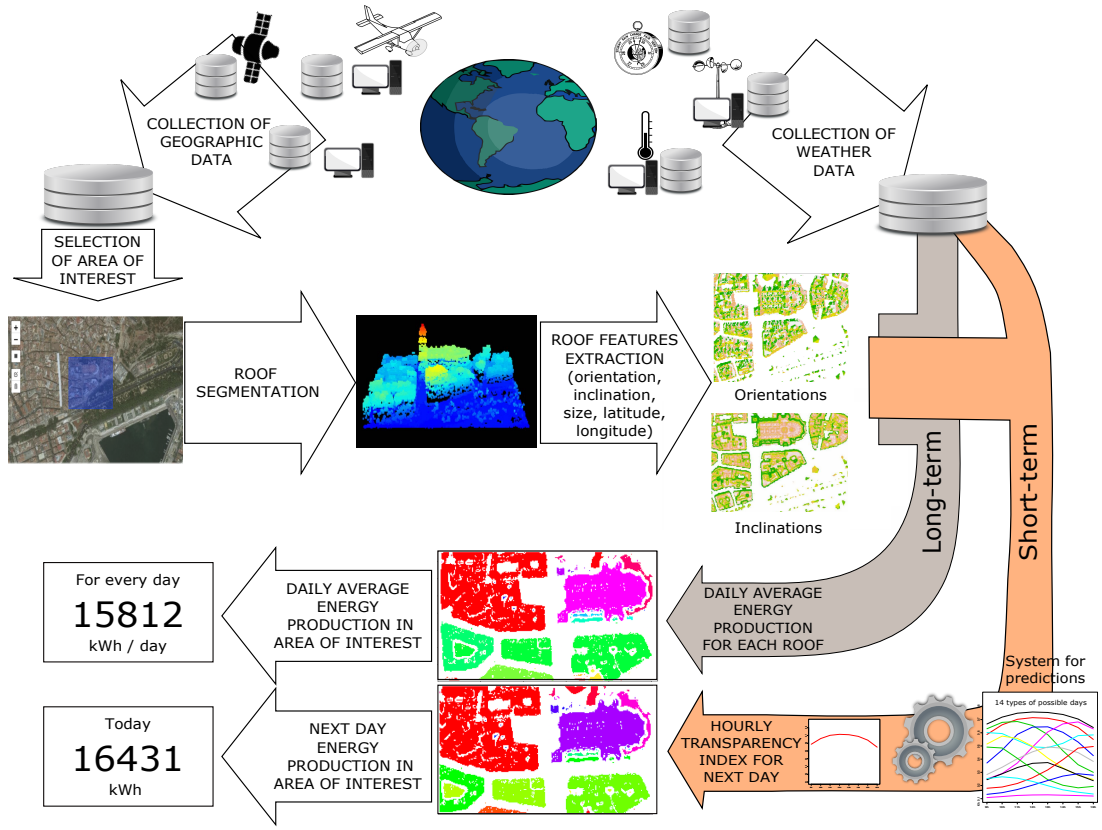


Figura 3.2: Metodología para la selección de emplazamientos para instalaciones fotovoltaicas. Fuente: (Rodríguez-Gómez y otros, 2022a).

La comprensión del problema (Fase 1) estableció el objetivo de esta herramienta: evaluar la producción potencial de energía que pueden generar las instalaciones fotovoltaicas instaladas en áreas urbanas a largo y corto plazo.

Durante la fase de comprensión de datos (Fase 2), se ha trabajado tanto con datos meteorológicos como con imágenes aéreas LiDAR del terreno urbano cubriendo toda el área geográfica de interés para el estudio. Los datos meteorológicos que utiliza el sistema son: observaciones convencionales (temperatura, humedad, velocidad del viento, ...), datos de radiación, y datos con parámetros predichos para el día siguiente.

Las imágenes LiDAR suelen estar disponibles para la mayoría de las ciudades y facilitan la segmentación semántica de los objetos urbanos, lo que posibilita la detección automática de los tejados. En dicha segmentación, cada punto de la imagen se asigna automáticamente a una clase específica (como edificio o vegetación). Las imágenes LiDAR también proporcionan modelos de las alturas de los objetos urbanos, lo que permite obtener las orientaciones e inclinaciones de los tejados en las áreas de interés.

La integración de datos y su preparación (Fase 3) consta principalmente de dos pasos que se han automatizado en gran medida: la selección del área de interés (ver la Subsección

3.2.1.1) y la segmentación de los tejados para la extracción de características (ver la Subsección 3.2.1.2).

En la fase de modelado (Fase 4), se necesitan dos modelos diferentes para satisfacer los objetivos de minería de datos establecidos en la primera fase: un modelo para predicciones a largo plazo y otro para predicciones a corto plazo. Aunque el segundo modelo ya estaba definido (del Campo-Ávila y otros, 2021), ha sido necesario un proceso de integración. Ambos modelos se describen en la Subsección 3.2.1.3.

La evaluación (Fase 5) se realizó contrastando los resultados obtenidos por nuestro sistema para puntos específicos en diferentes áreas con los resultados obtenidos por otros sistemas. Estos otros sistemas utilizan procesos manuales para calcular la producción fotovoltaica de un sistema específico. Una correcta evaluación permite integrar los modelos generados en el paso anterior (Fase 4) en el producto final (Fase 6).

En cuanto al despliegue (Fase 6), la herramienta puede estimar la energía solar producida en áreas de interés dentro de zonas urbanas a largo y corto plazo. Este es el resultado de integrar un modelo de estimación de energía fotovoltaica (largo plazo) con un sistema predictor horario de radiación solar existente (corto plazo).

3.2.1.1. Selección del área de interés

El sistema accede a la información geográfica a partir de imágenes LiDAR previamente descargadas, y crea un mapa con una cuadrícula que delimita las áreas disponibles para la estimación de energía. Permite al usuario seleccionar el área de interés definiendo un polígono en el mapa. A continuación, el sistema carga la imagen que cubre el área delimitada por las coordenadas del usuario a través del polígono en el mapa. Los pasos de procesamiento posteriores solo tienen en cuenta los datos filtrados en este paso.

Además de la información geográfica, se necesitan datos meteorológicos capturados en la ciudad o cerca de la ciudad. Sin embargo, la proximidad a la zona de interés es suficiente y se utiliza la información de las estaciones meteorológicas más cercanas.

3.2.1.2. Procesamiento de imágenes para segmentación de tejados y extracción de características

En la localización de cubiertas donde se pueden realizar instalaciones fotovoltaicas intervienen dos aspectos: en primer lugar, la detección de los lugares donde hay cubiertas en un área y, en segundo lugar, la determinación de sus características para generar una valoración precisa.

El primer paso se consigue a partir del procesamiento de la imagen LiDAR 3D que cubre

el área de interés de la ciudad seleccionada por el usuario. Se eliminan todos los elementos urbanos como la vegetación, manteniéndose únicamente el suelo y los tejados.

En el siguiente paso, se obtiene el modelo de altura normalizado (NDSM) eliminando la altura del terreno. En este modelo 2D, la altura de los objetos se mide al mismo nivel. Tras esta transformación, el sistema puede determinar los tejados disponibles y las alturas de los diferentes edificios detectados.

En este punto, el sistema permite al usuario seleccionar diferentes rangos de interés para diferentes valores de pendiente y orientación de los tejados.

A continuación se genera una capa ráster con una serie de componentes conectados obtenidos a partir de los tejados. Cada componente conexa consta de píxeles contiguos de tejados que cumplen los criterios del usuario en cuanto a pendientes y orientaciones. Para los cálculos de energía, cuando se habla de un tejado, en realidad se refiere a la componente conexa de algún tejado concreto.

3.2.1.3. Predicciones de energía solar

La estimación de la energía recibida en cada tejado depende del tamaño, orientación, inclinación y ubicación de éste y se puede modelar utilizando las expresiones propuestas por Iqbal (1983) y Coronas y Villarrubia (1983). En esta fase se ha utilizado la inclinación real de las cubiertas. También sería posible estimar la inclinación óptima, por ejemplo utilizando la propuesta realizada por Chang (2010), pero esto requeriría la evaluación individual de la integración de las instalaciones fotovoltaicas (PV) en la envolvente de los edificios.

La energía que podría producir un sistema fotovoltaico a largo plazo se estima utilizando el modelo propuesto por Osterwald (1986), teniendo en cuenta la información del paso anterior, según se detalla en la Subsección 2.2.1. Para esta estimación se calcula, en primer lugar, la energía horaria producida por cada sistema usando datos meteorológicos y datos del tejado. Los datos meteorológicos utilizados son la radiación solar horaria global y la temperatura horaria. Los datos del tejado son el tamaño, latitud, longitud, inclinación y orientación. La potencia generada por el sistema se estima mediante la expresión que se detalló en la Subsección 2.2.1.

Por lo tanto, la producción de energía promedio diaria se puede estimar para cada tejado, y la estimación para los tejados en un área deseada, se puede calcular por agregación de la estimación para cada tejado.

Por otro lado, en un escenario de corto plazo, la energía fotovoltaica producida se puede estimar con un día de antelación. Utilizando el modelo propuesto por del Campo-Ávila

y otros (2021) se calcula la predicción de la radiación solar horaria al día siguiente. El error absoluto medio (MAE) de este modelo es de 63 y 97 Wh/m^2, según se use o no la predicción del índice de transparencia, respectivamente. Esto lo hace similar en cuanto al error a otros modelos de minería de datos como puede ser el del trabajo de Cannizzaro y otros (2021) donde son esos errores son de 58 y 107 Wh/m^2, respectivamente. Una ventaja de este modelo es que utiliza como variables independientes la mayoría de las variables de entrada significativas seleccionadas en la propuesta de Castangia y otros (2021). Es decir, toma como entrada los datos meteorológicos registrados durante el día en curso y determinadas previsiones meteorológicas, y predice el tipo de radiación esperada para el día siguiente (soleado, nublado, parcialmente soleado, etc.). El modelo selecciona entre 14 tipos de patrones de radiación, que se identificaron en la fase de inducción del modelo, y proporciona el índice de transparencia horaria esperado para cada hora entre las 9:00 y las 16:00 para el día siguiente.

El proceso de predicción se resume en el lado derecho de la Figura 3.2. Como se explica en la propuesta de del Campo-Ávila y otros (2021), utilizando los datos meteorológicos disponibles para un día, el sistema desarrollado responde con el tipo de día estimado para el día siguiente. La radiación solar global por hora se estima utilizando el *centroide* que contiene los valores por hora del índice de claridad para ese día y los valores de radiación extraterrestre por hora. Esta estimación convenientemente transformada permite obtener la energía solar potencial calculada para el día siguiente para cada tejado y, por agregación, para el área de interés.

3.2.2. Integración y operación

Además del procesamiento de imágenes y la extracción de las características de las cubiertas, para el correcto funcionamiento del sistema se necesitan modelos de cálculo de energía fotovoltaica y el sistema predictor presentados en la sección previa. También son necesarios una serie de scripts para automatizar las descargas diarias de la AEMET con respecto a datos de radiación, observaciones convencionales y predicciones.

Aunque el procesamiento de imágenes se puede aplicar a todos los elementos, el usuario puede seleccionar las características de interés (como una orientación específica o la inclinación máxima o mínima de tejado para ser considerada) para reducir el esfuerzo computacional. La energía estimada se calcula para cada componente conexa que cumpla con los requisitos del usuario (promedio diario para la predicción a largo plazo o energía horaria para la predicción a corto plazo). Su cálculo agregado constituirá la energía fotovoltaica estimada en el área urbana de interés.

El esquema general de funcionamiento de la herramienta con la integración se muestra en

la Figura 3.3.

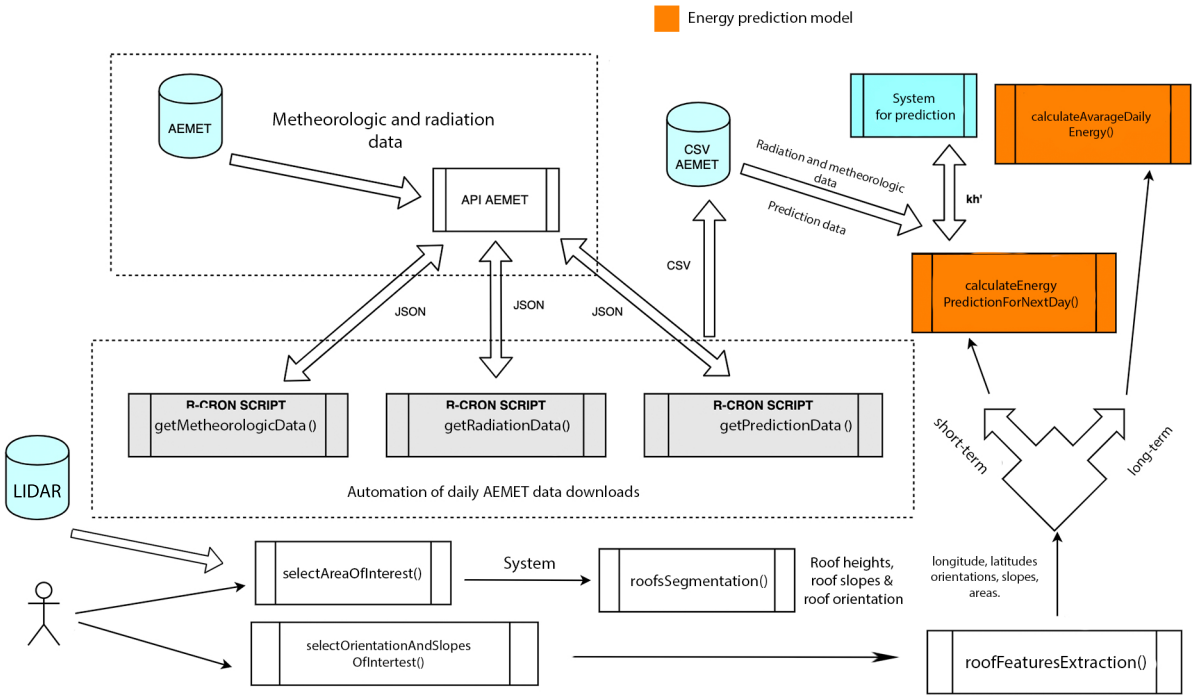


Figura 3.3: Esquema general de la herramienta con sus integraciones. Fuente: (Rodríguez-Gómez y otros, 2022a).

A continuación se describe el funcionamiento general del sistema desarrollado:

1. Selección del área de interés

   a) El sistema realiza la segmentación de los tejados.

2. Selección de la orientación e inclinación de interés para filtrar los tejados

   a) El sistema extrae las características de las componentes conexas (área, latitud media, longitud media, orientación media, e inclinación media).

3. Selección entre cálculo de energía solar a corto o largo plazo:

   a) Largo plazo:

      1) Predecir la energía solar promedio diaria para cada tejado.
      2) Predecir la energía solar promedio diaria total en el área.

   b) A corto plazo:

1) El sistema, mediante unos scripts que se ejecutan diariamente, obtiene los datos necesarios de la AEMET utilizando su API. En concreto, se descargan observaciones meteorológicas, de radiación y predicciones.
2) Se calcula la radiación esperada para el día siguiente usando un sistema de predicción que analiza datos previos. Para cada tejado se realiza una estimación de energía fotovoltaica horaria a corto plazo.
3) Se calcula la energía solar horaria que se espera que se produzca al día siguiente para el área de interés.

3.2.3. Tecnologías para la implementación del sistema desarrollado

Algunas de las herramientas que se usan más comúnmente para analizar y procesar imágenes urbanas son los Sistema de Información Geográfica (GIS) (Li y otros, 2019). Entre las herramientas más utilizadas para este propósito hay que destacar QGIS y ArcGIS. El procesamiento de imágenes de este tipo con lenguajes de ciencia de datos como `R` (R Core Team, 2019) ha cobrado impulso recientemente gracias al desarrollo de librerías como `lidR`, `rspatial` o `raster`. Estos incluyen una amplia gama de funciones que facilitan enormemente el trabajo con imágenes urbanas en 2D y 3D para obtener la información necesaria mediante su procesamiento.

Las principales tecnologías utilizadas para desarrollar la herramienta están basadas en el lenguaje `R`. Se pueden agrupar según la fase del proceso de minería de datos en la que se hayan utilizado.

En las primeras fases, como la adquisición y transformación de datos, se utilizó el paquete `dplyr` (Wickham y otros, 2020) para obtener conocimiento básico. El paquete `cron` (Wijffels, 2020) se ha utilizado para automatizar la descarga diaria de datos meteorológicos y de radiación.

Para la preparación de datos se usaron otros paquetes de R como `rspatial` (Hijmans, 2018), `raster` (Hijmans, 2020) y `lidR` (Roussel y otros, 2020). `lidR` procesa modelos 3D LiDAR y clasifica los objetos urbanos en él (edificios, suelo o vegetación) mientras que `raster` permite el cálculo de altura, orientación o inclinación.

Para automatizar las partes del proceso en las que se inducen modelos predictivos se han utilizado bibliotecas R para Machine Learning como el paquete `RWeka` (Hornik y otros, 2009), que integra algoritmos implementados en Weka (Witten y otros, 2017), o `caret` (Kuhn, 2020).

Finalmente, el paquete `shiny` (Chang y otros, 2021) ofrece un framework para el desarrollo

del DashBoard o aplicación web para el cálculo de energía fotovoltaica en áreas urbanas de interés. Se ha configurado un servidor Linux para el despliegue de la aplicación.

3.2.4. Resultados

En este apartado, se muestra un ejemplo de caso de uso real de la herramienta desarrollada para estimar energía fotovoltaica en una zona de interés urbano. La estimación se realizó para largo plazo (producción media diaria). La estimación a corto plazo seguiría un proceso similar.

Las comparaciones con métodos y herramientas anteriores también se enumeran al final de esta sección. Allí se destacan las ventajas presentadas en el sistema propuesto al que se ha denominado URSUS-PV [4], por haber sido desarrollado en el marco de un proyecto de investigación cuyo acrónimo es URSUS.

3.2.4.1. Ejemplo de validación

Una vez implementado URSUS-PV, se utilizó en un escenario real para probar sus capacidades. Se ha seleccionado la ciudad de Málaga, en España, y se recogieron datos meteorológicos de 10 años (de la Agencia Española de Meteorología, AEMET[5]), e imágenes aéreas LiDAR del terreno urbano cubriendo toda el área geográfica de interés para el estudio (del Centro Nacional de Información Geográfica, CNIG[6]).

El uso de la aplicación se describe a continuación.

Selección del área de interés y segmentación de tejados El primer paso consiste en seleccionar la zona urbana de interés para realizar las estimaciones de energía fotovoltaica. La herramienta muestra una interfaz que permite trazar polígonos sobre el mapa de la ciudad de estudio para definir el área de interés de entrada. El mapa muestra las imágenes LiDAR disponibles para esa ciudad como una cuadrícula superpuesta de color azul. El usuario puede ampliar y desplazar el mapa de la ciudad para ubicar el área de interés (AOI). A continuación, el sistema procesa la imagen LiDAR 3D del área seleccionada, filtra los objetos urbanos correspondientes a suelo y tejados. La librería `lidR` incluye algunas funciones que permiten la extracción de diferente información relacionada con imágenes LiDAR. Los tejados y el suelo se obtienen utilizando la función `lasfilter` que permite filtrar entre los elementos urbanos de la imagen LiDAR del área de interés. La

[4]Urban sustainability for photovoltaic energy prediction
[5]`http://www.aemet.es`
[6]`http://centrodedescargas.cnig.es`

función `lasNormalize` se utiliza para eliminar el suelo del modelo de alturas obteniendo así el modelo de alturas normalizado. Para obtener el modelo de alturas 2D normalizado se utiliza `gridCanopy`. Finalmente, el sistema muestra los tejados del área de interés y la altura de cada píxel de cada tejado. Las Figuras 3.4a, 3.4b y 3.4c muestran información calculada en diferentes momentos durante la selección del AOI y su segmentación.

(a) (b) (c)

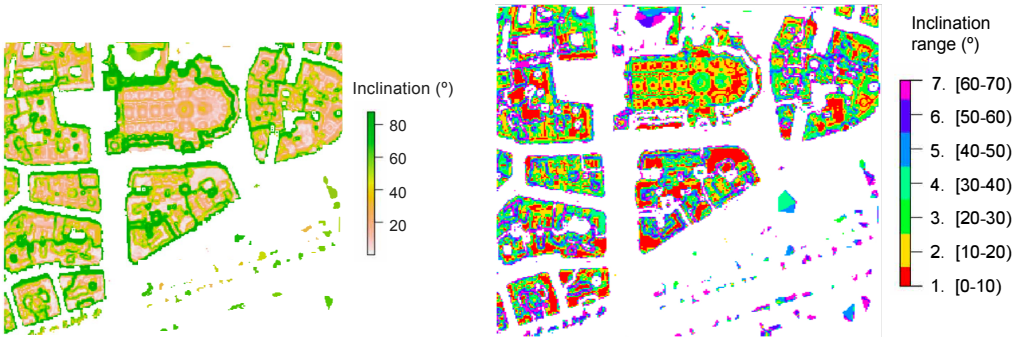
Figura 3.4: Segmentación de tejados: (a) selección del área urbana de interés, (b) modelo LiDAR 3D con segmentación de tejados y suelo, (c) Modelo de alturas normalizado de los tejados. Fuente: (Rodríguez-Gómez y otros, 2022a).

Filtrando tejados en base a la orientación e inclinación El usuario podría seleccionar la orientación y la inclinación (o pendiente) máximas de los tejados potenciales donde instalar sistemas fotovoltaicos.

El sistema permite al usuario seleccionar rangos de interés para pendientes y orientaciones agrupadas de 10 en 10 grados. Una vez seleccionadas, el sistema mostrará dos capas ráster; una con el valor de inclinación de cada píxel de cada tejado, y otro con el rango de pendientes de interés al que pertenece cada píxel.

La Figura 3.5 muestra la interfaz de usuario para los rangos de pendiente seleccionados. En este ejemplo, el usuario está interesado en todos los rangos de pendiente. 1.[0,10), 2.[10,20), ... 7.[60-70]. La Figura 3.5a muestra los valores de pendiente para cada píxel. La Figura 3.5b muestra los valores discretizados correspondientes a los rangos de inclinación en los que el usuario está interesado. El proceso para las orientaciones es el mismo.

La función `terrain` en la biblioteca `raster` ha sido utilizada para calcular las orientaciones y pendientes de los tejados de un modelo ráster 2D normalizado a partir de las alturas.



(a) (b)

Figura 3.5: Selección de inclinación: (a) Valores de inclinación para cada píxel, (b) Rango de inclinación seleccionado por el usuario para cada píxel. Fuente: (Rodríguez-Gómez y otros, 2022a).

Predicciones de energía fotovoltaica En este punto, el usuario, dependiendo de sus objetivos, puede seleccionar si realizar estimaciones a corto o largo plazo. En el ejemplo que se detalla, presentado en la Figura 3.6, el sistema realizó una estimación a largo plazo.

El primer paso fue calcular las componentes conexas de los tejados. Las componentes conexas (CC) de los tejados son trozos de tejados compuestos por píxeles contiguos que satisfacen los criterios de orientación e inclinación del usuario. Cada CC tiene una identificación única. El sistema genera una capa ráster de componentes conexas para cada trozo de tejado susceptible de albergar instalaciones fotovoltaicas. La función de la librería `raster` para generar la capa ráster de componentes conexas de los tejados es `clusp(formask, direcciones=8)`, donde `formask` es una capa ráster con valor 1 para los píxeles que satisfacen los criterios de orientación y pendiente, y direcciones=8 implica formar componentes conexas para los 8 vecinos con píxeles con valor 1.

Para cada componente conexa (parte de un tejado que satisface los criterios de orientación e inclinación), el sistema calcula la latitud media, la orientación media, la inclinación media y el tamaño (área en m^2). Con las características extraídas para cada componente conexa disponible, se añaden los datos meteorológicos y de radiación necesarios del municipio a partir de la estación meteorológica más cercana. Con los datos anteriores, y a partir de las las características de los sistemas fotovoltaicos susceptibles de ser instalados, el sistema predice la energía promedio diaria (kWh) para cada componente conexa. Finalmente, la suma de la energía diaria estimada para cada componente conexa, es la energía diaria estimada para todos los días en la zona urbana de interés analizada.

En el panel de estimación de energía (ver Figura 3.6), la herramienta muestra un mapa con las componentes que cumplen con los requerimientos de orientación e inclinación del usuario. Adicionalmente, dos etiquetas muestran el número de techos que cumplen con los

requisitos de las instalaciones y la estimación de energía fotovoltaica para el área urbana seleccionada inicialmente. El sistema también muestra información detallada sobre las componentes conexas procesadas (energía estimada, área, latitud, longitud, orientación e inclinación).

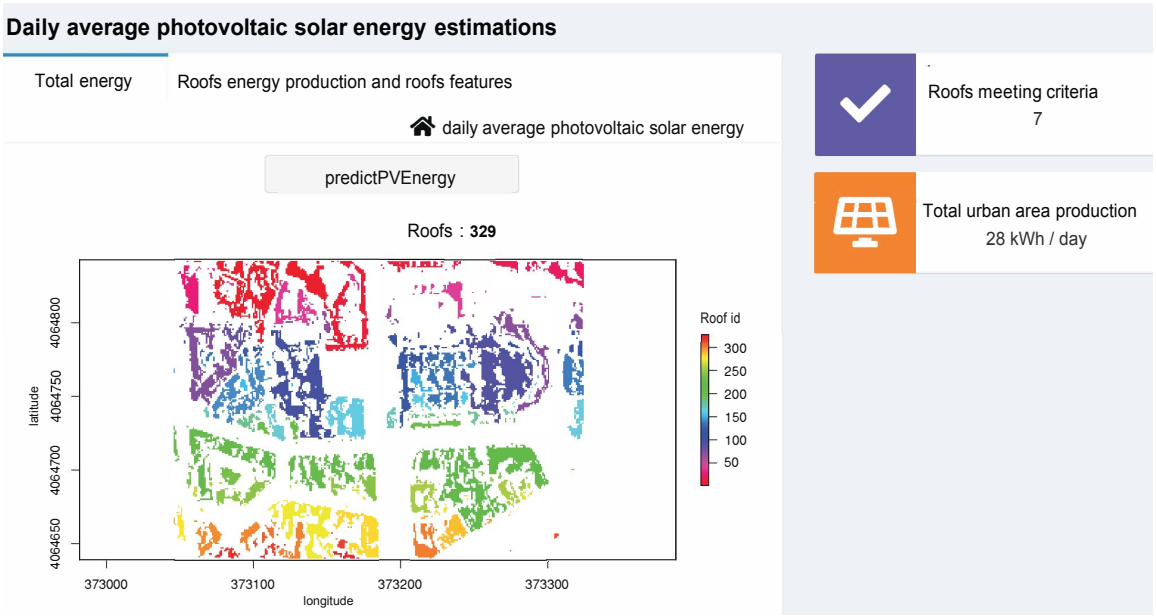


Figura 3.6: Panel de estimación de energía que muestra la cantidad de tejados que cumplen con los requisitos del usuario y los requisitos tecnológicos de los sistemas fotovoltaicos, así como la estimación diaria de energía solar fotovoltaica promedio. Fuente: (Rodríguez-Gómez y otros, 2022a).

3.2.4.2. Comparación con los modelos previos

Las aportaciones que ofrece la herramienta desarrollada en este trabajo de investigación con respecto a otras (como aquellas descritas anteriormente en la Subsección 3.1.1.3) están resumidas en la Tabla 3.1 y descritas a continuación:

- URSUS-PV utiliza imágenes LiDAR que permiten determinar con precisión los elementos urbanos de una zona. Esas imágenes son fáciles de descargar para muchos países y la disponibilidad está aumentando. Los datos meteorológicos necesarios en el sistema también son comunes (como la temperatura o la humedad) y suelen ser registrados por organismos nacionales. Por lo tanto, URSUS-PV es muy flexible en la incorporación de nuevas áreas urbanas. El proceso para incluir nuevas ciudades es fácil (y está documentado en la distribución del software).
- La herramienta es de código abierto y gratuita. Puede ser utilizada por cualquier persona: administraciones públicas, cooperativas, empresas distribuidoras de siste-

mas fotovoltaicos o particulares. El sistema se puede personalizar fácilmente para cualquier conjunto de ciudades.

- URSUS-PV realiza automáticamente una segmentación semántica de los tejados en el área de interés urbana permitiendo al usuario quedarse solo con las partes de los tejados del área urbana que satisfacen las necesidades de inclinación y orientación de los usuarios potenciales de las instalaciones fotovoltaicas.

- La herramienta permite tanto estimaciones a corto como a largo plazo para cualquier área urbana de interés (calles, barrios, urbanizaciones, ...)

	Estimación (largo-plazo) o Predicción (corto-plazo)	Fuente de datos actualizable	Dominio	Acceso
Deep photovoltaic nowcasting	corto-plazo (cinco minutos de antelación)	no disponible	instalaciones en Japón	no disponible
DeepRoof	largo-plazo (horas máximas de sol diarias durante el año)	no disponible	pocas ciudades en EE.UU	web (no mencionado en el artículo)
Google Sun Roof	largo-plazo (consumo medio diario anual)	no disponible	muchas ciudades en EE.UU / algunas localizaciones	web
PVWatts Calculator	largo-plazo (media diaria para los meses del año)	no necesario	mundial	web
Huella Solar	largo-plazo (media diaria para los meses del año)	sí, pero no es simple, ni intuitivo	pocas ciudades en España	web (requiere registro)
URSUS-PV	corto-plazo (predicciones horarias a un día vista) largo-plazo(consumo medio diario anual)	sí, sólo necesita nube de puntos LiDAR y datos meteorológicos	sistema preparado para algunas ciudades (por ahora de España)	web

Tabla 3.1: Comparativa de herramientas de predicción de energía solar urbana

3.3. Modelizado de perfiles de consumo eléctrico doméstico

En esta sección se describe la metodología propuesta para el modelizado de perfiles de consumo eléctrico doméstico. En el proceso de modelizado ha sido fundamental contar con información del dominio de aplicación que ha sido suministrada por expertos.

Se han propuesto dos algoritmos que permiten integrar esta información en la selección del número óptimo de tipos de días de consumo. Estos algoritmos pueden ser utilizados en dominios en los que sea necesario determinar el número de clústeres de una agrupación utilizando conocimiento experto.

3.3.1. Propuesta metodológica

El esquema de la Figura 3.7 resume la metodología de minería de datos propuesta para la caracterización de los perfiles de consumo eléctrico de los hogares de cualquier ciudad.

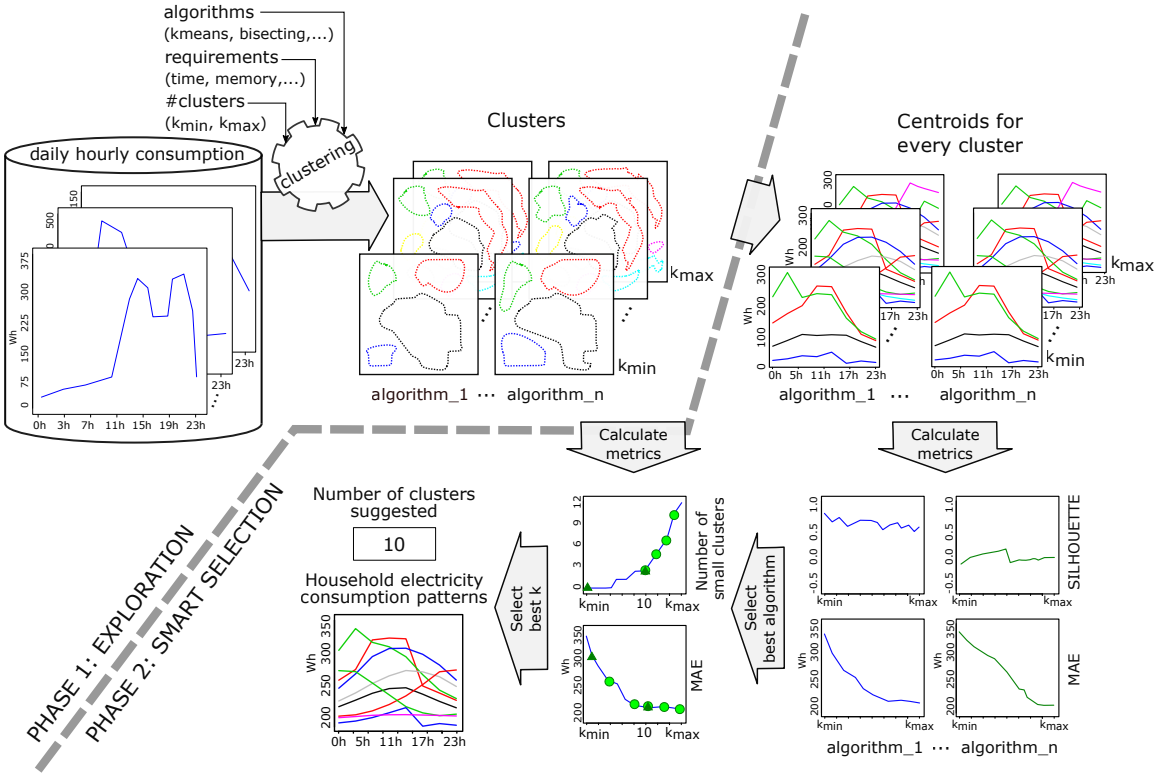


Figura 3.7: Metodología propuesta para caracterizar los perfiles de consumo eléctrico de los hogares. Fuente: (Rodríguez-Gómez y otros, unpublished_a)

Antes de explicar la metodología propuesta, se describen las transformaciones que se han aplicado a los conjuntos de datos con los que se ha validado la metodología, descritos en la Subsección 2.1.4. Se han seleccionado los datos con estos criterios:

- Se eliminaron todas las observaciones que correspondían a consumidores no domésticos. Se considera consumidor no doméstico todo consumidor que en alguna hora de cualquier día tenga un consumo de energía eléctrica superior a 15 kWh.
- Se han eliminado todas las observaciones (días) en las que falta el consumo de cualquier hora.
- Se han eliminado todas las observaciones con un consumo diario total inferior a 100 Wh.

Se propone una nueva metodología basada en el conocimiento de los expertos y la minería de datos para obtener modelos a partir de los datos y así poder caracterizar los perfiles de consumo de los hogares (o TLP, typical load profile). La metodología puede dividirse en dos fases: (a) inducción de modelos de agrupamiento (*clustering*), y (b) determinación del número óptimo de clústeres.

En la primera fase, se utilizan diferentes algoritmos de agrupamiento para generar modelos para un rango de número de clústeres definido por los expertos en el dominio ($k \in [k_{min}, k_{max}]$).

Una vez inducidos los diferentes modelos de agrupamiento para cada algoritmo, se calculan los *centroides* de cada modelo (uno para cada número de clústeres del modelo). Este paso permite calcular la distancia entre observaciones y *centroides*.

En la segunda fase, se propone el uso de una medida del error para evaluar cada uno de los modelos generados con cada uno de los algoritmos. El algoritmo que genere modelos con menor error absoluto medio (MAE), será seleccionado como algoritmo que mejor caracteriza los perfiles de consumo de los hogares de la ciudad de estudio.

Comparar dos secuencias numéricas de igual longitud (m variables) es una tarea de complejidad cuadrática si se usa la medida DTW ($O(m^2)$), pero calcular el error absoluto (AE) toma tiempo lineal ($O(m)$). Para el cálculo del MAE se ha utilizado la expresión descrita en la Sección 2.4, que en este caso puede reformularse como:

$$MAE = \frac{\sum_{i=1}^{k} \sum_{o_j \in C_i} AE(o_j, centroid_i)}{nm} \tag{3.1}$$

En caso de observar soluciones con resultados similares para el error (MAE), se utiliza el índice de Silhouette (SIL) para desempatar y seleccionar la mejor opción.

El siguiente paso es calcular una nueva información que los expertos en el dominio han considerado adecuada para evaluar la calidad de los modelos de agrupamiento: el número de clústeres con un porcentaje de observaciones menor a un porcentaje establecido por los expertos. En otras palabras, se pretende conocer el número de clústeres de pequeño tamaño, entendiendo por pequeños aquellos que no superan un umbral indicado por los expertos.

En este punto, el sistema dispone distintos valores para diferentes configuraciones de k, dando lugar a dos curvas que servirán de entrada a un nuevo algoritmo que se propone para la identificación del número óptimo de clústeres: (a) una curva con el error (MAE) de cada modelo, y (b) una curva con el número de clústeres pequeños de cada modelo.

Como se acaba de indicar, en esta metodología, se propone un nuevo procedimiento para identificar el número de clústeres más adecuado. Se llama K-ISAC_TLP y está basado en otro método, también nuevo, llamado ISAC. Ambos se describen a continuación.

Se ha seguido la recomendación sugerida por los expertos del dominio para alinear las medidas de evaluación de modelos de agrupamiento con el objetivo o dominio específico de la aplicación. Esta consideración es fundamental para generar modelos que sean útiles, como también indican otros autores (Aggarwal, 2015).

3.3.1.1. Nuevo procedimiento para determinar el número óptimo de clústeres

Como ya se adelantó en la Subsección 2.3.2.4, existen diferentes métodos para identificar el número más apropiado (u óptimo) de clústeres a considerar al realizar una tarea de agrupación en clústeres. Uno de los más utilizados es el método del codo (Shi y otros, 2021) (un enfoque manual y visual), que busca puntos en una curva donde ya no haya una mejora significativa (el codo de una curva). Pero este método adolece de algunas desventajas, como trabajar con curvas suaves o ruidosas (donde es difícil identificar el "codo correcto") o la incertidumbre de detectar un punto prematuro (cuando aún hay margen de mejora tras un codo elegido visualmente).

Esta subsección detalla un procedimiento, llamado K-ISAC_TLP, que se ha desarrollado para automatizar la detección del número más apropiado de clústeres (k) en el dominio de la caracterización de patrones de consumo eléctricos (TLP). Está basado en un método más genérico, también novedoso, llamado ISAC que evita una interpretación visual de las curvas y que se presenta en primer lugar.

ISAC (Identifier of Stable Areas in Curves) La idea principal de este método es inspeccionar una curva de forma que se detecten puntos a partir de los cuales se observe una estabilización en el comportamiento, al tiempo que cumple con unos requerimientos

mínimos. Se fundamenta en la construcción de triángulos consecutivos a lo largo del camino definido por una curva. A partir de estos triángulos es posible obtener información relevante como pueden ser las áreas de los triángulos y las pendientes definidas por los vértices más alejados de dichos triángulos.

Intuitivamente, el área de un triángulo puede mostrar alguna información sobre la alineación de los puntos que definen sus vértices: si estos puntos están alineados, el área tiende a ser cero. Por lo tanto, encontrar triángulos con áreas pequeñas implicará que sus vértices están alineados, y si esto sucede de forma consistente en sucesivos triángulos sugerirá la existencia de regiones estables en la curva. Una idea similar se ha utilizado previamente para identificar la estabilización en la curva de aprendizaje de algoritmos incrementales (Castillo y Gama, 2006; del Campo-Ávila y otros, 2008).

De la misma manera intuitiva, la pendiente entre los dos vértices más lejanos del triángulo puede indicar la velocidad de cambio en la curva: una pendiente cercana a cero sugiere una progresión mínima. Este tipo de progresión tiene una interpretación diferente según lo que represente la curva y lo que se pretenda optimizar:

a) Si la curva modela el comportamiento de una métrica que se pretende minimizar (como puede ser un error) y que típicamente comienza con valores altos que van descendiendo, una región con una pequeña pendiente indica que se espera poca mejora y sería un posible punto de parada.

b) Si la curva modela el comportamiento de una métrica que se pretende que sea estable (como puede ser la innecesaria complejidad de un modelo) y que normalmente empieza estabilizada para empezar a incrementar en algún momento, la aparición de una pendiente con un valor significativo indica que esa estabilidad inicial está desapareciendo y sería un posible punto de parada.

En la Figura 3.8 se presentan diferentes curvas en las que se pueden apreciar las casuisticas mencionadas anteriormente: triángulos con diferentes valores para el área y diferentes pendientes entre los vértices más alejados de los triángulos.

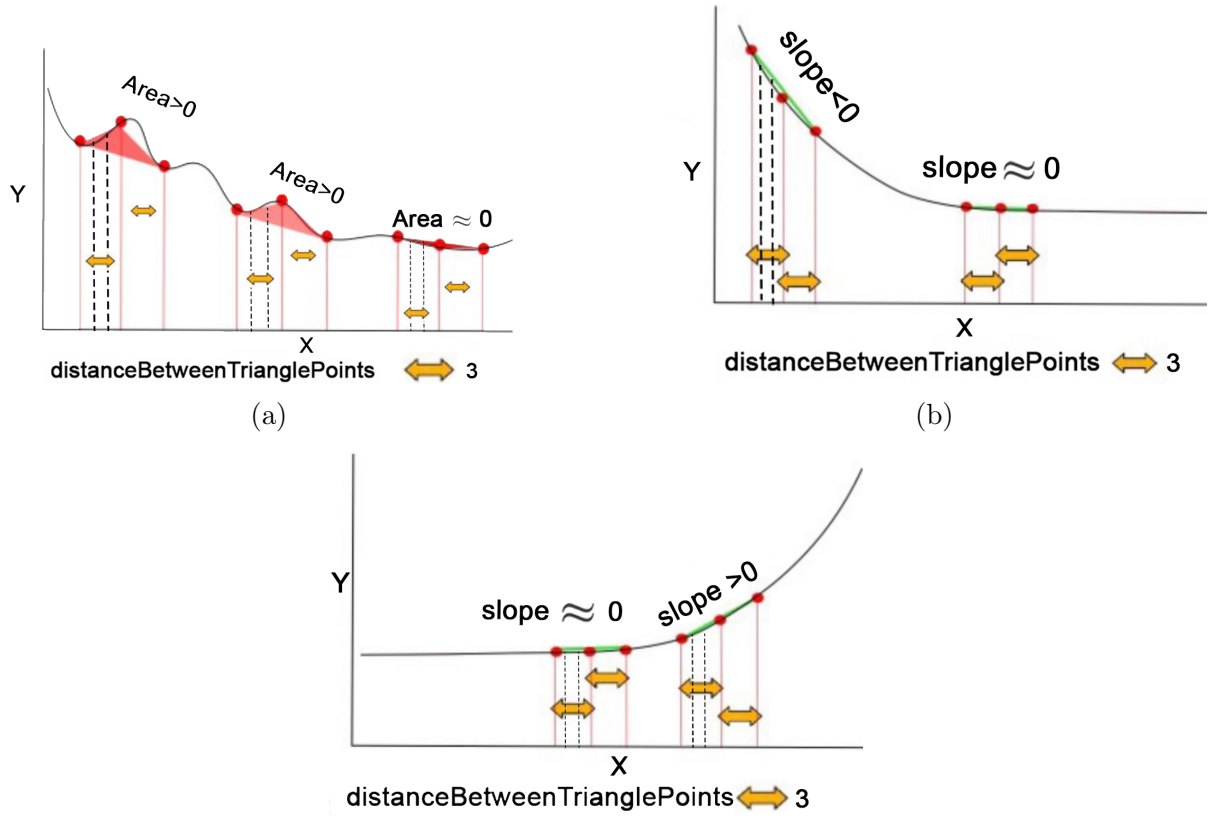


Figura 3.8: ISAC: (a) Curva con diferentes tamaños de áreas para los triángulos, (b) Curva que inicia con pendientes negativas para los vértices de los extremos de los triángulos (en algún momento las pendientes se aproximan a cero), (c) Curva que inicia estable (pendientes de los extremos de los triángulos cercanas a cero) y en algún momento las pendientes empiezan a ser positivas. Elaboración propia

En el Algoritmo 1 se detallan todos los pasos para identificar los puntos en los que una curva satisface unos criterios de estabilidad. La curva, con un total de *totalObs* puntos, se ha definido con x_values para los valores en el eje X y con $measure\_values$ para los valores en el eje Y. Desde cada punto de la curva se traza un triángulo cuyos vértices están definidos por el punto inicial y otros puntos que le siguen a una distancia preestablecida en el eje X. La distancia en cuestión se denomina $distanceBetweenTrianglePoints$ y es uno de los parámetros del algoritmo.

El algoritmo cuenta con otro parámetro que establece el número de triángulos consecutivos que tienen que seguir satisfaciendo los criterios de estabilidad para que la detección sea efectiva. Este parámetro se llama $consecutStability$ y sirve para evitar detecciones espurias de estabilidad en una región.

Dicha detección de estabilidad se concreta en dos situaciones que el método ISAC comprueba para cada punto de la curva :

- que una serie de triángulos consecutivos tengan un área menor o igual a un área máxima ($areaThreshold$), lo que denotaría que hay cierta linealidad entre los puntos, y
- que una serie de triángulos consecutivos tengan pendientes mayores que una pendiente mínima ($slopeThreshold$), lo que podría indicar un cambio en la tendencia.

K-ISAC_TLP es un nuevo procedimiento, basado en el algoritmo ISAC, específicamente configurado para la determinación del número óptimo de clústeres (k) durante la caracterización de perfiles de consumo eléctrico de los hogares.

Según el conocimiento de los expertos, existen dos curvas específicas que pueden ser significativas para seleccionar modelos que representen correctamente los perfiles característicos de consumo eléctrico. Se definen específicamente como: a) el error absoluto medio (MAE) de los modelos de agrupamiento para cada valor de k, y b) el número de clústeres irrelevantes de cada modelo para cada valor de k. Se considera que un clúster es irrelevante cuando el número de observaciones que lo forman es inferior al 1 % del conjunto de datos completo.

En estas curvas, los valores de k se organizan en el eje X de la curva, y los valores medidos para el correspondiente modelo (MAE o número de clústeres que se consideran irrelevantes por tener pocas observaciones), se colocan en el eje Y. La elección de estas medidas se corresponden con los dos tipos de curvas para las que ISAC está diseñado, de tal forma que:

a) cuando la curva describe una métrica que relaciona la complejidad de un modelo (como el número de centroides – k –) con el error del modelo (MAE), alcanzar una región estable y con pequeña pendiente supone encontrar un punto de interés.

b) cuando la curva describe una métrica que relaciona esa misma complejidad (k) con una métrica que mide qué parte del modelo es irrelevante (como el número de clústeres irrelevantes), pasar de una región sin pendiente a otra con una pendiente positiva señala otro punto de interés.

Por tanto, el método ISAC, configurado para el caso de perfiles de consumo eléctrico de los hogares, obtiene un conjunto con los valores de k relevantes para ambas curvas. La combinación de estos números de clústeres propuestos en ambas curvas, hace necesaria una priorización para determinar el valor k más apropiado. Se utilizan los siguientes criterios ordenados en importancia:

Algorithm 1: Método ISAC

Input: // DATA: List with totalObs values for independent variable x
$x\_values[1, totalObs]$
// DATA: List with measured values for totalObs measures
$measure\_values[1, totalObs]$
// number of values in X-axis between points of each triangle
$distanceBetweenTrianglePoints$
// number of consecutive triangles with stable areas
$consecutStability$
// minimum area value to consider that points in the triangle are aligned
$areaThreshold$
// maximum slope value to consider small progression
$slopeThreshold$

Output: // x values where the curve satisfies area and slope criteria
$stable\_x\_values[]$

```
// Variables to store areas and slopes for every triangle built on the curve
numberOfTriangles ← totalObs − (distanceBetweenTrianglePoints · 2);
areaTriangles[1, numberOfTriangles] ← initialize;
slopeTriangles[1, numberOfTriangles] ← initialize;
// for every triangle built in the curve
for i ← 1 to numberOfTriangles do
    // Update vertices for current triangle
    p_x ← x_values[i] ;
    p_y ← measure_values[i];
    q_x ← x_values[i + distanceBetweenTrianglePoints] ;
    q_y ← measure_values[i + distanceBetweenTrianglePoints];
    r_x ← x_values[i + (distanceBetweenTrianglePoints · 2)] ;
    r_y ← measure_values[i + (distanceBetweenTrianglePoints · 2)];
    // Calculate and store area and slope for current triangle
    areaTriangles[i] = calculateArea(p, q, r);
    slopeTriangles[i] = calculateSlope(p, r);
// for every triangle except last consecutStability triangles
for i ← 1 to (numberOfTriangles − consecutStability) do
    // Check for consecutive stable areas criteria
    isStable ← true ; j ← 0 ;
    while (isStable & (j < consecutStability)) do
        isStable ← isStable & (areaTriangles[i + j] ≤ areaThreshold)
                            & (slopeTriangles[i + j] ≥ slopeThreshold);
        j ← j + 1 ;
    // store the position of the first point for a relevant triangle
    if (isStable) then
        stable_x_values.add(x_values[i]) ;
```

1º: número mínimo común en ambas curvas;

2º: número mínimo detectado en la curva MAE; y

3º: número mínimo detectado en la curva que cuenta el número de clústeres irrelevantes.

La configuración de parámetros es el último problema antes de ejecutar el método ISAC para completar el procedimiento K-ISAC_TLP. Considerando la misma configuración determinada en investigaciones previas (Castillo y Gama, 2006; del Campo-Ávila y otros, 2008) el valor de *distanceBetweenTrianglePoints* se establece en 3. Con esa definición el método ISAC usa triángulos que no son ni demasiado grandes ni demasiado pequeños. El parámetro *consecutStable* se configura con un valor de 3 para evitar detecciones tempranas e inconsistentes, al tiempo que se reducen los requisitos de estabilidad demasiado estrictos. En cuanto a los umbrales de área y pendiente, en el contexto del consumo eléctrico de los hogares, se han seleccionado dos configuraciones diferentes en función de la curva:

- *Curva MAE*: como no puede estimarse el comportamiento de esta curva de forma genérica, se realiza una estimación del comportamiento global a partir de un triángulo hipotético que comienza al principio de la curva y termina al final (con un punto intermedio en el punto medio de la curva). Por lo tanto, los umbrales se definen de la siguiente manera:
  - El umbral para el área ($areaThreshold$) se define como el área de un triángulo proporcional al hipotético, pero considerando el tamaño de los triángulos que se utilizarán realmente en el método ISAC (según se configure el parámetro $distanceBetweenTrianglePoints$).
  - El umbral de la pendiente ($slopeThreshold$) se define como la pendiente entre el primer y el último punto de la curva.

- *Curva para el número de clústeres irrelevantes*: en este caso sí se conoce el comportamiento de esta curva. Por lo general, comienza con pocos clústeres irrelevantes (0 o casi 0) en los modelos iniciales (cuando se consideran los valores de k más pequeños) y, en algún momento, comienza a aumentar. La razón de ese incremento es la posibilidad de construir clústeres que en realidad no aportan poder descriptivo pero que en la práctica requieren unas pocas observaciones para formarse.

  Dado que se ha estudiado el comportamiento de esta curva y, puesto que el tamaño de los triángulos utilizados en el método ISAC se ha definido previamente, se han asignado valores constantes a los umbrales de la siguiente manera:

- El umbral para el área ($areaThreshold$) se define como $1,5$, que es el siguiente valor del área de los triángulos inmediatamente mayor que 0 (los valores de área aumentan en $1,5$ unidades a medida que los triángulos se hacen más grandes).
- El umbral de la pendiente ($slopeThreshold$) se define como $22,5°$, que es la mitad del ángulo de 45°. Cuando un modelo no se puede mejorar sustancialmente agregando más clústeres (aumentando k), cada nuevo clúster será irrelevante y la cantidad de clústeres irrelevantes crecerá al mismo ritmo que el valor de k. Esa tasa de crecimiento tiende a tener un ángulo de 45°.

El Algoritmo 2 ofrece una descripción detallada del procedimiento K-ISAC_TLP. El cálculo de los umbrales que necesita el algoritmo K-ISAC_TLP, considerando el consumo eléctrico de los hogares, se describen en los Algoritmos 3 y 4.

Algorithm 2: K-ISAC_TLP procedure

Input: // DATA: values for k to be considered in the range defined by the experts
$k\_values[k_{min}, k_{max}] = k_{min}, k_{min}+1, ..., k_{max}$
// DATA: List with MAE values measured for different k values
// For simplicity, assume that indexes are named $k_{min}, k_{min}+1, ..., k_{max}$
$MAE\_values[k_{min}, k_{max}]$
// DATA: List with the number of irrelevant clusters modelled for
// different k values
// For simplicity, assume that indexes are named $k_{min}, k_{min}+1, ..., k_{max}$
$irrelCluster\_values[k_{min}, k_{max}]$

Output: // values for k that meet the criteria defined by experts
$best\_k\_values[]$

```
// Configure parameters before calling ISAC method
distanceBetweenTrianglePoints ← 3;
consecutStability ← 3;
// Area threshold for MAE curve (specific depending on received curve)
MAE_areaThreshold ← estimateAdaptiveAreaThreshold(k_values,
                                                  MAE_values,
                                                  distanceBetweenTrianglePoints);
// Slope threshold for MAE curve (specific depending on received curve)
MAE_slopeThreshold ← estimateAdaptiveSlopeThreshold(k_values,
                                                    MAE_values);
// Area threshold for curve with number of irrelevant clusters (constant)
irrel_areaThreshold ← 1,5;
// Slope (in degrees) threshold for curve with number of irrelevant clusters (constant)
irrel_slopeThreshold ← 22,5;
// Variables to store best k values for every curve
best_k_values_MAE[] ← ISAC(k_values, MAE_values,
                           distanceBetweenTrianglePoints,
                           consecutStability,
                           MAE_areaThreshold, MAE_slopeThreshold);
best_k_values_irrel[] ← ISAC(k_values, irrelCluster_values,
                             distanceBetweenTrianglePoints,
                             consecutStability,
                             irrel_areaThreshold, irrel_slopeThreshold);
best_k_values[] ← commonValuesOn(best_k_values_MAE,
                                 best_k_values_irrel);
```

Algorithm 3: estimateAdaptiveAreaThreshold

Input: // DATA: values for k to be considered in the range defined by the experts
$k\_values[k_{min}, k_{max}] = k_{min}, k_{min}+1, ..., k_{max}$
// DATA: List with MAE values measured for different k values
// For simplicity, assume that indexes are named $k_{min}, k_{min}+1, ..., k_{max}$
$curve\_values[k_{min}, k_{max}]$
// number of values in X-axis between points of each triangle
$distanceBetweenTrianglePoints$

Output: // Area threshold
$areaThreshold$

// Largest triangle in the curve
$p'_x \leftarrow k_{min}$; $p'_y \leftarrow curve\_values[p'_x]$;
$q'_x \leftarrow \lceil (k_{min} + k_{max})/2 \rceil$; $q'_y \leftarrow curve\_values[q'_x]$;
$r'_x \leftarrow k_{max}$; $r'_y \leftarrow curve\_values[r'_x]$;
// Triangle proportional to largest triangle
// Share middle vertex (q'_x). Change extreme vertices (p''_x and r''_x)
$p''_x \leftarrow q'_x - distanceBetweenTrianglePoints$;
$p''_y \leftarrow q'_y + ((p'_y - q'_y) \cdot distanceBetweenTrianglePoints)/(q'_x - p'_x)$;
$r''_x \leftarrow q'_x + distanceBetweenTrianglePoints$;
$p''_y \leftarrow q'_y + ((r'_y - q'_y) \cdot distanceBetweenTrianglePoints)/(r'_x - q'_x)$;
$areaThreshold \leftarrow calculateArea(p'', q', r'')$;

Algorithm 4: estimateAdaptiveSlopeThreshold

Input: // DATA: values for k to be considered in the range defined by the experts
$k\_values[k_{min}, k_{max}] = k_{min}, k_{min}+1, ..., k_{max}$
// DATA: List with curve values measured for different k values
// For simplicity, assume that indexes are named $k_{min}, k_{min}+1, ..., k_{max}$
$curve\_values[k_{min}, k_{max}]$

Output: // Slope threshold
$slopeThreshold$

// Largest triangle in the curve
$p'_x \leftarrow k_{min}$; $p'_y \leftarrow curve\_values[p'_x]$;
$r'_x \leftarrow k_{max}$; $r'_y \leftarrow curve\_values[r'_x]$;
$slopeThreshold \leftarrow calculateSlope(p', r')$;

3.3.2. Resultados

A continuación, se muestran los resultados de aplicar la metodología automática propuesta para la caracterización de perfiles de consumo eléctrico a dos casos reales, descrita en la Sección 3.3, usando datos de consumo de España e Irlanda. La herramienta que implementa la metodología propuesta se ha llamado URSUS-TLP.

El objetivo de la primera fase es realizar una búsqueda exploratoria de grupos de datos. Esta búsqueda está guiada por expertos en el dominio del consumo de electricidad. Los expertos definieron un rango de clústeres entre 10 y 33 (k_{min}=10 y k_{max}=33), donde se deben descubrir modelos relevantes. Un valor menor podría perder grupos importantes y un valor mayor podría demorar innecesariamente la búsqueda (más de 30 o 35 grupos y sus patrones darían lugar a modelos demasiado complejos para ser analizados por los expertos).

En la primera fase, se aplicaron varios algoritmos a los conjuntos de datos de ambos países. Así, se generaron varios modelos para identificar perfiles de consumo utilizando diferentes valores de número de perfiles (k). La Tabla 3.2 detalla los algoritmos utilizados para inducir modelos de agrupamiento. También se incluyen los tiempos de procesamiento de los algoritmos que han sido capaces de generar modelos de agrupamiento para el rango de valores de k configurado por los expertos (o en su defecto la existencia de alguna limitación en tiempo o espacio).

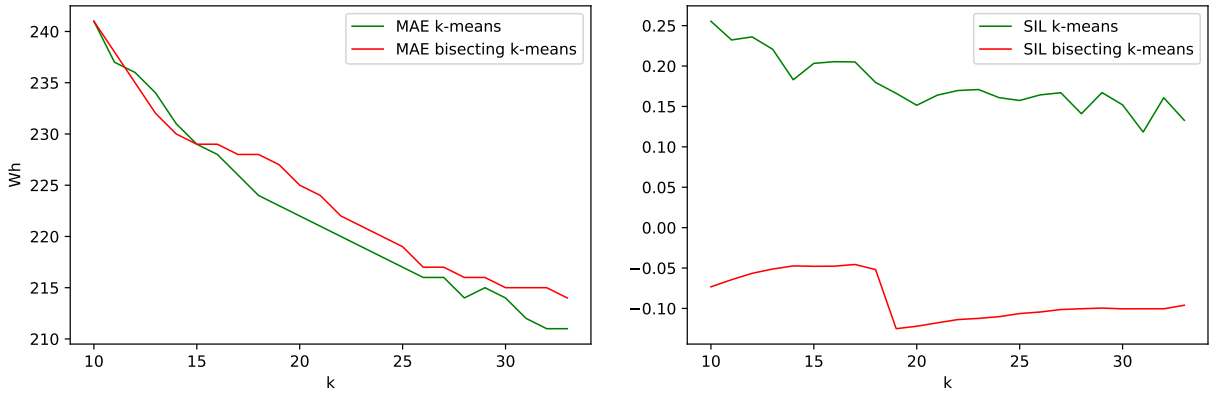
Los únicos algoritmos capaces de generar modelos de agrupamiento a partir de los 2 grandes conjuntos de datos (España e Irlanda), han sido k-means y bisecting k-means, tanto en computación paralela utilizando técnicas de agrupamiento con big data, como en computación no paralela. La reducción de tiempo al usar las versiones paralelas ha sido notable.

La segunda fase de la metodología comienza determinando el mejor algoritmo a usar y para ello se deben calcular dos métricas: el MAE y el índice de Silhouette. Se prefieren valores bajos de MAE, mientras que los valores altos de Silhouette son mejores. En la Tabla 3.2 se muestra el tiempo de ejecución de estos cálculos y se puede apreciar como, sin ser despreciable, es mucho menor que el tiempo para determinar la composición de los clústeres de la primera fase.

Para la curva MAE, los algoritmos k-means y bisecting k-means obtienen valores similares, por lo que el segundo criterio, el índice de Silhouette, ayudará con la selección. Esta métrica, calculada para ambos algoritmos, muestra que los modelos de agrupamiento obtenidos con k-means son mejores. La Figura 3.9 muestra los resultados para el conjunto de datos de España (se ha observado el mismo comportamiento en el conjunto de datos de Irlanda).

	Algoritmo	Tiempo (horas) fase 1 / fase 2	Librería / Lenguaje
Algoritmos no paralelizados	K-means	4.8 / 0.15	skicit-learn / Python
	Partitional (DTW)	Desbordamiento de memoria	tsclust / R
	Partitional (GAK)	Desbordamiento de memoria	tsclust / R
	Bisecting k-means	5.1 / 0.15	skicit-learn / Python
	BIRCH	Tiempo excedido	skicit-learn / Python
	DBSCAN	Tiempo excedido	skicit-learn / Python
	OPTICS	Tiempo excedido	skicit-learn / Python
Algoritmos de Big data	K-means	0.5 / 0.15	MLib / Python (Spark)
	Bisecting k-means	0.6 / 0.15	MLib / Python (Spark)

Tabla 3.2: Tiempos de ejecución para diferentes implementaciones de algoritmos de agrupamiento.



(a) Error absoluto medio (MAE)

(b) Índice de Silhouette

Figura 3.9: Valores de MAE y SIL para los diferentes modelos generados para cada valor de k con k-means y bisecting k-means. Fuente: (Rodríguez-Gómez y otros, unpublished_a)

La última etapa de la segunda fase de la metodología es determinar el número de clústeres más apropiado para modelar los datos utilizando el algoritmo propuesto K-ISAC_TLP. Como se explicó en la metodología, para evaluar los modelos, el algoritmo utiliza la curva de MAE y la curva con el número de clústeres pequeños (y potencialmente irrelevantes) presentes para diferentes valores de k (para cada modelo).

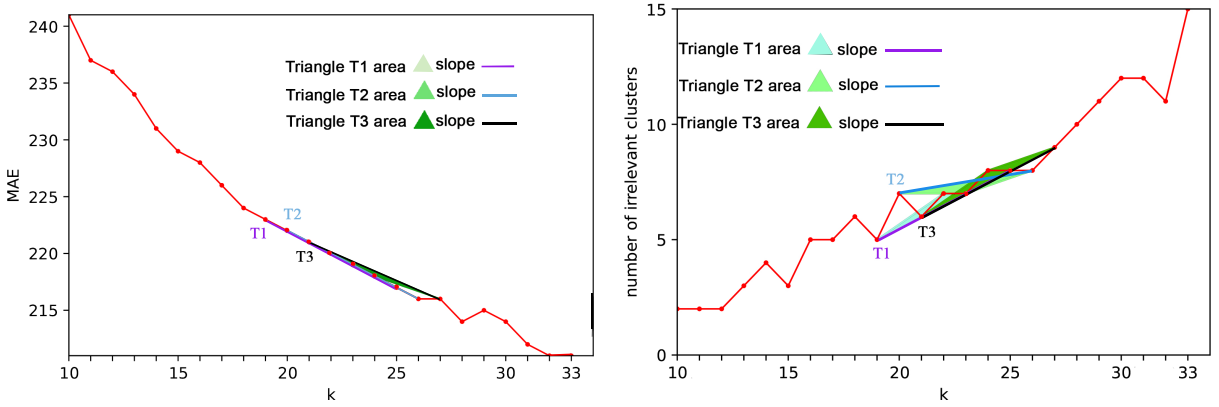
La ejecución del algoritmo K-ISAC_TLP en ambos conjuntos de datos produce resultados parcialmente similares y parcialmente diferentes. La diferencia se relaciona con la forma de los patrones detectados, que son diferentes para diferentes tipos de usuarios (lo que tiene sentido considerando dos conjuntos de datos de diferentes países).

La similitud se aprecia en la cantidad de patrones detectados, que es similar. En ambos casos, los valores de k propuestos se acercan a 20. La Figura 3.10 muestra una descripción visual del proceso para determinar el candidato como número óptimo de clústeres para el conjunto de datos de España usando el algoritmo K-ISAC_TLP. El valor k más pequeño

encontrado, y común a ambas curvas, es 19; por lo que se propone un modelo con 19 clústeres para España. Para el conjunto de datos de Irlanda, que se muestra en la Figura 3.11, el valor propuesto de k es 21.

Una vez determinados los valores de k más adecuados, los modelos específicos inducidos han sido evaluados por los expertos en el dominio para comprobar la validez de la metodología y el descubrimiento de nuevo conocimiento oculto en los datos.

Aunque el objetivo de este trabajo de investigación es elaborar una nueva metodología automatizable y no describir punto por punto cada patrón de consumo obtenido por los modelos, a continuación se presenta un resumen de los mismos. El objetivo es probar las capacidades de la propuesta, mostrar algunas mejoras logradas sobre otras metodologías y resaltar las ventajas identificadas por los expertos.



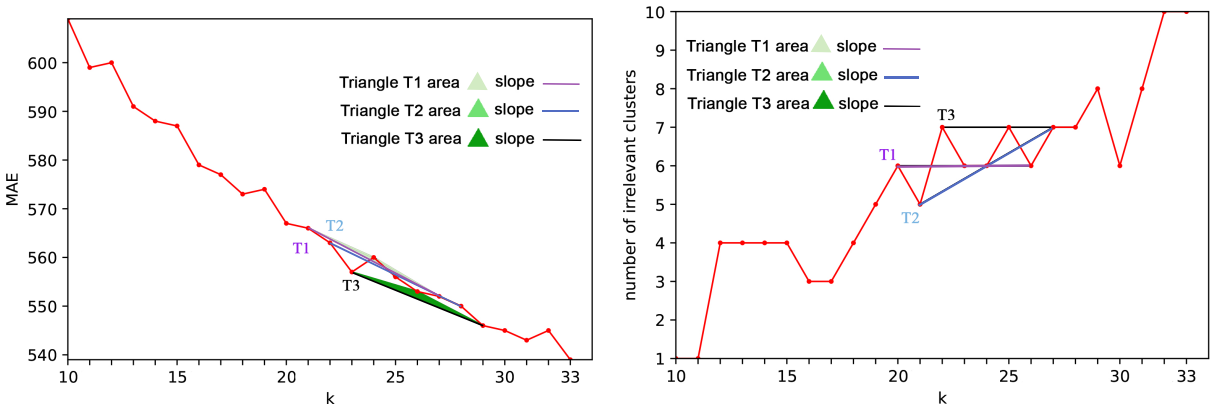
(a) Curve for MAE values (b) Curve for the number of irrelevant clusters

Figura 3.10: Descripción visual de la detección automática del valor de k propuesta por el procedimiento K-ISAC_TLP para el conjunto de datos de España. Fuente: (Rodríguez-Gómez y otros, unpublished_a)

La Figura 3.12 muestra el porcentaje de observaciones en cada clúster tanto de los datos de España como de Irlanda así como el perfil de consumo más común en España e Irlanda.

Las Figuras 3.14 y 3.13 muestran el conocimiento inducido en forma de perfiles de consumo eléctrico por el modelo k-means de 19 clústeres en España. Por otro lado, para el caso de Irlanda, las Figuras 3.16 y 3.15 muestran los 21 perfiles de consumo eléctrico inducidos, también con el método k-means.

Para los datos españoles, más del 28 % de las curvas se encuentran en el clúster 12; el consumo en este clúster oscila entre 70 y 120 Wh con valores superiores a partir de las 20:00 horas. Similar forma tiene el perfil de consumo del clúster 11, con más del 24 % de curvas; en este clúster el consumo es mayor y oscila entre 150 y 400 Wh. Los perfiles de consumo de los clústeres 10, 15, 17, 18 y 19 tienen formas similares aunque valores diferentes. Corresponden a hogares donde el consumo máximo se produce alrededor de



(a) Curve for MAE values

(b) Curve for the number of irrelevant clusters

Figura 3.11: Descripción visual de la detección automática del valor k propuesta por el procedimiento K-ISAC_TLP para el conjunto de datos de Irlanda. El mejor valor de k es 21, el k más bajo detectado por el método ISAC para la curva MAE. En este caso no existe coincidencia entre las curvas, por lo que se da prioridad a la curva MAE definida en la metodología. Fuente: (Rodríguez-Gómez y otros, unpublished_a)

las 20:00 horas, este valor máximo es de 1000 Wh para el clúster 10 (con el 8,6 % de las curvas), 1500, 3000, 2000 y 1500 Wh respectivamente para los clústeres 15, 17, 18 y 19. Los clústeres 1, 5, 8, 9 y 13 corresponden a perfiles de consumo con el valor máximo alrededor del mediodía; el consumo máximo oscila entre 1000 y 3000 Wh. Los clústeres 7 y 16 corresponden a viviendas donde el consumo máximo se registra a la 01:00 con un consumo máximo de 1600 y 1400 Wh respectivamente. El resto de clústeres tienen menos del 1 % de observaciones y, por tanto, no se consideran representativos de ningún patrón de consumo.

Para los datos de Irlanda, más del 24 % de las observaciones pertenecen al clúster 6. El consumo en este clúster tiene un pico de consumo de 500 Wh por la tarde alrededor de las 20:00 horas, y un pico de consumo de alrededor de 400 Wh por la mañana. El clúster 9 tiene una forma similar pero los valores de consumo son mayores, unos 3000 Wh por la mañana y unos 3300 Wh por la tarde (sólo el 1,5 % de las observaciones pertenece a este cluster). El segundo clúster con más observaciones (casi un 18 %) es el clúster 1. La forma de este perfil corresponde a un consumo que inicia por la mañana, con un pico alrededor de las 8:00 horas que comienza a aumentar a las 15:00 alcanzando un consumo máximo alrededor de las 20:00. Formas similares tienen los clústeres 11, 17, 19 y 21, con un consumo máximo de 4000, 3000, 5000 y 6000 Wh respectivamente. Los porcentajes de observaciones en estos clústeres varían entre 2,4 y 6,4. Las observaciones en el clúster 4 también tienen un consumo por la mañana y un pico de consumo de 3000 Wh por la tarde sobre las 17:00. Los clústeres 7, 10, 12, 13, 18 y 20 corresponden a perfiles de consumo que alcanzan un pico a lo largo de las mañanas (entre 2000 y 5000 Wh) y

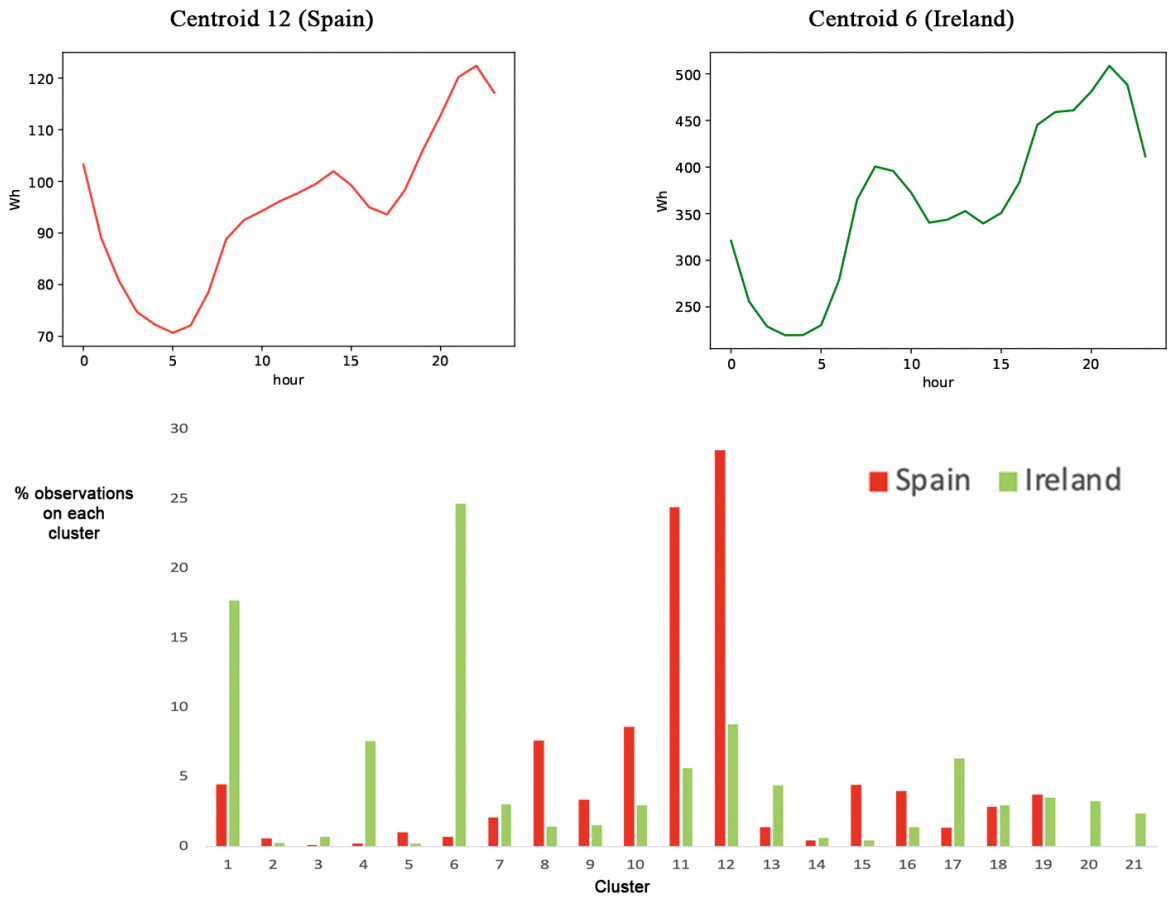


Figura 3.12: Porcentaje de observaciones en cada clúster. Fuente: (Rodríguez-Gómez y otros, unpublished_a)

tienen un consumo máximo de entre 1000 y 2000 Wh por las tardes; alrededor del 25 % de las observaciones se encuentran en estos perfiles. El perfil de consumo del clúster 5 corresponde a consumos nocturnos de hasta 8000 Wh; estas observaciones corresponden a consumidores no domésticos. El resto de clústeres (3, 2, 8, 14, 15 y 16) tienen un perfil de consumo realizado a lo largo del día, iniciando sobre las 9:00 o 10:00, y prolongándose aproximadamente hasta las 19:00 con valores de consumo que oscilan entre 6000 y 9000 Wh, lo que podría corresponder a comercios o empresas y no a usuarios domésticos.

En cuanto a resultados anteriores obtenidos para datos de Irlanda, en los resultados de McLoughlin y otros (2015) se utiliza una parte del conjunto de datos original, por lo que se puede hacer una comparación con nuestros resultados. En su caso, utilizaron datos de consumo recopilados durante seis meses para 3.941 consumidores. Proponen un total de 10 perfiles de consumo de electricidad como resultado de aplicar una serie de algoritmos de agrupamiento (k-means, SOM y k-medoids). Han utilizado el índice de Davies Bouldin para determinar el número óptimo de clústeres. Cada uno de estos perfiles está definido por 48 valores (consumo eléctrico en intervalos de media hora).

Aunque su enfoque es diferente al propuesto en este trabajo, el conjunto de datos que utilizan es más pequeño (en usuarios y meses), y los resultados son diferentes (hay 21 clústeres con 24 valores horarios en nuestro caso), los perfiles de consumo obtenidos como resultado de ambas metodologías, se pueden comparar para evaluar similitudes y diferencias. De acuerdo con los perfiles de consumo presentados en la Figura 6 del trabajo anterior (McLoughlin y otros, 2015), parecen ser coherentes con los obtenidos en este trabajo. En concreto, los clústeres 1, 4 y 9 que corresponden a perfiles de consumo con menor consumo por la mañana y mayor consumo por la tarde alrededor de las 19:00 o 20:00, son similares a los clústeres 1 y 6 obtenidos por nuestro modelo. Los clústeres 2 y 7 en (McLoughlin y otros, 2015) corresponden a un perfil de consumo donde los mayores consumos ocurren a mediodía; perfiles similares se encuentran en los perfiles 12, 13 y 18 descubiertos por nuestro modelo. De la misma forma, los perfiles de consumo 3 y 10 de dicho trabajo, corresponden a consumos con un pico a lo largo de la mañana y un menor consumo por la tarde; perfiles similares se encuentran en los clústeres 7, 10 y 20 de los propuestos en este trabajo.

Una diferencia importante con el trabajo de McLoughlin y otros (2015) consiste en que los algoritmos de agrupamiento y las métricas de evaluación, que ellos utilizan se han aplicado a una serie de muestras de consumo diario tomadas al azar del conjunto total de datos original. Dicho muestreo implica que los modelos de agrupamiento que generan tienen un número menor de perfiles, por lo que no pueden reflejar el conocimiento obtenido al trabajar con todo el conjunto de datos, como se ha hecho aplicando la metodología propuesta en este trabajo.

Considerar el conjunto de datos completo puede tener un costo computacional más alto, que incluso podría necesitar del uso de técnicas de *big data*, pero el resultado obtenido contemplará una mayor diversidad de patrones sin pérdida de información. Por ello, la aplicación de la metodología propuesta permite generar modelos que detectan más perfiles de consumo eléctrico que otras metodologías. Por ejemplo, se detectan patrones asociados a consumidores no domésticos (negocios o empresas).

Además de no tener que muestrear, algunas ventajas adicionales observadas por los expertos están relacionadas con el bajo nivel de preprocesamiento que debe realizarse:

- Solo se definen algunos filtros para eliminar observaciones incompletas o extremadamente raras.
- No se normalizan los datos. Esto puede revelar algunos patrones cuya forma puede ser similar pero que tengan diferentes rangos de consumo y estén relacionados con el mismo comportamiento de uso de la electricidad. Una ventaja adicional es la ausencia del paso de desnormalización, una tarea compleja que puede introducir

algunos errores en el proceso.

- No se aplica muestreo ni ningún proceso de selección o división de los datos originales. Esta idea reduce el sesgo introducido por la selección aleatoria de datos dejando algunos fuera del proceso, y por lo tanto evita las pérdidas de información en forma de nuevos clústeres que la representen. También evita procesos repetitivos y costosos para evaluar y validar la consistencia de los modelos generados a partir de datos de diferentes muestras.

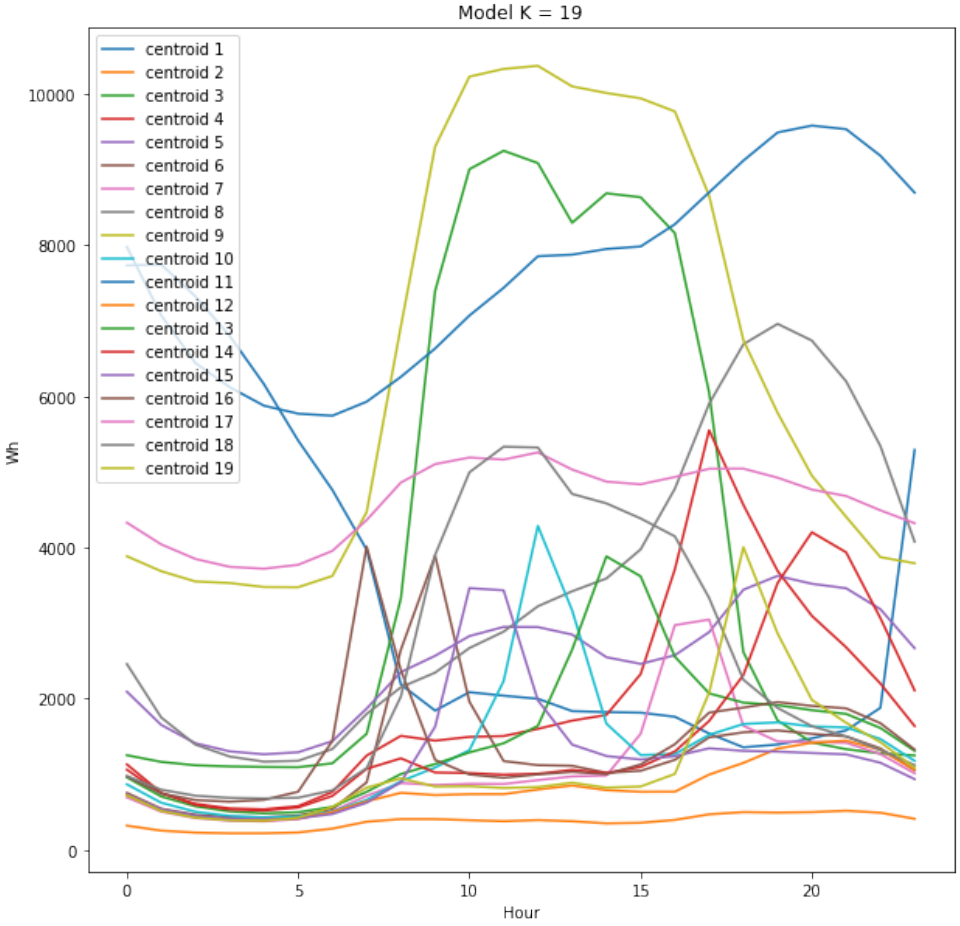


Figura 3.13: Centroides para los 19 perfiles de consumo eléctrico (España). Fuente: (Rodríguez-Gómez y otros, unpublished_a)

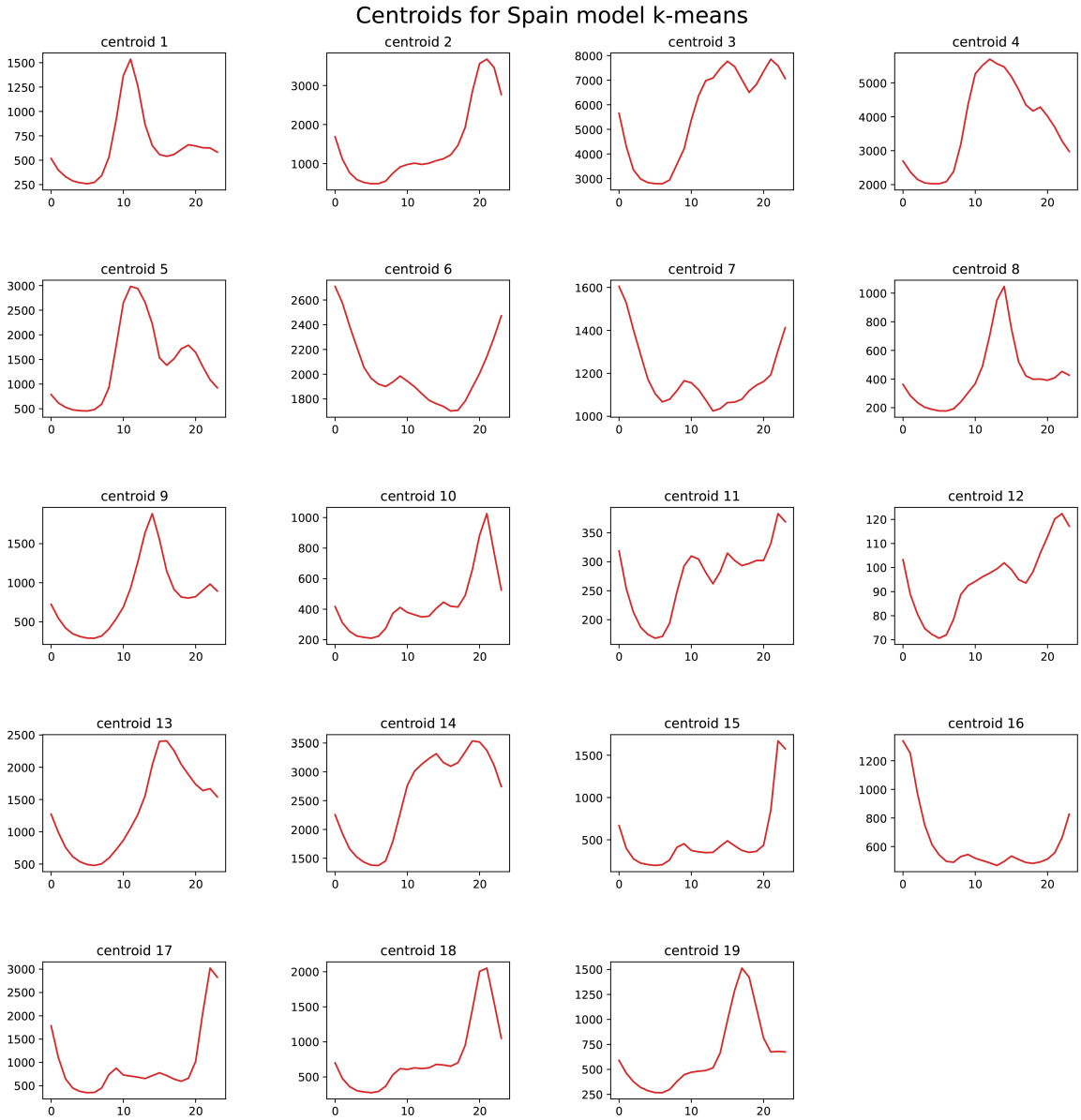


Figura 3.14: Perfiles de consumo eléctrico característicos (España). Fuente: (Rodríguez-Gómez y otros, unpublished_a)

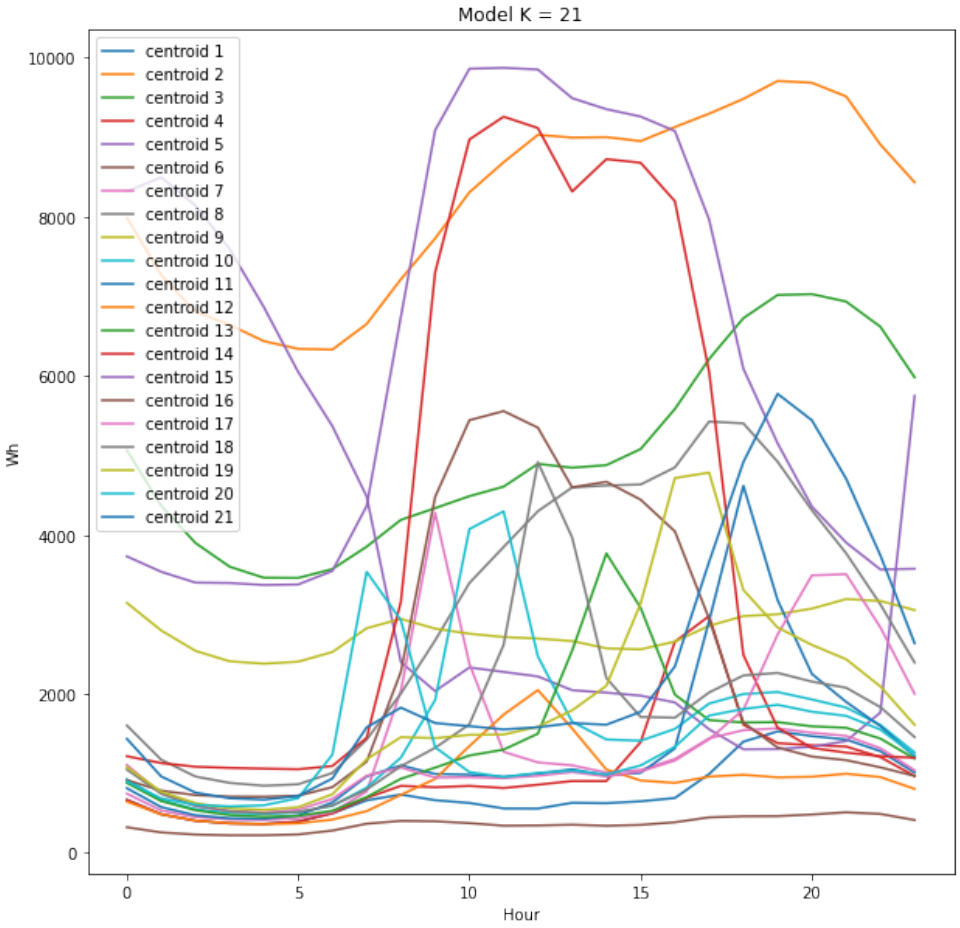


Figura 3.15: Centroides para los 21 perfiles de consumo eléctrico (Irlanda). Fuente: (Rodríguez-Gómez y otros, unpublished_a)

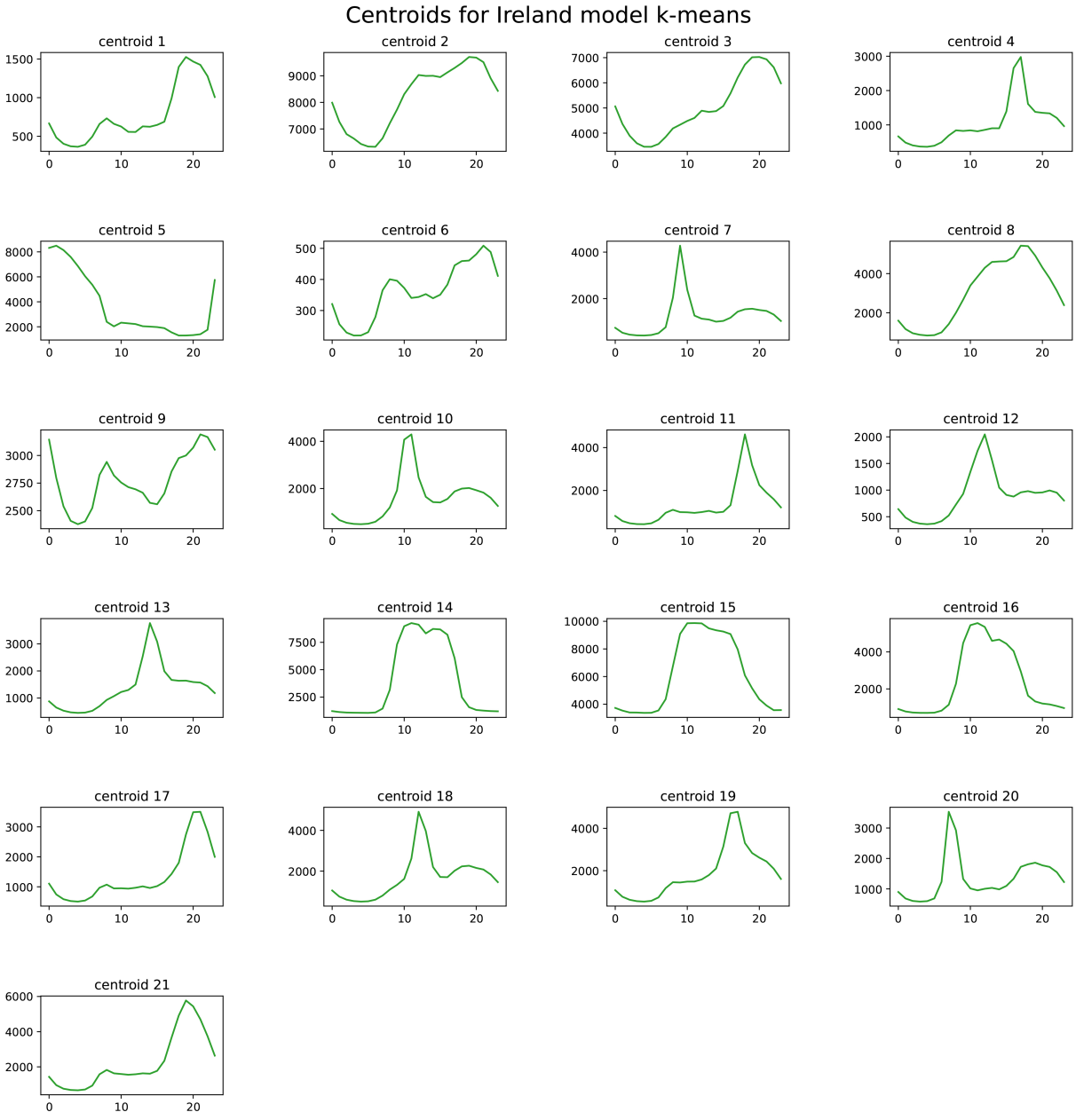


Figura 3.16: Perfiles de consumo eléctrico característicos (Irlanda). Fuente: (Rodríguez-Gómez y otros, unpublished_a)

3.4. Conclusiones

A continuación se detallan las principales conclusiones alcanzadas en este capítulo sobre generación y consumo de energía en entornos urbanos. Empezamos con los desarrollos alcanzados en el ámbito de la **selección automática de emplazamientos para instalaciones fotovoltaicas en entornos urbanos**.

Se ha desarrollado URSUS-PV, una herramienta simple, gratuita, y de código abierto que permite estimar el potencial de energía fotovoltaica que se puede generar en instalaciones fotovoltaicas (PV) a corto plazo (un día antes) y a largo plazo (promedio diario) en diferentes áreas urbana de interés (barrios, calles, urbanizaciones ...). El sistema inteligente podría ser potencialmente útil para múltiples tipos de usuarios, incluidos municipios, administraciones públicas, empresas del sector fotovoltaico, cooperativas o comunidades de vecinos.

Uno de los beneficios más significativos de la herramienta, consiste en la automatización de un proceso muy complejo y costoso en tiempo. Anteriormente, para obtener resultados globales en cuanto a potencial de energía fotovoltaica en áreas urbanas de interés a corto o largo plazo, se tenía que llevar a cabo un proceso manual.

URSUS-PV se puede ampliar fácilmente para incluir tantas ciudades como sea necesario. Solo se necesitan imágenes LiDAR y datos meteorológicos de la ciudad, y dichos datos suelen estar disponibles en las agencias nacionales de forma gratuita. El uso de este software podría mejorar la sostenibilidad energética urbana y ayudar en la lucha contra el cambio climático. Proporciona información de producción a largo plazo (media diaria), que es fundamental a la hora de evaluar, desde un punto de vista económico, la viabilidad de integrar instalaciones fotovoltaicas en entornos urbanos. Por tanto, antes de acometer la instalación de cualquier sistema fotovoltaico, se podría estimar en qué medida se satisfacen los requerimientos energéticos diarios y/o económicos demandados.

La información a corto plazo, ayudará a los gestores de grandes instalaciones a mejorar su integración en la red eléctrica. También puede ayudar a los propietarios de instalaciones de autoconsumo a determinar cuándo tendrán energía generada por sus instalaciones y desplazar su consumo a dichos horarios.

Esta herramienta ofrece excelentes oportunidades porque se puede actualizar fácilmente con nuevas funcionalidades implementadas por los mismos desarrolladores o por nuevos colaboradores ya que el código fuente es de acceso libre y se encuentra disponible de forma gratuita, al igual que la propia herramienta. Algunas de las nuevas características, podrían consistir en la inclusión de la detección automática de nuevos tipos de elementos en una ciudad diferentes a los tejados para ubicar las instalaciones, y en conseguir una

una estimación más precisa del potencial energético.

Algunos de los nuevos elementos urbanos que podrían considerarse (además de los tejados) para la localización óptima de emplazamientos fotovoltaicos podrían ser parcelas sin edificios, calles principales, o aparcamientos, que a su vez podrían tener un doble efecto positivo, ya que al mismo tiempo, generen sombras para los peatones.

El sombreado que se produce entre edificios es un punto no considerado en este trabajo, aunque no es imprescindible, ya que el sombreado se produce en las horas extremas del día, y las instalaciones fotovoltaicas obtienen su máximo rendimiento en las horas centrales del día.

La fase de estimaciones y cálculos de potencial energético también se ha automatizado completamente utilizando datos meteorológicos de la ciudad y la configuración de la instalación fotovoltaica que se podría integrar en cada cubierta.

A continuación, se procede a la descripción de las conclusiones alcanzadas como resultado del trabajos de investigación relacionado con la línea de investigación relativa a la **caracterización de perfiles de consumo eléctrico de los hogares**.

Se ha propuesto una nueva metodología que permite descubrir los perfiles de consumo eléctrico característicos de los hogares. Sus principales avances surgen del trabajo coordinado con expertos en la materia. En la primera fase, es fundamental delimitar un rango para el espacio de búsqueda en el que se debe encontrar el número de tipos de patrones de consumo (número de clústeres). Posteriormente, se pueden utilizar múltiples algoritmos de agrupamiento para generar diferentes modelos. Un punto crucial es la definición de las métricas para evaluar la calidad de modelos considerando el dominio de aplicación; en el ámbito de la caracterización de los perfiles de consumo eléctrico de los hogares, dichas métricas serán el MAE y el número de clústeres irrelevantes. Otro punto a destacar de la metodología propuesta, es el nuevo algoritmo que se propone para buscar automáticamente el número más adecuado de clústeres en el rango definido por los expertos.

La herramienta se ha implementado como software de código abierto y ha demostrado ser útil para ayudar a los expertos a obtener información oculta, útil, y relevante, a partir de grandes conjuntos de datos de consumidores domésticos minimizando los pasos de preprocesamiento y automatizando todo el proceso. Teniendo en cuenta todo el conjunto de datos, los modelos inducidos evitan la pérdida de información, tiempos de procesamiento, y problemas de inconsistencia que habitualmente ocurren con el muestreo. Como no se han normalizado los datos, todos los patrones se pueden comparar directamente ya que se identifican patrones con formas comunes pero diferentes rangos de consumo. Al prescindir del preprocesamiento de datos que suele utilizarse en otros trabajos

para dividir el conjunto de datos en consumos en función de diferentes tipos de días, meses o estaciones, la metodología que se propone identifica los patrones independientemente de estos aspectos temporales.

Investigaciones posteriores permitirán asignar características definitorias a cada patrón. Esas características pueden no ser únicas porque el mismo patrón puede repetirse en diferentes estaciones, meses o días para los mismos o diferentes usuarios. Por lo tanto, la clasificación típica del perfil de consumo para un usuario en un día específico será más flexible.

Los modelos obtenidos podrán ayudar a los expertos a adquirir rápidamente nuevos conocimientos en el dominio del consumo eléctrico de los hogares. Además, los resultados alcanzados podrán tener un impacto positivo en la lucha contra el cambio climático y en la mejora de la sostenibilidad urbana.

La metodología es fácilmente extensible a problemas de cualquier dominio donde los algoritmos de agrupamiento son de aplicación. Los expertos pueden hacer un buen uso de sus conocimientos definiendo el rango de búsqueda para acotar el número de clústeres, y estableciendo las métricas adecuadas con las que trabajará el algoritmo ISAC. Este algoritmo propuesto para identificar un número apropiado de clústeres también es extensible a otros campos y será objeto de trabajos futuros de investigación.

4

Herramientas inteligentes para asesorar en el enverdecimiento de ciudades

4.1. Estado del arte

Numerosos estudios sobre el cambio climático predicen un aumento generalizado de las temperaturas. Las consecuencias del aumento de las temperaturas son más preocupantes y significativas en las zonas urbanas que en las zonas rurales circundantes. Este fenómeno de calentamiento se debe principalmente al desarrollo en áreas urbanas y al aumento de las áreas edificadas (Santamouris, 2014; Lee y otros, 2013) y se conoce como fenómeno de isla de calor urbano (UHI, por sus siglas en inglés). El uso de materiales de construcción que absorben la mayor parte de la radiación solar y la liberan en forma de calor contribuye a reforzar este fenómeno. Los impactos del UHI afectan a la salud y a la calidad de vida de los residentes de forma directa o indirecta (Lin y otros, 2013), e incluso pueden incidir en la mortalidad (Goggins y otros, 2012; Heaviside y otros, 2016). La Agencia de Protección Ambiental de EE.UU. describe los problemas de salud respiratoria, los trastornos fisiológicos, la insolación y el aumento de la mortalidad como posibles consecuencias del fenómeno UHI (U.S. EPA, 2021). Esto representa un nuevo desafío para

la planificación urbana ya que es necesaria la implementación de diferentes alternativas para paliar estos problemas y hacer que las ciudades sean más resilientes.

En el trabajo de Susca y otros (2011) puede verse que existe una clara correlación entre la abundancia de vegetación y la temperatura de la superficie del suelo. Por lo tanto, la presencia de Infraestructuras Verdes Urbanas (UGI, por sus siglas en inglés) contribuye a la mitigación del fenómeno UHI (Hart y Sailor, 2009) y, como exponen Herrera-Gomez y otros (2017), se demuestra que incrementar el área con vegetación (tanto en número como en superficie), especialmente en las ubicaciones más desfavorables, constituye una solución interesante para la lucha contra tal fenómeno. Identificar las regiones potencialmente útiles, y decidir las acciones más convenientes a llevar a cabo, son parte del problema de ordenamiento territorial, que es una tarea de búsqueda y optimización potencialmente exigente (Behzadi y Alesheikh, 2013). Las características de los UGI (convencionales o integrados en los edificios), su abundancia, estado y distribución en las ciudades, influyen en su efectividad para mitigar el efecto UHI (Park y otros, 2017; Yang y otros, 2017). Por lo tanto, la presencia de estas infraestructuras verdes contribuyen a la mitigación del efecto UHI y puede llevarse a cabo mediante la instalación de elementos de diferente tipología: jardines verticales, techos verdes, parques públicos, reservas naturales, bosques, arroyos, vías verdes, senderos o jardines comunitarios. Una descripción de varias de estas infraestructuras, así como su contribución a mitigar el efecto UHI puede encontrarse en los trabajos de Hart y Sailor (2009) o Wolch y otros (2014).

A modo de resumen, un mecanismo importante para que las ciudades se adapten y sean reconducidas hacia modelos urbanos sostenibles, consiste en rediseñar sus infraestructuras incrementando la infraestructura verde urbana de las ciudades. Por eso mismo, cualquier espacio al aire libre que esté parcial o totalmente cubierto de vegetación, como hierba, arbustos, árboles u otro tipo de vegetación, se considera que es una infraestructura verde urbana (U.S. EPA, 2021).

Aunque la instalación de estas infraestructuras puede suponer un mayor coste y mantenimiento a las administraciones públicas, se están desarrollando políticas para potenciar su inclusión en las ciudades. La creación de parques y jardines es siempre deseable, pero no siempre es posible cuando las ciudades están densamente pobladas y no hay espacio para ubicarlos. En estos casos, se usan otras infraestructuras verdes como, por ejemplo, los jardines verticales o techos verdes. La Figura 4.1 muestra algunos ejemplos de los diferentes elementos de naturación que se pueden integrar en las ciudades. Su principal beneficio es la mitigación de las elevadas temperaturas en las zonas más desfavorables por el efecto isla de calor urbano. Otro de sus aspectos positivos se observa en la reducción de los consumos derivados de la climatización (European Commission, 2013) ya que algunas UGI, como los techos verdes o paredes verticales, producen un efecto

aislante en los edificios que envuelven. Al mitigar las temperaturas elevadas, provocan a su vez un mejor desempeño de las instalaciones fotovoltaicas, por lo que combinar su uso debería apreciarse un comportamiento doblemente positivo. El uso de infraestructuras verdes urbanas podría producir una reducción de la temperatura ambiente con valores entre 0,5°C y 2°C como se concluye en el estudio realizado por Susca y otros (2011).

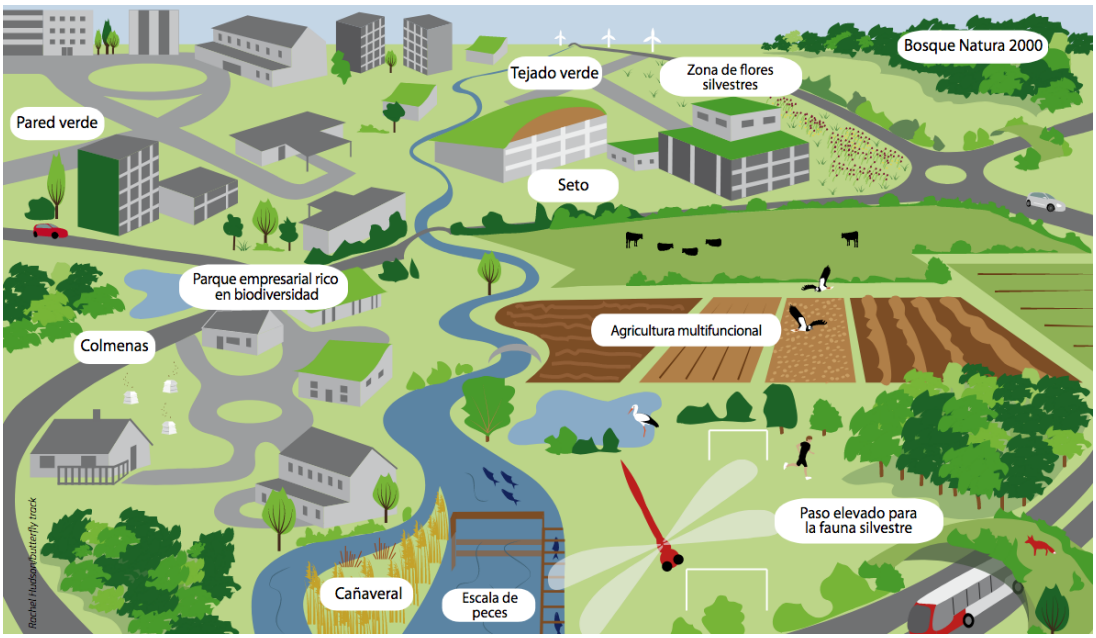


Figura 4.1: Posibles elementos de una infraestructura verde. Fuente: (EC)

Disponer de herramientas que permitan a los planificadores urbanos detectar automáticamente las áreas más desfavorables por el efecto UHI, y que permitan simular la temperatura que se obtendría en diferentes escenarios modificando la distancia, tipo, y cantidad de vegetación, supondría una mejora en la sostenibilidad urbana y un gran avance en la lucha contra el cambio climático si se aplicase de forma globalizada.

Las administraciones públicas están también concienciadas con la tarea de tener que enverdecer las ciudades para mitigar el efecto isla de calor urbano y reducir las temperaturas, ya que son conocedoras de su importancia. Esta tarea la suelen llevar a cabo los planificadores urbanos, pero conocer las áreas más apropiadas para ser enverdecidas es una tarea que demanda grandes recursos y puede no ofrecer los mejores resultados en caso de hacerse manualmente. El desarrollo de herramientas que automaticen los procesos les sería de gran utilidad porque podrían guiarles en la determinación de las zonas más desfavorables, serían capaces de realizar la predicción de la temperatura en cualquier área urbana analizando el entorno, e incluso les permitiría realizar simulaciones de temperatura en diferentes escenarios modificando la abundancia y la tipología de vegetación.

Disponer de esta herramienta, podría incrementar la incorporación más eficaz de

infraestructuras verdes en las ciudades, agilizando la tarea de reconversión hacia modelos urbanos sostenibles.

4.1.1. Determinación de zonas más desfavorables por el efecto UHI

En algunos trabajos se han estimado *manualmente* las ubicaciones óptimas para la instalación de UGI y el impacto que su implantación puede provocar en ciudades concretas con respecto al efecto UHI (Fernández y otros, 2015; Chun y Guldmann, 2018; Nesticò y otros, 2022). Esta forma de estimar ubicaciones óptimas no es escalable, ni rápida, ni reproducible o extensible fácilmente a otras ciudades. Estas limitaciones se pueden sortear gracias al potencial de la minería de datos espacial y a la sencillez que ofrece para la integración de datos heterogéneos, o para la extracción automática de información útil, novedosa y relevante. La implantación de un proceso se minería de datos adecuado se presenta como una estrategia apropiada para analizar datos de fuentes de sensores remotos e identificar automáticamente las áreas de una ciudad que son más desfavorables debido al efecto UHI.

Tal y como se ha descrito en la Sección 4.1, las zonas más desfavorables por el efecto UHI, se corresponden con áreas de escasa vegetación y elevada temperatura de la superficie terrestre.

Las imágenes satelitales suelen ser una de las fuentes de información actuales más empleadas para realizar análisis de variables climáticas de las ciudades tal y como se ha descrito en la Subsección 2.1.1, por lo que esta ha sido la alternativa seleccionada para la consecución de los objetivos en esta línea de investigación.

Se pueden utilizar diferentes técnicas analíticas para la teledetección de zonas verdes urbanas, como por ejemplo la segmentación semántica de elementos urbanos mediante técnicas de procesamiento de imágenes, o también se pueden calcular diferentes índices de cobertura terrestre. Por lo general, esta segunda técnica utiliza combinaciones de diferentes bandas de onda de sensores satelitales multiespectrales. El Índice de Vegetación de Diferencia Normalizada (NDVI) es el índice más conocido y aplicado para mapear las zonas verdes (Xue y Su, 2017). Es útil para diferenciar regiones verdes y no verdes dentro de áreas urbanas, así como para la caracterización del tipo de terreno.

Para la inducción de modelos que permitan clasificar de forma automática las zonas urbanas más desfavorables analizando la temperatura de la superficie terrestre y la vegetación, ha sido necesario llevar a cabo un estudio de las técnicas actuales más relevantes para calcular el NDVI y la LST. Se han utilizado los fundamentos teóricos

descritos en las Subsecciones 2.2.2 y 2.2.3 donde se detallan los pasos y expresiones para poder calcular el NDVI y la LST de un punto dado y así poder extenderlo a las diferentes ciudades a partir de imágenes satelitales.

Cada píxel en la imagen satelital está descrito con un par de valores, NDVI y LST, pero no existe una relación previa entre esas variables para categorizar una zona como desfavorable o no desfavorable. Tras realizar la revisión del estado del arte de la minería de datos en el contexto de la sostenibilidad urbana, el aprendizaje automático no supervisado (clustering) ha sido la técnica seleccionada como más adecuada para abordar este problema. Se han aprovechado los fundamentos teóricos, descritos en la Subsección 2.3.2, para decidir los algoritmos de clustering más adecuados, así como para seleccionar las métricas para determinar el número de clústeres más adecuado para abordar el problema.

Una vez están disponibles las variables LST y NDVI, y seleccionadas las técnicas y los algoritmos para realizar agrupaciones, es posible proceder al entrenamiento de modelos para la clasificación de las áreas urbanas en función de la combinación de ambos valores (NDVI, LST) en los diferentes puntos de las ciudades.

4.1.2. Predicción de la temperatura de la superficie terrestre (LST)

Estimar la temperatura de la superficie terrestre en un punto, analizando las características del entorno urbano, es fundamental para realizar simulaciones sobre la temperatura que se obtendría en diferentes escenarios modificando el tipo de vegetación, así como la distancia a zonas con vegetación o el tamaño de estas. Para poder inducir modelos basados en datos que permitan estimar esa temperatura en la superficie ha sido necesario llevar a cabo un proceso de análisis y revisión del estado del arte en cuanto a fundamentos teóricos, técnicas, y trabajos similares que perseguían el mismo objetivo.

El uso de imágenes satelitales para estudiar fenómenos climáticos o la distribución de UGI a escala de ciudad es muy útil, como se ha descrito en la Subsección 4.1.1. Existe una amplia gama de tecnologías de teledetección que pueden proporcionar una cobertura completa de las ciudades como nubes de puntos LiDAR o imágenes satelitales. La disponibilidad para un gran número de ciudades y el acceso gratuito a estos datos, hacen que la teledetección sea una herramienta valiosa como se describe en el trabajo de Zhu y otros (2019) para la extracción de información sobre el estado de las ciudades, incluso en entornos urbanos altamente heterogéneos y complejos. Programas como Copernicus [1] y Landsat [2] se encargan de tener actualizados los datos que permiten obtener información

[1] `https://www.copernicus.eu/en`
[2] `https://landsat.gsfc.nasa.gov`

sobre el estado de las ciudades, facilitando la descarga y acceso de series temporales o imágenes satelitales.

Se han realizado varios estudios sobre el papel que juega la vegetación en la disminución de las temperaturas en entornos urbanos en los últimos años (Fu y otros, 2022). Muchos de ellos, se basan en sensores remotos, pero la metodología utilizada para obtener los resultados es *manual* (Du y otros, 2017; Estoque y otros, 2017; Reis y Lopes, 2019; Su y otros, 2022; Liu y otros, 2022).

Un método usado frecuentemente para hacer predicciones de la temperatura de la superficie terrestre (LST) consiste en entrenar modelos de aprendizaje automático a partir de series temporales de LST a lo largo de los años en varias ubicaciones urbanas. Por ejemplo, en el trabajo de Mustafa y otros (2020) se clasifica el entorno urbano utilizando algoritmos de aprendizaje automático supervisado para extraer conclusiones sobre las variaciones de temperatura en función del entorno analizado en Beijing, China. Es fundamental señalar que los modelos fueron entrenados únicamente con datos de series temporales LST, sin incorporar información adicional sobre las características del entorno urbano para el entrenamiento. Otros enfoques que utilizan datos de series temporales de diferentes años aplican modelos ARIMA para predecir los valores de la LST, como se demuestra en el estudio de Chennai, India, realizado por Kesavan y otros (2021).

Examinar las características del entorno urbano para hacer predicciones de la LST es una forma de incorporar nueva información que podría proporcionar una mejor comprensión del problema. En esta línea, Khalil y otros (2021) proponen entrenar algoritmos de aprendizaje automático mediante el análisis de características ambientales (EVI: índice de vegetación mejorado, elevación y densidad de carreteras) junto con series temporales de LST recopiladas durante 20 años en Lahore, Pakistán. En otro trabajo para la predicción de LST en Irán (Karimi y Ghajari, 2022), se utilizan algoritmos genéticos para inducir modelos a través del entrenamiento a partir de características como la densidad de edificios o el nivel de contaminación del aire.

Identificar las relaciones entre los elementos urbanos y las temperaturas en las ciudades es una tarea compleja, en gran parte debido a la gran cantidad de datos y a su naturaleza heterogénea, ya que es un proceso que involucra muchas variables relacionadas con la sostenibilidad urbana. Las técnicas y aplicaciones de minería de datos, han demostrado su capacidad para descubrir conocimientos ocultos en muchos dominios (Liao y otros, 2012), lo que las convierte en un enfoque adecuado para aplicar al problema de determinar la temperatura de la superficie terrestre (LST) analizando las características del entorno cercano.

El uso de imágenes Landsat para extraer la LST para el entrenamiento de modelos es común, pero también se pueden utilizar otras fuentes de imágenes. Por ejemplo, Kartal y

Sekertekin (2022) utilizan imágenes satelitales MODIS de la región sur de Turquía. Las imágenes Landsat ofrecen una resolución temporal más baja, mientras que las imágenes MODIS ofrecen una resolución espacial más baja.

La investigación basada en el entrenamiento de modelos a partir de series temporales LST presenta una limitación cuando no tiene en cuenta las características urbanas de las áreas de estudio. Dichos estudios no permiten la simulación de nuevos escenarios para predecir la temperatura modificando las características del entorno, por lo que la incorporación de dichas características en la fase de entrenamiento de modelos será crucial. Las nubes de puntos LiDAR, combinadas con las imágenes satelitales, pueden proporcionar la información necesaria relativa al entorno urbano. Sharma y otros (2021) muestran todo el potencial de esta tecnología en términos de extracción de características urbanas. Una de las ventajas de usar imágenes LiDAR, es la simplicidad que aporta a la fase de preparación de datos, ya que incluye una clasificación urbana para cada punto de la imagen (por ejemplo, edificio, vegetación escasa, densa o moderada).

4.2. Determinación de zonas más desfavorables por el efecto UHI

4.2.1. Propuesta metodológica

Se propone una nueva metodología para determinar de forma automática las zonas más desfavorables por el efecto UHI (ausencia de vegetación y elevadas temperaturas). Esta metodología es aplicable a cualquier ciudad para la que se tengan los datos disponibles. La Figura 4.2 muestra de forma esquemática el resumen de la metodología que se describe a continuación.

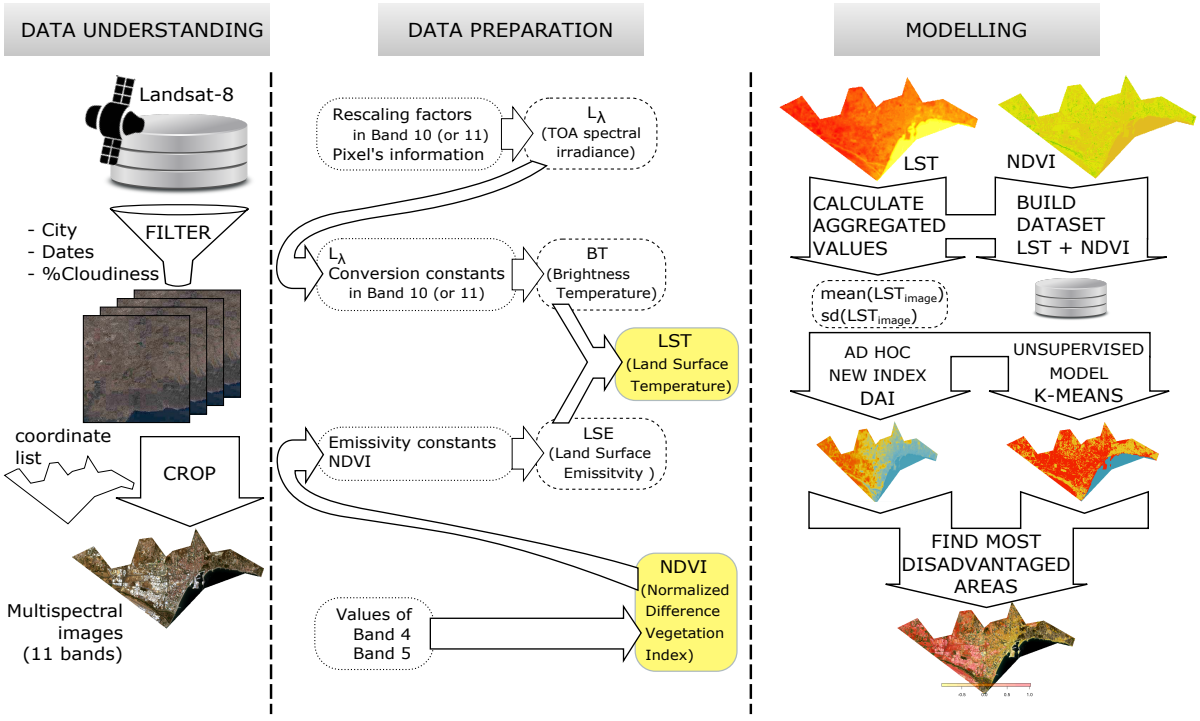


Figura 4.2: Resumen de la metodología donde se describen las principales fases para la identificación de áreas desfavorecidas. Fuente: (Rodríguez-Gómez y otros, unpublished_c)

Una vez se dispone de los datos necesarios y se ha definido el área de interés, el primer paso consiste en calcular el NDVI siguiendo los pasos descritos en la Subsección 2.2.2 a partir de la imagen Landsat-8 de la ciudad de estudio. A continuación, se calcula la LST a partir de la imagen satelital utilizada en el paso anterior siguiendo los pasos descritos en la Subsección 2.2.3 .

El siguiente paso consiste en el cálculo de un nuevo índice propuesto por los expertos. Este índice se utiliza para cuantificar el nivel de deterioro de las zonas más desfavorables

por el UHI analizando la información conjunta ofrecida por las variables LST y NDVI de la imagen de la ciudad de estudio. Se describe a continuación en la Subsección 4.2.1.1.

A continuación, se procede a la fase de modelizado para asignar automáticamente cada píxel de la ciudad a diferentes clústeres. Este proceso se detalla en la Subsección 4.2.1.2, y se centrará en la identificación de aquel clúster (o clústeres) donde se concentren los píxeles con escasa o nula vegetación y elevadas temperaturas.

En último lugar, el sistema utilizará la información calculada en los pasos previos para mostrar un mapa que informará de las zonas desfavorables y el valor de DAI para dichos píxeles, lo que permitirá identificar de forma visual las zonas que necesitan una actuación más urgente contra el efecto UHI.

4.2.1.1. Índice de áreas desfavorables. DAI

Se ha modelizado, siguiendo el conocimiento y recomendaciones de los expertos, un nuevo índice (Disadvantaged Area Index –DAI–) para determinar el grado de deterioro de un área en términos de exceso de temperatura y escasez de vegetación. Este índice está diseñado para asignar valores bajos a áreas que, comparadas con el resto de áreas, presentan una mayor cantidad de vegetación y unas temperaturas más bajas (NDVI mayores y LST menores). Por el contrario, los valores altos de DAI aparecen cuando se observan temperaturas más altas junto con poca o ninguna vegetación en esa región. Una vez se ha descrito el funcionamiento de este índice, se define de la siguiente forma:

$$\mathrm{DAI}_{\mathrm{pixel}} = \left(1 - \frac{\tanh(\mathrm{NDVI}_{\mathrm{pixel}})}{\tanh(1)}\right) \cdot \tanh\left(\frac{\mathrm{LST}_{\mathrm{pixel}} - \mathrm{media}(\mathrm{LST}_{\mathrm{imagen}})}{\mathrm{sd}(\mathrm{LST}_{\mathrm{imagen}})}\right) \tag{4.1}$$

donde $\mathrm{LST}_{\mathrm{image}}$ y $\mathrm{LST}_{\mathrm{pixel}}$ son el conjunto de todos los valores LST en la imagen y el valor LST de un píxel específico, respectivamente; $\mathrm{NDVI}_{\mathrm{pixel}}$ es el valor NDVI para ese mismo píxel y media y sd son las operaciones estadísticas para calcular la *media aritmética* y la *desviación típica*.

El uso de la función de tangente hiperbólica para la definición del índice DAI asegura que el valor esté entre -1 y 1 para valores NDVI no negativos. En el caso extremo en el que el NDVI de un píxel es exactamente 1, el DAI se convierte en 0. Además, el signo del DAI en un píxel determinado coincide con la temperatura de ese píxel en relación con la temperatura media de la región: el valor del DAI es positivo para píxeles con temperatura superior a la media y negativo para píxeles por debajo de la temperatura media. Podemos verificar visualmente la relación entre NDVI, temperaturas y DAI en la Figura 4.3.

Si profundizamos en el análisis formal del comportamiento del DAI podemos destacar que:

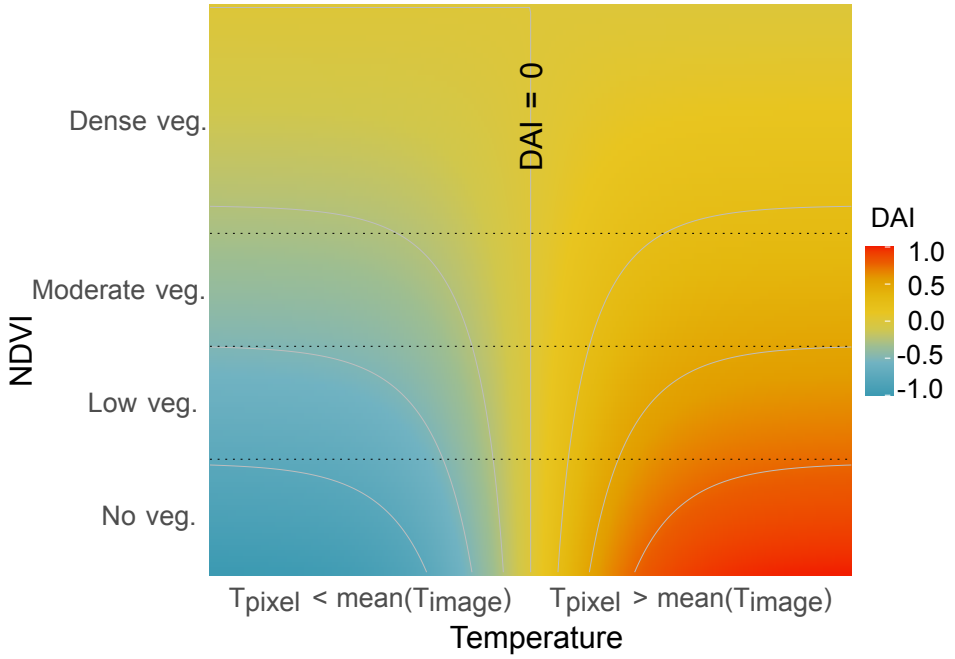


Figura 4.3: Mapa de calor de los valores DAI en función del NDVI y la diferencia de temperatura entre un píxel y el promedio de la imagen. Fuente: (Rodríguez-Gómez y otros, unpublished_c)

- El punto de equilibrio DAI = 0 se alcanza bajo dos condiciones: cuando NDVI = 1, de modo que el primer factor se cancela, o cuando $LST_{\text{pixel}} = \text{media}(LST_{\text{imagen}})$.
- El valor máximo DAI = 1 se alcanza cuando NDVI ≈ 0 (sin vegetación) junto con valores de temperatura muy altos ($LST_{\text{pixel}} >> \text{media}(LST_{\text{imagen}})$). Las regiones más desfavorecidas según el DAI serán aquellas que no tengan vegetación (y que serían susceptibles incluir nuevas infraestructuras verdes) y cuya temperatura supere en mayor medida la temperatura media de la región.
- El valor mínimo de DAI $= -1$ se alcanza cuando NDVI ≈ 0 y hay valores muy bajos de temperatura en el punto, en comparación con el promedio regional. Esta elección se debe a que las características topográficas y urbanísticas de estos sitios, a pesar de tener poca o ninguna vegetación, no tienen un LST alto y por lo tanto no sería necesario incluir vegetación allí.
- Los valores de DAI cercanos a 0 y ligeramente más altos corresponden al entorno urbano en el que la temperatura aumenta para los valores de NDVI en relación con áreas densamente plantadas. La razón para no considerar de interés para los planificadores urbanos a estas áreas urbanas radica en que son áreas donde es difícil añadir vegetación.
- Es en las zonas con poca o ninguna vegetación donde el término de temperatura

tiene mayor influencia en el cálculo del DAI, ya que en estos casos el primer factor se acerca a 1. Son precisamente en aquellas zonas con poca vegetación donde se puede actuar, por lo que los planificadores urbanos pueden usar el DAI para determinar cuáles de estas zonas son las más desfavorables.

- Como puede verse en la propia definición de DAI, el valor de NDVI es relevante para determinar las áreas menos favorecidas. Si dos puntos tienen la misma temperatura $T_0 >> \text{mean}(T_{\text{image}})$, el primer factor hace que los valores más bajos de NDVI (regiones con menos vegetación) se consideren más desfavorables. Del mismo modo, si la temperatura de ambos puntos es inferior a la media de la región, el segundo factor cambia de signo y la importancia del NDVI cambia de significado: los valores más altos de NDVI serán los más desfavorecidos, ya que modelan puntos que, a pesar de tener vegetación, no reducen tanto la temperatura como en las zonas donde hay menos vegetación para la misma temperatura.

4.2.1.2. Clustering de áreas urbanas más desfavorables

La segunda fase propuesta para la detección de áreas desfavorables se basa en algoritmos de agrupamiento. Este tipo de algoritmo se utiliza para crear grupos de píxeles en una imagen que comparten características similares. Estos algoritmos, al igual que la mayoría de los algoritmos de aprendizaje automático, son más difíciles de aplicar en conjuntos de datos de alta dimensionalidad (cantidad de observaciones o de variables). En este caso, la cantidad de variables no es un problema porque solo se usan dos variables para agrupar los píxeles: valores LST y NDVI. Pero, por otro lado, se debe considerar una gran cantidad de píxeles, y solo algunos algoritmos (como k-means (MacQueen, 1967) o CLARA (Kaufman y Rousseeuw, 1986)) pueden ser aplicados para la generación de modelos al tratar con imágenes satelitales de ciudades con una alta dimensionalidad en cuanto a número de píxeles. Si se consideran ciudades o pueblos más pequeños, sería posible utilizar otros algoritmos. En cualquier caso, y en línea con las recomendaciones de la literatura, k-means es un algoritmo comúnmente seleccionado debido a su simplicidad y eficiencia (De La Torre y Kanade, 2006). A pesar de algunos inconvenientes, como su sensibilidad a las condiciones iniciales (Mittal y otros, 2019), k-means ha logrado resultados prometedores en aplicaciones similares como se muestra en el trabajo de Martínez-Gordón y otros (2021). La idea principal de este procedimiento es descubrir automáticamente agrupaciones formadas por las áreas más desfavorables, en términos de escasez de vegetación y altas temperaturas.

El espacio de búsqueda, en cuanto a la determinación del número óptimo de clústeres, debe conocerse *a priori*, antes de aplicar los algoritmos de clustering. El número óptimo de clústeres generalmente se obtiene después de generar modelos para diferentes números

de clústeres y calcular una métrica de evaluación del agrupamiento realizado (Ezugwu y otros, 2021). Una de las métricas más utilizadas es el índice de Silhouette (Rousseeuw, 1987), que evalúa la similitud entre las observaciones de un mismo clúster (cohesión) y la diferencia entre las observaciones con respecto a otros clústeres (separación). En esta búsqueda, el conocimiento experto es muy útil para acotar el espacio de búsqueda.

4.2.2. Resultados

La metodología descrita ha sido implementada en la herramienta URSUS-UHI[3], y se ha aplicado a dieciséis ciudades de España para detectar las zonas más desfavorables por el efecto UHI y así identificar aquellas zonas en las que es más urgente intervenir. A continuación, se describen las ciudades seleccionadas para el estudio así como una justificación de la selección de las mismas.

Las dieciséis ciudades que se han seleccionado para el estudio cubren varios lugares de la geografía española con diferentes climas, incluyendo ciudades de varios tamaños y diferentes cantidades de infraestructuras verdes en su ámbito (UGI). En la Tabla 4.1 se enumeran dichas ciudades y, según sus características climáticas, se han categorizado en cuatro grupos: (I) ciudades del interior con veranos calurosos, (II) ciudades mediterráneas con costa, (III) ciudades del norte cercanas al mar, y (IV) ciudades del norte alejadas del mar. Las ciudades de los grupos I y II se caracterizan por un período cálido prolongado durante la primavera y el verano, aunque el primero alcanza temperaturas más altas. En cambio, en el grupo II, se observan temperaturas más suaves, con mayor humedad relativa. Las ciudades del grupo III tienen temperaturas más bajas en verano. Las ciudades del grupo IV también presentan temperaturas más bajas que las del sur, aunque pueden alcanzar temperaturas altas (superiores a los 30°C) en determinadas épocas del año.

Siguiendo la metodología descrita anteriormente, el primer paso sería el cálculo del NDVI y la LST para las ciudades de estudio. En la Figura 4.4 se muestra el mapa de NDVI y LST de dos de esas ciudades (Málaga y Sevilla). A continuación, se procedió al cálculo del índice DAI y a la inducción de los modelos de clustering para cada ciudad.

Para las ciudades a las que se les ha aplicado la metodología, los expertos han establecido acotar el espacio de búsqueda entre 3 y 10 clústeres. Tomando como criterio el índice de Silhouette se ha propuesto el uso de 3 clústeres como valor óptimo para las ciudades procesadas tras analizar su NDVI y LST y generar diferentes modelos para los diferentes números de clústeres. Esta decisión ha sido validada por los expertos. De este modo, un clúster representa las zonas más desfavorables, otro representa las áreas urbanas correspondientes a agua, y otro clúster, representa a las zonas menos desfavorables. La

[3]Urban sustainability for more disfavourable UHI areas detection

Tabla 4.1: Ciudades de España seleccionadas para el estudio de zonas desfavorables (UHI)

	Ciudad	**Habitantes(miles)**	**Área urbana (km^2)**	**Zonas verdes por habitante (m^2/habitante)**	**Temp. media (°C)**
I	Madrid	3183	274	15,78	26,9
	Sevilla	689	75	11,27	28,6
	Murcia	443	9,7	3,38	26,6
	Ciudad Real	75	9	31,44	27,8
	Alcobendas	116	18,7	15,13	26,7
II	Barcelona	1621	76,8	5,53	24,3
	Valencia	788	73	4,19	26,8
	Málaga	569	93,3	2,92	27,3
	Palma de Mallorca	416	27,7	3,34	26,2
III	Bilbao	345	17,5	5,74	21,6
	Vigo	293	5,9	3,65	20,5
	Vitoria	247	28,7	26,76	19,4
	Santiago Compost.	96	10,1	23,61	18,7
IV	Zaragoza	665	47,6	9,02	26,7
	Huesca	220	2,9	5,69	25,1
	Lleida	138	10,2	10,13	25,8

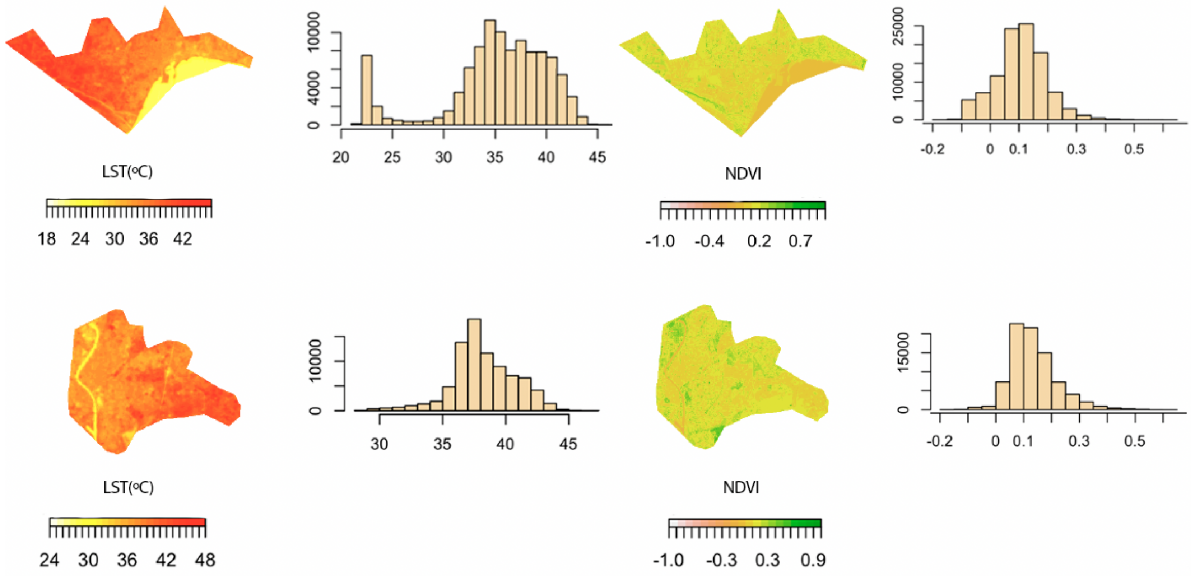


Figura 4.4: Imágenes NDVI y LST e histogramas para Málaga (arriba) y Sevilla (abajo). Fuente: (Rodríguez-Gómez y otros, 2022b).

Figura 4.5 muestra el número de clústeres óptimo para los modelos inducidos con el algoritmo k-means en las ciudades de Málaga y Sevilla para los diferentes números de clústeres.

La Figura 4.6 muestra, para Málaga y Sevilla, la relación entre los valores de LST y

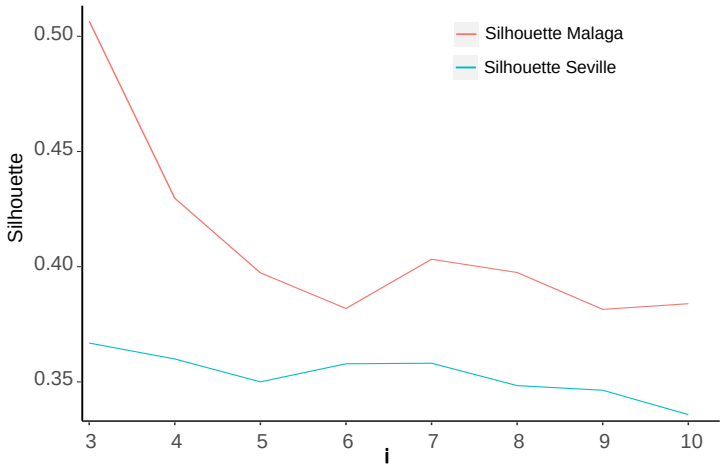


Figura 4.5: Índice de Silhouette para los modelos k-means según diferentes valores de k entre 3 y 10. Casos representados para Málaga y Sevilla. Un mayor valor del índice Silhouette indica un mejor desempeño del modelo. Fuente: (Rodríguez-Gómez y otros, unpublished_c)

NDVI con el índice DAI asociado y los clústeres detectados por el algoritmo. Se pone de manifiesto que las zonas con elevada temperatura (LST) y escasez de vegetación (bajo NDVI) suelen corresponderse con el clúster que representa las zonas desfavorables (cuyo DAI es mayor).

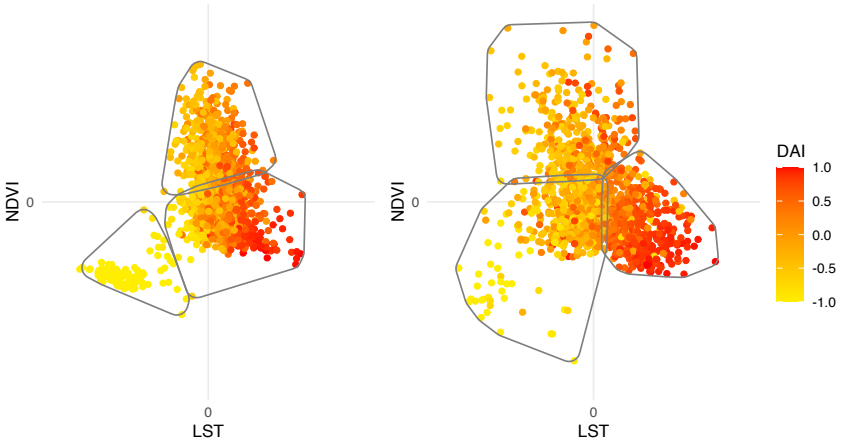


Figura 4.6: Relación entre los valores del DAI en función del clúster al que pertenecen y los valores de LST (eje X) y NDVI (eje Y) para Málaga (izquierda) y Sevilla (derecha).Fuente: (Rodríguez-Gómez y otros, unpublished_c)

La Figura 4.7 muestra el resultado de aplicar la metodología propuesta a algunas de las 16 ciudades de España. En el anexo A.4 se muestra el resultado de aplicar la metodología a las 16 ciudades. La escala de valores de DAI sólo se muestra para el clúster más desfavorable. En general, el porcentaje de áreas susceptibles de ser intervenidas de forma más urgente

mediante la instalación de infraestructuras verdes en las ciudades estudiadas, cambió sustancialmente de una ciudad a otra. Por ejemplo, solo el 13 % de la superficie total de Huesca se identificó como desfavorable, mientras que Bilbao y Valencia superan el 60 %.

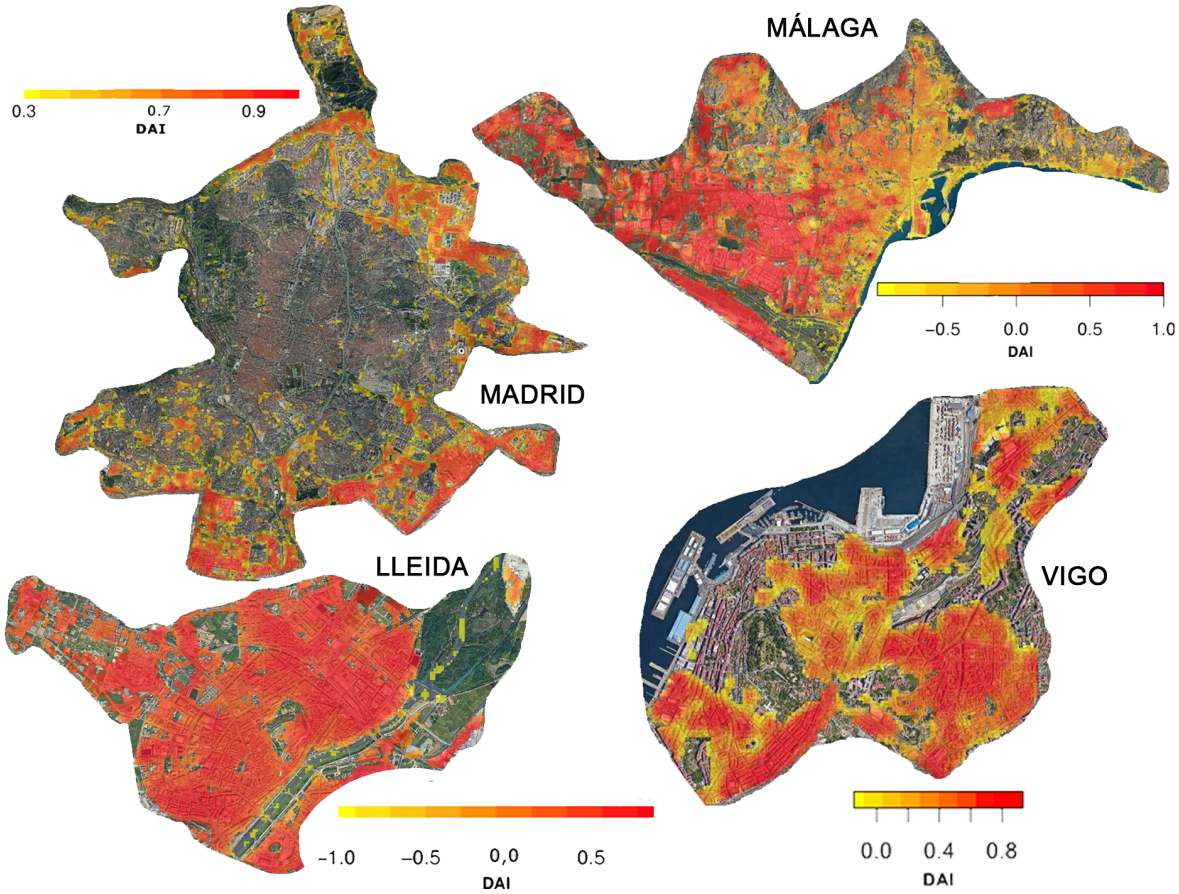


Figura 4.7: Zonas más desfavorables según el índice DAI para algunas de las 16 ciudades de estudio de España. Fuente: (Rodríguez-Gómez y otros, 2022b).

Las infraestructuras verdes (UGI) no son los únicos factores que afectan a la temperatura en un área particular de la ciudad, sino que también entran en juego otros factores. La influencia de la presencia de masas de agua es evidente como se muestra en el trabajo de Cai y otros (2018). Las zonas desfavorables suelen estar alejadas de la línea de costa tanto en Málaga, como en Palma de Mallorca y Barcelona. La influencia de la UGI en la LST también tiende a ser mayor lejos de la costa que cerca de ella (Ossola y otros, 2021).

Los espacios azules dinámicos, como los ríos, pueden absorber la radiación (Hathway y Sharples, 2012). Por ejemplo, las zonas más desfavorables no son las cercanas al río Guadalquivir en Sevilla o al río Ebro en Zaragoza. Las excepciones a esto se pueden ver, por ejemplo, en Lleida. Si bien se observa perfectamente el efecto favorable sobre las temperaturas urbanas que brinda el río que cruza la ciudad y el gran bosque de ribera ubicado en el sector noreste de la ciudad, se encuentran algunas zonas muy

desfavorables cerca del río y del bosque de ribera. Estas áreas corresponden principalmente a dos tipologías urbanas: (a) zonas con grandes superficies pavimentadas (por ejemplo, grandes plazas, estacionamientos de supermercados o patios de escuelas) y (b) grupos de grandes edificios industriales con techos metálicos o de grava, que son planas o ligeramente inclinadas. Por lo tanto, se deben plantar más árboles en grandes parcelas pavimentadas y se debe alentar a los edificios industriales a integrar techos verdes o sistemas de vegetación vertical. Esto también es aplicable a grandes edificios en el centro de la ciudad, como estaciones de tren (por ejemplo, la estación de Atocha en Madrid o la estación de Santa Justa en Sevilla).

Las áreas industriales suelen incluirse en el grupo de las zonas más desfavorecidas en todas las ciudades estudiadas, ya que están formadas por grandes edificios construidos con materiales que acumulan calor y rodeados de espacios impermeables con muy poca vegetación. Por ejemplo, las pocas zonas desfavorables que se encuentran en Vitoria y Huesca tienen precisamente esas características como puede verse en la Figura 4.8.

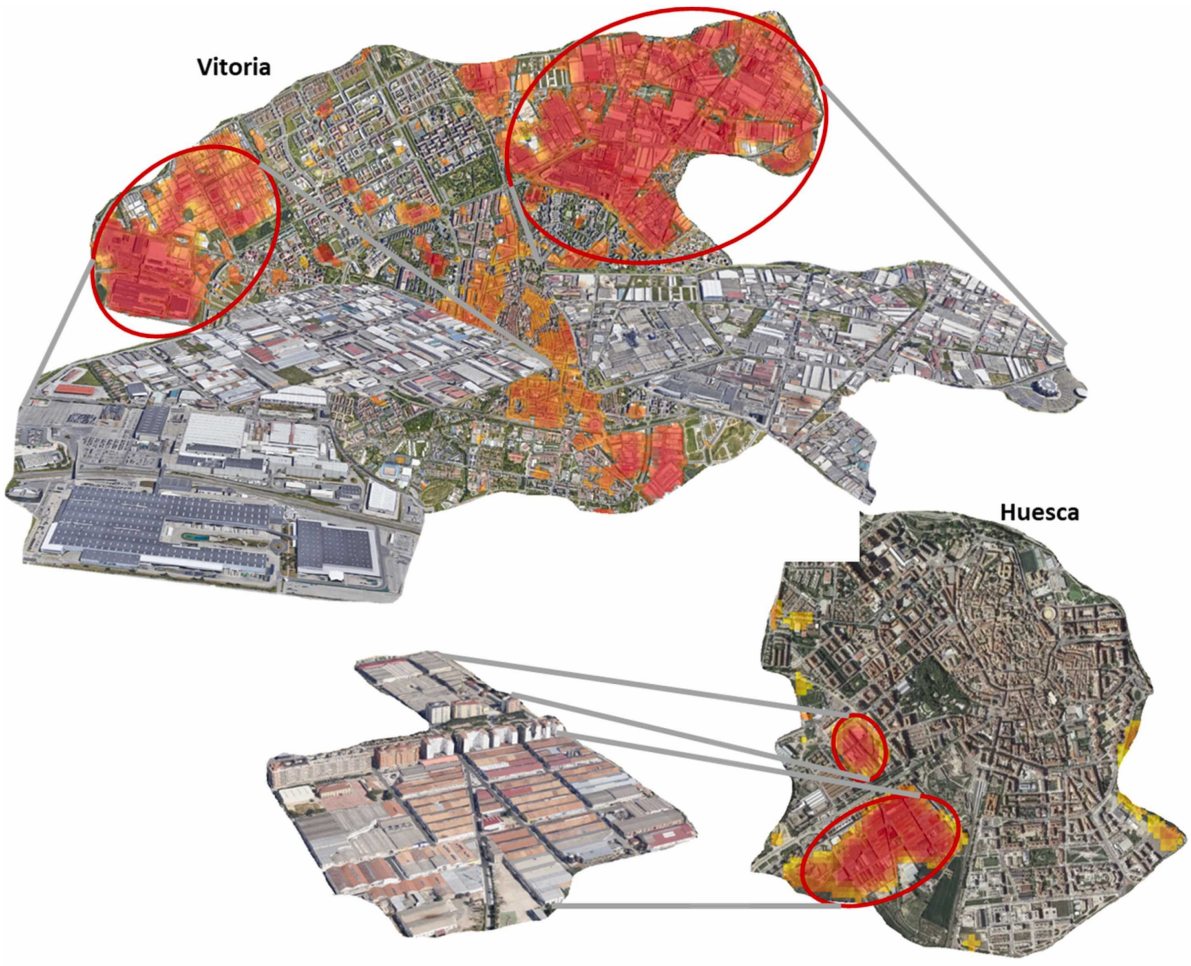


Figura 4.8: Zonas más desfavorables (UHI) en zonas pertenecientes a polígonos industriales. Fuente: (Rodríguez-Gómez y otros, 2022b).

En muchas áreas identificadas como desfavorables, no hay espacio disponible para incluir parques, jardines ni zonas arboladas. En estos casos, el papel de la vegetación integrada en la construcción, como los techos verdes y jardines verticales, juegan un papel importante para intervenir las zonas que requieran mayor urgencia.

Con la herramienta que se ha desarrollado, se han mejorado sustancialmente las funcionalidades que ofrecen otras herramientas previas. En la Tabla 4.2 se puede ver una comparativa entre la herramienta propuesta para la detección automática y los principales trabajos relacionados encontrados.

Tabla 4.2: Comparativa de trabajos para determinar las zonas más desfavorables (UHI)

	Temperatura	clasificación del terreno	Ubicación óptima	Metodología automática	Disponibilidad del software
(Chen y otros, 2014)	✓	✓	-	-	-
(Li y otros, 2016)	✓	✓	-	-	-
(Du y otros, 2017)	✓	✓	-	-	-
(Estoque y otros, 2017)	✓	✓	-	-	-
(Sun y otros, 2018)	✓	✓	-	-	-
(Li y Zhou, 2019)	✓	✓	-	-	-
(Masoudi y Tan, 2019)	✓	✓	-	-	-
(Asadi y otros, 2020)	✓	✓	-	-	-
(Rahaman y otros, 2022)	✓	✓	-	-	-
(Fernández y otros, 2015)	-	-	-	✓	-
(Bartesaghi-Koc y otros, 2019)	-	✓	-	✓	-
(Velázquez y otros, 2019)	-	-	✓	-	-
(Nesticò y otros, 2022)	-	-	✓	-	-
URSUS-UHI	✓	✓	✓	✓	código-abierto

4.3. Predicción de temperatura en la superficie terrestre (LST)

Se propone una novedosa metodología que permite ofrecer tres funcionalidades necesarias para la sostenibilidad urbana en esta línea de actuación: a) la extracción de las características urbanas del entorno cercano en las zonas más desfavorables de las ciudades, b) el entrenamiento de modelos de aprendizaje supervisado para la predicción de la LST conociendo las anteriores características, y c) la simulación de las temperaturas que se obtendrían en diferentes escenarios urbanos modificando las características relativas a la vegetación del entorno cercano (tipo, distancia, y cantidad).

4.3.1. Propuesta metodológica

Para una mejor comprensión de la metodología que se procede a detallar, esta ha sido dividida en una serie de subsecciones: la Subsección 4.3.1.1 describe la metodología para la extracción de las características del entorno urbano cercano, y la Subsección 4.3.1.2 describe la fase de modelizado que se ha llevado a cabo en el trabajo de investigación.

Las fases finales, evaluación y despliegue, donde la metodología comprueba y aplica los resultados atendiendo a los objetivos iniciales, se recogen en 4.3.2 y 4.3.3.

El resto de esta sección describe el procedimiento para obtener los datos necesarios para entrenar los modelos y un estudio del procesamiento realizado para seleccionar el modelo más apropiado y predecir LST.

4.3.1.1. Preparación de datos

Los datos necesarios para extraer conocimiento sobre la influencia del entorno cercano en la temperatura de la superficie terrestre se extraen principalmente a partir de tres fuentes:

- Las *imágenes Sentinel-2*, descritas en la Subsección 2.1.1, se utilizan para calcular los valores de NDVI de los píxeles de ciudad de estudio siguiendo los pasos descritos en la Subsección 2.2.2. A partir de la clasificación de los elementos urbanos en base al NDVI disponible en la Tabla 2.2, el sistema procede a la segmentación automática de elementos urbanos categorizados como masas de agua (píxeles contiguos asociados a agua con una superficie mayor a 5 km^2) y de las áreas urbanas sin vegetación. Según los expertos, realizar la segmentación del agua de las ciudades y de los espacios sin vegetación a partir de imágenes Sentinel-2, ofrece resultados de mayor calidad que a partir otras fuentes de información. Los expertos consideran que la presencia

de masas de agua en el entorno cercano de puntos urbanos es una característica importante para tener en cuenta, por lo que es necesario identificar las masas de agua. Por otro lado, las zonas sin vegetación son de especial interés porque serán las mejores candidatas en las que intervenir para reducir las temperaturas. Por lo tanto, es crucial comprender mejor cómo las características del entorno cercano, consideradas de relevancia por los expertos, afectan a la temperatura en estas áreas sin vegetación.

- Las *nubes de puntos LiDAR*. Como se ha descrito en la Subsección 2.1.2, las imágenes LiDAR clasifican a cada punto en una categoría de elemento urbano. La información relevante para esta investigación corresponde con las etiquetas LiDAR con los valores de 3 a 6. Los valores de 3 a 5 se utilizan para vegetación baja, moderada y densa, respectivamente, y el valor 6 indica la presencia de edificios. Además, la etiqueta LiDAR con valor 2 representa al suelo, elemento necesario para calcular la altura de los edificios sin tener en cuenta la altura del suelo.

- Las *imágenes Landsat-8* sirven para calcular los valores de LST en diferentes puntos de la ciudad siguiendo los pasos descritos en la Subsección 2.2.3. Como se mencionó anteriormente, el valor de la LST será la variable objetivo para realizar predicciones utilizando modelos de regresión.

La Figura 4.9 resume las fuentes de datos descritas así como las transformaciones realizadas para obtener el conjunto de datos final. Los detalles de estas transformaciones se describen a continuación.

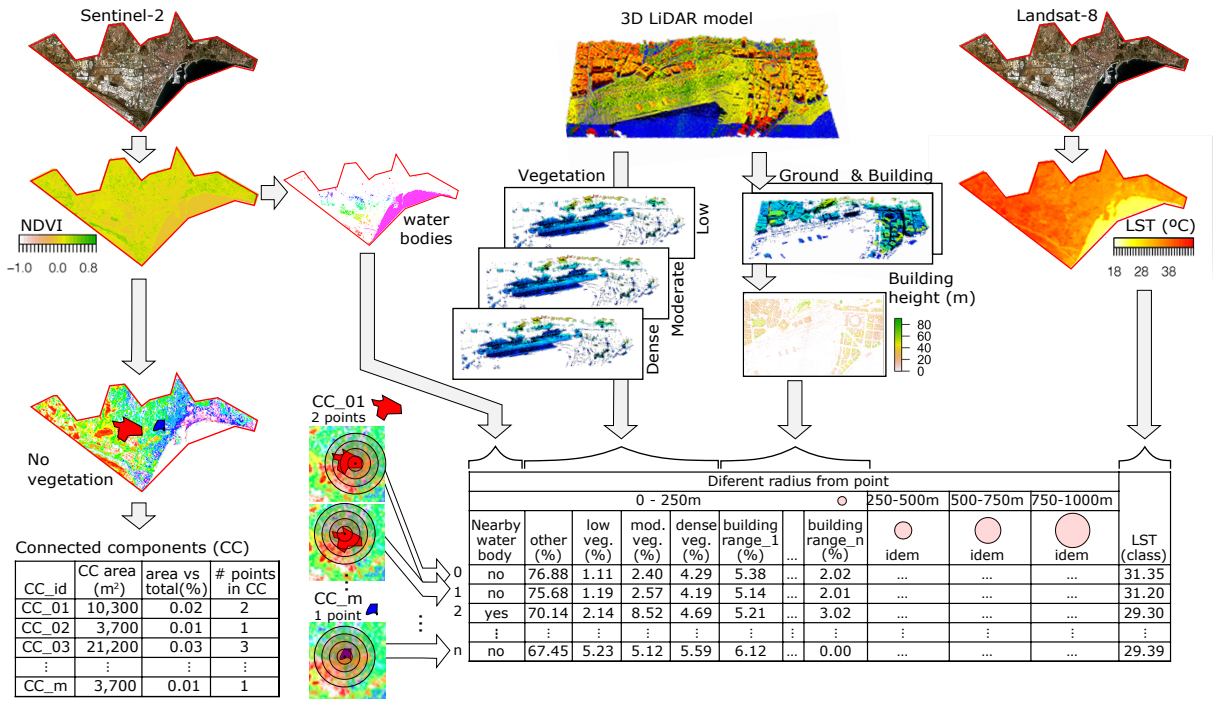


Figura 4.9: Metodología propuesta para extraer las características del entorno cercano.

Cada punto de estudio de la ciudad de interés será clasificado en una de las siguientes categorías: vegetación *baja*, *moderada* y *densa*, *edificio*, *agua*, u *otros*. En base a este criterio, el suelo urbano será clasificado en 6 categorías. La categoría *otros*, corresponde a elementos urbanos como carreteras, puentes, cables y otros elementos urbanos no incluidos en las clases anteriores.

En cuanto a los píxeles categorizados como *edificio*, la metodología propuesta incluye un paso para calcular nuevas subcategorías teniendo en cuenta sus alturas. La altura se calcula usando un modelo de superficie digital normalizado (NDSM) (Rao y otros, 2022), obtenido a partir de la imagen LiDAR filtrada con los puntos de clasificación (2: Suelo, 6: Edificio). El NDSM permite medir las alturas de los objetos en relación con el mismo suelo, ya que elimina el suelo de los diferentes edificios. Si es necesario, los expertos del dominio pueden discretizar este atributo numérico en múltiples rangos, de acuerdo con su conocimiento sobre la ciudad estudiada.

Los puntos de interés cuyo entorno cercano necesita ser estudiado son un factor determinante para un proceso de minería de datos exitoso. Los expertos en el dominio están interesados en cómo el medio ambiente afecta a la temperatura de las áreas sin vegetación. Por esa razón, esta metodología sigue un enfoque adaptativo para seleccionar puntos en el mapa representativos de cada zona sin vegetación. El etiquetado de componentes conexas (CCL), descrito en la Subsección 2.2.4, ha sido la técnica seleccionada para determinar las zonas no vegetadas en las ciudades. De esta manera,

se identifican áreas separadas sin vegetación y se seleccionan puntos aleatorios de las mismas. El número de puntos aleatorios tomados de cada área, depende de su extensión con respecto al área total. Así, se seleccionan pocos puntos para áreas pequeñas, mientras que a medida que aumenta el área, se seleccionan más puntos. Elegir el mismo número de puntos para cada área independientemente de su tamaño es ineficiente porque el entorno de muchos puntos en un área pequeña es el mismo y se puede representar con pocos puntos (o incluso con un solo punto).

Para cada uno de los puntos seleccionados al azar, se calculan algunas variables que describen su entorno cercano:

- la presencia (o no) de una masa de agua, y
- los porcentajes de suelo (respecto al área urbana total) ocupados por:
  - vegetación baja, moderada y densa,
  - edificios (que se pueden discretizar opcionalmente en rangos), y
  - otros elementos urbanos.

Estas variables se calculan en diferentes anillos de distancia: menor que 250 m, de 250 m a 500 m, de 500 m a 750 m, y finalmente, desde 750 m hasta 1 km. Todo el proceso de preparación se representa en la Figura 4.9.

4.3.1.2. Entrenamiento de modelos

El proceso llevado a cabo para la selección del mejor modelo para predecir LST, que es una de las principales contribuciones de este trabajo, sigue un entrenamiento de modelos con optimización de hiperparámetros (Bischl y otros, 2023).

Los algoritmos de regresión seleccionados para la generación de modelos, han sido descritos en la Subsección 2.3.3.1. Dado que todos estos algoritmos tienen diferentes *hiperparámetros* que controlan su ejecución (y por lo tanto pueden afectar su rendimiento), es necesario determinar cuál es la configuración óptima para cada uno.

Así, se realiza un ajuste fino de los hiperparámetros, realizando una búsqueda en maya para obtener la configuración que proporcione los mejores resultados para cada algoritmo. Esos resultados se determinan a partir de los errores MAE y RMSE obtenidos. En la Tabla 4.3 se pueden observar los hiperparámetros que se analizan para cada algoritmo, así como los valores que se examinan. Una estrategia de validación cruzada de 10 veces se repite 3 veces para garantizar la estabilidad de los resultados para cada modelo generado.

Tabla 4.3: Configuración de los parámetros de los algoritmos

Algoritmo	Hyperparametros	Valores analizados
RT	Profundidad máxima	$\{1, 2, 3, 4, 5\}$
ANN	Número de neuronas en la capa oculta	$\{2, 4, 8, 16, 24\}$
SVR	Coste	$\{0{,}25, 0{,}5, 1, 2, 4\}$
XGB	Profundidad × Número de rondas	$\{3, 7\} \times \{250, 500, 750\}$
Bagging	Número de iteraciones	$\{10, 20, 25, 35, 40\}$
RF	Número de árboles	$\{50, 100, 150, 200\}$

4.3.2. Resultados

Las fases de preparación y modelizado de datos, que forman parte de la metodología de extracción de datos presentada en la Sección 2.3, son genéricas y se pueden utilizar siempre que haya datos disponibles. En este apartado mostramos los resultados evaluados por expertos en un caso concreto y aplicando la metodología en una ciudad real como ejemplo. Por lo tanto, se muestra su proceso para validar el conocimiento antes de ser incorporado al sistema experto para apoyar la toma de decisiones.

4.3.2.1. Preparación de datos: ejemplo de validación

Se han recopilado y procesado los datos de la ciudad de Málaga, en el sur de España, siguiendo la descripción dada en la Subsección 4.3.1.1. Cada punto en el área de interés está definido por la clase que se le asigna; es decir, cada punto se etiqueta como agua, vegetación (baja, moderada o densa), edificio (incluida su altura) u otro. Cada punto corresponde a un píxel en la imagen LST, y su dimensión se define como un cuadrado de 10 m × 10 m.

A continuación, se procede a la segmentación de áreas no vegetadas utilizando un algoritmo de etiquetado de componentes conexas (CCL) y se identifican 8 754 componentes conexas. Los puntos que se seleccionan, garantizan que todas las componentes conexas estén representadas. Los componentes más pequeñas, con un área de 100 m^2, solo necesitan un punto para su estudio, mientras que las componentes más grandes, con un área de 1 865 300 m^2 se inspeccionan tomando 308 puntos. Finalmente, se obtienen 16 678 puntos representativos de las zonas no vegetadas de la ciudad de Málaga.

La temperatura de la superficie terrestre (LST) se conoce para cada punto de esos 16 678 puntos seleccionados, y se calculan 28 variables adicionales que describen su entorno cercano. Se definen cuatro regiones concéntricas de 250 m de ancho en el rango de 1 km, y se calculan siete variables para cada región. Esas variables indican la presencia de una masa de agua, el porcentaje de elementos no categorizados (otros), el porcentaje para cada

tipo de vegetación (baja, moderada o densa) y el porcentaje de dos tipos de edificaciones en base a la altura. En este caso, los expertos decidieron discretizar la altura de los edificios en dos rangos: más pequeños y más grandes que 24 m. Se consideraron otras opciones como prescindir de la discretización o utilizar una discretización en más grupos, pero en ambos casos resultaron alternativas menos adecuadas (por aumentar el error o por aumentar la complejidad sin disminuir el error).

4.3.2.2. Modelado: ejemplo de validación

Una vez creado el conjunto de datos con la información de las características del entorno cercano de puntos relevantes de la ciudad de Málaga (definidos por 16 678 ejemplos descritos por 28 atributos), se utiliza para entrenar diferentes modelos. Los algoritmos y configuraciones utilizados se enumeran en la Tabla 4.3 y los mejores resultados obtenidos se describen en la Tabla 4.4. Cada algoritmo (y configuración) se ha evaluado repitiendo 3 veces una validación cruzada de 10 veces y, por lo tanto, se han realizado 30 experimentos. Se ha realizado una validación estadística utilizando el test no paramétrico de Wilcoxon (Rey y Neuhäuser, 2011) sobre esos 30 experimentos para detectar diferencias significativas.

Tabla 4.4: Algoritmos, mejor configuración, y métricas

Algoritmo	Parámetros seleccionados.	MAE	RMSE	R^2	
RT	5	1.02 ±0.020	1.33 ±0.030	0.55 ±0.024	⊖
ANN	24	0.86 ±0.030	1.14 ±0.050	0.68 ±0.006	⊖
SVR	4	0.69 ±0.020	0.98 ±0.040	0.76 ±0.017	⊖
XGB	7×5	0.47 ±0.010	0.67 ±0.030	0.89 ±0.011	⊖
Bagging	25	1.00 ±0.020	1.30 ±0.030	0.58 ±0.018	⊖
RF	*50*	***0.44 ±0.010***	***0.63 ±0.020***	***0.9 ±0.007***	

Las configuraciones que obtienen las mejores métricas de evaluación (valores promedio y desviación estándar) para cada algoritmo se muestran en la Tabla 4.4. Hay una excepción con Random Forest ya que el modelo seleccionado usa 50 árboles en lugar de 200 . Los resultados logrados por 200 árboles son ligeramente mejores que los logrados por la versión con 50 árboles, pero las diferencias en precisión no son estadísticamente significativas. Por lo tanto, se ha optado por la versión con 50 árboles, ya que su complejidad computacional es mucho menor tanto en la fase de aprendizaje como en la de predicción.

Como se puede apreciar en la Tabla 4.4, el modelo más preciso es el inducido por Random Forest con 50 árboles. Para comprender la relevancia de esta afirmación, se han realizado pruebas estadísticas comparando este modelo con los demás, demostrando que el error

obtenido por el resto de modelos es significativamente mayor (p-valor $\ll 0,05$). Esto se representa con el símbolo $\ominus$.

4.3.2.3. Evaluación y ajuste del modelo: ejemplo de validación

Teniendo en cuenta los resultados obtenidos en la fase de modelado, y que el modelo ha sido entrenado con muchas variables (28), se ha decidido realizar un ajuste haciendo una selección de características para conservar los atributos más relevantes y descartar los que no lo son. Así, el modelo se puede adaptar para que sea menos complejo y se mantenga su capacidad predictiva.

Después de realizar una selección de características (Li y otros, 2017), utilizando un enfoque específico para el algoritmo de regresión seleccionado, se calcula la importancia general de cada variable (ver la Figura 4.10). Se puede observar que las variables más relevantes son aquellas que afectan a las regiones más próximas al punto de interés, en el rango entre 0 y 250 m. El porcentaje de edificios altos (por encima de 24 m) es particularmente relevante. Por otro lado, la presencia de las masas de agua no parece afectar a las diferencias en LST observadas entre diferentes localizaciones dentro de la ciudad estudiada.

El modelo final seleccionado que ha sido incorporado en el sistema experto ha tenido en cuenta la importancia de las características previamente calculadas. Solo se han considerado los 10 atributos más relevantes. Las diferencias en precisión entre modelos con 10 o más de 10 atributos son insignificantes para los requisitos de los expertos.

4.3.3. Sistema experto incluyendo conocimiento

Esta sección describe URSUS-LST[4], la aplicación implementada para utilizar el conocimiento adquirido durante la aplicación del proceso de minería de datos. El modelo puede cambiar de una ciudad a otra, aunque puede ser necesario repetir la fase de modelizado para ajustarlo mejor a una nueva ciudad. También podría repetirse la fase de modelizado para evaluar si algún otro algoritmo o configuración, ofrece mejores resultados en nuevas ciudades. Los pasos seguidos por la aplicación para predecir la temperatura de la superficie terrestre (LST) en un punto específico o para simular el nuevo valor de LST cuando se realizan cambios en el entorno cercano, son los mismos independientemente del modelo cargado.

[4]Urban sustainability for LST prediction

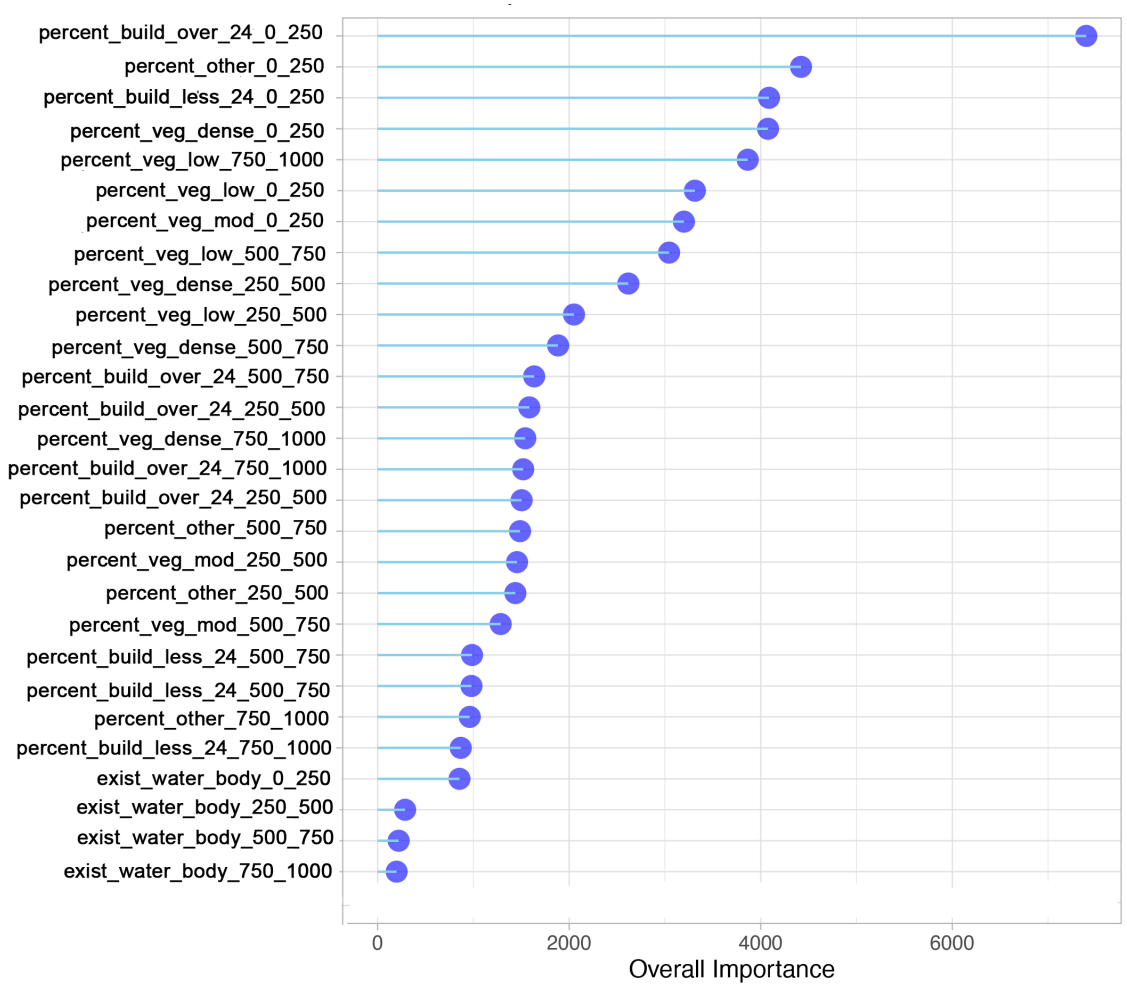


Figura 4.10: Importancia de las variables utilizadas durante un proceso de selección de características. El modelo utilizado es un Random Forest con 50 árboles. Los atributos se refieren al porcentaje de diferentes tipos de elementos urbanos en cuatro anillos de distancia.

4.3.3.1. Tecnologías para la implementación

Para este trabajo de investigación, la tecnología seleccionada ha sido `R`, ya que cuenta con las bibliotecas necesarias para procesar imágenes LiDAR, Landsat y Sentinel, y ofrece un entorno de trabajo que facilita la conversión de los resultados obtenidos de los procesos de minería de datos en herramientas web.

Las principales tecnologías de `R` utilizadas para desarrollar la herramienta se pueden agrupar según la fase del proceso de minería de datos en la que se han utilizado de forma intensiva.

Para la consecución de los objetivos, se han utilizado librerías de `R`, como `ráster` (Hijmans, 2020) y `lidR` (Roussel y otros, 2020) para la preparación de datos. `lidR` permite la

extracción de edificios y vegetación del entorno cercano, mientras que `ráster` facilita el cálculo de la altura de los edificios, del NDVI o de la LST, y la extracción de características del entorno cercano en anillos de 250 m. El paquete `dplyr` (Wickham y otros, 2020) se ha usado para procesar los datos para entrenar los modelos.

Las bibliotecas de `R` para el aprendizaje automático, como el paquete `caret` (Kuhn, 2020), se han utilizado para entrenar diferentes modelos mediante diferentes algoritmos y configuraciones, así como para evaluar los modelos.

Finalmente, el paquete `shiny` (Chang y otros, 2021) se ha utilizado para el desarrollo de la herramienta web para realizar predicciones y simulaciones de LST. Se ha configurado un servidor Linux para el despliegue y configuración de la aplicación web.

4.3.3.2. Predicción y simulación

Siguiendo con el ejemplo de validación del apartado anterior, las Figuras 4.11 y 4.12 presentan capturas de pantalla para ilustrar la aplicación. Se corresponden a la selección de un punto en el centro de la ciudad de Málaga (España), y en concreto, en la zona de la Catedral. El uso de la aplicación se describe a continuación.

Seleccionando el punto de estudio. Área de interés (a) El mapa muestra al usuario una serie de cuadrados azules que representan el área cubierta por las imágenes LiDAR donde se puede seleccionar el punto de estudio. A continuación, el sistema dibuja un círculo naranja de radio 1 km y comienza a extraer las características urbanas del entorno cercano. A partir del catálogo de imágenes LiDAR de la ciudad, la aplicación realiza el recorte del área circundante al punto de estudio para 1 km a su alrededor.

Extracción de edificios (b) A partir del modelo LiDAR del área de interés se obtiene una imagen que contiene únicamente los puntos LiDAR clasificados como edificios. El siguiente paso consiste en calcular el modelo de alturas normalizado eliminando el suelo del modelo para medir a todos los edificios sobre la misma base, y se calcula una capa ráster de los edificios categorizados por altura. La aplicación obtiene 2 imágenes ráster filtrando las alturas en la imagen con todos los edificios para determinar los edificios con alturas superiores e inferiores a 24 m.

Extracción de vegetación (c) Para el cálculo de las áreas con vegetación escasa, moderada y densa se sigue un proceso similar al descrito anteriormente para la segmentación de edificios. En este caso, la aplicación calcula 3 imágenes LiDAR filtrando

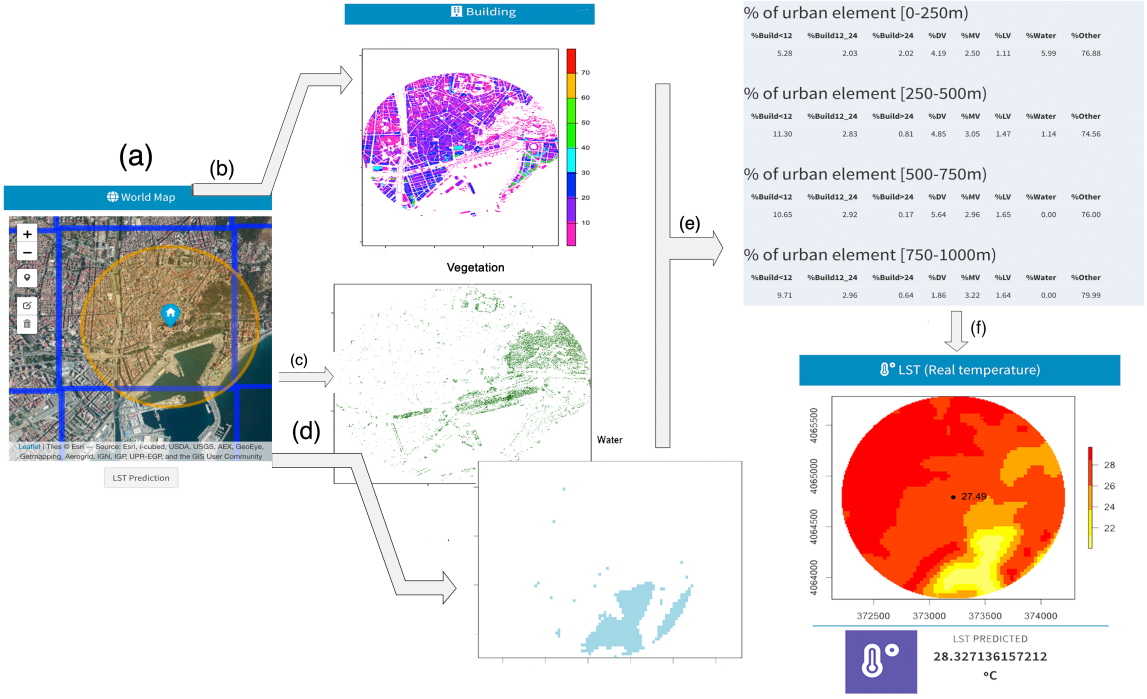


Figura 4.11: Capturas de pantalla con parte de la información generada por la herramienta para predecir la LST: (a) Selección del punto de estudio. Recorte del área de interés a 1 km. (b) Extracción de edificios. (c) Extracción de vegetación. (d) Existencia de masas de agua. (e) Cálculo del porcentaje de área de las características urbanas extraídas previamente. (f) Modelo de predicciones.

por puntos pertenecientes a vegetación baja, moderada y densa, y luego se obtienen sus modelos ráster 2D.

Existencia de masas de agua cercanas (d) Para obtener las áreas cubiertas por agua se utiliza una imagen ráster con el terreno clasificado según el NDVI de la ciudad como se ha descrito en 4.3.1.1. Esta imagen ráster se recorta a un radio 1 km desde el punto de estudio, al igual que se hizo con las imágenes del catálogo LiDAR. El siguiente paso consiste en realizar un etiquetado de componentes conexas sobre la capa ráster de agua previamente calculada, ya que es necesario determinar la extensión para verificar si existen o no masas de agua.

Calcular el porcentaje de área de las características urbanas (e) La división de cada capa ráster (agua, edificios o vegetación) en intervalos 250 m se realiza fácilmente usando la función `distanceFromPoints(point,raster)` de la biblioteca `raster`. La herramienta puede crear una máscara que represente el anillo de corte, eliminando puntos que están a distancias que no son de interés. La máscara de distancia (anillo de corte) se puede aplicar en diferentes capas ráster y sirve para calcular el área de cada característica

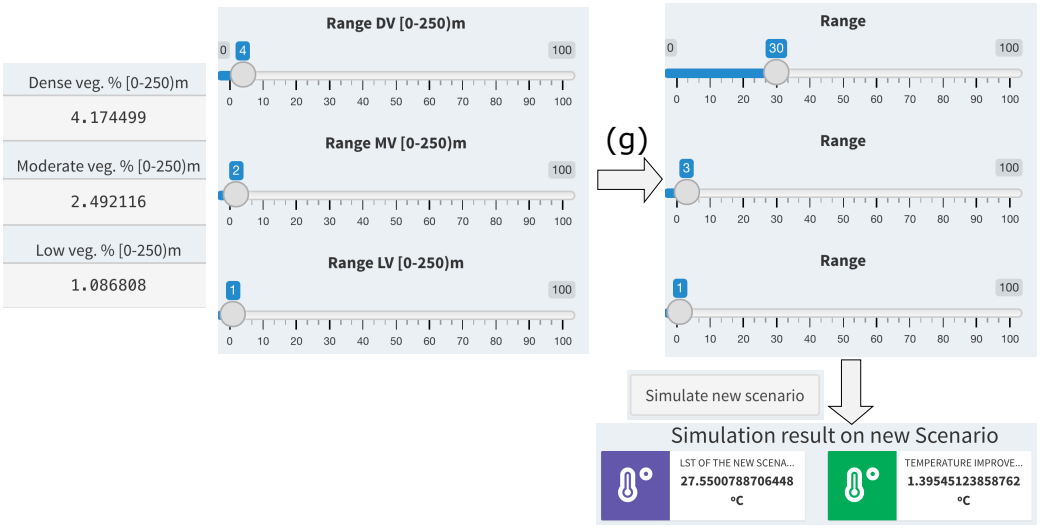


Figura 4.12: Capturas de pantalla con parte del proceso para simular nuevos escenarios (g). El nuevo valor LST se estima cambiando algunas variables en el entorno cercano del punto de estudio.

del entorno cercano.

Predicciones del modelo (f) Una vez calculados los valores de las variables que necesita el modelo propuesto para realizar las predicciones (presencia de masas de agua, %vegetación, %edificios y %otras características) para cada intervalo de 250 m, el modelo ya puede realizar predicciones de la LST en el punto seleccionado y la herramienta lo muestra. El modelo seleccionado es un Random Forest con 50 árboles entrenados con las 10 variables más relevantes como se describe en la Subsección 4.3.2.2. La Figura 4.11 muestra el ejemplo completo para el caso real que se está usando.

Simulación de nuevos escenarios en función de cambios en el entorno (g) El sistema es un medio para simular cómo una modificación de los porcentajes de vegetación en los intervalos podría afectar la temperatura de la superficie terrestre (LST). En esta nueva situación, la herramienta muestra la reducción de temperatura sobre el escenario real. La Figura 4.12 muestra la información real que describe el entorno de un punto y presenta los controles que un usuario puede configurar para configurar un nuevo escenario. En la parte inferior se muestra la temperatura en el nuevo escenario después de cambiar los porcentajes de vegetación.

4.4. Conclusiones

A continuación se detallan las principales conclusiones alcanzadas en este capítulo sobre herramientas inteligentes para asesorar en el enverdecimiento de ciudades. Empezamos con los desarrollos alcanzados en el ámbito de la **determinación de las zonas más desfavorables por el efecto UHI**.

Se ha desarrollado una metodología basada en el análisis de imágenes satélites para identificar qué partes de una ciudad tienen mayor necesidad de vegetación debido a las temperaturas excesivas y la falta de espacios verdes. La metodología se ha aplicado a dieciséis ciudades españolas con diferentes características donde se determinaron las zonas más desfavorables y se representaron en un mapa. Los resultados han sido validados por los expertos. El cálculo del Índice de Zonas Desfavorecidas (DAI), ofrece un indicador que permite definir el grado de deterioro de las zonas urbanas.

En un contexto de recursos limitados para aumentar la vegetación en las ciudades, la detección de áreas desfavorables constituye una herramienta muy interesante para las administraciones públicas, los responsables políticos municipales y los urbanistas, para definir la futura estrategia de planificación de espacios verdes y decidir qué ubicaciones deben abordarse e intervenirse añadiendo UGI de forma más urgente en una ciudad. En este sentido, es clara la utilidad de la información que proporciona el mapa que combina la región del clúster con zonas más desfavorables y DAI, ya que permite detectar aquellos que son realmente mucho más desfavorables, dentro de las zonas críticas.

A continuación, se procede a la descripción de las conclusiones alcanzadas como resultado del trabajos de investigación relacionado con la línea de investigación relativa a la **predicción y simulaciones de LST analizando las características del entorno urbano cercano**.

Se presenta una metodología novedosa para predecir la temperatura de la superficie terrestre (LST) y simular los efectos de agregar nueva infraestructura verde urbana sobre las temperaturas en áreas potencialmente afectadas por el efecto isla de calor urbano. La incorporación de características urbanas, además de los valores LST, da como resultado un modelo más preciso para predecir temperaturas. Esta metodología es extensible a cualquier ciudad y minimiza los pasos iniciales de preprocesamiento, lo que la hace eficiente y eficaz para los planificadores urbanos.

La metodología permite determinar y entrenar el modelo óptimo en la mencionada tarea de estimar los valores LST a partir de elementos urbanos y cómo se verán afectados si se introducen nuevos elementos urbanos.

El análisis de las características del entorno urbano en diferentes intervalos de distancia

para diferentes puntos de estudio ha demostrado que el impacto del factor de distancia en la temperatura LST se puede calibrar para proporcionar a los expertos del dominio conocimientos novedosos y relevantes. Además, la herramienta propuesta permite ajustar la región de estudio en intervalos de 250 m en un rango de 0 a 1 km, lo que permite a los usuarios determinar la distancia, la cantidad y el tipo óptimos de infraestructura verde urbana para reducir las temperaturas.

En la metodología propuesta, la identificación de las variables más importantes para predecir LST a partir de los elementos urbanos del entorno asegura que los modelos resultantes tengan en cuenta aquellas características con mayor impacto en la temperatura. A través de nuestro análisis, hemos encontrado que la cantidad de vegetación y edificios en menos de 250 m son las variables más influyentes en la determinación de la temperatura.

Además, podemos enumerar como contribución significativa de este trabajo, el desarrollo de una aplicación web de código abierto para planificadores urbanos que les permite simular el impacto en la temperatura a partir de la instalación de nuevos UGIS.

Según los expertos en la materia, el uso de la herramienta en diferentes ciudades tendría un impacto global en la mitigación del efecto UHI en las ciudades, y por tanto en la mejora de la sostenibilidad urbana y en la lucha contra el cambio climático. La principal limitación es la dependencia de la precisión del modelo de las imágenes LiDAR y satelitales utilizadas para entrenarlo.

En conclusión, la metodología y las herramientas propuestas tienen el potencial de facilitar el trabajo de los planificadores urbanos al proporcionar predicciones más precisas de LST y simular el impacto de agregar infraestructura verde urbana. Como tal, representa una valiosa contribución al desarrollo de entornos urbanos sostenibles. En trabajos futuros se podría explorar la integración de otros factores, como la contaminación del aire, para mejorar aún más la precisión de las predicciones de temperatura y el impacto de la infraestructura verde en la calidad ambiental urbana.

5

Conclusiones y trabajos futuros

En este capítulo se describen las principales conclusiones obtenidas en el trabajo de investigación y se presentan una serie de posibles trabajos futuros para continuar la investigación en las temáticas abordadas.

Como resultado de los trabajos desarrollados se han propuesto nuevos modelos y procedimientos basados en técnicas de minería de datos que permiten tratar y resolver diferentes problemas relacionados con la sostenibilidad urbana.

5.1. Modelos para la generación y consumo de energía en entornos urbanos

En el ámbito de la caracterización y predicción de la generación de energía en entornos urbanos, se ha desarrollado e implementado una herramienta siguiendo un proceso de minería de datos guiado por el conocimiento de los expertos, URSUS-PV, para poder hacer una **selección automática de emplazamientos para instalaciones fotovoltaicas en entornos urbanos**. Esta herramienta aporta una serie de ventajas a las que existían previamente, y que se enumeran a continuación:

- La herramienta realiza estimaciones de generación de energía con sistemas

fotovoltaicos a corto y largo plazo de cualquier área urbana, lo que permitirá detectar las áreas óptimas para ubicar instalaciones de este tipo.

- Mediante un proceso de minería de datos, guiado por el conocimiento de los expertos, se automatiza un proceso manual realmente complejo: cálculos fotovoltaicos a corto y largo plazo, extracción de características de edificios, etc.
- No se centra en tejados individuales, sino en áreas urbanas, pudiendo extenderse a calles, barrios, complejos de edificios, etc. Además, permite seleccionar el tipo de tejados de interés en las áreas de estudio en función de su orientación e inclinación.
- Gracias al uso de tecnologías LiDAR, es fácilmente adaptable para añadir nuevos tipos de ubicaciones (además de los tejados) en las que realizar las instalaciones: aparcamientos, avenidas, rotondas, etc.
- La herramienta es de código abierto y gratuita. Puede ser utilizada por cualquier persona o entidad: administraciones públicas, cooperativas, empresas distribuidoras de sistemas fotovoltaicos o particulares.
- Es extensible a cualquier ciudad ya que URSUS-PV utiliza imágenes LiDAR que permiten determinar con precisión los elementos urbanos de una zona. Esas imágenes son fáciles de descargar para muchos países y la disponibilidad está aumentando. Los datos meteorológicos necesarios en el sistema (como la temperatura o la humedad) también son comunes y suelen ser registrados por organismos nacionales. Por lo tanto, URSUS-PV es muy flexible en la incorporación de nuevas áreas urbanas y dicho proceso es sencillo.

El uso de está herramienta puede contribuir a acelerar el ritmo en el que se sustituye el uso de energías contaminantes por energías renovables como la solar en las ciudades, lo que repercutiría de forma directa en la mejora de la sostenibilidad urbana y en la lucha contra el cambio climático.

Además del contexto anterior, se ha abordado otro tema muy importante para aprovechar de forma más eficiente la energía eléctrica en las ciudades: la **caracterización y modelización de los perfiles de consumo eléctrico en hogares**. Cabe destacar diferentes contribuciones en esta temática:

- Hay una simplificación de las etapas de preprocesamiento de la información. No se precisa de un paso de normalización de los datos de consumo eléctrico, por lo que tampoco se necesita abordar la compleja tarea de desnormalización. Al mismo tiempo, evitando la normalización, todos los patrones de consumo se pueden comparar directamente, y se identifican patrones con formas comunes pero diferentes rangos. Otro efecto positivo de prescindir del preprocesamiento de datos es que

no es necesario segmentar los datos. Típicamente, este problema se ha abordado segmentando el conjunto de datos con distintos criterios, como sontipos de días (laborables, festivos, etc.), meses o estaciones. La metodología propuesta permite identificar los patrones independientemente de estos aspectos temporales, al no tener que estudiar por separado conjunto de datos segmentados.

- Tratamiento de grandes volúmenes de datos, ya que incluso con millones de perfiles de carga diarios, gracias a los algoritmos de *big data* que aprovechan los recursos computacionales, ha sido posible no perder información por tener que reducir el tamaño del conjuntos de datos por problemas computacionales.
- Integración del conocimiento de expertos en el dominio, desde las fases iniciales. En la primera fase de la metodología es fundamental delimitar un rango para el espacio de búsqueda en el que se debe encontrar el número de clústeres. Para ello, es fundamental determinar las medidas a optimizar, considerando el dominio; en el caso de los perfiles de consumo estos han sido el MAE y el número de clústeres irrelevantes de los modelos. Los resultados experimentales y el asesoramiento y proceso de validación llevado a cabo por los expertos, sugieren que esta metodología es una herramienta valiosa.
- Se ha propuesto un nuevo algoritmo para determinar el número óptimo de clústeres basado en la estabilidad de diferentes métricas registradas durante la fase de modelizado. Esta propuesta permite incorporar el conocimiento de los expertos, independientemente del dominio del problema.
- Se ofrece de forma gratuita, como código abierto, una implementación de esta metodología.

La metodología propuesta es fácilmente extensible a problemas de cualquier dominio donde los algoritmos de agrupamiento sean de aplicación, y los expertos pueden utilizar su conocimiento para reconducir la calidad de los modelos. El procedimiento propuesto para identificar un número apropiado de clústeres también es extensible a otros campos.

Como posibles **trabajos futuros en el ámbito de la modelización de la energía en entornos urbanos**, la herramienta URSUS-PV se podría extender para hacerla funcional con otros elementos urbanos además de los tejados, como pueden ser carreteras, aparcamientos, fachadas, etc., lo que podría favorecer el incremento en la disponibilidad de ubicaciones posibles para instalaciones fotovoltaicas al considerar un mayor abanico de emplazamientos urbanos.

Otra de las tareas futuras que se ha planificado es hacer extensible la herramienta a un mayor número de ciudades.

El algoritmo ISAC, que se ha propuesto como base para determinar el número óptimo de clústeres relacionados con el consumo eléctrico, puede ser aplicado a otros dominios. Para ello, como posibles trabajos de investigación futuros se plantean los siguientes:

- Acotar y mejorar sus posibles limitaciones.
- Comparar exhaustivamente sus capacidades con los resultados alcanzados por otros métodos (manuales o automáticos).
- Estudiar otras posibles curvas y parámetros en función del dominio de aplicación.
- Validación en nuevos contextos.

5.2. Herramientas inteligentes para asesorar en el enverdecimiento de ciudades

En el ámbito de las herramientas inteligentes para el asesoramiento en el enverdecimiento de ciudades, se han desarrollado e implementado una serie de herramientas y modelos siguiendo un proceso de minería de datos guiado por el conocimiento de los expertos. Para **determinar de forma automática las zonas más desfavorables por el UHI en entornos urbanos**, se ha desarrollado URSUS-UHI. Esta herramienta aporta una serie de ventajas a las herramientas que existían previamente, y que se enumeran a continuación:

- Se trata de una herramienta sencilla, intuitiva y gratuita que ha sido validada por los expertos tras su aplicación en 16 ciudades. Esto la convierte en una potencial aliada para la mejora de la sostenibilidad urbana, permitiendo identificar rápidamente las zonas que necesitan una intervención más inmediata.
- Se propone una novedosa metodología que combina técnicas de clustering y un índice (DAI) diseñado por los expertos, para determinar las zonas más desfavorecidas tras realizar un análisis del estado de la vegetación y de las temperaturas de las ciudades.
- La interfaz con el usuario permite dos formas diferentes de recoger las necesidades: una en la que se permite seleccionar manualmente el área de corte de la ciudad de estudio introduciendo las coordenadas del polígono; y otra en la que se puede realizar un corte de forma interactiva sobre el mapa de la ciudad, lo que facilita eliminar los puntos que puedan añadir ruido y delimitar el área a las zonas de interés.
- La herramienta es fácilmente extensible a cualquier ciudad ya que sólo requiere de imágenes Landsat-8 de la misma.

- URSUS_UHI puede utilizarse como punto de partida previo al uso de URSUS_LST, ya que en primer lugar se pueden determinar las zonas más desfavorables por el efecto isla de calor urbano, para a continuación poder simular cuál sería la combinación más adecuada en cuanto a distancia, abundancia, y tipo de infraestructura verde, para conseguir mejores resultados de cara a reducir las temperaturas de dichas zonas.

La combinación de las técnicas de clustering con el DAI ha demostrado su sinergia para determinar las áreas más desfavorables, como se puede deducir de la distribución de los valores de DAI dentro de los diferentes clústeres. En particular, se ha encontrado que un clúster específico generalmente se asigna a los píxeles con el DAI más alto. El resto de los clústeres representan áreas más favorables donde el efecto UHI no es tan fuerte.

La herramienta URSUS-UHI, implementada como resultado de la metodología propuesta, permite a los municipios, planificadores urbanos o administraciones públicas poder intervenir en las zonas más desfavorables reduciendo el efecto UHI. Dado que los recursos para aumentar la vegetación en las ciudades suelen ser limitados, gracias al nuevo indicador desarrollado (DAI) la herramienta ayudará a detectar cuáles de las zonas desfavorables deben priorizarse en cuánto a la incorporación de UGI. El uso de este recurso puede ser un mecanismo útil para aumentar la resiliencia de las ciudades frente al cambio climático.

En el contexto de la **predicción de LST analizando las características del entorno urbano cercano**, se ha desarrollado URSUS-LST. Esta herramienta aporta una serie de ventajas que se enumeran a continuación:

- Se propone una nueva metodología para el entrenamiento de modelos de predicción de LST fácilmente adaptable a cualquier ciudad.
- La herramienta permite la simulación de nuevos escenarios de naturación urbana (cantidad, tipo y distancia) y es capaz de predecir cual sería la mejora de la temperatura de las zonas más afectadas por el UHI tras la incorporación de nuevos UGI. Gracias a esta funcionalidad, los planificadores urbanos podrán cuantificar el impacto que tendría en cuanto a reducción de temperaturas, un cambio en la configuración de los UGI del entorno.
- Se ha realizado un estudio y análisis de las variables del entorno urbano más influyentes en la LST.
- La metodología propuesta minimiza los pasos iniciales de preprocesamiento. Se basa en la información recopilada de imágenes Landsat (para calcular valores LST), imágenes Sentinel (para segmentar masas de agua urbanas) y nubes de puntos LiDAR (para segmentar edificios y vegetación). El siguiente paso sería el

entrenamiento de modelos en la nueva ciudad, y finalmente la herramienta ya estaría preparada para utilizar los nuevos modelos adaptados a las características particulares de las nuevas ciudades, para así poder realizar predicciones y simulaciones de la LST.

- La herramienta que implementa dicha metodología es sencilla, intuitiva, y gratuita.

Calibrar el impacto del factor de distancia en LST es otro beneficio para los expertos del dominio, ya que les aporta conocimientos novedosos y relevantes, porque el sistema permite ajustar la vegetación en diferentes anillos de distancia.

Según los expertos en el dominio, el uso de la herramienta en diferentes ciudades tendría un impacto global en la mitigación del efecto UHI en las ciudades, y por tanto en la mejora de la sostenibilidad urbana.

El análisis de las características más relevantes del entorno urbano cercano para predecir la LST, podría suponer un punto de partida para que los planificadores urbanos puedan tener más control y conciencia sobre qué elementos urbanos deberían limitarse o prohibirse por ser más dañinos y tener una mayor repercusión negativa para la sostenibilidad urbana.

Una mejora futura en esta línea de investigación será introducir al sistema la posibilidad de determinar las mejores ubicaciones en las que realizar instalaciones de infraestructura verde (tejados, aparcamientos, fachadas ...) así como determinar el mejor tipo de infraestructura (techos verdes, parques, jardines verticales ...).

Extender el uso de la herramienta a otras ciudades, tendría un impacto global en la mejora de la sostenibilidad urbana y en la mitigación del cambio climático, por lo que sería otra acción futura interesante para intentar llevar a cabo a corto plazo.

Acrónimos

ANN Red de neuronas artificiales.

API Interfaz que permite la solicitud de una serie de servicios y permite la comunicación entre diferente software mediante peticiones.

GIS Sistema de información geográfica.

ISAC Identificación de áreas estables en curvas.

K-ISAC_TLP Algoritmo para determinar el número óptimo de clústeres para la caracterización de perfiles de consumo eléctrico.

LST Temperatura de la superficie terrestre.

NDVI Índice de Vegetación de Diferencia Normalizada.

PC Perfil de consumo eléctrico.

RF Bosques aleatorios.

SVM Máquina de vectores de soporte.

TLP Perfil de consumo eléctrico característico.

UGI Infraestructuras verdes urbanas.

UHI Islas de calor urbano.

URSUS-LST Herramienta para la estimación de LST analizando el entorno urbano cercano.

URSUS-PV Herramienta para la estimación de energía fotovoltaica en áreas urbanas.

URSUS-TLP Herramienta para la caracterización de perfiles de consumo eléctrico de los hogares.

URSUS-UHI Herramienta para la determinación automática de las zonas en las que es más necesario intervenir debido al efecto UHI.

XGB Incremento extremo de gradiente.

A

Anexos

A.1. Artículos publicados

Tabla A.1: Artículo sobre la localización de instalaciones fotovoltaicas en áreas urbanas.

Título original	Data driven tools to assess the location of photovoltaic facilities in urban areas
Revista	Expert System With Applications (ESWA)
Autores	Francisco Rodríguez-Gómez, José del Campo-Ávila, Marta Ferrer-Cuesta y Llanos Mora-López
Año	2022
Categoría	Computer Science, Artificial Intelligence (Q1)
DOI	`https://doi.org/10.1016/j.eswa.2022.117349`

Tabla A.2: Artículo para la detección de zonas desfavorecidas en diversas ciudades españolas

Título original	Detection of unfavourable urban areas with higher temperatures and lack of green spaces using satellite imagery in sixteen Spanish cities
Revista	Urban Forestry and Urban Greening (UFUG)
Autores	Francisco Rodríguez-Gómez, Rafael Fernández-Cañero, Gabriel Pérez, José del Campo-Ávila, Domingo López-Rodríguez y Luis Pérez-Urrestarazu
Año	2022
Categoría	Forestry (Q1) y Urban Studies (Q1)
DOI	`https://doi.org/10.1016/j.ufug.2022.127783`

A.2. Artículos en revisión

Tabla A.3: Artículo con una metodología para caracterizar patrones de consumo eléctrico en los hogares

Título original	Data-mining-based methodology for characterizing household electricity consumption patterns
Revista	Engineering Applications of Artificial Intelligence (EAAI)
Autores	Francisco Rodríguez-Gómez, José del Campo-Ávila y Llanos Mora-López
Año	2023
Categoría	Computer Science, Artificial Intelligence (Q1)

Tabla A.4: Artículo sobre la predicción de la temperatura en la superficie de zonas urbanas en función del entorno cercano

Título original	Data mining framework for near-environment based land surface temperature prediction
Revista	Expert Systems With Applications (ESWA)
Autores	Francisco Rodríguez-Gómez, José del Campo-Ávila, Luis Pérez-Urrestarazu y Domingo López-Rodríguez
Año	2023
Categoría	Computer Science, Artificial Intelligence (Q1)

Tabla A.5: Artículo sobre la identificación de áreas donde instalar infraestructuras verdes que reduzcan la temperatura ambiental.

Título original	Data-driven Identification of Urban Areas in Need of Green Infrastructure for Temperature Reduction
Revista	Applied Soft Computing (ASC)
Autores	Francisco Rodríguez-Gómez, Domingo López-Rodríguez, Luis Pérez-Urrestarazu y José del Campo-Ávila
Año	2023
Categoría	Computer Science, Artificial Intelligence (Q1)

A.3. Herramientas y repositorios

En el siguiente anexo, se adjuntan los enlaces a las herramientas descritas en la tesis, y al código de las mismas.

URSUS-PV

El software URSUS-PV está disponible de dos formas diferentes:

- La aplicación web gratuita para realizar estimaciones del potencial de energía fotovoltaica a corto o largo plazo en áreas urbanas, se encuentra disponible en el siguiente enlace:
  `http://ursus-shiny.uma.es/ursusdm_pv`.
- El código fuente de la herramienta se ha publicado bajo la Licencia Pública General GNU v3.0 y está disponible en el siguiente enlace:
  `https://github.com/ursusdm/ursusdm_pv`.

Las principales tecnologías utilizadas para desarrollar la herramienta propuesta están basadas en el lenguaje `R` (R Core Team, 2019) y algunas bibliotecas `R`.

URSUS-TLP

El código que implementa la metodología propuesta para caracterizar los perfiles de consumo en una ciudad, se encuentra disponible en el siguiente enlace:
`https://github.com/ursusdm/consumptionprofiles`.

El repositorio contiene un script con la metodología propuesta en la que se incluye el algoritmo k-ISAC-TLP para determinar el número óptimo de clústeres (k) para los modelos del mejor algoritmo, así como una serie de instrucciones con los pasos a seguir.

Por razones de privacidad, no ha sido posible compartir los conjuntos de datos de España, pero el conjunto de datos de Irlanda está disponible de forma pública (Commission for Energy Regulation (CER), 2012).

Las principales tecnologías utilizadas para desarrollar la herramienta propuesta están basadas en el lenguaje `python` (Python Core Team, 2019) y algunas bibliotecas `python`.

URSUS-UHI

El código fuente del script URSUS-UHI (publicado bajo la Licencia pública general de GNU v3.0) así como las instrucciones de uso, se encuentran disponibles en el siguiente repositorio:
`https://github.com/ursusdm/URSUS_UHI`.

Las imágenes satelitales Landsat-8 pueden descargarse para la ciudad de estudio de forma gratuita en Earth Explorer (USGS, Departamento del Interior, EE.UU.[1]). Las principales tecnologías utilizadas para desarrollar la herramienta propuesta están basadas en el lenguaje `R` (R Core Team, 2019) y algunas bibliotecas `R` que son gratuitas bajo licencias públicas.

URSUS-LST

Las principales tecnologías utilizadas para desarrollar la herramienta propuesta están basadas en el lenguaje `R` (R Core Team, 2019) y algunas bibliotecas `R`.

El código de URSUS-LST puede dividirse en 3 bloques:

- Script para la extracción de características del entorno urbano cercano de los puntos de estudio.
- Script para el entrenamiento y simulaciones con los modelos.
- Aplicación web para predicciones y simulaciones de LST (Licencia Pública General GNU v3.0)

El código fuente está disponible en
`https://github.com/ursusdm/URSUS_LST_PREDICTION`.

La aplicación web para simular la LST en diferentes escenarios modificando las características del entorno urbano, se encuentra disponible de forma gratuita en el

[1] `http://earthexplorer.usgs.gov`

siguiente enlace:
`http://ursus-shiny.uma.es/ursusdm_lst`.

Las imágenes satelitales Landsat-8 y Sentinel-2 pueden descargarse de forma gratuita en Earth Explorer (USGS, Departamento del Interior, EE.UU.)[2].

Las nubes de puntos LiDAR están disponibles de forma gratuita en el Centro Nacional de Información Geográfica (CNIG, España)[3]).

[2] `http://earthexplorer.usgs.gov`
[3] `https://centrodedescargas.cnig.es`

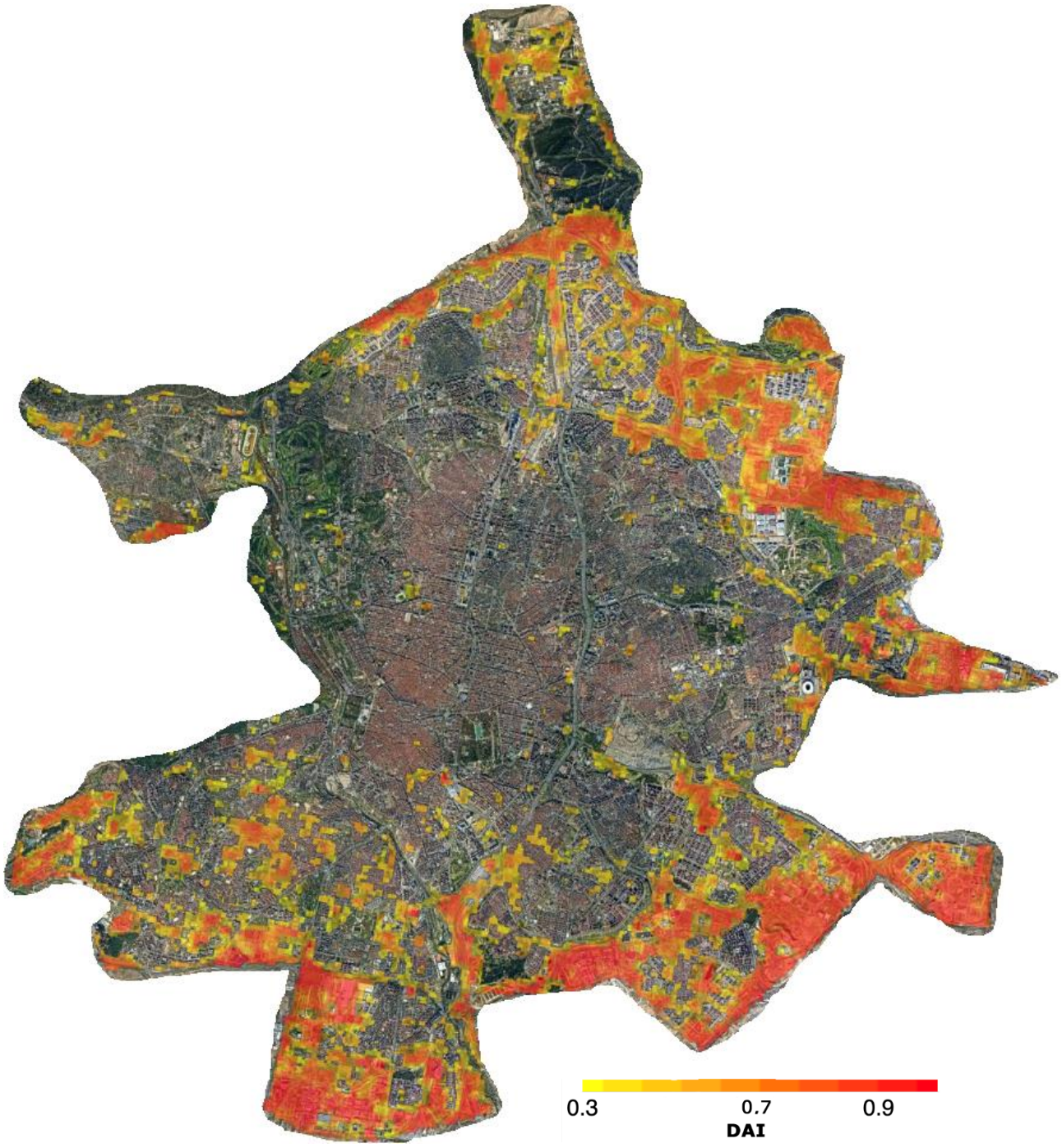
A.4. Zonas más desfavorables en 16 ciudades de España

Automatic detection of unfavourable urban areas with higher temperatures and lack of green spaces using satellite imagery

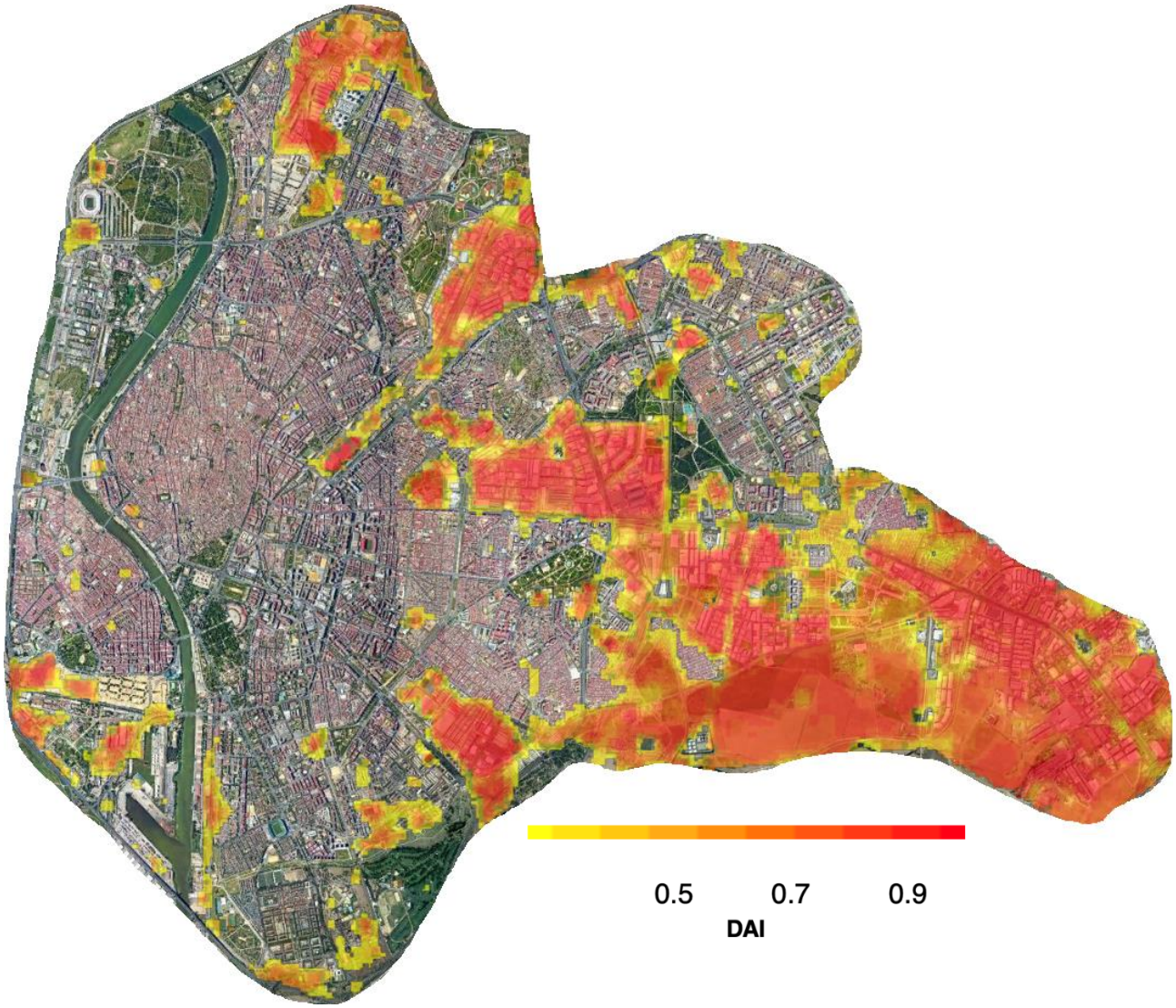
Appendix B: Maps with most unfavourable areas

In this appendix, maps with the most unfavourable areas (corresponding to those that belong to the worst cluster) are presented for each city.

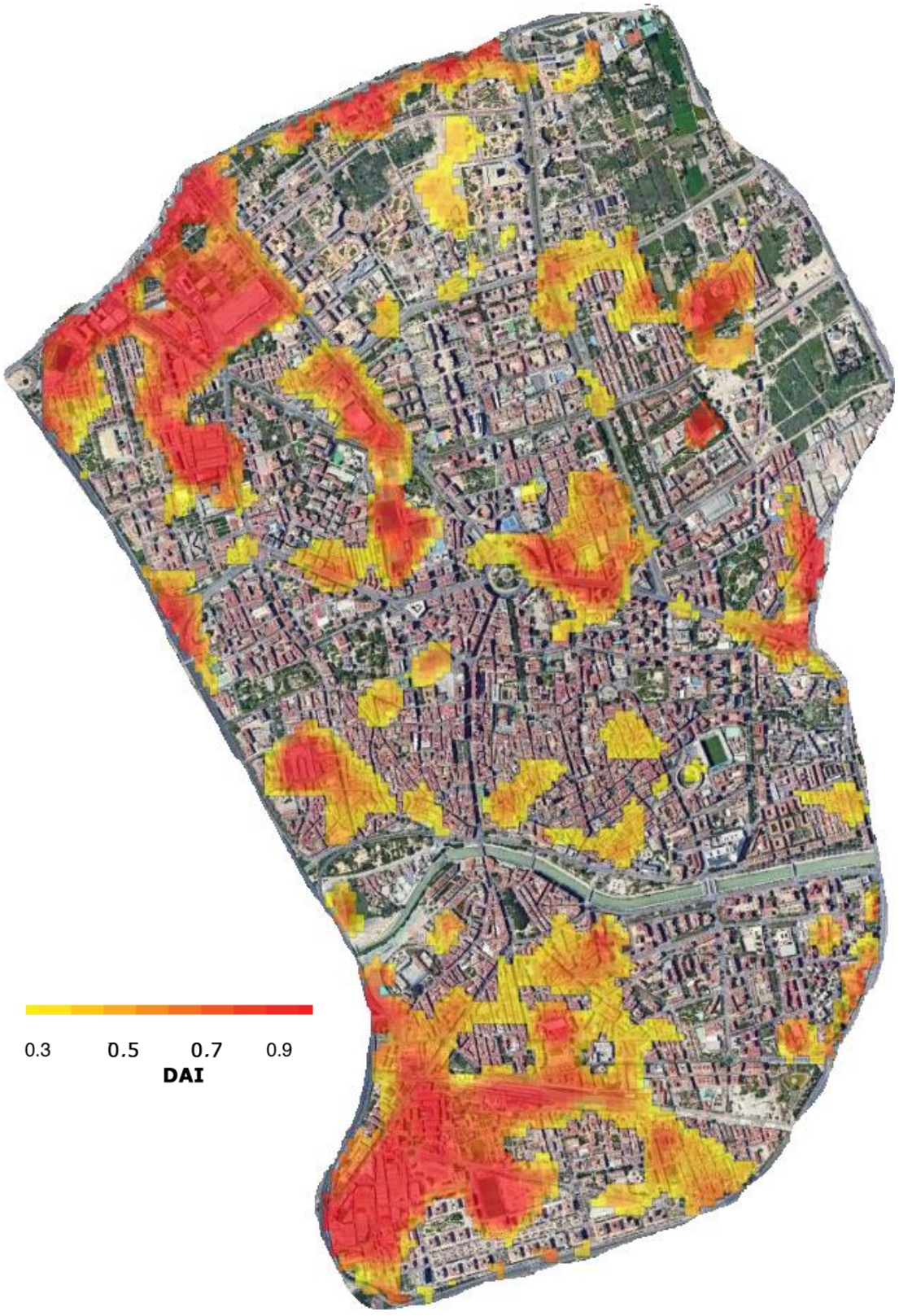
MADRID



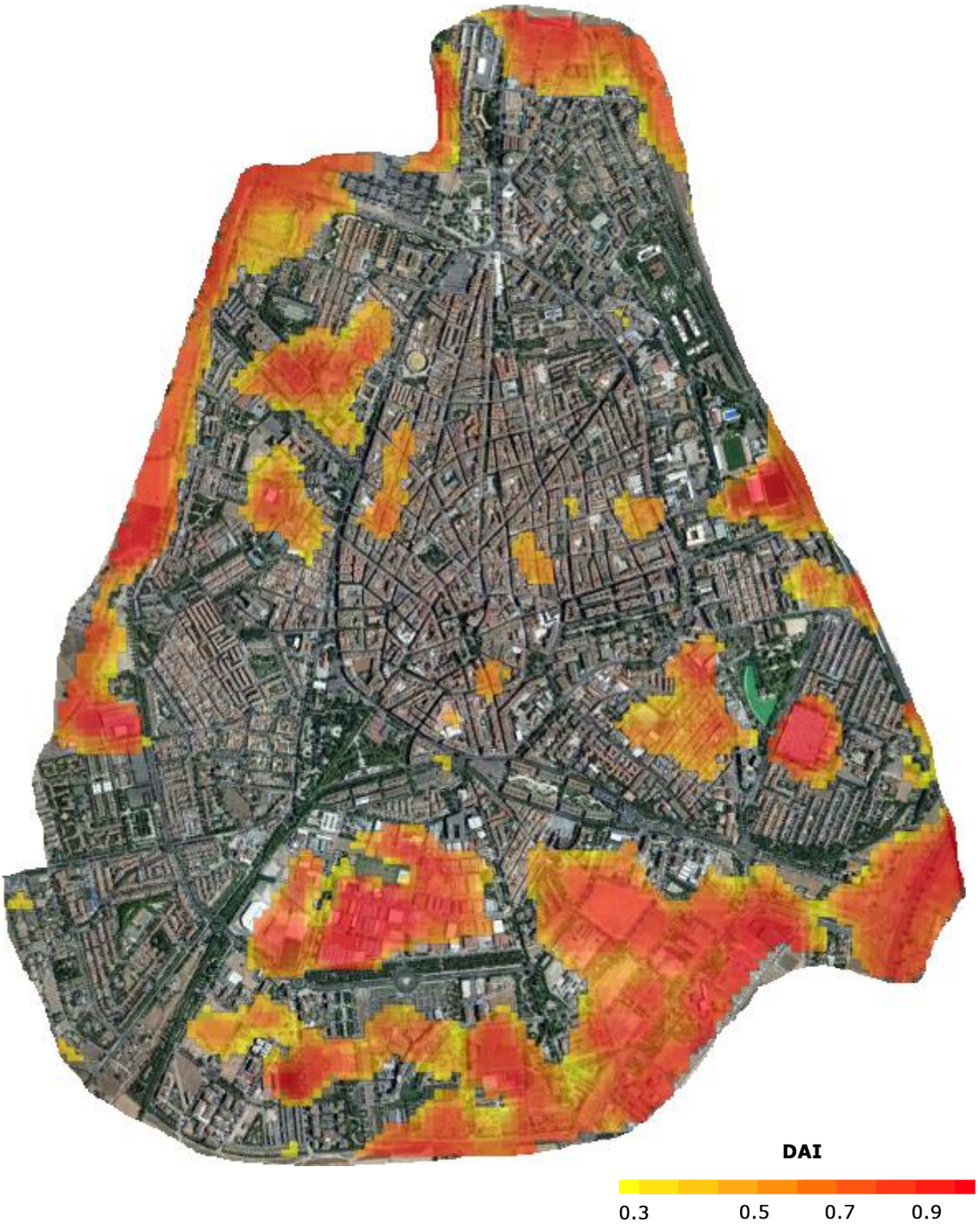
SEVILLE



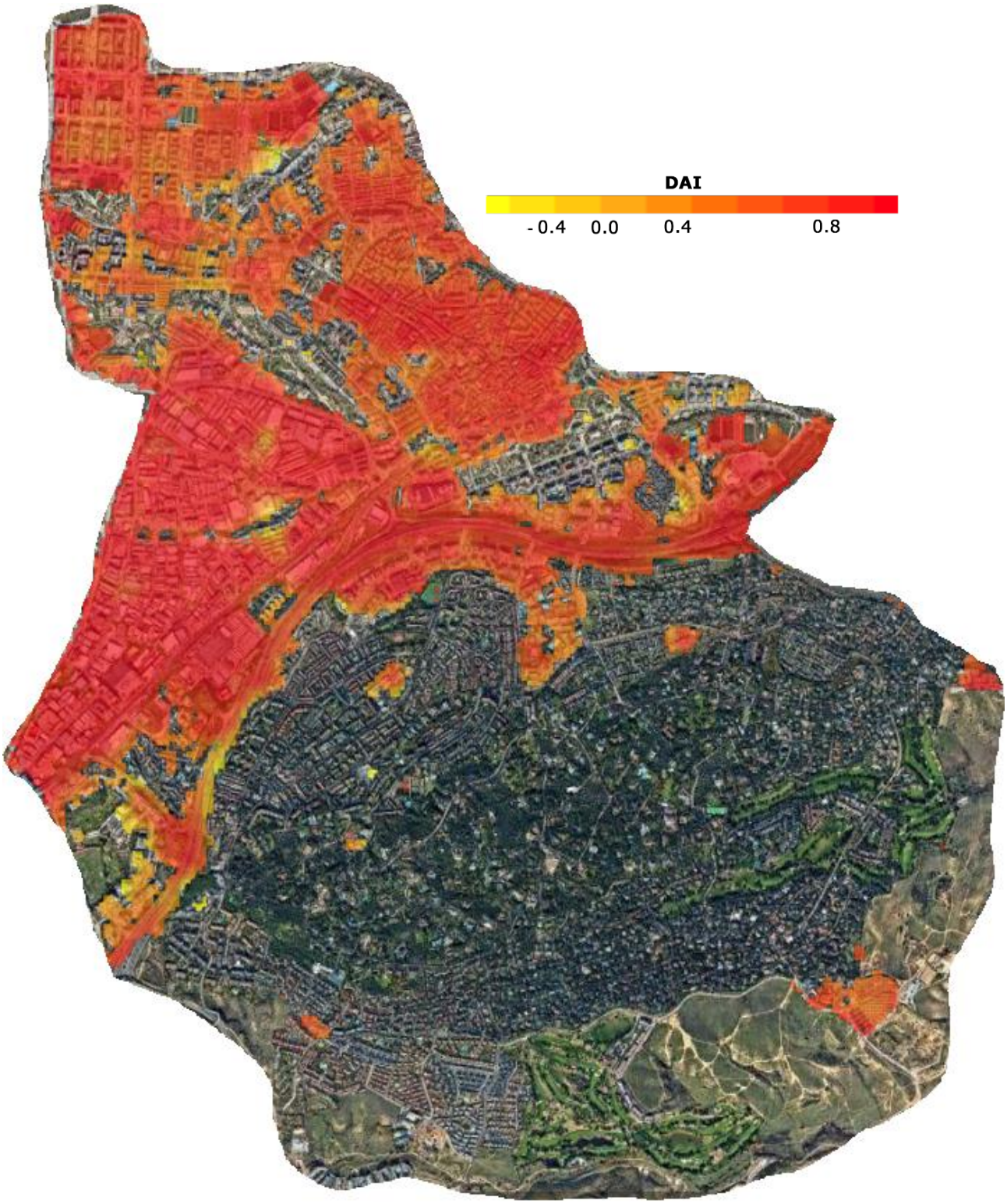
MURCIA



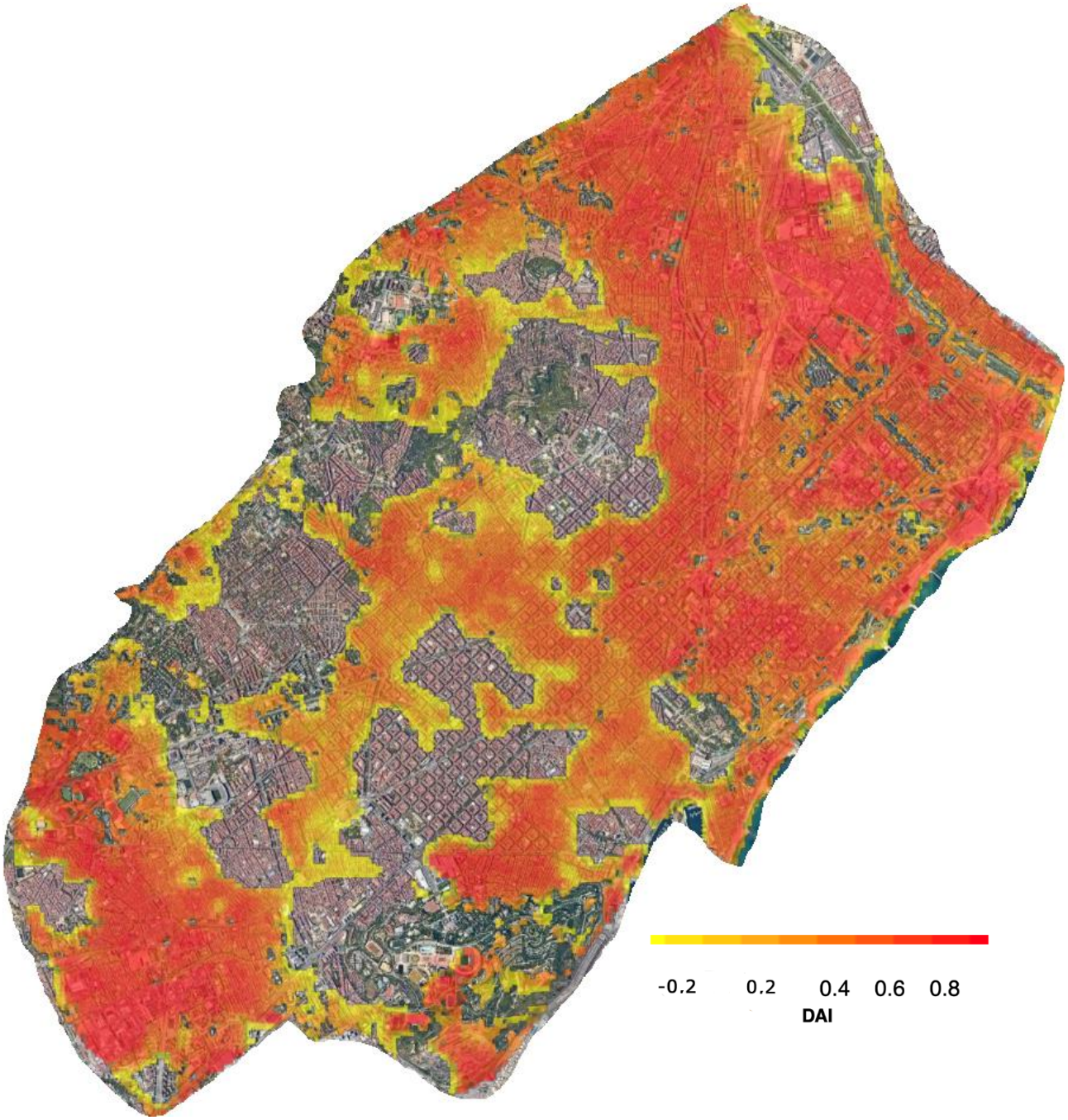
CIUDAD REAL



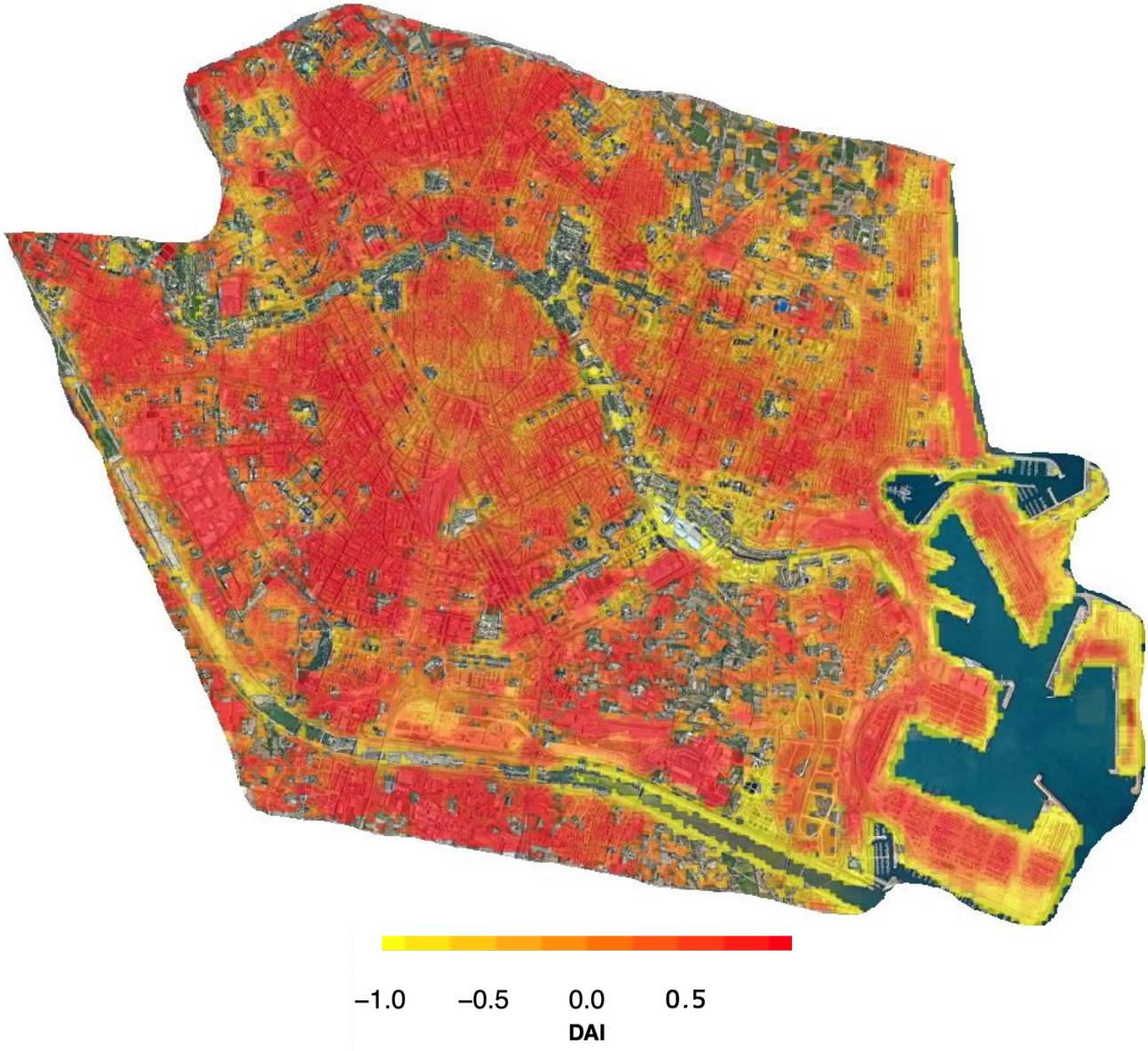
ALCOBENDAS



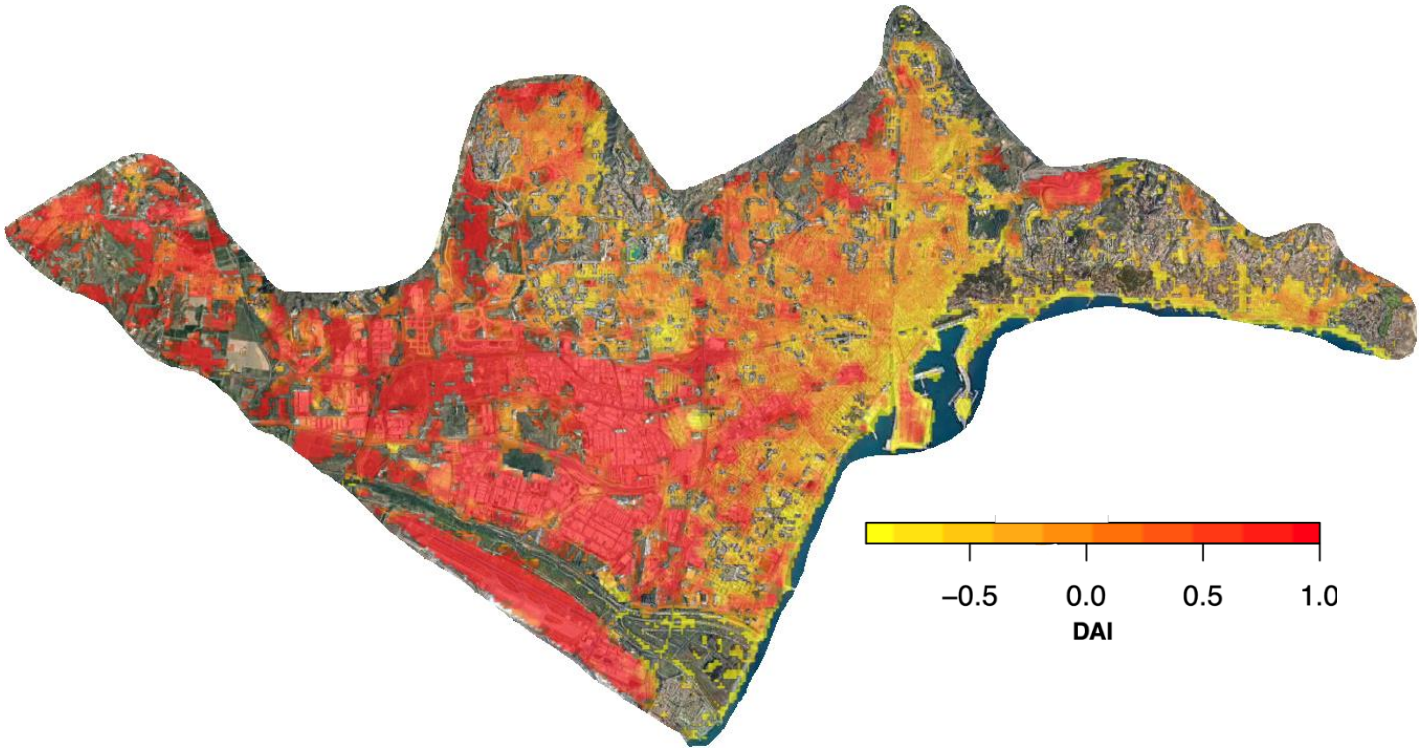
BARCELONA



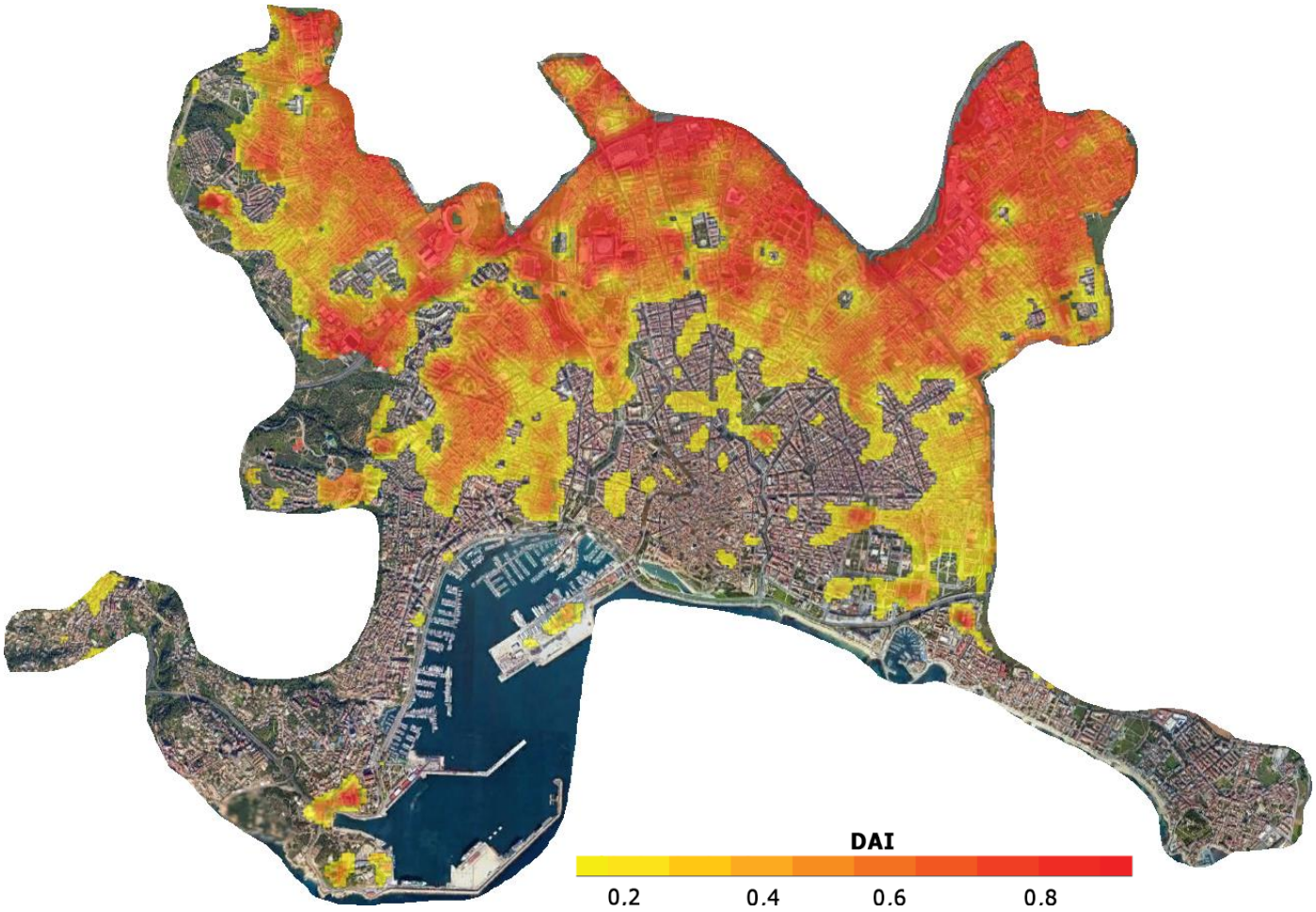
VALENCIA



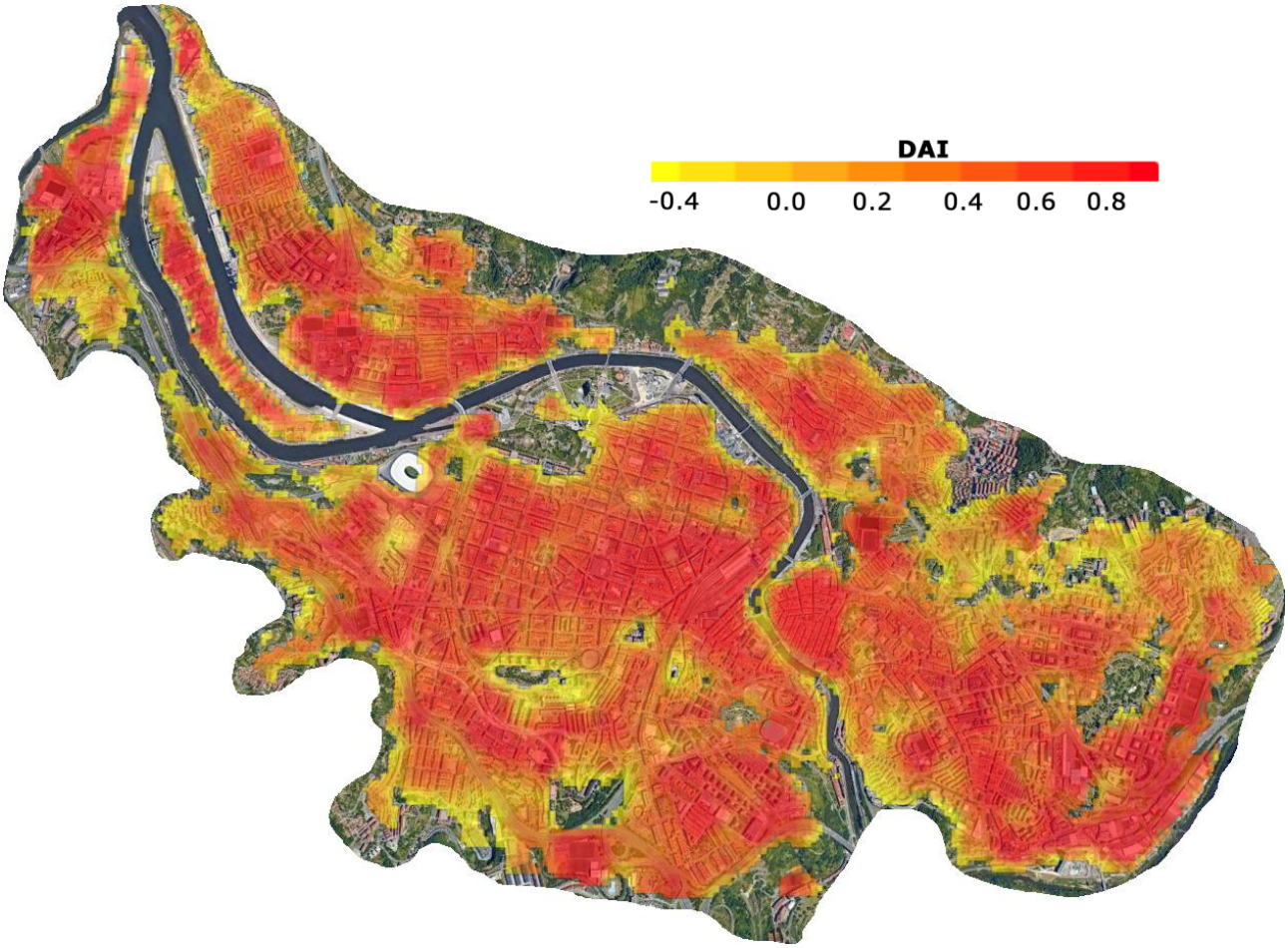
MALAGA



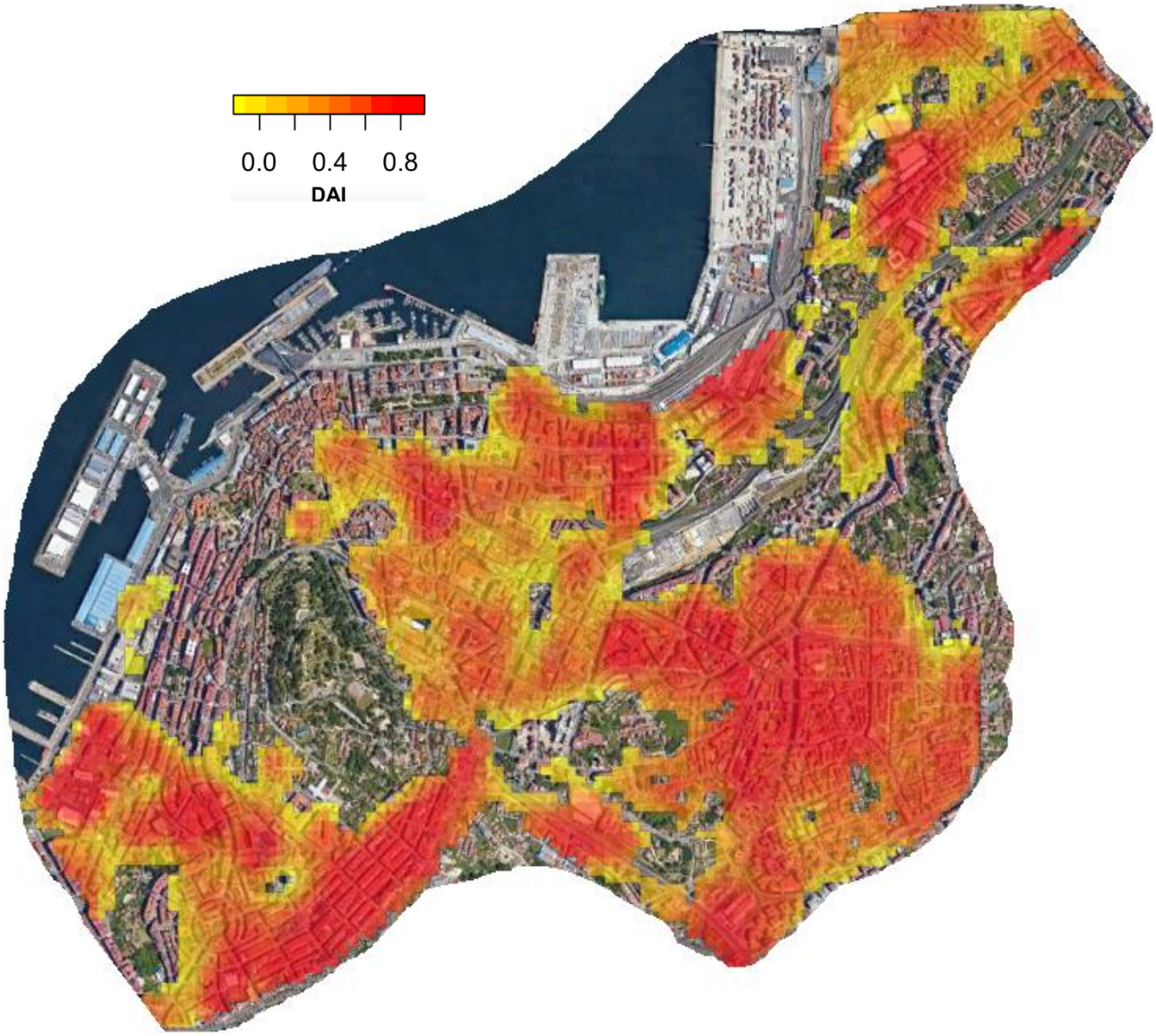
PALMA DE MALLORCA



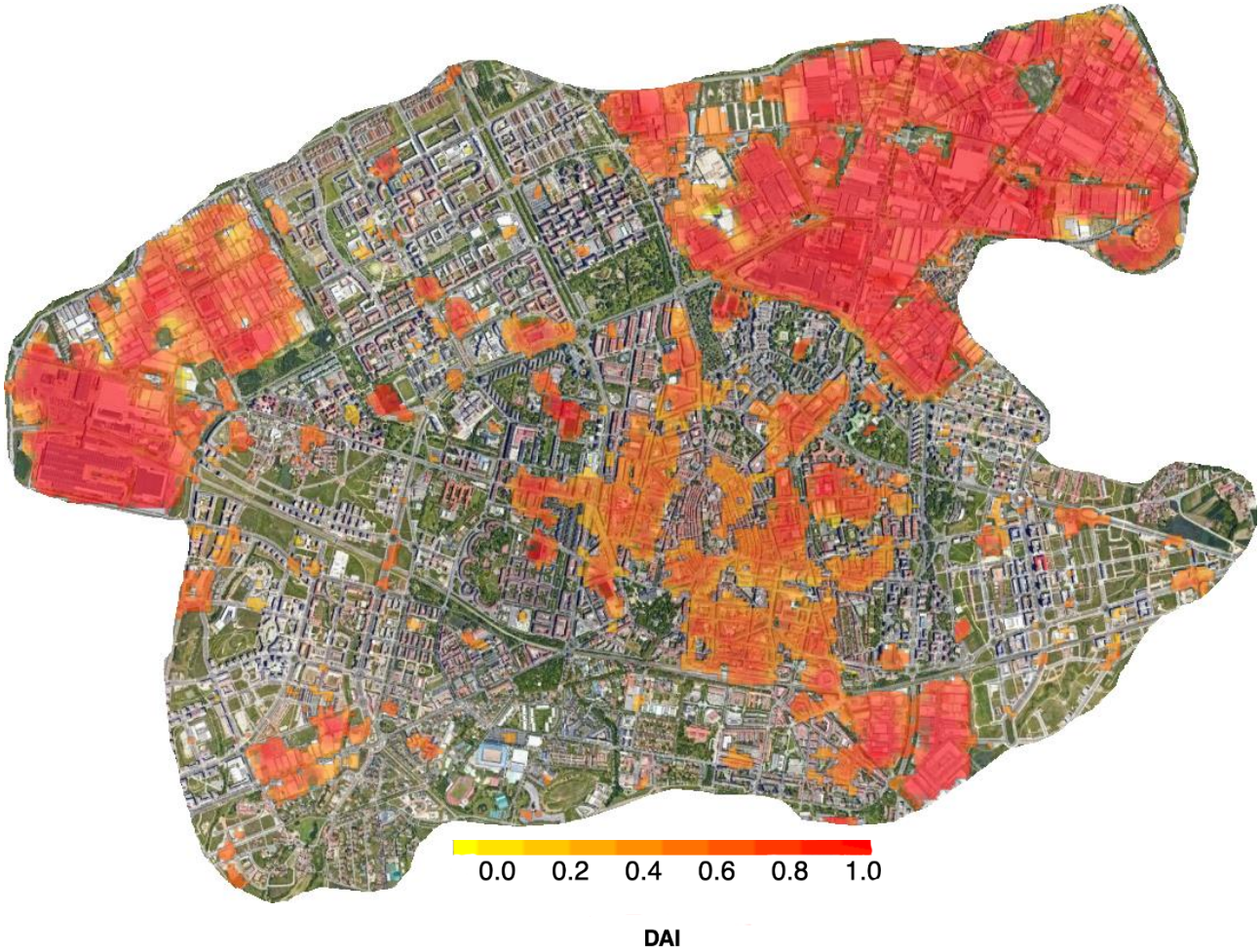
BILBAO



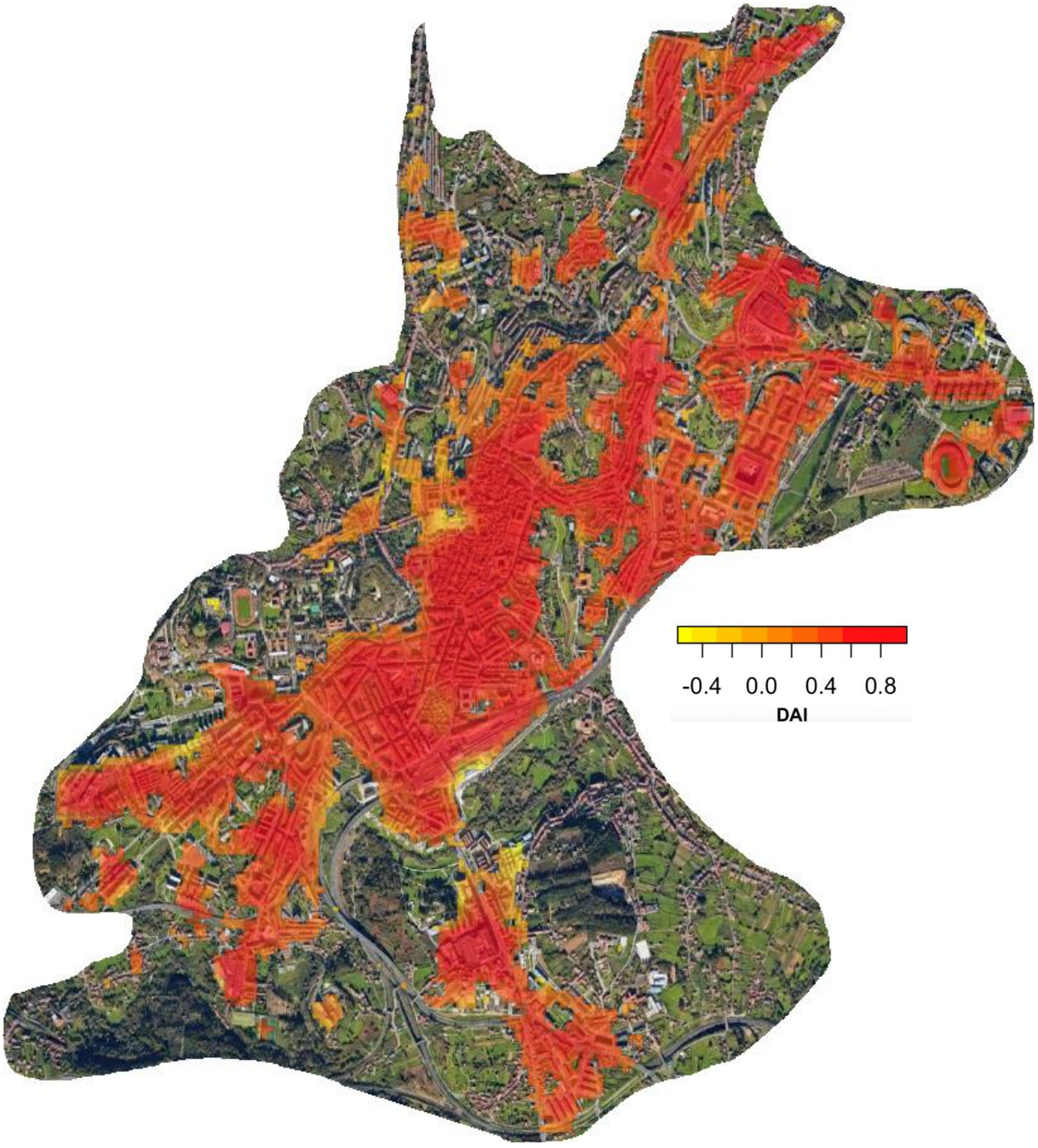
VIGO



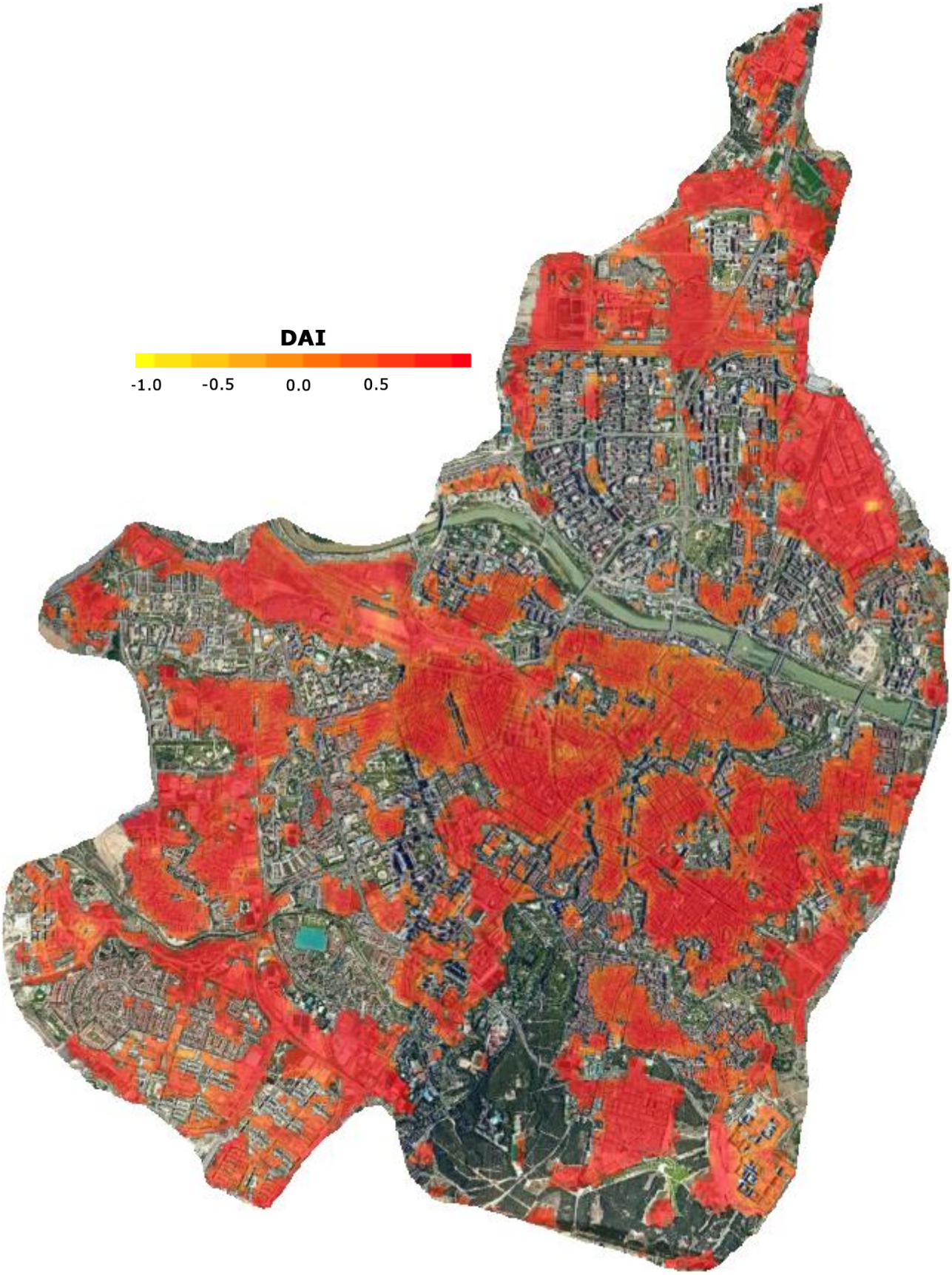
VITORIA



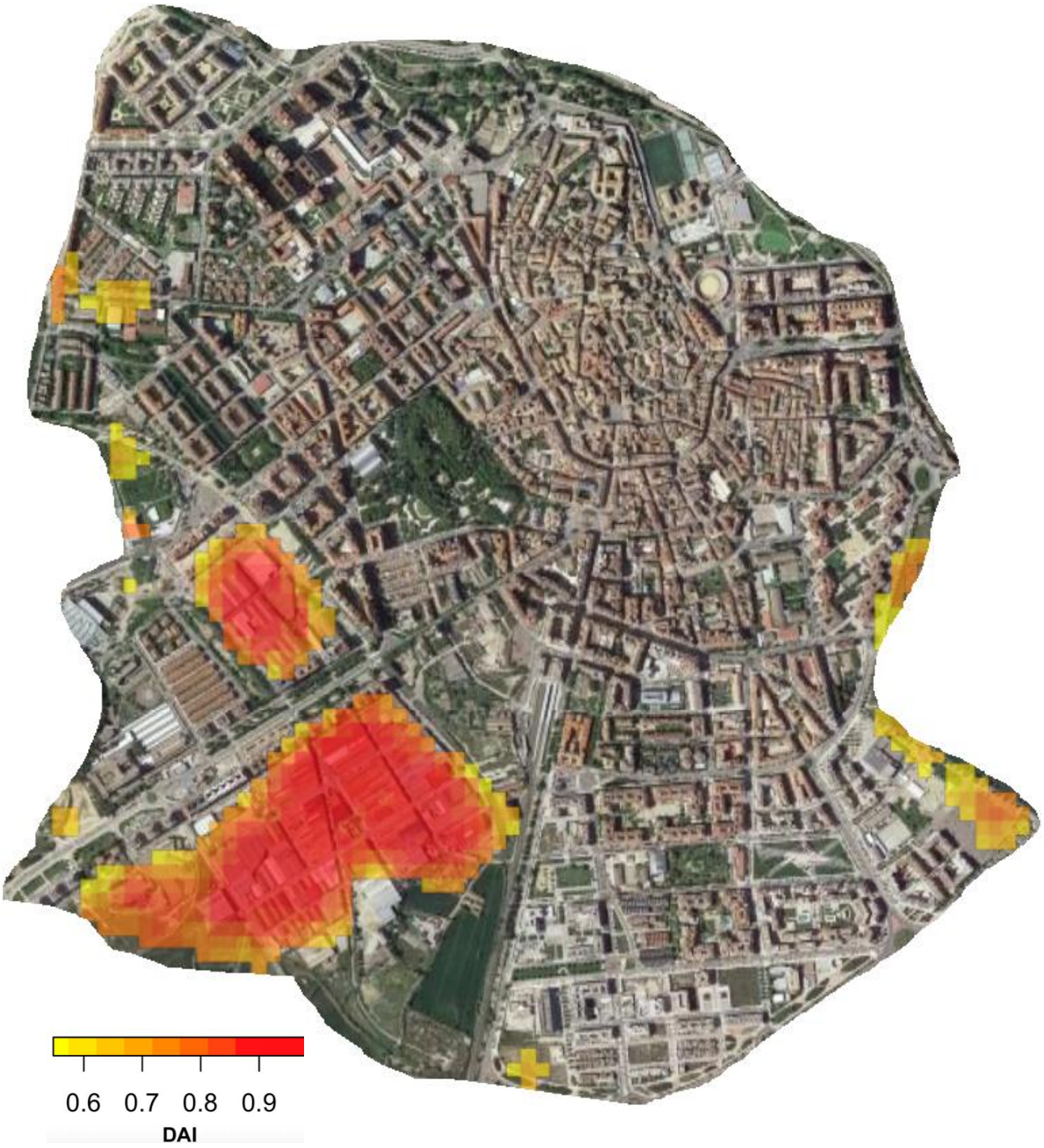
SANTIAGO DE COMPOSTELA



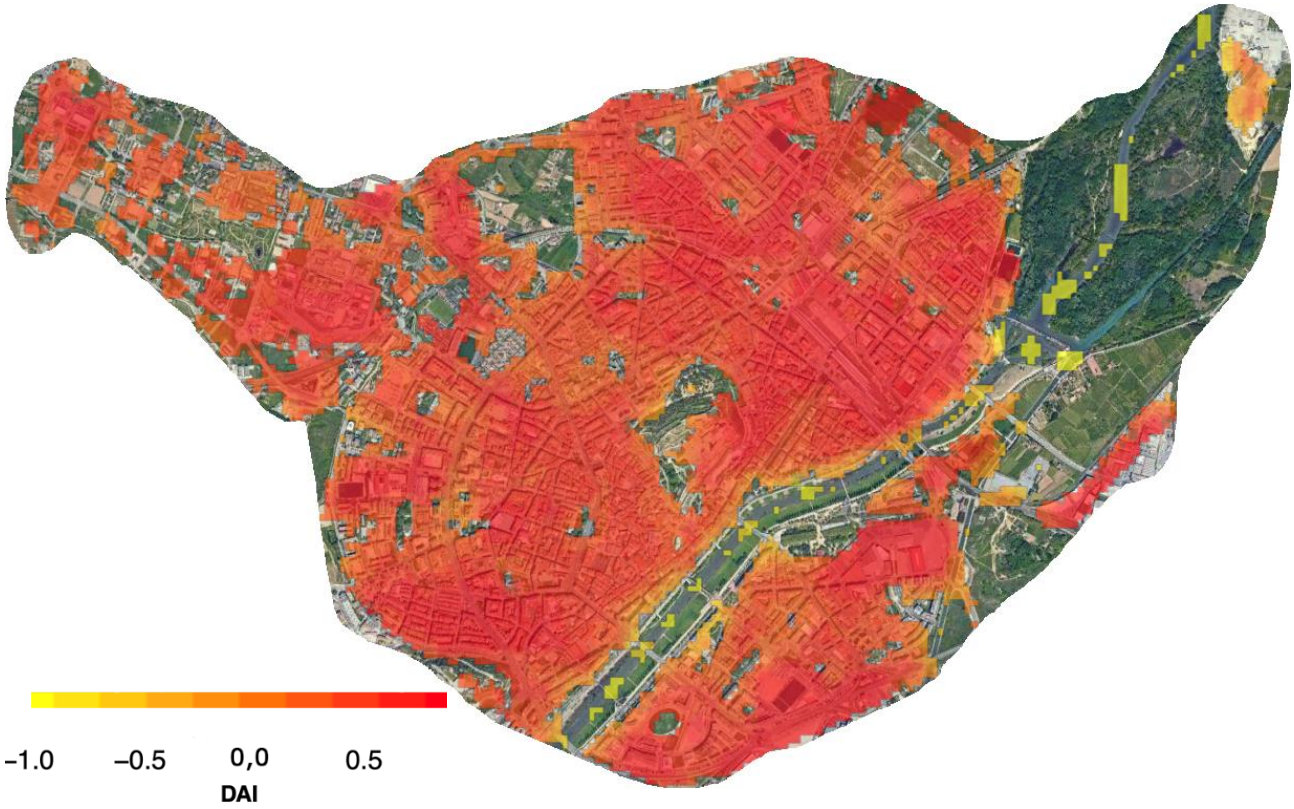
ZARAGOZA



HUESCA



LLEIDA



Bibliografía

European commission. URL https://ec.europa.eu/environment/nature/ecosystems/benefits/index_en.htm.

I. E. Agency. Solar PV. Technical report, Paris, 2020. URL https://www.iea.org/reports/solar-pv.

C. C. Aggarwal. *Data Mining*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-14141-1. doi:10.1007/978-3-319-14142-8.

H. Akbari y L. S. Rose. Urban Surfaces and Heat Island Mitigation Potentials. *Journal of the Human-Environment System*, 11(2):85–101, 2008. ISSN 1345-1324. doi:10.1618/jhes.11.85.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, y J. Sander. OPTICS: ordering points to identify the clustering structure. In *Proceedings on 1999 ACM SIGMOD international conference on management of data*, volume 28, pages 49–60. ACM PUB27 New York, NY, USA, jun 1999. doi:10.1145/304181.304187.

A. Asadi, H. Arefi, y H. Fathipoor. Simulation of green roofs and their potential mitigating effects on the urban heat island using an artificial neural network: A case study in Austin, Texas. *Advances in Space Research*, 66(8):1846–1862, 10 2020. ISSN 02731177. doi:10.1016/j.asr.2020.06.039.

M. Awad y R. Khanna. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi:10.1007/978-1-4302-5990-9_4.

M. Awrangjeb, C. Zhang, y C. S. Fraser. Automatic extraction of building roofs using LiDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 83:1–18, sep 2013. ISSN 09242716. doi:10.1016/j.isprsjprs.2013.05.006.

C. Bartesaghi-Koc, P. Osmond, y A. Peters. Mapping and classifying green infrastructure typologies for climate-related studies based on remote sensing data. *Urban Forestry & Urban Greening*, 37:154–167, 1 2019. ISSN 16188667. doi:10.1016/j.ufug.2018.11.008.

S. Behzadi y A. A. Alesheikh. Introducing a novel model of belief–desire–intention agent for urban land use planning. *Engineering Applications of Artificial Intelligence*, 26(9): 2028–2044, 10 2013. ISSN 09521976. doi:10.1016/j.engappai.2013.06.015.

B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A. Boulesteix, D. Deng, y M. Lindauer. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, page e1484, jan 2023. ISSN 1942-4787. doi:10.1002/widm.1484.

M. Bourdeau, P. Basset, S. Beauchêne, D. Da Silva, T. Guiot, D. Werner, y E. Nefzaoui. Classification of daily electric load profiles of non-residential buildings. *Energy and Buildings*, 233, 2021. doi:10.1016/j.enbuild.2020.110670.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 08856125. doi:10.1007/BF00058655.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi:10.1023/A:1010933404324.

L. Breiman, J. H. Friedman, R. A. Olshen, y C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Z. Cai, G. Han, y M. Chen. Do water bodies play an important role in the relationship between urban form and land surface temperature? *Sustainable Cities and Society*, 39: 487–498, 2018. ISSN 2210-6707. doi:10.1016/j.scs.2018.02.033.

D. Cannizzaro, A. Aliberti, L. Bottaccioli, E. Macii, A. Acquaviva, y E. Patti. Solar radiation forecasting based on convolutional neural network and ensemble learning. *Expert Systems with Applications*, 181:115167, 2021. ISSN 0957-4174. doi:10.1016/j.eswa.2021.115167.

M. Castangia, A. Aliberti, L. Bottaccioli, E. Macii, y E. Patti. A compound of feature selection techniques to improve solar radiation forecasting. *Expert Systems with Applications*, 178:114979, 2021. ISSN 0957-4174. doi:10.1016/j.eswa.2021.114979.

G. Castillo y J. Gama. An Adaptive Prequential Learning Framework for Bayesian Network Classifiers. *Lecture Notes in Artificial Intelligence*, 4213:67–78, 2006. doi:10.1007/11871637_11.

S. S. Cembranel, F. Lezama, J. Soares, S. Ramos, A. Gomes, y Z. Vale. A Short Review on Data Mining Techniques for Electricity Customers Characterization. In *2019 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia)*, pages 194–199. IEEE, mar 2019. ISBN 978-1-5386-7434-5. doi:10.1109/GTDAsia.2019.8715891.

W. Chang, J. Cheng, J. J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, y B. Borges. *shiny: Web Application Framework for R*, 2021. URL `https://cran.r-project.org/package=shiny`.

Y.-P. Chang. An ant direction hybrid differential evolution algorithm in determining the tilt angle for photovoltaic modules. *Expert Systems with Applications*, 37(7):5415–5422, 2010. ISSN 0957-4174. doi:10.1016/j.eswa.2010.01.015.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, y R. Wirth. *CRISP-DM 1.0*. 2000.

M. Charrad, N. Ghazzali, V. Boiteau, y A. Niknafs. {NbClust}: An {R} Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1–36, 2014.

A. Chen, X. A. Yao, R. Sun, y L. Chen. Effect of urban green patterns on surface urban cool islands and its seasonal variations. *Urban Forestry & Urban Greening*, 13(4):646–654, 1 2014. ISSN 16188667. doi:10.1016/j.ufug.2014.07.006.

T. Chen y C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. ISBN 9781450342322. doi:10.1145/2939672.2939785.

B. Chun y J. M. Guldmann. Impact of greening on the urban heat island: Seasonal variations and mitigation strategies. *Computers, Environment and Urban Systems*, 71: 165–176, 9 2018. ISSN 01989715. doi:10.1016/j.compenvurbsys.2018.05.006.

Commission for Energy Regulation (CER). CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010. 1st Edition. Irish Social Science Data Archive, 2012. URL `https://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

A. Coronas y M. Villarrubia. Radiación solar total y directa sobre superficies de cualquier inclinación y orientación en Barcelona. *Quaderns d'enginyeria*, 4(1):127–147, 1983.

C. Cortes, V. Vapnik, y L. Saitta. Support-vector networks. *Machine Learning 1995 20:3*, 20(3):273–297, sep 1995. ISSN 1573-0565. doi:10.1007/BF00994018.

M. Cuturi. Fast global alignment kernels. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 929–936, 2011.

D. L. Davies y D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, apr 1979. ISSN 0162-8828. doi:10.1109/TPAMI.1979.4766909.

F. De La Torre y T. Kanade. Discriminative Cluster Analysis. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006. doi:10.1145/1143844.

J. del Campo-Ávila, G. Ramos-Jiménez, J. Gama, y R. Morales-Bueno. Improving the performance of an incremental algorithm driven by error margins. *Intelligent Data Analysis*, 12(3):305–318, 2008. ISSN 15714128. doi:10.3233/ida-2008-12305.

J. del Campo-Ávila, A. Takilalte, A. Bifet, y L. Mora-López. Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation. *Expert Systems with Applications*, 167:114147, 2021. ISSN 0957-4174. doi:10.1016/j.eswa.2020.114147.

J. Dhanapal y T. Perumal. Inflated Power Iteration Clustering Algorithm to Optimize Convergence Using Lagrangian Constraint. In *Advances in Intelligent Systems and Computing*, volume 465, pages 227–237. Springer Verlag, 2016. ISBN 9783319336206. doi:10.1007/978-3-319-33622-0_21.

J. R. Doorga, S. D. Rughooputh, y R. Boojhawon. Multi-criteria GIS-based modelling technique for identifying potential solar farm sites: A case study in Mauritius. *Renewable Energy*, 2018. ISSN 18790682. doi:10.1016/j.renene.2018.08.105.

H. Du, W. Cai, Y. Xu, Z. Wang, Y. Wang, y Y. Cai. Quantifying the cool island effects of urban green spaces using remote sensing Data. *Urban Forestry & Urban Greening*, 27:24–31, 10 2017. ISSN 1618-8667. doi:10.1016/J.UFUG.2017.06.008.

R. Dutta, A. Morshed, J. Aryal, C. D'Este, y A. Das. Development of an intelligent environmental knowledge system for sustainable agricultural decision support. *Environmental Modelling and Software*, 52:264–272, 2 2014. ISSN 13648152. doi:10.1016/j.envsoft.2013.10.004.

S. El Joumani, S. E. Mechkouri, R. Zennouhi, O. El Kadmiri, y L. Masmoudi. Segmentation method based on multiobjective optimization for very high spatial resolution satellite images. *EURASIP Journal on Image and Video Processing*, 2017 (1):26, dec 2017. ISSN 1687-5281. doi:10.1186/s13640-016-0161-2.

M. Ester, H.-P. Kriegel, J. Sander, y X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

R. C. Estoque, Y. Murayama, y S. W. Myint. Effects of landscape composition and pattern on land surface temperature: An urban heat island study in the megacities of

Southeast Asia. *Science of the Total Environment*, 577:349–359, 2017. ISSN 18791026. doi:10.1016/j.scitotenv.2016.10.195.

European Commission. *Building a Green Infraestructure for Europe*. 2013. ISBN 9789279334283. doi:10.2779/54125.

A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, y J. O. Agushaka. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, 33(11):6247–6306, jun 2021. ISSN 0941-0643. doi:10.1007/s00521-020-05395-4.

S. Fan, L. Chen, y W. J. Lee. Machine learning based switching model for electricity load forecasting. *Energy Conversion and Management*, 49(6):1331–1344, 2008. ISSN 01968904. doi:10.1016/j.enconman.2008.01.008.

F. J. Fernández, L. J. Alvarez-Vázquez, N. García-Chan, A. Martínez, y M. E. Vázquez-Méndez. Optimal location of green zones in metropolitan areas to control the urban heat island. *Journal of Computational and Applied Mathematics*, 289:412–425, 2015. ISSN 03770427. doi:10.1016/j.cam.2014.10.023.

V. Figueiredo, F. J. Duarte, F. Rodrigues, Z. Vale, y J. Gouveia. Electric energy customer characterization by clustering. In *IEEE Intelligent Systems Applications to Power Systems*, 2003.

A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, Berlin, Heidelberg, 2002. ISBN 3540433317.

S. Freitas, C. Catita, P. Redweik, y M. Brito. Modelling solar potential in the urban environment: State-of-the-art review. *Renewable and Sustainable Energy Reviews*, 41: 915–931, jan 2015. ISSN 13640321. doi:10.1016/j.rser.2014.08.060.

J. Fu, K. Dupre, S. Tavares, D. King, y Z. Banhalmi-Zakar. Optimized greenery configuration to mitigate urban heat: A decade systematic review. *Frontiers of Architectural Research*, jan 2022. ISSN 20952635. doi:10.1016/j.foar.2021.12.005.

A. A. Fusami, O. C. Nweze, y R. Hassan. Comparing the Effect of Deforestation Result by NDVI and SAVI. *International Journal of Scientific and Research Publications (IJSRP)*, 10(06):918–925, 2020. doi:10.29322/ijsrp.10.06.2020.p102110.

N. L. Gavankar y S. K. Ghosh. Object based building footprint detection from high resolution multispectral satellite image using K -means clustering algorithm and shape parameters. *Geocarto International*, 34(6):626–643, may 2019. ISSN 1010-6049. doi:10.1080/10106049.2018.1425736.

P. Geyer, C. Koch, y P. Pauwels. Fusing data, engineering knowledge and artificial intelligence for the built environment. *Advanced Engineering Informatics*, 48:101242, 2021. ISSN 14740346. doi:10.1016/j.aei.2020.101242.

W. B. Goggins, E. Y. Chan, E. Ng, C. Ren, y L. Chen. Effect modification of the association between short-term meteorological factors and mortality by urban heat islands in Hong Kong. *PLoS ONE*, 7(6):9–14, 2012. ISSN 19326203. doi:10.1371/journal.pone.0038551.

Google. Google Project Sunroof, 2021, May 7. URL `https://www.google.com/get/sunroof`.

G. Hamerly y C. Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

M. a. Hart y D. J. Sailor. Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. *Theoretical and Applied Climatology*, 95 (3-4):397–406, 2009. ISSN 0177798X. doi:10.1007/s00704-008-0017-5.

E. Hathway y S. Sharples. The interaction of rivers and urban form in mitigating the urban heat island effect: A UK case study. *Building and Environment*, 58:14–22, 2012. ISSN 0360-1323. doi:10.1016/j.buildenv.2012.06.013.

L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, y Y. Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition*, 70:25–43, 2017. ISSN 00313203. doi:10.1016/j.patcog.2017.04.018.

C. Heaviside, S. Vardoulakis, y X. M. Cai. Attribution of mortality to the urban heat island during heatwaves in the West Midlands, UK. *Environmental Health: A Global Access Science Source*, 15(Suppl 1), 2016. ISSN 1476069X. doi:10.1186/s12940-016-0100-9.

J. Hernández Orallo, M. J. Ramírez Quintana, y C. Ferri Ramírez. *Introducción a la minería de datos*. Pearson Prentice Hall, 2004. ISBN 9788420540917.

S. S. Herrera-Gomez, A. Quevedo-Nolasco, y L. Pérez-Urrestarazu. The role of green roofs in climate change mitigation. A case study in Seville (Spain). *Building and Environment*, 123:575–584, 2017. ISSN 03601323. doi:10.1016/j.buildenv.2017.07.036.

R. J. Hijmans. *rspatial: rspatial.org data*, 2018. URL `https://rspatial.org/`.

R. J. Hijmans. raster: Geographic Data Analysis and Modeling, 2020. URL `https://cran.r-project.org/package=raster`.

K. Hornik, C. Buchta, y A. Zeileis. Open-Source Machine Learning: {R} Meets {Weka}. *Computational Statistics*, 24(2):225–232, 2009. doi:10.1007/s00180-008-0119-7.

HuellaSolar. Huella Solar, 2021, May 7. URL `http://www.huellasolar.com/`.

International Energy Agency. *Energy Technology Perspectives 2016*. Energy Technology Perspectives. OECD, Paris, 6 2016. ISBN 9789264252349. doi:10.1787/energy_tech-2016-en.

International Renewable Energy Agency, IRENA. Renewable Capacity Statistics. Technical report, 2023.

M. Iqbal. *An Introduction to Solar Radiation*. Academic Press, Inc.New York. London, 1983.

A. Karimi y Y. E. Ghajari. Improving land surface temperature prediction using spatiotemporal factors through a genetic-based selection procedure (case study: Tehran, iran). *Advances in Space Research*, 69(9):3258–3267, 2022. ISSN 0273-1177. doi:10.1016/j.asr.2022.02.004.

S. Kartal y A. Sekertekin. Prediction of MODIS land surface temperature using new hybrid models based on spatial interpolation techniques and deep learning models. *Environ Sci Pollut Res Int*, 29(44):67115–67134, May 2022. doi:10.1007/s11356-022-20572-9.

L. Kaufman y P. J. Rousseeuw. Clustering Large Data Sets. In *Pattern Recognition in Practice*, pages 425–437. Elsevier, 1986. doi:10.1016/B978-0-444-87877-9.50039-X.

R. Kesavan, M. Muthian, K. Sudalaimuthu, S. Sundarsingh, y S. Krishnan. ARIMA modeling for forecasting land surface temperature and determination of urban heat island using remote sensing techniques for chennai city, india. *Arabian Journal of Geosciences*, 14(11):1016, May 2021. ISSN 1866-7538. doi:10.1007/s12517-021-07351-5.

U. Khalil, B. Aslam, U. Azam, y H. M. D. Khalid. Time series analysis of land surface temperature and drivers of urban heat island effect based on remotely sensed data to develop a prediction model. *Applied Artificial Intelligence*, 35(15):1803–1828, 2021. doi:10.1080/08839514.2021.1993633.

M. Khoshboresh-Masouleh, F. Alidoost, y H. Arefi. Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors. *Journal of Applied Remote Sensing*, 14(03):1, jul 2020. ISSN 1931-3195. doi:10.1117/1.JRS.14.034503.

L. Kong, Z. Liu, y J. Wu. A systematic review of big data-based urban sustainability research: State-of-the-science and future directions. *Journal of Cleaner Production*, 273, 2020. ISSN 09596526. doi:10.1016/j.jclepro.2020.123142.

C. E. Kontokosta y C. Tull. A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, 197:303–317, 7 2017. ISSN 03062619. doi:10.1016/j.apenergy.2017.04.005.

M. Kuhn. *caret: Classification and Regression Training*, 2020. URL `https://cran.r-project.org/package=caret`.

J. S. Lee, J. T. Kim, y M. G. Lee. Mitigation of urban heat island effect and greenroofs. *Indoor and Built Environment*, 23(1):62–69, 2 2013. ISSN 1420-326X. doi:10.1177/1420326X12474483.

S. Lee, S. Iyengar, M. Feng, P. Shenoy, y S. Maji. DeepRoof: A Data-driven Approach For Solar Potential Estimation Using Rooftop Imagery. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2105–2113, New York, NY, USA, jul 2019. ACM. doi:10.1145/3292500.3330741.

W. Li, C. He, J. Fang, J. Zheng, H. Fu, y L. Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *Remote Sensing*, 11(4), 2019. ISSN 2072-4292. doi:10.3390/rs11040403.

X. Li y W. Zhou. Optimizing urban greenspace spatial pattern to mitigate urban heat island effects: Extending understanding from local to the city scale. *Urban Forestry & Urban Greening*, 41:255–263, 5 2019. ISSN 16188667. doi:10.1016/j.ufug.2019.04.008.

X. Li, W. Li, A. Middel, S. Harlan, A. Brazel, y B. Turner. Remote sensing of the surface urban heat island and land architecture in Phoenix, Arizona: Combined effects of land composition and configuration and cadastral–demographic–economic factors. *Remote Sensing of Environment*, 174:233–243, 3 2016. ISSN 00344257. doi:10.1016/j.rse.2015.12.022.

Y. Li, T. Li, y H. Liu. Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3):551–577, dec 2017. ISSN 0219-1377. doi:10.1007/s10115-017-1059-8.

S.-H. Liao, P.-H. Chu, y P.-Y. Hsiao. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311, sep 2012. ISSN 09574174. doi:10.1016/j.eswa.2012.02.063.

B. S. Lin, C. C. Yu, A. T. Su, y Y. J. Lin. Impact of climatic conditions on the thermal effectiveness of an extensive green roof. *Building and Environment*, 67:26–33, 2013. ISSN 03601323. doi:10.1016/j.buildenv.2013.04.026.

F. Lin y W. W. Cohen. Power Iteration Clustering. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 655–662, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

K. Liu, X. Li, S. Wang, y X. Gao. Assessing the effects of urban green landscape on urban thermal environment dynamic in a semiarid city by integrated use of airborne data, satellite imagery and land surface model. *International Journal of Applied Earth Observation and Geoinformation*, 107:102674, mar 2022. ISSN 1569-8432. doi:10.1016/J.JAG.2021.102674.

Q. Liu y Y. Wu. Supervised Learning. In *Encyclopedia of the Sciences of Learning*, pages 3243–3245. Springer US, Boston, MA, 2012. doi:10.1007/978-1-4419-1428-6_451.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

R. Martínez-Gordón, G. Morales-España, J. Sijm, y A. Faaij. A review of the role of spatial resolution in energy systems modelling: Lessons learned and applicability to the North Sea region. *Renewable and Sustainable Energy Reviews*, 141:110857, 5 2021. ISSN 13640321. doi:10.1016/j.rser.2021.110857.

M. Masoudi y P. Y. Tan. Multi-year comparison of the effects of spatial pattern of urban green spaces on urban land surface temperature. *Landscape and Urban Planning*, 184: 44–58, 4 2019. ISSN 01692046. doi:10.1016/j.landurbplan.2018.10.023.

F. McLoughlin, A. Duffy, y M. Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, 141:190–199, 2015. ISSN 0306-2619. doi:10.1016/j.apenergy.2014.12.039.

M. Mejbel Salih, O. Zakariya Jasim, K. I. Hassoon, y A. Jameel Abdalkadhum. Land Surface Temperature Retrieval from LANDSAT-8 Thermal Infrared Sensor Data and Validation with Infrared Thermometer Camera. *International Journal of Engineering & Technology*, 7(4.20):608, 11 2018. doi:10.14419/ijet.v7i4.20.27402.

K. Mets, F. Depuydt, y C. Develder. Two-stage load pattern clustering using fast wavelet transformation. *IEEE Transactions on Smart Grid*, 7(5):2250–2259, 2016. doi:10.1109/TSG.2015.2446935.

M. Mittal, L. M. Goyal, D. J. Hemanth, y J. K. Sethi. Clustering approaches for high-dimensional databases: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1300, 5 2019. ISSN 1942-4795. doi:10.1002/WIDM.1300.

Y. B. Moon. Simulation modelling for sustainability: a review of the literature. *International Journal of Sustainable Engineering*, 10(1):2–19, 2017. ISSN 19397046. doi:10.1080/19397038.2016.1220990.

R. Moreno-Sáez y L. Mora-López. Modelling the distribution of solar spectral irradiance using data mining techniques. *Environmental Modelling and Software*, 53:163–172, 2014. ISSN 13648152. doi:10.1016/j.envsoft.2013.12.002.

R. Moreno-Sáez, M. Sidrach-De-Cardona, y L. Mora-López. Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules. *Expert Systems with Applications*, 40(17):7141–7150, 2013. ISSN 09574174. doi:10.1016/j.eswa.2013.06.059.

H. Motieyan y M. S. Mesgari. An Agent-Based Modeling approach for sustainable urban planning from land use and public transit perspectives. *Cities*, 81:91–100, 2018. ISSN 02642751. doi:10.1016/j.cities.2018.03.018.

F. Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2 (5-6):183–197, 1991. doi:10.1016/0925-2312(91)90023-5.

E. K. Mustafa, Y. Co, G. Liu, M. R. Kaloop, A. A. Beshr, F. Zarzoura, y M. Sadek. Study for predicting land surface temperature (lst) using landsat data: A comparison of four algorithms. *Advances in Civil Engineering*, 2020:7363546, Mar 2020. ISSN 1687-8086. doi:10.1155/2020/7363546.

A. Nesticò, R. Passaro, G. Maselli, y P. Somma. Multi-criteria methods for the optimal localization of urban green areas. *Journal of Cleaner Production*, 374:133690, nov 2022. ISSN 0959-6526. doi:10.1016/J.JCLEPRO.2022.133690.

A. Ossola, G. D. Jenerette, A. McGrath, W. Chow, L. Hughes, y M. R. Leishman. Small vegetated patches greatly reduce urban surface temperature during a summer heatwave in adelaide, australia. *Landscape and Urban Planning*, 209:104046, 2021. ISSN 0169-2046. doi:10.1016/j.landurbplan.2021.104046.

C. Osterwald. Translation of device performance measurements to reference conditions. *Solar Cells*, 18(3-4):269–279, sep 1986. ISSN 03796787. doi:10.1016/0379-6787(86)90126-2.

I. Palomares, E. Martínez-Cámara, R. Montes, P. García-Moral, M. Chiachio, J. Chiachio, S. Alonso, F. J. Melero, D. Molina, B. Fernández, C. Moral, R. Marchena, J. P. de Vargas, y F. Herrera. A panoramic view and swot analysis of artificial intelligence for achieving the sustainable development goals by 2030: progress and prospects. *Applied Intelligence*, 6 2021. ISSN 0924-669X. doi:10.1007/s10489-021-02264-y.

Z. Pan, J. Xu, Y. Guo, Y. Hu, y G. Wang. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sensing*, 12(10):1574, may 2020. ISSN 2072-4292. doi:10.3390/rs12101574.

J. Park, J.-H. Kim, D. K. Lee, C. Y. Park, y S. G. Jeong. The influence of small green space type and structure at the street level on urban heat island mitigation. *Urban Forestry & Urban Greening*, 21:203–212, 1 2017. ISSN 1618-8667. doi:10.1016/J.UFUG.2016.12.005.

D. Pelleg y A. W. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019. URL `https://www.python.org/`.

J. R. J. R. Quinlan y J. Ross. *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers, 1993. ISBN 1558602380. URL `https://dl.acm.org/citation.cfm?id=152181`.

R Core Team. R: A Language and Environment for Statistical Computing, 2019. URL `https://www.r-project.org/`.

Z. A. Rahaman, A. A. Kafy, M. Saha, A. A. Rahim, A. I. Almulhim, S. N. Rahaman, M. A. Fattah, M. T. Rahman, K. S, A. A. Faisal, y A. Al Rakib. Assessing the impacts of vegetation cover loss on surface temperature, urban heat island and carbon emission in Penang city, Malaysia. *Building and Environment*, 222:109335, aug 2022. ISSN 0360-1323. doi:10.1016/J.BUILDENV.2022.109335.

A. Rajabi, M. Eskandari, M. Ghadi, L. Li, J. Zhang, y P. Siano. A comparative study of clustering techniques for electrical load pattern segmentation. *Renewable and Sustainable Energy Reviews*, 120, 2020. doi:10.1016/j.rser.2019.109628.

B. S. Rao, G. A. Kumar, C. Runjhun, C. V. K. V. P. J. Rao, y G. V. Babu. Improvement of Airborne LiDAR Intensity Image Content with Shaded nDSM and Assessment of Its Utility in Geospatial Data Generation. *Journal of the Indian Society of Remote Sensing*, 50(3):507–521, Mar 2022. ISSN 0974-3006. doi:10.1007/s12524-021-01468-6.

T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, y M. Kolehmainen. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*, 87(11):3538–3545, 2010. ISSN 03062619. doi:10.1016/j.apenergy.2010.05.015.

C. Reis y A. Lopes. Evaluating the cooling potential of urban green spaces to tackle urban climate change in Lisbon. *Sustainability (Switzerland)*, 11(9), 2019. ISSN 20711050. doi:10.3390/su11092480.

D. Rey y M. Neuhäuser. *Wilcoxon-Signed-Rank Test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi:10.1007/978-3-642-04898-2_616.

F. Rodríguez-Gómez, J. del Campo-Ávila, M. Ferrer-Cuesta, y L. Mora-López. Data driven tools to assess the location of photovoltaic facilities in urban areas. *Expert Systems with Applications*, 203:117349, 2022a. ISSN 0957-4174. doi:10.1016/j.eswa.2022.117349.

F. Rodríguez-Gómez, R. Fernández-Cañero, G. Pérez, J. del Campo-Ávila, D. López-Rodríguez, y L. Pérez-Urrestarazu. Detection of unfavourable urban areas with higher temperatures and lack of green spaces using satellite imagery in sixteen spanish cities. *Urban Forestry & Urban Greening*, 78:127783, 2022b. ISSN 1618-8667. doi:10.1016/j.ufug.2022.127783.

F. Rodríguez-Gómez, J. del Campo-Ávila, y L. Mora-López. A novel clustering based method for characterizing household electricity consumption patterns. *Engineering Applications of Artificial Intelligence*, unpublished_a. Under review.

F. Rodríguez-Gómez, D. López-Rodríguez, L. Pérez-Urrestarazu, y J. del Campo-Ávila. Data-driven identification of urban areas in need of green infrastructure for temperature reduction. *Applied Soft Computing*, unpublished_c. Under review.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi:10.1016/0377-0427(87)90125-7.

J.-R. Roussel, D. Auty, N. C. Coops, P. Tompalski, T. R. Goodbody, A. S. Meador, J.-F. Bourdon, F. de Boissieu, y A. Achim. lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment*, 251:112061, dec 2020. ISSN 00344257. doi:10.1016/j.rse.2020.112061.

S. Rueda. *Libro verde de sostenibilidad urbana y local en la era de la información*. 2012. ISBN 978-84-491-1233-1.

H. Sakoe y S. Chiba. Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics. *Speech and Signal Processing*, 26(1):43, 1978.

M. Santamouris. Cooling the cities – A review of reflective and green roof mitigation technologies to fight heat island and improve comfort in urban environments. *Solar Energy*, 103:682–703, 2014. ISSN 0038092X. doi:10.1016/j.solener.2012.07.003.

R. Sharan, A. Maron-Katz, y R. Shamir. CLICK and EXPANDER: A system for clustering and visualizing gene expression data. *Bioinformatics*, 19(14):1787–1799, sep 2003. ISSN 13674803. doi:10.1093/bioinformatics/btg232.

M. Sharma, R. D. Garg, V. Badenko, A. Fedotov, L. Min, y A. Yao. Potential of airborne LiDAR data for terrain parameters extraction. *Quaternary International*, 575-576:317–327, feb 2021. ISSN 10406182. doi:10.1016/j.quaint.2020.07.039.

C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, y J. Liu. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1):31, dec 2021. ISSN 1687-1499. doi:10.1186/s13638-021-01910-w.

A. J. Smola y B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug 2004. ISSN 1573-1375. doi:10.1023/B:STCO.0000035301.49549.88.

M. Steinbach, G. Karypis, y V. Kumar. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*, 2000.

Stroock, D. W. Probability theory : an analytic view, 2011. URL `https://archive.org/details/probabilitytheor00dwst`.

Y. Su, J. Wu, C. Zhang, X. Wu, Q. Li, L. Liu, C. Bi, H. Zhang, R. Lafortezza, y X. Chen. Estimating the cooling effect magnitude of urban vegetation in different climate zones using multi-source remote sensing. *Urban Climate*, 43:101155, may 2022. ISSN 2212-0955. doi:10.1016/J.UCLIM.2022.101155.

R. Sun, W. Xie, y L. Chen. A landscape connectivity model to quantify contributions of heat sources and sinks in urban regions. *Landscape and Urban Planning*, 178:43–50, 10 2018. ISSN 01692046. doi:10.1016/j.landurbplan.2018.05.015.

T. Susca, S. Gaffin, y G. Dell'Osso. Positive effects of vegetation: Urban heat island and green roofs. *Environmental Pollution*, 159(8-9):2119–2126, 8 2011. ISSN 02697491. doi:10.1016/j.envpol.2011.03.007.

W. Toussaint. Domestic Electrical Load Metering, Hourly Data 1994-2014 [dataset]. Version 1., 2019.

W. Toussaint y D. Moodley. Clustering Residential Electricity Consumption Data to Create Archetypes that Capture Household Behaviour in South Africa. *South African Computer Journal*, 32:1 – 34, 12 2020. ISSN 2313-7835. doi:10.18489/sacj.v32i2.845.

United Nations General Assembly. Transforming our World: the 2030 Agenda for Sustainable Development. Technical report, 2015.

U.S. EPA. Heat Island Impacts. Technical report, 2021.

J. Velázquez, P. Anza, J. Gutiérrez, B. Sánchez, A. Hernando, y A. García-Abril. Planning and selection of green roofs in large urban areas. Application to Madrid metropolitan area. *Urban Forestry & Urban Greening*, 40(March):323–334, 4 2019. ISSN 16188667. doi:10.1016/j.ufug.2018.06.020.

S. Verdu, M. Garcia, F. Franco, N. Encinas, A. Marin, A. Molina, y E. Lazaro. Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. In *IEEE PES Power Systems Conference and Exposition, 2004.*, volume 2, pages 1240–1247. IEEE, 2004. ISBN 0-7803-8718-X. doi:10.1109/PSCE.2004.1397641.

H. Wickham, R. François, L. Henry, y K. Müller. dplyr: A Grammar of Data Manipulation, 2020. URL `https://cran.r-project.org/package=dplyr`.

J. Wijffels. *cronR: Schedule R Scripts and Processes with the 'cron' Job Scheduler*, 2020. URL `https://cran.r-project.org/package=cronR`.

I. H. Witten, E. Frank, M. A. Hall, y C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, 4 edition, 2017. ISBN 978-0-12-804291-5.

J. R. Wolch, J. Byrne, y J. P. Newell. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape and Urban Planning*, 125:234–244, 2014. ISSN 0169-2046. doi:10.1016/j.landurbplan.2014.01.017.

A. J. Wyner, M. Olson, J. Bleich, y D. Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18: 1–33, 2017. ISSN 15337928.

D. Xu y Y. Tian. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2(2):165–193, 6 2015. ISSN 2198-5804. doi:10.1007/s40745-015-0040-1.

J. Xue y B. Su. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017:1353691, May 2017. ISSN 1687-725X. doi:10.1155/2017/1353691.

C. Yang, X. He, R. Wang, F. Yan, L. Yu, K. Bu, J. Yang, L. Chang, y S. Zhang. The effect of urban green spaces on the urban thermal environment and its seasonal variations. *Forests*, 8(5):1–19, 2017. ISSN 19994907. doi:10.3390/f8050153.

J. Zhang, R. Verschae, S. Nobuhara, y J.-F. Lalonde. Deep photovoltaic nowcasting. *Solar Energy*, 176:267–276, dec 2018. ISSN 0038092X. doi:10.1016/j.solener.2018.10.024.

T. Zhang, R. Ramakrishnan, y M. Livny. BIRCH: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, 25(2):103–114, jun 1996. ISSN 0163-5808. doi:10.1145/235968.233324.

Z. Zhu, M. A. Wulder, D. P. Roy, C. E. Woodcock, M. C. Hansen, V. C. Radeloff, S. P. Healey, C. Schaaf, P. Hostert, P. Strobl, J.-F. Pekel, L. Lymburner, N. Pahlevan, y T. A. Scambos. Benefits of the free and open landsat data policy. *Remote Sensing of Environment*, 224:382–385, 2019. ISSN 0034-4257. doi:10.1016/j.rse.2019.02.016.