Imagen que contiene Icono

Descripción generada automáticamente

Ciencia de datos para determinar visitantes en un destino turístico inteligente

Máster Universitario en Ciencia de Datos



Trabajo Fin de Máster

Autor:

Javier Ignacio Belmar Tevar

Tutor/es:

Jose Norberto Mazón López

Jose Jacobo Zubcoff Vallejo

Septiembre 2021

Texto

Descripción generada automáticamente

Índice

[1. Introducción 3](#_Toc82616393)

[2. Escenario de partida 5](#_Toc82616394)

[2.1. Recolección de SSIDs 7](#_Toc82616395)

[2.2. Recolección de SSIDs preferidos de dispositivos móviles 8](#_Toc82616396)

[3. Objetivos 9](#_Toc82616397)

[4. Tecnología y herramientas 10](#_Toc82616398)

[5. Trabajo relacionado 11](#_Toc82616399)

[6. Propuesta 11](#_Toc82616400)

[6.1. Valor añadido de la propuesta 11](#_Toc82616401)

[6.2. Descripción de los datos 11](#_Toc82616402)

[6.3. Recolección de los datos 11](#_Toc82616403)

[6.4. Análisis exploratorio 12](#_Toc82616404)

[6.4.1. Análisis diario 15](#_Toc82616405)

[6.5. El algoritmo: Pasos comunes 17](#_Toc82616406)

[6.5.1. Primer paso: discriminación con la lista de ssid’s 17](#_Toc82616407)

[6.5.2. Segundo paso: Clustering de frecuencias 18](#_Toc82616408)

[6.6. El algoritmo: Pasos de la solución escogida. 20](#_Toc82616409)

[6.6.1. Tercer paso: Obtención de un conjunto con conocimiento a priori. 20](#_Toc82616410)

[6.6.2. Cuarto paso: Aplicación a los datos del sensor el conocimiento a priori 21](#_Toc82616411)

[7. Resultados 21](#_Toc82616412)

[7.1. Discusión 25](#_Toc82616413)

[7.1.1. Otra forma de solventar la wifi inactiva 25](#_Toc82616414)

[8. Bibliografía 26](#_Toc82616415)

Tabla de ilustraciones

[Figura 1. Modelo DTI procedente de [CITA] 4](#_Toc82616418)

[Figura 2. Escenario de partida para el desarrollo de la solución que determine el número de visitantes en diversos puntos de la ciudad de Alcoy de manera diaria. 7](#_Toc82616419)

[Figura 3. Dispositivo HOPU Smart Spot IoT. 8](#_Toc82616420)

[Figura 4. Conjunto del cuestionario 12](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616421)

[Figura 5. Conjunto del cuestionario 12](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616422)

[Figura 6. Conjunto del sensor 12](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616423)

[Figura 7. Conjunto de ssid's de Alcoi 13](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616424)

[Figura 8. Gráfica de frecuencias de las ssid's 15](#_Toc82616425)

[Figura 9. Gráfica de frecuencias de las mac's 15](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616426)

[Figura 10. Serie del sensor inicial 16](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616427)

[Figura 11. Serie del cuestionario inicial 16](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616428)

[Figura 12. Rangos potencia del sensor 17](#_Toc82616429)

[Figura 13. Método del codo (SSE) 18](#_Toc82616430)

[Figura 15. Resultado del clustering para las mac's 19](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616431)

[Figura 14. Resultado del clustering para las ssid's 19](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616432)

[Figura 16. Mac's únicas (sensor) y personas en la oficina (cuest.) antes de exclusión 19](#_Toc82616433)

[Figura 17. Mac's únicas (sensor) y personas en la oficina (cuest.) después de exclusión 19](#_Toc82616434)

[Figura 18. Proporcion de personas con wifi en la oficina 21](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616435)

[Figura 19. Resultados Antes de algoritmo 22](#_Toc82616436)

[Figura 20. Resultados después de algoritmo 22](https://unialicante.sharepoint.com/sites/TFMJBelmar/Documentos%20compartidos/General/TFM.docx#_Toc82616437)

[Figura 21. Resultados antes de algorimo. Pot. Intermedia añadida 23](#_Toc82616438)

[Figura 22. Resultados después de algoritmo. Pot. Intermedia 24](#_Toc82616439)

[Figura 23. Resultados antes de algoritmo. Pot. Intermedia 24](#_Toc82616440)

# Introducción

Un destino turístico inteligente (DTI) se puede definir según la Sociedad de la Información a la Sociedad Estatal para la Gestión de la Innovación y las Tecnologías Turísticas (SEGITTUR) como “un espacio turístico innovador, accesible para todos, consolidado sobre una infraestructura tecnológica de vanguardia que garantiza el desarrollo sostenible del territorio, facilita la interacción e integración del visitante con el entorno e incrementa la calidad de su experiencia en el destino y la calidad de vida de los residentes” [6]. Por ello, se puede destacar que el aspecto clave de los DTI es la integración de las tecnologías de la información en el destino mediante la implantación de soluciones hardware y software innovadoras (desde sensores hasta herramientas de inteligencia de negocio) como se detalla en [4].

No obstante, la tecnología aplicada a un destino turístico no debe ser únicamente usada para mejorar la experiencia de las personas que visitan el destino (por ejemplo, para que puedan conectarse a Internet de manera gratuita mediante WiFi, para que puedan saber lo concurrida que está una playa o para que puedan conocer información de las atracciones turísticas por medio de una aplicación móvil), sino que también debe usarse para recopilar datos que permitan conseguir una mayor eficiencia en la gestión del destino [5].

En esta línea, se puede afirmar que un destino turístico se convierte en inteligente cuando hace un uso intensivo de la infraestructura tecnológica con el fin de (1) mejorar la experiencia del turista y (2) mejorar el proceso de toma de decisiones mediante la gestión de los datos.

Cabe destacar el modelo de DTI propuesto en [5] donde los autores reconocen las potencialidades de las tecnologías de la información para configurar destinos inteligentes, pero a partir de unas condiciones previas de carácter estratégico y relacional que determinan las acciones y el alcance de la estrategia. Su modelo se compone de tres niveles (ver Figura 1) que se detallan a continuación:

1. Un nivel estratégico-relacional (enfocado hacia una gobernanza sustentada en la colaboración público-privada que garantice la sostenibilidad y un entorno abierto e innovador).
2. Un nivel instrumental (centrado en la conectividad y la sensorización como precursora de un sistema de información que aporte inteligencia en el proceso de toma de decisiones).
3. Un nivel aplicado (orientado al desarrollo de soluciones inteligentes para la mejora de la experiencia, marketing y gestión del destino).

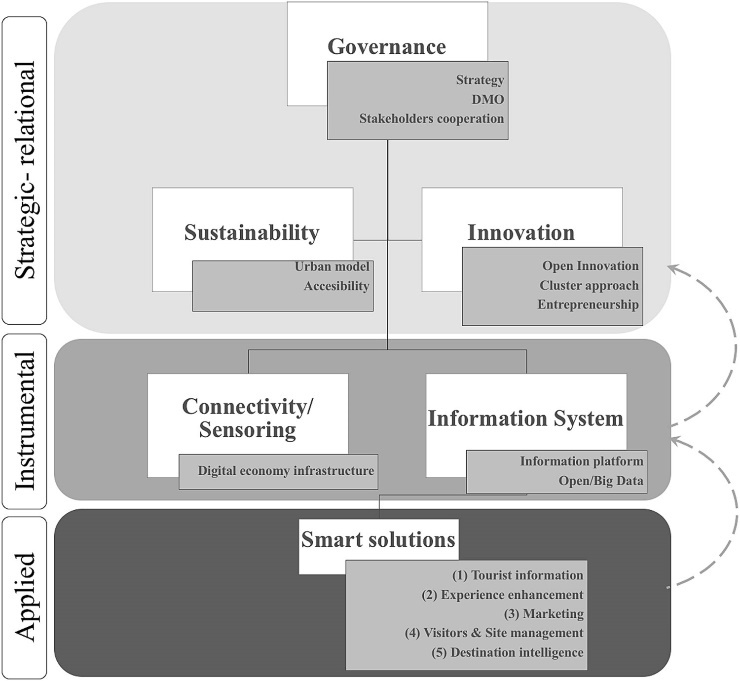


Figura 1. Modelo DTI procedente de [5]

Siguiendo este modelo de DTI, en este trabajo de fin de máster se sitúa en el nivel 3, es decir sobre la base de un problema específico de un destino turístico y una infraestructura hardware que nos proporciona ciertos datos, se propone la aplicación de técnicas de ciencias de datos para desarrollar una solución inteligente. Además, se debe destacar que este TFM realizará en un escenario real, en concreto en Alcoy, uno de los principales destinos turísticos de interior de la Comunitat Valenciana.

La localización geográfica de Alcoy, entre dos parques naturales, rango urbano, rico patrimonio histórico-cultural y tradición industrial le confieren interesantes atractivos que fueron objeto de una política y gestión turística a partir, fundamentalmente, de la creación de la Concejalía de Turismo y de la Oficina de Información Turística en 2000 y, sobre todo, del Plan de Dinamización Turística de Alcoi (2003), en cooperación con la Secretaría General de Turismo y la Agència Valenciana del Turisme. De hecho, las memorias anuales de la Oficina de Turismo de Alcoi[[1]](#footnote-2) reflejan una gestión turística avanzada, patente en la certificación con la Q de calidad de la Tourist Info, la participación del municipio en el Sistema Integral de Calidad Turística Española en Destinos (SICTED) o la reciente incorporación a la red de destinos turísticos inteligentes (red DTI) de la Comunitat Valenciana promovida por Invat·tur (Instituto Valenciano de Tecnologías Turísticas).

Este último hecho el que motiva que Alcoy sea el escenario del proyecto realizado en este trabajo de fin de máster. Alcoy trabaja desde hace tiempo en varios proyectos y estrategias centradas en la aplicación del concepto de ciudad inteligente (Alcoi Smart City[[2]](#footnote-3)) donde destacan proyectos para convertir a Alcoi en un destino turístico inteligente. Uno de estos proyectos se denomina Alcoi Tourist Lab, en el seno del cuál se desarrolla este TFM. Alcoi Tourist Lab está subvencionado por la Agencia Valenciana de la Innovación y tiene como uno de sus principales objetivos, el poder desarrollar y/o probar soluciones y proyectos piloto funcionales hardware y software, que permitan obtener y analizar patrones y niveles de gasto de visitantes y ciudadanos de Alcoy que disfrutan de la oferta cultural, turística, tecnológica, gastronómica y de ocio de la ciudad. El objetivo que se pretende es analizar patrones de visitas y preferencias (siempre respetando la privacidad y la legislación de protección de datos). Desde Alcoy, se pretende analizar estos datos para mejorar la gestión del destino turístico, así como de la propia ciudad, incluyendo la mejora de los servicios ofrecidos tanto por la administración pública como para los sectores de comercio, turismo y hostelería.

Para ello, una necesidad crucial es conocer el número de visitantes que se encuentran en Alcoy en un momento determinado. Esta necesidad correspondería al nivel 1 del modelo DTI descrito anteriormente (Figura 1) ya que se corresponde con un objetivo estratégico. A partir de esta necesidad el proyecto Alcoi Tourist Lab define la infraestructura de sensorización necesaria (nivel 2 del modelo DTI de la Figura 1) con el fin de recolectar los datos necesarios para desarrollar una solución que pueda determinar los visitantes de un destino turístico y diferenciarlos de los residentes (nivel 3 del modelo DTI de la Figura 1).

Como se menciona anteriormente, este TFM se centra en desarrollar esta solución para entre discernir visitantes y residentes en el municipio de Alcoy sobre la base de la infraestructura desarrollada en el proyecto Alcoi Tourist Lab. Además, se realiza una validación y ajuste de la solución mediante el uso de encuestas que se llevan a cabo en la Tourist Info (oficina de turismo) de Alcoy. De esta manera, el objetivo principal de este TFM es usar técnicas de ciencias de datos aprendidas durante el desarrollo del Máster en Ciencia de Datos de la Universidad de Alicante para poder hacer fiable la propuesta en este escenario real de la ciudad de Alcoy.

# Escenario de partida

Dentro del anteriormente mencionado proyecto Alcoi Tourist Lab, se ha desarrollado un sistema inteligente de recolección y procesamiento de datos de dispositivos móviles que permite obtener la afluencia de personas que se encuentran en Alcoy en un momento determinado. Este sistema se basa en el uso de SSID (de sus siglas en inglés Service Set Identifier) de los enrutadores (routers), es decir el nombre de una red WiFi. Cada SSID es una cadena de caracteres que consta de un máximo 32 caracteres. El objetivo del sistema es identificar a aquellos dispositivos móviles (para o cual se detecta también la MAC de cada dispositivo) cuya WiFi almacenada como preferida para la conexión (en concreto el SSID de esa WiFi) esté ubicado en Alcoy y diferenciarlos de aquellos dispositivos cuya SSID preferida no sea de Alcoy.

Con este fin se desarrollan dos artefactos software diferenciados, cada uno haciendo uso de un hardware específico, a saber:

* Software que recopila tramas de datos de dispositivos móviles mediante sensores de afluencia de personas (dispositivos Smart Spots de HOPU que se describen más adelante). Estas tramas se procesan con el fin de determinar cuál es el SSID de la WiFi de conexión preferida de cada dispositivo.
* Software que recopila los SSID de las redes WiFi existentes en Alcoi. Este software se implanta y ejecuta en un dispositivo hardware desarrollado ad-hoc basado en una Raspberry Pi.

Se debe hacer notar que la recolección de SSID se realiza por medio de dispositivos hardware donde no se almacena ningún dato. Por otro lado, las tecnologías empleadas son innovadoras pero lo suficientemente maduras para evitar cualquier riesgo de seguridad. En concreto se usa el protocolo MQTT, base de datos MongoDB y servidor que cumple todos los requisitos de seguridad según normativa europea.

Este es el escenario de partida sobre el que se desarrolla el presente TFM. Este escenario se describe a continuación.

La figura 2 muestra un diagrama de la arquitectura propuesta, cuyos elementos principales se describen a continuación:

* Un recolector de SSIDs de los puntos de acceso WiFi existentes en una ubicación específica (por ejemplo, una ciudad). Este recolector tiene como objetivo crear una base de datos precisa de los SSID disponibles en un destino turístico (parte superior izquierda de la figura, en naranja).
* Un recolector de SSID de puntos de acceso WiFi preferidos procedentes de tramas de datos de solicitud de sondeo de dispositivos móviles detectados por HOPU Smart Spots (parte superior derecha de la figura, en verde).
* A partir de los datos recolectados en el escenario de partida (recolector de SSIDs de la ciudad y recolector de SSIDs preferidos asociados a MACs de dispositivos móviles), se requiere el desarrollo de un algoritmo para determinar la cantidad diaria de visitantes en diferentes lugares alrededor del destino (parte inferior de la figura, en azul). Esta parte corresponde con el objetivo principal de este TFM.

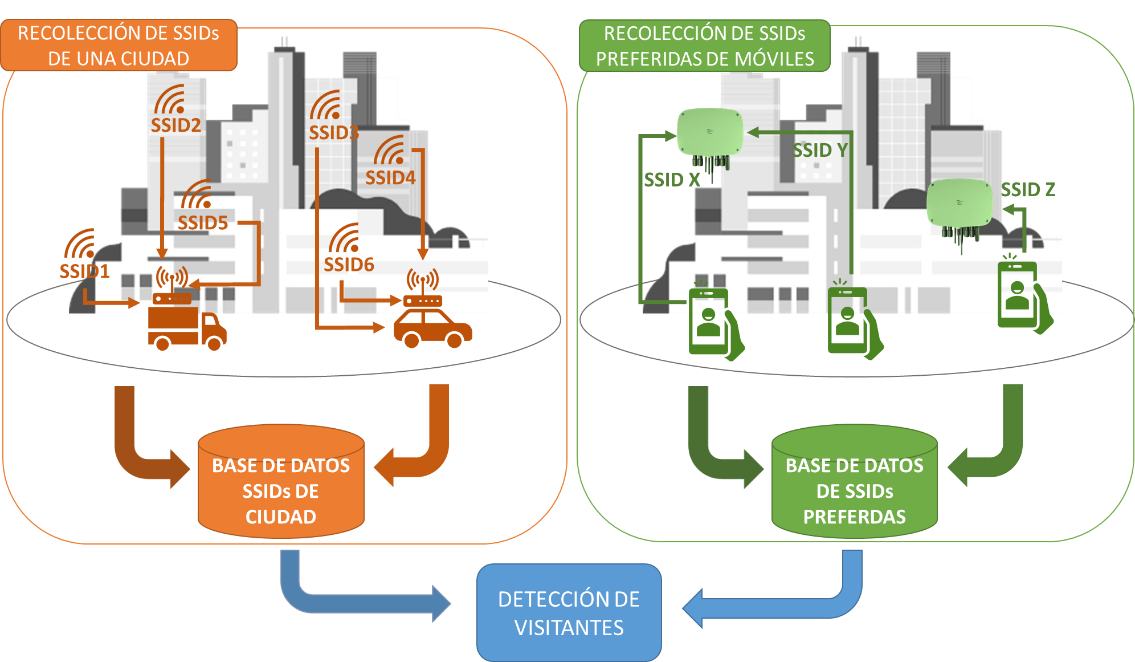


Figura 2. Escenario de partida para el desarrollo de la solución que determine el número de visitantes en diversos puntos de la ciudad de Alcoy de manera diaria.

## 2.1. Recolección de SSIDs

Los puntos de acceso WiFi suelen exponer sus SSID mediante el uso de tramas de balizas periódicas. Se trata de un anuncio periódico para informar a los dispositivos que están a la escucha de que ese SSID está disponible y tiene unas características determinadas. Los dispositivos clientes dependen de estas tramas de baliza para descubrir qué redes están disponibles (escaneo pasivo), y para asegurarse de que las redes a las que están asociadas están realmente presentes y disponibles.

Estos SSID pueden tener de 0 a 32 bytes de longitud y son, en general, una cadena de lenguaje natural de caracteres arbitrarios. Un análisis de los patrones expuestos en todo el mundo muestra que estas cadenas tienden a ser localmente identificables y únicas.

La recopilación de SSID es bastante inmediata y ha sido empleada masivamente por empresas como Google, Mozilla, etc. para mejorar sus sistemas de geolocalización de dispositivos móviles. Una tabla bastante completa de empresas que recogen esta información para el llamado WPS (WiFi Positioning System) y su disponibilidad se puede encontrar en [15]. La captura de SSIDs no está exenta de polémica, siendo destacable el escándalo en torno a la recogida subrepticia de datos WiFi mientras se capturaban imágenes de vídeo y datos cartográficos para el servicio Street View de Google [16]. A raíz de este escándalo, se han utilizado propuestas como la de añadir la etiqueta ”opt\_out” a los SSID para optar por la captura de este servicio.

La legalidad de la recogida de esta información, especialmente en el espacio europeo, no está clara. En el escenario donde se desarrolla este TFM, el proyecto forma parte de un servicio público municipal, por lo que se han extremado las medidas de legalidad y privacidad, de acuerdo con los servicios de asesoría jurídica del municipio. Básicamente, el enfoque que se ha seguido es el de encriptar inmediatamente cada uno de los SSID capturados y almacenar únicamente las marcas de tiempo y la ciudad en la que se ha capturado (no se requiere una localización más específica).

Aunque la implementación más inmediata del sistema de captura de SSID sería el uso de una aplicación para teléfonos móviles como la de Wigle [17], se decidió desarrollar un recolector de SSID a medida para simplificar el desarrollo y mejorar el rendimiento de la captura de tramas SSID utilizando antenas de alta ganancia.

Para probar la propuesta, se utilizó una Raspberry Pi 4 Modelo B junto con una antena WiFi USB de doble banda modelo TP-Link Archer T3U Plus con una sensibilidad de unos -75 dBm para la banda de 2,4 Ghz y -70 dBm para la banda de 5 GHz. La elección del tipo de antena para el escáner de SSIDs ha sido un elemento clave porque se pretendía captar los SSIDs de los edificios adyacentes desde el nivel del suelo. En este sentido, se ha seleccionado una antena omnidireccional con una ganancia de unos 5dBi para ofrecer un compromiso razonable entre alcance y un patrón de recepción adecuado para captar los SSID de los routers WiFi instalados en los edificios adyacentes.

La Raspberry Pi recoge los datos de los SSID de las redes WiFi de una ciudad y se aplica el algoritmo hash SHA1 al SSID de cualquier punto de acceso Wi-Fi encontrado, así como la fecha de captura en formato *timestamp* y el nombre de la ciudad donde se produjo.

Los datos recopilados se almacenan en una base de datos MongoDB donde son accesibles con el fin de desarrollar el presente TFM.

## 2.2. Recolección de SSIDs preferidos de dispositivos móviles

La recopilación de los SSID preferidos se realiza mediante una versión modificada del dispositivo HOPU Smart Spot que se muestra en la figura XXX. Se trata de un dispositivo IoT multisensor dedicado a aplicaciones de ciudades inteligentes. En nuestro caso, nos beneficiamos de su capacidad de monitorización de multitudes por WiFi que ha sido adecuadamente adaptada para conseguir las características deseadas que se describen a continuación.

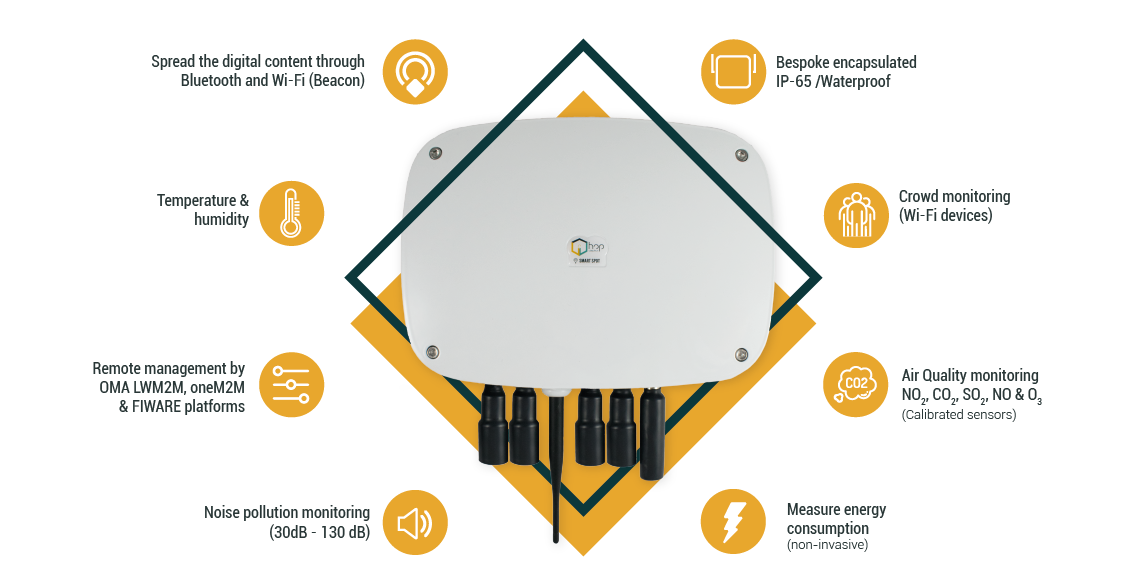


Figura 3. Dispositivo HOPU Smart Spot IoT.

El dispositivo HOPU Smart Spot escucha los paquetes de *probe request* enviados por los dispositivos móviles. Estos paquetes corresponden a los enviados por los teléfonos móviles cuando buscan un punto de acceso WiFi al que conectarse. Normalmente, la búsqueda de un punto de acceso WiFi puede ir acompañada del nombre del punto de acceso WiFi concreto que se busca. Una vez que se ha detectado un dispositivo móvil, dentro del propio HOPU Smart Spot, se aplica un algoritmo hash SHA1 a la MAC y al SSID (de la lista de redes preferidas o PNL) detectados, lo que imposibilita volver a obtener la MAC y el SSID, anonimizando así la información del usuario en el propio dispositivo después de haberla recibido.

Las MAC y los SSID anonimizados mediante el hash SHA1 se almacenan en la memoria volátil y se cuentan en intervalos de 1, 5 y 10 minutos internamente en el dispositivo. Las MAC (y el SSID correspondiente de la PNL) se descartan/borran de la memoria cuando no son detectadas por el dispositivo y están fuera del intervalo de tiempo máximo (10 minutos).

Cada minuto, el dispositivo HOPU Smart Spot cuenta los dispositivos detectados en cada uno de los intervalos (1, 5 y 10 minutos) y envía esta información a través de los protocolos de integración disponibles en el dispositivo (LwM2M/MQTT/LoRa/SENTILO/FIWARE).

Todo este proceso de hashing de MACs y SSIDs dentro del propio dispositivo y sin almacenarlos en la memoria no volátil se realiza para proteger los datos de los ciudadanos de acuerdo con la actual legislación de protección de datos.

Cabe destacar que la recogida de SSID se realiza mediante dispositivos hardware en los que no se almacenan datos. Por otro lado, las tecnologías utilizadas son innovadoras pero lo suficientemente maduras como para evitar cualquier riesgo de seguridad. En concreto, se utiliza FIWARE y, en concreto el protocolo MQTT para el *broker* de mensajes a través de la implementación Mosquitto, y bases de datos y servidores MongoDB que cumplen con todos los requisitos de seguridad según los estándares europeos.

# Objetivos

Una vez los datos se desarrolla el algoritmo de detección de visitantes se realiza un proceso de validación. Esto se consigue mediante la comparación de los resultados obtenidos por el sistema, con los datos obtenidos por encuestas manuales provenientes de algunos de los lugares donde estén los sensores. Estas encuestas recogen información cómo el lugar de residencia, la cantidad de personas que viajan juntas, o si tiene el Wifi del móvil activado o no. En concreto, se ha creado un formulario de Google Forms con el fin de obtener datos de usuarias y usuarios de la Tourist Info de Alcoi que nos permiten validar el sistema desarrollado

INVCUIR LA PARTE DE LOS FORMULARIOS DE LA OFICINA DE TURISMO…

# Tecnología y herramientas

Para llevar a cabo la propuesta se ha usado Python, con el se ha gestionado el manejo de los datos, la elaboración de las visualizaciones y la confección de los algoritmos. Se ha hecho uso de Visual Studio Code desarrollado por Microsoft para la edición del código. Además, este programa soporta una extensión de Jupyter notebook también desarrollado por Microsoft recientemente actualizada (v2021.8.2041215044).

A continuación, se muestran las librerías que se han considerado como herramientas empleadas:

#### Librerías generales

import pandas as pd # manejo de datasets

from pathlib import Path # relativización del path de los archivos

import numpy as np

import datetime as dt # para pandas gestion datetime

from statsmodels.tsa.seasonal import seasonal\_decompose

#### Visualizaciones

import plotly.express as px # gráficas interactivas 1

import plotly.graph\_objects as go # gráficas interactivas 2

import matplotlib.pyplot as plt

import plotly.figure\_factory as ff

#### Clustering / Dendrogramas

from sklearn.cluster import KMeans

from scipy.cluster.hierarchy import linkage # dendrogramas

Como se observa ha divido en tres partes las librerías utilizadas. Para el manejo de los conjuntos se ha utilizado pandas, que es la librería por excelencia de Python para el manejo de datos en forma de tablas. También se ha utilizado pathlib para trabajar con las rutas de los archivos y relativizar la búsqueda de estos a un nivel por encima de la carpeta donde se ejecuta este cuaderno. Se ha utilizado numpy para hacer sobre todo transformaciones de datos como por ejemplo el uso de np.nan o np.inf. Para la extracción de las estacionalidades se ha usado seasonal\_decompose.

Para las representaciones se ha utilizado las librerías de matplotlib y plotly. Se ha aprovechado la interactividad de esta última para obtener representaciones de mayor calidad. Por ejemplo, se puede hacer zoom a zonas en las que los valores parecen similares.

Por último para la detección de bajas y altas frecuencias se ha utilizado Kmeans de sklearn.cluster. Se ha hecho uso de linkage de scipy.cluster.hierarchy para apoyar la división de grupos de frecuencias.

# Trabajo relacionado

En los últimos años el estudio de la afluencia de gente mediante sensores ha ido creciendo indudablemente. Esto fue debido a que los dispositivos que más llevaba consigo la gente tenían conexión Wifi, y por ello surgieron investigaciones que aprovechaban dicha conexión como las de [1], [2] y [3]. En el estudio [1] se analiza la afluencia de personas en el aeropuerto usando dos monitores, estos monitores están situado a ambos extremos del control de seguridad, de esta manera en esta investigación el ground-truth serían los datos del propio control de seguridad. En el caso de nuestra propuesta el groun-truth sería los datos recogidos por los cuestionarios de la oficina de turismo.

Aeropuerto como ground-truth, comparan la eficacia del sistema con un simple contado de direcciones mac, estimación de densidad de peatones.

<https://dl.acm.org/doi/10.4108/icst.mobiquitous.2014.257870>

# Propuesta

## Valor añadido de la propuesta

El turismo es muy importante para España, visitantes

## Descripción de los datos

Los datos procedentes de los sensores son… Decir tamaño campos explicados, etc

Los datos procedentes del formulario usado en la oficina de turismo son… Decir tamaño campos explicados, etc

## Recolección de los datos

Se hace uso del API para la recolección de datos a partir de un proceso ETL realizado en Pentaho Data Integration que permite recolectar los datos de los sensores.

## Análisis exploratorio

Se ha realizado un análisis exploratorio para comprender y comprobar el contenido y la calidad de los datos. A lo largo del análisis se ha ido haciendo algunos arreglos para obtener unos datos de calidad. Además, se ha observado la distribución de los valores para los campos de los distintos conjuntos y se ha estudiado formas para la agregación de los datos, y así, obtener otros conjuntos que puedan sernos de gran ayuda para las soluciones que se presentarán más adelante.

Se ha iniciado el análisis con la carga de los conjuntos del cuestionario, sensor y ssid’s de Alcoi. Para el cuestionario, se hizo un arreglo para el número de personas en la oficina. Este arreglo consiste sumar 1 a dicho campo (num\_personas\_oficina), pues en el origen no se cuenta a la persona que realiza el cuestionario. En nuestro caso se contará, ya que será el campo que se utilice para contabilizar las personas que entran a la oficina de turismo.

Para futuras agregaciones (diarias) se ha creado un nuevo campo date que no estaba presente en los orígenes de datos. Esta operación se realiza tanto para el cuestionario como para el sensor.

Realizados estos dos arreglos iniciales, los dos conjuntos principales (sensor y cuestionario) tienen el siguiente aspecto:

1. Interfaz de usuario gráfica, Aplicación

   Descripción generada automáticamentePorción del conjunto del cuestionario:

Figura 4. Conjunto del cuestionario

Figura 5. Conjunto del cuestionario

1. Texto

   Descripción generada automáticamentePorción del conjunto del sensor:

Figura 6. Conjunto del sensor

Por otro lado, tenemos el conjunto de las ssid’s recogidas de Alcoi que nos servirá para hacer una primera aproximación a los visitantes recogidos por el sensor.

* Porción del conjunto de ssid’s:

Texto

Descripción generada automáticamente

Figura 7. Conjunto de ssid's de Alcoi

Más adelante se hablará de como este conjunto es crucial para de los visitantes detectados por el sensor.

La siguiente parte del análisis es la gestión de los nulos. Para empezar, se hace un barrido por los campos de los conjuntos y se encuentra que en el **cuestionario** el campo de origen tiene 520 nulos, en este caso no hay problema, pues que sea nulo depende directamente de la pregunta anterior del cuestionario. Concretamente, si en la anterior pregunta (segunda pregunta): "¿Cuál es su lugar de residencia?" se contesta Alcoi, la respuesta a la siguiente pregunta (tercera: “En el caso de que haya marcado anteriormente que su lugar de residencia NO es Alcoi. ¿Podría especificar de dónde es? (país, provincia, municipio...)”) figurará vacía, ya que se deberá especificar el origen en ese caso. Por tanto, lo que se va a realizar es un relleno de los campos nulos con "Alcoi".

Por otro lado, en el conjunto inicial del **sensor**, se observa que gran parte de los nulos cae sobre el campo visitor, de lo cual no hay de que preocuparse porque es nuestra tarea principal la estimación de este. También se ha encontrado algunos registros (3 en total) en los que en los campos de ssid y mac contienen algún nulo. Se procederá a eliminar dichos registros, la repercusión que genera esta eliminación no se puede comparar con la que generaría mantener los nulos.

Siguiendo con el análisis, nos damos cuenta de que es necesario realizar un filtro horario tanto para el sensor como para el cuestionario. El filtrado se realiza en dos pasos que consisten en:

* **Paso 1:** Se adaptan ambos conjuntos para que coincidan en fecha inicio y fecha fin. En este caso las fechas son 21 de diciembre del 2020 y 7 de julio del 2021
* **Paso 2:** Se realiza un filtrado de ambos conjuntos para que coincidan en horario. Esto se realiza porque el sensor siempre esta activo y la oficina de turismo no siempre está abierta, esta tiene sus **horarios**:
  + Lunes a viernes: 10-14h.
  + Sábados, domingos y festivos: 10-14h.
  + Los días 4 de enero y 5 de enero abre de 10-14 y 17-19h.
  + Festivos cerrados: 1 y 6 de enero, 1 de mayo y 25 de diciembre.

En todo momento se adapta el conjunto menos restrictivo (en lo que a fechas se refiere) al más restrictivo.

Hacer un análisis exploratorio también nos ayuda a conocer la distribución de los valores de los campos. Esto nos pueda ayudar a saber qué nos podremos encontrar más adelante e incluso nos pueda ayudar a detectar alguna anomalía.

En Python es bastante sencillo obtener la distribución de los valores de un campo, se realiza de siguiente forma: df[campo].value\_counts().sort\_values(). La última función sirve para ordenar los valores según la frecuencia en la que aparecen. A continuación, se comentarán algunas de las distribuciones de los campos más relevantes:

* Cuestionario
  + origen: Se observa una clara mayoría del número de registros cuyo origen Alcoi o Valencia (351 y 250 respectivamente), en tercer lugar, está Alicante con 65 registros.
  + num\_personas\_oficina: Nos damos cuenta de que las personas suelen entrar a la oficina en pareja o en solitario (522 y 454 respectivamente)
  + wifi\_abierta\_tlf: Con esta distribución sabremos cuantas personas de las que rellena el cuestionario tenían la wifi activa o no. Como se detallará más adelante, este campo formará parte del algoritmo discriminador. En este campo tiene 3 valores posibles, Sí (tiene wifi activa), No (no la tiene) y No entiendo la pregunta. La distribución de los valores es 675, 321 y 118 respectivamente
* Sensor
  + visitor\_v2: Forma parte del a **primera aproximación hacia el algoritmo**. Se detallará más adelante. La distribución es la siguiente: 18816 (Y) y 28995 (N)
  + ssid: campo que contiene la ssid asociada a cada mac en un instante determinado.
  + mac: Este es el campo que nos permitirá contar las personas detectadas por el sensor.

Antes de pasar con el algoritmo, se mostrarán dos gráficas que forman parte de este análisis, y que han ayudado a la decisión de realizar un clustering como siguiente paso del algoritmo.

Con la primera se observan las ssid’s ordenas en el eje x (donde figuran sus id’s) en función de sus frecuencias. Nótese que en las últimas ssid’s las frecuencias empiezan a subir de manera acentuada. **Probablemente** sea debido a que existen personas que estén constantemente en presencia del sensor, por lo que podrían trabajadores de la propia oficina, o si tuvieran mác’s dinámicas podrían pertenecer al mismo dispositivo que se conecta a la misma ssid.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 8. Gráfica de frecuencias de las ssid's

La siguiente gráfica se ha realizado con la misma idea que la anterior pero esta vez con las mac’s. Como se podrá observar también existen unos pocos registros que muestran una alta frecuencia.

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

Figura 9. Gráfica de frecuencias de las mac's

### Análisis diario

La manera de medir los sucesos de los conjuntos va a ser mediante el uso de series temporales. Estas series usarán como unidad tiempo el día. Para este proyecto estas series son una herramienta muy útil para saber lo que nos cuentan los datos.

Para empezar, se ha de adaptar los datos para que sean aptos para su incorporación a las series. Como ya se ha comentado antes, se va a medir principalmente las mac’s del sensor y las personas que entran en la oficina de turismo para rellenar el cuestionario.

Aunque inicialmente (como objetivo exploratorio) se va a tener en cuenta tanto residentes como visitantes, el objetivo fundamental y final de las series es la medición de los visitantes.

Antes de seguir, se va a mostrar las consideraciones que se han tomado a la hora de confeccionar las series. Estas dos series podría considerar un boceto de las que vendrán continuación.

Esta ha sido la forma de realizar el conteo de las personas detectadas por el sensor. Se ha realizado un agregado diario del número de mac’s **distintas** detectadas para día. No se va a considerar las mac’s repetidas en un mismo día porque se pretende contar tan solo los visitantes nuevos que acudan a la oficina.

"""

Con mac's únicas

"""

df\_sensor\_d2\_dist = df\_sensor\_d.groupby('date').mac.nunique()

df\_sensor\_d2\_dist = df\_sensor\_d2\_dist.to\_frame()

df\_sensor\_d2\_dist.reset\_index(level=0, inplace=True)

df\_sensor\_d2\_dist.head(5)

Se observa un pequeño fragmento de la serie inicial del sensor. En ella se observa las mac’s agregadas por día.

Imagen de la pantalla de un celular de un mensaje en letras blancas

Descripción generada automáticamente con confianza baja

Figura 10. Serie del sensor inicial

Por otro lado, la implementación para crear la serie del cuestionario se muestra en la porción código de abajo. Como se verá más adelante se ha cogido tan solo personas oficina, porque el estudio se centrará tan solo en las personas que entraron a esta.

"""

Se cuenta el número de personas en la oficina.

Como se verá más adelante, se va a escoger la potencia mínima e intermedia del sensor por lo que

no detectará a personas de fuera de la oficina.

"""

df\_cuest\_d2 = df\_cuest.copy()

df\_cuest\_d2 = df\_cuest\_d2[['date', 'num\_personas\_oficina']].groupby(['date'], as\_index=False).sum()

Una parte de la serie inicial del cuestionario es la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 11. Serie del cuestionario inicial

Aunque no lo parezca, estas dos series presentadas son muy importes para la presente investigación, pues conforman las bases de la lógica aplicada a la contabilización de campos objeto de estudio.

La decisión de escoger para el cuestionario tan solo las personas que entran en la oficina, viene condicionada por la siguiente gráfica y estudios posteriores a esta.

Gráfico de dispersión

Descripción generada automáticamente

Figura 12. Rangos potencia del sensor

Las áreas de distinto color representan diferentes niveles de potencia del sensor estudiado. Como se precia, estas áreas cambian a lo largo del tiempo.

Roja: Alta potencia.

Azul: Baja potencia.

Verde: Potencia intermedia.

Con esta visualización ya se puede escoger descartar aquellos registros recogidos con la potencia del sensor alta. Es decir, para los siguientes estudios se tomará como fecha mínima de serie el 10 de mayo (inicio zona azul).

## El algoritmo: Pasos comunes

A lo largo de la investigación se han desarrollado distintas soluciones (algoritmos) para poder aproximar las aportadas por el cuestionario. A pesar de obtener resultados distintos, todas las soluciones comparten una serie de pasos que veremos a continuación. Más adelante se explicará con detalle el algoritmo completo escogido.

### Primer paso: discriminación con la lista de ssid’s

Anteriormente se habló del conjunto de ssid’s de Alcoi, este conjunto es una pieza muy importante para la construcción del algoritmo de detección de visitantes del sensor. Recordemos que el conjunto contiene todas la ssid’s de Alcoi, y que como mediante el sensor también se recogen las ssid’s favoritas a las que se conectan, entonces es posible aproximar si un registro es visitante o no. Esto se realiza comprobando si la ssid que aparece en un registro pertenece a lista de ssid’s recogidas correspondientes a Alcoi (conjunto de ssid’s de Alcoi). Para señalar visitantes y residentes se ha creado un campo visitor\_v2 con Y si es visitante y N si es residente. Esto se podría implementar fácilmente de la siguiente forma:

list\_ssid = df\_ssid\_Alcoi['ssid'].tolist()

def f(row):

if row['ssid'] in list\_ssid:

val = 'N'

else:

val = 'Y'

return val

df\_sensor['visitor\_v2'] = df\_sensor.apply(f, axis=1)

### Segundo paso: Clustering de frecuencias

Imagen que contiene Gráfico

Descripción generada automáticamentePara las frecuencias de ambos campos (ssid y mac), se ha realizado el clustering mediante KMeans. Según la regla del codo implementada a partir de la obtención del SSE (Sum of squared error) o la inercia. El método es sencillo: se prueban distintas combinaciones en el parámetro K (número de clusters) y por cada una se calcula la inercia. Esta inercia es la distancia o el error de cada punto a su centroide.

Figura . Método del codo (SSE)

Como se observa, por el SSE obtenido y la naturaleza del problema (potenciales trabajadores o no) se escogen dos clusters. Como se verá más adelante se crearán dos clusters: Uno con frecuencias altas y otro con bajas frecuencias. El método del codo se utiliza en estos casos para dererminar el valor más óptimo de K, estes el apunto anterior la dismunición progresiva de la pendiente.

Este clustering se ha realizado simplemente para que no sea una persona la que escoja las frecuencias, a partir de las cuales se debe considerar desechar las mac’s o las ssid’s que potencialmente sean empleados de la oficina o alrededores.

A continuación se expone la implementación del clustering y una posterior implementación manual del aumento de la tolerancia. Esto es para que el cluster de altas frecuencias sea más permisivo:

km2 = KMeans(n\_clusters=2, init='k-means++', n\_init=10, algorithm='auto')

y\_km2 = km2.fit\_predict(df\_aux2[['counts']])

df\_y\_km2 = pd.DataFrame(y\_km2, columns = ['counts\_clust'])

# clusters

df\_y\_km2['counts\_clust'].value\_counts().sort\_values()

df\_y\_km2['index'] = np.arange(df\_y\_km2.shape[0])

df\_y\_km2.tail(10)

num = len(df\_y\_km2.query('counts\_clust != 0')['index'].to\_list())

################ Para relajar el cluster #################

tol = 0 # mayor tol, más macs eliminadas

##########################################################

num = num + tol

lista\_ids\_reps = df\_y\_km2.tail(num)['index'].to\_list()

lista\_macs\_reps = df\_aux2[df\_aux2['mac\_id'].isin(lista\_ids\_reps)].mac.to\_list()

En el ejemplo de arriba, la lista lista\_macs\_reps contiene las mac’s que se eliminarán del conjunto del sensor. Esta acción consigue rebajar lo suficiente (cantidad mínima) las cifras del sensor para que ahora se ajusten más a lo recogido por el cuestionario. Al estar el conjunto de frecuencias ordenado de menor a mayor, el resultado del clustering para las mac’s y las ssid’s resulta sencillo de comprobar, pues las id’s de las mac’s y las ssid’s con frecuencia elevada está situadas el final del conjunto y pertenecen al cluster 1.

Las siguientes figuras muestra la asignación por el clustering realizado:

Pantalla de computadora con letras

Descripción generada automáticamente con confianza mediaTabla

Descripción generada automáticamente con confianza media

Figura . Resultado del clustering para las mac's

Figura . Resultado del clustering para las ssid's

Con estos datos procede a elaorar una s lista de exclusion de mac’s y ssid’s. A continuación se muestra el pequeño cambio en los datos:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura . Mac's únicas (sensor) y personas en la oficina (cuest.) antes de exclusión

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura . Mac's únicas (sensor) y personas en la oficina (cuest.) después de exclusión

Se observa como la curva de las mac’s (color verde) dismunuye un poco.

## El algoritmo: Pasos de la solución escogida.

En esta sección se va a desarrollar los dos ultimos pasos de la solución escogida. Se seleccionó esta aproximación por la naturalidad de la misma, su justificación lógica y sus resultados.

### Tercer paso: Obtención de un conjunto con conocimiento a priori.

A continuación, se va a detallar las pequeñas etapas que componen este paso. El objetivo es la obtención de un conjunto que incorpore conocimiento a priori. Este conocimiento trata de las cifras de los visitantes mensuales recogidas en 2019. Como ya se ha visto anteriormente, se va a escoger los meses de mayo, junio y julio.

* Primera etapa: se obtienen las fechas de dos años atrás (2019), así como el número de mes de las fechas de ambos años.
* Segunda etapa: **Agregación mensual de los visitantes recogidos por el cuestionario,** esto se puede realizar por la etapa anterior. Recordemos que inicialmente el conocimiento nos proporciona los visitantes mensuales del año 2019.
* Tercera etapa: **Corrección de las medias mesuales del conjunto de 2019.** Se ha de adaptar aproximadamente a los días en los que realmente se recogen visitantes para el conjunto de 2021. Recalcar que estas series empiezan el 10 de mayo y acaban el 7 de julio, por lo que habría que realizar una pequeña corrección en 2019, pues en ese años las cifras pertenecen esos 3 meses completos.
* Cuarta etapa: **Uso del factor covid.** Es lógico pensar en la influencia que tiene el covid en las visitas. En esta etapa se trata de usar el factor mensual proporcionado por la oficina de turismo de Alcoy que pueda relacionar las visitas del 2019 mensuales con las 2021 mensuales. Este es 0.05, 0.1 y 0.1 en los meses de mayo, junio y julio.
* Quinta etapa: **Añadido de la componente seasonal.** Mediante la función de python seasonal\_decompose, se obtiene la componente estacionaria del cuestionario. Como se realizado mediante una descomposición aditiva, mediante algunos arreglos, el sensor se podrá beneficiar de ella.
* Sexta etapa: **Suma de la estacionalidad y arreglo diario del 2019 por el factor covid.** El arreglo diario no es más que la division entre la media mensual corregida (por el f. covid) y el número de días considerados para ese mes, a eso se le suma la componente estacionaria. Llamaremos a este campo obtenido visitantes\_dia\_media\_season\_corr.

### Cuarto paso: Aplicación a los datos del sensor el conocimiento a priori

Estas dos etapas aplican el conocimiento a priori a los registros del sensor.

* Séptima etapa: **Aplicación del ratio de las personas sin wifi.** En secciones anteriores se comentó que el sensor solo detectaba los dispositivos cuya conexión wifi estuviera activa. Se ha obtenido el ratio (a partir del cuestionario) aproximado de personas que usan y no usan dicha conexión. Gracias a estapas anteriores, tenenos una estimación de los visitantes del cuestionario con el campo visitantes\_dia\_media\_season\_corr. Por lo que se procede obtener el valor aproximado de las personas que entraron sin wifi activa a la oficina, y así sumárselo al sensor. A esta corrección se le suma los visitantes registrados por el sensor. A continuación se observa las proporciones obtenidas en el campo wifi\_abierta\_tlf:

Gráfico, Gráfico circular

Descripción generada automáticamenteInterfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Figura . Proporcion de personas con wifi en la oficina

* Octava etapa: **Corrección mediante factores externos.** Existen otros factores que influyen en la serie obtenida pero que con los medios disponibles son difíciles de calcular si no se dispone de la curva de referencia. Estos factores se aproxima por uno solo obtenido a partir de la media de los cocientes de la division entre los visitantes reales y el último valor de nuestra aproximación a ellos. Como se verá más adelante, al ser una media, la correlación entre los valores resultado (antes y despues de la opeación) y los valores reales no cambia.

# Resultados

En las secciones anteriores se han detallado los pasos y cada una de sus etapas que se han seguido para obtener los resultados que se verán continuación. Se realizará una comparación visual de ajuste y una comparación entre las correlaciones de las curvas obtenidas antes y después de los pasos seguidos. Las siguientes gráficas muestran los resultados alcanzados en el rango de fechas en el que el sensor estaba configurado con la potencia baja.

En esta gráfica se observa como el sensor no recibe apenas información de los dispositivos (más adelante se **discutirá** esto):

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura . Resultados Antes de algoritmo

Gráfico, Histograma

Descripción generada automáticamenteEn la siguiente visualización, vemos como se consigue lidiar con esa falta de sensibilidad del sensor en el **mismo** rango de fechas (algoritmo escogido).

Figura . Resultados después de algoritmo

Para comprobar los resultados de forma numérica, se ha utilizado la **correlación** de Pearson. La que se ha obtenido los datos la figura 13 *Resultados Antes de algoritmo* entre lo detectado por el cuestionario y lo detectado por el sensor es de 0.013.

Mientras que la correlación obtenida de la figura 14. *Resultados después de algoritmo* entre lo detectado por el cuestionario y lo detectado por el sensor es de **0.55**. Las gráficas que muestran estos resultados son:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 5. Plot de la correlación antes del algotimo. Pot. mínima

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 16. Plot de la correlación después del algotimo. Pot. mínima

Como se observa, la correlación es más acertada después del algoritmo. Ahora se mostrarán los resultados obtenidos si se añadiera la potencia intermedia (del 22 de junio al 7 de julio). Al añadir este rango de fechas en las visualizaciones se observa un gran cambio en esa parte en la gráfica análoga a la figura 13. De esta forma el sensor es capaz de detectar más información, pero existen unos inconvenientes que se **discutirán** más adelante.

Interfaz de usuario gráfica, Gráfico, Histograma

Descripción generada automáticamente

Figura . Resultados antes de algorimo. Pot. Intermedia añadida

En la gráfica mostrada se observa un crecimiento de los visitantes detectados por el sensor en el rango de fechas en el que la potencia es intermedia.

Si añadimos también ese rango de fechas los datos conseguidos por el algoritmo se obtendría la gráfica siguiente:

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

Figura . Resultados después de algoritmo. Pot. Intermedia

Los resultados de esta representación sugieren la idea de que existe una excedencia de visitantes detectados por el sensor tomando como referencia los recogidos por el cuestionario. Esto se **discutirá** también en la sección correspondiente.

En el primer rango de fechas (10 de mayo a 22 junio) parece existir días en los que no se detecte ningún dispositivo, pero en realidad sí que se están detectando. Inicialmente (antes de algoritmo) no existe una discriminación entre visitantes y residentes elaborada y esto también puede ocasionar las cifras tan bajas obtenidas en el primer rango. A continuación, se mostrará una grafica con los dos rangos de fecha en la que también se contabilizan los residentes iniciales:

Gráfico, Histograma

Descripción generada automáticamente

Figura . Resultados antes de algoritmo. Pot. Intermedia

Como se observa, en el primer rango hay una falta de detección de visitantes por parte del sensor, y es por esto por lo que nos decantamos por emplear el conocimiento a priori para compensar esta falta de información.

Al utilizar la correlación para medir los resultados de después de aplicar el algoritmo añadiendo la potencia intermedia, esta correlación baja su valor. Es decir, al aplicar la correlación entre los datos de la figura 18 obtenemos un valor de 0.24, y si se aplica solamente al rango en el que se obtienen los datos con potencia intermedia es de -0.062. Esto es debido a la excedencia de las cifras al usar dicha potencia. Más adelante habrá una **discusión** sobre una posible solución y la posibilidad de elección de un rango de potencia mayor en la **propagación** presente proyecto.

## Discusión

A partir de los resultados anteriores se elabora una discusión en la que también se mencionan posibles soluciones que podrían mejorar los resultados obtenidos en la elección final.

La falta de información en el primer rango sugiere la elección del segundo, pero la excedencia en las cifras del sensor respecto a las del cuestionario, disminuyen la calidad de la correlación. Como es de esperar, si añadimos el segundo rango al primero la correlación, es decir si tenemos en cuenta el periodo desde el 10 de mayo hasta el 7 julio, la correlación un aumenta lo suficiente. Para atajar el problema de la ya comentada excedencia se ha pensado en que se podría hacer de dos formas:

Forma 1: Eliminación de outliers en las cifras recogidas por el sensor, esto consiste en eliminar registros completos (días) en los que el sensor obtiene valores que podrías ser catalogados como anomalías.

Forma 2: Se podría añadir en el algoritmo que los outliers detectados en los datos del se remplacen por la media del día anterior y el siguiente.

Con estas dos formas se podrían eliminar o rebajar esos valores que exceden con creces los del cuestionario.

Una de las razones principales por la que se ha escogido el primer rango como solución final, es la de que al aplicar el algoritmo sólo en ese rango se obtiene una correlación aceptable. Otra de las razones es por el hecho de la practicidad que aporta no superar los valores del cuestionario, y es la obtención de una **cota inferior** de lo que realmente está sucediendo. Es decir, de esta forma se puede afirmar que como mínimo se va a obtener un número de visitantes determinado.

### Otra forma de solventar la wifi inactiva

De entre a las aproximaciones implementadas, merece la pena destacar una en la que se nos hicimos la siguiente pregunta: ¿Qué pasaría si todos los individuos tuvieran wifi? En teoría, si todos los individuos tuvieran wifi, el sensor sería capaz de detectarlos. Por ello, se hizo una aproximación que simulara este suceso. Esto se simuló simplemente de la siguiente forma:

df\_comparacion['ratio\_inc\_wifi\_vs'] = df\_comparacion['visitantes\_c'] \* 0.402

df\_comparacion['new\_visitantes\_s'] = df\_comparacion['visitantes\_s'] + df\_comparacion['ratio\_inc\_wifi\_vs']

Como se ve, y se discutirá más adelante, se recoge los encuestados que no tienen la wifi activa. Después, estas cifras se añaden a las resultantes del sensor. De esta forma, la estacionalidad no hace falta añadirla, pues ya forma parte de la adición realizada. Los pasos seguidos son los dos primeros: **Discriminación con lista de ssid’s y Clustering de frecuencias**. Después de estos pasos, se implementa la porción de código mostrado más arriba y más adelante una corrección de ajuste usando la media de los cocientes de nuevo (como en otras aproximaciones). Esto se hace para simular la perturbación en los datos que puedan acarrear otros factores externos.

Gráfico, Histograma

Descripción generada automáticamente

# Bibliografía

[1] L. Schauer, M. Werner, and P. Marcus, “Estimating crowd densities and pedestrian flows using Wi-Fi and bluetooth,” 2014, doi: 10.4108/icst.mobiquitous.2014.257870.

[2] X. Liu, J. Wen, S. Tang, J. Cao, and J. Shen, “City-Hunter: Hunting Smartphones in Urban Areas,” 2017, doi: 10.1109/ICDCS.2017.148.

[3] U. Singh, J. F. Determe, F. Horlin, and P. De Doncker, “Crowd Monitoring: State-of-the-Art and Future Directions,” *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*. 2020, doi: 10.1080/02564602.2020.1803152.

[4] U. Gretzel, M. Sigala, Z. Xiang, and C. Koo, “Smart tourism: foundations and developments,” *Electron. Mark.*, 2015, doi: 10.1007/s12525-015-0196-8.

[5] J. A. Ivars-Baidal, M. A. Celdrán-Bernabeu, J. N. Mazón, and Á. F. Perles-Ivars, “Smart destinations and the evolution of ICTs: a new scenario for destination management?,” *Curr. Issues Tour.*, 2019, doi: 10.1080/13683500.2017.1388771.

[6] “Libro Blanco de Destinos Turísticos Inteligentes SEGITTUR.” https://www.segittur.es/destinos-turisticos-inteligentes/proyectos-destinos/libro-blanco-destinos-turisticos-inteligentes/ (accessed Sep. 5, 2021).

[1][2][3][4][5] [6]

1. <https://www.alcoyturismo.com/> [↑](#footnote-ref-2)
2. <https://smartcity-alcoi.com/> [↑](#footnote-ref-3)