Universitat Oberta de Catalunya (UOC)

Màster Universitari en Ciència de Dades (*Data Science*)

PROJECTE DE FINAL DE MÀSTER

Àrea: Sistema de posicionament en interiors basat en l'aplicació de Machine Learning sobre senyals Wi-Fi

Implantació d'un sistema de geoposicionament de dispositius mòbils en interiors.

Cas d'ús: Monument de Casa de la Vall (Principat d'Andorra)

Autor: Josep Ribó Férriz

Responsable: Albert Solé Ribalta

Director: Joaquín Torres Sospedra

Andorra, 15 de gener de 2023

Crèdits/Copyright

FITXA DEL PROJECTE FINAL DE MÀSTER

Títol del projecte:	Implantació d'un sistema de geoposicionament de dispositius mòbils en interiors. Cas d'ús: Monument de Casa de la Vall (Principat d'Andorra)
Nom de l'autor:	Josep Ribó Férriz
Nom del director:	Joaquín Torres Sospedra
Nom del responsable:	Albert Solé Ribalta
Data de lliurament:	01/2023
Titulació:	Màster oficial en ciència de dades
Àrea del treball final:	Sistema de posicionament en interiors basat en l'aplicació de Machine Learning sobre senyals Wi-Fi
Llengua del treball:	Català
Paraules clau	Wi-Fi Sensors, Probe Request, IoT, Indoor Localization, Tracking, Fingerprint, Supervised Machine Learning, RSSI, ESP32, Smart Phones, Museum

Dedicatòria

- Als meus fills Alèxia i Pol i a la meva dona Carina, per la seva paciència, i per les hores que m'han permès dedicar a aquesta estranya afició.
- A la meva mare Meritxell pel seu suport, i al meu pare Pepito (EPD), perquè estic convençut que hauria gaudit llegint la memòria d'aquest projecte.

Agraïments

- A la Sindicatura (VIII Legislatura 2019-2023) del Consell General d'Andorra, per l'interès a apropar la ciència de dades i la innovació, a l'edifici que l'any 1702 va convertir-se en la seu d'un dels parlaments més antics i amb més continuïtat d'Europa, creat l'any 1419.
- Al Departament de museus i monuments nacionals del Ministeri de Cultura i Esports del Govern d'Andorra, per haver-me facilitat les estadístiques dels períodes analitzats en aquest estudi, i als seus/ves guies, per resoldre els dubtes que els hi he plantejat.

Abstract

The need to have Internet connectivity means that people usually use a mobile device with access to the network, using wireless connectivity (Wi-Fi).

To solve this need, our mobile devices are continuously searching for Wi-Fi signals to connect to, sending discovery frames, known as *Probe Request* (IEEE 802.11 wireless protocol).

The analysis of privacy weaknesses in discovery frames is well-known, and has been used on numerous occasions, to explore user tracking in different situations (museums, fairs, universities, urban mobility...).

This means that the analysis and evaluation of the discovery plots, allows conclusions of positioning and movement to be drawn, without the need to have access to the GPS systems of mobile devices. It should be remembered that the main technologies are devoting a lot of effort to avoid the vulnerability that the analysis of the set of factors surrounding the *Probe Request* frames can cause in the field of privacy.

Considering that the devices send this anonymous information constantly, the aim of this study is to explore the temporary presence of visitors in the different areas inside Casa de la Vall building, and the non-cooperative monitoring of the routes they take, using sensors and *Machine Learning* algorithms, field of artificial intelligence, to capture and geopositionate the frames emitted by their mobile devices, without the need to have access to the GPS service.

Keywords: Wi-Fi Sensors, Probe Request, IoT, Indoor Localization, Tracking, Fingerprint, Supervised Machine Learning, RSSI, ESP32, Smart Phones, Museum

Resum

La necessitat de disposar de connectivitat a Internet, provoca que habitualment, les persones fem ús d'un dispositiu mòbil amb accés a la xarxa, utilitzant connectivitat sense fil (wifi).

Per resoldre aquesta necessitat, els nostres dispositius mòbils estan cercant contínuament senyals wifi a les quals connectar-se, enviant trames de descobriment, conegudes com a *Probe Request* (protocol IEEE 802.11).

L'anàlisi de les debilitats de privadesa de les trames de descobriment és coneguda, i s'ha utilitzat en nombroses ocasions, per explorar el seguiment d'usuaris en diferents situacions (museus, fires, universitats, mobilitat urbana...).

Això vol dir que, l'anàlisi i avaluació de les trames de descobriment, permet extreure conclusions de posicionament i moviment, sense necessitat de disposar d'accés als sistemes GPS dels dispositius mòbils. Cal recordar que les principals tecnològiques estan dedicant molts esforços per evitar la vulnerabilitat que l'anàlisi del conjunt de factors que envolten a les trames *Probe Request*, pot provocar en l'àmbit de la privadesa.

Tenint en compte que els dispositius envien aquesta informació anònima de forma constant, l'objectiu d'aquest estudi és explorar la presència temporal de visitants en les diferents zones de l'interior de l'edifici de Casa de la Vall, i el seguiment no cooperatiu de les rutes que aquests realitzen, utilitzant sensors i algorismes de *Machine Learning*, camp de la intel·ligència artificial, per capturar i geolocalitzar les trames emeses pels seus dispositius mòbils, sense necessitat de disposar d'accés al servei GPS.

Paraules clau: Wi-Fi Sensors, Probe Request, IoT, Indoor Localization, Tracking, Fingerprint, Supervised Machine Learning, RSSI, ESP32, Smart Phones, Museum

Índex

Índex de figures

Índex de taules

List of Algorithms

Capítol 1

Introducció

1.1 Motivació

La problemàtica de geolocalitzar a les persones, es resol disposant utilitzant el servei GPS. Però en l'actualitat, aquest servei no és viable en espais interiors, i a més, requereix la infraestructura que permeti obtenir la seva ubicació, de forma precisa. Des de l'any 2000, amb el llançament de *RADAR* [7] per part de *Microsoft research*, s'ha fet ús de la tecnologia wifi per a aconseguir la ubicació de persones, en espais interiors. Hi ha interès, per part de la comunitat científica, no sols en posicionar persones de forma activa, sinó també de forma passiva.

L'interès d'aquest treball radica en la realització d'un estudi innovador, tant en el vessant tecnològic, com en l'ús de mètodes analítics d'intel·ligència artificial, dins d'un monument històric com és Casa de la Vall [4], per avaluar les possibilitats actuals de detecció, de seguiment de presència passiva, i posicionament de dispositius mòbils en interiors, basant-se en la hipòtesi que tot i que l'aleatorització de l'adreça MAC dels dispositius mòbils és suficient per protegir la privadesa [5] de l'usuari, no és suficient [8] per emmascarar informació estadística d'interès.

L'arquitectura amb solers de fusta del monument de Casa de la Vall, pot permetre geoposicionar els dispositius, no només en el pla horitzontal, sinó també de forma tridimensional. Aquest darrer pretén ser un element diferenciador d'aquest estudi, respecte a articles ja publicats.

Figura 1.1: Casa de la Vall. Seu de l'antic parlament d'Andorra

1.2 Objectius

Les fites en què es divideix l'estudi, ordenades de forma cronològica, són:

- L'obtenció d'una col·lecció de dades, extretes de les trames de descobriment *Probe Request*, emeses pels dispositius mòbils dels visitants, i recopilades pels sensors desplegats pel monument.

- La neteja [20], processament i anàlisi de les dades utilitzant algorismes de *Machine Learning*, que han de permetre correlacionar les diferents trames de descobriment, tot i l'aleatorietat introduïda de forma automàtica, per la major part dels dispositius.

- L'obtenció d'un model de detecció de presència i seguiment (temps mitjà de la visita, recorregut realitzat, temps dedicat a cada zona, zones més visitades, afluència [21] [20] segons horari i dia...) dels visitants del monument.

- L'avaluació del percentatge de dispositius que no adopten l'aleatorització MAC [1], perquè o no està implementada en els seus dispositius, o bé l'han desactivada, i la seva evolució els darrers anys. [9]

- La qualitat de la detecció de presència no consensuada, tot i emprar l'aleatorització MAC.

- La comparació entre la simplicitat de fer el seguiment a dispositius que no fan servir l'aleatorització d'adreces MAC, i la dificultat de realitzar el mateix seguiment a dispositius amb aleatorització MAC activa.
- La redacció d'una memòria de defensa de l'estudi, amb les conclusions tècniques i estadístiques que permetin avaluar la metodologia utilitzada i els resultats obtinguts.
- La redacció i publicació, en el cas d'assolir conclusions que permetin ampliar el coneixement en l'àmbit de l'estudi, d'un article científic.

1.3 Resum dels treballs relacionats

Tot i que, des de 2014, els fabricants estan emprant mesures relacionades amb la privadesa dels senyals dels dispositius mòbils, concretament de l'estàndard IEEE 802.11, com l'aleatorització [1] d'adreces MAC, existeix un gran interès dins de la comunitat científica de publicar articles relacionats en aquest àmbit. Cal destacar, entre altres dominis:

- Seguiment passiu i no cooperatiu de l'ocupació temporal, en espais interiors. [14] [26]
- Mesura dels indicadors de recepció del senyal RSSI. [10]
- Anàlisi de la mobilitat urbana. [12] [2]
- Riscos i amenaces a la privadesa dels usuaris, que utilitzen serveis de xarxes sense fils. [11]

Existeixen diferents factors que poden afectar l'efectivitat del monitoratge dels paquets de descobriment wifi, com són:

- Tipus d'aleatorització provoca soroll a les dades que intenta alterar la seva anàlisi.
- Freqüència de sondeig i nombre mitjà de sondes emeses.

També existeix una extensa bibliografia associada al microcontrolador *ESP32* per detectar els paquets de descobriment, recollir-los, i analitzar-los. [8]. En general, la bibliografia consultada, que té com a objectiu l'anàlisi de la informació que publiquen els dispositius mòbils, no pretén identificar o revelar la identitat dels usuaris, ni de cap manera, invertir l'aleatorització de l'adreça MAC, però evidencien que els esforços en desenvolupar alternatives per preservar la privadesa, no han aconseguit eliminar l'amenaça provocada pels paquets de descobriment wifi. Tot i que en la fase d'anàlisi, s'ha realitzat una extensa consulta a l'estat de l'art de projectes d'aquest àmbit, que han servit de guia per a desenvolupar aquest treball en un entorn real, cal tenir en compte que per l'especificitat del treball, s'ha efectuat un important esforç per adaptar, de forma dinàmica, els mètodes utilitzats, als resultats i situacions que s'han anat produint.

Capítol 2

Materials i mètodes

En aquest punt es presenta la problemàtica a resoldre, les tecnologies al nostre abast, la metodologia de recerca i algorismes de *Machine Learning*.

2.1 Posicionament en interiors

Durant l'última dècada, s'està avançant en el camp del posicionament i navegació en interiors, utilitzant tecnologies sense fil, tenint en compte que és més complex que en exteriors, ja que no es poden utilitzar els habituals sistemes globals de navegació per satèl·lit, donada la incapacitat dels senyals radioelèctrics de penetrar dins dels edificis. En aquest estudi, ens centrarem en la comunicació amb protocols wifi, analitzant les trames de gestió que fan servir els dispositius per descobrir, gestionar i optimitzar la connexió amb els punts d'accés (WAP).

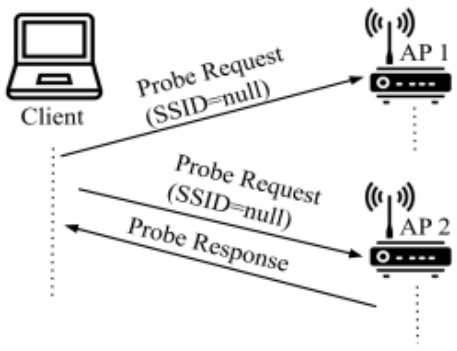

Figura 2.1: Esquema simplificat del protocol de descobriment dels dispositius mòbils

El principi del posicionament mitjançant WAP [18], es basa en el fet que el nivell del senyal percebut (RSSI) des dels sensors de l'entorn del visitant, és una funció de la seva posició. Cada sensor rep les sol·licituds *Probe Request* amb la intensitat del senyal corresponent, fins al punt

que alguns sensors no visibles, poden ignorar l'enviament del paquet. Com que el valor RSSI és un indicador de la posició relativa, respecte al sensor corresponent, esdevindrà el paràmetre principal, en què es basarà el càlcul de la ubicació del visitant.

Cal recordar que les xarxes wifi ja usades per un dispositiu, són afegides a la llista de xarxes conegudes o preferides del dispositiu. Un efecte secundari de fer servir aquests tipus d'exploració de xarxes conegudes, és la capacitat d'adversaris de recollir les trames d'intent de connexió, que contindran els paràmetres de les xarxes conegudes.

Un dels canvis més importants per a millorar la privacitat de l'intercanvi de trames de descobriment, va ser la introducció de l'aleatorització d'adreces MAC [1] l'any 2014, però la seva implementació encara no és perfecta, ja que encara es filtra informació que permet el seguiment dels dispositius. Alguns investigadors continuem centrar-se a explorar diferents maneres d'evitar les mesures de privadesa que es van introduint.

2.1.1 Mesures de prova

La situació física en la qual s'han distribuït els sensors d'aquest estudi, ha estat condicionada pels punts d'alimentació elèctrica existents, alhora que hem assegurat el menor l'impacte visual possible. Per aquest motiu, la distribució no respecta l'equidistància, tot i que sí que s'han tingut en compte criteris que permetessin al màxim el posicionament en les zones visitables.

El monument de Casa de la Vall està distribuït en tres plantes: planta baixa, primera planta i segona planta o sota coberta. Cal dir que només són d'accés públic, algunes zones de la planta baixa i la primera planta, d'aproximadament 300 metres quadrats, que han estat dividits en vuit zones o sales. S'han implantat deu sensors, distribuïts entre les tres plantes: Planta baixa dos sensors, primera planta cinc sensors i segona planta, tres sensors. Queden inclosos passadissos, escales i diferents vestíbuls. S'ha realitzat la calibració de quaranta-cinc llocs, amb un mínim de quatre calibracions per sala o zona.

Per les mesures associades a la calibració, s'ha utilitzat un dispositiu mòbil *Samsung* amb *Android* 13, situat a una alçada de vuitanta-cinc centímetres, sense una orientació concreta.

El càlcul de la ubicació d'un visitant, es realitza en tres fases: calibració, captura i posicionament.

2.1.2 Fase de calibració

La fase de calibració s'executarà només un cop, i permetrà desenvolupar l'anomenat ràdio mapa. Aquest ràdio mapa, representarà la col·lecció de punts de calibratge dels diferents llocs del museu, cadascun amb la seva llista de valors RSSI, també coneguda com a empremta digital. Els punts de calibratge, s'utilitzaran per calcular la ubicació més probable de l'usuari,

la ubicació real del qual, es desconeix. Es triarà, per l'interès estratègic (mida i distribució dels espais visitables) que comporten, les ubicacions d'interès a calibrar. Per cadascun d'aquests punts, es faran una sèrie de mesures de calibratge, a causa del fet que l'orientació de l'usuari afectarà el valor RSSI mesurat pel sensor. Per exemple, si la ubicació física del visitant està situada entre el sensor i el dispositiu mòbil, probablement la intensitat del senyal serà menor en comparació amb la situació en què el visitant es col·loca al costat oposat del dispositiu. Recordem que el senyal electromagnètic és atenuat pel cos humà.

L'objectiu de la col·lecció de mesures és determinar la intensitat del senyal rebut des de cada sensor *visible* en aquest lloc, amb aquesta orientació. A causa del fet que la força del senyal rebut és influït per molts factors, es faran una sèrie de mesures seqüencials per tal de recollir informació estadísticament més fiable, sobre quina intensitat mitjana del senyal es pot esperar.

Un cop realitzades les mesures, es fa un histograma amb els nivells de potència mesurats, on cada sensor produirà un histograma diferent.

2.1.3 Fase de captura

És la fase en què els sensors capturen els *Probe Request* provinents dels dispositius del visitant, identifiquen el nivell del senyal (i altra informació d'interès) amb què els paquets arriben als sensors, i temporalment emmagatzemen en local, per finalment transferir els fitxers resultant a una base de dades centralitzada, en els moments de sincronisme establerts. La ingesta i sincronització amb la base de dades no s'efectua en temps real, per la probable pèrdua de paquets que suposa commutar entre modes l'emissor/receptor de ràdio del microcontrolador. La freqüència d'emissió de paquets, per part del dispositiu mòbil, depèn del mode actiu en què es troba el sistema operatiu d'aquest: [14]

- Mode de repòs: redueix els serveis en segon pla, per reduir el consum de la bateria. Redueix l'activitat wifi *Probe Request*, fins al punt de deixar d'enviar-los, segons la configuració del dispositiu.

- Mode d'estalvi d'energia: amplia el temps de funcionament de la bateria, reduint les activitats de fons (en funció de l'activitat de l'usuari), així com la reducció de la freqüència d'enviament dels paquets *Probe Request*.

- Mode avió: desactiva el funcionament del wifi, desactivant el transmissor i receptor del dispositiu, segons la normativa de vol.

En general, el mode més habitual és el de repòs, que passa quan els usuaris apaguen la pantalla durant uns segons. Per aquest motiu el nivell d'activació d'aquests modes depèn, en gran manera, dels usuaris.

2.1.4 Fase de posicionament

És la fase en què el programari de càlcul de posició, rep totes les dades captades pels sensors, emeses pels dispositius mòbils. I utilitzant una tècnica concreta, i segons la informació disponible, donarà una posició teòrica del dispositiu en cada moment del temps.

Les tres tècniques usuals de posicionament són: [20]

- Trilateració: tècnica que estima la ubicació del visitant, calculant la distància entre el sensor i un dispositiu mòbil, mesurant la diferència de temps d'un dispositiu mòbil a un sensor. Permet estimar posicions relativament precises en distàncies curtes, però cal una total sincronització horària entre el dispositiu mòbil i el sensor. Com més gran sigui l'error de sincronització de temps, menor serà la precisió de la posició.

- Triangulació: mètode que estima la ubicació del dispositiu, mitjançant la mesura de l'angle trigonomètric entre el sensor i el dispositiu mòbil. Es té en compte la diferència de fase entre cada antena dels sensors. Per tal d'extreure la ubicació usant els angles, cal utilitzar més de dos sensors. Permet obtenir una precisió alta (segons els entorns).

- L'empremta dactilar: és una de les tècniques més utilitzades, per la seva alta estabilitat i precisió de posicionament. La tècnica d'empremta digital requereix un ràdio mapa que enregistra el valor d'RSS wifi, de cada sensor, per cada punt de referència identificat al mapa. Quan s'estima una posició, la tècnica cerca el patró wifi RSSI més proper al del ràdio mapa. La precisió és normalment inferior a la trilateració i la triangulació, però es veu menys afectat per la reflexió i l'atenuació del senyal.

En el cas d'aquest projecte, i tenint en compte la definició del sistema de sensors, i del maquinari desplegat, farem servir la tècnica d'emprenta d'actilar (*fingerprinting*), donat que es compararan amb els valors aconseguits durant la fase de calibració, amb les dades recopilades pels sensors.

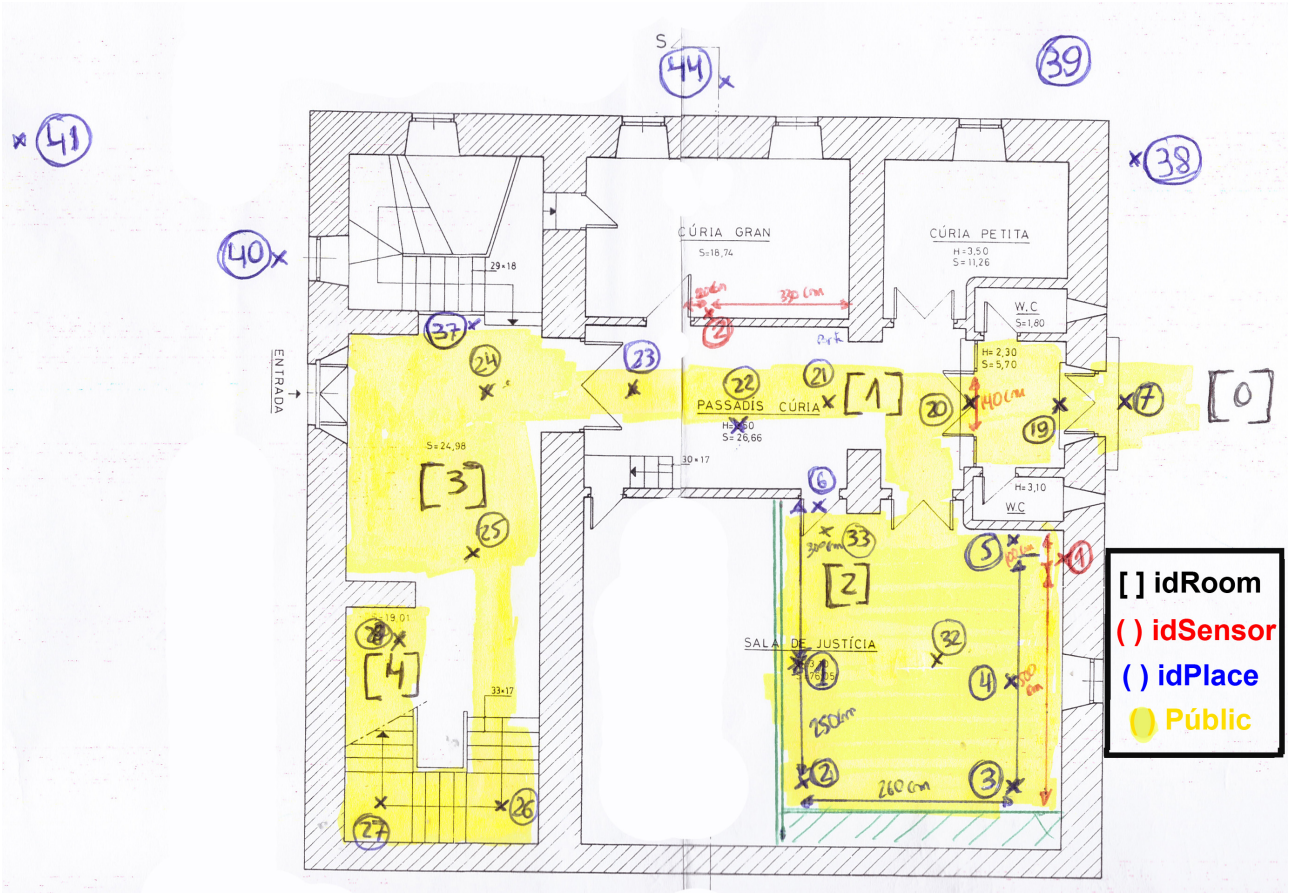

Figura 2.2: Ràdio mapa de la planta baixa

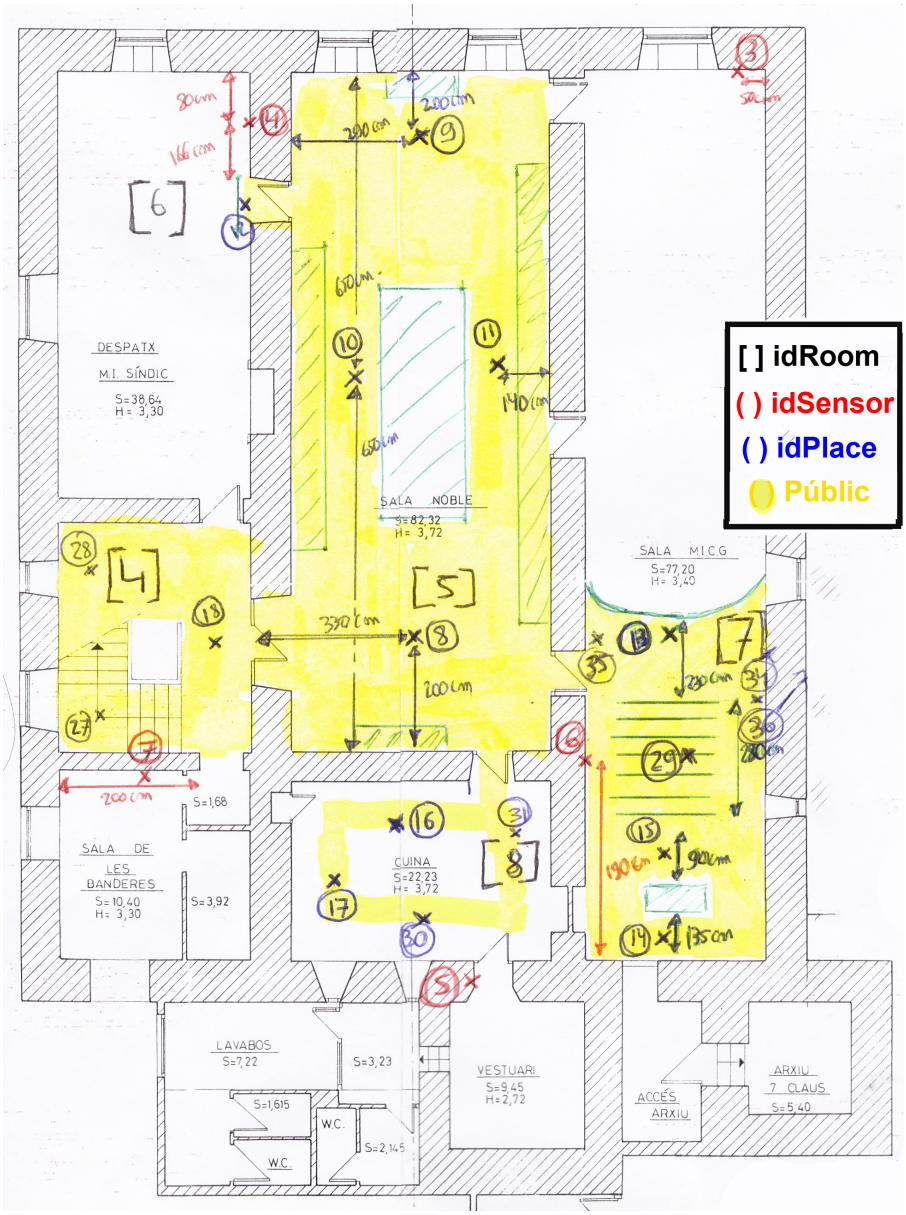

Figura 2.3: Ràdio mapa de la primera planta

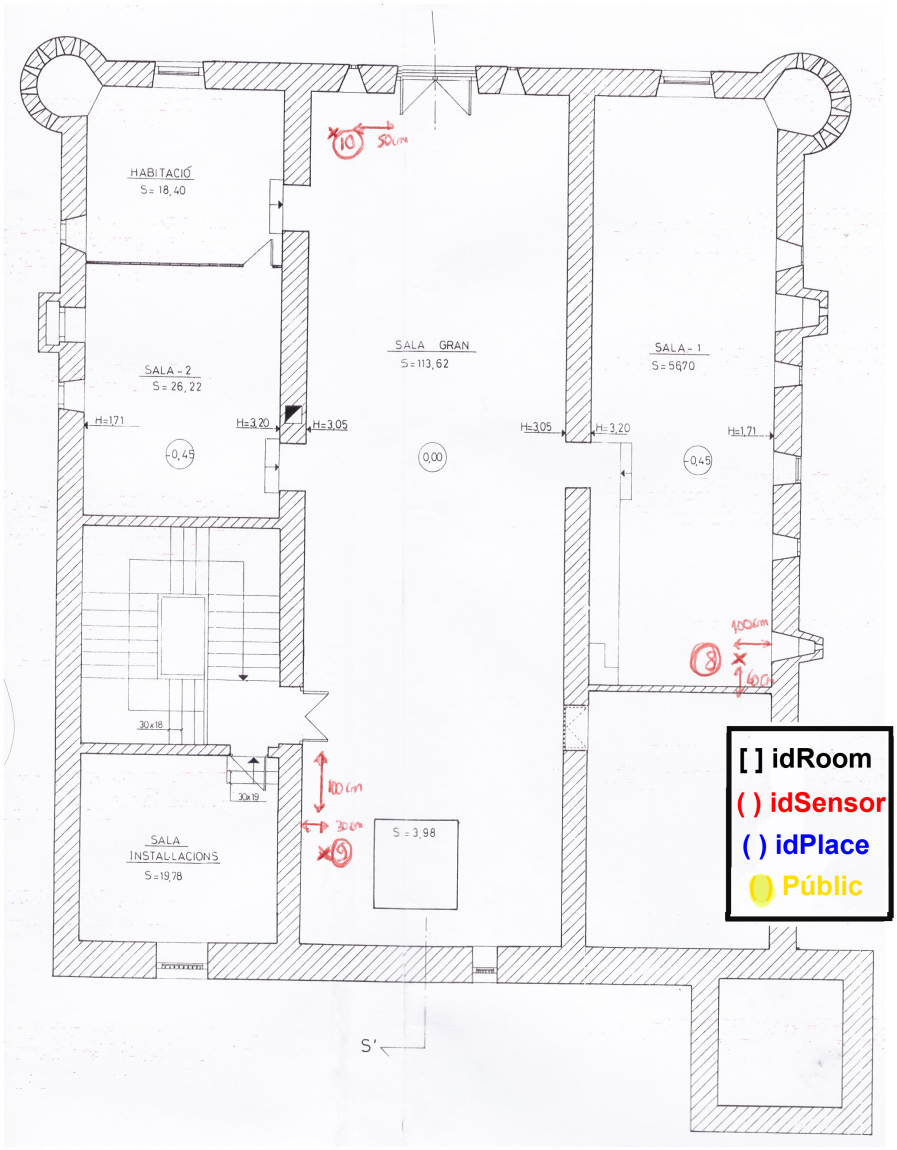

Figura 2.4: Ràdio mapa de la segona planta

2.2 Metodologia de recerca

Com a estratègia de recerca, utilitzarem *Design and Creation*, metodologia que té en compte el desenvolupament, el model, el mètode o la instància, com a contribució al coneixement, tot i que sovint serà fruit d'una combinació d'aquests.

Aquest projecte, i en general, un projectes de recerca en ciències de la computació, implica l'anàlisi, disseny i desenvolupament de productes basats en computadors, webs o sistemes. Són projectes que exploren i exhibeixen les possibilitats de les tecnologies de la informació en diferents aplicacions.

La metodologia *Design and Creation* inclou:

- L'aplicació de les tecnologies de la informació dins d'un nou domini, en el qual, són poc freqüents, i en què l'objectiu és demostrar la viabilitat tècnica, amb arguments.
- L'aplicació de tecnologia que permeti confirmar noves teories, en ciències de la computació, o en altres disciplines, amb una aportació rellevant o útil.
- L'artefacte informàtic és, en sí mateix, la principal contribució al coneixement.

En aquest àmbit, els investigadors solen investigar què passa quan un nou artefacte informàtic s'utilitza en un context de la vida real, amb persones reals, basant-se en un model d'aprendre fent (*Learning via making*). De forma genèrica, utilitza un procés iteratiu que inclou cinc passos [23]: consciència, suggeriment, desenvolupament, avaluació i conclusió.

- La consciència és el reconeixement i l'articulació d'un problema, que pot provenir de l'estudi de la literatura on els autors identifiquen àrees d'investigació, o de la lectura de noves troballes en una altra disciplina, o s'identifica la necessitat de nous desenvolupaments tecnològics.
- El suggeriment implica un salt creatiu des de la curiositat per un problema, a oferir una idea molt provisional de com es podria abordar una solució.
- El desenvolupament implementa la idea de disseny provisional, depenent del tipus d'artefacte informàtic que es proposi.
- L'avaluació examina l'artefacte desenvolupat i avalua la seva aportació i la desviació respecte a les expectatives.
- Finalment, en la conclusió, es consoliden i es redacten els resultats del procés de disseny, s'identifiquen els coneixements adquirits i punts forts. Fins i tot, els resultats inesperats o anòmals, podran esdevenir objecte d'investigacions posteriors.

Un altre punt d'interès, és que, dins de la metodologia *Design and Creation*, es té en compte la generació de dades. És important pensar en la possibilitat d'utilitzar les mètriques derivades de l'observació *digital* de la posició de les persones, per proposar canvis en el món real, que puguin ser de nou observats, i finalment avaluar els canvis provocats en el comportament de les persones. Tot i que en altres projectes de recerca és habitual la recollida de dades mitjançant entrevistes i qüestionaris, inicialment, pel que fa a aquest projecte, no està previst.

2.3 Implementació

Per aconseguir la detecció i ingesta dels paquets de descobriment enviats pels dispositius mòbils i el seu posterior posicionament, l'estudi proposa la instal·lació d'una desena de sensors, distribuïts de forma homogènia entre les tres plantes en què es divideix Casa de la Vall, durant aproximadament dos mesos.

2.3.1 Tecnologia

Utilitzant l'estàndard IEEE 802.11 i, concretament, la definició dedicada als paquets de descobriment, obtindrem les dades d'interès per aquest estudi. Els atributs de gestió que són utilitzats durant l'intercanvi d'informació de la connexió, i que capturarem i tractarem són:

- *Source Address*: Conté l'adreça MAC de la targeta de xarxa del dispositiu que envia la petició (habitualment, fictícia i aleatòria). Cal recordar que en els casos en què no s'utilitza l'aleatorietat MAC, aquesta adreça és única a escala global. Per aquest motiu, no s'emmagatzemarà en clar, i se li aplicarà l'algorisme criptogràfic *SHA256* (amb clau pròpia per augmentar l'entropia), que assegura la impossibilitat de recuperació de l'adreça MAC original.

- *Sequence Control*: Camp de 16 bits, que indica el nombre de seqüència de la trama.

- *Service Set Identifier*: Nom públic de la xarxa WLAN, provinent de la llista de xarxes wifi conegudes, que cerca el dispositiu mòbil, i a la qual s'intenta connectar (habitualment buit).

- *Received Signal Strength Indicator*: Indicador de la intensitat del senyal electromagnètic provinent del dispositiu mòbil, mesurat des del sensor que captura la trama.

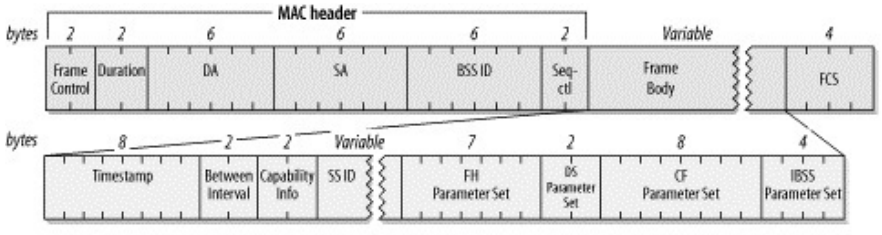

Figura 2.5: Esquema del paquet WLAN *Probe Request*

2.3.2 Maquinari

El sensor que utilitzarem per recollir les trames de descobriment és el microcontrolador *ESP32*, interconnectat amb una targeta *microSD* on s'emmagatzemaran les dades obtingudes. Per la correcta anàlisi de les dades, el circuit també incorpora un rellotge digital.

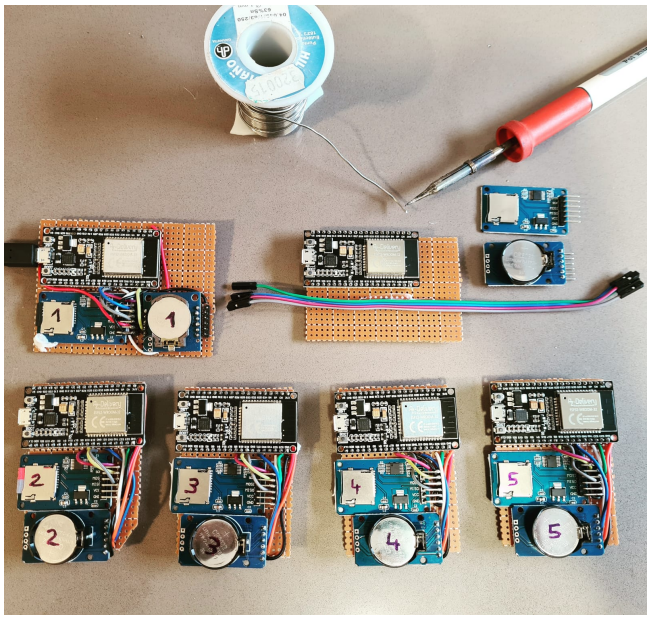

Figura 2.6: Ensamblat dels sensors *ESP32* amb *RTC DS3231* i *microSD Card*

La família *ESP32* es caracteritza pel baix cost i baix consum d'energia, i per disposar de capacitats wifi i *Bluetooth*. Els sensors s'alimenten via USB, connectant-se als endolls existents a les diferents sales, minimitzant l'impacte visual.

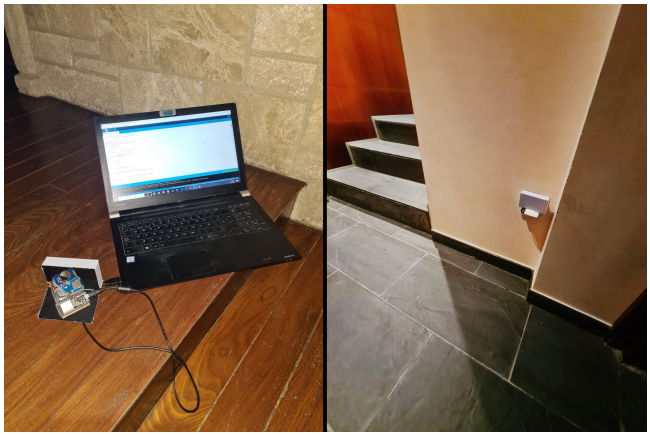

Figura 2.7: Actualització del *firmware in situ* del sensor núm.6 i sensor núm.5 connectat a l'alimentació elèctrica

Per ser respectuós amb el monument, i amb la fita d'ancorar els sensors sense impacte sobre el monument, s'aprofitaran les bases dels endolls existents. No es farà cap forat, ni s'utilitzarà cap d'adhesiu, que pugui provocar marques o desperfectes.

Figura 2.8: Punt de connexió de la sala de l'Hemicicle

Figura 2.9: Punt de connexió de la sala dels Passos perduts

Figura 2.10: Punt de connexió de la capella de l'Hemicicle

Figura 2.11: Punt de connexió del Tribunal de Corts

2.3.3 Programari

Per assolir els objectius descrits en aquest estudi, s'ha desenvolupat específicament el programari necessari. De forma resumida, les tasques que realitza el microcontrolador en iniciar-se són:

1. La targeta *microSD* serà muntada lògicament, amb el propòsit d'emmagatzemar les dades generades pel microcontrolador.

2. Vist que les trames de descobriment no contenen informació horària, el microcontrolador realitzarà un intent de connexió a un servidor *NTP* d'Internet, del qual obtenir l'hora en format *UTC*. En cas de no obtenir l'hora del servidor *NTP*, el microcontrolador recuperarà l'hora del rellotge digital integrat en el circuit, tot i estar mancat d'un sincronisme de qualitat.

3. La interfície sense fils del microcontrolador, canviarà a mode monitoratge.

4. Cada trama capturada serà filtrada, amb l'objectiu de descartar les mancades d'utilitat per aquest estudi, que no seran emmagatzemades.

5. En cas de capturar una trama de descobriment, el microcontrolador farà els càlculs necessaris per a la completa anonimització de la informació pseudo anònima [5], i l'emmagatzemarà en un fitxer dins de la targeta *microSD*.

6. Els fitxers resultants seran periòdicament transferits, mitjançant protocol *SFTP*, a un servidor que disposarà d'una base de dades que ha de permetre centralitzar els registres dels diferents sensors. En cas de no disposar de connectivitat d'Internet, el buidatge de les targetes *microSD Card* es farà manualment.

7. Finalment, s'exportaran les dades centralitzades en la base de dades, en un fitxer *.csv*, per tractar-lo amb el programari que es consideri més adequat.

2.4 Anàlisi de dades

L'estructura de dades que ha de permetre emmagatzemar les dades provinents dels sensors, contindrà els camps:

- Data i hora de la lectura
- Número de sensor
- Rellotge intern o extern
- Número de canal
- Nivell de recepció del senyal radioelèctric
- Resum *Hash*
- Aleatorització MAC activa
- Coincidència del prefix MAC DA:A1:19
- Número de *frame*
- SSID de destí

```
2022/10/07 00:17:11 01 03 02 -95 a9413b80c9041a8ede5b79499b1073a8356cc937dd63f9e9ddaee2f458585f89 0 0 0 11856 Sin seguridad
2022/10/07 00:18:02 01 03 13 -93 c10c3146f72b0fe4b48b154f6cf53fc9cce244456f120be7cf47b3f1f4d83543 1 0 0 37456 Apart Hotel Shusski
2022/10/07 00:20:24 01 03 09 -93 298b1c13d057231c6382c7144afe0632367fcd01af7eb607fc9af5c5d1f07e2c 1 0 0 49312
2022/10/07 00:20:39 01 03 01 -72 59eb685c01807a05e4a9c9fb119a9a97c5065c35d16dda512a41b6a37b53f259 1 0 0 13760
2022/10/07 00:20:44 01 03 10 -89 59a81d65ecfeceed72553de05fc244e40ad17c214e70fad288b28ed64c0caae3 0 0 0 63296
2022/10/07 00:21:34 01 03 06 -71 bc5a84383402ab57c33f80848d8b99fc72e6a91b98063455886l5bc797515e63 1 0 0 31584
2022/10/07 00:22:09 01 03 12 -72 2c1be72751ef81257df90d4eb4c1bc93836268e31c2910c9b044ed209e18c6fd 1 0 0 58752
2022/10/07 00:23:17 01 03 05 -90 ecae685cd83c8e5e7365b4223c20b6a04ed41bfba84a84828410dc752be16563 0 0 0 04352
2022/10/07 00:24:54 01 03 05 -83 475d4b6eb2605dab2f645ff29573162c61963fb38b7acd26b4f61d1c2fbd8fbc 1 0 0 07328
2022/10/07 00:24:54 01 03 05 -80 2c3786e3ca3af067f06d96ff2b2bda71e2a110c053382f741105ffbbc79956ef 1 0 0 10240
2022/10/07 00:25:51 01 03 02 -93 f6811a20876cc9fbe1fe58eb888837111a106dc43dd3d0dd6214fb3f6f097965 1 0 0 11808 Shusski Wifi
2022/10/07 00:26:24 01 03 03 -73 d54b27678fdf0ae0427bdd841781254a8e8e7d083cc87e20c1accd3cf8825dc0 0 0 0 41440
2022/10/07 00:26:39 01 03 06 -71 4e27f27196bf2b36e052493142ced3b42a7a1ac95d0b69e3c124f70ebc2135a1 1 0 0 52112
2022/10/07 00:26:39 01 03 06 -72 4e27f27196bf2b36e052493142ced3b42a7a1ac95d0b69e3c124f70ebc2135a1 1 0 0 52160
2022/10/07 00:27:20 01 03 10 -89 f7d937f5de6fa4ab22090e828d2b814036597ee0afcdb76cf72c36c7c5f1765e 0 0 0 32192 REDSYSWLAN
2022/10/07 00:28:01 01 03 01 -92 a9413b80c9041a8ede5b79499b1073a8356cc937dd63f9e9ddaee2f458585f89 0 0 0 57024 Sin seguridad
2022/10/07 00:29:23 01 03 09 -94 1db7e323c041052618105826cfa26837b4e70ba31e71bb448858166c874135df 0 0 0 32448
2022/10/07 00:29:39 01 03 02 -73 6f052a66090d49f57a360abf00a91d7a6da929ed1d539b5720948c070444db3a 1 0 0 18400
2022/10/07 00:30:34 01 03 07 -70 b90c092ff350a29faa247407ba598ff2a3f0f55b7dd891a462e9bcc058b75964 1 0 0 02560
2022/10/07 00:31:11 01 03 04 -87 960410d927431477fb9ab56dc769b59078f7451821370bb9fc83a40d435d0f79 0 0 0 30512
2022/10/07 00:31:47 01 03 11 -90 cb8fd3e38327ce1c472e043a45c4181bd40d34f41ce6522bf6dcb192bc503fa9 1 0 0 06656
2022/10/07 00:32:39 01 03 11 -70 9b8dc4283b3eae2c29d0b3eaf074f250cf7c5ae61fbc09bda75744d7ab58b0c4 1 0 0 48336
2022/10/07 00:34:29 01 03 09 -71 dd923310bb413f5befff097e5764f7830c6ae6ad699617c785ac791fb4818568 1 0 0 53296
2022/10/07 00:35:27 01 03 10 -91 59a81d65ecfeceed72553de05fc244e40ad17c214e70fad288b28ed64c0caae3 0 0 0 23328
2022/10/07 00:35:39 01 03 07 -72 2fba7b23e3ee5831dd9b7e94abaa515ef8cfe68ee77e2a4323b303960c75fc78 1 0 0 61568
2022/10/07 00:36:25 01 03 08 -89 0c4db7ef908d35fb3e28fb199e88d88a582cc1dc47b9622d348bb952bdd5b356 1 0 0 54928
2022/10/07 00:37:40 01 03 03 -79 71d025a37a7857c3e8ac9e8e0640da995e9ba6ec783e45265a1338cb95e7adb0 0 0 0 08640
2022/10/07 00:37:50 01 03 09 -90 f7d937f5de6fa4ab22090e828d2b814036597ee0afcdb76cf72c36c7c5f1765e 0 0 0 06096
2022/10/07 00:37:56 01 03 09 -93 0f26b783131676827de51231b112cc581e57a9702f9e237f0e40e37a1dd132e 0 0 0 00064 ARBRE DE NEU
2022/10/07 00:38:11 01 03 11 -90 59a81d65ecfeceed72553de05fc244e40ad17c214e70fad288b28ed64c0caae3 0 0 0 27824
2022/10/07 00:38:16 01 03 09 -97 8eaa81d94ec64b6d0ddeab609a5ded26de196ac9a55f72e16b9a5ba4ff6274c0 0 0 0 45120
2022/10/07 00:39:34 01 03 09 -74 c7e6c071f1b0c99373ecea0922aa3c5439ccae28d5f9ce2a2039bf76997aa720 1 0 0 63440
2022/10/07 00:40:09 01 03 01 -72 e9d417f1da5090f90618142d1ccbe043d8d6ede447e2a7a3703ae71c31f19937 1 0 0 62528
2022/10/07 00:41:01 01 03 02 -93 a9413b80c9041a8ede5b79499b1073a8356cc937dd63f9e9ddaee2f458585f89 0 0 0 53312 Sin seguridad
2022/10/07 00:41:10 01 03 07 -84 960410d927431477fb9ab56dc769b59078f7451821370bb9fc83a40d435d0f79 0 0 0 26416
2022/10/07 00:41:11 01 03 07 -83 960410d927431477fb9ab56dc769b59078f7451821370bb9fc83a40d435d0f79 0 0 0 26432
2022/10/07 00:41:39 01 03 12 -72 81463472b31d7206bb1b5cb0e26ce0ca55ffaf58584da06cc009c1a3783648df 1 0 0 29760
2022/10/07 00:42:55 01 03 09 -94 2dd05fcd8bdda9bcb69da6e285fdd2d2a4213452d46bb41d6d64fa6243f7befe 0 0 0 29072 AT-6DBC
```

Figura 2.12: Mostra del model de dades emmagatzemat per un sensor

El camp *Coincidència del prefix MAC DA:A1:19*, provés del CIDR propietat de Google. Tot i que l'adreça MAC indica en el segon caràcter hexadecimal del primer byte, mitjançant una

"A"que l'aleatorització MAC està activa, fins a la versió Android 8, no s'habilita l'aleatorització MAC mentre escaneja xarxes mitjançant paquets de descobriment. Per aquest motiu, capturarem els paquets amb aquest prefix, i en calcularem el seu resum Hash, per avaluar si l'adreça MAC dels paquets de descobriment és persistent, i la podem considerar com a MAC estàtica.

2.4.1 Models de predicció

És cert que com a estratègia pel posicionament en interiors, podríem haver utilitzat la distància Euclidiana per calcular la relació [13] entre el nivell de potència del senyal electromagnètic RSSI i la distància entre el dispositiu mòbil i els sensors, que té com a unitat el dBm (decibels per milivat).

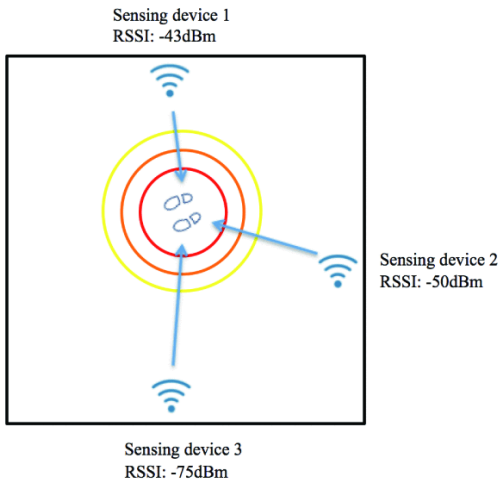

Figura 2.13: Exemple de nivells de senyal en dBm d'un dispositiu que propaga paquets de descobriment [9]

Però l'efecte de les estructures interiors i els cossos humans, provocarà reflexions, refraccions i difraccions, que afectaran de diferents formes a la propagació del senyal electromagnètic. Per aquest motiu no s'utilitzarà aquest de mètode posicionament, i la intenció és obtenir un conjunt de dades d'entrenament, que disposi de les característiques de cada punt, fruit del treball de camp.

Tenint en compte les característiques del conjunt de dades d'entrenament, els algorismes supervisats permetran crear un model predictiu de la localització per *fingerprinting* dels senyals dels dispositius mòbils, basat en dos passos:

1. Entrenament. Utilitzarem dades empíriques, recollides objectivament, i s'hi aplicarà un algorisme supervisat, amb l'objectiu de crear el model per posicionar les dades en els

punts de referència existents, segons les seves propietats.

2. Test. Les noves dades es posaran a prova dins del model construït en la fase d'entrenament, i ens revelarà l'eficàcia de la predicció.

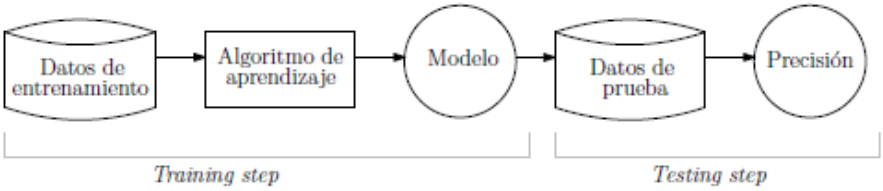

Figura 2.14: Procés de creació i validació d'un model basat en aprenentatge supervisat [16]

Els algorismes supervisats d'aprenentatge automàtic (ML), que utilitzarem per al posicionament en interiors són:

- *Naïve Bayes* (NB) [17]
- *Linear Regression* (LR) [15]
- *K-Nearest Neighbour* (kNN) [19] [25] [24]

Un cop disposem del model associat al ràdio mapa, i després d'obtenir una predicció eficaç, durant la fase de test, podem utilitzar el model per predir les posicions de la resta de mostres de dades, en quatre fases:

1. En la primera fase, utilitzant els algorismes supervisats de *Machine Learning*, es calcularan els punts de referència corresponents als *fingerprints*, tenint en compte el model obtingut de la mostra d'entrenament. Cal recordar que aquest procés tindrà en compte la multidimensionalitat dels nivells RSSI obtinguts. Una aproximació interessant, és la del processament del flux de desplaçaments. [22] El flux de desplaçaments s'estima en funció de la posició de les estimacions. El sistema pot estimar la posició dels dispositius amb una avaluació de qualitat de l'estimació, tenint en compte el número de sensors que participen en el posicionament. També es pot donar el cas d'intervals de temps sense captura de *Probe Request*. Per solucionar aquest problema, proposem el mètode de la interpolació del flux de visitants. El sistema estima la posició d'un dispositiu en una posició A en el temps X i D en el temps Z. No es disposa d'estimacions de la posició B ni C, però podem tenir en compte que la velocitat estimada de marxa d'una persona, és d'1,333 m/s. per estimar el temps necessari en anar d'A a B i de C a D. Si existeixen diverses rutes entre dues posicions, es considera que el dispositiu ha viatjat per la ruta més curta possible.

2. En la segona fase, es realitzarà l'extracció d'aquells registres identificats amb el servei d'aleatorització MAC desactivat. L'objectiu és disposar d'una mostra de qualitat, donat que en aquesta fase, s'utilitzarà com a clau d'agrupament de registres, el resum *Hash*.

3. En la tercera fase, es procedirà a agrupar les posicions de cada resum *Hash*, i s'utilitzarà l'atribut de temps per, de forma cronològica, identificar els diferents patrons de mobilitat dins del monument.

4. En la quarta fase, es realitzarà l'extracció dels registres que amb el servei d'aleatorització MAC actiu. El propòsit serà intentar correlacionar-los entre ells, mitjançant algorismes no supervisats, tenint en compte atributs com són:

   - Data i hora de la lectura
   - Coincidència del prefix MAC amb DA:A1:19
   - Número de frame
   - SSID de destí
   - Punt de referència

   Els patrons de mobilitat obtinguts en la tercera fase, també poden ser útils per millorar les prediccions d'aquesta darrera fase.

2.4.2 Visualització

L'eina *R Studio*, disposa de suficients llibreries per desenvolupar bones visualitzacions. En qualsevol cas, no es descarta utilitzar *Tableau*, en cas de necessitar gràfics complementaris.

2.4.3 Privacitat

Els dispositius mòbils (amb el servei wifi actiu) emeten trames de descobriment, que els sensors poden capturar, ingerir, i de les quals se'n poden extreure els atributs que es considerin d'interès per cada cas d'ús. L'atribut al qual dedicarem especial cura, és a l'adreça MAC, cadena que identifica de forma única, a escala mundial, la targeta de xarxa de cada dispositiu, i que generalment, no es pot modificar. L'adreça MAC està vinculada a un dispositiu, no a una persona, per tant, no revela la identitat del propietari del dispositiu, ni cap altra dada personal, sense l'ús d'informació addicional, com és el codi de la targeta SIM i el contracte de l'operadora de telecomunicacions (classificat com *informació confidencial* per l'operadora). Per aquest motiu, és gairebé impossible identificar a una determinada persona, únicament amb l'adreça MAC. Cercant jurisprudència relacionada amb aquest àmbit, i d'acord amb l'article 4, subsecció 5,

del GDPR, el processament de dades personals que requereix l'ús d'informació addicional per atribuir-los com dades personals a una persona física identificada o identificable, es considera *pseudonimització*. Per aquest motiu, tot i que l'adreça MAC és un identificador únic per identificar el dispositiu, no es classificaria com a dada personal, perquè no permet identificar a una determinada persona, sense l'ús d'informació addicional. Si bé és cert que l'Agència Espanyola de Protecció de Dades (AEPD) va publicar dos informes (0017/2019 i 0043/2019) [6] on s'analitza el tractament del senyal rebut pels dispositius mòbils, amb l'objectiu de conèixer els recorreguts que fan aquests dispositius dins d'un recinte determinat. Cada dispositiu mòbil disposa de diferents identificadors únics, que es poden transmetre i tractar posteriorment, en el context dels serveis de geolocalització. D'acord amb el Dictamen 4/2007, cal assenyalar que un identificador únic (com pot ser l'adreça MAC) permet fer un seguiment de l'usuari d'un dispositiu específic i, per tant, permet *singularitzar* l'usuari. En conseqüència, l'adreça MAC, és una dada de caràcter personal, i el tractament ha d'estar subjecte a aquesta normativa, assenyalant que el tractament conjunt de les dades relacionades amb un terminal mòbil, consistents en el TMSI, l'adreça MAC i el codi IMS (identificador de la targeta SIM de l'usuari), impliquen la recopilació d'informació suficient, perquè es pugui entendre que aquest tractament, està sotmès a la normativa de protecció de dades. Per evitar l'aplicació de la normativa de protecció de dades, cal produir una dissociació absoluta [3] de les dades de TMS, IMSI i l'adreça MAC del dispositiu de l'usuari, i que aquestes dades no puguin ser objecte de conservació. Per aquest motiu, en l'àmbit de l'estudi descrit en aquest document, l'adreça MAC serà processada immediatament després de la captura, aplicant-li l'algorisme criptogràfic *SHA256* unidireccional (irreversible), per convertir-la en una dada anonimitzada, i que, per tant, podrà ser emmagatzemada. Aquest procediment s'aplica com a capa addicional de privacitat, amb l'objectiu de convertir les dades pseudo anonimitzades en anònimes, assegurant la impossibilitat d'identificació de persones. Afirmem que, queda totalment descartada la possibilitat de revelació de la identitat dels propietaris dels dispositius dels quals es recopilaran dades (anònimes) durant l'estudi.

Com a garantia de la no necessitat d'aplicació de la normativa de protecció de dades en les dades capturades i emmagatzemades d'aquest projecte, el 6 d'octubre de 2022, es va presentar la documentació de l'avantprojecte a l'Agència Andorrana de Protecció de Dades. L'ADPA va confirmar que, sempre que s'anonimitzin les dades recollides i no quedin emmagatzemades les dades primigènies, i que no sigui possible revertir el procés d'anonimització, no es tractarà de dades personals.

2.4.4 Planificació

Presentem el detall de la planificació d'aquest projecte, amb les tasques i subtasques calendaritzades per dates, mitjançant un diagrama de *Gantt*:

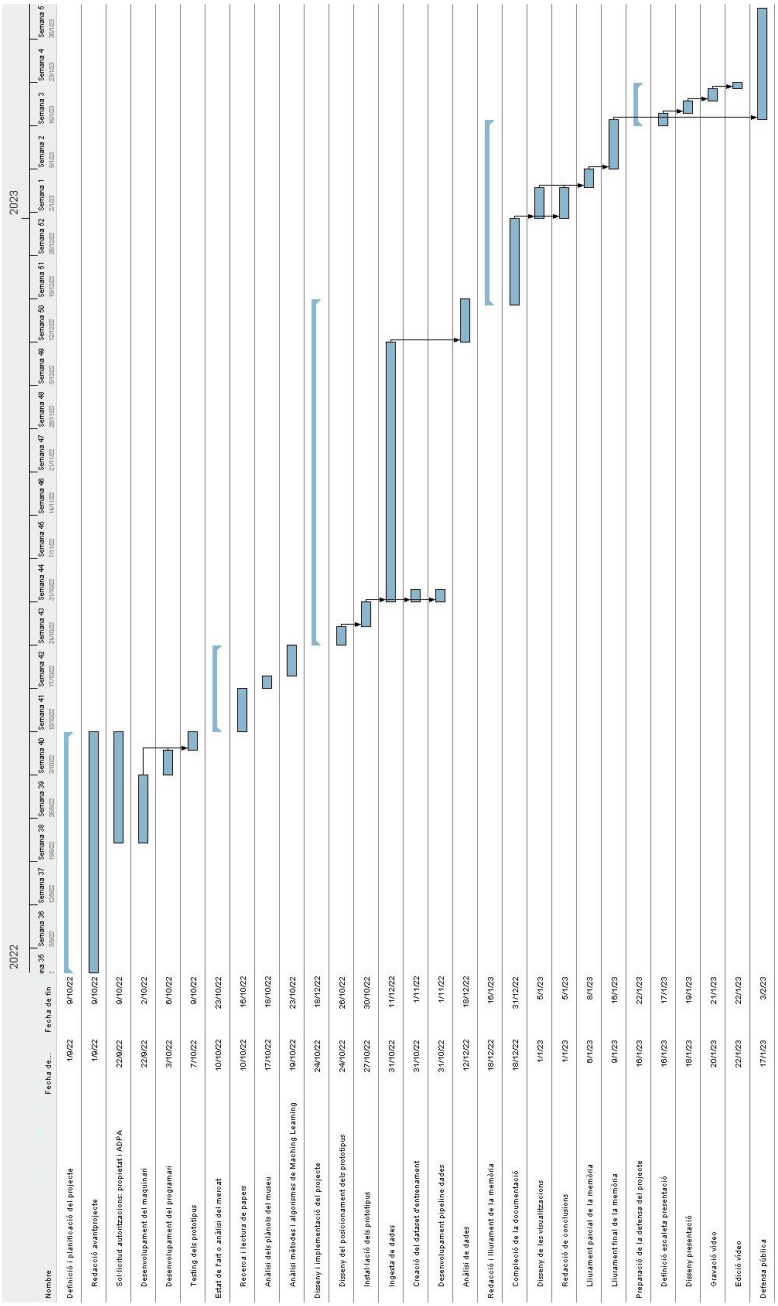

Figura 2.15: Fases en què es planifica el projecte

Capítol 3

Desenvolupament de l'estudi

3.1 Captura, postprocessament i anàlisi preliminar

Com aspectes interessants de la fase de captura de dades de l'estudi, cal indicar que el museu està perimetrat per una zona amb una intensa activitat humana, donat que fita amb dues places públiques (*plaça del Consell General i plaça de Casa de la Vall*), carrers públics (*C/ de la Vall, ascensor públic de l'Hble. Comú d'Andorra la Vella...*), amb una zona de restauració i d'oci amb bars i restaurants (*Nucli Antic*) i amb edificis residencials. Aquesta situació provoca que de forma constant, els sensors rebin paquets de descobriment no vinculats als visitants del monument.

Figura 3.1: Entorns del monument Casa de la Vall, configurats per dues places i dos carrers públics

Com a incidència a documentar, el dia 15 de novembre de 2022 a les 10h15, un simulacre

d'activació dels sistemes d'emergència del museu, va provocar la interrupció de l'alimentació de la majoria dels sensors. Tot i que el museu compta amb una infraestructura *SAI*, no tots els endolls utilitzats estaven connectats a aquest servei. Els sensors van recuperar l'hora del rellotge intern, però van aparèixer petites diferències de sincronisme entre ells, de fins a 3 segons. Per recuperar el sincronisme de totes les lectures, es va procedir a descarregar els fitxers de lectures dels sensors, i prenent com a referència un dels sensors connectats a endolls *SAI*, i mitjançant un *script python*, es van corregir les hores de les lectures dels fitxers amb errors de sincronisme, abans d'injectar-los a la base de dades del projecte.

Una de les accions efectuades en l'àmbit de la neteja de les dades, va ser crear dins de la base de dades, tres taules en les quals quedaven identificats els resums *Hash* (taula *blackhash*), prefix MAC DA:A1:19 (taula *blackidset1*) o noms SSID (taula *blackssid*), que per l'elevada quantitat aparicions, o per ser SSID molt habituals, de treballadors o veïns, calia evitar.

ssid	count(ssid)
Andorra Wifi	29210
Casa_Vall	7275
Barri Antic Hostel Pub	4658
TMobileWingman	2495
REDSYSWLAN	1740
CGtelip	1591
GUEST	1246
ardrone_095366	1080
CGA	1047
L'Orri Free	712
homerun1x	667
Hallak_5	648
AT-4A7A	569
CGA_CONVIDATS	560

Figura 3.2: Mostra de la taula de la base de dades *blackssid*

A mesura que es desenvolupava el projecte, es va considerar la necessitat de no efectuar cap tipus de preprocessament de dades utilitzant les taules anteriors, ja que resultava més efectiu realitzar el filtrat en altres fases de l'anàlisi. Un cop es van recopilar les lectures de la mostra dels 29 dies, es va realitzar una extracció de la base de dades, d'acord amb les necessitats de l'estudi (cal recordar que els sensors han capturat els paquets de descobriment les 24 hores, durant dos mesos), tenint en compte els canvis d'horari entre octubre i novembre, i les diferències d'horari obertura del monument entre setmana i caps de setmana, mitjançant una consulta *SQL* a la base de dades, que tenia en compte:

- Dies en el que el museu està obert (taula *workday*)
- Horari associat a cada tipus de dia (taula *timetable*)

- Lectures associades als dies i hores d'obertura del museu (taula *proberequest*)

El darrer pas d'aquesta fase va ser importar a *R Studio* els registres exportats des de la base de dades, i desenvolupar algunes visualitzacions:

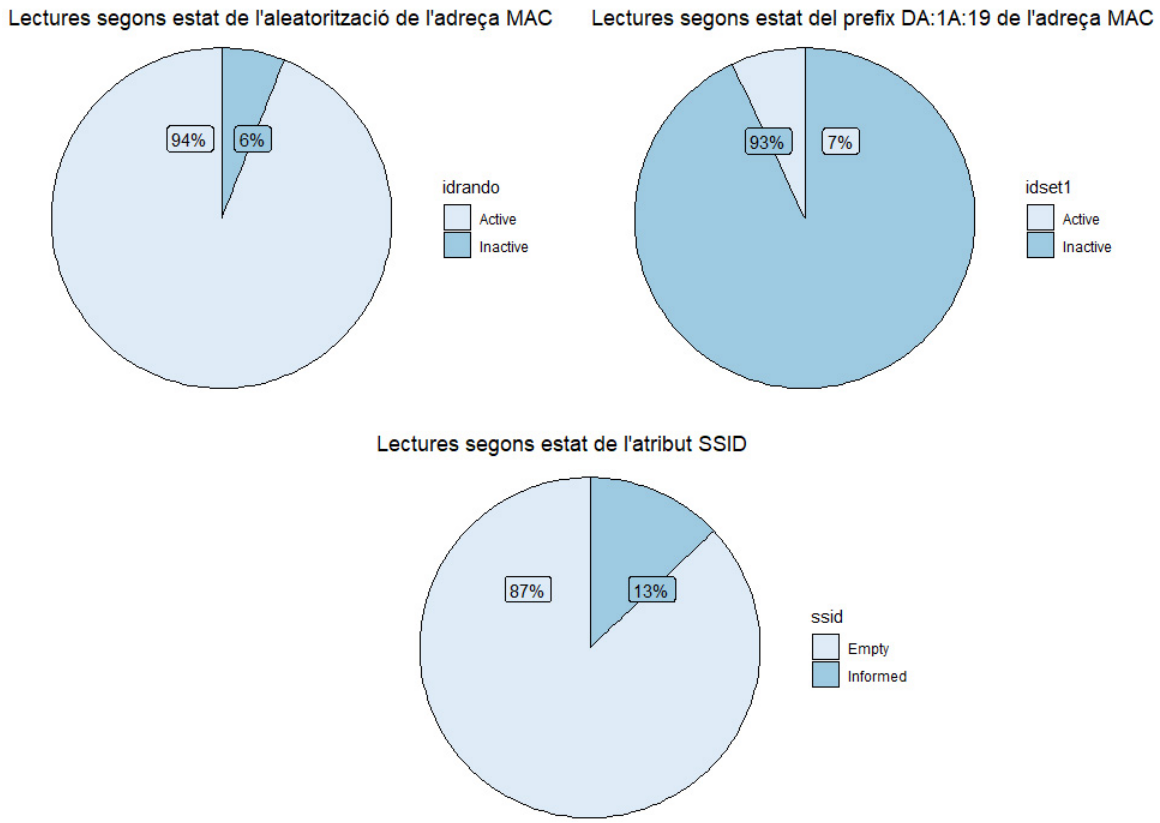

Figura 3.3: Gràfiques percentuals, segons el tipus d'aleatorització, estat del prefix MAC i l'atribut SSID

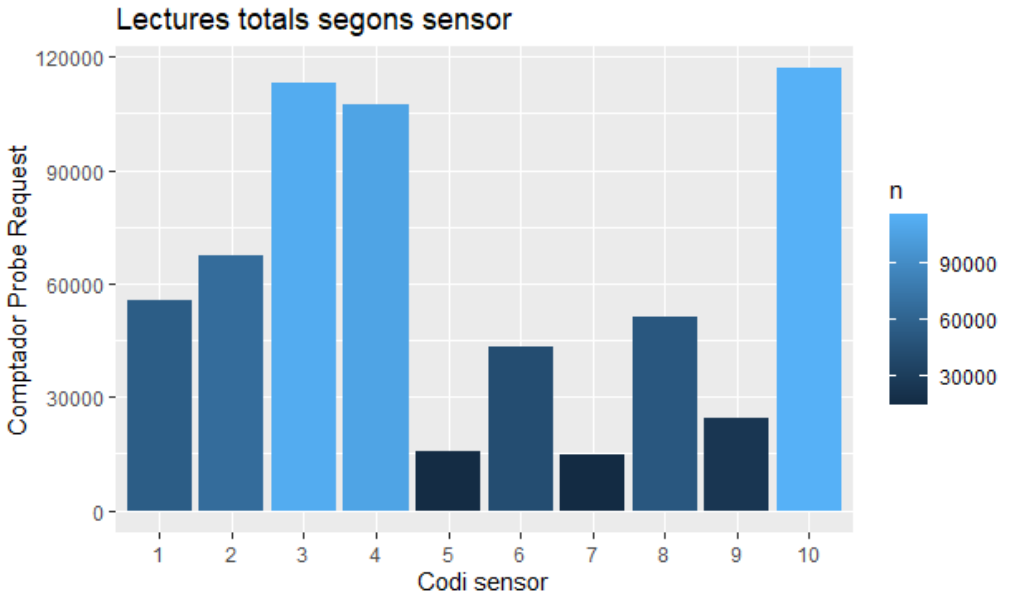

Figura 3.4: Nombre de paquets de descobriment capturats per cada sensor

La visualització següent, mostra com per evitar el *Channel overlapping*, i tenint en compte la configuració de la banda dels 2,4 GHz, els canals 1, 6 i 11 és la seqüència de canals que millor permet aprofitar l'espectre wifi, donat que permetre la coexistència de tres xarxes dins de l'espectre, sense que s'encavalquin, sempre que no es produeixi l'encavalcament amb altres xarxes ja existents, o altres tipus d'interferències.

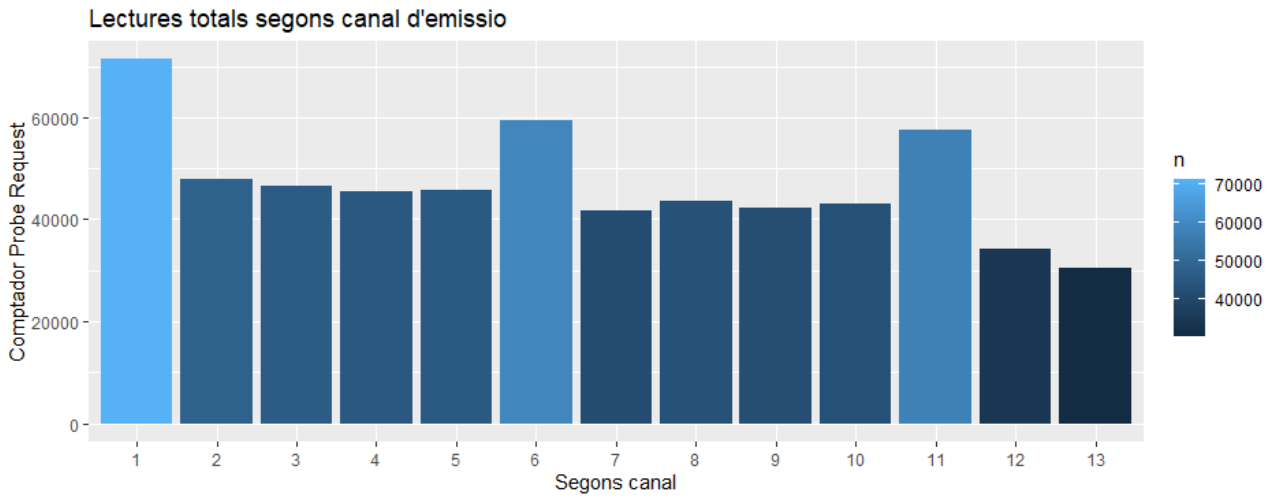

Figura 3.5: Nombre de paquets de descobriment capturats, segons el canal radioelèctric de recepció

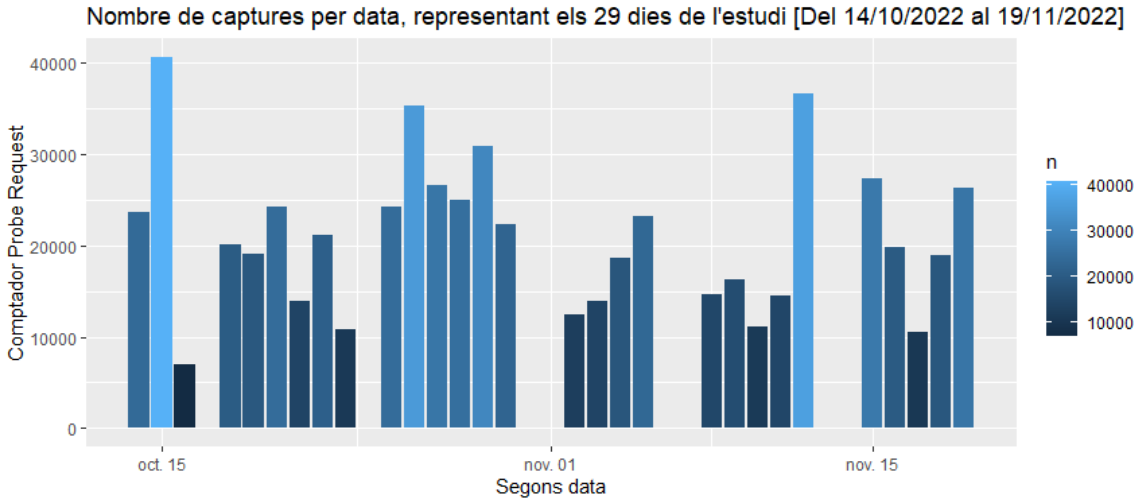

Figura 3.6: Nombre de paquets de descobriment capturats segons data

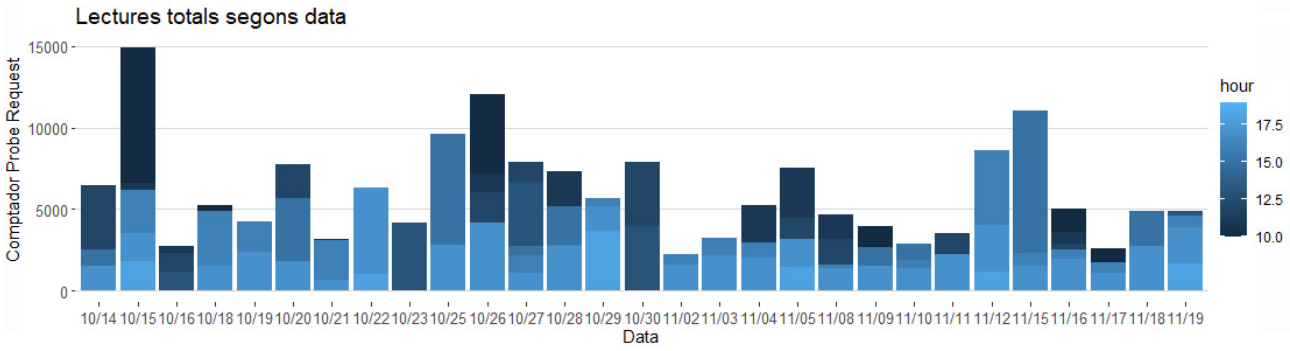

Figura 3.7: Nombre de paquets de descobriment capturats, segons la data i hora

3.2 Georeferenciació de les posicions i calibrat del model

El procés de georeferenciació de les posicions i calibrat del model, és una de les fases més importants del projecte, donat que permet construir els conjunts de dades d'entrenament i test del model, imprescindible per predir la geoposició de les lectures capturades durant les diferents rutes que realitzen els visitants del museu. El procés de georeferenciació es realitza amb el museu completament buit, aprofitant els dies de tancament, amb les condicions d'il·luminació (llums i projectors) i d'accés (portes) amb els mateixos nivells i posició, que trobaríem en un dia d'obertura. La intenció és reproduir, de la forma el més fidel possible, les condicions d'interferències físiques, elèctriques i/o electromagnètiques d'un dia d'obertura normal del monument. Els punts a georeferenciar han estat plantejats sobre els plànols del museu (i els seus accessos i

entorns), cercant distàncies raonables entre ells, de pocs metres. Denominarem aquests punts *idplace*, amb un codi correlatiu de l'u al quaranta-cinc. Cal dir que no han estat atorgats de forma físicament coherent, donat que s'han efectuat dues calibracions diferents, i durant la segona es van afegir alguns que no s'havien identificat inicialment. A més, agruparem en vuit zones diferents (sala, passadís, escales..) denominades *idroom*, el conjunt dels quaranta-cinc punts d'interès, amb l'objectiu d'analitzar la coherència en la navegació entre els diferents espais. En aquest cas, la codificació sí que s'ha realitzat utilitzant la lògica d'arribada a les diferents zones del monument. Tècnicament, per georeferenciar un punt d'interès, es posiciona el mòbil a una alçada de vuitanta-cinc centímetres (simulant que està en una butxaca, mà o bossa del visitant), i aprofitant un rellotge sincronitzat amb el rellotge dels sensors, s'encén el servei wifi del dispositiu mòbil, forçant-lo a enviar paquets de descobriment de forma constant, durant un minut, dedicant (aproximadament) 15 segons a emetre paquets, situant el cos amb orientació nord, est, sud i oest. Passat el minut d'emissió, s'atura el servei wifi del dispositiu mòbil, s'anota en un full la posició, moment d'inici i fi de cada emissió, i es deixa aproximadament un minut de distància entre cada prova, per evitar l'encavallament de diferents proves. Un cop acabada la fase de georeferenciació, es recopilen els fitxers de lectures generats a cada sensor, mitjançant el servei *SFTP*. Tenint en compte que el dispositiu que s'ha utilitzat per generar tràfic de paquets de descobriment, no permetia configurar l'adreça MAC com a estàtica, ha estat necessari realitzar un processament manual d'aquests, utilitzant com a criteri de selecció, les lectures que contenien una SSID molt concreta, configurada amb aquest objectiu. Les referències interessants per a la georeferenciació, seleccionades a mà, es van introduir dins de la taula de la base de dades *positionvalue*, creant uns registres formats pels valors de cadascun dels 10 sensors, i el codi de lloc que caracteritza el *fingerprint*. Per garantir la traçabilitat, s'afegeix a aquest registre el resum *Hash* de la lectura original.

S'han realitzat dues sessions de georeferenciació, en dates diferents:

- La sessió número 1, es va realitzar el dia 24 d'octubre de 2022, a les 18h. En aquesta primera sessió, no es va realitzar cap tipus de rotació de l'usuari (nord, sud, est i oest) durant la georeferenciació dels llocs d'interès. Es va utilitzar de forma aleatòria, la posició lògica de l'usuari. L'enviament de paquets de descobriment, es va realitzar durant aproximadament 30 segons, des de cada punt d'interès. Inicialment, es van identificar fins a 31 llocs d'interès, però després d'analitzar els resultats, tot i no ser gens descoratjadors, es va decidir realitzar una segona sessió de georeferenciació més exhaustiva i rigorosa, aprofitant l'experiència anterior, i el nou coneixement assolit mitjançant la lectura d'articles d'aquest àmbit. La primera sessió va assolir 281 registres, generats a partir d'unes 1.821 lectures.

- La sessió número 2, es va realitzar el dia 6 de novembre de 2022, a les 11h, aprofitant

que era un dia festiu (diumenge). Després d'haver analitzat més extensament l'estat de l'art, es va repetir la georeferenciació dels punts d'interès ja identificats per la sessió número 1, afegint-ne de nous (arribant als 45), i dedicant aproximadament vint segons a cadascuna de les quatre orientacions, amb l'objectiu de situar el cos humà de l'usuari, entre el dispositiu emissor i els sensors, i provocar l'apantallament (debilitació del senyal). Cal recordar que, la freqüència utilitzada pel senyal wifi (en aquest estudi) és de 2,4GHz, freqüència que provoca la vibració de les molècules d'aigua del cos humà i, per tant, l'apantallament provoca una distorsió i dissipació del senyal emès. La segona sessió va assolir 590 registres, generats a partir d'unes 3.756 lectures.

Un cop identificats els nivells de senyal dels respectius sensors, per cada punt d'interès, es crea una taula *positionvalue* dins de la base de dades del projecte, en la que s'introdueixen els 281 registres de la sessió núm. 1 i 590 registres de la sessió núm. 2, en format vector d'n valors dBm, un per cadascun dels deu sensors, de les vuit sales (set interiors i una d'exterior), georeferenciant quaranta-cinc llocs diferents. La col·lecció dels n registres obtinguts per cada lloc d'interès, representen el seu wifi *fingerprint*. Un cop assolida la col·lecció dels *fingerprints* característics de cada lloc d'interès, i emmagatzemada dins de la base de dades del projecte, s'exporten en un fitxer.

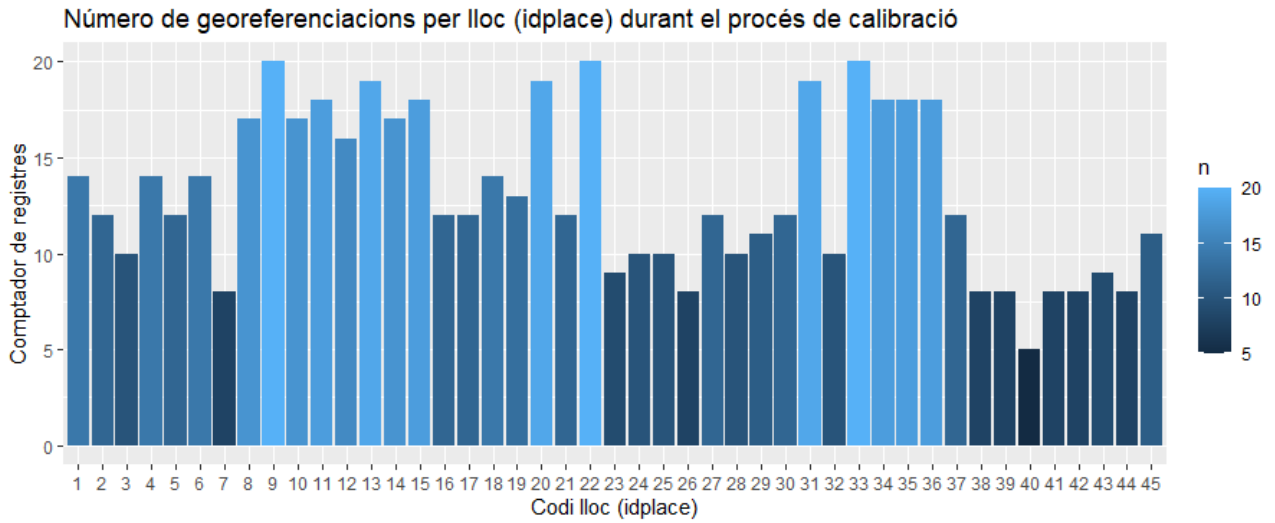

Figura 3.8: Detall del nombre de mostres de registres de georeferenciació per cada lloc (*idplace*)

Un cop importades les dades des d'*R Studio* es realitza la correcció dels valors en decibels que caracteritzen el *fingerprint* de cada sensor, donat que per defecte, el valor nul (manca de senyal) està identificat amb un zero. Tenint en compte que l'escala dels valors RSSI (entre -30dBm i -100dBm) atorga als valors negatius propers a zero el significat d'alt nivell de senyal

(proximitat al sensor), resultava incoherent, mantenir el zero com a límit, i per aquest motiu, és reemplaçat per un valor no possible, i seguint la lògica, molt llunyà al sensor (-200 dBm).

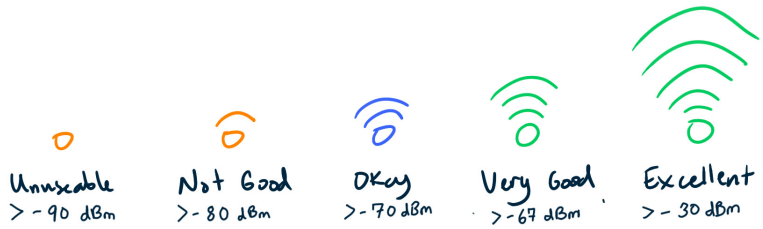

Figura 3.9: Descripció dels nivells de senyal relacionats amb la seva qualitat

Considerant que la sessió de georeferenciació núm. 2 s'ha desenvolupat amb més rigor metodològic, i gaudeix de més precisió, és la sessió de col·lecció de lectures seleccionada per generar dues mostres: la d'entrenament (*test*) i la de prova (*train*), dividides al 90% i 10% respectivament, i que s'utilitzaran per entrenar i validar el model, utilitzant els tres algorismes de *Machine Learning* seleccionats.

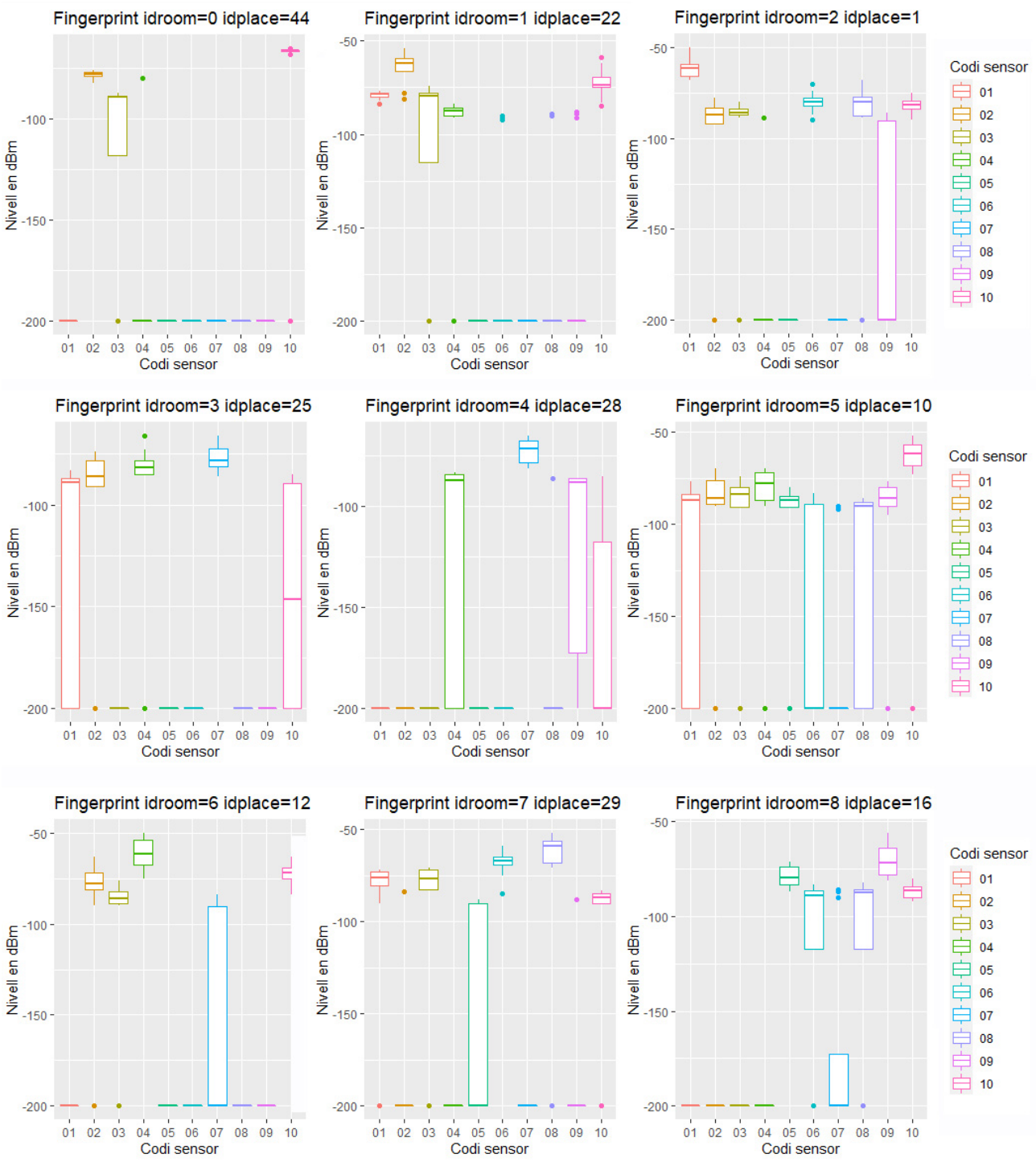

Figura 3.10: Visualització dels *fingerprint* de llocs (*idplace*), representatius de les vuit sales en què s'han dividit els espais del museu

Amb l'objectiu de validar la predicció del model, s'extreu l'atribut *idplace* del subconjunt

de *test*, però es conserva de forma paral·lela, per realitzar la verificació final. L'atribut *idplace* permet generalitzar diferents punts d'un mateix espai mitjançant l'identificador *idroom*. Un cop realitzada la predicció, es compara l'atribut *idroom* georeferenciat, amb el predit pel model.

3.2.1 Predicció del posicionament interior mitjançant *Naïve Bayes* (NB)

Mitjançant l'algorisme *Naïve Bayes*, s'ha obtingut un baix percentatge d'encert (38,98%), i la matriu de confusió generada, presenta una diagonal central amb certa difusió, amb dispersió de valors, representativa de força errors d'encert entre les posicions de referència i predicció:

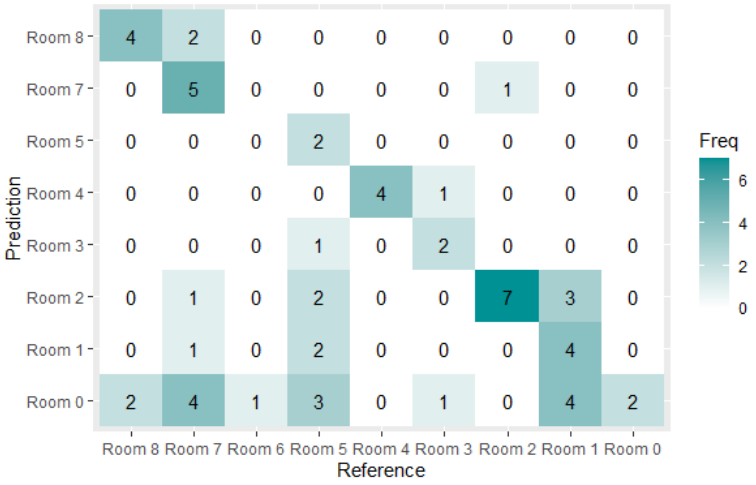

Figura 3.11: *Confusion Matrix* de la predicció sobre la sessió de georeferenciació núm. 2, mitjançant l'algorisme NB

3.2.2 Predicció del posicionament interior mitjançant *Linear Regression* (LR)

Mitjançant l'algorisme *Linear Regression*, s'ha obtingut un molt baix percentatge d'encert (3,38%), i la matriu de confusió generada, presenta una diagonal central pràcticament inexistent, amb molta dispersió de valors, representativa de la marca d'encert entre les posicions de referència i predicció:

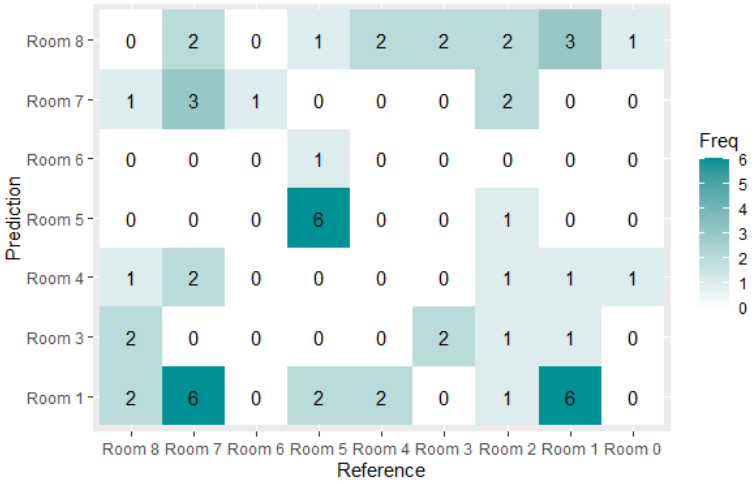

Figura 3.12: *Confusion Matrix* de la predicció sobre la sessió de georeferenciació núm. 2, mitjançant l'algorisme LR

3.2.3 Predicció del posicionament interior mitjançant *K-Nearest Neighbors* (kNN)

Mitjançant l'algorisme *K-Nearest Neighbors*, amb el paràmetre k per defecte (k=1), s'ha obtingut un elevat percentatge d'encert (86,44%), i la matriu de confusió generada, presenta una diagonal molt ben definida:

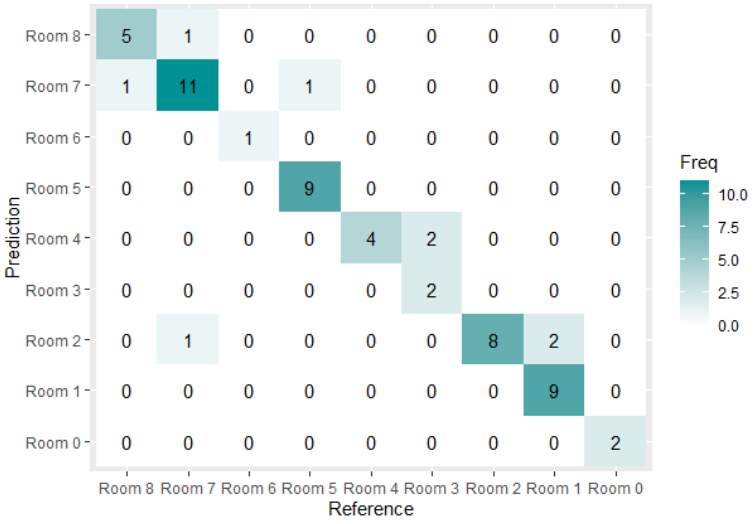

Figura 3.13: *Confusion Matrix* de la predicció sobre la sessió de georeferenciació núm. 2, mitjançant l'algorisme kNN amb el valor k per defecte

El plànol demostra que, els únics errors comesos, han estat entre sales contigües de la planta

baixa (sales 1 i 2, sales 2 i 4, sales 3 i 4):

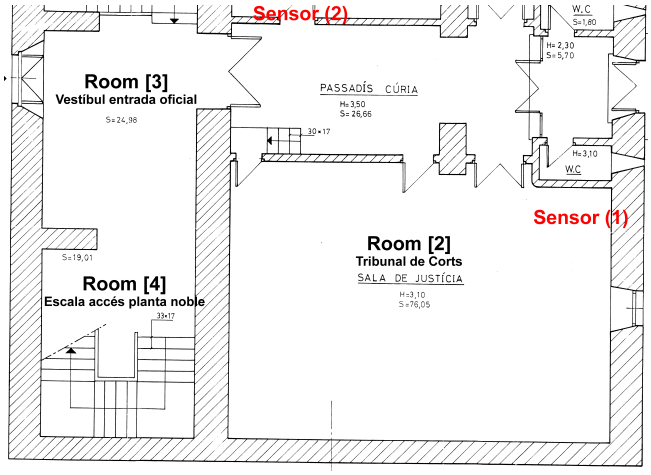

Figura 3.14: Plànol de la situació de les sales i sensors de la planta baixa

També entre les sales contigües de la primera planta (sales 5 i 7, sales 8 i 7), o entre les sales superposades entre la planta baixa i la primera planta (sales 2 i 7):

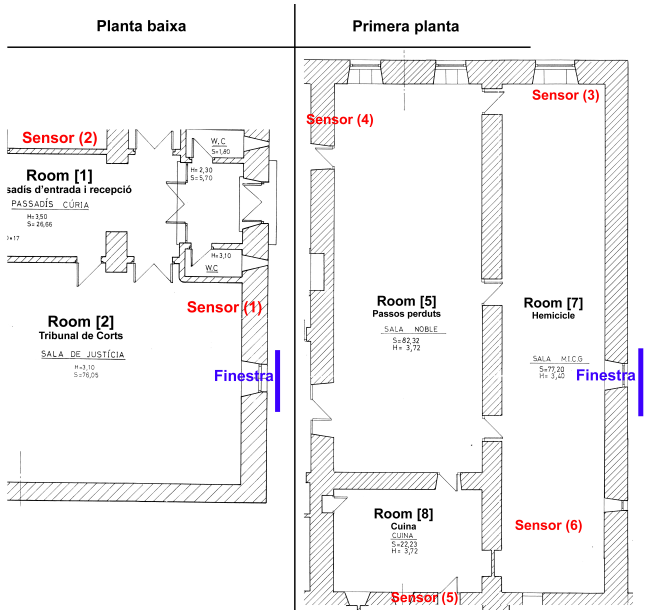

Figura 3.15: Plànol de la situació de les sales (sense la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

Una millora en el model de predicció, és indicar el nombre de veïns que l'algorisme ha de revisar. Si la k és massa baixa, l'algoritme utilitzarà pocs veïns, i provoca *overfitting*. Si la k és massa alta, la predicció tendirà, cada cop més, a aproximar-se a la mitjana, i per aquest motiu,

apareixerà el problema d'*underfitting*. Un possible valor, per una k òptima, s'obté calculant l'arrel quadrada del nombre de mostres de referència disponibles (sempre que aquest nombre no sigui massa elevat). En el nostre cas, disposem de 531 mostres de referència i, per tant, una possible k òptima és 23. Un cop realitzat l'ajustament de la k òptima, s'obté una millora en el percentatge d'encert de l'1,7%, assolint el 88,14%, i la matriu de confusió generada presenta una diagonal molt ben definida, en la que es resol l'error entre sales de plantes superposades (sala 2 i 7), sense afectar pràcticament als errors ja identificats entre sales de la mateixa planta, visualitzats en la matriu de confusió anterior.

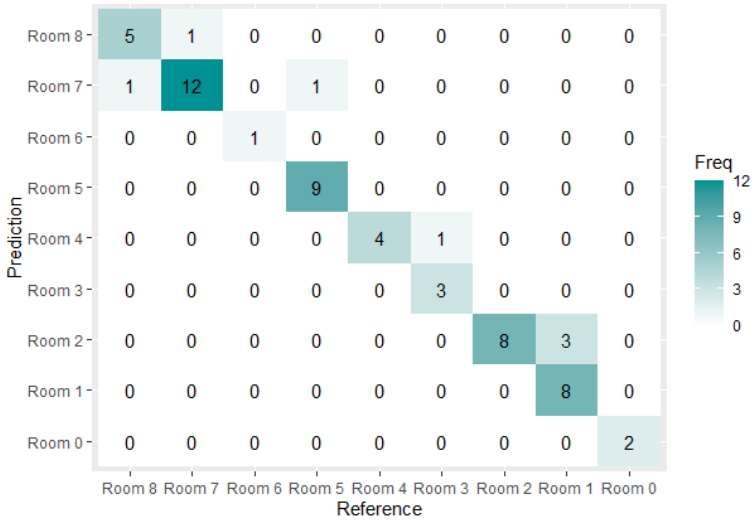

Figura 3.16: *Confusion Matrix* de la predicció sobre la sessió de georeferenciació núm. 2, mitjançant l'algorisme kNN amb valor k òptim

L'ajustament del valor k òptim, aconsegueix eliminar l'error representats en l'anterior matriu de confusió, figura 3.16, entre les sales 2 i 7, sales superposades entres la planta baixa i la primera planta. Alhora, ha reduït l'error anteriorment detectat entre les sales 3 i 4, però ha incrementat l'error entre les sales 1 i 2, contigües en la mateixa planta:

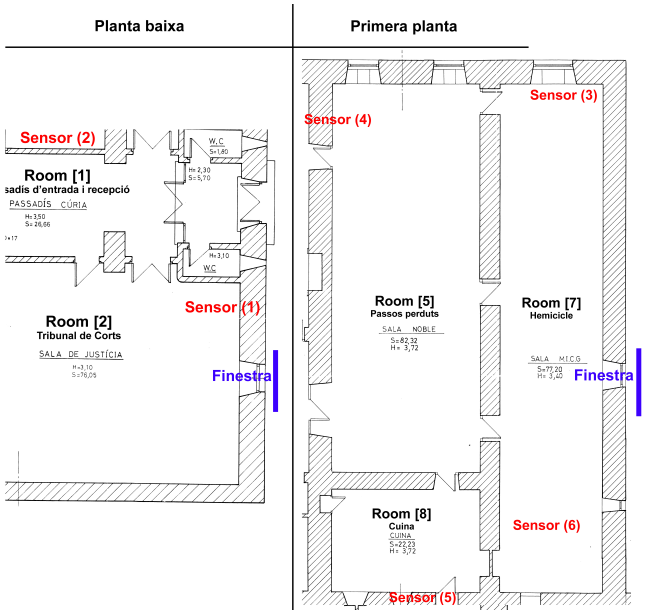

Figura 3.17: Plànol de la situació de les sales (sense la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

Vist el bon resultat d'encert en la predicció d'aquest darrer model, kNN amb k=23, serà el mètode seleccionat per desenvolupar la resta de prediccions d'aquest projecte.

Mètode ML	***k***	***Encert***
Naïve Bayes		38,98%
Linear Regression		3,38%
K-Nearest Neighbors	1	86,44%
K-Nearest Neighbors	23	88,14%

Taula 3.1: Taula de detall de l'encert entre els diferents algorismes supervisats d'aprenentatge automàtic

3.3 Generació d'una ruta monitoritzada

Els visitants del monument disposen de dues opcions per fer la visita del museu:

- Utilitzar una audioguia, que és un dispositiu electrònic que es cedeix als visitants, i amb suport d'un mapa en paper que condueix als visitants entre els diferents llocs d'interès, permet escoltar de forma individual, la història i peculiaritats de cada espai. L'ús de l'audioguia permet igualment, rutes aleatòries i durades diferents. L'audioguia disposa de 15 pistes d'àudio, de durades d'entre 30 segons, i 3 minuts amb 14 segons, i disposa de tres àudios tipus *bonus track* per a l'ampliació dels coneixements.

- Acompanyats d'un/a guia oficial. En aquest sentit, els guies oficials del Ministeri de Cultura i Esports del Govern d'Andorra, per coherència amb l'*storytelling*, solen seguir la mateixa ruta que identifica l'audioguia, però depenent del públic o del seu interès, poden realitzar adaptacions (de la ruta o de la durada de l'explicació dels espais) a una necessitat o interès concret.

L'opció més habitual (fins i tot amb grups) és la utilització d'audioguies, motiu pel qual podem afirmar que un percentatge molt elevat dels visitants, hauria de seguir la ruta:

1. Exterior del monument
2. Passadís d'entrada i recepció
3. Vestíbul de l'entrada oficial
4. Escales d'accés a la planta noble
5. Passos perduts
6. Despatx del de/la Síndic/a
7. Passos perduts
8. Cuina
9. Passos perduts
10. Hemicicle
11. Passos perduts
12. Escala d'accés a la planta baixa
13. Vestíbul de l'entrada oficial
14. Passadís d'entrada i recepció
15. Tribunal de Corts
16. Exterior
    (a) Monument als Pareatges
    (b) Monument al Manual Digest
    (c) Porta oficial

(d) Monument a la Constitució

(e) 7 poetes

17. Recepció (retorn de l'audioguia)

Amb l'objectiu de provar el funcionament del model, de la mateixa forma que amb el procés de georeferenciació, s'han realitzat i capturat dues rutes, amb el museu buit, seguint la lògica d'un visitant (esquema anterior), documentant l'hora d'inici i fi en cada lloc, i mantenint una alta activitat del servei de xarxa wifi del dispositiu mòbil:

1. El dia 1 de novembre de 2022, a les 11h, es va aprofitar el dia festiu per realitzar una ruta monitoritzada, utilitzant l'audioguia que el museu proporciona als visitants. Cal tenir en compte que, es van realitzar amb els mateixos projectors de llum, portes obertes i dispositius informàtics encesos, amb l'objectiu de reproduir de la forma més fidel possible, les condicions d'interferències físiques, elèctriques i/o ràdio elèctriques d'un dia d'obertura del monument. És cert que, tenint en compte que el museu es trobava tancat al públic, no va ser possible realitzar la ruta exterior de forma contínua, i es van utilitzar una sortida secundària no habitual.

2. El dia 14 de novembre de 2022, es va generar la segona ruta monitoritzada, utilitzant l'audioguia, però sense forçar una publicació activa de paquets de descobriment per part del dispositiu mòbil. Aquest fet va provocar una important reducció de paquets de descobriment capturats, reduint igualment la possibilitat d'agrupar les trames segons l'SSID específic. Aquesta segona ruta no s'ha utilitzat, donat que resultava complex identificar els paquets de descobriment que pertanyien al visitant monitoritzat.

Un cop es disposen de les lectures dels diferents sensors, i després de passar la fase de classificació del tipus de captura, segons els atributs ja detallats, s'utilitzarà un algorisme de conversió de les lectures (un sensor capta una lectura en un moment concret) en registres, agrupant en un rang de temps determinat, els millors valors (més propers a zero) de cada sensor, en un sol registre (per cada fracció de temps). Tot i que la diferència detectada entre els diferents rellotges dels deu sensors, era d'un màxim de dos segons, s'acorda definir una finestra de temps de 5 segons per cada registre, amb la finalitat d'assolir registres alimentats per suficients lectures.

Algorithm 1 Conversió del dataframe *df3* de lectures, a un dataframe *df1* de registres

$dfl1 \leftarrow data.frame(-200,-200,-200,-200,-200,-200,-200,-200,-200,-200,"","")$
$colnames(dfl1 \leftarrow ('idsensor1','idsensor2','idsensor3','idsensor4',...,'lecturecut','hash')$
$dfl2 \leftarrow data.frame(-200,-200,-200,-200,-200,-200,-200,-200,-200,-200,"","")$
$sensor \leftarrow df3[1,].idsensor$
$value \leftarrow df3[1,].rssi$
$datetime \leftarrow df3[1,].datetime$
$hash \leftarrow df3[1,].hash$
for i **to** nrow(dfl3) **do**
  if dfl2[sensor] < value **then**
    $dfl2[sensor] \leftarrow value$
    $dfl2[11] \leftarrow datetime$
    $dfl2[12] \leftarrow hash$
  end if
  $temp\_datetime \leftarrow dfl3[i,].datetime$
  if (difftime(datetime_aux, temp) < 5 || hash != dfl3[i,].hash **then**
    $dfl1[nrow(dfl1)+1,] \leftarrow dfl2$
    $dfl2 \leftarrow data.frame(-200,-200,-200,-200,-200,-200,-200,-200,-200,-200,"","")$
  end if
  $sensor \leftarrow df3[1,].idsensor$
  $value \leftarrow df3[1,].rssi$
  $datetime \leftarrow df3[1,].datetime$
  $hash \leftarrow df3[1,].hash$
end for
if dfl2[sensor] < value **then**
  $dfl2[sensor] \leftarrow value$
  $dfl2[11] \leftarrow datetime$
  $dfl2[12] \leftarrow hash$
end if
$dfl1[nrow(dfl1)+1,] \leftarrow dfl2$

Capítol 4

Resultats

4.1 Predicció i validació de la geoposició de la ruta monitoritzada

Utilitzant els recursos i mètodes detallats en el capítol 2, i després de monitoritzar la ruta del dia 1 de novembre de 2022, en la qual es va mantenir el dispositiu mòbil amb molta activitat de xarxa, per assolir l'enviament del màxim número de paquets de descobriment, es va realitzar un dur treball per relacionar, de forma manual, els paquets de descobriment, donat que no es va poder configurar el dispositiu mòbil en modalitat MAC estàtica. Com a criteri de selecció es va utilitzar una SSID específica que permetia associar els paquets de descobriment.

Un cop seleccionades les lectures associades a la ruta del visitant monitoritzat, utilitzarem l'algorisme 1, de transformació de lectures en registres. De les 997 lectures capturades, l'algorisme les transforma en 164 registres. El següent pas var ser associar de forma manual, tenint en compte l'hora identificada en la documentació, els registres obtinguts per l'algorisme anterior, als llocs físics realment visitats durant la ruta monitoritzada.

Un cop resolta la fase anterior, s'introdueix al model la ruta monitoritzada, i es comparen els *idrooms* reals, amb els predits, obtenint un *accuracy* del 76,22% i la *Confusion Matrix*:

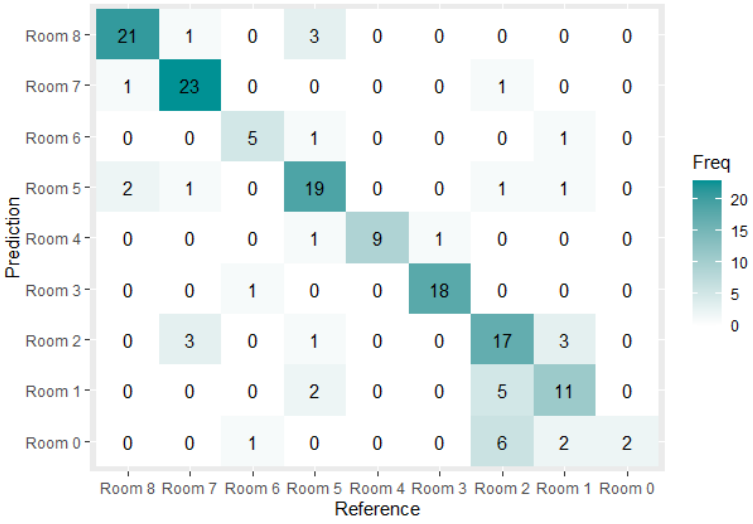

Figura 4.1: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, sense cap refinament qualitatiu

En aquesta matriu de confusió apareixen, respecte a la matriu de confusió de la figura 3.13, nous errors entre les sales properes a la porta d'accés i l'exterior del monument, sales 0 i 1, les sales 0 i 2 i les sales 0 i 6:

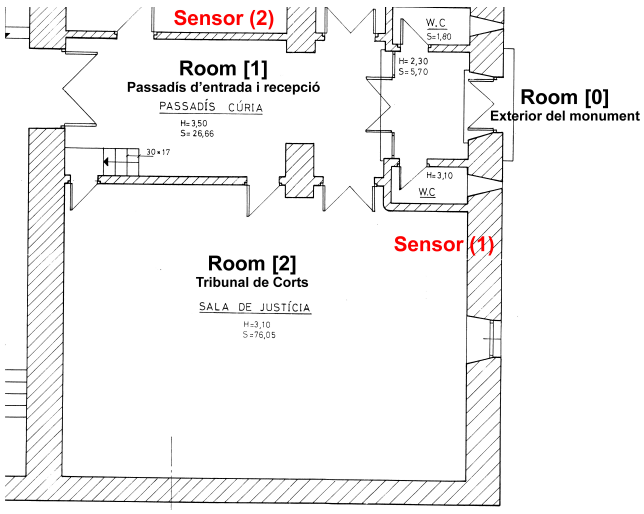

Figura 4.2: Plànol de la situació de les sales i sensors de la planta baixa

També entre les sales contigües de la primera planta (sales 5 i 8 i sales 7 i 8), o entre les sales superposades entre la planta baixa i la primera planta (sales 1 i 6, sales 1 i 5 i sales 2 i 5):

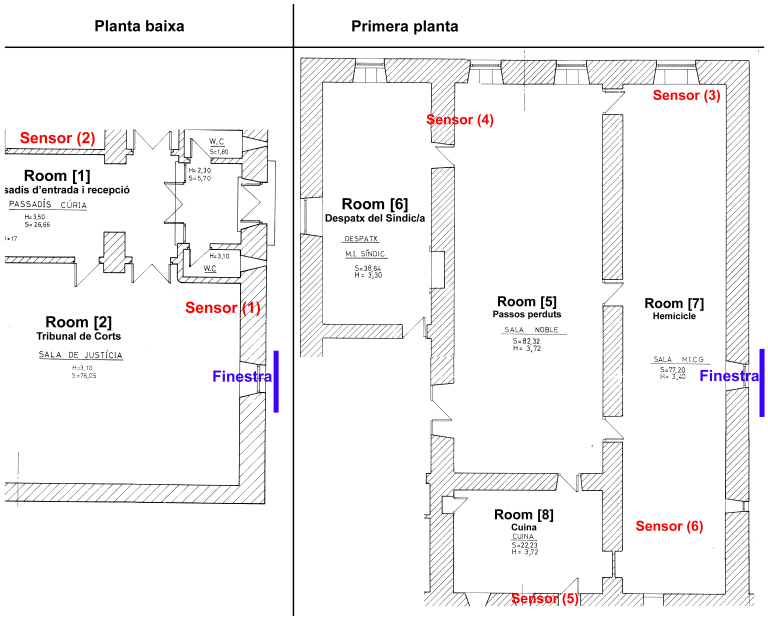

Figura 4.3: Plànol de la situació de les sales (inclosa la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

Amb l'objectiu de millorar la qualitat de la col·lecció de registres generats per aquesta ruta monitoritzada, i tornar a analitzar el resultat de la predicció, per avaluar si les accions efectuades han estat efectives, es realitza un procés de neteja de registres, utilitzant dos criteris objectius:

- La qualitat de les lectures: Es crea un indicador de la qualitat de cada registre, tenint en compte els diferents nivells en dBm de les lectures capturades, el nombre de sensors representats i el tipus de sensor (més o menys exposat a l'exterior). L'indicador de qualitat es calcula multiplicant per un factor que s'incrementa segons el nivell de qualitat, el sumatori de senyals associat a cada nivell qualitat del senyal:
  - $dataquality = lessthan100\ x\ 3$
  - $dataquality = dataquality + lessthan90\ x\ 4$
  - $dataquality = dataquality + lessthan80\ x\ 5$
  - $dataquality = dataquality + lessthan80\ x\ 6$
  - ...

  També es considera el fet que, les lectures hagin estat capturades pels sensors (5, 7 i 8) menys exposats a l'exterior del monument. A l'indicador de qualitat se li agrega un valor, tenint en compte la dificultat d'assolir el senyal d'algun dels tres sensors 5, 7 i 8. Aquest

valor és, per exemple, el valor de *dataquality* inicial més 10 unitats, si el registre conté un valor, pel senyal del sensor 5 (el menys exposat), i s'incrementa en 7 unitats, en el cas que el registres també contingui el senyal 7 (més exposar que el sensor 5) i s'incrementa en 5 unitats, en el cas que contingui algun valor pel senyal del sensor 8 (més exposat que el sensor 7):

- *dataquality = ifelse(idsensor5 > -90, dataquality + 10,dataquality)*
- *dataquality = ifelse(idsensor7 > -90, dataquality + 7, dataquality)*
- *dataquality = ifelse(idsensor8 > -90, dataquality + 5, dataquality)*

Detall d'un exemple, del càlcul del *dataquality* d'un registre x:

	Sensor										
	1	2	3	4	5	6	7	8	9	10	
dBm value	-79	-66	-89	-200	-200	-200	-200	-88	-200	-61	***Totals***
lessthan100											=0
lessthan90			1								=1
lessthan80	1										=1
lessthan70		1								1	=2
lessthan60											=0
if sensor5 dBm >-90											=0
if sensor7 dBm >-90											=0
if sensor8 dBm>-90								1			=1

dataquality(x) = (1x4 + 1x5 + 2x6)+(5)=26

Taula 4.1: Taula de detall d'un exemple del mètode de càlcul del nivell de qualitat d'un registre

Un cop obtingut el nivell de qualitat de cada registre, s'efectuarà un procés de neteja mitjançant els *dataquality* iguals o menors a 9. És el valor de qualitat equivalent a representar tres senyals de sensors exposats a l'exterior del museu, amb nivells de senyal d'entre -90dBm i -100dBm.

- El nombre de sensors representats en cada lectura mitjançant el nivell de senyal: S'eliminen aquells registres que, representen senyals de dos sensors (o menys) dels deu possibles. L'objectiu és gaudir d'un mínim de tres sensors (del nivell de senyal) diferents, amb la finalitat d'avaluar la triangulació d'una posició (*idplace*). En les posicions exteriors, força

allunyades, no ha estat possible disposar de les lectures de tres sensors i excepcionalment, s'ha permès un mínim de dos sensors (els senyals capturats) per registre.

Segons el procés metodològic descrit, s'aconsegueix reduir el nombre de registres un 12% (de 164 registres passem a 144 registres) i s'assoleix una millora aproximadament d'un 0,17% en l'*accuracy* (76,39%). En l'àmbit de la matriu de confusió, la millora no és massa visible:

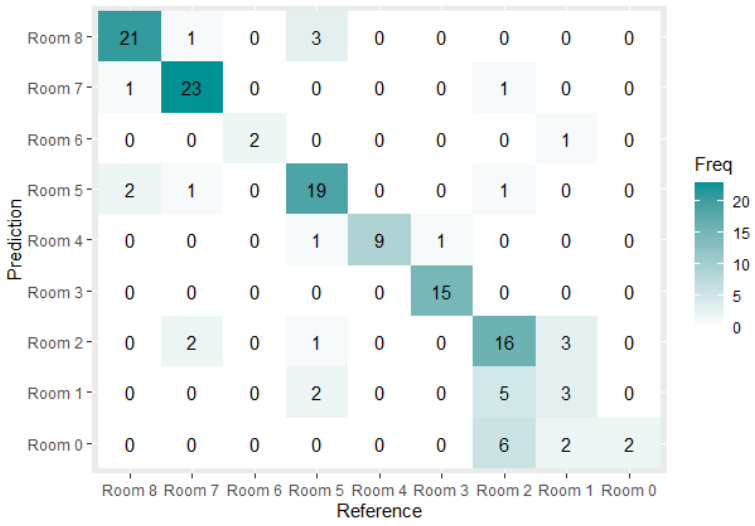

Figura 4.4: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, i amb un primer refinament segons la qualitat mínima exigible i del nombre mínim de sensors representats

El refinament per la qualitat mínima, del nombre mínim de sensors representats i de la reducció de registres correlatius per visitant i sala, aconsegueix eliminar els errors representats en l'anterior matriu de confusió, figura 4.1, entre les sales 0 i 6, les sales 3 i 6, les sales 5 i 6 i les sales 1 i 5, un cop més, sales contigües d'una mateixa planta, o superposades entre la planta baixa i la primera planta:

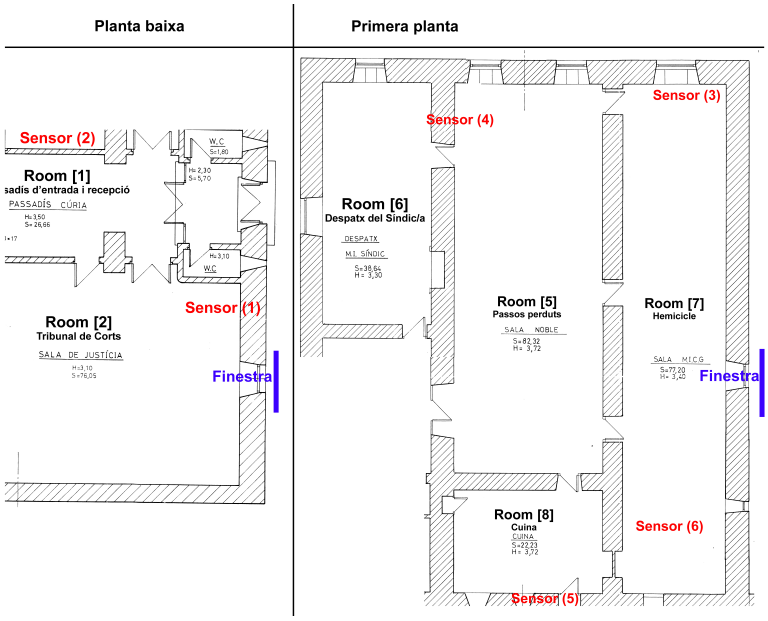

Figura 4.5: Plànol de la situació de les sales (inclosa la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

Després d'analitzar la relació de les sales realment visitades, amb les sales predites pel model, s'identifiquen salts poc coherents, provocats per *fingeprints* propers, que no tenen massa sentit en una seqüència com la que presenta la taula següent:

lecturecut	hash	idroom	idroom_predict	dataquality
2022-11-01 10:33:01	012345678910abcdefghi09876543210	5	5	22
2022-11-01 10:33:04	012345678910abcdefghi09876543210	5	5	27
2022-11-01 10:33:06	012345678910abcdefghi09876543210	5	5	25
2022-11-01 10:33:09	012345678910abcdefghi09876543210	5	5	14
2022-11-01 10:33:15	012345678910abcdefghi09876543210	5	1	17
2022-11-01 10:33:18	012345678910abcdefghi09876543210	5	5	14
2022-11-01 10:33:25	012345678910abcdefghi09876543210	5	5	27
2022-11-01 10:33:31	012345678910abcdefghi09876543210	5	5	25
2022-11-01 10:34:48	012345678910abcdefghi09876543210	6	6	16
2022-11-01 10:34:56	012345678910abcdefghi09876543210	6	6	19
2022-11-01 10:35:41	012345678910abcdefghi09876543210	5	8	27
2022-11-01 10:35:48	012345678910abcdefghi09876543210	5	5	30
2022-11-01 10:35:55	012345678910abcdefghi09876543210	8	5	25

Figura 4.6: Mostra un salt poc coherent (en vermell), de la predicció de sala (*idroom_predict*) del model, respecte a sales (*idroom*) consecutives (en verd) de la ruta monitoritzada

El model ofereix la predicció registre a registre, tenint en compte el criteri dels diferents senyals obtinguts per diferents sensors, però no té cap mecanisme per detectar si aquesta

predicció és viable, en l'àmbit dels desplaçaments dins del museu. Per millorar la coherència de la ruta predita pel model ML, utilitzarem dues estratègies de postprocessament dels registres:

4.1.1 Eliminació d'*outliers* mitjançant detecció del temps mínim real de desplaçament entre sales

La finalitat del primer tractament és fusionar els registres que, de forma correlativa, identifiquen al mateix visitant dins de la mateixa sala (*idroom_predict*), donat que de forma separada no aporten cap informació de valor afegit a la descripció de la ruta. A més, d'aquesta estratègia se'n deriva una aproximació als registres que probablement són *outliers*.

lecturecut	hash	idroom	idroom_predict	dataquality	idroom_tdist
2022-11-01 10:33:01	012345678910abcdefghi09876543210	5	5	22	2
2022-11-01 10:33:15	012345678910abcdefghi09876543210	5	1	17	14
2022-11-01 10:33:18	012345678910abcdefghi09876543210	5	5	14	3
2022-11-01 10:34:48	012345678910abcdefghi09876543210	6	6	16	90
2022-11-01 10:35:41	012345678910abcdefghi09876543210	5	8	27	53
2022-11-01 10:35:48	012345678910abcdefghi09876543210	5	5	30	7
2022-11-01 10:36:00	012345678910abcdefghi09876543210	8	8	34	12
2022-11-01 10:36:27	012345678910abcdefghi09876543210	8	5	30	147
2022-11-01 10:38:32	012345678910abcdefghi09876543210	8	8	22	5
2022-11-01 10:38:56	012345678910abcdefghi09876543210	8	7	25	24
2022-11-01 10:39:08	012345678910abcdefghi09876543210	5	5	30	12
2022-11-01 10:39:18	012345678910abcdefghi09876543210	7	7	14	10
2022-11-01 10:40:56	012345678910abcdefghi09876543210	7	2	14	98

Figura 4.7: Mostra representativa d'un salt poc coherent (en vermell) dins d'una ruta, després d'haver fusionat els registres *idroom_predict* correlatius (en verd)

Aquesta acció ha permès reduir el nombre de registres de 144 a 57 registres (una reducció del 60%). És cert que l'*accuracy* es redueix fins al 50,88%, donat que ha eliminat una part important de coincidències entre la sala de referència i la sala predita, per aquest motiu la diagonal central de la matriu de confusió ha vist reduïts els seus valors, mentre que es mantenen els valors de les prediccions errònies:

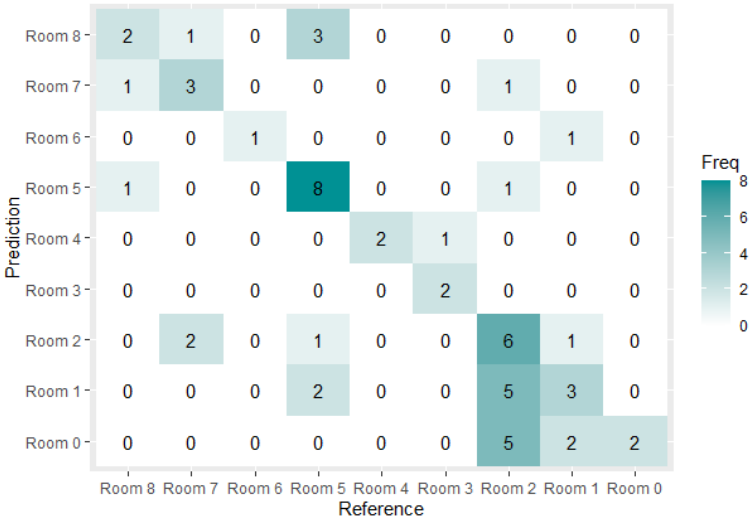

Figura 4.8: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, i refinament per la qualitat mínima, del nombre mínim de sensors representats i de la reducció de registres correlatius per visitant i sala

El refinament per la qualitat mínima, del nombre mínim de sensors representats i de la reducció de registres correlatius, aconsegueix eliminar els errors representats en l'anterior matriu de confusió, figura 4.4, entre les sales 5 i 7 i les sales 4 i 5, un cop més, sales contigües d'una mateixa planta. Alhora, ha reduït l'error anteriorment detectat entre les sales 1 i 2 i sales 5 i 8 contigües en la mateixa planta:

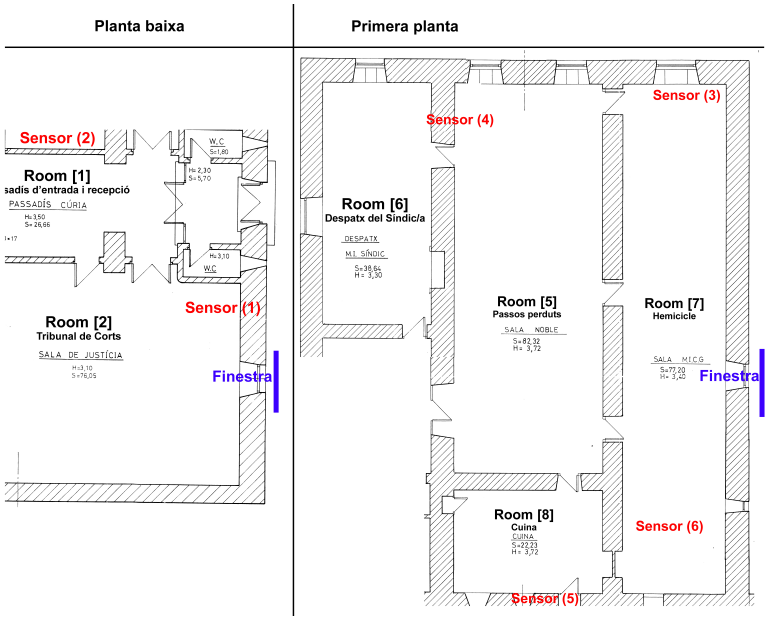

Figura 4.9: Plànol de la situació de les sales (inclosa la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

El següent pas és calcular la diferència de temps (en segons) entre els registres, d'un mateix visitant (en aquest cas, és així, donat que es tracta de la ruta monitoritzada d'un sol visitant):

lecturecut	hash	idroom	idroom_predict	dataquality	idroom_tdist	idroom_predict_tdist
2022-11-01 10:33:01	012345678910abcdefghi09876543210	5	5	22	2	0
2022-11-01 10:33:15	012345678910abcdefghi09876543210	5	1	17	14	15
2022-11-01 10:33:18	012345678910abcdefghi09876543210	5	5	14	3	15
2022-11-01 10:34:48	012345678910abcdefghi09876543210	6	6	16	90	0
2022-11-01 10:35:41	012345678910abcdefghi09876543210	5	8	27	53	5
2022-11-01 10:35:48	012345678910abcdefghi09876543210	5	5	30	7	0
2022-11-01 10:36:00	012345678910abcdefghi09876543210	8	8	34	12	0
2022-11-01 10:38:27	012345678910abcdefghi09876543210	8	5	30	147	0
2022-11-01 10:38:32	012345678910abcdefghi09876543210	8	8	22	5	0
2022-11-01 10:38:56	012345678910abcdefghi09876543210	8	7	25	24	2
2022-11-01 10:39:08	012345678910abcdefghi09876543210	5	5	30	12	0
2022-11-01 10:39:18	012345678910abcdefghi09876543210	7	7	14	10	0
2022-11-01 10:40:56	012345678910abcdefghi09876543210	7	2	14	98	25

Figura 4.10: Mostra del recàlcul del temps de desplaçament (en segons) entre sales predites

S'utilitza una nova matriu, calculada prèviament al museu, en la qual s'identifiquen dues a dues, la distància de temps (en segons) real, que un humà dedica en el desplaçament entre sales:

Rooms	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
0	0s	0s	2s	4s	6s	20s	25s	25s	25s
1	0s	0s	0s	0s	0s	15s	21s	18s	18s
2	2s	0s	0s	5s	3s	18s	28s	25s	25s
3	4s	0s	5s	0s	0s	10s	18s	15s	15s
4	6s	0s	3s	0s	0s	0s	6s	5s	2s
5	20s	15s	18s	10s	0s	0s	0s	0s	0s
6	25s	21s	28s	18s	6s	0s	0s	5s	5s
7	25s	18s	25s	15s	5s	0s	5s	0s	2s
8	25s	18s	25s	15s	2s	0s	5s	2s	0s

Taula 4.2: Matriu que identifica la distància (en segons) real entre les diferents sales, necessària per al desplaçament d'un visitant

La finalitat és, confirmar que la necessitat de temps per efectuar el desplaçament d'un visitant entre dues sales associades a un salt, és major al temps associat a una predicció errònia. Si s'aconsegueix demostrar aquesta hipòtesi, s'eliminaran el registres afectats, donat que representen desplaçaments físicament impossible.

lecturecut	hash	idroom	idroom_predict	dataquality	idroom_tdist	idroom_predict_tdist
2022-11-01 10:33:01	012345678910abcdefghi09876543210	5	5	22	2	0
2022-11-01 10:34:48	012345678910abcdefghi09876543210	6	6	16	90	0
2022-11-01 10:35:41	012345678910abcdefghi09876543210	5	8	27	53	5
2022-11-01 10:35:48	012345678910abcdefghi09876543210	5	5	30	7	0
2022-11-01 10:36:00	012345678910abcdefghi09876543210	8	8	34	12	0
2022-11-01 10:38:27	012345678910abcdefghi09876543210	8	5	30	147	0
2022-11-01 10:38:32	012345678910abcdefghi09876543210	8	8	22	5	0
2022-11-01 10:38:56	012345678910abcdefghi09876543210	8	7	25	24	2
2022-11-01 10:39:08	012345678910abcdefghi09876543210	5	5	30	12	0
2022-11-01 10:39:18	012345678910abcdefghi09876543210	7	7	14	10	0
2022-11-01 10:40:56	012345678910abcdefghi09876543210	7	2	14	98	25
2022-11-01 10:41:32	012345678910abcdefghi09876543210	7	8	17	22	2
2022-11-01 10:42:15	012345678910abcdefghi09876543210	7	7	20	43	2

Figura 4.11: Mostra del refinament per temps mínim de desplaçament, efectuat un cop s'elimina el salt

Un cop realitzat aquest procés de neteja, s'han reduït els 57 registres anteriors, fins als 42 registres. És cert que l'eliminació dels registres de dubtosa viabilitat, es podria substituir per la mutació de sales, segons l'error més habitual. Una millora d'aquest mètode seria cercar quines són les mutacions estadísticament més comunes, i quina de les considerades temporalment viables, maximitzaria la probabilitat de representar la sala correcta.

Per finalitzar aquest procés, es realitza una actualització de la matriu de confusió, i queda confirmat que tot i tenir en compte un nombre molt més baix (-26%) de registres, l'encert es manté pràcticament estable (millora un 1,5%), situant-se al 52,38%:

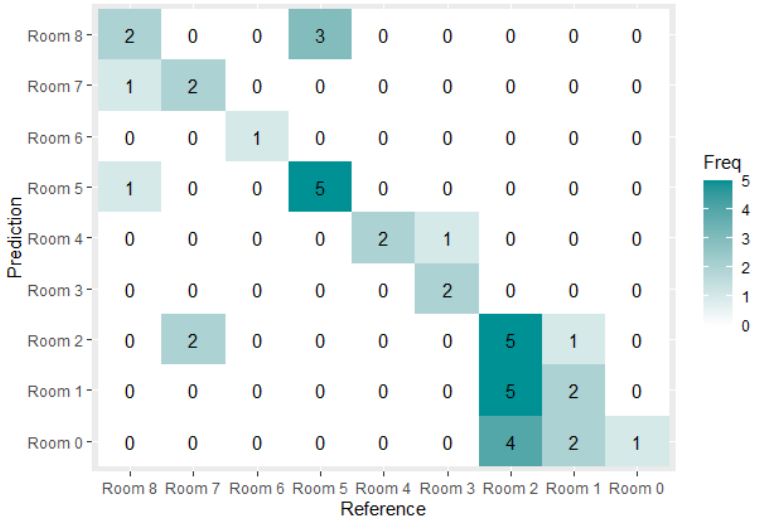

Figura 4.12: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, i amb un segon refinament mitjançant el filtre de desplaçaments temporalment impossibles

El refinament pel filtre de desplaçaments temporalment impossible, aconsegueix eliminar els errors representats en l'anterior matriu de confusió, figura 4.8, entre les sales 7 i 8, les sales 5 i 1, les sales 5 i 2, les sales 5 i 2, les sales 2 i 5 i les sales 2 i 7 i les sales 1 i 6:

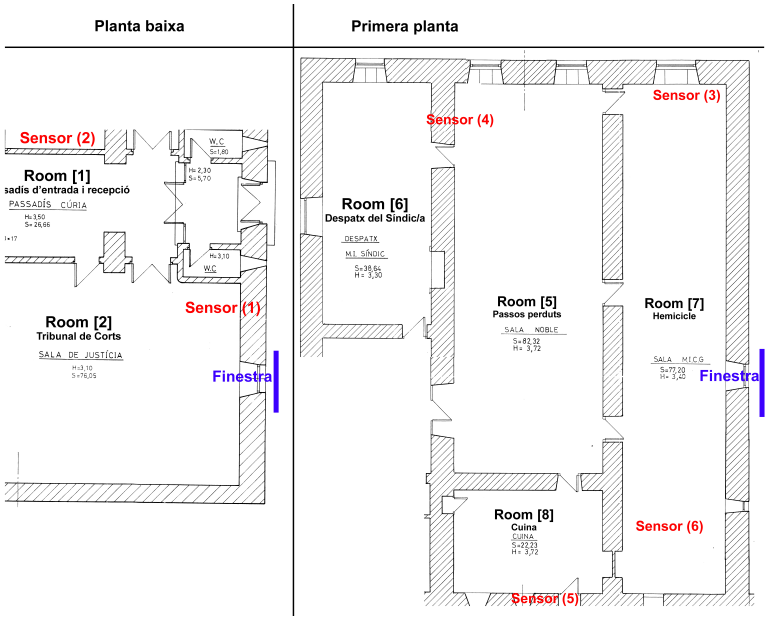

Figura 4.13: Plànol de la situació de les sales (inclosa la sala 6) i sensors de la planta baixa i primera planta, alineades mitjançant la finestra exterior d'ambdues plantes

Un cop realitzat aquest procés de neteja, poden tornar a aparèixer registres consecutius geoposicionats en la mateixa sala. Per resoldre aquesta situació, es fa una nova compactació d'aquests, i dels 42 registres previs, la mostra es redueix mínimament, fins a situar-se amb 40 registres. L'encert millora un 2,62% i queda situat en un 55,00%.

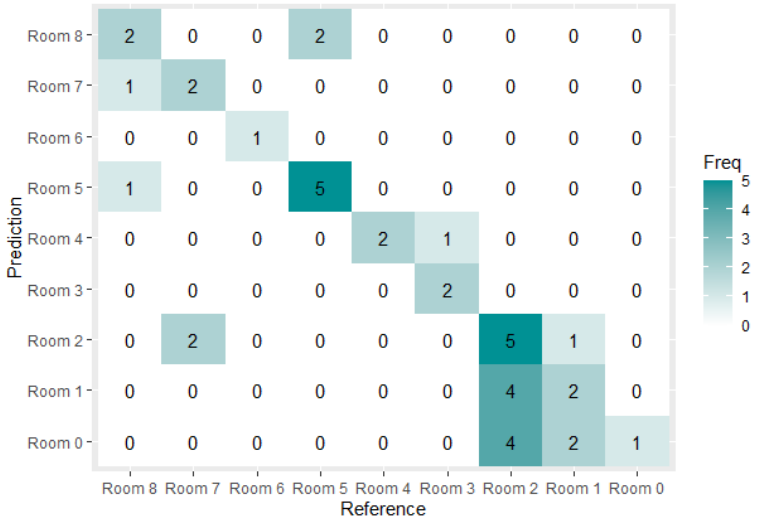

Figura 4.14: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, amb el segon refinament mitjançant el filtre de desplaçaments temporalment impossibles i eliminació de registres correlatius d'un mateix visitant, en una mateixa sala

El darrer procés de neteja, no aconsegueix eliminar errors de predicció respecte a l'anterior matriu de confusió, figura 4.12, però sí reduir-ne l'error entre les sales 5 i 9 i les sales 2 i 1: Es podria realitzar una crida recursiva, fins a assolir una situació de convergència, però a partir de prediccions errònies, la recursivitat d'aquest tractament podria provocar la canibalització dels registres de la sessió.

Es pren la ruta predita i filtrada, es torna a calcular la diferència de temps (en segons) entre els registres consecutius, per calcular el temps mitjà d'ocupació predit (*mean_predict_min*) del visitant monitoritzat per cada sala (*idroom*), i es compara amb el temps d'ocupació real per sala (*mean_real_min*) del visitant monitoritzat (detallat en la documentació realitzada durant la ruta), per obtenir l'error aproximat (50,77%):

idroom	***mean_predict_min***	***mean_real_min***	***%_error***
0	1,75	1,00	75%
1	1,33	2,86	53%
2	4,50	2,25	100%
3	4,28	3,85	15%
4	1,63	1,71	5%
5	2,81	1,20	150%
6	0,88	1,20	26%
7	3,98	5,51	28%
8	3,33	3,16	5%

Taula 4.3: Taula de detall de l'error entre el temps mitjà predit (*mean_predict_min*) d'ocupació per cada sala (*idroom*) del visitant monitoritzat pel mètode del temps mínim real de desplaçament entre sales, i el temps mitjà real (*mean_real_min*) d'ocupació per cada sala del visitant monitoritzat.

Finalment, es genera una matriu descriptiva de les rutes obtingudes, que s'utilitzarà al final de l'estudi, per analitzar les diferents rutes predites i extreure'n conclusions.

4.1.2 Dissolució d'*outliers* mitjançant finestra

L'objectiu d'aquest tractament alternatiu és, l'avaluació *minut a minut*, de les prediccions del model i la selecció de la sala més probable, utilitzant la moda estadística, per minut. És un mètode que en el cas de disposar de diversos registres per minut, pretén dissoldre els *outliers*, donat que la seva aparició sol ser residual.

Partint del conjunt de dades previ al mètode d'eliminació de registres *outliers*, mitjançant la detecció del temps mínim real d'un desplaçament entre sales, i previ a la neteja de registres

pels dos criteris de nivell de qualitat i nombre de sensors partícips del registre detallat en el punt 4.1 (164 registres), aquest mètode assoleix un encert del 96% i una matriu de confusió amb una diagonal molt neta:

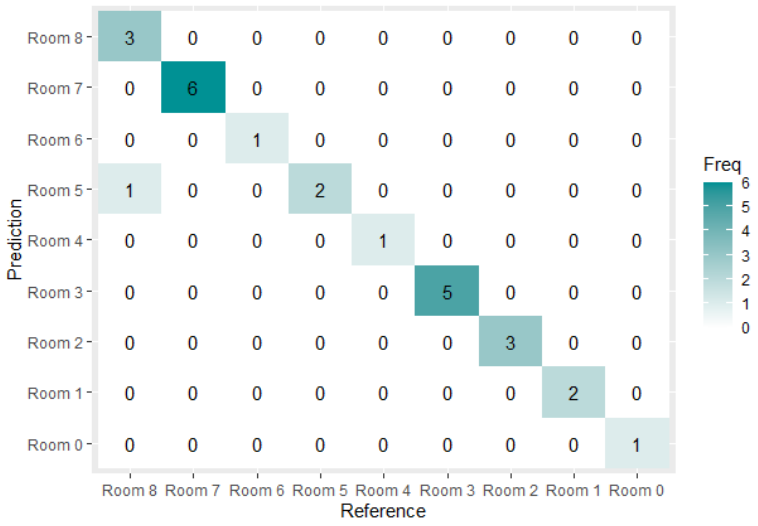

Figura 4.15: *Confusion Matrix* de la predicció sobre la ruta monitoritzada, utilitzant finestra

El darrer mètode de neteja utilitzant finestra, aconsegueix eliminar errors de predicció respecte a l'anterior matriu de confusió, figura 4.14, entre les sales 8 i 7, les sales 7 i 2, les sales 5 i 8, les sales 3 i 4, les sales 2 i 0, les sales 2 i 1, les sales 1 i 0, les sales 1 i 2, en resum, només manté un únic error de predicció entre les sales 8 i 5, consecutives en la primera planta:

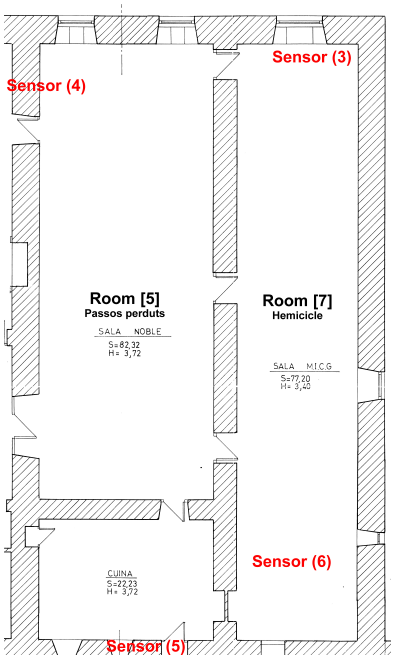

Figura 4.16: Plànol de la situació de les sales 5 i 7 i els seus sensors, de la primera planta

En l'àmbit de la predicció del temps de permanència a cada sala, aquest darrer mètode, respecte al mètode anterior, millora els percentatges d'encert i redueix un 50% l'error mitjà la predicció del temps, fins al 28,22%. És curiós, donat que l'arrodoniment a minut, lògicament hauria d'augmentar error de la predicció.

idroom	*mean_predict_min*	*mean_real_min*	*%_error*
0	1	1,00	0%
1	2	2,86	30%
2	3	2,25	34%
3	5	3,85	30%
4	1	1,71	41%
5	2	1,20	67%
6	1	1,20	16%
7	6	5,51	9%
8	4	3,16	27%

Taula 4.4: Taula de detall de l'error entre el temps mitjà predit (*mean_predict_min*) d'ocupació per cada sala (*idroom*) del visitant monitoritzat pel mètode de dissolució d'*outliers* mitjançant finestra, i el temps mitjà real (*mean_real_min*) d'ocupació per cada sala del visitant monitoritzat

4.2 Estimació de durada de sessions i ocupació per franja horària, de rutes tipus MAC estàtica o prefix MAC

La característica principal de les lectures seleccionades en aquesta fase, és que mantenen la MAC estàtica o el prefix MAC DA:A1:19 durant tota la ruta (o fins i tot, durant diferents dies). Aquesta característica assegura l'estabilitat del resum *Hash* durant tota la ruta, i la relació d'un a un, entre cadascun dels resums *Hash* del dispositiu del visitant que el publica. L'objectiu és obtenir l'estimació del temps mitjà per sessió i l'ocupació del museu en franges horàries per data.

S'ha realitzat una preselecció de les sessions d'entre 10 i 150 minuts, que inclouen lectures de 9 dels 10 sensors desplegats pel museu. Es considera aquest rang de temps, el mínim i màxim raonable d'una visita al monument. D'aquesta preselecció, s'han obtingut 10.776 lectures. Sobre les lectures preseleccionades, s'aplica l'algorisme descrit en el punt 3.3, algorisme 1, que transforma les lectures en registres. De les 10.776 lectures, s'obtenen 3.643 registres. Tenint en compte que els registres mantindran l'atribut resum *Hash*, i amb la finalitat d'obtenir la durada del temps (en segons) de la sessió, es recupera el moment d'inici i fi de cada sessió, i s'afegeix com nou atribut de cada registre. Dels 3.643 registres, s'obtenen 147 sessions. Els estadístics determinats per les durades (en temps) de les 147 sessions són:

min	10m 27s
1r. qu	31m 53s
median	41m 20s
mean	46m 42s
3r. qu	56m 26s
max	2h 24m 40s

Taula 4.5: Taula d'estadístics d'estimació de la durada mitjana de les sessions de les rutes de lectures de tipus MAC estàtica o prefix MAC DA:A1:19

S'aprofita l'estimació de la durada mitjana de les diferents sessions, i s'assumeix que respecta la distribució normal. Partint de les lectures del tipus MAC estàtica o amb prefix MAC DA:A1:19, es calcula un factor de lectures per visitant i minut. Es pot realitzar la correlació lineal de l'ocupació de monument, aplicant el factor al número de lectures complementàries (MAC aleatòria o SSID informada) a les utilitzades per calcular el factor (MAC estàtica o amb prefix MAC DA:A1:19). Apareix un cas més complex, de lectures produïdes per sessions de diferents individus de forma paral·lela, sense que les diferents sessions siguin coincidents en l'hora d'inici i fi.

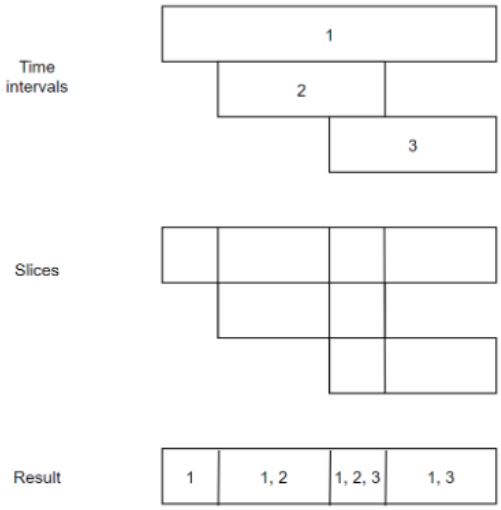

Figura 4.17: Exemple de *session overlapping* en un mateix *slot* de temps

Per aquest motiu, s'utilitza un algorisme que impacta el nombre de minuts i lectures a la porció (*slot*) d'hora que li correspon a cada sessió. Un cop efectuat aquest agregat, es calcula la mitjana de lectures per hora, tenint en compte el nombre de lectures (tipus MAC estàtica o prefix MAC DA:A1:19), agregant els minuts i nombre de sessions concurrents (*session overlapping*).

Una millora al mètode proposat, seria calcular el percentatge de participació (en minuts) que cada sessió suposa a cada porció horària, donat que tot i que participi de forma residual (pocs minuts), la divisió es realitzaria de forma proporcional al nombre de sessions concurrents, i no al nombre de minuts aportats per sessió. S'identifiquen 112 porcions horàries (*slots*). No totes les franges horàries de tots els dies estan representades, donat que poden no estar participades per sessions del tipus tractat. Un cop es disposa de la mitjana de lectures per hora, capturades de cada visitant amb MAC estàtica o prefix MAC DA:A1:19, s'obtenen agrupats per hora, el nombre de lectures dels registres complementaris (MAC aleatòria o SSID informada).

La taula següent, es mostren les lectures de nou franges horàries, del tipus MAC estàtica i prefix MAC DA:A1:19, amb els minuts d'*overlapping* de les diferents sessions i el nombre de sessions individuals:

slot	*lect_by_hour_hash_or_prefix*	*min_overlapping*	*uniq_hash*
2022-10-14 09h	5	3,01	1
2022-10-14 10h	28	60,00	1
2022-10-14 11h	111	59,76	3
2022-10-14 12h	65	34,00	2
2022-10-15 09h	1	1,90	1
2022-10-15 10h	549	329,31	6
2022-10-15 11h	193	133,03	6
2022-10-15 16h	14	2,81	1
2022-10-15 17h	64	28,43	1

Taula 4.6: Taula de mostra del nombre de lectures, per franja horària, del tipus MAC estàtica i prefix MAC DA:A1:19, amb els minuts d'*overlapping* de les diferents sessions i el nombre de sessions individuals

Tenint en compte aquests valors, es pot extrapolar el nombre de visitants (MAC aleatòria o SSID informada), i sumant ambdues estimacions, s'ha d'obtenir el total de visitants concurrents del museu (sense geolocalització).

Cal afegir la consideració que, el soroll provocat en l'entorn del museu, i captat pels sensors (majoritàriament exteriors, però també interiors) afegeix un error de paquets de descobriment tractats com a visitants, que realment no ho són. Es poden aproximar, tenint en compte els dies de tancament del monument.

Com a informació complementària, es mostren els percentatges que suposen les lectures capturades pels sensors menys exposats (5, 7 i 9) als exteriors, respecte al total de lectures capturades pels deu sensors desplegats. S'han agregat per hora del dia, donat que els diferents patrons de mobilitat, varien segons les hores i els dies (entrada o sortida de l'escola, feina, oci...). Cal dir que només es mostren les hores que afecten l'objectiu de l'estudi.

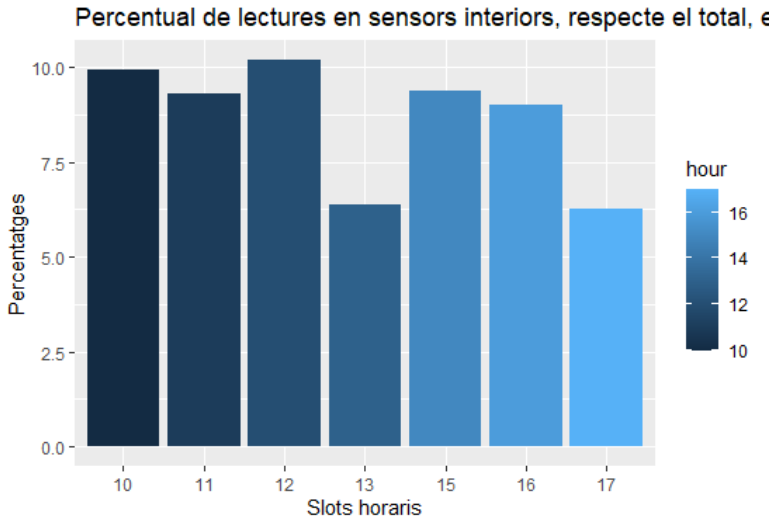

Figura 4.18: Representació percentual entre els sensors menys exposats a l'exterior i els més exposats, segons l'hora

Es confirma que el percentatge retrocedeix, en els moments d'altra freqüentació del carrer i/o places adjacents, donat que són els moments habituals de desplaçaments, fruit de l'inici, interrupció o fi, de la jornada laboral i/o escolar. De forma genèrica, s'identifica una disminució del 10% al 6% de mitjana, de lectures capturades pels sensors menys exposats, respecte al total de sensors, a les 13h i a les 17h. Per obtenir una aproximació del nivell mitjà de soroll, que ha de permetre refinar les estimacions d'ocupació per hora del museu, i assolir la xifres de lectures originades únicament pels visitants, es fa una selecció dels dilluns de tancament del museu, en els quals no s'identifiquen visites privades (o tasques de manteniment).

Es podria realitzar un càlcul equivalent, pel que fa als caps de setmana, donat que el patró de mobilitat exterior variarà radicalment, però s'ha considerat que aportarà poca millora al refinament, en relació a la dedicació requerida, donat que els horaris en dissabte i diumenge pateixen variacions els dos mesos pels quals s'han obtingut lectures (octubre i novembre).

Per aquest motiu, s'utilitzaran els dilluns 24/10/2022 i 14/11/2022, ja que després d'analitzar les lectures dels diferents dilluns dels mesos tractats, són aquests dos dilluns els que ofereixen els criteris qualitatius descrits anteriorment. S'agrupen les lectures per franja horària, i es calcula la mitjana. Les captures de paquets de descobriment capturats en dilluns, assoleix la xifra de 43.716 lectures. Un cop filtrades pels horaris i dilluns objectiu, s'obtenen 9.518 lectures. El següent pas és calcular la mitjana dels dos dilluns, i extreure els indicadors per:

- Lectures de tots els tipus
- Lectures tipus MAC estàtica o prefix MAC DA:A1:19

- Lectures tipus MAC aleatòria o SSID informat

- Lectures de tots els tipus, dels sensors menys exposats a l'exterior (5, 7 i 9)

Els valors representatius de la correcció de soroll mitjà, segons els tipus de registres anteriors són:

slot	***correction_all***	***correction_hashprefix***	***correction_aleassid***	***correction_579***
10h	371	75	296	296
11h	715	101	614	614
12h	1.740	72	1.668	1.668
13h	1.294	181	1.113	1.113
15h	754	105	648	648
16h	1.081	45	1.035	1.035
17h	640	108	532	532

Taula 4.7: Taula de valors representatius de la correcció de soroll mitjà

Es considera, com a valor de lectures per hora d'un dia d'obertura (tipus MAC aleatòria o SSID informada), la diferència entre la mitjana de lectures total per hora i la correcció del soroll per hora, quantificades segons els dilluns de tancament del museu. Cal recordar que les lectures tipus MAC estàtica o prefix MAC DA:A1:19 es comptabilitzaran directament com a visitants, i no caldrà considerar-les en el prorrateig de lectures per hora.

La taula següent, mostra un exemple d'estimació dels visitants totals per dia i franja horària, tenint en compte la mitjana de lectures tipus MAC estàtica o prefix MAC DA:A1:19, aplicada a les lectures complementàries, menys la correcció del soroll tipus MAC aleatòria o SSID informada:

slot	*uniq_hash*	*estima_visitors_alea_ssid*	*estima_visitors_all*
2022-10-14 10h	1	105	106
2022-10-14 11h	3	16	19
2022-10-14 12h	2	19	21
2022-10-15 10h	6	23	29
2022-10-15 11h	6	11	17
2022-10-15 16h	1	14	15
2022-10-15 17h	1	20	21

Taula 4.8: Taula de mostra de l'estimació dels visitants totals per dia i franja horària, tenint en compte la mitjana de lectures tipus MAC estàtica o prefix MAC DA:A1:19, aplicada a les lectures complementàries, menys la correcció del soroll tipus MAC aleatòria o SSID informada

4.3 Validació de la durada mitjana de sessions i ocupació per franja horària, de rutes de lectures tipus MAC estàtica o prefix

No es disposa de cap font de dades fiable, que permeti la validació de la durada mitjana de les sessions de rutes. Tècnicament, es pot considerar com a temps de durada de la sessió, al temps entre el primer i el darrer paquet de descobriment capturat per a un visitant concret. En aquest cas, apareixerà l'error provocat per visitants que poden romandre en els entorns del monument, més temps del que dedicaran efectivament a fer estrictament la visita. Una millora del mètode, hauria de permetre quantificar aquest tipus de desviació de temps, previ o posterior a la visita del monument.

Existeix l'oportunitat de realitzar la validació de l'estimació de l'ocupació del museu per franja horària, però prèviament, cal contextualitzar els indicadors estadístics externs a aquest estudi, que s'utilitzaran.

Després de presentar aquest projecte de recerca al Departament de museus i monuments nacionals del Ministeri de Cultura i Esports del Govern d'Andorra, es va sol·licitar accés a les dades estadístiques de les reserves i vendes d'aquest museu. La documentació facilitada per aquest departament conté indicadors quantitatius estructurats per:

- Mes i dia, per franja horària (en divisions d'una hora)
- Individuals o per grups (amb el nombre de persones del grup)
- Nacionalitat

- Franja d'edat

La informació compartida pel departament en qüestió, ha estat de molta utilitat pel procés de validació de les estimacions efectuades. Es va estructurar en un format normalitzat (220 registres) i va ser introduïda a la base de dades del projecte. Els mesos d'octubre i novembre, es dóna una casuística interessant. Són els mesos en què es desenvolupen les visites culturals dels alumnes dels diferents sistemes educatius d'Andorra. Curiosament, aquest *target* de visitant, en l'àmbit de l'estadística facilitada, queda classificat en un rang preestablert d'entre 10 i 18 anys. Tot i que inicialment no semblava que pogués ser una classificació problemàtica, es considera la premissa que un alumne de 10 anys difícilment durà un dispositiu mòbil al centre educatiu, mentre que un de 16 anys, és molt probable que sí que el dugui. Aquesta situació afegeix un element de desviació que cal considerar, durant aquesta fase de validació entre les dades estimades i les dades reals. Altres casuístiques detectades són, a nivell estadístic, disposar d'indicadors de grups en franges acotades a una hora. A nivell de les lectures dels paquets de descobriment, es confirma empíricament, que per motiu d'organització (o altres), apareixen grups que s'encavallen entre franges horàries correlatives.

En una primera fase, el procediment de validació dels indicadors d'ocupació del museu, considerava totes les lectures dels deu sensors instal·lats. El càlcul de la mitjana, entre la predicció d'ocupació i l'ocupació real, pateix un error d'aproximadament un 163,48%. Es mostra un exemple de les desviacions entre l'ocupació real, segons les estadístiques i les estimacions de visitants, obtingudes del mètode de càlcul anteriorment detallat:

slot	***occup_real***	***estima_visitors_all***	***%_error***
2022-10-14 10h	6	106	1.666,66
2022-10-14 11h	57	19	66,66
2022-10-14 12h	30	21	30,00
2022-10-15 10h	53	29	45,28
2022-10-15 11h	35	17	51,42
2022-10-15 16h	11	15	36,36
2022-10-15 17h	6	21	250,00

Taula 4.9: Taula de mostra de les desviacions entre l'ocupació real segons les estadístiques i les estimacions de visitants

Es considera un error massa elevat, i per aquest motiu es realitza una segona fase en la qual es pren en consideració la hipòtesi que, la desviació està condicionada per la posició dels sensors, l'hora i el tipus de dia (dies entre setmana o caps de setmana) i que de forma majoritària, l'error és provocat per les lectures originades pels no visitants del monument. Per aquest motiu, es

realitzarà una segona estimació limitant l'anàlisi de lectures capturades pels tres sensors menys exposats a l'exterior del monument i, per tant, que es veuen menys afectats per la circulació de persones (carrers i/o places). Partint de l'ocupació real (segons estadístiques), la mitjana de les lectures capturades per hora, dels sensors menys exposats, menys el nivell de soroll dels sensors menys exposats, permetrà obtenir una ràtio de paquets per minut i visitant. Un cop assolida aquesta ràtio, s'efectua la predicció de l'ocupació, utilitzant les lectures capturades pels sensors menys exposats (5, 7 i 9). La següent taula presenta una mostra de l'ocupació real (segons les estadístiques facilitades), les estimacions sense correcció de soroll (*no correction*) i amb correcció de soroll (*with correction*) i els dos errors percentuals associats a cada tipus de correcció:

occup_real	***estima_visitors_579nc***	***estima_visitors_579wc***	***%_errornc***	***%_errorwc***
6	12	12	100,00	100,00
53	54	57	1,88	7,54
6	2	2	66,66	66,66
58	12	13	79,31	77,58
27	14	15	48,14	44,44
0	21	22	21,00	22,00
4	3	3	25,00	25,00
1	1	1	0,00	0,00

Taula 4.10: Taula de mostra de l'ocupació real (segons les estadístiques facilitades), les estimacions sense correcció de soroll (*no correction*) i amb correcció de soroll (*with correction*) i els dos errors percentuals associats a cada tipus de correcció

Com a validació d'aquest mètode d'estimació de l'ocupació, respecte a les estadístiques facilitades pel Departament de museus i monuments nacionals, utilitzant únicament els sensors menys exposats a l'exterior (5, 7 i 9), s'obtenen ràtios d'error del:

- 60,98% si no es té en compte el soroll mitjà (*nc*).
- 63,62% si es té en compte el soroll mitjà (*wc*).

Sembla contradictori que l'extracció del soroll mitjà, respecte a la no extracció del soroll mitjà, incrementi l'error, i no es disposa d'una justificació concreta d'aquest fet.

4.4 Geoposició i càlcul de la duració mitjana de permanència per sala de MAC estàtica o prefix MAC

L'objectiu d'aquesta fase és, geoposicionar i calcular el temps d'ocupació en les diferents sales, utilitzant els paquets de descobriment tipus MAC estàtica o prefix MAC DA:A1:19 . El nombre de registres processats és de 3.643.

S'utilitzarà el mètode detallat en el punt 4.1. que realitza un procés de neteja de registres, utilitzant dos criteris objectius: la qualitat de les lectures (redueix la mostra de les 3.643 a 1.410 registres) i el número mínim de senyals de diferents sensors (redueix la mostra de 1.410 a 1.366 registres), representades en cada lectura. La col·lecció de registres ja filtrada, s'introdueix en el model entrenat del punt 4.6, i obté la predicció d'un lloc (*idplace*) per cada registre, i se l'infereix la sala (*idroom*) a la qual pertany el lloc. Tenint en compte el concepte de sessió d'un visitant, es simplifica la permanència en una mateixa sala, realitzant compactació dels registres consecutius, sempre que pertanyin a la mateixa sala. Aquest procés redueix el nombre de registres dels 1.366 anteriors, fins als 1.093.

S'efectua una nova tria, amb l'objectiu de maximitzar la qualitat de les sessions obtingudes. S'eliminaran aquelles sessions, per les quals la predicció no hagi identificat cinc de les nou sales existents (més de la meitat). Aquest filtre reduirà dels 1.093 registres anteriors, fins als 859 actuals.

És el moment de realitzar el càlcul del temps que cada visitant roman en les diferents sales, calculant la distància (en segons) basada en els geoposicionaments. En aquest punt, s'aplica el mètode de neteja de dades descrit en el punt 4.1.1, que té en compte la distància real (en segons) entre sales, i que elimina aquells registres pels quals, la distància real (en segons), és major que la distància predita (desplaçament físicament impossible). Aquesta operació de neteja, permet reduir el nombre de registres fins als 758. Un cop realitzat aquest procés de neteja, poden tornar a aparèixer registres consecutius geoposicionats en la mateixa sala. Per resoldre aquesta situació, es realitza una nova compactació d'aquests, i dels 758 registres previs, es redueix fins als 674 registres.

idroom	*mean_predict_min*
0	2,93
1	2,83
2	3,15
3	3,72
4	2,80
5	4,33
6	2,54
7	3,79
8	3,89

Taula 4.11: Taula del temps mitjà d'ocupació dels visitants, amb geoposicionament basat en MAC estàtica o prefix MAC, per sala

Donat que en aquest cas, no es tracten rutes monitoritzades, no es disposa d'una mitjana de temps per sala que ens permeti la validació de la predicció. Finalment, s'afegeixen les rutes predites a la matriu descriptiva de les rutes ja obtingudes pels mètodes anteriors, i s'utilitzarà al final de l'estudi, per analitzar-les i extreure'n conclusions.

4.5 Duració de la mitjana de sessions amb SSID informada

S'utilitza el mateix criteri de preselecció de sessions del tipus SSID informada, que en punt 4.2. (rang de temps de sessió d'entre 10 i 150 minuts, i lectures de nou dels deu sensors instal·lats). Un cop seleccionades les 14.283 lectures d'aquest tipus, s'utilitza l'algorisme de transformació de lectures (un sensor capta una lectura en un moment concret) en registres, agrupant en un rang de temps determinat, els millors valors (més propers a zero) de cada sensor, en un sol registre, descrit en el punt 4.1, però s'executa en aquest apartat adaptant l'algorisme 1, al nom d'SSID en comptes del resum *Hash*. S'obtenen 4.775 registres que, un cop agrupats utilitzant les respectives SSID, formen 148 sessions. Cal preveure que, la qualitat dels indicadors obtinguts pels tipus de sessions amb lectures tipus SSID informada, probablement serà menor, ja que diferents dispositius poden emetre el mateix nom d'SSID. Es calcula la duració mitjana de les sessions, i s'obté un valor de 51 minuts i 34 segons, equivalent a les estimacions d'altres tipus de lectures, però amb un lleuger increment de temps (d'un +11% respecte a les rutes produïdes per lectures del tipus MAC estàtica o prefix MAC DA:A1:19), que pot ser conseqüència que, diferents dispositius d'un mateix grup de visitants, publiquin el mateix nom d'SSID. Aquest

fet augmentaria la densitat de publicacions de paquets de descobriment per sessió, afectant en excés, al temps de permanència.

min	13m 1s
1r. qu	34m 40s
median	43m 52s
mean	51m 34s
3r. qu	1h 1m 41s
max	2h 21m 26s

Taula 4.12: Taula d'estadístics de la duració mitjana de les sessions de visitants amb SSID informada

S'aprofita el detall d'estimacions de les duracions de sessions, i s'assumeix que respecten la distribució normal. S'afegeixen les rutes obtingudes per aquest mètode, a la matriu de rutes ja obtingudes pels mètodes anteriors, que s'utilitzarà al final de l'estudi, per analitzar les rutes predites i extreure'n conclusions.

Finalment, es realitza la comparació entre els temps mitjà i medià de duració de les sessions, segons els tipus de registres MAC estàtica o prefix MAC, i MAC aleatòria i SSID informada. La variació entre ambdós tipus és relativament baixa, per tant, la qualitat de l'estimació és probable que sigui elevada.

	MAC estàtica i prefix	***MAC aleatòria i SSID informada***	***Error***
median	41m 20s	43m 52s	6,1%
mean	46m 42s	51m 34s	10,4%

4.6 Geoposició i càlcul de duració mitjana de permanència per sala amb SSID informada

Es recuperen els 4.775 registres del punt anterior 4.5, però es realitza el procés de neteja de registres pels dos criteris de nivell de qualitat (redueix la mostra fins als 1.983 registres) i nombre mínim de sensors partícips del registre (redueix la mostra fins als 1.876 registres) detallat en el punt 4.1. No era un procés interessant, en l'àmbit del càlcul de la durada mitjana de les sessions, però ha quedat demostrada la millora de la qualitat en els processos de geoposicionament dels visitants. S'introdueix en resultat en el model, i s'obté la predicció dels llocs (*idplace*) associats als *fingerprints*, i s'infereix la sala corresponent (*idroom*).

S'efectua una nova tria, amb l'objectiu de maximitzar la qualitat de les sessions obtingudes. S'eliminaran aquelles sessions per les quals, la predicció no hagi identificat cinc de les nou sales existents (més de la meitat). Aquest filtre reduirà els 1.876 registres anteriors, fins als 1.639 actuals.

Tenint en compte el concepte de sessió d'un visitant, es simplifica la permanència en una mateixa sala, realitzant compactació dels registres consecutius, sempre que pertanyin a la mateixa sala. Aquest procés redueix el nombre de registres fins als 1.138. És el moment de realitzar el càlcul del temps que cada visitant roman en les diferents sales, calculant la distància (en segons) basada en els geoposicionaments. En aquest punt, s'aplica el mètode de neteja de dades descrit en el punt 4.1.1, que té en compte la distància real (en segons) entre sales, i que elimina aquells registres pels quals la distància real (en segons), és major que la distància predita (desplaçament físicament impossible). Aquesta operació de neteja, permet reduir el nombre de registres fins als 1.002.

Un cop realitzat aquest procés de neteja, poden tornar a aparèixer registres consecutius geoposicionats en la mateixa sala. Per resoldre aquesta situació, es realitza una nova compactació d'aquests, i dels 1.002 registres previs, es redueix a 943 registres. Es podria realitzar una crida recursiva, fins a assolir una situació de convergència, però a partir de prediccions errònies, la recursivitat d'aquest tractament, podria provocar la canibalització dels registres de la sessió. Finalment, s'utilitza la ruta predita i filtrada, per calcular el temps mitjà d'ocupació del visitant amb lectures de tipus SSID informada, per cada sala:

idroom	*mean_predict_min*
0	2,31
1	2,62
2	1,75
3	2,43
4	2,58
5	2,22
6	2,33
7	3,06
8	2,51

Taula 4.13: Taula del temps mitjà d'ocupació dels visitants, amb geoposicionament basat en MAC aleatòria i SSID informada, per sala

Donat que en aquest cas no es tracten rutes monitoritzades, no es disposa d'una mitjana de temps per sala, que ens permeti la validació de la predicció. Però es pot comparar amb les

duracions mitjanes del tipus MAC estàtica o prefix DA:A1:19, i s'obté una desviació mitjana del 40%:

idroom	*mean_predict_min_hashprefix*	*mean_predict_min_ssid*	*%_error_compar*
0	2,93	2,31	27%
1	2,83	2,62	8%
2	3,15	1,75	80%
3	3,72	2,43	53%
4	2,80	2,58	9%
5	4,33	2,22	95%
6	2,54	2,33	9%
7	3,79	3,06	24%
8	3,89	2,51	55%

Taula 4.14: Taula de comparació entre el temps mitjà d'ocupació dels visitants del tipus MAC estàtica o prefix MAC, i el tipus MAC aleatòria i SSID informada, i càlcul de la desviació

No es disposa de cap font d'informació que permeti la validació de la predicció obtinguda pel model, utilitzant el tipus de lectures amb SSID informada. S'afegeixen les rutes obtingudes per aquest mètode, a la matriu de rutes existent, que s'utilitzarà al final de l'estudi, per analitzar les rutes predites i extreure'n conclusions.

4.7 Geoposicionament amb MAC aleatòria

De la mostra total de registres amb MAC aleatòria, s'extreuen els registres de tipus SSID informada, ja tractats en el punt anterior 4.6. Es genera una col·lecció de registres, utilitzant els mecanismes descrits en els punts 4.4 i 4.6. El resultat, s'introdueix en el model, i s'obté la predicció dels llocs (*idplace*) associats als *fingerprints*, i s'infereix la sala corresponent (*idroom*). De les 513.610 lectures inicials, s'obtenen 289.935 registres. Es detecta que els registres obtinguts seran d'una qualitat baixa, donat que majoritàriament, contenen dos nivells de senyal de sensors, per registre. Aplicant els filtres del punt 4.1, en l'àmbit de la quantitat de sensors que permeten la triangulació del registre (més de dos) i una qualitat de dades per sobre de 9, dels 289.935 registres, s'obté una mostra de 38.228 registres. S'aplica el model sobre aquests registres, amb l'objectiu d'obtenir la predicció de la seva geolocalització.

Tenint en compte la dificultat d'associar els registres (i/o lectures) del tipus MAC aleatòria, no s'ha identificat cap criteri que permeti l'agrupament dels registres en sessions. Aquest fet provoca que no es podrà obtenir els temps mitjans de duració de les sessions, o temps mitjà

de permanència en les diferents sales. Per aquest motiu, les diferents triangulacions que ML predigui, no podran agrupar-se en rutes, tot i que sí permeten identificar els llocs o sales més impactades. Generant els agregats per sala (*idroom*), de les 8 sales possibles, la que acumula més prediccions és l'hemicicle (amb 9.371 registres), seguit dels registres que provenen de l'exterior (*idroom* 0) del monument (amb 9.086 registres) i dels que provenen del passadís de recepció (amb 4.612 registres). Es mostra la taula amb la predicció de les sales visitades:

idroom_predict	***counter***
7	9.371
0	9.086
1	4.612
8	3.661
5	3.073
2	2,961
4	2.349
3	1.952
6	1.164

Taula 4.15: Taula de resum de les prediccions de visitants amb MAC aleatòria, per sala

La posició (*idplace*) amb més prediccions, correspon a la zona de les pintures murals de la capella (4.221 registres), seguit de dues posicions exteriors del museu (amb 3.732 i 2.875 registres), seguit de la zona de recepció (2.517 registres) i de l'armari de les Set Claus (1.855 registres). Es mostra la taula dels 10 llocs amb més geolocalitzacions de visitants, predites pel model:

idplace_predict	*counter*
14	4.221
45	3.732
38	2.875
22	2.517
34	1.855
31	1.448
8	1.309
12	1.164
35	1.132
4	1.015

Taula 4.16: Taula de resum de les prediccions de visitants amb MAC aleatòria, per lloc

Tot i no haver associat les lectures de tipus MAC aleatòria en sessions, la següent visualització pretén demostrar que, sense massa dificultat, es podrien correlacionar amb les geoposicions de les rutes provinents dels registres tipus MAC no aleatòria. En els moments d'alta densitat de geolocalitzacions per una sala concreta, de registres de tipus MAC aleatòria, es troben força coincidències de registres tipus MAC no aleatòria, en aquest cas, agrupats en rutes. Com a oportunitat de millora d'aquest projecte, es podria utilitzar la densitat de geoposicions dels registres tipus MAC aleatòria, per augmentar la qualitat de les geoposicions dubtoses, de registres tipus MAC no aleatòria.

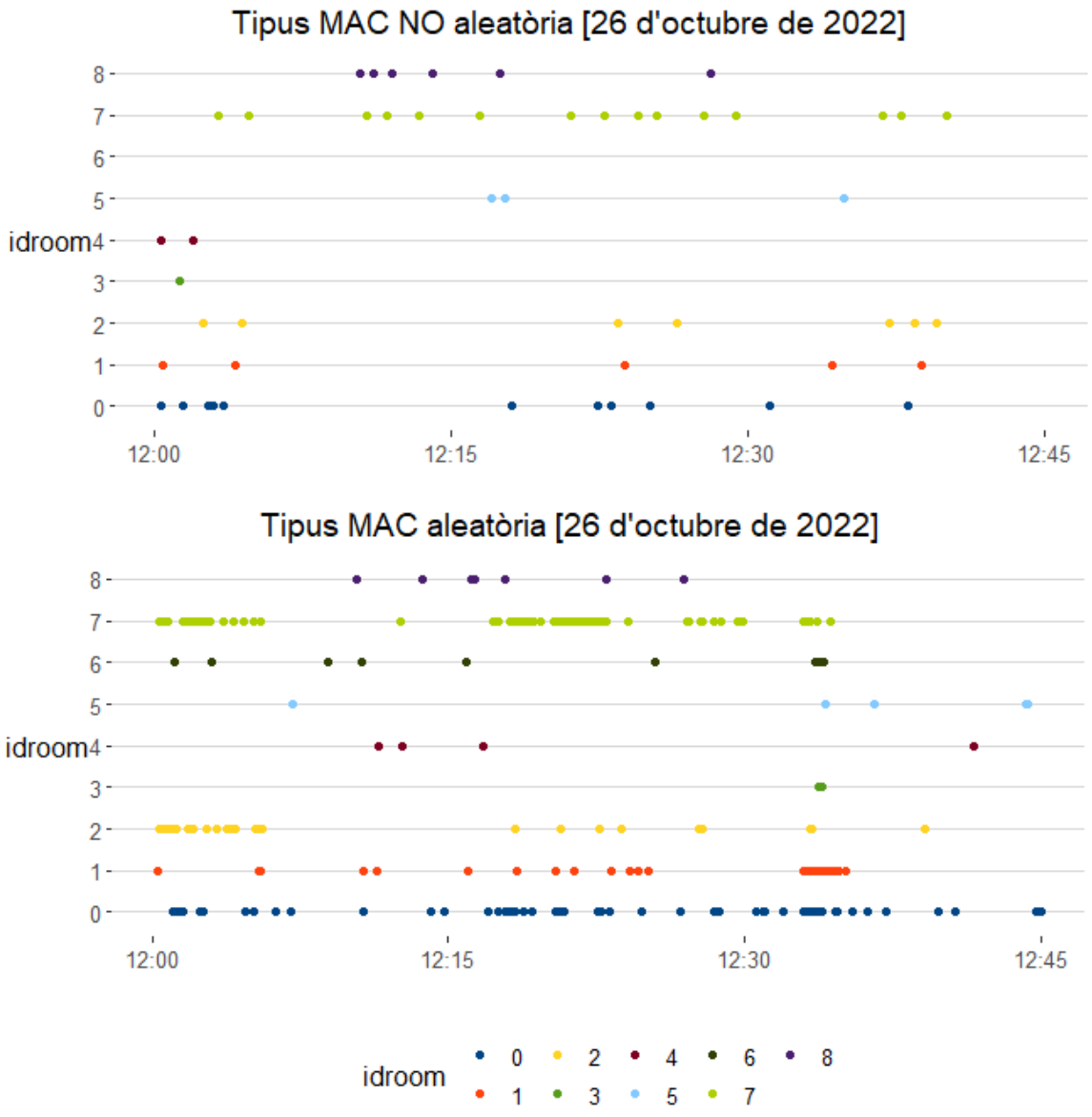

Figura 4.19: Visualització comparativa, per un dia i hora concrets, entre les geoposicions predites de registres tipus MAC no aleatòria agrupats en rutes, i els registres tipus MAC aleatòria no agrupats en rutes

4.8 Anàlisi de les rutes predites pel model ML

Durant les diferents fases de l'estudi, s'han recopilat en una matriu, les diferents rutes:

- Ruta monitoritzada: La realitzada amb l'audioguia, sense predicció. (Una sola ruta, formada per 16 geolocalitzacions, en sales consecutives diferents).
- Ruta predita: La realitzada amb l'audioguia, predita segons el model de ML. (Una sola

ruta, formada per 46 geolocalitzacions, en sales consecutives diferents).

- Rutes MAC estàtica o prefix MAC: Les rutes predites, segons els registres amb MAC estàtica o prefix MAC DA:A1:19. (85 rutes, amb diferent nombre de geolocalitzacions, en sales consecutives diferents).

- Rutes SSID: Les rutes predites, segons els registres amb SSID informada. (81 rutes, amb diferent nombre de geolocalitzacions, en sales consecutives diferents).

Cal recordar que dins de l'estudi, no s'ha desenvolupat la tasca de relacionar els registres de tipus MAC aleatòria amb SSID no informada. Una oportunitat de millora del projecte, seria utilitzar la resta d'atributs capturats, per relacionar registres de tipus MAC aleatòria i SSID no informada.

hash	idroom	dataquality	idroom_tdist	order	ssid	test
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	1	18	0	1		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	5	21	120	2		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	1	18	134	3		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	5	18	62	4		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	1	29	74	5		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	7	20	129	6		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	2	23	120	7		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	6	12	323	8		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	7	20	60	9		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	6	12	123	10		2
31e1de84736f6142c269ae4091b95a72737f712056d0140515...	2	12	75	11		2
33bf1a180dfaab50a27439205fc6e8844b0c0ac8c8ba85fc64d...	5	15	0	1		2
33bf1a180dfaab50a27439205fc6e8844b0c0ac8c8ba85fc64d...	6	12	20	2		2
33bf1a180dfaab50a27439205fc6e8844b0c0ac8c8ba85fc64d...	0	12	45	3		2
33bf1a180dfaab50a27439205fc6e8844b0c0ac8c8ba85fc64d...	6	12	80	4		2

Figura 4.20: Mostra de diverses rutes de tipus 2, descrites registre a registre

La següent tasca ha estat convertir cadascuna de les diferents rutes, en una cadena de posicions (tenint en compte l'identificador de cada sala) i emmagatzemar cadascuna de les cadenes resultants en un vector, on queda reflectit el detall de la navegació dels visitants, entre les diferents sales de cada ruta. El primer anàlisi de la mostra, té com a objectiu, verificar si existeix coincidència entre algunes de les 175 rutes identificades. Un cop confirmat que no existeix cap coincidència entre elles, la intenció ha estat comparar-les respecte a la ruta real no

predita, i obtenir un indicador de similitud. Per aquest motiu, es decideix avaluar el vector de rutes, mitjançant el mètode de la distància de *Levenshtein*. Aquest mètode retorna un índex que comptabilitza el nombre mínim d'operacions requerides per transformar una cadena de caràcters, en una altra. S'entén per operació, una inserció, eliminació o la substitució d'un caràcter. La comparació entre les rutes descrites dóna com a resultat:

Ruta	***Mostra***	***Min***	***Median***	***Mean***	***Max***	***Example (Min case)***
Ruta màster	1					0134565857543121010
Ruta predita	1	23	23	23	23	17454585168585875728..
Rutes MAC i prefix	87	8	13	13,47	21	01345758565210101
Rutes SSID	86	10	13	14,68	38	3452757870210

Taula 4.17: Taula de comparació entre la ruta monitoritzada (no predita) i les rutes predites, segons el tipus de registre

És cert que visualment, les rutes detallades a la taula 4.17, no són prou descriptives del nivell de similitud entre elles. L'indicador de similitud del mètode de la distància de *Levenshtein*, no sembla ser prou representatiu de la proximitat real entre les diferents rutes. Per aquest motiu, presentem mitjançant una visualització, la comparació entre la ruta monitoritzada i la ruta predita (tipus MAC estàtica o prefix MAC) més propera, segons l'indicador de *Levenshtein*, amb valor 8.

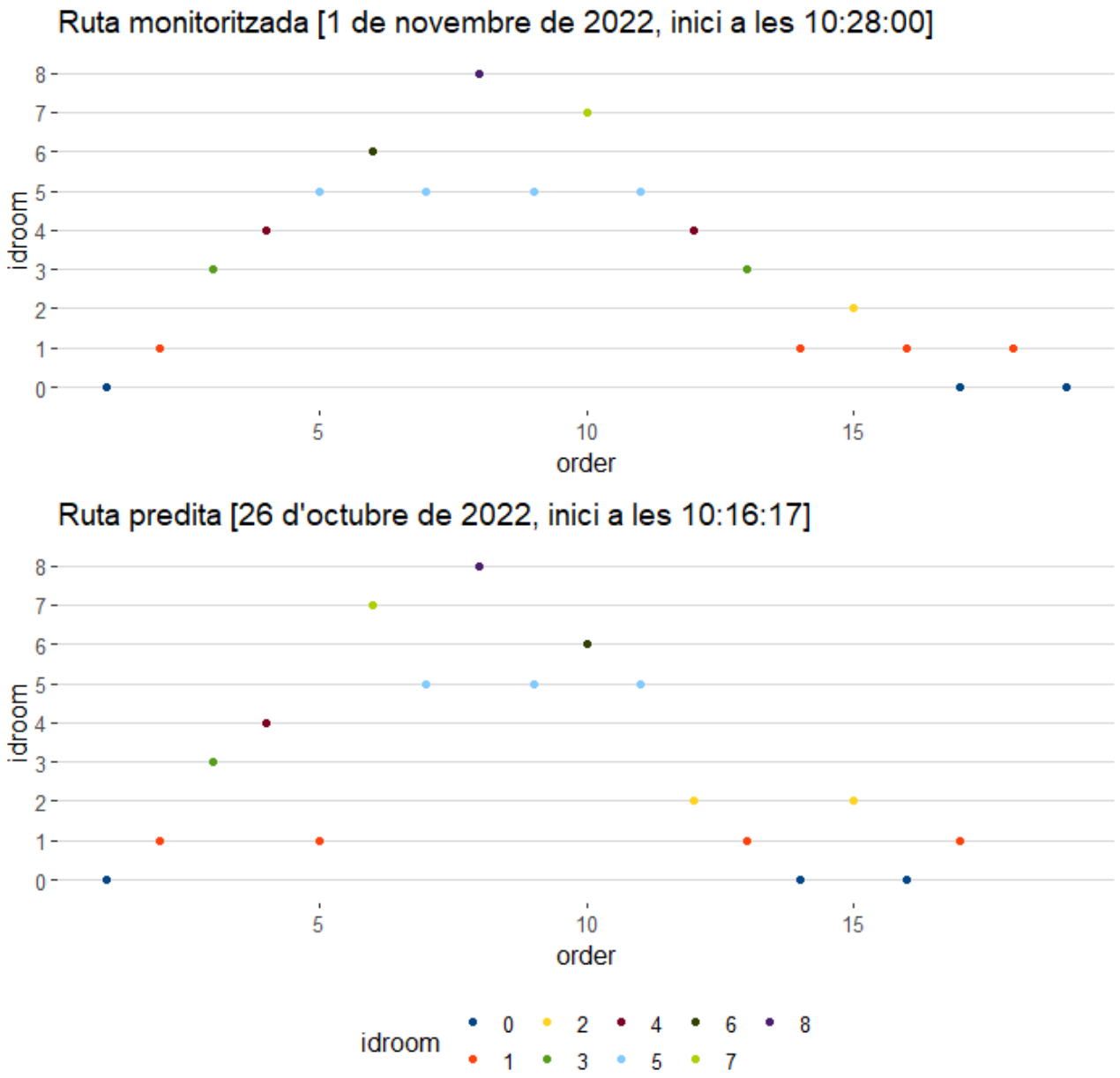

Figura 4.21: Visualització comparativa, entre la ruta monitoritzada (no predita) i la ruta predita més propera, segons el mètode de la distància de *Levenshtein*

Podem observar, les dues rutes són, durant les quatre primeres posicions, exactament iguals, amb l'únic error de la posició 5, que idealment s'hauria de situar en la sala 5 (Passos perduts), però que la predicció situa en la sala 1 (Passadís d'entrada i recepció). L'error queda justificat pels *fingerprints* similars en dues sales superposades entre plantes. La següent desviació situa la següent sala lògica, teòricament la sala 6 (Despatx del Síndic), com a sala 7 (Hemicicle), però curiosament detectem que coincideixen en visitar en posició 8, la sala 8 (Cuina), i tornen a

desviar-se, entre la sala 7 (Hemicicle) i la sala 6 (Despatx del Síndic). Està clar, que el visitant ha pogut canviar l'ordre lògic de visita entre aquestes dues sales, donat que el retorn a la sala 5 (Passos perduts) és totalment simètric entre ambdues rutes, i aquesta sala 5 és la zona d'accés entre les dues sales, per les quals es produeix la inversió de l'ordre. La següent desviació és no predir les sales 4 (Escala) i 3 (Vestíbul entrada oficial) i 1 (Passadís i recepció), però donat que són zones de pas, és probable que el temps del desplaçament no va ser suficient per a identificar al visitant. Torna a aparèixer coincidència, en posicions diferents (donat que la predicció ha obviat les tres anteriors), en les sales 2 (Tribunal de corts), sala 1 (Passadís i recepció) i la sala 0 (Exterior del monument). El final predit manté la lògica, donat que en haver de retornar l'audiòfon, el visitant ha de tornar a la sala 1 (Passadís i recepció), això sí, la sala 2 identificada en posició 15, sembla un *outlier*, molt probablement provocat pels desplaçaments exteriors, propers a la sala 2 (Tribunal de Corts).

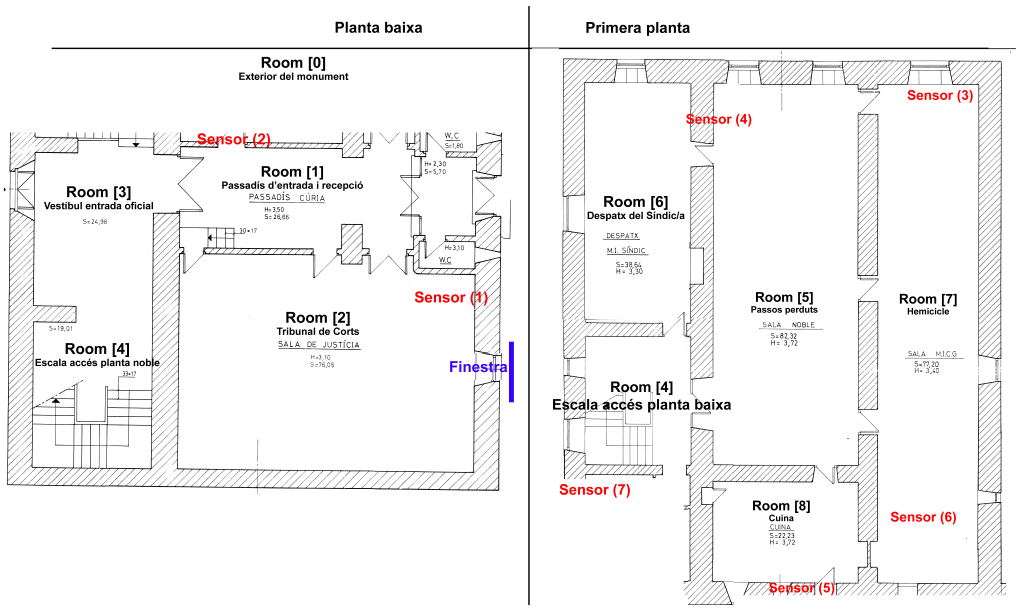

Figura 4.22: Plànol de la situació de les sales i tots els sensors, de la planta baixa i primera planta

Capítol 5

Conclusions

Tot i disposar de l'avantprojecte i d'un primer prototipus dels sensors ensamblats i programats, el projecte va començar amb dues setmanes de retard, fruit de la gestió de l'autorització requerida per part de la Sindicatura del Consell General d'Andorra, així com per la consulta a l'ADPA que va ser considerada prerequisit.

El desplegament dels sensors va presentar incidències relacionades amb un baix nivell de cobertura wifi, en les zones en les quals, es disposava de connexió elèctrica. Vista la dificultat de trobar emplaçaments alternatius, es va optar per fer la descàrrega manual (utilitzant un sistema de *hotspot* wifi) sensor a sensor, cada tres dies. El canvi en el procediment inicialment dissenyat, va suposar adaptar i carregar el *firmware* dels microcontroladors, un cop ja desplegats dins del monument.

La fase de calibració va presentar algunes dificultats, donat que amb l'objectiu de ser el més fidel possible a la realitat, es va voler prioritzar l'ús un dispositiu mòbil com a client (i no un microcontrolador ESP32 programat per aquesta finalitat). Cap dels tres dispositius mòbils que es varen emprar, disposava de capacitat de mantenir l'adreça MAC estàtica. Per aquest motiu, es va declarar dins del dispositiu mòbil, el nom d'una xarxa SSID específica. Un cop realitzat el procés físic de calibració, es va haver de seleccionar manualment aquells paquets de descobriment, amb l'SSID específica.

L'estudi posa de manifest que, encara existeix una important quantitat de dispositius que no adopten l'aleatorització MAC. Cal afegir que, una rellevant quantitat de dispositius mòbils, es poden rastrejar fàcilment, mitjançant altres atributs, com són el prefix MAC o les SSID específiques.

Els dispositius sense aleatorització MAC, prefix MAC i SSID específiques, se'n pot fer un seguiment senzill, mentre que els dispositius que usen l'aleatorització d'adreces MAC, són difícils de rastrejar. A mesura que passi el temps, es preveu una millora en la dificultat de fer el seguiment dels dispositius mòbils, donada la substitució dels més antics (amb pitjors

implementacions de l'aleatorització d'adreces MAC) per dispositius més nous, després d'arribar al final de la seva vida útil.

En l'àmbit de la geolocalització en interiors, l'acció de situar els sensors en els extrems de les diferents plantes del museu per assolir la millor triangulació dels dispositius mòbils, ha provocat l'efecte no desitjat d'haver d'assumir un alt soroll provocat pels mòbils dels no visitants del monument, donat que els entorns del museu, gaudeixen de molta activitat social i la circulació de ciutadans i turistes és molt elevada.

Un altre aspecte que ha provocat cert error en la geolocalització en interiors, ha estat el fet que la construcció dels solers està formada per bigues d'acer i fusta. Aquesta configuració de materials dissipa menys el senyal radioelèctric, del que podria representar el formigó armat. Aquesta casuística ha provocat certa confusió entre sales sobreposades entre plantes, i menys sovint, l'habitual confusió entre sales contigües d'una mateixa planta.

L'anàlisi del nivell d'encert del model construït pels diferents mètodes d'algorismes supervisats d'aprenentatge automàtic, determinen que, amb molta diferència, el millor mètode de predicció per aquesta tipologia de dades, on és estratègic calcular de la distància euclidiana entre un punt objectiu i els diferents veïns que formen grups, és el *K-Nearest Neighbors*.

S'ha desenvolupat un algorisme que realitza la conversió entre les lectures captades pels diferents sensors, i els registres que per cada unitat de temps, relacionen les lectures a un mateix dispositiu, mitjançant un camp clau (adreça MAC o SSID específica). S'han desenvolupat diferents tipus de neteja de dades, tenint en compte diferents paràmetres (qualitat del registre, nombre de sensors participants, distància en temps...). També s'ha desenvolupat un mètode d'ajustament de les lectures, tenint en compte el soroll provocat pels dispositius mòbils que tot i no ser visitants, són detectats pels sensors més exposats a la façana de l'edifici.

Els indicadors obtinguts, en l'àmbit dels temps mitjans de durada de les sessions, coincideixen (10% de desviació de la mitjana) entre els diferents tipus de registres analitzats, i són molt propers al temps de duració de l'àudio de l'audioguia. Pel que fa als indicadors d'ocupació, respecte a les estadístiques oficials, presenta una desviació mitjana del 30%. El càlcul d'aquesta variable està fortament influïda per la quantitat d'emissions de paquets de descobriment, dels dispositius mòbils, associada al mode actiu configurat (mode avió, estalvi d'energia, actiu...).

Cal dir que, en algunes situacions, els indicadors assolits mitjançant aquest estudi, podrien complementar i contextualitzar l'estadística oficial, tant pel que fa a l'inici i fi real de les visites de grups, com per identificar la càrrega real de cada zona o sala.

En l'àmbit de la predicció de rutes (unes 167), s'ha avaluat respecte a la ruta monitoritzada, però ha mancat disposar d'altres modes de validació, donat que els visitants, tot i seguir la configuració de punts d'interès determinat per l'audioguia, es poden moure lliurement, repetint sales o passant de forma més ràpida per altres.

En resum, en confirma que tot i partir d'un desplegament de sensors força limitat, s'han aconseguit indicadors de qualitat, que podrien complementar les estadístiques que de forma manual es mantenen als museus. A més, considerem que la captura i l'anàlisi de paquets de descobriment wifi, representa un anàlisi desatès i poc intervencionista, que permetria obtenir indicadors equivalents als desenvolupats en aquest projecte, de monuments dels quals no es disposa d'estadístiques oficials. És cert que, un procés més acurat de calibració i validació presencial, augmentaria de forma significativa la qualitat de les prediccions.

En l'àmbit de la mobilització de les competències de les diferents assignatures del màster, durant aquest treball final, ha estat fonamental "Tipologia i cicle de vida de les dades", que ha permès identificar tots els elements de planificació i de tipologia, útils per dissenyar cada etapa, amb els instruments corresponents. Cal dir que, després de sol·licitar dades *raw* a diferents institucions del país, i de rebre'n negatives (evidentment, pels dubtes legals que per a ells suposava cedir dades a tercers), vaig plantejar l'objectiu de proveir-me, des de l'origen, de les dades necessàries per desenvolupar aquest projecte. D'aquesta finalitat, n'han derivat tasques, inicialment no previstes, com són sol·licitar les autoritzacions necessàries per al desplegament sensòric, així com garantir que les dades generades, quedaven fora de la llei de protecció de dades. D'altra banda, i tenint en compte els entorns on es realitzava la captura, s'han dissenyat, ensamblat i programat els sensors necessaris per a la captura de les dades, així com els mecanismes de sincronització, emmagatzematge i transferència d'aquestes. És cert que, aquestes darreres tasques, no són competència directa d'aquest màster, però resulten essencials dins de la fase de captura. Un cop assolit un volum de dades interessant per a la seva anàlisi, l'assignatura "Models avançats de mineria de dades", m'ha aportat les eines i les competències necessàries per entendre la informació existent dins d'aquesta col·lecció. En una primera fase, per entendre i extreure'n coneixements, i més endavant, utilitzant tècniques i metodologies, per realitzar prediccions. Finalment, l'assignatura que suposa la clau de volta d'aquest projecte és "Visualització de dades". És rellevant obtenir aquelles visualitzacions que, han de permetre entendre els comportaments o conclusions d'un procés eminentment tècnic i molt ampli, als consumidors finals de la informació. El desenvolupament de bones visualitzacions i quadres de comandament, és del tot estratègic per l'explotació del coneixement que se'n desprèn del treball amb les dades.

Com a tasques pendents de desenvolupar, en un treball futur, i des del punt de vista que durant el treball s'ha prioritzat l'anàlisi de les dades, que permetessin desenvolupar hipòtesis objectives, amb la mínima dependència de valoracions subjectives, restaria pendent extreure informació de les dades amb resum *Hash* aleatori, donat que s'han de plantejar hipòtesis utilitzant atributs que suposen relacions dèbils, amb menors índex de confiança.

Bibliografia

[1] Cisco dna center for randomized mac addresses.

[2] How we tracked and analyzed over 200,000 people's footsteps at mit.

[3] Interpretació del reglament general de protecció de dades (gdpr).

[4] La casa de la vall, seu del consell general.

[5] Privacidad de wi-fi i aleatorización de la dirección mac.

[6] Tratamiento de la señal recibida por los teléfonos móviles para conocer el recorrido realizado por los clientes.

[7] Paramvir Bahl and Venkata N Padmanabhan. Radar: An in-building rf-based user location and tracking system.

[8] Tomas Bravenec, Joaquín Torres-Sospedra, Michael Gould, Tomas Fryza, and Centro Algoritmi. Exploration of user privacy in 802.11 probe requests with mac address randomization using temporal pattern analysis.

[9] Ante Dageli and Toni Perkovi. Location privacy and changes in wifi probe request based connection protocols usage through years.

[10] Saandeep Depatla, Arjun Muralidharan, and Yasamin Mostofi. Occupancy estimation using only wifi power measurements.

[11] Audit Association for Computing Machinery. Special Interest Group on Security, ACM SIGMOBILE, National Science Foundation (U.S.), United States. Army Research Office, and Association for Computing Machinery. *Proceedings of the 8th ACM Conference on Security and Privacy in Wireless and Mobile Networks.*

[12] Yuhan Gao and Jan Dirk Schmöcker. Estimation of walking patterns in a touristic area with wi-fi packet sensors. *Transportation Research Part C: Emerging Technologies*, 128, 10 2021.

[13] M W P Maduranga and Ruvan Abeysekara. Supervised machine learning for rssi based indoor localization in iot applications. *International Journal of Computer Applications*, 183:26–32, 10 2021.

[14] Vikrom Maikaensarn. The review of modes of operation and randomization in smartphone passive monitoring using wi-fi probe context.

[15] Luiz Oliveira, Daniel Schneider, Jano De Souza, and Weiming Shen. Mobile device detection through wifi probe request analysis. *IEEE Access*, 7:98579–98588, 2019.

[16] Jordi Gironés Roig Jordi Casas Roma Julià Minguillón Alfonso Ramon Caihuelas Quiles. MinerÍa de datos modelos y algoritmos.

[17] Alessandro E.C. Redondi and Matteo Cesana. Building up knowledge through passive wifi probes. *Computer Communications*, 117:1–12, 2 2018.

[18] Guenther Retscher, T U Wien, Dirk Heberling, Eva Moser, and Dennis Vredeveld. Performance and accuracy test of the wlan indoor positioning systemïpos lbs2its-curricula enrichment delivered through the application of location-based services to intelligent transport view project performance and accuracy test of the wlan indoor positioning system ïpos".

[19] Alfatta Rezqa, Winnersyah Feri, and Nenny Anggraini. Identification and position estimation method with k-nearest neighbour and home occupants activity pattern.

[20] IEEE Communications Society, Institute of Electrical, and Electronics Engineers. *2020 International Conference on Communication Systems Networks (COMSNETS).*

[21] Xiaoyong Tang, Bin Xiao, and Kenli Li. Indoor crowd density estimation through mobile smartphone wi-fi probes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50:2638–2649, 7 2020.

[22] Hiroaki Togashi, Hiroshi Furukawa, Yuki Yamaguchi, Ryuta Abe, and Junpei Shimamura. Network-based positioning and pedestrian flow measurement system utilizing densely placed wireless access points. Institute of Electrical and Electronics Engineers Inc., 11 2016.

[23] Vijay Vaishnavi and Bill Kuechler. Design science research in information systems.

[24] Edwin Vattapparamban, Bekir Sait Çiftler, Ismail Güvenç, Kemal Akkaya, and Abdullah Kadri. Indoor occupancy tracking in smart buildings using passive sniffing of probe requests. pages 38–44. Institute of Electrical and Electronics Engineers Inc., 7 2016.

[25] Lei Yang, Hao Chen, Qimei Cui, Xuan Fu, and Yifan Zhang. Probabilistic-knn: A novel algorithm for passive indoor-localization scenario. volume 2015. Institute of Electrical and Electronics Engineers Inc., 7 2015.

[26] Yuji Yoshimura, Stanislav Sobolevsky, Carlo Ratti, Fabien Girardin, Juan Pablo Carrascal, and Josep Blat. An analysis of visitors' behavior in the louvre museum: A study using bluetooth data.