Universidad de Málaga

Programa de Doctorado en Ingeniería de Telecomunicación

Escuela Técnica Superior de Ingeniería de Telecomunicación

TESIS DOCTORAL

Modelos de aprendizaje automático basados en información de contexto espaciotemporal para redes celulares autoorganizadas

Autor:

Juan Luis Bejarano Luque

Directores:

Matías Toril Genovés

Mariano Fernández Navarro

Málaga, 2023

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

D./Dña JUAN LUIS BEJARANO LUQUE
Estudiante del programa de doctorado INGENIERÍA DE TELECOMUNICACIÓN de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: MODELOS DE APRENDIZAJE AUTOMÁTICO BASADOS EN INFORMACIÓN DE CONTEXTO ESPACIOTEMPORAL PARA REDES CELULARES AUTOORGANIZADAS.

Realizada bajo la tutorización de MARIANO FERNÁNDEZ NAVARRO y dirección de MATÍAS TORIL Y MARIANO FERNÁNDEZ NAVARRO (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

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

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

En Málaga, a 26 de JUNIO de 2023

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Firmado por BEJARANO LUQUE JUAN LUIS - 78986978X el día 27/06/2023 con un certificado emitido por AC Fdo.: JUAN LUIS BEJARANO LUQUE Doctorando/a	FERNANDEZ NAVARRO MARIANO - 25079628Z Firmado digitalmente por FERNANDEZ NAVARRO MARIANO - 25079628Z Fecha: 2023.06.27 14:46:03 +02'00' Fdo.: MARIANO FERNÁNDEZ NAVARRO Tutor/a
TORIL GENOVES MATIAS - 33393202P Firmado digitalmente por TORIL GENOVES MATIAS - 33393202P Fecha: 2023.06.27 13:13:11 +02'00' FERNANDEZ NAVARRO MARIANO - 25079628Z Firmado digitalmente por FERNANDEZ NAVARRO MARIANO - 25079628Z Fecha: 2023.06.27 14:46:59 +02'00' Fdo.: MATÍAS TORIL Y MARIANO FERNÁNDEZ NAVARRO Director/es de tesis	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

UNIVERSIDAD DE MÁLAGA

Ingeniería de Comunicaciones

AUTORIZACIÓN PARA LECTURA DE LA TESIS DOCTORAL

Por la presente, Dr. D. Matías Toril Genovés y Dr. D. Mariano Fernández Navarro, profesores doctores del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga, certifican que el doctorando Juan Luis Bejarano Luque ha realizado en el Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga, bajo su dirección, el trabajo de investigación correspondiente a su TESIS DOCTORAL titulada

Modelos de aprendizaje automático basados en información de contexto espaciotemporal para redes celulares autoorganizadas

Dicho trabajo ha dado lugar a las siguientes publicaciones en revistas y aportaciones a congresos que no han sido utilizadas en tesis anteriores:

1. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, R. Acedo-Hernández, S. Luna-Ramírez, "A Data-Driven Algorithm for Indoor/Outdoor Detection Based on Connection Traces in a LTE Network," IEEE Access, vol. 7, pp. 65877–65888, mayo 2019.
2. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, A. J. García, S. Luna-Ramírez, "A Context-Aware Data-Driven Algorithm for Small Cell Site Selection in Cellular Networks," IEEE Access, vol. 8, pp. 105335–105350, junio 2020.
3. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, "A Deep-Learning Model for Estimating the Impact of Social Events on Traffic Demand on a Cell Basis," IEEE Access, vol. 9, pp. 2169-3536, mayo 2021.
4. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, L.R. Jiménez, S. Luna-Ramírez, "Statistical model for mobile user positioning based on social information," MDPI Electronics, vol. 10, junio 2021.
5. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, R. Acedo-Hernández, S. Luna-Ramírez, "Algoritmo de detección interior/exterior para trazas de conexión de una red LTE," XXXIII Simposium Nacional de la Unión Científica Internacional de Radio URSI 2018, Granada, España, septiembre 2018.
6. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, A.J. García, S. Luna-Ramírez, "Algoritmo de detección de huecos de cobertura usando información de redes sociales," XXXIV Simposium Nacional de la Unión Científica Internacional de Radio, Sevilla, España, septiembre 2019.
7. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, "Análisis de la influencia de eventos sociales en redes celulares," XXXV Simposium Nacional de la Unión Científica Internacional de Radio, Málaga (evento en formato remoto por COVID-19), septiembre 2020.
8. J. L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, L.R. Jiménez, S. Luna-Ramírez, "Statistical model for mobile user positioning based on social information," 3rd Post-Iracon Meeting, Online Meeting, 31 mayo - 1 junio, 2021.

9. J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, “Comparativa de modelos de aprendizaje profundo para series temporales de alta resolución,” XXXVI Simposium Nacional de la Unión Científica Internacional de Radio, Vigo (evento en formato remoto por COVID-19), septiembre 2021.

Por todo ello, consideran que esta Tesis es apta para su presentación al Tribunal que ha de juzgarla. Y para que conste a efectos de lo establecido, AUTORIZAN la presentación de esta Tesis en la Universidad de Málaga.

En Málaga, a 26 de junio de 2023.

TORIL GENOVES MATIAS - 33393202P

Firmado digitalmente por TORIL GENOVES MATIAS - 33393202P
Fecha: 2023.06.27 13:13:30 +02'00'

Fdo.: Dr. D. Matías Toril Genovés

FERNANDEZ NAVARRO MARIANO - 25079628Z

Firmado digitalmente por FERNANDEZ NAVARRO MARIANO - 25079628Z
Fecha: 2023.06.27 14:48:04 +02'00'

Fdo.: Dr. D. Mariano Fernández Navarro

Una vez descartado lo imposible, lo que queda,
por improbable que parezca, debe ser la verdad

Arthur Conan Doyle en "La aventura de la diadema de berilo".

Agradecimientos

El desarrollo de una tesis doctoral es un viaje largo y, en ocasiones, muy difícil, pero a su vez muy gratificante cuando lo finalizas. En un proyecto de estas dimensiones siempre vives experiencias en el camino que te hacen crecer tanto profesional como personalmente. Gran parte de estas vivencias y enseñanzas han sido gracias a las personas que me han acompañado en el viaje y que sin ellas no hubiera sido tan divertido.

No podía comenzar este apartado de otra forma que agradeciendo a los dos principales pilares en mi vida. Primero, mi familia, Chari, Juan Jesús y Víctor. Aunque sus nombres no aparezcan en la portada, esta tesis es también suya, ya que sin su constante apoyo y total sacrificio sin reservas no habría podido llegar hasta donde estoy ahora. Igualmente, a Noelia, la mejor compañera que he podido encontrar en este viaje, que, aunque se ha incorporado a mitad del camino, lo ha recorrido conmigo sin dudar, a pesar de las largas tardes en el laboratorio o las noches trabajando. Muchas gracias por vuestra comprensión y ánimos.

De la misma forma, mi más sincero agradecimiento va hacia mis directores, Matías y Mariano, porque desde el primer momento me han ofrecido su ayuda y consejo en todo lo que he necesitado. Sin su invaluable guía y su desprendida dedicación no sería el profesional en el que me he convertido. Ha sido un verdadero honor y un completo placer aprender y crecer bajo vuestro manto.

En este tipo de aventura, es tan importante el destino como las personas que te acompañan a lo largo del camino. En este sentido he tenido mucha suerte, ya que no hubiera deseado mejores compañeros que los que he conocido. En especial quiero dar las gracias a los amigos que he hecho en la que ha sido mi casa estos años, el laboratorio 2.3.3.a. A Luis, el patrón del laboratorio, que me acogió desde que entré por la puerta y me guio en mis primeros pasos en esta etapa. A Carol, hace 5 años que comenzamos este viaje a la vez y hemos crecido juntos. A Joaquín, ese nuevo que no hablaba pero que llegó para alegrarnos los días tanto dentro como fuera del laboratorio. A Nuria,

que a pesar de sus propios problemas siempre está ahí para echar una mano en lo que pueda. Por último, a Luisi y Pablo, que, aunque terminaron su propia etapa antes que yo, me han acompañado parte del camino. También quiero agradecer al resto de mis compañeros del doctorado y, en concreto, al 1.3.3 (Mario, Alberto, Pablo, Fran, Aarón...) y al 1.3.1 (Hao, Jesús, Jéssica, Carlos...) por los buenos ratos en los almuerzos, congresos y, en general, hacer más amena la vida en la Escuela.

Por último, me gustaría agradecer al Instituto Universitario de Investigación en Telecomunicación (TELMA) y la Universidad de Málaga, así como las entidades responsables que han permitido apoyar financieramente el desarrollo de esta tesis, sin lo cual no hubiera sido posible: el Ministerio de Economía y Competitividad (TEC2015-69982-R), el Ministerio de Ciencia, Innovación y Universidades (RTI2018099148-B-I00) y la Junta de Andalucía (UMA18-FEDERJA-256).

Resumen

Los constantes avances tecnológicos de los últimos años han revolucionado los servicios prestados por las redes móviles, que hoy en día demandan mayores volúmenes de información y menores tiempos de respuesta para su correcto funcionamiento. De la misma forma, estas mejoras técnicas han elevado las expectativas que tienen los usuarios de la red, esperando unas velocidades de conexión y una estabilidad muy por encima de las prestadas en las redes móviles actuales.

La experiencia que disfruta el usuario típico de las redes de telefonía está altamente influenciada por el entorno, condicionando el uso que hace de los diferentes servicios. Por ello, incluir el contexto de los usuarios en las tareas de gestión de la red es una de las principales tendencias en el desarrollo de nuevas técnicas de autoorganización, como se deduce de las especificaciones de la nueva generación de redes móviles, las redes 5G.

La posición del usuario y el instante de tiempo en el que hace uso de la red destacan especialmente entre las características del contexto de los usuarios que más impacto tienen en el comportamiento de la red. Por ejemplo, identificar si en ciertas zonas de la red los servicios se demandan desde localizaciones de interior o detectar grandes aglomeraciones de usuarios navegando simultáneamente, ya sean persistentes o esporádicas, permite mejorar las soluciones aplicadas a problemas de rendimiento. Por desgracia, esta información de contexto no suele ser accesible para los operadores de red, lo que dificulta su aplicación en las tareas de gestión de red.

El objetivo de esta tesis doctoral es desarrollar modelos construídos con diferentes técnicas de aprendizaje automático, que permitan estimar aspectos del contexto espaciotemporal de los usuarios para aplicarlos a tareas de autogestión de la red. El enfoque práctico seguido a lo largo de esta tesis ha motivado que, como restricciones de partida, los datos de entrada empleados sean de fácil acceso y los modelos desarrollados sean eficientes computacionalmente, para que puedan ser integrados fácilmente en las herramientas de gestión radio.

El trabajo desarrollado se inicia con el estudio de la dimensión espacial del contexto. Así, la primera de las contribuciones de esta tesis es un modelo que estima uno de los aspectos más importantes del contexto espacial del usuario, que es si este ha accedido a la red desde una localización de interior o de exterior. Dicha estimación se realiza únicamente a partir de indicadores de rendimiento de la conexión registrados en el sistema de gestión de red.

Otro de los aspectos fundamentales del contexto espacial de los usuarios es la posición geográfica de los mismos, es decir, dónde se sitúan los terminales en el escenario. Para ello, la segunda contribución aprovecha la ubicuidad actual de las redes sociales para mejorar el posicionamiento de los terminales, combinando la información obtenida de redes sociales, enriquecida con metadatos georreferenciados, con la ubicación interior o exterior de los usuarios deducida con el modelo anterior.

Para demostrar la aplicabilidad de los modelos desarrollados, la tercera contribución de esta tesis propone un método de selección de emplazamientos para picoceldas en redes celulares heterogéneas. A diferencia de otros métodos en la bibliografía, el algoritmo propuesto basa la localización de las nuevas celdas pequeñas en la mejora que se obtendría, desde el punto de vista del usuario final, en términos de volumen de descarga generado en cada zona de la red, basándose en el contexto espacial de los usuarios.

Por último, se extiende el estudio con la dimensión temporal del contexto de los usuarios. Con este fin, la última contribución de esta tesis doctoral aborda uno de los problemas clásicos en la gestión de redes móviles, como es la predicción del tráfico futuro de una celda del sistema. La principal dificultad de este problema radica en la gran variabilidad del comportamiento de los usuarios, que impide una previsión precisa. Una de las causas principales de esta variabilidad son los eventos sociales, en los que una gran cantidad de usuarios se concentran en un mismo establecimiento durante un periodo de tiempo limitado. En el modelo propuesto, se predice la demanda futura de tráfico de datos de una celda con resolución horaria del día siguiente, aplicándole una corrección dependiente del tipo de evento que va a tener lugar. Para ello, se recolectan, de varios repositorios, datos de diferentes tipos de eventos sociales ocurridos en el pasado para entrenar un modelo de aprendizaje profundo.

Índice general

Agradecimientos **III**

Resumen **V**

Índice de tablas **XI**

Índice de figuras **XIII**

Lista de Acrónimos **XVII**

1. Introducción **1**

1.1. Motivación 1

1.2. Objetivos y Contribuciones 5

1.3. Metodología de trabajo 7

1.4. Estructura del documento 8

2. Marco teórico **11**

2.1. Redes de acceso móvil 11

2.1.1. Información registrada de la red 13

2.2. Redes AutoOrganizadas 15

2.2.1. Autoplanificación 15

2.2.2. Autodespliegue 16

2.2.3. Autooptimización 17

2.2.4. Autoreparación 18

2.3. Posicionamiento de usuarios en redes móviles 19
2.3.1. Medidas empleadas en la localización 19
2.3.2. Clasificación de los métodos de posicionamiento en redes celulares 21
2.3.3. Métodos de posicionamiento 21
2.4. Técnicas de aprendizaje automático 26
2.4.1. Técnicas clásicas 29
2.4.2. Redes neuronales 34

3. Clasificación probabilística de conexiones de interior 43
3.1. Estado de la técnica 44
3.2. Formulación del problema 45
3.2.1. Datos del sistema de gestión de red 47
3.2.2. Datos de uso de suelo 49
3.3. Modelo de clasificación de conexiones en interior 50
3.3.1. Etiquetado artificial 50
3.3.2. Modelo de regresión 54
3.3.3. Selección de variables 56
3.4. Evaluación del modelo 56
3.4.1. Metodología experimental 57
3.4.2. Resultados 60
3.4.3. Complejidad computacional 66
3.5. Conclusiones 67

4. Posicionamiento estadístico de conexiones basado en información social 69
4.1. Estado de la técnica 70
4.2. Formulación del problema 71
4.3. Modelo de posicionamiento estadístico 73
4.3.1. Preprocesado 75

4.3.2. Distribución espacial de conexiones 78
4.4. Evaluación del modelo 80
4.4.1. Metodología experimental 80
4.4.2. Resultados 83
4.4.3. Complejidad computacional 88
4.5. Conclusiones 89

5. Selección automática de emplazamientos de picoceldas con datos multimodales 91

5.1. Estado de la técnica 92
5.2. Formulación del problema 93
5.2.1. Impacto de la mejora de tasa binaria con búfer completo 94
5.2.2. Impacto de la mejora de tasa binaria con búfer completo en conexiones de tráfico a ráfagas 95
5.2.3. Efecto del contexto de interior en el impacto de la mejora de tasa binaria con búfer completo 97
5.3. Modelo automático de selección de nuevos emplazamientos 99
5.3.1. Paso 1: Construcción del mapa de ganancia de tráfico en la red . 100
5.3.2. Paso 2: Selección de emplazamientos 105
5.4. Evaluación del modelo 107
5.4.1. Metodología experimental 108
5.4.2. Resultados 109
5.4.3. Complejidad computacional 119
5.5. Conclusiones 120

6. Predicción de la demanda de tráfico en eventos sociales 123

6.1. Estado de la técnica 124
6.2. Análisis preliminar del impacto de eventos sociales en el tráfico celular . 127
6.3. Modelo de predicción de tráfico en eventos sociales 132

6.3.1. Paso 1: Recolección de datos y preprocesado 133
6.3.2. Paso 2: Modelado de serie temporal 134
6.3.3. Paso 3: Corrección del impacto de eventos 135
6.4. Validación del modelo 139
6.4.1. Metodología experimental 139
6.4.2. Resultados 143
6.4.3. Complejidad computacional 150
6.5. Conclusiones 151

7. Conclusiones finales **155**
7.1. Resumen del trabajo 155
7.2. Publicaciones 158
7.3. Líneas futuras 161

Apéndice A. Herramientas de recolección y tratamiento de datos **165**
A.1. Herramientas de recolección 165
A.1.1. Recolección a través de APIs 166
A.1.2. Recolección mediante técnicas de raspado web 171
A.2. Herramientas de procesado 173
A.2.1. Construcción de conexiones 173
A.2.2. Construcción de mapas de uso de suelo 175

Bibliografia **177**

Índice de tablas

3.1. Descriptores de conexión usados en el modelo completo. 47

3.2. Descripción de los tipos de uso de suelo. 50

3.3. Densidad de conexiones por tipo de uso de suelo. 58

3.4. Estadísticas de los descriptores de conexión estudiados a nivel de celda. 59

3.5. Análisis de regresión para el modelo simplificado. 65

4.1. Tipos de uso de suelo en el escenario. 81

4.2. Estadísticas del número de conexiones por cuadrícula con los distintos modelos distribución espacial de conexiones comparados. 84

5.1. Tipos de uso de suelo en el escenario. 108

5.2. Parámetros de regresión por contexto. 109

5.3. Estadísticas de ganancia de volumen de datos por conexión y anillo existentes (en bytes). 110

5.4. Estimación del incremento potencial de volumen de datos (en bytes). . 113

5.5. Ejemplos de nuevos emplazamientos en los que los modelos se comportan de forma diferente. 116

6.1. Categorías de eventos 129

6.2. Selección de hiperparámetros para los modelos clásicos de series temporales. 140

6.3. Selección de hiperparámetros para los modelos de aprendizaje profundo. 142

6.4. Selección de hiperparámetros para los modelos en la etapa de corrección por evento. 144

6.5. Error medio absoluto de los diferentes modelos por sentido del enlace [kbps]. 145

6.6. Tiempo total de entrenamiento del modelo completo [s]. 145

6.7. Error Medio Absoluto global tras la etapa de corrección por evento [kbps]. 148

6.8. Tiempos de entrenamiento en DL [s]. 151

Índice de figuras

2.1. Arquitectura conceptual de una red de acceso. 13

2.2. Perceptrón. 35

2.3. Ejemplo de perceptrón multicapa. 36

2.4. Neurona convolucional. 37

2.5. Ejemplo de red neuronal convolucional. 37

2.6. Ejemplo de red autocodificadora. 38

2.7. Neurona LSTM. 39

2.8. Ejemplo de LSTM apilada. 41

2.9. Ejemplo de estructura para aprendizaje multitarea. 42

3.1. Ejemplo de cálculo de la superficie de uso del suelo por anillo. 52

3.2. Diagrama de flujo del modelo de clasificación. 55

3.3. Ratios de uso de suelo en el escenario. 58

3.4. Ratios de conexiones en interior computados con los usos de suelo vs. obtenidos de las trazas (modelo completo, todas las celdas). 61

3.5. Ratio de conexiones de interior por celda computados con los usos de suelo frente a los obtenidos de las trazas (modelo completo, segregado en 3 grupos). 62

3.6. Percentiles de la probabilidad de conexiones en interior obtenidos de las trazas vs. ratios de conexiones en interior computados con los usos de suelo (modelo completo, segregado en 3 grupos). 62

3.7. Importancia de cada variable en cada grupo de celdas (modelo completo). 63

3.8. Impacto del número de variables por grupo de celdas. 64

3.9. Ratios de conexiones en interior por celda computados con los usos de suelo vs. obtenidos de las trazas (modelo simplificado). 64

3.10. Percentiles de la probabilidad de conexiones en interior obtenidos de las trazas vs. ratios de conexiones en interior computados con los usos de suelo (modelo simplificado). 66

3.11. Resultados obtenidos con el juego de validación. 67

4.1. Diagrama de flujo del modelo de distribución espacial de conexiones. . . 73

4.2. Ejemplo de variograma esférico. 76

4.3. Ejemplo de kriging para el cálculo del mapa de publicaciones sociales subyacentes. 77

4.4. Ejemplo de obtención de datos de contexto y de publicaciones de un anillo. 79

4.5. Función de distribución acumulada de la probabilidad de conexión en interior en el escenario. 82

4.6. Función de distribución acumulada del número de publicaciones geolocalizadas por cuadrícula en el escenario de distintas redes sociales. . . . 82

4.7. Secciones del mapa de distribución espacial de publicaciones sociales en el escenario. 83

4.8. Número de conexiones por cuadrícula para cada modelo de distribución espacial de conexiones. 84

4.9. Comparación de modelos de distribución de conexiones en el escenario. 86

5.1. Gráfico de dispersión del volumen de datos frente al caudal del enlace de radio por conexión. 96

5.2. Gráfico de dispersión del ratio de conexiones de búfer completo frente al caudal por anillo de avance temporal (TA). 96

5.3. Gráfico de dispersión del volumen de datos frente al caudal del enlace de radio por conexión, separado por contexto. 97

5.4. Gráfico de dispersión del ratio de conexiones de búfer completo frente al caudal por anillo de TA, separado por contexto. 98

5.5. Flujograma del modelo propuesto. 100
5.6. Flujograma del modelo propuesto. Paso 1: Construcción del mapa de ganancia de tráfico en la red. 101
5.7. Método de estimación de valores óptimos interpolando curvas de regresión. 102
5.8. Flujograma del modelo propuesto. Paso 2: Selección de emplazamientos. 105
5.9. Pseudocódigo del algoritmo para la selección de emplazamientos. 106
5.10. Distribución acumulada de probabilidad de la ganancia de volumen estimada por conexión y por anillo. 110
5.11. Distribución estadística de los indicadores por cuadrícula del mapa. . . 111
5.12. Distribución espacial de la mejora potencial de volumen de tráfico, $M_{\Delta V}(x, y)$. 112
5.13. Incremento de volumen para las mejores 100 nuevas ubicaciones. 114
5.14. Nivel de señal recibida (OpenSignal) en ejemplo de selección de emplazamiento. 114
5.15. Histograma del ratio de ganancia en interior para los mejores 100 emplazamientos candidatos. 115
5.16. Distribución de la ganancia de volumen de tráfico de los 100 mejores nuevos emplazamientos candidatos para los métodos basados en red social (SM), trazas (TM) y trazas+red social (TSM). 115
5.17. Propuesta de emplazamiento para solucionar problemas de cobertura. . 117
5.18. Análisis detallado de TM y TSM (caso 2). 117
5.19. Propuesta de emplazamiento para solucionar problemas de capacidad. . 118
5.20. Análisis detallado de TM y TSM (caso 4). 118

6.1. Distribución de la intensidad de tráfico adicional diaria por tipo de evento. 130
6.2. Diagrama de bloques del modelo de previsión de tráfico basado en aprendizaje profundo. 132
6.3. Modelo multitarea Long Short-Term Memory Convolucional para la predicción del tráfico normal (sin evento). 136
6.4. Modelos de aprendizaje profundo para la estimación del impacto de los eventos. 138

6.5. Distribución de los tipos de evento en días laborables y fines de semana. 141

6.6. Proceso de convergencia durante el entrenamiento de los modelos de aprendizaje profundo. 147

6.7. Precisión de los modelos por enlace y tipo de evento. 149

6.8. Patrón horario del tráfico predicho en torno al evento. 150

A.1. Esquema de la herramienta de colección de datos de redes sociales en tiempo real. 168

A.2. Esquema de la herramienta de colección de datos históricos de eventos. 169

A.3. Esquema de la herramienta de colección mediante raspado web. 172

A.4. Esquema de la herramienta construcción de conexiones. 173

A.5. Esquema de la herramienta de construcción de mapas de uso de suelo. . 175

Acrónimos

3GPP	**3**rd **G**eneration **P**artnership **P**roject
5GPPP	**5G** **P**ublic **P**rivate **P**artnership
A-GNSS	**A**ssisted **G**lobal **N**avigation **S**atellite **S**ystem
AE	**A**uto**E**ncoder
ANN	**A**rtificial **N**eural **N**etwork
AOA	**A**ngle **O**f **A**rrival
APP	**A**ntenna **P**lacement **P**roblem
ARIMA	**A**uto-**R**egresive **I**ntegrated **M**oving **A**verage
CAPEX	**CAP**ital **EX**penditure
CDF	**C**umulative **D**istribution **F**unction
CEM	**C**ustomer **E**xperience **M**anagement
CEP	**C**omplex **E**vent **P**rocessing
CID	**C**ell **ID**entifier
CM	**C**onfiguration **M**anagement
CNN	**C**onvolutional **N**eural **N**etwork
CQI	**C**hannel **Q**uality **I**ndicator
CSV	**C**omma-**S**eparated **V**alues
CTR	**C**ell **T**raffic **R**ecording
DBN	**D**eep **B**elief **N**etwork
DeL	**De**ep **L**earning
DL	**D**own**l**ink
DTF	**D**ata **T**race **F**ile
E-CID	**E**nhanced **C**ell **ID**entifier
eMBB	**e**nhanced **M**obile **B**road**B**and
GARCH	**G**eneralized **A**uto-**R**egressive **C**onditional **H**eteroskedasticity
GNN	**G**raph **N**eural **N**etwork
GPS	**G**lobal **P**ositioning **S**ystem

IoT	**I**nternet **o**f **T**hings
KPI	**K**ey **P**erformance **I**ndicator
LSTM	**L**ong **S**hort-**T**erm **M**emory
LTE	**L**ong **T**erm **E**volution
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MDT	**M**inimization **D**rive **T**est
MIMO	**M**ultiple **I**nput **M**ultiple **O**utput
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
mMTC	**m**assive **M**achine **T**ype **C**ommunications
MSE	**M**ean **S**quare **E**rror
MT	**M**ulti**T**ask
MTL	**M**ulti**T**ask **L**earning
NLOS	**N**on-**L**ine-**O**f-**S**ight
NR	**N**ew **R**adio
O&M	**O**peration & **M**aintenance
OSS	**O**perations **S**upport **S**ystem
OTDOA	**O**bserved **T**ime **D**ifference **O**f **A**rrival
PCH	**P**aging **CH**annel
PDCP	**P**acket **D**ata **C**onvergence **P**rotocol
PM	**P**erformance **M**anagement
PRS	**P**ositioning **R**eference **S**ymbols
QoE	**Q**uality **o**f **E**xperience
OPEX	**OP**erational **EX**penditure
QoS	**Q**uality **o**f **S**ervice
RAC	**R**adio **A**dmission **C**ontrol
RACH	**R**andom **A**ccess **CH**annel
RAT	**R**adio **A**ccess **T**echnology
ReLU	**Re**gularized **L**inear **U**nit
REM	**R**adio **E**nvironment **M**ap
RL	**R**einforcement **L**earning
RNN	**R**ecurrent **N**eural **N**etwork
ROP	**R**eporting **O**utput **P**eriod
RRC	**R**adio **R**esource **C**ontrol
RSRP	**R**eference **S**ignal **R**eceived **P**ower

RSS	**R**eceived **S**ignal **S**trength
RSTD	**R**eference **S**ignal **T**ime **D**ifference
RTT	**R**ound-**T**rip **T**ime
SA	**S**imulated **A**nnealing
SARIMA	**S**easonal **A**uto-**R**egresive **I**ntegrated **M**oving **A**verage
SC	**S**mall **C**ell
SINR	**S**ignal to **I**nterference plus **N**oise **R**atio
SL	**S**upervised **L**earning
SM	**S**ocial **M**odel
SON	**S**elf-**O**rganizing **N**etwork
SSL	**S**emi-**S**upervised **L**earning
SVR	**S**upport **V**ector **R**egressors
TA	**T**iming **A**dvance
TBS	**T**errestrial **B**eacon **S**ystem
TCP	**T**ransmission **C**ontrol **P**rotocol
TDOA	**T**ime **D**ifference **O**f **A**rrival
TOA	**T**ime **O**f **A**rrival
TM	**T**race **M**odel
TSM	**T**race and **S**ocial **M**odel
TTI	**T**ime **T**ransmission **I**nterval
UE	**U**ser **E**quipment
UETR	**U**ser **E**quipment **T**raffic **R**ecording
UL	**U**p**L**ink
URLLC	**U**ltra **R**eliable **L**ow **L**atency **C**ommunications
USL	**U**n**S**upervised **L**earning
UTDOA	**U**plink **T**ime **D**ifference **O**f **A**rrival
VLAN	**V**irtual **L**ocal **A**rea **N**etwork

Capítulo 1

Introducción

Este capítulo introduce el contenido de esta tesis doctoral. En primer lugar, se presenta el marco conceptual del trabajo, definiendo las motivaciones que han llevado al desarrollo del mismo. A continuación, se enumeran los objetivos de esta tesis. Por último, se muestra la estructura del resto del documento y se discuten las principales contribuciones de esta tesis doctoral.

1.1. Motivación

Es evidente que, en los últimos años, la evolución y crecimiento de las redes móviles ha sufrido una progresión imparable. El constante desarrollo de nuevos equipos, más económicos, accesibles y potentes, ha democratizado el acceso a las redes celulares, disparando las peticiones de servicios ya afianzados, como servicios de transmisión en directo, acceso a redes sociales, juegos en tiempo real, etc. Estas mejoras han provocado que tanto el tráfico generado en la red como el número de dispositivos conectados simultáneamente hayan sufrido un crecimiento exponencial. En paralelo, han ido apareciendo novedosos servicios desde el punto de vista del usuario (p.ej., vídeos en 360 grados, realidad virtual, comunicaciones entre máquinas en entornos de Internet de las Cosas, trabajos o juegos en la nube, ciudades inteligentes, etc.) que se prevé que multiplicarán este crecimiento, estando previsto un incremento de 3 veces del tráfico actual y un total de 9.100 millones de dispositivos móviles activos para 2028 [1].

Para cubrir las necesidades de este crecimiento masivo, nacen las nuevas tecnologías englobadas en la quinta generación de comunicaciones móviles (5G) [2]. Los nuevos

servicios que han ido apareciendo en los últimos tiempos se comportan de forma completamente distinta entre sí, por lo que requieren de distintas configuraciones para su óptimo funcionamiento. En esta quinta generación, para segregar de forma eficiente estos servicios y así poder optimizar su desempeño, se definen tres grupos de servicios distintos según sus requerimientos: a) servicios de banda ancha móvil mejorada (*enhanced Mobile Broadband*, eMBB), que engloba aquellos servicios que precisan de altas tasas de transferencia (como la realidad virtual, el videostreaming en altas resoluciones o las transmisiones en directo), b) comunicaciones masivas tipo máquina (*massive Machine Type Communications*, mMTC), que presenta escenarios en los que hay una ingente cantidad de dispositivos conectados entre sí, como en equipos de monitorización de salud o ciudades inteligentes, y c) comunicaciones ultra fiables de baja latencia (*Ultra Reliable Low Latency Communications*, URLLC), en las que una comunicación casi inmediata y estable entre dispositivos es vital, como en las comunicaciones entre vehículos o la cirugía a distancia.

Este crecimiento exponencial ha resultado en redes celulares tan grandes y complejas que es imposible realizar tareas de mantenimiento y optimización de forma manual. Esta limitación se verá acentuada con las soluciones implementadas en 5G para cubrir las necesidades de los distintos casos de uso. Así, la complejidad de la gestión de red se ha identificado en la industria como uno de los principales cuellos de botella en las futuras redes 5G. En este contexto, nació el concepto de Redes Autoorganizadas (*Self-Organizing Network*, SON) [3, 4], con el objetivo de automatizar estas tareas de mantenimiento. Estas técnicas se pueden dividir en cuatro bloques: Autoplanificación (*Self-Planning*), Autodespliegue (*Self-Deployment*), Autooptimización (*Self-Optimization*) y Autoreparación (*Self-Healing*).

Los operadores de redes móviles tradicionalmente se basaban en parámetros de Calidad de Servicio (Quality of Service, QoS) para realizar las tareas de mantenimiento de la red. Hoy en día, este enfoque, centrado en aspectos de la red, se ha sustituido por un enfoque más centrado en el usuario final, orientado a determinar con mayor precisión la percepción que el usuario tiene del comportamiento del servicio que se le ofrece, en lo que se conoce como la Calidad de Experiencia (Quality of Experience, QoE) [5]. Parametrizar la experiencia que los usuarios tienen al hacer uso de los diferentes servicios permite optimizar los procesos de planificación, optimización y reparación de la red. Por este motivo, uno de los factores principales que destacan a los diferentes operadores de redes y servicios frente a sus competidores es la Gestión de Experiencia del Usuario/Cliente (Customer Experience Management, CEM) [6]. To-

do ello hace que la comprensión de los servicios, los parámetros más relevantes en la experiencia del usuario y el contexto sean un pilar fundamental en el desarrollo del sector. Desafortunadamente, por lo general, la relación entre los diferentes indicadores de rendimiento que registra la red y la calidad que percibe el usuario no es trivial [7]. Por ello, es complicado conocer la opinión de los usuarios sobre los diferentes servicios, siendo la principal herramienta la realización de costosas encuestas, de las que se obtiene información estadística que puede no ser precisa. Estas limitaciones dificultan la realización de modelos y mediciones de la QoE.

El contexto del usuario, formado por las características propias o del entorno que enmarcan al usuario en el uso de la red móvil (como, p.ej., el tipo de terminal, la ubicación interior/exterior, la hora del día, la geolocalización o la actividad que el usuario esté realizando), es uno de los factores que más influyen en la percepción del servicio [8]. Por ello, los modelos de QoE más sofisticados tienen en cuenta el contexto del usuario (p.ej., la hora del día o su ubicación). Para reconocer la actividad del usuario, los proveedores de servicio pueden utilizar las mediciones activas de los sensores en el cuerpo a través del aprendizaje combinado (*ensemble learning*) [9], pero esta información no se registra en la red y, por tanto, no puede usarse en los procesos de autogestión. Como alternativa, los operadores de red pueden inferir el contexto del usuario aprovechando los eventos de señalización radio registrados por la red en cada conexión [10]. Una vez obtenida, esta información puede utilizarse para desarrollar algoritmos SON que hagan uso del contexto para mejorar su desempeño [11].

Para cumplir con las ambiciosas especificaciones de los nuevos casos de uso de 5G, los operadores llevarán el rendimiento de la red al siguiente nivel mediante la combinación de múltiples técnicas. Entre ellas, la densificación de la red se ha reconocido como una forma eficiente de proporcionar una mayor capacidad de red y una mejor cobertura. En zonas densamente pobladas, la densificación se consigue combinando la infraestructura macrocelular con despliegues de celdas pequeñas (*Small Cells*, SC). En estas redes heterogéneas, compuestas por macroceldas y celdas pequeñas, decidir la mejor ubicación para nuevas SC es clave para aprovechar al máximo la nueva infraestructura. Sin embargo, la mayoría de los enfoques actuales de selección de emplazamientos sólo tienen en cuenta indicadores simples de cobertura de red y calidad de señal, por la dificultad de modelar la planificación dinámica de paquetes con usuarios de múltiples servicios y diferentes condiciones de enlace radio en una herramienta de planificación de red. En ausencia de dicho modelo de rendimiento, la función de minimización de pruebas en ruta (*Minimization Drive Test*, MDT) [12] permite la recopilación de mediciones

geolocalizadas que pueden utilizarse para construir mapas precisos de rendimiento de la red (*Radio Environment Map*, REM). Dichos mapas pueden utilizarse después para detectar huecos de cobertura [13]. Lamentablemente, MDT rara vez se activa en las redes reales por la sobrecarga de trabajo que supone el procesamiento de estas mediciones. Como consecuencia, las tareas de replanificación y optimización de red a menudo tienen que realizarse basándose en mediciones sólo posicionadas geográficamente con el identificador de la celda y las estadísticas de avance temporal (*Timing Advance*, TA). Este enfoque conduce a grandes errores de localización, más aún en zonas con un TA elevado, donde la combinación de identificador de celda y TA produce un anillo de localización de una gran superficie, lo que impide estimar el contexto del usuario [14]. Con estos errores de posicionamiento, es muy difícil deducir el contexto de la comunicación. Esta limitación se agrava cuando, para reducir la necesidad de almacenamiento, el sistema de gestión no registra las medidas de posicionamiento de forma individual (por usuario). sino agregada (por celda o, en el mejor de los casos, por anillo).

Con los recientes avances en computación, que permiten un procesado eficiente de una ingente cantidad de información, el interés por la ciencia de datos ha crecido tanto en el ámbito científico como empresarial. Como resultado, se han puesto en marcha muchas iniciativas de datos abiertos en todo el mundo. Los portales de datos abiertos ofrecen ahora un acceso directo y automatizado a valiosos activos que pueden utilizarse para mejorar la gestión de redes celulares. Algunas empresas de desarrollo de agentes de usuario (*Terminal Agents*) [15] ofrecen mediciones reales recogidas por usuarios anónimos, que pueden utilizarse para evaluar los despliegues actuales [16]. Igualmente, algunos portales de datos abiertos, por lo general de naturaleza gubernamental, publican información de la principal actividad socioeconómica que acontece en cada zona del terreno (denominadas uso de suelo en la literatura), que influye enormemente en la forma en la que cada persona hace uso de la red.

En este contexto de la sociedad de la información, las redes sociales se han afianzado en los últimos años como un servicio casi omnipresente y de fácil acceso, que son otra fuente interesante para entender el comportamiento de los usuarios. Así, la actividad en las redes sociales puede ser un indicador relevante para, p.ej., predecir el tráfico celular, independientemente de la tecnología de acceso o del proveedor de red [17], o detectar zonas de escasa cobertura de la señal o rendimiento del servicio [18]. Por otro lado, las aglomeraciones puntuales de usuarios son un potencial problema en el rendimiento de la red y en muchos casos eran, a priori, impredecibles. En general, estas aglomeraciones se deben a eventos sociales o laborales, que en muchos casos se planifican, quedando

registrados en repositorios de acceso libre en forma de calendarios. Esta información puede utilizarse posteriormente para explicar el comportamiento anormal de la red durante los procedimientos de diagnosis de problemas.

Este entorno, en el que se facilita enormemente el acceso al estado y el contexto espaciotemporal de las personas, propicia que los operadores de red empiecen a implementar algoritmos en los que esta información juega un papel relevante. Estos algoritmos, usualmente basados en técnicas de aprendizaje automático, se enfocan en el mantenimiento preventivo de la red, mejorando la eficacia de los procesos de optimización y planificación para que el usuario perciba la mejor experiencia posible.

1.2. Objetivos y Contribuciones

El objetivo principal de esta tesis doctoral es desarrollar algoritmos que, aprovechando las técnicas de tratamiento de datos y de aprendizaje automático, mejoren la eficacia de procesos de planificación y optimización automática de la red, apoyándose en información de contexto espaciotemporal de los usuarios. Los operadores de la red tienen acceso a una gran variedad de indicadores que describen el comportamiento de la red y los usuarios, pero la información del contexto no está disponible directamente. Por ello, primero es necesario mejorar la detección de distintos aspectos de la situación del usuario en la red móvil (como la ubicación geográfica, el entorno, etc.) mediante técnicas de tratamiento de datos masivos. Con este objetivo genérico, los objetivos específicos de esta tesis son:

a) Desarrollar algoritmos para la clasificación a posteriori de conexiones de interior o exterior a partir de los descriptores radio de la conexión.

b) Desarrollar algoritmos para la mejora en el posicionamiento de conexiones a partir de los descriptores de la conexión y mensajes geolocalizados de redes sociales.

c) Desarrollar algoritmos para identificar áreas en la red con mal comportamiento, debido a malas condiciones de propagación o a una alta demanda simultánea, aprovechando la información de contexto espacial en el escenario.

d) Desarrollar algoritmos para la predicción del tráfico en la red, teniendo en cuenta las condiciones espaciotemporales en cada celda, como, p.ej., aglomeraciones de usuarios debidas a eventos de carácter social.

Como resultado del trabajo, se han generado las siguientes contribuciones:

- Un conjunto de herramientas que, a través de varias fuentes, obtienen datos de distinta naturaleza con información espaciotemporal (como publicaciones geolocalizadas de redes sociales, datos sobre los usos de suelo en cada zona de los escenarios utilizados, la localización y el horario de eventos sociales masivos, ...), que pueden usarse para modelar el comportamiento de usuarios.

- Un modelo que aprovecha la información de uso de suelo para estimar la probabilidad de que una conexión haya ocurrido en interior a partir de sus indicadores de red.

- Un algoritmo que combina la información de la probabilidad de que cada conexión haya ocurrido en interior, los usos de suelo y la distribución espacial de publicaciones de redes sociales en el escenario para distribuir estadísticamente las conexiones en el terreno.

- Una metodología para, a partir del comportamiento de los usuarios, estimar la mejora potencial de la red en términos de volumen de tráfico con óptimas condiciones de propagación y recursos disponibles, tanto a nivel de conexión como a nivel de celda, y un algoritmo que, haciendo uso del modelo de posicionamiento anterior, detecta áreas de la red en las que ocurren problemas, tanto de capacidad como de cobertura.

- Un modelo de aprendizaje profundo para predecir el tráfico generado en una celda, corrigiendo dicha predicción en los eventos sociales futuros aprovechando la información de su horario y su naturaleza (conciertos, partidos de fútbol, etc.).

Para asegurar la aplicabilidad de los modelos, todas las contribuciones se basan en datos reales, combinando tanto datos de acceso abiertos como datos propios de la red de telefonía, accesible únicamente por los operadores. Estos datos permiten desarrollar modelos más realistas, mejorando la detección de diferentes aspectos espaciotemporales (como el geoposicionamiento de los usuarios, las franjas horarias críticas en la red, etc.) que permiten afinar los procesos de gestión de red. Al emplear datos reales de diferentes fuentes, es necesario procesarlos en tiempo diferido para sincronizarlos en el dominio espacial y temporal.

1.3. Metodología de trabajo

En esta tesis se sigue una línea de trabajo continuo incremental, en la que cada contribución se basa en los modelos desarrollados anteriormente, demostrando la aplicabilidad de los mismos. Para cada una de las contribuciones, se ha seguido la misma metodología, descrita a continuación:

1. El trabajo comienza con un estudio detallado del problema que se pretende abordar, revisando el estado de la tecnología de acceso radio más usada actualmente (*Long Term Evolution*, LTE), además de las diferentes técnicas de tratamiento de datos y de aprendizaje automático (como modelos de clasificación o arquitecturas específicas de redes neuronales para predicción de series temporales aplicadas en cada uno de los problemas).

2. Seguidamente, se formula el problema específico tratado en el capítulo. Para ello, se describen los puntos clave que caracterizan el problema, junto a los posibles impedimentos que explican el por qué no se ha abordado hasta ahora con el enfoque propuesto. Además, en el caso de existir una hipótesis de partida, se comprueba la veracidad de la misma. En esta fase se realiza la especificación de requisitos, definiendo las entradas (datos de partida, tanto propios de la red celular, como son las trazas de conexión, como externos, p. ej. publicaciones geolocalizadas de redes sociales o usos de suelo) y salidas del algoritmo. Una vez definidos los datos necesarios, se desarrollan las diferentes herramientas para su recolección y tratamiento.

3. Una vez definido el problema, se concibe el modelo para solventarlo. Primero se plantea el diagrama de flujo del algoritmo de modelado desarrollado, en el que se muestran las diferentes etapas. Después, se detalla cada uno de los procesos en cada etapa, seleccionando, si es necesario, el modelo de aprendizaje automático que mejor se adapta al problema a resolver con el tamaño del conjunto de datos disponible.

4. A continuación, se detallan los datos de partida para el desarrollo del modelo (características de la red celular de la que se extrae la información, estadísticas de los datos extraídos, información empleada ajena a la red...), así como los resultados obtenidos, comparándolos, si es posible, con otros algoritmos ya existentes para determinar la mejora obtenida respecto al estado de la tecnología. Además,

para demostrar la viabilidad de la implementación de los modelos en herramientas reales de planificación radio, se valoran aspectos computacionales de cada modelo.

5. Por último, se extraen las conclusiones, tanto del desarrollo de los modelos como de sus resultados.

1.4. Estructura del documento

En esta sección se presenta una breve descripción del contenido de cada capítulo, destacando las principales contribuciones en cada uno de ellos.

Para enmarcar el desarrollo de los modelos descritos en los capítulos posteriores, el ***Capítulo 2*** presenta una descripción teórica de los conceptos y técnicas básicas que se emplean en el trabajo. Por un lado, se presenta un breve resumen de conceptos relacionados con las redes de acceso de telefonía móvil, así como la información generada en ellas. Al mismo tiempo, se describen los conceptos relacionados con las redes autoorganizadas, finalizando con una revisión de las técnicas de geolocalización en redes de telefonía. Para facilitar la lectura de los siguientes capítulos, se presentan también las diferentes técnicas de aprendizaje automático empleadas, tanto las clásicas como algunas más avanzadas, basadas en aprendizaje profundo.

Una vez asentadas las bases, en los siguientes cuatro capítulos se describen cada uno de los modelos desarrollados. El primero de ellos, que sirve como punto de partida, se presenta en el ***Capítulo 3***. Este capítulo explica el modelo que estima el contexto de los usuarios, clasificando las conexiones como de interior o exterior en base a sus descriptores registrados.

El contexto del usuario es una información extremadamente útil. Por ejemplo, un mismo usuario puede hacer un uso distinto del mismo servicio, dependiendo si está en interior (p.ej., en su casa o en la oficina) o en exterior (p.ej., en parques o calles), por lo que puede usarse para estimar mejor la experiencia del usuario. Otra utilidad de esta información puede ser como soporte en procesos de posicionamiento, descartando localizaciones que no concuerden con el contexto del usuario. En este sentido, el ***Capítulo 4*** aplica los resultados del modelo descrito en el capítulo anterior, junto a información espacial de publicaciones georreferenciadas obtenidas de redes sociales, para deducir estadísticamente la distribución espacial de los usuarios en la red.

Una vez conocida la distribución espacial de los usuarios, los procesos de planifica-

ción y optimización de red pueden ser llevados al siguiente nivel, identificando potenciales cuellos de botella de capacidad causados por la aglomeración de usuarios. Este es el objetivo del modelo presentado en el ***Capítulo 5***, donde se usan los resultados de los dos modelos anteriores para detectar en qué zonas de la red se generaría más tráfico al mejorar las condiciones de la red mediante el despliegue de una nueva picocelda.

La previsión de futuros cuellos de botella de capacidad en la red es vital para cumplir los objetivos marcados por los diferentes casos de uso en la nueva generación de comunicaciones móviles. Estas degradaciones del servicio ofrecido por la red no son siempre persistentes en el tiempo, sino que muchas veces son de naturaleza esporádica. Un claro ejemplo son las aglomeraciones puntuales de usuarios en eventos de carácter social. En el ***Capítulo 6*** se presenta la última aportación de esta tesis, que consiste en un modelo de aprendizaje profundo para predecir el tráfico en la red en estos eventos masivos.

Finalmente, el ***Capítulo 7*** engloba las conclusiones finales del trabajo, remarcando las principales contribuciones de esta tesis doctoral y explorando las posibles líneas futuras de investigación.

De forma adicional, en el ***Apéndice A*** se explican las diferentes herramientas de recolección y procesado de datos abiertos con información del contexto espaciotemporal de los usuarios, empleadas en el desarrollo de todos los modelos contenidos en esta tesis.

Capítulo 2

Marco teórico

Este capítulo contextualiza los modelos propuestos en los capítulos posteriores, explicando tanto la tecnología en las que se ubican como las técnicas de aprendizaje automático empleadas a lo largo de la tesis. Con este fin, el capítulo se divide en cuatro secciones. En la sección 2.1, se explican brevemente las redes de acceso de telefonía móvil, así como la diversa información recolectada en ellas. En la sección 2.2 se hace un breve resumen del paradigma de las redes Autoorganizadas, describiendo cada una de las familias de tareas que las componen. Una vez determinada la importancia del contexto espaciotemporal, tanto en las redes actuales como en las futuras, la sección 2.3 profundiza en las diferentes técnicas de localización en redes móviles. Por último, la sección 2.4 se dedica a las técnicas de aprendizaje automático empleadas a lo largo de esta tesis doctoral.

2.1. Redes de acceso móvil

Las redes de telefonía móvil son una de las tecnologías más relevantes en la sociedad actual, permitiendo a los usuarios acceder a servicios y contenidos de forma prácticamente ubicua. Para alcanzar estos niveles de cobertura, las redes móviles se componen de una red de acceso y un núcleo de red.

La red de acceso es una estructura no jerárquica formada por un gran número de estaciones base que, a través de la interfaz radio, dan servicio a los equipos de usuario (*User Equipment*, UE). Desde la introducción del estándar LTE [19], la tendencia es dotar de mayor funcionalidad a las estaciones base, siendo capaces de realizar gran

parte de las tareas de gestión de sus recursos radio para reducir la latencia en la red. Dependiendo de los recursos de los que disponen y la potencia de transmisión de sus antenas, las estaciones base se reparten el área geográfica de la red en secciones, denominadas celdas, permitiendo a los UE cambiar de celda de forma transparente al usuario conforme este se desplaza por el terreno. En función de su alcance, estas celdas pueden clasificarse como:

- Macroceldas. Son celdas destinadas principalmente a asegurar la cobertura en las ciudades, así como entornos rurales o poco poblados, por lo que tienen un rango de cobertura muy amplio, llegando a los pocos kilómetros.

- Microceldas. Las microceldas tienen como objetivo cubrir zonas con alta densidad de población, como pueden ser núcleos urbanos. En estos casos la gran concentración de usuarios hace que los recursos disponibles en la celda sean ocupados por los usuarios más cercanos, por lo que no tienen unos requisitos de potencia demasiado elevados y su rango de cobertura alcanza las centenas de metros.

- Picoceldas. Estas celdas se sitúan habitualmente en entornos públicos de interior (p.ej., palacio de congresos, oficinas...), aunque también pueden ser ubicadas en exterior. El objetivo de las picoceldas es solucionar problemas de cobertura o capacidad en puntos específicos de la red, por lo que el rango de este tipo de celdas alcanza las pocas decenas de metros.

- Femtoceldas. Este tipo de celdas son un caso muy particular de picocelda, empleadas para mejorar la cobertura en entornos de interior privado, como viviendas u oficinas, con un área de cobertura de varios metros y que utilizan como red de retorno el bucle de abonado.

El núcleo de red se encarga de realizar el resto de las tareas de gestión de recursos de la red de las que no pueden encargarse las estaciones base y recopilar la información generada en la red. Además, gestiona los procesos de suscripción y autenticación de los usuarios en la red, al igual que su movilidad entre celdas. Por último, el núcleo de red actúa como pasarela de servicios y datos entre los usuarios y la red troncal, así como con equipos ajenos a la red a través de internet.

La Fig. 2.1 ejemplifica una red de acceso, donde un conjunto de celdas cubre diferentes servicios solicitados por usuarios en la red. Estas celdas reciben del sistema de soporte de operaciones (*Operations Support System*, OSS) la configuración de la red

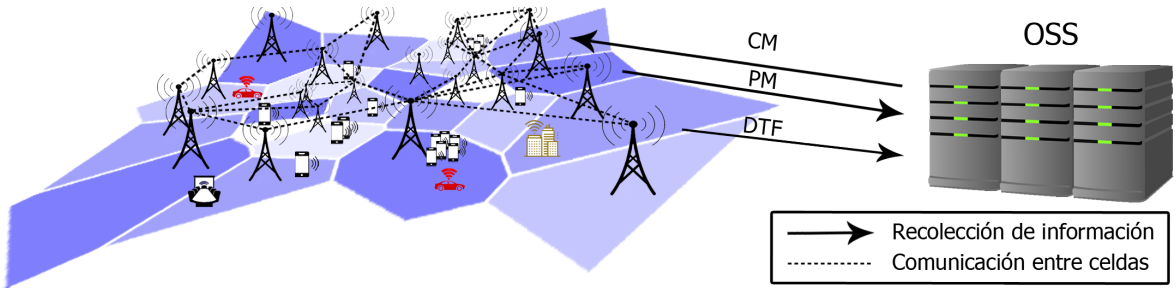


Figura 2.1: Arquitectura conceptual de una red de acceso.

en forma de ficheros de información de gestión de configuración (*Configuration Management*, CM). Una vez configuradas, las celdas reportan la información recopilada en forma de ficheros de información de gestión de rendimiento (*Performance Management*, PM) y ficheros de datos de trazas (*Data Trace Files*, DTF), descritos más adelante.

Para optimizar la gestión de recursos de red, es de vital importancia conocer el rendimiento de las celdas que la componen, así como el comportamiento de los usuarios a los que dan servicio. En este sentido, la recolección de información procedente de las estaciones base cobra una gran relevancia, permitiendo a los operadores adelantarse a los posibles problemas de rendimiento e, incluso, desarrollar modelos que permitan a la red autogestionarse, evitando así fallos introducidos por el factor humano.

2.1.1. Información registrada de la red

Una red celular genera una gran cantidad de información en forma de alarmas, contadores y trazas [10]. Esta información puede clasificarse en tres grupos, dependiendo de su naturaleza:

- La información de gestión de configuración (CM), donde se almacenan los parámetros de configuración de la red.
- La información de gestión de rendimiento (PM), que recolecta estadísticas agregadas, con diferentes resoluciones temporales, respectivas al comportamiento y rendimiento de los diferentes elementos de la red de acceso durante un periodo de tiempo, definido como periodo de producción de informes (*Reporting Output Period*, ROP).
- Los ficheros de datos de trazas, donde se recogen los mensajes de señalización entre los componentes de la red, así como los mensajes de comunicación entre los equipos de usuario y las estaciones base y, si está activa, la información recogida

con la funcionalidad de minimización de pruebas de conducción (*Minimization Drive Test*, MDT). Estas trazas radio pueden, a su vez, clasificarse en:

- Registros de tráfico de celda (*Cell Traffic Recordings*, CTR), que monitorizan el comportamiento de cada celda a partir de mediciones anónimas de todas las conexiones ocurridas en ellas.
- Registros de tráfico de equipos de usuario (*User Equipment Traffic Recordings*, UETR), que contienen la información de una sola conexión en concreto, tras ser anonimizada.

Las trazas radio consisten en archivos donde se registran eventos de interacciones de control, generados por procedimientos comunes (p.ej., establecimiento y liberación de llamadas/sesiones) o funcionalidades avanzadas de automatización de la red (p.ej., equilibrio de carga). En el pasado, esta información tan detallada (estadísticas de TA, mediciones de RSRP, volúmenes de datos por canal, velocidad de descarga/subida, etc.) sólo se utilizaba para la diagnosis de problemas de forma directa por los operadores de la red. Sin embargo, con los recientes avances en el análisis de grandes datos, las trazas se han convertido en un activo muy valioso para la optimización automática de la red.

Los eventos registrados pueden ser externos e internos. Los eventos externos consisten en mensajes de señalización estandarizados que las estaciones base intercambian con otros elementos de la red. Por contra, los eventos internos incluyen información dependiente del proveedor sobre el rendimiento de la estación base. Asimismo, los eventos pueden ser periódicos (activados por tiempo) o no periódicos (activados por evento). Todos estos eventos se registran en archivos generados por la estación base y se envían periódicamente al OSS. A continuación, los archivos de trazas se sincronizan y decodifican mediante técnicas de procesado de eventos complejos (*Complex Event Processing*, CEP) [20] en un servidor de procesamiento de trazas.

El registro de todos los eventos generados hace que las trazas de conexión tengan una riqueza de información superior a los contadores recogidos en los PM y, por tanto, son mucho más interesantes a la hora de determinar el contexto de cada usuario, por lo que se emplean como fuente de información de la red a lo largo de gran parte de la tesis. Por desgracia, esta riqueza de información conlleva un gran requisito de almacenamiento, por lo que los operadores de red no siempre activan la recolección de todos los eventos disponibles y, cuando lo hacen, es por un corto periodo de tiempo. Por tanto, sigue siendo necesario recurrir a la información de los PM para estudiar la evolución de la red a medio-largo plazo.

2.2. Redes AutoOrganizadas

El desarrollo de la tecnología móvil ha llevado a la aparición de terminales más potentes y asequibles, derivando en un incremento de los equipos conectados y, en última instancia, en una red de telefonía tan sofisticada que dificulta la tarea de los operadores de red. La complejidad introducida por las mejoras planteadas en las redes 5G no hace más que acentuar este problema, por lo que automatizar las tareas de mantenimiento y gestión de la red es clave para su correcto funcionamiento. Estos factores hacen que las redes Autoorganizadas [3] sean de especial interés, facilitando enormemente la gestión de los recursos y reduciendo los costes de implementación (*Capital Expenditure*, CAPEX) y operación (*Operational Expenditure*, OPEX).

Las tareas de mantenimiento y gestión de la red abarcan un amplio abanico de cometidos de distinta naturaleza. Por claridad, en la bibliografía se distinguen 4 grupos de tareas principales, expuestos a continuación.

2.2.1. Autoplanificación

El primero de los grupos comprende la planificación automática de elementos en la red. En este grupo, el objetivo es determinar de forma dinámica el ajuste y la localización de nuevos elementos de la red troncal y de acceso para solucionar carencias en la red celular, producidas por cambios en la topología del escenario o los comportamientos de los usuarios. Para ello, dentro de la Autoplanificación se distinguen los siguientes casos de uso:

- **Planificación de la ubicación de emplazamientos**. Una de las principales tareas en la planificación de la red es determinar la posición de nuevos emplazamientos en la red de acceso, para obtener el máximo rendimiento con el menor coste, así como las características del nuevo emplazamiento (p.ej., el número de celdas emplazadas, los recursos de cada una de ellas, sus características físicas, etc.).

- **Planificación de parámetros**. Una vez fijado el emplazamiento, es necesario configurar el nuevo nodo para su correcto funcionamiento. En este sentido, es necesario ajustar dos tipos de parámetros. Por un lado, se configuran los parámetros radio, incluyendo tanto los nuevos indicadores de celda y la lista de celdas vecinas, como ajustes de los canales de acceso aleatorio (Random Access CHan-

nel, RACH) y aviso (Paging CHannel, PCH) y ajustes de funcionamiento de las celdas (potencia, traspaso y gestión de recursos radio). Por otro lado, es necesario configurar la comunicación del nuevo nodo con el resto de los equipos de red a través de los parámetros de transporte, como las direcciones IP, la inclusión en redes virtuales locales (Virtual Local Area Network, VLAN) o los ajustes de Calidad de Servicio.

- **Planificación de estructura de red**. Una vez fijadas tanto la ubicación como la configuración de los nuevos nodos, estos se incluyen en las listas de las celdas cercanas, para habilitar los traspasos y que el nodo esté definitivamente integrado en la red.

2.2.2. Autodespliegue

Una vez planificado un nuevo nodo en la red, se prepara para su explotación comercial. Para minimizar la intervención humana, que puede llevar a errores de configuración, es interesante automatizar estas labores de implementación. De esto se encargan las tareas de Autodespliegue, cubriendo las necesidades de preparación, instalación, autentificación y verificación de los componentes que conforman el nuevo nodo, tanto virtuales (software) como físicos (hardware), presentadas a continuación.

- **Preparación**. La primera tarea para desplegar los nuevos nodos es preparar todos los componentes necesarios para su correcto funcionamiento. Uno de los elementos clave del sistema es el software que habilita la operación del nodo, por lo que es vital que no haya errores en su instalación. Para asegurar la integridad del software descargado, es necesario implementar mecanismos de apoyo en los procesos de descarga, instalación y actualización de los componentes software del nodo.

- **Instalación**. Tras preparar todos los elementos necesarios, se instalan y conectan cada uno de ellos, preparados para autoconfigurarse con la mínima intervención humana, con las configuraciones planificadas en las tareas del grupo anterior, generando de forma automática una lista con todos los componentes hardware, así como su estado y configuración, en el Sistema de Inventario de la red.

- **Autentificación**. Una vez instalados y configurados los componentes, es preciso dotar al nodo de la seguridad suficiente para mantener la integridad de la red.

Para ello, primero se necesita autentificar el nodo en la red, autorizándolo para acceder a las entidades de red, tanto internas como externas. Tras este paso, se configuran los procedimientos necesarios para establecer un túnel bidireccional y estable a las puertas de enlace de acceso y de operación y mantenimiento (Operation and Maintenance, O&M).

- **Verificación**. Con los procesos de instalación y autentificación terminados, es necesario comprobar si el nodo está preparado para su explotación comercial. En esta tarea se verifica el correcto funcionamiento y configuración de los distintos componentes instalados, examinando el estado del nuevo nodo y su preparación para ser explotado de forma comercial. En caso negativo, se debe automatizar la generación de un informe indicando el origen de los fallos.

2.2.3. Autooptimización

En las tareas de Autooptimización se busca mejorar el rendimiento actual de algún aspecto en la red móvil de forma dinámica. Este objetivo se consigue a partir de medidas realizadas en tiempo real de indicadores de rendimiento, tanto a nivel de usuario (medidas de los terminales) como a nivel de celda (obtenidas de las estaciones base). Para esta labor, el contexto de los usuarios es un factor de extrema relevancia. Los procesos de optimización automática comprenden:

- **Optimización de cobertura**. Los cambios introducidos continuamente en la arquitectura de la red, debidos tanto a tareas de Autoplanificación como a otras tareas de Autooptimización, pueden derivar en una pérdida de cobertura en algunas zonas respecto a la planificación inicial. Por tanto, optimizar las áreas de servicio de las celdas en la red, continuamente y de forma automática, es clave para que el usuario final reciba el mejor servicio posible. La mejora de cobertura incluye la actualización de la lista de celdas vecinas en cada uno de los nodos de la red o el control de interferencia entre celdas vecinas para mejorar la calidad de señal (y, con ello, la eficiencia espectral) en la red.

- **Optimización de capacidad**. Otro de los objetivos de optimización es aprovechar al máximo los recursos ya disponibles en la red, para reducir la necesidad de nuevos equipos. Áreas de la red altamente transitadas pueden derivar en celdas completamente cargadas, junto a celdas vecinas relativamente desocupadas y, por tanto, en un uso ineficiente de la capacidad de la red. De solucionar este proble-

ma se encargan las tareas de optimización de capacidad, balanceando la carga de las celdas cercanas, modificando los márgenes de traspaso u optimizando la asignación de recursos.

- **Optimización de calidad de experiencia**. En algunas situaciones, los recursos disponibles en la red son suficientes para cubrir las necesidades de tráfico, pero, debido a desajustes en la configuración de la red de retorno, no se obtiene el mejor rendimiento de los servicios prestados. Las tareas de optimización de calidad de experiencia resuelven estos problemas analizando el rendimiento de la comunicación extremo a extremo, mejorando la respuesta de los elementos de la red de transporte y del núcleo de red o enfocando la optimización de parámetros radio desde el punto de vista de la experiencia final del usuario.
- **Optimización de la eficiencia operacional del sistema**. Las tareas de Autooptimización suelen buscar mejorar algún aspecto de la red de acceso. En algunos casos, estos aspectos no solo tienen relación con la calidad del servicio que se presta al usuario, sino con la eficiencia con la que funcionan los equipos de la red, pudiendo conllevar sobrecostes energéticos innecesarios cuando los equipos no operan eficientemente. Hay ocasiones en las que partes de la red que no son usadas, debido al bajo tráfico que soportan durante un periodo de tiempo. En este escenario, las tareas de ahorro de energía buscan aumentar la eficiencia energética, apagando los equipos sin uso durante los periodos de baja actividad.

2.2.4. Autoreparación

Este grupo de tareas persigue solucionar, tanto reactiva como proactivamente, los diferentes problemas de rendimiento y operación que aparezcan en la red, incluyendo las mejoras o reemplazos de todos los componentes, tanto virtuales como físicos.

- **Monitorización de la red**. Para conseguir recuperar servicios interrumpidos o solventar problemas sin que afecten significativamente a la experiencia de los usuarios, el sistema debe detectar cualquier fallo en el menor espacio de tiempo. Con este objetivo, la red debe ser capaz de monitorizarse de forma autónoma, integrando las medidas de rendimiento necesarias y correlacionándolas para la gestión de fallos o la detección de interrupción de algún servicio.
- **Mantenimiento de recursos**. Las tareas de mantenimiento de los componentes de la red deben ser lo más autónomas posible, para evitar largos periodos de

inactividad por descuidos humanos. En el caso del hardware, no es posible realizar de forma automática el reemplazo de componentes defectuosos, pero debe requerir la mínima intervención manual. Por otro lado, el mantenimiento de los componentes software puede llevar un mayor grado de autonomía, permitiendo la descarga y actualización automática de nuevo software o gestores de elementos de red.

- **Recuperación de fallos**. Con la monitorización automática de la red, los fallos en diferentes puntos de la arquitectura pueden ser detectados de forma dinámica. Esto permite la recuperación autónoma, si es posible, de interrupciones del funcionamiento de distintos elementos en la red o, en caso negativo, la mitigación de los problemas causados.

2.3. Posicionamiento de usuarios en redes móviles

Conocer la posición del usuario en una red celular es de importancia capital para optimizar la arquitectura y funcionamiento de la red y, por tanto, la experiencia final del usuario. En la actualidad, existen tecnologías de localización con una gran precisión (p.ej. *Global Navigation Satellite System*, GNSS), pero son externas a las redes móviles y, por tanto, no siempre están disponibles para los operadores de red. Esto hace que sean necesarios métodos de posicionamiento dentro de la red móvil, aprovechando su infraestructura y explotando diferentes características de propagación de las señales de comunicación. En esta sección se presenta un resumen de los fundamentos de las principales técnicas de posicionamiento en las redes celulares.

2.3.1. Medidas empleadas en la localización

En las redes celulares, los diferentes métodos de localización [21] se basan en la realización de medidas que dependan fuertemente de la posición del usuario. Estas mediciones pueden ser dependientes tanto de la estación base servidora como de las propiedades del canal de propagación. En concreto, las medidas de localización más habituales son:

– **Identificador de la celda servidora**. Para poder acceder a los servicios que ofrecen las redes celulares, los terminales deben conectarse a una celda de la red. Por tanto, conocer el número de identificación de la celda que da servicio al

terminal, o identificador de celda, permite usar la posición física de la estación base como punto de referencia fijado, acotando dónde se encuentra el equipo al área de cobertura de la misma.

- **Intensidad de la señal recibida**. La potencia medida por los usuarios de la señal emitida por las estaciones base depende de la distancia que los separa, aunque queda muy condicionada a las características del entorno en el que la señal radio se propaga. La atenuación introducida por el medio, las superficies reflectantes o los obstáculos de visión en el camino de propagación de la señal hacen que la intensidad en recepción tenga una gran variabilidad espacial y, por tanto, sea especialmente interesante a la hora de localizar un terminal.

- **Tiempo de llegada**. En contraposición a la intensidad de la señal, una propiedad que se mantiene relativamente constante, independientemente de las condiciones, es la velocidad de propagación de la onda en el medio. En consecuencia, una buena estimación de la distancia entre el terminal y la estación base servidora es el tiempo que tarda la señal en llegar de un equipo al otro. Esto se mide en las redes celulares de dos formas,

  a) *Time Of Arrival* (TOA), como el instante de tiempo marcado por el reloj interno del terminal en el momento de recepción de una señal enviada por la estación base servidora.

  b) *Time Difference Of Arrival* (TDOA), como la diferencia entre dos medidas de TOA obtenidas a partir de, al menos, dos señales de la misma naturaleza transmitidas desde estaciones base distintas en el mismo instante de tiempo.

- **Ángulo de llegada**. En un escenario donde las estaciones base que componen la red se apoyan en antenas direccionales, es posible medir el ángulo que forman el terminal que recibe la señal y la estación base servidora respecto a una referencia que, por lo general, está basada en la topología de la red.

- **Información externa**. Las medidas anteriores permiten realizar una primera estimación de la posición de un terminal en la red, pero en algunos casos no es suficiente (p.ej., en entornos de interior, donde las condiciones de propagación introducen demasiado ruido en la estima con GNSS). En estos casos, el uso de técnicas de minería de datos con información independiente de la red móvil (p.ej., datos de cámaras de vigilancia, direcciones MAC o IP, tipo de actividad socio-económica en la zona, publicaciones de redes sociales...) permite localizar los

terminales o actuar como ayuda en los métodos de posicionamiento basados en las medidas antes descritas.

2.3.2. Clasificación de los métodos de posicionamiento en redes celulares

Los métodos de posicionamiento pueden clasificarse de múltiples formas, en función de las propiedades del enlace de comunicación, el tipo de integración en la tecnología de comunicación... La clasificación más común es en función de la metodología usada para la recolección y procesado de las medidas empleadas en el posicionamiento. En este sentido, existen 4 tipos de posicionamiento, dependiendo de dónde se realizan las medidas y dónde se procesan,

a) **Autoposicionamiento**. Estas técnicas engloban aquellos métodos donde tanto la medición como el procesamiento recae en el terminal móvil.

b) **Posicionamiento remoto**. Al contrario que en las técnicas de autoposicionamiento, el posicionamiento remoto basa su funcionamiento en la realización de medidas y procesamiento de estas por parte de los equipos de la red.

c) **Autoposicionamiento indirecto**. Cuando los equipos de la red realizan las mediciones que se emplean en la localización, pero el procesado recae en el terminal móvil.

d) **Posicionamiento remoto indirecto**. Incluye técnicas donde las medidas son realizadas por el terminal móvil, pero se procesan en un equipo de la red.

2.3.3. Métodos de posicionamiento

Partiendo de las medidas descritas anteriormente, se han desarrollado un amplio abanico de métodos de posicionamiento en redes celulares. Debido a que en este tipo de redes el principal objetivo de las técnicas de localización es conocer la posición de un determinado terminal de usuario, estos métodos suelen ser de posicionamiento remoto o, más comúnmente, de posicionamiento remoto indirecto. Por simplicidad, en esta subsección se presentan primero las técnicas básicas de posicionamiento, continuando con una descripción del proceso de posicionamiento en las redes LTE actuales y los requerimientos para las futuras redes 5G.

Técnicas básicas de posicionamiento

a) Detección de proximidad

Las técnicas de detección de proximidad son aquellas que basan la localización de los terminales en su cercanía a unos puntos fijados en la red. En estos métodos de posicionamiento, se distinguen 3 técnicas,

- **Contacto físico**. El método más simple de posicionamiento mediante detección de proximidad es el caso en el que el terminal de usuario enlaza de forma física, ya sea con sensores de presión, contacto o distancia, con la estación base. En estos casos, si se conoce la posición de los sensores, es posible localizar con gran precisión al terminal.
- **Métodos identitarios**. En estos métodos, la posición de los equipos de usuario se estima en función del rango de comunicación de las estaciones base de la red. En la forma más sencilla de los métodos identitarios, la posición se define como la de la estación base servidora, a través del identificador de celda. Esta forma de posicionar puede introducir una gran imprecisión, siendo el rango de la celda el mejor estimador de la precisión del modelo. Con el objetivo de reducir este error, otras variaciones del método identitario básico incluyen la información de las celdas vecinas que es capaz de detectar el terminal, usándolas como centro de masas dentro del área de cobertura de la celda servidora.
- **Macro posicionamiento**. En las técnicas de macro posicionamiento se usan medidas de información externa a la red, aplicando técnicas de minería de datos para posicionar los terminales de usuarios.

b) Análisis del entorno (*fingerprinting*)

En las técnicas de análisis del entorno, el proceso de localización se fundamenta en la comparación de las medidas realizadas por el terminal con un mapa de mediciones, permitiendo conocer dónde se sitúa el usuario. Este mapa de medidas condiciona de forma determinante la precisión del modelo de posicionamiento, ya que, mientras mayor riqueza (tanto a nivel espacial como en la cantidad de propiedades medidas) posea el mapa, menor es el error cometido en la localización. Por tanto, en las técnicas de análisis de entorno se diferencian dos fases:

- **Fase de calibración**. Es la fase más crítica en este tipo de técnicas, ya que

construye el mapa de mediciones. Este mapa se puede construir de dos formas, a) mediante una batería de medidas en el escenario de localización, cubriendo la mayor cantidad y resolución de terreno posible, o b) mediante la simulación del entorno usando métodos de propagación.

– **Fase de explotación**. Una vez obtenido el mapa de caracterización del escenario, en la fase de explotación se compara este mapa con las medidas que realiza el equipo de usuario, obteniendo una aproximación de la ubicación del terminal.

c) Triangulación

Las técnicas de triangulación se basan en la aplicación de propiedades trigonométricas sobre medidas de intensidad, tiempo o ángulo en recepción para determinar la posición del usuario. Estas técnicas suelen ser apoyadas por métodos identitarios, ya que es necesario conocer la posición de la estación base de la que se miden las características de propagación. Dentro de esta categoría de métodos, se diferencian 3 técnicas,

– **Lateración**. Las técnicas de lateración localizan a los usuarios a partir de la distancia entre el terminal y, al menos, dos estaciones base de la red. Esta distancia se estima a partir de las medidas realizadas de la señal recibida, o bien del tiempo de llegada de la señal entre la estación base y el terminal.

– **Localización hiperbólica**. La técnica de localización hiperbólica es una variación de la lateración mediante tiempo de llegada. En este caso, la triangulación se realiza a partir de las medidas de diferencia de tiempo de llegada, TDOA. Estas medidas forman una hipérbole desde la posición de un par de estaciones base, por lo que es necesario, al menos, recibir las mediciones de 3 estaciones base diferentes para realizar la localización.

– **Angulación**. La última técnica de triangulación se basa en las mediciones del ángulo de llegada. En el método de angulación, se determina la posición del usuario como la intersección de, al menos, dos rectas, cada una de ellas en la dirección del ángulo que forma el terminal con cada estación base.

d) Métodos híbridos

Cada uno de los métodos anteriores se basan en alguna medida que puede introducir imprecisiones en la localización, ya sea por la naturaleza de la medida (métodos

identitarios), por los equipos de medición o las condiciones de propagación (triangulación) o la necesidad de una etapa de calibración (técnicas de análisis de entorno). Para solventar estos problemas nacen los acercamientos híbridos, que combinan medidas y/o técnicas de posicionamiento diferentes para aumentar la precisión del método.

Posicionamiento en redes de telefonía actuales

Actualmente, la manera más extendida de posicionar usuarios en las redes de telefonía es a través de mediciones GNSS asistido y, en concreto, la tecnología de GPS asistido (A-GPS) [22], debido a la integración de esta tecnología en prácticamente todos los terminales del mercado. Para que los terminales puedan procesar las señales GPS, se necesita una conexión de línea de visión con varios satélites y el conocimiento de los parámetros orbitales. Estos datos se descargan antes de la primera fijación de posición, que puede tardar varios minutos (*cold fix*). Para acelerar el proceso, un servidor de red puede proporcionar datos de asistencia (p.ej., sincronización precisa, desplazamiento Doppler del satélite, posición aproximada del dispositivo móvil o correcciones diferenciales) para obtener una medición GPS más rápida y fiable (*warm fix*), en lo que se conoce como tecnología A-GPS. No obstante, estas mediciones GPS solo suelen reportarse al OSS a través de la funcionalidad MDT, que rara vez está activada por el consumo de batería de los terminales y la necesidad de capacidad de almacenamiento y procesado que supone. Por tanto, los operadores necesitan una manera alternativa de localizar usuarios en la red, apoyada principalmente por las medidas obtenidas de la propia red, dependientes de la Tecnología de acceso radio (*Radio Access Technology*, RAT) empleada.

a) Redes 4G

En las redes LTE actuales, los requisitos de localización definidos por el Proyecto de Colaboración de Tercera Generación (*3rd Generation Partnership Project*, 3GPP) vienen determinados por las llamadas de emergencia, que necesitan una precisión de hasta 50 metros en el 67 % y de hasta 150m en el 95 % de las conexiones en un país.

Para cumplir ese objetivo, los fabricantes introdujeron una serie de técnicas de localización. Por un lado, métodos independientes de las técnicas RAT, como los métodos A-GNSS (A-GPS, Galileo, BeiDou...) o la mejora de otros métodos usando métricas de sensores barométricos, Wi-Fi/Bluetooth o Sistemas de Balizas Terrestres (*Terrestrial Beacon System*, TBS), principalmente ideados para la localización en interiores.

Por otro lado, métodos que se basen en métricas de la red, descritas a continuación,

- **E-CID**. El Identificador de Celda Mejorado (*Enhanced Cell ID*, E-CID) es una ampliación del método identitario basado en el ID de la celda servidora. En esta técnica, se añaden al identificador de celda otras métricas de la red, como el Avance de Tiempo, el Tiempo de Viaje o el Ángulo de Llegada de la señal, para reducir el área de localización en el rango de cobertura de la celda servidora.

- **OTDOA**. Una de las técnicas de posicionamiento más comunes es la Diferencia de Tiempo Observada (*Observed Time Difference Of Arrival*, OTDOA). Esta técnica envía desde varias estaciones base cercanas unas señales específicas, llamadas Símbolos de Referencia de Posicionamiento (*Positioning Reference Symbols*, PRS), usadas por los terminales de usuario para medir la diferencia entre los tiempos de llegada desde dos estaciones base. Estas diferencias de tiempo, denominadas medidas de Diferencia de Tiempo de Señales de Referencia (*Reference Signal Time Difference*, RSTD), se envían al servidor de localización para computar la posición del usuario en cuestión.

- **UTDOA**. La Diferencia de Tiempo en el Enlace Ascendente (*Uplink Time Difference Of Arrival*, UTDOA) es una variación de la técnica OTDOA, que emplea las señales del enlace ascendente recibidas desde el equipo del usuario en varias estaciones base para realizar la localización.

b) Redes 5G

En las nuevas redes 5G la posición del usuario cobra mucho más valor, permitiendo adaptar el uso de los recursos de la red a las necesidades del usuario, dependiendo del caso de uso específico. Esto hace que los requerimientos de precisión en la localización de los equipos sean más estrictos, desde pocos metros (p.ej., servicios de banda ancha móvil mejorada, *enhanced Mobile BroadBand* -eMBB-) a decímetros (p.ej., comunicación entre vehículos) o incluso centímetros (p.ej., Industria 4.0 o servicios en interior). Por tanto, es necesario implementar métodos de posicionamiento más avanzados, que aprovechen las mejoras tecnológicas de 5G para incrementar la exactitud de la localización [23]. Algunas de estas mejoras son:

- **Redes heterogéneas**. La densificación de estaciones base introducida en las redes heterogéneas restringe el área de cobertura de cada celda en la red y, por

lo tanto, reduce el error de posicionamiento con técnicas de métodos identitarios, como el E-CID usado en LTE.

– **Frecuencias de portadora más altas y anchos de banda de la señal mayores**. Por un lado, el uso de portadoras de mayor frecuencia hace que la propagación de la señal se vea más influenciada por el escenario, bloqueando las transmisiones indirectas (*Non-Line-Of-Sight transmissions*, NLOS) y mejorando así la precisión de medidas que dependan de la línea de visión entre el terminal y la estación base, como el TOA o el AOA. Por otro lado, aumentar los anchos de banda usados mejora la resolución temporal del canal, optimizando medidas multitrayectoria, como el TOA o el tiempo de viaje (*Round-Trip Time*, RTT).

– **Antenas multidireccionales**. La introducción de antenas multidireccionales permite la transmisión en haces dentro del rango de cobertura de la celda. Esto permite mejorar la precisión de mediciones de potencia, angulares y de tiempo para los PRS usados en métodos de posicionamiento por OTDOA.

2.4. Técnicas de aprendizaje automático

Las redes celulares registran una cantidad ingente de información en forma de interacciones de los terminales con la red [10]. Tradicionalmente, no se ha podido sacar el máximo partido de esta información, debido a la incapacidad de poder procesarla. Con los últimos avances en tecnologías de la información, hoy en día es posible procesar grandes volúmenes de información en tiempo real. Esta capacidad justifica el interés de la industria de las comunicaciones móviles en el procesado de datos masivos, englobado bajo el término *Big Data Analytics* (BDA) [24]. En este ámbito, uno de los principales paradigmas de programación, que ha ganado notoriedad en los últimos años, es el aprendizaje automático (*Machine Learning*, ML). Tradicionalmente, el desarrollo de algoritmos implicaba programar manualmente las reglas que regían la relación esperada entre los datos de entrada y la salida. En cambio, actualmente se busca que el algoritmo desarrollado aprenda por sí mismo estas reglas, a partir de los datos que se suministran.

En el aprendizaje automático se busca dotar a las máquinas de capacidad de aprendizaje, entendiendo este como la generalización, a partir de un conjunto de experimentos, del conocimiento o la habilidad que se pretende aprender, pudiendo ser aplicado posteriormente a datos que no han sido vistos hasta el momento. Dentro de este campo

de la computación, se definen tres tipos de paradigmas para conferir dicha capacidad al sistema:

- **Aprendizaje Supervisado** [25]. El paradigma del Aprendizaje Supervisado (*Supervised Learning*, SL) se basa en descubrir la relación entre unas variables de entrada, o regresores, y una variable de salida, o respuestas, suministrando al algoritmo una gran cantidad de ejemplos. Este tipo de aprendizaje se emplea para solucionar un conjunto de problemas específicos que, dependiendo de su naturaleza, pueden diferenciarse en dos tipos distintos:
  - *Problemas de regresión.* En ocasiones, es interesante estimar el valor que toma un determinado parámetro, a priori desconocido, en función de una serie de variables con las que presenta una clara correlación. Esta es la definición de los problemas de regresión, donde se tiene una(s) variable(s) de entrada y, a partir de ejemplos, se busca determinar las relaciones inherentes (lineales o no) con la variable dependiente, para realizar una predicción del valor de esta última.
  - *Problemas de clasificación.* Al contrario que en los ejercicios de regresión, en este tipo de problemas no se busca encontrar las ecuaciones que rigen las relaciones entre las entradas y la salida, sino que se pretende determinar la clase (salida) de cada uno de los experimentos a partir de sus características (entradas del modelo).
- **Aprendizaje No Supervisado** [26]. A diferencia del SL, el Aprendizaje No Supervisado (*Unsupervised Learning*, USL) pretende descubrir el conocimiento intrínseco al conjunto de datos, sin una salida que sirva como ejemplo. Esto se consigue estudiando la estructura de los datos, permitiendo la construcción de los llamados espacios latentes, un espacio vectorial donde cada dato ocupa una posición y las separaciones entre datos determinan sus relaciones, dependiendo de la dirección, sentido y distancia. De la misma forma que en SL, el USL se aplica a diferentes tipos de problemas:
  - *Problemas de agrupamiento.* Este tipo de cuestiones pueden considerarse como un equivalente a los problemas de clasificación comentados anteriormente. La diferencia reside en que no se tiene un conjunto de medidas etiquetadas para determinar la clase, sino que se exploran las concentraciones naturales de las muestras en un conjunto de datos, identificando los grupos característicos de los datos.

  - *Detección de anomalías.* Otro problema típico dentro del USL es la exploración de las características de los datos, pudiendo identificarse aquellos puntos dentro del conjunto que presentan un comportamiento anómalo. Esto es interesante en aplicaciones como, por ejemplo, la detección de valores extremos en el juego de datos, que lleven a otros modelos de ML a comportarse de forma errática.
  - *Problemas de asociación.* En ocasiones, es interesante explorar las relaciones entre variables dentro del mismo conjunto de datos, con el objetivo de, por ejemplo, encontrar dependencias causales ocultas entre diferentes características. Este es el objetivo de los problemas de asociación.

- **Aprendizaje por Refuerzo** [27]. El tercer paradigma dentro del campo del ML es el Aprendizaje por Refuerzo (*Reinforcement Learning*, RL). En esta forma de entrenamiento, el aprendizaje se lleva a cabo explorando todas las opciones y eligiendo aquellas que optimicen una función de recompensa definida.

Para entrenar correctamente los modelos de aprendizaje automático y poder asegurar la generalización del aprendizaje, en lugar de la memorización el juego de datos empleado para su entrenamiento (lo que se conoce como sobreajuste (*overfitting*)), el juego de datos disponible se divide aleatoriamente en tres juegos de datos distintos, cada uno de ellos en distinta proporción, descritos a continuación:

- **Juego de datos de entrenamiento**. El juego de datos de entrenamiento es el de mayor tamaño, ya que es el empleado para el ajuste de los parámetros del modelo. Usualmente se suele construir usando entre un 40 % y un 80 % del juego de datos original.

- **Juego de datos de validación**. Para controlar que el modelo no se ha sobreajustado, el juego de validación verifica el rendimiento del modelo, comparando la pérdida obtenida con la del juego de entrenamiento. Si estas pérdidas difieren mucho es probable que sea debido a un sobreajuste en el entrenamiento. El juego de validación suele ocupar entre el 20 % y el 30 % del juego de datos inicial, aunque si originalmente se dispone de un tamaño de muestra muy reducido este juego de datos no se crea, recayendo la validación del modelo en el juego de evaluación.

- **Juego de datos de evaluación**. La última etapa en el proceso de ajuste de un modelo de aprendizaje automático es la evaluación del aprendizaje, es decir, cómo

de bien ha aprendido a resolver el problema en cuestión. Para ello, es necesario que el rendimiento del modelo se evalúe con un conjunto de datos que no haya sido empleado durante el entrenamiento, por lo que usualmente se reserva entre un 20 % y un 30 % del juego de datos inicial para esta tarea.

Para contextualizar los modelos propuestos a lo largo de la tesis, a continuación se explican las diferentes técnicas que se emplean en la bibliografía para los problemas planteados, tanto los modelos más clásicos como arquitecturas más complejas, basadas en aprendizaje profundo.

2.4.1. Técnicas clásicas

La gran atención que ha atraído el aprendizaje automático estos años ha derivado en un gran abanico de técnicas y algoritmos que permiten solventar problemas de diversa naturaleza. A lo largo de esta tesis doctoral se han escogido algoritmos basados en el Aprendizaje Supervisado para desarrollar los modelos propuestos, por la mayor fiabilidad que aportan al tener los datos etiquetados. A continuación se explican dos de los conjuntos de técnicas más habituales.

Regresiones

Las regresiones son uno de los problemas más clásicos dentro del aprendizaje automático. En estos modelos, el objetivo es encontrar la ecuación que determina la relación entre las variables de entrada, o descriptores, y la variable de salida, o respuesta. Existe un sinfín de modelos de regresión pero, para resumir, se explican solo los 2 tipos de regresiones usados en esta tesis.

a) Regresión lineal

La regresión lineal es la aproximación más simple al problema, buscando encontrar la ecuación del hiperplano que modele la respuesta, $\hat{y}$, en función de los descriptores usados, X_i, como,

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N \,, \tag{2.1}$$

donde β_i es el coeficiente de regresión del predictor i, que representa su pendiente en el hiperplano de regresión. Este tipo de modelos son ampliamente usados en la bibliografía, permitiendo explorar la relación lineal entre variables que se pretenden

estudiar.

b) Regresión logística

En ocasiones, las relaciones entre los descriptores y la respuesta siguen reglas más complejas, haciendo que la regresión lineal no se adecúe correctamente a los datos. En estos casos, estudiar regresiones no lineales permite analizar este tipo de comportamientos. Un ejemplo claro es la clasificación binaria (p.ej., si se cumple o no un requisito o si un experimento es positivo o negativo) en función de una serie de características. Para este tipo de problemas se emplea la regresión logística, donde se aplica la función logística para linealizar la respuesta de la regresión, siguiendo la fórmula

$$P_y = \frac{1}{1 + e^{y_p}} \,, \tag{2.2}$$

donde P_y es la probabilidad de que la respuesta pertenezca a una de las dos clases binarias, mientras que y_p es la probabilidad linealizada. Con esta función, se aplica un modelo de regresión lineal, como

$$\hat{y_p} \approx \ln\left(\frac{P_y}{1 - P_y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N \tag{2.3}$$

Análisis de series temporales

Otra tarea clásica dentro del paradigma del aprendizaje supervisado es modelar cómo evoluciona una variable a lo largo del tiempo, prediciendo qué valores adoptará en un horizonte deseado. Este es el ámbito de aplicación del análisis de las series temporales, donde algunos de los modelos más conocidos son Holt-Winters [28] y los Estacionales AutoRregresivos Integrados de Media Móvil (*Seasonal Auto-Regressive Integrated Moving Average*, SARIMA) [29].

a) Holt-Winters

Los modelos de suavizado exponencial son uno de los acercamientos clásicos al modelado de series temporales. Estos modelos se basan en la caracterización de la serie temporal mediante el suavizado de los valores observados. Un modelo básico de series temporales aplicando suavizado exponencial es el desarrollado por Holt, que consigue modelar tanto el nivel como la tendencia de una serie temporal mediante

un doble suavizado. El modelo de Holt-Winters es una evolución de este método, en el que se consigue capturar también la estacionalidad de la serie estudiada. En este modelo, también conocido como suavizado exponencial triple, la serie temporal se define como una combinación de tres componentes: a) el nivel, l_t, b) la tendencia, b_t, y c) la estacionalidad, s_t, junto a sus parámetros de suavizado (α, β y γ, respectivamente).

Dependiendo de la forma en la que se comportan las componentes de nivel y estacionalidad, se distinguen 2 variantes del modelo:

- **Modelo aditivo**. En esta variante, la predicción de cada instante temporal se define como la suma de las componentes de nivel y tendencia en el instante anterior y la componente de estacionalidad en el correspondiente instante de tiempo del periodo anterior. El modelo aditivo se aplica cuando las componentes de tendencia y estacionalidad evolucionan de forma lineal en el tiempo. Las ecuaciones que rigen este modelo son

$$\hat{y}_{t+h} = l_t + h \cdot b_t + s_{t+h-m(k+1)} \,, \tag{2.4}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \,, \tag{2.5}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \,, \tag{2.6}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \,, \tag{2.7}$$

  donde t es el instante de tiempo actual, h es el horizonte con el que se pretende realizar la predicción, m es el periodo de la estacionalidad (p.ej., 24 para una serie horaria o 12 para una serie mensual) y k es el número de periodos hasta el instante actual t.

  En esta ecuación, la componente de tendencia viene heredada del modelo desarrollado por Holt. Por otro lado, la componente de nivel definida en el modelo de Holt es corregida por la estacionalidad introducida en este modelo.

- **Modelo multiplicativo**. A diferencia de la variante anterior, en el modelo multiplicativo se estudia el caso en que la magnitud de la componente estacional crece con el tiempo, de forma proporcional al nivel y tendencia de la serie temporal. Por tanto, la predicción del modelo multiplicativo se define como la multiplicación de la componente estacional en el instante correspondiente del periodo anterior con la suma de las componentes de nivel y tendencia del instante anterior. Este

modelo sigue las siguientes ecuaciones,

$$\hat{y}_{t+h|t} = (l_t + h \cdot b_t) s_{t+h-m(k+1)} \,, \tag{2.8}$$

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \,, \tag{2.9}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} \,, \tag{2.10}$$

$$s_t = \gamma \frac{y_t}{l_{t-1} - b_{t-1}} + (1-\gamma) s_{t-m} \,. \tag{2.11}$$

b) Modelo Estacional AutoRregresivo Integrado de Media Móvil

Otro de los modelos clásicos en el estudio de series temporales es el Estacional AutoRregresivo Integrado de Media Móvil. Al igual que el método de Holt-Winters es una ampliación del modelo Holt, los modelos SARIMA son una evolución de los AutoRregresivos Integrados de Media Móvil (ARIMA), introduciendo la componente de estacionalidad. Los modelos ARIMA son un subconjunto de las regresiones lineales, donde se usan mediciones anteriores de la propia variable de salida para estimar el futuro. Para ello, el método se construye en torno a las 3 características que le dan nombre:

- **Autorregresivo**. Como se ha comentado anteriormente, los modelos ARIMA usan las mediciones anteriores de la propia variable para estimar el futuro. Este es el concepto de autorregresión, siguiendo la ecuación

$$\phi_p(B) = (1 - \phi_1(B) - \phi_2(B^2) - \ldots - \phi_P(B^p)) \,, \tag{2.12}$$

  donde ϕ_i es el operador de autorregresión en el instante $i \in [0, p]$, B^i es el coeficiente de regresión i de las medidas anteriores y p es el orden de autorregresión, indicando el número de muestras anteriores que se usan para predecir el futuro.

- **Integrado**. El modelo es integrado porque las predicciones no se hacen respecto a las mediciones de la variable, sino a la diferencia entre mediciones consecutivas, como

$$\nabla^d = (1-B)^d \,, \tag{2.13}$$

  donde ∇^d es el operador diferencial de medidas consecutivas en un orden de integración d.

- **Media móvil**. La última componente de los modelos ARIMA es la media móvil.

En las regresiones lineales, se encuentra la recta que mejor define una nube de medidas, por lo que para cada una de ellas se comete un error, o residuo, ϵ. La componente de media móvil pretende modelar este error, mediante una regresión de estos residuos de la forma

$$\theta_q(B) = (1 - \theta_1(B) - \theta_2(B^2) - \ldots - \theta_q(B^q)) \,, \tag{2.14}$$

donde θ_e es el operador de media móvil en el instante $e \in [0, q]$ de los errores y q es el orden de media móvil, determinando el número de residuos usados en la estimación.

De esta forma, la predicción realizada por un modelo ARIMA se realiza como

$$\phi_p(B)\, \nabla^d y_t = \theta_q(B)\epsilon_t \,. \tag{2.15}$$

El modelo ARIMA descrito (2.15) no es capaz de capturar la estacionalidad de una serie temporal. Esto se soluciona en los modelos SARIMA, siguiendo la misma filosofía para ajustar la periodicidad, introduciendo los parámetros P, D, Q, siendo los órdenes de autoregresión, integración y media móvil de la estacionalidad, respectivamente, además del periodo s. Las ecuaciones estacionales, equivalentes a 2.12, 2.13 y 2.14, son

$$\Phi_P(B^s) = (1 - \Phi_1(B^s) - \Phi_2(B^{2s}) - \ldots - \Phi_P(B^{Ps})) \,, \tag{2.16}$$

$$\nabla_s^D = (1 - B^s)^D \,, \tag{2.17}$$

$$\Theta_Q(B^s) = (1 - \Theta_1(B^s) - \Theta_2(B^{2s}) - \ldots - \Theta_Q(B^{Ps})) \,, \tag{2.18}$$

donde Φ, ∇_s y Θ son los operadores de autoregresión, integración y media móvil con una separación estacional de periodo s, respectivamente. Una vez definidas todas las componentes, el modelo SARIMA completo se compone de la siguiente forma,

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d y_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t \,. \tag{2.19}$$

2.4.2. Redes neuronales

Las llamadas redes neuronales artificiales (*Artificial Neural Networks*, ANN) están cobrando un gran interés dentro del aprendizaje automático en los últimos años, donde se aplican modelos de arquitectura de aprendizaje jerárquico mediante la concatenación de capas.

Las estructuras clásicas de ANN están compuestas por varias unidades de computación no lineal (comúnmente denominadas "neuronas") distribuidas en capas consecutivas y conectadas entre sí, permitiendo a la red modelar una gran variedad de comportamientos muy complejos. Estas unidades de computación pueden ser de diferentes tipos, en función del problema que se pretenda abordar, pudiendo combinar varios de ellos en estructuras más complejas.

Estas arquitecturas se suelen definir como prealimentadas, o *feed-forward*, donde los datos fluyen consecutivamente de una capa a la siguiente, y sus parámetros internos se suelen ajustar mediante algoritmos de aprendizaje supervisado. Así, para entrenar este tipo de redes neuronales, se aplica el método de retropropagación [30] para optimizar una función de pérdida (p.ej., el Error Cuadrático Medio -MSE-, el Error Absoluto Medio -MAE-, el Error Porcentual Absoluto Medio -MAPE-...) mediante el algoritmo de descenso de gradiente durante un número épocas predefinido o hasta que la función de pérdida se estabiliza en el mínimo global.

Una época se define como el ciclo en el que el modelo ha procesado todo el juego de entrenamiento una vez, es decir, una iteración completa a través de todo el juego de datos de entrenamiento. Durante cada época, el juego de entrenamiento se divide en lotes (*batches*) más pequeños, que son procesados individualmente, reajustando los parámetros del modelo después de cada uno de los lotes. De esta forma, se consigue mejorar la eficiencia computacional durante el entrenamiento, aliviando la carga en memoria de los datos, y ayudar a evitar mínimos locales en la función de pérdida. Tanto el número de épocas necesario como el tamaño de los lotes depende de la complejidad del modelo, así como del tamaño del juego de datos disponible.

Debido a la creciente cantidad de datos disponibles, así como al gran rendimiento obtenido, los modelos ANN han ido evolucionando, creando el campo del aprendizaje profundo (*Deep Learning*, DeL), una de las ramas más novedosas dentro del aprendizaje automático. En los modelos de DeL, la profundidad otorgada por la concatenación de numerosas capas de distinta naturaleza permite el modelado de problemas mucho más complejos, aprovechando al máximo el potencial de los datos recopilados.

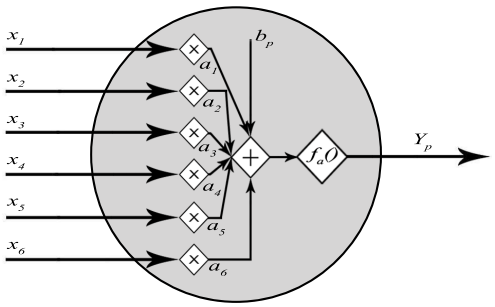


Figura 2.2: Perceptrón.

A continuación, se describen las arquitecturas básicas que pueden conformar una ANN, tras lo cual se describe una de las técnicas de aprendizaje profundo empleada en esta tesis.

a) Perceptrón multicapa

La estructura más común para abordar los problemas de regresión en el ámbito del aprendizaje profundo es el perceptrón multicapa (*Multi-Layer Perceptron*, MLP) [31]. El elemento fundamental de las estructuras MLP es la neurona, o perceptrón, mostrado en la Fig. 2.2. Un perceptrón tiene un número N de entradas, $x_i, i \in [1, N]$, y proporciona una salida Y_p realizando una suma ponderada de las entradas con un parámetro constante, llamado sesgo. En el proceso de entrenamiento, un perceptrón aprende los valores de estos pesos, $a_i, i \in [1, N]$, y el sesgo, b_p, que mejor se ajustan al conjunto de entrenamiento. Con esta definición, cada perceptrón se comporta como una regresión multivariante por lo que, para evitar que todas las neuronas de un modelo MLP colapsen en una única regresión multivariante, se modifica la salida de cada perceptrón para que pierda su comportamiento lineal, utilizando una función de activación f_a (p.ej., *Regularized Linear Unit* -ReLU-, sigmoide...) como

$$Y_p = f_a \left(b_p + \sum_{i \in [1,N]} x_i a_i \right) . \tag{2.20}$$

La arquitectura general de un MLP, mostrada en la Fig. 2.3 se compone de la capa de entrada, formada por los valores de entrada utilizados para calcular la salida de la red, la capa de salida, formada por las neuronas que calculan las salidas de la red, y, opcionalmente, una o más capas ocultas, que realizan los cálculos intermedios para

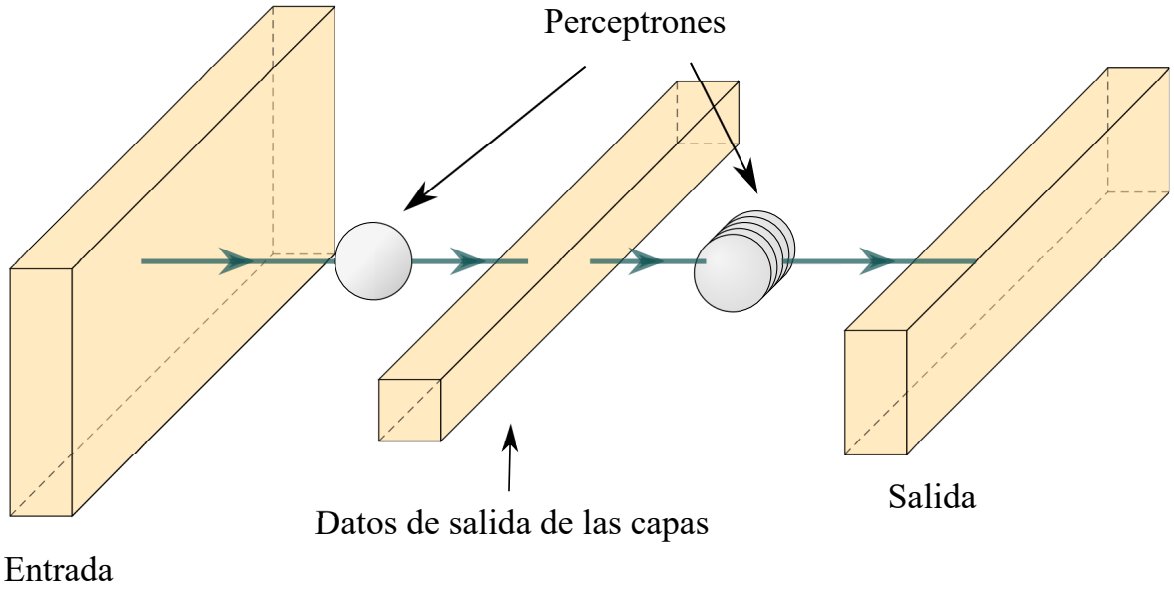


Figura 2.3: Ejemplo de perceptrón multicapa.

llevar a cabo operaciones más complejas.

b) Redes convolucionales

Otra estructura habitual en el aprendizaje profundo es la red neuronal convolucional (*Convolutional Neural Network*, CNN) [32]. Esta arquitectura introduce una nueva unidad computacional, la neurona convolucional (también conocida como filtro, kernel o mapa de activación), mostrada en la Fig. 2.4. A diferencia del perceptrón, esta neurona no tiene un peso para cada entrada, sino que entrena una ventana k_c, utilizada para convolucionar la entrada en diferentes dimensiones (p.ej., la dimensión temporal en una serie de tiempo o las dimensiones horizontal y vertical en una imagen). Así, a diferencia del valor único devuelto por el perceptrón, la salida Y_c del filtro es una matriz con la información dimensional relevante comprimida, calculada como

$$Y_c = f_a \left(b_c + \sum_{i \in [1, N_x]} x_i * k_c \right) . \tag{2.21}$$

La estructura de la CNN está formada por una o más capas ocultas compuestas por varias neuronas de este tipo, denominadas capas convolucionales, seguidas de una capa de regresión construida con perceptrones, tal como se muestra en la Fig. 2.5. Las capas convolucionales pretenden abarcar la entrada, manteniendo la información relevante en menos datos y, por tanto, reduciendo la cantidad de parámetros de entrenamiento en

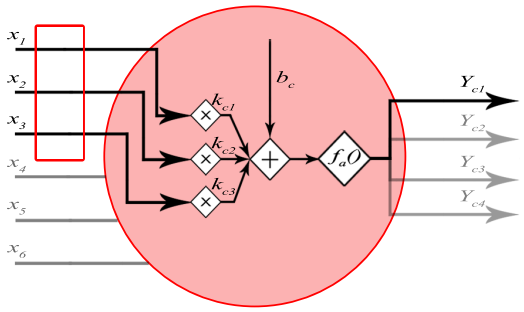


Figura 2.4: Neurona convolucional.

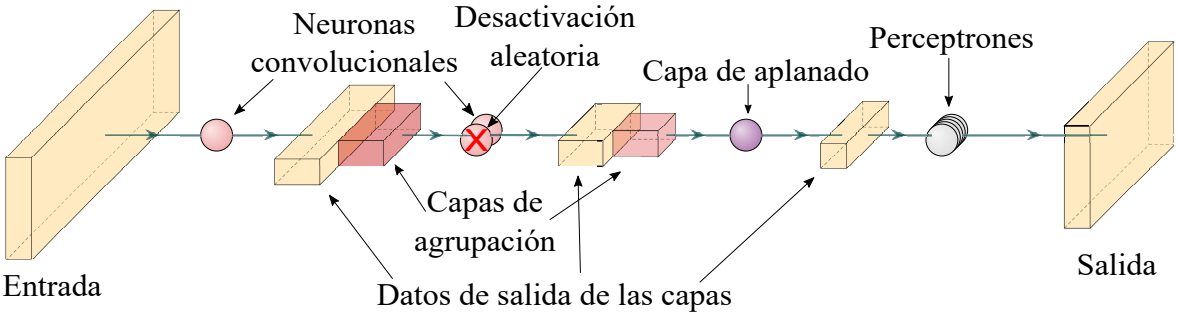


Figura 2.5: Ejemplo de red neuronal convolucional.

la capa de regresión. Sin embargo, esta estructura presenta problemas de sobreajuste con juegos limitados de datos. Para evitarlo, existe un conjunto de capas intermedias, como las capas de agrupación (*pooling*) [33] o desactivación aleatoria de neuronas (*dropout*) [34], que se sitúan entre las capas convolucionales para ayudar a prevenir este sobreajuste. Debido a la diferencia entre las dimensiones con las que trabajan distintos tipos de neuronas, en ocasiones es necesario reajustar la salida de una capa para adaptarla a la siguiente, aplicando capas como, por ejemplo, de aplanado (*flatten*) [35].

Una variación de este tipo de arquitectura son los denominados autocodificadores (o *Autoencoders*, AE) [36]. Un AE es una red de tipo *feed-forward* que aprende una codificación eficiente de su entrada. Los AE explotan las operaciones no lineales para aprender representaciones de baja dimensión con un error menor que los esquemas de reducción de características lineales (p.ej., el análisis de componentes principales).

La primera parte, llamada ruta de contracción, realiza la operación de convolución a través de una red convolucional tradicional. La segunda parte, llamada ruta de expansión, reconstruye la salida a partir de la información recibida por la ruta de contracción. Para ello, se apoya en las neuronas convolucionales traspuestas, que realizan

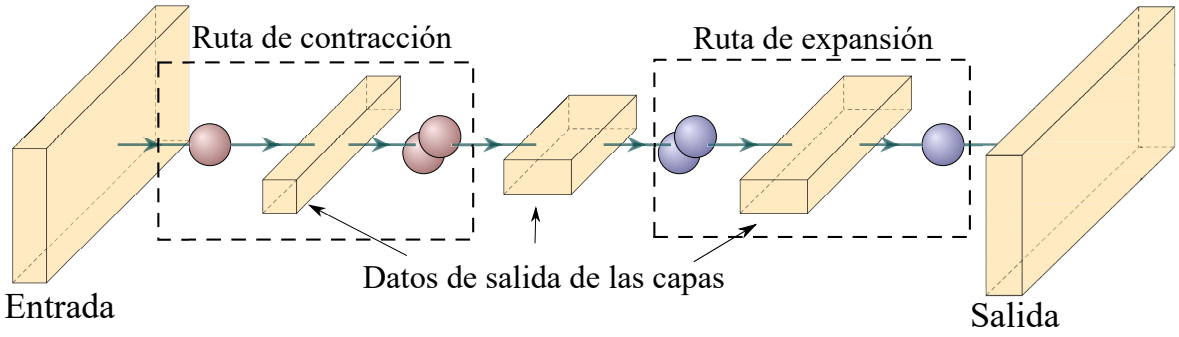


Figura 2.6: Ejemplo de red autocodificadora.

una transformación en la dirección opuesta a la sentido de la convolución, realizando un muestreo ascendente de la entrada a través de una ventana, k_{tc}, para obtener una salida Y_{tc} de mayor tamaño, como

$$Y_{tc} = f_a \left(b_{tc} + \sum_{i \in [1,N]} x_i *' k_{tc} \right) , \tag{2.22}$$

donde *' es el operador de convolución traspuesta. La Fig. 2.6 ejemplifica una arquitectura de AE simple, con 2 capas de 1 y 2 neuronas convolucionales respectivamente en la etapa de contracción y otras 2 capas de 2 y 1 neuronas respectivamente en la etapa de expansión.

Al igual que la arquitectura MLP, las estructuras CNN son *feed-forward* y se clasifican como algoritmos de aprendizaje supervisado, por lo que se emplea el mismo procedimiento de retropropagación para optimizar la función de pérdida seleccionada en la etapa de entrenamiento.

c) Redes recurrentes

El tercer tipo de estructura clásica en redes neuronales artificiales son las redes neuronales recurrentes (*Recurrent Neural Network*, RNN). En estas arquitecturas, el elemento principal es la neurona recurrente, cuya particularidad respecto al perceptrón es que se retroalimenta, añadiendo a las entradas la predicción anterior y otorgando así una memoria a la neurona, siendo especialmente interesante para modelar datos dependientes de una dimensión (p.ej., series temporales).

Las neuronas recurrentes tradicionales son capaces de aprender las dependencias a corto plazo, pero no pueden modelar las relaciones a largo plazo [37]. Esta limitación se resuelve en las neuronas *Long Short-Term Memory* (LSTM) [38], también conocidas

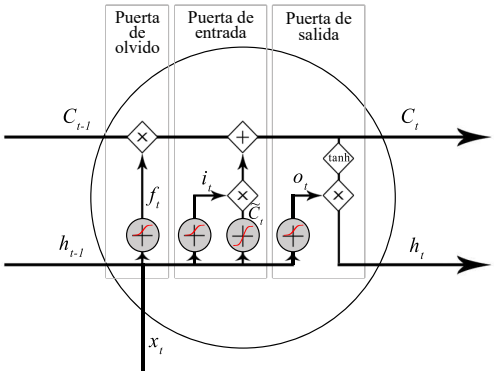


Figura 2.7: Neurona LSTM.

como células de memoria, mostradas en la Fig.2.7, introduciendo el estado de la célula, C_t. Este estado modela la "memoria"de la neurona, que define la relevancia de la entrada, x_t, y el estado oculto anterior de la neurona (también conocido como vector de salida), h_{t-1}, en el estado oculto actual, h_t. El proceso completo de la neurona LSTM viene determinado por tres puertas:

- Puerta de olvido. En esta puerta, la neurona LSTM fija qué información almacenada en el estado de la neurona se descarta. Para ello, se utiliza la etapa de olvido, compuesta por una capa de perceptrón con función de activación sigmoide, σ. La puerta de olvido toma como entrada x_t y h_{t-1} y da como salida el vector de activación de la puerta de olvido, f_t, que libera la relevancia del estado anterior, C'_{t-1}, como

$$f_t = \sigma\left(b_{f_t} + W_{f_t}[h_{t-1}, x_t]\right) \ , \tag{2.23}$$

  donde W_{f_t} y b_{f_t} son la matriz de pesos y el sesgo de la capa de olvido. De esta manera, cada valor del estado de la neurona anterior se pondera en función de su relevancia en el siguiente estado.

- Puerta de entrada. En este paso se calcula el nuevo estado de la neurona. En primer lugar, se calcula un estado preliminar, $\tilde{C}_t$, sobre la base de x_t y h_{t-1}, empleando una capa de perceptrón con tangente hiperbólica, $tanh$, como función de activación, como

$$\tilde{C}_t = \tanh\left(b_{\tilde{C}_t} + W_{\tilde{C}_t}[h_{t-1}, x_t]\right) \ , \tag{2.24}$$

donde $W_{\tilde{C}_t}$ y $b_{\tilde{C}_t}$ corresponden a la matriz de pesos y al sesgo de la capa de estado de la neurona preliminar. A continuación, el vector de activación de una puerta de entrada, i_t, se obtiene de x_t y h_{t-1} para ponderar el impacto de $\tilde{C}_t$ en el nuevo estado, como

$$i_t = \sigma\left(b_{i_t} + W_{i_t}[h_{t-1}, x_t]\right) , \tag{2.25}$$

donde W_{i_t} y b_{i_t} son, respectivamente, la matriz de pesos y el sesgo de la capa de entrada. Por último, el estado anterior y el estado preliminar de la neurona se combinan para obtener el estado actual real de la neurona, calculando una multiplicación puntual de cada estado con su vector de activación, como

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t . \tag{2.26}$$

- Puerta de salida. Por último, la salida de la neurona, h_t, se calcula en la puerta de salida. Para ello, el vector de activación de la puerta de salida, o_t, se obtiene de la misma forma que i_t, como

$$o_t = \sigma\left(b_{o_t} + W_{o_t}[h_{t-1}, x_t]\right) , \tag{2.27}$$

donde W_{o_t} y b_{o_t} representan la matriz de pesos y el sesgo de la capa de salida. Este vector se utiliza para ponderar el resultado de realizar una operación de tangente hiperbólica puntual sobre C_t para obtener h_t, como

$$h_t = o_t \tanh(C_t) . \tag{2.28}$$

La red LSTM básica consta de una sola capa oculta, con varias neuronas LSTM para modelar múltiples estados ocultos, y una capa de salida, que consiste en una capa totalmente conectada de perceptrones para derivar el valor o valores previstos. Además, se pueden apilar múltiples capas ocultas en una estructura profunda en la que cada capa proporciona una nueva secuencia a la siguiente capa [39]. Estas capas adicionales pueden utilizarse para representar el problema en diferentes escalas de tiempo [40]. La Fig. 2.8 presenta un ejemplo de estructura de LSTM con varias capas ocultas.

Las arquitecturas RNN convencionales son sensibles a los problemas de desvanecimiento del gradiente [41]. Para resolverlo, las arquitecturas LSTM incluyen una unidad

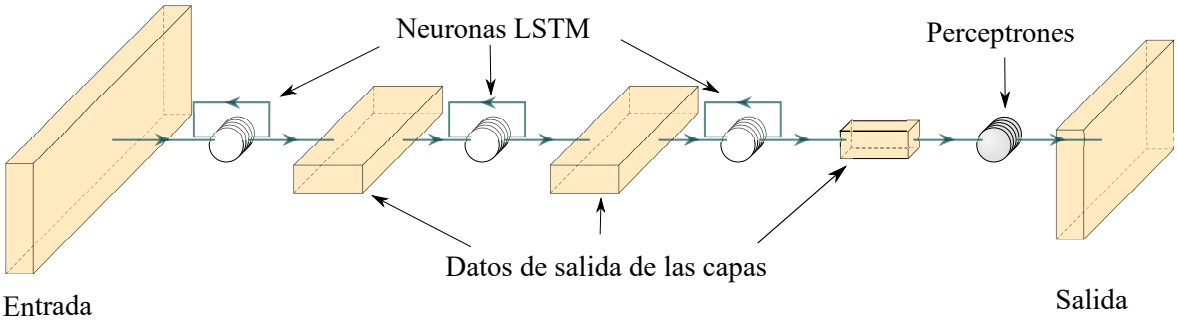


Figura 2.8: Ejemplo de LSTM apilada.

lineal con una autoconexión de peso fijo en cada celda de memoria [42], que controla el flujo de error en el descenso de gradiente.

d) Arquitecturas multitarea

En las arquitecturas descritas hasta ahora, se suele emplear un solo modelo para resolver un problema determinado. Por ejemplo, si se pretende predecir el comportamiento futuro de una serie temporal, suele ser buena idea recurrir a arquitecturas basadas en RNN, mientras que si se busca resolver problemas de regresión la mejor opción suele ser una arquitectura basada en MLP. En cambio, hay ocasiones en las que el problema que se busca modelar es muy complejo, influenciado por múltiples factores que afectan significativamente al rendimiento de un solo modelo, dificultando drásticamente su entrenamiento. En este contexto nacen las arquitecturas de aprendizaje multitarea (*MultiTask Learning*, MTL) [43]. Las arquitecturas MTL abordan este tipo de problemas dedicando una estructura ANN para modelar cada factor (o tarea) por separado. Entrenando conjuntamente esta combinación de arquitecturas ANN, el aprendizaje de cada una de las tareas individuales influye en el aprendizaje del resto de la arquitectura MTL, llegando a modelar patrones mucho más complejos que si se entrenaran por separado. De esta forma, se consigue una representación individual de cada tarea dentro del modelo, permitiendo explotar la relación entre ellas para mejorar la generalización en el aprendizaje del problema que se buscaba modelar desde un principio.

Existen varias aproximaciones para la implementación de MTL en arquitecturas de redes neuronales profundas, que pueden dividirse en tres grandes grupos: a) el modelado simultáneo de un conjunto de características a partir de una misma entrada, b) el entrenamiento de varias arquitecturas simultáneas cuyos aprendizajes se afectan entre sí y c) el modelado de un solo problema complejo afectado por varias tareas

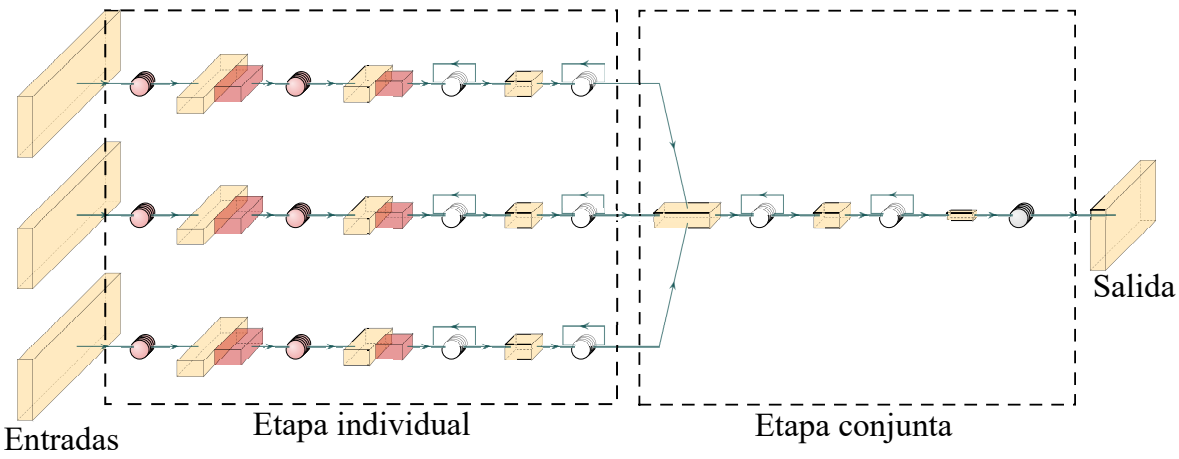


Figura 2.9: Ejemplo de estructura para aprendizaje multitarea.

definidas como entradas. En esta tesis, se ha utilizado esta última aproximación.

En ese tipo de arquitecturas, se definen dos etapas diferentes [44]. En la primera etapa, destinada al aprendizaje individual de cada una de las variables de entrada, se aplican tantas redes neuronales como tareas independientes estudiadas, cada una de ellas entrenadas por separado. En la segunda etapa, donde se modela la influencia de las entradas en cada una de las tareas de salida, las salidas de la etapa individual se combinan en un aprendizaje conjunto para alimentar tantas redes neuronales como características a modelar. Para ilustrar este tipo de arquitecturas, la Fig. 2.9 presenta un ejemplo con 3 tareas independientes y una única salida. En este ejemplo, la etapa individual está compuesta por las 3 tareas independientes, cada una de ellas modelada con la misma arquitectura de procesado. Las salidas de las tres tareas se concatenan para formar la entrada de la etapa conjunta, que tiene una arquitectura diferente y es responsable de obtener la salida del modelo completo.

Un claro caso de aplicación de este tipo de arquitecturas es la predicción de series temporales de indicadores de rendimiento en una celda. En una red celular, los indicadores de las celdas vecinas tienen una clara correlación entre sí, por lo que predecir de forma conjunta las series temporales de los indicadores de celdas vecinas permite optimizar el aprendizaje del modelo y mejorar así la predicción. Como cada celda puede tener características distintas, es interesante modelarlas por separado para combinarlas posteriormente en la predicción final, para lo que las arquitecturas multitarea son especialmente adecuadas.

Capítulo 3

Clasificación probabilística de conexiones de interior

Como se ha comentado anteriormente, el objetivo principal de esta tesis es desarrollar algoritmos que aprovechen la información de dónde y/o cuándo se distribuyen los usuarios en una red móvil. En la mayoría de los casos, esta información no se registra, por lo que es necesario inferirla. En este capítulo se plantea el problema de saber si una conexión ocurrió en zonas de interior o en espacio abierto. Para solucionar este problema, se presenta un algoritmo que, partiendo de los descriptores de una conexión registrados en el sistema de gestión de red, deduce un valor de probabilidad de que la conexión se haya generado en interior.

Para explicar el modelo, este capítulo se compone de cinco secciones. La sección 3.1 revisa los trabajos previos más relevantes en el ámbito de la estimación del contexto de usuarios, focalizándose en la discriminación del entorno de interior/exterior en el que se generan las peticiones de servicio de una red celular. Seguidamente, en la sección 3.2 se formula el problema de conocer la información del contexto del usuario sin datos externos a la red. A continuación, la sección 3.3 presenta la primera contribución de esta tesis, el algoritmo propuesto para clasificar de forma probabilística cada conexión como de interior a partir de sus indicadores de red. La sección 3.4 muestra el escenario usado para su desarrollo, así como los resultados obtenidos y la carga computacional que ha conllevado su obtención. Finalmente, en la sección 3.5 se exponen las principales conclusiones de este capítulo.

3.1. Estado de la técnica

El entorno en el que tiene lugar la comunicación influye mucho en las expectativas del usuario. Por ello, estimar el contexto de una sesión es clave para evaluar la experiencia del usuario en la gestión de redes móviles. Algunos ejemplos de factores contextuales son el tipo de terminal, la edad, el sexo, las experiencias previas, la hora del día y la ubicación [45]. En particular, la localización del usuario juega un papel fundamental, ya que determina los servicios demandados, las condiciones del enlace o el rendimiento esperado del servicio [8]. Así, una estimación precisa del entorno del usuario es fundamental para evaluar su experiencia.

Con las nuevas aplicaciones conscientes del contexto, el interés por detectar la actividad de un usuario en interiores ha crecido considerablemente. Esta información puede utilizarse para mejorar la usabilidad, cambiando los ajustes de la interfaz (p.ej., el brillo de la pantalla, el volumen del audio...), ahorrar energía apagando los sensores innecesarios (p.ej., receptor de satélite en interiores, Wi-Fi en exteriores) o mejorar el posicionamiento seleccionando el mejor método de localización (p.ej., A-GNSS en exteriores, huella digital basada en Wi-Fi en interiores) [46].

Los enfoques tradicionales para la detección en interiores se basan en el posicionamiento del usuario con la suficiente precisión como para localizarlo dentro de un edificio en el mapa. Sin embargo, el posicionamiento A-GNSS no es preciso en interiores debido a la debilidad de la señal de los satélites, y los métodos basados en Wi-Fi han demostrado ser poco fiables para aislar los entornos interiores y exteriores [47]. Como alternativa, un método simple, pero inexacto, para aislar la actividad en interiores se basa en la detección de la falta de señal GNSS [48].

La integración de múltiples sensores en los smartphones ha aportado mejores métodos para esta tarea. Un primer grupo de esquemas basados en la detección continua mejora la precisión del posicionamiento combinando las señales de los receptores A-GNSS y los sensores inerciales (p.ej., acelerómetro, brújula y/o giroscopio) [49, 50, 51]. Un segundo grupo de esquemas analizan el entorno con sensores alternativos, como un sensor de temperatura [52], una cámara [53], un sensor geomagnético [54], un sensor de luz [55] o un micrófono [56]. También se han propuesto métodos híbridos para combinar la información de diferentes receptores y sensores en el terminal. En [55], la detección en interiores se lleva a cabo mediante un modelo de cadena de Markov oculta que agrega la información de un sensor de nivel de luz, un sensor geomagnético y mediciones de la potencia de señal recibida (*Received Signal Strength*, RSS). Un método similar se

propone en [57] para la detección en interiores a partir de mediciones del acelerómetro y del RSS celular/Wi-Fi.

La agregación de datos puede realizarse de forma más eficaz mediante el aprendizaje automático [9, 55, 58]. En [9], el reconocimiento de la actividad se lleva a cabo combinando los datos de varios sensores en el cuerpo mediante el aprendizaje con métodos combinados. En [58], se prueban varios clasificadores de ML para aislar entornos exteriores abiertos, semiexteriores, interiores ligeros e interiores profundos a partir de mediciones RSS de cuatro estaciones base de telefonía móvil. Con un alcance más limitado, se propone un algoritmo de ML en [59] para detectar si un usuario está dentro o fuera de un edificio concreto a partir de las mediciones RSS de Wi-Fi. Todos estos métodos, basados en aprendizaje supervisado, dependen de la disponibilidad de un gran conjunto de casos etiquetados para entrenar y validar el modelo de clasificación.

En [46], se muestra que adaptar el clasificador usado a nuevos entornos y dispositivos es esencial para lograr una detección robusta en interiores. Para ello, se propone un algoritmo de aprendizaje semi-supervisado basado en el co-entrenamiento [60] que utiliza dos clasificadores entrenados con datos no etiquetados de diferentes sensores para mejorar la detección en interiores.

Algo importante a destacar es que muchos de los métodos anteriores utilizan datos a los que sólo puede acceder desde la capa de aplicación, por lo que no están disponibles para el operador de la red. Otros modelos se basan en medidas de nivel de señal, obviando la información de comportamiento del usuario, que puede variar en función de su entorno. En este sentido, no se ha encontrado ningún estudio anterior que haya comprobado si se puede detectar la actividad en interiores a partir de los descriptores de tráfico de una llamada recogidos en la interfaz radio.

3.2. Formulación del problema

Como se explicó en el ***Capítulo 2***, los problemas de clasificación se engloban en el campo del Aprendizaje Supervisado. En este caso, a diferencia del Aprendizaje No Supervisado, se tiene un conjunto de medidas (también llamados puntos de datos) etiquetadas, que se utilizan para identificar las reglas internas que relacionan cada punto de datos con su etiqueta. Lamentablemente, el sistema de gestión de las redes celulares no registra la información relacionada con el entorno del usuario, por lo que no existe un juego de datos que pueda usarse como referencia en el entrenamiento de

los modelos de SL.

Para solventar este problema, en esta tesis se recurre a un paradigma que nace como combinación de SL y USL, llamado Aprendizaje Semi-Supervisado (*Semi-Supervised Learning*, SSL) [60]. En este tipo de aprendizaje, se usan tanto datos etiquetados como no etiquetados para mejorar el aprendizaje del algoritmo.

En este capítulo, se plantea un modelo de aprendizaje automático, donde las entradas son una serie de indicadores obtenidos de las trazas y la salida es la probabilidad de que una conexión haya ocurrido en interior. Con este objetivo, se usa SSL para entrenar el modelo con los indicadores seleccionados agregados a nivel de celda, generando un juego de datos con tantos puntos de datos como celdas tenga el escenario. Para solucionar el problema del etiquetado de datos, el juego de datos se etiqueta artificialmente con un ratio de conexiones en interior, computado a partir de los tipos de uso de suelo en el área de cobertura de la celda. Debido a este proceso de etiquetado artificial, necesario para entrenar el modelo propuesto, el tamaño del juego de datos disponible es muy limitado, impidiendo el correcto entrenamiento de modelos complejos, como los basados en redes neuronales. Por ello, se propone un modelo basado en una regresión logística multivariable. Una vez entrenado, este modelo se explotará a nivel de conexión (usuario), determinando la probabilidad de que cada conexión haya ocurrido en interior usando los coeficientes correspondientes a la celda servidora.

Con este enfoque, se espera que el modelo aprenda las relaciones globales entre los diferentes indicadores y el contexto de interior, siendo capaz de modelar el problema como un hiperplano, donde cada indicador influye linealmente y de forma independiente en la estimación del contexto. Una vez aprendidas estas relaciones globales a nivel de celda, la explotación del modelo a nivel de conexión busca aprovechar la variabilidad de los valores de los indicadores por usuario para identificar su contexto. De esta forma, se espera que el ajuste a nivel de celda obtenga unos ratios de conexión de usuarios de interior muy cercanos a los calculados con los tipos de uso de suelo, mientras que la explotación a nivel de conexión otorgue, para conexiones ocurridas en la misma celda, probabilidades de interior muy dispares, pudiendo llegar a detectar conexiones totalmente de interior o de exterior. Para obtener estos resultados, es necesario que todas las celdas del escenario tengan perfiles similares, es decir, que los usuarios de interior y de exterior se comporten de igual forma en la red. De no ser así, es necesario detectar los diferentes patrones de celdas presentes en la red y aplicar un modelo distinto a cada uno.

Tabla 3.1: Descriptores de conexión usados en el modelo completo.

Tiempo de conexión RRC	[ms]
Volumen de datos en el enlace descendente	[bytes]
Ratio de dirección de enlace	[-]
Ratio de continuidad de la transmisión en el enlace descendente	[-]
Ratio de continuidad de la transmisión en el enlace ascendente	[-]
Ratio de volumen en el último TTI en el enlace descendente	[-]
Throughput en el enlace descendente	[kbps]
Throughput en el enlace ascendente	[kbps]
RSRP medio	[dBm]
Error de estimación de potencia	[dB]
Diferencia de RSRP entre la celda servidora y la 1ª vecina	[dB]
Número de celdas vecinas con RSRP > -108 dBm	[-]
Número de celdas vecinas con una diferencia de RSRP con la celda servidora < 10 dB	-]
RSRP - $\overline{\text{RSRP}}$ (anillo)	[dB]
RSRP - $\overline{\text{RSRP}}$ (celda)	[dB]
Porcentaje de medidas limitadas en potencia en el canal ascendente	[%]
CQI medio	[-]
CQI medio en Rango 1	[-]
CQI medio en Rango 2	[-]
SINR en el canal PUSCH	[dB]
SINR en el canal PUCCH	[dB]
Eficiencia espectral en el enlace descendente	[bits/RE]
Eficiencia espectral en el enlace ascendente	[bits/RE]
Porcentaje de medidas en Rango 2	[%]
Porcentaje de medidas en Modo de Transmisión 3, Rango 1	[%]
Porcentaje de medidas en Modo de Transmisión 3, Rango 2	[%]

A continuación se aporta un breve resumen de la información disponible de la red, tras lo cual se describe en mayor detalle el proceso de etiquetado.

3.2.1. Datos del sistema de gestión de red

En el modelo propuesto, la fuente de información son las trazas de conexión, a partir de las que se calculan varios descriptores de rendimiento de las trazas (*Key Performance Indicators*,KPI), que reflejan el comportamiento de los usuarios y los elementos de red en una conexión. La Tabla 3.1 resume el conjunto de KPI preseleccionados a priori para la realización del estudio completo. A continuación se incluye una breve descripción de los menos intuitivos. Salvo que se indique lo contrario, todos los indicadores se calculan tanto en el enlace ascendente (*Uplink*, UL) como en el enlace descendente (*Downlink*, DL).

1. **Ratio de continuidad de la transmisión.** Este indicador refleja el grado de continuidad (es decir, la ausencia de ráfagas) de la transmisión de datos en una conexión k, como

$$R_{cont}(k) = \frac{T_{act}(k)}{T_{RRC}(k)} \, , \tag{3.1}$$

donde $T_{act}(k)$ es el tiempo en el que había datos en el búfer para transmitir y $T_{RRC}(k)$ es el tiempo de conexión radio (*Radio Resource Control*, RRC) en la conexión k, ambos en milisegundos. En dicha expresión, se aprecia cómo la relación de continuidad es cercana a 1 cuando la transmisión no tiene interrupciones, o cercana a 0 cuando la transmisión es extremadamente a ráfagas.

2. **Ratio de volumen en el último intervalo de transmisión.** El volumen de datos generado en una conexión es, en muchos casos, demasiado grande para ser transmitido por completo con los recursos disponibles en la red, por lo que es necesario dividirlo en segmentos, cada uno de ellos transmitido en un Intervalo Temporal de Transmisión (*Time Transmission Interval*, TTI). Un último TTI es aquel en el que se vacía el búfer de transmisión del usuario [61]. Obsérvese que una conexión con una única petición solo llena el búfer una vez, por lo que solo tiene un último TTI, mientras que, en una conexión con datos en ráfaga, el búfer se llena y vacía varias veces a lo largo de la conexión y, por lo tanto, el número de últimos TTI es mayor. En consecuencia, el ratio de volumen del último TTI se define como

$$R_{Vol_{ultTTI}}(k) = \frac{Vol_{ultTTI}(k)}{Vol_{total}(k)} \, , \tag{3.2}$$

donde $Vol_{ultTTI}(k)$ es el volumen de datos transmitido en los últimos TTI y $Vol_{total}(k)$ es el volumen total de datos transmitidos en la conexión k. Este ratio da una pista sobre la discontinuidad de la transmisión. Estudiando los casos extremos, aquellas conexiones en las que un único TTI es suficiente para vaciar el búfer, como en la señalización o la mensajería instantánea, mostrarán valores de $R_{Vol_{ultTTI}}$ cercanos a 1, mientras que si el volumen de datos de la conexión es muy alto, el ratio tendrá valores próximos a 0.

3. **Ratio de dirección del enlace.** Este ratio da información sobre la dirección principal de la comunicación, calculando

$$R_{enlace}(k) = \frac{Vol_{UL}(k)}{Vol_{UL}(k) + Vol_{DL}(k)} , \tag{3.3}$$

donde $Vol_{UL}(k)$ y $Vol_{DL}(k)$ son el volumen de datos a nivel PDCP en UL y DL, respectivamente. Por lo tanto, esta relación tiende a 1 si la mayoría de los datos se transmiten en UL, y a 0 si prevalece en DL.

4. **Error de estimación de potencia.** La propagación de la señal emitida por cada estación base se ve muy influenciada por las características del entorno en el que se propaga. Tradicionalmente, las tareas de planificación se han apoyado en modelos de propagación, como Okumura-Hata, para estimar la potencia recibida por los usuarios en el terreno. Estos modelos obtienen buenos resultados en exterior, pero la mayoría fallan con la atenuación o reflexiones de la señal en el interior de los edificios, por lo que el error cometido puede dar pistas de la ubicación del usuario. La medida de error usada en esta tesis se define como la diferencia entre la potencia recibida de la señal de referencia (RSRP) medida por el terminal y la estimada por el modelo Okumura-Hata, como

$$E_{potencia}(k) = RSRP_{medida}(k) - RSRP_{estimada}(d(k)) , \tag{3.4}$$

donde $RSRP_{medida}$ es el RSRP comunicado por el terminal y $RSRP_{estimada}(d(k))$ es el estimado a partir de la distancia a la estación base servidora, d_k, derivado de las medidas de TA para la conexión k.

3.2.2. Datos de uso de suelo

Además de los datos sobre la actividad de los usuarios, el modelo descrito en este capítulo necesita información sobre el entorno en el que se sitúan los usuarios. Algunas iniciativas de datos abiertos promovidas por instituciones gubernamentales (como ayuntamientos o gobiernos autonómicos), así como plataformas populares de *crowdsourcing* (p.ej., OpenStreetMap [62]), recopilan información sobre la actividad socioeconómica de cada área del terreno con un gran nivel de detalle, aportando información sobre el contexto de cada zona. Esta información es de acceso público, por lo que los operadores de red pueden obtenerla fácilmente para aplicarla en los procesos de gestión de red.

Para simplificar la información contextual, se ha generado un conjunto de clases

Tabla 3.2: Descripción de los tipos de uso de suelo.

Clase	Descripción	Interior/Exterior
Servicios	Áreas destinadas a servicios institucionales, culturales, médicos o académicos.	Interior
Oficinas	Edificios de oficinas destinados a servicios gestión, información o trabajo.	Interior
Uso mixto	Zonas sin un uso de suelo definido o mezcla de varios tipos de uso de suelo.	Interior
Residencial	Áreas de casas, hoteles y servicios para visitantes.	Interior
Entretenimiento	Zonas de comercios y entretenimiento.	Interior
Industrial	Áreas de servicios industriales y de manufacturación.	Interior
Espacio abierto	Campos, parques y zonas verdes.	Exterior
Carreteras	Caminos, áreas bajo derecho de paso y senderos, tanto peatonales como utilizados por vehículos.	Exterior
Ríos/Lagos	Zonas de agua en el terreno.	Exterior

que resume la actividad socio-económica en cada sección del terreno, denominados a lo largo de la tesis como tipos de uso de suelo. En la Tabla 3.2 se ofrece una breve descripción de los usos del suelo empleados, así como su clasificación en tipos de uso de interior y tipos de exterior.

3.3. Modelo de clasificación de conexiones en interior

Una vez formulado el problema, se presenta el modelo propuesto en este capítulo. Para ello, primero se explica el proceso de etiquetado artificial de datos, necesario para poder entrenar un modelo de regresión en ausencia de un conjunto de datos etiquetados. A continuación, se presenta el modelo de regresión logística propuesto para la obtención de la probabilidad de que una conexión sea de interior/exterior.

3.3.1. Etiquetado artificial

Debido a que en la red sólo se realizan mediciones pasivas (es decir, no se hace un geoetiquetado de las conexiones a priori), y los métodos basados en E-CID o RSRP utilizados durante el procesado de las trazas no tienen la suficiente precisión para localizar a un usuario en el interior, no es posible disponer de casos reales etiquetados. Por lo tanto, el único medio de generar casos etiquetados a posteriori es encontrar celdas

(o anillos) que cubran sólo los usos del suelo en interiores o exteriores. Un análisis más detallado muestra que hay muy pocos casos de celdas (o anillos) que cubran sólo zonas exteriores, pero ninguno que cubra sólo zonas interiores. En ausencia de etiquetas de conexiones contrastadas a nivel de celdas (o anillo), el etiquetado de datos se realiza de forma probabilística, asignando como salida el ratio de ubicaciones de interior de la celda (o anillo).

Para determinar qué etiqueta corresponde a cada punto del juego de datos, se establece una relación entre los usos del suelo y los indicadores obtenidos de las trazas, ambos agregados por celdas. Para ello, el escenario de la red se divide en una malla de puntos, de forma que el área de servicio de una celda se define como el conjunto de cuadrículas que representa la zona geográfica a la que da servicio una estación base.

Como el objetivo es clasificar las conexiones como interior/exterior, se agrupan los usos del suelo en interiores y exteriores, y se calcula un ratio de uso del suelo en interiores por cada celda c, $R_{LU_{in}}(c)$. Este ratio sirve para aproximar el ratio de conexiones generadas en interiores en esa celda c, $R_{k_{in}}(c)$, que es una estimación de la probabilidad de que una conexión k se genere en interiores en la celda, $P_{in}(c)$ (en adelante, se definirá como probabilidad de interior por brevedad).

El ratio de conexiones interiores en una celda, $R_{k_{in}}(c)$, se obtiene mediante 3 pasos, detallados a continuación.

Paso 1: Estimación del área de servicio de la celda/anillo

Como no se conoce la ubicación exacta del usuario, hay que aproximarla a partir del área de servicio de cada celda en el mapa. En las redes celulares, el radio de las celdas varía drásticamente, desde cientos de metros hasta decenas de kilómetros. En este trabajo, el área de las celdas omnidireccionales se representa de forma circular, mientras que en emplazamientos multisectoriales se define como un sector, cuyo ancho varía con el número de celdas del sector. En ambos casos, el radio del área de cobertura de la celda se deriva de las estadísticas de TA. En concreto, el radio de la celda se calcula como el percentil 90 % de la distribución de TA de las conexiones en la celda. Con esta configuración, se pretende considerar sólo los anillos de TA con cobertura sin fisuras de una celda. Así, se descartan los anillos de TA más alejados de una celda, que tienden a dar servicio de forma discontinua debido al solapamiento de las celdas. Se debe observar que la inclusión de estos últimos podría introducir algún sesgo en las estadísticas de uso del suelo procedentes de las cuadrículas del anillo de TA servido por otras celdas.

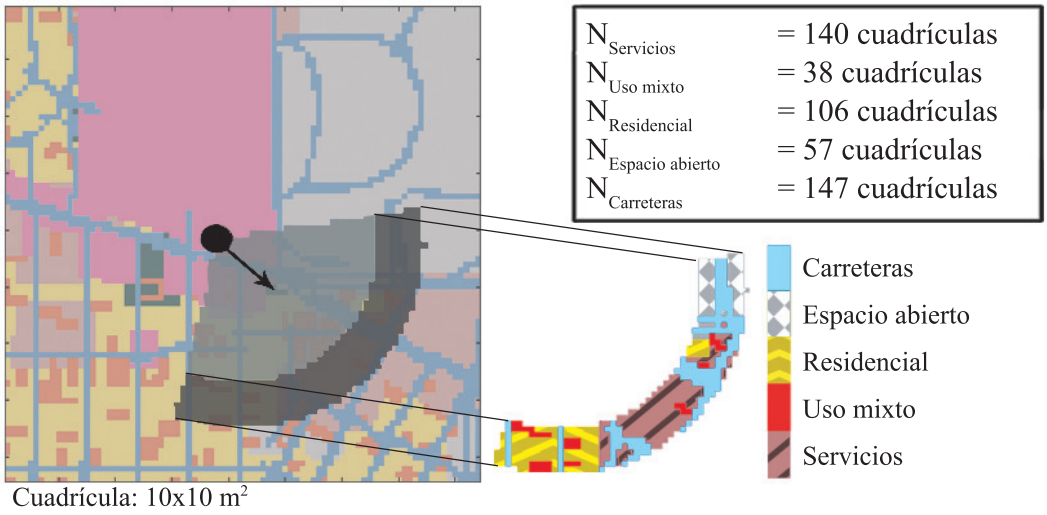


Figura 3.1: Ejemplo de cálculo de la superficie de uso del suelo por anillo.

Por conveniencia, el área de servicio de la celda se divide en anillos de 78 metros, que corresponden al ancho de espacio cubierto por el mismo valor de TA en LTE [63]. A continuación, las áreas de celdas/anillos se representan mediante una máscara binaria que muestra las cuadrículas servidas por cada celda/anillo.

Paso 2: Estimación de la superficie de uso del suelo por anillo

A continuación, el número de cuadrículas de un determinado uso del suelo por anillo se estima superponiendo la máscara del área del anillo sobre el mapa de uso del suelo. La Fig. 3.1 muestra un ejemplo de este proceso. La figura de la izquierda muestra la posición (punto) y el acimut (flecha) de la antena de una estación base servidora en el mapa. El área cubierta por una celda en un emplazamiento trisectorial se representa sombreado claro, mientras que el cuarto anillo de TA de la celda se representa sombreado de forma más oscura. La figura geométrica de la derecha muestra la parte del mapa de uso del suelo que se encuentra debajo del anillo, donde cada uso del suelo está representado por un color diferente. Para mayor claridad, también se muestra el número de cuadrículas por uso del suelo en el anillo.

Paso 3: Estimación de los ratios de uso efectivo del suelo por celda

Por simplicidad, el ratio de conexiones interiores por celda podría aproximarse por la proporción de cuadrículas de usos del suelo interior en la celda (es decir, el ratio de área interior). Sin embargo, esta estimación no tendría en cuenta que el tráfico no se distribuye uniformemente entre los diferentes tipos de suelo dentro de una celda. Por

un lado, algunos anillos de TA dan servicio a más conexiones que otros, por lo que su ratio de área interior debería prevalecer a la hora de calcular las estadísticas a nivel de celda. Del mismo modo, algunas cuadrículas dentro de un anillo tienden a servir más conexiones que otras en función del uso del suelo. Por ejemplo, una cuadrícula de uso de suelo de campo abierto presenta una menor densidad de usuarios que una cuadrícula de oficinas. Por esta razón, es importante evitar el sesgo en la clasificación interior/exterior introducido por grandes áreas geográficas con bajo volumen de tráfico (p.ej., agua o campo abierto).

Para tener en cuenta la diferencia entre cuadrículas, se estima una densidad media de conexiones por uso del suelo, β_{LU_n}, que muestra el número medio de conexiones por unidad de tiempo y cuadrícula para cada uso del suelo n (siendo $n \in \{$senderos, residencial, servicios,$\dots\}$). Para ello, se formula un sistema de ecuaciones lineales sobredeterminado que relaciona el número total de conexiones por celda c, $N_{con}(c)$, con las densidades de conexión, $\hat{\beta}_{LU_n}$, y el número de cuadrículas por uso del suelo en la celda, $N_{cuad}(LU_n, c)$, como

$$N_{con}(c) = \sum_n \hat{\beta}_{LU_n} N_{cuad}(LU_n, c) \quad \forall\, c\,, \tag{3.5}$$

$$\hat{\beta}_{LU_n} \geq 0 \quad \forall\, n\,. \tag{3.6}$$

El sistema de ecuaciones en (3.5)-(3.6), que incluye tantas ecuaciones como celdas y usos de suelo, y tantas incógnitas como usos del suelo, puede resolverse como un problema de optimización restringido, minimizando la suma del error cuadrado y bajo la restricción de que todas las densidades de conexión sean positivas. A continuación, las densidades nulas, asociadas a los usos del suelo menos frecuentes, se refinan utilizando regresiones parciales sobre un subconjunto de celdas con la mayor proporción de esos usos del suelo.

Para tener en cuenta la distribución desigual del tráfico entre los anillos, se calcula un peso de tráfico por anillo r en la celda c, $\omega(r, c)$, como

$$\omega(r, c) = \frac{N_{con}(r, c)}{\sum_r N_{con}(r, c)}\,, \tag{3.7}$$

donde $N_{con}(r, c)$ es el número de conexiones en el anillo r de la celda c, tomado de las medidas de TA.

En paralelo, se define el ratio de conexiones de interior en un anillo, determinado

a partir de la estima del número de conexiones en las cuadrículas de usos de suelo de interior ($n \in LU_{in}$) respecto a la estima de conexiones en todas las cuadrículas del anillo,

$$\kappa_{in}(r,c) = \frac{\sum\limits_{n \in LU_{in}} \hat{\beta}_{LU_n} N_{cuad}(LU_n, r, c)}{\sum\limits_{n} \hat{\beta}_{LU_n} N_{cuad}(LU_n, r, c)} , \tag{3.8}$$

donde $N_{cuad}(LU_n, r, c)$ es el número de cuadrículas de uso de suelo n en el anillo r de la celda c.

Por último, el coeficiente de conexión interior efectivo (en contraposición al coeficiente de área de interiores) se calcula por celda, combinando las dos ecuaciones anteriores, como el ratio entre la suma de todas las conexiones en interior en cada anillo y el número total de conexiones

$$R_{k_{in}}(c) = \sum_{r \in c} \kappa_{in}(r,c)\, \omega(r,c) = \frac{N_{con}(r,c) \sum\limits_{r \in c} \left[\frac{\sum\limits_{n \in LU_{in}} \hat{\beta}_{LU_n} N_{cuad}(LU_n, r, c)}{\sum\limits_{n} \hat{\beta}_{LU_n} N_{cuad}(LU_n, r, c)} \right]}{\sum\limits_{r} N_{con}(r,c)} . \tag{3.9}$$

Como se ha explicado anteriormente, el ratio resultante se utiliza para aproximar la probabilidad de conexión en interiores, $P_{in}(c)$, que es la variable dependiente en el modelo de clasificación, explicado en la siguiente subsección.

3.3.2. Modelo de regresión

Una vez etiquetados los datos, se utiliza un algoritmo de aprendizaje supervisado para encontrar el mejor modelo para relacionar los descriptores de rendimiento de las trazas, agregados a nivel de celda, $I_i(c)$, y la probabilidad de conexiones de interior en dicha celda, $P_{in}(c)$. A continuación, se utiliza dicho modelo para relacionar los KPI de conexión, $I_i(k)$, con su probabilidad de interior, $P_{in}(k)$.

El diagrama de flujo de este modelo se muestra en la Fig. 3.2. En él, se aprecia cómo las entradas consisten en dos fuentes de datos: a) el mapa de usos del suelo y b) la información registrada en las trazas de conexión radio. El mapa consiste en una malla de puntos que representa la actividad comercial de cada pequeño trozo de terreno (cuadrícula).

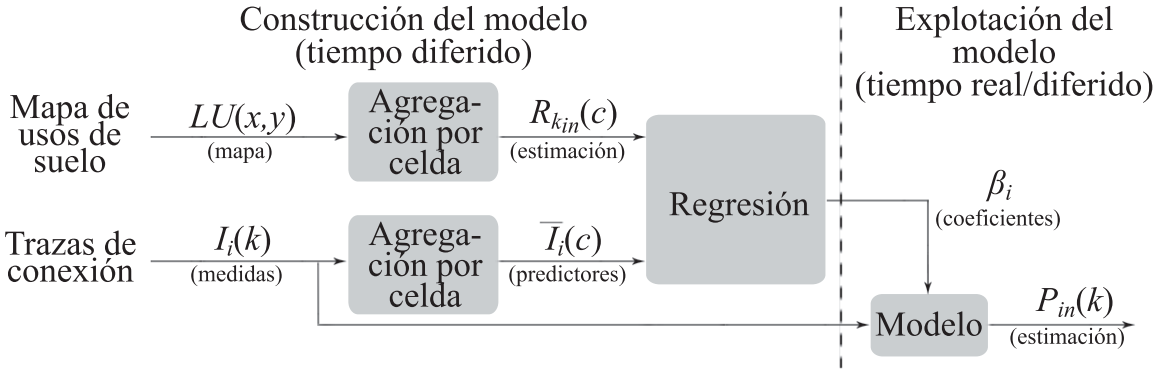


Figura 3.2: Diagrama de flujo del modelo de clasificación.

Con los descriptores definidos en la sección anterior, se construye un modelo de regresión para estimar la probabilidad media de conexión en interiores celda por celda a partir de los valores medios de los descriptores de las conexiones ocurridas en ellas. El objetivo del análisis de regresión es: a) encontrar la combinación mínima de indicadores que estimen con precisión la probabilidad de conexión en interiores, y b) calcular los coeficientes de regresión.

Como se explicó en la Sección 3.2, la probabilidad de conexión en interiores (que es la variable dependiente en el modelo) se aproxima durante el proceso de entrenamiento mediante el ratio de conexiones en interiores. Como este ratio oscila entre 0 y 1, se selecciona un modelo de regresión logística. La regresión logística es un modelo lineal generalizado [64] que refleja cómo una variable de respuesta binaria (en este caso, si una conexión es interior o exterior) depende de un conjunto de variables explicativas (KPIs de conexión). Con este objetivo, se estima una versión transformada logarítmica del cociente entre las probabilidades de que la conexión sea interior y la de que sea exterior aproximados por los ratios construidos con medidas, mediante una regresión lineal múltiple, como

$$\begin{aligned} \ln\left(\frac{P_{in}(c)}{1-P_{in}(c)}\right) \approx \ln\left(\frac{R_{k_{in}}(c)}{1-R_{k_{in}}(c)}\right) = \\ = \beta_0 + \beta_1\overline{I_1}(c) + \beta_2\overline{I_2}(c) + ... + \beta_{N_i}\overline{I_{N_i}}(c)\ , \end{aligned} \tag{3.10}$$

donde β_i es el coeficiente de regresión para el KPI I_i, $\overline{I_i}(c)$ es el valor medio del KPI I_i para las conexiones servidas por la celda c, N_i es el número de variables (KPIs) en el modelo, y $P_{in}(c)$ y $R_{k_{in}}(c)$ son la probabilidad de conexión en interiores y el ratio calculado de conexiones en interiores en la celda, respectivamente.

Una vez obtenidos los coeficientes de regresión para cada indicador en el análisis a nivel de celda, el modelo puede explotarse a nivel de conexión simplemente aplicando los coeficientes de regresión correspondientes a los KPI de la conexión a clasificar, como

$$P_{in}(k) = \beta_0 + \beta_1 I_1(k) + \beta_2 I_2(k) + ... + \beta_{N_i} I_{N_i}(k) \,, \tag{3.11}$$

donde el índice k denota una conexión particular.

3.3.3. Selección de variables

La disminución del número de variables en un modelo de regresión reduce la carga computacional, acelera el proceso de aprendizaje, mejora la capacidad de generalización y facilita la interpretación. En este trabajo, la selección de variables se lleva a cabo mediante el método de eliminación recursiva de características descrito en [65]. El proceso comienza realizando una regresión con todas las variables y clasificando las características en función de sus valores p, que reflejan la probabilidad de que cada coeficiente de regresión sea cero [66]. A continuación, se elimina la variable menos relevante (la que tiene el mayor valor p), obteniendo un modelo más sencillo de una variable menos a costa de reducir la precisión del modelo. Este paso de simplificación se repite hasta que la precisión del modelo disminuye por debajo de un determinado umbral (normalmente, cuando el coeficiente de determinación, R^2, del modelo simplificado es 0,1 menor que el del modelo completo).

3.4. Evaluación del modelo

Para comprobar la precisión del modelo propuesto, se evalúa su comportamiento con un gran juego de trazas de conexión, obtenidas de una red LTE comercial. Para poner en contexto la evaluación del modelo, primero se describen las características del escenario en el que se sitúa la red de la que se obtuvieron los datos, así como los datos de uso de suelo empleados para la obtención del ratio de conexiones interiores. A continuación, se muestra el rendimiento del modelo, tanto a nivel de celda como a nivel de conexión. Por último, para determinar la viabilidad del modelo en un entorno real de gestión de red, se presenta su eficiencia computacional.

3.4.1. Metodología experimental

El escenario considerado cubre un área geográfica de 125 km^2, correspondiente al área metropolitana de una ciudad de 800.000 habitantes. Para tener una resolución que permita identificar edificios, la zona se divide en cuadrículas de 10x10 m^2. El uso principal de suelo de cada cuadrícula se obtiene a partir de mapas construidos con datos abiertos obtenidos de entidades públicas. La Fig. 3.3 muestra un diagrama de sectores con la proporción de cada uso del suelo en el escenario. En este gráfico, los usos de suelo cuya naturaleza corresponde a zonas exteriores se representan con bordes discontinuos, mientras que los usos de suelo en interiores se representan con bordes continuos.

En ese escenario, se sitúan 400 celdas LTE de un único operador, agrupadas en 175 emplazamientos trisectorizados, con una frecuencia de portadora de 2.325 MHz y 15 MHz de ancho de banda del sistema. En estas celdas, la recogida de trazas se activó durante 2 horas por la mañana, periodo en el cual las celdas mencionadas sirvieron 236.357 conexiones.

La tabla 3.3 muestra la densidad de conexiones por uso del suelo n, estimada a partir del número de conexiones y los ratios de uso del suelo calculados por celda, utilizando el modelo de regresión de (3.5)-(3.6). Como se podría esperar a priori, se observa que la densidad de conexiones es mayor en las oficinas, los servicios y las carreteras. Es notable que algunos tipos de uso de suelo que, a priori, deben estar menos poblados, como los ríos/lagos, presentan una densidad de conexiones mayor que otros tipos de uso de suelo más susceptibles de albergar conexiones, como los usos de suelo de entretenimiento o industrial. Esto puede ser debido a la baja representación de estos tipos de uso de suelo en el escenario (2 %, 4 % y 3 %, respectivamente), que podría introducir un sesgo en la obtención de sus densidades de conexiones. Otro factor puede ser la franja horaria de recolección de los datos empleados, que podría derivar en una baja población en suelos de tipo entretenimiento, siendo un horario principalmente laboral.

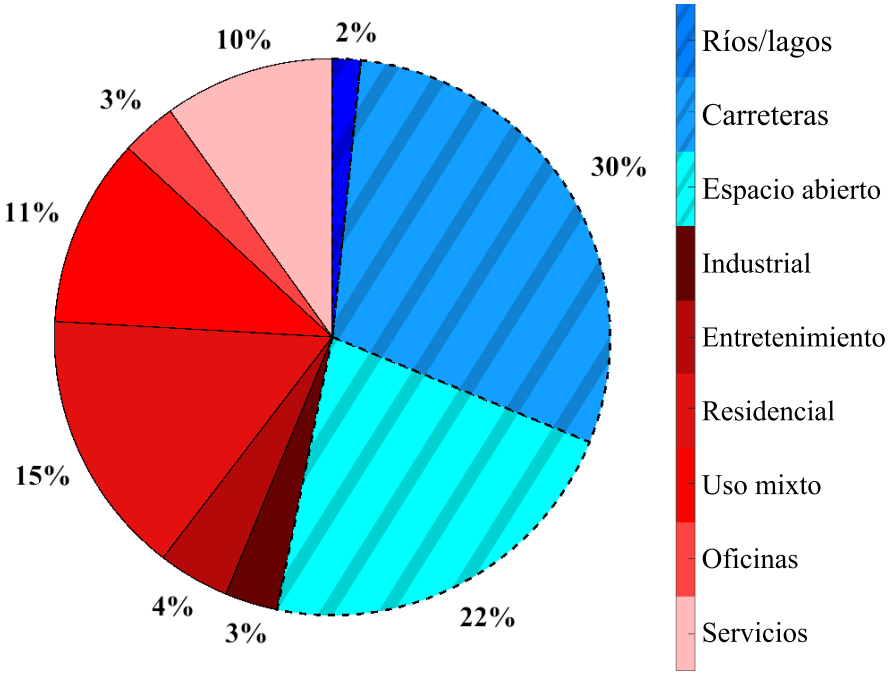


Figura 3.3: Ratios de uso de suelo en el escenario.

Densidades de conexión	
Uso de suelo, n	$\hat{\beta}_{LU_n}$ $\left[\frac{\text{Nº conexiones}}{\text{cuadrícula} \times \text{hora}}\right]$
Servicios	0,8176
Oficinas	3,0417
Uso mixto	$3,741 \times 10^{-9}$
Residencial	0,6401
Entretenimiento	$1,107 \times 10^{-9}$
Industrial	$8,395 \times 10^{-9}$
Espacio abierto	0,2340
Carreteras	0,8286
Ríos/Lagos	0,0588

Tabla 3.3: Densidad de conexiones por tipo de uso de suelo.

El modelo de regresión logística se entrena y valida con las trazas de estas 400 celdas. Para conocer mejor los datos de las trazas, la tabla 3.4 presenta los valores mínimos, medios y máximos de los KPI seleccionados en todas las conexiones del conjunto de datos.

Debido a que la información sobre la localización en interiores de las conexiones

Tabla 3.4: Estadísticas de los descriptores de conexión estudiados a nivel de celda.

Descriptor	Mínimo	Media	Máximo
Tiempo de conexión RRC [ms]	262	$2,49\times10^{4}$	$5,64\times10^{6}$
Volumen de datos en DL [bytes]	0	$5,53\times10^{5}$	$1,45\times10^{9}$
Ratio de dirección de enlace [-]	0	0,42	1
Ratio de continuidad de la transmisión en DL [-]	$1,57\times10^{-6}$	0,01	0,90
Ratio de continuidad de la transmisión en UL [-]	$4,71\times10^{-6}$	0,03	0,98
Ratio de volumen en el último TTI en DL [-]	0	0,62	1
Throughput en DL [kbps]	0	11,52	2.820,58
Throughput en UL [kbps]	0	1,04	884,95
RSRP medio [dBm]	-141	-99,81	-44
Error de estimación de potencia [dB]	-184,11	-70,40	5,29
ΔRSRP entre servidora y 1ª vecina [dB]	-40	11,42	98
Nº de vecinas con RSRP > -108 dBm	0	1,40	5
Nº de vecinas con ΔRSRP con servidora < 10 dB	0	0,95	8
RSRP - $\overline{\text{RSRP}}$ (anillo) [dB]	-122,98	-35,52	50,85
RSRP - $\overline{\text{RSRP}}$ (celda) [dB]	-122,98	-37,06	50,85
% medidas limitadas en potencia en UL [%]	0	55	100
CQI medio	0	8,56	15
CQI medio en Rango 1	0	8,70	15
CQI medio en Rango 2	0	5,13	15
SINR en el canal PUSCH [dB]	-7,92	9,12	20
SINR en el canal PUCCH [dB]	-11,75	2,76	5
Eficiencia espectral en DL [bits/RE]	0	0,75	5,20
Eficiencia espectral en UL [bits/RE]	0	1,19	3,63
% medidas en Rango 2 [%]	0	33,33	100
% medidas en Modo de Trans. 3, Rango 1 [%]	0	70,40	100
% medidas en Modo de Trans. 3, Rango 2 [%]	0	29,60	100

no está disponible en los conjuntos de datos recogidos por los operadores de telefonía móvil, el algoritmo descrito no puede compararse con los métodos clásicos de detección de actividad en interiores. Nótese que, aunque se realice una campaña de medición con un conjunto de usuarios geolocalizados, no cubriría todas las posibles condiciones de tráfico y propagación en el conjunto de datos de la red utilizado en este capítulo.

Para evitar el sobreajuste del modelo, el conjunto de celdas se divide en dos grupos, uno para el entrenamiento del algoritmo y otro para la validación del aprendizaje obtenido. El juego de entrenamiento incluye el 80 % de las celdas (320 celdas), mientras que el juego de validación incluye el otro 20 % (80 celdas). Ambos conjuntos se construyen de forma aleatoria para evitar sesgos producidos por la selección de puntos de datos, compartiendo la misma distribución espacial de las celdas. La evaluación del modelo se basa en la bondad del ajuste entre los ratios de interior computados con los datos de uso de suelo (es decir, los obtenidos con la ecuación 3.9) y los obtenidos del modelo

(ecuación 3.11) en el conjunto de validación, medido por el coeficiente de determinación de la muestra, R^2.

3.4.2. Resultados

Una vez descrito el escenario, en esta subsección se muestran los resultados obtenidos por el modelo de clasificación propuesto. Para clarificar los resultados, primero se ilustra la construcción del modelo, discutiendo algunos resultados intermedios obtenidos con el juego de entrenamiento. A continuación, se discute la evaluación del modelo, basándose en los resultados obtenidos con el juego de validación.

Construcción del modelo

El primer acercamiento al problema es construir el modelo para el escenario completo, usando la división del juego de datos antes descrita para entrenar y validar el modelo, teniendo así un único método aplicable a toda la red. En este sentido, la Fig. 3.4 muestra un gráfico de dispersión de los ratios de conexión interior calculados a partir del mapa de uso del suelo, frente a los derivados con el modelo de regresión logística con todas las variables (26 KPI). En este diagrama, cada punto corresponde a una de las 320 celdas del conjunto de entrenamiento. Puede observarse cómo el modelo capta la tendencia general, pero no es capaz de ajustarse de forma precisa a los datos de entrenamiento ($R^2 = 0,435$). Un análisis más detallado revela que esta falta de ajuste se debe en parte a los errores en los ratios de conexiones por uso de suelo basados en el mapa (sistema de ecuaciones en (3.5)-(3.6)) causados por las grandes desviaciones en las densidades de conexión ($\hat{\beta}_{LU_n}$) dentro de la misma clase de uso del suelo. Por ejemplo, la densidad de población/conexión no es la misma en todas las cuadrículas clasificadas como residenciales, ya que, por ejemplo, la altura de los edificios es completamente diferente en el centro de la ciudad y en los suburbios. Esto deriva en la existencia de diferentes perfiles de celdas en el escenario, por lo que es necesario detectarlos para poder modelarlos correctamente, ya que no se disponen de datos suficientes para entrenar un modelo más complejo.

Para corregir este problema y aumentar la precisión del modelo, se ejecuta un proceso iterativo de detección de valores atípicos basado en el algoritmo propuesto en [67] en las 400 celdas que componen el juego de datos. En cada iteración, se eliminan del conjunto de datos las muestras (celdas) clasificadas como anómalas en la parte superior (celdas sobreestimadas) y en la parte inferior (celdas subestimadas), hasta que

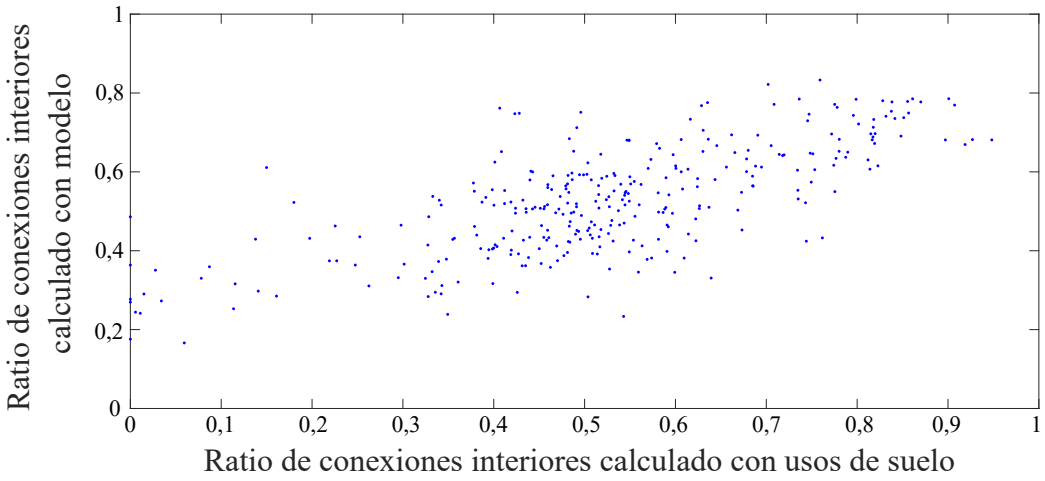


Figura 3.4: Ratios de conexiones en interior computados con los usos de suelo vs. obtenidos de las trazas (modelo completo, todas las celdas).

no se detecten más valores atípicos. Como resultado, se crean tres grupos de celdas: un primer grupo con celdas estimadas correctamente, un segundo con celdas subestimadas y un tercero con celdas sobreestimadas. Una vez obtenidos los tres grupos, se realiza la división de cada uno en entrenamiento y validación y se entrenan tres modelos diferentes, obteniendo así un modelo de regresión diferente para cada grupo.

La Fig. 3.5 evalúa la bondad del ajuste del modelo de regresión logística construido por grupos, comparando los ratios de conexión en interior basados en el mapa de uso de suelo y los obtenidos de las trazas en cada grupo. Como era de esperar, ambos ratios se ajustan mejor que con el modelo único, dando lugar a coeficientes de determinación más altos ($R^2_{grupo1} = 0{,}857$, $R^2_{grupo2} = 0{,}993$ y $R^2_{grupo3} = 0{,}786$, en comparación con $R^2 = 0{,}435$ para el modelo de toda la red), evidenciando así el correcto modelado del problema a nivel de celda.

Los modelos resultantes, derivados de las estadísticas a nivel de celda, pueden utilizarse para calcular la probabilidad de interior para cada conexión de forma independiente en función de sus descriptores. La Fig. 3.6 muestra los resultados de los tres modelos cuando se ejecutan por conexión. En cada subfigura, las tres curvas muestran los percentiles 5, 50 y 95 de la distribución de los valores de la probabilidad de interior calculados en base a las conexiones en cada celda de cada grupo. En esta figura, cada celda está representada por 3 puntos (1 por curva) alineados verticalmente, con una abscisa igual al ratio de conexiones en interiores basada en el mapa de usos de suelo en la celda. Para mayor claridad, los puntos de cada curva están conectados por un seg-

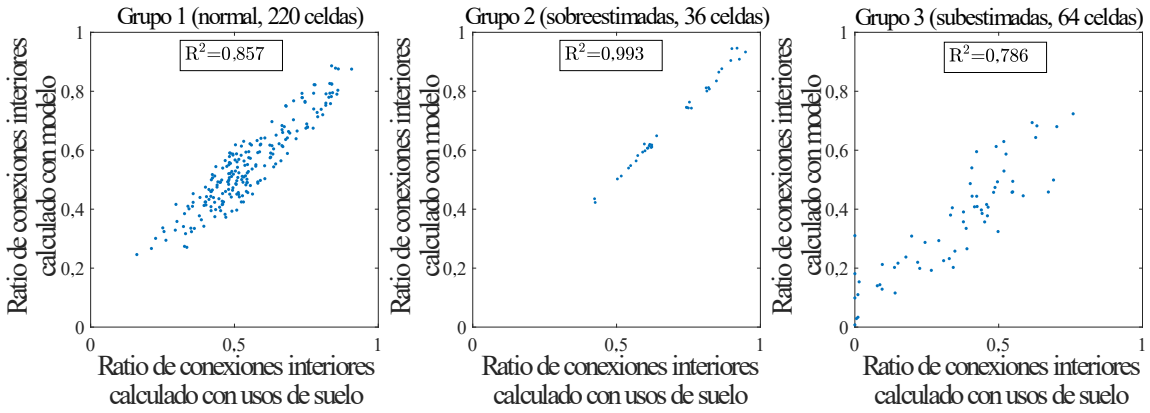


Figura 3.5: Ratio de conexiones de interior por celda computados con los usos de suelo frente a los obtenidos de las trazas (modelo completo, segregado en 3 grupos).

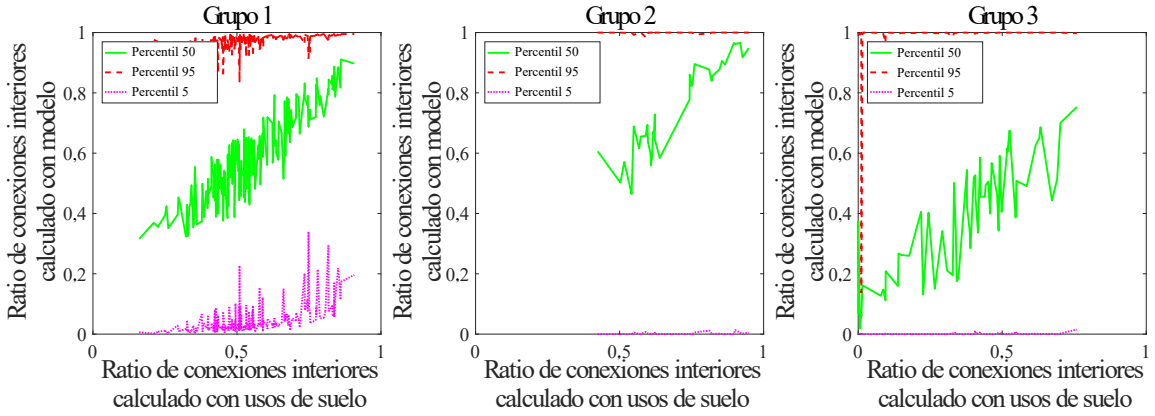


Figura 3.6: Percentiles de la probabilidad de conexiones en interior obtenidos de las trazas vs. ratios de conexiones en interior computados con los usos de suelo (modelo completo, segregado en 3 grupos).

mento de línea. Al observar la curva del percentil 50, se confirma que los tres modelos tienden a dar más conexiones con una alta probabilidad de ser interiores en las celdas con un mayor ratio de conexión interior basado en el mapa. Asimismo, las curvas de los percentiles 5 y 95 muestran que, como se esperaba, en las celdas con un ratio elevado de conexiones en interior basado en el mapa, algunas conexiones individuales siguen teniendo una probabilidad de interior baja. Este hecho demuestra que el modelo está segregando las conexiones en función de sus descriptores.

Para comprobar la importancia de las variables de entrada, la Fig. 3.7 muestra los valores p de cada descriptor en el modelo completo de los 3 grupos de celdas. También se representan los valores p medios, máximos y mínimos de los 3 modelos. Para facilitar el análisis, los descriptores en el eje de las abscisas están ordenados por su valor p medio en todos los grupos (cuanto más bajo, más significativo). Al comparar

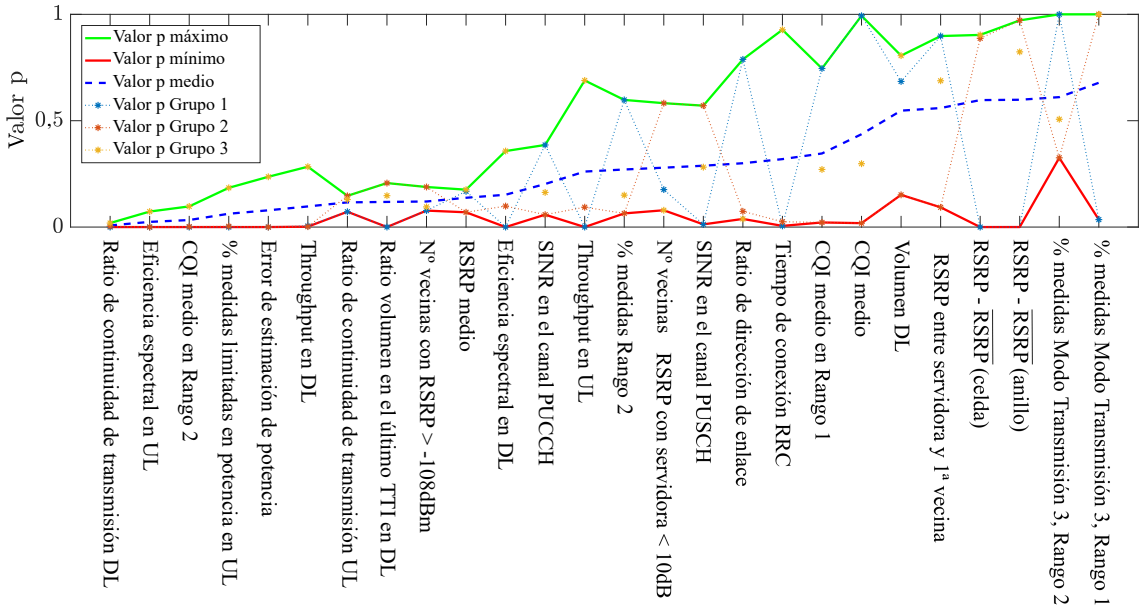


Figura 3.7: Importancia de cada variable en cada grupo de celdas (modelo completo).

las curvas de los grupos 1 a 3, se observa que cada grupo de celdas tiene sus propios descriptores relevantes, lo que justifica la necesidad de separar en grupos. Centrándose en los valores p medios, se concluye que los indicadores más relevantes a priori son el ratio de continuidad de transmisión DL, la eficiencia espectral UL, el CQI medio en el rango 2, el porcentaje de medidas de potencia limitada y el error de estimación de potencia.

En la Fig. 3.8, se muestra la pérdida de ajuste experimentada al reducir el número de variables en los modelos de regresión. Se observa cómo disminuyendo el número de descriptores desde los 26 del modelo completo a 7 en los grupos 1 y 2 y a 11 en el 3, la pérdida de precisión sigue siendo pequeña. En concreto, se obtiene un $R^2 = 0{,}794$, 0,908 y 0,713 en los modelos simplificados, frente a 0,857, 0,993 y 0,786 de los modelos completos en los grupos 1, 2 y 3, respectivamente.

En la tabla 3.5 se desglosan los resultados del análisis de regresión con los modelos simplificados. La tabla incluye las 13 características de conexión más relevantes en los tres grupos, junto con el valor de su coeficiente de regresión, $\hat{\beta}_i$, la desviación estándar, $s_{\hat{\beta}_i}$, el estadístico t de Student, T, y el valor p, P. Analizando estos últimos, se infiere que los indicadores más importantes son el ratio de continuidad de la transmisión DL, el error de estimación de potencia, el porcentaje de mediciones de potencia limitada UL reportadas por los terminales, la eficiencia espectral UL y el porcentaje de mediciones en el rango 2.

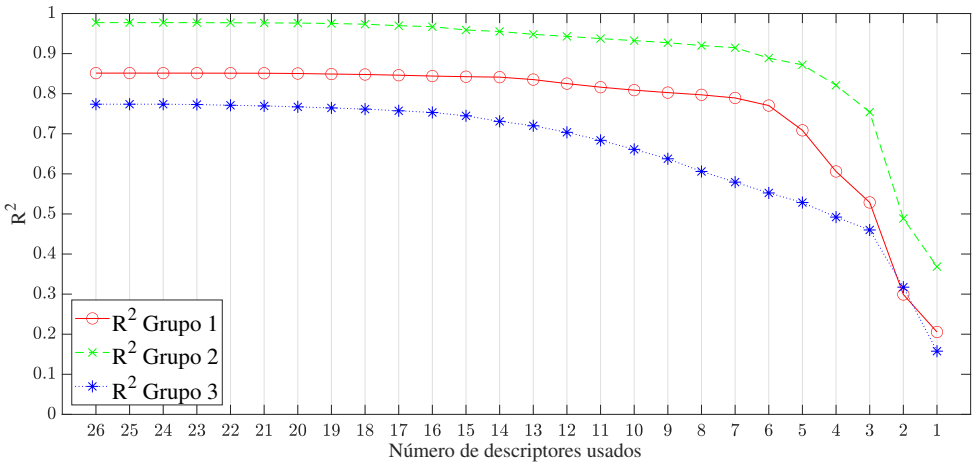


Figura 3.8: Impacto del número de variables por grupo de celdas.

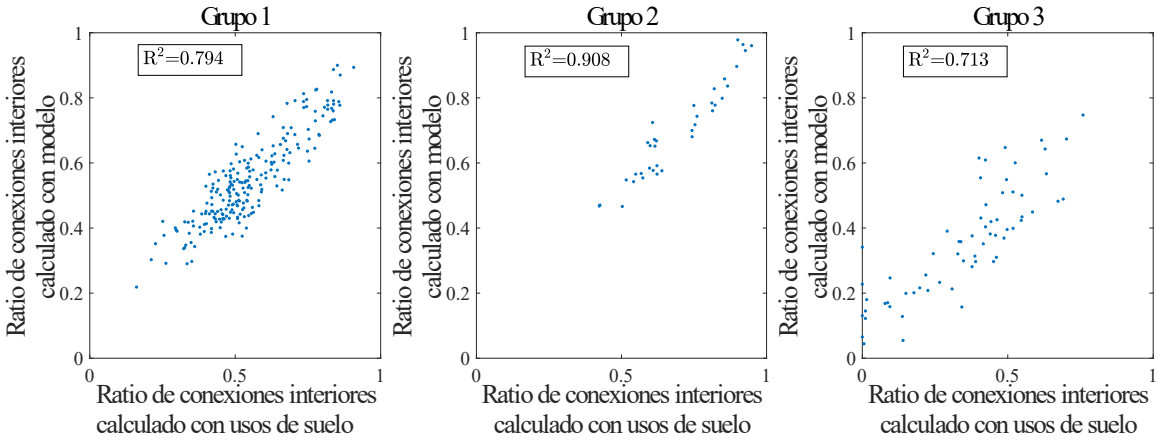


Figura 3.9: Ratios de conexiones en interior por celda computados con los usos de suelo vs. obtenidos de las trazas (modelo simplificado).

Para comprobar el comportamiento de los modelos simplificados, en la Fig. 3.9 se comparan los ratios de conexión interior por celda obtenidos con los modelos simplificados frente a los valores basados en el mapa. Asimismo, la Fig. 3.10 muestra los percentiles de la distribución de la probabilidad de interior de las celdas de cada grupo obtenidos con los modelos simplificados. Comparando las figuras Fig. 3.5-3.6 y Fig. 3.9-3.10, se observa que los modelos simplificados se comportan como los modelos completos en los tres grupos.

Validación del modelo

Una vez construidos los modelos con las 320 celdas del conjunto de entrenamiento, los modelos resultantes se prueban con las conexiones originadas en las 80 celdas del

Tabla 3.5: Análisis de regresión para el modelo simplificado.

Estadísticas de regresión				
Grupo de celdas		Grupo 1	Grupo 2	Grupo 3
Coeficiente de determinación, R^2		0,794	0,908	0,713
Análisis de varianza				
Grupo 1				
Descriptor de conexión	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$	T	P
Constante	0,34	0,02	14,08	$3,1 \times 10^{-32}$
Ratio cont. en trans. DL	-2,55	0,23	-11,27	$2,1 \times 10^{-23}$
Error de est. potencia	0,76	0,05	15,50	$9,8 \times 10^{-37}$
RSRP - $\overline{\text{RSRP}}$(celda)	-0.37	0.03	-12,81	$3,3 \times 10^{-28}$
% medidas limitadas en potencia en UL	-1.08	0.08	-12,87	$2,2 \times 10^{-28}$
CQI medio en Rango 2	1,47	0,30	4,83	$2,6 \times 10^{-6}$
Ef. espectral en UL	-0,90	0,08	-11,09	$7,5 \times 10^{-23}$
% medidas Rango 2	-2,02	0,34	-6,01	$7,9 \times 10^{-9}$
Grupo 2				
Descriptor de conexión	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$	T	P
Constante	1,13	0,07	16,41	$6,8 \times 10^{-16}$
Ratio cont. en trans. DL	-4,99	1,04	-4,77	$5,2 \times 10^{-5}$
Throughput DL	2.45	0.71	3.46	$1,8 \times 10^{-3}$
Error de est. potencia	1.25	0.13	9.78	$1,6 \times 10^{-10}$
RSRP - $\overline{\text{RSRP}}$(celda)	-0,41	0,09	-4,72	$6,0 \times 10^{-5}$
% medidas limitadas en potencia en UL	-0,88	0,19	-4,69	$6,5 \times 10^{-5}$
Ef. espectral en UL	-0,87	0,16	-5,36	$1,0 \times 10^{-5}$
% medidas Rango 2	-0,80	0,14	-5,92	$2,3 \times 10^{-6}$
Grupo 3				
Descriptor de conexión	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$	T	P
Constante	-0,65	0,09	-6,81	$1,0 \times 10^{-8}$
Ratio dir. de enlace	3,79	1,38	2,74	$8,4e{-}3$
Ratio cont. en trans. DL	-2,10	0,93	-2,27	$2,7 \times 10^{-2}$
Error de est. potencia	0,17	0,09	1,90	$6,3 \times 10^{-2}$
% medidas limitadas en potencia en UL	-2,63	0,60	-4,40	$5,4 \times 10^{-5}$
CQI medio	16,94	6,83	2,48	$1,7 \times 10^{-2}$
CQI medio en Rango 1	-18,11	7,05	-2,57	$1,3 \times 10^{-2}$
CQI medio en Rango 2	4,39	1,13	3,88	$2,9 \times 10^{-4}$
SINR en PUCCH	0,82	0,28	2,96	$4,7 \times 10^{-3}$
Ef. espectral en DL	2,87	0,63	4,52	$3,63 \times 10^{-5}$
Ef. espectral en UL	-2,68	0,74	-3,62	$6,7 \times 10^{-4}$
% medidas Rango 2	-4,79	1,17	-4,09	$1,5 \times 10^{-4}$

conjunto de validación. La Fig. 3.11 muestra diferentes percentiles de la distribución de la probabilidad de conexión en interiores en las celdas del conjunto de datos de validación. De la figura se concluye que los tres modelos dan lugar a distribuciones de

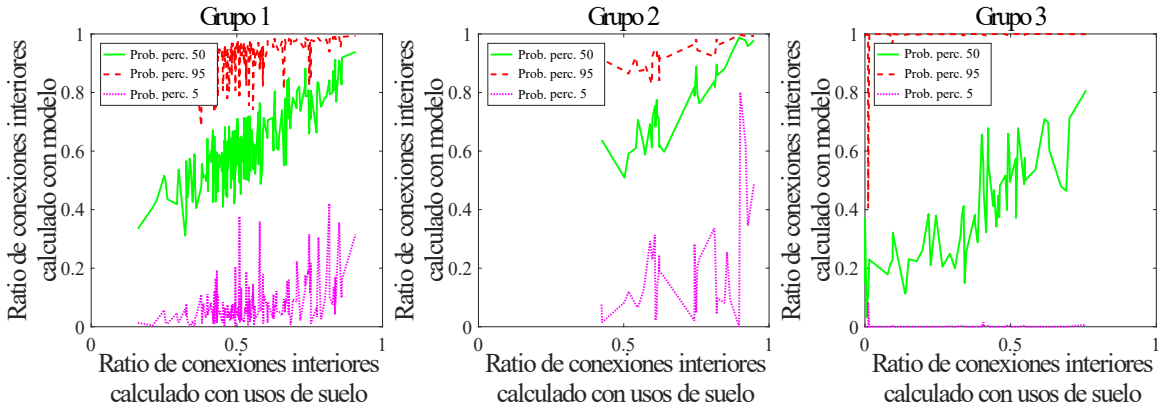


Figura 3.10: Percentiles de la probabilidad de conexiones en interior obtenidos de las trazas vs. ratios de conexiones en interior computados con los usos de suelo (modelo simplificado).

la probabilidad de conexión en interiores que son coherentes con los ratios de interiores basados en mapas. Las mayores desviaciones se observan en el Grupo 3, formado principalmente por celdas con bajos ratios de interior (zonas abiertas). Esto se debe posiblemente a errores en los ratios de interior basados en los mapas, causados por inexactitudes en las densidades de conexión estimadas de los usos del suelo no poblados (p.ej., espacios abiertos, que por su gran extensión, producen sesgos importantes en los ratios).

El análisis de los residuos confirma la bondad de los modelos de regresión logística propuestos. La prueba de Koenker-Basset [68] descarta la heteroscedasticidad en los 3 grupos, lo que demuestra que los residuos no dependen del valor de los descriptores de conexión (es decir, la homocedasticidad). Aunque la prueba de Shapiro-Wilk [69], utilizada para comprobar que los residuos tienen una distribución normal, sólo da resultados satisfactorios en uno de los modelos, un análisis más detallado de la distribución de los residuos en los demás grupos muestra una distribución casi normal. Asimismo, la prueba de validación cruzada [31], realizada dividiendo repetidamente el conjunto de datos de las celdas en diferentes conjuntos de entrenamiento y validación, confirma que los modelos de regresión resultantes no cambian significativamente.

3.4.3. Complejidad computacional

Para desarrollar el método estadístico propuesto en este capítulo, es necesario realizar un trabajo previo, que incluye la recogida/preprocesamiento de trazas de conexión y la construcción del mapa de usos del suelo. El tiempo de ejecución del preprocesado

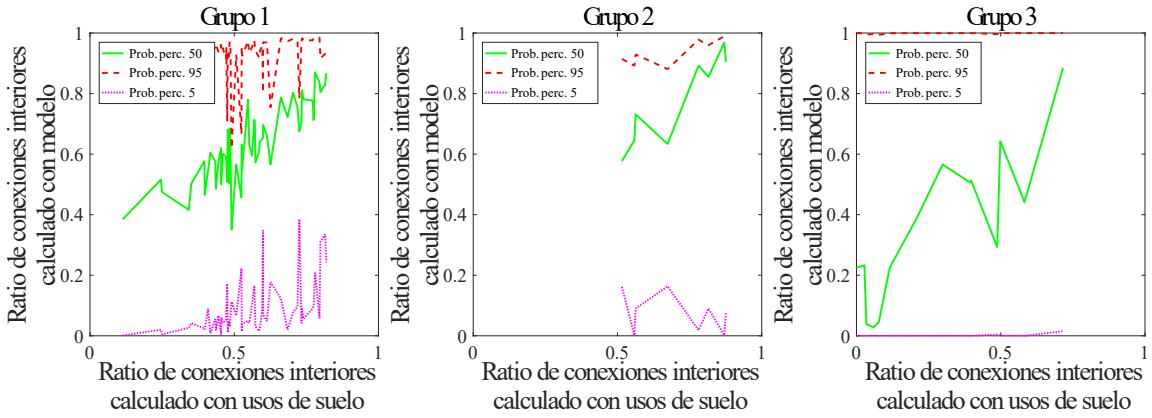


Figura 3.11: Resultados obtenidos con el juego de validación.

de trazas crece de forma lineal con el número de conexiones y descriptores, mientras que la construcción del mapa crece de forma lineal con el número de casillas del mapa.

Una vez obtenida y procesada la información, la complejidad computacional del método viene dada por la complejidad del algoritmo de selección de variables. El algoritmo construye modelos de regresión logística con un número decreciente de variables, donde N_i es el número de descriptores de conexión. Así, la complejidad temporal en el peor de los casos es $\mathcal{O}(N_i^3\, N_{con})$, donde N_{con} es el número de conexiones.

El modelo estadístico propuesto se ha implementado con las *toolbox* de estadística y aprendizaje automático de Matlab. En este entorno, y para el juego de datos de entrenamiento, compuesto por 320 celdas (125 m^2 de área geográfica), 189.510 conexiones (en un periodo de 2 horas) y 26 KPIs, el tiempo de ejecución del modelo completo, tanto para el entrenamiento/construcción como para la predicción, es de 48,3 segundos en un servidor con un procesador de 8 núcleos de 2,4 GHz y 64 GB de RAM. En estos datos no se incluye el tiempo necesario para decodificar las trazas de conexión.

3.5. Conclusiones

En este capítulo, se ha propuesto un modelo basado en datos para detectar si una determinada conexión se origina en una ubicación interior a partir de los atributos recogidos por la red móvil para cada conexión. El modelo se basa en una regresión logística, aplicada sobre los datos de las trazas de las conexiones radio almacenadas en el sistema de gestión de la red. Aunque el modelo se deriva de datos agregados por celda, el modelo derivado puede ejecutarse a nivel de conexión, dando diferentes probabilidades de interior para las conexiones en la misma celda. Para asegurar la

generalidad del modelo, se ha probado con un gran conjunto de trazas de una red LTE comercial.

Los resultados han demostrado que los ratios de conexión en interiores obtenidos con el modelo propuesto, basado en trazas, son coherentes con los ratios de interior observados en los mapas de uso de suelo, obteniendo un coeficiente de determinación de 0,79, 0,91 y 0,71 en el conjunto de datos de validación con el modelo simplificado de cada grupo de celdas. Este hecho indica que la precisión del modelo es suficiente para explicar el contexto interior/exterior de cada conexión en la mayoría de los casos. No obstante, se observan grandes desviaciones en celdas aisladas, que podrían deberse a errores en las densidades de conexión estimadas para algunos usos del suelo, o imprecisiones en los datos de uso de suelo, debidas a la resolución insuficiente del mapa o la falta de información precisa suministrada por las autoridades.

El análisis de las variables del modelo ha confirmado que las medidas relacionadas con la potencia del piloto son los indicadores que mejor muestran que una conexión puede haberse originado en el interior (p.ej., el error de estimación de potencia, con un valor p de $9,8 \times 10^{-37}, 1,6 \times 10^{-10}$ y $6,3 \times 10^{-2}$ en los 3 grupos). Además, se ha demostrado que las mediciones del enlace ascendente, como la eficiencia espectral en dicho enlace o el porcentaje de medidas limitadas en potencia, son indicadores valiosos para definir la situación en interiores. Inesperadamente, los resultados han mostrado que, aunque algunos indicadores relacionados con el servicio (p.ej., la relación de continuidad del tráfico) pueden ayudar a detectar conexiones en interiores, otros indicadores más significativos (p.ej., el ratio de dirección del enlace) son irrelevantes en la mayoría de los grupos. Por tanto, no hay pruebas claras de que el tipo de servicio demandado condicione la ubicación interior/exterior del usuario. Esto podría deberse a la gruesa granularidad espacial del conjunto de datos utilizado para construir el modelo, en el que se agregan los descriptores de conexión a nivel de celda, combinando el área cubierta por los anillos de avance temporal.

Los modelos desarrollados en este capítulo no se conciben como un sustituto de los modelos de actividad en interiores de la capa de aplicación, sino como una alternativa que permite a los operadores de redes aprovechar al máximo los datos de rastreo. Estos modelos pueden utilizarse en la replanificación de la red basada en medidas para mejorar la precisión de los algoritmos de geolocalización, localizando al usuario en mosaicos interiores/exteriores del mapa en función de las características de la conexión. Una vez entrenado, la baja complejidad computacional del modelo de regresión facilita su integración en las herramientas de planificación radio.

Capítulo 4

Posicionamiento estadístico de conexiones basado en información social

La ubicación en interiores de los usuarios es una de las características contextuales más relevantes, que puede aprovecharse para seleccionar modelos de QoE conscientes de contexto u optmizar los métodos de geolocalización en la red móvil. Este último problema es el que motiva este capítulo, donde se estudia el problema del posicionamiento estadístico de usuarios en la red. Para obtener esta distribución estadística, se usa la probabilidad de conexión en interior, obtenida con el algoritmo del capítulo anterior, como contexto de los usuarios y una estimación de la distribución de la población en el escenario obtenida de la información geo-posicionada de redes sociales de uso masivo.

Para explicar el proceso, este capítulo se compone de 5 secciones. Para contextualizar el algoritmo, en la sección 4.1 se comentan diferentes contribuciones de interés en el campo del posicionamiento de usuarios. A continuación, la sección 4.2 explora el problema de posicionamiento en las redes móviles. La sección 4.3 presenta la segunda contribución de esta tesis, el modelo de distribución estadística de usuarios en la red, mientras que la sección 4.4 presenta las pruebas realizadas para validar el algoritmo propuesto. Para finalizar el capítulo, la sección 4.5 resume las conclusiones extraídas del trabajo.

4.1. Estado de la técnica

Hoy en día, las redes móviles incluyen métodos de posicionamiento del usuario muy sofisticados basados en el terminal o la red [70]. Actualmente en LTE, el principal método de posicionamiento es el GPS asistido (A-GPS) [22], en el que el UE determina su ubicación procesando las señales GPS. Sin embargo, la información registrada para la gestión de la red es muy limitada, donde la única forma de que los operadores tengan acceso a la información GPS del usuario es a través de la funcionalidad MDT [71], que amplía el marco de gestión de trazas definido por el 3GPP para dotar al sistema de Operación, Administración y Mantenimiento de la red (*Operation, Administration, and Maintenance*, OAM) con la capacidad de recopilar la señalización entre diferentes nodos a todos los niveles de la red (usuarios, red de acceso y núcleo). Desgraciadamente, MDT rara vez se activa para todos los usuarios y de forma continua en el tiempo, debido, entre otras cosas, al consumo de batería que requiere del usuario, lo que hace que las trazas de llamadas anónimas proporcionadas por los equipos de la red carezcan a menudo de información detallada sobre la ubicación [72]. Por ello, la replanificación y optimización de la red debe hacerse utilizando otros métodos de posicionamiento. Estos métodos pueden ser basados o asistidos, tanto por la red como por el equipo de usuario (*User Equipment*, UE).

El principal método de posicionamiento alternativo en las redes móviles es por E-CID, que combina el Identificador de Celda y el Avance de Tiempo del enlace ascendente para localizar el terminal. Lamnetablemente, el alto margen de error cometido deshabilita este método para la detección de problemas en la red. Otros métodos, como la Diferencia de Tiempo de Llegada o el Ángulo de Llegada, que triangulan la posición del usuario a partir de mediciones (p.ej., la diferencia de tiempo entre señales de referencia o el ángulo de llegada) entre el terminal y las estaciones base observables, también son ampliamente utilizados, pero son muy sensibles a las condiciones de entorno, principalmente en entornos urbanos.

En contraposición, los métodos de localización basados en mapeo capturan firmas que se comparan con un conjunto de mediciones geoetiquetadas para identificar la ubicación de un dispositivo [73]. Estas técnicas, inicialmente diseñadas para localizar al usuario en un plano 2D, se han ampliado con la introducción de la tecnología MIMO 3D para escenarios interiores 3D [55]. Además, el uso de diferentes sensores (p.ej., barométricos para determinarla altura en interiores [74], de proximidad [75] o inerciales [76]) se ha estandarizado recientemente, mejorando así su precisión. Sin embargo,

las estimas de posición con este tipo de técnicas se ven afectadas por las imprecisiones introducidas por el canal radioeléctrico (p.ej., la ausencia de línea de visión, las sombras, el multitrayecto, etc.). Además, suponen un alto coste de preparación, ya que hay que realizar un gran número de mediciones para generar los mapas de comparación.

En paralelo, el interés por la ciencia de los datos ha ido aumentando en los últimos años debido a los recientes avances en la tecnología de la información, que posibilitan un procesado mucho más eficiente de grandes conjuntos de datos. Como resultado, se han lanzado muchas iniciativas de datos abiertos en todo el mundo. Los portales de datos abiertos ofrecen ahora un acceso directo y automatizado a valiosos activos que pueden utilizarse para mejorar la gestión de las redes celulares. Algunas empresas (p.ej., OpenSignal [77], WeFi [78], Ookla [79] o CellRebel [80]) ofrecen mediciones reales recogidas por usuarios anónimos (*crowdsourced*), que pueden utilizarse para evaluar el rendimiento de los despliegues actuales [16]. Las redes sociales son otra fuente de información para entender el comportamiento de los usuarios. La actividad en las redes sociales puede utilizarse para predecir el tráfico celular, independientemente de la tecnología de acceso utilizada por el proveedor de red [17]. Al mismo tiempo, la información sobre eventos sociales, obtenida de los resultados de búsqueda de los navegadores o de los repositorios de datos abiertos, puede utilizarse para explicar el comportamiento anormal de la red durante la resolución de problemas [81]. Del mismo modo, las áreas de mala cobertura de la señal o el rendimiento del servicio (vulgarmente conocidos como puntos negros) pueden detectarse mediante el procesamiento de mensajes de texto geoetiquetados en las redes sociales [18].

Desde el punto de vista de los operadores de red, la posición exacta de cada usuario no es tan relevante en las tareas de mantenimiento. En cambio, conocer la distribución espacial de los usuarios en un escenario es vital para optimizar las tareas SON de la red (p.ej., detectar zonas con problemas de capacidad). Con este objetivo, la información geolocalizada de redes sociales puede ser una fuente de datos interesante. No obstante, no se ha encontrado en la bibliografía un estudio que emplee esta información de redes sociales para estimar la distribución espacial de conexiones en una red móvil.

4.2. Formulación del problema

Conocer dónde se encuentran las agregaciones de usuarios que puedan suponer un cuello de botella es vital en las tareas de gestión de una red celular. Lamentablemente, la única forma de que los terminales reporten la posición del usuario al OSS es activar

la funcionalidad MDT, lo que no suele hacerse por el consumo de recursos de batería y almacenamiento que conlleva. Por fortuna, desde el punto de vista de la gestión de la red móvil, no es tan necesario conocer la posición precisa de los UE en el escenario, sino que lo verdaderamente importante es conocer su distribución estadística. El E-CID puede ser un acercamiento decente al problema, mayoritariamente en los anillos de menor superficie cercanos a la estación base servidora, pero la precisión es insuficiente en los anillos de mayor tamaño lejos de la estación base.

Una forma de reducir esta imprecisión en el proceso de localización es añadir información sobre el contexto del usuario. En concreto, la ubicación del usuario en entornos de interior o exterior es una información valiosa para perfilar su geoposición, pudiendo localizar al usuario en los edificios cubiertos por el anillo o en las calles/parques de la zona, reduciendo enormemente el área de imprecisión. Aun así, una vez identificado el contexto, seguiría siendo necesario precisar la posición el usuario.

Es un hecho que los usuarios cada vez utilizan más aplicaciones de carácter social, compartiendo constantemente su experiencia e impresiones con el resto de las personas en su entorno. De esta forma, las redes sociales juegan un papel fundamental, facilitando el acceso del público general a un portal donde publicar sus ideas, fotos o vídeos de forma masiva y, en muchos casos (como en Twitter [82] o Flickr [83]), permitiendo la inclusión de metadatos con la localización en la que se encuentran. Esta última información puede servir para refinar las estimas de la distribución espacial de conexiones realizadas con las medidas de red. Conviene resaltar que la información de localización que suministran las redes sociales, cuando está disponible, ofrece una resolución mucho mayor que la ofrecida por la técnica E-CID. Así, combinando ambas fuentes, es posible dotar de mayor granularidad al proceso de distribución espacial de conexiones, aumentando su precisión global principalmente en anillos lejanos.

En este capítulo, se plantea un modelo de posicionamiento estadístico de conexiones en una red celular, donde las entradas son las trazas de conexión y el fichero de planificación de la red, junto a la información de uso de suelo y una base de datos de publicaciones geolocalizadas en el área, y la salida es el mapa de distribución espacial de conexiones. La principal novedad del modelo presentado es la combinación de las diferentes fuentes de información descritas en esta sección.

En este modelo, al igual que en el modelo descrito en el capítulo anterior, el escenario de la red se define como una malla de puntos, mientras que las áreas de los anillos pertenecientes a cada celda se construyen de forma geométrica, dependiendo de las

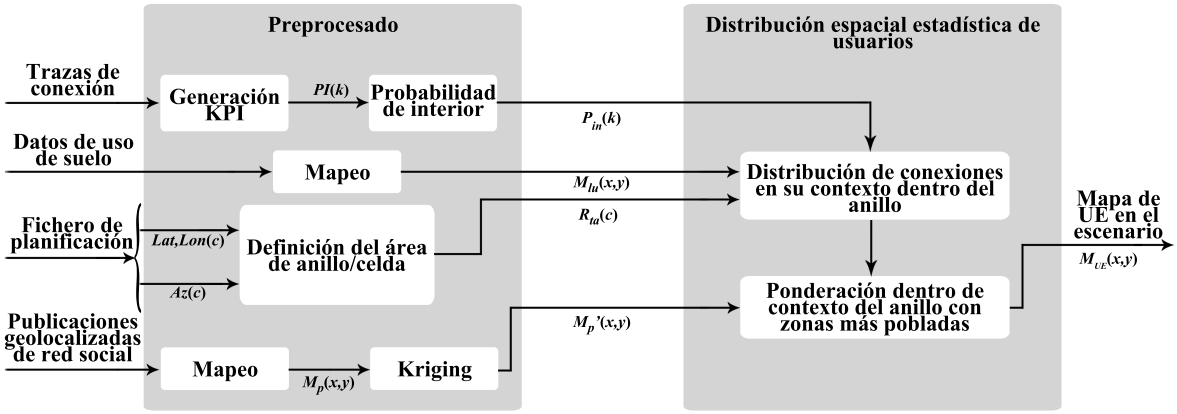


Figura 4.1: Diagrama de flujo del modelo de distribución espacial de conexiones.

estadísticas de TA y del número de celdas que comparten el emplazamiento. Debido a esta forma de definir el área de los anillos, puede existir un solapamiento entre anillos de diferentes celdas y, por tanto, algunas de las cuadrículas del mapa pueden ser cubiertas por varios anillos.

4.3. Modelo de posicionamiento estadístico

Para estimar desde dónde se establecen las conexiones en un escenario, primero hay que preprocesar las entradas descritas anteriormente. En esta sección, primero se especifica el procesado de los datos de entrada para conocer las características espaciales del escenario. A continuación, se explica el proceso de distribución estadística de conexiones en el espacio, que reparte las conexiones en un escenario en las áreas donde es más probable que se establezcan.

El diagrama de flujo del modelo propuesto en este capítulo aparece en la Fig. 4.1. El modelo recibe 4 entradas: (a) la información registrada en las trazas de conexión radio, (b) datos de usos de suelo del escenario, (c) información sobre la ubicación y orientación de las celdas en el escenario y (d) mensajes de las redes sociales (conocidos como publicaciones) que tienen asociados metadatos de geoposicionamiento. Estas fuentes de datos se han seleccionado por la importante información que aportan. Por un lado, la localización y orientación de las celdas en el escenario permite delimitar la zona geográfica en la que se encuentra el usuario, mientras que las trazas de conexión y los mapas de uso del suelo proporcionan información sobre el usuario y su contexto. Por último, las publicaciones geoposicionadas de las redes sociales se utilizan como una aproximación a la distribución espacial de la población dentro de cada celda o anillo

de TA.

El proceso de posicionamiento descrito a continuación se basa en el contexto espacial de los usuarios (es decir, interior/exterior), en lugar de los métodos tradicionales basados en medidas, que son susceptibles a las condiciones de propagación radio. En el modelo descrito, se distinguen dos pasos principales. El primer paso, la distribución de conexiones en su contexto dentro del anillo, se divide a su vez en dos etapas. Primero, se obtiene una delimitación de la zona en la que se produce la conexión a partir de la disposición de los emplazamientos en la red y de las estadísticas de TA, obtenida del fichero de planificación con la ubicación exacta de emplazamientos y sus sectores. Después, este área se reduce aún más, haciendo coincidir el contexto del usuario con las ubicaciones interiores/exteriores del escenario, es decir, localizando las conexiones interiores en edificios y las conexiones exteriores en parques y calles. El contexto interior del usuario se obtiene a partir de los descriptores registrados en las trazas radio, utilizando el modelo descrito en el capítulo anterior. Al mismo tiempo, las ubicaciones interiores/exteriores se obtienen a partir de los datos de uso de suelo. Por último, en la ponderación con zonas más pobladas, se realiza la distribución final de conexiones en la red. Esta distribución se obtiene ponderando qué partes de la zona limitada en el paso anterior tienen más probabilidades de acoger la conexión, usando una estimación de la distribución poblacional obtenida a partir de las publicaciones geolocalizadas recopiladas de los medios sociales.

Con estas entradas se obtiene un mapa con la distribución estadística de las conexiones en el escenario, donde cada cuadrícula corresponde a una agregación de las conexiones ocurridas en todos los anillos que la cubren.

Debido al creciente interés por la explotación de enormes cantidades de datos, la preocupación por la privacidad ha cobrado gran importancia. El tratamiento de los datos debe garantizar la privacidad del usuario durante todo el proceso de gestión de los mismos. Esta privacidad suele conseguirse mediante métodos de anonimización ejecutados en la primera fase por los propietarios de las redes sociales. De esta forma, el proceso de anonimización sustituye los datos del usuario final por una identificación de usuario única. La asociación de ID de cliente y usuario sólo está disponible para los gestores de las redes sociales, pero no para los operadores de red móvil, por lo que es posible utilizar la información geoposicionada de redes sociales en el modelo de posicionamiento, asegurando en todo momento la privacidad de los usuarios.

4.3.1. Preprocesado

El primer elemento imprescindible para el desarrollo de este capítulo es conocer el contexto de cada usuario. Como esta información no se registra en el OSS, los operadores necesitan estimarla. Esto se consigue con el algoritmo descrito en el capítulo anterior, donde se estima la probabilidad de que un usuario haya establecido la conexión en interior (es decir, $P_{in}(k)$) o exterior (es decir, $P_{ext}(k) = 1 - P_{in}(k)$) a partir de sus indicadores radio.

A continuación, se requiere saber la información de dónde se encuentran los edificios (es decir, interior) y qué áreas del escenario son descubiertas (es decir, exterior). Para ello, al igual que en el capítulo anterior, se construye un mapa de usos de suelo, $M_{lu}(x, y)$, a partir de información obtenida, de portales de datos aportados por fuentes oficiales, o bien de plataformas de datos abiertos (p.ej., OpenStreetMap). La información procedente de estas plataformas suele estar compuesta de una lista de elementos geométricos (polígonos, líneas y puntos), definidos por las coordenadas geográficas de los vértices del elemento, y una etiqueta que define su principal actividad social. Debido a la gran cantidad de valores que pueden adoptar las etiquetas de estos polígonos, estas se agrupan en un número inferior de clases con los diferentes tipos de uso de suelo.

En este trabajo, para determinar en qué área del escenario es posible que ocurra la conexión, se recurre al E-CID. Mediante el E-CID, se puede reducir la posible área de establecimiento de la conexión a la que cubre el anillo definido por el valor de TA en el que se ubica la conexión dentro de la celda servidora. En este capítulo, al igual que en el capítulo anterior, el área de un anillo en el mapa se define como todas aquellas cuadrículas cubiertas por la sección del área de la celda servidora que está a una distancia entre $78(i_{TA} - 1)$ y $78i_{TA}$, donde i_{TA} es el índice de anillo y la distancia entre anillos consecutivos es de 78 metros.

Por último, para estimar la dispersión espacial de la población, se construye una matriz de las mismas dimensiones que el mapa de usos de suelo y, a partir de la información de localización proporcionada por las redes sociales, se agrega el número de publicaciones geolocalizadas por cuadrícula, generando un mapa de distribución de publicaciones, $M_p(x, y)$. Esta matriz tiene el problema que, en zonas de baja densidad de población (p.ej., en campo abierto), el número medio de publicaciones por cuadrícula es muy inferior a uno, lo que provoca que la mayoría de las cuadrículas no tengan publicaciones o, en algunos casos, aparezcan solo unas pocas. Para estimar mejor estos pequeños valores de densidad, se aplica una regresión en procesos gaussianos (deno-

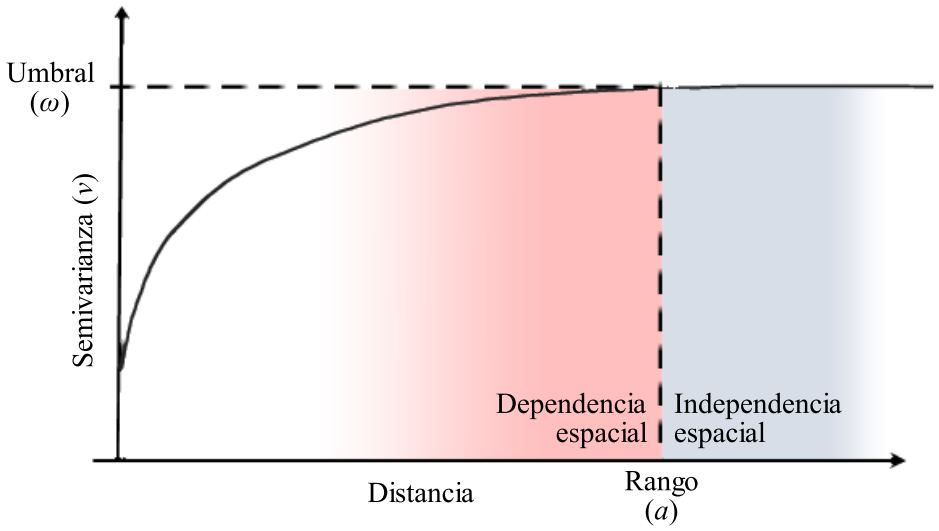


Figura 4.2: Ejemplo de variograma esférico.

minado krigeaje) para obtener la distribución espacial de publicaciones subyacente, $M'_p(x, y)$. El krigeaje (o *kriging*) [84] es un proceso de interpolación geoestadístico que, partiendo de unos puntos de referencia en un mapa, usa un modelo de variograma para estimar los puntos intermedios, modelando la semivarianza[1] entre pares de muestras según la distancia que los separa. El ejemplo clásico de aplicación de estas técnicas es en cartografía, donde, a partir de unas cotas de altura en el terreno, se estima la altitud del resto de puntos del escenario. En procesos de krigeaje simple, uno de los modelos de variograma más empleados es el variograma esférico [85], definido como

$$v = \begin{cases} \omega \left[\frac{3}{2} \left(\frac{|h|}{a} \right) - \frac{1}{2} \left(\frac{|h|}{a} \right)^3 \right] & \text{si} \quad h < a \,, \\ \omega & \text{si} \quad h > a \,, \end{cases} \,, \tag{4.1}$$

donde v es el valor de semivarianza, ω es el umbral o meseta, que determina el nivel al que se regula la semivarianza, y a es el rango del variograma, es decir, la distancia a la que la muestra deja de verse afectada por las vecinas. Por último, h es el vector de distancias a los puntos de referencia. La Fig. 4.2 ejemplifica una curva típica del modelado de la varianza con un variograma esférico.

Este variograma se emplea para ajustar los pesos usados en la estimación de las áreas sin mediciones, siguiendo el siguiente sistema de ecuaciones

[1] La semivarianza es una medida de la dispersión de una variable aleatoria en un conjunto de datos espaciales. Se define como la media de las diferencias cuadráticas entre los valores de la variable aleatoria en dos puntos espaciales separados por una distancia h

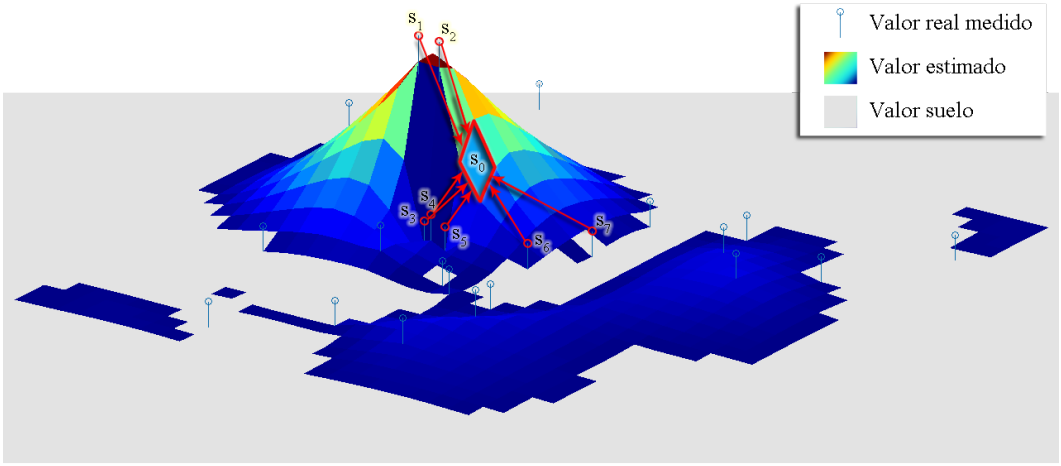


Figura 4.3: Ejemplo de kriging para el cálculo del mapa de publicaciones sociales subyacentes.

$$\begin{cases} v(h_{1p}) = \sum_{i=1}^{N} \lambda_i v(h_{1i}) \\ v(h_{2p}) = \sum_{i=1}^{N} \lambda_i v(h_{2i}) \\ \vdots \\ v(h_{np}) = \sum_{i=1}^{N} \lambda_i v(h_{ni}) \end{cases}, \tag{4.2}$$

donde N es el número de muestras reales medidas, $v(h_{jp})$ es el valor de semivarianza entre los puntos medidos j y p y λ_i es el peso correspondiente a la semivarianza entre los puntos j y i. Una vez ajustados los pesos que relacionan la semivarianza de las medidas reales, estos se aplican en la estimación de los valores en puntos sin mediciones reales de la siguiente forma,

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i), \tag{4.3}$$

donde $Z(s_i)$ es el valor medido en la ubicación s_i y $\hat{Z}(s_0)$ es el valor estimado en la ubicación que se desea predecir, s_0. Para facilitar la comprensión del impacto de las ecuaciones en el mapa de publicaciones sociales subyacente, $M'_p(x, y)$, la Fig. 4.3 muestra un ejemplo de la obtención del valor estimado de publicaciones sociales en s_0 a partir de 7 medidas reales (s_1, s_2, ..., s_7).

4.3.2. Distribución espacial de conexiones

Una vez procesados los datos de entrada del algoritmo, se construye un mapa con la localización de las conexiones. Desgraciadamente, las conexiones en las trazas rara vez están geolocalizadas, por lo que hay que localizarlas por el método E-CID. Esta técnica presenta grandes errores de posicionamiento en los anillos alejados de la celda servidora. Para eludir este problema, la distribución espacial de las conexiones dentro de un anillo puede inferirse a partir de la distribución de los mensajes geoetiquetados tomados de las redes sociales, como en [17]. En esta operación, se asume que la distribución de conexiones será proporcional a la distribución de mensajes de redes sociales, ya que la transmisión de mensajes cortos no está condicionada a un buen enlace de radio.

El proceso de geolocalización comienza con la creación de una malla con las mismas dimensiones de cuadrícula que el mapa de usos del suelo, $M_{lu}(x, y)$. En adelante, los índices (x, y) se refieren a los índices horizontales y verticales de las cuadrículas. A partir de los usos del suelo, se obtiene una matriz de contexto del usuario con el mismo tamaño, $M_{ct}(x, y)$, que indica si cada cuadrícula es interior o exterior (es decir, $M_{ct}(x, y) \in \{interior, exterior\}$). Este mapa se usa para distribuir cada conexión, dentro del anillo en el que ocurrieron (es decir, $(x, y) \in r$), en las áreas correspondientes a su contexto, en función de la probabilidad de que la conexión haya ocurrido en un contexto de interior/exterior. Para refinar aún más el área donde situar la conexión, se utiliza la distribución espacial de publicaciones subyacente, $M'_p(x, y)$, como estimación de la distribución de la población en el escenario, pudiendo aplicarse para ponderar en qué zonas del anillo es más probable que una conexión ocurra (es decir, zonas con mayor población). En la Fig. 4.4 se muestra un ejemplo de la distinción de cuadrículas interior/exterior dentro de un anillo, así como la sección equivalente de la distribución espacial de publicaciones subyacente. En la leyenda de la figura, se distinguen las cuadrículas de interior con un patron de tablero de ajedrez, frente a las cuadrículas de exterior, presentadas con un tono gris constante. Además, el mapa de distribución espacial de publicaciones subyacente se ilustra con un mapa de colores, donde valores más altos obtienen tonos rojos, mientras que las áreas menos pobladas se representan con tonos azules. En el anillo de ejemplo mostrado en la figura se pueden observar dos principales áreas de concentración de publicaciones de redes sociales, una situada en exteriores, en la parte central derecha del anillo, y otra de interiores, localizada en el extremo izquierdo.

De esta forma, los datos descritos se utilizan para calcular la probabilidad de que se

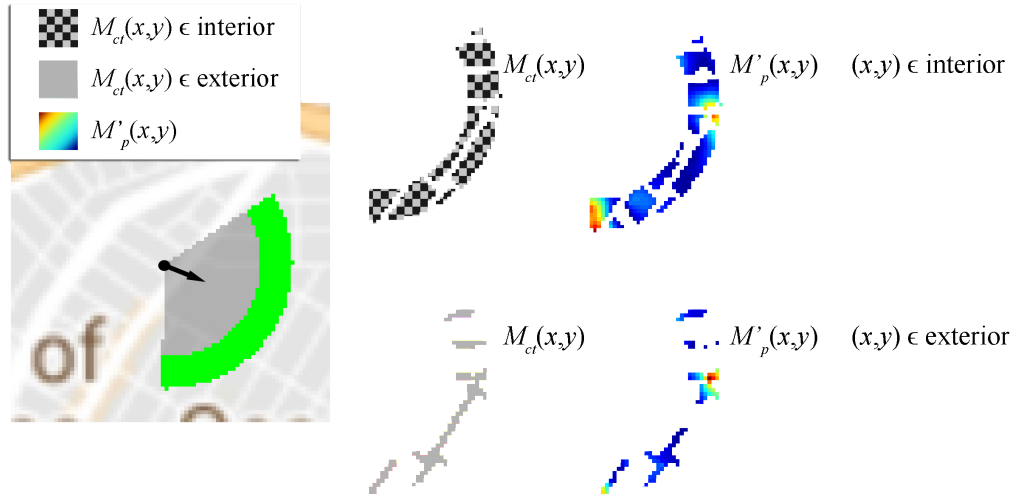


Figura 4.4: Ejemplo de obtención de datos de contexto y de publicaciones de un anillo.

produzca una conexión en una cuadrícula (x, y) etiquetada como contexto ct (es decir, interior o exterior) en el anillo r. Para realizar esta estimación, se tiene en cuenta cómo de poblada está cada cuadrícula individual y el tamaño de la superficie del contexto en el anillo a la que pertenece. La estimación de la probabilidad por cuadrícula se resuelve como

$$P(r,x,y) = \frac{1 + M'_p(x,y)}{N_{ct}(r) + \sum_{(x,y)\in r,ct} M'_p(x,y)} \qquad (x,y) \in r, ct\ , \tag{4.4}$$

donde $N_{ct}(r)$ es el número de cuadrículas etiquetadas como contexto ct en el anillo r y $M'_p(x,y)$ es la distribución espacial de las publicaciones sociales. Obsérvese que el mismo índice de cuadrícula (x, y) puede ser servido por anillos de diferentes celdas y, por tanto, se asocian diferentes valores de $P(r, x, y)$ a cada anillo que sirve a la misma cuadrícula. Un análisis más detallado de (4.4) muestra que, en los anillos en los que el número de publicaciones geoetiquetadas es 0 (como podría ser en zonas despobladas), $M'_p(x,y) = 0$, causando que las conexiones registradas en las trazas se distribuyan uniformemente en el área del anillo (es decir, $P(r,x,y) = 1/N_{ct}(r)$). En cambio, en los anillos con un gran número de publicaciones geoetiquetadas, $M'_p(x,y) >> 1$, causando que las conexiones se distribuyan siguiendo la distribución de publicaciones en el anillo (es decir, $P(r,x,y) \approx (M'_p(x,y))/\sum_{(x,y)\in r} M'_p(x,y)$).

Una vez calculadas las probabilidades de las conexiones del anillo en cada cuadrícula, estas probabilidades se proyectan en el mapa. Para ello, se distribuyen las conexiones

de interior/exterior en sus respectivas cuadrículas dentro del anillo según las probabilidades en (4.4) y luego se agrega entre los anillos que dan servicio a la misma cuadrícula, como

$$M_{UE}(x,y) = \sum_{r/(x,y)\in r} \sum_{k\in r} P_{ct}(k)P(r,x,y) \qquad (x,y) \in ct, \tag{4.5}$$

donde $M_{UE}(x,y)$ es el mapa de distribución estadística de conexiones en el escenario y $P_{ct}(k)$ es la probabilidad de que la conexión k haya ocurrido en el contexto ct.

4.4. Evaluación del modelo

En esta sección se comprueba la validez del modelo en un escenario real. Para ello, primero se explican los datos del escenario usados para el desarrollo del modelo. A continuación, se muestran los resultados de la distribución estadística de conexiones en la red obtenida con el modelo descrito en este capítulo, que hace uso de la información de contexto de interior y las publicaciones geolocalizadas para mejorar la precisión (en adelante, modelo completo). Las estimas obtenidas se comparan con las obtenidas con dos métodos de referencia. En el primero de ellos, con una distribución de conexiones basada solo en el E-CID, en el que las conexiones se distribuyen equitativamente en toda la superficie del anillo en el que ocurren, es decir, $M_{UE}(x,y) = \sum_{r/(x,y)\in r} N_k(r)/N_{(x,y)}(r)$, donde $N_{(x,y)}(r)$ es el número de cuadrículas en el anillo y $N_k(r)$ es el número de conexiones en el anillo (llamado en adelante modelo E-CID). En el segundo de ellos, con un modelo intermedio que añade únicamente información del contexto, es decir, $M'_p(x,y) = 0 \quad \forall(x,y)$, (llamado modelo de contexto). Por último, se evalúa la complejidad computacional del modelo.

4.4.1. Metodología experimental

El modelo desarrollado en este capítulo es una aplicación práctica del algoritmo propuesto en el capítulo anterior, por lo que se usa el mismo escenario para su validación. Este escenario corresponde a una ciudad de 800.000 habitantes y su extrarradio, que cubre un área geográfica de 125 km^2. En este área se han desplegado 400 celdas LTE, agrupadas en 175 emplazamientos trisectoriales, con una frecuencia de portadora de 2.325 MHz y un ancho de banda del sistema de 15 MHz. En estas celdas se activa la recogida de trazas de conexión durante 2 horas, registrándose 166.561 conexiones.

Tabla 4.1: Tipos de uso de suelo en el escenario.

Uso de suelo	Descripción	Interior/Exterior	Ocupación
Servicios	Áreas de uso institucional, cultural, médico o académico.	Interior	10 %
Oficinas	Oficinas de gestión, información o trabajo.	Interior	3 %
Uso mixto	Zonas sin uso de suelo principal definido.	Interior	11 %
Residencial	Casas, hoteles y alojamientos.	Interior	15 %
Entretenimiento	Comercios y entretenimiento.	Interior	4 %
Industrial	Zonas para trabajo industrial o de manufacturación.	Interior	3 %
Espacio abierto	Campos, parques y zonas verdes.	Exterior	22 %
Carreteras	Caminos, carreteras y senderos.	Exterior	30 %
Ríos/Lagos	Zonas de agua en el terreno.	Exterior	2 %

Tanto el mapa de usos de suelo como el mapa de distribución de publicaciones geolocalizadas se definen como una malla de cuadrículas de 10x10m^2 para tener una resolución suficiente. En la Tabla 4.1 se muestra un desglose de los tipos de uso de suelo, su descripción, la clasificación de interior o exterior y el porcentaje de ocupación en el escenario.

Por último, las publicaciones geoetiquetadas de las redes sociales en la zona se recogen en tiempo real, empleando la herramienta descrita en el Apéndice A.1, durante 16 meses para Twitter y 12 meses para Flickr, lo que da como resultado 785.515 y 33.519 publicaciones, respectivamente.

La Fig. 4.5 muestra la Función de Distribución Acumulada (*Cumulative Distribution Function*, CDF) de las probabilidades de cada conexión de haber ocurrido en interior. Se aprecia cómo, a partir de sus descriptores, el 80 % de las conexiones se etiquetan como llamadas de interior con una probabilidad superior al 92,5 %.

La Fig. 4.6 muestra las CDF del número de publicaciones por cuadrícula del escenario, $M_p(x, y)$, desglosadas por aplicación (línea de guiones para Twitter y línea de puntos para Flickr). Se puede observar que la mayoría de los mensajes provienen de publicaciones en Twitter, ya que coincide su CDF con la distribución de publicaciones totales. Por ello, será la principal fuente de información, dejando a las publicaciones de Flickr como una fuente auxiliar para reforzar determinados puntos del escenario con poco tráfico de redes sociales. Igualmente, se aprecia cómo el 24,5 % de las cuadrículas del escenario tienen más de 100 publicaciones.

Para observar la resolución de la distribución espacial de publicaciones subyacen-

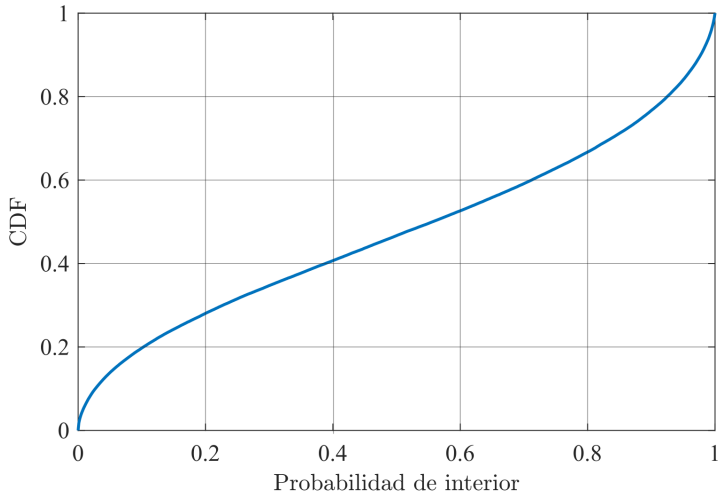


Figura 4.5: Función de distribución acumulada de la probabilidad de conexión en interior en el escenario.

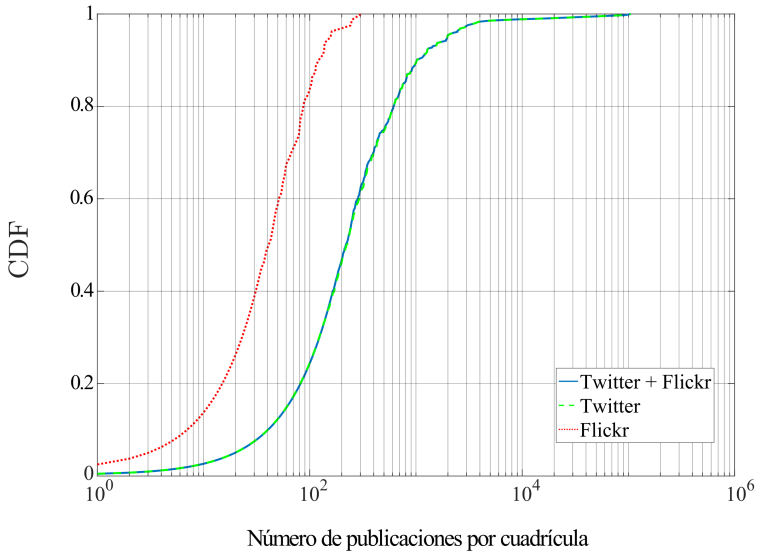
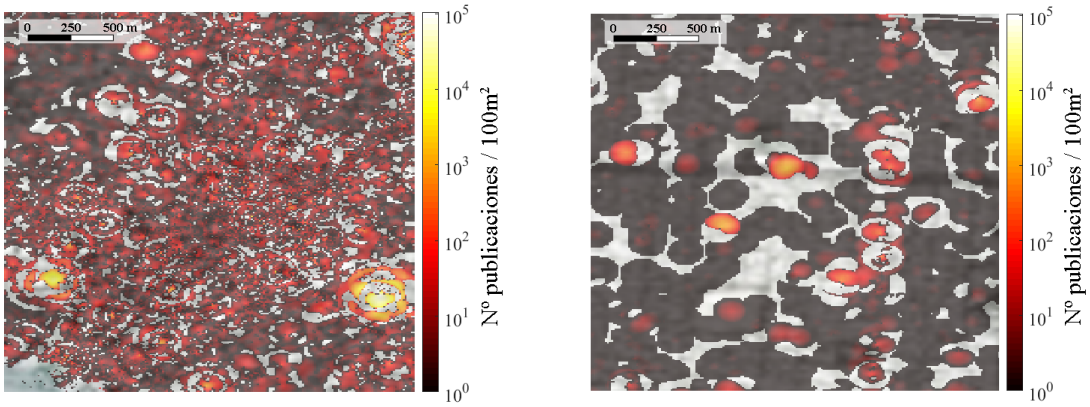


Figura 4.6: Función de distribución acumulada del número de publicaciones geolocalizadas por cuadrícula en el escenario de distintas redes sociales.

te, $M'_p(x, y)$, la Fig. 4.7 presenta dos secciones de $M'_p(x, y)$ en el escenario, uno en el que la densidad de población por metro cuadrado es elevada (centro de la ciudad, Fig. 4.7(a)) y otro que se encuentra relativamente despoblada (el extrarradio de la ciudad, Fig. 4.7(b)). En ambas figuras se puede observar el efecto del variograma esférico empleado en el proceso de kriging simple durante la estimación de $M'_p(x, y)$, apareciendo artefactos circulares en las áreas sin medidas reales cercanas. Por otro lado, las zonas huecas en el mapa representan las áreas con una densidad menor a 1 publicación



(a) Zona de alta densidad de población. (b) Zona de baja densidad de población.

Figura 4.7: Secciones del mapa de distribución espacial de publicaciones sociales en el escenario.

por cuadrícula (100 m^2). Además, se aprecia cómo, en zonas de alta densidad de población, $M'_p(x, y)$ toma valores muy heterogéneos, abarcando amplio rango dinámico. Por el contrario, en zonas de baja densidad poblacional, $M'_p(x, y)$ adopta, salvo en algún punto aislado, valores mucho más constantes y de menor magnitud.

El método propuesto pretende resolver la falta de precisión en la localización de conexiones en la red, mediante la explotación del contexto espacial tanto de las conexiones como del escenario. La única forma de obtener mediciones GPS de los usuarios es utilizando la funcionalidad MDT, que rara vez está activa y nunca a todas las conexiones de la red. Esta falta de información imposibilita la estimación del error en el proceso de posicionamiento, por lo que es necesaria una evaluación más exploratoria, comparando los resultados de la distribución estadística de conexiones en la red con los modelos de E-CID y de contexto anteriormente comentados.

4.4.2. Resultados

En el escenario descrito, se prueba el algoritmo de posicionamiento estadístico. Para comprobar el funcionamiento del modelo, la Fig. 4.8 muestra un diagrama de cajas del número de conexiones por cuadrícula en el escenario completo obtenido mediante el modelo E-CID, el modelo de contexto y el modelo completo. Este tipo de gráficas muestra los percentiles 25, 50 y 75 de los valores estudiados mediante el límite inferior, la línea central y el límite superior de la caja, respectivamente, junto a los valores extremos no atípicos (presentados con las líneas discontinuas) y los valores atípicos (indicados con cruces rojas). Para optimizar la representación del rango de valores

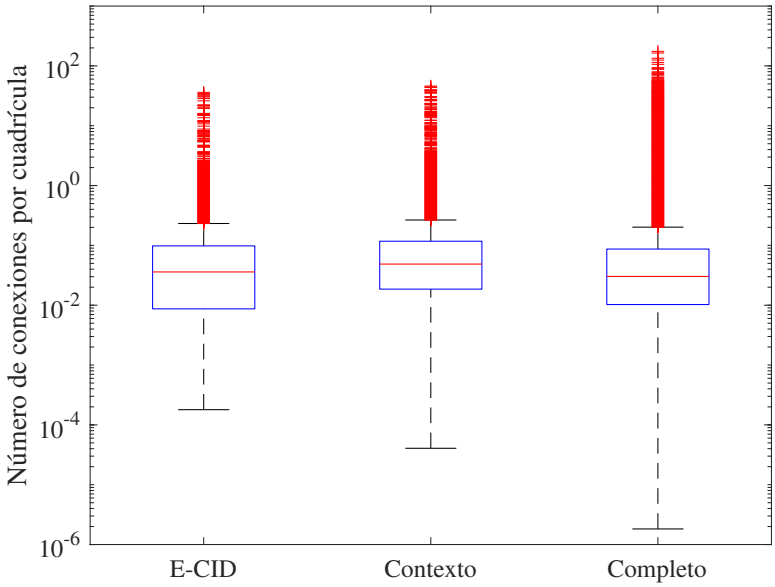


Figura 4.8: Número de conexiones por cuadrícula para cada modelo de distribución espacial de conexiones.

Tabla 4.2: Estadísticas del número de conexiones por cuadrícula con los distintos modelos distribución espacial de conexiones comparados.

Método	**Mínimo**	**Perc. 25 %**	**Perc. 50 %**	**Perc. 75 %**	**Máximo**
Modelo E-CID	$1,79 \times 10^{-4}$	$8,69 \times 10^{-3}$	$3,60 \times 10^{-2}$	$9,81 \times 10^{-2}$	36,29
Modelo de contexto	$4,06 \times 10^{-5}$	$1,86 \times 10^{-2}$	$4,87 \times 10^{-2}$	$1,17 \times 10^{-1}$	45,99
Modelo completo	$1,82 \times 10^{-6}$	$1,03 \times 10^{-2}$	$3,03 \times 10^{-2}$	$8,68 \times 10^{-2}$	175,19

obtenidos, el eje de ordenadas se presenta en escala logarítmica. Como complemento, la Tabla 4.2 presenta los datos numéricos de los estadísticos en cada modelo.

La Fig. 4.8 ilustra cómo, al añadir el contexto interior de las conexiones (es decir, el modelo de contexto), se detectan agrupaciones de usuarios en interior, vaciando así las zonas exteriores en las que no había conexiones y, por tanto, disminuyendo la cantidad de conexiones en esas cuadrículas, ampliando así la distribución estadística de valores de densidad en la figura en comparación con el modelo E-CID. Este hecho se ve favorecido por la inclusión de los publicaciones geoposicionadas de redes sociales en el modelo completo, que permite identificar cuadrículas con alta densidad de conexiones y zonas despobladas, enmascaradas en los modelos anteriores. Como consecuencia, la distribución estadística del modelo completo se amplía adicionalmente. Estos hechos se traducen en que los modelos completo y de contexto, que utilizan información adicional, mejoran las posiciones de las conexiones en el área del anillo, obteniendo un diagrama

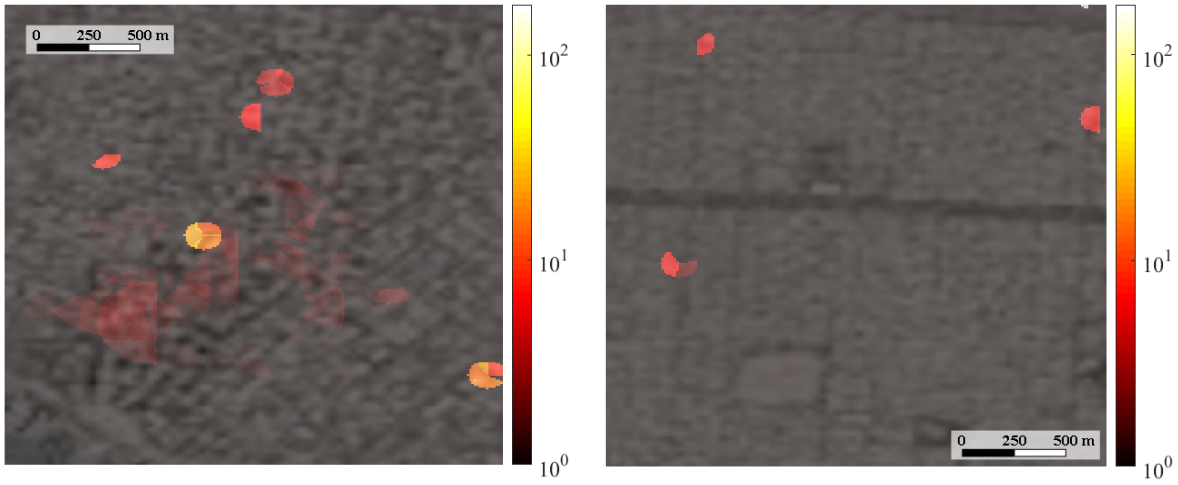
de cajas más disperso que el modelo E-CID.

Para comprobar la resolución en la distribución de conexiones en la red, la Fig. 4.9 presenta varios mapas de calor con los resultados obtenidos en los ejemplos de densidad de población en la Fig. 4.7. En estos casos, las conexiones se han distribuido con los 3 modelos descritos anteriormente: a) y b) con el modelo E-CID, c) y d) con el modelo de contexto y, e) y f) con el modelo completo, donde se combinan el E-CID, el contexto del usuario y las publicaciones geolocalizadas.

La Fig. 4.9(a)(b) muestra la distribución espacial de conexiones aplicando el método tradicional E-CID en zonas de alta y baja densidad de población. Las formas definidas que aparecen en la figura provienen de la distribución uniforme en cada anillo de la celda, principalmente de los anillos más cercanos a la estación base. En cambio, la forma dispersa de la parte inferior de la Fig. 4.9(a) es debida a la superposición de polígonos de muchos anillos de celdas distintas. Estas subfiguras evidencian que el método distribuye por igual todas las conexiones en el área cubierta por el anillo de servicio, independientemente de la topología del área o de la alta (Fig. 4.9(a)) o baja (Fig. 4.9(b)) densidad de población. La única variabilidad en la distribución de conexiones se observa en la Fig. 4.9(a), donde el mayor número de conexiones y los anillos superpuestos de TA muestran algunas irregularidades en las áreas de solapamiento de los anillos.

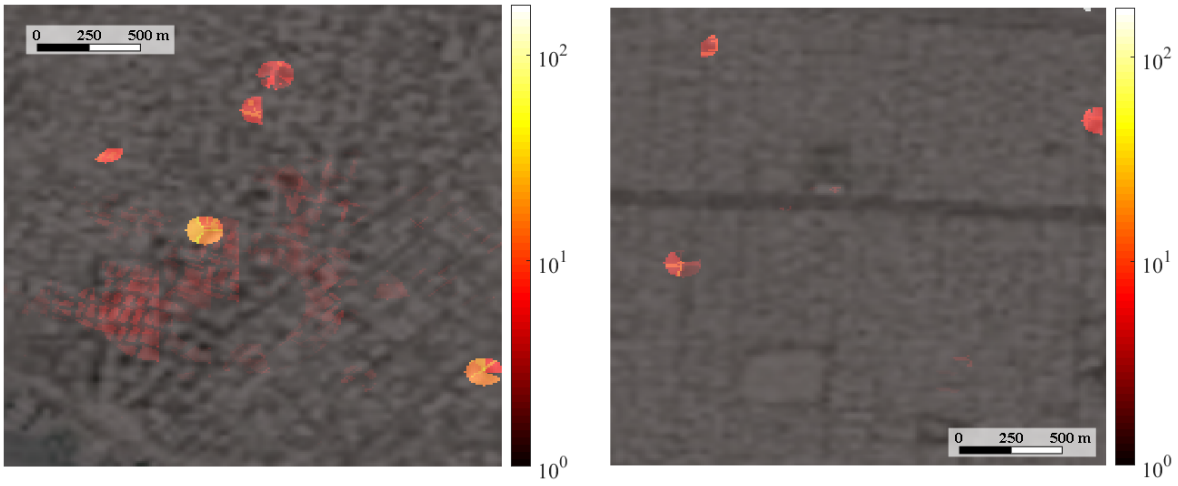
Del mismo modo, la Fig. 4.9(c)(d) presenta los resultados obtenidos en las mismas zonas con el modelo de contexto (es decir, añadiendo el contexto interior al método E-CID). Esta metodología de posicionamiento aprovecha el contexto interior para segregar las conexiones entre edificios y espacios abiertos. Aunque este comportamiento se puede observar en áreas con picos de conexiones de ambas subfiguras, esto es más evidente en la Fig. 4.9(c), donde la cantidad de conexiones permitió diferenciar las conexiones realizadas en los edificios (mucho más numerosas) frente a las de las calles (mucho menos frecuentes), pudiendo distingir los polígonos de los bloques de edificios definidos en $M_{lu}(x, y)$.

Finalmente, la Fig. 4.9(e)(f) ilustra los resultados del modelo completo, donde el método E-CID se mejora tanto con la información espacial del contexto interior como con la de las redes sociales. Con la inclusión de la información social, el modelo de contexto mejora mucho en las zonas muy pobladas (Fig. 4.9(e)), permitiendo obtener una resolución mucho mayor en el proceso de localización. Este incremento de resolución se observa en la granuralidad del mapa de calor, perdiendo en la mayoría de los casos las



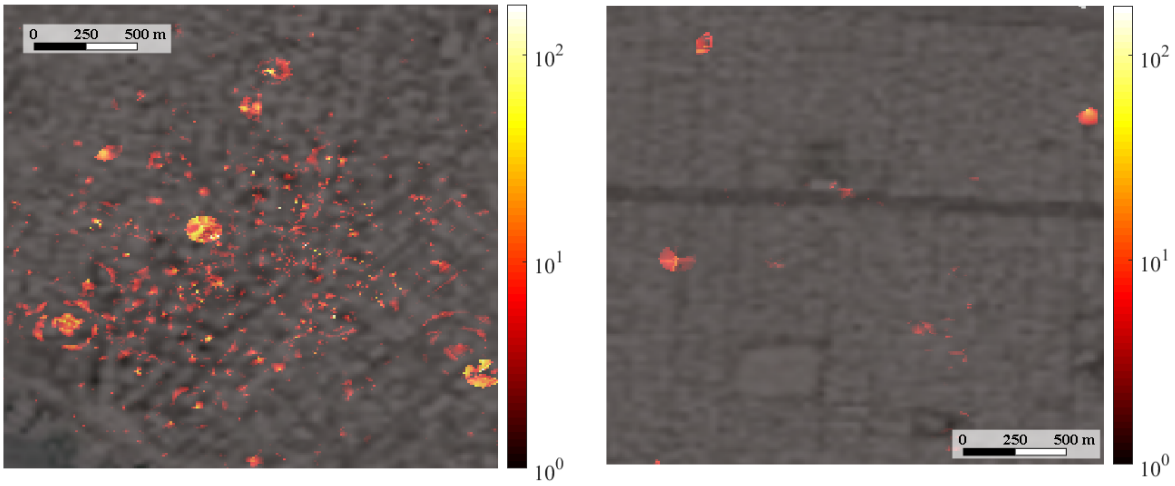
(a) E-CID, alta densidad.

(b) E-CID, baja densidad.



(c) Contexto, alta densidad.

(d) Contexto, baja densidad.



(e) Completo, alta densidad.

(f) Completo, baja densidad.

Figura 4.9: Comparación de modelos de distribución de conexiones en el escenario.

formas de anillo vistas en Fig. 4.9(a) y Fig. 4.9(c) y apareciendo picos de conexiones en zonas en las que no se podía observar ningún anillo en los métodos anteriores. Gracias a esta mejora de la resolución, es posible identificar el edificio dentro del anillo o la celda desde donde se genera la mayor parte del tráfico. Por el contrario, las zonas relativamente despobladas (Fig. 4.9(f)) mejoran mucho menos con la información social, comportándose de forma similar al modelo de contexto (Fig. 4.9(d)).

Para comprobar que la inclusión de la componente de publicaciones sociales en el modelo de posicionamiento no fuerza los resultados a realizar una copia del mapa de distribución de publicaciones sociales, se ha realizado un análisis comparativo entre el mapa de densidad de publicaciones subyacente y los resultados obtenidos por el modelo completo. Este estudio se ha realizado tanto para la zona de alta densidad de población (Fig. 4.7(a) y Fig. 4.9(e), respectivamente) como para la zona de baja densidad de población (Fig. 4.7(b) y Fig. 4.9(f)).

En la zona de alta densidad de población, se observan dos núcleos poblacionales, ambos en la parte inferior de la Fig. 4.7(a), uno en el borde izquierdo y otro en el borde derecho. Estos núcleos se ven reflejados, en parte, en el mapa de distribución de conexiones obtenido por el modelo completo, mostrado en la Fig. 4.9(e), donde el pico de agregación de publicaciones sociales en la zona inferior derecha coincide aproximadamente con el primer anillo de las celdas que forman un emplazamiento trisectorial, abarcando un elevado número de conexiones. Se observa que las conexiones se agregan heterogéneamente dentro del área de dichos anillos, respetando la distribución de publicaciones sociales en dichas áreas.

Por el contrario, el núcleo de publicaciones sociales observado a la izquierda no coincide con un pico de conexiones, sino con una superposición de varios anillos de celdas cercanas. En este caso, el modelo distribuye las conexiones en función del mapa de publicaciones sociales subyacente y del contexto de interior de las conexiones, desdibujando las áreas de los anillos en la zona, con unos niveles de agregación de conexiones más homogéneos que en el mapa de publicaciones sociales subyacente. Además, en la Fig. 4.9(e) se detecta un núcleo de agregación de conexiones que no está presente en la Fig. 4.7(a), en torno al centro de la figura, similar al primer caso estudiado. Aquí, la distribución espacial de publicaciones, aun teniendo unos niveles de agregación mucho más homogéneos, ayuda a obtener unos niveles de agregación de conexiones mucho más heterogéneos, detectando zonas relativamente vacías en anillos altamente ocupados.

En cambio, en zonas con una baja densidad de población, no se observa una corre-

lación clara entre el mapa de densidad de publicaciones subyacente (Fig. 4.7(b)) y el mapa de distribución de usuarios con el modelo completo (Fig. 4.9(f)), a pesar de la leve mejora con la inclusión de la información social respecto al modelo de contexto (Fig. 4.9(d)). Este comportamiento puede deberse al bajo número tanto de conexiones como de publicaciones sociales en el área.

Ante la falta de un conjunto de datos geolocalizados para demostrar la validez de este modelo, las figuras mostradas ponen de manifiesto que la información proporcionada por las publicaciones geolocalizadas, cuando está disponible, mejora significativamente el proceso de localización de conexiones sin condicionar completamente los resultados. Para las zonas de alta densidad de población mostradas en las Fig. 4.9(a)(c)(e), donde la actividad en redes sociales es lo suficientemente alta como para generar un número significativo de publicaciones geolocalizadas, el modelo completo consigue un aumento significativo de la resolución de la distribución de las conexiones. La granularidad del modelo completo (Fig. 4.9(e)) se aprecia en puntos concretos del mapa, donde la concentración de usuarios se incrementa notablemente respecto a los modelos E-CID y de contexto (Fig. 4.9(a)(c)), en los que la distribución de conexiones es más uniforme. En cambio, en las zonas de baja densidad de población, los métodos de contexto y completo se comportan de forma similar (Fig. 4.9(d)(f)), por la falta de publicaciones geoposicionadas. Estas subfiguras también señalan que, como era de esperar, los escenarios de alta densidad de población (Fig. 4.9(a)(c)(e)) abarcan un mayor número de conexiones que las zonas relativamente despobladas, lo que justifica el uso de la densidad de población para ponderar las zonas del anillo donde es más probable que se establezca una conexión.

4.4.3. Complejidad computacional

Para realizar el posicionamiento estadístico de conexiones, primero es necesario preprocesar los datos de entrada. En esta fase, la tarea que mayor carga computacional necesita es el proceso de *kriging*, cuya complejidad temporal es $\mathcal{O}(N^3)$ [86] para resolver $N + 1$ ecuaciones lineales para N puntos de origen. Una vez que se dispone de los datos de entrada, la complejidad computacional del método completo viene dada por el algoritmo de construcción de la distribución espacial de las conexiones. El algoritmo distribuye $N_{con}(r)$ conexiones originadas en el área $R_{ta}(r)$ cubierta por un anillo r de los N_r anillos en el escenario. Así, la complejidad temporal en el peor de los casos es $\mathcal{O}(N_r\, R_{ta}\, N_{con})$.

En la implementación del método, se han combinado diferentes herramientas. El procesamiento de las trazas se realiza mediante el procesado de eventos complejos con rutinas *Esper* [87]. Por otro lado, los usos del suelo se procesan con Matlab, mientras que los datos de redes sociales se obtienen con la Streaming API de Twitter usando Java a través de la librería Twitter4j y la API de Flickr usando Python[82, 88, 83], procesándolos con Matlab. Por último, el modelo de posicionamiento estadístico se ha implementado usando el *toolbox* de procesado de imagen en Matlab.

Todos estos procesos se han ejecutado en un servidor con un procesador de 2,4 GHz y 64 GB de RAM. Para procesar las trazas de conexión (400 celdas, 2 horas de trazas, 166.561 conexiones), el tiempo necesario es de 282 segundos. Asimismo, los tiempos para construir los mapas de uso de suelo, $M_{lu}(x, y)$, y densidad de publicaciones, $M'_p(x, y)$, (125 km^2, 1.222.787 casillas) son de 50.444 y 57.743 segundos, respectivamente. Para concluir, el tiempo empleado en el posicionamiento de 166.561 conexiones es de 160 segundos.

4.5. Conclusiones

El conocimiento de la distribución espacial de los terminales de usuario es vital a la hora de realizar tareas de gestión de las redes celulares. En este capítulo, la distribución de conexiones se infiere mediante un algoritmo que, apoyándose en la información de contexto del usuario y publicaciones geolocalizadas de redes sociales, determina en qué cuadrículas del escenario es más probable que ocurra una conexión.

Los resultados obtenidos indican que la distribución de publicaciones geoposicionadas de redes sociales suministra información muy relevante de las áreas donde se produce una aglomeración de usuarios, principalmente en zonas de alta densidad de población, aportando una granularidad al proceso de posicionamiento de conexiones imposible con el posicionamiento por E-CID. Además, la información del contexto de los usuarios, aunque en menor medida, mejora también la precisión del posicionamiento, pudiendo segregar las conexiones en edificios o espacios abiertos. Esto hace que la inclusión de ambas fuentes de datos sea interesante a la hora de conocer la distribución de los usuarios en la red.

El análisis de la distribución de publicaciones en el escenario refleja que, como era de esperar, la información espacial que aporta cada red social depende enormemente de su popularidad. En el caso estudiado, la distribución de publicaciones viene deter-

minada casi por completo por Twitter, suministrando un $95,9\,\%$ de las publicaciones sociales de la base de datos, dejando a Flickr como un refuerzo en determinados puntos del escenario. Queda por comprobar si, añadiendo redes sociales de popularidad similar, pero con una política de publicación de datos más restrictiva (p.ej., Instagram o Facebook), se mejoraría la granularidad en áreas del escenario donde una sola red social no aporta suficiente información.

El modelo presentado en este capítulo sirve como apoyo en tareas de planificación y operación de red donde la posición de los usuarios es un factor relevante, como la implementación de nuevas estaciones base o la gestión de recursos radio en áreas con problemas de capacidad debidos a aglomeraciones de usuarios. Las tareas de construcción de mapas de usos de suelo y de publicaciones son las que mayor carga computacional tienen, pero, una vez realizadas, los resultados pueden almacenarse para futuras aplicaciones del proceso de posicionamiento. Por otro lado, la baja complejidad computacional del método permite una fácil integración en las herramientas de planificación radio. Además, la diversidad de las fuentes de datos utilizadas permite aprovechar el potencial de los últimos sistemas de gestión de redes basados en procesado de datos masivos.

Capítulo 5

Selección automática de emplazamientos de picoceldas con datos multimodales

Conocida la distribución de los usuarios en la red, puede evaluarse la eficiencia de la utilización de los recursos radio en cada punto del escenario. Esta evaluación permite detectar problemas de falta de cobertura o capacidad. Para solventar dichos problemas, una posible solución es la inclusión de nuevas estaciones base de corto alcance (picoceldas) o celdas en interior de edificios (femtoceldas) en áreas donde las estaciones base actuales no son capaces de servir correctamente a los usuarios, ya sea por problemas de propagación o por falta de recursos debido a la gran cantidad de conexiones servidas. Este es el objetivo de la contribución descrita en este capítulo, donde se determina la posición óptima de nuevas picoceldas/femtoceldas basándose en el comportamiento de los usuarios espacialmente distribuidos en el escenario.

De la misma forma que en capítulos anteriores, en la sección 5.1 se comentan diferentes aproximaciones al problema de colocación de antenas nuevas en un escenario. En la sección 5.2 se formulan las hipótesis de partida. A continuación, la sección 5.3 presenta la tercera contribución de esta tesis, un algoritmo de detección automática de la posición óptima de nuevas celdas pequeñas utilizando la distribución espacial de conexiones descrita en el capítulo anterior. La sección 5.4 presenta la validación del algoritmo en un escenario real. Para concluir, en la sección 5.5 se extraen una serie de conclusiones del trabajo desarrollado.

5.1. Estado de la técnica

El problema de la colocación de antenas (*Antenna Placement Problem*, APP) puede formularse como un problema de optimización clásico. Las variables de decisión son las coordenadas de la estación base, a menudo restringidas a un conjunto limitado de ubicaciones candidatas, mientras que la función objetivo puede ser cualquier combinación de indicadores de rendimiento global de la red. Por razones computacionales, este problema de optimización combinatoria suele resolverse mediante enfoques heurísticos. Los trabajos anteriores pueden clasificarse por tipo de entorno, criterios de diseño y algoritmo de solución.

En cuanto al entorno radioeléctrico, un primer grupo de trabajos se ocupa de escenarios exteriores macrocelulares y microcelulares. En [89], se utiliza el recocido simulado (*Simulated Annealing*, SA) para resolver el APP en un entorno de microceldas TDMA/FDMA basado en indicadores de relación señal-interferencia-más-ruido (SINR) y pérdidas de trayecto. En [90], el APP en WCDMA se formula como un problema de programación lineal entera, resuelto mediante búsqueda tabú. En [91], el APP se formula para encontrar el número mínimo de antenas para un nivel de cobertura deseado. En [92], el objetivo del APP es maximizar la cobertura en GSM mientras se satisface un requisito mínimo de SINR, lo que se consigue mediante algoritmos genéticos. En [93] y [94], se realiza un análisis de sensibilidad para comprobar el impacto de la ubicación del emplazamiento y los ángulos de inclinación de la antena en la capacidad de una red WCDMA con una distribución desigual del tráfico. En [95], se utilizan la búsqueda local aleatoria y la búsqueda tabú para resolver el APP con el fin de optimizar conjuntamente los costes de instalación, la calidad de la señal y la cobertura del tráfico en WCDMA.

Un segundo grupo de trabajos extiende los métodos anteriores a escenarios de interior. En algunos de ellos, el APP se formula para minimizar la pérdida de trayecto (o maximizar la cobertura) con un algoritmo de optimización de propósito general (p.ej., genético [96], búsqueda directa [97], SA [98] o heurístico [99]). De forma similar a [91], [100] propone la programación entera binaria para encontrar el número mínimo de puntos de acceso que garanticen una SINR mínima en el escenario. En [101], se propone un método heurístico para colocar puntos de acceso en interiores en WCDMA con restricciones en la SINR del enlace ascendente (UL) y descendente (DL). Estudios posteriores se centran en la optimización de la SINR mediante diferentes métodos (p.ej., la enumeración por fuerza bruta en WCDMA [102] y LTE [103], el enjambre

de partículas en WCDMA [104] y la aproximación por reducción en WCDMA [105]). En [106], se propone un método para la colocación de femtoceldas para minimizar la potencia de transmisión de los usuarios móviles. En [107], se propone una metodología para ubicar las femtoceldas empresariales en un edificio para maximizar la eficacia de los esquemas de equilibrio de carga de movilidad. Por otro lado, en [108, 109], el objetivo es encontrar la mejor ubicación del punto de acceso Wi-Fi para el posicionamiento óptimo del usuario en entornos interiores. Más recientemente, en [110] se presenta una visión general de las estrategias de despliegue de picoceldas o *Small Cells* (SC) para entornos de Internet de las Cosas (IoT) 5G.

Las redes sociales pueden proporcionar información valiosa para la APP. En [17], se muestra que la actividad de las redes sociales agregada a nivel de distrito puede utilizarse para predecir el tráfico celular con una resolución espaciotemporal superior a la de los enfoques actuales, basados en datos censales. Igualmente importante es el hecho de que los datos de los medios sociales reflejan la demanda global de tráfico a través de las tecnologías de acceso radioeléctrico o los proveedores de red. Del mismo modo, las cargas de los sistemas celulares pueden estimarse mediante un modelo de colas ajustado con la distribución de mensajes geoetiquetados de Twitter [111]. No obstante, aún está por comprobar si los datos de las redes sociales pueden utilizarse para predecir el tráfico a una escala inferior (p.ej., a nivel de edificio).

El enfoque clásico en los problemas de APP está basado en indicadores de propagación radio, dejando a un lado la experiencia del usuario. Este hecho motiva el desarrollo del algoritmo descrito en este capítulo. Las principales contribuciones del modelo descrito son: a) combinar las mediciones de la red celular con el contexto del usuario y los datos de las redes sociales para detectar los puntos problemáticos de la red con resolución de edificio, y b) cuantificar el beneficio de rendimiento esperado de cada nueva SC, basándose en el comportamiento observado del usuario y el que previsiblemente tendrá en condiciones óptimas del enlace radio.

5.2. Formulación del problema

El despliegue de una nueva SC puede aumentar el nivel de señal (cobertura), la calidad de la señal (eficiencia espectral) y los recursos disponibles (ancho de banda), mejorando en última instancia la capacidad del enlace percibida por el usuario. Es de esperar que, con la mejora de los recursos radio disponibles, la ganancia de rendimiento obtenida sea mucho mayor para los usuarios de servicios que requieren un gran volumen

de datos (p.ej., videostreaming, descarga de aplicaciones...) que para los usuarios de servicios que solicitan poco volumen de datos (p.ej., mensajería instantánea, llamadas de voz...). Por desgracia, en las redes de acceso radio actuales no se realiza una clasificación de las conexiones [112]. A falta de una clasificación precisa de los servicios, se pueden utilizar las estadísticas de transmisión del último TTI para identificar los servicios de gran volumen de datos. Como se comentó en el Capítulo 3, un último TTI es el último intervalo de transmisión de una ráfaga de datos que vacía temporalmente el búfer de transmisión del usuario en el enlace descendente o ascendente [113]. Por lo tanto, los servicios que requieren un uso intensivo de datos (es decir, que precisan de una o más ráfagas grandes de datos, ocupando un elevado número de recursos) suelen tener bajos ratios de último TTI (ecuación (3.2)). Por el contrario, los servicios no intensivos en datos, que consisten en pocas ráfagas cortas de datos, el ratio de último TTI se dispara. Por estas razones, los servicios que requieren una gran cantidad de datos tienden a comportarse como un servicio ideal de búfer completo, mientras que los servicios no intensivos en datos pueden modelarse como tráfico a ráfagas cortas.

Sobre esta hipótesis, se identifican dos fuentes principales de ganancia de tráfico en la red. La primera componente de ganancia procede de los usuarios que ya utilizan servicios intensivos en datos (es decir, de búfer completo) sin la SC, que aumentarán el tráfico que generan como resultado de una mayor duración de la sesión y/o la calidad del contenido que solicitan, debido a una mejor experiencia de usuario. La segunda fuente de ganancia de tráfico proviene de los usuarios que, en el pasado, sólo solicitaban servicios no intensivos en datos (es decir, conexiones de último TTI) debido al mal comportamiento de la red, pero que, con mejores condiciones, solicitan también servicios de búfer completo debido a la mejor experiencia de usuario. Además, es de esperar que el comportamiento de los usuarios dependa de su contexto, por lo que ambas fuentes de ganancia de tráfico se deben estudiar en función de la situación del usuario (interior o exterior). A continuación, se realiza un análisis en mayor detalle de cada una de las fuentes de ganancia potencial en la red para comprobar su validez.

5.2.1. Impacto de la mejora de tasa binaria con búfer completo

Para comprobar cómo se ve afectado el volumen de datos de los usuarios de los servicios de búfer completo por las condiciones de la red, se estudia la relación entre el volumen de datos, V, y el caudal de los usuarios (*throughput*), T, por conexión k. La

Fig. 5.1 muestra un gráfico de dispersión de estas dos variables observadas en conexiones móviles reales (un punto por conexión). El valor del caudal en el eje de las abscisas corresponde al caudal medio de la conexión en la capa del Protocolo de Convergencia de Datos de Paquetes (*Packet Data Convergence Protocol*, PDCP) excluyendo los últimos TTI (lo mismo para el volumen de datos en el eje de las ordenadas) [114]. Dado que la atención se centra en los servicios que no son de último TTI, sólo se consideran en el análisis las conexiones con un ratio de volumen en último TTI inferior al 10 % (es decir, las conexiones que pueden modelarse como búfer completo). Para mayor claridad, en la figura se incluyen tres curvas de regresión. La curva central se obtiene por regresión lineal en todas las muestras (nótese la escala logarítmica en el eje x), y por tanto refleja la tendencia general. Las otras dos curvas se calculan mediante regresión cuantílica para reflejar la tendencia de los casos extremos. En concreto, la curva inferior/superior se obtiene mediante una regresión lineal sobre el percentil 10 %/90 % del volumen de datos de las conexiones en diferentes rangos de caudal. En las tres curvas, se puede observar que el volumen de datos por conexión aumenta con el caudal medio de los usuarios, lo que demuestra la hipótesis de que la mejora de las condiciones del enlace radioeléctrico suele conducir a un mayor volumen de datos por conexión. Como era de esperar, la correlación no es fuerte, ya que dos conexiones pueden tener volúmenes de datos muy diferentes para el mismo valor de caudal debido a las diferentes longitudes de las sesiones o a las diferentes calidades/servicios disponibles. Esta observación justifica la necesidad de varias curvas de regresión para modelar la relación entre el caudal y el volumen de datos por conexión. Un análisis más detallado de los residuos muestra que el error absoluto medio se reduce de $4,935 \times 10^6$ con una sola curva a $2,25 \times 10^6$ con 3 curvas, consiguiendo una reducción del 45,64 %.

5.2.2. Impacto de la mejora de tasa binaria con búfer completo en conexiones de tráfico a ráfagas

De la misma forma, para comprobar si el número de solicitudes de servicios de último TTI de un usuario se ve afectado por las condiciones de la red, se examina la correlación entre el ratio de conexiones de búfer completo respecto a las conexiones a ráfagas, R_{FB}, y el caudal. Al igual que en el caso anterior, se considera que una conexión es una conexión de último TTI si se transmite más del 90 % de los datos en los últimos TTI de las ráfagas, mientras que la conexión se considera de búfer completo si menos del 10 % de los datos transmitidos son en últimos TTI. Todas las demás conexiones que

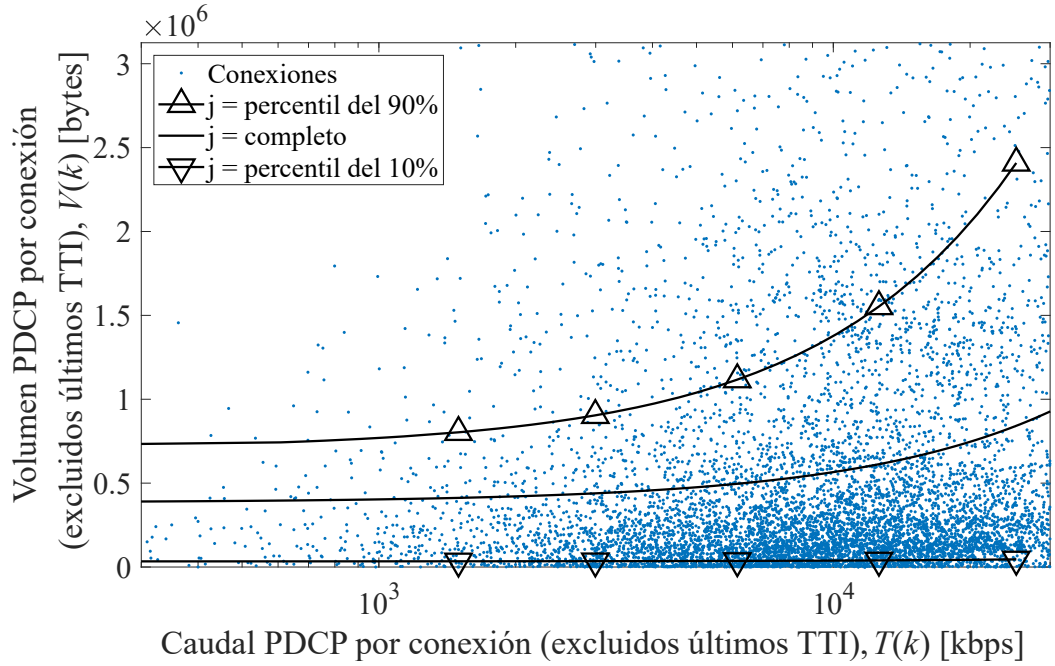


Figura 5.1: Gráfico de dispersión del volumen de datos frente al caudal del enlace de radio por conexión.

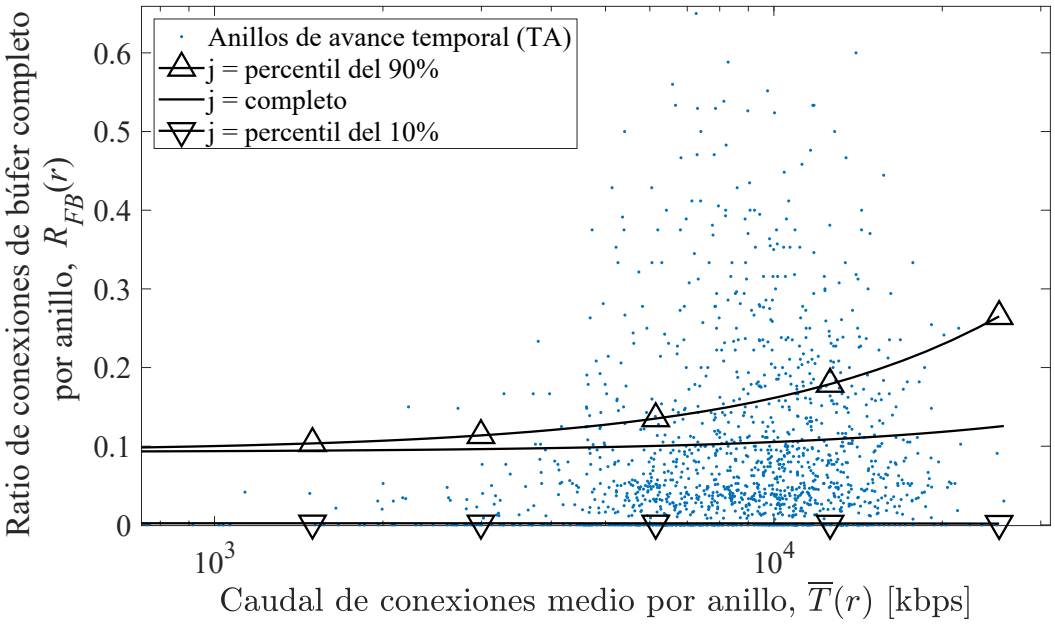


Figura 5.2: Gráfico de dispersión del ratio de conexiones de búfer completo frente al caudal por anillo de avance temporal (TA).

se encuentran entre estos dos umbrales se descartan en el análisis. Como se explicó en el Capítulo 4, en ausencia de trazas geolocalizadas, las conexiones se posicionan en función del E-CID. Así, el análisis puede realizarse agregando todas las conexiones dentro de

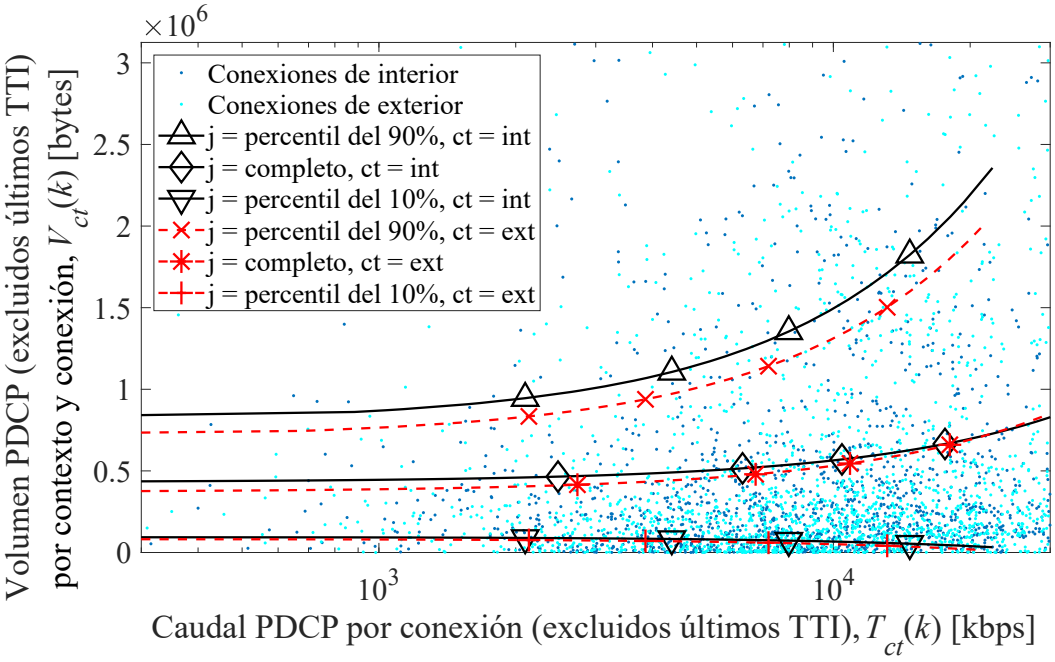


Figura 5.3: Gráfico de dispersión del volumen de datos frente al caudal del enlace de radio por conexión, separado por contexto.

un anillo de TA. Para que los datos estudiados sean estadísticamente relevantes, sólo se consideran los anillos con más de 25 conexiones, representando aproximadamente el 20 % de los anillos y cubriendo el 70 % de las conexiones en la red. La Fig. 5.2 muestra un gráfico de dispersión del ratio de conexiones de búfer completo/último TTI frente al caudal medio de las conexiones en el anillo, $\bar{T}$ (1 punto por celda-anillo). De nuevo, se superponen tres curvas de regresión, obtenidas por regresión lineal y cuantílica. En las curvas de tendencia general y del percentil 90 %, se confirma que los anillos con un mayor caudal medio tienden a tener mayores ratios de conexiones de búfer completo. Asimismo, la gran dispersión justifica la necesidad de varias curvas de regresión también en este caso (en este caso, el error medio absoluto con 3 curvas se reduce en un 53,28 %).

5.2.3. Efecto del contexto de interior en el impacto de la mejora de tasa binaria con búfer completo

Los análisis anteriores pueden realizarse para las conexiones interiores y exteriores por separado, para comprobar el impacto del contexto del usuario en la ganancia de tráfico al mejorar los recursos disponibles. A falta de un método preciso de posicionamiento del usuario, las conexiones se etiquetan como interiores o exteriores basándose en el algoritmo descrito en el Capítulo 3. Para evitar datos confusos que lleven a im-

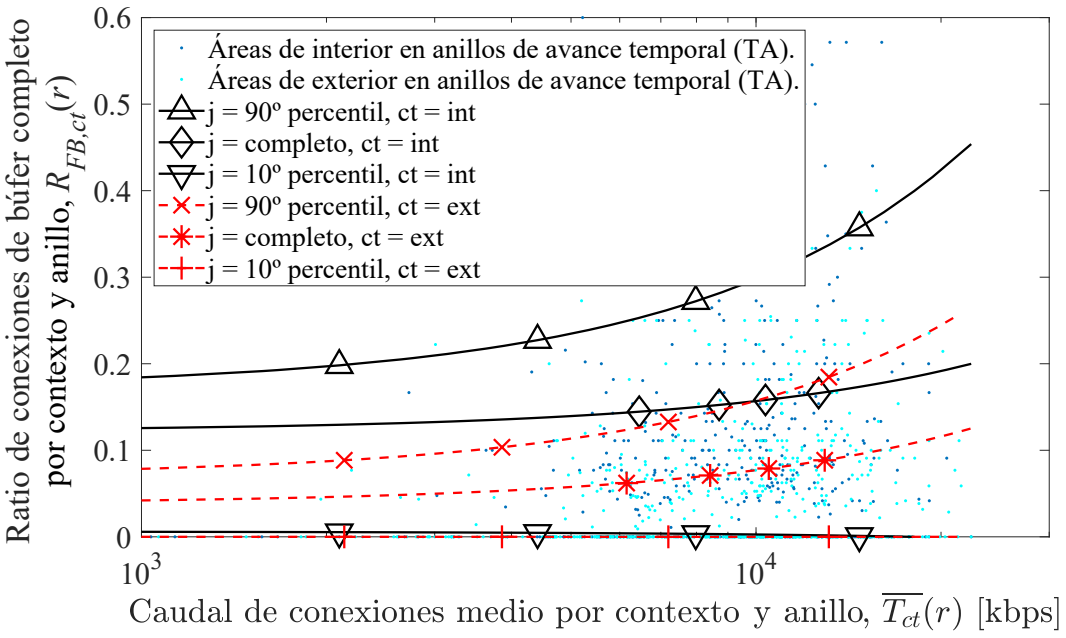


Figura 5.4: Gráfico de dispersión del ratio de conexiones de búfer completo frente al caudal por anillo de TA, separado por contexto.

precisiones, solo las conexiones con una gran confianza de ser interiores o exteriores (probabilidad de interior >90 % o <10 %) se consideran en el análisis. En la Fig. 5.3 se presentan los resultados del desglose de las curvas de regresión de la Fig. 5.1 por contexto. En esta figura, las curvas de exterior e interior difieren solo ligeramente para el volumen de datos de conexión, lo que indica que el volumen de datos por conexión no se ve muy afectado por el contexto del usuario.

Al igual que con la mejora potencial en conexiones de búfer completo, se repite el estudio incluyendo el contexto de las conexiones. En este caso, al contrario que pasa en la mejora obtenida de las conexiones de búfer completo, en la Fig. 5.4 las curvas de interior superan ampliamente a las de exterior. De ello, se deduce que, para las mismas condiciones de caudal de los usuarios, la proporción de servicios intensivos en datos frente a los no intensivos es mayor en interior que en exterior. Esta observación justifica la necesidad de un conjunto diferente de curvas de regresión para interiores y exteriores.

De este último análisis se derivan dos modelos de regresión, cada uno de los cuales consta de 6 curvas de regresión. El primer modelo define la relación entre el volumen de datos, V, y el caudal medio de la conexión (en periodos de búfer completo), T, en una conexión k que tuvo lugar en el contexto ct como

$$V^{(j)}(k) = \beta^{(j)}_{1,V,ct} T(k) + \beta^{(j)}_{0,V,ct} \quad k \in ct\,, \tag{5.1}$$

donde $\beta^{(j)}_{i,V,ct}$ es el coeficiente de regresión i de la curva de regresión j ($j \in \{10\,\%, \textit{completo}, 90\,\%\}$) y el contexto de usuario ct ($ct \in \{int, ext\}$ para interiores y exteriores, respectivamente). La regresión $j = \textit{completo}$ es una regresión lineal simple con todas las muestras, mientras que j=10 % y 90 % corresponden a una regresión cuantílica con el percentil 10 % y el percentil 90 % de los valores de volumen de datos por franja de caudal, respectivamente. De la misma forma, el segundo modelo define la relación entre la proporción de conexiones de búfer completo frente a conexiones a ráfagas, R_{FB}, y el caudal medio de todas las conexiones, $\overline{T}$, en un anillo TA r en el contexto ct como

$$R^{(j)}_{FB,ct}(r) = \beta^{(j)}_{1,R,ct} \overline{T}_{ct}(r) + \beta^{(j)}_{0,R,ct}\,, \tag{5.2}$$

donde$\beta^{(j)}_{i,R,ct}$ es el coeficiente de regresión i para la curva de regresión j y el contexto de usuario ct.

5.3. Modelo automático de selección de nuevos emplazamientos

El objetivo del modelo presentado en la Fig. 5.5 es determinar, en una red celular operativa, las zonas en las que puede aumentarse la demanda de tráfico mejorando las condiciones de los enlaces radio. Para ello, se estudian los factores más influyentes en la demanda de tráfico, que son: a) la distribución espacial de los usuarios, b) el caudal máximo de la red por ubicación y c) el comportamiento de los usuarios en función del contexto.

Al igual que el modelo descrito en el Capítulo 4, el modelo recibe 4 entradas: (a) la información registrada en las trazas de conexión radio, (b) datos de usos del suelo en el escenario, (c) información de la ubicación y orientación de las celdas en el escenario y (d) publicaciones de redes sociales que tienen metadatos de geoposicionamiento. Con estos datos, el modelo devuelve como salida una lista de emplazamientos de nuevas picoceldas, definiendo su posición geográfica (latitud y longitud), la ganancia de tráfico

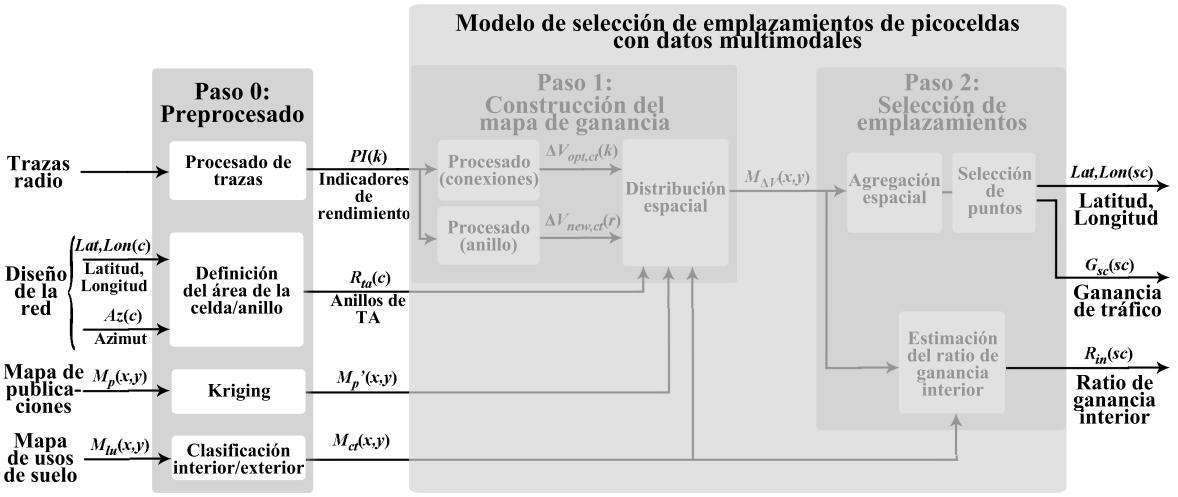


Figura 5.5: Flujograma del modelo propuesto.

que se espera obtener y un ratio de conexión interior, que determina su localización dentro o fuera de los edificios.

El modelo está compuesto por dos pasos: a) un primer paso para construir un mapa de cómo se incrementaría el tráfico en cada cuadrícula del escenario si se despliega una nueva SC y, b) un segundo paso para seleccionar las posiciones que mayor impacto tengan en la generación de tráfico. Para que el modelo funcione correctamente, es necesario un paso previo que adapte la información de entrada a la forma deseada. Este paso de preprocesado es el mismo que el preprocesado del capítulo anterior, por lo que no se describe en este capítulo.

5.3.1. Paso 1: Construcción del mapa de ganancia de tráfico en la red

Como se ha demostrado en la sección 5.2, la ganancia de tráfico obtenida por el despliegue de una nueva SC se divide en dos componentes: a) el aumento de tráfico procedente de las conexiones existentes, calculado por conexión (sólo aplicable a las conexiones que necesitaban más recursos, en las que no prima el último TTI de cada ráfaga), y b) el aumento procedente de nuevas conexiones de servicios de uso intensivo de datos, calculado por anillo. Por tanto, para detectar la posición óptima de la nueva SC, es necesario conocer en qué áreas del escenario se alcanza un máximo en la combinación de ambas componentes de ganancia. Para ello, se sigue el procedimiento mostrado en la Fig. 5.6, donde primero se procesan los indicadores de rendimiento para

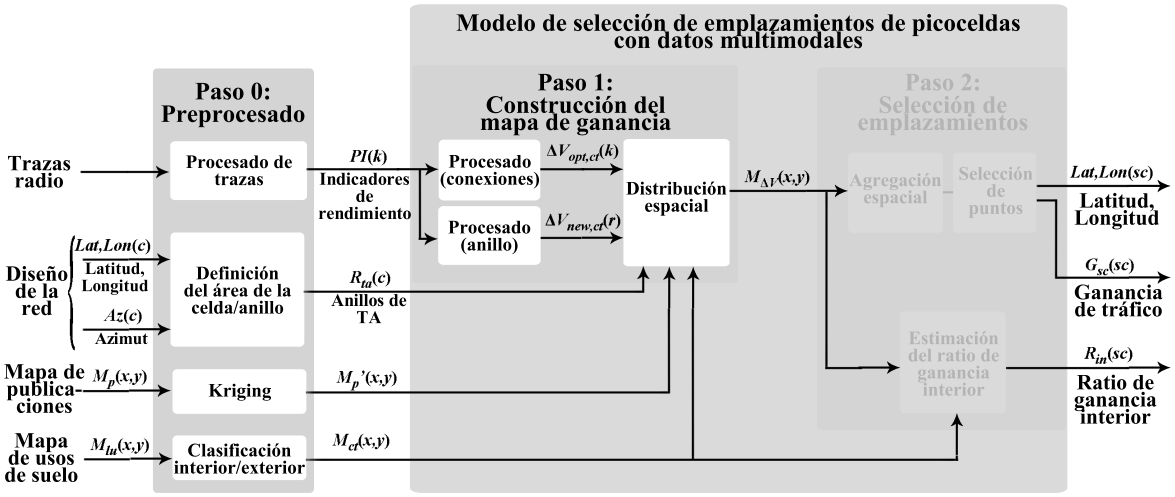


Figura 5.6: Flujograma del modelo propuesto. Paso 1: Construcción del mapa de ganancia de tráfico en la red.

obtener las componentes de ganancia de tráfico y después se distribuyen en el mapa, de forma similar al posicionamiento de conexiones descrito en el Capítulo 4.

Como primer paso para obtener las componentes de ganancia de tráfico en la red, se define un valor de caudal óptimo, T_{opt}, como el mejor caudal que puede alcanzarse con unas condiciones de propagación óptimas. Para evitar valores anómalos debidos a conexiones ocurridas en celdas prácticamente vacías, donde todos los recursos están disponibles, T_{opt} se fija en el percentil 90 % del caudal medio de todas las conexiones de la red.

A continuación, las conexiones se clasifican como interiores o exteriores. Para ello, la probabilidad de que una determinada conexión se genere en un contexto determinado, $P_{ct}(k)$, se estima a partir de los datos de traza con el algoritmo descrito en el Capítulo 3. Como resultado, una conexión podría ser etiquetada tanto como interior como exterior, con diferentes pesos dados por la probabilidad de interior.

Siguiendo la clasificación anterior, la ganancia potencial de volumen de datos de una conexión existente se divide en una componente de exterior y otra de interior, calculadas de la siguiente forma

$$\Delta V_{FB,ct}(k) = P_{ct}(k)(V'_{ct}(k) - V(k))\,, \tag{5.3}$$

donde $P_{ct}(k)$ es la probabilidad interior/exterior de la conexión, $V(k)$ es el volumen de datos a nivel PDCP de la conexión excluyendo su último TTI, tomado de las trazas, y

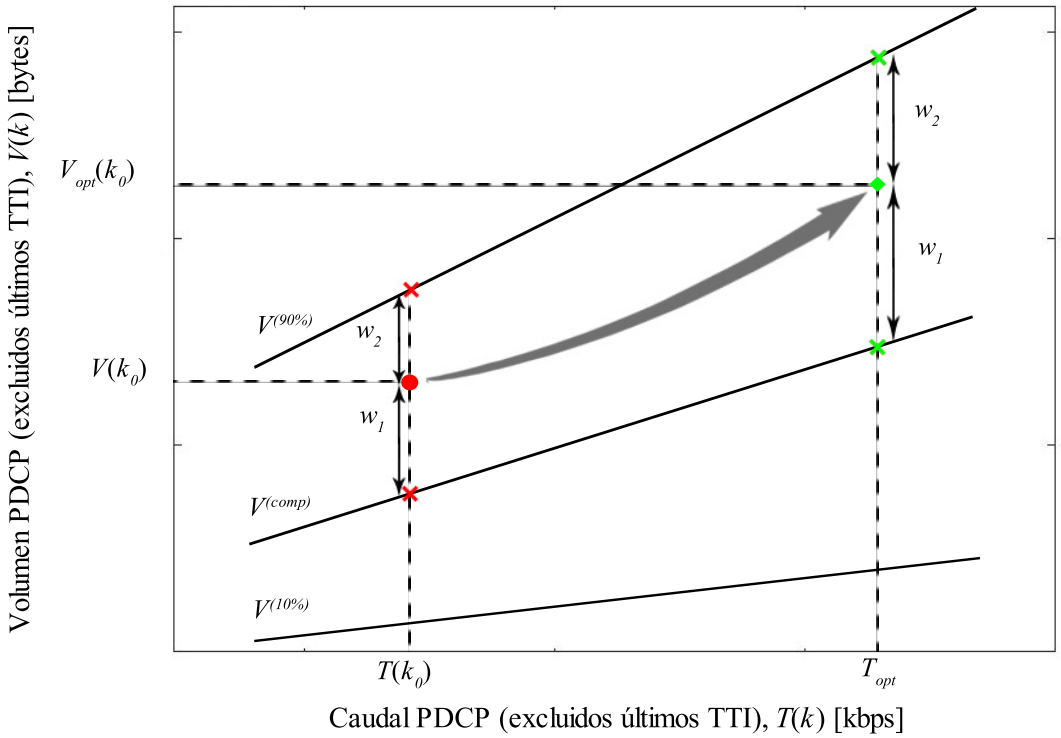


Figura 5.7: Método de estimación de valores óptimos interpolando curvas de regresión.

$V'_{ct}(k)$ es el volumen potencial de datos interior/exterior de la conexión con condiciones óptimas radio (es decir, después de desplegar la nueva SC). Este último valor se estima mediante el modelo de regresión en la ecuación (5.1), usando como valor de caudal $T(k) = T_{opt}$. Aun así, hay que decidir la curva de regresión concreta entre las tres descritas en la subsección 5.2.3, j=10 %, *completa* o 90 %. Para resolver este problema, se utiliza la interpolación lineal entre las curvas más cercanas, tal como se muestra en la Fig. 5.7. En concreto, los coeficientes de ponderación se determinan como

$$w_1 = \begin{cases} 1 & \text{si} \quad V(k) \geq \widehat{V_{90}}(T(k)) \,, \\ \frac{V(k)-\widehat{V_{comp}}(T(k))}{\widehat{V_{90}}(T(k))-\widehat{V_{comp}}(T(k))} & \text{si} \quad \widehat{V_{comp}}(T(k)) \geq V(k) > \widehat{V_{90}}(T(k)) \,, \\ \frac{V(k)-\widehat{V_{10}}(T(k))}{\widehat{V_{comp}}(T(k))-\widehat{V_{10}}(T(k))} & \text{si} \quad \widehat{V_{10}}(T(k)) \geq V(k) > \widehat{V_{comp}}(T(k)) \,, \\ \frac{V(k)}{\widehat{V_{10}}(T(k))} & \text{si} \quad V(k) < \widehat{V_{10}}(T(k)) \,, \end{cases} \tag{5.4}$$

$$w_2 = \begin{cases} 0 & \text{si} \quad V(k) \geq \widehat{V_{90}}(T(k)) \,, \\ 1 - w_1 & \text{si} \quad \widehat{V_{10}}(T(k)) \geq V(k) > \widehat{V_{90}}(T(k)) \,, \\ 0 & \text{si} \quad V(k) < \widehat{V_{10}}(T(k)) \,, \end{cases} \tag{5.5}$$

donde w_1 y w_2 son los coeficientes de ponderación de las curvas más cercanas al valor del volumen generado por la conexión, definidos en las ecuaciones (5.4)-(5.5), respectivamente. De esta forma, si el volumen generado se encuentra entre el percentil 90 % y la media de todas las conexiones de la red, se usa el w_1 definido en la segunda condición, mientras que si está entre la media y el percentil 10 % se aplica el w_1 definido en la tercera condición. En ambos casos se aplica w_2 definido en la condición central. Para reducir la influencia de valores de volumen anómalos en los datos de las trazas, estos coeficientes están delimitados por los valores de las curvas de los percentiles 90 y 10 (primera y última condición en las ecuaciones (5.4) y (5.5)).

Una vez obtenida la ganancia potencial de volumen que cada conexión de búfer completo generaría en condiciones óptimas, se realiza el mismo proceso para la ganancia obtenida por la aparición de nuevas conexiones de este tipo en la red. Para ello, primero se calcula el número de nuevas conexiones de este tipo que aparecerían en cada anillo de la red con mejores condiciones radio. Esto se realiza aplicando el mismo proceso de ponderación mostrado en la Fig. 5.7, pero aplicado al ratio de conexiones de búfer completo respecto a conexiones a ráfagas. Con este ratio óptimo, se obtiene el número de nuevas conexiones de búfer completo. En concreto, primero se calculan el número de conexiones equivalentes de búfer completo, $N_{FB,ct}(r)$, y de último TTI, $N_{ultTTI,ct}(r)$, por contexto ct en el anillo r sin la nueva SC, obtenidos a partir de las características de las llamadas efectuadas en el anillo como

$$N_{FB,ct}(r) = \sum_{k \in r} P_{ct}(k) R_{Vol_{ultTTI}}(k), \tag{5.6}$$

$$N_{ultTTI,ct}(r) = \sum_{k \in r} P_{ct}(k)(1 - R_{Vol_{ultTTI}}(k)) \,, \tag{5.7}$$

donde $R_{Vol_{ultTTI}}(k)$ es la parte del volumen de datos transmitida en los TTI de búfer completo en la conexión k (p.ej., 4 conexiones en el mismo anillo con $P_{ct}(k) = 0,5$ y $R_{Vol_{ultTTI}}(k) = 0,5$ equivalen a 1 conexión ideal de búfer completo). A partir del

número de conexiones de búfer completo y de tráfico a ráfagas en cada anillo de la red, se calcula el aumento del número de conexiones de servicios intensivos de datos en condiciones ideales en cada contexto por anillo como

$$\begin{aligned}\Delta N_{new,ct}(r) &= N'_{FB,ct}(r) - N_{FB,ct}(r) \\ &= N_{ultTTI,ct}(r) R'_{FB,ct}(r) - N_{FB,ct}(r)\,,\end{aligned} \tag{5.8}$$

donde $N'_{FB,ct}(r)$ es el número previsto de conexiones de búfer completo previstas con una nueva SC en el contexto ct del anillo r, estimado a partir de la nueva relación de conexiones FB/ráfagas con la SC, $R'_{FB,ct}(r)$. Este último se estima mediante el modelo de regresión en (5.2) con el proceso de interpolación de curvas de la Fig. 5.7 y $\overline{T}_{ct}(r) = T_{opt}$. Por tanto, el aumento del volumen de datos en el anillo r se calcula como

$$\Delta V_{new,ct}(r) = \Delta N_{new,ct}(r)\, \overline{V_{opt,ct}}(r)\,, \tag{5.9}$$

donde $\overline{V_{opt,ct}}(r)$ es el volumen medio de datos de una conexión de búfer completo ideal en el anillo r y contexto ct con la nueva SC, calculado como

$$\overline{V_{opt,ct}}(r) = \frac{\sum_{k \in r,ct} V'_{ct}(k)}{N_{FB,ct}(r)}\,. \tag{5.10}$$

Una vez calculadas las dos componentes de ganancia de tráfico por conexión y por anillo, hay que proyectarlas en el mapa utilizando la localización de las conexiones. Para ello, se hace uso de una pequeña variación del algoritmo presentado en el Capítulo 4, donde, en lugar de distribuir estadísticamente las conexiones en las cuadrículas del mapa, se distribuye la parte porcentual de las ganancias de tráfico calculadas anteriormente. Para ello, el aumento del volumen de datos debido a las conexiones existentes (por conexión) o a las nuevas conexiones (por anillo) se distribuye en cuadrículas interiores/exteriores dentro del anillo según las probabilidades en (4.4) y luego se agrega entre los anillos que dan servicio a la misma cuadrícula.

En concreto, el aumento del volumen de datos debido a las conexiones existentes o nuevas en la cuadrícula (x, y), $\Delta V_{opt}(x, y)$ y $\Delta V_{new}(x, y)$, respectivamente, se calcula como

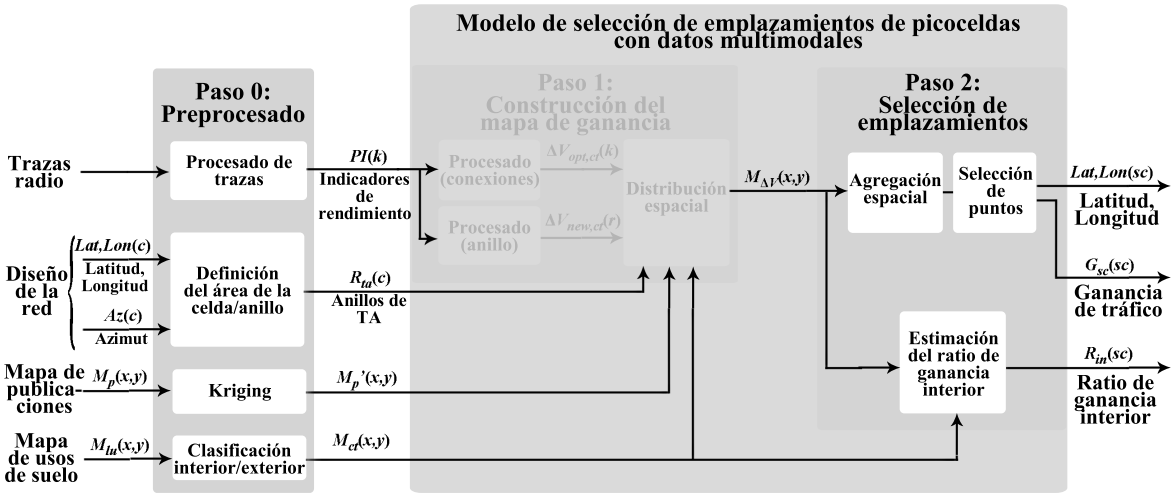


Figura 5.8: Flujograma del modelo propuesto. Paso 2: Selección de emplazamientos.

$$\Delta V_{opt}(x,y) = \sum_{r/(x,y)\in r} \sum_{k\in r} \Delta V_{opt,ct}(k) P(r,x,y) \quad (x,y) \in ct \,, \tag{5.11}$$

$$\Delta V_{new}(x,y) = \sum_{r/(x,y)\in r} \Delta V_{new,ct}(r) P(r,x,y) \quad (x,y) \in ct \,. \tag{5.12}$$

Ambos términos se agregan para construir el mapa de ganancia de tráfico como

$$M_{\Delta V}(x,y) = \Delta V_{opt}(x,y) + \Delta V_{new}(x,y) \,. \tag{5.13}$$

5.3.2. Paso 2: Selección de emplazamientos

Una vez obtenido el mapa de ganancia de tráfico potencial, debe decidirse dónde ubicar la nueva SC, siguiendo el proceso mostrado en la Fig. 5.8. Primero se calcula la ganancia de volumen de datos agregada asociada a una ubicación de SC candidata (x,y), $G_{sc}(x,y)$, como

$$G_{sc}(x,y) = \sum_{(i,j)\in A_{sc}(x,y)} M_{\Delta V}(i,j) \,, \tag{5.14}$$

donde $A_{sc}(x,y)$ es el área de servicio de una hipotética SC situada en (x,y). En este capítulo, se asume que una SC es omnidireccional, de modo que $A_{sc}(x,y)$ es un círculo de radio 50 metros centrado en (x,y) [115, 116].

Entrada: Mapa de la ganancia de volumen de datos agregada asociada a una candidata a ubicación de SC ($G_{sc}(x, y)$), ganancia mínima aceptable (G_{min}) y mapa de la ganancia potencial de volumen de datos por cuadrícula ($M_{\Delta V}(x, y)$)
Salida: Lista de nuevas ubicaciones de SC (x_{sc}, y_{sc}) y ganancia conseguida ($G_{sc}(sc)$)

Mientras $\max(G_{sc}(x, y)) > G_{min}$ **se repite:**
\\Obtener la posición de ganancia máxima:

$$(x_{sc}, y_{sc}) = \arg\max_{(x,y)} (G_{sc}(x, y))$$

\\Obtener ganancia máxima:

$$G_{sc}(sc) = G_{sc}(x_{sc}, y_{sc})$$

\\Establecer en 0 todas las cuadrículas cubiertas por la nueva SC:

$$M_{\Delta V}(x, y) = 0 \quad \forall (x, y) \in A_{sc}(x_{sc}, y_{sc})$$

\\Recalcular el mapa de ganancia de volumen por SC:

$$G_{sc}(x, y) = \textstyle\sum_{(i,j)\in A_{sc}(x,y)} M_{\Delta V}(i, j) \quad \forall (x, y)$$

Termina

Figura 5.9: Pseudocódigo del algoritmo para la selección de emplazamientos.

A continuación, se identifican las mejores ubicaciones candidatas para una nueva SC detectando los máximos locales en $G_{sc}(x, y)$ de forma iterativa. La Fig. 5.9 presenta el pseudocódigo del algoritmo de selección de ubicaciones candidatas. Las entradas del algoritmo son (a) el mapa de agregación de la ganancia de tráfico asociada a una ubicación de SC candidata, $G_{sc}(x, y)$, (b) el valor de ganancia mínima aceptable, G_{min}, por debajo del cual no se considera la implementación de una nueva SC y (c) el mapa de ganancia potencial de tráfico por cuadrícula, $M_{\Delta V}(x, y)$. Las salidas resultantes son (a) una lista de coordenadas, correspondiente a las nuevas ubicaciones seleccionadas para una nueva SC y (b) la previsión de ganancia de tráfico conseguida por cada una de ellas. Se busca de forma iterativa el valor máximo y sus coordenadas en $G_{sc}(x, y)$, añadiendo dicha posición en la lista de ubicaciones seleccionadas. Para evitar la selección de sitios demasiado cercanos que puedan solaparse, el mapa de ganancia potencial de tráfico, $M_{\Delta V}(x, y)$, se actualiza cada vez que se selecciona una nueva SC, forzando a 0 el valor de los puntos bajo la nueva SC seleccionada, y se vuelve a generar el mapa de agregación de ganancia para seleccionar la siguiente ubicación. El resultado del proceso anterior es una lista de nuevos emplazamientos de SC, incluyendo la cuadrícula que indica el centro de sus áreas de servicio y la ganancia potencial esperada.

Hay que tener en cuenta que, en áreas heterogéneas que comprenden cuadrículas interiores y exteriores contiguas, el contexto de la cuadrícula central puede no ser el

mismo que el contexto (interior o exterior) del que procede la mayor parte de la ganancia de tráfico. Por ejemplo, la ubicación sugerida para una SC que cubra cuatro edificios cercanos (donde la mayor parte de la ganancia de tráfico proviene de los usuarios de interior) podría estar situada en la calle (exterior). Para evitar esta situación, se estima un ratio de mejora en interiores para cada SC, $R_{int}(x_{sc}, y_{sc})$, calculado a partir del mapa de ganancia de tráfico como

$$R_{int}(x_{sc}, y_{sc}) = \frac{\sum_{(x,y)\in A_{sc,int}(x_{sc},y_{sc})} M_{\Delta V}(x,y)}{\sum_{(x,y)\in A_{sc}(x_{sc},y_{sc})} M_{\Delta V}(x,y)} \ , \tag{5.15}$$

donde $A_{sc,int}(x_{sc}, y_{sc})$ es el área clasificada como interior dentro del área de servicio de la celda. En este indicador, valores bajos revelan que la mayor parte de la ganancia de tráfico proviene de zonas exteriores, mientras que valores cercanos a 1 indican que la ganancia de tráfico proviene de interior. En el primer caso, la nueva SC debería estar situada en el exterior, mientras que, en el segundo, la SC debería estar situada en zonas de interior. En este trabajo, se establece heurísticamente un umbral de $R_{int,th} = 0,5$ para decidir si una SC debe ubicarse en interiores (es decir, $R_{int}(x_{sc}, y_{sc}) \geq 0,5$) o en exteriores (en caso contrario). En la práctica, el valor del umbral se establece por parte de los técnicos de la red. Si la cuadrícula de la ubicación sugerida para la SC no coincide con ese contexto, es necesario un análisis más detallado para determinar la mejor ubicación según el contexto de ganancia predominante. Este análisis puede acabar con el despliegue de varias SC para cubrir por separado las ubicaciones interiores y exteriores si el área de servicio incluye diferentes contextos. Este proceso está pensado para tener en cuenta todos los máximos locales presentes en el mapa de ganancia potencial de volumen por SC. Se busca evitar que, debido a la geografía del escenario, la aplicación de una máscara que limite la búsqueda en la detección de ubicaciones pase por alto zonas críticas de la red, no llegando a detectar picos altos de ganancia máxima.

5.4. Evaluación del modelo

Para mostrar la validez del modelo descrito en este capítulo, en primer lugar se describe el escenario empleado en su desarrollo. A continuación, se muestran los resultados del modelo en el escenario de desarrollo. Tras esto, se muestra una comparación con otros enfoques para determinar las mejoras que presenta el modelo. Por último, se discuten los aspectos computacionales.

Tabla 5.1: Tipos de uso de suelo en el escenario.

Uso de suelo	Descripción	Interior/Exterior	Ocupación
Servicios	Áreas de uso institucional, cultural, médico o académico.	Interior	10 %
Oficinas	Oficinas de gestión, información o trabajo.	Interior	3 %
Uso mixto	Zonas sin uso de suelo principal definido.	Interior	11 %
Residencial	Casas, hoteles y alojamientos.	Interior	15 %
Entretenimiento	Comercios y entretenimiento.	Interior	4 %
Industrial	Zonas para trabajo industrial o de manufacturación.	Interior	3 %
Espacio abierto	Campos, parques y zonas verdes.	Exterior	22 %
Carreteras	Caminos, carreteras y senderos.	Exterior	30 %
Ríos/Lagos	Zonas de agua en el terreno.	Exterior	2 %

5.4.1. Metodología experimental

El modelo propuesto es también una aplicación práctica de los modelos de los Capítulos 3 y 4, por lo que se ha vuelto a emplear el mismo escenario para su desarrollo. En concreto, este escenario corresponde a un área geográfica de 125 km^2, donde se sitúa una ciudad y su extrarradio con 800.000 habitantes. En este área se han desplegado 400 celdas LTE, agrupadas en 175 emplazamientos trisectoriales, con una frecuencia de portadora de 2.325 MHz y un ancho de banda del sistema de 15 MHz. En estas celdas se registraron 166.561 conexiones en un periodo de recolección de trazas de 2 horas.

Para tener una resolución espacial suficiente, que distinga al menos nivel de edificio, todos los mapas empleados se definen como una malla de cuadrículas de 10x10m^2. En la Tabla 5.1 se muestra un resumen de los tipos de uso de suelo empleados, así como su descripción, la clasificación de interior o exterior y el porcentaje de ocupación en el escenario.

Por último, se ha generado una base de datos con las publicaciones geoetiquetadas de las redes sociales en la zona, recogidas en tiempo real durante 16 meses para Twitter y 12 meses para Flickr, acumulando un total de 785.515 y 33.519 publicaciones, respectivamente.

Cabe destacar que el proceso de selección de nuevos emplazamientos propuesto en este capítulo no tiene restricciones espaciales y, por tanto, todas las cuadrículas del mapa son ubicaciones candidatas. En la práctica, esto no es posible, debido a posibles restricciones físicas (p.ej., ubicación en altura imposible), de acceso a la red de

Tabla 5.2: Parámetros de regresión por contexto.

Contexto, ct	interior			exterior		
Percentil, j	90 %	50 %	10 %	90 %	50 %	10 %
$\beta^{(j)}_{0,V,ct}$	$9,41\times10^{9}$	$3,46\times10^{6}$	$5,32\times10^{4}$	$7,35\times10^{6}$	$2,97\times10^{6}$	$7,59\times10^{4}$
$\beta^{(j)}_{1,V,ct}$	284,21	105,81	21,98	365,96	127,89	12,76
$\beta^{(j)}_{0,R,ct}$	0,241	0,122	0	0,097	0,038	0
$\beta^{(j)}_{1,R,ct}$	$1,86\times10^{-5}$	$3,47\times10^{-6}$	0	$1,48\times10^{-5}$	$3,89\times10^{-6}$	0

transmisión (p.ej., puntos de fibra disponibles), legales (p.ej., colegio cercano o domicilio privado en la ubicación seleccionada), etc.

Para comprobar el comportamiento del modelo, se prueban tres modelos: a) una versión simplificada del modelo descrito en este capítulo, basada únicamente en trazas, que segrega las conexiones en exteriores/interiores para afinar la distribución espacial de usuarios a partir del avance temporal con el mapa de uso del suelo (denominado *modelo de trazas*, TM), b) una variante del modelo, inspirada en el modelo para detectar puntos conflictivos de tráfico propuesto en [17], que deriva la distribución espacial de los usuarios sólo a partir de los datos de las redes sociales (denominado *modelo social*, SM), y c) el modelo propuesto para seleccionar las mejores ubicaciones candidatas a alojar una nueva SC, que combina trazas y datos de redes sociales (denominado *modelo de trazas sociales*, TSM). Los dos primeros son enfoques que se utilizan como referencia para comprobar las ventajas frente a utilizar únicamente publicaciones de redes sociales o trazas, respectivamente.

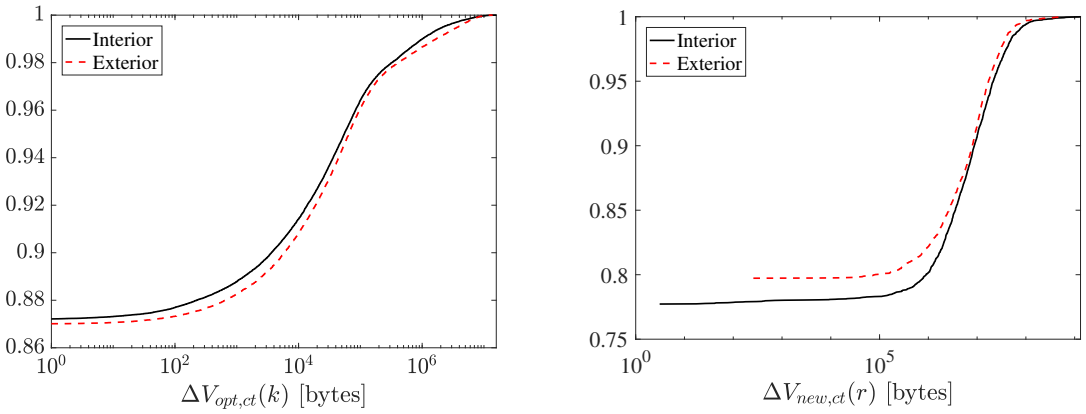
Idealmente, se evaluaría el rendimiento del modelo desplegando los emplazamientos sugeridos por los modelos en la red real. Como esto no ha sido posible, la comparación se basa en el mapa de ganancia de tráfico potencial por cuadrícula derivado por el modelo TSM. Así, el análisis sólo muestra dónde (y en qué medida) los tres modelos se comportan de forma diferente.

5.4.2. Resultados

Una vez descrito el escenario de validación, se presentan los resultados obtenidos con el modelo. Las Fig. 5.3 y 5.4 muestran el diagrama de dispersión del volumen de datos y el ratio de conexiones de búfer completo frente al caudal, construido a partir del conjunto de datos de trazas. Estas figuras se utilizan para derivar el conjunto de curvas de regresión por contexto para cada indicador. Por otro lado, la Tabla 5.2 presenta los coeficientes de regresión resultantes del ajuste de estas curvas de regresión por contexto,

Tabla 5.3: Estadísticas de ganancia de volumen de datos por conexión y anillo existentes (en bytes).

Ganancia	Percentil 10 %	Media	Percentil 90 %
$\Delta V_{opt,int}(k)$	0	$6,077 \times 10^4$	$5,450 \times 10^7$
$\Delta V_{opt,out}(k)$	0	$1,279 \times 10^5$	$1,304 \times 10^8$
$\Delta V_{new,int}(r)$	0	$7,104 \times 10^6$	$1.,980 \times 10^9$
$\Delta V_{new,out}(r)$	0	$4,860 \times 10^6$	$6,430 \times 10^8$



(a) Conexiones existentes.

(b) Nuevas conexiones

Figura 5.10: Distribución acumulada de probabilidad de la ganancia de volumen estimada por conexión y por anillo.

descritas con las ecuaciones 5.1 y 5.2.

Una vez obtenidos los modelos de regresión, se estima la ganancia potencial de volumen de datos por conexión y anillo. La Tabla 5.3 resume los resultados presentando los valores del percentil 10, la media y el percentil 90 de las ganancias de volumen de datos de las conexiones, $\Delta V_{opt,ct}(k)$, y los anillos, $\Delta V_{new,ct}(r)$, en el escenario. La Fig. 5.10 muestra sus CDF, desglosadas por contexto. Como era de esperar, en la Tabla 5.3, se observa que los valores de ganancia son menores para las conexiones individuales que para los anillos, que agregan varias ubicaciones. Más interesante aún, de la Fig. 5.10(a), se deduce que, para las conexiones existentes, se consiguen mayores ganancias de volumen de datos por conexión en el exterior. Por el contrario, de la Fig. 5.10(b), se deduce que los anillos de interior muestran un mayor aumento de tráfico debido al incremento de las nuevas conexiones de datos.

Las ganancias potenciales estimadas por conexión y anillo se proyectan en un mapa mediante la geolocalización de los datos de la red. La Fig. 5.11 muestra la distribución estadística de varios indicadores (número de conexiones, caudal medio, volumen total de datos y volumen medio de datos por conexión) calculados por cuadrícula, mediante

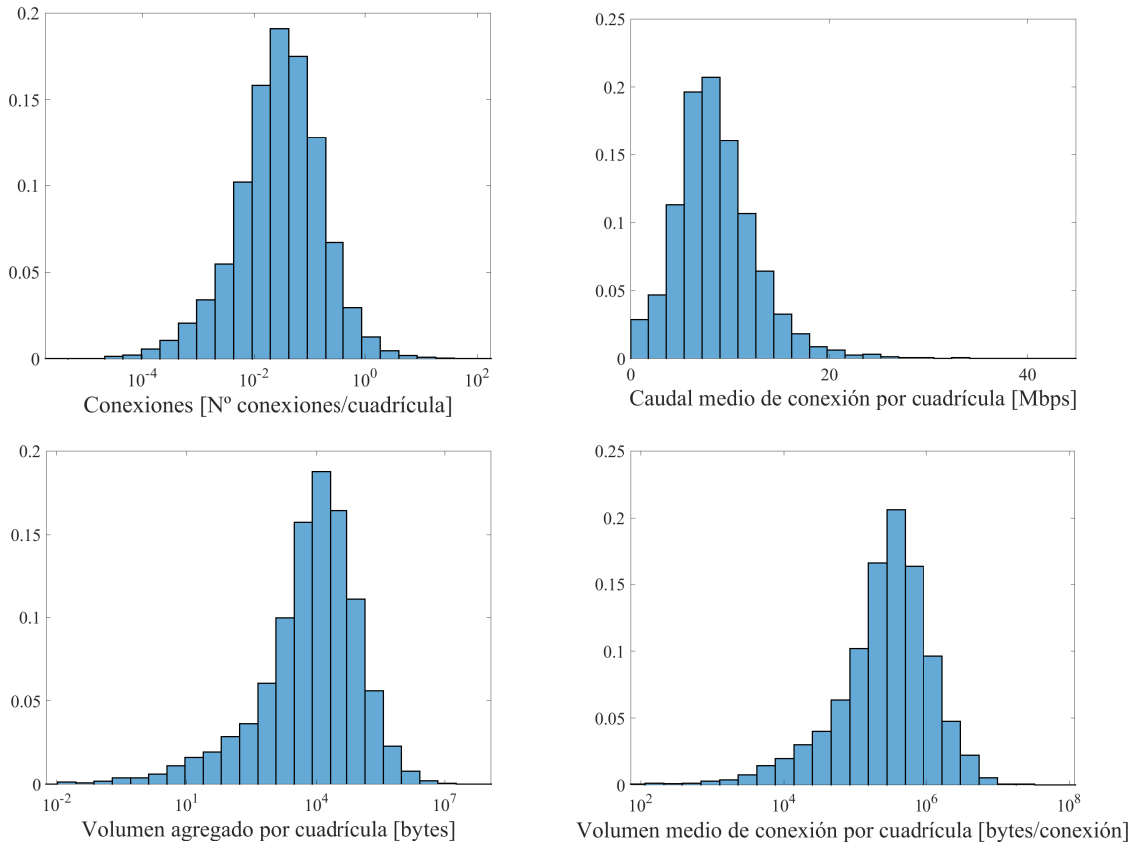
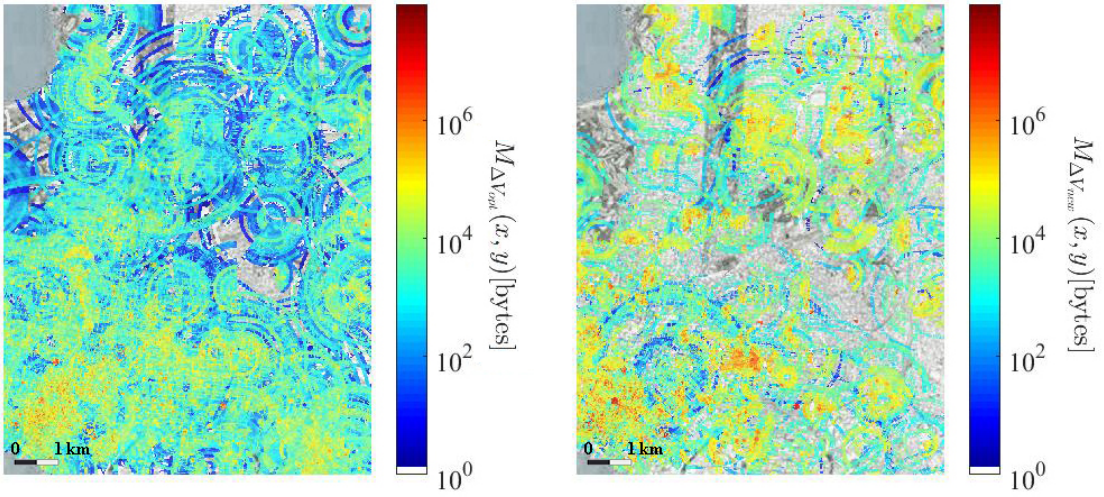


Figura 5.11: Distribución estadística de los indicadores por cuadrícula del mapa.

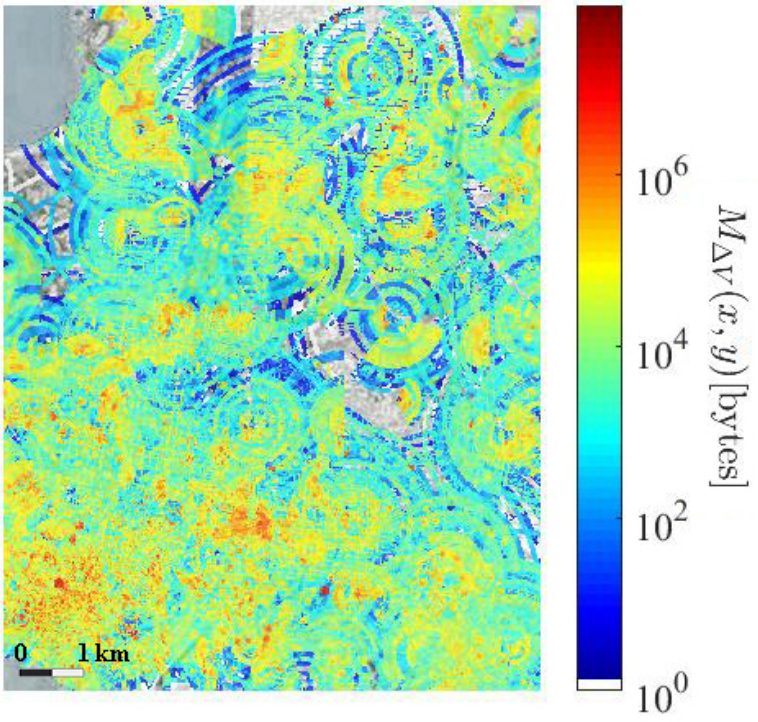
la distribución de conexiones en el mapa descrita en el Capítulo 4. La gran variabilidad observada en todos los indicadores no es más que la consecuencia de la heterogeneidad del escenario, que comprende zonas de densidad de población y condiciones de propagación muy diferentes (p.ej., el número de conexiones por cuadrícula de 100 m^2 oscila entre 0 y 175 conexiones). Esto justifica la necesidad de un modelo preciso que considere todos los factores mencionados.

Por otro lado, las Fig. 5.12 (a)-(c) representan el mapa de ganancia de tráfico potencial, desglosado por sus dos componentes, sobre la ortofoto de la zona. De izquierda a derecha, las Fig. 5.12 (a), (b) y (c) muestran la ganancia de volumen potencial por cuadrícula de las conexiones existentes, los nuevos servicios y la ganancia potencial completa, respectivamente. Se observa que el volumen potencial ganado por la mejora de las conexiones ya establecidas, $M_{\Delta V_{opt}}(x, y)$, está más distribuido por el mapa. Por el contrario, la ganancia por nuevos servicios, $M_{\Delta V_{new}}(x, y)$, está más concentrada en zonas concretas. Esto se debe al hecho de que muchos de los anillos que no tienen conexiones de alto volumen de datos (conexiones no de último TTI) se encuentran en zonas con altas concentraciones de usuarios y, por tanto, la congestión resultante en



(a) Conexiones existentes

(b) Nuevas conexiones



(c) Completa

Figura 5.12: Distribución espacial de la mejora potencial de volumen de tráfico, $M_{\Delta V}(x, y)$.

la red disuade a los usuarios de usar servicios demandantes de datos. En ambos mapas, las cuadrículas transparentes muestran las zonas con ganancia cero, en las que no se han establecido conexiones. La mayoría de estos mosaicos se encuentran en zonas despobladas, fuera de la región de cobertura prevista. La Tabla 5.4 confirma estos resultados al presentar algunas estadísticas de las tres distribuciones espaciales. Co-

Tabla 5.4: Estimación del incremento potencial de volumen de datos (en bytes).

KPI	Percentil 10	Mediana	Percentil 90
$M_{\Delta V}(x,y)$	0	$1,030 \times 10^4$	$1,903 \times 10^8$
$M_{\Delta V_{opt}}(x,y)$	0	$2,309 \times 10^3$	$6,839 \times 10^7$
$M_{\Delta V_{new}}(x,y)$	0	$3,187 \times 10^3$	$1,818 \times 10^8$

mo muestran las figuras y la tabla, las mayores ganancias de volumen por cuadrícula proceden de las nuevas conexiones (nótese los colores más cálidos en la Fig. 5.12(b)). Asimismo, el mapa de ganancia global puede utilizarse para detectar zonas que ya funcionan en condiciones óptimas (es decir, con un valor bajo de $M_{\Delta V}(x,y)$) y otras con malas condiciones de enlace radio y/o muchos usuarios (es decir, con un valor alto de $M_{\Delta V}(x,y)$). Nótese que, al tener distribuciones espaciales muy diferentes, la suma de los estadísticos de ambas componentes de ganancia no tiene por qué coincidir con los estadísticos de la ganancia completa.

Una vez se conoce dónde puede aumentarse el tráfico de la red, se decide dónde ubicar las nuevas SC. Para ello, se agregan las ganancias potenciales de volumen de datos por cuadrícula, calculando la ganancia total de tráfico en las diferentes ubicaciones candidatas de SC, $G_{sc}(x,y)$. En la Fig. 5.13 se presentan los 100 mayores valores de ganancia total obtenidos para una ubicación, ordenados de mayor a menor. Para un análisis más detallado, las ganancias se desglosan para las conexiones existentes y las nuevas. Se observa que, en la mayoría de sitios, las ganancias de tráfico provienen de las nuevas conexiones de búfer completo.

Para mostrar la capacidad del modelo para detectar problemas de cobertura, en la Fig. 5.14 se representan dos ejemplos de ubicaciones de SC seleccionadas (marcadas con un punto en la figura) y sus áreas de cobertura ideales (marcado con una circunferencia de 50 metros de radio) sobre un mapa de cobertura extraído de la plataforma OpenSignal [77]. Esta plataforma recoge mediciones geolocalizadas del nivel de señal de forma anónima de los usuarios de móviles suscritos a esta iniciativa. En la figura, se observa que los emplazamientos SC propuestos cubren zonas donde se reporta un nivel de señal débil en OpenSignal (zonas rojas en la figura).

Por último, el análisis se centra en el indicador del ratio de ganancia en interiores, $R_{int}(x_{sc}, y_{sc})$, que refleja qué parte de la ganancia de volumen de datos de una SC procede de las cuadrículas de interior. La Fig. 5.15 muestra el histograma de $R_{int}(x_{sc}, y_{sc})$ para los 100 mejores SC candidatos. Se observa que el ratio de ganancia en interiores de los mejores emplazamientos tiende a ser superior a 0,5, lo que demuestra que la

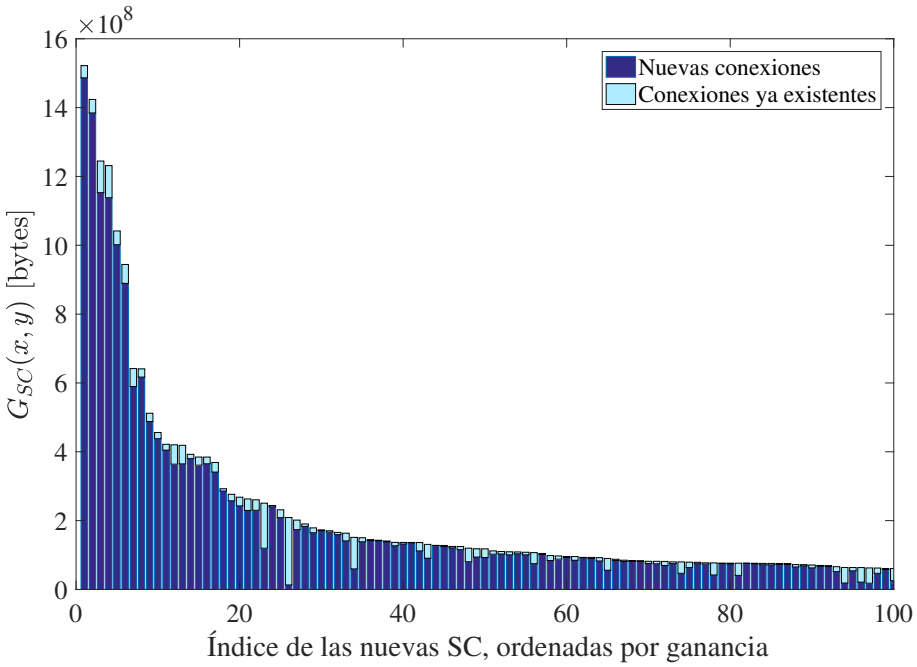
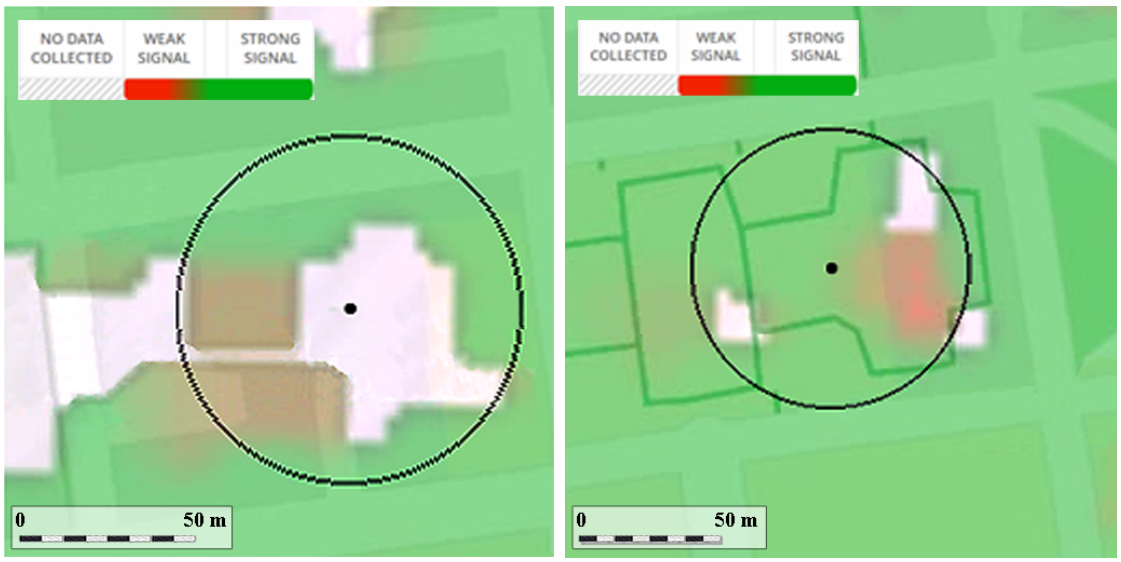


Figura 5.13: Incremento de volumen para las mejores 100 nuevas ubicaciones.



(a) Emplazamiento 1 (b) Emplazamiento 2

Figura 5.14: Nivel de señal recibida (OpenSignal) en ejemplo de selección de emplazamiento.

ganancia de tráfico se origina principalmente en los usuarios de interiores. En concreto, 73 de los 100 mejores emplazamientos tienen un ratio de ganancia en interiores superior a 0,5, por lo que deberían ubicarse en interiores.

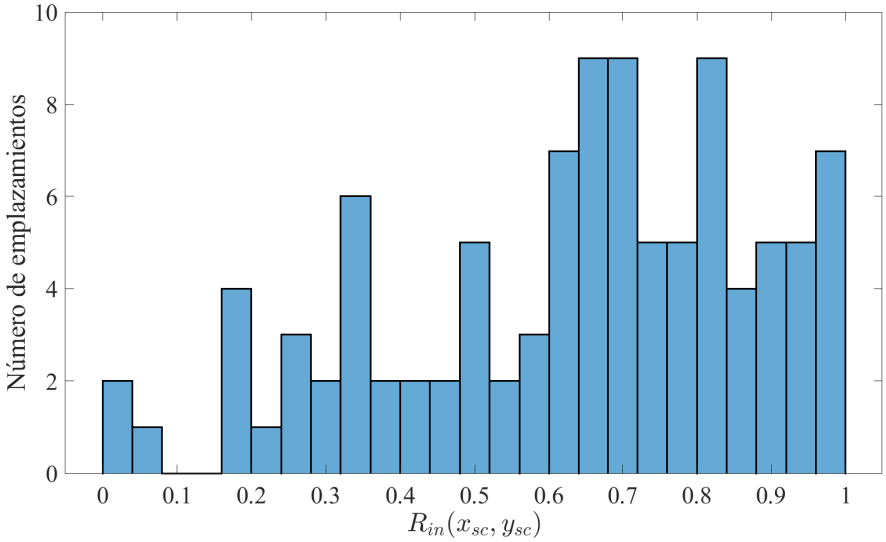


Figura 5.15: Histograma del ratio de ganancia en interior para los mejores 100 emplazamientos candidatos.

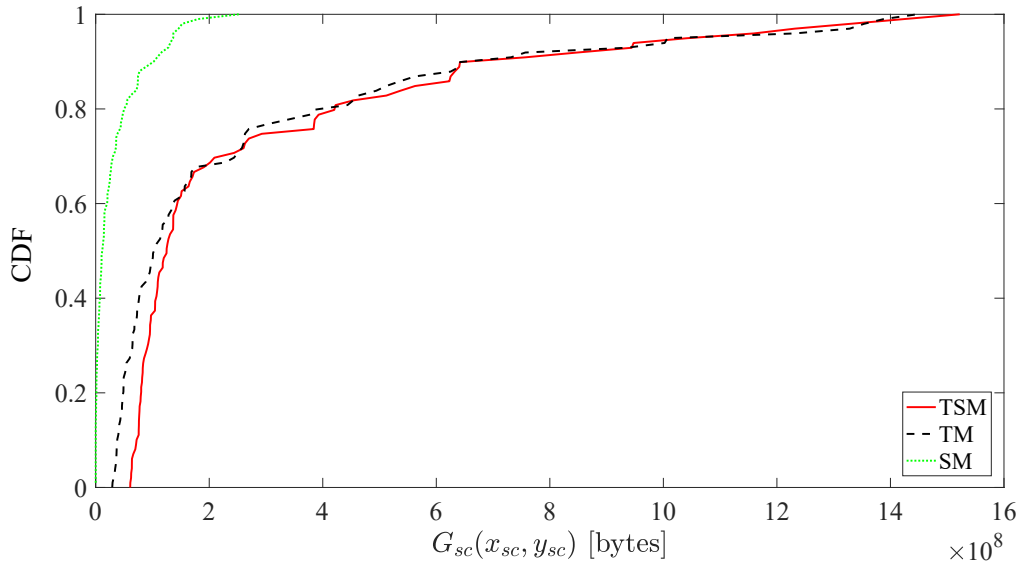


Figura 5.16: Distribución de la ganancia de volumen de tráfico de los 100 mejores nuevos emplazamientos candidatos para los métodos basados en red social (SM), trazas (TM) y trazas+red social (TSM).

Comparación del modelo con otras propuestas

La Fig. 5.16 muestra la CDF de las ganancias de volumen de datos obtenidas de los mejores 100 nuevos emplazamientos sugeridos por los tres modelos, evaluados con el mapa de ganancia de volumen de datos potencial de TSM. Como se esperaba, TSM obtiene mayores ganancias de volumen de datos por emplazamiento. Agregando

Tabla 5.5: Ejemplos de nuevos emplazamientos en los que los modelos se comportan de forma diferente.

Caso	Descripción	Ejemplo
Caso 1	Emplazamiento con problemas de cobertura detectados tanto por TSM como por TM/SM, pero con variaciones de posición.	Fig. 5.17(a)
Caso 2	Emplazamiento con problemas de cobertura detectados solo por TSM.	Fig. 5.17(b)
Caso 3	Emplazamiento con problemas de capacidad detectados tanto por TSM como por TM/SM, pero con variaciones de posición.	Fig. 5.19(a)
Caso 4	Emplazamiento con problemas de capacidad detectados solo por TSM.	Fig. 5.19(b)
Caso 5	Emplazamiento detectado solo por TM/SM (falso emplazamiento).	Fig. 5.19(c)

la ganancia de los 100 emplazamientos candidatos de la Fig. 5.16, TSM consigue una ganancia total de volumen de datos en la red de 27,58 GB, TM de 25,40 GB (8 % menos) y SM sólo de 3,22 GB (88 % menos). Este resultado pone de manifiesto que la solución obtenida por SM difiere mucho de la de TSM y TM. Un análisis detallado muestra que la diferencia entre TM y TSM no es demasiado grande, ya que muchos anillos de TA tienen pocas publicaciones debido al tamaño limitado del conjunto de datos de redes sociales, lo que hace que el tráfico en estos anillos se distribuya uniformemente (por contexto) en TSM, como en TM. Se espera que se observen mayores diferencias con un conjunto de datos de publicaciones en redes sociales más amplio.

Un análisis más detallado de las localizaciones de nuevas SC en un mapa muestra importantes diferencias entre los modelos. A continuación, se muestran casos en los que el modelo TSM propuesto se comporta de forma diferente a los modelos de comparación, TM y SM. Por claridad, las nuevas SC se dividen en dos grupos, dependiendo de si su objetivo es cubrir zonas con malas condiciones de propagación (es decir, huecos de cobertura) o con una gran demanda de tráfico, derivando en problemas de capacidad (es decir, puntos de alta demanda de tráfico), detectados de forma exploratoria a partir de las condiciones de su entorno, de la cantidad de usuarios en la zona y del nivel de señal recibida. En la tabla 5.5 se resumen los cinco casos identificados. En ambos grupos, se analizan varios casos en los que: a) el punto problemático es detectado por varios modelos, pero en ubicaciones ligeramente diferentes, o b) el punto problemático sólo se detecta por uno de los modelos.

En la Fig. 5.17(a) se muestra el primer caso de un hueco de cobertura en exterior detectado por TSM y TM. Para obtener una imagen completa, la figura muestra las ubicaciones de las SC sugeridas por TSM y TM, las estaciones base cercanas y las

(a) Caso 1. (b) Caso 2.

Figura 5.17: Propuesta de emplazamiento para solucionar problemas de cobertura.

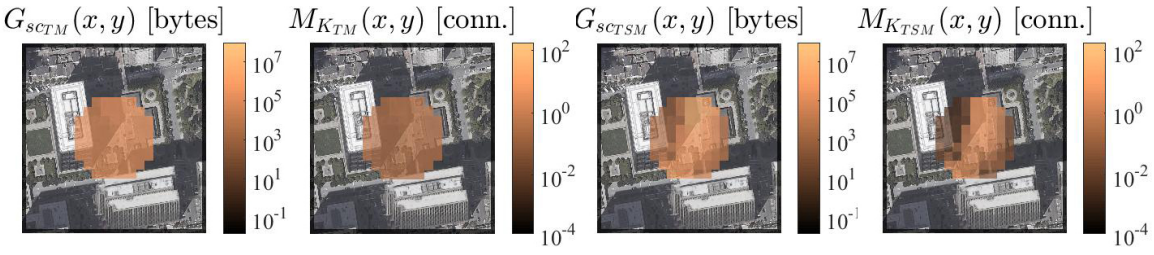
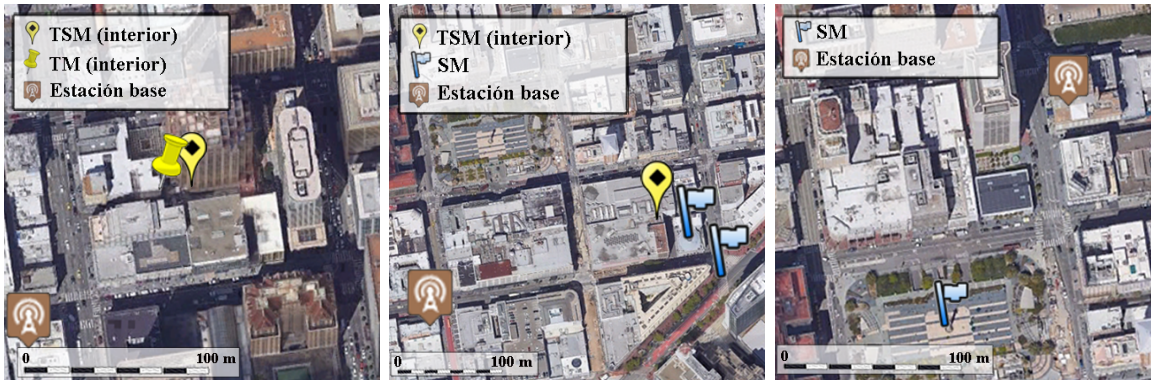


Figura 5.18: Análisis detallado de TM y TSM (caso 2).

publicaciones geoposicionadas. Ambas ubicaciones de SC cubren un área que no está en línea de visión con la estación base macrocelular debido a un edificio alto (señalado con una circunferencia semitransparente), que genera un hueco de cobertura. Se observa que la ubicación SC de TSM se desplaza hacia la derecha, siguiendo la distribución de las publicaciones, cubriendo un edificio de mayor tamaño que, previsiblemente, contendrá un mayor número de personas. A pesar de este desplazamiento, la ganancia total de volumen de datos estimada por TSM y TM es casi la misma (1437,4 MB y 1423,7 MB, respectivamente). En ambos casos, la mayor parte de la ganancia procede de las cuadrículas de exterior, coincidiendo con el contexto de gran parte de las publicaciones sociales en el área.

La Fig. 5.17(b) muestra el segundo caso de un hueco de cobertura en interior detectado únicamente por TSM. Un análisis preliminar muestra que el edificio donde se encuentra la SC es un aparcamiento subterráneo, que no tiene línea de visión directa con la estación base macrocelular cercana debido a un rascacielos. Para entender por



(a) Caso 3. (b) Caso 4. (c) Caso 5.

Figura 5.19: Propuesta de emplazamiento para solucionar problemas de capacidad.

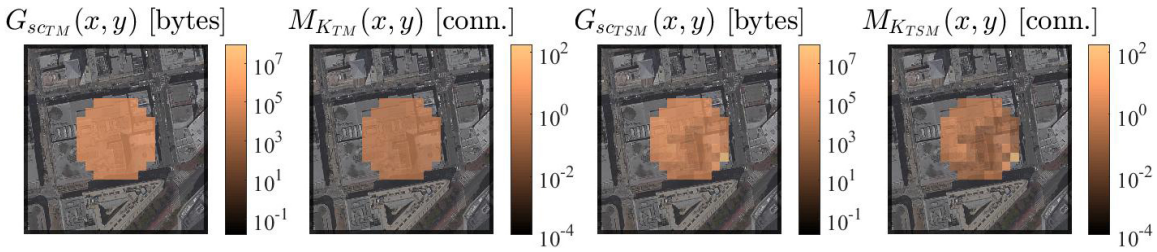


Figura 5.20: Análisis detallado de TM y TSM (caso 4).

qué TM no sugiere un emplazamiento para esta zona, la Fig. 5.18 muestra el número de conexiones por cuadrícula, $M_k(x, y)$, y la ganancia de tráfico asociada, $G_{sc}(x, y)$, utilizados por TSM y TM en la zona de la SC sugerida por TSM. A la izquierda, se observa que la densidad de conexiones en TM es baja y regular, lo cual es el resultado de distribuir las conexiones en los anillos de TA de la macrocelda de manera uniforme en el espacio (manteniendo la segregación por contexto). En cambio, la densidad de conexiones en TSM es grande e irregular, debido a la concentración de publicaciones en el interior del aparcamiento. Esta diferencia justifica la razón de las grandes desviaciones en las ganancias de tráfico esperadas. Además, nótese que, a pesar del gran número de publicaciones en la zona, SM no sugiere una nueva SC debido a su incapacidad para detectar problemas de cobertura al no manejar medidas de nivel de señal (que sí aportan las trazas en TM y TSM).

En la Fig. 5.19(a) se muestra el tercer caso de una zona con problemas de capacidad en interior detectado por TSM y TM. En este caso, TSM localiza la nueva SC en la esquina de un rascacielos que alberga una importante empresa gracias a las publicaciones geolocalizadas, mientras que TM localiza la nueva SC en un edificio cercano más pequeño (y posiblemente menos poblado). Como resultado, la ganancia total conseguí-

da por la nueva SC (estimada con el mapa de tráfico de TSM) es de 125,81 MB para la posición sugerida por TSM y de sólo 67,59 MB para la de TM. En este caso, SM no consigue incluir el sitio en la lista de las 100 mejores ubicaciones de SC porque el número de publicaciones sociales en la zona no es lo suficientemente alto. Este hecho causaría que el emplazamiento no se priorizase.

La Fig. 5.19(b) ilustra el caso de una zona con problemas de capacidad en interior detectado sólo por TSM y SM. Aunque no se aprecia en la figura, la SC sugerida por TSM está dentro de un centro comercial, lo que enfatiza la importancia de considerar las publicaciones geolocalizadas. En la Fig. 5.20 se presenta el mismo análisis detallado, mostrando que TM no detecta la zona altamente cargada debido a su dificultad para geolocalizar las conexiones con precisión dentro de los anillos de TA. En este caso, SM detecta dos puntos de concentración de usuarios buscando picos en la distribución espacial de las publicaciones de redes sociales.

Por último, la Fig. 5.19(c) muestra un caso en el que SM detecta una zona altamente poblada que no es detectada por TSM ni TM. Este caso puede explicarse por un pico de publicaciones en una región donde los usuarios móviles tienen una cobertura adecuada y suficientes recursos de radio disponibles.

5.4.3. Complejidad computacional

El modelo propuesto necesita un trabajo previo de recopilación/preprocesamiento de trazas y construcción del mapa de usos del suelo/publicaciones, descrito en las secciones 3.4.3 4.4.3, antes de poner en marcha el procedimiento de la Fig. 5.5. Al igual que el modelo del Capítulo 4, la complejidad computacional del modelo viene dada por el algoritmo de construcción de la distribución espacial de conexiones. En este caso, el algoritmo distribuye N_i indicadores a partir de $N_{con}(r)$ conexiones originadas en el área cubierta por un anillo r de los N_r anillos del escenario. Por lo tanto, la complejidad temporal en el peor de los casos es $\mathcal{O}(N_r\, N_{con}\, N_i)$.

El modelo propuesto se implementa con la *toolbox* de estadística, aprendizaje automático y procesado de imagen de Matlab. Todos los procesos se probaron en un servidor con un procesador de 8 núcleos de 2,4 GHz y 64 GB de RAM. El tiempo necesario para decodificar las trazas de conexión y construir los mapas de uso del suelo y de publicaciones se comenta en la sección 4.4.3. El tiempo para detectar los 100 mejores sitios candidatos con el conjunto de datos anterior (166.561 conexiones servidas por 400 celdas, repartidas en 125 km^2 divididos en 1.222.787 cuadrículas) es de 238

segundos, de los que el 67 % se emplea en la construcción de la distribución espacial de conexiones.

5.5. Conclusiones

En la actualidad, la selección de nuevos emplazamientos de picoceldas es un proceso muy laborioso que requiere muchas tareas que han de realizarse de forma manual. En este capítulo se propone un modelo automático basado en datos contextuales para detectar con precisión zonas con problemas de cobertura o capacidad en una red móvil. El indicador principal para detectar dichos problemas es el caudal de las conexiones registrado en las trazas de la interfaz radio. El núcleo del modelo es el posicionamiento de las conexiones basado en la probabilidad de interior de cada conexión y en la distribución de las publicaciones geolocalizadas de las redes sociales, descrito en el capítulo anterior. El modelo se ha probado con un gran conjunto de datos de trazas de una red LTE comercial y una base de datos de publicaciones geolocalizadas de Twitter y Flickr.

Los resultados obtenidos con el modelo propuesto han demostrado que la inclusión de la distribución espacial de las publicaciones de redes sociales, así como el contexto de los usuarios, aporta una información muy relevante a la hora de localizar puntos problemáticos en la red, llegando a detectar ganancias potenciales de tráfico por encima de 27 GB por nueva SC en el periodo de recolección de trazas. Además, los problemas detectados en la red combinando datos de conexión y los posts geoetiquetados son coherentes con su contexto, ya que los sitios detectados debido a una mala cobertura presentan malas condiciones de propagación desde la macrocelda servidora, mientras que los puntos con problemas de capacidad se encuentran en lugares muy poblados (p.ej., museos, escuelas, centros comerciales, etc.). Asimismo, la distinción interior/exterior para la nueva SC es coherente con el escenario, es decir, los emplazamientos etiquetados como interiores se encuentran en ubicaciones interiores, mientras que los puntos que cubren zonas abiertas de la ciudad se clasifican como exteriores. Además, la comparación con otros enfoques similares ha mostrado importantes diferencias en los emplazamientos seleccionados.

Un componente clave del modelo propuesto son las curvas de regresión que modelan el impacto del caudal de los usuarios en el volumen de tráfico y la combinación de servicios. Las cifras han demostrado que la precisión de la regresión aún puede mejorarse. Para ello, se pueden derivar modelos de regresión más sofisticados que consideren más predictores con técnicas de aprendizaje automático, siempre que se disponga de un

conjunto de datos de medición amplio y diverso.

La baja complejidad computacional del modelo permite una fácil integración en las herramientas de planificación radio. De la misma forma que el modelo del capítulo anterior, al combinar diferentes fuentes de datos, el modelo puede aprovechar al máximo el potencial de los últimos sistemas de gestión de red basados en datos masivos.

Capítulo 6

Predicción de la demanda de tráfico en eventos sociales

En capítulos anteriores, solo se ha estudiado la influencia del contexto espacial en las redes celulares. Pero el dónde no es el único factor relevante en el impacto de los usuarios en el rendimiento de la red. Conocer cuándo se va a producir una avalancha de peticiones de recursos es vital para mejorar la experiencia del usuario. Un ejemplo claro es el caso de los eventos de masas, donde se concentra una gran cantidad de usuarios debido a un acontecimiento de carácter social (p.ej., conciertos, encuentros deportivos, ferias y congresos...). En dichos eventos, los asistentes quieren compartir su experiencia con sus contactos, generando un ingente volumen de tráfico concentrado en un corto periodo de tiempo en una zona específica. Observando solo la demanda de tráfico en la red, estos eventos parecen esporádicos y, por lo general, aleatorios. Sin embargo, la información espaciotemporal de los próximos eventos es de acceso público, por lo que puede usarse para estimar el impacto de los mismos en la red y solucionar de forma proactiva problemas de capacidad por falta de recursos.

Este capítulo sigue el mismo esquema que los anteriores. La sección 6.1 detalla diferentes aproximaciones al problema de la predicción de tráfico en redes celulares. Seguidamente, en la sección 6.2 se realiza un análisis preliminar del impacto de los eventos sociales en la red. A continuación, la sección 6.3 presenta la cuarta y última contribución de esta tesis, un modelo de aprendizaje profundo que predice el tráfico en una celda considerando el impacto de futuros eventos sociales. La sección 6.4 presenta los resultados obtenidos por el modelo de predicción propuesto sobre un juego de datos reales. Finalmente, en la sección 6.5 se comentan las conclusiones extraídas del

comportamiento del modelo.

6.1. Estado de la técnica

La predicción de la demanda de tráfico forma parte de las redes proactivas, cuyo objetivo es caracterizar el comportamiento humano y la dinámica de la red para optimizar la asignación de recursos. Como estos recursos deben gestionarse a diferentes niveles, se han realizado numerosas investigaciones para prever el tráfico celular con diferentes resoluciones espaciales (p.ej., área de celda [117, 118], ciudad [119], provincia [120] o red [121]) y horizontes temporales (p.ej., segundos [117], minutos [118, 122], horas [123, 119], días [121] o meses [120]). La predicción a largo plazo tiene como objetivo encontrar la tendencia general de la serie temporal de tráfico para la replanificación de la red, mientras que la predicción a corto plazo rastrea las fluctuaciones rápidas de la demanda de tráfico para la gestión dinámica de los recursos radio.

Tradicionalmente, la previsión de la demanda de tráfico se ha tratado como un problema clásico de análisis de series temporales. En esta línea, se proponen modelos lineales, como el suavizado exponencial de Holt-Winters o la media móvil integrada autorregresiva (ARIMA) estacional (*Seasonal Auto-Regressive Integrated Moving Average*, SARIMA), para predecir el tráfico celular. Estos modelos definen las variables dependientes del tiempo como una combinación de componentes de tendencia, estacionalidad y ruido. En [121], se utiliza un modelo SARIMA para prever el tráfico diario en una red GSM para los próximos 28 días. En [120], este enfoque se amplía para predecir el tráfico de las horas de mayor afluencia mensualmente en una provincia. En [123], se presenta un modelo de Holt-Winters para predecir el tráfico cursado por celda durante 7 días sobre una base horaria. En [118], las series temporales se descomponen primero en componentes de regularidad y aleatoriedad mediante el análisis de componentes principales y luego se aplica ARIMA para predecir el tráfico de 9.000 estaciones de telefonía móvil. En [117], se propone un modelo ARIMA/GARCH para predecir el tráfico del Protocolo de Control de Transmisión (TCP) entre el Laboratorio Lawrence Berkeley y el resto del mundo con diferentes resoluciones temporales (1, 10 y 100 segundos). En [124], se predice con ARIMA la evolución de la tasa de datos alcanzable por segundo para usuarios móviles específicos. Estos enfoques lineales pueden ampliarse con modelos no lineales de heterocedasticidad condicional autorregresiva generalizada (*Generalized AutoRegressive Conditional Heteroskedasticity*, GARCH) para series no estacionarias cuya varianza de error cambia con el tiempo [123]. Como alternativa, los

modelos basados en actividad y construidos a partir de los registros de llamadas capturan los patrones de desplazamiento en las áreas metropolitanas, que pueden utilizarse para predecir las fluctuaciones diarias del tráfico sobre una base horaria [125][126]. Otra opción es basar la predicción del tráfico en el procesamiento de señales para captar las variaciones a corto plazo (p.ej., el filtrado de Kalman [127] o la detección compresiva temporal [128]).

Con los recientes avances en inteligencia artificial y el interés por explotar los datos masivos recogidos por las redes móviles, la previsión moderna del tráfico se basa en algoritmos de aprendizaje automático como, por ejemplo, la regresión de vectores de soporte (*Support Vector Regressors*, SVR) [129] o las redes neuronales artificiales (ANN) [130, 131]). Entre las ANN, las Redes Neuronales Recurrentes (RNN) son especialmente adecuadas para el procesamiento de secuencias de datos, ya que alimentan la salida de una capa a las neuronas de una capa anterior para formar ciclos dirigidos que proporcionan memoria, de forma análoga a los modelos autorregresivos, como (S)ARIMA. En [132], el tráfico celular se predice mediante la agrupación de grupos de celdas similares, la eliminación de la información redundante mediante la descomposición *wavelet* y la captura de las dependencias temporales con redes neuronales de Elman. Estas últimas dependencias pueden derivarse con mayor precisión con redes de memoria a largo y corto plazo (*Long Short-Term Memory*, LSTM), que retienen la información durante períodos de tiempo más largos [38]. Con LSTM, se puede prever el tráfico celular por celdas a diferentes escalas de tiempo [133, 134, 135]. Alternativamente, se pueden utilizar redes de creencia profundas (*Deep Belief Networks*, DBN) no supervisadas para capturar la dependencia de largo alcance en el tráfico de red cuando se predice con resoluciones temporales más altas [136]. Todos los esquemas anteriores centrados en las variaciones temporales pueden ampliarse añadiendo dependencias espaciales entre regiones adyacentes, derivadas con autocodificadores (*autoencoders*, AE) [129] o Redes Neuronales Convolucionales (CNN) [137, 138]. Para ello, se suele dividir el escenario en una cuadrícula regular de regiones agregando la demanda de tráfico por región. Las dependencias más complejas a nivel de celda pueden modelarse de forma más eficaz mediante la extracción de características basadas en la correlación de tráfico [139] o en las redes neuronales de grafos (GNN) [122, 140].

En este marco de predicción del tráfico, el contexto del usuario influye mucho en las aplicaciones y servicios móviles que solicitan las personas [141]. En consecuencia, las fluctuaciones diarias del tráfico a nivel de celda dependen en gran medida de la hora del día (p.ej., las horas de trabajo, las horas de desplazamiento, la noche) y del tipo

de actividad humana en la zona (p.ej., zonas residenciales, centros de transporte, áreas de ocio nocturno...) [8] [142] [143]. El contexto del usuario, como se ha comentado en capítulos anteriores, puede derivarse de las mediciones activas recogidas por los sensores del dispositivo del usuario final. De la misma forma, algunas redes sociales (p.ej., Twitter) comparten datos por colaboración abierta distribuida con la localización de los mensajes de texto, que pueden utilizarse para inferir la distribución espacial de los usuarios [144][145]. Una vez obtenida la información de contexto, puede aplicarse para mejorar los algoritmos SON [146, 147, 148]. En [149], se presenta un estudio de los métodos de predicción y optimización que explotan la información de contexto para prever la evolución de las condiciones de la red y asignar recursos de red de forma proactiva.

Un ejemplo de cómo el contexto del usuario influye en el rendimiento de la red son los eventos sociales masivos, que atraen a muchas personas en el mismo lugar y momento. Por este motivo, existe un creciente interés en anticipar y caracterizar este tipo de eventos. En [150], se propone una plataforma de minería de datos para detectar eventos sociales geolocalizados a partir de publicaciones geoetiquetadas en redes sociales. Otros estudios tienen en cuenta las conexiones de las redes sociales para predecir la participación en un evento social concreto [151, 152]. Afortunadamente, muchos eventos sociales se programan con antelación, lo que podría servir para anticiparse a los picos de tráfico que requieren acciones correctivas [153]. Sin embargo, estos eventos aislados quedan excluidos del análisis que realizan los operadores para la replanificación de la red, ya que ésta se centra en un horizonte temporal mucho más largo (típicamente, de 3 a 6 meses).

Es bien sabido que las arquitecturas de aprendizaje profundo superan a los modelos clásicos de análisis de series temporales para la previsión del tráfico celular, especialmente cuando las mediciones disponibles presentan una granularidad temporal y espacial fina [135] [154]. Sin embargo, todos los trabajos mencionados consideran exclusivamente la información de la red o simples factores basados en la localización independiente del tiempo (p.ej., puntos de interés en [139]) como entradas a sus modelos. Queda por comprobar si la precisión de la predicción puede mejorarse añadiendo información relacionada con factores externos, como los eventos sociales. La información sobre eventos sociales suele estar disponible públicamente para eventos futuros, de modo que los operadores de telefonía móvil pueden recopilarla fácilmente para la predicción del tráfico. Sin embargo, una vez que el evento ha tenido lugar, dicha información suele eliminarse, lo que dificulta la creación de un gran conjunto de datos

para entrenar un modelo de aprendizaje profundo. Esta dificultad puede explicar por qué la información sobre eventos sociales aún no se ha tenido en cuenta en los modelos de predicción de tráfico más avanzados. No se ha encontrado en la bibliografía ningún método para la previsión del tráfico en redes móviles basado en el aprendizaje profundo que considere los eventos sociales como entrada.

6.2. Análisis preliminar del impacto de eventos sociales en el tráfico celular

Con el objetivo de determinar si los eventos sociales tienen un impacto notable en la red, se realiza un análisis preliminar del impacto de eventos sociales en el volumen de tráfico que generan los usuarios en la red. El principal servicio usado en estos eventos son las redes sociales, tanto para compartir la experiencia que tiene el usuario como para consultar las publicaciones de su círculo de contactos, generando tráfico de descarga y de subida. Debido a esto, el estudio se realiza tanto para el enlace descendente como para el enlace ascendente. Además, para tener una mejor visión del impacto de los eventos en la red, se estudian diferentes tipos de eventos, clasificados según su naturaleza. Por último, para determinar si el impacto de cada tipo de evento depende del día de la semana, se diferencia entre influencia en días laborables y en fines de semana.

El primer problema que presenta este análisis es la obtención de suficientes datos sobre eventos sociales para que el estudio sea estadísticamente relevante. Debido al interés que suscitan y a la publicidad que reciben, la información sobre los eventos sociales futuros es fácilmente accesible antes del evento. Sin embargo, una vez realizado el evento, este pierde interés para el público, por lo que las plataformas y calendarios suelen eliminarlos de sus bases de datos para liberar espacio. Esto hace que la recolección de suficientes muestras sea una tarea ardua. Por esta razón, en este trabajo, la información sobre los eventos sociales pasados se recoge combinando diferentes fuentes:

1. Plataformas de anuncio de eventos, calendarios en línea y aplicaciones de venta de entradas. Estas plataformas en línea ofrecen información de los próximos eventos en cada ciudad. Algunas de ellas, como Eventful [155] o Yelp [156], guardan toda esta información en una base de datos, a la que se puede acceder fácilmente a través de una interfaz de programación de aplicaciones (*Application Programming Interface*, API).

2. Páginas web de ayuntamientos, periódicos locales y revistas especializadas. Las

páginas web de los ayuntamientos de las principales ciudades, así como algunos periódicos en línea y revistas especializadas, contienen un calendario con los eventos que se celebran en las ciudades. Esta información puede recogerse mediante técnicas de raspado web (*web scraping*). En algunos casos, estas fuentes no proporcionan explícitamente la latitud y la longitud del lugar del evento, por lo que deben obtenerse a partir de la dirección postal mediante técnicas de geodecodificación.

3. Organizaciones y páginas web empresariales. Las entidades dedicadas a la organización de eventos públicos (p.ej., las federaciones deportivas) y las empresas que los acogen (p.ej., las salas de conciertos o los teatros) suelen publicar los eventos programados en sus sitios web. En algunos casos, la información sobre eventos pasados también está disponible públicamente, pero en formato de texto plano, por lo que puede recogerse mediante *web scraping*.

4. Plataformas de datos abiertos. En los últimos años, el creciente interés por las técnicas de Big Data y Machine Learning ha fomentado el acceso a los datos públicos. Este hecho ha llevado a las administraciones públicas (ayuntamientos, gobiernos estatales, etc.) a crear plataformas abiertas que proporcionan datos de diversa naturaleza fácilmente accesibles para los desarrolladores. Algunas de estas plataformas incluyen una base de datos con información sobre eventos sociales.

Una vez recopilados suficientes eventos sociales, se combina esta información espaciotemporal con las medidas de la red móvil para conocer cómo influyen los eventos en el tráfico. Por un lado, se usa la información espacial para conocer qué celda da servicio al evento, mientras que, por otro, la información temporal se emplea para correlacionar las medidas de la red con el evento.

Una vez correlacionada la información de los eventos con las medidas de la red, se estima la influencia de los eventos en el tráfico celular cuantificando el tráfico adicional en un día generado por los eventos. Esto se consigue obteniendo el patrón de tráfico diario típico por celda, promediando la serie de tráfico diario en días sin eventos. A continuación, se estima la intensidad media de tráfico diario adicional causada por un evento en una celda, agregando la diferencia de tráfico en las horas del evento y redistribuyéndolo a lo largo de las 24 horas del día, como

$$\Delta T_{e,s} = \frac{\sum_{c} \sum_{t \in [1,24]} \left(T_{c,e}(t) - \overline{T_{c,s}}(t) \right)}{24} \text{ [bps]}, \tag{6.1}$$

Tabla 6.1: Categorías de eventos

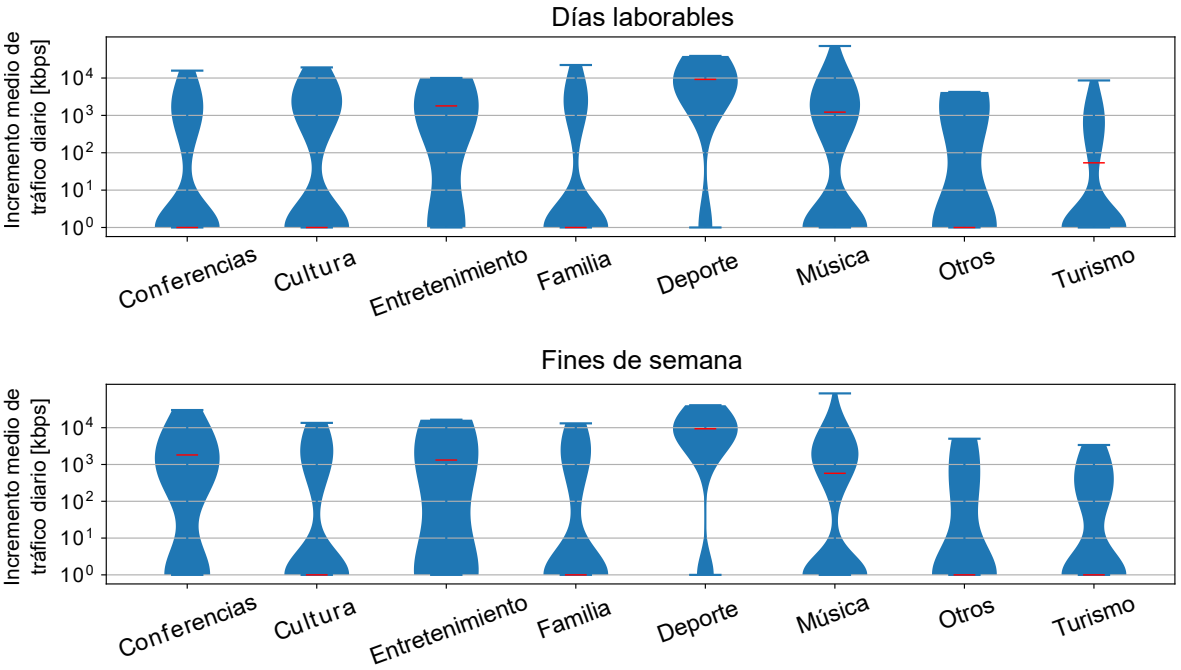
Clase	Definición
Conferencias	Eventos de carácter profesional donde se realizan ponencias de temas específicos, como cursos, talleres, conferencias, seminarios, congresos, etc.
Cultura	Eventos de contenido cultural, relacionados con cualquier variedad de arte, como exposiciones, recitales, lecturas, etc.
Entretenimiento	Eventos relacionados con el ocio y el deporte practicado por los asistentes.
Familia	Eventos de carácter social o cuyo contenido está destinado para todos los públicos.
Música	Conciertos y eventos relacionados con la música.
Deporte	Acontecimientos deportivos de carácter profesional con público.
Turismo	Eventos relacionados con la movilidad vacacional o turismo, como cruceros o partidas/llegadas de trenes.
Otros	Eventos sin una categoría específica.

donde $\Delta T_{e,s}$ es la intensidad de tráfico media adicional del tipo de evento e y el tipo de día s (es decir, día laboral/fin de semana), $\overline{T_{c,s}}(t)$ representa el patrón de tráfico diario típico de la celda c y el tipo de día s, y $T_{c,e}(t)$ es el tráfico en las horas del día en que ocurrió un evento del tipo e en la celda c. De la misma forma que se realizó con los usos de suelo en capítulos anteriores, para resumir la gran cantidad de posibles valores que adoptan las etiquetas de los eventos proporcionadas por los directorios de eventos, estos se agrupan en 8 clases, descritas en la Tabla 6.1.

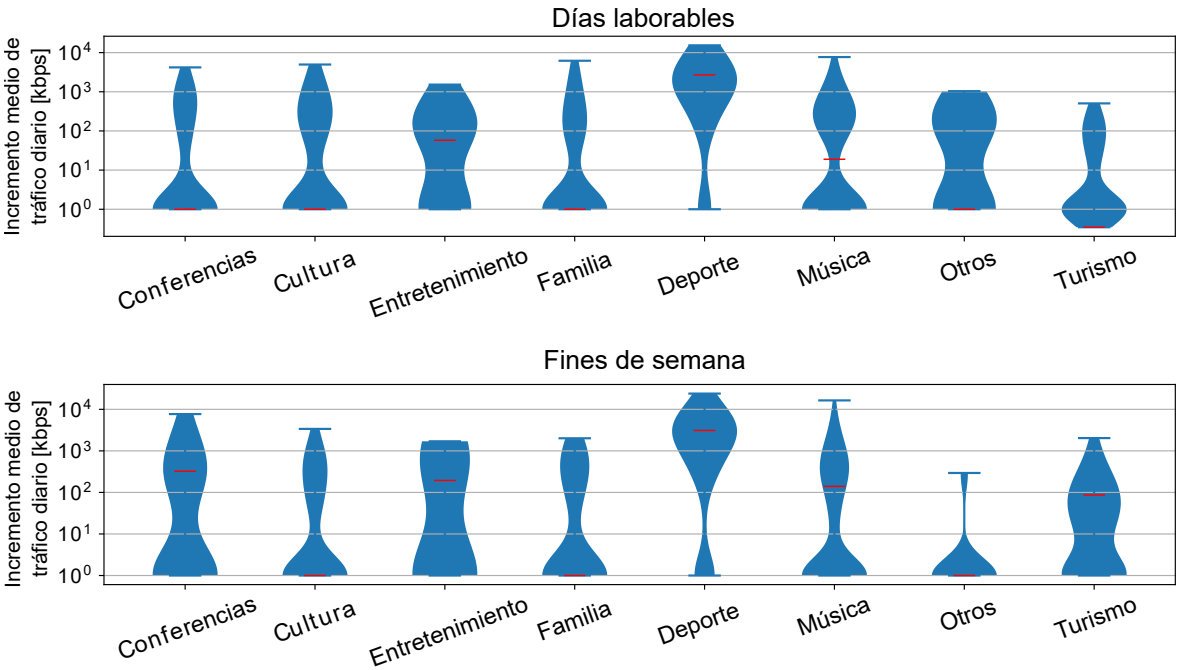
En las Fig. 6.1(a) y Fig. 6.1(b) se presentan gráficos de violín de la distribución de la intensidad de tráfico adicional por celda y día producida por diferentes eventos, segregados para días laborables/fin de semana y enlaces descendente/ascendente. Para mayor claridad, los valores de tráfico se representan en escala logarítmica, fijando a 1kbps los valores inferiores. En estas figuras, se presenta con una línea roja horizontal la mediana de la intensidad de tráfico adicional en cada tipo de evento.

En ambas figuras, se observa que la mayoría de las clases tienen una base amplia, lo que indica que un número importante de los eventos recogidos tienen una influencia insignificante en el tráfico de la red. Esto puede deberse a imprecisiones en la estimación del patrón de tráfico diario típico en la celda servidora, debidas a la problemática en la obtención de eventos, haciendo que no se recojan todos los eventos ocurridos. Otro motivo puede ser la influencia de factores en el tráfico de la red que no se han tenido en cuenta, como festividades (p.ej., la Semana Santa o verbenas de barrio) que incrementan el tráfico en celdas donde los eventos no son tan relevantes.

Sin embargo, algunas clases tienen valores extremos superiores a 10 Mbps, lo que



(a) Enlace descendente.



(b) Enlace ascendente.

Figura 6.1: Distribución de la intensidad de tráfico adicional diaria por tipo de evento.

confirma que algunos eventos tienen un fuerte impacto en el tráfico de la red. Asimismo, algunas clases (p.ej., Otros y Entretenimiento) tienen un gráfico amplio para todo el eje de ordenadas, lo que muestra una gran dispersión del impacto de esos eventos. Esta diversidad dificulta la estimación exacta del impacto de cada uno de esos eventos.

Por el contrario, unas pocas clases (p.ej., eventos deportivos) muestran dos intervalos claramente diferenciados en la distribución, con la mayoría de las muestras concentradas en torno a un valor medio superior a 1 Mbps, independientemente del canal de enlace y el día de la semana. Estas clases son las candidatas ideales para la predicción del tráfico, ya que tienen sistemáticamente un gran impacto.

Respecto al tipo de día, algunas clases de eventos (p.ej., Conferencias o Música) parecen tener un impacto similar en días laborables y en fines de semana, pero en general se aprecia un mayor impacto en fines de semana, tanto para el canal ascendente como para el descendente. En estos casos, los eventos suelen ocurrir en horarios incompatibles con la actividad laboral impidiendo la asistencia en estos días, por lo que el impacto de este tipo de eventos se centra en los fines de semana.

Comparando la Fig. 6.1(a) y la Fig. 6.1(b), se observa que el valor de la mediana en el enlace descendente es mayor en la mayoría de las clases. Por lo tanto, la influencia de los eventos en términos absolutos es mayor en el enlace descendente, lo que era de esperar ya que el tráfico de enlace descendente es, generalmente, mayor en todos los casos. Por otro lado, la comparación de la Fig. 6.1(a) y la Fig. 6.1(b) muestra que, aunque el tráfico generado en las redes celulares en los fines de semana tiende a ser menor, el impacto de los eventos en general es equivalente entre fines de semana y días laborables, siendo incluso en algunos tipos de evento mayor en los fines de semana debido a una mayor audiencia. La excepción es la clase de Turismo, cuyos eventos en fin de semana tienen un impacto menor en el enlace descendente y mayor en el enlace ascendente. Este resultado podría indicar que los asistentes a estos eventos en fin de semana tienden a subir (en lugar de descargar) contenidos (p.ej., compartir fotos y vídeos).

Aunque el análisis preliminar no muestra que todos los tipos de eventos influyan significativamente, es evidente que algunas clases de eventos sociales tienen un gran impacto en el tráfico generado en la red, por lo que es necesario tenerlos en cuenta a la hora de predecir el tráfico futuro en tareas de planificación y optimización. Debido al mayor tamaño de muestra y al mayor impacto en la red, el desarrollo del modelo de predicción de tráfico presentado a continuación se centra en las clases Deportes y Música.

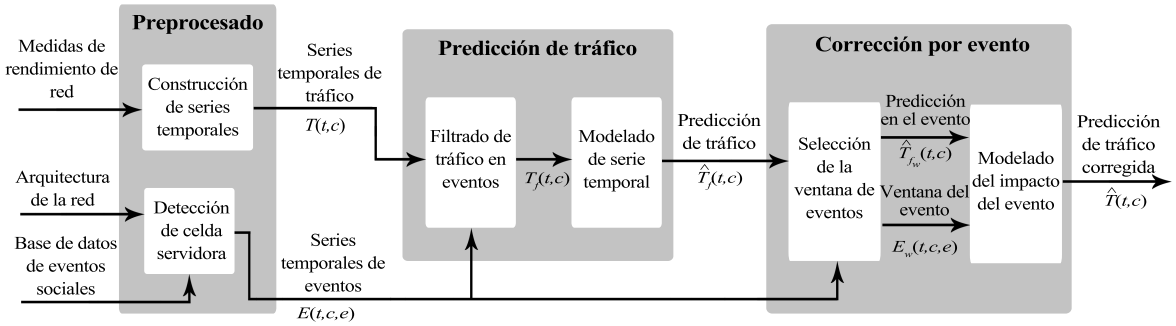


Figura 6.2: Diagrama de bloques del modelo de previsión de tráfico basado en aprendizaje profundo.

6.3. Modelo de predicción de tráfico en eventos sociales

Una vez demostrado el impacto de algunos eventos sociales en el tráfico celular, es interesante incluir esta información en la predicción de la demanda de tráfico por celda. Este es el objetivo del método propuesto en este capítulo, que pretende aumentar la precisión en la previsión del tráfico celular, corrigiendo un modelo de aprendizaje profundo para series temporales con información espaciotemporal de los eventos programados. Para ello, es necesario conocer:

a. El patrón de tráfico diario habitual de cada celda de la red.

b. La distribución espaciotemporal de los eventos en el escenario (es decir, dónde, cuándo y durante cuánto tiempo tiene lugar un evento).

c. El impacto de cada tipo de evento (p.ej., partido, concierto de música...) en el tráfico celular.

La Fig. 6.2 muestra un diagrama de bloques del modelo de predicción propuesto. Las entradas del método son: a) las mediciones de tráfico celular, recopiladas por celda y hora, b) la disposición de las estaciones base en la red celular, incluidas las coordenadas de los emplazamientos y los acimuts de las antenas, y c) la información de los eventos sociales programados que tienen lugar en el escenario (es decir, ubicación, fecha, duración y tipo de evento). Todas estas entradas proporcionan información valiosa sobre el tráfico que se transporta en la red. Por un lado, las mediciones de tráfico recolectadas dan una noción del tráfico habitual generado en cada celda. Por otro lado,

la disposición de la red y la ubicación de los eventos sociales determinan las celdas que dan servicio a los asistentes a dichos eventos. Por último, la hora programada, la duración y el tipo de evento ayudan a modelar el impacto de eventos específicos en el tráfico de las celdas. La salida del modelo es la serie temporal prevista del tráfico de cada celda del sistema desglosado por hora.

En su interior, el método consta de tres pasos. En el primero se recogen y preprocesan los datos de entrada. En el segundo paso se predice el tráfico transportado por celda en condiciones normales, es decir, sin eventos sociales. Por último, en el tercer paso se estiman las series de tráfico corregidas, incluyendo el impacto de los eventos sociales. A continuación se ofrece una explicación más detallada de cada paso.

6.3.1. Paso 1: Recolección de datos y preprocesado

En las redes celulares, las mediciones de tráfico de las celdas se recogen periódicamente en el OSS con fines de gestión de la red. Para desarrollar el modelo descrito en este capítulo, se han seleccionado las mediciones de tráfico agregadas por horas, ya que la mayoría de los eventos sociales duran un par de horas. No es necesaria una mayor resolución temporal, ya que la hora exacta de inicio y fin de los eventos no se registra en las bases de datos de eventos. Además, el periodo de recogida de datos y el área geográfica deben ser suficientemente grandes para que el tamaño de la muestra sea estadísticamente relevante, y así poder entrenar un modelo de aprendizaje profundo sin problemas de infrajuste (o *underfitting*). Más adelante se demostrará que un conjunto de medidas a nivel nacional de varios miles de celdas durante un par de meses es suficiente para obtener resultados fiables. Dichos datos deben aportarse por el operador de la red. Por último, la información sobre eventos sociales pasados se obtiene combinando las fuentes descritas en la sección 6.2.

Las mediciones de tráfico se utilizan para construir una serie temporal de tráfico por celda c y hora t, $T(t, c)$, medida en bits por segundo. Para evitar que las celdas con mayor tráfico sesguen el término de error en la etapa de entrenamiento, todas las series temporales se normalizan individualmente con la hora de máxima y mínima ocupación de la celda. A continuación, cada evento se asigna a una celda de servicio en función de sus coordenadas. Para ello, se aproxima el área de servicio de cada celda mediante la teselación de Voronoi y la sectorización basada en las coordenadas del emplazamiento y el ángulo acimutal [157]. A continuación, se utiliza la información de fecha/hora del evento para construir una serie temporal binaria de ocurrencia de eventos por celda y

tipo de evento, $E(t, c, e)$, con la misma resolución temporal que $T(t, c)$. En concreto, $E(t, c, e)$ es igual a 1 en la hora t si un suceso del tipo e tiene lugar en la zona servida por la celda c, y 0 en caso contrario.

6.3.2. Paso 2: Modelado de serie temporal

El modelado de serie temporal persigue predecir el patrón de tráfico diario que se produce en cada celda en condiciones normales, es decir, sin eventos sociales. Para ello, se selecciona una red neuronal profunda que aúna tres de las técnicas de modelado más potentes, como son el procesado mediante redes convolucionales, la realimentación con redes recurrentes LSTM y la predicción conjunta mediante estructuras multitarea (*multitask*).

Para modelar el impacto de los eventos sociales en el tráfico de las celdas antes, después y durante el evento, es necesario primero inferir de forma precisa el patrón de tráfico típico en un día sin evento. En la Fig. 6.3 se detalla la estructura de la red Convolucional-LSTM multitarea (MT-ConvLSTM), utilizado para caracterizar la fluctuación diaria normal del tráfico en una celda. Por claridad, la figura no sólo representa los bloques de construcción, sino también el tamaño de las estructuras de datos que se intercambian entre las capas. Se aprecia que la arquitectura MT-ConvLSTM consta de dos fases. En la primera fase, se utilizan tres bloques ConvLSTM independientes para procesar las medidas de tráfico de la celda analizada (celda servidora) y de sus celdas vecinas (de ahí el nombre de multitarea). En el ejemplo, se parte de un emplazamiento trisectorial, compuesto por una celda servidora y dos vecinas del mismo emplazamiento que comparten banda de frecuencia. En una segunda fase, se utiliza un cuarto bloque para predecir el tráfico diario en la celda servidora, combinando la salida de los bloques de la fase anterior. En concreto, los bloques de entrada comprenden 4 capas ocultas (2 convolucionales y 2 LSTM), cada una de ellas con 24 neuronas, para derivar características a diferentes escalas de tiempo. El bloque de salida comprende 2 capas ocultas LSTM y una capa de perceptrones de salida, cada una de 24 neuronas, generando 24 salidas (1 por hora del día) para evitar la necesidad de iterar recursivamente el modelo para cubrir el horizonte de 24 horas.

El tamaño de la entrada en cada uno de los modelos de la primera fase viene dado por la ventana de observación, que debe ser lo suficientemente grande como para capturar varios ciclos de repetición diarios y semanales. Dado que la correlación del tráfico se degrada después de algunas semanas [158], se selecciona una ventana de

observación de 3 semanas.

Cabe señalar que los modelos de series temporales están concebidos para construir un modelo por serie temporal (es decir, por celda). Sin embargo, la corta duración de las series de datos recogidas por celda (1.440 muestras para un periodo de 60 días) limita el número de puntos de datos (lapsos de tiempo) que pueden utilizarse para entrenar el modelo por celda. Para sortear este problema, se entrena un único modelo MT-ConvLSTM a nivel de red, como en [159]. Para ello, se considera el conjunto de celdas estudiado (aquellas celdas susceptibles de dar servicio en un evento social) de forma combinada, como si las series temporales de las diferentes celdas estuvieran concatenadas en una única serie temporal. Este proceso se implementa generando 3 series temporales artificiales: una primera realizada mediante la concatenación de series temporales de todas las celdas que dan servicio a las sedes en las que se producen los eventos, y otras dos series correspondientes a las vecinas coubicadas en la misma banda de frecuencias del emplazamiento trisectorial. A continuación, los conjuntos de datos utilizados para entrenar el MT-ConvLSTM se generan con todos los lapsos de tiempo de cada serie temporal de las celdas de cada grupo, evitando aquellas ventanas de observación en la unión de las series temporales de dos celdas. Así, la entrada para entrenar el modelo MT-ConvLSTM son 3 series temporales con los valores de las series temporales de tráfico $T(t, c)$ de cada celda c en el emplazamiento durante la ventana de observación t_{obs}, $\forall\, t \in [t - t_{obs} + 1, t]$. Para evitar el impacto de los eventos, los días-horas en los que se produjeron eventos sociales en el pasado se descartan dentro de la red MT-ConvLSTM mediante una capa de enmascaramiento. Debe tenerse en cuenta que, aunque el modelo se haya entrenado en toda la red, el modelo se explota a nivel de celda para prever el tráfico en celdas individuales. Así, la salida del modelo es la predicción de la serie temporal para las próximas 24 horas en la celda de servicio, es decir, $\{\hat{T}_f(t+1, c), \cdots, \hat{T}_f(t+24, c)\}\ \forall\, c$ (f de filtrado).

6.3.3. Paso 3: Corrección del impacto de eventos

A diferencia de la predicción del tráfico típico diario, la estimación del impacto de los eventos sociales se resuelve como un problema de regresión. Para simplificar el problema, se supone que el impacto de un evento se restringe a una sola celda (la que sirve al lugar donde se realiza el evento) y a una ventana temporal limitada (11 horas centradas en la hora central del evento). Esta suposición se deriva del hecho de que los eventos sociales suelen celebrarse en un único lugar con un horario específico. Así,

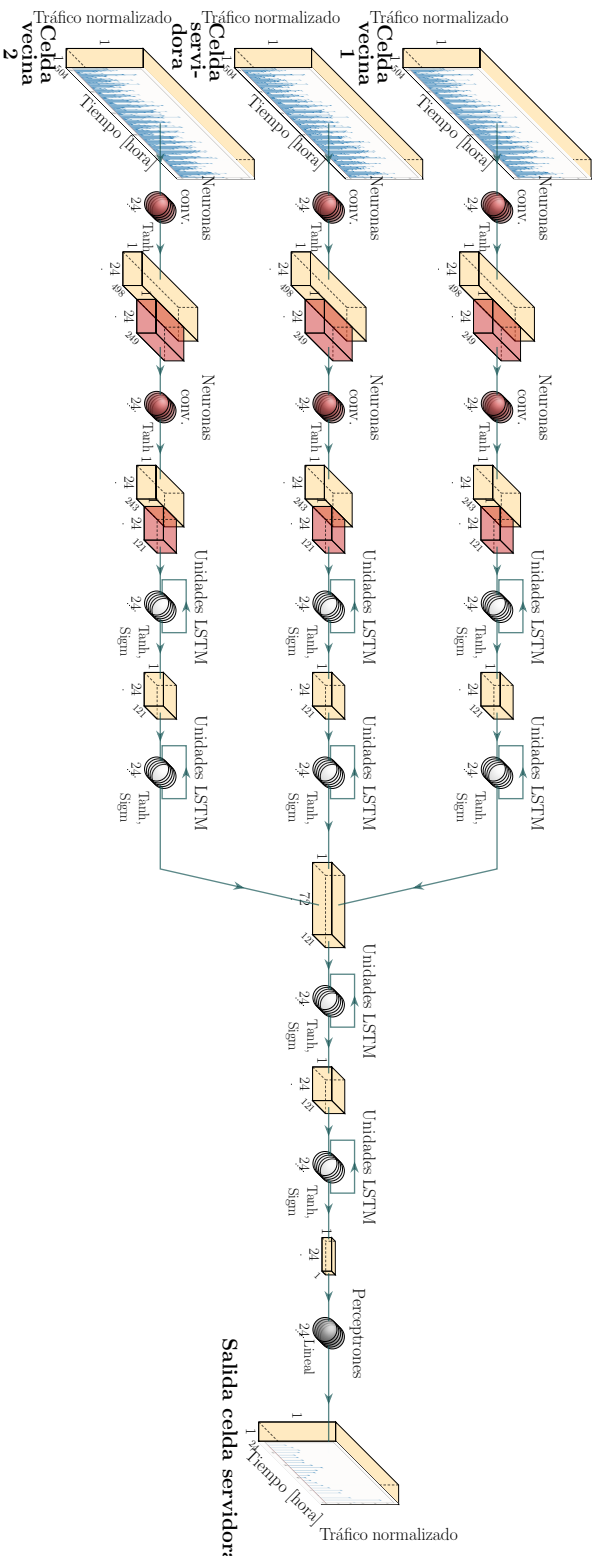


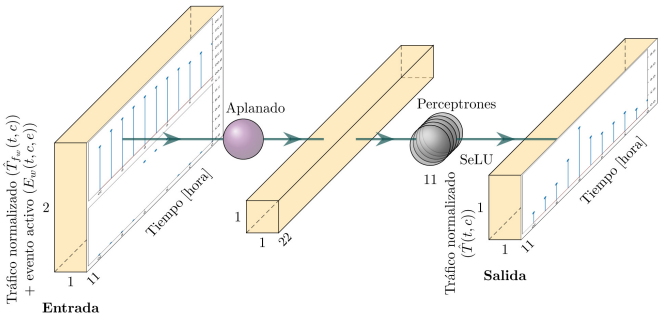
Figura 6.3: Modelo multitarea Long Short-Term Memory Convolucional para la predicción del tráfico normal (sin evento).

el pico de tráfico generado por los asistentes suele limitarse a la celda que da servicio al local mientras se celebra el evento y, tal vez, en las horas anteriores y posteriores, cuando la gente llega o se va del evento. Por lo tanto, el modelo sólo necesita actualizar la previsión de tráfico diario en la ventana temporal seleccionada en torno al evento en la celda de servicio.

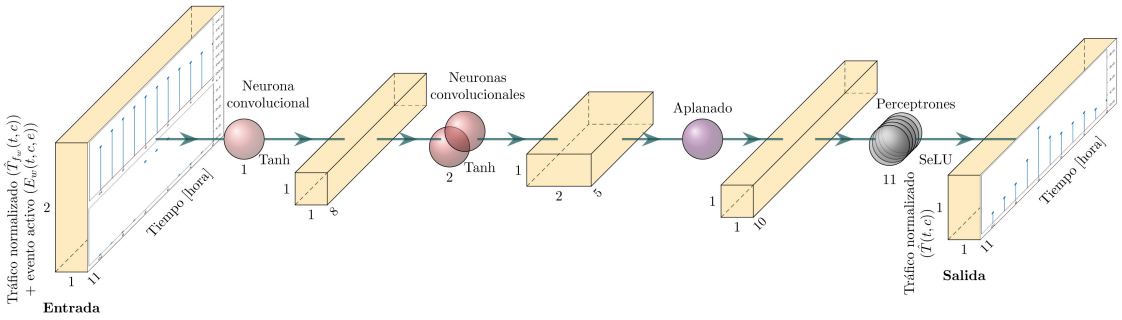
Para ello, las series temporales de tráfico normal (predicho), de ocurrencia de eventos (programados) y de tráfico real (medido) se acotan en un intervalo en torno al evento que sirven (centrado en la ventana w), generando las series temporales acotadas $\hat{T_{f}}_{w}(t,c)$, $E_w(t,c,e)$ y $T_w(t,c)$, respectivamente. Como resultado, se generan $2+N_{et}$ series temporales de 11 muestras por celda que cubren un evento, donde N_{et} es el número de tipos de eventos estudiados, es decir, el número de series temporales en $E_w(t,c,e)$. Con estas series temporales cortas, se entrena un modelo para predecir el tráfico real en la celda durante la ventana temporal considerada, $\hat{T}_w(t,c)$. En la fase de entrenamiento, cada punto del conjunto de datos corresponde a un evento que ocurre en la red, caracterizado por los valores de las entradas, $\hat{T_{f}}_{w}(t,c)$ y $E_w(t,c,e)$, y la salida, $T_w(t,c)$. Debido al diferente comportamiento observado por tipo de evento en la sección 6.2, se entrena un modelo distinto por tipo de evento para toda la red. En la fase de explotación, los valores de tráfico predichos, $\hat{T}_w(t,c)$, se utilizan para sustituir los valores de tráfico normales en las 11 horas alrededor del evento, $\hat{T}_f(t,c)$, para obtener el patrón de tráfico diario de las siguientes 24 horas, $\{\hat{T}(t+1,c),\cdots,\hat{T}(t+24,c)\}\,\forall\, c$.

Para modelar el impacto de los eventos, se prueban tres arquitecturas de red neuronal diferentes, mostradas en la Fig.6.4. La primera arquitectura, mostrada en la Fig. 6.4(a), es un perceptrón multicapa (*Multi-Layer Perceptron*, MLP). La entrada es una matriz bidimensional, que consiste en la previsión de tráfico, $\hat{T_{f}}_{w}(t,c)$, y el vector de ocurrencia de eventos, $E_w(t,c,e)$, en la ventana de tiempo del evento. Una capa de aplanamiento convierte la matriz de entrada de 2x11 en una matriz unidimensional de 1x22. La capa de salida consiste en una capa de 11 perceptrones totalmente conectados para modelar los diferentes desfases temporales en la ventana de eventos.

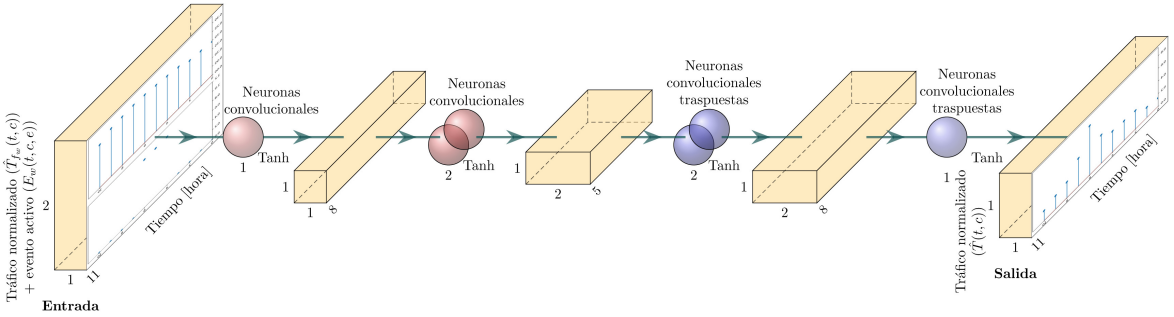
La segunda arquitectura, mostrada en la Fig. 6.4(b), es una red neuronal convolucional (*Convolutional Neural Network*, CNN) de dos capas ocultas. La primera capa convolucional fusiona la información de tráfico y de ocurrencia de eventos, mientras que la segunda capa convolucional comprime aún más la información. Para ello, se utiliza un núcleo (*kernel*) de diferente tamaño en cada capa. La capa de regresión consta de una capa que aplana la salida de la convolución y una capa de 11 perceptrones totalmente conectados.



(a) Modelo de Perceptrón Multicapa.



(b) Modelo de Red Neuronal Convolucional.



(c) Modelo de Autoencoder.

Figura 6.4: Modelos de aprendizaje profundo para la estimación del impacto de los eventos.

La tercera arquitectura, mostrada en la Fig. 6.4(c), es un modelo basado en modelos convolucionales, llamado autocodificador, (*Autoencoder*, AE). En esta arquitectura, el tráfico previsto en una celda en una determinada ventana temporal (p.ej., un día) se comprime en su información relevante a través de la ruta de contracción. Después, dicha serie temporal se remuestrea de nuevo en la ruta de expansión según el impacto de los eventos que tienen lugar en la celda.

En las dos últimas arquitecturas, se insertan capas intermedias de dilución (*dropout*) entre las capas ocultas. Esta técnica de regularización consigue no tener en cuenta algunas neuronas durante la etapa de entrenamiento, seleccionadas al azar, para reducir la sensibilidad al peso de neuronas específicas. Así, se evita el sobreajuste, mejorando

la capacidad de generalización del modelo.

6.4. Validación del modelo

El modelo descrito se evalúa con un conjunto de datos reales construido con información de eventos geolocalizados, recopilados a partir de directorios de eventos públicos, y medidas horarias de tráfico celular de una red LTE comercial durante dos meses. Para mejorar la comprensión del conjunto, primero se presenta el juego de datos usado, después se describen las características de los modelos y finalmente se presentan los resultados obtenidos por etapa del modelo. Por último, se discuten los aspectos computacionales del modelo completo.

6.4.1. Metodología experimental

Hasta ahora, las diferentes contribuciones de esta tesis se han validado sobre el mismo escenario. Lamentablemente, ese escenario no es lo suficientemente amplio (tanto espacial como temporalmente) como para tener suficiente información sobre los eventos. Por lo tanto, el nuevo escenario que contextualiza este capítulo comprende una red LTE comercial que cubre un área geográfica de más de 500.000 km^2 con 50 millones de habitantes. Los datos de tráfico se obtienen de 17.775 celdas situadas en 4.869 emplazamientos. Se recogen mediciones de tráfico por celda y hora durante 2 meses (febrero y marzo de 2019) tanto en el enlace descendente (es decir, *downlink*, DL) como en el ascendente (es decir, *uplink*, UL). Para evitar el ruido estocástico introducido por celdas de diferente comportamiento (p.ej., ubicadas en el centro de un núcleo urbano, que sirven áreas despobladas, ...), se seleccionan aquellas celdas que hayan servido, al menos, un evento social, resultando solo 323 celdas. De esta forma, se obtienen 2x323 series temporales (1 por celda y enlace) con $60 \times 24 = 1440$ mediciones de volumen de tráfico horario.

Por otro lado, combinando las fuentes enumeradas en la sección 6.2, se recogió un total de 1.367 eventos sociales categorizados para la zona y el periodo considerados. Las áreas de servicio de las celdas, obtenidas mediante el teselado de Voronoi, indican que estos eventos fueron servidos por las 323 celdas comentadas anteriormente, lo que demuestra que algunas celdas atendieron múltiples eventos. Para reducir el número de modelos, los eventos se agrupan en las 8 clases descritas en la Tabla 6.1. Para resumir la distribución de eventos, las Fig. 6.5(a)-(b) muestran 2 diagramas de sectores con

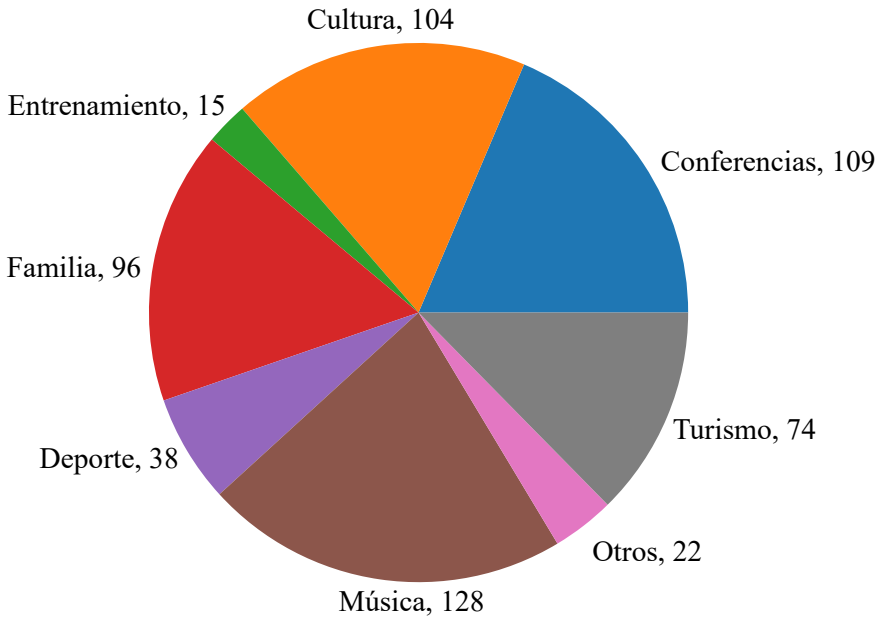
Tabla 6.2: Selección de hiperparámetros para los modelos clásicos de series temporales.

Holt-Winters	
Tendencia	Aditiva
Estacionalidad	Aditiva
Periodicidad	24
SARIMA	
(p, d, q)	(1, 0, 1)
(P, D, Q)	(1, 1, 1)
m	24

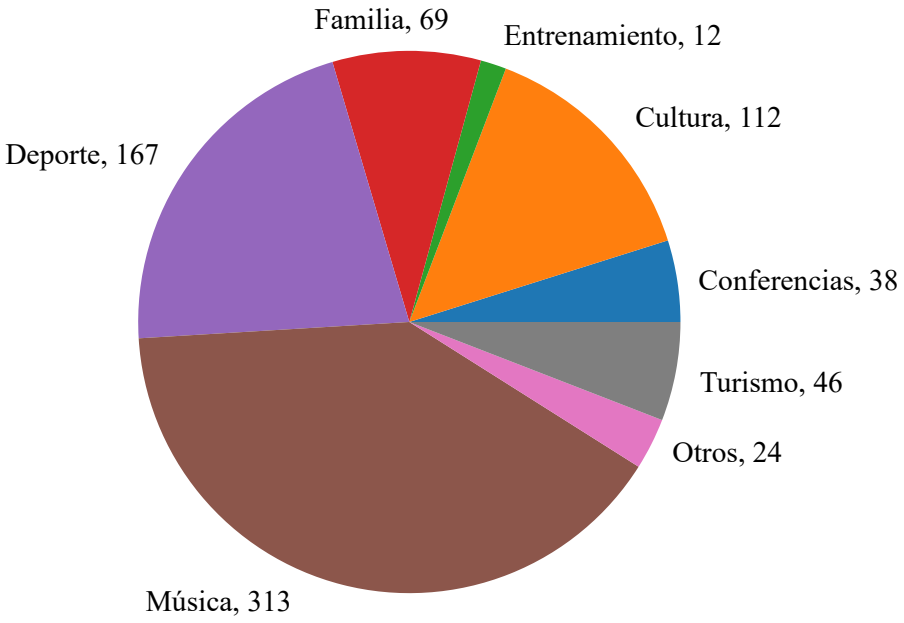
el recuento de eventos por clase, tanto en días laborables como en fines de semana. Se observa que la categoría más común en el conjunto de datos es la música, con 441 eventos.

El análisis preliminar presentado en la Sección 6.2 demuestra la influencia de los eventos en el tráfico celular cuantificando el tráfico adicional agregado en un día generado por los eventos. Aun así, el procedimiento seguido no es capaz de estimar de forma precisa el patrón horario típico de tráfico en cada celda, por lo que se realiza un análisis más detallado de dicho patrón con un modelo de aprendizaje profundo más sofisticado. Para ello, se utiliza un único modelo MT-ConvLSTM entrenado con las series temporales de las 323 celdas del escenario para pronosticar las fluctuaciones normales del tráfico por celdas.

Para justificar la elección de la red MT-ConvLSTM, el modelo se compara con dos esquemas clásicos de análisis de series temporales, a saber, Holt-Winters [121] y SARIMA [123]. Como estos últimos no consiguen captar la periodicidad a largo plazo, sólo para estos dos métodos se genera un modelo independiente por celda para predecir el tráfico normal en días laborables y fines de semana, respectivamente. Para ello, se crean dos series temporales artificiales con días laborables o fines de semana para entrenar los modelos por separado. En ambos casos, la ventana de observación es de 3 semanas y los algoritmos se aplican recursivamente para predecir el tráfico en las 24 horas del día siguiente (es decir, el horizonte de 24 horas). El modelo propuesto también se compara con enfoques más modernos que los esquemas clásicos de análisis, como una LSTM básica, una Convolutional-LSTM (ConvLSTM) [160], una red Convolucional Dilatada (D-Conv) [161] y una LSTM Multitarea (MT-LSTM) [44]. Las Tablas 6.2 y 6.3 presentan los ajustes finales de hiperparámetros seleccionados para los algoritmos clásicos y de aprendizaje profundo en la etapa de predicción de tráfico normal, respectivamente. Para determinar qué modelo obtiene los mejores resultados, la cifra de mérito escogida es el Error Medio Absoluto (*Mean Absolute Error*, MAE).



(a) Días laborables.



(b) Fines de semana.

Figura 6.5: Distribución de los tipos de evento en días laborables y fines de semana.

Tabla 6.3: Selección de hiperparámetros para los modelos de aprendizaje profundo.

Hiperparámetros comunes	
Función de error	*Mean Absolute Error* (MAE)
Algoritmo de entrenamiento	*Adaptive Moment Estimation* (Adam)
Número de perceptrones en capas densas	24
Función de activación en capas densas	*Lineal*
Algoritmo de inicialización de pesos	Glorot constante
División de datos	80 % entrenamiento / 20 % evaluación
Regla de interrupción	5 épocas sin mejorar
Tamaño de lote	128
LSTM	
Número de capas	4 (3 LSTM, 1 densa)
Número de celdas LSTM por capa	24
Número de parámetros en el modelo	12.504
LSTM convolucional	
Número de capas	6 (2 convolucionales, 3 LSTM, 1 densa)
Número de neuronas convolucionales por capa	24
Número de celdas LSTM por capa	24
Función de activación en capas convolucionales	Tangente hiperbólica (*tanh*)
Número de parámetros en el modelo	18.960
Convolucional dilatada	
Número de capas	4 (3 convolucionales, 1 densa)
Número de neuronas convolucionales por capa	24
Ratio de dilatación	2
Función de activación en capas convolucionales	*tanh*
Número de parámetros en el modelo	38.280
LSTM multitarea	
Número de capas etapa de entrada	2 LSTM
Número de capas etapa de salida	3 (2 LSTM, 1 densa)
Número de celdas LSTM por capa	24
Número de parámetros en el modelo	36.216
LSTM convolucional multitarea	
Número de capas etapa de entrada	4 (2 convolucionales, 2 LSTM)
Número de capas etapa de salida	3 (2 LSTM, 1 densa)
Número de neuronas convolucionales por capa	24
Número de celdas LSTM por capa	24
Función de activación en capas convolucionales	*tanh*
Número de parámetros en el modelo	55.584

Con estos hiperparámetros, las redes mencionadas se entrenan con las series temporales de tráfico de las 323 celdas con eventos sociales, tanto del enlace descendente como del enlace ascendente.

Una vez demostrada la superioridad de la arquitectura MT-ConvLSTM en el modelado del tráfico usual de las celdas, se entrenan los modelos de corrección de la serie temporal con el impacto de los eventos. La mejora derivada de añadir el impacto de

los eventos se cuantifica comparando la red básica MT-ConvLSTM, que proporciona la previsión de tráfico normal, con el sistema que amplía la MT-ConvLSTM con un MLP, CNN o AE para modelar el impacto de los sucesos. Dado que los diferentes eventos tienen un impacto diferente, se deriva un modelo separado por clase de evento. Para simplificar, el análisis se centra en dos de los eventos más comunes (deportivos y musicales). En ambos modelos, cada punto de datos consta de un vector de entrada de 2x11 valores con el tráfico normal y los valores de ocurrencia del evento y un vector de salida de 11 valores con el tráfico corregido.

De la misma forma que las Tablas 6.2 y 6.3, la Tabla 6.4 presenta la selección final de hiperparámetros de los modelos testeados para la etapa de la corrección por evento.

Las estructuras construidas con estos hiperparámetros se emplean para entrenar 2 modelos, uno para los eventos clasificados como deportivos (142 eventos) y otro para los musicales (160 eventos), tanto para DL como para UL, resultando en un total de 4 modelos.

Todos los modelos se implementan con las bibliotecas *scikit-learn*, *statsmodels* y *Keras*. El ajuste de los hiperparámetros (p.ej., número de neuronas por capa, funciones de activación...) se realiza mediante una búsqueda en malla dentro del espacio de parámetros [162]. Dicha acción se realiza explorando los posibles valores que pueda adoptar cada uno de los hiperparámetros ajustados, probando cada combinación y seleccionando aquellos que optimicen los resultados del modelo. En concreto, se exploraron diferentes funciones de activación (lineal, sigmoide, tanh, *Regularized Linear Unit* -ReLU-, y SELU), diferente número de capas convolucionales y convolucionales traspuestas (en ambos casos, 1 y 2), y diferente número de capas densas (1, 2). Por simplicidad, el criterio de selección de hiperparámetros óptimos es la combinación que mejor resultado da en uno de los cuatro modelos (en concreto, el modelo para eventos deportivos en DL).

6.4.2. Resultados

Modelado de serie temporal

Primero, se evalúa el aprendizaje del modelo con el juego de datos, sin normalizar. Para ello, la Tabla 6.5 muestra el MAE obtenido con los diferentes modelos, tanto para DL como para UL. Se puede comprobar cómo los modelos de aprendizaje profundo (LSTM, ConvLSTM, D-Conv, MT-LSTM y MT-ConvLSTM) obtienen resultados mu-

Tabla 6.4: Selección de hiperparámetros para los modelos en la etapa de corrección por evento.

Hiperparámetros comunes	
Función de error	MAE
Algoritmo de entrenamiento	Adam
Función de activación en capas convolucionales	*tanh*
Función de activación en capas densas	*Scaled Exponential Linear Unit* ($SELU$)
Algoritmo de inicialización de pesos	Glorot constante
División de juegos	80 % entrenamiento / 20 % evaluación
Regla de interrupción	70 épocas
Tamaño de lote	10 (deporte) / 3 (música)
Perceptrón Multicapa	
Número de capas	2 (aplanado y de salida)
Número de neuronas en la capa de salida	11
Número de parámetros del modelo	253
Red Neuronal Convolucional	
Número de capas	4 (2 convolucionales, 1 aplanado y 1 densa)
Número de núcleos por capa convolucional	1 (primera capa), 2 (segunda capa)
Tamaño de los núcleos	(2,4), (1,4)
Ratio de *Dropout*	0,2
Número de neuronas en capa densa	11
Número de parámetros del modelo	140
Autocodificador	
Número de capas	5 (2 convolucionales, 2 convolucionales transpuestas y 1 aplanado)
Número de núcleos por capa convolucional	1, 2
Tamaño de los núcleos por capa convolucional	(2,4), (1,4)
Número de núcleos por capa convolucional transpuesta	1, 1
Tamaño de los núcleos por capa convolucional transpuesta	(1,4), (1,4)
Ratio de *Dropout*	0,2
Función de activación en capas convolucionales transpuestas	$tanh, SELU$
Número de parámetros del modelo	46

cho mejores que los modelos clásicos de series temporales, obteniendo una reducción del 23,6 % en DL y 6,2 % en UL entre el peor modelo de aprendizaje profundo (LSTM) y el mejor de los modelos clásicos (Holt-Winters). Por otro lado, dentro del aprendizaje profundo, los modelos multitarea obtienen mejores resultados que los modelos individuales, reduciendo un 6,5 % y un 2,8 % el MAE para DL y UL entre los modelos más cercanos (ConvLSTM y MT-LSTM), por lo que se confirma la importancia de incluir las celdas vecinas en el proceso de predicción del tráfico de celda. En concreto, la arquitectura LSTM convolucional multitarea obtiene los mejores resultados para ambos

Tabla 6.5: Error medio absoluto de los diferentes modelos por sentido del enlace [kbps].

Modelo	Descendente	Ascendente
Holt-Winters	1092.9	182.3
SARIMA	2474.8	365.1
LSTM	835,0	171,0
ConvLSTM	747,3	130,2
D-Conv	833,8	140,6
MT-LSTM	698,9	126,5
MT-ConvLSTM	692,0	114,6

Tabla 6.6: Tiempo total de entrenamiento del modelo completo [s].

Modelo	Downlink	Uplink
LSTM	45.606	55.883
ConvLSTM	34.799	19.760
D-Conv	4.815	4.724
MT-LSTM	592.384	282.267
MT-ConvLSTM	93.900	173.313

enlaces, mejorando el MAE un 1 % en DL y un 9,4 % en UL respecto a la arquitectura LSTM multitarea. Esto demuestra que la LSTM convolucional multitarea es el modelo que mejor se adecúa a la complejidad de las fluctuaciones de tráfico en la red celular.

Por último, para comparar la adecuación de los modelos de aprendizaje profundo, se presenta el proceso de entrenamiento en la Tabla 6.6 y en la Fig. 6.6. Específicamente, la Tabla 6.6 muestra el tiempo de entrenamiento de los modelos evaluados en ambos enlaces. Se puede observar que los modelos construidos únicamente con neuronas convolucionales tardan mucho menos en entrenarse que los que se basan en redes recurrentes. Aun así, tanto en los modelos individuales como en los modelos multitarea, la inclusión de capas convolucionales reduce significativamente el tiempo de entrenamiento de las redes recurrentes. Para el experimento realizado, el tiempo de entrenamiento se reduce un 23,7 %/64,6 % en los modelos individuales y un 84,1 %/38,6 % en los modelos multitarea, aunque estos valores cambian en cada proceso de entrenamiento, debido al inicio aleatorio de los pesos de las neuronas al principio del entrenamiento con el algoritmo de inicialización Glorot constante [163].

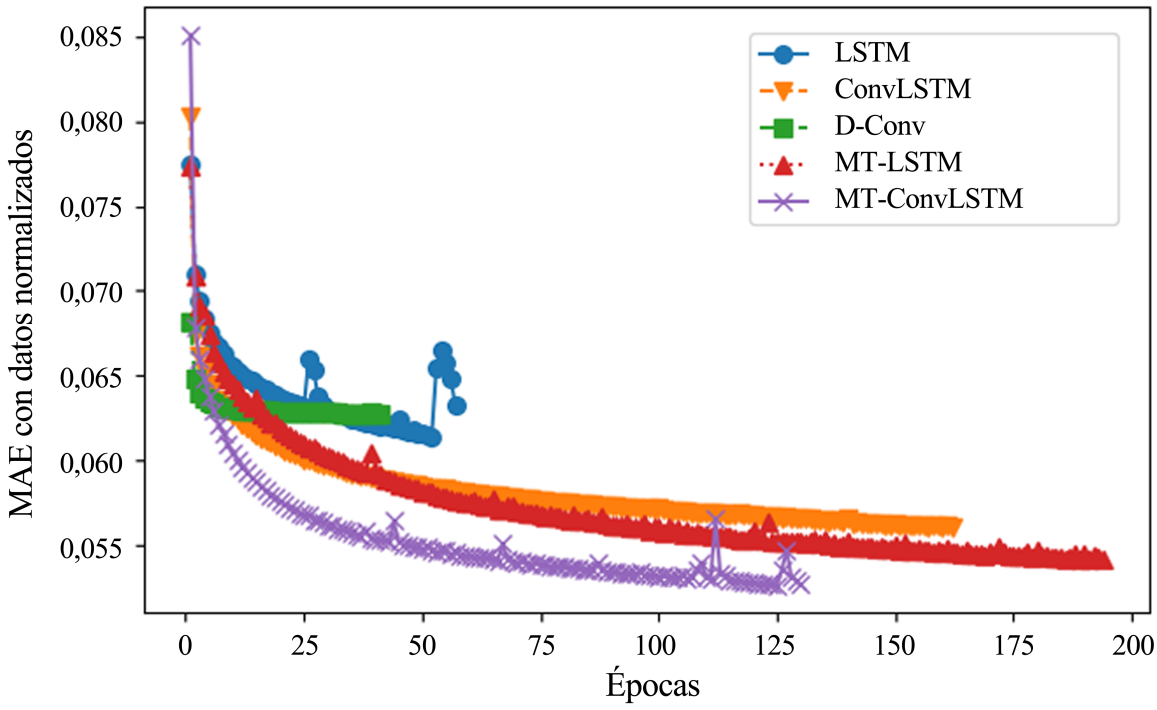
La Fig. 6.6 presenta la evolución de la función de error por época de entrenamiento. Hay que destacar que durante el proceso de entrenamiento, los datos de las series temporales se normalizan, por lo que los datos de MAE en estas gráficas se obtienen de valores en el rango entre 0 y 1. En estas figuras se evidencia que los modelos multitarea obtienen mejores resultados tanto en el enlace descendente (Fig. 6.6(a)) como en el enlace ascendente (Fig. 6.6(b)), mostrando la influencia de las celdas vecinas en el

tráfico. En concreto, los modelos de redes convolucionales dilatadas son los que más rápido convergen, pero peores resultados obtienen. Por otro lado, la inclusión de capas convolucionales mejora notablemente el entrenamiento de la red LSTM simple, tanto en modelos individuales como multitarea, haciendo que converjan más rápido y obtengan mejores resultados. La mejora en los resultados obtenidos se debe a la extracción de características que realizan las capas convolucionales, obviando información innecesaria de la serie temporal, mientras que la reducción en el tiempo de entrenamiento se debe a la reducción de la longitud de cada punto de datos a la entrada de las capas LSTM, reduciendo el tiempo necesario en el proceso de actualización de los estados de memoria. Nótese el caso del enlace ascendente, donde la reducción global de tiempo entre el modelo MT-ConvLSTM y la MT-LSTM es mucho menor que en el enlace descendente, pero debido a la regla de finalización del entrenamiento fijada, en este experimento MT-ConvLSTM duplica el número de épocas de MT-LSTM.

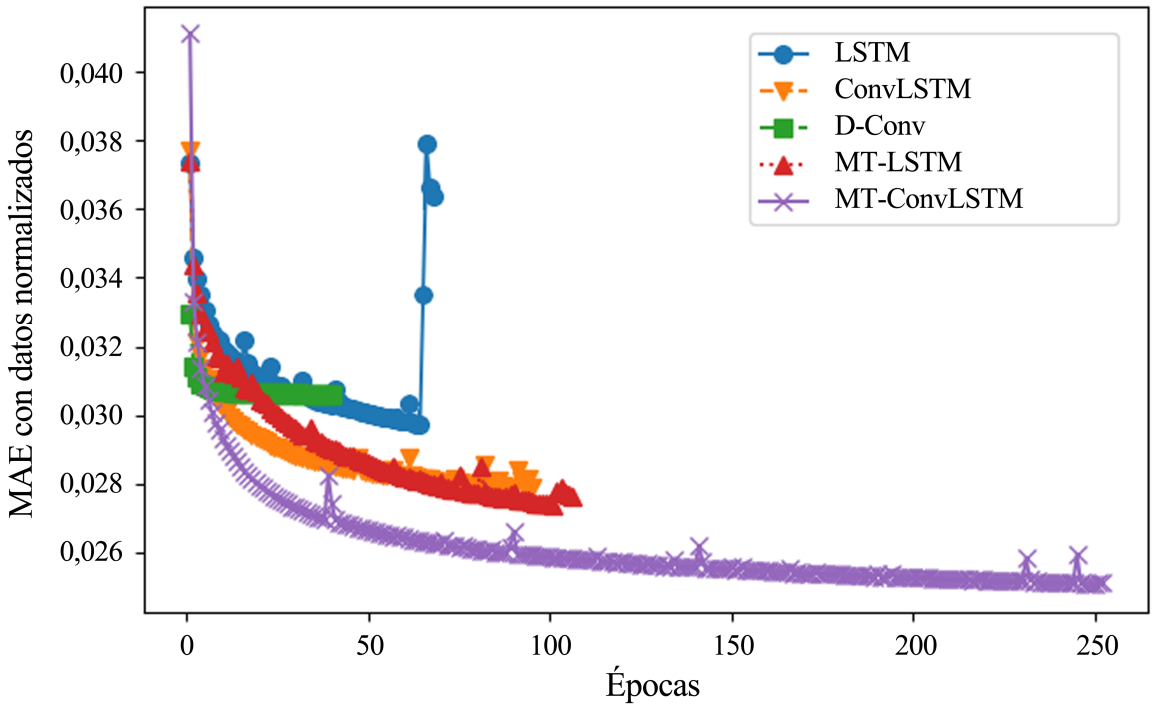
Corrección de predicción de tráfico debido a eventos

La Tabla 6.7 presenta el MAE global alcanzado por los diferentes modelos de corrección de eventos para las celdas/días con eventos deportivos y musicales, respectivamente. Cabe precisar que el análisis se limita a las horas que rodean al evento. Se observa que los tres modelos que consideran el impacto de los eventos mejoran al modelo básico MT-ConvLSTM en ambos tipos de eventos, ya que obtienen un error inferior. Con cualquiera de los tres modelos de corrección de eventos se obtiene una reducción de MAE de una tercera parte para los eventos deportivos, mientras que, para los eventos musicales, la reducción de MAE es mucho menor (no llegando al 5 % en el mejor de los casos). Este resultado indica claramente que los eventos deportivos recopilados en las bases de datos tienen una gran influencia en el tráfico celular y su impacto es predecible, mientras que la menor audiencia y a la mayor heterogeneidad de los eventos musicales almacenados en las bases de datos de eventos atenúan su impacto en la red.

Concretamente, en el enlace descendente la reducción de MAE es del 36,7 % y 4,7 % con MLP, 37,5 % y 4,5 % con CNN y 34,1 % y 1,6 % con AE para los eventos deportivos y musicales, respectivamente. Por otro lado, en el enlace ascendente se obtienen resultados similares, detallados en la Tabla 6.7, donde la reducción del MAE para los eventos deportivos varía del 22 % (AE) al 31,2 % (MLP), y del 2,6 % (AE) al 4,1 % (MLP) para los eventos musicales. En concreto, el modelo que muestra el mejor rendimiento de MAE en la mayoría de los casos es MT-ConvLSTM+MLP, obteniendo



(a) Enlace de bajada.



(b) Enlace de subida.

Figura 6.6: Proceso de convergencia durante el entrenamiento de los modelos de aprendizaje profundo.

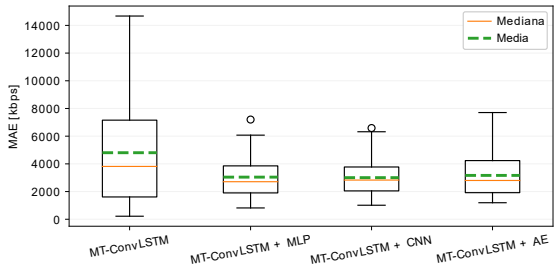
Tabla 6.7: Error Medio Absoluto global tras la etapa de corrección por evento [kbps].

Enlace descendente		
	Deporte	Música
MT-ConvLSTM	4805,5	1511,7
MT-ConvLSTM + MLP	3042,1	1440,2
MT-ConvLSTM + CNN	3005,0	1443,4
MT-ConvLSTM + AE	3166,5	1487,3
Enlace ascendente		
	Deporte	Música
MT-ConvLSTM	1849,8	449,3
MT-ConvLSTM + MLP	1272,9	430,9
MT-ConvLSTM + CNN	1301,7	431,0
MT-ConvLSTM + AE	1443,2	437,5

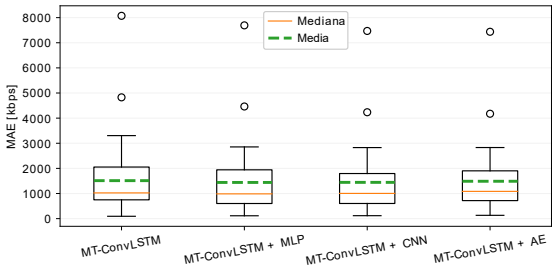
un MAE de 1440,2 kbps para los eventos musicales en DL y 1272,9 y 430,9 kbps para los eventos deportivos y musicales, respectivamente, en UL. En cambio, MT-ConvLSTM+CNN consigue el MAE más bajo para los eventos deportivos en DL, con 3005,0 kbps, obteniendo un rendimiento muy cercano a MT-ConvLSTM+MLP en el resto de los casos. En todos los casos, AE muestra el peor rendimiento tanto en DL como en UL, lo que demuestra que, con los datos disponibles, AE no es capaz de modelar el impacto de los eventos tan bien como otras estructuras.

Para comprobar las diferencias entre eventos, las Fig. 6.7(a)-(d) muestran un diagrama de cajas de la distribución de MAE por evento con los diferentes modelos, desglosado para eventos deportivos y musicales en DL y UL. Los límites inferior/superior de los recuadros reflejan los percentiles 25/75, mientras que las líneas sólidas/discontinuas del centro representan los valores medios y medianos, respectivamente. Los círculos determinan los valores atípicos. Los valores medios/medianos que se observan confirman la mencionada reducción de MAE en DL y UL. Igualmente importante es que la desviación máxima se reduce significativamente en todos los casos al considerar el impacto de los eventos. Por ejemplo, el error experimentado en el peor evento deportivo en el enlace descendente disminuye casi a la mitad con cualquiera de los modelos de corrección de eventos propuestos. En el enlace ascendente, se observa una reducción menor en todos los eventos deportivos. Como se observa en la Tabla 6.7, el método con mayor reducción de desviaciones extremas en la mayoría de los casos es MT-ConvLSTM+MLP, mientras que MT-ConvLSTM+CNN tiene un rendimiento similar.

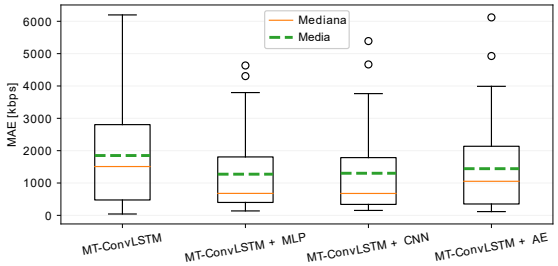
Para comprobar el impacto de los eventos sobre una base horaria, las Fig. 6.8(a)-(d) comparan el patrón típico de tráfico horario para eventos deportivos y musicales en DL y UL derivado con los diferentes modelos. Dicho patrón se define como el valor de tráfico



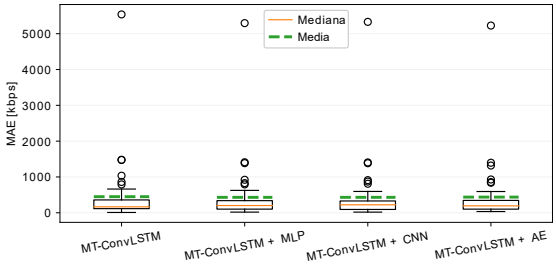
(a) Enlace descendente en eventos deportivos.



(b) Enlace descendente en eventos musicales.



(c) Enlace ascendente en eventos deportivos.



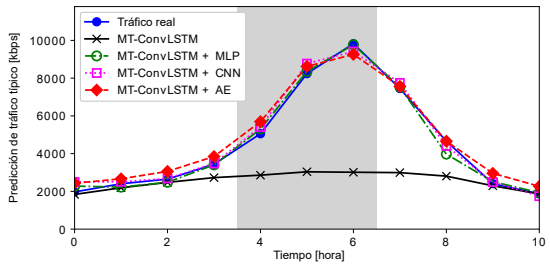
(d) Enlace ascendente en eventos musicales.

Figura 6.7: Precisión de los modelos por enlace y tipo de evento.

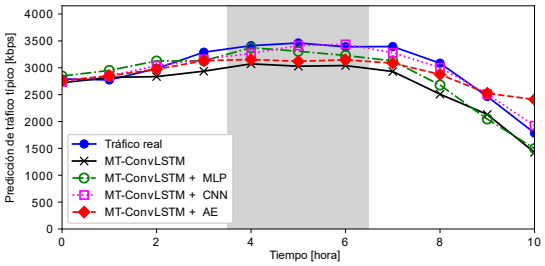
medio entre las instancias de eventos por clase y enlace. A efectos de comparación, el intervalo típico de duración del evento, extraído del vector de ocurrencia del evento, está sombreado en gris. En la Fig. 6.8(a), se muestra que el patrón de tráfico real de DL durante los eventos deportivos difiere en gran medida del tráfico normal predicho por MT-ConvLSTM. En cambio, todos los modelos de corrección de eventos se aproximan al patrón de tráfico real con precisión. En la Fig. 6.8(b), se observa que lo mismo ocurre con los conciertos, aunque el tráfico normal predicho por MT-ConvLSTM se acerca más a los valores reales ya que el impacto de los eventos musicales es más débil.

También se observan algunas diferencias importantes en el perfil de tráfico asociado a cada evento. En los eventos deportivos, se observa un claro pico de tráfico al final del partido, con una corta cola sólo después del mismo, originada, probablemente, por la gente que se queda en los alrededores del recinto. En cambio, en los conciertos, el tráfico se dispersa en un periodo más largo alrededor del evento, posiblemente debido a que los asistentes llegan antes y se marchan más tarde del evento. Se observan tendencias similares en el UL, mostradas en las Fig. 6.8(c)-(d). Este hecho justifica la adopción de modelo distinto para cada tipo de evento.

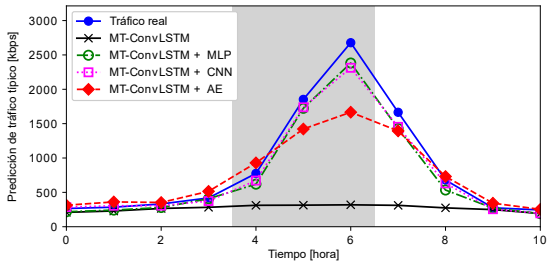
Una inspección más detallada de los patrones de tráfico por hora de los tres modelos de corrección de eventos revela que MT-ConvLSTM+CNN proporciona patrones más suaves debido a la operación de convolución en el eje temporal. La forma del patrón resultante se aproxima más a la real, especialmente en los eventos musicales,



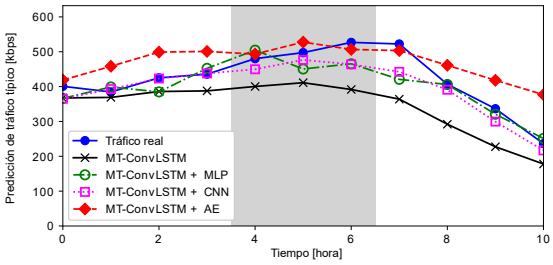
(a) Enlace de bajada en eventos deportivos.



(b) Enlace de bajada en eventos musicales.



(c) Enlace de subida en eventos deportivos.



(d) Enlace de subida en eventos musicales.

Figura 6.8: Patrón horario del tráfico predicho en torno al evento.

que se presentan en la Fig. 6.8(b) y (d). Por el contrario, nótese las fluctuaciones en el patrón UL obtenido por MT-ConvLSTM+MLP. Queda por comprobar si, con un conjunto de datos de eventos más grande o una arquitectura CNN más sofisticada, MT-ConvLSTM+CNN podría superar a MT-ConvLSTM+MLP. Para ello, habría que confirmar que el patrón de tráfico real también es suave evento por evento (y no sólo el patrón de tráfico típico).

6.4.3. Complejidad computacional

El método propuesto necesita recoger y preprocesar las medidas de tráfico de las celdas y los datos de los eventos. El tiempo de ejecución del preprocesamiento de las mediciones de tráfico crece linealmente con el número de celdas y las horas por celda. Del mismo modo, el tiempo de ejecución del preprocesamiento de eventos crece linealmente con el número de eventos en la zona geográfica considerada.

Para el modelo completo, compuesto por la etapa de predicción de tráfico típico y corrección de eventos, la complejidad temporal del algoritmo de entrenamiento en el peor de los casos viene dada por la red MT-ConvLSTM utilizada para predecir el patrón de tráfico horario normal. La complejidad computacional de las capas convolucionales es $\mathcal{O}(k \cdot n \cdot d \cdot f)$, donde k es el tamaño del *kernel* de las convoluciones, n es la longitud de la secuencia (ventana), d es la dimensión de representación (es decir, eje temporal)

Tabla 6.8: Tiempos de entrenamiento en DL [s].

Etapa	Arquitectura	Tipo de evento	Tiempo
Predicción del tráfico típico	MT-ConvLSTM	-	29.870,00
Impacto de los eventos	MLP	Deportes	2,32
		Música	4,98
	CNN	Deportes	3,66
		Música	7,25
	AE	Deportes	4,85
		Música	8,41

y f es el número de filtros de la capa [164]. Por otro lado, una LSTM simple es local en el espacio (porque el tamaño de la red no influye en la complejidad de actualización de la red por peso y muestra temporal) y en el tiempo (porque la longitud de la secuencia de entrada no influye en los requisitos de almacenamiento del método), por lo que su complejidad computacional es $\mathcal{O}(N_w)$, donde N_w es el número de pesos [42].

Como ejemplo del coste temporal del modelo, en la Tabla 6.8 se presentan los tiempos de entrenamiento sólo en un enlace (DL) con un conjunto de datos de 226.176 ventanas de 3 semanas de tráfico celular por hora, y un conjunto de eventos que comprende 142 partidos deportivos y 160 conciertos con una ventana de tiempo de 11 horas, empleando un ordenador portátil con un procesador Intel de cuatro núcleos, frecuencia de reloj de 1,8 GHz y 16 GB de RAM. Se observa que las capas convolucionales añadidas en las estructuras más complejas (CNN y AE) aumentan ligeramente los tiempos de entrenamiento. No obstante, los tiempos de ejecución son insignificantes para las tareas de replanificación, en las que las previsiones de tráfico deben obtenerse con varios días de antelación para planificar las acciones correctoras de los problemas de capacidad derivados de los eventos programados.

6.5. Conclusiones

Una previsión precisa del tráfico futuro será clave para gestionar las futuras redes 5G, en las que las predicciones tendrán que hacerse con más frecuencia y mayor granularidad debido al menor tamaño de las celdas y a las características de virtualización, como en el caso de la funcionalidad de *network slicing*. Como se ha podido comprobar, el tráfico de celda se ve fuertemente afectado por los eventos sociales. En este capítulo, se presenta un nuevo método que, siendo consciente del contexto social de los usuarios, mejora la previsión del tráfico con información de eventos programados. La arquitectura de aprendizaje profundo propuesta utiliza la información de eventos

sociales públicos para corregir las previsiones horarias de tráfico por celda, teniendo en cuenta los picos inusuales de tráfico generados por la concentración espaciotemporal de usuarios. La evaluación del modelo se ha realizado sobre un conjunto de datos reales de 2 meses de mediciones de tráfico y eventos en una red LTE en funcionamiento.

Un análisis preliminar ha mostrado que los eventos sociales tienen un impacto significativo en el tráfico generado en las celdas, tanto en el enlace descendente como en el ascendente, obteniendo en ambos enlaces picos de intensidad de tráfico adicional en torno a 1-10 Mbps para todos los tipos de evento y día de la semana. También se ha confirmado que una red Convolucional LSTM multitarea es la mejor técnica para capturar las fluctuaciones normales a lo largo de un día, incluso con un único modelo compartido por todas las celdas de la red, obteniendo una reducción de error de 7,4 % (DL) y 12 % (UL) respecto a la mejor arquitectura simple y de 1 % (DL) y 9,4 % (UL) respecto a la otra arquitectura multitarea.

Se han probado tres arquitecturas diferentes para modelar los picos de tráfico generados por eventos deportivos y musicales: un perceptrón multicapa, una red convolucional básica y un autocodificador. El perceptrón multicapa ha obtenido el mínimo error de predicción en la mayoría de los casos (obteniendo un error de 1440,2 en DL para eventos musicales y de 1272,9 y 430,9 en UL para eventos deportivos y musicales), pero la red convolucional ha logrado resultados similares con menos parámetros, mostrando su potencial para conjuntos de datos de entrenamiento pequeños. Una vez entrenado el modelo, la baja carga computacional de las operaciones subyacentes permite una fácil integración en las herramientas de planificación radio.

El método propuesto puede predecir el patrón de tráfico local generado por los eventos sociales con 24 horas de antelación, lo cual es vital para planificar las infraestructuras de redes móviles (p.ej., celdas temporales). Enfoques anteriores se basan en planificar las acciones a tomar en función de los horarios de los eventos compartidos por las autoridades locales o en la experiencia de los operadores en el pasado. Esta forma de abordar el problema no se adapta bien a los sistemas 5G, en los que el menor tamaño de las celdas reducirá el tamaño de los eventos a tener en cuenta. En este contexto, las predicciones de tráfico diario pueden utilizarse para reconfigurar automáticamente la cantidad de recursos radio asignados a los distintos segmentos (*slices*) de la red por celda y por día.

Uno de los principales inconvenientes del método es la necesidad de una gran base de datos de eventos. Esta información sólo puede conseguirse combinando múltiples

fuentes, algunas de las cuales requieren conocimientos de *web scraping*. Aun así, sólo unas pocas clases de eventos (p.ej., deporte y música) tienen suficientes eventos para entrenar modelos de aprendizaje profundo. Para mejorar la precisión de la predicción, líneas futuras de desarrollo pueden incluir nuevas entradas en el módulo de corrección de eventos, como el número esperado de asistentes por evento, estimado por el tamaño del recinto, o el uso del suelo del lugar que alberga el evento (p.ej., interior/exterior). Asimismo, el modelo puede ampliarse fácilmente para predecir otros indicadores de rendimiento diferentes del volumen de tráfico, como la carga media de celda o la tasa de transmisión media de los usuarios.

Capítulo 7

Conclusiones finales

En este capítulo final se resume el trabajo realizado en esta tesis, remarcando sus principales contribuciones. Esta tesis se cierra con las posibles mejoras y líneas futuras de ampliación.

7.1. Resumen del trabajo

Los constantes avances en la tecnología, que propician la aparición de nuevos terminales más potentes y la creación de nuevos servicios, han generado unas expectativas en los usuarios de la red móvil difíciles de alcanzar. Uno de los principales factores que condicionan la experiencia del usuario es la situación en la que se encuentra. Esta tesis doctoral ha explorado la influencia del contexto espaciotemporal del usuario en el uso que hace este de la red, demostrando el beneficio de incluir esta información en los procesos de autoorganización del sistema.

El estudio ha comenzado con un análisis de las redes de acceso de telefonía móvil y la diversa información que pueden proporcionar, así como las técnicas de automatización de gestión de la red. Por otro lado, se ha realizado un repaso de las técnicas de aprendizaje automático aplicadas a diferentes problemas, desde las clásicas hasta las más modernas y complejas. Una vez afianzadas las bases teóricas, se ha estudiado el estado de la técnica y la tecnología en los diferentes problemas planteados en esta tesis. El análisis de la bibliografía mostró que, en la mayoría de los casos, la información relacionada con el contexto de los usuarios no se contempla, aunque en algunos enfoques más modernos empieza a cobrar interés. Este hecho puede explicarse por la dificultad

de obtener esta información, lo que motiva la búsqueda de una forma de estimarla a partir de los datos que se disponen en las herramientas de gestión radio, ampliándola si es necesario con información de portales externos.

En esta búsqueda se ha comprobado que la situación del usuario está estrechamente relacionada con el uso que hace de la red. Por tanto, conociendo este uso es posible inferir elementos del contexto de la conexión, como, por ejemplo, si esta ha ocurrido en interiores o en zonas abiertas. Para ello, en la primera de las contribuciones de esta tesis se ha planteado un modelo de regresión logística multivariable para clasificar conexiones móviles en función de sus indicadores de rendimiento recopilados de las trazas, empleando datos de la principal actividad socioeconómica del terreno para etiquetar los elemento del conjunto de datos en el proceso de entrenamiento. El análisis de los resultados demuestra que la información empleada es suficiente para segregar las conexiones de interior con un elevado nivel de confianza.

Una vez comprobada la posibilidad de deducir el contexto espacial de los usuarios, se ha planteado su aplicación en la gestión de la red. Conocer la distribución espacial de los usuarios en la red es especialmente interesante, permitiendo, por un lado, conocer de antemano a cuántos usuarios afectarían los cambios introducidos en los procesos de replanificación de la red celular, y, por otro, estimar el alcance necesario de dichos cambios. La clasificación de conexiones en interior/exterior es extremadamente útil para este fin, permitiendo adaptar la estimación de la mejora en la generación de tráfico producida por la replanificación de la red en función del contexto de los usuarios, o discriminar localizaciones en el proceso de posicionamiento de conexiones. Otra valiosa fuente de información a la hora de localizar conexiones es la densidad de población en el escenario, ayudando a conocer dónde es más probable que ocurra una conexión. Aunque esta información es fácilmente accesible a través del catastro en barrios residenciales, es poco fiable, porque no refleja la situación real de la población, pudiendo estar a lo largo del día en sus puestos de trabajo o en centros recreativos. Como alternativa, en la segunda contribución de esta tesis se plantea estimar la densidad de población a partir de las publicaciones geolocalizadas de redes sociales, debido a su ubicuidad actual. Añadiendo ambas fuentes de datos al proceso de localización, se ha podido observar una gran mejora en la precisión y resolución respecto a métodos tradicionales basados únicamente en medidas radio, obteniendo una granularidad vital para mejorar la eficiencia de los procesos de gestión de la red.

Conocida la distribución de los usuarios, es posible perfeccionar la solución de problemas en tareas de autoplanificación y autooptimización. Por ejemplo, en la planifica-

ción de redes heterogéneas planteadas para 5G, uno de los problemas típicos a la hora de determinar la posición de nuevas estaciones base de corto alcance es si la mejora de rendimiento introducida compensa los gastos derivados del despliegue. Las redes heterogéneas están diseñadas para solucionar diferentes tipos de problemas (huecos de cobertura y falta de capacidad). Los planteamientos actuales de selección de nuevos emplazamientos solventan los problemas de propagación de forma eficiente, pero en muchos casos depende de la experiencia del operador para corregir la falta de capacidad, siendo este el principal punto débil de los métodos actuales. La tercera contribución de esta tesis ha demostrado que enfocar el emplazamiento de las nuevas picoceldas desde el punto de vista del usuario, estimando el incremento de tráfico esperado en cada posible emplazamiento, es una alternativa con gran potencial para poder identificar áreas problemáticas de la red debidas tanto a problemas de propagación como de capacidad.

Hasta ese punto, el estudio se había centrado en el contexto espacial de los usuarios. Sin embargo, dada la naturaleza dinámica de la vida urbana, el contexto temporal es también un factor determinante. En este trabajo, se ha demostrado con datos reales que las aglomeraciones de usuarios que derivan en problemas de capacidad no siempre se mantienen en el tiempo, como puede ser en zonas donde siempre suelen coexistir una gran cantidad de personas (como en grandes centros comerciales); también pueden ser esporádicos pero significativos para la capacidad de la red y, por tanto, impactar negativamente en la experiencia de los usuarios. Un claro ejemplo de estos cuellos de botella son algunos eventos sociales masivos, tratados en la última contribución de esta tesis, por lo que conocer el impacto que estos tienen en la red es crucial para poder aplicar soluciones de forma proactiva. Por la complejidad de los patrones de comportamiento de los usuarios, las arquitecturas de aprendizaje profundo son especialmente interesantes para abordar este problema, demostrando una precisión mucho mayor que los modelos clásicos basados en técnicas de aprendizaje automático demasiado simples.

Uno de los principales criterios en el desarrollo de los modelos propuestos en esta tesis es la aplicabilidad real de los mismos, por lo que han sido diseñados para que, una vez entrenados, tengan un bajo consumo de recursos, lo que facilita su integración en herramientas de gestión de red. De esta forma, la baja carga computacional de los modelos permite, a su vez, su inclusión en plataformas de *edge computing* empleadas en la tecnología 5G, donde los recursos suelen ser limitados. Además, toda la información empleada en el desarrollo proviene, o bien del sistema de gestión de red, o bien de portales de datos públicos, lo que facilita su recolección.

7.2. Publicaciones

Como resultado del trabajo, se han elaborado las siguientes publicaciones:

Artículos de revista

[I] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, R. Acedo-Hernández, S. Luna-Ramírez, "A Data-Driven Algorithm for Indoor/Outdoor Detection Based on Connection Traces in a LTE Network," IEEE Access, vol. 7, pp. 65877–65888, mayo 2019.

[II] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, A. J. García, S. Luna-Ramírez, "A Context-Aware Data-Driven Algorithm for Small Cell Site Selection in Cellular Networks," IEEE Access, vol. 8, pp. 105335–105350, junio 2020.

[III] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, "A Deep-Learning Model for Estimating the Impact of Social Events on Traffic Demand on a Cell Basis," IEEE Access, vol. 9, pp. 2169-3536, mayo 2021.

[IV] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, L.R. Jiménez, S. Luna-Ramírez, "Statistical model for mobile user positioning based on social information," MDPI Electronics, vol. 10, junio 2021.

Congresos y reuniones científicas

[V] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, R. Acedo-Hernández, S. Luna-Ramírez, "Algoritmo de detección interior/exterior para trazas de conexión de una red LTE," XXXIII Simposium Nacional de la Unión Científica Internacional de Radio URSI 2018, Granada, España, septiembre 2018.

[VI] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, A.J. García, S. Luna-Ramírez, "Algoritmo de detección de huecos de cobertura usando información de redes sociales," XXXIV Simposium Nacional de la Unión Científica Internacional de Radio, Sevilla, España, septiembre 2019.

[VII] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, "Análisis de la influencia de eventos sociales en redes celulares," XXXV Simposium Nacional de la Unión Científica Internacional de Radio, Málaga (evento en

formato remoto por COVID-19), septiembre 2020.

[VIII] J. L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, L.R. Jiménez, S. Luna-Ramírez, "Statistical model for mobile user positioning based on social information," 3rd Post-Iracon Meeting, Online Meeting, 31 mayo - 1 junio, 2021.

[IX] J.L. Bejarano-Luque, M. Toril, M. Fernández-Navarro, C. Gijón, S. Luna-Ramírez, "Comparativa de modelos de aprendizaje profundo para series temporales de alta resolución," XXXVI Simposium Nacional de la Unión Científica Internacional de Radio, Vigo (evento en formato remoto por COVID-19), septiembre 2021.

Las publicaciones generadas en esta tesis pueden englobarse en tres grandes grupos. En el primero de ellos se sitúan las publicaciones en las que se infiere el contexto de los usuarios, donde entrarían las publicaciones [I], [V], que presentan el modelo de clasificación de conexiones de interior, explicado en el *Capítulo 3*.

El segundo grupo abarca las contribuciones en las que se aplica la información de contexto para mejorar procesos en herramientas de gestión de red, como [IV], [VIII], donde se exponen los resultados del modelo de posicionamiento estadístico de usuarios con información espacial de redes sociales, descrito en el *Capítulo 4*, o [III], [VII], [IX], que aportan los resultados del modelo de predicción de tráfico corregido por eventos sociales, expuesto en el *Capítulo 6*.

Por último, el tercer grupo de publicaciones contiene aquellas que, haciendo uso de estos procesos mejorados con información de contexto, desarrollan algoritmos de autoorganización de la red, por ejemplo, [II], [VI], donde se explica la solución del algoritmo de autoplanificación de celdas pequeñas, detallado en el **Capítulo 5**.

El desarrollo de estas contribuciones se ha enmarcado en varios proyectos de investigación, enumerados a continuación:

- QoESON - QoE-aware self-planning/optimization in B4G networks (TEC2015-69982-R), financiado por el Ministerio de Economía y Competitividad.
- 5GPLAN - Self-planning of virtualized 5G networks (RTI2018-099148-B-I00), financiado por el Ministerio de Ciencia, Innovación y Universidades.
- Predicting Quality of Experience (UMA-FEDERJA-256), financiado por la Junta de Andalucía.

- LOCUS – LOCation and analytics on-demand embedded in the 5G ecosystem for Ubiquitous vertical applicationS (H2020-ICT-20-2019, 5G-PPP Phase 3 - Part 4, 2019), financiado por el marco europeo del Horizonte 2020.

Durante esta tesis, también se ha colaborado en los siguientes trabajos:

Otras aportaciones

[X] P.A. Sánchez, J.L. Bejarano-Luque, J.F. Cantón, M. Toril, and S. Luna-Ramírez, "Herramienta de planificación radio para evaluar la calidad de experiencia en redes móviles," XXXIII Simposium Nacional de la Unión Científica Internacional de Radio URSI 2018, Granada, España, septiembre 2018.

[XI] J.F. Cantón, J.L. Bejarano-Luque, P.A. Sánchez, and M. Toril, "Clasificación de tráfico encriptado mediante aprendizaje no supervisado en redes LTE," XXXIII Simposium Nacional de la Unión Científica Internacional de Radio URSI 2018, Granada, España, septiembre 2018.

[XII] P.A. Sánchez, S. Luna-Ramírez, M. Toril, C. Gijón, and J.L. Bejarano-Luque, "A data-driven scheduler performance model for QoE assessment in a LTE radio network planning tool," Computer Networks, vol. 173, pp. 107228, mayo 2020.

[XIII] C. Baena, S. Fortes, E. Baena, J. L. Bejarano-Luque, M. Tranle, and R. Barco, "GPPP y SDR como una potente herramienta científica," XXXV Simposium Nacional de la Unión Científica Internacional de Radio, Málaga (evento en formato remoto por COVID-19), septiembre 2020.

[XIV] C. Gijón, M. Toril, S. Luna-Ramírez, J.L. Bejarano-Luque, and M.L. Marí-Altozano, "Estimación de capacidad en redes LTE mediante aprendizaje supervisado," XXXV Simposium Nacional de la Unión Científica Internacional de Radio, Málaga (evento en formato remoto por COVID-19), septiembre 2020.

[XV] J. Sánchez, M. Toril, V. Wille, M. Fernández-Navarro and J. L. Bejarano-Luque, "Algoritmo de filtrado heurístico para medidas de geolocalización en trazas MDT," XXXV Simposium Nacional de la Unión Científica Internacional de Radio, Málaga (evento en formato remoto por COVID-19), septiembre 2020.

[XVI] C. Gijón, M. Toril, S. Luna-Ramírez, J. L. Bejarano-Luque and M. L. Marí-Altozano, "Estimating Pole Capacity From Radio Network Performance Statistics

by Supervised Learning," IEEE Transactions on Network and Service Management, vol. 17, no. 4, pp. 2090-2101, diciembre 2020

[XVII] L. Jiménez, M. Solera, M. Toril, S. Luna-Ramírez and J. L. Bejarano-Luque, "Impact of the Uplink on the performance of YouTube 360º Live Video Streaming Service in LTE," IEEE Access, vol. 9, pp. 2169-3536, septiembre 2021.

7.3. Líneas futuras

Durante el desarrollo de las contribuciones contenidas en esta tesis, han ido surgiendo posibles líneas futuras para mejorar y ampliar los resultados obtenidos. A continuación se incluye una breve descripción de algunas de ellas.

Un primer enfoque para extender la información relacionada con el contexto es recurrir a modelos más sofisticados o fuentes de información que no han sido explotadas en esta tesis:

- **Exploración de algoritmos/modelos más complejos**

  En la última contribución de esta tesis, se ha podido comprobar cómo, con suficiente tamaño de muestra, los modelos de aprendizaje profundo superan ampliamente los resultados obtenidos con modelos clásicos. Por lo tanto, una interesante línea de investigación puede ser aplicar estos modelos, basados en redes neuronales, en la estimación del contexto de los usuarios, como en la clasificación de conexiones como interiores o exteriores. Sustituyendo el modelo de regresión logística por una red neuronal, se espera que la mejora obtenida en este aspecto no solo beneficiaría al proceso de identificación de conexiones de interior, sino que también impactaría positivamente en todas las herramientas que hacen uso de esta información, como los modelos propuestos en los *Capítulos 4* y *5*. El inconveniente de las redes neuronales es que, al ser modelos más complejos, necesitan una mayor cantidad de datos para ser entrenados, que no siempre es posible de obtener. En caso de no disponer de juegos de datos suficientemente extensos para entrenar modelos de redes neuronales, es posible aplicar modelos clásicos más robustos frente a la falta de información, como los árboles aleatorios o *Random Forest*.

  Además, algunas técnicas de preprocesado de datos, como la gestión de valores anómalos o la búsqueda de patrones naturales en los datos mediante agrupamiento no supervisado, han demostrado mejorar el aprendizaje de cualquier tipo de

modelo, por lo que sería interesante estudiar la inclusión de estas técnicas en el modelo de estimación de contexto. Aprovechando estas técnicas de preprocesado, es posible que, aumentando la resolución espacial en el cálculo de los ratios de conexiones de interior (modelo basado en anillos en lugar de celdas), se puedan detectar perfiles de usuario más complejos, mejorando los resultados de clasificación. Otra línea de ampliación del modelo de clasificación podría ser la inclusión de la información de configuración de la celda servidora (p.ej., ancho de banda o frecuencia de portadora), permitiendo contextualizar indicadores relativos al servicio del usuario, esperando mejorar los resultados obtenidos.

Siguiendo la misma filosofía que en el caso anterior, la contribución presentada en el *Capítulo 5* se basa, no tanto en un modelo de aprendizaje automático, sino en un algoritmo que puede optimizarse. El algoritmo de detección de emplazamientos propuesto, una vez calculado el mapa de ganancia potencial de tráfico en el escenario, es de tipo voraz. Este tipo de algoritmos, en ocasiones, se estanca en un óptimo local, siendo incapaz de detectar la mejor solución global. Otra línea de ampliación del trabajo es el estudio de otros tipos de algoritmos, como algoritmos genéticos o recocido simulado, para comprobar la idoneidad de los emplazamientos seleccionados.

- **Enriquecimiento de la información contextual**

  A lo largo del trabajo se ha podido comprobar que la información del entorno del usuario es un componente clave en la gestión de la red móvil. Este entorno cambia de forma dinámica, por lo que añadir la componente temporal a la distribución estadística de los usuarios en la red podría mejorar los procesos de posicionamiento de las conexiones. De esta forma se podría aprovechar la movilidad de los usuarios en su posicionamiento, siendo capaces de detectar aglomeraciones periódicas de usuarios (p.ej., oficinas en las mañanas de días laborables, zonas de ocio nocturno en fines de semana...) para refinar los resultados obtenidos por el modelo del *Capítulo 4*.

  Al mismo tiempo, otro de los aspectos clave que afectan al comportamiento de los usuarios es el tipo de servicio que se demanda. Cada servicio tiene unos requerimientos completamente distintos, más aún en los nuevos casos de uso de las redes 5G. Por tanto, conocer el principal servicio solicitado en la celda es vital para mejorar la eficacia de los procesos de optimización de red. En esta línea, determinar qué servicio domina en cada tipo de uso de suelo puede servir para aprovechar al máximo la asignación de recursos, adaptando la solución automática de proble-

mas en función del servicio principal. Por ejemplo, servicios de videostreaming en conciertos/partidos generarán un pico de tráfico que puede derivar en problemas de capacidad, necesitando una mayor cantidad de recursos radio disponibles, mientras que servicios de poco tráfico, típicamente a ráfagas (como conversaciones, redes sociales o búsqueda web) en entornos de alta concentración de usuarios (p.ej., centros comerciales o culturales) pueden causar problemas de congestión por una gran cantidad de usuarios simultáneos, necesitando una asignación más eficiente de los recursos disponibles.

Otra forma de ampliar el trabajo es explorar aplicaciones de la información contextual espaciotemporal de los usuarios en el mantenimiento de la red. Algunas de las posibles líneas de investigación en este sentido son:

- **Desarrollo de modelos conscientes del contexto para casos de uso en 5G**

  Una de las principales áreas de investigación es desarrollar algoritmos que, aprovechando la información contextual de los usuarios (dónde, cuándo y cómo están usando la red), saquen el máximo partido del particionado de red (*network slicing*). En este ámbito, se plantea la predicción de tráfico por arrendatario (*tenant*) corregida por la ocurrencia de eventos sociales para la mejora de diferentes situaciones, como la asignación de recursos a los diferentes servicios o el diseño de redes heterogéneas para solventar problemas de capacidad en diferentes casos de uso dentro de la tecnología 5G (como en aplicaciones de comunicaciones masivas tipo máquina o de banda ampliada, donde la densidad de dispositivos o la gran cantidad de tráfico generado pueden saturar los recursos de las celdas servidoras).

- **Estudio de distintas actividades sociales masivas**

  Otra de las futuras ampliaciones del trabajo realizado es estudiar nuevas situaciones en las que el contexto del usuario sea un factor determinante en el correcto funcionamiento de la red. Un ejemplo podría ser el análisis de las horas de máxima afluencia en estaciones de transporte de elevado tránsito, por ejemplo, los principales aeropuertos o estaciones de trenes.

Apéndice A

Herramientas de recolección y tratamiento de datos

En los diferentes modelos explicados en esta tesis doctoral, la principal aportación es la aplicación de información relativa al contexto espaciotemporal de los usuarios en los procesos de gestión de red. Por este motivo, para desarrollar estos modelos es necesario realizar un trabajo previo para la obtención y el tratamiento de una ingente cantidad de datos de distinta naturaleza, tanto relativos a la red celular como datos externos. En este apéndice, se explican las herramientas desarrolladas para recolectar y procesar los datos necesarios. En concreto, en la Sección A.1 se detallan las herramientas de recolección de datos ajenos a la red, aplicando técnicas de raspado web o accediendo mediante interfaces de programación de aplicaciones (*Application Programming Interfaces*, APIs). Una vez obtenidos los datos, es necesario procesarlos para poder utilizarlos en los distintos modelos. En la Sección A.2 se describen las herramientas desarrolladas para el procesado de la información, tanto de la red como de las herramientas de recolección o de los portales de datos abiertos.

A.1. Herramientas de recolección

La información relativa a la posición y al momento en el que los usuarios hacen uso de la red celular es vital para los modelos descritos en esta tesis, pero por desgracia los sistemas de gestión de red no son capaces de recopilar dicha información. Por este motivo, es crucial obtenerla de otras fuentes, siendo imprescindible que sean públicamente accesibles para que sea comercialmente viable implementar los modelos desarrollados

en los sistemas de gestión de red. En algunos casos, la información necesaria es accesible a través de portales de datos abiertos, pero no siempre es así, por lo que es necesario un trabajo de recolección previo. En esta sección se describen las diferentes herramientas desarrolladas para la recolección de datos de redes sociales (aplicados en el Capítulo 4) y plataformas de venta de entradas (usados en el Capítulo 6), a través de sus respectivas APIs, y de historial de eventos sociales (Capítulo 6), mediante raspado web.

A.1.1. Recolección a través de APIs

En la actualidad, la ingente cantidad de información que se genera continuamente es de vital importancia en múltiples ámbitos y las empresas son conscientes de ello. Por este motivo, muchas de estas empresas monetizan la información que se origina en sus plataformas facilitando el acceso a través de APIs. Como norma general, las APIs pueden ser privadas, cuya aplicación suele ser la interconexión entre equipos internos de una empresa, o públicas, a las que cualquier persona puede acceder, necesitando en algunas ocasiones un registro previo. Las empresas que ofrecen datos generados en sus plataformas optan por la implementación de APIs públicas, lo cual les permite limitar el flujo de información accesible por cada desarrollador en función del plan de pago que contraten, siendo gratuito el plan más básico. Estas APIs son interfaces que habilitan la comunicación entre los desarrolladores (clientes) y la base de datos de las empresas (servidores) mediante un conjunto de solicitudes y respuestas definidas.

Existen varios tipos de APIs en función de su objetivo y arquitectura, de los cuales el más común actualmente es el tipo API de transferencia de estado representacional (*Representational State Transfer*, REST) [165]. Las APIs de REST basan su funcionamiento en el intercambio de mensajes sobre el protocolo de transferencia de hipertexto (*HyperText Transfer Protocol*, HTTP), que facilitan el envío de solicitudes y respuestas incluyendo datos en múltiples formatos de texto, como XML o JSON. La popularidad de las APIs de REST viene dada por su fácil uso e integración en los sistemas software de los clientes, debido principalmente a sus características:

- Se definen un conjunto de operaciones, comunes para todas las APIs de REST, que facilitan el acceso y la interacción con todos los recursos que ofrece el servidor. Las operaciones más comunes son:
    - Peticiones POST. Este tipo de peticiones son empleadas para añadir un nue-

vo objeto a un recurso específico del equipo origen (normalmente el servidor) al equipo destino (cliente).

  - Peticiones GET. Estas peticiones se realizan desde el equipo de origen (usualmente el cliente) para solicitar información sobre un recurso concreto, contenida en el equipo destino (servidor).
  - Peticiones PUT. Este tipo de peticiones son similares a las POST, con la salvedad de que no añade un nuevo objeto, sino que sustituye el valor de un objeto ya existente en el recurso del equipo destino.
  - Peticiones DELETE. Las peticiones DELETE borran del equipo destino la información especificada en la petición.

- El protocolo de comunicación entre los clientes y el servidor no necesita guardar el estado de la conexión, es decir, cada mensaje que intercambian cliente/servidor contiene toda la información necesaria para completar la petición.

- Debido al uso del protocolo HTTP, es posible navegar entre los recursos que ofrece el servidor únicamente conociendo su identificador (*Uniform Resource Identifier*, URI), siguiendo enlaces que direccionen a estos recursos y sin la necesidad de infraestructuras adicionales.

Debido al fácil acceso por parte de los usuarios, la seguridad en las APIs de REST es crucial. Para conseguir esto, la API solo acepta peticiones recibidas de un usuario que previamente ha sido registrado y que se ha autenticado correctamente. Este registro permite, además, definir perfiles de acceso, limitando los recursos disponibles en función de la cuota de suscripción. Existen dos técnicas de autenticación en las APIs de REST:

- Autenticación mediante códigos de seguridad. Los códigos de seguridad, o *tokens*, son códigos de acceso único generados en el momento de autenticación del usuario con sus credenciales. Este código es temporal y habilita al cliente realizar peticiones a los recursos del servidor para los que tiene derechos de acceso, sin tener que volver a identificarse mientras dure la vida útil del token.

- Autenticación mediante claves de API. Estas claves son códigos generados en el momento de registro del usuario e identifican un recurso concreto al que tiene derecho de acceso. Este tipo de autenticación es menos segura que la autenticación mediante tokens, pero permite monitorizar el uso de la API y restringir el acceso en función del número de peticiones.

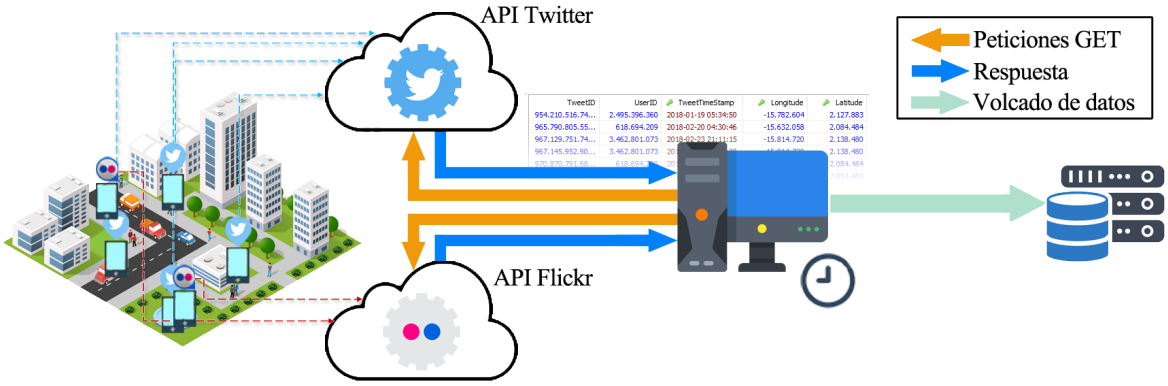


Figura A.1: Esquema de la herramienta de colección de datos de redes sociales en tiempo real.

Todas las APIs empleadas en las herramientas desarrolladas son públicas y de tipo REST con suscripción gratuita, por lo que la autenticación se realiza mediante claves de API.

APIs de acceso a datos de redes sociales en tiempo real

La herramienta basada en APIs de acceso en tiempo real es la empleada en la recolección de publicaciones geolocalizadas de redes sociales, usadas en el modelo descrito en el Capítulo 4. Esta herramienta consta de 2 procesos, el primero usa la Streaming API de Twitter, donde el cliente se implementa en Java a través de la librería Twitter4j [82, 88], mientras que el segundo proceso se basa en la API de Flickr, implementando el cliente en Python [83]. El funcionamiento de la herramienta está representado en la Figura A.1, donde ambos procesos corren en paralelo de forma ininterrumpida.

El primer paso es configurar ambos procesos, para ello es necesario introducir la siguiente información:

- Las claves de la API a la que va a acceder el proceso.
- El área geográfica donde se va a recopilar la información geolocalizada de Twitter y Flickr, respectivamente, delimitada como un área rectangular que se define por las coordenadas de la esquina superior derecha y la esquina inferior izquierda.
- El tiempo del contador entre peticiones. Este contador duerme el proceso después de cada petición durante el tiempo establecido, para no sobrecargar de llamadas la API y no sobrepasar el límite de la suscripción gratuita. Por defecto, este contador se ha configurado en 3.600 segundos (1 hora).

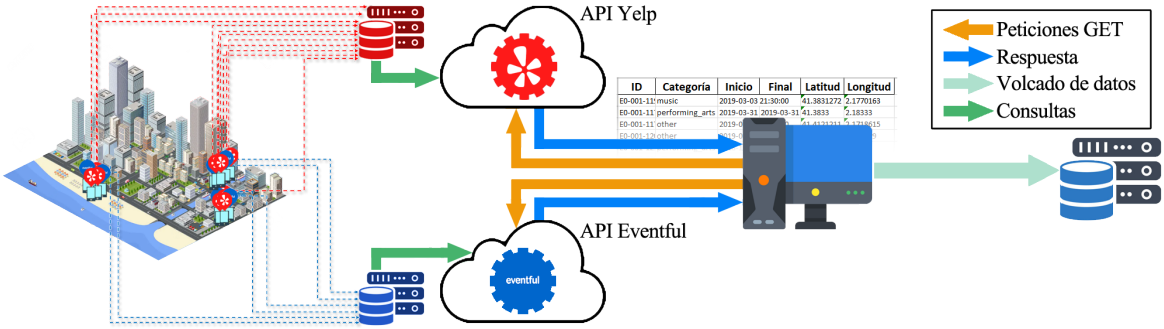


Figura A.2: Esquema de la herramienta de colección de datos históricos de eventos.

Una vez configurados, se lanzan ambos procesos para que empiece la colección de publicaciones geolocalizadas de redes sociales. Primero, cada proceso realiza una primera petición GET a su respectiva API, introduciendo las claves en la cabecera y como filtros de búsqueda las coordenadas geográficas y el rango de tiempo de búsqueda, definido como la marca de tiempo inicial (instante actual menos la duración del contador) y la marca de tiempo final (instante actual). Una vez obtenidas las respuestas de cada API, se vuelca el resultado en su respectiva base de datos y se duerme el proceso.

Los parámetros que se guardan por publicación de redes sociales son: a) identificador único de la publicación de redes sociales, b) coordenadas GPS de la publicación y c) la marca de tiempo del instante que se generó. Cuando finalice el contador, se repite el procedimiento completo de forma continuada hasta que se cierre la herramienta.

APIs de acceso a datos históricos de eventos

La herramienta basada en APIs de acceso a datos históricos es una de las tres fuentes de información de eventos sociales masivos, usada en el modelo descrito en el Capítulo 6. De la misma forma que la herramienta de APIs en tiempo real, esta herramienta consta de 2 procesos, el primero usa la API de Yelp [156], mientras que el segundo proceso se basa en la API de Eventful [155]. En ambos casos el cliente se implementa en Python. La Figura A.2 presenta un esquema de la herramienta de APIs de datos históricos. Al contrario que la herramienta de APIs en tiempo real, en este caso solo es necesario ejecutar la herramienta una vez, sin quedar los procesos corriendo en segundo plano.

Primero, se configuran los procesos de la herramienta, introduciendo los siguientes datos:

- Las claves de la API a la que va a acceder el proceso.

- Las áreas geográficas donde se va a buscar la información de eventos sociales, tanto en Yelp como en Eventful. En este caso, el área geográfica se delimita por las coordenadas centrales de cada zona a recopilar y un radio en torno a dichas coordenadas, definiendo un área circular en la que se van a buscar los datos históricos de eventos sociales. Las áreas seleccionadas para obtener la información de eventos sociales corresponden a las cubiertas por un círculo de 30 kilómetros de radio en torno al centro geográfico de 10 ciudades densamente pobladas.

- El rango de tiempo de búsqueda de eventos sociales, definido por la marca de tiempo inicial y final del rango. Para el desarrollo del modelo, estos parámetros se configuraron para recoger los eventos sociales de los meses de marzo y abril de 2019.

- La cantidad máxima de resultados que se obtienen por búsqueda. En estas APIs, a diferencia de las APIs de la herramienta anterior, se acceden a datos históricos, por lo que las empresas limitan la cantidad de eventos que se adjuntan para controlar el tamaño de la respuesta. En el caso de esta herramienta, interesa configurar este parámetro al máximo permitido para realizar el mínimo de peticiones posibles, siendo 1.000 en Yelp y 250 en Eventful.

- El desfase inicial en la búsqueda. Es posible que la cantidad de eventos disponibles en la base de datos sea mayor que el número máximo de resultados añadidos en una petición. En estos casos, para acceder a los eventos que no se añadirían en la búsqueda, es necesario incluir en la petición un desfase, u *offset*, que descarta los primeros eventos de la base de datos en la petición. Por defecto, el desfase inicial de la herramienta está configurado en 0, incrementándolo en cada iteración con la cantidad de resultados obtenidos en la iteración anterior.

Al ejecutar la herramienta de colección de datos históricos, se itera en las diferentes áreas geográficas configuradas. Por cada una de estas áreas, cada proceso realiza una primera petición GET a su respectiva API de forma similar a la herramienta de datos en tiempo real, incluyendo en la cabecera la petición las claves de la API y como filtros de búsqueda el área geográfica, el rango de tiempo y el límite máximo de resultados. Una vez realizada la consulta, la API devuelve los primeros eventos sociales registrados en su base de datos dentro del rango seleccionado, además del número total de eventos disponibles. Si este número es mayor que el límite de resultados, se incrementa el *offset* con el número máximo de resultados por búsqueda y se vuelve a realizar la petición GET en la misma área geográfica, repitiendo el procedimiento hasta obtener todos

los resultados. Cuando se recopilan todos los eventos sociales en el área geográfica, se vuelcan a la base de datos de eventos sociales y se empieza el procedimiento para la siguiente área.

Por cada evento recopilado, la información que se almacena en la base de datos es: a) un identificador único por evento, b) las coordenadas geográficas de la ubicación del evento, c) las marcas de tiempo de inicio y final del evento, d) el nombre oficial del evento y e) una etiqueta que determina la categoría del evento (p.ej., concierto, taller, partido, etc.).

A.1.2. Recolección mediante técnicas de raspado web

Por desgracia, no toda la información generada en Internet es accesible a través de APIs, por lo que para aprovechar por completo el potencial de esta información es necesario buscar otras formas de recolección. En este contexto aparece el raspado web [166], o *web scraping*, englobando un conjunto de técnicas cuyo objetivo es la extracción y almacenamiento de información contenida en los sitios web, ya sea de forma manual o empleando programas, denominados *web scrawlers*, que automaticen el proceso. Para poder manejar las distintas formas que adoptan los datos contenidos en los sitios web, estos programas pueden ser desde procedimientos personalizados para el sitio concreto, ayudados por humanos, hasta grandes sistemas capaces de estructurar por completo cualquier sitio web.

Para extraer la información deseada de un sitio web, los *web scrawlers* suelen analizar su código fuente en lenguaje de marcado de hipertexto (*Hypertext Markup Language*, HTML). Este código fuente puede ser obtenido de dos formas: a) utilizando el protocolo HTTP, realizando una petición GET con la dirección URL (*Uniform Resource Locator*) y obteniendo el código HTML en su respuesta o b) simulando la navegación humana a través de un buscador, empleando herramientas como Selenium [167], un ecosistema que habilita el manejo de navegadores web desde aplicaciones programadas en diferentes lenguajes de programación.

Una vez obtenido el código fuente, los *web scrawlers* estructuran la información para su posterior almacenado en el formato deseado. Para ello, analizan el código en búsqueda de diferentes patrones de texto que aparezcan en los elementos que componen el sitio web (como encabezados, distintos complementos dentro del cuerpo de la página, etc.) o en sus metadatos, pudiendo interactuar con elementos responsivos y extrayendo la información deseada.

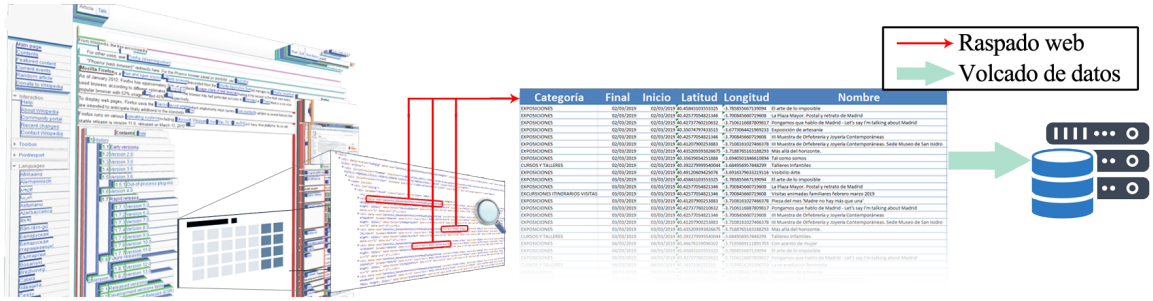


Figura A.3: Esquema de la herramienta de colección mediante raspado web.

Usando el raspado web, se desarrolla la última herramienta para la recolección de información ajena a la red móvil, destinada a la recolección de eventos sociales masivos. Esta herramienta es la segunda fuente de información de eventos sociales del modelo descrito en el Capítulo 6. De la misma forma que en las herramientas anteriores, la Figura A.3 muestra un esquema de la herramienta de recolección mediante raspado web.

La herramienta de raspado web consta de un conjunto de procesos desarrollados *ad hoc* para cada sitio web que ha sido seleccionado para recopilar información de eventos sociales. El primer paso de desarrollo es una inspección manual del sitio, detectando la página o elemento que contiene la información de eventos sociales pasados, así como los elementos que permiten, si es necesario, navegar en el sitio web para recopilar toda la información disponible. Además, se localizan las palabras clave que identifican cada elemento en el código HTML. Una vez realizada la inspección, se accede a cada sitio web a través del proceso correspondiente. Debido a la presencia de algunos componentes JavaScript que solo generan la información necesaria cuando se accede desde un navegador, se hace uso de las rutinas de la librería Selenium que manejan el navegador Mozilla Firefox para acceder a cada sitio web y obtener e interactuar con el código fuente. En este punto, se hace uso de las palabras clave para filtrar los elementos necesarios y obtener la información de cada evento social almacenado en el sitio.

De la misma forma que en la herramienta anterior, la información que se guarda en la base de datos por cada evento recopilado es: a) el identificador de evento, b) su dirección postal, de la que se extraen las coordenadas GPS mediante geocodificación, c) las marcas de tiempo de inicio y final, d) el nombre del evento y e) una etiqueta que determina la categoría del evento.

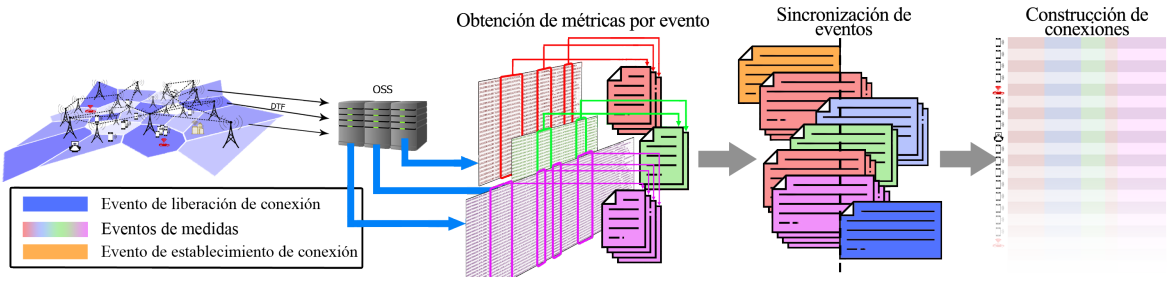


Figura A.4: Esquema de la herramienta construcción de conexiones.

A.2. Herramientas de procesado

La información necesaria para el desarrollo de los modelos descritos en esta tesis, tanto las métricas y configuración de la red celular como la información contextual, descargada de portales de datos abiertos o recogida con las herramientas de la sección anterior, muchas veces necesita un procesado que permita adaptarla a un formato común. Para ello, se han desarrollado una serie de herramientas que construyen las conexiones que han ocurrido en la red móvil a partir de sus trazas, así como la caracterización contextual del escenario que las acoge. En esta sección se describe primero la herramienta de construcción de conexiones, pilar de los modelos descritos en los Capítulos 3, 4 y 5. A continuación, se explica la herramienta de construcción de mapas de uso de suelo, que define el contexto de interior en cada área geográfica servida por la red en los mismos modelos.

A.2.1. Construcción de conexiones

Como ya se comentó en el Capítulo 2, las trazas de conexión son la fuente de información más rica que puede obtenerse de una red móvil debido al gran número de eventos registrados de distinta naturaleza. Estos eventos contienen información a nivel de usuario y pueden ser internos, que incluyen información de rendimiento, o externos, con mensajes estandarizados entre estaciones base y otros elementos de la red. A su vez, los eventos pueden ser periódicos, con medidas equiespaciadas en tiempo, o asíncronos, disparados por una acción en la red. Con esta información es posible conocer el comportamiento de las conexiones registradas, pero es necesario realizar un tratamiento de los archivos de trazas. Para ello, la herramienta de construcción de conexiones, presentada en la Fig. A.4, se basa en técnicas de procesado de eventos complejos (*Complex Event Processing*, CEP) [20], definiendo cada conexión como el conjunto de indicadores rescatados de cada tipo de evento y asociados al mismo indicador único.

Para construir las conexiones ocurridas en el periodo de recolección de trazas, la herramienta desarrollada se compone de tres procesos:

1. **Obtención de métricas por evento.** El primer paso para la construcción de conexiones es la obtención de las medidas que la componen. Para ello, se analizan los ficheros de cada tipo de evento por separado. Estos ficheros almacenan la información recopilada de la red en formato de valores separados por comas (*Comma-Separated Values*, CSV), donde, independientemente de la conexión que genere el mensaje, cada métrica ocupa siempre la misma posición. Por tanto, una vez estudiados los indicadores que se desean obtener de cada conexión, se identifican las posiciones de sus respectivas métricas y se extraen sus valores.

2. **Sincronización de eventos.** Aunque sean generados por una misma conexión, los mensajes de cada tipo de evento están completamente desacoplados entre sí, por lo que es necesario sincronizar los diferentes eventos para construir cada conexión. Una vez obtenidos las medidas necesarias por mensaje, cada mensaje se agrupa por el identificador único de la conexión que lo ha generado. Este identificador de conexión se define como la combinación de dos indicadores: a) el de la celda que le da servicio y b) un identificador de usuario dentro de la celda, obtenido de un registro limitado, que identifica la conexión en la comunicación entre la función de control de admisión radio (*Radio Admission Control*, RAC) y la banda base.

   Ese último identificador queda reservado por un periodo de tiempo para el usuario al que ha sido asignado, pero una vez liberado puede ser asignado a una conexión distinta, por lo que es necesario sincronizar también los mensajes en tiempo. Para ello, se hacen uso de los eventos de establecimiento y liberación de conexión radio para identificar el inicio y fin de las conexiones, agrupando los mensajes del resto eventos con el mismo identificador de usuario y cuya marca de tiempo se encuentre entre las marcas de tiempo de los mensajes de establecimiento y liberación de conexión correspondientes.

3. **Construcción de conexiones.** Una vez sincronizados los mensajes de cada conexión en todos los eventos recopilados, se construyen los indicadores que componen cada conexión. Estos indicadores pueden ser simples o compuestos. Los indicadores simples son aquellos que se obtienen directamente de las medidas obtenidas de los eventos o que solo requieren un leve procesado, como una agregación o promediado de valores de eventos asíncronos o la obtención la moda o

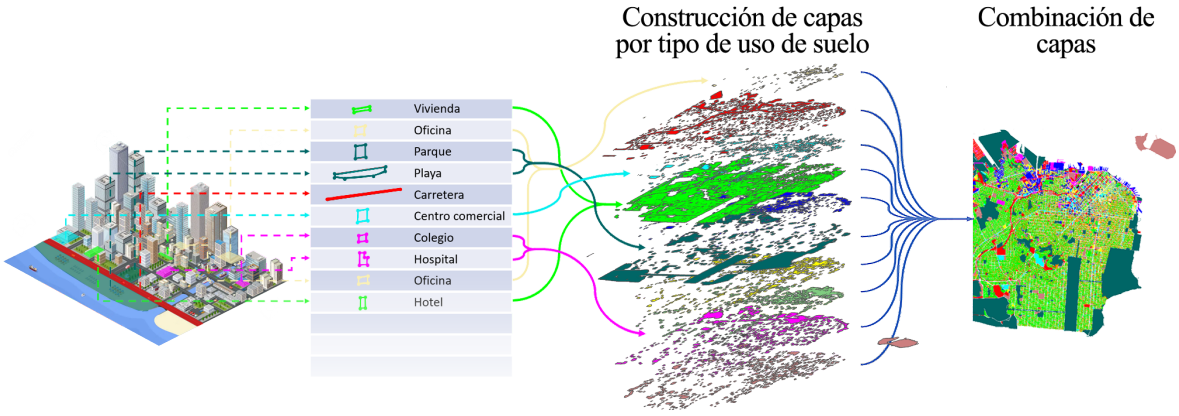


Figura A.5: Esquema de la herramienta de construcción de mapas de uso de suelo.

mediana de valores en eventos periódicos. Los indicadores compuestos son aquellos que necesitan un procesado más complejo y están basados en indicadores simples. La concatenación de todos los indicadores de conexión, tanto simples como compuestos, forman el juego de datos de conexiones basado en trazas.

A.2.2. Construcción de mapas de uso de suelo

El segundo pilar de los modelos desarrollados en esta tesis es el contexto espacial de los usuarios en la red. En concreto, el contexto de interior cobra especial relevancia en los modelos de los Capítulos 3, 4 y 5. Esta información no es propia de la red móvil, sino que es una propiedad del escenario físico que la alberga y, por tanto, es necesario recurrir a fuentes de información externas a la red para conocerla. En esta tesis, el contexto de interior se define a partir de la principal actividad socio-económica en cada zona geográfica, definida a lo largo de la tesis como uso de suelo. Debido al creciente interés en la ciencia de los datos, varias plataformas de datos abiertos (como OpenStreetMap [62]), así como algunas instituciones gubernamentales, han generado grandes conjuntos de datos para importantes núcleos urbanos y lugares de interés, formados por líneas y polígonos geolocalizados, que son enriquecidos con información adicional, entre la cual se añade una etiqueta que define la característica física específica del polígono (p. ej., hotel, vivienda, hospital, parque, etc.). Con estos juegos de datos como entrada, la herramienta descrita en esta subsección, esquematizada en la Fig. A.5, construye un mapa de usos de suelo que define la actividad socio-económica en cada área de la red móvil.

Para construir el mapa de usos de suelo, la herramienta desarrollada está formada por dos procesos:

1. **Construcción de capas por tipo de uso de suelo.** El proceso de construcción del mapa de usos de suelo comienza por la generación de un mapa por cada tipo de uso de suelo definido. Para ello, el primer paso es preprocesar la información del conjunto de datos para poder generar estos mapas. En este sentido, primero se clasifica el gran abanico de etiquetas que aparecen en el conjunto de datos, agrupando cada una de ellas en unas pocas clases que las resuma y que posteriormente serán los tipos de uso de suelo en el escenario. A continuación, se define el área que limita el escenario, se obtienen las coordenadas geográficas de sus esquinas opuestas y, aplicando una transformación basada en la ortodrómica [168], las coordenadas geográficas de cada elemento del conjunto de datos se transforman a coordenadas cartesianas, proyectando todos los polígonos en el mismo plano.

   Una vez se han preprocesado los datos, se construyen los mapas por cada tipo de uso de suelo, que formarán las diferentes capas del resultado final. Cada uno de estos mapas se define como una malla que representa el área que limita el escenario, compuesta por cuadrículas que definen secciones de misma superficie. Para construir estas capas, primero se inicializan las mallas de todos los tipos de uso de suelo con valores nulos. A continuación, se itera en los elementos que componen el conjunto de datos y se localizan las cuadrículas que contienen las coordenadas cartesianas de los vértices de cada polígono dentro de la malla correspondiente. Una vez posicionado en su respectivo mapa, se asigna el mismo valor a todas las cuadrículas de la malla del tipo de uso de suelo que entran dentro del área definida por los vértices del polígono.

2. **Combinación de capas.** La generación del juego de datos se realiza de forma manual por seres humanos y, por tanto, existe un riesgo de error, así como imprecisiones en los polígonos almacenados. Estas imprecisiones pueden derivar en solapamientos de las capas generadas en el proceso anterior, haciendo así que la misma zona sea clasificada con dos usos de suelo simultáneamente. Por este motivo, el mapa de usos de suelo se genera combinando las capas de cada tipo en un orden específico, definido por la calidad y la precisión de la información de cada capa. Así, capas con información poco relevante o con polígonos simplificados son usadas como base del mapa de usos de suelo, agregando por orden de prioridad el resto de las capas y, en caso de solape, sustituyendo los valores anteriores por los de la nueva capa. Una vez terminada la combinación, el resultado final será una malla formada por cuadrículas que representan una sección del terreno y cuyos valores determinan el uso de suelo en cada área del escenario.

Bibliografía

[1] Ericsson, "Ericsson Mobility Report June 2023," pp. 3–13, 2023, [En línea]. [Último acceso: Jun. 21, 2023].

[2] 5G-PPP, "5G empowering vertical industries," *5G-PPP White Papers*, no. 9, 2016.

[3] J. Ramiro y K. Hamied, *Self-organizing networks: self-planning, self-optimization and self-healing for GSM, UMTS and LTE.* John Wiley & Sons, 2011.

[4] NGMN Alliance, "NGMN recommendation on SON and O&M requirements," *Next Generation Mobile Networks, White paper, December*, 2008.

[5] S. Malisuwan, D. Milindavanij, y W. Kaewphanuekrungsi, "Quality of Service (QoS) and Quality of Experience (QoE) of the 4G LTE perspective," *International Journal of Future Computer and Communication*, vol. 5, no. 3, p. 158, 2016.

[6] A. Banerjee, "Revolutionizing CEM with subscriber-centric network operations and QoE strategy," *White paper, Heavy Reading*, 2014.

[7] S. Baraković y L. Skorin-Kapov, "Survey and challenges of QoE management issues in wireless networks," *Journal of Computer Networks and Communications*, vol. 2013, 2013.

[8] J. Reades, F. Calabrese, A. Sevtsuk, y C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing*, vol. 6, no. 3, 2007.

[9] M. Keally, G. Zhou, G. Xing, J. Wu, y A. Pyles, "Pbn: towards practical activity recognition using smartphone-based body sensor networks," en *9th ACM Conference on Embedded Networked Sensor Systems (SenSys 2011)*, pp. 246–259, 2011.

[10] N. Baldo, L. Giupponi, y J. Mangues, "Big Data Empowered Self Organized Networks," en *20th European Wireless Conference (European Wireless 2014)*, May. 2014.

[11] 5G-PPP, "5G architecture," *5G-PPP White Papers*, no. 1, 2016.

[12] W. A. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, y B. Sebire, "Minimization of drive tests solution in 3GPP," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 28–36, 2012.

[13] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, y P. Mähönen, "Automated coverage hole detection for cellular networks using radio environment maps," en *11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt 2013)*, pp. 35–40, 2013.

[14] V. Wille, M. Toril, y R. Barco, "Impact of antenna downtilting on network performance in geran systems," *IEEE Communications Letters*, vol. 9, no. 7, pp. 598–600, 2005.

[15] A. J. Garcia, M. Toril, P. Oliver, S. Luna-Ramirez, y R. Garcia, "Big data analytics for automated QoE management in mobile networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 91–97, 2019.

[16] F. Malandrino, C.-F. Chiasserini, y S. Kirkpatrick, "Cellular network traces towards 5g: Usage, analysis and generation," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 529–542, 2017.

[17] B. Yang, W. Guo, B. Chen, G. Yang, y J. Zhang, "Estimating mobile traffic demand using twitter," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 380–383, 2016.

[18] W. Guo y J. Zhang, "Uncovering wireless blackspots using twitter data," *Electronics Letters*, vol. 53, no. 12, pp. 814–816, 2017.

[19] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," Technical Report (TR) 125.913, 2013, version 7.3.0, Release 7.

[20] I. de la Bandera, M. Toril, S. Luna-Ramírez, V. Buenestado, y J. M. Ruiz-Avilés, "Complex Event Processing for Self-Optimizing Cellular Networks," *Sensors*, vol. 20, no. 7, p. 1937, 2020.

[21] S. Frattasi y F. Della Rosa, *Mobile positioning and tracking: from conventional to cooperative techniques*. John Wiley & Sons, 2017.

[22] F. S. T. Van Diggelen, *A-GPS: Assisted GPS, GNSS, and SBAS*. Artech House, 2009.

[23] S. Dwivedi, R. Shreevastav, F. Munier, J. Nygren, I. Siomina, Y. Lyazidi, D. Shrestha, G. Lindmark, P. Ernström, E. Stare *et al.*, "Positioning in 5g networks," *arXiv preprint arXiv:2102.03361*, 2021.

[24] A. Imran, A. Zoha, y A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[25] P. Cunningham, M. Cord, y S. J. Delany, "Supervised learning," en *Machine learning techniques for multimedia*. Springer, 2008, pp. 21–49.

[26] P. Cunningham, M. Cord, y S. J. Delany, "Unsupervised learning and clustering," en *Machine learning techniques for multimedia*. Springer, 2008, pp. 51–90.

[27] R. S. Sutton y A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[28] E. S. Gardner Jr, "Exponential smoothing: The state of the art—part ii," *International journal of forecasting*, vol. 22, no. 4, pp. 637–666, 2006.

[29] P. J. Brockwell y R. A. Davis, *Time series: theory and methods*. Springer science & business media, 2009.

[30] Y. Chauvin y D. E. Rumelhart, *Backpropagation: theory, architectures, and applications*. Psychology press, 1995.

[31] I. H. Witten y E. Frank, "Data mining: practical machine learning tools and techniques with java implementations," *Acm Sigmod Record*, vol. 31.1, no. 1, pp. 76–77, 2002.

[32] R. Yamashita, M. Nishio, R. K. G. Do, y K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[33] S. Albawi, T. A. Mohammed, y S. Al-Zawi, "Understanding of a convolutional neural network," en *International Conference on Engineering and Technology (ICET 2017)*, pp. 1–6, 2017.

[34] P. Baldi y P. J. Sadowski, "Understanding dropout," en *Advances in Neural Information Processing Systems (NIPS 2013)*, pp. 2814–2822, 2013.

[35] X. Zhang, T. Lin, J. Xu, X. Luo, y Y. Ying, "Deepspectra: An end-to-end deep learning approach for quantitative spectral analysis," *Analytica chimica acta*, vol. 1058, pp. 48–57, 2019.

[36] I. Goodfellow, Y. Bengio, y A. Courville, *Deep Learning*. MIT Press, 2016, [En línea]. [Último acceso: Ene. 21, 2021].

[37] Y. Bengio, P. Simard, y P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[38] S. Hochreiter y J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] R. Pascanu, C. Gulcehre, K. Cho, y Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.

[40] A. Graves, A.-R. Mohamed, y G. Hinton, "Speech recognition with deep recurrent neural networks," en *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 6645–6649, 2013.

[41] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[42] S. Hochreiter y J. Schmidhuber, "LSTM can solve hard long time lag problems," en *Advances in Neural Information Processing Systems (NIPS 1997)*, pp. 473–479, 1997.

[43] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017.

[44] K. Zhang, Z. Liu, y L. Zheng, "Short-term prediction of passenger demand in multi-zone level: Temporal convolutional neural network with multi-task learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1480–1490, 2019.

[45] K. Mitra, A. B. Zaslavsky, y C. Åhlund, "Context-aware QoE modelling, measurement, and prediction in mobile computing systems," *IEEE Transactions on Mobile Computing*, vol. 14, no. 5, pp. 920–936, 2015.

[46] V. Radu, P. Katsikouli, R. Sarkar, y M. K. Marina, "A semi-supervised learning approach for robust indoor-outdoor detection with smartphones," en *12th ACM Conference on Embedded Network Sensor Systems (SenSys 2014)*, pp. 280–294, 2014.

[47] P. D. Groves, H. Martin, K. Voutsis, D. Walter, y L. Wang, "Context detection, categorization and connectivity for advanced adaptive integrated navigation," en *26th International Technical Meeting of the Satellite Division of the institute of navigation (ION GNSS+ 2013)*. The Institute of Navigation, 2013.

[48] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, y R. R. Choudhury, "No need to war-drive: Unsupervised indoor localization," en *10th International Conference on Mobile Systems, Applications, and Services (MobiSys 2012)*, pp. 197–210, 2012.

[49] M. Alzantot y M. Youssef, "Crowdinside: automatic construction of indoor floorplans," en *20th International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2012)*, pp. 99–108, 2012.

[50] W. Xu, R. Chen, T. Chu, L. Kuang, Y. Yang, X. Li, J. Liu, y Y. Chen, "A context detection approach using GPS module and emerging sensors in smartphone platform," en *IEEE Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS 2014)*, pp. 156–163, 2014.

[51] A. Martinelli, H. Gao, P. D. Groves, y S. Morosi, "Probabilistic context-aware step length estimation for pedestrian dead reckoning," *IEEE Sensors Journal*, 2017.

[52] J. Krumm y R. Hariharan, "Tempio: inside/outside classification with temperature," en *2nd International Workshop on Man-Machine Symbiotic Systems*, pp. 241–250, 2004.

[53] A. Payne y S. Singh, "Indoor vs. outdoor scene classification in digital photographs," *Pattern Recognition*, vol. 38, no. 10, pp. 1533–1545, 2005.

[54] J. Chung, M. Donahoe, C. Schmandt, I.-J. Kim, P. Razavai, y M. Wiseman, "Indoor location sensing using geo-magnetism," en *9th International Conference on Mobile Systems, Applications, and Services (MobiSys 2011)*, pp. 141–154, 2011.

[55] P. Zhou, Y. Zheng, Z. Li, M. Li, y G. Shen, "Iodetector: A generic service for indoor outdoor detection," en *10th AMC Conference on Embedded Network Sensor Systems (SenSys 2012)*, pp. 113–126, 2012.

[56] R. Sung, S.-h. Jung, y D. Han, "Sound based indoor and outdoor environment detection for seamless positioning handover," *ICT Express*, vol. 1, no. 3, pp. 106–109, 2015.

[57] M. Ali, T. ElBatt, y M. Youssef, "SenseIO: Realistic ubiquitous indoor outdoor detection system using smartphones," *IEEE Sensors Journal*, vol. 18, no. 9, pp. 3684–3693, 2018.

[58] W. Wang, Q. Chang, Q. Li, Z. Shi, y W. Chen, "Indoor-outdoor detection using a smart phone sensor," *MDPI Sensors*, vol. 16, no. 10, p. 1563, 2016.

[59] O. Canovas, P. E. Lopez-de Teruel, y A. Ruiz, "Detecting indoor/outdoor places using WiFi signals and AdaBoost," *IEEE Sensors Journal*, vol. 17, no. 5, pp. 1443–1453, 2017.

[60] X. Zhu, A. B. Goldberg, R. Brachman, y T. Dietterich, *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

[61] P. A. Sánchez, S. Luna-Ramírez, M. Toril, C. Gijón, y J. L. Bejarano-Luque, "A data-driven scheduler performance model for qoe assessment in a LTE radio network planning tool," *Computer Networks*, vol. 173, p. 107186, 2020.

[62] S. Coast, "OpenStreetMap," Disponible en: https://www.openstreetmap.org/, 2004, [En línea]. [Último acceso: Nov. 4, 2018].

[63] L. Jarvis, J. McEachen, y H. Loomis, "Geolocation of LTE subscriber stations based on the timing advance ranging parameter," en *Military Communications Conference (MILCOM 2011)*, pp. 180–187, 2011.

[64] P. McCullagh y J. A. Nelder, *Generalized linear models*, vol. 37. CRC press, 1989.

[65] V. Wille, M. Toril, y S. Luna-Ramirez, "Estimating pole capacity in a live hsdpa network," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1260–1263, 2013.

[66] W. C. Navidi, *Statistics for engineers and scientists*. McGraw-Hill Higher Education New York, NY, USA, 2008.

[67] S. Chatterjee y A. S. Hadi, "Influential observations, high leverage points, and outliers in linear regression," *Statistical Science*, pp. 379–393, 1986.

[68] R. Koenker y G. Bassett, "Robust tests for heteroscedasticity based on regression quantiles," *Econometrica*, vol. 50, no. 1, pp. 43–61, 1982.

[69] S. S. Shapiro y M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 12 1965.

[70] H. Liu, H. Darabi, P. Banerjee, y J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[71] J. Johansson, W. A. Hapsari, S. Kelley, y G. Bodog, "Minimization of drive tests in 3GPP release 11," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 36–43, 2012.

[72] 3GPP, "Digital cellular telecommunications system (phase 2+); Universal Mobile Telecommunications System (UMTS); Telecommunication management; Subscriber and equipment trace; Trace data definition and management," Technical Specification (TS) 132.423, 2007, version 7.4.0, Release 7.

[73] Q. D. Vo y P. De, "A survey of fingerprint-based outdoor localization," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 491–506, 2016.

[74] 3GPP, "Stage 2 functional specification of UE positioning in E-UTRAN," Technical Specification (TS) 36.305, 2016, version 13.0.0, Release 13.

[75] S. Holm, "Hybrid ultrasound-rfid indoor positioning: Combining the best of both worlds," en *IEEE International Conference on RFID Technology and Applications (RFID-TA 2009)*, pp. 155–162, 2009.

[76] P. D. Groves, "Navigation using inertial sensors [tutorial]," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 2, pp. 42–69, 2015.

[77] B. Gill, S. Khanifar, J. Robinson, y S. Westwood, "OpenSignal," Disponible en: https://www.opensignal.com/, 2010, [En línea]. [Último acceso: Jun. 20, 2019].

[78] Y. Vardi, T. Scherzer, A. Kohavi, y S. Scherzer, "WeFi," Disponible en: https://www.wefi.com/, 2006, [En línea]. [Último acceso: Ene. 10, 2021].

[79] M. Apgar, "Ookla," Disponible en: https://www.speedtest.net/, 2006, [En línea]. [Último acceso: Jun. 21, 2023].

[80] T. Rathonyi, "CellRebel," Disponible en: https://www.cellrebel.com/, 2018, [En línea]. [Último acceso: May. 8, 2023].

[81] S. Fortes, D. Palacios, I. Serrano, y R. Barco, "Applying social event data for the management of cellular networks," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 36–43, 2018.

[82] "Twitter API documentation," Disponible en: https://dev.twitter.com/docs, 2006, [En línea]. [Último acceso: Abr. 2, 2019].

[83] "Flickr API documentation," Disponible en: https://www.flickr.com/services/api/, 2004, [En línea]. [Último acceso: Abr. 2, 2019].

[84] M. A. Oliver y R. Webster, "Kriging: a method of interpolation for geographical information systems," *International Journal of Geographical Information System*, vol. 4, no. 3, pp. 313–332, 1990.

[85] J. Delhomme, "Kriging in the hydrosciences," *Advances in Water Resources*, vol. 1, no. 5, pp. 251–266, 1978.

[86] X. Zhong, A. Kealy, y M. Duckham, "Stream kriging: Incremental and recursive ordinary kriging over spatiotemporal data streams," *Computers & Geosciences*, vol. 90, pp. 134–143, 2016.

[87] E. Inc., "Esper," Disponible en: http://www.espertech.com/esper/, 2006, [En línea]. [Último acceso: Ene. 14, 2020].

[88] Y. Yamamoto, "Twitter4j library," Disponible en: http://twitter4j.org/en/, 2007, [En línea]. [Último acceso: Abr. 2, 2019].

[89] H. R. Anderson y J. P. McGeehan, "Optimizing microcell base station locations using simulated annealing techniques," en *IEEE Vehicular Technology Conference (VTC 1994)*, vol. 44, pp. 858–858, 1994.

[90] C. Y. Lee y H. G. Kang, "Cell planning with capacity expansion in mobile communications: A tabu search approach," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 5, pp. 1678–1691, 2000.

[91] A. J. Nebro, E. Alba, G. Molina, F. Chicano, F. Luna, y J. J. Durillo, "Optimal antenna placement using a new multi-objective chc algorithm," en *9th Annual Conference on Genetic and Evolutionary Computation (GECCO 2007)*, pp. 876–883. ACM, 2007.

[92] L. Raisanen y R. M. Whitaker, "Comparison and evaluation of multiple objective genetic algorithms for the antenna placement problem," *Mobile Networks and Applications*, vol. 10, no. 1-2, pp. 79–88, 2005.

[93] M. J. Nawrocki y T. W. Wieckowski, "Optimal site and antenna location for UMTS output results of 3g network simulation software," en *IEEE 14th International Conference on Microwaves, Radar and Wireless Communications (MIKON 2002)*, vol. 3, pp. 890–893, 2002.

[94] J. Niemelä y J. Lempiäinen, "Impact of base station locations and antenna orientations on UMTS radio network capacity and coverage evolution," en *IEEE International Symposium On Wireless Personal Multimedia Communications (WPMC 2003)*, 2003.

[95] E. Amaldi, A. Capone, y F. Malucelli, "Planning UMTS base station location: Optimization models with power control and algorithms," *IEEE Transactions on wireless Communications*, vol. 2, no. 5, pp. 939–952, 2003.

[96] L. Nagy y L. Farkas, "Indoor base station location optimization using genetic algorithms," en *IEEE 11th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC 2000)*, vol. 2, pp. 843–846, 2000.

[97] Z. Ji, T. K. Sarkar, y B.-H. Li, "Methods for optimizing the location of base stations for indoor wireless communications," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 10, pp. 1481–1483, 2002.

[98] L. Nagy, "Global optimization of indoor radio coverage," *Automatika*, vol. 53, no. 1, pp. 69–79, 2012.

[99] S. Rodd, A. Prof, y A. H. Kulkarni, "Optimization algorithms for access point deployment in wireless networks," *Journal of Computer Applications*, vol. 2, no. 1, pp. 2–2, 2009.

[100] J. Wong, A. Mason, M. Neve, y K. Sowerby, "Base station placement in indoor wireless systems using binary integer programming," *IEE Proceedings-Communications*, vol. 153, no. 5, pp. 771–778, 2006.

[101] Y. Ngadiman, Y. H. Chew, y B. S. Yeo, "A new approach for finding optimal base stations configuration for cdma systems jointly with uplink and downlink constraints," en *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2005)*, vol. 4, pp. 2751–2755, 2005.

[102] L. K. Pujji, K. W. Sowerby, y M. J. Neve, "A new algorithm for efficient optimisation of base station placement in indoor wireless communication systems," en *7th Annual Communication Networks and Services Research Conference (CNSR 2009)*, pp. 425–427, 2009.

[103] S. Wang, W. Guo, y T. O'Farrell, "Optimising femtocell placement in an interference limited network: Theory and simulation," en *IEEE Vehicular Technology Conference (VTC Fall 2012)*, pp. 1–6, 2012.

[104] M. Talau, E. C. Wille, y H. S. Lopes, "Solving the base station placement problem by means of swarm intelligence," en *IEEE Symposium on Computational Intelligence for Communication Systems and Networks (CIComms 2013)*, pp. 39–44, 2013.

[105] L. K. Pujji, K. W. Sowerby, y M. J. Neve, "Development of a hybrid algorithm for efficient optimisation of base station placement for indoor wireless communication systems," *Wireless personal communications*, vol. 69, no. 1, pp. 471–486, 2013.

[106] J. Liu, Q. Chen, y H. D. Sherali, "Algorithm design for femtocell base station placement in commercial building environments," en *IEEE International Conference on Computer Communications (INFOCOM 2012)*, pp. 2951–2955, 2012.

[107] J. M. Ruiz, M. Toril, y S. Luna-Ramírez, "A femtocell location strategy for improving adaptive traffic sharing in heterogeneous LTE networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 38, 2015.

[108] C. Sharma, Y. F. Wong, W.-S. Soh, y W.-C. Wong, "Access point placement for fingerprint-based localization," en *IEEE International Conference on Communication Systems (ICCS 2010)*, pp. 238–243, 2010.

[109] K. Farkas, Á. Huszák, y G. Gódor, "Optimization of Wi-Fi access point placement for indoor localization," *Journal IIT (Informatics & IT Today)*, vol. 1, no. 1, pp. 28–33, 2013.

[110] F. Al-Turjman, E. Ever, y H. Zahmatkesh, "Small cells in the forthcoming 5G/IoT: Traffic modelling and deployment overview," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 28–65, 2018.

[111] H. Klessig, H. Kuntzschmann, L. Scheuvens, B. Almeroth, P. Schulz, y G. Fettweis, "Twitter as a Source for Spatial Traffic Information in Big Data-Enabled Self-Organizing Networks," en *IEEE Wireless Communications and Networking Conference (WCNC 2017)*, pp. 1–5, 2017.

[112] C. Gijón, M. Toril, M. Solera, S. Luna-Ramírez, y L. R. Jiménez, "Encrypted traffic classification based on unsupervised learning in cellular radio access networks," *IEEE Access*, vol. 8, pp. 167 252–167 263, 2020.

[113] V. Buenestado, J. M. Ruiz-Aviles, M. Toril, S. Luna-Ramírez, y A. Mendo, "Analysis of Throughput Performance Statistics for Benchmarking LTE Networks," *IEEE Communications Letters*, vol. 18, no. 9, pp. 1607–1610, 2014.

[114] 3GPP, "Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E UTRAN): Definitions," Technical Specification (TS) 32.450, Jun. 2010, version 9.1.0 Release 9.

[115] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, y L. Nan, "Trends in Small Cell Enhancements in LTE Advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98–105, 2013.

[116] ICT-671650 mmMAGIC, *D1.1 "Use case characterization, KPIs and preferred suitable frequency ranges for future 5G systems between 6 GHz and 100 GHz"*, Nov. 2015, mmMAGIC Deliverable 1.1 Version 1.

[117] D. Tikunov y T. Nishimura, "Traffic prediction for mobile network using Holt-Winter's exponential smoothing," en *15th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2007)*, pp. 1–5, 2007.

[118] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, y Y. Li, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 796–805, 2016.

[119] C. Zhang, H. Zhang, D. Yuan, y M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656–1659, 2018.

[120] Y. Yu, J. Wang, M. Song, y J. Song, "Network traffic prediction and result analysis based on seasonal ARIMA and correlation coefficient," en *International Conference on Intelligent System Design and Engineering Application (ISDEA 2010)*, vol. 1, pp. 980–983, 2010.

[121] Y. Shu, M. Yu, O. Yang, J. Liu, y H. Feng, "Wireless traffic modeling and prediction using seasonal ARIMA models," *IEICE Transactions on Communications*, vol. 88, no. 10, pp. 3992–3999, 2005.

[122] L. Fang, X. Cheng, H. Wang, y L. Yang, "Mobile demand forecasting via deep graph-sequence spatiotemporal modeling in cellular networks," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3091–3101, 2018.

[123] B. Zhou, D. He, y Z. Sun, "Traffic modeling and prediction using ARIMA/GARCH model," *Modeling and Simulation Tools for Emerging Telecommunication Networks*, pp. 101–121, 2006.

[124] N. Bui y J. Widmer, "Data-driven evaluation of anticipatory networking in lte networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 10, pp. 2252–2265, 2018.

[125] J. Scourias y T. Kunz, "A dynamic individualized location management algorithm," en *8th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 1997)*, vol. 3, pp. 1004–1008, 1997.

[126] M. Yin, M. Sheehan, S. Feygin, J. Paiement, y A. Pozdnoukhov, "A generative model of urban activities from cellular data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, pp. 1682–1696, 2018.

[127] M. D. Jnr, J. D. Gadze, y D. K. Anipa, "Short-term traffic volume prediction in UMTS networks using the kalman filter algorithm," *International Journal of Mobile Network Communications & Telematic*, vol. 3, no. 6, pp. 31–40, 2013.

[128] R. Li, Z. Zhao, X. Zhou, y H. Zhang, "Energy savings scheme in radio access networks via compressive sensing-based traffic load prediction," *Transactions on emerging telecommunications technologies*, vol. 25, no. 4, pp. 468–478, 2014.

[129] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, y D. Yang, "Spatio-temporal modeling and prediction in cellular networks: A big data enabled deep learning approach," en *IEEE International Conference on Computer Communications (INFOCOM 2017)*, pp. 1–9, 2017.

[130] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, y H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless communications*, vol. 24, no. 5, pp. 175–183, 2017.

[131] C. Zhang, P. Patras, y H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[132] Y. Zang, F. Ni, Z. Feng, S. Cui, y Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks," en *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP 2015)*, pp. 458–462, 2015.

[133] Y. Hua, Z. Zhao, Z. Liu, X. Chen, R. Li, y H. Zhang, "Traffic prediction based on random connectivity in deep learning with long short-term memory," en *IEEE 88th Vehicular Technology Conference (VTC-Fall 2018)*, pp. 1–6, 2018.

[134] H. D. Trinh, L. Giupponi, y P. Dini, "Mobile traffic prediction from raw data using LSTM networks," en *IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2018)*, pp. 1827–1832, 2018.

[135] A. Azari, P. Papapetrou, S. Denic, y G. Peters, "Cellular traffic prediction and classification: a comparative evaluation of LSTM and ARIMA," en *International Conference on Discovery Science (DS 2019)*, pp. 129–144. Springer, 2019.

[136] L. Nie, D. Jiang, S. Yu, y H. Song, "Network traffic prediction based on deep belief network in wireless mesh backbone networks," en *IEEE Wireless Communications and Networking Conference (WCNC 2017)*, pp. 1–5, 2017.

[137] C.-W. Huang, C.-T. Chiang, y Q. Li, "A study of deep learning networks on mobile traffic forecasting," en *IEEE 28th Annual International Symposium on*

Personal, Indoor, and Mobile Radio Communications (PIMRC 2017), pp. 1–6, 2017.

[138] C. Zhang y P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," en *18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2018)*, pp. 231–240, 2018.

[139] J. Feng, X. Chen, R. Gao, M. Zeng, y Y. Li, "DeepTP: An end-to-end neural network for mobile cellular traffic prediction," *IEEE Network*, vol. 32, no. 6, pp. 108–115, 2018.

[140] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, y C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 2190–2202, 2018.

[141] P. Makris, D. N. Skoutas, y C. Skianis, "A survey on context-aware mobile and wireless networking: On networking and computing environments' integration," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 362–386, 2012.

[142] S. Almeida, J. Queijo, y L. M. Correia, "Spatial and temporal traffic distribution models for gsm," en *IEEE 50th Vehicular Technology Conference (VTS 1999)*, vol. 1, pp. 131–135 vol.1, 1999.

[143] H. Wang, F. Xu, Y. Li, P. Zhang, y D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," en *Internet Measurement Conference (IMC 2015)*, pp. 225–238, 2015.

[144] B. Yang, W. Guo, B. Chen, G. Yang, y J. Zhang, "Estimating mobile traffic demand using twitter," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 380–383, 2016.

[145] H. Klessig, H. Kuntzschmann, L. Scheuvens, B. Almeroth, P. Schulz, y G. Fettweis, "Twitter as a source for spatial traffic information in big data-enabled self-organizing networks," en *IEEE Wireless Communications and Networking Conference (WCNC 2017)*, pp. 1–5, 2017.

[146] S. Fortes, A. Aguilar-García, R. Barco, F. B. Barba, J. A. Fernández-Luque, y A. Fernández-Durán, "Management architecture for location-aware self-organizing LTE/LTE-A small cell networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 294–302, 2015.

[147] B. Ma, B. Yang, Y. Zhu, y J. Zhang, "Context-Aware Proactive 5G Load Balancing and Optimization for Urban Areas," *IEEE Access*, vol. 8, pp. 8405–8417, 2020.

[148] S. Fortes, A. A. Garcia, J. A. Fernandez-Luque, A. Garrido, y R. Barco, "Context-aware self-healing: user equipment as the main source of information for small-cell indoor networks," *IEEE Vehicular Technology Magazine*, vol. 11, no. 1, pp. 76–85, 2016.

[149] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, y J. Widmer, "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1790–1821, 2017.

[150] A. S. A. Khan, I. Afyouni, y Z. A. Aghbari, "A big data platform for spatio-temporal social event discovery," en *IEEE 21st International Conference on Mobile Data Management (MDM 2020)*, pp. 248–249, 2020.

[151] Z. Yu, F. Yi, Q. Lv, y B. Guo, "Identifying on-site users for social events: Mobility, content, and social relationship," *IEEE Transactions on Mobile Computing*, vol. 17, no. 9, pp. 2055–2068, 2018.

[152] T. Xu, H. Zhu, H. Zhong, G. Liu, H. Xiong, y E. Chen, "Exploiting the dynamic mutual influence for predicting social event participation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1122–1135, 2019.

[153] A. Aguilar-Garcia, S. Fortes, A. Fernandez Duran, y R. Barco, "Context-aware self-optimization: Evolution based on the use case of load balancing in small-cell networks," *IEEE Vehicular Technology Magazine*, vol. 11, no. 1, pp. 86–95, 2016.

[154] Q. He, G. P. Koudouridis, y G. Dán, "A comparison of machine and statistical time series learning for encrypted traffic prediction," en *International Conference on Computing, Networking and Communications (ICNC 2020)*, pp. 714–718, 2020.

[155] "Eventful API documentation," Disponible en: https://api.eventful.com/docs, 2004, [En línea]. [Último acceso: Abr. 24, 2020].

[156] "Yelp API documentation," Disponible en: https://www.yelp.com/developers/, 2004, [En línea]. [Último acceso: Abr. 24, 2020].

[157] A. J. García, V. Buenestado, M. Toril, S. Luna-Ramírez, y J. Ruiz, "A geometric method for estimating the nominal cell range in cellular networks," *Mobile Information Systems*, vol. 2018, 2018.

[158] M. Toril, S. Luna-Ramírez, y V. Wille, "Automatic replanning of tracking areas in cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2005–2013, Jun. 2013.

[159] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, y S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 554–557, 2018.

[160] Y. Sun, F. P.-W. Lo, y B. Lo, "EEG-based user identification system using 1d-convolutional long short-term memory neural networks," *Expert Systems with Applications*, vol. 125, pp. 259–267, 2019.

[161] B. Rekabdar y C. Mousas, "Dilated convolutional neural network for predicting driver's activity," en *21st International Conference on Intelligent Transportation Systems (ITSC 2018)*, pp. 3245–3250, 2018.

[162] M. Claesen y B. De Moor, "Hyperparameter search in machine learning," *arXiv preprint arXiv:1502.02127*, 2015.

[163] X. Glorot y Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," en *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh y M. Titterington, Eds., vol. 9, pp. 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010.

[164] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, y I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[165] M. Masse, *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. "O'Reilly Media, Inc.", 2011.

[166] B. Zhao, "Web scraping," *Encyclopedia of big data*, pp. 1–3, 2017.

[167] J. Huggins, "Selenium," Disponible en: https://www.selenium.dev/, 2004, [En línea]. [Último acceso: Feb. 15, 2023].

[168] C. Carter, "Great circle distances," *SiRF White Paper*, 2002.