Universidad Zaragoza

Trabajo Fin de Máster

Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección

Modelado, publicación y análisis de la calidad de los datos de monitorización de calidad del aire en el contexto del proyecto TRAFAIR

Autor

Manuel Ángel Jáñez García

Director

Javier Nogueras Iso

Facultad de Filosofía y Letras

Curso 2019-2020

Resumen

Los desafíos relativos a la calidad del aire urbano representan la segunda prioridad ambiental de la Comisión Europea por sus efectos en la salud humana. Con esa premisa, este trabajo desarrollado en el contexto del proyecto TRAFAIR: *Understanding traffic flows to improve air quality* (2017-EU-IA-0167) se ha centrado en el tratamiento y publicación de los conjuntos de datos de monitorización de calidad del aire en el prototipo de la infraestructura desplegada por el proyecto para la ciudad de Zaragoza.

En primer lugar, se han identificado las redes de sensores de los sistemas de monitorización para la toma de datos: las estaciones reguladas del Ayuntamiento de Zaragoza y los sensores de bajo coste desplegados por el proyecto TRAFAIR. A partir de los datos proporcionados por los sensores, se ha implantado un modelo de base de datos para su almacenamiento.

En segundo lugar, se ha propuesto una arquitectura para la publicación de parte de los datos almacenados en la base de datos de TRAFAIR a través del servidor de datos espaciales Geoserver. Para ello, se han modelado tanto capas de fenómenos discretos, vistas parciales de la información almacenada en la base de datos, como coberturas espaciales generadas a partir de algoritmos de interpolación. En cuanto a su publicación en abierto, se ha implantado y adaptado un software para la creación de las capas a través del API Rest de Geoserver, lo que ha permitido a su vez acceder a los datos mediante servicios estándar definidos por el Open Geospatial Consortium (OGC). Para la difusión de los datos en portales de datos abiertos, se ha participado también en la implantación de un software para exportar la información de las capas de Geoserver como metadatos de tipo GeoDCAT-AP en servidores de tipo CKAN. Además, se ha propuesto una estrategia para distribuir algunos conjuntos de datos en formatos específicos (CSV, RDF, NetCDF).

Finalmente, se ha analizado la calidad de los datos espaciales publicados desde distintos puntos de vista: conformidad con las especificaciones de datos de INSPIRE, calidad de los metadatos GeoDCAT-AP a partir de una metodología basada en el procedimiento empleado por el Portal Europeo de Datos; análisis de la completitud de las observaciones; y análisis de la exactitud de las observaciones de los sensores del proyecto TRAFAIR.

Abstract

Urban air quality challenges represent the second environmental priority of the European Commission because of its effects on human health. With this premise, this work developed in the context of the research project TRAFAIR: *Understanding traffic flows to improve air quality* (2017-EU-IA-0167) has focused on the processing and publication of air quality monitoring data sets in the prototype of the infrastructure deployed by the project for the city of Zaragoza.

Firstly, the sensor networks of the monitoring systems for data collection have been identified: the regulated stations of the Zaragoza City Council and the low-cost sensors deployed by the TRAFAIR project. Based on the data provided by the sensors, a database model has been implemented for their storage.

Secondly, an architecture has been proposed for publishing part of the data stored in the TRAFAIR database through the Geoserver spatial data server. For this purpose, both discrete phenomenon layers, partial views of the information stored in the database, and spatial coverage generated from interpolation algorithms have been modeled. As for open access publishing, a software has been implemented and adapted for the creation of the layers through the Geoserver Rest API, which in turn has allowed access to the data through standard services defined by the Open Geospatial Consortium (OGC). For the dissemination of the data in open data portals, we have also participated in the implementation of a software to export the information of the Geoserver layers as GeoDCAT-AP type metadata in CKAN type servers. In addition, a strategy has been proposed to distribute some data sets in specific formats (CSV, RDF, NetCDF).

Finally, the quality of the published spatial data has been analyzed from different points of view: conformity with the INSPIRE data specifications, quality of the GeoDCAT-AP metadata from a methodology based on the procedure used by the European Data Portal; analysis of the completeness of the observations; and analysis of the accuracy of the observations from the TRAFAIR project sensors.

Agradecimientos

A mi pareja, Virginia, por su paciencia y apoyo durante tantos años.

A mis padres, María Jesús y Miguel Ángel, y mi hermano Álvaro, por estar siempre ahí.

Y al tutor del presente TFM y director dentro del proyecto, Javier Nogueras, por su incalculable ayuda y dedicación. También a Raquel Trillo, coordinadora del mismo, por la oportunidad concedida al formar parte de todo ello, así como al resto de compañeros de TRAFAIR por su paciencia y colaboración.

Índice general

Índice de figuras

Índice de cuadros

Índice de códigos

1. Introducción

El proyecto TRAFAIR: *Understanding traffic flows to improve air quality* (2017-EU-IA-0167)[1] es una acción financiada por el instrumento de la Unión Europea CEF (Connecting Europe Facility),[2] cuyos objetivos principales son la monitorización en tiempo real de los niveles de polución, el desarrollo de un servicio de predicción de la calidad del aire basado en datos resultantes de las condiciones variables de los flujos de tráfico y predicciones meteorológicas, así como la definición y publicación de los datos, y metadatos, de calidad del aire como datos abiertos.

En la Unión Europea, los esfuerzos para mejorar la calidad del aire urbano durante las últimas décadas han llevado a autoridades nacionales y comunitarias a intensificar sus actuaciones para lograr la reducción de los límites previstos en la legislación. En concreto, España sigue bajo la atenta mirada de las instituciones europeas por no respetar los valores límite de dióxido de nitrógeno (NO_2).[3] Este contaminante se genera fundamentalmente en áreas urbanas donde el tráfico es más intenso y existen instalaciones domésticas de calefacción, y también, donde están presentes centrales térmicas de generación eléctrica y áreas industriales. Sin embargo, gran parte de los ciudadanos y administraciones no tienen acceso a información sobre niveles de contaminación en sus localidades. Tomando como base esta situación, el proyecto TRAFAIR persigue el desarrollo de un conjunto de herramientas comunes para el seguimiento de la calidad del aire en tiempo real, los flujos de tráfico urbano, así como los pronósticos de dispersión de contaminación en las próximas 24 ó 48 horas [1].

[1] `http://trafair.eu/`
[2] `https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom`
[3] CE, Comunicado del.25 de julio de 2019:
`https://ec.europa.eu/commission/presscorner/detail/es/IP_19_4256`

Espoleando la consecución de este objetivo se encuentra la amenaza para la salud que representa la polución del aire en toda Europa: 412.000 muertes en 2016;[4] en parte atribuibles a los efectos derivados del transporte terrestre masivo. En ese sentido, se busca contrastar las posibilidades de implantar herramientas a diversas escalas para el análisis de los niveles de polución, de manera cooperativa, a través de la experiencia piloto en 6 ciudades europeas de diferente tamaño: Zaragoza (670.000 hb.), Florencia (380.000 hb.), Módena (187.000 hb.), Livorno (158.000 hb.), Santiago de Compostela (97.000 hb.) y Pisa (88.000 hb.).

TRAFAIR aspira a proporcionar a la sociedad un conjunto coherente de servicios con información útil para analizar la calidad de sus lugares de residencia, en base a estándares comunes, que ayuden a sustentar las decisiones políticas, y concienciar acerca de sus entornos para contribuir en última instancia a mejorar la calidad de vida de los ciudadanos europeos.

1.1. Justificación del trabajo

El presente Trabajo de Fin de Máster (TFM) se enmarca en el contexto del proyecto TRAFAIR en la ciudad de Zaragoza, y forma parte del currículo académico del Máster Universitario en Tecnologías de la Información Geográfica para la Ordenación del Territorio: Sistemas de Información Geográfica y Teledetección, de la Universidad de Zaragoza.

Desde el punto de vista de las Tecnologías de la Información Geográfica (TIG), TRAFAIR ofrece un conjunto diverso de productos finales (servicios OGC, *web mapping*, IDEs, datos abiertos, cartografía de polución y predicción, aplicaciones móviles, etc.), en los que la información geoespacial y el potencial uso de Sistemas de Información Geográfica (SIG) es fundamental; explicitando la robusta conexión existente entre las geociencias y las tecnologías de la computación. Particularmente, en su vertiente de desarrollo e implementación de soluciones geoespaciales, hace uso de procesos de modelado, población y gestión de la información geográfica, y sus metadatos, mediante Infraestructuras de Datos Espaciales (IDE) en conformidad con las

[4] *"Las estimaciones de los efectos en la salud atribuibles a la exposición a la contaminación atmosférica indican que las concentraciones de PM2,5 en 2016 fueron responsables de unas 412.000 muertes prematuras originadas por la exposición a largo plazo en Europa (más de 41 países), de las cuales unas 374.000 se produjeron en la UE-28"* EEA Report No 10/2019, pág. 8: `https://www.eea.europa.eu/publications/air-quality-in-europe-2019`

directivas europeas en la materia (INSPIRE)[5] y los estándares de calidad necesarios para su publicación.

Existen múltiples aplicaciones prácticas que ponen en valor su utilidad, como pueden ser:

1. Potenciar la actual generación de datos puntuales de estaciones reguladas de calidad del aire existentes en grandes ciudades, con conjuntos de datos interpolados para cartografiar los niveles de polución.
2. Ofrecer herramientas de código abierto que fomenten la interoperabilidad de los datos y servicios, faciliten la continuidad de los proyectos locales y reduzcan los costes de implantación.
3. Contribuir a paliar las actuales carencias de monitorización ambiental urbana en España, proponiendo marcos comunes para la ampliación y estandarización de la red de análisis de calidad del aire en países que disponen de una extensa red de pequeñas y medianas urbes que pueden no ser capaces de asumir sus costes. En España residían en 2019, 19.032.829 personas en municipios urbanos de entre 20.000 y 200.000 habitantes, un 40,5 % del total nacional.[6]
4. Potenciar el Portal Europeo de Datos (European Data Portal),[7] así como su uso, generalizando la visualización y descarga de información para favorecer el desarrollo eficiente de políticas de movilidad, ordenación territorial, salud pública, medio ambiente, etc., más específicas y vinculadas al ámbito de ejecución.
5. Publicar servicios web y aplicaciones móviles para usuarios finales, que permitan el seguimiento, análisis y evaluación de las distintas estrategias públicas por parte de las instituciones académicas, ciudadanos y empresas.

[5] Directiva 2007/2/CE del Parlamento Europeo y del Consejo, de 14 de marzo de 2007. `https://eur-lex.europa.eu/eli/dir/2007/2/oj`

[6] Instituto Nacional de Estadística, Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero de 2019. `https://www.ine.es/up/cD2Tzg9t` (accedido 13 jul, 2020)

[7] `https://www.europeandataportal.eu/es`

1.2. Estado de la cuestión

Tradicionalmente, el acceso a los datos oficiales era una de las asignaturas pendientes de la administración pública por su complejidad, restricciones de acceso y limitaciones técnicas. Sin embargo, las últimas dos décadas han visto incrementada de manera intensa la disponibilidad de los mismos debido a la evolución tecnológica, siendo la propia ciudadanía la que demanda que los conjuntos de datos y temáticas disponibles cada vez sean mayores, y además, lo sean abiertos.

La definición comúnmente aceptada que la Open Knowledge Foundation utiliza para los datos abiertos como "aquellos que pueden ser utilizados, reutilizados y redistribuidos libremente";[8] se puede concretar aún más, como los datos gubernamentales que normalmente se proporcionan de forma gratuita, en un formato legible por la máquina y con restricciones mínimas para su utilización [2]. La creciente disponibilidad de datos en abierto por parte de las instituciones se ha visto correspondida con un aumento de usuarios, en términos generales, como se ejemplifica en la Figura 1.1 que hace referencia a la evolución reciente del portal de datos abiertos de la República de Irlanda.[9]

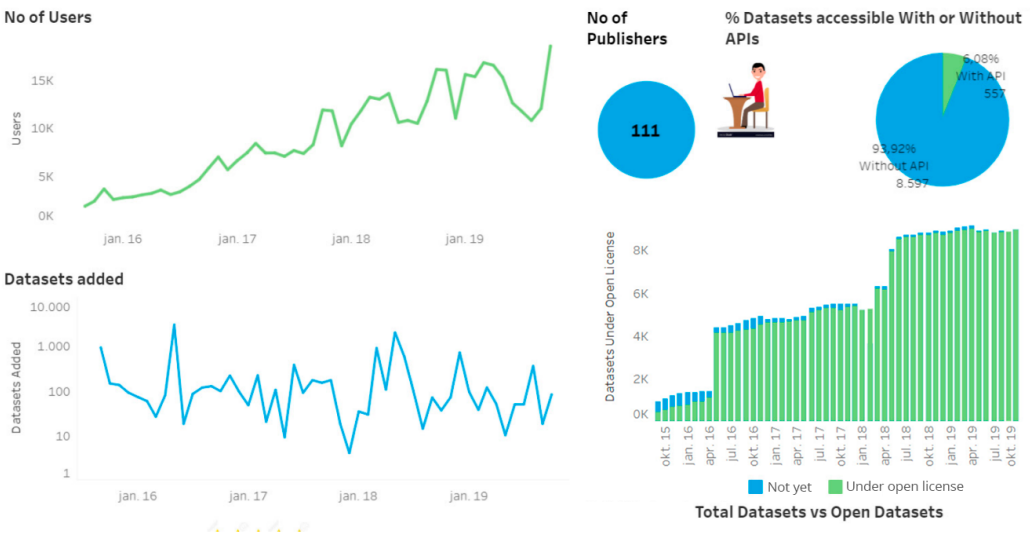

Figura 1.1: Evolución del portal irlandés de datos abiertos. European Data Portal.

[8] Open Knowledge Foundation, Open Data Handbook. https://opendatahandbook.org/guide/es/what-is-open-data/ (accedido 13 jul, 2020

[9] https://data.gov.ie/

A pesar de todo, los informes de la Unión Europea (Open Data Maturity Report 2019 [3]), exponen un crecimiento discreto, con un acceso restringido a nichos muy específicos. Sintetizando, en buena parte los datos son reutilizados por usuarios provenientes de la propia administración que los emite, el 54,8 %, u otras administraciones y personal académico (8,3 %), siendo apenas un 9,7 % de los usuarios encuestados en 2019 en España, ciudadanos sin vinculación alguna con las anteriores y un 24,2 % empresas [4].

No es de extrañar, dado que menos del 0,005 % de la población de la mayoría de los estados miembros acceden como usuarios únicos a sus portales, alrededor de 60.000 para el caso de España (≈0,00125 % respecto a la población nacional, Figura 1.2). Esta situación pone de manifiesto un potencial grado de crecimiento, por múltiples vías, como por ejemplo el incremento de la visibilidad, exposición de las bondades del servicio a escala europea o la publicación de datos más valiosos acordes con las demandas de los posibles usuarios.

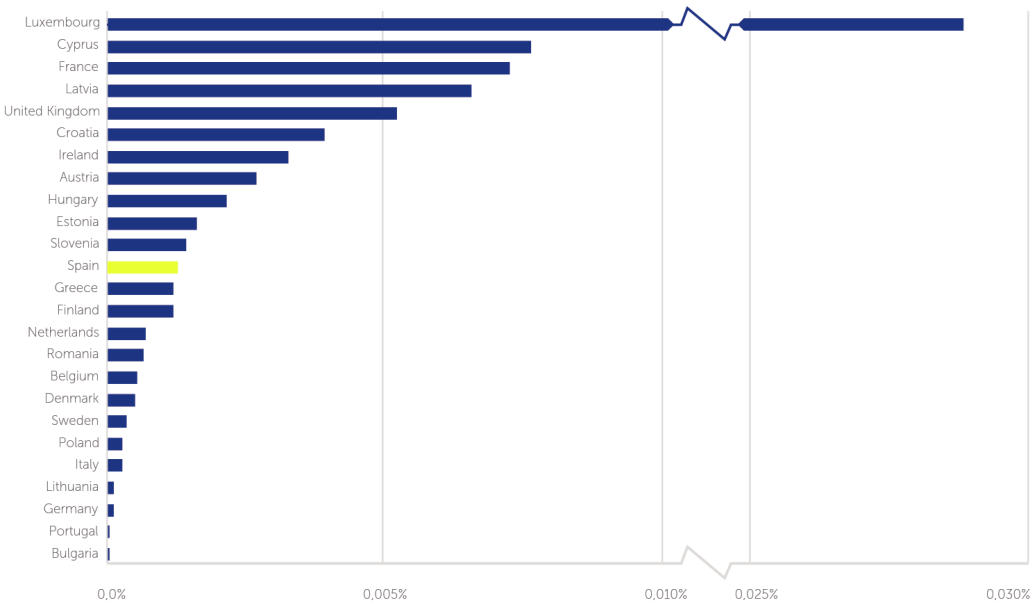

Figura 1.2: Usuarios únicos mensuales, EU-28. Open Data Maturity Report 2019.

Dentro del conjunto de datos que se publican en la UE, los datos abiertos vinculados a temáticas ambientales suponen la segunda categoría más popular de cuantas se documentan como tema de los metadatos que se encuentran presentes en los portales europeos y que son conformes con el perfil diseñado por la Comisión Europea para describir conjuntos de datos del sector público

(DCAT-AP),[10] tal y como se puede observar en la Figura 1.3. En relación con eso, la documentación de los datos para su posterior registro como datos abiertos, debe realizarse garantizando su compatibilidad con los modelos de metadatos utilizados. En particular, la compatibilidad con las normas de ejecución de la Directiva INSPIRE se puede conseguir mediante la adopción de perfiles como GeoDCAT-AP [5], el cual permite una correspondencia directa con el estándar internacional ISO 19115 [6] para metadatos geográficos. La descripción de los conjuntos de datos según este perfil se centra principalmente en proporcionar información sobre tres entidades principales: un catálogo con los conjuntos de datos y las formas de distribución asociadas a cada uno de ellos que se publica a través de un portal de datos abiertos; y una distinción entre las propiedades principales, aquellas propiedades de metadatos de DCAT-AP con vinculación directa con los metadatos ISO 19115 e INSPIRE, y las ampliadas, con propiedades adicionales para proporcionar una unión completa con los metadatos ISO 19115 e INSPIRE [7].

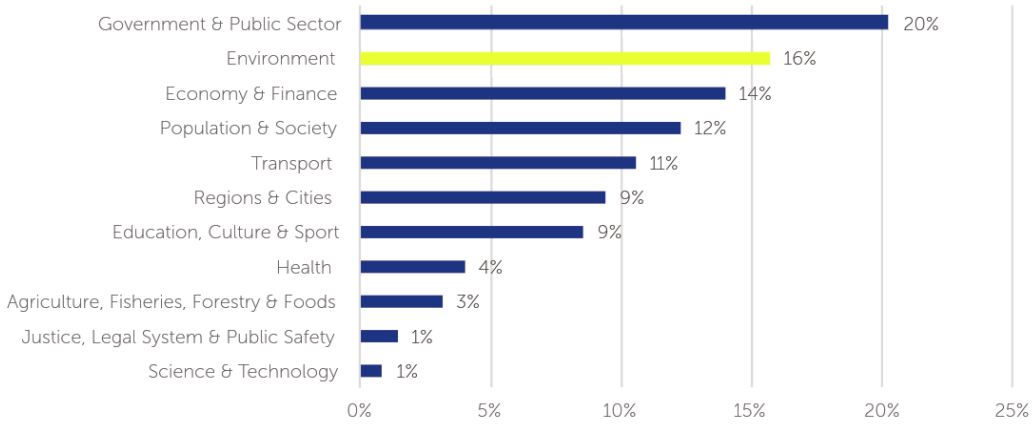

Figura 1.3: Categorías DCAT-AP 2019, UE-28. Open Data Maturity Report 2019.

A pesar de la generalización del uso de datos espaciales en acciones cotidianas, como las aplicaciones que hacen uso del geoprocesamiento para trazar rutas de navegación, la búsqueda de precios y alquileres inmobiliarios o el seguimiento de datos meteorológicos, todavía hoy la difusión de la información espacial es problemática. Existe documentación que trata de paliar la heterogeneidad en la difusión, publicación y reutilización de los datos como las Spatial Data on the Web Best Practices del W3C [8] abordando cuestiones como la representación de la geometría, el uso de sistemas de referencia de

[10]DCAT Application profile for data portals in Europe. `https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/about`

coordenadas y la publicación de meetadatos espaciales en lo que respecta a una ontología normalizada para la información geográfica [9]. En el contexto del presente trabajo estas cuestiones son de especial interés ya que surgen diversos problemas que se deben superar durante la integración de datos ambientales de monitorización y predicción en infraestructuras de datos semánticos abiertos [10].

En lo que respecta a la calidad de los datos geográficos y su nivel de completitud a fin de simular los posibles problemas, existen dos grandes opciones: procedimientos de evaluación automática de la calidad, como aquellos basados en servicios web con el objetivo de poder informar sobre la completitud (utilizando el emparejamiento a nivel de feature), la evaluación de la exactitud posicional (emparejamiento a nivel interno) y sobre la consistencia topológica utilizando unos métodos directos internos [11]; o procedimientos manuales más tradicionales en base a los principios descritos en la normativa ISO (actualmente 19157:2013 [12]), mediante pruebas estadísticas y cuadros de ensayo de muestras, inspección de atributos para valorar la exactitud temática, evaluación de la conformidad, etc. tal y como se realizaba, en algunas agencias geográficas europeas desde finales del siglo XX como el Institut Géographique National (IGN)[11] francés [13].

1.3. Objetivos

El objetivo principal del presente trabajo es el modelado, publicación y análisis de la calidad de datos abiertos de calidad del aire, que puedan contribuir a facilitar la evaluación de los niveles de polución urbana y el diseño de políticas territoriales específicas para el cumplimiento de los requerimientos en materia ambiental.

Buscando lograr el objetivo general, se definieron una serie de metas parciales que se detallan a continuación:

- Identificación de los sistemas de monitorización de la calidad del aire desde estaciones reguladas del Ayuntamiento de Zaragoza y la red de sensores de bajo coste dispuestas por el equipo de TRAFAIR.
- Descripción, modelado y proceso de población de la base de datos local mediante la información recogida por las estaciones reguladas y sensores de bajo coste.

[11] Actualmente Institut national de l'information géographique et forestière (Instituto Nacional de Información Geográfica y Forestal), `http://www.ign.fr/`

– Integración y definición de los modelos diseñados para la publicación de los fenómenos discretos mediante *Feature Types*,[12] así como para los mapas de interpolación a través de coberturas.

– Publicación de los conjuntos de datos en servidores de mapas web, portales de datos abiertos e integración de nuevos formatos de datos que mejoren su accesibilidad.

– Evaluación de la calidad de la cartografía generada, de las diferencias presentes en los datos obtenidos a partir de las estaciones legales y los sensores, y de los modelos conceptuales diseñados en TRAFAIR con los de los temas propuestos por INSPIRE.

1.4. Metodología y tecnologías aplicadas

Figura 1.4: Diagrama de Gantt con la planificación de las tareas.

[12]Recursos espaciales basados en datos vectoriales.

Para el desarrollo del TFM se ha hecho uso de una metodología dividida en tres fases diferenciadas. La primera vinculada a la definición de los objetivos, actividades del trabajo y la documentación sobre el mismo; una segunda para el análisis conjunto del modelado y población de la base de datos, y también, la definición de los conjuntos de datos presentes; y por último, una etapa de implementación-validación para la publicación de los datos abiertos y la evaluación de su calidad.

Todo ello se puede observar en la Figura 1.4, que muestra la distribución de los tiempos y tareas más concretas del trabajo. Hay que señalar que las partes más costosas temporalmente han sido tanto la publicación de los conjuntos de datos abiertos como la evaluación de la calidad de los mismos. Estas partes se superponen durante el tramo final debido a la revisión, y corrección de errores, durante la continua actualización de los conjuntos en el servidor de mapas y el portal de datos abiertos.

Respecto a las tecnologías usadas, estas han sido variadas en función de la fase del trabajo. De forma transversal, para la gestión, población y publicación de los conjuntos de datos se ha utilizado fundamentalmente el lenguaje de programación SQL [14] como herramienta de gestión de las bases de datos relacionales, concretamente mediante PostgreSQL [15], un SGDB[13] libre y de código abierto, que ofrece una gran cantidad de funcionalidades como la extensión espacial PostGIS [16] utilizada para almacenar objetos con características geográficas.

En previsión de garantizar la publicación y descarga de los datos en la web, se ha utilizado el software libre Geoserver [17], una aplicación modular que facilita la incorporación de extensiones y permite el cumplimiento de las especificaciones del Open Geospatial Consortium (OGC). Mediante Geoserver se han ajustado los servicios proporcionados a los estándares de descarga de fenómenos continuos, Web Coverage Service (WCS), discretos, Web Feature Service (WFS) y de visualización, Web Mapping Service (WMS). También, para la distribución de datos abiertos y la gestión de los metadatos geográficos se ha utilizado el software CKAN [18], que facilita la puesta en marcha de portales de distribución de conjuntos de datos.

Con el fin de evaluar la calidad de los modelos de interpolación de datos observados, así como la cobertura espacial, se han utilizado las librerías de análisis de datos pandas [19] mediante Python 3 [20], y el software SIG de código libre QGIS [21], para todo lo relativo a la verificación, tratamiento, análisis y modificación de la cartografía y estilos necesarios para su publica-

[13]Sistema gestor de base de datos.

ción web. Añadir también, las diversas librerías y software necesarios para la conversión y suministro de los datos en diversos formatos, y las ontologías descriptivas, como Tarql (SPARQL for Tables [22]), o NetCDF-4 [23], etc.

De forma resumida, se ha hecho uso de cinco lenguajes de programación diferentes: SQL, para gestionar las bases de datos, Python, para la validación de la calidad y automatización de tareas en QGIS, Java para la publicación de los conjuntos de datos, R para la interpolación de los mapas de calidad del aire y extracción de datos de POIs, y por último, para la documentación se ha utilizado LaTeX [24]. Respecto al entorno de desarrollo se ha utilizado PyCharm, IntelliJ, RStudio y el editor Visual Studio Code.

2. Modelado y población de datos

El modelado y población de los conjuntos de datos en la base de datos espacial del proyecto hace referencia al flujo de trabajo necesario para construir un almacén con información suficiente para el propósito deseado y de acuerdo a un modelo lógico coherente. Para ello es primero necesario detallar los distintos tipos de entidades existentes, así como sus atributos, las agrupaciones, interconexiones y relaciones estructurales existentes entre ellas.

La estructura básica de generación de la información y servicios OGC previstos en el proyecto TRAFAIR se puede visualizar en la Figura 2.1, con una base de datos central PostgreSQL junto con una extensión PostGIS, que la transforma de una base de datos relacional clásica en una base de datos espacial, permitiendo la adición de geometrías, nuevos tipos de datos, y también funciones e índices diseñados para facilitar el trabajo con información geográfica. Esta base de datos es alimentada por fuentes diversas: estaciones reguladas del Ayuntamiento de Zaragoza, sensores de bajo coste posicionados por los miembros del proyecto y un conjunto de datos espaciales (edificios, arcos de calles, información meteorológica, etc.). Estos últimos son necesarios para la posterior generación de información de predicción, y que junto con los datos de monitorización, procuran una serie de salidas con información sobre la calidad del aire en diferentes entornos, ya sean aplicaciones web o portales de datos abiertos.

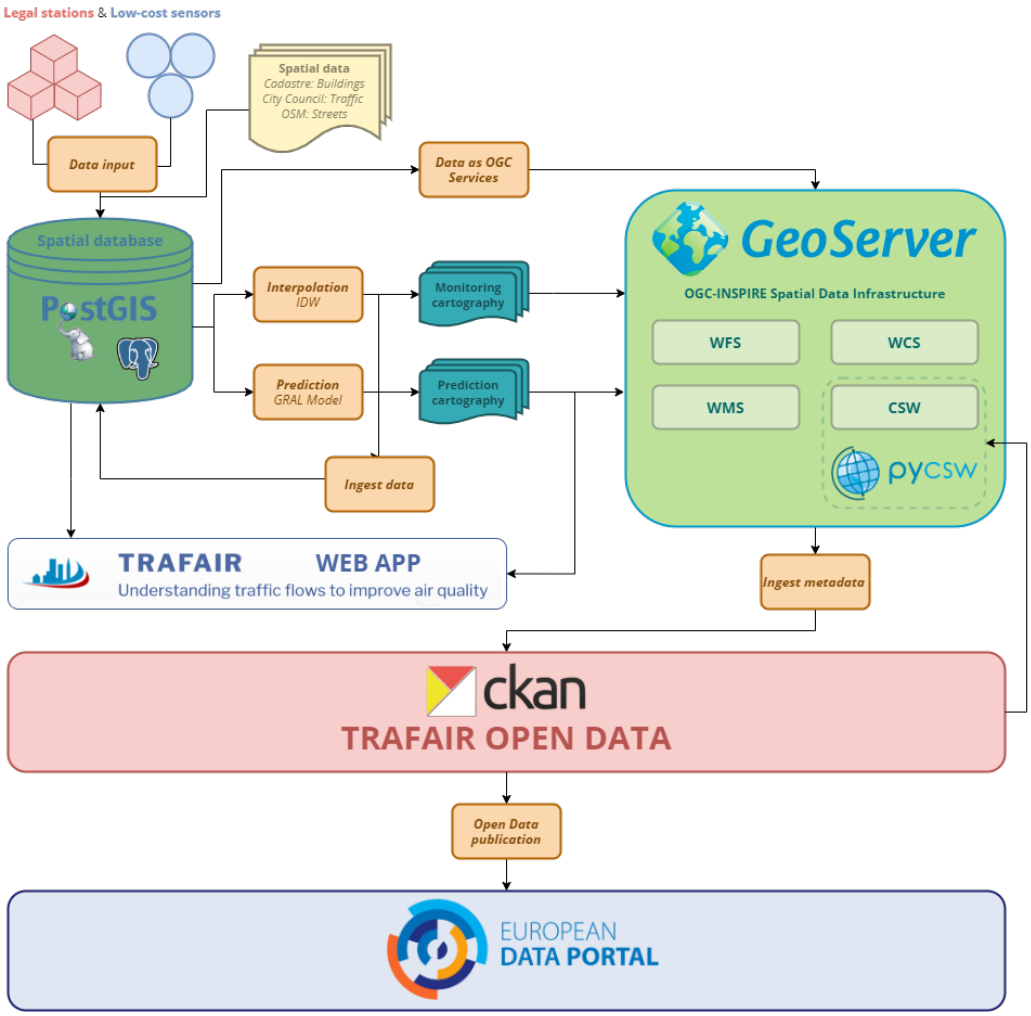

Figura 2.1: Estructura de datos y servicios del proyecto TRAFAIR.

2.1. Monitorización desde estaciones reguladas

En la ciudad de Zaragoza se hayan distribuidos diversos sensores (Cuadro 2.1) para el seguimiento de variables ambientales. En concreto, el Ayuntamiento de Zaragoza[1] proporciona datos sobre la calidad del aire obtenidos en ocho estaciones legales de monitorización divididas en dos redes de ámbitos diferentes, una Europea (EUROAIRNET, Red Europea de Control y

[1] `https://www.zaragoza.es/sede/portal/medioambiente/calidad-aire/red/estaciones`

vigilancia de la Calidad del aire ambiente) y la red nacional española[2].

Los datos de los contaminantes recogidos por las estaciones se envían en "tiempo real", cada hora, aunque debido a que están sujetos a un proceso continuo de validación, estos no son oficiales hasta que transcurran 9 meses desde que finaliza el año en que se tomaron.[3]

Nombre	Red	Contaminantes	Zona
El Picarral	EUROAIRNET	PM_{10}, NO_2, CO, H_2S, O_3	Urbana
Renovales	EUROAIRNET	PM_{10}, O_3, NO_2, SO_2, CO	Urbana fondo
Roger de Flor	EUROAIRNET	PM_{10}, O_3, NO_2, SO_2, CO	Urbana
Actur	Nacional	PM_{10}, O_3, NO_2, CO	Urbana
Avda. de Soria	Nacional	PM_{10}, O_3, NO_2, CO	Urbana
Centro	Nacional	SO_2, O_3, NO_2, CO	Urbana comercial residencial
Jaime Ferrán	Nacional	SO_2, O_3, NO_2, CO, H_2S	Suburbana
Las Fuentes	Nacional	PM_{10}, O_3, NO_2, SO_2, CO	Urbana tráfico

Cuadro 2.1: Estaciones reguladas del Ayuntamiento de Zaragoza

Figura 2.2: Distribución de las estaciones del Ayuntamiento de Zaragoza.

[2] https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/redes/

[3] D2011/850/UE: Decisión de Ejecución de la Comisión, de 12 de diciembre de 2011. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32011D0850

El proceso de carga de la información disponible, validada y no validada, en la Base de Datos, podía recogerse a partir de dos orígenes:

1. Servicio SPARQL end-point para la realización de consultas sobre datos almacenados en RDF del portal de Datos Abiertos del Ayuntamiento de Zaragoza.[4]

2. Los metadatos históricos anuales validados de todas las estaciones españolas están disponibles en formato de valores separados por comas (CSV) en el Ministerio para la Transición Ecológica y el Reto Demográfico.[5]

```
PREFIX ssn:<http://purl.oclc.org/NET/ssnx/ssn#>
PREFIX ssnext:<http://vocab.linkeddata.es/datosabiertos/def/medio-ambiente/
      calidad-aire/ssn-ext#>
PREFIX dul:<http://www.loa-cnr.it/ontologies/DUL.owl#>
SELECT DISTINCT ?magnitud ?estacion ?fecha ?valor ?verif

WHERE {
      ?uri a ssn:Observation;
      ssn:observedBy ?estacion;
      ssn:observationResult ?res;
      ssn:observedProperty ?magnitud;
      ssnext:observationStatus ?verif;
      ssn:observationResultTime ?fecha.
      ?res ssn:hasValue/dul:hasDataValue ?valor
      FILTER  (?fecha>="2020-07-24"^^xsd:date and ?fecha<="2020-07-25"^^xsd:
     date and regex(?uri, "horaria"))
}
ORDER BY ?fecha ?estacion ?magnitud LIMIT 10
```

Código 2.1: Ejemplo de consulta en el punto de acceso SPARQL del Ayuntamiento.

magnitud	estacion	fecha	valor
http://es.dbpedia.org/resource/%3%3xidos_e_itr%3%3geno	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/26	2020-07-23T00:00:00	32.43
http://es.dbpedia.org/resource/Di%3%3xido_e_itr%3%3geno	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/26	2020-07-23T00:00:00	22.85
http://es.dbpedia.org/resource/Mon%3%3xido_e_arbono	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/26	2020-07-23T00:00:00	0.17
http://es.dbpedia.org/resource/Ozono	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/26	2020-07-23T00:00:00	57.5
http://es.dbpedia.org/resource/Sulfuro_e_idr%3%3geno	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/26	2020-07-23T00:00:00	33.5
http://es.dbpedia.org/resource/%3%3xidos_e_itr%3%3geno	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/29	2020-07-23T00:00:00	21.72
http://es.dbpedia.org/resource/Di%3%3xido_e_zufre	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/29	2020-07-23T00:00:00	4.11
http://es.dbpedia.org/resource/Di%3%3xido_e_itr%3%3geno	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/29	2020-07-23T00:00:00	19.1
http://es.dbpedia.org/resource/Mon%3%3xido_e_arbono	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/29	2020-07-23T00:00:00	0.23
http://es.dbpedia.org/resource/Ozono	http://www.zaragoza.es/api/recurso/medio-ambiente/calidad-aire/estacion/29	2020-07-23T00:00:00	55.05

Cuadro 2.2: Resultado consulta punto de acceso SPARQL del Ayuntamiento.

Mediante la implementación de servicios periódicos que actualizan la información existente, se van poblando las tablas de la base de datos de información sobre estaciones reguladas validadas y no validadas. Aunque el Ayuntamiento de Zaragoza sí que ofrece algunos datos verificados, o al menos preliminarmente en el punto SPARQL, los datos de más de un año de antigüedad dejan de estar accesibles. Por tanto, es necesario obtener y almacenar los datos de los dos orígenes diferentes como respaldo.

[4] `https://www.zaragoza.es/sede/portal/datos-abiertos/servicio/sparql`

[5] `https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/datos/`

Dado que el Portal de Datos Abiertos de Zaragoza carga con periodicidad la información necesaria, estos tienen que recogerse en consonancia con la fecha última de los datos ya presentes en la base de datos de TRAFAIR. El proceso de carga se puede observar en la Figura 2.3; en primer lugar se realiza una consulta a la base de datos de TRAFAIR para saber cuándo es la última fecha de la que tenemos datos. Mediante el uso de diversas librerías de Python (PyGreSQL [25] y SPARQLWrapper [26]), que permiten establecer la conexión, se comunica al punto SPARQL del Ayuntamiento una fecha determinada y la del momento de la petición como límites, para posteriormente recuperarse y guardarse los nuevos datos de calidad del aire que no estén ya presentes.

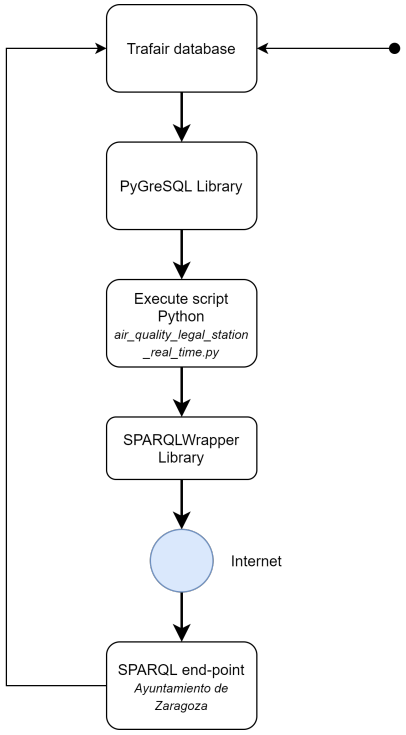

Figura 2.3: Proceso de carga de datos procedentes de las estaciones.

Este proceso ETL[6] de carga y almacenamiento se apoya en la definición de una base de datos relacional (sobre un sistema de gestión de bases de datos PostgreSQL), cuyo modelo lógico se presenta en la Figura 2.4 mediante un diagrama UML. En concreto, hay una entidad principal (tabla *aq_legal_station*) que contiene los atributos identificativos y las geometrías

[6]Procedimiento de construcción de un almacén de datos: los datos se recuperan (*Extract*), se convierten (*Transform*) en un formato que puede ser analizado, y se almacenan (*Load).*

de las 8 estaciones reguladas y sobre la que pivotan una serie de entidades (tablas) asociadas con los valores de los contaminantes y partículas recogidas, tanto para aquellas validadas como las que no han sido validadas.

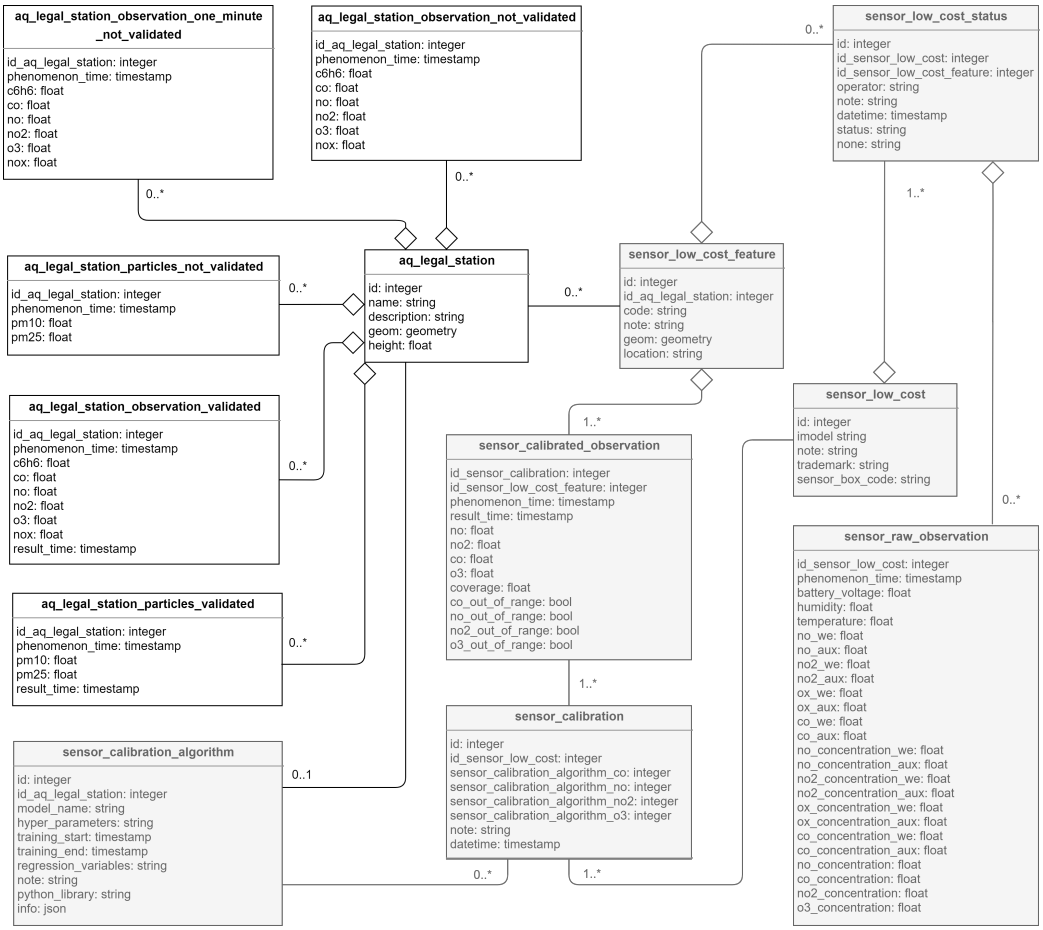

Figura 2.4: Modelo UML de las estaciones reguladas en la base de datos de Trafair.

2.2. Monitorización desde sensores low-cost

Con respecto al seguimiento alternativo de los datos de calidad del aire, los miembros de TRAFAIR se han encargado de desplegar una red de sensores de bajo coste *(low-cost)* encargados de recoger datos de contaminación en la ciudad de Zaragoza, como se puede observar en la Figura 2.5. Estos dispositivos generan los datos brutos que posteriormente, mediante una serie de algoritmos de calibración, resultan en concentraciones de gas de los cuatro

contaminantes recolectados, CO, NO, NO_2 y O_3.

En concreto, para cada uno de los sensores se obtienen dos valores de voltaje (mV), correspondientes a dos electrodos diferentes, un *working electrodo* (we) y otro *auxiliar* (aux) (Cuadro 2.3) que se utilizan para corregir desviaciones de las medidas. El fabricante aplica su propia calibración y ofrece valores de concentraciones (ppb) para estos dos electrodos. No obstante, el equipo de TRAFAIR genera un algoritmo de calibración para cada sensor y por lo tanto obtiene un valor de concentración para cada uno de los contaminantes. La información es recogida por cada sensor y mediante una red inalámbrica se transmiten los datos a la Base de Datos PostgreSQL de TRAFAIR, donde cada observación es almacenada en "tiempo real" con una periodicidad de un minuto.

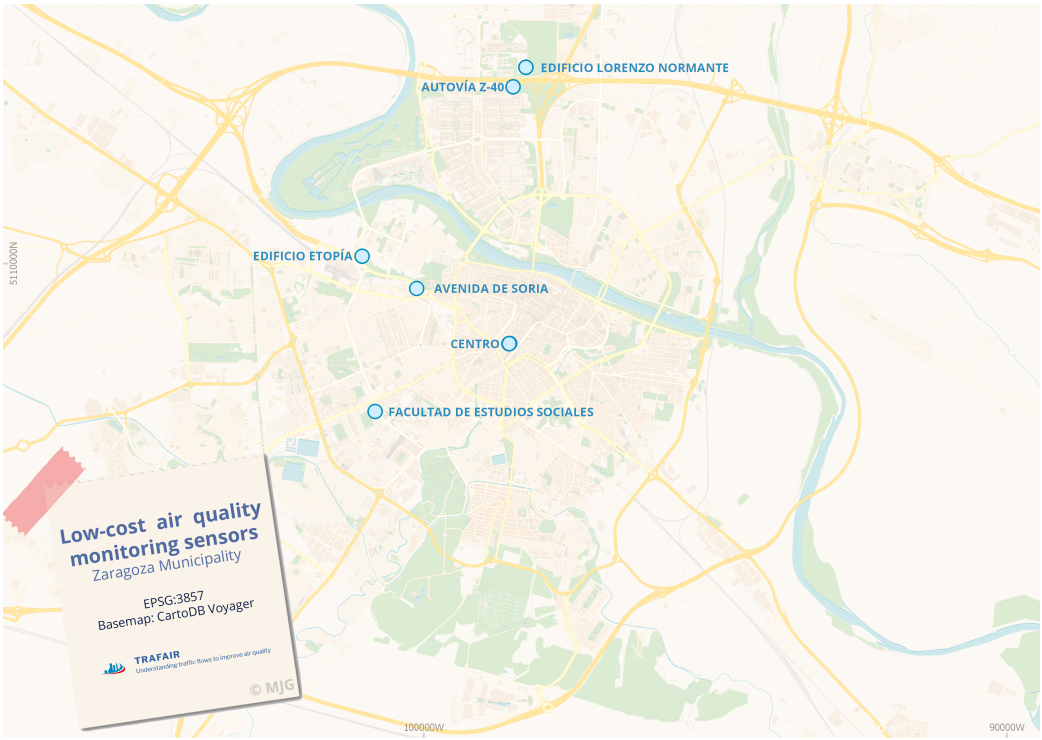

Figura 2.5: Distribución de los sensores low-cost.

Sensor	Información	Zona
Avda. de Soria	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana
Autovía Z-40	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana tráfico
Centro	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana comercial residencial
Edificio Etiopía	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana
Edificio Lorenzo Normante	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana
Facultad de Estudios Sociales	Temperatura, Humedad, Voltaje, NO_2, CO, NO_2 O_3	Urbana

Cuadro 2.3: Sensores low-cost del proyecto TRAFAIR.

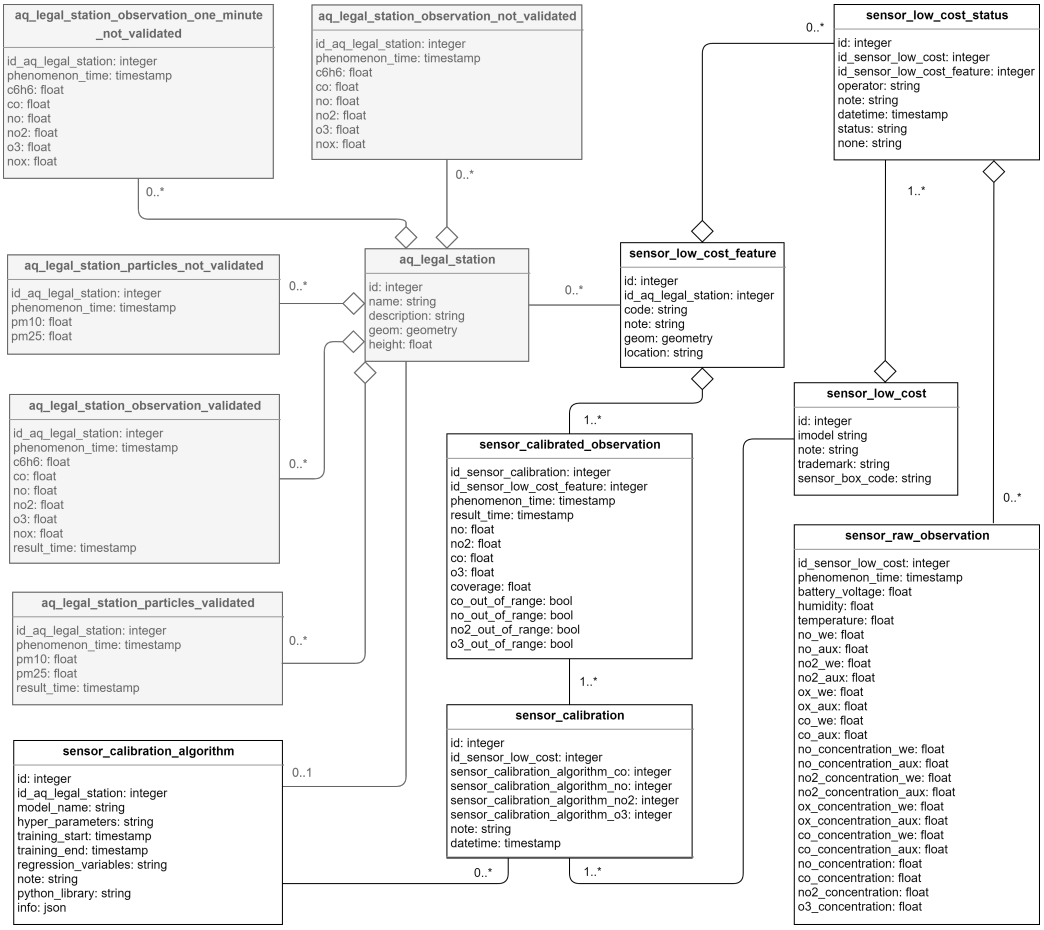

Figura 2.6: Modelo UML de los sensores low-cost en la base de datos de Trafair.

Al visualizar el diagrama UML relativo a los sensores de bajo coste en la Figura 2.6, se puede ver como existe una relación entre las estaciones reguladas (*aq_legal_station*) y los sensores (*sensor_low_cost_feature*), a través del [id] del primero. Esto es necesario para asignar los algoritmos de calibración de la información de los sensores, a una de las estaciones existentes, permitiendo transformar los datos brutos recogidos (*sensor_raw_observation*), en

datos calibrados (*sensor_ calibrated_ observation*). Estos últimos son compatibles con los almacenados para cada una de las estaciones reguladas cada 10 minutos (validados y no validados), facilitando que se puedan realizar operaciones posteriormente mediante la combinación de ambos conjuntos de información.

En definitiva, a partir de los distintos tipos de fuentes, se extraen los siguientes datos de calidad del aire de la ciudad de Zaragoza que pueblan la base de datos de TRAFAIR:

- Ayuntamiento de Zaragoza
  Estaciones reguladas (no validados-validados preliminarmente):
  - Datos no validados de CO, NO, NO_2, O_3 y PM_{10}.[7]
  - Extensión temporal máxima de un año y resolución temporal horaria.
  - Proceso: Recolección de datos horaria y carga en base de datos.
- Ministerio de Transición Ecológica
  Estaciones reguladas (validados):
  - Datos no validados de CO, NO, NO_2, O_3 y PM_{10}
  - Disponibilidad temporal, 9 meses tras la finalización del año. Resolución temporal horaria.
  - Proceso: Recolección JSON o CSV tras publicación de datos históricos y carga en base de datos.
- TRAFAIR
  Sensores low-cost (sin calibrar):
  - Datos no calibrados de CO, NO, NO_2 y O_3.
  - Disponibilidad temporal desde la puesta en marcha del sensor. Resolución temporal en "tiempo real".
  - Proceso: Recolección continua de datos y carga en base de datos.

[7] Partículas de menos de 10 μm de diámetro.
`https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/emisiones/prob-amb/particulas.aspx`

3. Definición y modelado de los conjuntos de datos accesibles como datos abiertos

Dado que la publicación de la información de calidad del aire como datos abiertos es uno de los principales objetivos del proyecto TRAFAIR, la definición y modelado de los conjuntos de datos accesibles como datos abiertos es un paso imprescindible para lograrlo. En ese sentido, la base de datos relacional que la sustenta requiere un esquema lógico y un conjunto de funciones, que construyan las tablas-vistas necesarias para la publicación de los datos abiertos. La colección de conjuntos de datos de TRAFAIR se compone tanto de fenómenos discretos (*Feature Types*) como de coberturas (*Coverages*).

A la definición y publicación de los datos de calidad del aire ya calibrados, se le suman informaciones recogidas por los sensores, como las medidas obtenidas cerca de las ubicaciones de referencia con fines de calibración, o los datos de voltaje y temperatura. También, con el objetivo de permitir la reutilización de los datos por la comunidad científica, se publican los datos brutos. Los datos calibrados de los sensores de calidad del aire se publican en todas las ciudades adscritas al proyecto, incluidos los datos primitivos calibrados con una resolución temporal de diez minutos y los datos agregados con una resolución temporal de una hora. Las coberturas de la calidad del aire en tiempo real se actualizan cada 10 minutos en cada ciudad. Además, los conjuntos agregados de coberturas históricas de una hora también son cargados en la base de datos. Por último, en cada ciudad se construyen las últimas coberturas espacio-temporales y los resultados históricos del modelo de dispersión de la contaminación atmosférica que se utiliza para la predicción de la calidad del aire.

3.1. Modelos para los fenómenos discretos

Los fenómenos discretos (en inglés *Feature Types*), representan datos temáticos o categorizados, discretizados espacialmente dado que presentan unos límites definidos. Para el caso de TRAFAIR, en lo que respecta a calidad del aire, estos hacen referencia a datos presentes en la base de datos que han sido recogidos en un punto concreto de la superficie terrestre (sensor, estaciones), y que se guardan en forma de tablas y vistas con información espacial. Del mismo modo, también hacen referencia a los servicios asociados (WFS) con los que se publica cada una de las entidades.

El software Geoserver permite definir los *Feature Types* tomando como fuente de datos las tablas y vistas disponibles en una base de datos. Para establecer los atributos deseados de los *Feature Types* ha sido necesario definir varios scripts SQL para la creación de tablas y vistas, así como disparadores y funciones que faciliten su creación y rellenado. En concreto, a través de `airQualitySensors_structuresandtriggers.sql` se crean las tablas de observaciones en tiempo real, horarias y con los metadatos de calibración. En el fragmento de Código 3.1 se puede ver un extracto de las tablas y vistas temporales que se deben borrar inicialmente, si existen, previa creación de las mismas en el desarrollo de las diferentes funciones.

```
1  ----------------------------------------
2  -- AIR QUALITY CALIBRATED SENSOR DATA --
3  ----------------------------------------
4
5  drop trigger if exists after_air_quality_observation on
       public.sensor_calibrated_observation;
6  drop function if exists open_data.
       update_real_time_air_quality_observation();
7  drop table if exists open_data.
       hourly_air_quality_observations;
8  drop view if exists open_data.
       real_time_hourly_air_quality_observations;
9  drop table if exists open_data.
       air_quality_observations_last_hour;
10 drop table if exists open_data.
       real_time_air_quality_observations;
11 drop view if exists open_data.
       air_quality_observation_provenance_metadata;
12 ...
```

Código 3.1: Extracto script SQL airQualitySensors_structuresandtriggers.

También existe un script denominado `airQualitySensors_historicmanagement.sql` para la creación de las funciones que generan las vistas históricas en donde se almacenan las observaciones, tal y como se pueden ver en el

extracto del Código 3.2 en el que se enumeran las distintas funciones creadas en la base de datos.

```
1  ----------------------------------------
2  -------- AIR QUALITY SENSOR DATA --------
3  ----------------------------------------
4
5  create or replace function open_data.
       delete_air_quality_sensor_open_data(start_time
       timestamp, end_time timestamp) ...
6  create or replace function open_data.
       insert_air_quality_sensor_open_data(start_time
       timestamp, end_time timestamp) ...
7  create or replace function open_data.
       delete_raw_air_quality_sensor_open_data(start_time
       timestamp, end_time timestamp) ...
8  create or replace function open_data.
       insert_raw_air_quality_sensor_open_data(start_time
       timestamp, end_time timestamp) ...
9  create or replace function open_data.
       delete_sensor_calibration_view(sensor_status integer)
        ...
10 create or replace function insert_sensor_calibration_view
       (sensor_status integer) ...
```

Código 3.2: Extracto script SQL airQualitySensors_historicmanagement.

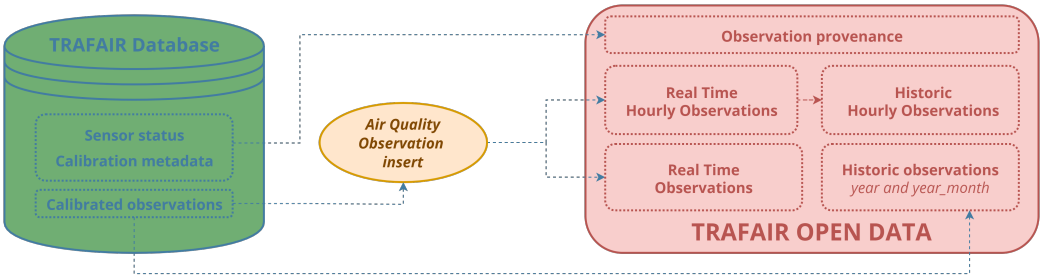

Figura 3.1: Esquema de creación de fenómenos discretos.

Como se ilustra en la Figura 3.1, cada vez que se genera una nueva observación calibrada, se inserta en la base de datos una observación en tiempo real en *real_time_air_quality_observations*. Además, si el instante de tiempo es todavía de la última hora, entonces se inserta en una tabla temporal (vista) para la última hora a partir de (*real_time_hourly_air_quality_observations*). De lo contrario, si es ya una nueva hora, primero se agregan los últimos datos disponibles en la vista y se insertan los valores promediados horarios en la tabla *hourly_air_quality_observations* donde se almacenarán permanentemente, luego se borra la tabla temporal con los datos de la última hora a la espera de siguientes observaciones. Si el instante de tiempo corresponde a un nuevo año, se crea una nueva vista con observaciones calibradas para un año

especifico (*historic_air_quality_observations_year*). También existen funciones concretas para construir los conjuntos calibrados históricos mensuales (*historic_air_quality_sensor_raw_observations_year_month*).

Con respecto a la vista *air_quality_observation_provenance_metadata*, esta hace referencia a los metadatos de los sensores de bajo costo y de los algoritmos de calibración utilizados. Cada inserción describe un proceso completo, que incluye un sensor y el algoritmo de calibración que se utilizó para generar las concentraciones de contaminante observadas.

Name	Title	Description
air_quality_observation_provenance_metadata	Air quality observation metadata	This dataset provides provenance metadata of each calibration model that has been used to generate calibrated air quality.
real_time_air_quality_observations	Real time air quality observation	This dataset provides real time air quality observation data, for all the pollutants, at the location of each sensor.
real_time_hourly_air_quality_observations	Real time hourly air quality observation	This dataset provides average values of the last hour.
historic_air_quality_observations_year	Historic air quality observations by year	This dataset provides primitive calibrated air quality observation data for every year.
historic_air_quality_sensor_raw_observations_year_month	Historic air quality sensor raw observations by year and month	This dataset provides primitive calibrated air quality observation data for every year and month.
hourly_air_quality_observations	Hourly air quality observations	This dataset provides hourly average air quality concentration data.

Cuadro 3.1: Título y descripción humana de los fenómenos discretos.

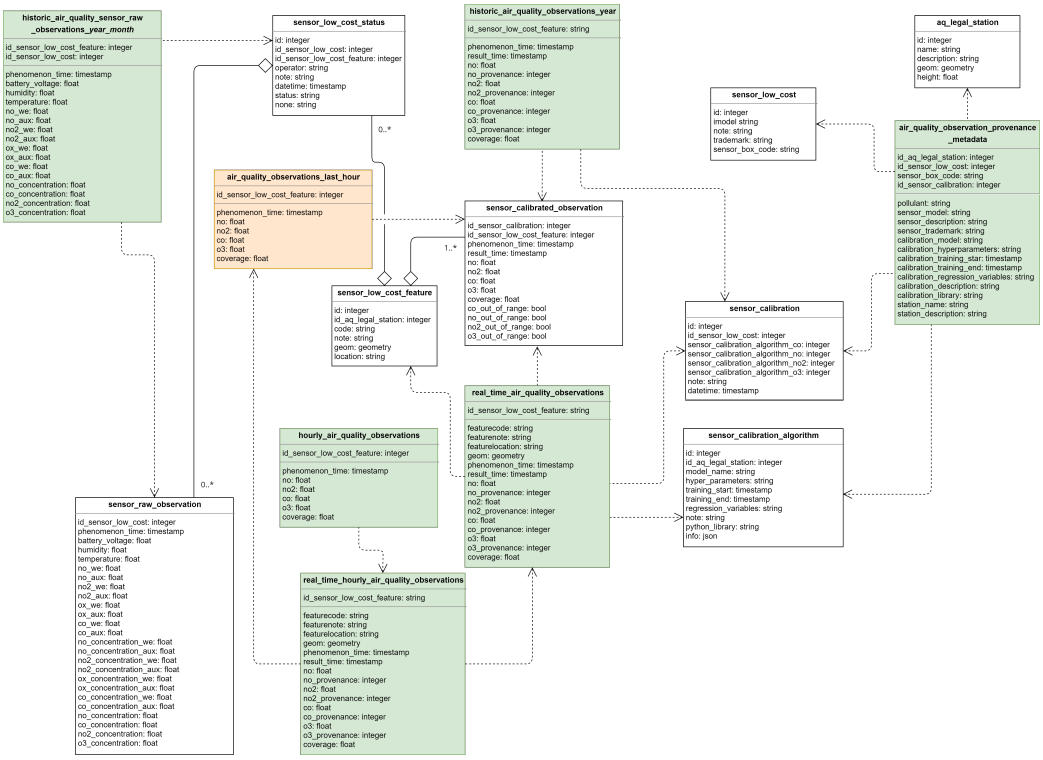

Figura 3.2: Modelo UML de los fenómenos discretos en la base de datos de Trafair.

En el Cuadro 3.1 se resumen los *Feature Types* generados, su descripción

humana y título. Del mismo modo, se describe su origen en el diagrama de clases UML que se muestra en la Figura 3.2. En este diagrama las clases representan tres tipos de tablas/vistas de la base de datos: las tablas originales de la base de datos (en color blanco), las vistas auxiliares (en color naranja), y las tablas/vistas (en color verde) que se publican como *Feature Types*. Además se muestran relaciones de dependencia entre las tablas/vistas derivadas y las tablas originales.

3.2. Modelos para las coberturas

Los conjuntos de datos de coberturas espaciales son construidos a partir de los mapas interpolados de calidad del aire (monitorización). También se generan coberturas mediante los productos de predicción generados con el modelo de dispersión de Lagrange, GRAL, de la Universidad Tecnológica de Graz,[1] pero este segundo tipo de coberturas no es objeto principal del TFM.

Una cobertura (o *coverage* en inglés) es un conjunto de datos ráster que puede ser almacenado en un servidor de mapas, en concreto, en el proyecto TRAFAIR la infraestructura utilizada es Geoserver, que almacena los archivos GeoTIFF[2] generados, y también, los mosaicos de imágenes,[3] de las diferentes coberturas espacio-temporales.

En la Figura 3.3 se expone el procedimiento de creación de las coberturas. Cada 10 minutos se produce una nueva interpolación ráster de los datos promediados de monitorización de calidad del aire obtenidos por los sensores (observaciones calibradas) y por las estaciones reguladas (datos no validados) para el período introducido, que genera una cobertura en formato GeoTIFF: *real_time_air_quality_observations_coverage*. Las coberturas en tiempo real reemplazan constantemente a la inmediatamente anterior. También existen un mosaico de coberturas horarias y otro de diarias. Para la primera se almacena el promedio de las observaciones para la última hora (*hourly_air_quality_observation_coverage_date*), y la segunda (*daily_air_quality_observation_coverage_year_month*) recoge los valores agregados máximo, mínimo y promedio del día, guardándolos también en un mosaico de manera mensual .

[1] GRAL Dispersion Model. `https://github.com/GralDispersionModel/GRAL`

[2] Estándar de metadatos OGC para información georreferenciada en archivos de imagen TIFF. `https://www.ogc.org/standards/geotiff`

[3] Almacenamiento conjunto de modelos ráster. `https://docs.geoserver.org/stable/en/user/data/raster/imagemosaic/`

Con respecto a los datos producto de GRAL, se generan un conjunto de datos interpolado de cobertura espacio-temporal con la predicción de las próximas 48 horas (*latest_air_quality_prediction_coverage*), que va sobrescribiéndose, y un mosaico de coberturas de predicción histórica con la última generada, así como las 7 anteriores (*air_quality_prediction_coverage_date*).

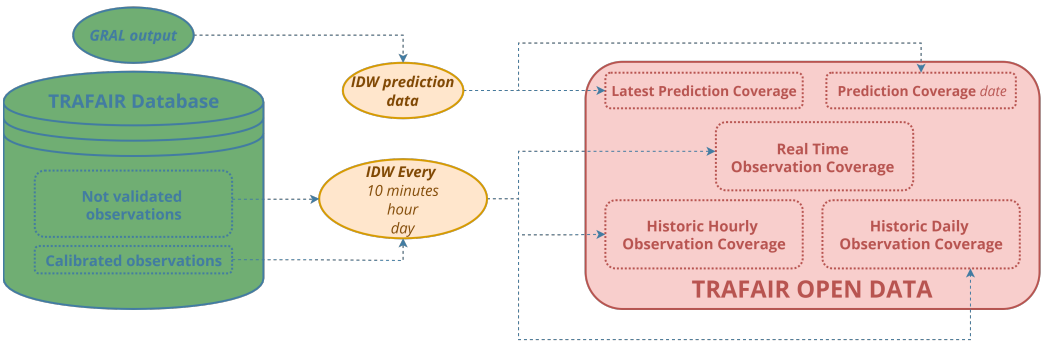

Figura 3.3: Esquema de creación de coberturas de monitorización.

En lo que respecta a las coberturas de observación de calidad del aire, estás se generan mediante la interpolación de los datos obtenidos por las estaciones y sensores. En general, los procesos de interpolación consisten en la estimación de los valores que alcanza una variable "V" en un conjunto de puntos definidos por un par de coordenadas (x, y), partiendo de los valores medidos en una muestra de puntos situados en un área de estudio. Obteniendo con ello valores predichos de atributos para los que en realidad no existe tal información.

En concreto, en el seno del proyecto TRAFAIR, se decidió optar por el interpolador Inverso a la Distancia, IDW (*Inverse Distance Weighting*), con un factor de ponderación $p = 2$, para la obtención de los mapas históricos y en tiempo real. Esta decisión vino motivada por las observaciones y evaluación de los resultados preliminares respecto de cada contaminante y en función del valor de RMSE (*Root Mean Square Error*) mediante la técnica de dejar un punto excluido, utilizando para ello $n-1$ puntos para evaluar la superficie de interpolación y luego calcular el estadístico de error RMSE entre el valor predicho en la ubicación del punto excluido y los valores medidos en el mismo punto. Hay que tener en cuenta el hecho de que los diferentes rendimientos para cada contaminante, y método de interpolación, son sensibles a los mismos patrones espaciales de concentración; pues para un mismo ámbito solo pueden ser sensibles a las diferentes pautas de emisión, dado que todos los contaminantes se difunden siguiendo las mismas condiciones meteorológicas.

Los interpoladores propuestos fueron los siguientes:

- OK: Método ordinario de Kriging.
- IDW2: Método de ponderación de distancia inversa con p=2.
- IDW1: Método de ponderación de distancia inversa con p=1.
- OAK: Kriging automático ordinario (el variograma se evalúa automáticamente a partir de los datos).
- NN-3: Vecino más próximo con 3 elementos vecinos.
- NN-6: Vecino más próximo con 6 elementos vecinos.
- SPLINE: Regresión por el método splines.
- AVG: Método promediado.
- IDW2-OK: Combinación de ponderación de distancia inversa con p=2 y Kriging ordinario.

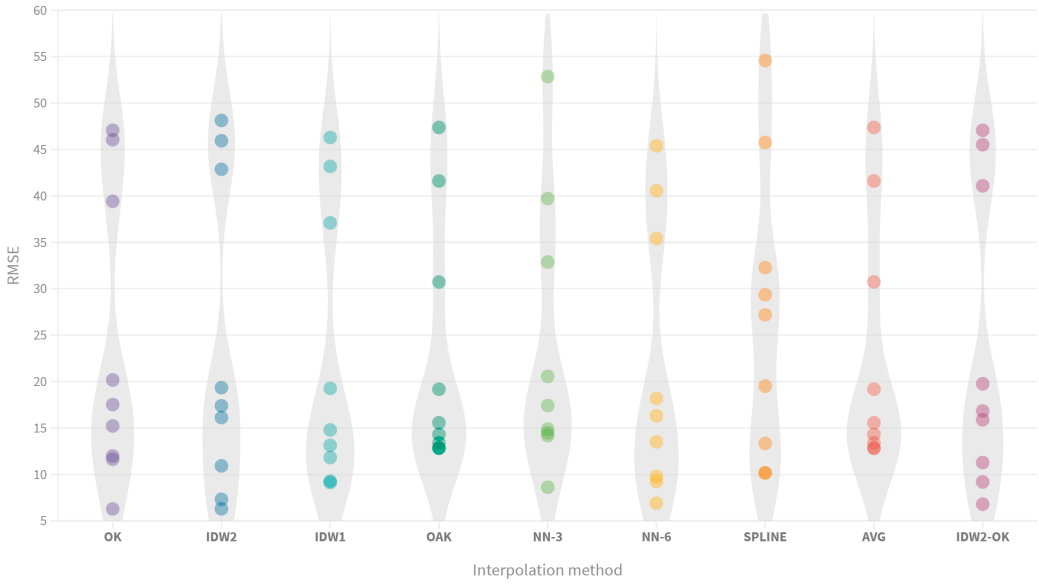

Figura 3.4: Distribución del RMSE para cada método, en 9 puntos.

Finalmente fue elegido IDW2 por la combinación de su menor RMSE (se puede ver en la Figura 3.4 a partir de los datos del Cuadro A.1 anexado) en el promedio de diferentes días, y fundamentalmente, en sus mejores resultados en diversos tipos de contaminantes (O_3, NO_2 y CO). También por el hecho

de preferirse una superficie continua y lisa (IDW1, IDW2, OK), que de coherencia al fenómeno a representar, frente a una superficie segmentada (NN3 o NN6) como se puede ver en la Figura A.1. El método AVG simplemente se utiliza como métrica de comparación debido a que es el resultado de promediar los valores predichos intermedios en función de los puntos muestreados presentes.

El funcionamiento del interpolador elegido es el siguiente, el promedio de todas las ubicaciones no observadas utilizando valores de ubicaciones cercanas producen valores que son proporcionales a la proximidad de los puntos, al mismo tiempo que puede especificarse mediante el coeficiente de potencia del IDW un factor de ponderación, permitiendo modificar el peso de los puntos cercanos. Si este es muy grande, los valores cercanos ejercen una influencia desmesurada sobre los puntos no muestreados, dando como lugar una distribución poligonal. Si por el contrario es muy pequeño, la mayor parte de los puntos tendrán valores similares, promediando en función del radio de búsqueda. Por ese motivo, habitualmente se utilizan valores p de entre 1 y 4, en concreto para el proyecto TRAFAIR, es $p = 2$.

Name	Title	Description
real_time_air_quality_observations_coverage	Real time air quality observations coverage	This dataset provides a spatial coverage update every 10 minutes to contain a later estimation of hte concentration of each pollulant in each point of the city
hourly_air_quality_observation_coverage_date	Hourly air quality observations coverage	This dataset provides the hourly evolution during the day of the estimation of the observed concentration of each pollutant.
daily_air_quality_observation_coverage_year_month	Daily air quality observations coverage	The dataset provides the daily evolution of the estimation of the maximum, minimum and average concentrations of each pollulant.
latest_air_quality_prediction_coverage	Latest air quality prediction coverage	This dataset provides the latest air quality prediction generated by GRAL model, one time instant for each of the following 48 hours.
air_quality_prediction_coverage_date	Air quality prediction coverage by date	This dataset provides the last 8 prediction coverages, the very last one and also the previous 7.

Cuadro 3.2: Título y descripción humana de las coberturas.

```
#Database parameters and queries
dsn = "PG:dbname='db' host='ip' port='xxxx' user='user'
    password='pass'"

querystring = paste("select f.geom, ",aggFun,"(o.co) as
    co, ",aggFun,"(o.no) as no, ",aggFun,"(o.no2) as no2,
     ",aggFun,"(o.o3) as o3
                    from sensor_calibrated_observation o,
     sensor_low_cost_feature f
                    where o.id_sensor_low_cost_feature=f.
    id
                    and phenomenon_time >= '",startstamp,
    "'
                    and phenomenon_time < '",stamp,"'
                    group by f.id
                    union all
                    select f.geom, ",aggFun,"(o.co) as co
    , ",aggFun,"(o.no) as no, ",aggFun,"(o.no2) as no2, "
    ,aggFun,"(o.o3) as o3
```

```
12                      from aq_legal_station_observation_not
    _validated o, aq_legal_station f
13                      where o.id_aq_legal_station=f.id
14                      and phenomenon_time >= '",startstamp,
    "'
15                      and phenomenon_time < '",stamp,"'
16                      group by f.id"
17                      , sep="")
```

Código 3.3: Extracto script R de interpolación de las coberturas, consulta SQL.

```
1  ...
2
3  if (nrow(data)>0) {
4    data = sf::as_Spatial(data)
5
6    #Zaragoza
7    crsutm = crs('+proj=utm +zone=30 +ellps=WGS84 +towgs84
       =0,0,0,0,0,0,0 +units=m +no_defs')
8    datautm = spTransform(data,crsutm)
9
10   #Create raster for interpolation
11   utmRasterExtent=extent(spTransform(boxPoints, crsutm))
12   rasterout = raster(utmRasterExtent)
13   res(rasterout) = 4
14   crs(rasterout)=crsutm
15   gridraster <- as(rasterout, 'SpatialGrid')
16
17   #interpolate co
18   modelco = gstat(formula = co ~ 1, # intercept only
       model
19                   data = datautm[!is.na(datautm$co),],
20                   set = list(idp = 2))
21   idwco = predict(modelco, newdata=gridraster)
22   rasterco=raster(idwco)
23
24  ...
25
26   # Create multibandraster
27   multibandraster = brick(rasterco,rasterno2,rastero3,
       rasterno)
28   multibandraster = projectRaster(multibandraster, crs=
       crdref)
29
30   # Export raster
31   writeRaster(multibandraster, paste("output/", fileName,
       ".tif",sep=""), overwrite=TRUE, format="GTiff",
       options="COMPRESS=LZW")
32 }
```

Código 3.4: Extracto script R de interpolación de las coberturas, multibanda.

En el Cuadro 3.2 se describen las coberturas. Estas son generadas a partir de un script en R que se conecta a la base de datos para solicitar todos los conjuntos de datos disponibles que permitan realizar la interpolación. Este proceso se puede observar en el extracto de `interpolateaggregate.R` (Código 3.3) que muestra la consulta SQL necesaria para solicitar aquellas

observaciones calibradas y no calibradas (líneas 5 y 12), que permitan combinar la información de los sensores y estaciones legales que se corresponden con un intervalo temporal preestablecido; esto posteriormente se guarda en un *dataframe*[4] con atributos espaciales (línea 4 del Código 3.4), que nutrirán a los archivos ráster con la información ya interpolada mediante el método IDW para cada uno de los contaminantes. Se puede ver un ejemplo para el CO a partir de la línea 17, posteriormente se combinan mediante la generación de un multibanda (línea 27) y se exporta en formato GeoTIFF (línea 31).

Estos archivos ráster que contienen la información interpolada de los cuatro contaminantes son la base de la cartografía publicada. En el Anexo E se pueden ver ejemplos más detallados de la cartografía generada por el método de interpolación elegido y las observaciones recogidas en el proyecto.

[4] `https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Data-frames`

4. Publicación de datos abiertos

Como se ha visto en los capítulos previos, la publicación de los conjuntos de datos accesibles como datos abiertos representa el flujo de trabajo necesario para la puesta a disposición del público general de aquellos conjuntos de datos que han sido definidos a tal efecto. Por ello, en el presente capítulo se tratará de describir el flujo necesario para construir y publicar en el servidor de mapas tanto los fenómenos discretos (existentes en la base de datos) como las coberturas, haciendo uso de la API Rest de Geoserver.[1] Del mismo modo, se hará referencia a la publicación en portal de datos abiertos tipo CKAN de los metadatos y diversos formatos de distribución de datos (CSV, RDF y NetCDF).

Todo el código fuente utilizado por los miembros del proyecto para las diferentes tareas esta disponible en un repositorio colaborativo en Gitlab.[2] Del mismo modo, los conjuntos de datos publicados, y sus actualizaciones (ver Anexo B), se encuentran disponibles en el portal de datos abiertos (también a través del EDP, Aragon Open Data, etc.) del proyecto TRAFAIR[3]

4.1. Publicación de las capas en Geoserver

4.1.1. Publicación de fenómenos discretos

Una vez almacenados en la base de datos los fenómenos discretos en tiempo real e históricos mediante los scripts SQL (ver Cuadro 4.1), hay que trasladarlos al servidor de mapas que se encargará de generar los servicios de

[1] `https://docs.geoserver.org/stable/en/user/rest/`
[2] A fecha del 27/07/2020 el repositorio es privado, pero esta previsto compartir públicamente parte del mismo. `https://gitlab.com/trafair`
[3] `http://atila.unizar.es:3394/organization/universidad-de-zaragoza`

publicación compatibles con las especificaciones OGC. Para ello es necesario seguir un flujo de trabajo determinado que aparece descrito en la Figura 4.1.

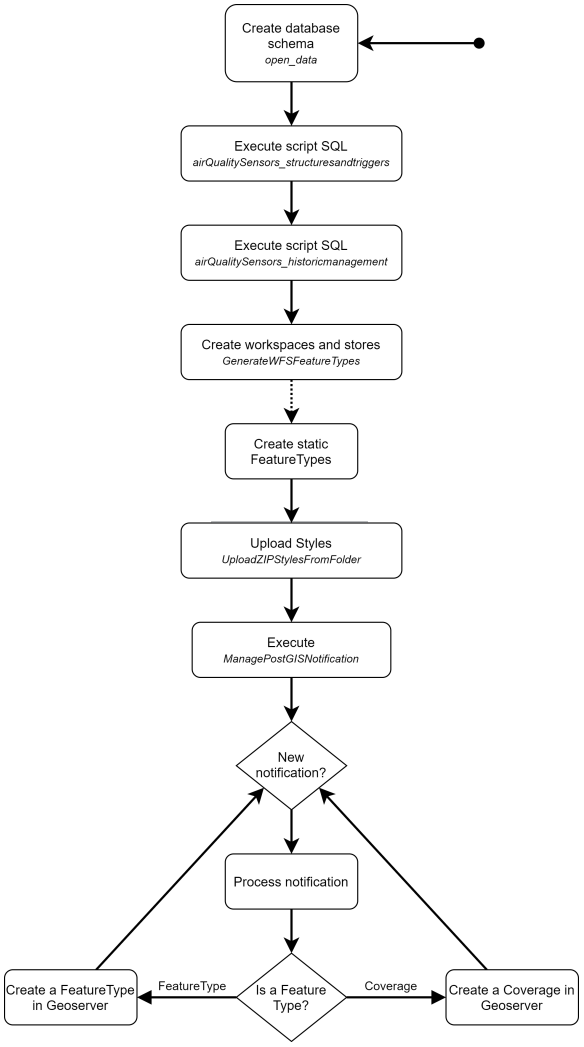

Figura 4.1: Flujo de trabajo para la creación de fenómenos discretos.

Subtype	Name	WFS	Database	SQL script
Real-Time	air_quality_observation_provenance_metadata	X	View	airQualitySensors_structuresandtriggers
Real-Time	real_time_air_quality_observations	X	Table	airQualitySensors_structuresandtriggers
Real-Time	real_time_hourly_air_quality_observations	X	View	airQualitySensors_structuresandtriggers
Historic	historic_air_quality_observations_year	X	View	airQualitySensors_historicmanagement, airQualitySensors_structuresandtriggers
Historic	historic_air_quality_sensor_raw_observations_year_month	X	View	airQualitySensors_historicmanagement, airQualitySensors_structuresandtriggers
Historic	hourly_air_quality_observations	X	Table	airQualitySensors_historicmanagement, airQualitySensors_structuresandtriggers

Cuadro 4.1: Servicios y script SQL origen de los fenómenos discretos.

Los componentes básicos que se relacionan son el propio Geoserver y la base de datos de TRAFAIR, mediante un proyecto Java denominado *GeoserverRestManagerMobile*, que contiene los archivos y clases necesarias para el proceso de publicación de los fenómenos discretos.

En primer lugar, han de establecerse en la base de datos los esquemas[4] necesarios. En concreto, para la publicación de los datos de calidad del aire es necesario el esquema propio `open_data`.

```
1 CREATE SCHEMA open_data
2     AUTHORIZATION write_user;
3     GRANT ALL ON SCHEMA open_data TO write_user;
4     GRANT USAGE ON SCHEMA open_data TO read_user;
```

Código 4.1: Consulta SQL para la creación de un nuevo esquema.

Debido a que de forma predeterminada no se asignan privilegios a ninguno de los usuarios, estos han de ser otorgados manualmente (líneas 3 y 4 del Código 4.1). Una vez se han establecido los esquemas básicos, se deben ejecutar los scripts que fueron vistos en el pasado capítulo. Ambos listados de instrucciones, `airQualitySensors_structuresandtriggers.sql` y `airQualitySensors_historicmanagement.sql` definen las tablas, vistas, funciones y disparadores[5] que permiten almacenar los fenómenos discretos en la base de datos.

En el momento en que la estructura de la base de datos esta correctamente configurada, es necesario replicar la estructura de almacenes y espacios de trabajo, y cargar los *Feature Types* estáticos en Geoserver, para ello se hace uso de la clase Java `GenerateWFSFeatureTypes`. En esencia, contiene cuatro clases privadas que permiten crear los contenedores de capas, o espacios de trabajo[6] básicos que almacenarán las distintas entidades adscritas, mediante `createDefaultWorkspace`. Si fuera necesario, se pueden limpiar los espacios preexistentes con `deleteOtherWorspaces` y así evitar problemas futuros, a través de la clase `createPostgisDataStore` se diseña el almacén de datos[7] que establecerá la conexión con el esquema en la base de datos de TRAFAIR.

[4]En inglés *schema*, contenedores de tablas y vistas dentro de la base de datos.
`https://www.postgresql.org/docs/11/ddl-schemas.html`
[5]En inglés *triggers*, funciones que se activan cuando ocurre otro evento.
`https://www.postgresql.org/docs/11/plpgsql-trigger.html`
[6]En inglés *workspace*, contenedores de entidades de características similares.
`https://docs.geoserver.org/latest/en/user/data/webadmin/workspaces.html`
[7]En inglés *datastore*, almacén de datos.
`https://docs.geoserver.org/stable/en/user/rest/api/datastores.html`

Y por último, `createStaticFeatureTypes` genera las *Feature Types* estáticas a partir de la información en base de datos y las almacena en el espacio de trabajo, todo ello se puede ver resumido en el Código 4.2

```
...
public class GenerateWFSFeatureTypes {

    private static final String TRAFFIC_TRAFAIR_STYLE = "traffic_trafair";
    private static final String OPEN_DATA = "open_data";
    private static final String WEBAPP = "webapp";
    private static final String MOBILE = "mobile";
    private static final String TRAFAIR = "trafair";
    private static final String PUBLIC = "public";

    public static void main(String[] args) throws IOException {
        createDefaultWorkspace();
        deleteOtherWorspaces();
        createPostgisDataStore();
        createStaticFeatureTypes();
    }

...

    private static void createPostgisDataStore() throws IOException {
        DataStoreService dataStoreService = new DataStoreService();

        PostgisConfig postgisConfig = GeoserverUtils.getPostgisConfig();
        String workspace = OPEN_DATA;
        postgisConfig.setSchema(OPEN_DATA);
        String store = OPEN_DATA;
        String result = dataStoreService.createPostgisDataStore(workspace, store, postgisConfig);
        System.out.println("Created datastore: " + result);
        );

    private static void createStaticFeatureTypes() throws IOException {
        FeatureTypeService featureTypeService = new FeatureTypeService();

        List<String> features = Arrays.asList(
                "air_quality_observation_provenance_metadata", "real_time_air_quality_observations",
                "real_time_hourly_air_quality_observations", "hourly_air_quality_observations"
        );
...
```

Código 4.2: Extractos del código de la clase Java GenerateWFSFeatureType.

Antes, o después, de cargar los *Feature Types* y sus conexiones con la base de datos, se pueden almacenar los estilos personalizados que controlarán la apariencia de los datos geoespaciales representados mediante servicios WMS. Estos están en formato .sld, que es un estándar OGC,[8] y permiten establecer escalas de color diferenciadas en función de las unidades y necesidades de representación en el proyecto. En concreto en TRAFAIR, mediante la clase Java `UploadZIPStylesFromFolder`, se cargan los ficheros comprimidos que contienen los 9 estilos diferentes para calidad del aire al espacio de trabajo en Geoserver, para que posteriormente puedan ser asignados a las diferentes capas. Todos se pueden ver en la Figura 4.2: cuatro de acuerdo a las normas de la EEA (European Environment Agency) para visualizar los contaminantes (CO, NO, NO_2 y O_3); otros cuatro estilos personalizados para TRAFAIR; y un último estilo utilizado para las salidas de predicción con los umbrales de NO_x.

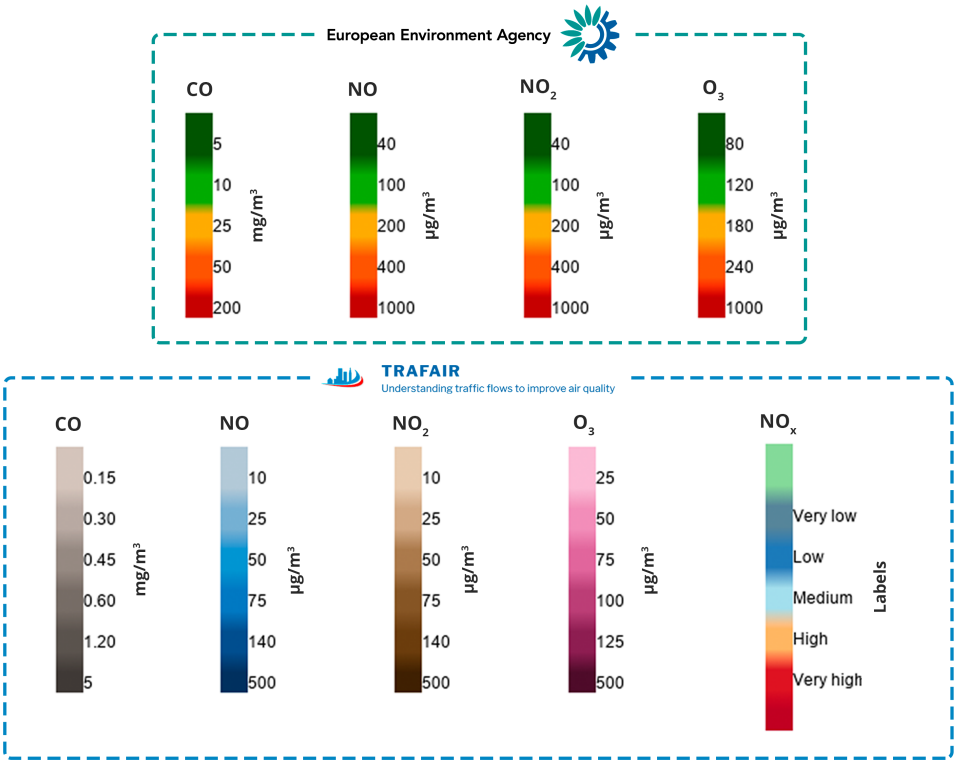

Figura 4.2: Estilos cargados en el Geoserver de TRAFAIR.

[8] https://www.ogc.org/standards/sld

El último paso se corresponde con la generación de las vistas que dan lugar a los *Feature Types* históricos. Mediante una ejecución manual de una de las funciones presentes en el script `airQualitySensors_historicmanagement.sql` en concreto `insert_raw_air_quality_sensor_open_data` (ver Código 4.3). Esta función va creando las vistas mensuales en la base de datos, conforme a un período introducido (`(start_time timestamp, end_time timestamp)`) de carácter mensual, que van rellenándose con información de las tablas con los datos de observación sin calibrar y el estado de los sensores. Posteriormente, se emite una notificación al código Java para que cree los *Feature Type historic_air_quality_sensor_raw_observations_year_month*, de tipo histórico, en el Geoserver, y que es captada por la clase Java que las gestiona (`ManagePostgisNotification`) que genera los servicios según corresponda.

```
create or replace function open_data.insert_raw_air_quality_sensor_open_data(start_time timestamp, end_time timestamp) returns void as $$
declare
  query text;
  ym record;
  payload text;
BEGIN
   -- generate views for historic_air_quality_sensor_raw_observations_year_month
    for ym in
      select distinct to_char(extract(year from phenomenon_time), 'FM0000') as y, to_char(extract(month from phenomenon_time),'FM00') as m
    from public.sensor_raw_observation obs, public.sensor_low_cost_status s
    where obs.id_sensor_low_cost_status=s.id
      -- and s.status = 'running'
        and phenomenon_time >= start_time and phenomenon_time < end_time
    order by y,m
...

   execute query;
    -- notify to register the WFS layer in geoserver
    payload := ''
        || '{'
        || '"featuretype":"'  || 'historic_air_quality_sensor_raw_observations_'||ym.y||'_'||ym.m   ||'",'
        || '"services":'        || '["WFS"]'
                                   ||','
        || '"operation":"'      || 'CREATE'
                                    ||'"'
        || '}';
    PERFORM pg_notify('db_notifications', payload);
...
```

Código 4.3: Extracto de la función SQL de creación de vistas históricas.

4.1.2. Publicación de coberturas

La publicación de las coberturas de monitorización es distinta a la de los fenómenos discretos, dado que el segundo proceso sí almacena sus salidas en la base de datos, o se construyen directamente en la misma, mientras que las coberturas han de ser generadas externamente a partir de información de la base de datos, y después, cargadas en Geoserver para que las gestione.

A través de un proyecto Java denominado `SensorDataInterpolation`, se realiza la creación periódica de las imágenes georreferenciadas con el método de interpolación IDW2, comentado en el capítulo previo, y la creación de las coberturas en Geoserver. Hace uso de diferentes tecnologías para la configuración de tareas periódicas (ejecución cada 10 minutos, horaria, diaria, etc.) y consta de varias clases principales que lanzan la ejecución del script R (`interpolateaggregate`) que interpola las observaciones y genera las imágenes que se sirven después como coberturas.

Las clases `CreateDailyImageMosaic` y `CreateHourlyImageMosaic` crean los mosaicos iniciales para las interpolaciones diarias, se genera una nueva para cada mes, y con todas las horarias.

`DailyHistoricalService` es una tarea que inicialmente va creando las coberturas históricas agregadas por día entre dos fechas especificadas en un archivo de configuración.

`SensorDataInterpolationService` es la clase que contiene las funciones que se ejecutan de forma periódica y generan las diversas coberturas que pueden verse resumidamente en el Cuadro 4.2.

El proceso, tal y como se observa en la Figura 4.3, y como en el flujo de trabajo de construcción de los fenómenos discretos, comienza con la creación de los espacios de trabajo y almacenes necesarios para la publicación de las coberturas mediante la clase Java `GenerateWFSFeatureTypes`. Una vez están presentes en Geoserver, es hora de incorporar los estilos, un paso que sería prescindible si ya fueron cargados anteriormente. Una vez están los elementos necesarios para sustentar la publicación de las coberturas en Geoserver, deben iniciarse los servicios periódicos que irán nutriendo los mosaicos de imágenes de monitorización. En primer lugar se evalúa el tiempo actual, y en función del mismo, resultan imágenes cada 10 minutos (tiempo real), o mosaicos horarios y diarios para almacenar información histórica, como se detalla a continuación:

– Cada diez minutos se invoca al script R de interpolación que genera la de tiempo real: *real_time_air_quality_observation_coverage*, que es reemplazada constantemente por la cobertura más actual.

– Cada hora se invoca al script R de interpolación y se genera la imagen con la media de la última hora. Se insertan al mosaico de la interpolaciones horarias: *hourly_air_quality_observation_coverage_date*.

– Cada día se invoca al script R para que cree las imágenes con los valores agregados (promedio, mínimo y máximo) y las incorpora al mosaico de coberturas diarias mensuales donde se almacenan, *daily_air_quality_observation_coverage_year_month*. Al finalizar el mes se genera un nuevo mosaico mensual.

Subtype	Name	WCS	WMS	Java class
Real-Time	real_time_air_quality_observations_coverage	X	X	SensorDataInterpolation
Historic	hourly_air_quality_observation_coverage_date	X	X	CreateHourlyImageMosaic
Historic	daily_air_quality_observation_coverage_year_month	X	X	CreateDailyImageMosaic
Latest prediction	latest_air_quality_prediction_coverage	X	X	GRALInterpolation
Historic prediction	air_quality_prediction_coverage_date	X	X	GRALInterpolation

Cuadro 4.2: Servicios y clase Java origen de las coberturas.

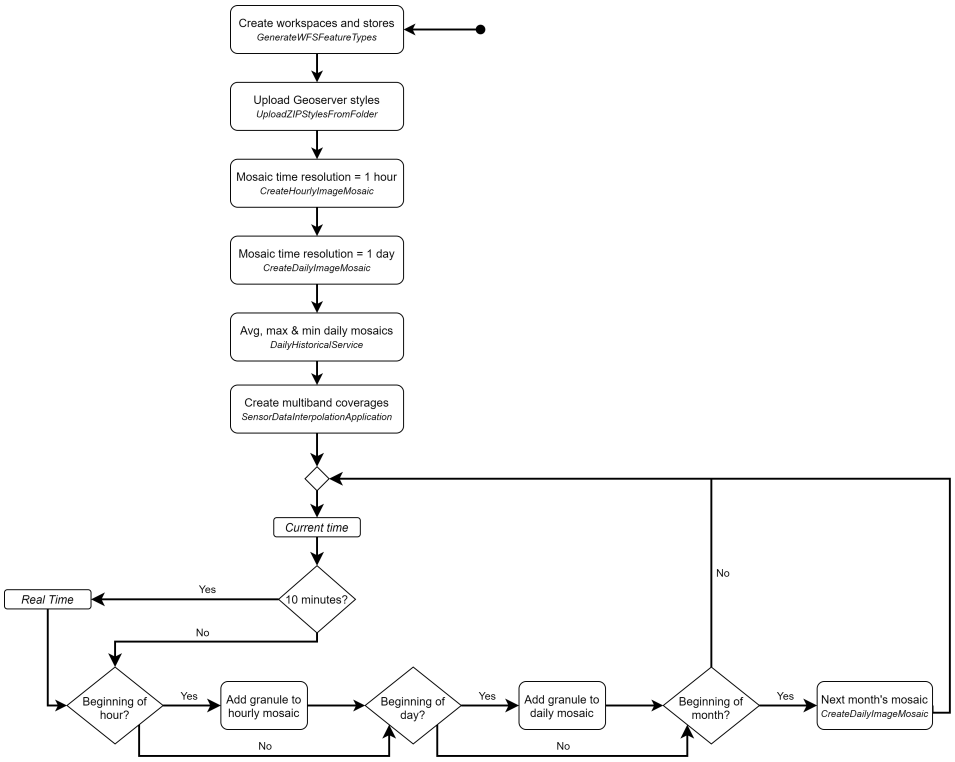

Figura 4.3: Flujo de trabajo para la creación de coberturas de monitorización.

4.1.3. Creación de servicios OGC

Los proyectos Java descritos en las secciones 4.1.1 y 4.1.2 permite configurar distintas capas en Geoserver, que posteriormente pueden ser accedidas mediante interfaces conformes con las especificaciones de servicios establecidas por el OGC. Los fenómenos discretos son accesibles por medio de servicios WFS, y opcionalmente también mediante servicios WMS si requieren visualización. Las coberturas son accesibles mediante servicios WCS, y opcionalmente también por medio de servicios WMS si requieren visualización. A continuación se muestran algunos ejemplos de las capas accesibles mediante estos servicios.

– Web Feature Service (WFS): Capas de fenómenos discretos de calidad del aire. Ejemplo: *historic_air_quality_observations_2019*

```
1 FID,featureid,phenomenon_time,result_time,no,
      no_provenance,no2,no2_provenance,co,
      co_provenance,o3,o3_provenance,coverage
2 historic_air_quality_observations_2019.fid--1
      e16f3b5_1738fd06e16_-1847,24,2019-09-11
      T00:20:00,2020-05-21T10:38:13
      .015,4.62059293,21,9.51384812,22,
      0.12829769,23,65.21304963,24,1
3 historic_air_quality_observations_2019.fid--1
      e16f3b5_1738fd06e16_-1846,24,2019-09-11
      T00:50:00,2020-05-21T10:38:13
      .015,4.8580541,21,10.08335984,22,
      0.12860563,23,66.51418208,24,1
4 historic_air_quality_observations_2019.fid--1
      e16f3b5_1738fd06e16_-1845,24,2019-09-11
      T01:30:00,2020-05-21T10:38:13
      .015,4.43070818,21,13.99431626,22,
      0.12962319,23,56.19352817,24,1
5 historic_air_quality_observations_2019.fid--1
      e16f3b5_1738fd06e16_-1844,24,2019-09-11
      T01:40:00,2020-05-21T10:38:13
      .015,4.29735092,21,13.10910851,22,
      0.13152616,23,62.85603664,24,1
6 historic_air_quality_observations_2019.fid--1
      e16f3b5_1738fd06e16_-1843,24,2019-09-11
      T01:50:00,2020-05-21T10:38:13
      .015,4.61006881,21,14.22249203,22,
      0.14372467,23,61.30412811,24,1
```

Código 4.4: Resultado CSV de 5 elementos tras Petición *GetFeature*.

– Web Coverage Service (WCS): Capas de coberturas espacio-temporales de calidad del aire. Ejemplo: *real_time_air_quality_observation_coverage_CO*

```
1  <?xml version="1.0" encoding="UTF-8"?><
     GetCoverage version="1.1.1" service="WCS"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-
     instance" xmlns="http://www.opengis.net/wcs
```

```
    /1.1.1" xmlns:ows="http://www.opengis.net/ows
    /1.1" xmlns:gml="http://www.opengis.net/gml"
    xmlns:ogc="http://www.opengis.net/ogc"
    xsi:schemaLocation="http://www.opengis.net/wcs
    /1.1.1 http://schemas.opengis.net/wcs/1.1.1/
    wcsAll.xsd">
  <ows:Identifier>
    open_data:real_time_air_quality_observation_cove
  rage_CO</ows:Identifier>
  <DomainSubset>
    <ows:BoundingBox crs="
    urn:ogc:def:crs:EPSG::4326">
      <ows:LowerCorner>41.617799277315
    -0.937511507198367</ows:LowerCorner>
      <ows:UpperCorner>41.692319277315
    -0.843095507198367</ows:UpperCorner>
    </ows:BoundingBox>
  </DomainSubset>
  <Output store="true" format="image/tiff">
    <GridCRS>
      <GridBaseCRS>urn:ogc:def:crs:EPSG::4326</
    GridBaseCRS>
      <GridType>urn:ogc:def:method:WCS:1.1
    :2dSimpleGrid</GridType>
      <GridOffsets>4.7999999999999974E-5
    -3.599999999999985E-5</GridOffsets>
      <GridCS>urn:ogc:def:cs:OGC:0.0:Grid2dSquareCS
    </GridCS>
    </GridCRS>
  </Output>
</GetCoverage>
```

Código 4.5: Resultado XML tras petición *GetCoverage*.

– Web Map Service Interface Standard (WMS): Instancias de representación de mapas de coberturas. Ejemplo: *hourly_ air_ quality_ observation _ coverage_ NO2*

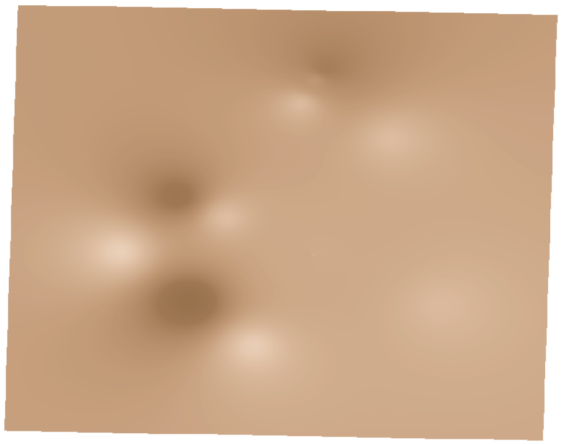

Figura 4.4: Resultado GeoTIFF tras petición *GetMap*.

4.2. Publicación en portales de datos abiertos e integración de nuevos formatos

En esta sección se describen los últimos pasos que son necesarios para publicar los conjuntos de datos en un portal de tipo CKAN una vez que hemos creado y configurado las capas en Geoserver. En la figura 4.5 se puede ver el flujo de trabajo para la publicación de datos abiertos. En la sección 4.1 nos hemos centrado en la creación de capas de Geoserver accesibles mediante OGC. En esta sección describimos los pasos siguientes: extracción de metadatos facilitados por los servicios OGC mediante su operación *GetCapabilites* (descripción de las capacidades del servicio) , *harvest OGC metadata* en la figura; ingestión de metadatos en un servidor de tipo CKAN, *ingest metadata* en la figura; y difusión de los metadatos del servidor CKAN a través de otros servidores de carácter superior asociados, *harvest metadata* en la figura.

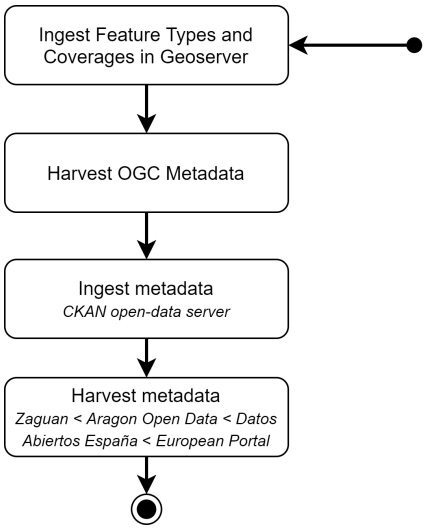

Figura 4.5: Flujo de trabajo para la publicación de datos abiertos.

Para recuperar los metadatos disponibles de las capacidades del servicio y convertirlos en unos metadatos conformes con el estándar GeoDCAT-AP se desarrolló un programa Python a través de un Trabajo Fin de Grado asociado también al proyecto [27].

El diagrama UML expuesto en la Figura 4.6 muestra las propiedades necesarias para describir los conjuntos de datos y distribuciones en base al estándar GeoDCAT-AP.

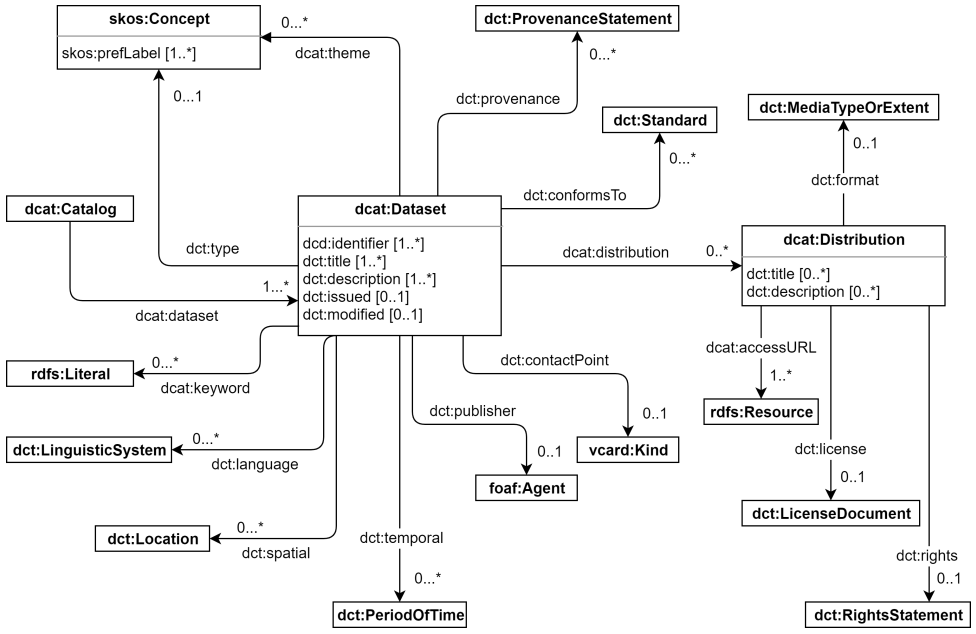

Figura 4.6: Modelo UML con las propiedades en base a GeoDCAT-AP.

Como el proceso de generación automática de metadatos parte de la escasa información sobre las capacidades de los servicios OGC y las capas a las cuales facilitan acceso, una tarea asumida por este TFM dentro de la publicación de datos en portales abiertos ha sido la revisión de los metadatos generados de forma automática. En primer lugar, los conjuntos de datos publicados como capas de Geoserver (vistas en tiempo real, históricas, etc.) mediante los proyectos Java descritos en la sección 4.1 no incluyen información sobre su extensión temporal. Esta información sobre la extensión temporal es un requisito para que los metadatos del servidor CKAN puedan ser cargados en otros portales de datos abiertos donde está federado el portal CKAN. Para solucionarlo deben incorporarse las extensiones temporales de manera manual en CKAN.

Un segundo elemento que ha sido necesario revisar ha sido la idoneidad de los títulos y descripciones generados para las capas en los proyectos descritos en la sección 4.1. Aunque los proyectos Java intentaban detectar patrones y generar títulos humanos según el estilo de los Cuadros 3.1 y 3.2, hay títulos y descripciones de capas que no se ajustan totalmente a estos patrones o ha habido que modificar esta información de las capas que ya existían previamente en Geoserver.

Por último, un tercer elemento de la revisión de los metadatos en el portal de

CKAN ha sido la integración de mejoras en la publicación, añadiendo diferentes distribuciones que faciliten el tratamiento e interoperabilidad de los datos mediante la diseminación de otros formatos como CSV, RDF y NetCDF. Estos nuevos formatos de distribución se describen con mayor detalle en las secciones 4.2.1, 4.2.2 y 4.2.3

Por último, debemos remarcar que la publicación en el portal de datos abiertos del proyecto TRAFAIR (CKAN) permite que los datos abiertos escalen, y es el paso previo al volcado en el repositorio institucional de documentos de la Universidad de Zaragoza (Zaguan).[9] Desde Zaguan, los metadatos son recolectados por el portal de datos abiertos del Gobierno de Aragón (Aragon Open Data),[10] pasan al español (Iniciativa de datos abiertos del Gobierno de España),[11] y por último, al portal de datos abiertos de la Unión Europea (European Data Portal).

4.2.1. Datos tabulares CSV

El formato tabular CSV hace referencia a los archivos de texto cuyos campos (atributos) están delimitados por comas y que contienen una línea para cada registro (característica); los valores también pueden estar separados por punto y coma, tabulación o espacio.

Representan un estándar en el intercambio de información alfanumérica para su tratamiento estadístico y como también permiten el almacenamiento de las geometrías en el formato OGC WKT (*Well Known Text*)[12] son un buen método para compartir datos vectoriales. Como Geoserver soporta de forma nativa la exportación de fenómenos discretos, se puede implementar el recurso de forma dinámica en CKAN a través de la adicción de una petición GetFeature (ver Código 4.6) a los conjuntos de datos que se quieran distribuir en formato tabular.

```
1 http://atila.unizar.es:8081/geoserver/open_data/ows?
      service=WFS&version=1.0.0&request=GetFeature&typeName
      =open_data%3Ahistoric_air_quality_observations_2019&
      outputFormat=csv
```

Código 4.6: Petición *GetFeature* de un CSV al servidor.

[9] https://zaguan.unizar.es/
[10] https://opendata.aragon.es/
[11] https://datos.gob.es/
[12] Representación compacta de objetos geométricos legible por máquina y por humanos. https://www.ogc.org/standards/wkt-crs

4.2.2. Datos semánticos RDF

Los archivos descriptivos RDF o Marco de Descripción de Recursos (Resource Description Framework)[13] son un modelo estándar para la descripción de cualquier recurso de Internet que hace uso del lenguaje XML como sistema de comunicación. Facilitan el intercambio de información y la catalogación de bibliotecas de recursos al homologar, mediante ontologías, los metadatos para las distintas propiedades, como descripciones, origen, autores, etc.

En el contexto del proyecto TRAFAIR en Zaragoza, se decidió incorporar el formato de datos abiertos RDF para la publicación de fenómenos discretos, en particular, para las observaciones históricas distribuidas bajo el nombre *historic_ air_ quality_ observations_ year*, cuya estructura se detalla a continuación:

- featureid (*numeric*): Identificador del elemento (localización) donde han sido observadas las concentraciones.
- phenomenon_time (*time instant*): Instante temporal donde las observaciones calibradas son validas en el elemento correspondiente.
- result_time (*time instant*): Instante de tiempo en el que el proceso de calibración es ejecutado para generar observaciones calibradas desde los voltajes brutos del sensor.
- NO (*numeric*): Concentración de NO en microgramos por metro cúbico.
- NO_provenance (*numeric*): Identificador del proceso de calibración para NO en el conjunto de datos air_quality_observation_- provenance_metadata.
- NO2 (*numeric*): Concentración de NO_2 en microgramos por metro cúbico.
- NO2_provenance (*numeric*): Identificador del proceso de calibración para NO_2 en el conjunto de datos air_quality_observation_provenance_metadata.
- CO (*numeric*): Concentración de CO en miligramos por metro cúbico.
- CO_provenance (*numeric*): Identificador del proceso de calibración para CO en el conjunto de datos air_quality_observation_provenance_-metadata.
- O3 (*numeric*): Concentración de O_3 en microgramos por metro cúbico.

[13] `https://www.w3.org/RDF/`

- O3_provenance (*numeric*): Identificador del proceso de calibración para O_3 en el conjunto de datos air_quality_observation_provenance_-metadata.
- coverage (*numeric*): Porcentaje de datos brutos del sensor disponibles para el proceso de calibración durante el período pertinente, con respecto al número máximo de datos posibles.

Se adapta la información definida para el conjunto de datos históricos en su proceso de conversión de CSV a RDF mediante el programa Java SPARQL for Tables (detallado en el Anexo C), a partir de un vocabulario descriptivo ya existente (ESAIR) [28], que extiende la ontología W3C Semantic Sensor Network Ontology (SOSA)[14] para datos de calidad del aire en áreas urbanas y mantiene los tesauros de contaminantes publicados por la European Environment Agency (EIONET vocabulary)[15]. Como se puede ver en el Código 4.8 se convierten los datos tabulares en tripletas semánticas mediante una consulta `CONSTRUCT`.

En función de las diferentes propiedades se transforman los atributos de las observaciones históricas de 2019 en sus clases correspondientes, y también se otorga a cada registro un identificador único universal (UUID).

```
PREFIX esair: <http://vocab.linkeddata.es/datosabiertos/
    def/medio-ambiente/calidad-aire>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
PREFIX sosa: <http://www.w3.org/ns/sosa>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

CONSTRUCT {
  ?u geo:Feature ?FID;
  esair:AirQualitySensor ?featureid;
  sosa:phenomenonTime ?phenomenon_time;
  sosa:Result ?result_time;
  esair:monoxidoDeNitrogeno ?no;
  sosa:Procedure ?no_provenance;
  esair:dioxidoDeNitrogeno ?no2;
  sosa:Procedure ?no2_provenance;
  esair:monoxidoDeCarbono ?co;
  sosa:Procedure ?co_provenance;
  esair:ozono ?o3;
  sosa:Procedure ?o3_provenance;
  sosa:Sampling ?coverage;

  }
  FROM <file:historic_air_quality_observations_2019.csv>
  WHERE {
    BIND (UUID() AS ?u)
  }
```

Código 4.7: Consulta SPARQL para el mapeo CSV a RDF.

[14] `https://www.w3.org/TR/vocab-ssn/`
[15] Air Quality Pollutants. `http://dd.eionet.europa.eu/vocabulary/aq/pollutant/`

Se expone el resultado para un registro en el extracto del Código 4.8, donde se puede observar como la sintaxis de Turtle[16] permite compactar diversas sentencias separadas por ".", e internamente los sujetos, por diversos predicados (columnas del CSV) con ";".

En primer lugar se hace referencia a los vocabularios utilizados (`@prefix`) y después se detalla la información del primer registro (línea 7), su identificador único, el sujeto, (`<urn:uuid:...`), y posteriormente el resto de clases correspondientes a la información presente en el CSV de origen (`geo:, esair:, sosa:`). Un ejemplo esquemático para dos de los registros existentes se puede ver en el grafo incluido en la Figura C.1 del Anexo C.

```
1  @prefix geo:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
       .
2  @prefix xsd:  <http://www.w3.org/2001/XMLSchema> .
3  @prefix esair:  <http://vocab.linkeddata.es/datosabiertos
       /def/medio-ambiente/calidad-aire> .
4  @prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns
       #> .
5  @prefix sosa:  <http://www.w3.org/ns/sosa> .
6
7  <urn:uuid:090cca67-60ad-4beb-890b-59ff0c645034>
8          geo:Feature                "
       historic_air_quality_observations_2019.fid--1
       e16f3b5_1738fd06e16_-17cf" ;
9          esair:AirQualitySensor     "24" ;
10         sosa:phenomenonTime        "2019-09-11T00:20:00"
       ;
11         sosa:Result                "2020-05-21T10
       :38:13.015" ;
12         esair:monoxidoDeNitrogeno  "4.62059293" ;
13         sosa:Procedure             "21" ;
14         esair:dioxidoDeNitrogeno   "9.51384812" ;
15         sosa:Procedure             "22" ;
16         esair:monoxidoDeCarbono    "0.12829769" ;
17         sosa:Procedure             "23" ;
18         esair:ozono                "65.21304963" ;
19         sosa:Procedure             "24" ;
20         sosa:Sampling              "1" .
```

Código 4.8: Extracto en notación Turtle del RDF resultante.

4.2.3. Datos matriciales NetCDF

El formato NetCDF permite la creación, acceso e intercambio de datos matriciales en conjuntos multidimensionales, a los que se pueden incluir metadatos, lo que facilita la publicación de información espacial en formato ráster. Mediante las bibliotecas de software y formatos de datos independientes y

[16]Tripletas conformadas por un sujeto, predicado y objeto. `https://www.w3.org/TR/turtle/`

gratuitos,[17] utilidades gratuitas como Panoply,[18] o mediante Python, R o Matlab, se puede crear y manipular conjuntos de datos NetCDF.

Geoserver permite la exportación de datos en NetCDF, siempre y cuando el servidor disponga de las librerías necesarias y se instale correctamente la extensión NetCDF Output Format.[19] Debido al gran tamaño de las coberturas históricas (hourly, daily), la petición *GetCoverage* del formato, requiere mucho tiempo de espera, e implicaría servidores de mayores capacidades que el utilizado en el proyecto en Zaragoza. Con lo que se ha implementado de manera experimental, y para generar información con series temporales en una sola petición, pero no va a ser un formato de publicación disponible en los portales de datos abiertos.

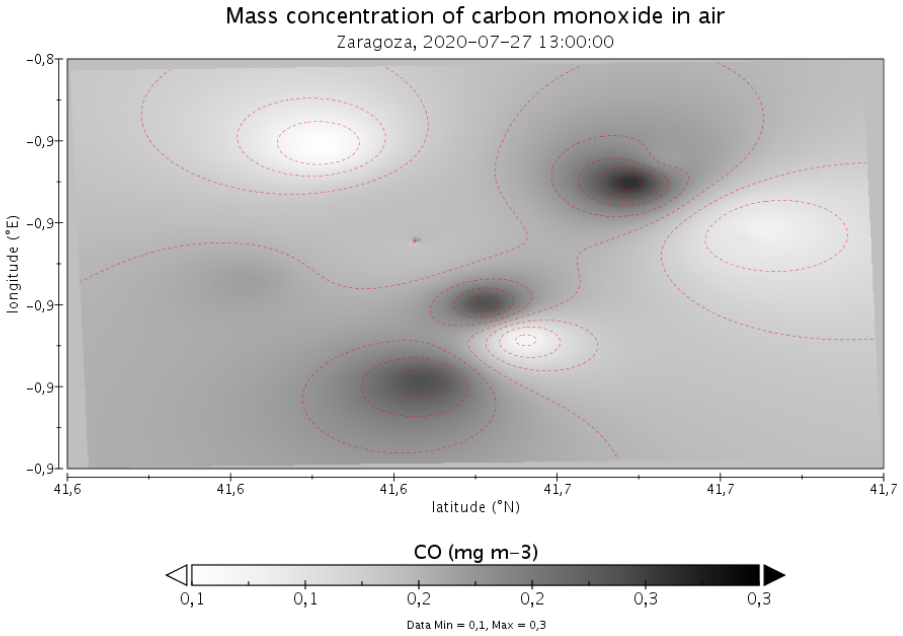

Figura 4.7: Visualización de archivo NetCDF graficado.

La Figura 4.7 representa *hourly_air_quality_observation_coverage_CO.nc*[20] que contiene 178 (time = 178) coberturas horarias de observaciones de monoxido de carbono, variable *mass_concentration_of_carbon_monoxide_in_air*, medida en mg/m^3 (units: mg m-3), entre el 27 de julio de 2020 y el 3 de agosto

[17] `https://www.unidata.ucar.edu/software/netcdf/docs/getting_and_building_netcdf.html`
[18] `https://www.giss.nasa.gov/tools/panoply/`
[19] `https://docs.geoserver.org/maintain/en/user/extensions/netcdf-out/index.html`
[20] *.nc* es la extensión de los ficheros en formato NetCDF.

de 2020 (no viene explicitado en este archivo de metadatos) y las propiedades de las referencias espaciales, tal y como se puede observar en los metadatos expuestos en el Código 4.9.

```
netcdf file:/D:/TRAFAIR/hourly_air_quality_observation_coverage_CO.nc {
  dimensions:
    time = 178;
    lat = 2070;
    lon = 1967;
  variables:
    double time(time=178);
      :long_name = "time";
      :description = "time";
      :units = "seconds since 1970-01-01 00:00:00 UTC";

    float lat(lat=2070);
      :long_name = "latitude";
      :units = "degrees_north";
      :standard_name = "latitude";

    float lon(lon=1967);
      :long_name = "longitude";
      :units = "degrees_east";
      :standard_name = "longitude";

    float mass_concentration_of_carbon_monoxide_in_air(time=178, lat=2070, lon=1967);
      :long_name = "mass_concentration_of_carbon_monoxide_in_air";
      :units = "mg m-3";
      :_FillValue = -3.4E38f; // float

  // global attributes:
  :spatial_ref = "GEOGCS[\"WGS 84\", \n DATUM[\"World Geodetic System 1984\", SPHEROID[\"WGS 84\", 6378137.0, 298.257223563, AUTHORITY[\"EPSG\",\"7030\"]], AUTHORITY[\"EPSG\",\"6326\"]], \n PRIMEM[\"Greenwich\", 0.0, AUTHORITY[\"EPSG\",\"8901\"]], \n UNIT[\"degree\", 0.017453292519943295], AXIS[\"Geodetic longitude\", EAST], \n AXIS[\"Geodetic latitude\", NORTH], AUTHORITY[\"EPSG\",\"4326\"]]";
  :GeoTransform = "-0.9374875071983719 4.800000000000003E-5 0.0 41.6923012773151 0.0 -3.599999999999985E-5";
}
```

Código 4.9: Metadatos del archivo NetCDF exportado de Geoserver.

5. Análisis de la calidad

Al abordar la calidad de los datos espaciales se puede hacer mención a la literatura que cubre aspectos generales sobre cómo detectar los valores atípicos y los errores de medición en objetos con información espacial destinados al análisis de la calidad del aire [29], como son los datos recogidos por los sensores o estaciones legales. Y también a otros aspectos más específicos relacionados con la precisión posicional, de sus atributos y de su semántica, así como la integridad, o completitud, temporal y espacial [30]. Existen anomalías comunes en la publicación de los datos espaciales que pueden ser identificadas, o solucionadas, de manera más rutinaria mediante scripts en R o Python. Por ejemplo, aquellas vinculadas a la presencia de duplicados, vacíos en los datos recogidos o la ausencia de identificadores únicos y/o atributos espaciales en todos los registros. Sea cuál sea el tamaño o complejidad de la arquitectura de un proyecto, profundizar en la evaluación de los potenciales errores permitirá mejorar la calidad general de los conjuntos de datos.

Debido a que la magnitud de un análisis exhaustivo de la calidad supera los límites del presente trabajo, en este capítulo se hará referencia a tres aspectos básicos: la comparación de los modelos lógicos desarrollados para los sensores/estaciones y los fenómenos discretos con las especificaciones de datos propuestas para los temas tratados por la directiva INSPIRE, la metodología de evaluación de la calidad de los metadatos (MQA)[1] basada en el método empleado por el Portal Europeo de Datos Abiertos para medir la calidad de los datos recolectados; y en último lugar, la evaluación de la completitud temporal, tratando de analizar si existen vacíos en la cobertura temporal de los datos recogidos, y en el mismo sentido, analizar la cobertura espacial de los sensores y estaciones reguladas en función de su localización en la ciudad de Zaragoza.

[1]Metadata Quality Assessment.
`https://www.europeandataportal.eu/mqa/methodology?locale=es`

Adicionalmente se analiza la exactitud de los datos recogidos por la red de sensores desplegados por TRAFAIR comparándolos con los datos de las estaciones legales del Ayuntamiento de Zaragoza, en particular para comprobar si la presencia de algunos de los conjuntos de datos pueden suplir errores en otros, o si la combinación utilizada para generar la cartografía de interpolación (información calibrada de los sensores y no validada de las estaciones) puede ser sustituida exclusivamente por las observaciones de los sensores.

5.1. Comparativa con otros modelos de referencia

Los temas de datos espaciales definidos en los Anexos de la Directiva INSPIRE (Directiva 2007/2/EC) relevantes para los conjuntos de datos publicados son: instalaciones de monitorización ambiental (*Environmental monitoring Facilities* [31]) y condiciones atmosféricas y características geográficas meteorológicas (*Atmospheric conditions* [32]). Ambos esquemas también se especifican en UML, y detallan la lista de los objetos, incluidas sus propiedades y tipos asociados. Este tipo de modelos genéricos otorgan la libertad requerida para introducir necesidades temáticas específicas al mismo tiempo que se mantiene una estructura consistente de datos compartidos.

Mediante la comparación de los modelos construidos con los referidos de INSPIRE se pueden identificar aquellos objetos ausentes, pero que son requeridos, las propiedades de datos que se deben transformar de acuerdo con las definiciones de INSPIRE; e incluso, si fuera necesario, ampliar las extensiones de temas para cumplir con las especificaciones del conjunto de datos.

Primeramente, para las instalaciones de monitorización ambiental (Figura 5.1), el tema INSPIRE aborda dos aspectos principales: el objeto en sí de monitorización ambiental (estación o sensor) y las observaciones y mediciones recogidas en cada uno de los mismos de acuerdo a la normativa ISO 19156.[2] También se incorporan otros elementos como la información administrativa vinculada a las instalaciones, las capacidades de recogida de datos o si forma parte de alguna red. En ese sentido, el diagrama de la Figura 5.2 expone las características atmosféricas y sus atributos meteorológicos, vinculadas ambas temáticas a sus aspectos geográficos. En general, se plantea que la información publicada para los usuarios contemple, al menos, información

[2]ISO 19156: 2011, Geographic information - Observations and measurements. `https://www.iso.org/obp/ui/#iso:std:iso:19156:ed-1:v1:en`

sobre la precipitación, temperatura, evapotranspiración y el viento.

En resumen, que se incorpore información sobre las condiciones físicas del entorno en el que han sido captados los datos.

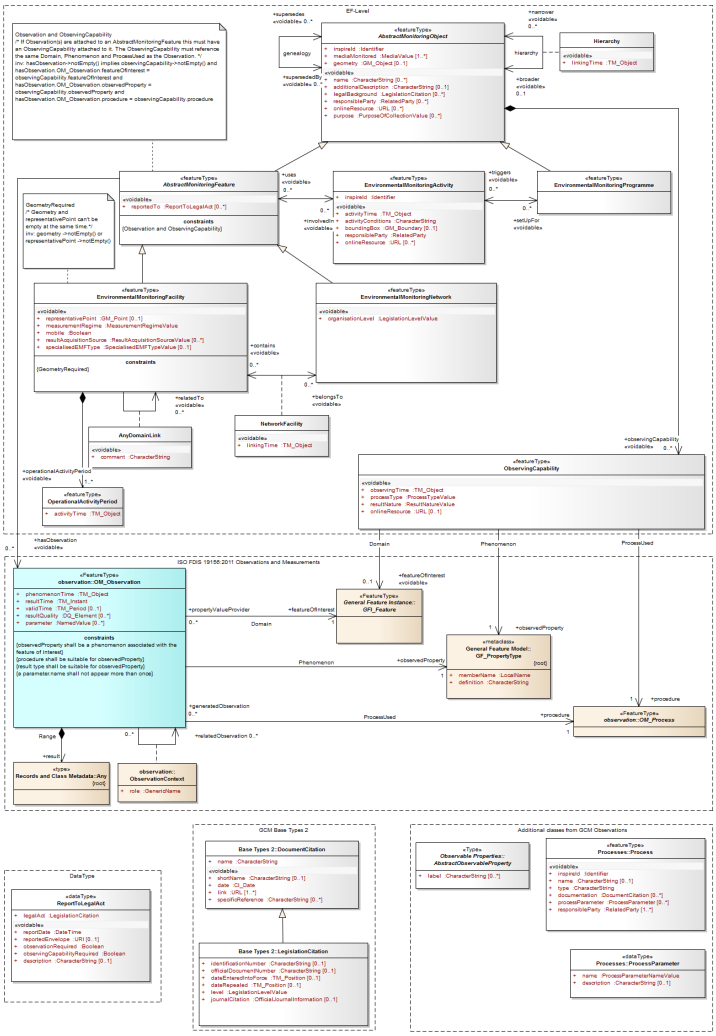

Figura 5.1: Esquema general tema *Environmental monitoring Facilities*. INSPIRE.

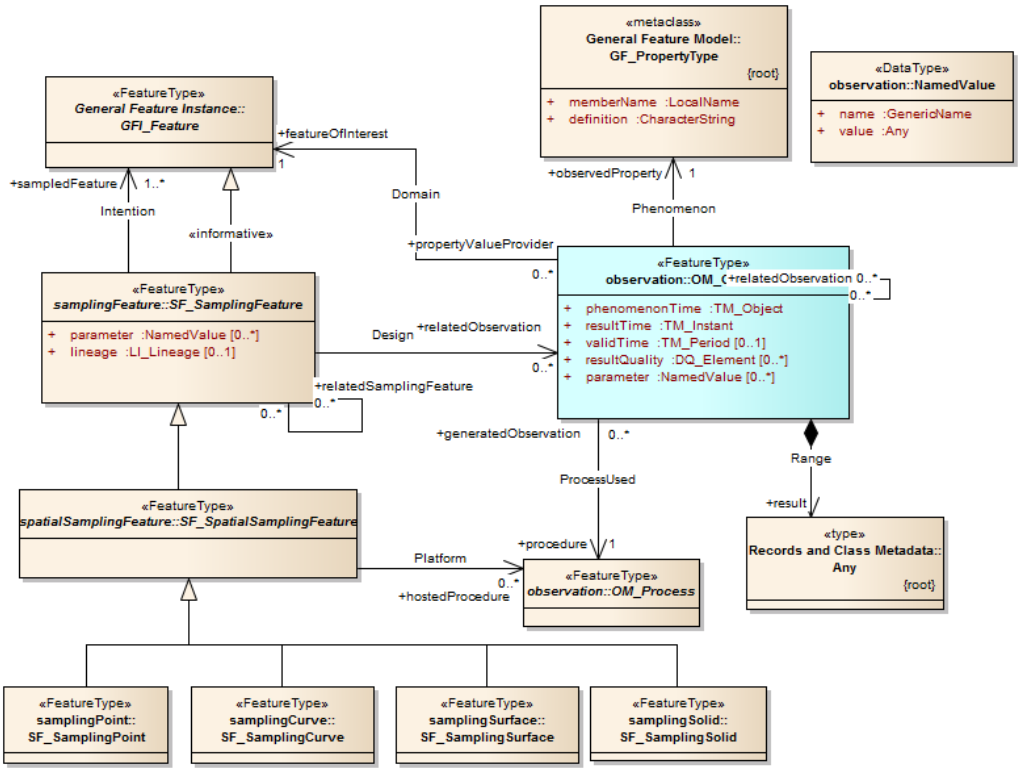

Figura 5.2: Esquema general tema *Atmospheric Conditions*. INSPIRE.

Tal y como se puede apreciar en el modelo del Capítulo 2 para las estaciones reguladas (*Air quality monitoring stations*) y los sensores de bajo coste (*Air quality sensors low-cost*) anexado en las Figuras 2.4 y 2.6, y comparativamente, en los Cuadros 5.1 y 5.2; existen posibles correspondencias entre las propiedades de ambos modelos y las que proponen los temas de INSPIRE, pero centrados en los aspectos temáticos que se han de publicar acerca de datos de calidad del aire.

Las propuestas de INSPIRE plantean el contenido de la especificación de los datos, tanto obligados como recomendados, que deben contener los conjuntos de datos publicados. En especial, se incluyen aquellas características, como los atributos geográficos que informan sobre la geometría de representación, los atributos temporales o las propiedades de las instalaciones. Estos elementos constituyen una fuente de información fundamental para entender el funcionamiento de las instalaciones y analizar los datos atmosféricos recogidos.

Las tablas recogen como existe una predominancia en TRAFAIR de los obje-

tos vinculados a la información sobre los datos recogidos por las instalaciones ambientales (*SensorCalibratedObservation, AQ_ LegalStationObservationValidated*, ...), y ninguna información acerca de 3 de los 4 tipos de objetos espaciales relativos al esquema de instalaciones ambientales (*EnvironmentalMonitoring...*):

Asimismo, si se tuviese que ampliar el modelo de la base de datos de TRAFAIR, la comparación con la especificación de datos de INSPIRE permite identificar algunos elementos que convendría tener en cuenta: las actividades (*...Activity*) desempeñadas, que expresa la necesidad de describir la actividad de las campañas de monitorización ; las redes (*...Network*) de las que forman parte, dado que una instalación de vigilancia ambiental puede pertenecer a varias redes de vigilancia ambiental por ejemplo, como en el caso de las estaciones del Ayuntamiento de Zaragoza, formando parte de una red nacional y una europea; y programas (*...Programme*), que definen el objetivo de la recogida de observaciones y/o el despliegue. La falta de consideración de estos elementos en el modelo de TRAFAIR explica también que en los conjuntos de datos (*Feature Types*) que se han publicado como datos abiertos predomine la información acerca de las observaciones. Tampoco aparece correspondencia entre la clase GFI_Feature (conjunto de todas las clases que son *Feature Types*) y ambos modelos, esta última representa una instancia de la "metaclase" GF_FeatureType que si está referenciada como OM_Observation.

En resumen, si el proyecto TRAFAIR tuviese la obligación de publicar los datos en conformidad con las especificaciones de datos de INSPIRE, sería necesario ampliar el modelo de la base de datos de TRAFAIR para incorporar información adicional a los datos sobre los sensores/estaciones, la red a la que están adscritos, actividades de las que forman parte, y otra serie de características que ayuden a informar sobre el cometido y las funciones de las instalaciones.

Air quality sensors low-cost	Air quality monitoring stations	Environmental monitoring Facilities
SensorLowCostFeature	AQ_LegalStation	EnvironmentalMonitoringFacility
SensorLowCost	-	AbstractMonitoringFeature
SensorRawObservation	AQ_LegalStationObservationNotValidated	OM_Observation
SensorCalibratedObservation	AQ_LegalStationObservationValidated	OM_Observation
-	AQ_LegalStationObservationOneMinuteNotValidated	OM_Observation
-	AQ_LegalStationParticlesNotValidated	OM_Observation
-	AQ_LegalStationParticlesValidated	OM_Observation
SensorLowCostStatus	-	OperationalActivityPeriod
SensorCalibration	-	ObservingCapability
SensorCalibrationAlgorithm	-	AbstractMonitoringObject
-	-	EnvironmentalMonitoringActivity
-	-	EnvironmentalMonitoringNetwork
-	-	EnvironmentalMonitoringProgramme

Cuadro 5.1: Correspondencias con *Environmental monitoring Facilities*, INSPIRE.

Air quality sensors low-cost	Air quality monitoring stations	Atmospheric Conditions
SensorLowCostFeature	AQ_LegalStation	SF_SpatialSamplingFeature
SensorLowCost	-	SF_SamplingFeature
SensorRawObservation	AQ_LegalStationObservationNotValidated	OM_Observation
SensorCalibratedObservation	AQ_LegalStationObservationValidated	OM_Observation
-	AQ_LegalStationObservationOneMinuteNotValidated	OM_Observation
-	AQ_LegalStationParticlesNotValidated	OM_Observation
-	AQ_LegalStationParticlesValidated	OM_Observation
SensorLowCostStatus	GF_Property	-
SensorCalibration	-	OM_Process
SensorCalibrationAlgorithm	-	GF_Property
-	-	GFI_Feature

Cuadro 5.2: Correspondencias con *Atmospheric conditions*, INSPIRE.

5.2. Evaluación de la Calidad de los Metadatos (MQA)

En lo que respecta a la calidad de los metadatos generados, se ha utilizado la metodología de evaluación de la calidad (*Metadata Quality Assurance* o MQA) propuesta dentro del contexto del Portal Europeo de Datos (EDP). En TRAFAIR Zaragoza se ha desarrollado un programa Python que implementa el MQA ya que la versión del código fuente facilitada por EDP no está actualizada de acuerdo a los criterios actuales del MQA. [3]

A partir de un catalogo de metadatos RDF DCAT-AP recolectados a través de un end-point RDF del servidor CKAN se evalúa un conjunto de dimensiones, cada cual con distintos indicadores, basados en los principios rectores FAIR[4] para la gestión y administración de datos científicos [33].

La Figura 5.3 expone de forma resumida las dimensiones y métricas que contemplan la herramienta MQA, en ella, las dimensiones se muestran de izquierda a derecha según el peso ascendente que suponen. En primer lugar, se valora que los conjuntos dispongan de propiedades que brinden más contexto al usuario (*Contextuality*, Contextualidad), como por ejemplo los derechos de uso, el tamaño o fecha de creación y modificación.

[3] Metadata Quality Assurance, monitoring tool for metadata quality. `https://gitlab.com/european-data-portal/deprecated-components/metadata-quality-assurance`

[4] *Findability, Accessibility, Interoperability, and Reuse of digital assets.*

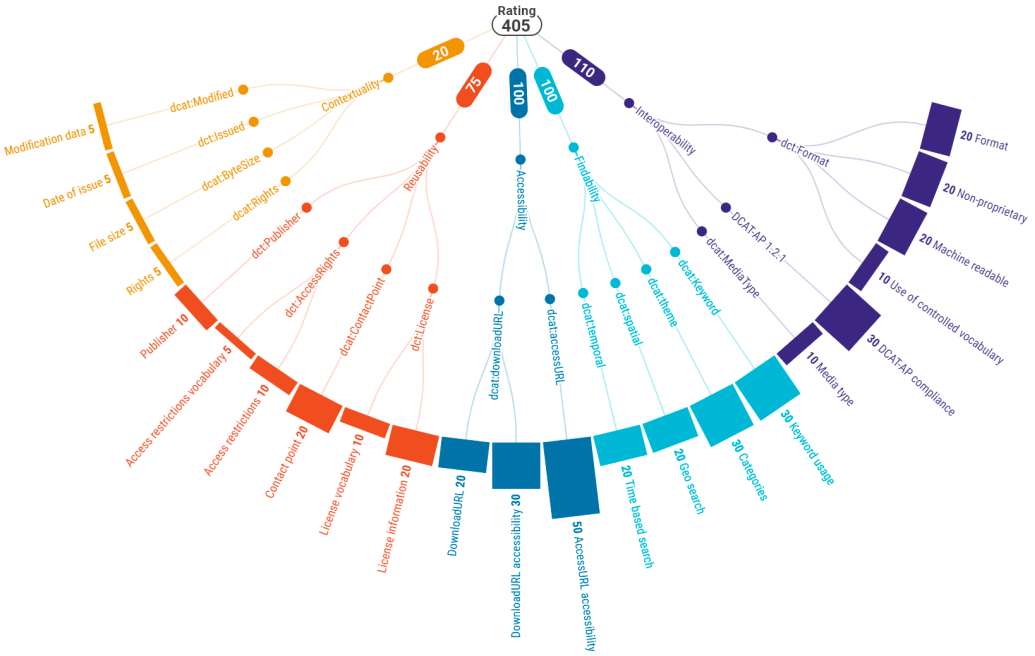

Figura 5.3: Dimensiones y métricas de la herramienta MQA.

También se valora que estén correctamente descritos para poder reutilizarse, o combinarse posteriormente (*Reusability*, Reusabilidad) mediante el establecimiento de licencias y restricciones de uso, la incorporación de vocabularios controlados para las mismas, o la definición de los puntos de contacto y propiedad de los conjuntos.

Para evaluar la Accesibilidad (*Accessibility*) se valora la capacidad y garantías de recuperación y acceso a los datos mediante protocolos abiertos, gratuitos y de implementación universal.

También se evalúa que los conjuntos de datos sean fáciles de buscar para las personas y las máquinas (*Findability*, Facilidad de búsqueda), ya sea mediante palabras clave (*keywords*), categorías, y si disponen de información georreferenciada o temporalmente acotada.

Por último, la dimensión de mayor peso es la que determina si una distribución se considera interoperable, o lo que es lo mismo, si es capaz de integrarse con otros conjuntos de datos, aplicaciones o flujos de trabajo (*Interoperability*, Interoperabilidad), puntuando: el establecimiento de formatos de distribución y tipos de medios, así como si ambos pertenecen a vocabularios controlados; si se usan lenguajes formales compartidos y legibles por máquinas, o los metadatos son conformes con la especificación DCAT-AP o un derivado válido;

y otras características como si los formatos de las distribuciones no están sujetos a propiedad intelectual.

Para conocer la calificación de los metadatos de los conjuntos de datos publicados por TRAFAIR, se recolectan tanto del portal de datos abiertos CKAN del proyecto en Zaragoza, como de los metadatos ya disponibles en el Portal Europeo de Datos. Estos últimos se descargan mediante un filtrado por la palabra clave "Trafair" con otro programa Python específicamente diseñado para esa tarea para recabar los registros uno a uno en formato RDF Turtle.

Para el caso de los metadatos recogidos del portal CKAN de TRAFAIR en Zaragoza, la fecha de evaluación es del 31 de Julio de 2020, para los recolectados del EDP, el 4 de Septiembre. En ambas evaluaciones se obtienen valores adecuados, rangos bueno y suficiente, con 244,04 (Cuadro 5.4) y 219,01 puntos (Cuadro 5.5) respectivamente, como se puede apreciar en el Cuadro 5.3.

CKAN Zaragoza
European Data Portal

Calificación	Rango de puntos
Excelente	351 - 405
Buena	**221 – 350**
Suficiente	**121 – 220**
Mala	0 - 120

Cuadro 5.3: Resumen de la puntuación MQA del European Data Portal.

Es interesante destacar como los metadatos a su paso por diferentes portales (Zaguan, Aragón, España), llegan al European Data Portal lo suficientemente transformados como para presentar calificaciones muy dispares, pese a estar todos ellos inicialmente rellenados en el portal CKAN del proyecto.

Existen diferentes explicaciones. Por ejemplo, algunas métricas evalúan campos que no son considerados en el perfil de metadatos GeoDCAT-AP, centrado en el dominio geoespacial, como `dct:accessRights` o `dcat:downloadURL`, dado que ambas propiedades pertenecen a vocabularios propuestos para recursos de dominios más generales dentro del procedimiento MQA.

Otro motivo ya mencionado anteriormente es que las extensiones temporales en su escalada a otros portales (`dct:temporal`), se pierden debido a reconversiones en formatos de metadatos pensados para otra tipología de datos, no específicamente geográficos. Este es el caso del estándar MARC21 [34] que usa Zaguan, pensado principalmente para la gestión de información bibliográfica. Debido a que los metadatos exigen una transformación de los

registros y propiedades de GeoDCAT-AP a MARC, previa carga al portal de la Universidad, básicamente todos aquellos registros que no cumplen los requerimientos básicos del conjunto de elementos de Dublin Core [35] se pierden y no llegan a Aragon Open Data, ni por ende, al resto de portales.

También es llamativo ver como el cumplimiento del estándar DCAT-AP falla en todos los conjuntos de datos, en CKAN y el EDP, pese a que GeoDCAT-AP es una extensión del estándar DCAT-AP para describir conjuntos de datos, series y servicios geoespaciales. Esto puede ser debido a la sensibilidad de la herramienta MQA, dado que si se identifica al menos un problema en los metadatos, estos se consideran no conformes. Habría que reconocer en donde se produce la violación y el motivo de que cada conjunto de datos no sea conforme para el catálogo seleccionado.

Dimension	Indicator/property	Count	Population	Percentage	Points	Weight
Findability	dcat:keyword	58	58	1	30	30
Findability	dcat:theme	58	58	1	30	30
Findability	dct:spatial	34	58	0.59	11.72	20
Findability	dct:temporal	58	58	1	20	20
Accesibility	dcat:accessURL code=200	93	93	1	50	50
Accesibility	dcat:downloadURL	0	93	0	0	20
Accesibility	dcat:downloadURL code=200	0	93	0	0	30
Interoperability	dct:format	93	93	1	20	20
Interoperability	dcat:mediaType	0	93	0	0	10
Interoperability	dct:format from vocabulary	4	93	0.04	0.43	10
Interoperability	dct:format non proprietary	35	93	0.38	7.53	20
Interoperability	dct:format machine readable	4	93	0.04	0.86	20
Interoperability	DCAT-AP compliance	0	58	0	0	30
Reusability	dct:license	89	93	0.96	19.14	20
Reusability	dct:licensefromvocabulary	89	93	0.96	9.57	10
Reusability	dct:accessRights	0	58	0	0	10
Reusability	dct:accessRightsfromvocabulary	0	58	0	0	5
Reusability	dcat:contactPoint	58	58	1	20	20
Reusability	dct:publisher	58	58	1	10	10
Contextuality	dct:rights	89	93	0.95	4.78	5
Contextuality	dcat:byteSize	0	93	0	0	5
Contextuality	dct:issued	58	58	1	5	5
Contextuality	dct:modified	58	58	1	5	5
Total points					244.04	

Cuadro 5.4: Resultado MQA CKAN TRAFAIR Zaragoza, 31/07/2020.

Dimension	Indicator/property	Count	Population	Percentage	Points	Weight
Findability	dcat:keyword	58	58	1	30	30
Findability	dcat:theme	58	58	1	30	30
Findability	dct:spatial	58	58	1	20	20
Findability	dct:temporal	0	58	0	0	20
Accesibility	dcat:accessURL code=200	147	152	0.97	48.36	50
Accesibility	dcat:downloadURL	0	152	0	0	20
Accesibility	dcat:downloadURL code=200	0	152	0	0	30
Interoperability	dct:format	152	152	1	20	20
Interoperability	dcat:mediaType	0	152	0	0	10
Interoperability	dct:format from vocabulary	152	152	1	10	10
Interoperability	dct:format non proprietary	152	152	1	20	20
Interoperability	dct:format machine readable	5	152	0.03	0.66	20
Interoperability	DCAT-AP compliance	0	58	0	0	30
Reusability	dct:license	152	152	1	20	20
Reusability	dct:licensefromvocabulary	152	152	1	10	10
Reusability	dct:accessRights	0	58	0	0	10
Reusability	dct:accessRightsfromvocabulary	0	58	0	0	5
Reusability	dcat:contactPoint	0	58	0	0	20
Reusability	dct:publisher	59	58	1	10	10
Contextuality	dct:rights	0	152	0	0	5
Contextuality	dcat:byteSize	0	152	0	0	5
Contextuality	dct:issued	0	58	0	0	5
Contextuality	dct:modified	0	58	0	0	5
Total points					219.01	

Cuadro 5.5: Resultado MQA European Data Portal, 04/09/2020.

Al explorar algunos errores se constatan las complicaciones existentes en la exportación de los metadatos que impiden el cumplimiento de algunas propiedades debido a las limitaciones de funcionamiento del software del portal tipo CKAN. Si visualizamos un extracto del informe de error SHACL[5] que aparece acotado a un conjunto en el Código 5.1; para un mismo conjunto de datos se produce una violación en varios puntos: `dct:rights` y `dct:format`. En el primero, el valor del nodo debería ser un enlace a un recurso de tipo `dct:LicenseDocument`, pero CKAN no permite describir el tipo. De manera similar, el formato, "WMS" debería aparecer como una URL enlazando a un recurso de tipo `dct:MediaTypeOrExtent`, pero CKAN tampoco permite describir el tipo. [6]

```
1 Validation Report
2 Conforms: False
3 Results (1449):
4 Constraint Violation in ClassConstraintComponent (http://
      www.w3.org/ns/shacl#ClassConstraintComponent):
5   Severity: sh:Violation
```

[5] *Shapes Constraint Language*, un lenguaje para validar datos RDF confrontando una serie de condiciones.
`https://www.w3.org/TR/shacl/`

[6] https://publications.europa.eu/resource/authority/file-type

```
6   Source Shape: [ sh:class dct:RightsStatement ; sh:
      maxCount Literal("1", datatype=xsd:integer) ; sh:path
       dct:rights ; sh:severity sh:Violation ]
7   Focus Node: <http://atila.unizar.es:3394/dataset/2
      ee599c9-a837-4101-9bb7-f7287cd40a40/resource/468baef9
      -67ba-4b80-8978-3325d6bc4361>
8   Value Node: <http://inspire.ec.europa.eu/metadata-
      codelist/LimitationsOnPublicAccess/noLimitations>
9   Result Path: dct:rights
10 ...
11     Constraint Violation in NodeKindConstraintComponent (
      http://www.w3.org/ns/shacl#
      NodeKindConstraintComponent):
12  Severity: sh:Violation
13  Source Shape: [ sh:class dct:MediaTypeOrExtent ; sh:
      maxCount Literal("1", datatype=xsd:integer) ; sh:
      nodeKind sh:IRI ; sh:path dct:format ; sh:severity sh
      :Violation ]
14  Focus Node: <http://atila.unizar.es:3394/dataset/2
      ee599c9-a837-4101-9bb7-f7287cd40a40/resource/468baef9
      -67ba-4b80-8978-3325d6bc4361>
15  Value Node: Literal("WMS")
16  Result Path: dct:format
17 ...
```

Código 5.1: Extracto del informe de error SHACL para metadatos del EDP.

5.3. Análisis de la completitud de las observaciones de calidad del aire

La evaluación de la completitud de las observaciones trata de determinar si hay *gaps* (huecos o lagunas) temporales, o de observaciones, en algunas estaciones y sensores, pero también, la cobertura espacial de las instalaciones en función de su localización. La ausencia de datos puede deberse a aquellos que nunca se recogieron o a los que fueron extraviados, en ambos casos implica una pérdida de información que tiene un impacto en todos los procesos sucesivos.

```
# Import libraries
import csv
import pandas as pd
from datetime import datetime, timedelta, date, time
import sys

...

# Original phenomenon_times into a pandas df
time_file = pd.DataFrame(data, columns=[timestamp_col,
    featureid_col])

# phenomenon_time to timestamps in year-month-day hour:
    minute:seconds format
time_file[timestamp_col] = pd.to_datetime(time_file[
    timestamp_col], format='%Y-%m-%d %H:%M:%S')
print(time_file.info())

# Take the diff of the first column (drop 1st row since
    it's undefined)
deltas = time_file[timestamp_col].diff()[1:]

# Filter diffs (here days > 1, but could be seconds,
    hours, etc)
gaps = deltas[deltas > timedelta(hours=1)]
print(gaps)

# Timedelta format
def strfdelta(gaps, fmt):
    d = {"days": gaps.days}
    d["hours"], rem = divmod(gaps.seconds, 3600)
    return fmt.format(**d)

#-- OUTPUT FILES --#
## Output files
data_csv = (data_folder + "output/" + "report-" +
    data_file + ".csv")

# Write gaps report to csv file
print(f'---Export gap report---\n',f'{ data_file} has {
    len(gaps)} gaps, with an average gap duration of: {
    strfdelta(gaps.mean(), "{days} days and {hours} hours
    ")}')
```

```
35 org_stdout = sys.stdout
36 output_csv = open((data_csv), "w")
37 sys.stdout = output_csv
38 print(f'start,'
39       f'hours_duration,'
40       f'end')
41 for start, duration in gaps.iteritems():
42     gap_start = time_file[timestamp_col][start - 1]
43     gap_end = time_file[timestamp_col][start]
44     print(f'{datetime.strftime(gap_start, "%Y-%m-%d %H:%M
       :%S")},'
45                       f'{(duration.days*24) + (duration.
       seconds//3600)},'
46                       f'{ datetime.strftime(gap_end, "%Y-%
       m-%d %H:%M:%S")}')
47 sys.stdout = org_stdout
48 output_csv.close()
49 print(" Output file: " + data_csv)
50
```

Código 5.2: Extracto script Python timestamp_gaps_report.py.

Para analizar la cobertura temporal se han desarrollado varios scripts en Python. El primero comprueba la existencia de huecos temporales en los datos de observaciones desde estaciones-sensores (Código 5.2), haciendo uso de las librerías `pandas` y `datetime`,[7] que permiten trabajar con los datos obtenidos de las tablas presentes en la base de datos que contienen los atributos temporales (*phenomenon_time*) en que se recogieron las observaciones. Se filtran los *timestamps*, o registros temporales, y se analiza si existen diferencias, en horas, entre cada uno de los registros existentes. La recolección de los *gaps*, expone todas aquellas horas (o días, minutos, segundos, etc.) en las que no se han cargado observaciones. Se resume la información, como se puede ver a partir de la línea 34, en un archivo csv que contiene el inicio y el fin del hueco y las horas de duración. Esto permite trabajar posteriormente con ese tipo de datos o presentarlos de forma informativa (Informe de cobertura temporal, Cuadros 5.6 y 5.7).

El segundo script completa el resumen de los vacíos temporales cuantificando esas lagunas en forma de registros, con la duración única en unidades de tiempo. Esto facilita la representación de los resultados mediante un diagrama generado utilizando la librería `calplot`[8] en el que se visualiza el comportamiento de los datos anuales (Figuras 5.4 y 5.5) como un mapa de calor.

[7]Module for manipulating dates and times. `https://docs.python.org/3/library/datetime.html`
[8]Calendar heatmaps from Pandas time series data. `https://pythonhosted.org/calmap/`

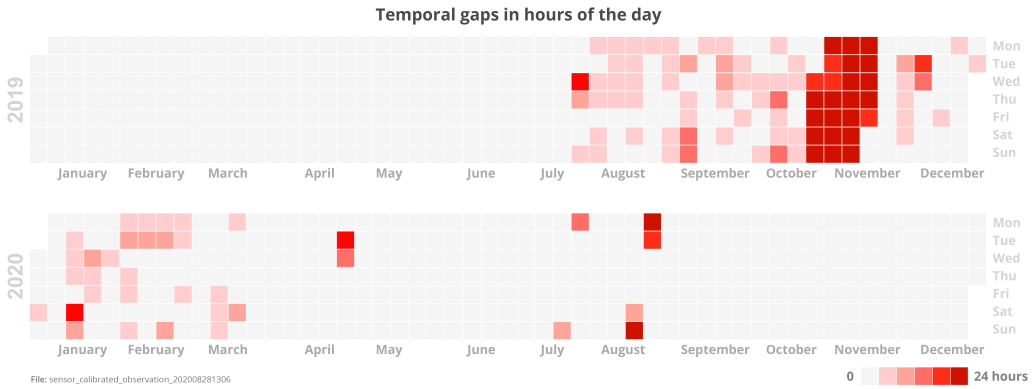

Figura 5.4: Cobertura temporal, observaciones calibradas.

sensor_calibrated_observation_202008281306
has 151 gaps, with an average gap duration of: 0 days and 12 hours

start	hours_duration	end
2019-07-31 08:00:00	22	2019-08-01 06:10:00
2019-08-04 10:50:00	1	2019-08-04 12:10:00
2019-08-05 08:30:00	2	2019-08-05 11:00:00
2019-08-05 20:00:00	1	2019-08-05 21:10:00
...	...	...

Cuadro 5.6: Extracto informe cobertura temporal, observaciones calibradas.

Las observaciones calibradas, provenientes de los sensores de bajo coste, presentan una menor regularidad en su captura; destacando el periodo inicial del proyecto (22-07-2019), donde existen amplios vacíos de datos de más de varios días de duración. Esto es debido a que fue el momento en que se desplegaron los sensores, observándose una mayor continuidad en la carga de datos a partir de Marzo de 2020. También es interesante apreciar que la puesta en marcha del nuevo proceso de calibración (finales de Junio) ha producido algunos errores en los meses de Julio y agosto.

Esta intervención que afectó a los conjuntos de datos de observaciones esta vinculada al propio proceso de prueba y error del proyecto. Aunque genera algunas lagunas en la cobertura temporal de los datos, estas se compensarán con las implicaciones que puede suponer una recogida y tratamiento de los datos más estable y precisa en el futuro.

En el caso de las observaciones no validadas recogidas por la red de estaciones legales del Ayuntamiento de Zaragoza, y teniendo en cuenta que el proyecto se inicio en Julio de 2019, solo existe una hora sin datos cargados en la base de datos de TRAFAIR. Bien pudiera ser producida por un error en el proceso de carga utilizado, o por ser, efectivamente, una ausencia de datos

en la fuente de origen por procesos de calibración o mantenimiento. Sea cual sea el motivo, el rendimiento en la recogida de los datos es muy alto, lo que permite construir unos conjuntos de datos muy robustos para sustentar la generación de cartografía de interpolación.

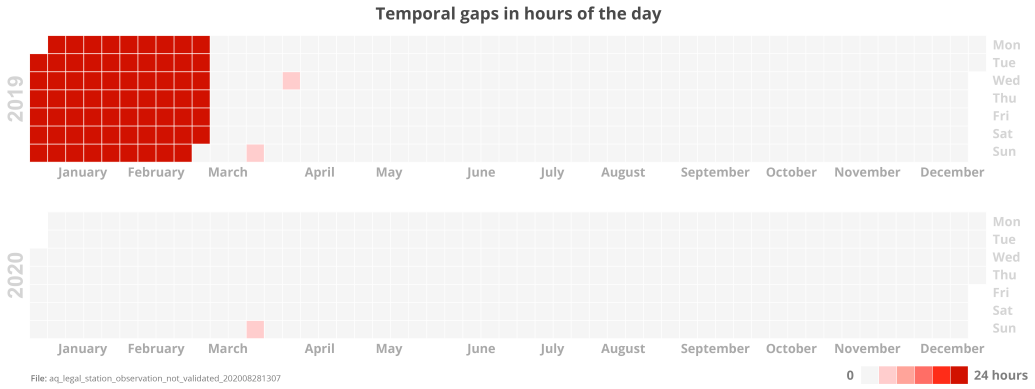

Figura 5.5: Cobertura temporal, observaciones no validadas.

aq_legal_station_observation_not_validated_202008281307
has 5 gaps, with an average gap duration of: 14 days and 18 hours

start	hours_duration	end
2019-01-01 00:00:00	1632	2019-03-10 00:00:00
2019-03-31 01:00:00	2	2019-03-31 03:00:00
2019-04-10 11:00:00	2	2019-04-10 13:00:00
2019-05-12 09:00:00	131	2019-05-12 09:00:00
2020-03-29 02:00:00	1	2020-03-29 03:00:00

Cuadro 5.7: Informe de cobertura temporal, observaciones no validadas.

La combinación de ambos conjuntos de datos (calibrados y no validados), desde el momento en que se estabilizó en el tiempo la captación de los datos desde los sensores, permite aumentar las capacidades de los segundos con una mayor extensión espacial. En el Anexo D, de calidad de los datos, se expone más información acerca de este tema; principalmente en lo que respecta a la cobertura espacial de las observaciones desde ambas instalaciones (ver Figura D.1) y respecto del total teórico temporal en el que estuvieron activas (Cuadros D.1 y D.2). Al centrarse en la distribución espacial de las mediciones de los sensores (Figura D.4), se puede comprobar como existe un predominio de las observaciones recogidas en los sensores del Centro y el Edificio Lorenzo Normante en el Campus Río Ebro, al norte de la ciudad. También se observa que un gran número de los datos recogidos entre junio de 2019 y agosto de 2020 pertenecen a sensores situados en el centro de la

ciudad de Zaragoza. Esto conlleva que los mapas interpolados generados durante ese período contasen con muy poca aportación de los sensores de bajo coste.

Sin embargo, desde inicios de Julio, todos los sensores parece mantener un rendimiento aceptable, completando la gran mayoría de las fracciones de 10 minutos diarias, lo que resultará en conjuntos de datos más completos. Los datos disponibles desde las estaciones reguladas en la base de datos, en general, muestran porcentajes de horas completadas superiores al 80 % (ver Anexo D), con la excepción de la estación de Jaime Ferrán, situada al noreste de la ciudad, pero al disponer de una estación cercana en el Picarral, ese sector podría cubrir la falta de sensores desplegados.

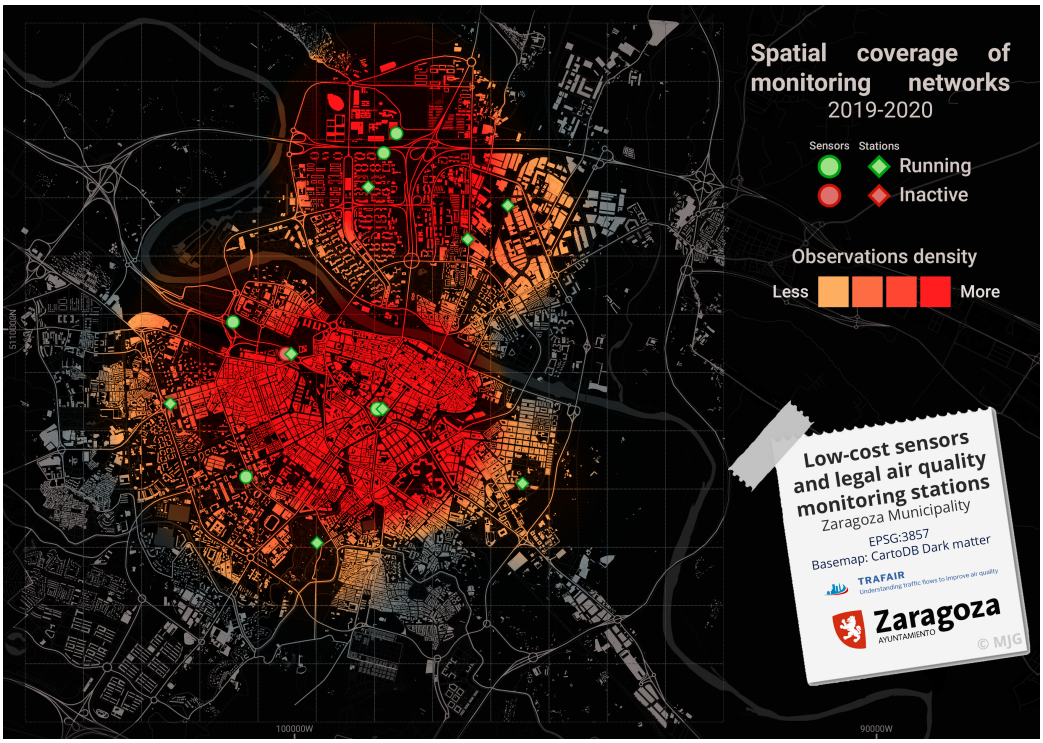

Figura 5.6: Cobertura espacial de la red combinada, 2019-2020.

La distribución de ambas redes se ha diseñado para ser parcialmente complementaria, pues los sensores buscan ampliar la cobertura de datos, pero también, calibrar los mismos con la información de las estaciones cercanas. En la Figura 5.6 se puede ver como la densidad de las observaciones aumenta ostensiblemente en comparación con la que presentan por separado las no validadas (Figura D.2) y calibradas (Figura D.3).

5.4. Comparativa de datos entre red de sensores del proyecto y estaciones legales

El propósito de esta sección es analizar la exactitud de las observaciones mediante la comparación de mapas de interpolación a partir de las diferentes fuentes de datos disponibles, haciendo uso de estadísticos de error, como el ya mencionado RMSE o la cartografía de los residuales, y la validación cruzada.

Las fuentes disponibles en la ciudad de Zaragoza son la red de estaciones legales del Ayuntamiento y los sensores desplegados por TRAFAIR. Ambas instalaciones generan diferentes conjuntos de datos con información sobre concentración de contaminantes; de los cuales se pueden utilizar para generar cartografía de calidad del aire los siguientes:

Estaciones reguladas

- *aq_legal_station_observation_not_validated*, observaciones no validadas recogidas horariamente por las estaciones reguladas, con disponibilidad casi inmediata. Sin embargo, no siempre cuentan con todos los contaminantes, habitualmente solo CO, NO_2 y O_3.
- *aq_legal_station_observation_validated*, observaciones validadas recogidas horariamente por las estaciones reguladas, es decir, datos en origen no validados que han sido evaluados, contrastados y verificados. Tienen una disponibilidad más retrasada, de varias semanas o meses, pero ya disponen de información para los cuatro contaminantes incluido NO.
- *aq_legal_station_observation_one_minute_not_validated*, observaciones no validadas recuperadas cada minuto. Se descarta dado que no se dispone de ningún dato en la BB.DD.

Sensores de bajo coste

- *sensor_raw_observation*, observaciones brutas obtenidas desde los sensores con disponibilidad inmediata. Al no estar calibradas, no pueden utilizarse para generar mapas de calidad del aire coherentes.
- *sensor_calibrated_observation*, observaciones tras aplicar los algoritmos de calibración. Tienen disponibilidad inmediata y se almacenan en tiempo real.

La metodología actual recupera los datos almacenados para las observaciones calibradas de los sensores y las no validadas de las estaciones legales, los

combina y genera la cartografía interpolada para cada contaminante.

Para poder analizar la exactitud de las observaciones calibradas de la red de sensores de bajo coste es necesario utilizar una fuente de referencia, que en este caso serán los datos validados de las estaciones.

En nuestro estudio, vamos a realizar dos comparaciones de los datos observados en la red de sensores del proyecto. Por un lado, se va a comparar las interpolaciones utilizando exclusivamente los datos calibrados de la red de sensores con la fuente de referencia. Por otro lado, también es interesante comparar las interpolaciones que se generan actualmente y que combinan datos calibrados de los sensores con datos no validados de las estaciones.

Para poder realizar el estudio comparativo se ha utilizado una fecha en la que estuvieran disponibles datos de todos los conjuntos, y un contaminante, NO_2, haciendo uso de los 14 puntos de observación de la red combinada de estaciones y sensores.

El motivo de restringir a un contaminante y fecha muy concreta (23/06/2020 10:00-11:00) es que desde TRAFAIR Zaragoza se han estado implementando los algoritmos de calibración a principios de Junio, con lo cual, a fecha de realización del presente trabajo solo hay datos calibrados a partir de ese momento. da también la casualidad de que los únicos datos de observaciones validadas de las estaciones legales disponibles son de Junio del 2020, y no para todos los días, concretamente solo existe una breve coincidencia entre ambos conjuntos de datos, la hora seleccionada. Por el mismo motivo, ya que no se disponen de todos los contaminantes del estudio cargados en observaciones validadas, solo se ha evaluado con NO_2. En resumen, el marco de los datos analizados se caracteriza por los siguientes parámetros.

- Fecha de observación: 2020/06/23
- Localización: Zaragoza, España
- 14 localizaciones de la red de sensores y estaciones legales, (POINTS).
- Contaminantes: NO_2 ($\mu g/m^3$).
- Período horario: 10:00-11:00

En las Figuras 5.7 y 5.8 se puede visualizar el comportamiento de los residuales respecto de los datos validados en la extensión de las coberturas ráster, calibrada y con datos combinados. Definido como la diferencia entre el valor observado y el estimado (interpolado), el valor residual permite cuantificar el error local y cartografiar su distribución.

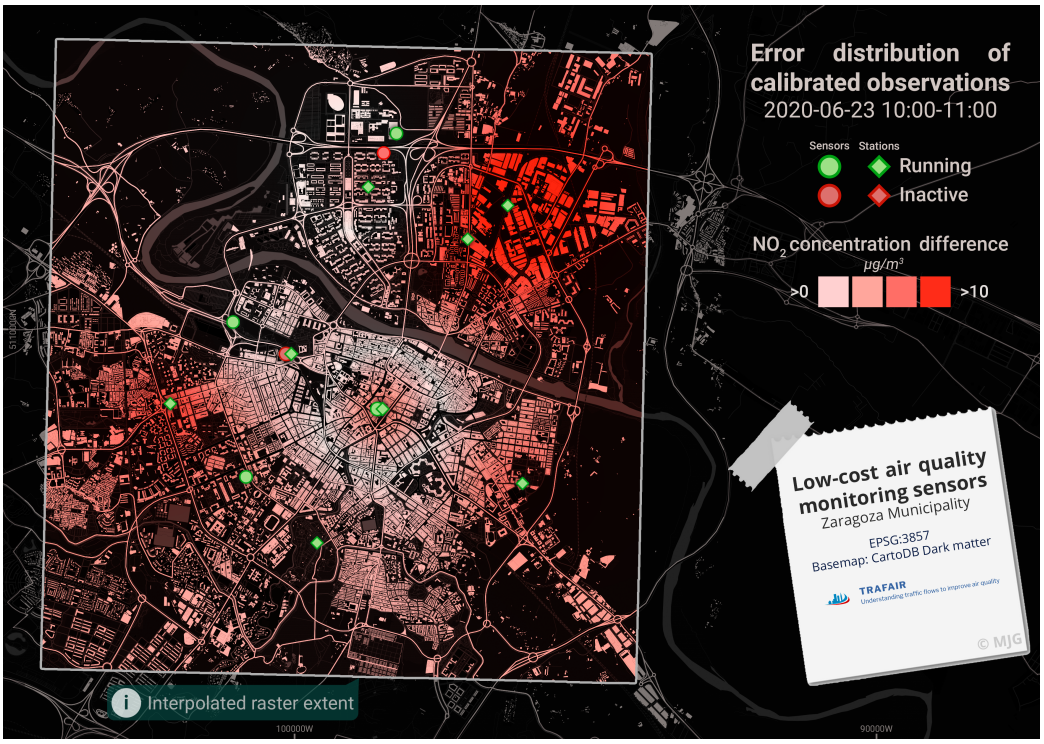

Figura 5.7: Mapa de error de las observaciones calibradas, sensores low-cost.

Figura 5.8: Mapa de error de las observaciones combinadas, sensor-estación.

POINT	CALIBRATED	COMBINED	MEAN RMSE
1	5.88	4.01	4.94
2	5.91	4.02	4.96
3	4.80	3.45	4.13
4	5.01	3.60	4.30
5	5.10	3.52	4.31
6	4.66	3.60	4.13
7	4.79	3.74	4.27
8	5.01	3.93	4.47
9	4.91	3.67	4.29
10	4.64	3.38	4.01
11	4.75	3.43	4.09
12	4.94	3.70	4.32
13	3.94	2.06	3.00
14	3.30	2.24	2.77
MEAN RMSE	4.83	3.45	4.14

Cuadro 5.8: RMSE de las observaciones de NO_2 calibradas y combinadas.

Como es lógico, a partir de las observaciones calibradas se predicen con dificultad datos más allá de su área de cobertura espacial. Generalmente para las localizaciones en las que no se encuentran sensores de bajo coste, al noreste, suroeste y sureste, la cuantificación del error es superior hasta alcanzar los 10 $\mu g/m^3$. Poniendo en perspectiva las magnitudes, un error de 10 $\mu g/m^3$, supone 14 veces menos el valor límite horario para la protección de la salud humana del dióxido de nitrógeno, que se sitúa en 140 $\mu g/m^3$. Esto permitiría, a priori, usar cartografía interpolada utilizando exclusivamente como fuente de origen los datos de los sensores, y con mayor motivo si se amplia la red de sensores desplegando los mismos en áreas con mayor error comparativo.

Las observaciones combinadas al contar con los datos no validados, que en su mayor parte pasan a ser verificados, responden de una manera más precisa tanto en el mapeo del error como en las pruebas de validación cruzada mediante el método de dejar un punto excluido (Cuadro 5.8).

Se puede observar que el error RMSE no se evaluó fácilmente donde se encuentran las instalaciones 1 y 2. Observando la cartografía de la Figura 5.9 podemos ver que estos sensores están localizados relativamente cerca uno del otro, ambos al norte de la ciudad. En general los valores extremos de los valores medidos en estas áreas de la ciudad difieren sensiblemente del resto. Por otro lado, se obtienen los valores más bajos en las instalaciones situadas en el centro (13 y 8).

La reubicación de los sensores en función de los resultados, y también, de la localización urbana de los mismos (carreteras, espacios industriales abiertos con poco tránsito, residenciales, etc.) requeriría un estudio más pormenorizado. Algunos datos evaluados simplemente pueden deberse al principio de

autocorrelación espacial derivado de su proximidad. Aunque el RMSE puede considerarse "bajo" en ambos supuestos, hay que tener en cuenta que se esta intentando abarcar un área bastante amplia (61 km^2) con 14 puntos de observación, pudiendo no resultar suficientes si se desean precisiones por debajo del $\mu g/m^3$.

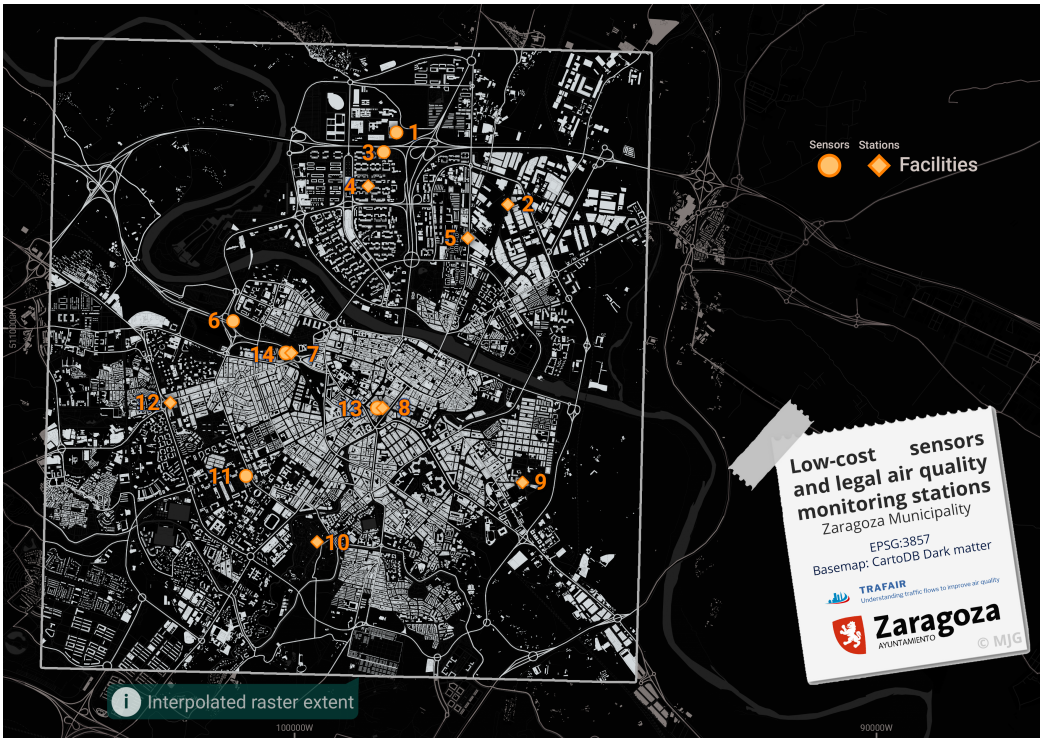

Figura 5.9: Identificadores de las instalaciones en el mapa de error.

6. Conclusiones y trabajo futuro

6.1. Conclusiones

El objetivo principal del presente trabajo era la realización de un flujo de trabajo coherente para el modelado, publicación y evaluación de los datos abiertos de monitorización de la calidad del aire. Este objetivo se considera satisfecho al generarse la cartografía de calidad del aire, estar presentes los conjuntos de datos en el European Data Portal y haberse constatado que los datos recolectados se adecuan a las especificaciones de datos propuestas para los temas tratados por la directiva INSPIRE y las métricas detalladas en la metodología MQA.

Aunque existen aspectos susceptibles de mejora, tal y como se indican en el siguiente apartado, en general se considera que el flujo puede ser personalizado, y extendido, para aplicarse en casos reales de otras ciudades o áreas urbanas diferentes al contexto del proyecto TRAFAIR en Zaragoza. La infraestructura planteada por el proyecto TRAFAIR y el flujo propuesto en este TFM permiten generar información y servicios que contribuyen a paliar las necesidades de seguimiento de los niveles de polución estipuladas por la legislación ambiental. De esta forma, se contribuye, en última instancia, a la mejora de la calidad de los entornos urbanos y de las comunidades que los habitan.

6.2. Trabajo futuro

En esta sección se plantean algunos puntos para la aplicación de mejoras o abrir nuevas líneas de trabajo a partir del presente TFM.

En primer lugar, se podrían remodelar la base de datos para incluir en los conjuntos de datos publicados aquellos aspectos propuestos por INSPIRE, dentro de las temáticas que engloban la monitorización de calidad del aire, que han quedado excluidos total o parcialmente, en particular, aquellos relativos a la información adicional sobre actividades, programas y redes adscritas.

En segundo lugar, respecto a la distribución de datos abiertos en formato RDF, convendría mejorar la definición de los contaminantes mediante la incorporación de vocabulario de (O_3, NO, NO_2, CO y NO_x) procedente del EIONET AQD air quality pollutants vocabulary.[1]

En tercer lugar, sería conveniente ampliar el estudio comparativo de los datos obtenidos a partir de los sensores de bajo coste y las estaciones legales con períodos temporales más amplios y distribuidos, también, con todos los contaminantes. Somos conscientes de que la metodología habitual requiere elegir días y horas específicas para presentar las capacidades del mapa de interpolación y sus resultados estadísticos. También hay que tener en cuenta el hecho de que han de elegirse franjas horarias concretas para cada contaminante, dado que siguen una tendencia diaria diferente en un entorno urbano [36]. Por ejemplo, se podrían elegir las 18:00-19:00 para el NO_2, frente a las 10:00-11:00 seleccionadas como único recurso.

También sería interesante evaluar el comportamiento de las observaciones calibradas de los sensores para el NO, al ser el contaminante que más tarda en estar disponible para las observaciones de las estaciones reguladas validadas. Actualmente, las observaciones de NO de los sensores de bajo coste son la única fuente de datos de la mayor parte de la cartografía de interpolación generada. Por ejemplo, los últimos datos validados de NO provenientes de estaciones reguladas en la base de datos son del 27 de Abril de 2020, cuando todavía no se habían aplicado procesos de calibración correctos a la información bruta de los sensores (junio de 2020).

En cuarto lugar, sería útil simplificar los procesos de evaluación de la cobertura temporal incorporando una conexión directa a la información disponible en la base de datos mediante las librerías de Python, psycopg2 o PyGreSQL, y con la implementación de tareas periódicas mensuales, cada trimestre o año, para realizar un seguimiento de la calidad de los datos. En el mismo sentido, se deberían desarrollar scripts o plugins en Python, haciendo uso de PyQGIS[2], para analizar la evolución de la cobertura espacial de los datos de

[1] `http://dd.eionet.europa.eu/vocabulary/aq/pollutant/`

[2] Python scripting in QGIS, Developer Cookbook. `https://docs.qgis.org/3.10/en/docs/pyqgis_developer_cookbook/index.html`

manera periódica.

Finalmente, otra línea de trabajo futura podría ser la evaluación de la calidad de los datos espaciales mediante experimentos destinados a valorar la consistencia topológica y la exactitud posicional de los registros.

Bibliografía

[1] L. M. Bretón, R. Trillo, J. Fabra, J. Nogueras y M. U. Alzueta, «TRAFAIR: Análisis de los flujos de tráfico para mejorar la calidad del aire urbano», es, *Jornada de Jóvenes Investigadores del I3A*, vol. 7, mayo de 2019, ISSN: 2341-4790. DOI: 10.26754/jji-i3a.003603.

[2] P. A. Johnson, R. Sieber, T. Scassa, M. Stephens y P. Robinson, «The Cost(s) of Geospatial Open Data», en, *Transactions in GIS*, vol. 21, n.º 3, págs. 434-445, 2017, ISSN: 1467-9671. DOI: 10.1111/tgis.12283.

[3] European Union y Publications Office, «Open data maturity report 2019.», en, inf. téc. 978-92-78-42052-9, 2020, OCLC: 1140696689. dirección: https://op.europa.eu/publication/manifestation_identifier/PUB_OABE19001ENN.

[4] A. Abella, «La reutilización de datos abiertos en España II», Universidad Rey Juan Cuarlos, inf. téc., 2019. dirección: https://www.desidedatum.com/wp-content/uploads/2019/12/La_reutilizacio%CC%81n_datos_abiertos_en_espan%CC%83a_2019.pdf.

[5] E. Commission, *GeoDCAT Application profile for data portals in Europe, GeoDCAT-AP v1.0.1*, 2016. dirección: https://joinup.ec.europa.eu/release/geodcat-ap/101.

[6] I. O. for Standardization (ISO), «ISO 19115-1:2014, Geographic information — Metadata — Part 1: Fundamentals», Geneva, CH, inf. téc., 2014. dirección: https://www.iso.org/obp/ui/fr/#iso:std:iso:19115:-1:ed-1:v1:en.

[7] J. Nogueras-Iso, H. Ochoa-Ortiz, M. Á. Jañez, J. R. R. Viqueira, L. Po y R. Trillo-Lado, «Automatic publication of Open Data from OGC services: the use case of TRAFAIR project», Enviado a *The Twelfth International Conference on Advanced Geographic Information Systems, Applications, and Services - GEOProcessing 2020*, 2020.

[8] J. Tandy, L. van den Brink y P. Barnaghi, «Spatial data on the web best practices», *W3C Working Group Note*, 2017. dirección: https://www.w3.org/TR/sdw-bp/.

[9] L. van den Brink, P. Barnaghi, J. Tandy, G. Atemezing, R. Atkinson, B. Cochrane, Y. Fathy, R. García Castro, A. Haller, A. Harth, K. Janowicz, Kolozali, B. van Leeuwen, M. Lefrançois, J. Lieberman, A. Perego, D. Le-Phuoc, B. Roberts, K. Taylor y R. Troncy, «Best practices for publishing, retrieving, and using spatial data on the web», en, *Semantic Web*, vol. 10, n.º 1, págs. 95-114, ene. de 2019, Publisher: IOS Press, ISSN: 1570-0844. DOI: 10.3233/SW-180305.

[10] J. R. R. Viqueira, S. Villarroya, D. Mera y J. A. Taboada, «Smart Environmental Data Infrastructures: Bridging the Gap between Earth Sciences and Citizens», en, *Applied Sciences*, vol. 10, n.º 3, pág. 856, ene. de 2020, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app10030856.

[11] d. A. Xavier y E. Magnus, *Automatic evaluation of geospatial data quality using web services*, spa. Jaén : Universidad de Jaén, jul. de 2018, Accepted: 2018-07-05T07:54:22Z. dirección: http://hdl.handle.net/10953/877.

[12] International Organization for Standardization (ISO), «ISO 19157:2013, Geographic information — Data quality», en, Geneva, CH, inf. téc., 2013, Library Catalog: www.iso.org. dirección: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/25/32575.html.

[13] A. Jakobsson, «Data Quality and Quality Management -Examples of Quality Evaluation Procedures and Quality Management in European National Mapping Agencies», ene. de 2002. dirección: http://lib.tkk.fi/Diss/2006/isbn9512282062/article3.pdf.

[14] T. P. G. D. Group, *PostgreSQL 11 documentation, SQL Conformance.* dirección: https://www.postgresql.org/docs/11/features.html.

[15] ——, *PostgreSQL 11 documentation.* dirección: https://www.postgresql.org/docs/11/index.html.

[16] P. P. S. Committee, *PostGIS 2.5 documentation.* dirección: https://postgis.net/docs/manual-2.5/.

[17] O. S. G. Foundation, *Geoserver 2.16.0 documentation.* dirección: http://geoserver.org/release/2.16.0/.

[18] C. Association, *CKAN 2.8 documentation.* dirección: https://docs.ckan.org/en/2.8/.

[19] T. P. D. Team, *Pandas 1.0.5 documentation.* dirección: https://pandas.pydata.org/docs/.

[20] T. P. S. Foundation, *Python 3.8 documentation.* dirección: `https://docs.python.org/3.8/`.

[21] Q. D. Team, *QGIS 3.10 documentation.* dirección: `https://docs.qgis.org/3.10/es/docs/user_manual/`.

[22] R. Cyganiak, *Tarql: SPARQL for Tables documentation.* dirección: `https://tarql.github.io/`.

[23] U. P. Center, *NetCDF (network Common Data Form) documentation.* dirección: `https://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit`.

[24] *LaTeX - A document preparation system.* dirección: `https://www.latex-project.org/`.

[25] T. P. team, *PyGreSQL 5.2 documentation.* dirección: `https://pygresql.org/contents/index.html`.

[26] P. S. Foundation, *SPARQLWrapper documentation.* dirección: `https://pypi.org/project/SPARQLWrapper/`.

[27] H. Ochoa Ortiz, «Desarrollo de mecanismos de publicación de datos para el estudio de la calidad del aire en entornos urbanos», 2020. dirección: `https://deposita.unizar.es/record/54739?ln=en`.

[28] O. Corcho, *Vocabulario sobre calidad del aire.* dirección: `http://vocab.linkeddata.es/datosabiertos/def/medio-ambiente/calidad-aire`.

[29] S. Janssen, C. Guerreiro, P Viane, E. Georgieva, P. Thunis, K Cuvelier, E Trimpeneers, J Wesseling, A Montero, A Miranda y col., *Guidance Document on Modelling Quality Objectives and Benchmarking– FAIRMODE WG1*, 2017. dirección: `https://op.europa.eu/s/oflA`.

[30] M. F. Goodchild, *Fundamentals of spatial data quality.* John Wiley & Sons, 2010, vol. 662.

[31] *INSPIRE Data Specification for the spatial data theme Environmental Monitoring Facilities*, 10 de dic. de 2013. dirección: `https://inspire.ec.europa.eu/Themes/120/2892`.

[32] *INSPIRE Data Specification for the spatial data theme Atmospheric Conditions and Meteorological Geographical Features*, 10 de dic. de 2013. dirección: `https://inspire.ec.europa.eu/Themes/141/2892`.

[33] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne y col., «The FAIR Guiding Principles for scientific data management and stewardship», *Scientific data*, vol. 3, n.º 1, págs. 1-9, 2016. DOI: `10.1038/sdata.2016.18`.

[34] Biblioteca del Congreso de los EEUU - Oficina de desarrollo de Redes y Normas MARC, *Normas MARC.* dirección: `https://www.loc.gov/marc/marcspa.html`.

[35] DCMI Usage Board, «DCMI Metadata Terms», 2020. dirección: `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`.

[36] C Borrego, A. Costa, J Ginja, M Amorim, M Coutinho, K Karatzas, T. Sioumis, N Katsifarakis, K Konstantinidis, S De Vito y col., «Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise», *Atmospheric Environment*, vol. 147, págs. 246-263, 2016.

Anexos

A. Estadístico de Error RMSE de las metodologías de interpolación

- Fecha de observación: 2020/03/19
- Localización: Módena, Italia
- 9 Estaciones reguladas, (*Points*).
- Contaminantes: NO_2 ($\mu g/m^3$), O_3 ($\mu g/m^3$) y CO (mg/m^3).
- Períodos horarios: 7:00-8:00, 13:00-14:00 y 18:00-19:00

POINT	OK	IDW2	IDW1	OAK	NN-3	NN-6	SPLINE	AVG	IDW2-OK	MEAN RMSE
1	15.22	17.39	13.13	14.31	14.47	9.25	27.19	14.31	15.88	15.68
2	20.17	19.34	19.26	19.18	20.56	18.19	19.51	19.18	19.76	19.46
3	6.28	7.32	9.28	12.81	8.62	9.77	10.16	12.81	6.79	9.31
4	17.52	16.12	14.79	12.85	17.41	16.32	29.35	12.85	16.83	17.12
5	11.98	6.29	9.13	13.40	14.17	6.91	13.33	13.40	9.19	10.87
6	47.06	45.93	46.29	47.37	39.71	45.39	32.27	47.37	45.51	44.10
7	39.42	42.86	37.09	30.72	52.85	35.42	45.75	30.73	41.08	39.55
8	11.63	10.92	11.82	15.56	14.88	13.51	10.16	15.56	11.27	12.81
9	46.04	48.12	43.17	41.60	32.87	40.55	54.58	41.60	47.06	43.95
MEAN RMSE	23.92	23.81	22.66	23.09	23.95	21.70	26.92	23.09	23.71	23.65

Cuadro A.1: RMSE de los contaminantes promediados para los diferentes métodos.

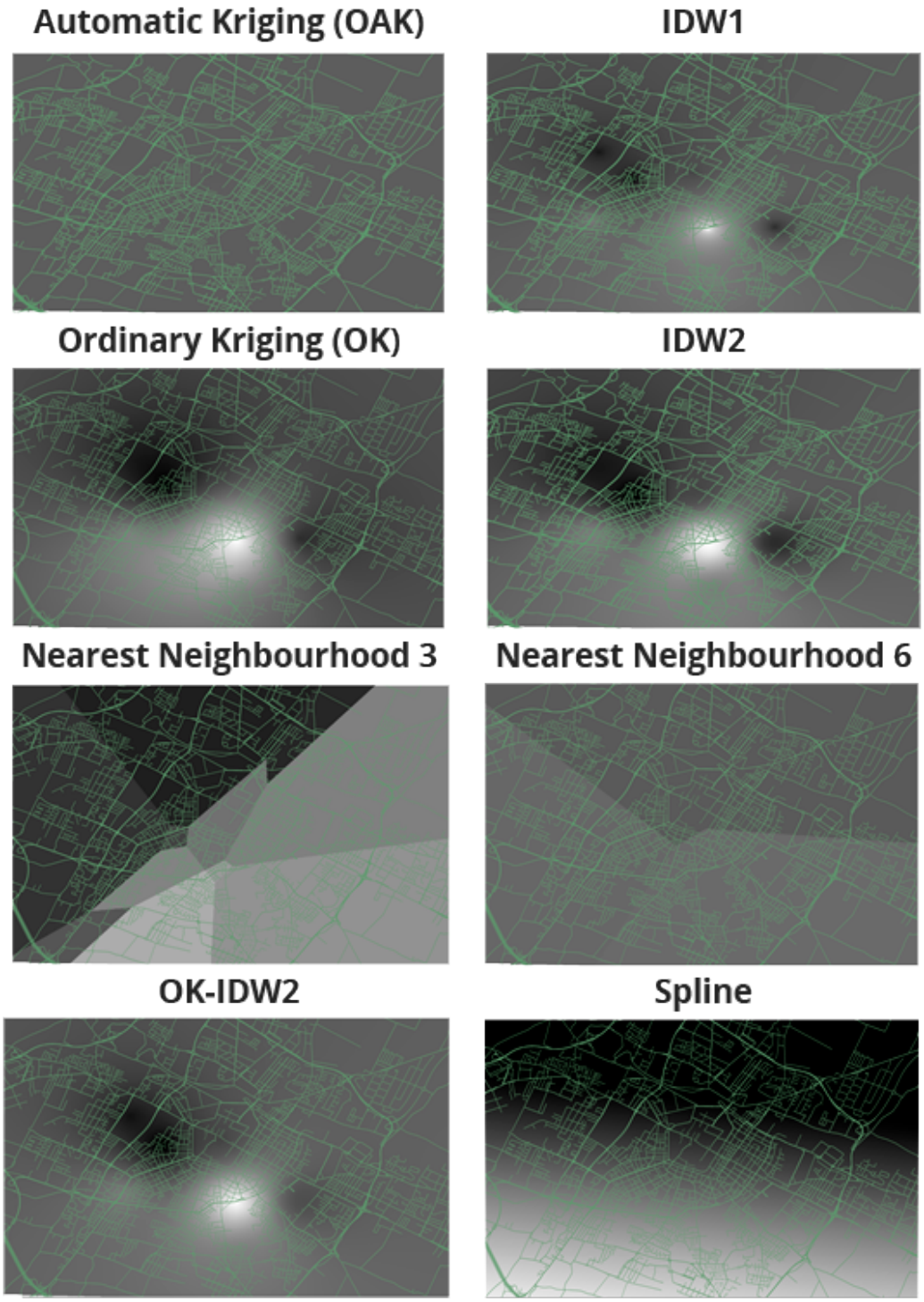

Figura A.1: Comparación superficies de interpolación del NO_2, TRAFAIR.

B. Resumen de la actualización de los datos de calidad del aire

B.1. Tiempo real

Datasets	Type	General update policy
Air quality observation metadata	Feature type	This dataset is updated every time a new calibration process is performed, by adding information about that new calibration process applied.
Real time air quality observation	Feature type	This dataset is updated every time new calibrated data is inserted.
Real time hourly air quality observation	Feature type	This dataset is updated every time new calibrated data is inserted, recomputing the average values during the current fraction of the last hour.
Real time air quality observations coverage	Coverage	This spatial coverage is updated every 10 minutes to reflect an up-to-date estimation of the concentrations.

Cuadro B.1: Conjuntos de datos en tiempo real, Zaragoza. TRAFAIR.

B.2. Históricos

Datasets	Type	General update policy
Historic air quality observations by year	Feature type	A dataset is created once per year (the year is appended with a "_" to the name of the dataset, with format "yyyy").
Hourly air quality observations	Feature type	This dataset is updated once every hour, by inserting the average values of pollutants measured in the corresponding location during the previous hour. The hourly air quality observations will be kept at least during 2 days, in such a way that the hourly daily observations for the last 48 hours will be available.
Hourly air quality observations coverage	Coverage	There is a spatio-temporal dataset consisting of an image mosaic to store the hourly aggregate values of pollutants in the form of coverages. The hourly coverages for the last 8 days are maintained.
Daily air quality observations coverage *year month*	Coverage	A dataset is created once per month (the aggregate function, year and month are appended with a "_" to the name of the dataset, with format "yyyy" and "mm", respectively).

Cuadro B.2: Conjuntos de datos históricos, Zaragoza. TRAFAIR.

B.3. Predicción

Datasets	Type	General update policy
Latest air quality prediction coverage	Coverage	This spatio-temporal dataset is updated once a day, with the hourly air quality predictions for the following 48 hours.
Air quality prediction coverage by date	Coverage	A dataset is created once per day (the date is appended with a "_" to the name of the dataset, with format "yyyymmdd"). We will keep the spatio-temporal coverages with the GRAL predictions of the last 7 days, in addition to the coverage Latest air quality prediction coverage. So, a total of 8 days of predictions (the coverages of the last 8 days) will be maintained at any given moment.

Cuadro B.3: Conjuntos de datos de predicción, Zaragoza. TRAFAIR.

C. Conversión de archivos CSV en datos semánticos RDF

Se descarga la herramienta *Tarql: SPARQL for Tables* de:
`https://github.com/tarql/tarql/releases`

Se obtiene un archivo CSV del Geoserver mediante una petición *GetFeature*:
`http://atila.unizar.es:8081/geoserver/open_data/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=open_data%3Ahistoric_air_quality_observations_2019&outputFormat=csv`

Se genera un archivo `query.sparql` con la consulta CONSTRUCT para aplicar en el archivo de entrada:

```
PREFIX esair: <http://vocab.linkeddata.es/datosabiertos/
    def/medio-ambiente/calidad-aire>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
PREFIX sosa: <http://www.w3.org/ns/sosa>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

CONSTRUCT {
  ?u geo:Feature ?FID;
  esair:AirQualitySensor ?featureid;
  sosa:phenomenonTime ?phenomenon_time;
  sosa:Result ?result_time;
  esair:monoxidoDeNitrogeno ?no;
  sosa:Procedure ?no_provenance;
  esair:dioxidoDeNitrogeno ?no2;
  sosa:Procedure ?no2_provenance;
  esair:monoxidoDeCarbono ?co;
  sosa:Procedure ?co_provenance;
  esair:ozono ?o3;
  sosa:Procedure ?o3_provenance;
  sosa:Sampling ?coverage;

  }
  FROM <file:historic_air_quality_observations_2019.csv>
  WHERE {
    BIND (UUID() AS ?u)
  }
```

Se ejecuta en el símbolo de sistema y se exporta el resultado a un archivo de texto:

```
tarql query.sparql historic_air_quality_observations_2019.csv > resultado.txt
```

Mediante RDF Grapher (`http://www.ldf.fi/service/rdf-grapher`) podemos visualizar los datos RDF en un gráfico. Por ejemplo, la Figura C.1 muestra el gráfico para 2 registros.

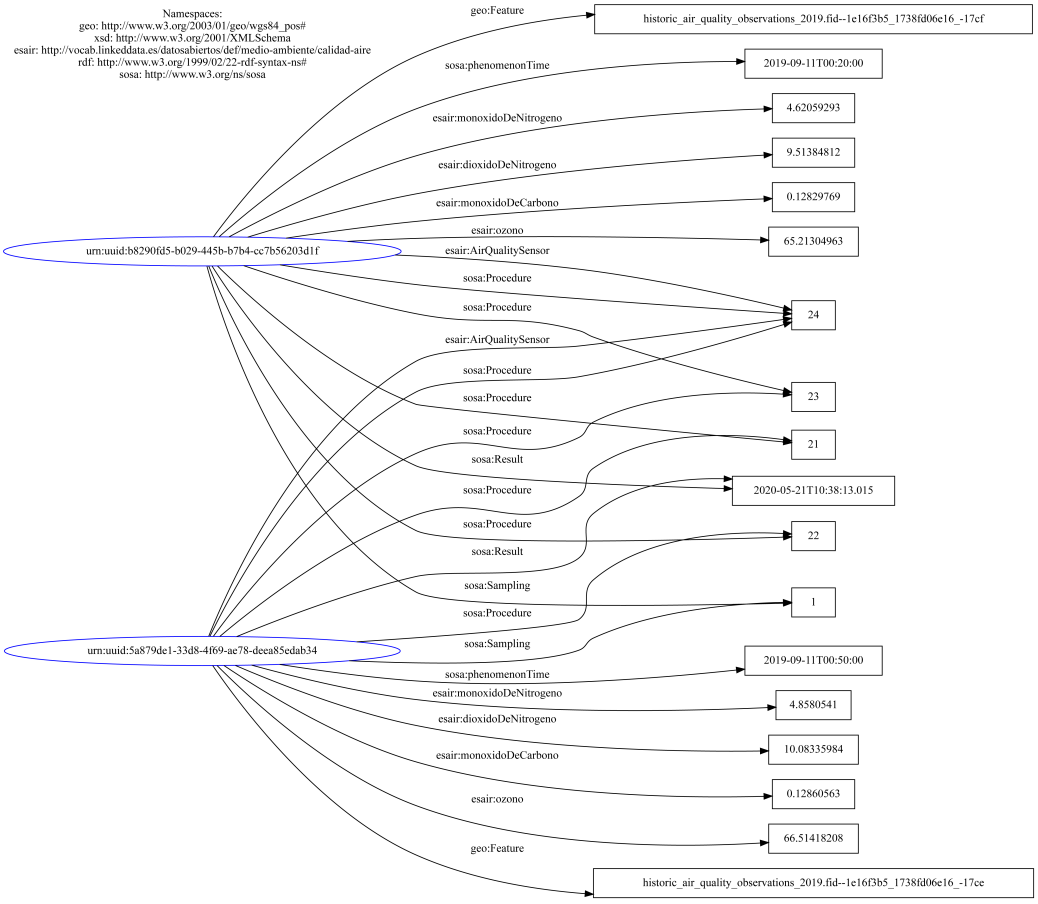

Figura C.1: Visualización de 2 registros RDF.

D. Anexo de calidad de los datos

Figura D.1: Distribución de la red combinada de estaciones y sensores.

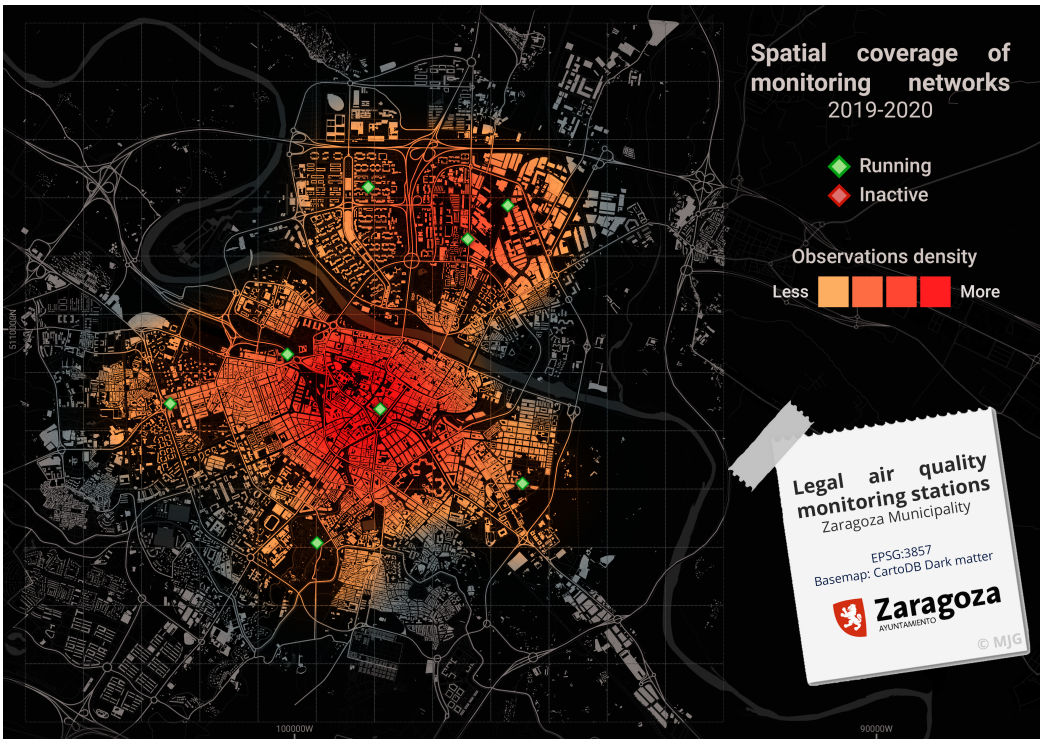

Figura D.2: Cobertura espacial estaciones legales, 2019-2020.

Air Quality Legal Station	% Total hours	Observations	Days with observations	Theoretical days
El Picarral	94.1 %	12,126	520	537
Roger de Flor	91.7 %	11,813	500	537
Jaime Ferrán	64.8 %	8,352	400	537
Renovales	86.5 %	11,148	473	537
Las Fuentes	84.7 %	10,915	466	537
Centro	242.3 %	31,229	473	537
Avenida Soria	108.8 %	14,024	471	537
Actur	87.1 %	11,224	477	537

Cuadro D.1: Observaciones y total teórico de estaciones legales, 2019-2020.

Nota: Las estaciones de Avenida de Soria y Centro a veces incluyen observaciones cada 15 mins.

Las mediciones fijas requieren una captura mínima de datos del 90 % de los valores teóricos según lo establecido en la Directiva 2008/50/CE, [3] sin contar procesos de calibración o mantenimiento (-5 %). Se ha tenido en cuenta el valor del 85 % como referencia en los formatos de los Cuadros D.1 y D.2.

[3] Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008 , relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa. Anexo I. `https://eur-lex.europa.eu/eli/dir/2008/50/oj`

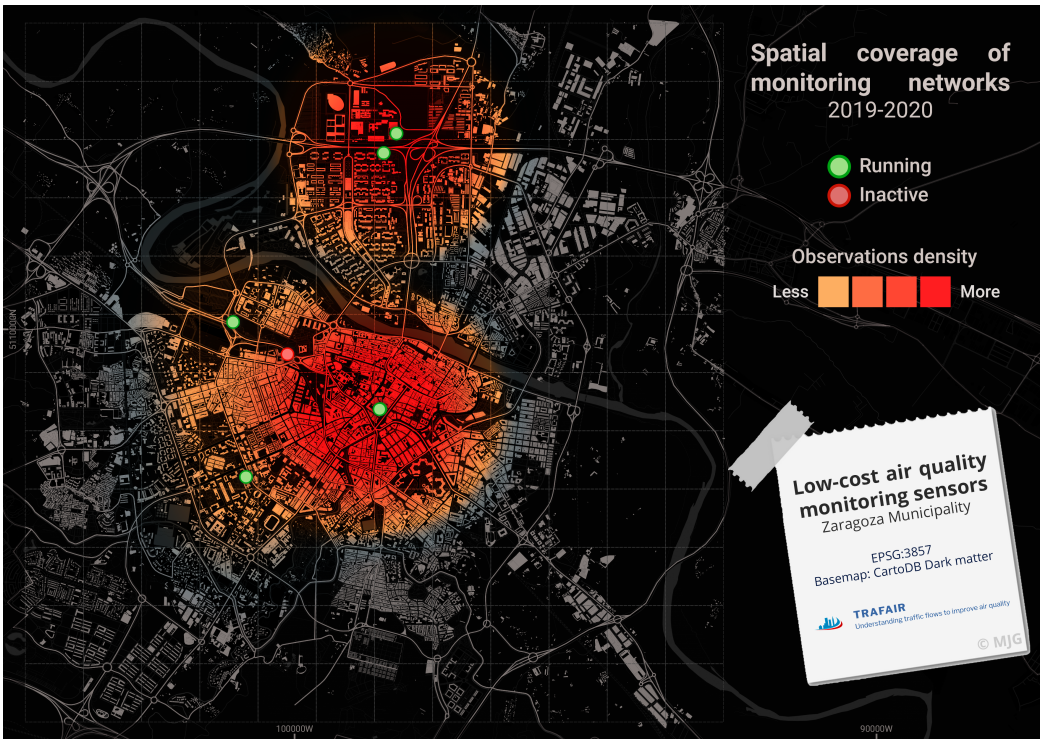

Figura D.3: Cobertura espacial sensores low-cost, 2019-2020.

Air Quality Low-cost Sensor	% Total 10 mins	10 mins observations	Days with observations	Theoretical days
Centro	67.0 %	43,439	286	324
Edif. Lorenzo Normante	75.8 %	23,289	140	177
Edificio Etiopía	81.7 %	7,694	79	99
Facultad de Estudios Sociales	71.4 %	7,189	79	99
Autovía Z-40	77.1 %	7,917	79	99

Cuadro D.2: Observaciones y total teórico de sensores low-cost, 2019-2020.

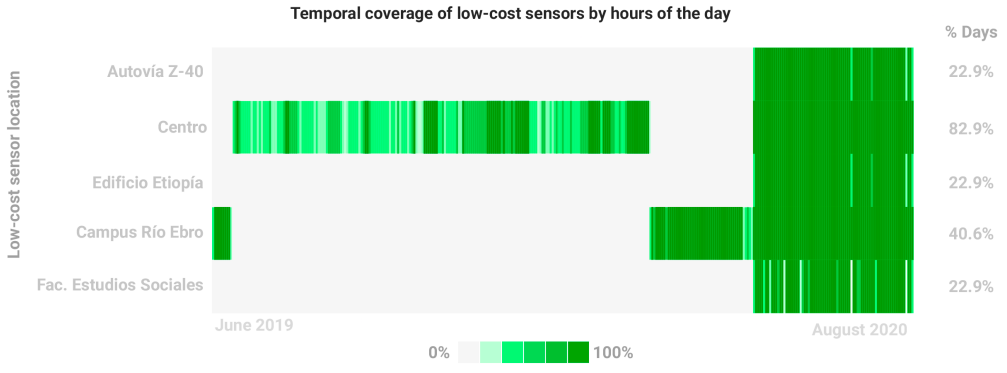

Figura D.4: Cobertura temporal de sensores low-cost, 2019-2020.

D.1. Comparación estadísticas de validación

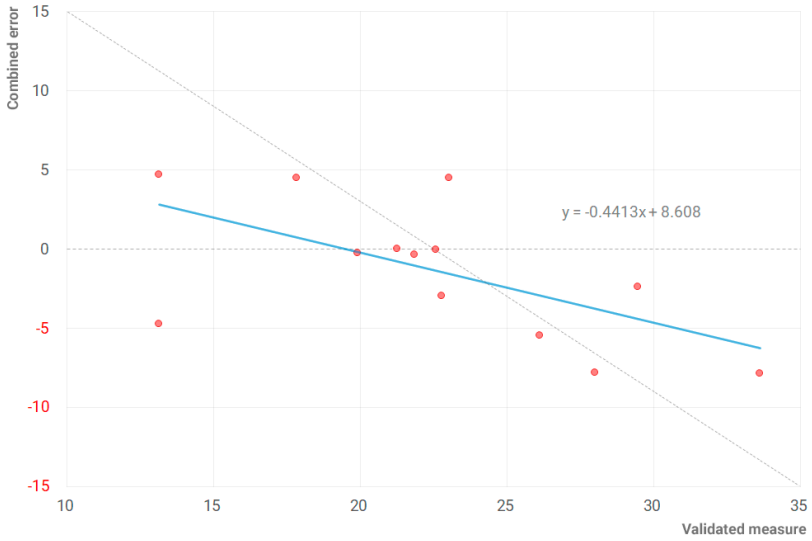

Figura D.5: Validación cruzada del error de observaciones combinadas y validadas.

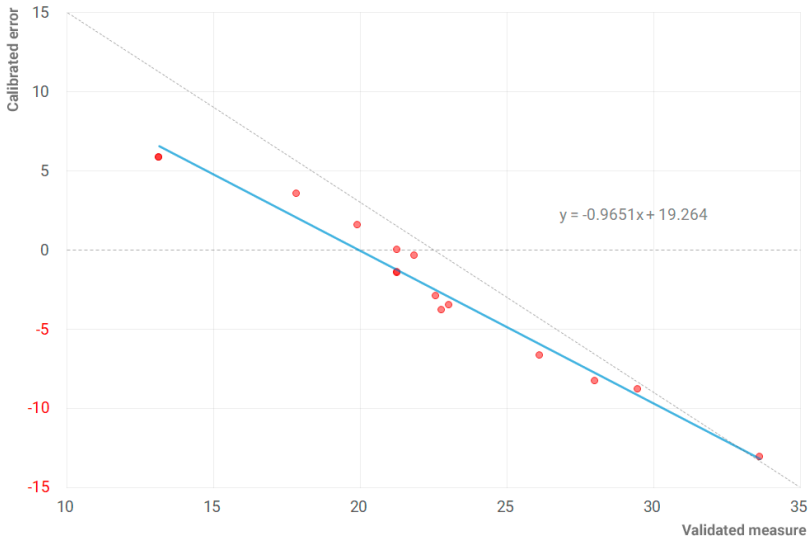

Figura D.6: Validación cruzada del error de observaciones calibradas y validadas.

D.2. Estado de los sensores low-cost

Location	ID Sensor low cost	Date time	Status
Centro	34	22/07/2019 7:00	calibration
Centro	38	22/07/2019 7:00	calibration
Centro	42	22/07/2019 7:00	calibration
Edificio Lorenzo Normante, Campus Río Ebro	47	22/07/2019 7:00	calibration
Edificio Lorenzo Normante, Campus Río Ebro	23	22/07/2019 10:45	running
Edificio Lorenzo Normante, Campus Río Ebro	26	22/07/2019 10:45	running
Edificio Lorenzo Normante, Campus Río Ebro	31	22/07/2019 10:45	running
Centro	52	22/07/2019 10:45	running
Centro	55	22/07/2019 10:45	running
Centro	24	01/08/2019 6:00	calibration
Centro	27	01/08/2019 6:00	calibration
Centro	32	01/08/2019 6:00	calibration
Edificio Lorenzo Normante, Campus Río Ebro	53	01/08/2019 6:00	calibration
Centro	56	01/08/2019 6:00	calibration
Edificio Lorenzo Normante, Campus Río Ebro	35	01/08/2019 7:30	running
Edificio Lorenzo Normante, Campus Río Ebro	39	01/08/2019 7:30	running
Edificio Lorenzo Normante, Campus Río Ebro	43	01/08/2019 7:30	running
Autovía Z-40	48	01/08/2019 7:30	running
Edificio Lorenzo Normante, Campus Río Ebro	18	01/08/2019 7:30	running
Edificio Etiopía	36	27/09/2019 7:30	running
Facultad de Estudios Sociales	44	27/09/2019 7:30	calibration
Facultad de Estudios Sociales	20	06/11/2019 16:00	running
Centro	45	06/11/2019 16:00	running
Centro	21	03/12/2019 7:30	calibration
Autovía Z-40	57	03/12/2019 7:30	calibration
Edificio Lorenzo Normante, Campus Río Ebro	28	04/12/2019 14:00	running
Autovía Z-40	49	09/12/2019 18:00	running
Centro	50	06/02/2020 18:45	running
Edificio Lorenzo Normante, Campus Río Ebro	29	06/03/2020 8:30	running
Edificio Lorenzo Normante, Campus Río Ebro	40	06/03/2020 8:30	running
Edificio Etiopía	30	11/03/2020 10:30	running
Centro	37	13/03/2020 7:30	calibration
Centro	41	13/03/2020 7:30	calibration
Centro	46	13/03/2020 7:30	calibration
Edificio Lorenzo Normante, Campus Río Ebro	51	13/03/2020 7:30	calibration
Centro	22	13/03/2020 8:00	calibration
Facultad de Estudios Sociales	33	13/03/2020 8:45	running
Edificio Lorenzo Normante, Campus Río Ebro	54	13/03/2020 9:45	running
Edificio Lorenzo Normante, Campus Río Ebro	25	13/03/2020 10:30	running
Autovía Z-40	58	13/03/2020 13:30	running
Autovía Z-40	59	30/07/2020 9:35	running

Cuadro D.3: Tabla con el estado de los sensores de bajo coste, 2019-2020.

E. Anexo de cartografía de interpolación

Simple feature collection with 19 features

Date: 2020/07/28 00:00:00 to 2020/08/28 00:00:00

Location: Zaragoza, España

Geometry type: Points

Dimension: XY

Geographic CRS: WGS-84

Bounding box:

xmin: -0.9175043

ymin: 41.63415

xmax: -0.8631285

ymax: 41.68132

id	type	location	co	no	no2	o3	geom
54	sensor	Edificio Lorenzo Normante	0.1364489	14.11076	22.46895	58.94089	POINT(-0.88259 41.68133)
41	sensor	Centro	0.1946712	5.547519	21.31474	64.72341	POINT(-0.885197 41.64955)
51	sensor	Centro	0.1898862	5.346895	24.91606	61.03284	POINT(-0.885197 41.64955)
46	sensor	Centro	0.2123179	9.867591	21.11018	55.3795	POINT(-0.885197 41.64955)
25	sensor	Edificio Lorenzo Normante	0.1474715	14.07382	22.42816	63.8355	POINT(-0.88259 41.68133)
30	sensor	Edificio Etiopía	0.1539886	12.3515	23.29018	58.08917	POINT(-0.907822 41.6596)
22	sensor	Centro	0.2064424	9.238617	20.43076	56.53538	POINT(-0.885197 41.64955)
37	sensor	Centro	0.1910651	6.634216	24.26547	58.1343	POINT(-0.885197 41.64955)
59	sensor	Autovía Z-40	0.1500486	10.81631	27.29588	60.27651	POINT(-0.88456 41.67908)
33	sensor	Facultad de Estudios Sociales	0.1403186	13.14429	21.11212	66.98023	POINT(-0.90582 41.64175)
58	sensor	Autovía Z-40	0.1460237	11.27459	25.86849	61.89861	POINT(-0.88456 41.67908)
36	legal station	Renovales	0.2006534	NA	11.22631	68.14441	POINT(-0.8948568 41.63415)
29	legal station	Roger de Flor	0.223338	NA	16.47555	59.18572	POINT(-0.9175043 41.65019)
38	legal station	Centro	0.2159666	NA	10.83983	66.15015	POINT(-0.8850941 41.64957)
40	legal station	Actur	NA	NA	11.25917	63.50748	POINT(-0.886945 41.67516)
26	legal station	El Picarral	0.190195	NA	19.71806	58.81607	POINT(-0.8715919 41.66916)
37	legal station	Las Fuentes	0.155375	NA	16.43519	65.91849	POINT(-0.8631285 41.641)
32	legal station	Jaime Ferrán	0.1375	NA	16.39186	56.04435	POINT(-0.8654079 41.67303)
39	legal station	Avenida de Soria	0.1802639	NA	16.6519	65.82586	POINT(-0.899425 41.65588)

Cuadro E.1: Promedio medidas no validadas-calibradas, 28/07-28/08/2020.

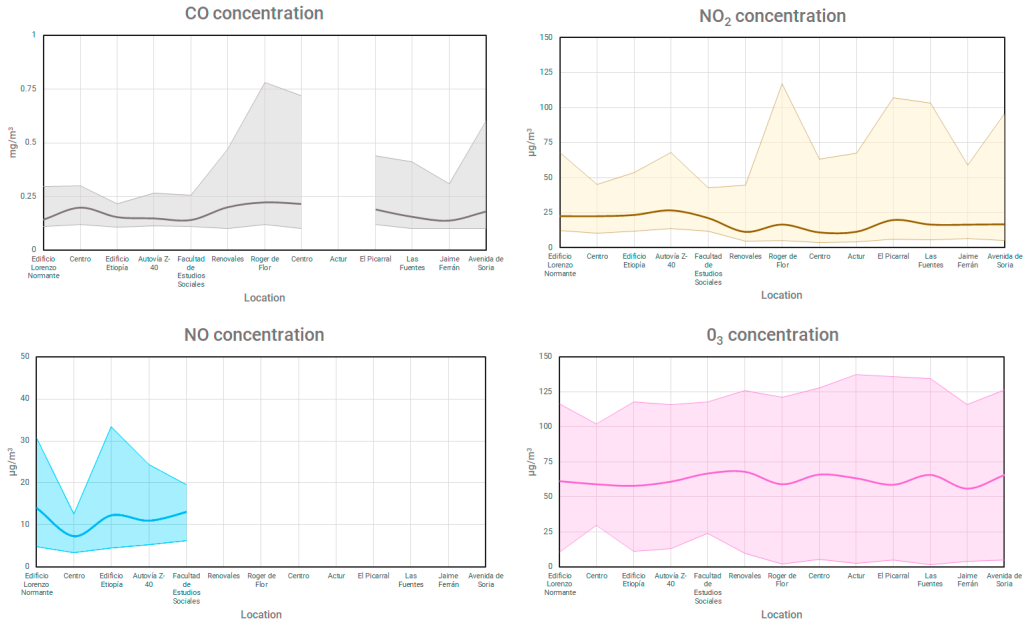

Figura E.1: Valor promedio y extremos de los contaminantes, 28/07-28/08 de 2020.

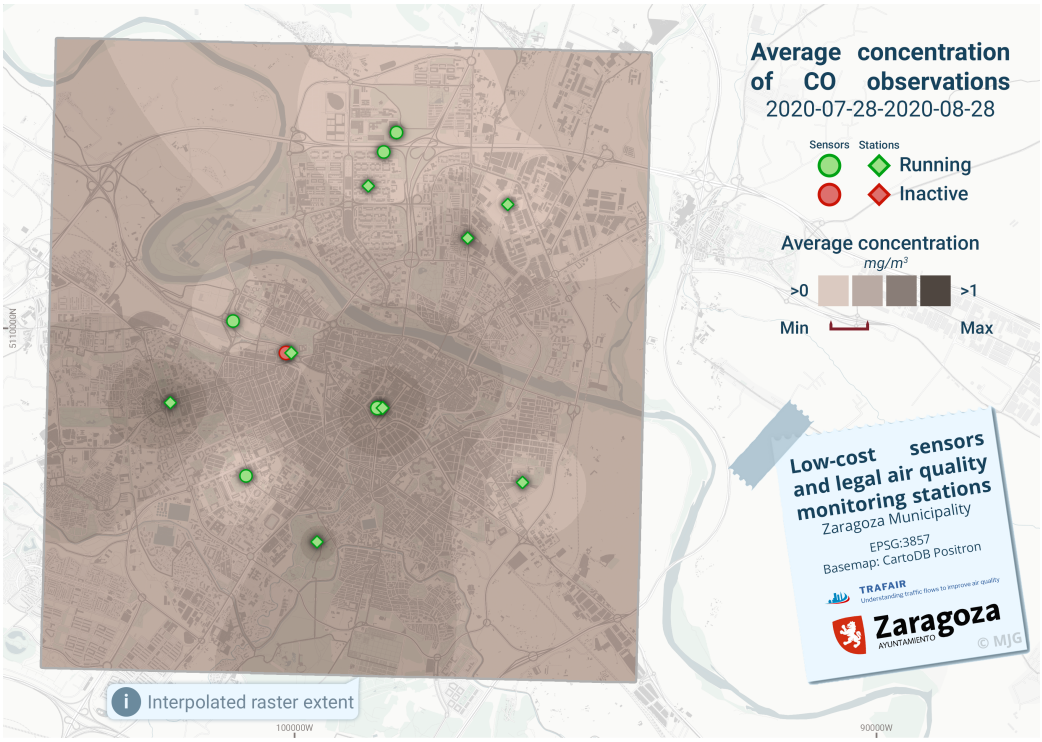

Figura E.2: Valor promedio de observaciones de CO, 28/07-28/08 de 2020.

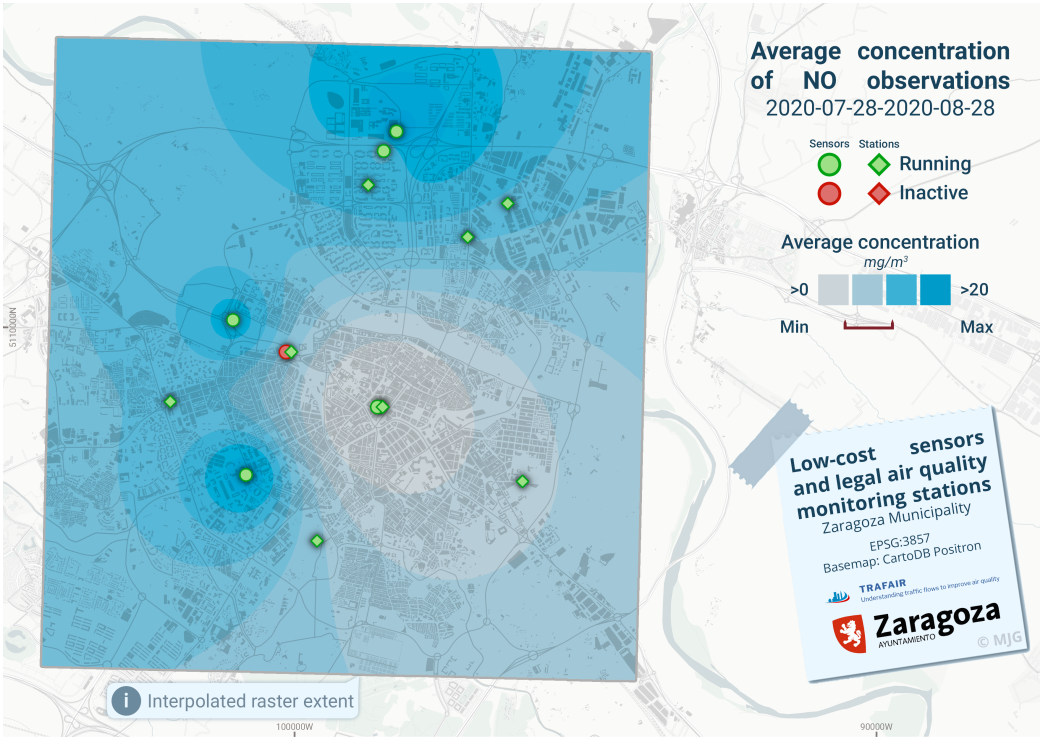

Figura E.3: Valor promedio de observaciones de NO, 28/07-28/08 de 2020.

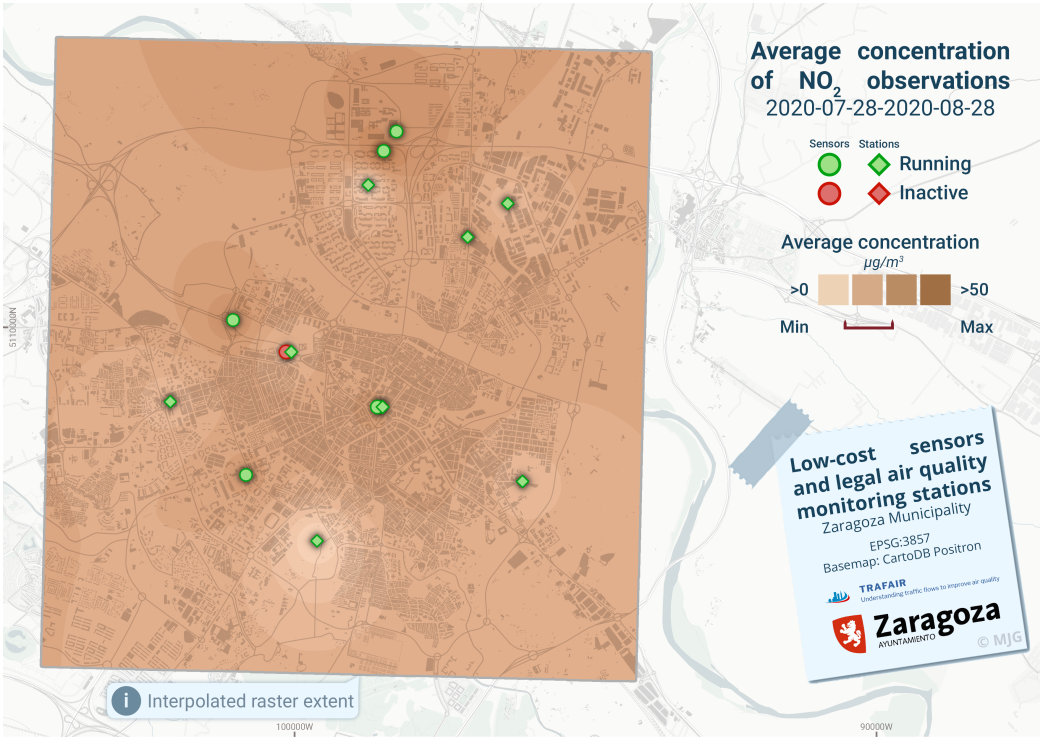

Figura E.4: Valor promedio de observaciones de NO_2, 28/07-28/08 de 2020.

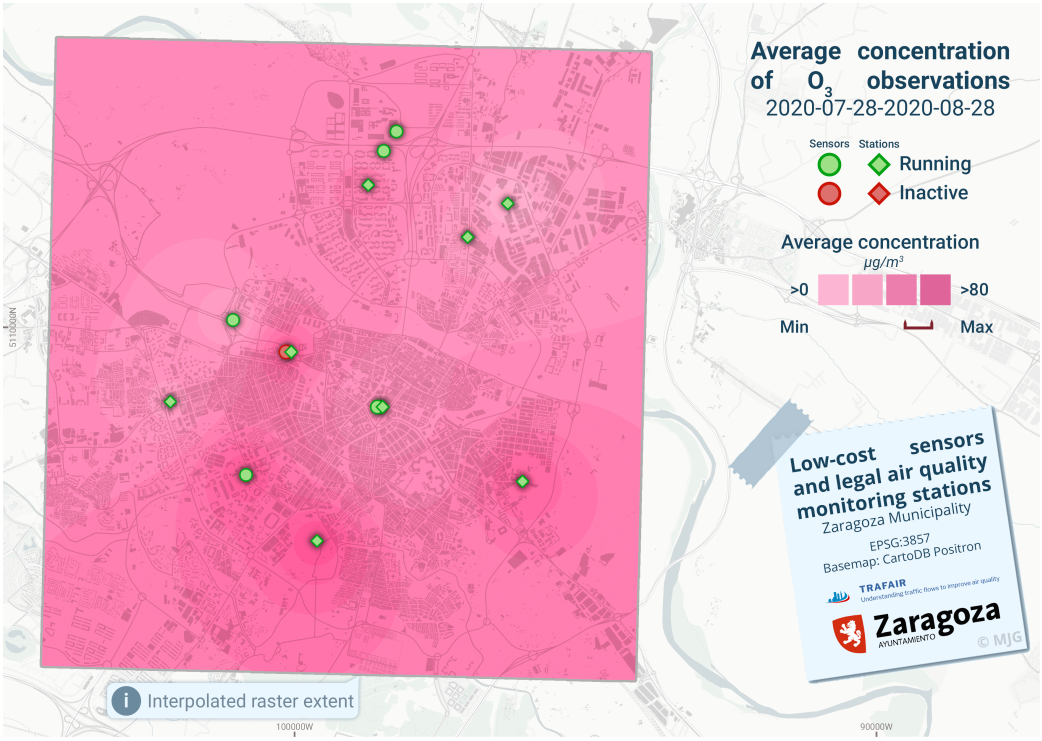

Figura E.5: Valor promedio de observaciones de O_3, 28/07-28/08 de 2020.