UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Development of artificial intelligence systems for
signal processing and signal enhancement in particulate matter sensors

Desarrollo de sistemas de inteligencia artificial para la calibración y
mejora de la señal en sensores destinados al análisis de partículas
en suspensión presentes en el aire

D. Eduardo Illueca Fernández

2024

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Development of artificial intelligence systems for
signal processing and signal enhancement in particulate matter sensors

Desarrollo de sistemas de inteligencia artificial para la calibración y mejora de la señal en sensores destinados al análisis de partículas en suspensión presentes en el aire

Autor: D. Eduardo Ilueca Fernández

Director/es: D. Jesualdo Tomás Fernández Breis y D. Antonio Jesús Jara Valera

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO DE DOCTOR

Aprobado por la Comisión General de Doctorado el 19-10-2022

D. Eduardo Illueca Fernández

doctorando del Programa de Doctorado en

Informática

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Development of artificial intelligence systems for signal processing and signal enhancement in particulate matter sensors / Desarrollo de sistemas de inteligencia artificial para la calibración y mejora de la señal en sensores destinados al análisis de partículas en suspensión presentes en el aire.

y dirigida por,

D. Jesualdo Tomás Fernández Breis

D. Antonio Jesús Jara Valera

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones o, tal y como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 6 de Mayo de 2024

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Información básica sobre protección de sus datos personales aportados	
Responsable:	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación:	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad:	Gestionar su declaración de autoría y originalidad
Destinatarios:	No se prevén comunicaciones de datos
Derechos:	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Eduardo Illueca Fernández ha realizado esta Tesis Doctoral financiado por una ayuda competitiva de la Fundación Séneca (21300/FPI/19), y opta a la Mención de Doctorado Internacional gracias a la estancia realizada en Karolinska Institute (Suecia) durante 3 meses, financiada con una ayuda para estancias externas destinada a los investigadores predoctorales de la Fundación Séneca (21681/EFPI/21)

Dare to think!
Immanuel Kant

Acknowledgements

Este tesis doctoral no es únicamente el final de una etapa académica, sino también de una etapa personal de mi vida. Según Balzac, las mejores historias son las de aprendizaje, y la mía ha sido sin duda una de aprendizaje continuo, que comenzó en los primeros días como estudiante de bioquímica, que despertaron mi vocación científica. En este viaje que culmina con esta tesis he tenido la ocasión de conocer a personas maravillosas y amigos a las que les debo mi más sincero agradecimiento.

Comenzando por la carrera, agradecer sinceramente a todos los compañeros que me han apoyado durante esos cuatro años de sacrificio, y a algunos el año extra de master. Entre ellos, tengo que mencionar al que ha sido mi compañero en esta etapa, Erik, por la paciencia demostrada durante una época complicada y por actuar siempre como modelo de buena persona.

Al equipo de Libelium, por haberme acogido como una familia, y sin el cuál este proyecto no habría sido posible, aportándome conocimiento, tiempo y ayuda, que me han permitido seguir adelante con mi investigación. Especialmente, quiero agradecer a mis compañeros del departamento de I+D por todo el apoyo, en concreto a Noel, Pedro, Alejandro, Arturo, Iris, Isa, Jaime y Jorge por el trabajo compartido.

No puedo olvidarme tampoco de mi familia, mi madre, mi padre, Nati y Antonio, por todo el cariño que me habéis transmitido estos años, dándome el soporte y apoyo que necesitaba para seguir siempre adelante y confiando en mí, así como por ser las personas más cercanas a las que acudir cuando las cosas iban mal. Gracias de corazón por los daños colaterales que hayáis podido sufrir en el proceso de escritura de esta tesis.

No obstante, el mayor reconocimiento quiero transmitirselo a mis directores de tesis, ya que al final esta aventura ha sido tan suya como mía. En particular, a Antonio quiero agradecerle su capacidad de escucha y ayuda, y el enseñarme que para ser un gran investigador la empatía es una cualidad fundamental, siendo capaz de otorgarme la libertad necesaria para transformar mis ideas en algo real, así como a centrar los esfuerzos en lo que es realmente relevante y no caer en lo superfluo.

A Jesualdo le debo agradecer su confianza desde el primer día y el haberme acompañado desde mi trabajo fin de grado, cuando apenas sabía programar, hasta el final de la tesis, orientándome con claridad y objetividad en cada momento. También agradecerle las revisiones en el último momento y los consejos que me han permitido aprender poco a poco, con trabajo e ilusión, la profesión investigadora.

Abstract

Resumen

La Organización Mundial de la Salud (OMS) ha reportado en sus últimos informes relativos a la calidad del aire que la contaminación atmosférica seguirá siendo uno de los principales problemas ambientales cruciales en los próximos años y a resolver durante el transcurso del siglo debido al efecto que está provocando en la salud de los ciudadanos. La contaminación del aire, según la OMS, no solo representa un desafío ambiental significativo, sino que también tiene profundas implicaciones para la salud pública global. Por tanto, este pronóstico requiere un enfoque más exhaustivo en la implementación de acciones dirigidas a reducir los niveles de contaminantes atmosféricos. Entre los diversos contaminantes del aire, la materia particulada y los aerosoles (PM) destacan como los contaminantes más perjudiciales para la salud humana debido a sus efectos cardiovasculares y respiratorios. Las partículas más pequeñas, conocidas como PM2.5, son especialmente peligrosas porque pueden atravesar la barrera alveolar en los pulmones y acceder al flujo sanguíneo, causando una serie de problemas de salud que incluyen enfermedades respiratorias y cardíacas. Estas partículas finas se generan a partir de diversas fuentes, incluyendo la combustión de combustibles fósiles en vehículos, industrias, y plantas de energía, así como de fuentes naturales como incendios forestales y erupciones volcánicas. En este contexto, una parte significativa de la población mundial se encuentra expuesta a concentraciones de PM que superan los estándares recomendados por la OMS. Las grandes ciudades, en particular, son las áreas donde este efecto se generaliza más, debido a la alta densidad de tráfico vehicular, la actividad industrial, y población que se encuentra expuesta a estos niveles. En este contexto, la monitorización y estudio de la calidad del aire se presentan como actividades esenciales que deben ser realizadas de manera continua para evaluar los niveles de polución ya que esta información es crucial para los gobiernos y autoridades públicas, ya que les permite aplicar las restricciones adecuadas para mitigar los efectos adversos de la contaminación.

A pesar de la importancia que tiene la medición de la calidad del aire y los diferentes contaminantes atmosféricos, el uso de equipos de medición tradicionales presenta ciertos desafíos. El primero de ellos es que estos equipos suelen ser costosos y su implementación y operación pueden ser complejas. Por otro lado, además de no ser fácilmente automatizables presentan sistemas analógicos que dificultan la extracción y almacenamiento de datos. Esto hace necesario explorar nuevos enfoques y tecnologías que permitan una monitorización más eficiente y accesible. Aquí es donde el Internet de las Cosas (IoT) entra en juego, ofreciendo una alternativa más económica y con potencial para ser ampliamente distribuida, permitiendo el desarrollo de ciudades inteligentes, donde los diferentes componentes sensoriales de la ciudad reportan a una arquitectura centralizada - en mayor o menor grado - que permite el procesamiento de los datos. Por tanto, los dispositivos IoT pueden proporcionar mediciones en tiempo real y a un costo reducido, lo cual es fundamental para una monitorización del aire que sea continua en el tiempo. No obstante, los dispositivos IoT también tienen sus limitaciones, especialmente en la precisión de las mediciones de partículas y la calidad del dato. Estos dispositivos pueden estar sujetos a sesgos significativos debido a la influencia de factores externos como la humedad, la forma de las partículas o la composición química, lo cual puede afectar la fiabilidad

de los datos recogidos. Para abordar este problema, es necesario desarrollar soluciones que mejoren la fiabilidad de estos dispositivos. Desde el punto de vista de la transición digital, la emergencia de los algoritmos de inteligencia artificial y las arquitecturas desarrolladas para ciudades inteligentes permiten ejecutar y entrenar diferentes modelos basados en técnicas computacionales clásicas o en *deep learning*.

Por ello, en esta tesis se exploran diversas soluciones para mejorar la señal en sensores de partículas mediante una metodología bottom-up. En primer lugar, se ha desarrollado un sistema de hardware corrector de humedad basado en columnas de silicona. Este sistema ha sido evaluado en comparación con métodos de referencia para asegurar su efectividad. La corrección de la humedad es esencial ya que esta puede afectar considerablemente las mediciones de partículas, llevando a resultados inexactos que pueden distorsionar la evaluación de la calidad del aire. Además del desarrollo de hardware, se ha implementado una capa *edge* que permite el procesamiento de datos mediante modelos de machine learning. Esta capa es capaz de realizar predicciones de concentración de partículas a corto plazo, lo que es fundamental para una respuesta rápida ante episodios de alta contaminación. El uso de *machine learning* en este contexto permite analizar grandes volúmenes de datos de manera eficiente y obtener insights valiosos para la toma de decisiones. Posteriormente, los datos hiperlocales recogidos por estos dispositivos IoT se utilizan para calibrar el modelo de transporte químico CHIMERE-WRF. La calibración de este modelo con datos locales mejora significativamente su precisión y fiabilidad. Los modelos de transporte químico son herramientas esenciales para predecir la dispersión y concentración de contaminantes en el aire, y su calibración con datos precisos es crucial para obtener resultados confiables. Este enfoque mejora la interoperabilidad de los datos, permitiendo que diferentes sistemas y aplicaciones puedan acceder y utilizar estos datos de manera eficiente. La interoperabilidad es fundamental en el contexto de las Ciudades Inteligentes, donde múltiples sistemas necesitan trabajar juntos para proporcionar servicios eficientes y mejorar la calidad de vida de los ciudadanos.

El primero de los artículos describe y evalúa un sistema de medición de nanopartículas basado en dos columnas de silicio, demostrando que el secado de las muestras de aire mejora sustancialmente las mediciones realizadas en comparación con el sensor OPC-N3. Además, los resultados obtenidos con respecto a las mediciones de referencia son bastante buenos, con valores de $R^2 = 0.83$ para PM2.5 en el mejor escenario para promedios horarios y un coeficiente de Pearson por encima de 0.80 para promedios y máximos diarios. Adicionalmente, se ha validado que estos dispositivos pueden integrarse en una arquitectura de IoT e interactuar con un servicio de calibración en tiempo real para mejorar la calidad de los datos. Los resultados de esta investigación demuestran que el sistema de secado desarrollado mejora significativamente la precisión de las mediciones de PM2.5 en comparación con los sensores existentes.

El segundo de los trabajos escala una capa en la arquitectura proponiendo una plataforma IoT-Edge-Cloud basada en el dispositivo Smart Spot para el monitoreo de partículas en interiores en entornos industriales y la predicción de riesgos para promover la detección temprana de superaciones. Los resultados obtenidos en relación con los operadores de predicción muestran una precisión del 87 % y un AUC de 0.81. Además, hemos demostrado que esta solución es disruptiva en los siguientes puntos con respecto al estado del arte: (i) es la primera plataforma para el monitoreo de la calidad del aire en interiores basada en una arquitectura IoT-Edge-Cloud, mejorando la escalabilidad en comparación con las arquitecturas tradicionales de la nube; (ii) está basada en FogFlow y FIWARE, lo que facilita su reutilización en otros entornos; (iii) mejora el rendimiento en términos de ancho de banda y tiempo de procesamiento, reduciendo la latencia en un 26 %; y (iv) está basada en un sistema de calidad del aire hiperloco y de alta calidad con un R^2 de 0.91 para PM10 y un R^2 de 0.74 para PM2.5 en comparación con las mediciones de referencia.

Por otro lado, el tercer artículo se centra en los modelos de transporte químico que se ejecutan en la nube, y requieren la mayor parte de recursos computacionales. Se ha propuesto una calibración del modelo CHIMERE-WRF que ha permitido mejorar la correlación del sistema en un 63 % para NO2 y un 25% para O_3. Estos contaminantes son de particular interés debido a sus efectos adversos en la salud humana y el medio ambiente. Mejorar la precisión de las predicciones de estos contaminantes permite implementar políticas más efectivas para su reducción y control. Una propuesta innovadora de esta investigación es la nueva metodología automatizada para la zonificación de la calidad del aire en la Región de Murcia. Esta metodología permite identificar nuevos puntos de monitorización de manera eficiente, lo cual es crucial para una evaluación exhaustiva de la calidad del aire en diferentes zonas. Gracias a ello, se ha propuesto una nueva zonificación para la Región de Murcia y la instalación de un nuevo punto de medición en la zona norte, aunque esta propuesta debe ser validada por los expertos en la materia, y tener en cuenta factores socioeconómicos. Sin embargo, los resultados obtenidos abren nuevas posibilidades de investigación y destacan diferentes limitaciones que deberían abordarse en distintos estudios: i) expandir el área de estudio para recopilar datos más representativos geográficamente; ii) comprender la replicación de los modelos de redes neuronales en otras ubicaciones, ya que se han entrenado con datos de Murcia y iii) mejorar la calidad de los conjuntos de datos de referencia proporcionados por las autoridades públicas, especialmente en lo que respecta al número de valores faltantes para algunos contaminantes. Por otro lado, se necesita más investigación en el campo de la zonificación automática de la calidad del aire para entender el papel de las redes neuronales en la reducción de dimensionalidad y proporcionar diferentes arquitecturas para la calibración y el preprocesamiento.

Finalmente, el último artículo ha desarrollado una metodología basada en la Minería de Procesos Interactiva para realizar un análisis sensible orientado a secuencias, demostrando que el uso de *interactive process mining* mejora sustancialmente la identificación de actividades objetivo relacionadas con la alta exposición y permite a los expertos tomar las mejores políticas urbanas. Además, los resultados obtenidos en relación con los KPIs mostraron una reducción del 26 % en la exposición a PM2.5 y una disminución del 16 % en el riesgo relativo de mortalidad en el mejor escenario. Por lo tanto, los resultados sugieren que nuestra hipótesis inicial es correcta y que el análisis sensible orientado a secuencias podría ser una herramienta poderosa para mejorar el medio ambiente en las ciudades. El uso de esta metodología propuesta permite avanzar en las siguientes líneas que hasta ahora han sido limitadas en el estado del arte: i) la evaluación de políticas de salud pública para reducir la exposición de la población; ii) la mejora del conocimiento en la relación entre exposición y riesgo, permitiendo a los investigadores definir modelos precisos de respuesta a la exposición y iii) las aplicaciones de la *interactive process mining* en la modelización ambiental y la mitigación de la contaminación del aire.

En conclusión, la combinación de mediciones hiperlocales de calidad del aire con simulaciones basadas en modelos de transporte químico ofrece una mejora significativa en la precisión de las mediciones de partículas y otros contaminantes. Esto permite generar productos valiosos para evaluar el impacto de diferentes políticas en la sostenibilidad, en lo que se conoce como sustainability impact assessment. Estos conocimientos son cruciales en el sector de las ciudades inteligentes, donde la gestión eficiente de recursos y la mejora de la calidad de vida de los ciudadanos son objetivos primordiales. El desafío futuro radica en la aplicación de las grandes cantidades de datos generados por los dispositivos IoT en escenarios más complejos y en el desarrollo de soluciones digitales para ciudades. La integración de estos datos en sistemas de gestión urbanos permitirá tomar decisiones más informadas y efectivas, mejorando así la sostenibilidad y la resiliencia de las ciudades frente a los desafíos ambientales. Además, la interoperabilidad y el uso de estándares abiertos serán clave para facilitar la colaboración entre diferentes sistemas y plataformas, maximizando el potencial de las tecnologías IoT en la monitorización y gestión de la calidad del aire. Por ello, esta tesis presenta un enfoque integral para mejorar la monitorización y gestión de la calidad del aire utilizando tecnologías IoT y modelos de machine learning. Los resultados obtenidos no solo demuestran la viabilidad de estas tecnologías, sino también su potencial para transformar la manera en que se gestiona la calidad del aire en nuestras ciudades. Con un enfoque basado en la precisión, la interoperabilidad y la sostenibilidad, este trabajo sienta las bases para futuras investigaciones y desarrollos en el campo de la monitorización ambiental y el desarrollo de ciudades inteligentes.

Summary

The World Health Organization (WHO) claims that air pollution will be a significant environmental concern in the coming years, leading to increased emphasis on actions focused on reducing pollutant levels in the air, from which particulate matter (PM) is the most harmful to health, with a large percentage of the population exposed to levels exceeding WHO standards. In this sense, air quality monitoring is essential to know pollutant levels in the air. However, traditional measuring devices are expensive and not automatic, making necessary to develop solutions based on new paradigms such as the Internet of Things (IoT) that are able to reduce costs.

However, IoT devices present significant biases in measurement processes, especially concerning particles, due to the influence of external factors such as humidity. Therefore, this thesis explores solutions to improve signal in particle sensors using a bottom-up methodology. Firstly, a dryer system based on silicone columns has been developed, which has been evaluated with respect to reference methods. Next, an edge computing layer has been deployed allowing data processing using machine learning models to make short-term particle concentration predictions. Later, hyperlocal data from devices is used to calibrate the CHIMERE-WRF chemical transport model.

The results demonstrate that the developed drying system improves PM2.5 accuracy compared to existing sensors, with a coefficient of determination (R^2) equal to 0.83. Additionally, these devices can be easily integrated into an IoT-Edge-Cloud architecture for risk assessment in work environments, using machine learning to predict exceedances with 80 % accuracy. On the other hand, calibration of the CHIMERE-WRF model allows for a 63 % improvement in system correlation for NO_2 and a 25 % improvement for O_3. Lastly, a new proposal for air quality zoning for the Murcia Region, using an automated methodology, allows for the identification of new monitoring points.

In conclusion, the combination of hyperlocal air quality measurements and simulations with chemical transport models allows for improved particle measurement and the generation of products to assess the impact on sustainability of different policies, known as sustainability impact assessment. This knowledge is crucial in the Smart Cities sector. The future challenge lies in applying large amounts of data generated by IoT in more complex scenarios and digital solutions for cities.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and State of the Art

1.1 — Context

Sustainability Impact Assessment (SIA) is a central instrument for evidence-based policy-making in EU trade policy, which seeks to understand and optimize the effect of social and political actions in different environmental domains [1]. One of the most important targets in sustainability impact assessment is atmospheric pollution. Poor air quality in a major urban city is considered a critical environmental issue because air quality and air pollution are some of the consequences and disadvantages of large human agglomerations. Thus, real-time air pollution data collection and analysis in a smart city is essential for urban sustainability. It is necessary to provide cities with the infrastructure to measure this pollutant in the air. Hence, cost-effective means are needed to measure air pollution [2]. These infrastructures and resources have traditionally been built on expensive regulatory monitoring, enhanced with low-resolution satellite data and mathematical forecasting models. This solution does not cover the requirements of a dynamic *Smart City*. They can harness ubiquitous technologies - Internet of Things (IoT), big data, artificial intelligence, smartphones, edge computing, and high-performance architectures - in concert with traditional methods for a better understanding of air quality and propose sustainability policies to improve quality of life [3].

However, before discussing the different infrastructures, services, and architectures that can enhance air quality, it is necessary to clarify what a *Smart City* is. A first definition highlights the role of digital transformation, defining a *Smart City* as *"An innovative city that uses information and communication technologies and other means to improve the quality of life, the efficiency of urban operation and services, and competitiveness, ensuring that it meets the needs of present and future generations concerning economic, social, and environmental aspects"* [4]. However, it could be wrong to define an *Smart City* only from a technological point of view because the creation of a *Smart City* requires the understanding of the evolution and experiences of urban development and have a clear grasp of the concept, emphasis, and developmental trajectory of the city [5]. Just like a traditional city, this process is based on six social axes: smart economy, smart mobility, smart environment, smart population, smart lifestyle, and smart government. In short, in an ideal *Smart Cities*, all the elements that make up the city can communicate with each other [6].

During the last years, many cities in the world have tried to implement a *Smart City* architecture to solve the issues of urban planning and governance using the capacity of modern information and communication technologies. Past experiences in this regard indicate one of the biggest challenges for the implementation of smart cities, which is generally not related to the technology itself, but to the problems that occur during the implementation of the idea, which is mainly due to the lack of a common and clear strategy between the stakeholders [7]. However, many success stories in the scientific literature demonstrate that *Smart Cities* are feasible and can improve urban quality of life. To measure the effectiveness of a *Smart City*, the motion index indicator (CIMI) was defined to compare 77 city indicators among ten dominant categories in urban life - economy, technology, human capital, social cohesion, international outreach, environment, mobility and transportation, urban planning, public management, and governance [8]. The results obtained draw a map with the following cities in the *Top Ten of Smartness*: New York (USA), London (UK), Paris (France), San Francisco (USA), Boston (USA), Amsterdam (Netherlands), Chicago (USA), Seoul (South Korea), Geneva (Switzerland), and Sydney (Australia). [9].

The concept of *Smart City* is evolving to a new approach based on *Smart and Connected Communities*, that describes better how the technologies and communications can connect and put together all the elements in a city, to address synergistically the needs of remembering what happened in the past, the present needs, and the future needs of planning [10]. This dynamic context points out the role of technology and the need for robust architectures in *Smart Cities*. In general terms, *Smart Cities* follows a bottom-up architecture which is composed of four layers: i) sensing layer, in charge of data collection from IoT physical devices; ii) transmission layer, which integrates various communication protocols available in the IoT paradigm, that allows to send sensor data to the cloud/edge; iii) data management layer, whose responsibility is to processes data analytics workflows and to store valuable information; and iv) application layer, that are useful for the end user interface provision offered by various applications at the top layer for end user [11]. In Figure 1.1 an specific *Smart City* architecture is provided as example. Thus, it should be noted that some of the components can change to others with similar applications.

In this context, monitoring and communicating air quality data is a key element of a *Smart City*, as it encompasses multi-sectorial functions, data, and information exchanges on pollution sources, mitigation, health burden, socioeconomic impacts, and policy formulation. For its effectiveness, real-time and near-real-time data communication is required. However, these data are often created in silos or not available in cross-sectorial usable form, necessitating integrated data governance translating data into knowledgeable information [12]. In addition, the emergence of climate change mitigation put pollution management in the center of the *Smart Cities* services. It highlighted the need for interaction with citizens through urban living labs and citizen sciences initiatives and public authorities, companies, and research centers as critical stakeholders, which should be synchronized in the urban environment [13].

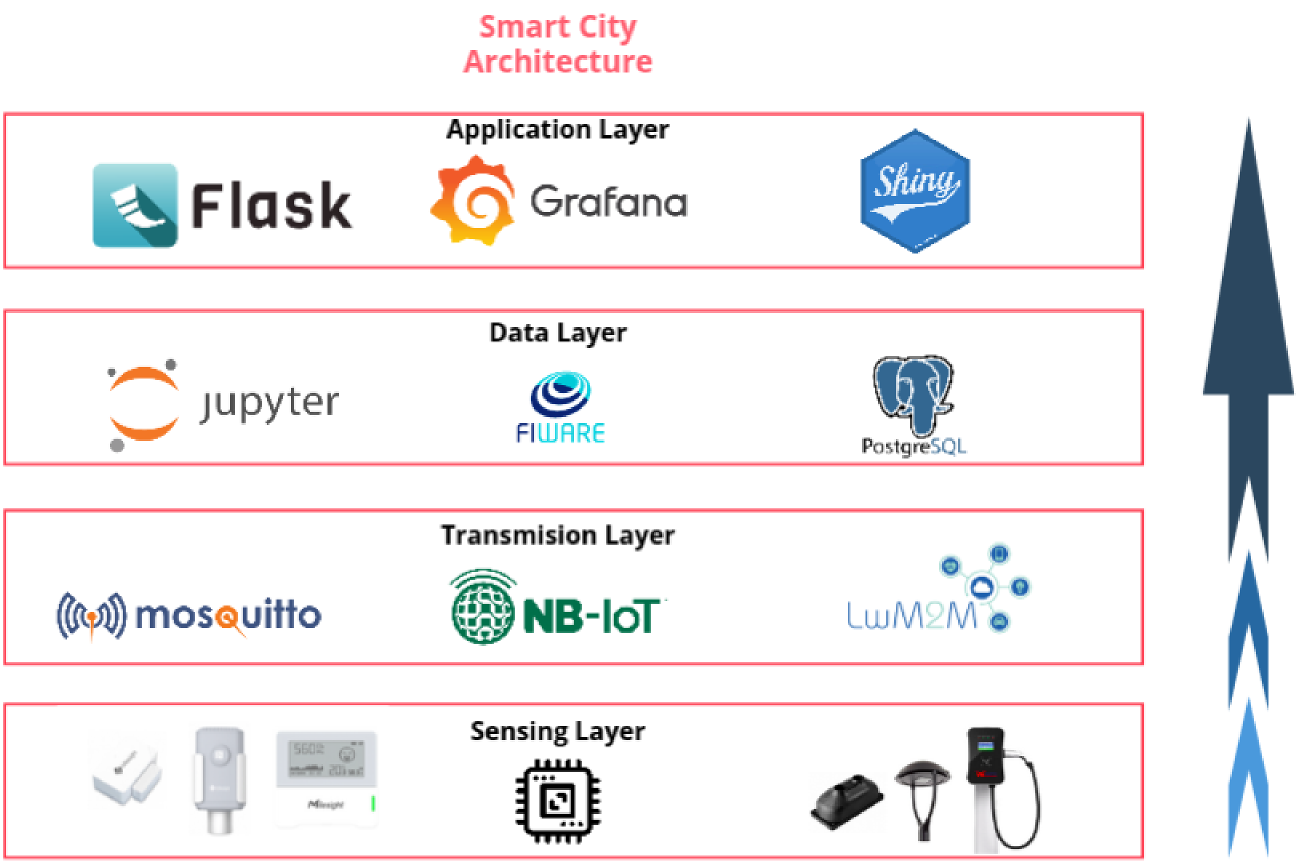

Figure 1.1: Example of a Smart City architecture, composed by a sensing layer, a transmision layer, a data management layer and an application layer.

However, as suggested from the previous description, there is a gap between the need for sustainability impact assessment services in a *Smart City* and air quality monitoring strategies claimed by cities. These challenges are not only in the sensing layer - where the expensive reference measurements are a clear limitation - but also in the data management layer, which has to deal with complex data that requires scientific expertise higher than other domains, and in the application layer that should provide this complexity in an understandable way to the end user, to create engage in citizens and allow public authorities to make decisions in climate change mitigation policies.

Thus, this thesis will explore these gaps and provide solutions, especially based on artificial intelligence methods, but also in improving the actual hardware or the mathematical models used to gather air quality predictions.

1.2 — Motivation

According to the World Health Organisation, air pollution causes more than seven million deaths per year, making it one of the major public health problems of recent times. Regarding Chemical

Pollution, it represents a paradigm of risk factors associated with relatively modest increases in individual risk but substantial disease burden at the population level, especially with the mortality and morbidity of different diseases [14]. In addition, exposure to these particles is associated with various cardiovascular and respiratory diseases [15, 16]. Furthermore, certain gases such as O_3, NO_2 or SO_2 are associated with chronic obstructive pulmonary syndromes. In particular, ozone - a highly oxidising agent - is correlated with lung inflammation, the alveolar epithelium's destruction and the pleura's chemical composition [17]. Thus, one of the key aspects of digital transformation and the challenges facing society today related to climate change is air quality monitoring, evaluation, prediction and mitigation plans for cleaner air. This problem is higher in cities, perceived by citizens as unhealthy places. For this reason, air quality and its management are among the most important services in Smart Cities [18].

For this reason, this work is motivated to provide a Smart City architecture and platform for Sustainable Impact Assessment in Smart Cities. For this purpose, it is necessary to define this term coherently with the context described in the previous section. The interpretation followed in this work can be summarized in Figure 1.2. If we start from the top, the atmosphere of an urban environment presents a certain amount of pollutants, whose concentration is measured by monitoring devices. The concentration observed in the atmosphere is produced by pollutants emissions from different sources. These emissions are quantified in inventories that collect the emission flux of a certain chemical species [19]. On the other hand, the concentration observed in the atmosphere impacts citizens living in the city. This impact can be measured according to different *Key Performance Indicators* (KPIs) related to health outcomes, climate change effect or environmental issues [20]. To mitigate these impacts, public authorities and policymakers should take decisions to reduce the emissions and, therefore, pollutant concentrations and impacts. However, policymakers need a comprehensive tool that simulates the effect in the whole loop of emissions reduction. The set of tools, methodologies and models that allows for assessing the effectiveness of policies and scenarios in concentration and impacts to help authorities make decisions based on data is called Sustainability Impact Assessment.

This approach brings together multiple variables and components operating across three interconnected scales: micro (individual and interpersonal factors including disciplinary background and risk perceptions); meso (network, organizational and institutional); and macro (environmental, social, cultural, political, and economic factors)[21]. All these scales are present in a Smart City. The problem that arises is that current urban digitalization does not reflect the whole real-life context and dynamics of a living city, and cross-sectoral platforms are required to be composed of multiple components [22].

To be able to represent this dynamics, there is a need of integrating and connecting all the components specified in Figure 1.2 into a smart cities architecture proposed in Figure 1.1. These architectures should have the following components: i) an extensive network of accurate IoT devices; ii) a contextual intelligence component that normalize and standardize all the data from different

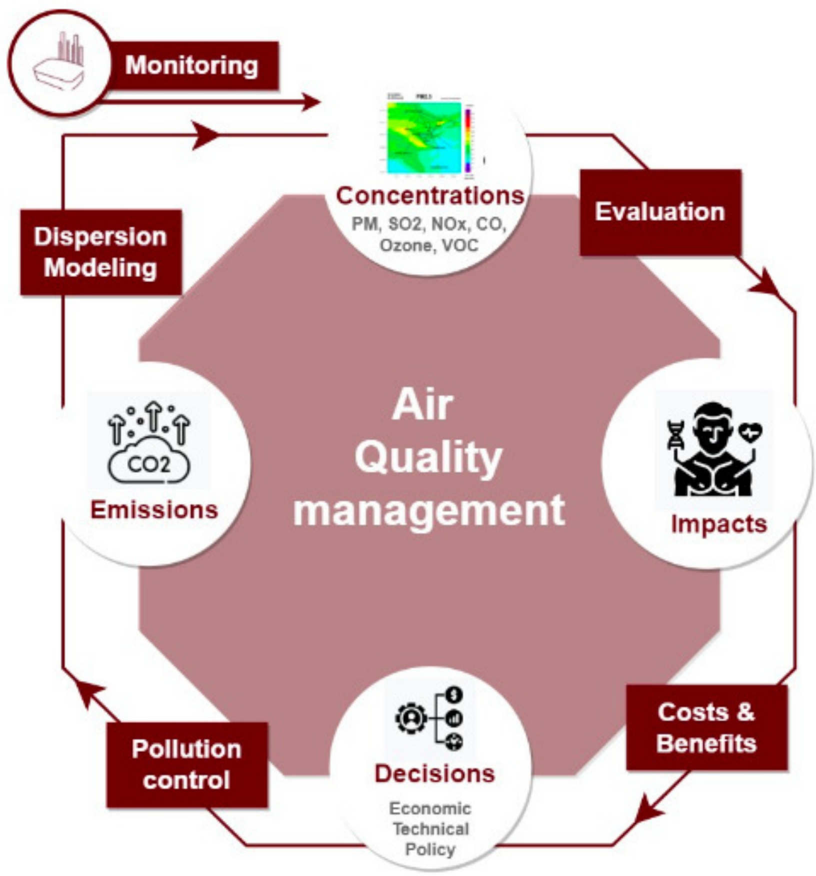

Figure 1.2: Definition of Sustainability Impact Assessment loop for emissions reductions through hyperlocal measurements and simulations.

sources and standardise to guarantee the compliance with FAIR guidelines [23]; iii) an analytic layer with statistical models; and iv) a modelling component that integrates dispersion and forecasting models; and v) a knowledge layer in charge .

However, the IoT devices which are building blocks of the first layer present important errors in measurement processes, especially concerning particles, due to the influence of external factors such as humidity. Therefore, this thesis aims to develop solutions to correct the signal in hyperlocal particulate sensors. In addition, the need of real-time data could be limited by traditional cloud approaches making necessary to explore other paradigm such edge computing or federated learning.

The next section will explore the state of the art of the different layers proposed in Figure 1.1. The goal of this review is to understand the different solutions available in the literature and in the market to identify research gaps and how to connect the different components. The review methodology has followed a divide and conquer approach, first exploring each layer separately and then discussing the different platforms available on the market and the comparison with the proposed architecture.

1.3 — State of The Art

The starting point of this thesis are two new concepts that have emerged recently in the state of the art, named *hyperlocal air quality measurements*, which are taken by *hyperlocal IoT devices*, which are IoT devices that integrate a set of low-medium cost sensors, allowing to design dense air quality networks, going beyond the traditional reference stations. This new approach could improve the quality and the resolution of the sensing layer in a *Smart City* architecture. In addition, using these devices can improve the services offered by smart cities, improving the quality of the collected data, allowing the data management models to obtain valuable knowledge from the data, and enhancing the user experience in the application layer. In the context of air quality, these services are based on physical and chemical models that predict the concentrations in pollutants. Last, some approaches used the edge computing approach to improve the performance in the transmission layer and the data management layer.

In this sense, this section will describe all the developments found in the state of the art, providing the main insights and discussing their impact on implementing sustainability impact assessment services. In concrete, in the sensing layer this section explore the different methodologies for measuring pollutants in the air. Then, regarding the transmission layer an exhaustive description of all the FIWARE components, followed by a review of the cutting edge data analysis frameworks. Regarding this, a full subsection is dedicated to air quality modelling strategies because of its relevance in sustainability impact assessment. Last, the use of semantic technologies in the air quality domain is explored.

1.3.1. Air Quality Devices

With the development of new technologies, particularly Internet of Things (IoT), there has been an increase in the deployment of low-cost air quality monitoring systems. These systems use inexpensive and hyperlocal IoT sensors with lower accuracy as compared to robust systems. This fact has raised some concern regarding the quality of the data gathered by the IoT systems, which may compromise the performance of the environmental models [24]. In this context, we define an air quality device as a system that integrates one or more pollutant sensors with a centralised electronic intelligence. Table 1.1 compares reference stations with hyperlocal IoT systems based on several key factors. The left-hand column presents attributes of reference stations, while the right-hand column represents corresponding characteristics of hyperlocal IoT systems. In terms of accuracy, reference stations are rated positively for high accuracy but are associated with higher cost. On the other hand, hyperlocal IoT systems are acknowledged for providing low sampling times [25], but their quality assessment can be challenging [26]. Reference stations are considered official data sources but may have lower resolution capabilities, while hyperlocal IoT systems offer high-resolution data but are not officially recognized [27]. Additionally, reference stations have a longer operational lifetime. Still, they may not be as compatible with IoT technology. Hyperlocal IoT systems are designed to be IoT-friendly, but have a shorter operational lifespan [28]. Table 1.1 concisely compares these two air quality monitoring systems, highlighting their strengths and limitations.

Table 1.1: Comparison between hyperlocal IoT systems and reference air quality devices

Reference Stations		Hyperlocal IoT Systems	
Pros	Cons	Pros	Cons
High Accuracy	Expensive	Low Sampling Times	Difficult to Assess Quality
Official	Low Resolution	High Resolution	Not Official
Long Lifetime	Not IoT-friendly	IoT-friendly	Short Lifetime

Reference Technology

According to *Directive 2008/50/EC*, a reference methodology is accredited according to EN/ISO 17025 and should be applied by establishing quality assurance and quality control system which provides for regular maintenance to assure the accuracy of measuring devices. Furthermore, this regulation define an analytic method for each pollutant, which should be used in the reference laboratories. Thus, it is easy to define reference technology as the devices that are based on the methodologies validated by *Directive 2008/50/EC*

- **Carbon Monoxide (CO).** The reference method established in the European Legislation for this pollutant is the defined in the *EN 14626:2012 standard Ambient air. The standard method for the measurement of the concentration of carbon monoxide by non-dispersive infrared spectroscopy*. The technology is based on an infra-red beam that passes through the sample and each gas component in the sample absorbs at a concrete frequency in the infrared region. If we know this frequency, it is possible to quantify the concentration of CO as it is proportional to the absorbed radiation. The main advantage of this approach is that it is selective, achieving an accuracy < 1 %, a limit of detection < 0.04 ppmv and a range between 0-1000 ppmv [29]. Another technology found in the state of the art is Cavity Ring-Down Spectroscopy. The beam from a single-frequency laser diode enters a cavity defined by two or more high-reflectivity mirrors. The cavity quickly fills with circulating laser light when the laser is on. A fast photodetector senses the small amount of light leaking through one of the mirrors to produce a signal directly proportional to the intensity in the cavity. This technology is more used for greenhouse gases [30], but it has been tested with a precision of 1.5 ppbv, a limit of detection < 2 ppbv, and an operation range covering 0-5 ppmv [31]. Last, electrochemical methods have also been used for portable CO analyzers with a precision of 3 ppmv and a range of 0-500 ppmv as well as diffusive sampler based on Focaltin Reaction o photometric processes [32].

- **Ozone (O_3).** This gas is one of the oxygen allotrope and plays an essential role in atmospheric chemistry. Thus, there are a lot of strategies in the state of the art in quantifying the concentration of O_3 in the atmosphere. The most important is the automatic method defined in the European Legislation and the standards *EN 14625:2012 Standard method for the measurement of the concentration of ozone by ultraviolet photometry* and *ISO 13964*. According to the O_3 absorbing spectrum, a 254 nm light absorbed by O_3 is proportional to the concentration in air, according to the Lambert-Bee formulation. This means a precision of 0.5 %, a detection limit between 0.2-0.4 ppbv and an operation range between 0-10 ppmv [33]. In addition, nitric-oxide chemiluminescence exhibits similar performance in terms of accuracy and detection limit but a narrow operation range (0-2 ppmv) [32]. However, this pollutant presents different standards for indoor concentrations, in concrete *EN 14412:2004 Indoor air quality - Diffusive samplers for the determination of concentrations of gases and vapours - Guide for selection, use and maintenance* that proposes the use of ion chromatography combined with a conductivity/absorbance [34] or a UV-Vis spectrophotometer [35].

- **Nitrogen Oxides (NO_x).** From a chemical point of view, nitrogen oxides comprises a set of chemical compounds containing nitrogen and oxygen. However, in the context of air quality it is important to measure two of them: nitrogen monoxide (NO) and nitrogen dioxide (NO_2). The reason to group both pollutants is that their quantification methods are quite similar. An example is the reference method proposed in the legislation and described in the *EN 14211:2012 Standard method for measuring the concentration of nitrogen dioxide and nitrogen monoxide by chemiluminescence*. This approach is based on the following photochemical chemical reaction principle, which produces photons: $NO+O_3 \rightarrow NO_2+O_2+h\nu$. By measuring the luminescence produced by the reaction, it is possible to quantify NO_2 and NO. This method presents the same accuracy for both gases (0.4 ppbv) and a detection limit of 50 pptv and 40 pptv for NO_2 and NO, respectively, and operation range between 50 pptv-1 ppmv for NO_2 and 50 pptv-100 ppmv for NO [36, 37]. In addition, Cavity Attenuated Phase Shift Spectroscopy is used to measure both pollutants with a precision of 0.5 % and detection limit of $<$ 10 pptv and 0-1000 ppbv for both gases [38]. For NO_2 there are other strategies in state of the art as the chemoluminescence determination with luminol [39], laser-induced fluorescence spectroscopy [40] and Griess-Saltzman method [41], but these methods are less used in practice due to the difficulty to couple with NO measurements.

- **Sulphur Dioxide (SO_2).** This pollutant presents some challenges in the measurements compared to the compounds exposed before. The most important is the concentration range which is present in the atmosphere. The SO_2 concentration is generally below that of other inorganic species. For this reason, European legislation promotes using highly sensitive methods such as ultraviolet fluorescence as a reference. This approach presents a precision of 0.5 %, a detection limit of 0.4 ppbv and a range between 0-2000 ppbv. This method is defined in the standard *EN 14212:2012 Ambient air - Standard method for the measurement of the* *concentration of sulphur dioxide by ultraviolet fluorescence* [31]. In addition, ion chromatography using triethanolamine can determine SO_2 in a range between 0.1-200 $\mu g/m^3$ [42].

- **Particle Matter (**$PM10$**,** $PM2.5$ **and** $PM1$**).** According to the related *2008/50/EC* and the standard *EN 12341:2015 Ambient air - Standard gravimetric measurement method for the determination of the PM10 or PM2.5 mass concentration of suspended particulate matter*, particulate matter measurements are performed by taking a sample onto the filter and then weighing it on a balance equipped with a scale. The sampling uses the so-called PM collectors (or sequential PM collectors), with a nominal flow rate of 2.3 m^3/h. Measurement results are expressed in $\mu g/m^3$, where the air volume corresponds to the volume at ambient conditions, near the inlet, at sampling time. However, applying such a solution requires proper conditioning of the filters, which after 24 hours of exposure to the sample, is waiting for further analysis by the operator [43]. Other approaches are based on optical methods, which measure the optical diameter of particles and infer the particle mass by assuming a density and refractive index of the particles being measured to calculate particle mass [44]. However, tapered element oscillating microbalance, beta attenuators and optical PM sensors all measure different particle diameters/masses and do not align exactly with the gravimetric method [45].

Hyperlocal IoT Sensors

In recent years, a new generation of IoT-enabled devices for air pollution monitoring has emerged to resolve traditional monitoring systems' limitations and reduce the overall cost [46]. Due to developments in instrumentation and communications, deploying networks of sensors at high spatial density is now feasible. The term 'sensor' refers to the assembly of the detection element, measurement electronics, air-inlet, air-sampling and communications systems, and housing and mounting that deliver the measurement result together [27]. This thesis proposes the term hyperlocal, which concerns the ability to generate high-resolution data compared to regulatory stations with certified instruments. The technology underlying these sensors depends on the pollutant to be measured and the specific application. In contrast to the methods discussed above, approximations are required to be capable of being compressed into an IoT device, which implies a loss of accuracy. This section will review the main analytical methodologies and the performance of the main devices on the market against reference stations.

Regarding inorganic gas sensors to measure gaseous air pollutants, there are currently two types of low-cost sensors available in the market: (i) metal-oxide-semiconductor (MOS) sensors and (ii) electrochemical (EC) sensors [28]. The most important are the amperometric sensors inside the electrochemical methods. Amperometry is a conventional electroanalytical technique widely used to identify and quantify electroactive species in the gas phase. Applying amperometry to gas-phase analytes involves a unique gas/liquid/solid boundary (analyte-electrolyte-electrode) and an interfacial transport process that frequently controls the AGS's response characteristics and analytical performance. A simple amperometric cell with a two-electrode configuration is illustrated in Figure 1.3 wherein the electro-chemical system consists of two electrodes, i.e., a working electrode and a counter electrode, and the electrolyte solution in which the two electrodes are immersed. When applying a proper potential between the two electrodes, electroactive species in the electrolyte solution will participate in electrochemical reactions on the surfaces of each of the electrodes. A reduction reaction occurs at the cathode, takes electrons from it, and combines them with an oxidized species. According to Faraday's Law, each analyte molecule generates or uses an exact number of electrons. So the charge (the number of electrons or Coulombs of charge) is exactly related to the number of analyte molecules reacting in the system and the current (electron flow) at the electrode is exactly related to the rate of the electrochemical reactions (current is Coulombs/s) [47].

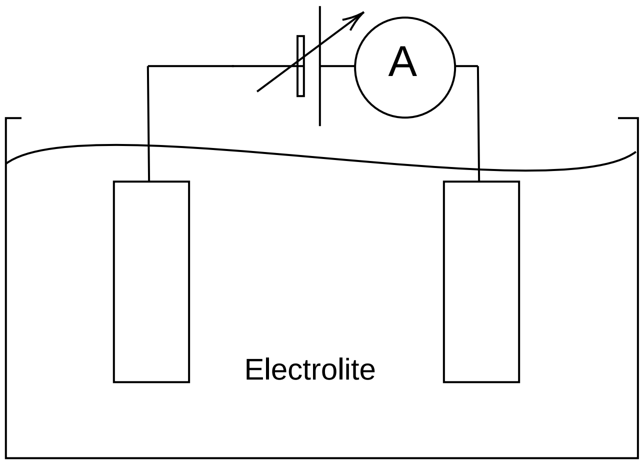

Figure 1.3: Principle of working of amperometric gas sensors with two-electrode configuration for liquids. Adapted from [47].

On the other hand, the light scattering method is used in hyperlocal PM sensors since the sensors based on this principle are cheap to manufacture, have low power requirements, and quick response times [48]. Commercially available hyperlocal particle sensors are based on optical scattering (Figure 1.4). In these devices, a flow of air is generated by employing a fan or convection due to a heating element. Particles in the airflow travel through a beam of light, propagating in the z-direction, and scatter light in all directions. The amount of scattered light detected in a particular direction Φ depends on the particle properties (size, shape, absorption, refractive index), wavelength, and polarisation. In general terms, high intensity means a large particle and vice versa. The particle size is determined based on the measured intensity. Thus, these devices are known as optical particle counters because they calculate the diameter of the particle and classify it into one bin, a size range in which all particles with a diameter within this size range are contained [49]. Then, the mass concentration is computed by assuming that the particle is spherical, so the mass of a single particle is defined by $m = \frac{4}{3}\pi r^2 \rho$ [50].

In this context, and because of the technological limitations of the methods described above, it is critical to assess the quality of reference methods against reference stations. Figure 1.5 summarises the results found in the literature regarding the performance against official devices, using the R^2

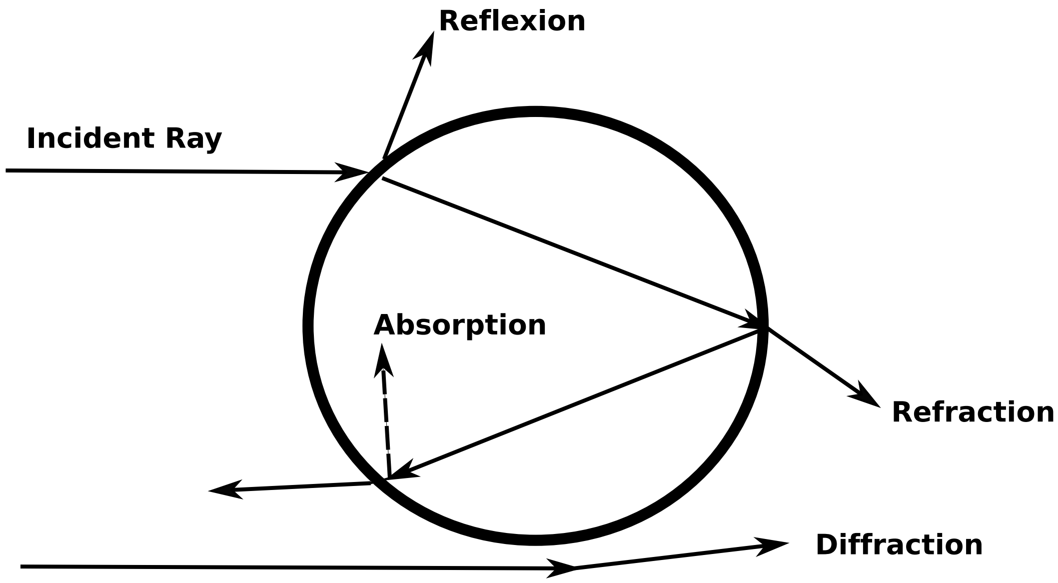

Figure 1.4: Phenomena and physic processes induced by the interaction of a single particle with a light beam. Adapted from [49]

metric. It is important to note that the particle matter subplot summarizes the results for the fractions PM10, PM2.5 and PM1. The first conclusion that could be extracted from this figure is that for gas sensors, there is a high difference between laboratory studies and field experiments - this last kind of tests are conducted in real environments. For gases, the results are quite good in laboratory condition whereas there is wide range of accuracies for field tests. This highlights two insights: i) the laboratory results are not extrapolable to field conditions and ii) the replicability of the field tests is low because there is a wide range of results in function of several factors. The trend for aerosols is similar, but it is true that the range of results in laboratory test is wider, because of the difficulty to replicate the experiments due to the fact that aerosols are a mixture of particles with different size and chemical composition.

Looking into detail the performances of the particle optical sensors, the following results are collected from the state of the art in Table 1.2, which separates the results in function of the sensor brand and the methodology used. Laboratory test refers to experiments done in controlled conditions while field test are done with sensors working in a real environment. In both case, a reference device is required to compute accuracy as R^2.

Regarding O_3 sensors, Aeroequal SM50 presents R^2 values ranging from 0.77 to 0.94 in field tests [57, 65], whereas the sensor UnitecSens 3000 was only tested in laboratory conditions with $Res < 2.0$ ppb [66]. The SGS MICS series present R^2 values in the laboratory between 0.88-0.95 [67]. Regarding electrochemical O_3 sensors, the reported R^2 values range from 0.13 to 0.70 during field testing of the Alphasense O3B4 sensor [67], but a $R^2 > 0.99$ is obtained in laboratory conditions [68]. The linear models tested showed R^2 greater than 0.99 during laboratory testing of the O3_3E1F sensor [68]. However, during field deployment, multiple linear regression models improve the results, and the R^2 values ranged from 0.85 to 0.94 with the calibration. After 4.5 months, the sensor performance

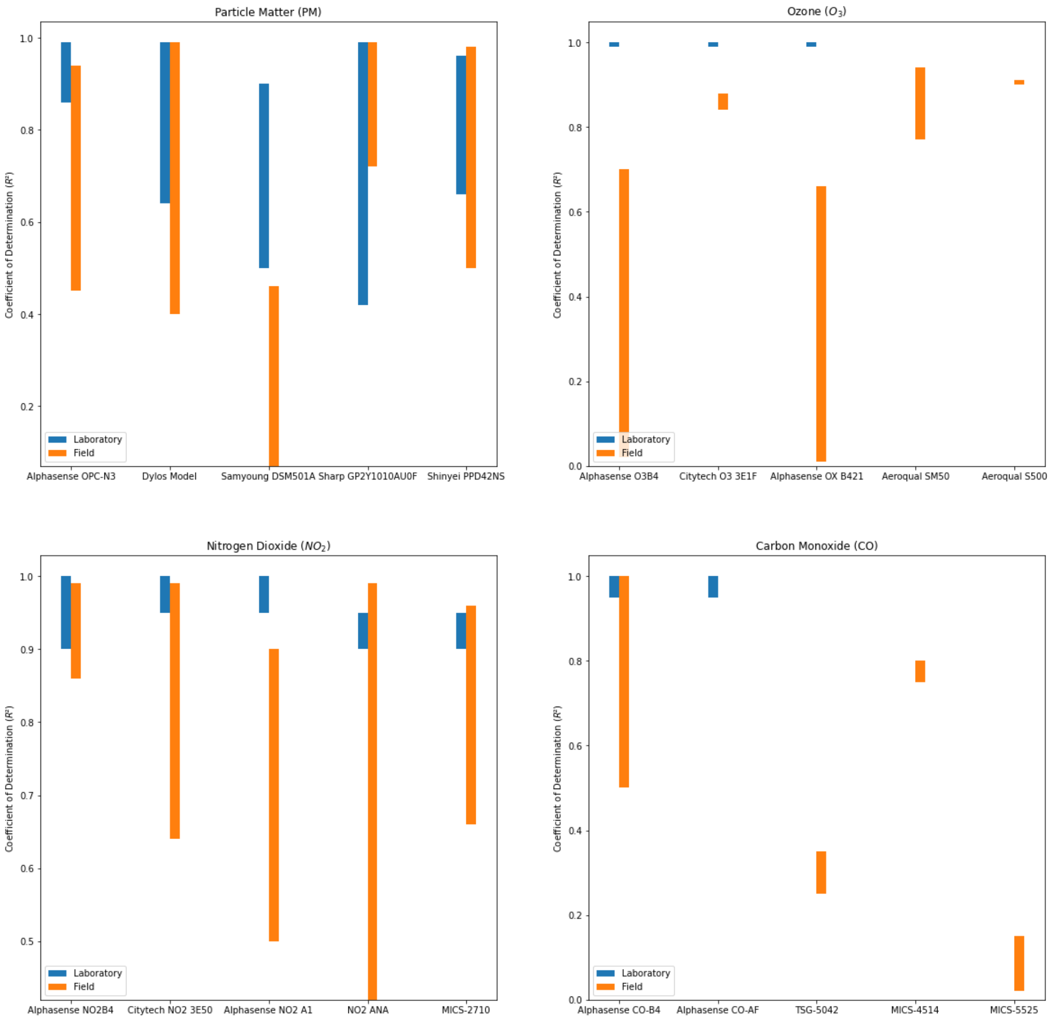

Figure 1.5: Performance of Hyperlocal Air Quality Sensors against reference measurements found in the state of the art.

Table 1.2: Performance of different machine learning models used for sensor calibration

Sensor	**Environment**	R^2 **Test**	**References**
Alphasense OPC-N2	Laboratory	0.94-0.99	[51]
Alphasense OPC-N3	Field	0.45-0.94	[52]
Dylos 1100 Pro and 1700	Laboratory	0.97-0.99	[53, 54, 55]
Dylos 1100 Pro and 1700	Field	0.48-0.99	[56, 57]
Plantower PMS 1003	Laboratory	0.69-0.99	[58]
Plantower PMS 1003	Field	0.82-0.93	[58]
Plantower PMS 3003	Laboratory	0.73-0.97	[58]
Samyoung DSM501A	Laboratory	0.88-0.90	[48, 59, 54]
Samyoung DSM501A	Field	0.07-0.97	[60]
Sharp DN7C3CA006-GP2Y1010AU0F	Laboratory	0.42-0.99	[54, 55, 48, 59, 61, 62]
Sharp DN7C3CA006-GP2Y1010AU0F	Field	0.72-0.99	[54, 55, 48, 59, 61, 62]
Shinyei PPD42NS	Laboratory	0.66-0.99	[63, 64, 58, 48]

deteriorated, and the R^2 values dropped from 0.67 to 0.82 with the simple linear regression model and 0.58 to 0.82 with the multiple linear regression model [69]. These results indicate that the response curves of the sensors were time-variable, possibly due to sensor ageing and/or dust accumulation.

The hyperlocal NO_2 sensors exhibit excellent performance in controlled laboratory conditions. However, their performance degrades significantly in real-world settings for reasons similar to those discussed for O_3 sensors. Moreover, field investigations report considerable variation in R^2 values. For instance, two studies employing metal oxide NO_2 sensors yield contradictory results. For the MICS-2710 sensor, part of the Air Quality Egg platform, there were observed R^2 values under 0.1 between sensor outputs and reference measurements, showing a poor performance [57]. In contrast, the RMSE values reported with the same sensor presented reasonable values ranging from 6.9–9.5 ppb, using a multiple linear regression model for calibration, considering temperature and humidity effects on the sensor's response [70]. Regarding electrochemical sensors, some studies have reported R2 close to 0.90 between the sensor response and the reference measurements after applying correction algorithms for interference by O_3 or humidity [71, 72, 73]. A more exhaustive study tested 24 electrochemical NO_2 hyperlocal sensors and reported R^2 = 0.04–0.52 during field tests at a reference station [26]. These results highlight that even for identical sensors and platforms, different results can be obtained, remarking the need of strong quality control in the manufacturing process for both sensors and platforms, and the corresponding calibration [28].

The use of electrochemical CO sensors in literature is lower than in NO_2 and O_3 sensors. In chamber conditions, there is an excellent agreement between the sensor output and reference measurements with $R^2 = 0.99$ [26, 72, 73]. However, the field research shows different insights and variations in sensor performances, according to *Figure 1.5*. Two field studies reported moderate to excellent R2 values (0.53–0.97) for the CO-B4 sensor [67, 73]. However, two other field studies have reported poor R2 values (0.17–0.45) for the CO-B4 and TGS-5042 sensors when calibrating them with reference measurements [26, 74]. Regarding metal oxide sensors, MICS-5525 sensor's response decreased linearly when the temperature was increased from 19 °C to 40 °C during chamber testing. The MICS-5525 sensor's reproducibility was moderate with R^2 between 0.38 and 0.60 [70]. A similar device MICS-4514 was tested under field conditions, reporting R^2 ranging between 0.76–0.78 compared to reference measurements when calibrated using simple or multiple linear regression models. However, the same models and device performed poorly after 4-5 months ($R^2 < 0.1$), pointing out the effect of ageing in sensor response [74].

Last, electrochemical Alphasense SO_2 sensors from the B series were tested in field conditions, obtaining R^2 values between 0.00 and 0.05. Authors argue that this result is logical since normal environments had very low SO_2 concentration [75]. However, these results change when the same sensor is tested after neural network calibration in an environment with SO_2 values ranging from 0-400 ppb. In this case, high R^2 values were obtained, ranging from 0.68 to 0.99, but a slight decrease in the sensor's sensitivity was observed after 18 weeks of validation [76]. From these results, it could be concluded that SO_2 sensors are more challenging in calibration and data improvement because of the low concentration ranges in normal conditions, and it is necessary to perform the validation in extreme environments.

Calibration Methodologies

The calibration of a sensor is a key step in data quality assessment. However, there is a wide range of methodologies for this purpose. The method used to calibrate hyperlocal air quality sensors is generally considered confidential information by most manufacturers. Thus, little information can be extracted about the calibration of devices that fall under the “black box” category compared to those that fall under the *open source* category. In contrast, several studies can be found on calibrating *open source* hyperlocal air quality sensors, both with laboratory and field tests. Calibration consists of setting a mathematical model describing the relationship between sensor data and reference measurements. The methods used to infer these relationships belong to the field of *machine learning*, in concrete supervised models for regression analysis. However, most of the calibrations were carried out during field tests, while only a limited number of laboratory-based calibration experiments were found. In this section, we summarize the results of a literature review in calibration methodologies for air quality sensors, as shown in Table 1.3, adapted from this work [77]. In addition, more inputs have been added to the original table to cover the most recent developments in the state of the art.

Next, the different models proposed in Table 1.3 are described.

- **Algebraic Regression Models**. This set of models includes algebraic functions that explains the relationship between the real values and the measure ones. In Table 1.3, *simple linear regression* (linear), *multiple linear regression* (MLR), *logarithmic regression* (log), quadratic regression (quad) and exponential regression (exp). The simplest model is the *simple linear regression* using the least squares method, in which one target value is modelled as a linear function of a single predictor. An extension of this model is the *multiple linear regression*, whose goal is, given an input vector $X^T = (X_1, X_2, X_3, ..., X_p)$, to predict the real value of the output vector Y, as defined in the following equation, where p is the number of features and β is the coefficient associated with each X_j. The quadratic, exponential or logarithmic extension are based on generate predictor by applying the quadratic, exponential or logarithmic functions to one or more of the original predictors. At the end, this is a multiple linear regression model in whihch some predictors are functions of other [99].

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- **Support Vector Regressors**. This kind of regressors are based on the *support vector machine* algorithm formulated at the end of 20th century and marked in 1.3 as SVR. Mathematically, let us define a training vector $(x_1, y_1), ..., (x_l, y_l) \subset \chi\mathbb{R}$, and the objective of the problem is to find a predictor function $f(x)$ that has at maximum deviation of *varepsilon* compared to the *target* value. Thus, it is possible to define a linear problem $f(x) =< w, x > +b$, where we seek to minimise the scalar product $< w, w >$, so the following optimisation problem can be posed,

Table 1.3: Performance of different machine learning models used for sensor calibration

Pollutant	Calibration Model	Records	References	R^2 Model	R^2 Comparison
CO	ANN	2	[74]	-	0.58
CO	linear	12	[73], [26], [78], [79], [74] , [80]	0.85	0.15
CO	MLR	21	[57], [77], [81], [70], [74] , [80]	0.89	0.83
CO	RF	1	[80]	0.91	-
NO	ANN	2	[74]	-	0.57
NO	linear	8	[26], [78], [74] , [77]	0.96	0.032
NO	MLR	20	[57], [82], [77], [74] , [81]	0.92	0.91
NO	RF	2	[82]	-	0.9
NO	SVR	2	[82]	-	0.90
NO_2	ANN	7	[83], [69]	0.87	0.94
NO_2	linear	25	[73], [69], [26], [78], [77], [80], [71]	0.25	0.17
NO_2	log	1	[84]	0.89	-
NO_2	MLR	48	[57], [85], [86], [69], [87], [82], [83], [77], [70], [81], [80]	0.81	0.81
NO_2	RF	7	[82], [83], [80]	0.86	0.91
NO_2	SVR	4	[83]	0.85	0.94
NO_2	SVR	2	[82]	-	0.78
O_3	ANN	2	[69]	-	0.89
O_3	linear	13	[73], [69], [26], [78], [77]	0.84	0.53
O_3	log	1	[84]	0.88	-
O_3	MLR	20	[57], [69], [77], [68], [81]	0.91	0.88
PM1	Kholer	2	[49]	-	0.74
PM10	linear	3	[88], [89]	0.77	0.73
PM10	quad	1	[59]	0.65	-
PM10-2.5	linear	4	[90], [91], [89]	0.63	0.98
PM2.5	exp	3	[92], [58], [63]	0.91	0.97
PM2.5	Kholer	2	[93], [49]	-	0.78
PM2.5	linear	37	[94], [48], [59], [88], [89], [61], [58], [95], [56]	0.84	0.64
PM2.5	MLR	17	[57], [85], [95], [56], [96]	0.81	0.65
PM2.5	quad	8	[97], [59], [95] [64]	0.87	0.88
PM2.5	RF	3	[96]	-	0.79
PM2.5–0.5	linear	9	[53], [98], [91], [89]	0.84	0.98
PM2.5–0.5	MLR	1	[57]	0.6	0.45
PM2.5–0.5	quad	6	[54]	0.82	-

which can be solved using quadratic programming [100].

$$\frac{1}{2}||w||^2$$

$$y_i - < w, x > -b \geq \varepsilon$$

$$< w, x > +b - y_i \geq \varepsilon$$

- **Random Forest Regressors**. This regression methods - RF in Table 1.3 - are based on *random forest* models, belonging to the boostrap aggregation methods, which seek to reduce the variance of the prediction function. It is a collection of several decision trees - Boolean structures that return an output value from an input value. Therefore, for a dataset of B variables $1, 2, ..., b$ a prediction function can be defined, which is represented in the following equation where X is the observation to be calibrated [99]. Random Forest has advantages over other regression methods because it is faster; it is robust to outliers and noise; it has good accuracy; and it is easy to parallelise computationally. However, its main drawback is that it is highly dependent on the training dataset (overfitting) [101].

$$f_{rf}^{b}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

- **Artificial Neural Networks**. Marked in Table 1.3 as ANN, artificial neural networks are models based on the combination of different neurons in complex architectures. A neuron is a linear function that receives a set for input predictors X pondered by a vector of weights W. To avoid linearinity, the output of the neuron is tranoformed by a non-linear function (g) called activation function. Thus, the output of the neuron can be defined by the following equation [102].

$$a = g(\sum_{i=1}^{D} w_i x_i + w_0) = g(\sum_{i=0}^{D} w_i x_i)$$

The training of a neural network consists on calculate the weights that minimizes the error compared to a target value. The *backpropagation* algorithm is based on calculating the weights recursively, starting at the output layer from the observed error and backpropagating the error to the previous layers. The method of updating the weights is the gradient descent method [103].

- **Kölher Correction**. Unlike machine learning methods, this strategy is based on the physical theory proposed by Köhler [104], that describes the relationship between the dry diameter of a particle and the wet diameter. Thus, it is an humidity correction approach. This algorithm is based in a parameter κ that is dependent on the aerosol chemical composition, so *a priori* in this aspect is needed. This method is only applicable for particles [105].

$$D_{Dry} = D_{Wet}\frac{RH}{1 + \kappa D_{Wet}\frac{RH}{D_{Dry}}}$$

Edge Computing in IoT Networks

The use of extensive hyperlocal air quality networks allows to apply complex architectures that divides the processing work in different nodes, that we will call as IoT-Edge-Cloud architectures. Several studies have demonstrated that machine learning models can improve the real-time data reported by IoT devices by increasing the accuracy of the raw data and generating more reliable datasets [80, 106]. In addition, this strategy could be used to forecast air quality data using sensor data, especially with deep learning approaches based on neural networks. The main problem that presents this machine learning approach is its integration into an edge computing architecture. Some studies overcame this limitation in other fields, such as intrusion detection in IoT systems [107]. In the smart cities domain, distributed machine learning has been applied to energy management [41] but not - to our knowledge - in air quality monitoring. One of the main functionalities that provides machine learning is the ability to perform pollutant predictions in time [108]. There are several techniques and strategies to apply machine learning in edge nodes.

- **Federated Learning**. This approach is based on the Distributed Selective Stochastic Gradient Descent (DSSGD), initially oreinted to preserve privacy in deep learning [109]. Federated learning allows each part of the system to keep its local model private while iteratively updating it by integrating gradients of others through a parameter server [110]. The main advantage of FL is that edge nodes exchange and aggregate local models, avoiding extra computation, and reducing communication overhead when ML model sizes are sufficiently smaller than data sizes [111].

- **Model Partitioning**. This paradigm consists in segmenting a deep learning model into different packages that are located into different nodes in the edge layer [110]. This strategy is linked to Kubernetes approaches that orchestrates from the cloud the partitioning and synchronization between models [112].

- **Right-Sizing**. Although the initial layers of the DNN reduce the size of the data, sometimes the intermediate result generated is still large, or the whole process is too slow. Therefore, a method for rapid inference is necessary. Thus, the network must be trained in a different way to generate good results before reaching the end of the network (early-exit), being able to give a result in real-time, without the need to send it to the cloud, thereby reducing the traffic. This method is known as model right-sizing [113], where the DNN model has different exit points and a shorter branch implies a smaller size and thus a shorter runtime. This mechanism focuses on adjusting its size to the limitation of the existing environment. Despite of the number of devices employed, the right-sizing methodology is oriented towards optimizing the utilization of external resources to expedite computational processes. In the context of Deep Neural Network (DNN) right-sizing, the primary focus lies in customizing the model dimensions to align with the limitations of the prevailing computational infrastructure. This necessitates the application of advanced training techniques to iteratively create a newly adapted model derived from the initial one [112].

- **Edge Only**. This approach is the most extreme in distributed learning, and the computations are all performed in edge devices or nodes, with the cloud server, when used, acting only as host for data storage [113]. Inside this paradigm, several framework has been developed, as the recent OpenEI, that offers a lightweight deep learning Package Manager similar to TensorFlow Lite, optimized to run AI algorithms at the edge, which guarantees low power consumption and low memory footprint. In adition, the model selection algorithm tool provided by OpenEI looks for the most suitable one for a specific edge platform based on users' requirements and the capabilities of the edge platform [114].

- **Model Compression**. Model compression is an efficient technique for encapsulating models on edge nodes with limited resources by altering the network architecture itself in an attempt to reduce its parameters, and therefore its size [113]. In this context, several frameworks using reinforcement learning has been developed for achieving model compression in a fully automated way, as AutoML [115] or TinyML [116]. The extreme use of compression leads to Binary NNs (BNNs), where the activations and weights are reduced to binary representations, with massive reductions in resource usage and costs for edge computing. Nonetheless, binarization must be applied very carefully to not prompt drastic performance and scalability issues in complex tasks [113].

To implement these functionalities in an edge ecosystem, tools like Foglets have been developed to optimise task deployment over distributed edges to save bandwidth and reduce latency [117]. However, more research on exploring the programming model for existing frameworks (e.g., Apache Storm and Spark) or developing new programming models with non-standardized interfaces is needed. FogFlow's programming has emerged in this context, allowing IoT service developers to program elastic IoT services easily over the cloud and edges. In addition, it supports standard interfaces to share and use contextual data across services, obtaining good benchmarking results for latency, throughput, and scalability [118].

1.3.2. FIWARE: Contextual Intelligence in Smarts Cities

Contextual intelligence has become more crucial for achieving IoT intelligence, efficiency, effectiveness, performance, and sustainability by enabling interactions between IoT devices such as sensors/actuators, smartphones and connected vehicles, to name but a few. Context management platforms (CMP) are emerging as a promising solution to deliver contextual intelligence for IoT [119]. Belonging to CMP, FIWARE platform is an open-source framework designed to enable the development of innovative and scalable smart solutions for various domains, such as smart cities, Industry 4.0, agriculture, and healthcare. It provides a set of core technologies and standards that facilitate the creation of interoperable and customizable applications [120]. Figure 1.6 summarizes the different components that can be found in a FIWARE architecture.

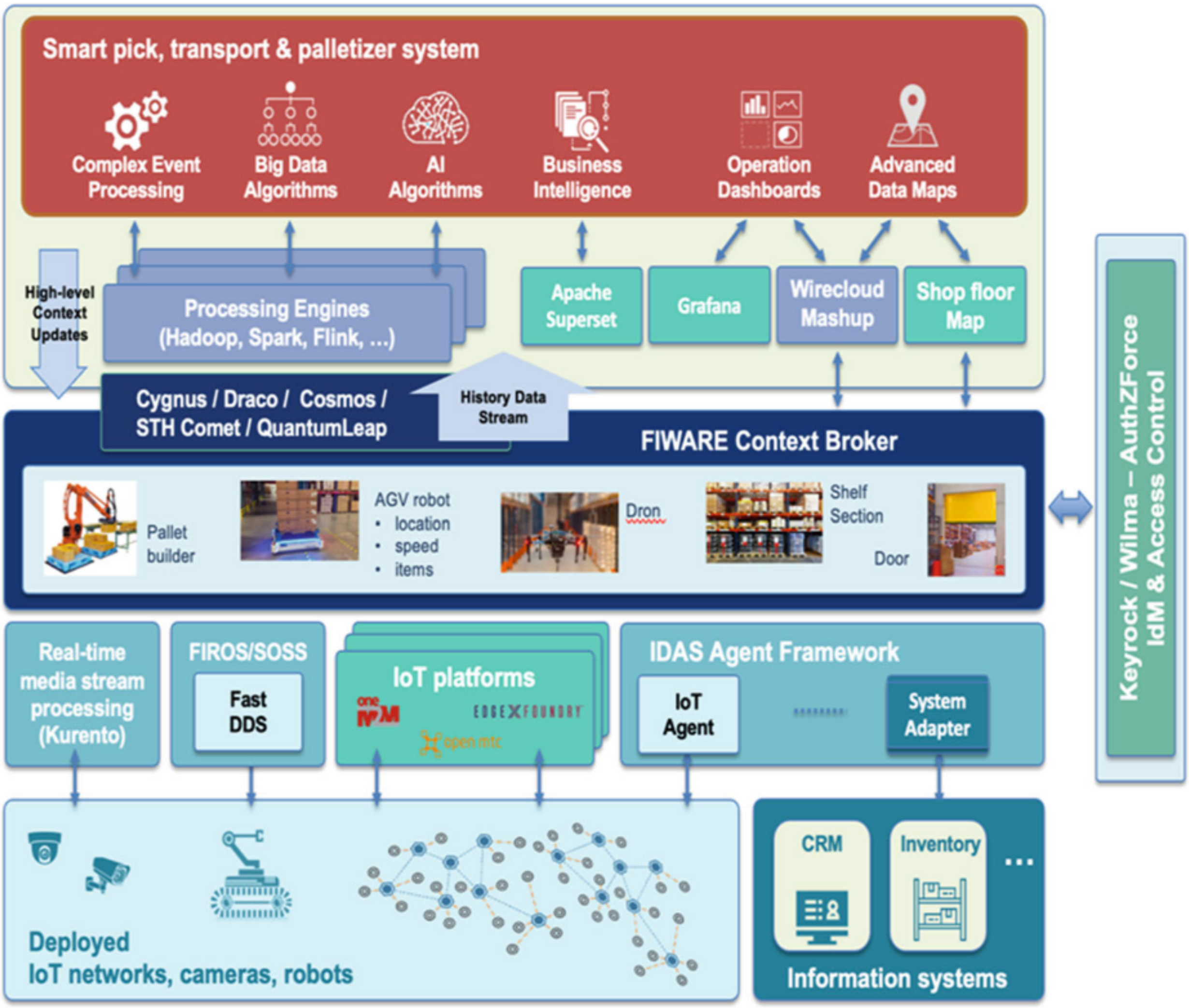

Figure 1.6: Catalogue of smart cities technological components (©2021, FIWARE).

The FIWARE architecture is based on generic enablers (GE), reusable software modules that extend the platform's functionality and facilitate the development of smart applications. FIWARE GEs cover various areas, including data processing, security, and visualization. The principal GEs that are part of the FIWARE platform can be broadly categorized as follows:

- FIWARE Context Broker: The Context Broker is at the heart of the FIWARE platform, managing and handling context information in real-time, using a publish-subscribe pattern providing an NGSI interface. Clients can create context elements, query and update them, and subscribe to changes in context information to receive notifications. Other elements interact with Orion through HTTP/HTTPs requests. Orion only saves the latest information state. It allows applications to access and update context data related to entities, such as devices, sensors, and users. The Context Broker uses the NGSI (Next Generation Service Interface) API, which supports simple query-based interactions and subscription-based mechanisms for real-time updates [121]. There are three main open source components used for FIWARE publish-subscribe systems: Orion-LD, Stellio and Scorpio [122].

- FIWARE NGSI-LD: The ETSI Industry Specification Group for cross-cutting Context Information Management (ISG CIM) named its API using the text string "NGSI-LD" with the formal agreement of OMA which originally defined NGSI. This greatly helps to avoid confusion with the IEC CIM/Common Information Model specifications. The rationale is to reinforce the fact that it leverages on the former OMA NGSI 9 and 10 interfaces [1] and FIWARE NGSIv2 [2] to incorporate the latest advances from Linked Data. NGSI-LD is an evolution of the NGSI API, aiming to align FIWARE with the Linked Data standards and Semantic Web technologies. NGSI-LD is designed to offer more flexibility, extensibility, and data interoperability, making it easier to integrate and exchange data with other systems [123].

- FIWARE IoT Agent: IoT Agents are a set of software modules handling South IoT Specific protocols and North OMA NGSI interaction. These agents allow to work with the IoT devices that use communication protocols like LWM2M over CoaP, JSON or UltraLight over HTTP/MQTT, OPC-UA, Sigfox, or LoRaWAN. This component will represent the IoT devices in a FIWARE platform as NGSI entities in a Context Broker [121]. Thus, they are crucial for integrating IoT devices and sensors with the FIWARE platform. They act as translators between the IoT protocols used by devices and NGSI data formats required by the Context Broker. These IoT agents could be connected to the context broker through a MQTT broker, as Mosquitto [124].

- FIWARE DRACO: Draco is a data management system[3], which builds on Apache NiFi[4]. It empowers users to create sophisticated and scalable directed graphs for data routing, transformation, and system mediation. The system utilizes an array of processors and controllers, enabling the conversion of incoming data into NGSI entities and attributes essential for publishing in the Context Broker. Additionally, the Draco system serves a dual purpose, functioning not only in the data acquisition block of the DT architecture but also in the persistence aspect [121]. Within Draco, specialised processors are responsible for persisting NGSI context data into external storage systems, enabling the creation of a historical data view within the Context Broker [125].

- FIWARE Keyrok: In the FIWARE Ecosystem, several GEs manage the authorization and authentication of users and devices. Generally, every request sent to a FIWARE GE must be

[1] Open Mobile Alliance, Next Generation Service Interfaces Architecture Approved Version 1.0 Published 29 May 2012

[2] Fiware, FIWARE-NGSI v2 Specification. See e.g. http://fiware.github.io/specifications/ngsiv2/stable/

[3] https://fiware-draco.readthedocs.io/

[4] https://nifi.apache.org/

authenticated using a token [121]. The Keyrok GE [5] is an Identity Management component based on OAuth 2.0. It is responsible for issuing the tokens for previously registered and authenticated subjects (users and devices). Keyrock relies on the OpenStack IdM implementation called Keystone [6], extending Keystone by implementing the SCIM standard. SCIM is intended to reduce the cost and complexity of user management operations through a common user schema, extension model, and REST API with a rich but simple set of operations [126].

In summary, the FIWARE platform offers a comprehensive and versatile ecosystem of tools and components that empower developers to build smart applications with a focus on interoperability, scalability, and openness. By adhering to open standards and facilitating data sharing, FIWARE encourages collaboration and accelerates the adoption of smart solutions across various industries.

1.3.3. Data Analytics Tools and Environments

In general, in our proposed *smart city* architectures, time-series analysis has been performed using relational databases, especially time-series databases. However, non-SQL databases have been used that incorporate a spacious range of distinct database technologies [127]. These are developed for designing modern applications. A non-relational or non-SQL database that delivers a process for procuring and retrieving data has been depicted, using data mining to address the tasks of gathering and processing sensor data [128]. Different technologies for different purposes in big data platforms are used in our reviewed papers, such as API, Navitia.io, CitySDK, SPARQL Query, R Programming, Predictive Analytics or Blockchain [129].

Relational Databases

A relational database is a set of data linked together following a logic, which is represented in tables with rows and columns to store information in these objects. The data is defined in the first column, the header column, and each row is associated with primary key, and each column has a datatype. In this model, each row corresponds to an observation or measurement that is collected in the database. SQL query language is the most used in relational schemas, so these databases are also known as SQL databases. In fact, some of them as Apache, PostgreSQL, CrateDB or BigQuery implement SQL dialects that come from the SQL standard [130]. In the context of *Smart Cities*, the following SQL databases are used in the state of the art.

[5] https://fiware-idm.readthedocs.io/
[6] https://docs.openstack.org/developer/keystone/

- **CrateDB**[7]. CrateDB is a distributed SQL database built on top of a cloud-native architecture. It combines the familiarity of SQL with NoSQL's scalability and data flexibility, enabling developers to: i) use SQL to process any data, structured or unstructured; ii) perform SQL queries at real-time speed; and iii) scale simply. It is widely used in the context of Smart Cities because it is FIWARE friendly and could be easily connected to the Context Broker through the QuantumLeap interface [131]. CrateDB has also been used in digital twin applications to manage cities or building resources [132]. CrateDB is available in the cloud, at the edge, and on premise to fit everyone's needs. Customers often use CrateDB to store and query real-time data. This is because CrateDB makes handling the velocity, volume, and diversity of machine and log data easy and economical [133].

- **PostgreSQL**[8]. This database system plays a crucial role in the development and implementation of Smart Cities around the world. PostgreSQL is a powerful, open-source object-relational database system based on the SQL language with over 35 years of active development. It has earned a strong reputation for reliability, feature robustness, and performance. It serves as a reliable platform for storing and retrieving data related to traffic management, energy consumption, waste management, public safety, environmental monitoring, and more. With its support for spatial data and advanced querying capabilities, PostgreSQL enables the integration of geospatial information, enabling smart city applications such as real-time traffic optimization, urban planning, and location-based services [134, 18]. Additionally, PostgreSQL allows for integrating various sensors and devices, facilitating data-driven decision-making and fostering innovation in smart city projects [135].

- **MariaDB**[9]. MariaDB Server is one of the most popular open-source relational databases. It was made by the original developers of MySQL and guaranteed to stay open source. It is part of most cloud offerings and the default in most Linux distributions. MariaDB has been used in some smart city applications, such as crowd monitoring, to manage the information connected to the user's system credentials, built throughout the development process, using the RESTful methodology [136].

- **MySQL**[10]. MySQL Database is a fully-managed database service powered by the integrated HeatWave in-memory query accelerator. It is the only cloud-native database service that combines transactions, analytics, and machine learning services into MySQL Database, delivering real-time, secure analytics without the complexity, latency, and cost of ETL duplication. It has been used in platforms to monitor radon in indoor environments, such as homes and workplaces [137].

[7] https://crate.io/
[8] https://www.postgresql.org/
[9] https://mariadb.org/
[10] https://www.mysql.com/

Non-Relational Databases

Non-relational databases comprise a high number of database systems that are not based on the relational model and data is queried using a non-SQL language, usually specific to each database. Some notable open-source NoSQL databases are Apache Cassandra, Apache HBase, and Apache CouchDB. Others are more focused on time series It is worth noting that the two largely popular databases Elasticsearch [88] and MongoDB [89] were open-source projects up to 2018 and 2021, respectively, when they switched their licenses to the Server Side Public License (SSPL), a source-available but not open-source license [135]. A special group inside non-relational databases is focused on time series management, as InfluxDB, TimescaleDB, and OpenTSDB [138]. Graph databases are another important group in non-relational databases, a scalable option for dealing with complex, semi-structured, and densely connected data. It is very fast in terms of queries and gives a response in milliseconds. Graph databases are highly useful in communication, healthcare, retail, financial, social networks, online business solutions and smart cities [139]. Graph database system follows CRUD (create, read, update, delete) methods used in a graph data model and uses index-free adjacency [140].

- **InfluxDB**[11]. InfluxDB is a NoSQL database designed specifically for efficiently managing time series data, whereas relational databases are more suitable for applications requiring complex relationships between datasets. InfluxDB has increased its presence in IoT architectures because it is compatible with dynamics visualization tools such as Graphana [141] and FIWARE friendly [142]. In addition, because of its target in time series, it is strongly recommended for architectures that will feed artificial intelligence services [143]. In the context of Smart Cities, recent studies demonstrate that time series databases could be used to improve the city system's performance and reduce storage requirements, as well as reducing the costs of smart city platforms [144].

- **Neo4j**[12]. This is a disk-based transactional graph database. Neo4j also supports another language like Python except for Java for graph operations. Neo4j is an open-source project in a GPLv3 Community edition, with Advanced and Enterprise editions available under both the AGPLv3 and a commercial license. Neo4j is the best graph database for enterprise deployment. It scales to billions of nodes and relationships in a network [145]. Regarding its feasibility in Smart Cities, recent studies demonstrate that Neo4j can be used in FIWARE platforms for different IoT solutions [146].

- **GraphDB**[13]. The model permits an explicit representation of graphs by defining object classes whose objects can be viewed as nodes, edges, and explicitly stored paths of a graph (which

[11] https://www.influxdata.com/
[12] https://neo4j.com/
[13] https://graphdb.ontotext.com

is the whole database instance). A database in GraphDB is a collection of object classes partitioned into three kinds of classes: simple, link, and path classes. There are also data types, object types, and tuple types [147, 148]. This database is a specialized in-store for RDF triples, which has been used in different initiatives in the state of the art as the validation of the proposed Smart City RDF Benchmark (http://www.disit.org/smartcityrdfbenchmark) based on Florence Smart City accessible as Km4City [149].

Analytic Environments

In the rapidly evolving data analysis landscape, the availability of robust and versatile tools has become paramount for efficiently extracting insights from complex datasets. Among the forefront contenders, JupyterLab, Apache Spark, and Airflow are indispensable pillars of modern data analysis workflows. JupyterLab, an evolution of Jupyter Notebooks, provides an interactive and user-friendly environment for data exploration and visualization. On the other hand, Apache Spark empowers data professionals with lightning-fast distributed processing capabilities, enabling them to tackle big data challenges easily. Apache Airflow orchestrates data pipelines by complementing these tools, offering automation and scheduling capabilities that streamline the end-to-end analytical process. This section provides a quick overview to the tools that will be used in this thesis as main work environments.

- **JupyterLab**[14]. It is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality. It is based on the concept of notebook, a shareable document based on the JSON formats, combining code, plain language descriptions and data.
- **Apache Spark**[15]. It is a multi-language engine for executing complex processes and performing data engineering, data science, and machine learning tasks on single-node machines or clusters. Apache Spark is a key environment in scaling machine learning applications and optimizing cost-consuming algorithms and big data processing queries.
- **AirFlow**[16]. It is an open source platform created by the community to programmatically author, schedule, and monitor workflows. It allows users to implement standard Python features to create their workflows, including date-time formats for scheduling and loops to generate tasks dynamically. This allows you to maintain full flexibility when building your workflows. For this reason, the combintion of AirFlow with Apache Spark is quite common in current big data applications.

[14] https://jupyter.org/
[15] https://spark.apache.org/
[16] https://airflow.apache.org/

1.3.4. Air Quality Modelling and Forecasting

An air quality model is a mathematical representation that combines and synthesises all the factors involved in air quality. The output of all these elements results in pollutant concentration's spatial distribution and temporal evolution. The importance of mathematical models in air quality prediction has increased, and there are two different approaches. The first one is based on statistical and empirical observation and is called *statistical models* while the second group aims to model the physical and chemical process that occurs in the atmosphere. These last models are named *deterministic* and typically solve the differential equations that represent the processes controlling the atmospheric processes and are used for a range of tasks, including developing new scientific understanding and environmental policies [150].

Statistical Models

This section includes a wide range of methods, some of them based on basic statistical principles and others framed in recent artificial intelligence technologies and the emerging field of machine learning. One of the most basic methods to predict air quality is based on decision trees, concretely in the CART algorithm (*Clasification and Regression Trees*). This method develops a decision tree by continuously splitting peak pollutant concentration data into two groups based on a single value of a selected predictor variable [151]. The selected predictor variable and the threshold cutoff value are determined by the CART algorithm, after being trained with reference data. The software identifies the variables with the highest correlation with the pollutant. It seeks to split the data set into the two most dissimilar groups. The splitting of the data set and tree development continues until the data in each group are sufficiently uniform. Predictor variables used in CART typically include meteorological data (i.e., temperature, wind speed, cloud cover, etc.), but may also include air quality data or other data such as the day of week or length of day. *Figure 1* shows a simple decision tree modelled by CART to predict O_3 daily maximum concentrations in Los Angeles. It is quite simple to forecast pollutant concentrations using the decision tree created by the CART analysis.

However, CART models are so inefficient to perform an accurate prediction. For example, slight changes in input variables may produce large changes in the output concentrations. This is an objective tool that can only predict pollutant concentrations based on information contained within the observed and forecasted data. Changes into weather conditions or emissions may not be reflected in the predictor variables and may cause uncertainty in the pollutant predictions. CART may not predict pollutant concentrations during periods of unusual emissions patterns due to holidays or other events; however, human forecasters can account for these changes and their potential impact on pollutant concentrations. But the main limitation is the fact that decision trees should be updated along the time. For example, the tree shown in *Figure 1* was developed in 1988, so it is not valid to predict real concentration nowadays.

Thus, the limitation of this model demonstrated the necessity to implement more complex models and treat the air quality prediction as a regression model. Regression is a statistical method for describing the relationship among variables. Regression equations are developed for air quality forecasting to describe the relationship between pollutant concentration (the predict and what is being predicted) and other predictor variables (e.g., temperature, wind speed, etc.) Regression equations have been successfully used to forecast pollutant concentrations in many areas [152, 153, 154].

Chemistry Transport Models

Deterministic air quality models describe the atmospheric dispersion of trace species by modelling different processes such as chemical transport, reactions, emissions and deposition. This methodology attempts to mathematically represent these important processes that affect ambient air quality. Air quality modelling actually requires a system of models that work together to simulate the emission, transport, diffusion, transformation, and removal of air pollution. These models have three modules or submodels: meteorology, emissions and air quality. The deterministic mathematical models calculate the pollutant concentrations from emission inventory and meteorological variables according to the solution of various equations representing the relevant physical processes [155].

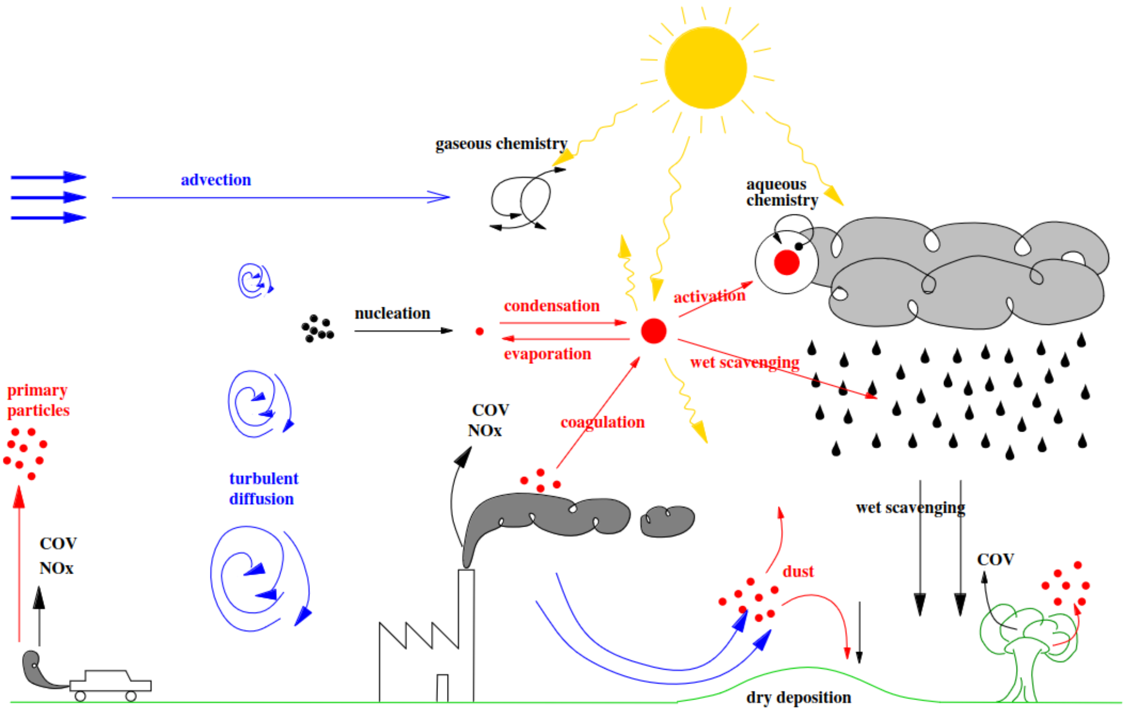

Figure 1.7: Proceses governing chemical concentration in atmosphere. Adapted from [155].

Moreover, deterministic models can be used to define and implement suitable control strategies for emissions and are a key tool in Sustainability Impact Assessment. Air Quality Models can be considered integrated into atmospheric models to make the governing equations, modelling, and computational algorithms reliable and compatible. The most important are *chemistry transport models*, which solves the mass continuity equation for each pollutant, where C represents the concentration vector (mass/volume) or density number (molecules/volume) of the trace species and uC is mass flux AND u is the wind velocity vector. This equation represents the mass conservation law.

$$\frac{\partial C}{\partial t} + \nabla uC = 0$$

This general equation should be adapted to if c_i stands for the concentration of a gas-phase species labelled by i, the time evolution is governed by a reaction-diffusion-advection partial differential equation, where x represents the vector with the three-dimensional coordinates, ρ is the air density, K the eddy coefficient matrix, χ_i is a function that represents the chemical reactions of the specie i, $Lambda_i$ is the scavenging function - it represents the amount of pollutant that are removed by rain and another meteorological process, and S_i is a function that represents the emissions of the specie i.

$$\frac{\partial c_i}{\partial t} + div(Vc_i) = div(\rho K \nabla \frac{c_i}{\rho}) + \chi_i(C, x, t) - \Lambda_i(x, t)c_i + S_i(x, t)$$

Also, initial and boundary conditions have to be defined. In differential equations, a boundary condition expresses the behaviour of a function on the boundary (border) of its area of definition. An initial condition is like a boundary condition, but then for the behaviour when it is derivated against time. In *chemistry transport models*, it is common to use the following definition of boundary conditions, where n is the vertical unitary vector (upward oriented), E_i the surface emissions and v_i is the velocity of dry deposition, computed as a parameterization of meteorological fields and the land use cover.

$$-K\Lambda c_i n = E_i(x, t) - v_i^{dep}(x, t)c_i$$

Thus, the aim of *chemistry transport model* is to solve this last equation for all the chemical species defined in the C vector. Note that more components are involved in the previous equation for aerosol species- multiphase chemistry. Different strategies exist to solve this equation, so it is possible to classify 3D AQM according to this criteria. The two principal approaches are *lagrangian* and *eulerian*. A *gaussian* method also exists, but this does not exactly solve this equation. The strategy selected depends on the desired application. The most important are the following: (1) risk assessment of chemical, biological or radiological releases (operational context); (2) environmental forecast; (3) impact studies or process studies [156]. Table 1.4 shows the principal 3D Air Quality Models and the classification of CHIMERE into the group 3D eulerian multiphase chemistry transport models.

Table 1.4: Clasification of different chemistry transport models in the sate of the art

Model	Type	Application	Reference
WRF-Chem	Eulerian	Air Quality Forecast	[157, 158]
CMAQ	Eulerian	Multiphase chemistry	[159]
CHIMERE-WRF	Eulerian	Air Quality Dispersion	[160, 161]
CALPUFF	Lagrangian	Pollutant Transport	[162]
AERMOD	Eulerian	Local Dispersion	[163]
HYSPLIT	Semi-Lagrangian	Trajectory Model	[164]
CAMx	Eulerian	Photechemical Modelling	[165]
GEOS-Chem	Eulerian	Atmospheric Composition	[166]
LOTOS-EUROS	Eulerian	Assimilation and Forecast	[167]
MATCH	Eulerian	Urban Modelling	[168]
MOZART	Eulerian	Ozone Modelling	[169]
TOMCAT/SLIMCAT	Eulerian	Aerosol Micropysics	[170]
POLYPHEMUS	Eulerian	Air Quality Forecast	[171]
TCAM	Eulerian	Aerosol Photochemistry	[172]
CLaMS	Lagrangian	Reverse Domain Modelling	[173]
FLEXPART	Lagrangian	Trajectory Model	[174, 175]
MOCAGE	Semi-Lagrangian	Air Quality Forecasting	[176]
GEM-MACH	Semi-Lagrangian	Aerosol and Ozone	[177]

Street Canyon Models

The fundamental process of enhancing urban environments to be both healthier and sustainable relies on effectively modeling air quality and its adjustable parameters. Consequently, investigating the influence of urban morphology on the dispersion of pollutants within cities and their subsequent interaction with the atmosphere emerges as a pivotal aspect of urban planning [178]. In this context, computational fluid dynamics (CFD) offers an exhaustive description of mean flow and turbulence, making them well-suited for dealing with complex scenarios like urban settings. Nonetheless, their applicability decreases when extended to prolonged simulation in time and when expanding the full spectrum of urban diversity across wide regions. These limitations come from constraints in computational resources and potential delays in processing times [179].

1.3.5. Semantic Technologies in the Air Quality Domain

Smart cities are complex socio-technical-economical systems in which different devices, smartphones, technological systems, services, and applications interact and exchange data through the Internet of Things (IoT) paradigm [180]. This development requires infrastructures and applications that manage big data generated from sensors and other sources [181]. Semantic models are still an effective strategy to address data and system complexity. Semantic interoperability in this context refers to the ability of various agents, services, and applications to share their large amounts of data, knowledge, and contextual awareness to benefit the city [182, 183].

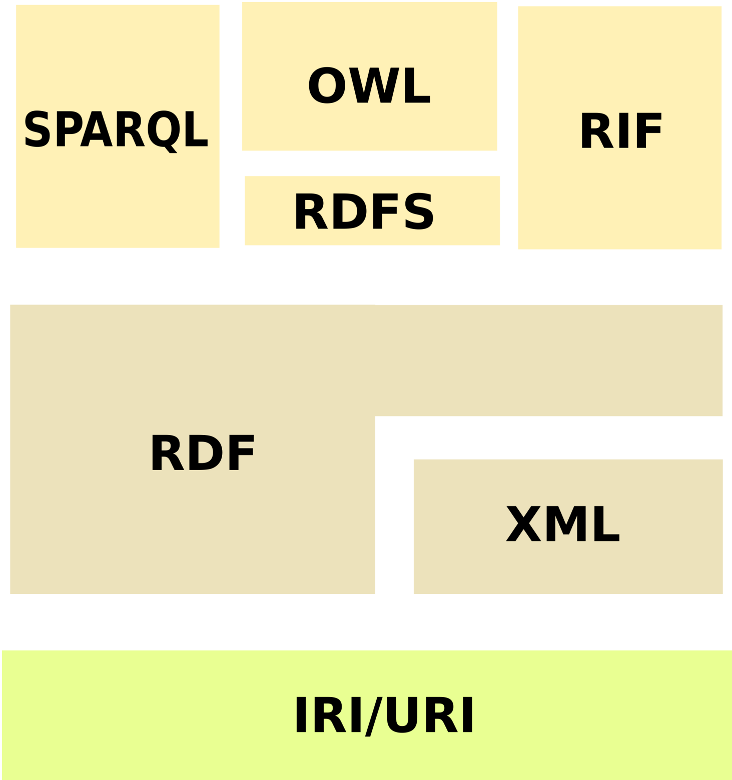

Figure 1.8: Semantic Web Architecture. On the bottom, the IRI/URI base provides an identifier for each resource. The next layer defines the RDF language to connect the resources using triplets as well as the XML scheme. On the top, are the languages that provide conceptualization (OWL) and logic (RIF) as well as the interface to query semantic data.

The Semantic Web is an extension of the web that aims to automate document processing and retrieval, providing information with a well-defined meaning that allows computers and people to work cooperatively [184]. The meaning of web resources is established by means of formal metadata, defined by new metadata using IRI/URI as unique identifiers, to be managed by software agents to managed by software agents to discover and integrate information. The semantic web architecture is a pile of language and technologies based on the IRI concept, as shown in Figure 1.8. The next layer defines the RDF language to connect the resources using triplets and the XML scheme. On top, there are languages that provide conceptualization (OWL) and logics (RIF) as well as the interface to query semantic data. All these languages will be defined in this section.

To explain this new application of semantic technologies, a new concept has arisen in the state of the art. The Semantic Web of Things (SWoT) merges Semantic Web technologies with the Internet of Things (IoT) domain to establish knowledge-based environments that facilitate advanced ontology-based functionalities [183]. The SWoT paradigm connects devices, objects, and phenomena with ontology-based annotations, enabling detailed, formal, and well-structured definitions. This approach provides unambiguous semantics, empowering automated reasoning tools to deduce new knowledge from comprehensive and integrated data views. As we move from IoT to IoE (Internet of Everything), the transition entails a corresponding progression from SWoT to the Semantic Web of Everything (SWoE) [185]. It is easy to note that in this new framework, semantic technologies forge synergies and interconnections between individuals, objects, processes, and data via the Web.

Semantic interoperability between IoE entities goes beyond the data exchange format or the explicit description of data models. It relates to a technique that automates this process without needing any configuration or programming [186]. Not only must the data be shared between two or more systems, but it must also be understood by each system. Sensors use self-defined data formats like JSON, CSV or XML to store their data. However, the data and schemas used by different hyperlocal IoT devices are not compatible. Also, in the air quality domain, information is also presented in many different units of measurement (ppb, ppm, $\mu g/m^3$). Hyperlocal device systems face interoperability issues as a consequence of disparate data description formats and incompatibility in the semantics of data representations. When data exchange is standardized and given a common meaning, regardless of the structure and semantics of the source content, semantic interoperability is accomplished [182].

Semantic Web Languages

One of the main goals of the semantic web is to enable computers to understand the implicit semantics of the concepts and resources defined among the *world wide web*. In other words, the semantic web seeks to extend the network of human-readable resources by adding machine-understandable metadata about resources and their relationships with other resources. For this purpose, semantic web languages formally describe concepts, terms, and relationships within a given knowledge domain to be used to write the metadata. There are essentially three types of languages: i) data representation languages (e.g., RDF), ii) knowledge representation languages (e.g., OWL) and iii) rule-based languages (e.g., RIF) [187].

- **Resource Description Framework (RDF).** This language is a standard for Web metadata that the World Wide Web Consortium (W3C) developed. It is based on expanding the traditional notion of document metadata, being a suitable language for describing any Web resource. As such it provides interoperability between applications that exchange machine-understandable information on the Web. RDF is based on concatenating resources in subject-predicate-object triplets, allowing the connection of the resources in complex knowledge graphs [188]. RDF Schema (RDFS), which is an extension of RDF, provides additional constructs to increase the expressivity of the language [187].
- **Web Ontology Language**. The OWL family of languages are "object-oriented" languages for defining and instantiating ontologies [189, 190, 191]. This language provides specifications on how to derive logical consequences. For example, if an individual *Peter* is an instance of the class Student, and Student is a subclass of *Person*, then one can derive that *Peter* is also an instance of *Person* in a similar way as it happens in RDFS [187]. However, OWL is much more expressive than RDFS [192].

- **Rule Interchange Format**. The RIF language aims to become a standard in the interchange of logical information, especially rules. A rule is perhaps one of the simplest notions in computer science. It is a Horn formula based on an IF-THEN construct. If some condition (the IF part) that is checkable in some dataset holds, then the conclusion (the THEN part) is processed. Deriving somewhat from its roots in Logics, rule systems use a notion of predicates that there are many rules languages in existence, and what is needed is to exchange rules between them [187].

  That is, RIF focuses on exchange rather than trying to develop a single one that fits all rule language because, in contrast to other Semantic Web standards, such as RDF, OWL, and SPARQL, it was immediately clear that a single language would not satisfy the needs of many popular paradigms for using rules in knowledge representation and data modelling [193]

- **The Query Language SPARQL**. This language is an approximation to perform queries in RDFS datasets and knowledge graphs. From a logical perspective, it can be considered a notion of the *conjuctive/disjunctive* query model [194]. For this reason, SPARQL is a query language for databases and has much in common with SQL. This language is widely extended in biology and chemical datasets, as most of them are based on RDF triples. This approach makes developing a dedicated server that presents the data and supports data querying easy [195].

Ontologies

In artificial intelligence, an ontology is an explicit model of the concrete knowledge domain. There are several definitions of ontologies in literature and across different fields of informatics. In knowledge representation and information modelling, an ontology is a formal and structured representation of a conceptual framework [196]. In addition, ontologies can also be defined as collections of computer-readable term definitions of the domain, specifying sets of instances/data (concepts/classes), their properties, and binary relationships between them [197]. In any case, ontologies are basic pieces in the semantic process and are the cornerstone technology for linked data, knowledge graphs and annotation procedures.

An ontology is a formal conceptualization of a domain of interest and consists of the following three different syntactic categories: i) entities, such as classes, properties, and individuals, are identified by URIs and can be thought of as primitive terms or names. Entities represent basic elements of the domain being modelled; ii) Expressions represent complex notions in the domain being modelled. For example, a class expression describes a set of individuals in terms of the restrictions on the individuals' features; and iii) axioms are statements that are asserted to be true in the domain being modelled [187].

- **Linked Data (LD)**. It serves as the foundational element for the semantic web, with the aim of establishing human-machine understandable links between datasets. LD provides the necessary methodologies to create and leverage these relationships effectively. It involves a set of design principles that facilitate the exchange of machine-readable, interconnected data across the Web [182]. By utilizing Semantic Web technologies, the Open Government Data (OGD) standards of integrity, participation, and collaboration are harnessed to integrate citizens into the smart city paradigm [198].

  The combination of LD with Open Data gives rise to Linked Open Data (LOD), where data is freely available for use and distribution [199]. Tim Berners-Lee and colleagues have highlighted publishing and linking techniques for structured content on the World Wide Web, involving the use of URIs as labels to identify entities. HTTP URIs enable individuals to search for these labels, providing valuable information when a URI is queried. Additionally, standards like RDF, NGSI-LD and SPARQL are employed, and links to other URIs are included to facilitate further exploration of related entities [200].

- **Knowledge Graphs**. Ontologies are the basis of the recent knowledge graphs (KG). A general ontology specifies the schema of a domain but does not include information about specific domain individuals. By adding data instances to ontological terms (to semantically describe the data with suitable metadata), a knowledge graph is generated. Thus, it is possible to define a knowledge graph as an instance-populated ontology with data and individuals [183]. In this process, ontology engineering is a key point in a methodological approach to the development of knowledge graphs [201].

  Knowledge graphs use a graph-based data model to capture knowledge in application scenarios that involve integrating, managing and extracting value from diverse sources of data at large scale [202]. Through this abstraction, the graph provide a concise representation where edges describe complex relations between the nodes that represent the entities of the domain [147], allowing to apply several artificial intelligence algorithm as *rule mining* or *graph neural networks* [203, 204]. Scalable frameworks for graph analytics can be leveraged for computing centrality, clustering, summarising, and regression, to gain insights about the domain being described [205].

  Recent studies proposes climate change mitigation solutions by proposing a climate knowledge graph for the integration of multiple climate data and other data sources into one service, leveraging Web technologies (e.g. HTTP) for multi-source climate data analysis. One of the challenges in this approach is the need to integrate heterogenous data from different sources in the knowledge graphs, that means to work with not FAIR formats as netCDF arrays, aiming to boost the cross-domain climate related research [206].

- **Semantic Annotation**. Enhancing the interoperability of dynamic environments like smart cities can be achieved by adding semantic annotations to the descriptions of devices, objects, and phenomena [207]. This involves annotating data resources with semantic metadata. The process begins with entity identification, followed by entity disambiguation, and finally, the annotation of data to the resources. These steps may vary depending on the resource type. Within a service-oriented architecture, the functionality of real-world objects and data are characterized by semantic services accessible through a unified interface, thereby enhancing the capability of diverse entities to interoperate seamlessly [208]. By semantizing and creating services over data, automation and dynamic provisioning of entity search, selection, and negotiations, among other activities, become feasible [209].

  Various semantic annotation methods primarily focus on mapping sensor information to established ontologies using specific mapping languages like D2RQ [210] and R2RML [211]. Additionally, mapping languages like RML (an extension of R2RML) aim to translate heterogeneous and multimodal data into RDF, independent of the input data format [212]. SPARQL-Generate facilitates RDF production from any RDF dataset and a collection of documents in arbitrary formats [213]. A recent work, RDF-Gen, proposes a consistent, efficient, and scalable method for constructing RDF knowledge graphs from diverse streaming and archived data, outperforming RML and SPARQL-Generate in terms of speed, scalability, and usability [214].

  The process of data annotation in essential in different science domains. For instance, trying to capture all information on samples, instruments, reagents, and research objectives, for analytical chemistry methods is hot topic in the state of the art. A simple approach consists on capturing those key steps of the method that are either transformative or generative, that is, materially change a sample, generate information on the sample (by an analytical procedure such as mass spectrometry), or transform the data from one form into another [215]. This is quite advanced in genomic science, in which the Gene Ontology effort (GO; https://geneontology.org) is the most comprehensive and widely used knowledge concerning the functions of genes. In GO, all functional knowledge is structured and represented in a form amenable to computational analysis, essential to support modern biological research. The GO knowledgebase is structured using a formal ontology by defining classes of gene functions (GO terms) with specified relations to each other [216].

Vocabulary Standardisation

There is an increasing interest in the development of data spaces, for which the interoperability of heterogeneous data need to be achieve.

This is a key principle observed in the design of the World Wide Web: content providers publish web pages on web servers (endpoints), knowing that web browsers can connect to them and retrieve web pages whose content they can render and display to end users. It means that all data-spaces

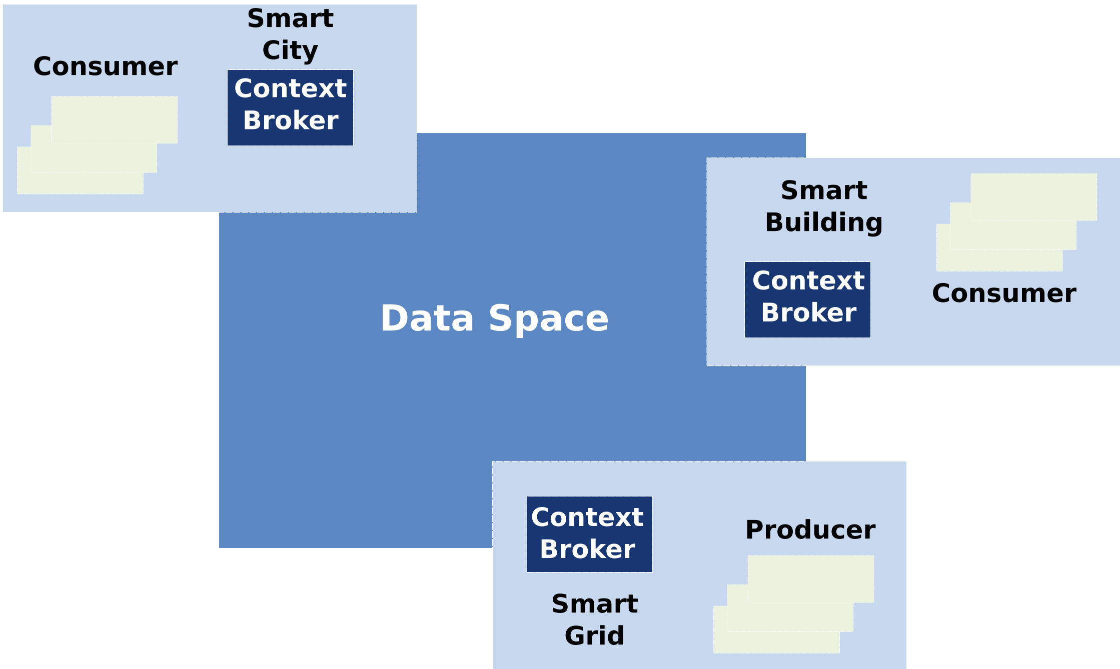

Figure 1.9: Data exchange in a data space

participants should “speak the same language,” translating into adopting domain-agnostic common APIs and security schemas for data exchange.

Thus, it is essential to use standardised vocabularies for all communities participating in a data space. For instance, if all the communities define, let say, ”no2” as it is done in a web resource identified by an IRI, it is possible to guarantee interoperability between applications and algorithms. In the air quality domain, Table 1.5 lists several ontologies along with their corresponding descriptions. The first ontology, CHEBI[17], describes chemical entities and small molecules of biological interest. The second, ENVO[18], is centered around describing environments and pollutants. On the other hand, FIWARE Smart Data Models[19] are not an ontology, but provide standardised vocabularies related to Smart Cities. GEO[20] deals with describing geographical entities. MAMO’s[21] scope is to describe mathematical models. Lastly, the SAREF ontology[22] is oriented towards describing IoT ecosystems.

From this table, it is easy to note that there is a wide range of vocabularies available in the air quality domain. In addition there is some redundancy as some terms are defined by different IRIs. For example, the ”no2” is defined by FIWARE and also in the CHEBI ontology. For this reason, a process of standardisation and unification is needed in the domain. Next, the main vocabulaires are explored in more detail.

[17]https://www.ebi.ac.uk/chebi/
[18]https://sites.google.com/site/environmentontology/
[19]https://smartdatamodels.org/
[20]http://purl.obolibrary.org/obo/geo.owl
[21]http://identifiers.org/mamo
[22]https://saref.etsi.org/

Table 1.5: Selected ontologies for the air quality domain

Ontology	Description
CHEBI	Describe chemical entities and small molecules.
ENVO	Describe environment and pollutants.
Smart Data Models	Describe smart cities vocabularies
GEO	Describe geographical entities
MAMO	Describe mathematical models
SAREF	Describe IoT ecosystems

- **CHEBI**. Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on "small" chemical compounds. The scope of this ontology are molecular entities that are either natural products or synthetic products, which have an effect on the processes of living organisms [217]. ChEBI finds extensive utility across diverse applications, serving as a reliable reservoir of distinctive and unchanging identifiers for chemical substances in annotations within a broad spectrum of bioinformatics databases [218]. Furthermore, it plays a crucial role within the Semantic Web. A notable instance is the recent translation of PubChem database content into RDF format, where ChEBI classes were effectively employed to furnish the rdf:type classification for the chemicals contained in PubChem's RDF representation [219]. Figure 1.10 shows the individual in the CHEBI ontology and how it is modelled.

  It could seem that CHEBI is not an ontology related with air quality, but all the main pollutants described in *Directive 2008/50/EC* are defined in CHEBI ontology as small molecules that has an effect on metabolomics - this could be considered as an additional definition of pollutant, but it does not apply to particles. In addition, in metabolomics field, a substantial improvement has been achieved in detecting and identifying new biomarkers of environmental pollutants through molecules identification and annotation [220]. For instance, CHEBI has been used to identify effects on metabolism related to emissions from refineries [221]. Thus, subsequent efforts should focus on investigating these ontologies applied to social determinants of health. This encompasses ontologies like ENVO , CHEBI , and ECTO (Environmental conditions, treatments, and exposures ontology) [222].

  However, CHEBI is more than an ontology. It is also public web application that has undergone a comprehensive overhaul and enhancement, focusing on aspects such as page arrangement, adaptability to various devices, menu architecture, and an overall augmentation of the interactive user encounter. An integral facet of this modernization was the replacement of all Java applets with suitable JavaScript components and libraries. Historically, Java applets led to persistent user dissatisfaction due to their inconsistent performance across diverse platforms and their tendency to be cumbersome for downloading, particularly impeding usage under sluggish connections. Additionally, concerns regarding the security of Java in web browsers emerged, resulting in occasional blocking of certain applets due to browser security settings [218]. Inside the services provided by ChEBI web application, it is worth to note BiNChE, an

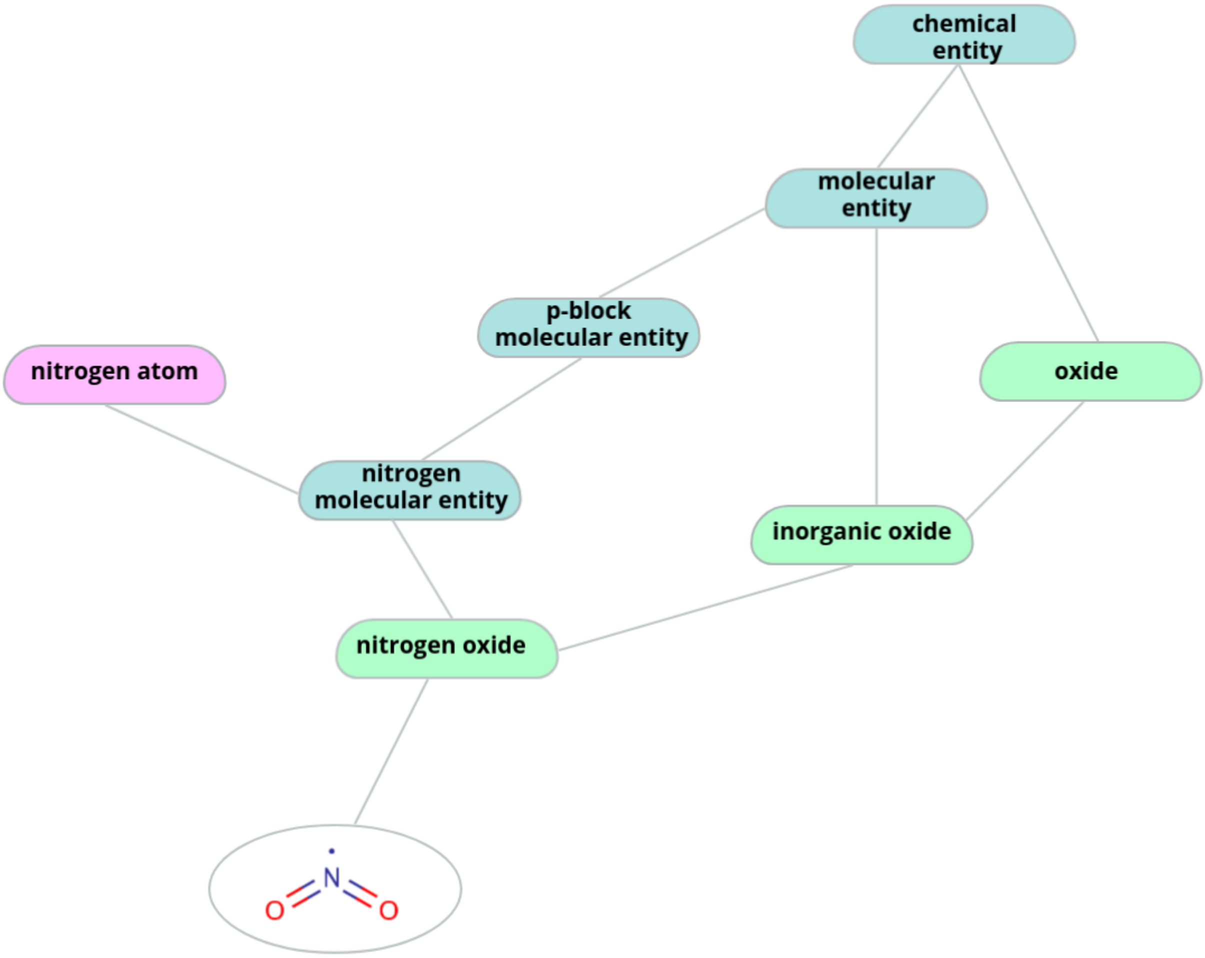

Figure 1.10: Model proposed for nitrogen dioxide in the CHeBI ontology.

enrichment analysis tool [23], which is also available as a software library. The tool offers plain or weighted analysis options against the ChEBI role, structure or combined ontology [223]. Also, OntoQuery is another usefull serice which allows Description Logic queries in the easy to use Manchester syntax to be executed against the pre-loaded and pre-reasoned ChEBI ontology [224].

- **ENVO**. The Environment Ontology is a community-led, open project which seeks to provide an ontology for specifying a wide range of environments relevant to multiple life science disciplines and, through an open participation model, to accommodate the terminological requirements of all those needing to annotate data using ontology classes [225]. ENVO provides descriptions of habitats, environmental processes, human-influenced surroundings, and entities pertinent to endeavors in environmental health and the worldwide Sustainable Development Agenda for the year 2030. Various segments of ENVO have served as platforms for fostering and initiating novel ontologies in hitherto unaddressed areas, such as the domains of food and agronomy [226].

[23] http://www.ebi.ac.uk/chebi/tools/binche/

- **FIWARE Smart Data Models**. This initiative launched by the FIWARE foundation aims to provide a library of standardised data models described in JSON-LD format and compatible with NGSIv2/NGSI-LD APIs and other RESTful interfaces, always in compliance with the Open API specifications[24]. Data Models published under the initiative are compatible with schema.org and other standards as ontologies. Since its creation, more than 500 data models have been published, and the number of organizations contributing data model descriptions is constantly growing [227]. All published data models and schemas are available in the following link[25].

  The strength of the NGSI-API is the fact that is domain-agnostic. In other words, many different systems have been developed using NGSI APIs in different domains transversal to smart cities, as smart energy, agriculture, ports or health, to mention a few. It is easy to note that this approach facilitates data sharing, because each system participating in the architecture will be publishing data that enriches a digital twin data representation of the city. These systems participating in digital twins do not know what other systems may consume their published data. Still, they will be able to assess concrete terms and conditions for accessing/using this data in a FAIR way [227].

  Data models are grouped into domain subjects, which are referred from repositories associated with the multiple application domains being considered. It should be highlighted that subjects are very specific to a given application domain. In contrast, others may be relevant to multiple domains. For instance, *weather* that is relevant to almost every domain or *sewage* that is relevant to the Smart Cities and Smart Water domains [227]. Figure 1.11 shows the publication process of a data model as well as the twelve domains available at this moment - at the end of 2023.

[24] https://smartdatamodels.org/
[25] https://github.com/smart-data-models

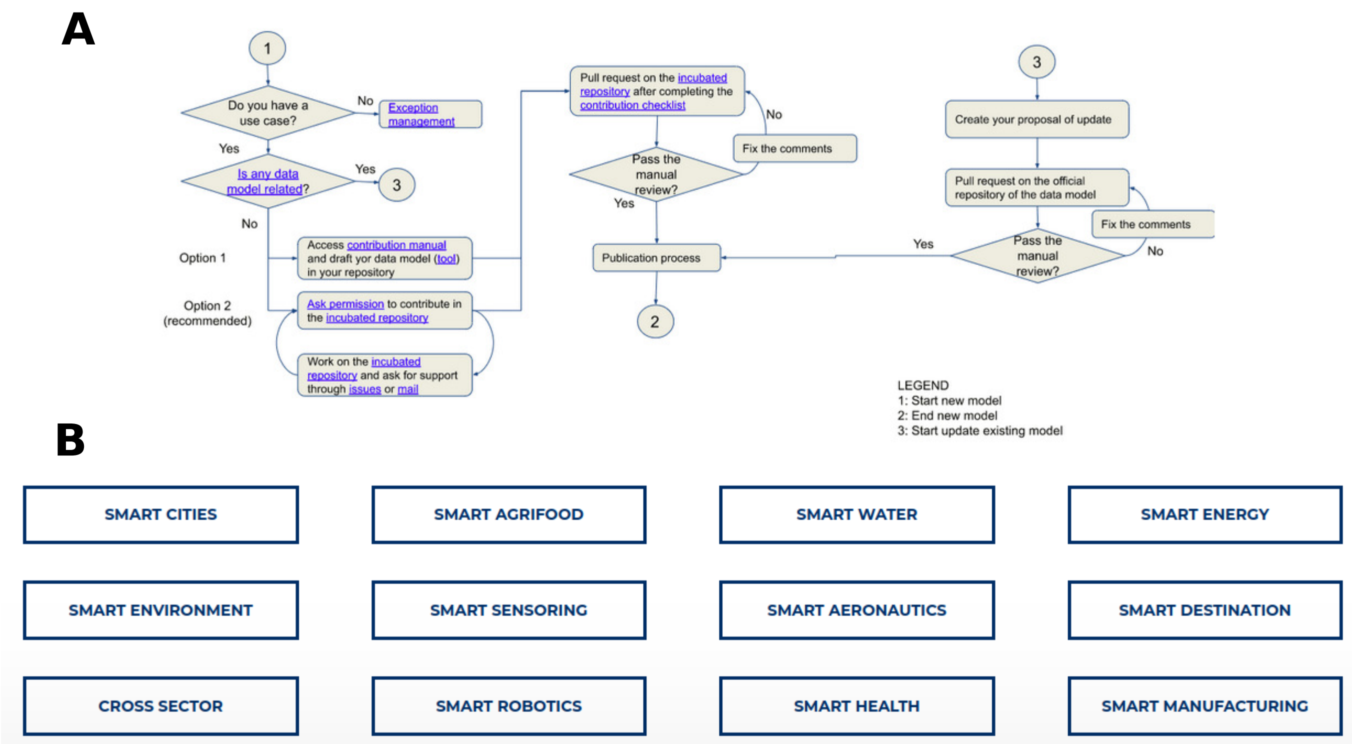

Figure 1.11: A. Workflow that should follow a smart data model to be published in the GitHub repository. B. List of available domains at the end of 2023 (©2021, FIWARE)

Chapter 2

Challenges, Objectives and Hypothesis

2.1 — Overview

The *Challenges, Objectives, and Hypothesis* section is a key component of this work focused on sustainability impact assessment. This section lays the foundation for the entire research, outlining the key problems, goals, and anticipated outcomes. The primary purpose of this section is to provide readers with a comprehensive understanding of the research's context and to set the stage for the subsequent chapters. This is a key point in assessing the potential contributions of the research to the field of sustainability impact assessment. Emphasis is placed on how the study's findings could advance knowledge, inform policy decisions, or provide practical insights to industry stakeholders.

However, reaching this milestone is not easy because of the complexity of sustainability assessment. It should be noted that sustainability is a multidimensional concept, encompassing economic, social, and environmental aspects, making its assessment inherently intricate. In addition, there is a lack of standardized methodologies due to the innovation related to this field and the continuous changes of the understanding of sustainability and climate change mitigation policies in the society and the legislation, which make it difficult to run comparisons and benchmarking. Other key point is that gathering reliable and comprehensive data on sustainability indicators can be challenging due to the variety of data sources and the quality of available information. Last, it is essential that the solutions proposed by this thesis are able to engage relevant stakeholders in the assessment process, but ensuring their participation and cooperation is a challenge.

If we analyse this in a holistic point of view, the thesis goals and hypotheses should develop a comprehensive sustainability impact assessment framework, aiming to create an integrated and adaptable framework that considers economic, social, and environmental dimensions of sustainability. This framework should be able to evaluate the sustainability performance of a specific action like a project, policy, or industry taken by a stakeholder, such as a city, an industry or a policy maker (KPIs). This framework should be able to compare different sustainability strategies or scenarios and recommend sustainable practices to stakeholders for enhancing sustainability outcomes.

From a technical point of view, the objectives and hypotheses of this thesis should be related to: i) improving the quality of the particulate matter sensors to allow stakeholders to make decisions based on data, ii) implementing algorithms that allow computing the impact of concrete actions, iii) assuring the interoperability of the different components of the platform, iv) guaranteeing the trustability and reliability of the solutions developed and v) trying to achieve replicability of the different components in different uses cases.

Thus, this chapter provides a comprehensive overview of the research scope, purpose, and expected outcomes. This chapter sets the initial point for the subsequent ones by addressing the challenges found in the state of the art, outlining clear and feasible objectives, and formulating testable hypotheses to validate the objectives proposed. It establishes the relevance of this work in the broader context of sustainability impact assessment.

The next sections define the main challenges, formulate the hypothesis to be validated in this thesis and detail the objectives to be achieved in this work.

2.2 — Challenges

The main challenge for this thesis is connecting and integrating the different heterogeneous elements described in state of the art into a sustainability impact assessment framework. These elements include the connection of hyperlocal air quality devices, FIWARE general enablers (GE), persistence databases systems and air quality models. Despite these elements being developed independently in state of the art, this thesis is the first approach to integrating all these components in a unique framework for sustainability impact assessment.

In concrete, the following challenges have been addressed during this work, named *CHX* to guarantee traceability in the conclusions chapter.

- **Low reliability, accuracy and replicability of hyperlocal air quality IoT devices in Smart City applications (CH1).**

  As has been exposed in the state of the art, electrochemical (toxic gases) and optical sensors present several problems to be solved. For particles, the main challenges are to deal with i) the influence of particle size and size distribution in the sample, ii) the effect of the chemical composition of the particle because organic compounds have a higher capacity to absorb light than inorganic compounds; and iii) the effect of humidity due to an increase of the aerodynamic radius of the particle [228, 229]. Regarding this last point, it is easy to note that particle mass scales with particle size to the power three, so small errors in particle classification led to large errors in calculated mass.

$$m = \sum_{i=0}^{n} \frac{4}{3}\pi r^2 \rho$$

This error could be quantified through a growth factor:

$$g(RH) = D_{Wet}\frac{RG}{D_{Dry}}$$

Thus, the challenge here is to reduce this hygroscopic factor by implementing drying systems in hyperlocal devices or quantifying this factor and correcting the signal obtained by the sensor.

- **Integration of chemistry transport models in sustainability impact assessment frameworks (CH2).**

  Chemistry transport models are powerful tools to compute air quality dispersion and forecasting in sustainability impact assessment. However, the main limitation of this model is the computational cost when increasing the resolution. This is due to the Courant-Friedrichs-Lewy (CFL) condition. This condition is mandatory for convergence when partial differential equations are solved by finite difference. The CFL for a grid box, as presented in the following equation, is a ratio between the mean wind speed v_{xi}, the time-step Δt and the cell size in all three directions i as Δx_i. The more the wind speed is high, the more the time-step has to decrease. The wind speed varies every minute. The best way to optimize the numerical cost of a simulation is to adapt the time step [230].

$$CFL = \Delta t \sum_{i=1}^{3} \frac{v_{xi}}{\Delta x_i}$$

  From this equation, it is easy to deduce that when the cell size decreases, the CFL increases, and the computational time increases. Thus, one key challenge is developing superresolution chemistry transport models by simplifying the model by developing metamodels based on machine learning methodologies. However, this approach should keep a compromise between high resolution and computational accuracy. This is also important by implementing sustainability impact assessment applications like air quality zoning [231] or impact health computation [232], which requires high-resolution data.

2.3 — Hypothesis

The overall hypothesis of this doctoral thesis is that by following a bottom-up methodology and integrating the different components proposed in Figure 1.1 - the sensing layer, the transmission layer, the data layer and the application layer - it will be possible to implement a sustainability impact assessment platform for smart cities.

This global hypothesis can be split into the following technical hypotheses, related which each one of the different components in the platform

- **The combination of hardware improvements, artificial intelligence algorithms and chemistry transport models will improve the accuracy of hyperlocal IoT sensors (H1).**

  This hypothesis states that the convergence of three key factors - hardware improvements, artificial intelligence algorithms, and chemistry transport models - will significantly enhance the precision and reliability of hyperlocal Internet of Things (IoT) sensors. Firstly, hardware improvements such as humidity correction systems will likely provide these sensors with enhanced capabilities, allowing them to capture and process data more efficiently and accurately. Integrating advanced artificial intelligence algorithms will enable the sensors to correct and calibrate data in real time, helping stakeholders in decision-making. Last, incorporating chemistry transport models will introduce the possibility of creating soft reference stations, enabling the sensors to be calibrated in areas without reference methods. Thus, this combination has synergy with hyperlocal IoT sensing, providing valuable and precise information for various applications, including environmental monitoring, smart cities, and personalized health tracking.

- **The use of an IoT-Edge-Cloud architecture will improve the performance of the system by reducing latency and bandwidth usage (H4).**

  This hypothesis proposes that the distribution of computational work in an IoT-Edge-Cloud platform will produce faster processing and a lower bandwidth. The use of the edge computing paradigm will allow to integrate machine learning models in a way that the computation cost is reduced and it is feasible to use in real environments. In this thesis, we propose a short-term forecasting algorithm to predict particulate matter excess in indoor environments, that will be benefited of the use of the edge paradigm. To test this hypothesis, an Smart Spot has been installed in the center of the Libelium laboratory, to assess the level of particles and create alerts when the recommended levels are overcome.

- **It is possible to integrate chemistry transport models in IoT platforms to validate sustainability impact assessment applications in different scenarios (H3).**

  This hypothesis suggests the feasibility of integrating chemistry transport models into Internet of Things (IoT) platforms to validate sustainability impact assessment applications across various scenarios. Incorporating these complex models into IoT frameworks makes it possible to simulate and analyze complex chemical interactions and transport processes within the environment. This integration would enable the assessment of sustainability impacts in diverse situations, such as urban environments, smart cities, air quality zoning or health impact assessment. By leveraging real-time data from IoT sensors and chemistry transport models, it is possible to improve the reliability of these models by comparing them with more information sources and improving resolution.

2.4 — Objectives

The main objective of this thesis is to implement a sustainability impact assessment platform that helps stakeholders implement sustainable policies based on data. This platform comprises different layers presented in the state of the art. For each layer, specific objectives and sub-objectives are defined.

In concrete, the following technical objectives and sub-objectives are defined according to te hypothesis defined before, named as OX and $OX.Y$ for objectives and sub-objectives, respectively.

- **Improve the quality and reliability of hyperlocal air quality systems (O1).**
  - *Improve data measurement by implementing humidity correction systems (O1.1).* This objective aims to increase the performance of air quality sensors compared to reference measurements in terms of regression metrics. This objective will be achieved if the values of those metrics improve the values exhibited by the state of the art approaches, and if the humidity inside the device is reduced in comparison with the humidity outside the system.
  - *Implement architectures based on edge computing to implement short-term forecasting services (O1.2).* This objective aims to use edge computing to implement accurate short-term predictions by using soft machine learning models. This objective will be achieved if the values of the prediction accuracy improve the value presented in the state of the art, and the edge paradigm improves the performance of the system in terms of latency and bandwidth usage.
- **Integrate chemistry transport models in IoT platforms and improve its performance by using hyperlocal ground measurements (O2).**
  - *Improve performance of chemistry transport models (O2.1).* This objective aims to increase the accuracy of chemistry transport models, by applying machine learning methodologies. This objective will be achieved if the precision of simulations in terms of correlations

with ground measurements increase compared to the state of the art, keeping a good performance against reference measurements.

- *Improve the application of chemistry transport models to air quality zoning problems (O2.2)* . This objective aims to perform air quality zoning over a certain region using chemistry transport models, hyperlocal sensors and machine learning models. This objective will be achieved if automatic coherent zoning is achieved, and it is similar to the European Air Quality Portal proposed.
- *Use chemistry transport models to assess personal exposure and health impacts to citizens (O2.3).* This objective aims to assess personal exposure and health impacts and identify risky activities using air quality models. This objective will be achieved if a methodology for exposure assessment can be defined taken as input air quality gridded data, and sequential activity data.

Chapter 3

Overall Methodology

This thesis explores solutions to improve signal in particle sensors to achieve a sustainability impact assessment framework. For this reason, the methodology followed uses a bottom-up methodology according to the architecture presented in Figure 1.1. This methodology is summarized in Figure 3.1, based on the following steps: i) first, the laboratory calibration and quality assurance of the IoT devices, including the design of a dryer system described in Chapter 4, corresponding to the sensing-IoT layer; ii) then, the second step focuses on the deployment of the edge layer and as described in Chapter 5, allowing data processing using machine learning models to make short-term particle concentration predictions, covering this step the transmission layer and part of the data management layer and; iii) the execution of complex air quality models to provide some services in the application layer, as the air quality zoning presented in Chapter 6 and the sequence oriented sensitive analysis for health assessment validated in Chapter 7.

However, this chapter will cover only the generic methodology that is transversal to this thesis. In other words, the focus is to describe the building blocks of the sustainability impact assessment that are present in the state of the art, and we decided to use as a starting point of our thesis. This includes the following points: i) the IoT devices used for air quality monitoring as well as the laboratory calibration procedure before the installation in the final location; ii) the CHIMERE-WRF chemistry transport model used as engine for feeding the specific services used in this thesis; and iii) the IoT-Edge-Cloud architecture used to support the framework developed. Also, this section describe the pollutants to be measured and the kind to collected, as well as the reason to choose this tools compared with other alternatives in the state of the art.

In contrast, the followed for the specific innovations proposed in this thesis as the development of the dryer system, the embedded machine learning models as well as the air quality zoning service and the sequence oriented sensitive analysis methodology are described on the respective chapters of this thesis.

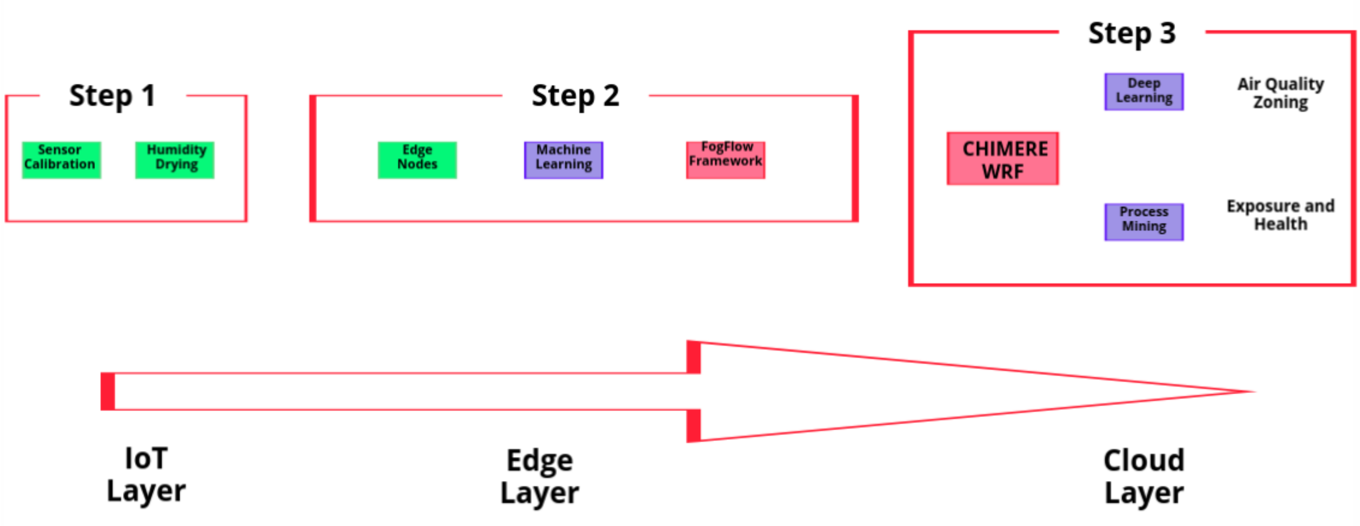

Figure 3.1: Thesis methodological framework

3.1 — Smart Spot Sensor System

IoT Smart Spot devices enable air quality monitoring by integrating gas sensors, categorised by the World Health Organization (WHO) as harmful gases for humans in certain concentrations, such as H_2S, NH_3, NO, NO_2, SO_2, CO and O_3. In addition, they can measure other parameters, such as particulate matter for PM1, PM2.5 and PM10, ambient noise, and meteorological parameters. These devices have certified emission and inmission monitoring validity due to the wide ranges allowed by the integrated sensor technology from the supplier Alphasense. Figure 3.2 presents the main characteristics of the Smart Spot Air Quality System.

The Smart Spot can measure different gases with a concentration between 0 and 10.000 ppb allowed. It will also be able to provide the information in $\mu g/m^3$ autonomously and in real-time to match the unit used in official air quality regulations. In addition to gases, the system incorporates sensors for dust or particulate matter, temperature, humidity and atmospheric pressure. The whole system is optimised to work between the maximum values established by the different regulations and between the typical values of the gases present to obtain greater precision in this range. This deployment allows real-time and accurate environmental pollution monitoring, making it an ideal solution for Smart Cities and other applications. Table 3.1 presents the main technical features of different pollutants regarding range, accuracy, resolution and interfering species.

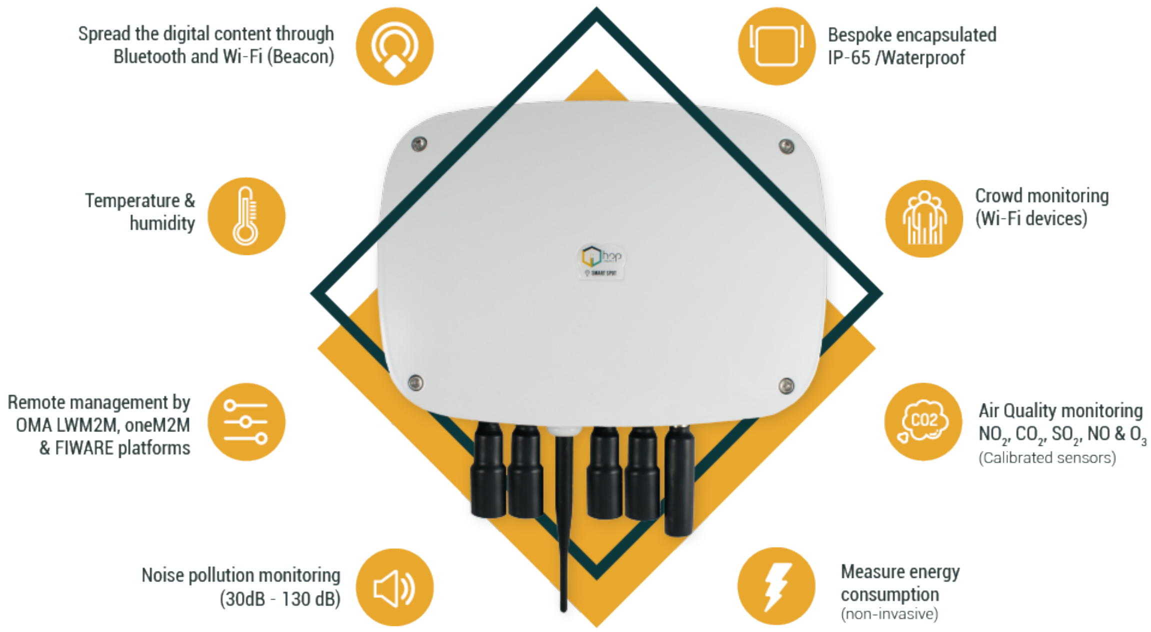

Figure 3.2: Smart Spot Air Quality System Overview

Pollutant	Range	Accuracy	Resolution	Cross Sensitivity
NO_2	0-5 pmm	± 2 ppb	1 ppb	$Cl > H_2S > NO > SO_2 > CO$
NO	0-5 pmm	± 10 ppb	1 ppb	$H_2S > NO > CO > Cl$
O_3	0-5 pmm	± 3 ppb	1 ppb	$Cl > H_2S > NO > SO_2 > CO$
SO_2	0-5 pmm	± 3 ppb	1 ppb	$NO > Cl > H_2S$
CO	0-5 pmm	± 10 ppb	1 ppb	$H_2S > NO$
PM10	0-2 mg/m^3	90 %	0.1 ug/m^3	Humidity
PM2.5	0-2 mg/m^3	90 %	0.1 ug/m^3	Humidity
PM1	0-5 mg/m^3	90 %	0.1 ug/m^3	Humidity

The gas sensors used in this device are electrochemical and have ISO 9001:2015 certification, which regulates the verification process of all its manufactured units. To ensure the reliability of the measurements made with these sensors, calibration and assembly in the probe is carried out according to the manufacturer's technical specifications in an ENAC laboratory certified ISO 17025:2005. The final product complies with the European directive 2008/50/EC on air quality. As one of the great differential values of the product, all the gas measurement sensors used have a high precision calibration carried out with certified gas standards (Linde brand) in the laboratory.

Including all these sensors and capacities in a single device provides savings in installation, maintenance and management, and communications. Regarding the connection possibilities of this device, it is offered in multiple versions, including Wi-Fi, LoRa, GSM/GPRS and NB-IoT. On the other hand, they allow the use of communication protocols such as LwM2M, MQTT and Modbus TCP (industrial environments), allowing Smart Spot to be a FIWARE-Ready device. Smart Spots allow different types of power sources, high-capacity batteries and solar panel recharging. Thanks to this versatility, these devices can work in remote natural environments without the need for an electrical installation and operate in city environments where it is not possible to receive a continuous electrical connection.

3.2 — Laboratory Calibration

To guarantee the trustability and accuracy of the results, all laboratories worldwide should demonstrate traceability of the results with the standards and methodologies considered as references. The standards founding the prerequisites for the expertise of laboratories and additional conformity assessment bodies (CABs) are the ones employed by accreditation bodies to assess CABs for their accreditation [233]. Nowadays, the greatest common standard for the expertise of testing and calibration laboratories is ISO/IEX 17025:2017. At the same time, additional documents give advice and helpful clarifications to laboratories and accreditation bodies at the international, national, and regional levels - as shown in Figure 3.3. Regarding ISO/IEC 17025:2017, *big data* is becoming a fundamental actor as it helps manage the huge amount of data a laboratory can generate [234].

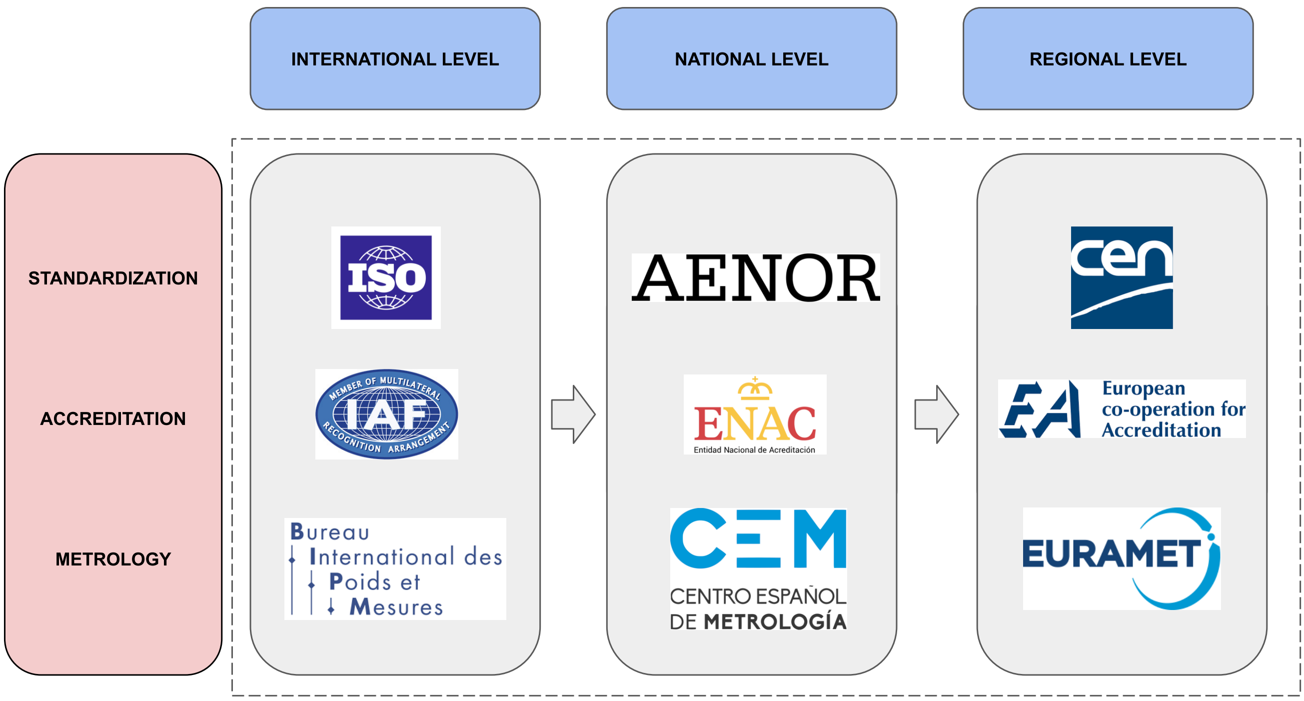

Figure 3.3: Accreditation and certification bodies at international, national (Spain) and regional (EU) levels

Table 3.1: Data quality requeriments from CEN/TS 17660

Pollutant	Range	Uncertainty	Standard / Procedure
SO_2	$25\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.041C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
NO	$25\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.041C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
NO_2	$25\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.055C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
NO_2	$25\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.041C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
CO	$0.4\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 30\ x\ 10^{-9}\ mol/mol$	$0.041C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
O_3	$20\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$2.6\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
O_3	$50\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.033C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
O_3	$500\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.067C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212
SH_2	$25\ x\ 10^{-9}\ mol/mol\ \leq\ C\ \leq 1500\ x\ 10^{-9}\ mol/mol$	$0.041C\ +\ 0.7\ x\ 10^{-9}\ mol/mol$	UNE-EN 14212

The IoT air quality systems employed in this thesis have been calibrated within the Libelium laboratory, holding certification of ISO 17025. It is imperative to underline that ISO 17025 certification demonstrates its validity upon a concrete and determined set of pollutant concentrations and specific chemical species. To ensure absolute clarity and precision, we have thoughtfully documented these specified concentration ranges, systematically organizing them in Table 3.1 for easy reference and comprehension. This table presents the aforementioned concentration ranges and meticulously delineates the associated uncertainties for each interval [233]. These uncertainties are characterized by a linear function dependent on each range's concentration (C). Furthermore, it is paramount to recognize that these calibrations adhere to the national standards prescribed by UNE-EN 14212 and ISO 17025, ensuring the highest level of traceability in the air quality measurements.

Regarding equipment, the analyzers used in the laboratory are provided by Teledyne, being the concrete models T300U for CO, T265 for O_3, T200U for $NO_x, NO - NO_2$ and T100Y for SO_2, SH_2. however, particle matter is not included in the laboratory's certification, but it has devices for calibrating particulate matter, such as the GRIMM 11-D.

3.3 — Field Tests

After laboratory calibration, the sensor system should be validated through field tests. We followed the methodology described in the standard CEN/TS 17660 to do this (Figure 3.3). According to this, the co-location tests should take place where sample air can be collected and performed according to the *Guide to demonstrating the equivalence of ambient air monitoring methods*. The sensor systems shall be installed in a control station by the manufacturer's requirements, with a minimum of 192h of concentrations higher than 50 % of the average of 98th of percentiles of hourly values in the last 5 years (or in all years available, for recent stations) required.

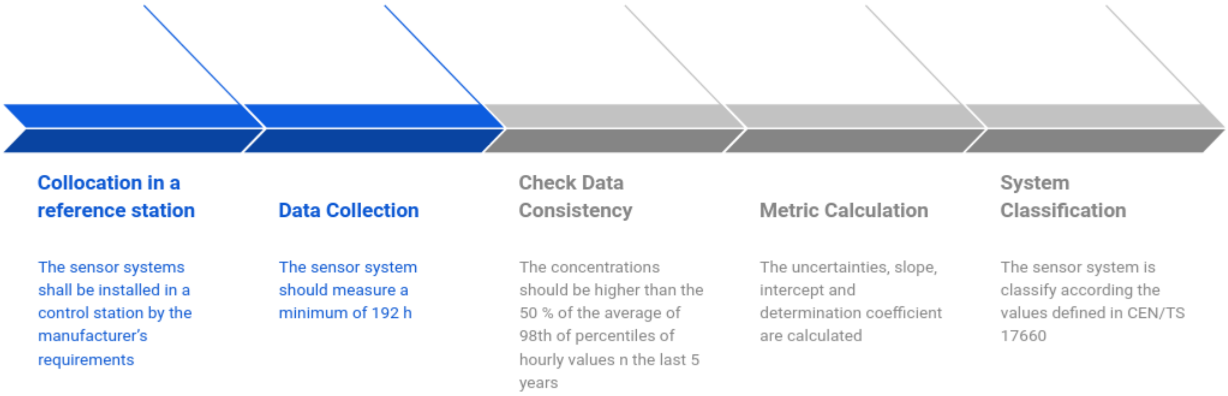

Figure 3.4: Field test steps according to the methodology defined in the standard CEN/TS 17660

- **Pearson Coefficient**: real coefficient, which can take values between -1 and 1, defined as the quotient of the covariance and the standard deviations. Values close to one mean a positive correlation, and close to -1 mean a negative correlation. Values close to 0 mean no dependencies between variables.

$$r = \frac{\sum_{i=1}^{n}(y_i - m_y)(\hat{y}_i - m_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - m_y)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - m_{\hat{y}})^2}} \tag{3.1}$$

- **Mean Absolute Error**: average value of the difference in absolute value between predictions and observations, where $\hat{y}_i$ is the prediction of the parameter under study, and y_i is the real value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.2}$$

- **Determination Coefficient**: the coefficient of determination measures how well a statistical model predicts an outcome. The lowest possible value is 0, and the highest is 1. The better a model makes predictions, the closer the value will be to 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - m_y)^2} \tag{3.3}$$

- **Uncertainty**: this parameter models the uncertainty between the sensor system and the reference station.

$$u(bs, s) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_s}(y_{i,j} - \bar{y}_i)^2}{n(n_s - 1)}} \tag{3.4}$$

3.4 IoT-Edge-Cloud Architecture

3.4.1. Cloud Layer

The cloud architecture is one of the main components of our solution, orchestrating and balancing the workload and processes between the different edge nodes. This component optimises the workload distribution, ensuring that each edge node is utilised effectively. In addition, by balancing the workload, the cloud node can prevent any single-edge node from becoming overloaded and potentially causing system instability. Last, the containers in the cloud are responsible for providing the graphical interface that allows users to configure and manage the workloads and resources of the network and for monitoring and troubleshooting issues that may arise.

The cloud architecture developed for this work is composed of the following services: i) a designer module that provides an easy-to-use web-based graphical user interface for designing and defining the tasks that will be executed on the platform; ii) a discovery container, which provides an API for resource discovery, allowing other components to discover and interact with available resources; iii) a master of the FogFlow platform, responsible for managing the distributed execution of tasks and services across multiple nodes, receiving task and service requests from various sources and ensuring that they are executed on the appropriate edge nodes or cloud resources; iv) an Orion LD context broker, which is an intermediary between cloud and edge services, routing requests from cloud services to the corresponding services in the edge; v) persistence components as CrateDB or Timestamp connected to the Context Broker through QuantumLeap; vi) a cloud worker responsible for executing tasks on the cloud, managing containerised services, and facilitating communication between cloud nodes and edge nodes; and vii) the RabbitMQ software [31], enabling communication between the edge nodes and edge cloud by acting as a middleman for exchanging messages between the different components of the platform.

3.4.2. Edge Layer

The edge node is responsible for receiving raw data from the Smart Spot, processing the data, and publishing the result without going to the cloud. This decentralises the data processing, increasing efficiency and allowing for data processing to be done as close to the source as possible, reducing the latency associated with sending data to the cloud. By processing the data locally, the edge node reduces the amount of data that must be sent to the cloud, resulting in faster processing and lower bandwidth usage. This approach benefits applications with critical low latency, such as real-time control systems.

The edge node is a complex system that typically consists of three services working together to process and manage data locally: i) The Eclipse MQTT broker acts as a mediator between the devices and the edge node, managing the topics and subscriptions, ensuring that messages are sent to the correct destinations with a high reliability and scalability [32]; ii) the FogFlow IoT worker, which receives data from the Eclipse MQTT broker, and then it sends this data to the Edge Worker for processing; iii) the edge worker, which acts as a data processing unit that takes the MQTT format and transforms it to NGSI-LD; iv) the FogFLow Operator that performs the embedded machine learning tasks; and v) connector with the cloud.

As the edge node, we used SECO solutions for edge computing [1], equipped with powerful processors, sufficient memory, and various connectivity options to communicate with other devices and the central data center. The operating system and platform was provided by Red Hat Enterprise Linux [2], offering offer features like containerization, orchestration, and management tools necessary for deploying and managing applications at the edge efficiently.

3.4.3. Sensing Layer

Data processing is a key step in our solution, covering IoT data collection to knowledge generation through machine learning. The data exchange between IoT devices, processing routines in nodes and the Orion-LD context broker in the cloud layer is done through the NGSI-LD communication protocol, as it is the ideal method for facilitating data exchange between the different components of the architecture defined before [35], as it is shown in Figure 3. This choice was made because NGSI-LD provides a standardised data model for the Internet of Things (IoT) that enables the description and exchange of data in a structured and semantically rich format, being structured using a set of metadata that includes information about the data's source, quality, and context. This metadata is linked using a context IRI that points to a JSON-LD file with the corresponding vocabulary. This allows for more precise and accurate data analysis, which is critical for successful AI applications. Additionally, the NGSI-LD standard is highly scalable, accommodating large and complex datasets without sacrificing performance.

3.5 — Air Quality Modelling

The engine of the sustainability impact assessment platform is the CHIMERE-WRF, an open-source multi-scale chemistry-transport model over a range of spatial scales, from the hemispheric to the urban scale, with resolutions from 1km to several degrees. The model comprises several modules: i) the most important is the CHIMERE itself, which computes the chemistry concentrations using as

[1] https://edge.seco.com/
[2] https://www.redhat.com/en/technologies/linux-platforms/enterprise-linux/edge-computing

input the ii) emissions models and iii) the meteorological model. Lastly, iv) IoT sensors and satellite observations could be used to calibrate the model. We have selected this model for the thesis as it can cover a wide range of resolutions, until 1 km x 1 km, and it is widely used in the scientific literature. The key processes affecting the chemical concentrations in the atmosphere are represented in Figure 3.5, being emissions, transport (advection and mixing), chemistry and deposition [235].

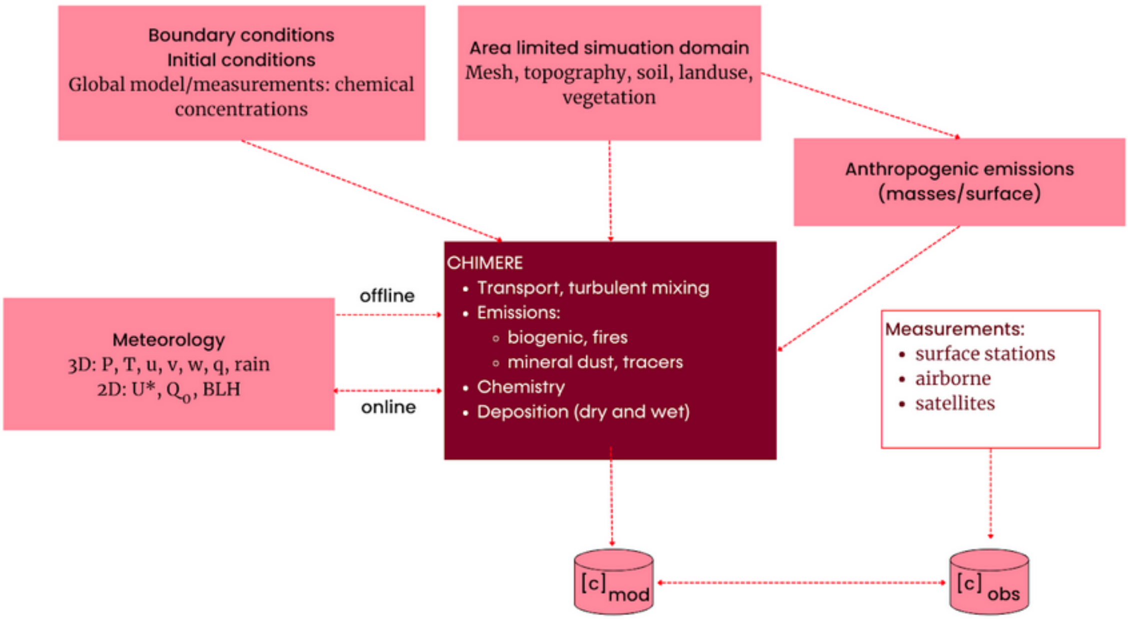

Figure 3.5: Different modules and connection of the CHIMERE-WRF chemistry transport model. Adapted from [237]

CHIMERE is an Eulerian chemistry transport model, meaning that the simulation domain is split into fixed bins called bins [156]. The numerical method used to estimate the temporal solution of the stiff system of partial differential equations is adapted from the second-order TWOSTEP algorithm for gas-phase chemistry only [236]. It is based on applying a Gauss–Seidel iteration scheme to the 2-step implicit backward differentiation (BDF2) formula,

$$c^{n+1} = \frac{4}{3}c^n - \frac{1}{3}c^{n-1} + \frac{2}{3}\Delta t R(c^{n+1}) \tag{3.5}$$

where c^n is the vector of chemical concentrations at time t_n, $R(c) = P(c) - L(c)$ being the difference between the production of c ($R(c)$) and the destruction of c ($L(c)$) and Δt being the time step between t_n and t_{n+1}. As seen in Figure 3.5, the production and destruction of a pollutant is governed by emissions, transport, chemical reactions and deposition. CHIMERE models the loss and production for each term and computes the total loss and production. Then, these last values are used in Equation 3.5, and the proposed system equations are solved. Thus, it is easy to note that the main strategy in CHIMERE is to model all the factors that affect concentrations as kinetic processes.

Table 3.2: CHIMERE-WRF performance in the literature

Simulation	Model	Obs	RMSE	Bias	r
O3 daily mean, all stations					
v2017r5	66.34	58.12	18.64	8.22	0.75
v2020-offline	62.28	58.12	17.23	4.16	0.75
O3 daily max, all stations					
v2017r5	86.56	81.53	18.88	5.03	0.76
v2020-offline	80.74	81.53	18.38	−0.79	0.76
NO2 daily mean, all stations					
v2017r5	8.45	10.14	6.43	−1.69	0.51
v2020-offline	6.58	10.14	6.67	−3.56	0.51
NO2 daily max, all stations					
v2017r5	14.88	19.62	13.34	−4.74	0.38
v2020-offline	11.84	19.62	14.15	−7.78	0.34
PM10 daily mean, all stations					
v2017r5	13.66	17.16	11.12	−3.50	0.41
v2020-offline	11.1	17.16	11.23	−6.06	0.63
PM10 daily max, all stations					
v2017r5	18.07	32.28	25.95	−14.21	0.33
v2020-offline	17.2	32.28	25.45	−15.08	0.5
PM2.5 daily mean, all stations					
v2017r5	11.75	11.01	7.87	0.75	0.5
v2020-offline	10.01	11.01	7.93	−1.00	0.66
PM2.5 daily max, all stations					
v2017r5	15.69	20.28	15.8	−4.59	0.45
v2020-offline	15.71	20.28	16.31	−4.58	0.57

As Figure 3.5 highlights, several inputs are needed to compute the product and loss terms. There are four inputs: i) the meteorology, modelled with Weather Research Forecast (WRF), fed by the Global Forecast System (GFS); ii) the emission fluxes, that can be computed by *emisurf*, a preprocessor included in CHIMERE interpolates emission inventories into a grid domain; iii) the boundary conditions, using MEGAN 30 s and iv) the domain with grid points. Last, the chemical reactions have been modelled using the MELCHIOR kinetic mechanism [237].

The reason to choose CHIMERE is the good correlation against reference measurements, as stated in Table 3.2, which presents simulation results for various air quality parameters, including O3, NO2, PM10, PM2.5, and AOD, measured in $\mu g/m^3$. It compares two models, "v2017r5" and "v2020-offline," by reporting the observed values (Obs), the Root Mean Square Error (RMSE), the Bias, and the correlation coefficient (R) between the observed and simulated data. In summary, for O3 daily mean and daily max, v2020-offline outperforms v2017r5 with lower RMSE and Bias values while maintaining a similar correlation coefficient. For NO2, PM10, and PM2.5, v2020-offline generally exhibits lower RMSE and Bias values than v2017r5. Finally, for AOD, both models show improved performance in terms of RMSE and Bias in v2020-offline compared to v2017r5.

The integration of hyperlocal Internet of Things (IoT) devices represents this thesis's significant and innovative aspect. This disruption is facilitated through the workflow implementation detailed in Figure 3.6, which encompasses several key stages. First, the initial phase involves collecting hyperlocal data from various sources, thus enabling a comprehensive and granular understanding of the environmental conditions at specific locations. This data acquisition process is the foundation for subsequent stages and plays a pivotal role in enhancing the precision and accuracy of the entire system. The next crucial step involves using this hyperlocal data to calibrate dispersion models. Calibration ensures that the simulation results align closely with real-world measurements, which is instrumental in improving the overall reliability and effectiveness of the system. This calibration process is integral to achieving accurate and relevant insights into the local air quality and environmental conditions.

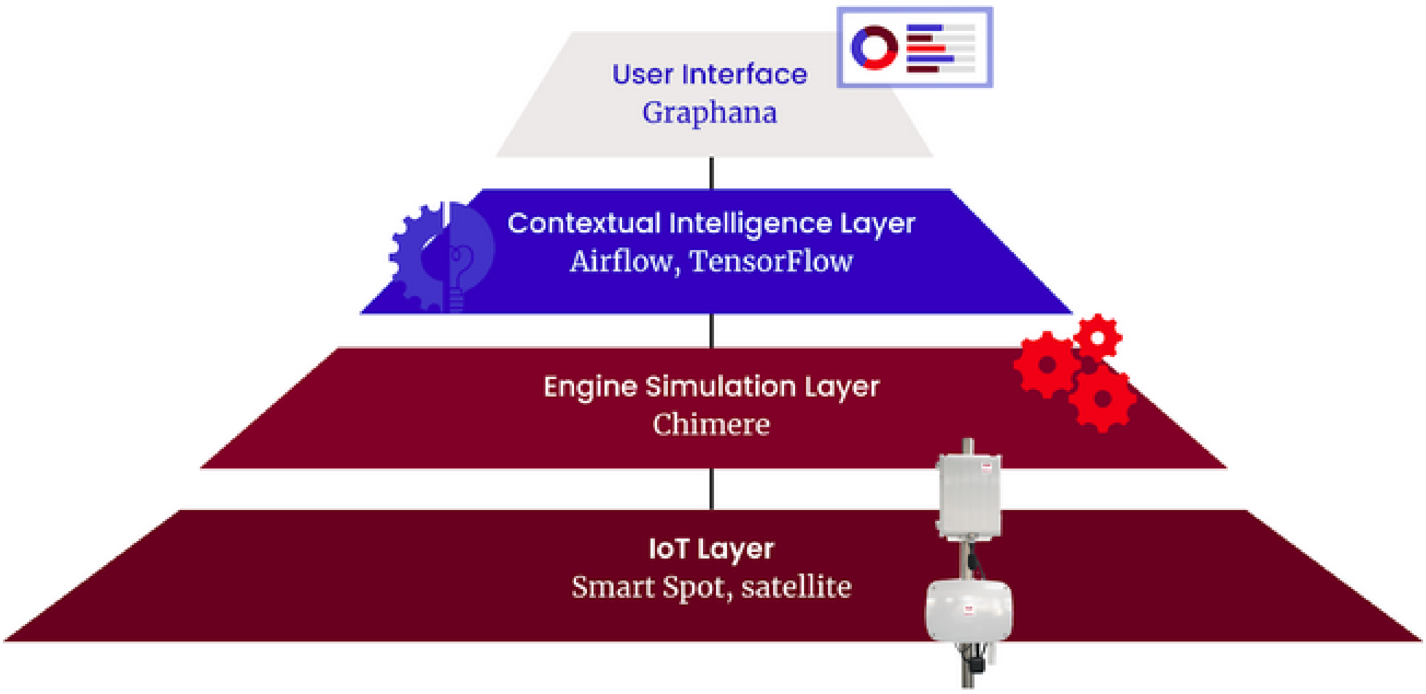

Figure 3.6: Workflow for data collecting and processing

Furthermore, the results obtained from the dispersion models serve as valuable input for the analytical layer, incorporating cutting-edge artificial intelligence techniques. This analytical layer is instrumental in extracting meaningful patterns, trends, and actionable insights from the vast and complex dataset. By harnessing the power of artificial intelligence, it becomes possible to uncover hidden correlations and make predictions that can substantially impact environmental decision-making. This thesis has provided some analytics focused on personal exposure assessment, risk prevention and air quality zoning in this context.

Finally, the integration of hyperlocal IoT data and the insights gained from the analytical layer culminate in the computation of Key Performance Indicators (KPIs). These KPIs act as vital metrics, providing a clear and concise means of evaluating and quantifying the effectiveness of environmental measures and interventions. Visualizing these KPIs through a user-friendly interface empowers stakeholders, researchers, and decision-makers to access and interpret the data efficiently, thus fostering a more informed and proactive approach to addressing environmental challenges.

Chapter 4

Design and Evaluation of a Drying System for IoT Hyperlocal Nanoparticles Monitoring Devices

Title
Design and evaluation of a dryer system for IoT hyperlocal particulate matter monitoring device
Authors
<u>Eduardo Illueca Fernández</u>, Iris Cuevas Martínez, Jesualdo Tomás Fernández Breis and Antonio Jesús Jara Valera
Details
<u>Journal</u>: IEEE Sensors Journal, <u>Publisher</u>: IEEE, <u>Volume</u>: 24, <u>Number</u>: 7, <u>Pages</u>: 11152 - 11165, <u>Year</u>: 2024, <u>JIF</u>: 4.3, <u>Rank</u>: Q1, <u>Status</u>: Published, <u>DOI</u>: 10.1109/JSEN.2024.3364537
Abstract
Particulate matter (PM) monitoring and climate change mitigation actions have been promoted due to the Paris agreement because of their impact on health and high mortality rate. High-resolution networks based on hyperlocal Internet of Things (IoT) sensors can play a fundamental role in improving data quality. In this context, hyperlocal refers to air quality IoT systems that allow collecting data in real time and in the cheapest way in comparison with local reference stations. Despite these methods are powerful and widely used by the scientific community, the signal is highly affected by relative humidity (RH). In this article, we present a system for measuring nanoparticles based on drying the air sampled and avoiding the hygroscopic growing of the particles. To the best of our knowledge, this is the first dryer system approach developed for IoT hyperlocal sensors. In addition, the relevance of our solution is supported by the following points: 1) we propose a new dryer system that has been patented; 2) our solution can be integrated into an IoT infrastructure that allows it to interact with other services; and 3) our solution has been validated in a real scenario in the city of Madrid. We have observed that the integration of a dryer system improves the performance of the OPC-N3 sensor and that we can measure the PM10 and PM2.5 fractions with high precision, R2=0.83 . In addition, our solution can measure small particles, such as PM1, with a good correlation against the reference air quality stations. Thus, our work contributes by improving high-spatial-resolution nanoparticle monitoring in correlation to official measurements to mitigate climate change.

Chapter 5

Embedded machine learning of IoT streams to promote early detection of unsafe environments

Title
Embedded machine learning of IoT streams to promote early detection of unsafe environments
Authors
Eduardo Illueca Fernández, Antonio Jesús Jara and Jesualdo Tomás Fernández Breis
Details
Journal: Internet of Things, Publisher: Elsevier, Volume: 25, Number: -, Pages: 101128, Year: 2024, JIF: 5.9, Rank: Q1, Status: Published, DOI: 10.1016/j.iot.2024.101128
Abstract
Indoor particulate matter (PM) are small solid and liquid particles present in the air, and its monitoring is one of the key challenges regarding workplace safety because of its impact on human health. To address this issue, the Internet of Things (IoT) paradigm allows the implementation of hyperlocal monitoring systems, typically using traditional cloud architectures, which can be enhanced using edge computing architectures. For this reason, we propose an IoT-Edge-Cloud architecture for a platform which promotes the early detection of unsafe environments through machine learning, composed of a sensing layer that collects all the data, an edge layer that performs the artificial intelligence tasks and a cloud layer orchestrating. This architecture is based on the FogFlow framework and the FIWARE components. Our solution proposes an embedded model that can predict the occurrence of PM values higher than the recommended ones—according to the Occupational Safety and Health Administration (OSHA) indicators—with an 87 % of accuracy and a reduction of latency of 26 %. Our platform is innovative because it is based on the FogFlow framework for edge computing and supported by the Smart Spot device validated in a field test. This step is missing from similar state-of-the-art platforms. Thus, we believe that this work contributes to demonstrating the usefulness of AIoT to monitor workplace safety and make trustable predictions, avoiding risky environments.

Chapter 6

Improving Air Quality Zoning through Deep Learning and Hyperlocal Measurements.

Title
Improving air quality zoning through deep learning and hyperlocal measurements
Authors
Eduardo Illueca Fernández, Antonio Jesús Jara Valera and Jesualdo Tomás Fernández Breis
Details
Journal: IEEE Access, Publisher: IEEE, Volume: 12, Number: -, Pages: 38700 - 38716, Year: 2024, JIF: 3.9, Rank: Q2, Status: Published, DOI: 10.1109/ACCESS.2024.3374208
Abstract
According to the Air Quality Directive 2008/50/EC, air quality zoning divides a territory into air quality zones where pollution and citizen exposure are similar and can be monitored using similar strategies. However, there is no standardized computational methodology to solve this problem, and only a few experiences in the Comunidad of Madrid based on chemistry transport models. In this study, we propose a methodological improvement based on the application of deep learning. Our method uses the CHIMERE-WRF air quality modelling system and adds a step that uses neural networks architectures to calibrate the simulations. We have validated our method in the Region of Murcia. The results obtained are promising given the values of the Pearson coefficient, obtaining r=0.94 for NO_2 and r=0.95 for O_3 , improving 86 % and 29 % the performances reported in the state of the art. In addition, the cluster score improves after applying neural networks, demonstrating that neural networks improve the consistency of clusters compared to the current air quality zoning. This opened new research opportunities based on the use of neural networks for dimension reduction in spatial clustering problems, and we were able to provide recommendations for a new measurement point in the Region of Murcia Air Quality Network.

Chapter 7

Sequence-Oriented Sensitive Analysis for PM2.5 Exposure and Risk Assessment using Interactive Process Mining.

Title
Sequence-oriented sensitive analysis for PM2.5 exposure and risk assessment using interactive process mining
Authors
Eduardo Illueca Fernández, Carlos Fernández Llatas, Antonio Jesús Jara Valera, Jesualdo Tomás Fernández Breis and Fernando Seoane Martínez
Details
Journal: PLoS ONE, Publisher: PLoS, Volume: -, Number: -, Pages: e0290372, Year: 2023, JIF: 3.7, Rank: Q2, Status: Published, DOI: 10.1371/journal.pone.0290372
Abstract
The World Health Organization has estimated that air pollution will be one of the most significant challenges related to the environment in the following years, and air quality monitoring and climate change mitigation actions have been promoted due to the Paris Agreement because of their impact on mortality risk. Thus, generating a methodology that supports experts in making decisions based on exposure data, identifying exposure-related activities, and proposing mitigation scenarios is essential. In this context, the emergence of Interactive Process Mining—a discipline that has progressed in the last years in healthcare—could help to develop a methodology based on human knowledge. For this reason, we propose a new methodology for a sequence-oriented sensitive analysis to identify the best activities and parameters to offer a mitigation policy. This methodology is innovative in the following points: i) we present in this paper the first application of Interactive Process Mining pollution personal exposure mitigation; ii) our solution reduces the computation cost and time of the traditional sensitive analysis; iii) the methodology is human-oriented in the sense that the process should be done with the environmental expert; and iv) our solution has been tested with synthetic data to explore the viability before the move to physical exposure measurements, taking the city of Valencia as the use case, and overcoming the difficulty of performing exposure measurements. This dataset has been generated with a model that considers the city of Valencia's demographic and epidemiological statistics. We have demonstrated that the assessments done using sequence-oriented sensitive analysis can identify target activities. The proposed scenarios can improve the initial KPIs—in the best scenario; we reduce the population exposure by 18 % and the relative risk by 12 %. Consequently, our proposal could be used with real data in future steps, becoming an innovative point for air pollution mitigation and environmental improvement.

Chapter 8

Discussion and Conclusions

One of the key aspects within the digital transformation, and in the challenges faced by today's society such as climate change, is the monitoring of air quality. The World Health Organization has estimated that air pollution will be one of the most significant challenges related to the environment in the following years, and air quality monitoring and climate change mitigation actions have been promoted due to the Paris Agreement because of their impact on mortality risk. The particular importance is the effect of aerosol and particle matter (PM). In 2019, approximately 74 % of the urban population was still exposed to PM2.5 concentrations exceeding the World Health Organization (WHO). Thus, generating methodologies and tools that supports experts in making decisions based on sustainability impact assessment, monitoring actions, identifying exposure-related activities, and proposing mitigation scenarios is essential. For this reason, the development of solutions for air quality monitoring is a heat research topic in the context of Internet of Things and Smart Cities.

8.1 — Synthesis of the Thesis

This thesis has investigated solutions for sustainability impact assessment of mitigation policies, by following a bottom-up methodology. The first step was to perform a benchmarking study of the hyperlocal air quality sensors in the market according to the standard CENT/TS 17660, that was later validated in Madrid, by using machine learning techniques that allow an improvement in the analysis and correction of the signal for PM and toxic gases, as well as their evaluation and standardisation with respect to the reference methods. Finally, once the quality of the measurements has been assessed, it was possible to use this hyperlocal data to feed chemistry transport model that computes the dispersion of pollutants across the space. In this work, the CHIMERE model has been improved by using reference stations and IoT devices. The last step is to manage all this data by applying FAIR guidelines and to transform the CHIMERE simulation datasets into a knowledge graph by using an ontology driven approach – we have reused vocabularies from FIWARE Smart Data Models and CHEBI ontology - to improve the interoperability of data and results, and make it more trustable for end users and policy makers.

The proposed solutions act at the different layers in an IoT architecture, implementing innovation in both hardware and software sides. By acting at different levels, it was possible to apply different use cases goals according to the needs of different stakeholder and policy makers.

For the sensing layer, IoT devices integrates a wide range of hyperlocal pollutant sensors. In the case of particles these sensors are based on light scattering technology, and one of the main limitations of optical particle counters (OPC) is the error related to high humidity levels because of the hygroscopic growth of the particles. For this reason, we have proposed in Chapter 4 a new dryer system for IoT sensors, which has been validated using co-location tests. The system has been tested in the city of Madrid, in concrete in three measurement points inside the low emissions zone area, in the *Center* district. We demonstrated in this Chapter that our system can remove the humidity from the air sampling, improving the performance of the system.

Next, Chapter 5 provides insights in the context of architecture. The main challenges in traditional IoT-Cloud architectures are latency and processing times, not allowing for a real time processing of the information, required for most of the use cases in Smart Cities and IoT applications. In consequence, we have proposed an IoT-Edge-Cloud architecture for promoting safe workplaces, demonstrating that the edge approach reduce latency in processing information and allows to emedd machine learning models using interfaces as FogFLOW.

On the top of the architecture proposed for sustainability impact assessment are placed artificial intelligence services. In this sense, one of the solutions that this thesis has developed is an air quality zoning algorithm based on deep learning and clustering. One of the key step of this algorithm is the calibration through hyperlocal measurements, making a clear connection between the sensing layer. In Chapter 6 we proposed the CHIMERE-WRF chemistry transport model as pollutant dispersion engine, followed by a calibration using a ResNET architecture based on hyperlocal sensors, finishing with a clustering to classify the domain in different air quality zones. This study has been done in the Region of Murcia, proposing a new measurement point in the north of the region.

Last, other service implemented during this thesis is the personal exposure assessment, starting from gridded air quality data from a model, which in this case was bi-linear interpolation. We developed a new methodology based on interactive process mining that for each activity compute different KPI related to exposure and health, and select the target activities in which policy makers can act. This methodology has been validated using a synthetic population, and it has been observed that the measurements taken improve the KPIs in all proposed scenarions, becoming a promising technology that should be test with real data.

8.2 — Hypothesis and Objective Compliance

8.2.1. Achievement of Objectives

The main objective of this thesis has been addressed, as new devices, methodologies and models has been developed inside the proposed sustainability impact assessment framework. Next, we discuss the achievement of the objectives defined in Chapter 2.

- **Improve the quality and reliability of hyperlocal air quality systems (O1).**
  - *Improve data measurement by implementing humidity correction systems (O1.1).* This objective has been fulfilled by the development of a new IoT device that includes a dryer system. We have measured humidity and outside the device, and it is observed that the humidity decrease. However, there is still a strong correlation between internal and external humidity, showing that the device is still affected by humidity changes, being important to continue researching on this topic.
  - *Implement architectures based on edge computing to implement short-term forecasting services (O1.2).* This objective has been completed by deploying the IoT-Edge-Cloud architecture in a real working environment - Libelium's laboratory - and demonstrating that the system has an 87 % of accuracy.
- **Integrate chemistry transport models in IoT platforms and improve its performance by using hyperlocal ground measurements (O2).**

- *Improve performance of chemistry transport models (O2.1).* The CHIMERE performance improves after applying neural network calibration, obtaining better performance for all pollutants, increasing all the values of r and R^2 and improving the performance reported by CHIMERE [161]. This goal is specially achieved for NO_2 and O_3, by improving correlation in a 25 % and a 73 %. The same occurs with mean absolute error, reducing the bias of CHIMERE due to the uncertainities in the emission inventories.
- *Improve the application of chemistry transport models to air quality zoning problems (O2.2).* In this thesis, an automatic air quality zoning methodology has been extended and improved by proposing ANN as an additional step to calibrate air quality simulations, as the zoning proposed for the Region of Murcia is more coherent that the proposed by the current one according to the regulation criteria.
- *Use chemistry transport models to assess personal exposure and health impacts to citizens (O2.3).* The sequence-oriented methodology proposed in this thesis part from air quality models, and compute exposures by multiplying the air quality in a geographical feature by the time spent by an individual in this location. Moreover, our methodology goes beyond and allows to compute relative risk and identify target activities for policy makers.

- **Implement a knowledge layer for FAIRification based on semantic web technologies (O3).**

  - *Define an ontology that describes the air quality domain (O3.1).* A new ontology has been created in this thesis. The core of ontology is the concept of Simulation, which refers to the result of a running of CHIMERE. Each simulation is composed of a set of instances of Air Quality Forecast, which is the part of the simulation linked to a geospatial unit and a timestamp, and it is possible to infer an Air Quality Index. Each air quality instance can have several Concentration instances, that is linked to an unique Chemical Entity. The ontology is published in the following link [1]
  - *Transform netCDF data into RDF triples (O3.2).* The generation of RDF triples has been performed for short term simulations, using the SWIT framework, with a linear time cost. This transformation goes beyond and allows to store the triples and a GraphDB database providing an SPARQL endpoint. We have achieved not only store the air quality simulations in a GraphDB database, but also access the data in a proper way. This is an alternative to the traditional relational and time series approach, which present limitations regarding the flexibility to represent complex relationships in air quality data because of the underlying complex schemas.

8.2.2. Hypothesis Validation

The overall hypothesis of this doctoral thesis was that by following a bottom-up methodology and integrating the different components proposed in Figure 1.1 it will be possible to implement a sustainability impact assessment platform for smart cities. In general, we consider that this hypothesis has been validated as this thesis has contributed to each of the mentioned layers providing valuables insights and applications in the sustainability impact assessment platform.

In concrete, this global hypothesis was divided into the following technical hypotheses, that has been analysed in a separated way, to understand the how each one of the specific contributions helped us to validate the hypothesis proposed.

[1] https://purl.org/chimere-ontology

- **The combination of hardware improvements, artificial intelligence algorithms and chemistry transport models will improve the accuracy of hyperlocal IoT sensors (H1).**

  We can conclude that this hypothesis has been validated, as the dryer system developed during this thesis improves the correlations of the particulate matter hyperlocal air quality sensors. In concrete, our system increase correlation of OPC-N3 compared to the study performed by the South Coast AQMD without humidity system. For PM10, we obtained R^2 ranging from 0.62 to 0.83 for hourly measurements, contrasting with $R^2 = 0.48 - 0.53$ for the single sensor. A similar pattern occurred for the PM2.5 fraction, achieving values ranging from 0.34 to 0.82 compared to $R^2 = 0.61 - 0.69$. In this case, we achieved a greater score in the best scenario but a lower score in the worst one. However, the PM2.5 results were more coherent with the 0.38-0.67 proposed for OPC-N2 [95]. Finally, $R^2 = 0.50$ for PM1, which was lower than the obtained in the validation done by the South Coast AQMD (0.78-0.82). It is important to notice that we could only validate PM1 against one station, so there is a lack of statistical evidence to compare with the proposed ranges. On the other hand, we achieved similar performances to those obtained in the state of the art for PM2.5 using machine learning approaches, with R2 ranging from 0.78 to 0.83, compared with 0.78 and 0.88 [238].

- **The use of an IoT-Edge-Cloud architecture will improve the performance of the system by reducing latency and bandwidth usage (H2).**

  This hypothesis has been demonstrated thanks to the effectiveness of the IoT-Edge-Cloud approach in enhancing performance within a real indoor workplace. It is necessary to emphasise that we improved state of the art approaches by reducing latency and bandwidth usage, using a standard FIWARE IoT-Cloud architecture as control, and comparing it with the solution proposed using the edge computing approach. The results obtained demonstrates that the edge approach is more efficient with a latency reduction of 26 % and an accuracy in the prediction of the 87 %, improving other proposals without edge computing in the melting sector [239].

- **It is possible to integrate chemistry transport models in IoT platforms to validate sustainability impact assessment applications in different scenarios (H3).**

  Regarding the air quality zoning methodology proposed in this thesis, it is easy to note that hyperlocal data plays a key role in this methodology, essential to guarantee the completeness of the data. Some of the results proposed in this paper are based on the data provided by public authorities. Thus, the proposed methodology should be validated with more complete data in other locations, especially concerning particles where the number of missing values is around 80 % for PM2.5. We have demonstrated that IoT devices are essential to improve chemistry transport models, as they improve their application to sustainability impact assessment use cases, such as air quality zoning or personal exposure assessment.

- **It is possible to model the air quality domain using ontologies and knowledge graphs, increasing the explainability of the models used in the platform (H4).**

  Taking into consideration the results presented, we can conclude that the hypothesis has been validated for short term air quality forecasting by using SWIT. A RDF file in turtle format has been obtained, and this can be automated for a simulation of 72 h. In addition, a scalability analysis has been performed, obtaining a linear complexity when individuals increase, which suggests that our approach could be extended to long term simulations as a future research line. This is coherent with the complexity analysis performed in the state of the art, which assesses that in the best scenario (without collisions between individuals) the complexity of SWIT is linear [240]. One of the reasons for the successful mapping is the use of a proper ontology. In this work, a new ontology has been defined to model a CHIMERE simulation, providing a new insight in the state of the art, and complaining with the objectives proposed for this thesis.

8.3 — Contributions

The main contribution of this thesis has been the study of artificial intelligence technologies to support the development of a framework for sustainability impact assessment in Smart Cities. We have demonstrated that this framework improves the limitations found in the state of the art in terms of accuracy, processing time, computational complexity and explainability. In concrete, the following contributions were achieved during this research:

- **An analysis of the CEN/TS 17660 standard regarding air quality systems**. This thesis has analysed the effect of intersensor variability on the uncertainty calculated using the new to obtain conclusions on the effect of this variability in the new methodology. For this purpose, a sensitivity analysis has been carried out, and the classification has been computed for each of the inputs. Our results show that CEN/TS 17660-1:2021 is a robust measure that can accurately assess the quality of the sensor systems and could determine their ability to meet data quality objectives.
- **Development of a new dryer systems for particulate matter IoT devices**. One of the main contribution of this thesis is the development of a new IoT device that integrates a dryer system based on two silica columns and an infrared heating system. We have observed that the integration of a dryer system improves the performance of the OPC-N3 sensor and that we can measure the PM10 and PM2.5 fractions with high precision, $R^2 = 0.83$, in comparison with other IoT sensors in the state of the art. In addition, the relative humidity is reduced inside the device compared to the ambient values provided by meteorological stations. Additionally, we have demonstrated that these devices can be integrated into an IoT architecture and provide real time particulate matter data, which allows to implement high dense sensor networks for sustainability impact assessment.
- **Implementation of an IoT-Edge-Cloud architecture for risk assessment in work environments**. During this research we have deployed an IoT-Edge-Cloud architecture for a platform which promotes the early detection of unsafe environments through machine learning, composed of a sensing layer that captures particulate matter data, an edge layer that encapsulates a machine learning model and a cloud layer orchestrating. This architecture is based on the FogFlow framework and the FIWARE components. We demonstrated that the proposed architecture can predict the occurrence of PM values higher than the recommended ones with an 87 % of accuracy and a reduction of latency of 26 %.
- **Improve CHIMERE-WRF accuracy thank to the calibration through deep learning models**. We have developed several deep learning calibration models for different pollutants, using hyperlocal ground measurements as reference. The performance of CHIMERE improves after applying ANN calibration, obtaining better performance for all pollutants, increasing all the values of r and R^2 and improving the performance reported by CHIMERE according to the validation performed in 2021 [161]. For O_3, a Pearson Coefficient of 0.94 is obtained, improving our performance by 32 % and official performance by 25 %. The same occurs for NO_2, improving our performance by 73 % and official performance by 63 % with an $r = 0.88$. In contrast, for PM10, the performance is 0.89 and for PM2.5, $r = 0.75$, but the elevated number of missing values indicates that it is necessary to be careful in interpreting these results. NO_2 and CO present good correlations -$r = 0.81$ and $r = 0.68$, respectively - but the number of missing values for CO is around 60 %. At the same time, for NO, it is possible to infer an improvement as the time series for training is complete. For SO_2, the performance is poor after ANN calibration.
- **A new air quality zoning proposal for the Region of Murcia, applying an automatic methodology**. We propose a new air quality zoning based on the clusters obtained after applying ANN, following the recommendation of functional air quality zoning and the European regulation. The main difference compared to the Ministry for the Ecological Transition and the Demographic Challenge (MITECO) zoning for the Region of Murcia is the

creation of a new zone in the north part of the region, with a new monitoring point. This is quite important because, at this moment, the ES1401 zone covers more than 50 % of the territory with a unique monitoring point, so it is impossible to define all the exposure in this zone. Thus, the new zoning is based on splitting this huge area into two different ones and proposing a new monitoring point in Cieza or Jumilla.

- **Definition of a new methodology for particulate matter exposure and health impact assessment based on interactive process mining**. This thesis has defined a new methodology for particulate matter personal exposure assessment in collaboration with Karolinska Institute in Sweden. This new methodology can identify target activities and tasks for exposure mitigation, being a valuable tool for policy makers. This sequence-oriented sensitive analysis presents several advantages concerning the classical methodologies. As an example, by studying the sequences it is possible to assess that acting on building infrastructure is effective because it is a common place that appears in all the sequences, and the same applies to studying any scenario. Thus, the interactive process mining workflow is more agile, explainable, and human-oriented, avoiding the black box effect. Indeed, this new methodology shows the target parameters, and it is only necessary to iterate on these few inputs.
- **A new ontology that describes the domain of chemistry transport models**. In this work, a new ontology has been defined to model a CHIMERE simulation, providing a new insight in the state of the art, and complaining with the second objective proposed for this paper. In addition, this ontology meets with the quality standard according to the OQuaRE metrics and the ROBOT consistency tests. This ontology has been also compared with others present in the state of the art, showing that the ontology is more similar to the chemical ontologies than the environmental ones. This is logical because this ontology is focused on describing chemical components in the atmosphere as well as mathematical processes in simulations. However, this ontology covers a domain that is not explored in the state of the art, and should be part of future research.

8.4 — Limitations in the Work Performed

This thesis has found different limitations in the technologies applied, which encompass challenges related to the hardware of the sensors, as the miniaturization drying systems and the detection limits of particulate matter sensors, as well as limitations regarding software scalability and the unavailability of datasets for training the artificial intelligence models. Understanding these limitations is essential for interpreting the findings and outlining potential avenues for future research in the field.

- **Miniaturisation of particulate matter sensors and drying systems**. The atmospheric science community frequently utilizes optical particle counters (OPCs) and Mie scattering theory for sizing individual aerosol particles. However, measurements of aerosols by small systems have many advantages, such as low cost, ease and cost of deployment, and ease of access to inaccessible areas such as those close to urban conurbations, as well as integrate it in wearable devices. The dimensions of the solution proposed in this thesis is not easy to miniaturise, being important to research on how to reduce the size of the different components or explore other technologies for particulate matter quantification.

- **Effect of particle shape and chemical composition in the measurements**. One of the limitations of the field test performed is that it is not possible to quantify the effect of chemical composition and particle size [48]. However, it is known that this physical parameters strongly affect the optical principle behind the technology of particulate matter sensors. Regarding chemical composition, it is possible to standardize urban environments, as particles in these environments present similar compositions [241]. Thus, to test this effect, comparing urban with rural locations is necessary. In contrast, it is more complicated to assess and correct the effect of particle size, and the literature suggests that this can be done by applying other technologies based on gravimetry [242].

- **Scalability of the IoT-Edge-Cloud architecture**. The proposed platform in this thesis presents limitations in terms of scalability regarding the replication of the platform, according to the following lines: i) decrease the dependency of edge nodes, allowing to replicate of the platform without increasing the number of nodes and allowing to connect more IoT devices; ii) test the latency times to identify bottlenecks in the architecture and demonstrate an improvement in comparison with cloud architecture; and iii) scaling the machine learning models to guarantee the applicability in different environments, and the replication of models in different nodes.

- **Spatial resolution of the simulations performed with CHIMERE-WRF**. The air quality zoning performed in this thesis starts from the simulations performed using the chemistry transport model CHIMERE-WRF. The spatial resolution used for these simulations was 10 x 10 km. However, the literature suggests to cover a major spatial resolution of 1 km x 1 km [231]. This resolution is possible with CHIMERE, but requires a high computational cost and infrastructures that were not available during this thesis. For an air quality zoning, a resolution of 10 km x 10 km is acceptable as we worked over a regional scale, but personal exposure assessment requires a higher resolution. To overcome this limitation, we used bi-linear interpolation as air quality model, which is less complex in computational terms, but it will be interesting to assess exposure using data from more complex models.

- **Unavailability of sequential datasets for highly exposed activities identification**. The main limitation of the validation of the sequence oriented methodology developed in this thesis is the use of synthetic data. It is true that this is enough to demonstrate the effectiveness of the methodology but it is necessary to test it in a real pilot. In this context, the use of wearable sensors can allow activity monitoring, creating a new cultural phenomenon in citizen science, whereby members of the general population voluntarily wear tracking devices that continuously log their data in exchange for potential improvements in the quality of life or physical performance. This paradigm can be applied to the real-time monitoring of exposure, and used for medical purposes [243]. Thus, the next step is to apply the proposed methodology to real data collected by citizens. In addition, the use of the architectures developed in this thesis allows to easily integrate all these processes and data.

8.5 — Future Works and Vision

Possible research paths to follow up this thesis include taking into consideration research in each one of the components. As we proposed an architecture based on an IoT-Edge-Cloud strategy, it is necessary to research on each one of the levels in the architecture proposed. In this sense, the development of more accurate particulate matter sensors by exploring different physical technologies combined with more complex distributed computing strategies at edge level seems to be one of the next steps. In addition to this, more research is need in the top layer to improve the resolution of chemistry transport models and implementing semantic mechanisms to make the architecture more compliant with FAIR guidelines.

8.5.1. Towards Nanoparticles Detection Devices

Nanoparticles are particles having a diameter lower than 100 nm. These particles are in the ultrafine size range, meaning that they can be deposited in the lower areas of the respiratory tract and can be translocated directly into the blood through the alveolar barrier [244]. The chemical composition of these particles is complex, including metallic traces, organic compounds, carbonaceous species or biological agents as virus [245]. Therefore, it is necessary to know their properties, impact on the environment, effect on human health, methods of protection against them, and to provide sensing methodologies to monitor and detect their presence in air [246]. Currently, there are two conceptual approaches based on different physical principles to reach nanoparticles scale.

The first approach are based on suspended microchannel resonators, an emerging technology that allows a powerful characterization tool for nano-scale particles [247]. This methodology allows for measuring the mass and concentration of particles suspended in fluid in a flow-through mode [248]. The buoyant mass is defined as $m_b = V \times (\rho_o - \rho_l)$, where V is the volume of the object, ρ_o is the density of the object and ρ_l is the density of the liquid in which the object is immersed [249]. Here, the main challenge is to apply this principle to particulate matter detection in air. In these sense, there are few theoretical studies in the state of the art that the demonstrate that solidly mounted resonator can detect PM in air improving detection limit to 50 ng thanks to temperature modulation [250].

The second approach is based on thermophoresis, where particles move thanks to a transport force that occurs due to a temperature gradient. In consequence, it is easy to note that particles with larger sizes travel a shorter path, implying that it is possible to discriminate particles in function of the covered distance [251]. This is shown in Figure 8.1, and different electrodes can be used to detect different size particles. The signal transmission is done by impedance variations due to particle/electric field interactions occurring while a single particle precipitates towards an electrode surface, increasing its capacitance, being the concentrations/mass of this particle proportional to the impedance [252]. This concept has been applied in a prototype which employs thermophoresis and capacitance measurement for particulate matter separation and detection, allowing to identify airbone pathogens [253]. The strength of this approach is that can separate also particles from differents compositions.

However, the state of these technologies is not mature enough to be integrated in real IoT infrastructures. These approaches should deal with different challenges that are not explored yet in the state in the art, and are a subject of future research in the field. The first one is the miniaturization of the technology, a requirement for any IoT particulate matter sensor. The second one is the lack of validation of these methodologies with the official ones. To the best of my knowledge, there are not comparisons between termophoresis and resonators with manual gravimetry, because of the premature state of the technology. In this sense, a validation in terms of accuracy will be required for demonstrate usability in IoT infrastructures.

8.5.2. Towards Hyperresolution Air Quality Dispersion Models

Chemistry transport models have limitations in terms of resolution, primarily due to the trade-off between achieving more accurate and detailed information and incurring higher computational costs associated with the Courant-Friedrichs-Lewy condition. In this context, it is impractical to attain a resolution higher than 1 km x 1 km. As explained in section 6, the simulations performed for air quality zoning had a resolution of 10 km x 10 km, but lower resolutions are required for more accurate zoning or use of models in other applications. Additionally, uncertainties

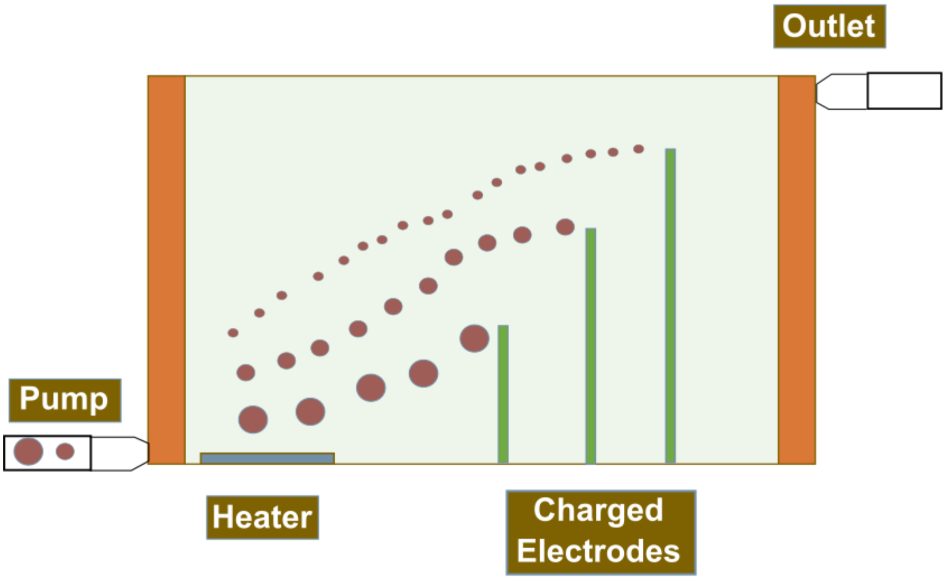

Figure 8.1: Nanoparticle theoretical sensor design and model illustrating size discrimination via thermophoresis. Adapted from [254]

in initial conditions, input variables, and parameterizations introduce biases that restrict their utility for certain tasks [254]. Last, another limitation of CTMs is excessive numerical diffusion, one of the major limitations in the representation of long-range transport by chemistry transport models [255].

To overcome this limitation, the most explored solution in the state of the art to deal with the Courant-Friedrichs-Lewy condition is to model pollution at a street level [30]. Different types of models may be used to represent the pollution in street canyons. The first approach is based on computational fluid dynamics (CFD) models [179], which describe the urban geometry, the air flow, and the pollutant concentrations in the streets. The main limitation of this strategy is the computational costs, requesting log time simulations in a city scale with a large street network. Some recurring CFD models in the literature are OpenFOAM [256], STAR-CCM+ [257], and the PALM model [258].

On the other hand, parametric models are more suitable for smart cities purposes because of their low computational cost, and are combined with statistical assumptions. Some parametric models, e.g. Polyphemus [259] and CALINE4 [260], are based on the use of a Gaussian dispersion methodology to represent emitted traffic-related pollutants, such as a Gaussian plume or puff. Other parametric models use parameterisations based on CFD modeling or wind-tunnel experiments to describe the flow in each street and the exchange from street to street and between streets and the overlying atmosphere. The transport of pollutants from one street to another is taken into account through intersections, e.g. SIRANE [261] and the Model of Urban Network of Intersecting Canyons and Highways MUNICH [262]. The flow above the street network is represented by a Gaussian dispersion methodology (SIRANE) or by one- or two-way nesting in a regional model (MUNICH).

8.5.3. Towards Knowledge Graph Representation of Chemistry Transport Models Data

To conduct reliable data-driven research, it is crucial to utilize large-scale data and shift away from isolated, closed methods of investigation towards a more interconnected scholarly approach [263]. In this context, a group of internationally recognised experts in data science and bioinformatics co-authored the FAIR Principles — a set of recommendations to ensure that data resources and scientific software are *Findable*, *Accessible*, *Interoperable*, and *Reusable* [23]. FAIR compliance is a milestone in enabling new prediction analytics to access data for forecasting computation, understanding the semantic context providing provenance of the information, and assessing the rules for data transformation [264]. Chemistry Transport Models (CTMs) are one of the computational scientific models for air quality modelling and were explored in this thesis. Thus, future research steps should work on the FAIRification of these mathematical models [265].

However, most part of the data generated by CTMs are not FAIR, delaying the use of interoperable data in air quality systems [266]. In concrete, CHIMERE works with the netCDF, a scientific format for representing multidimensional arrays in low-level, to allow the fast processing using programming languages such as C, C++ or FORTRAN [267]. This implies that the information provided does not meet with FAIR guidelines, and makes it difficult to guarantee the re-usability, meaning a lack of research in this field. For this reason, this thesis has developed the CHIMERE ontology which describes the information present in a netCDF output file using terms from ChEBI and the FIWARE Smart Data Models. Future research will aim to generate knowledge graphs representing the outputs of a chemistry transport model simulation keeping the relationships between pollutants and entities. Our hypothesis states that it is possible to transform netCDF files into RDF/XML knowledge graphs with semantic information as the CHIMERE ontologys by using an ontology-driven methodologies [268]. In addition, the knowledge graphs generated will be stored in an RDF database, from where stakeholders can extract relevant information through the SPARQL endpoint.

8.5.4. Towards Distributed Algorithms in IoT Sensor Networks

The development and research on IoT devices enable the proliferation of new services and applications, in which *edge computing* encompasses a highly complex scenario with a huge amount of different devices that must cooperate with each other. However, this approach requires effective orchestration mechanisms to guarantee the smooth performance of applications and services [269]. However, mechanisms typically applied to the *cloud* can not directly be migrated to the *edge* given its particular characteristics. These problems highlight the importance of the design and development of new orchestration mechanisms for the *edge*, particularly to handle the demands of low latency.

In this line, this thesis has demonstrated that using *edge computing* reduces the latency of machine learning services, but it is important to address more complex orchestrating mechanism for addressing more complex use cases. The first approach proposed in the literature is service migration according final user movement, which can be predicted using machine learning methodologies [270]. Another approach is exploring machine learning techniques with alternative heuristics for the service placement and priority assignment. This requires to have access to datasets that contain enough information for the learning process. It is important to note that in this case the machine learning model will be placed in the orchestrator side and would be in charge off coordinating other machine learning models that performs different tasks [271].

8.5.5. Towards Individual Personal Exposure Identification

By measuring and modeling pollutant exposure concentrations and by characterizing spatial and temporal sequential human activities, an exposure surveillance system that provides real-time exposure data can be developed based on health risk indicators [272]. The methodology proposed in this thesis allows to improve this surveillance by identifying target activities, but requires a lot of data regarding air quality, personal activities and individual geolocalization, which are sensible according to regulations and are difficult to collect [273].

In this context, citizen science approaches are emerging to engage citizens for collecting environmental data. This is specially relevant as a new type of devices, called wearable air quality sensors, which allow citizen scientists to collect exposure data from its daily activities [274]. Some projects as SOCIO-BEE (101037648) or GreenGage (101086530) are acting on promoting community engagement and social innovation to bridge this gap via wearable modules for air-quality observation, providing an affordable, accessible and open wearable with a low-cost ICT-based solution, combining trusted service modules for co-creation activities and micro-volunteering crowdsourcing tools [275].

For this reason, the next step in scientific research is to combine the interactive process mining methodology proposed in this thesis with data collected in different citizen sciences projects. By using wearable devices, it is possible to monitor exposure in different areas, and citizens can label the activity or task that they are performing in the moment of measurement. This methodology will allow to collect all the input data required for 7, as well as to create awareness among the most important stakeholders involved.

8.5.6. Towards Implementation of Sustainability Impact Assessment in Smart Cities

Regarding digital climate mitigation tools and sustainability impact assessment, this thesis has developed improvements on particle matter measurements and data analysis by combining hardware improvements and machine learning algorithms, allowing end users and public authorities to perform scenario simulations, KPIs computation, and decision making. Our proposal is disruptive in the state of the art because it is the first approach, to our knowledge, that combines IoT measurements with dispersion simulations of a wide combination of scenarios. In this thesis, we have combined our scientific contribution with a high user experience so that public authorities can efficiently use our tools.

However, it is necessary to add improvements to the prototypes solution by integrating new data sources to improve the quality of the models. The most important point is generating hyper local data by installing devices in the pilot cities contributing in the following aspects: i) it will improve the interpolation and dispersion models, as they act as air quality stations; ii) it will allow obtaining data with a higher temporal resolution, allowing adjusting the system to small changes, iii) it will allow validating the generated models more reliably and iv) the measurement of new variables such as new pollutants, noise, and crowd monitoring. The hardware improvements and algorithms developed during this thesis will help to increase the quality of the data and going to dense IoT networks thanks to the combination of the humidity correction systems and the distributed computing approaches.

In addition, integrating new data on parking - in real-time -, public transport, bikes, and all the information provided from the city level will make it possible to implement new functionalities in the solution, complementing the air quality and particulate matter data explored in this thesis. Some suggestions proposed for development in smart cities platforms that could be benefited from the solutions developed in this thesis are the following: i) a model that predicts the availability of bicycles and public transport in the city, ii) a system for calculating routes that minimize the impact on air quality, and iii) artificial intelligence models that predict the movement of people and predict high traffic events.

8.6 — Publications, Conferences and Projects

8.6.1. Publications Indexed in Journal Citation Report (JCR)

- Eduardo Illueca Fernández, Antonio Jesús Jara Valera , and Jesualdo Tomás Fernández Breis, "Improving Air quality zoning through deep learning and hyperlocal measurements," IEEE Access, 2024. Available: https://doi.org/10.1109/access.2024.3374208
- Eduardo Illueca Fernández, Iris Cuevas Martínez, Jesualdo Tomás Fernández Breis, and Antonio Jesús Jara Valera, "Design and evaluation of a dryer system for IoT hyperlocal particulate matter monitoring device," IEEE Sensors Journal, 2024. Available: https://doi.org/10.1109/JSEN.2024.3364537
- Eduardo Illueca Fernández, Antonio Jesús Jara Valera , and Jesualdo Tomás Fernández Breis, "Embedded machine learning of IoT streams to promote early detection of unsafe environments," Internet of Things, 2024. Available: https://doi.org/10.1016/j.iot.2024.101128
- Eduardo Illueca Fernández, Carlos Martinez Llatas, Antonio Jesús Jara Valera, Jesualdo Tomás Fernández Breis, and Fernando Seoane Martínez, "Sequence-oriented sensitive analysis for PM2.5 exposure and risk assessment using interactive process mining," PLoS ONE, 2023. Available: https://doi.org/10.1371/journal.pone.0290372

8.6.2. Articles Submitted to JCR Journals

- Noel Gomáriz Kühne, Eduardo Illueca Fernández, Youngseob Kim, Jesualdo Tomás Fernández Breis, Yelva Roustan and Antonio Jesús Jara Valera, "Towards Sustainability Impact Assessment: Air Quality Modelling for Low Emissions Zones and Energy Management in Smart Cities," Energy.

8.6.3. Publications at Conferences

- Eduardo Illueca Fernández, Nuria Bernabé Mulero, Alejandro Pujante Pérez, Jorge María Merino García, Iris Cuevas Martínez , and Antonio Jesús Jara Valera, "CEN/TS 17660 in air quality systems for data quality validation and certification over smart spot air quality systems," In International Conference on Ubiquitous Computing and Ambient Intelligence, 2022. Available: https://doi.org/10.1007/978-3-031-21333-5_65
- Eduardo Illueca Fernández, Noel Gomáriz Kühne, Nuria Bernabé Mulero, and Antonio Jesús Jara Valera, Advancing Sustainability Impact Assessment: A Comprehensive Tool for Low Emissions Zone Management," In 8th International Conference on Smart and Sustainable Technologies, 2023. Available: https://doi.org/10.23919/SpliTech58164.2023.10193202

8.6.4. Projects

The studies performed during this thesis have contributed to the project *Knowledge graph methods for process mining models and clinical guideline models in support of the Learning Health System (MINEGRAPH)*, led by Dr. Jesualdo Tomás Fernández Breis, funded through grant PID2020-113723RB-C22 funded by MCIN/AEI/10.13039/501100011033/

Bibliography

[1] T. Udrea and A. Bauer, “Between control and independence: computational modelling within ec’s trade sustainability impact assessments,” *Impact Assessment and Project Appraisal*, vol. 41, no. 1, pp. 21–34, 2023.

[2] I. A. T. Hashem, R. S. A. Usmani, M. S. Almutairi, A. O. Ibrahim, A. Zakari, F. Alotaibi, S. M. Alhashmi, and H. Chiroma, “Urban computing for sustainable smart cities: Recent advances, taxonomy, and open research challenges,” *Sustainability*, vol. 15, no. 5, p. 3916, 2023.

[3] A. Kaginalkar, S. Kumar, P. Gargava, and D. Niyogi, “Review of urban computing in air quality management as smart city service: An integrated iot, ai, and cloud technology perspective,” *URBAN CLIMATE*, vol. 39, SEP 2021.

[4] Y. Alshamaila, S. Papagiannidis, H. Alsawalqah, and I. Aljarah, “Effective use of smart cities in crisis cases: A systematic review of the literature,” *International Journal of Disaster Risk Reduction*, p. 103521, 2023.

[5] Z. Liu and J. Wu, “A review of the theory and practice of smart city construction in china,” *Sustainability*, vol. 15, no. 9, p. 7161, 2023.

[6] A. Caragliu, C. Del Bo, and P. Nijkamp, “Smart Cities in Europe,” *Journal of Urban Technology*, vol. 18, no. 2, pp. 65–82, 2011. [Online]. Available: ⟨GotoISI⟩://WOS:000299210100005

[7] B. Mohamadrezapour, S. Rafiepour, and M. Ali Pour, “A review of urban living labs experiences with emphasis on the implementation of environment-based smart cities,” *Science and Technology Policy Letters*, 2023.

[8] P. Berrone and J. Ricart, “New york edges out london as the world’s smartest city,” *IESE insight review*, 2016.

[9] P. Berrone, J. E. Ricart, A. Duch, and C. Carrasco, “Iese cities in motion index 2019,” *IESE, ST-509-E*, vol. 5, p. 2019, 2019.

[10] Y. Sun, H. Song, A. J. Jara, and R. Bie, “Internet of things and big data analytics for smart and connected communities,” *IEEE access*, vol. 4, pp. 766–773, 2016.

[11] B. N. Silva, M. Khan, and K. Han, “Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities,” *Sustainable cities and society*, vol. 38, pp. 697–713, 2018.

[12] A. Kaginalkar, S. Kumar, P. Gargava, N. Kharkar, and D. Niyogi, “Smartairq: A big data governance framework for urban air quality management in smart cities,” *FRONTIERS IN ENVIRONMENTAL SCIENCE*, vol. 10, APR 8 2022.

[13] A. Tanda and A. De Marco, “A review of an urban living lab initiative,” *Review of Policy Research*, vol. 38, no. 3, pp. 370–390, 2021.

[14] S. Conti, C. Fornari, P. Ferrara, I. C. Antonazzo, F. Madotto, E. Traini, M. Levi, A. Cernigliaro, B. Armocida, N. L. Bragazzi *et al.*, “Time-trends in air pollution impact on health in italy, 1990–2019: An analysis from the global burden of disease study 2019,” *International Journal of Public Health*, vol. 68, p. 1605959, 2023.

[15] M. Hadei and K. Naddafi, “Cardiovascular effects of airborne particulate matter: A review of rodent model studies,” *Chemosphere*, vol. 242, p. UNSP 125204, Mar. 2020.

[16] T. F. Mar, T. V. Larson, R. A. Stier, C. Claiborn, and J. Q. Koenig, "An analysis of the association between respiratory symptoms in subjects with asthma and daily air pollution in Spokane, Washington," *Inhal Toxicol*, vol. 16, no. 13, pp. 809–815, Dec. 2004.

[17] N. N. Hansel, M. C. McCormack, and V. Kim, "The Effects of Air Pollution and Temperature on COPD," *COPD*, vol. 13, no. 3, pp. 372–379, Jun. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878829/

[18] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *Ieee Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014. [Online]. Available: ⟨GotoISI⟩://WOS:000209672000004

[19] S. Darras, C. Granier, C. Liousse, D. Boulanger, N. Elguindi, and H. Le Vu, "The eccad database, version 2: Emissions of atmospheric compounds & compilation of ancillary data," *IGAC Newsletter*, 2018.

[20] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Frontiers in public health*, p. 14, 2020.

[21] L. Rickards, J. Wiseman, and Y. Kashima, "Barriers to effective climate change mitigation: the case of senior government and business decision makers," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 5, no. 6, pp. 753–773, 2014.

[22] I. M. Ciumasu, "Smart solutions for sustainability: Rdi for urban and societal transitions requires cross-sectoral experimentation platforms," in *Sustainable, Innovative and Intelligent Societies and Cities*. Springer, 2023, pp. 203–238.

[23] 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 *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[24] J. Buelvas, D. Múnera, D. P. Tobón V, J. Aguirre, and N. Gaviria, "Data quality in iot-based air quality monitoring systems: a systematic mapping study," *Water, Air, & Soil Pollution*, vol. 234, no. 4, p. 248, 2023.

[25] X. Dai, W. Shang, J. Liu, M. Xue, and C. Wang, "Achieving better indoor air quality with iot systems for future buildings: Opportunities and challenges," *Science of The Total Environment*, p. 164858, 2023.

[26] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environment international*, vol. 99, pp. 293–302, 2017.

[27] G. Miskell, K. Alberti, B. Feenstra, G. S. Henshaw, V. Papapostolou, H. Patel, A. Polidori, J. A. Salmond, L. Weissert, and D. E. Williams, "Reliable data from low cost ozone sensors in a hierarchical network," *Atmospheric Environment*, vol. 214, p. 116870, 2019.

[28] A. C. Rai, P. Kumar, F. Pilla, A. N. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, and D. Rickerby, "End-user perspective of low-cost sensors for outdoor air pollution monitoring," *Science of The Total Environment*, vol. 607, pp. 691–705, 2017.

[29] C.-F. Ou Yang, Y.-C. Lin, N.-H. Lin, C.-T. Lee, G.-R. Sheu, S.-H. Kam, and J.-L. Wang, "Inter-comparison of three instruments for measuring regional background carbon monoxide," *Atmospheric Environment*, vol. 43, no. 40, pp. 6449–6453, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231009008012

[30] N. Wang, G. Pugliese, M. Carrito, C. Moura, P. Vasconcelos, N. Cera, M. Li, P. Nobre, J. Georgiadis, J. Schubert *et al.*, "Breath chemical markers of sexual arousal in humans," *Scientific reports*, vol. 12, no. 1, p. 6267, 2022.

[31] D. K. Farmer, M. E. Vance, J. P. Abbatt, A. Abeleira, M. R. Alves, C. Arata, E. Boedicker, S. Bourne, F. Cardoso-Saldaña, R. Corsi *et al.*, "Overview of homechem: House observations of microbial and environmental chemistry," *Environmental Science: Processes & Impacts*, vol. 21, no. 8, pp. 1280–1300, 2019.

[32] F. Villanueva, M. Ródenas, A. Ruus, J. Saffell, and M. F. Gabriel, "Sampling and analysis techniques for inorganic air pollutants in indoor air," *Applied Spectroscopy Reviews*, vol. 57, no. 7, pp. 531–579, 2022.

[33] O. Poupard, P. Blondeau, V. Iordache, and F. Allard, "Statistical analysis of parameters influencing the relationship between outdoor and indoor air quality in schools," *Atmospheric Environment*, vol. 39, no. 11, pp. 2071–2080, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231005000166

[34] S. Karthikeyan, S. V. Perumal, R. Balasubramanian, M. S. Zuraimi, and K. W. Tham, "Determination of ozone in outdoor and indoor environments using nitrite-impregnated passive samplers followed by ion chromatography," *Journal of the Air & Waste Management Association*, vol. 57, no. 8, pp. 974–980, 2007.

[35] P. Martin, B. Cabañas, F. Villanueva, M. P. Gallego, I. Colmenar, and S. Salgado, "Ozone and nitrogen dioxide levels monitored in an urban area (ciudad real) in central-southern spain," *Water, Air, and Soil Pollution*, vol. 208, pp. 305–316, 2010.

[36] E. I. Tolis, G. Panaras, and J. G. Bartzis, "A comprehensive air quality investigation at an aquatic centre: Indoor/outdoor comparisons," *Environmental Science and Pollution Research*, vol. 25, pp. 16 710–16 719, 2018.

[37] D. B. Collins, R. F. Hems, S. Zhou, C. Wang, E. Grignon, M. Alavy, J. A. Siegel, and J. P. Abbatt, "Evidence for gas–surface equilibrium control of indoor nitrous acid," *Environmental science & technology*, vol. 52, no. 21, pp. 12 419–12 427, 2018.

[38] L. Chatzidiakou, A. Krause, O. A. Popoola, A. Di Antonio, M. Kellaway, Y. Han, F. A. Squires, T. Wang, H. Zhang, Q. Wang *et al.*, "Characterising low-cost sensors in highly portable platforms to quantify personal exposure in diverse environments," *Atmospheric measurement techniques*, vol. 12, no. 8, pp. 4643–4657, 2019.

[39] P. L. Kebabian, E. C. Wood, S. C. Herndon, and A. Freedman, "A practical alternative to chemiluminescence-based detection of nitrogen dioxide: Cavity attenuated phase shift spectroscopy," *Environmental science & technology*, vol. 42, no. 16, pp. 6040–6045, 2008.

[40] J. A. Thornton, P. J. Wooldridge, and R. C. Cohen, "Atmospheric no2: In situ laser-induced fluorescence detection at parts per trillion mixing ratios," *Analytical Chemistry*, vol. 72, no. 3, pp. 528–539, 2000.

[41] V. Földváry, G. Bekö, S. Langer, K. Arrhenius, and D. Petráš, "Effect of energy renovation on indoor air quality in multifamily residential buildings in slovakia," *Building and Environment*, vol. 122, pp. 363–372, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132317302433

[42] W. J. Seow, G. S. Downward, H. Wei, N. Rothman, B. Reiss, J. Xu, B. A. Bassig, J. Li, J. He, H. D. Hosgood *et al.*, "Indoor concentrations of nitrogen dioxide and sulfur dioxide from burning solid fuels for cooking and heating in yunnan province, china," *Indoor Air*, vol. 26, no. 5, pp. 776–783, 2016.

[43] T. Mach and J. S. Bihałowicz, "How to effectively analyse the impact of air quality on society–review of modern measurement techniques and apparatus: Particulates," *Scientific Papers of the Main School of Fire Service*, vol. 84, pp. 55–71, 2022.

[44] F. M. J. Bulot, H. S. Russell, M. Rezaei, M. S. Johnson, S. J. J. Ossont, A. K. R. Morris, P. J. Basford, N. H. C. Easton, G. L. Foster, M. Loxham, and S. J. Cox, "Laboratory comparison of low-cost particulate matter sensors to measure transient events of pollution," *Sensors*, vol. 20, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2219

[45] M. Ródenas García, A. Spinazzé, P. T. Branco, F. Borghi, G. Villena, A. Cattaneo, A. Di Gilio, V. G. Mihucz, E. Gómez Álvarez, S. I. Lopes *et al.*, “Review of low-cost sensors for indoor air quality: Features and applications,” *Applied Spectroscopy Reviews*, vol. 57, no. 9-10, pp. 747–779, 2022.

[46] P. Asha, L. Natrayan, B. Geetha, J. R. Beulah, R. Sumathy, G. Varalakshmi, and S. Neelakandan, “Iot enabled environmental toxicology for air pollution monitoring using ai techniques,” *Environmental research*, vol. 205, p. 112574, 2022.

[47] J. R. Stetter and J. Li, “Amperometric gas sensors a review,” *Chemical reviews*, vol. 108, no. 2, pp. 352–366, 2008.

[48] Y. Wang, J. Li, H. Jing, Q. Zhang, J. Jiang, and P. Biswas, “Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement,” *Aerosol science and technology*, vol. 49, no. 11, pp. 1063–1077, 2015.

[49] A. Di Antonio, O. A. Popoola, B. Ouyang, J. Saffell, and R. L. Jones, “Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter,” *Sensors*, vol. 18, no. 9, p. 2790, 2018.

[50] P. Görner, X. Simon, D. Bémer, and G. Lidén, “Workplace aerosol mass concentration measurement using optical particle counters,” *Journal of Environmental Monitoring*, vol. 14, no. 2, pp. 420–428, 2012.

[51] S. Sousan, K. Koehler, L. Hallett, and T. M. Peters, “Evaluation of the alphasense optical particle counter (opc-n2) and the grimm portable aerosol spectrometer (pas-1.108),” *Aerosol Science and Technology*, vol. 50, no. 12, pp. 1352–1365, 2016, pMID: 28871213. [Online]. Available: https://doi.org/10.1080/02786826.2016.1232859

[52] M. Vogt, P. Schneider, N. Castell, and P. Hamer, “Assessment of low-cost particulate matter sensor systems against optical and gravimetric methods in a field co-location in norway,” *Atmosphere*, vol. 12, no. 8, p. 961, 2021.

[53] A. L. Northcross, R. J. Edwards, M. A. Johnson, Z.-M. Wang, K. Zhu, T. Allen, and K. R. Smith, “A low-cost particle counter as a realtime fine-particle mass monitor,” *Environmental Science: Processes & Impacts*, vol. 15, no. 2, pp. 433–439, 2013.

[54] A. Manikonda, N. Zíková, P. K. Hopke, and A. R. Ferro, “Laboratory assessment of low-cost pm monitors,” *Journal of Aerosol Science*, vol. 102, pp. 29–40, 2016.

[55] S. Sousan, K. Koehler, G. Thomas, J. H. Park, M. Hillman, A. Halterman, and T. M. Peters, “Inter-comparison of low-cost sensors for measuring the mass concentration of occupational aerosols,” *Aerosol Science and Technology*, vol. 50, no. 5, pp. 462–473, 2016.

[56] D. M. Holstius, A. Pillarisetti, K. Smith, and E. Seto, “Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in california,” *Atmospheric Measurement Techniques*, vol. 7, no. 4, pp. 1121–1131, 2014.

[57] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis *et al.*, “Community air sensor network (cairsense) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern united states,” *Atmospheric measurement techniques*, vol. 9, no. 11, pp. 5281–5292, 2016.

[58] K. Kelly, J. Whitaker, A. Petty, C. Widmer, A. Dybwad, D. Sleeth, R. Martin, and A. Butterfield, “Ambient and laboratory evaluation of a low-cost particulate matter sensor,” *Environmental pollution*, vol. 221, pp. 491–500, 2017.

[59] M. Alvarado, F. Gonzalez, A. Fletcher, and A. Doshi, “Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites,” *Sensors*, vol. 15, no. 8, pp. 19 667–19 687, 2015.

[60] N. Zikova, P. K. Hopke, and A. R. Ferro, "Evaluation of new low-cost particle monitors for pm2. 5 concentrations measurements," *Journal of Aerosol Science*, vol. 105, pp. 24–34, 2017.

[61] G. Olivares and S. Edwards, "The outdoor dust information node (odin)–development and performance assessment of a low cost ambient dust sensor," *Atmospheric Measurement Techniques Discussions*, vol. 8, no. 7, pp. 7511–7533, 2015.

[62] G. Olivares, I. Longley, and G. Coulson, "Development of a low-cost device for observing indoor particle levels associated with source activities in the home," *International Society of Exposure Science (ISES), Seattle, WA*, 2012.

[63] E. Austin, I. Novosselov, E. Seto, and M. G. Yost, "Correction: Laboratory evaluation of the shinyei ppd42ns low-cost particulate matter sensor," *PLoS One*, vol. 10, no. 10, p. e0141928, 2015.

[64] M. Gao, J. Cao, and E. Seto, "A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of pm2. 5 in xi'an, china," *Environmental pollution*, vol. 199, pp. 56–65, 2015.

[65] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain, "On the feasibility of measuring urban air pollution by wireless distributed sensor networks," *Science of The Total Environment*, vol. 502, pp. 537–547, 2015.

[66] L. Spinelle, M. Gerboles, M. Aleixandre, and F. Bonavitacola, "Evaluation of metal oxides sensors for the monitoring of o 3 in ambient air at ppb level," *Chemical Engineering Transactions*, vol. 54, pp. 319–324, 2016.

[67] C. Borrego, A. Costa, J. Ginja, M. Amorim, M. Coutinho, K. Karatzas, T. Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito *et al.*, "Assessment of air quality microsensors versus reference methods: The eunetair joint exercise," *Atmospheric Environment*, vol. 147, pp. 246–263, 2016.

[68] L. Spinelle, M. Gerboles, and M. Aleixandre, "Performance evaluation of amperometric sensors for the monitoring of o3 and no2 in ambient air at ppb level," *Procedia engineering*, vol. 120, pp. 480–483, 2015.

[69] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost available sensors for air quality monitoring. part a: Ozone and nitrogen dioxide," *Sensors and Actuators B: Chemical*, vol. 215, pp. 249–257, 2015.

[70] R. Piedrahita, Y. Xiang, N. Masson, J. Ortega, A. Collier, Y. Jiang, K. Li, R. P. Dick, Q. Lv, M. Hannigan *et al.*, "The next generation of low-cost personal air quality sensors for quantitative exposure monitoring," *Atmospheric Measurement Techniques*, vol. 7, no. 10, pp. 3325–3336, 2014.

[71] C. Lin, J. Gillespie, M. Schuder, W. Duberstein, I. J. Beverland, and M. R. Heal, "Evaluation and calibration of aeroqual series 500 portable gas sensors for accurate measurement of ambient ozone and nitrogen dioxide," *Atmospheric Environment*, vol. 100, pp. 111–116, 2015.

[72] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmospheric Environment*, vol. 70, pp. 186–203, 2013.

[73] L. Sun, K. C. Wong, P. Wei, S. Ye, H. Huang, F. Yang, D. Westerdahl, P. K. Louie, C. W. Luk, and Z. Ning, "Development and application of a next generation air sensor network for the hong kong marathon 2015 air quality monitoring," *Sensors*, vol. 16, no. 2, p. 211, 2016.

[74] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. part b: No, co and co2," *Sensors and Actuators B: Chemical*, vol. 238, pp. 706–715, 2017.

[75] P. Han, H. Mei, D. Liu, N. Zeng, X. Tang, Y. Wang, and Y. Pan, "Calibrations of low-cost air pollution monitoring sensors for co, no2, o3, and so2," *Sensors*, vol. 21, no. 1, p. 256, 2021.

[76] D. H. Hagan, G. Isaacman-VanWertz, J. P. Franklin, L. M. Wallace, B. D. Kocar, C. L. Heald, and J. H. Kroll, “Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments,” *Atmospheric Measurement Techniques*, vol. 11, no. 1, pp. 315–328, 2018.

[77] F. Karagulian, M. Barbiere, A. Kotsev, L. Spinelle, M. Gerboles, F. Lagler, N. Redon, S. Crunaire, and A. Borowiak, “Review of the performance of low-cost sensors for air quality monitoring,” *Atmosphere*, vol. 10, no. 9, p. 506, 2019.

[78] E. S. Cross, L. R. Williams, D. K. Lewis, G. R. Magoon, T. B. Onasch, M. L. Kaminsky, D. R. Worsnop, and J. T. Jayne, “Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements,” *Atmospheric Measurement Techniques*, vol. 10, no. 9, pp. 3575–3588, 2017.

[79] M. Gerboles, L. Spinelle, M. Signorini *et al.*, “Airsenseur: an open data/software/hardware multi-sensor platform for air quality monitoring. part a: sensor shield,” *JRC Technical Report*, 2015.

[80] N. Zimmerman, A. A. Presto, S. P. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian, “A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring,” *Atmospheric Measurement Techniques*, vol. 11, no. 1, pp. 291–313, 2018.

[81] P. Wei, Z. Ning, S. Ye, L. Sun, F. Yang, K. C. Wong, D. Westerdahl, and P. K. Louie, “Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring,” *Sensors*, vol. 18, no. 2, p. 59, 2018.

[82] A. Bigi, M. Mueller, S. K. Grange, G. Ghermandi, and C. Hueglin, “Performance of no, no 2 low cost sensors and three calibration approaches within a real world application,” *Atmospheric Measurement Techniques*, vol. 11, no. 6, pp. 3717–3735, 2018.

[83] J. M. Cordero, R. Borge, and A. Narros, “Using statistical methods to carry out in field calibrations of low cost air quality sensors,” *Sensors and Actuators B: Chemical*, vol. 267, pp. 245–254, 2018.

[84] D. Vaughn, T. Dye, P. Roberts, A. Ray, and J. DeWinter, “Characterization of low-cost no2 sensors,” *US Environmental Protection Agency: Washington, DC, USA*, 2010.

[85] L. Sun, D. Westerdahl, and Z. Ning, “Development and evaluation of a novel and cost-effective approach for low-cost no2 sensor drift correction,” *Sensors*, vol. 17, no. 8, p. 1916, 2017.

[86] B. Mijling, Q. Jiang, D. de Jonge, and S. Bocconi, “Practical field calibration of electrochemical no2 sensors for urban air quality applications,” 2017.

[87] M. Mueller, J. Meyer, and C. Hueglin, “Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of zurich,” *Atmospheric Measurement Techniques*, vol. 10, no. 10, pp. 3783–3799, 2017. [Online]. Available: https://amt.copernicus.org/articles/10/3783/2017/

[88] A. Cavaliere, F. Carotenuto, F. Di Gennaro, B. Gioli, G. Gualtieri, F. Martelli, A. Matese, P. Toscano, C. Vagnoli, and A. Zaldei, “Development of low-cost air quality stations for next generation monitoring networks: Calibration and validation of pm2. 5 and pm10 sensors,” *Sensors*, vol. 18, no. 9, p. 2843, 2018.

[89] M. Jovašević-Stojanović, A. Bartonova, D. Topalović, I. Lazović, B. Pokrić, and Z. Ristovski, “On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter,” *Environmental Pollution*, vol. 206, pp. 696–704, 2015.

[90] S. Sousan, K. Koehler, L. Hallett, and T. M. Peters, “Evaluation of the Alphasense optical particle counter (OPC-N2) and the Grimm portable aerosol spectrometer (PAS-1.108),” *Aerosol Science and Technology*, vol. 50, no. 12, pp. 1352–1365, 2016. [Online]. Available: ⟨GotoISI⟩://WOS:000393463100008

[91] I. Han, E. Symanski, and T. H. Stock, "Feasibility of using low-cost portable particle monitors for measurement of fine and coarse particulate matter in urban ambient air," *Journal of the Air & Waste Management Association*, vol. 67, no. 3, pp. 330–340, 2017.

[92] P. J. Dacunto, N. E. Klepeis, K.-C. Cheng, V. Acevedo-Bolton, R.-T. Jiang, J. L. Repace, W. R. Ott, and L. M. Hildemann, "Determining pm 2.5 calibration curves for a low-cost particle monitor: common indoor residential aerosols," *Environmental Science: Processes & Impacts*, vol. 17, no. 11, pp. 1959–1966, 2015.

[93] L. R. Crilley, M. Shaw, R. Pound, L. J. Kramer, R. Price, S. Young, A. C. Lewis, and F. D. Pope, "Evaluation of a low-cost optical particle counter (alphasense opc-n2) for ambient air monitoring," *Atmospheric Measurement Techniques*, vol. 11, no. 2, pp. 709–720, 2018.

[94] A. Mukherjee, L. G. Stanton, A. R. Graham, and P. T. Roberts, "Assessing the utility of low-cost particulate matter sensors over a 12-week period in the cuyama valley of california," *Sensors*, vol. 17, no. 8, p. 1805, 2017.

[95] T. Zheng, M. H. Bergin, K. K. Johnson, S. N. Tripathi, S. Shirodkar, M. S. Landis, R. Sutaria, and D. E. Carlson, "Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments," *Atmospheric Measurement Techniques*, vol. 11, no. 8, pp. 4823–4846, 2018.

[96] H.-Y. Liu, P. Schneider, R. Haugen, and M. Vogt, "Performance assessment of a low-cost pm2. 5 sensor for a near four-month period in oslo, norway," *Atmosphere*, vol. 10, no. 2, p. 41, 2019.

[97] B. Chakrabarti, P. M. Fine, R. Delfino, and C. Sioutas, "Performance evaluation of the active-flow personal dataram pm2. 5 mass monitor (thermo anderson pdr-1200) designed for continuous personal exposure measurements," *Atmospheric Environment*, vol. 38, no. 20, pp. 3329–3340, 2004.

[98] S. Steinle, S. Reis, C. E. Sabel, S. Semple, M. M. Twigg, C. F. Braban, S. R. Leeson, M. R. Heal, D. Harrison, C. Lin *et al.*, "Personal exposure monitoring of pm2. 5 in indoor and outdoor microenvironments," *Science of the Total Environment*, vol. 508, pp. 383–394, 2015.

[99] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statical Learning*, second edition ed., ser. Springer Series in Statics: Data Mining, Inference and Prediction. Springer, 2017.

[100] A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," STATISTICS AND COMPUTING, Tech. Rep., 2003.

[101] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[102] R. M. Morales and J. T. P. Méndez, *Inteligencia artificial: Métodos, técnicas y aplicaciones.* McGraw-Hill España, 2011.

[103] R. Rojas and R. Rojas, "The backpropagation algorithm," *Neural networks: a systematic introduction*, pp. 149–182, 1996.

[104] H. Köhler, "The nucleus in and the growth of hygroscopic droplets," *Transactions of the Faraday Society*, vol. 32, pp. 1152–1161, 1936.

[105] P. Zieger, R. Fierz-Schmidhauser, E. Weingartner, and U. Baltensperger, "Effects of relative humidity on aerosol light scattering: results from different european sites," *Atmospheric Chemistry and Physics*, vol. 13, no. 21, pp. 10 609–10 631, 2013.

[106] C. Malings, R. Tanzer, A. Hauryliuk, S. P. Kumar, N. Zimmerman, L. B. Kara, A. A. Presto, and R. Subramanian, "Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring," *Atmospheric Measurement Techniques*, vol. 12, no. 2, pp. 903–920, 2019.

[107] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep learning for air quality forecasts: a review," *Current Pollution Reports*, pp. 1–11, 2020.

[108] J. Saini, M. Dutta, and G. Marques, "Modeling indoor pm2. 5 using adaptive dynamic fuzzy inference system tree (adfist) on internet of things-based sensor network data," *Internet of Things*, vol. 20, p. 100628, 2022.

[109] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[110] C. P. Filho, E. Marques Jr, V. Chang, L. Dos Santos, F. Bernardini, P. F. Pires, L. Ochi, and F. C. Delicato, "A systematic literature review on distributed machine learning in edge computing," *Sensors*, vol. 22, no. 7, p. 2665, 2022.

[111] M. S. Elbamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, "Wireless edge computing with latency and reliability guarantees," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1717–1737, 2019.

[112] S. Zhang, W. Li, Y. Wu, P. Watson, and A. Zomaya, "Enabling edge intelligence for activity recognition in smart homes," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2018, pp. 228–236.

[113] S. Lin, Z. Zhou, Z. Zhang, X. Chen, and J. Zhang, "On-demand accelerating deep neural network inference via edge computing," in *Edge Intelligence in the Making: Optimization, Deep Learning, and Applications*. Springer, 2021, pp. 151–168.

[114] X. Zhang, Y. Wang, S. Lu, L. Liu, W. Shi *et al.*, "Openei: An open framework for edge intelligence," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1840–1851.

[115] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "Amc: Automl for model compression and acceleration on mobile devices," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 784–800.

[116] P. P. Ray, "A review on tinyml: State-of-the-art and prospects," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1595–1623, 2022.

[117] E. Saurez, K. Hong, D. Lillethun, U. Ramachandran, and B. Ottenwälder, "Incremental deployment and migration of geo-distributed situation awareness applications in the fog," in *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*, 2016, pp. 258–269.

[118] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[119] A. Hassani, A. Medvedev, A. Zaslavsky, P. Delir Haghighi, P. P. Jayaraman, and S. Ling, "Efficient execution of complex context queries to enable near real-time smart iot applications," *Sensors*, vol. 19, no. 24, p. 5457, 2019.

[120] F. Cirillo, G. Solmaz, E. L. Berz, M. Bauer, B. Cheng, and E. Kovacs, "A standard-based open source iot platform: Fiware," *IEEE Internet of Things Magazine*, vol. 2, no. 3, pp. 12–18, 2019.

[121] J. Conde, A. Munoz-Arcentales, A. Alonso, S. Lopez-Pernas, and J. Salvachua, "Modeling digital twin data and architecture: A building guide with fiware as enabling technology," *IEEE Internet Computing*, vol. 26, no. 3, pp. 7–14, 2022.

[122] A. Lazidis, K. Tsakos, and E. G. Petrakis, "Publish–subscribe approaches for the iot and the cloud: Functional and performance evaluation of open-source systems," *Internet of Things*, vol. 19, p. 100538, 2022.

[123] D. Bees, L. Frost, M. Bauer, M. Fisher, and W. Li, "Ngsi-ld api: For context information management," *ETSI White Paper*, no. 31, 2019.

[124] R. A. Light, "Mosquitto: server and client implementation of the mqtt protocol," *Journal of Open Source Software*, vol. 2, no. 13, p. 265, 2017.

[125] J. Conde, A. Munoz-Arcentales, A. Alonso, G. Huecas, and J. Salvachúa, "Collaboration of digital twins through linked open data: Architecture with fiware as enabling technology," *IT Professional*, vol. 24, no. 6, pp. 41–46, 2022.

[126] J. B. Bernabe, J. L. Hernandez-Ramos, and A. F. S. Gomez, "Holistic privacy-preserving identity management system for the internet of things." *Mobile Information Systems*, 2017.

[127] B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, "Building a big data platform for smart cities: Experience and lessons from santander," in *2015 IEEE International Congress on Big Data*. IEEE, 2015, pp. 592–599.

[128] A. Finogeev, A. Finogeev, L. Fionova, A. Lyapin, and K. Lychagin, "Intelligent monitoring system for smart road environment," *Journal of Industrial Information Integration*, vol. 15, pp. 15–20, 2019.

[129] T. Jahan, S. B. Hasan, N. Nafisa, A. A. Chowdhury, R. Uddin, and M. S. Arefin, "Big data for smart cities and smart villages: A review," in *Intelligent Computing & Optimization: Proceedings of the 4th International Conference on Intelligent Computing and Optimization 2021 (ICO2021) 3*. Springer, 2022, pp. 427–439.

[130] D. Graur, I. Müller, M. Proffitt, G. Fourny, G. T. Watts, and G. Alonso, "Evaluating query languages and systems for high-energy physics data," in *Journal of Physics: Conference Series*, vol. 2438, no. 1. IOP Publishing, 2023, p. 012034.

[131] R. Plamanescu, C. Deaconescu, G. Stamatescu, and M. Albu, "High-reporting rate smart metering framework using fiware technology," in *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2023, pp. 1–6.

[132] A. Z. Román-García, R. Nieto, P. V. Hernandez, M. C. Rodriguez-Sanchez, and M. G. Carrillo, "Smart building digital twin: Wireless sensing and actuation architecture at rey juan carlos university," in *2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*. IEEE, 2023, pp. 65–69.

[133] M. Toure, F. Gandon, K. Guidedi, M. Lo, and C. Guéret, "Survey of models and architectures to ensure linked data access," Ph.D. dissertation, Inria & Université Nice Sophia Antipolis, CNRS, I3S, Sophia Antipolis, France, 2023.

[134] M. A. Şahin, M. Yakar, A. Ulvi, and A. Y. Yiğit, "Development of open-source applications using webgis technology," *Intercontinental Geoinformation Days*, vol. 5, pp. 179–182, 2022.

[135] T. Domínguez-Bolaño, O. Campos, V. Barral, C. J. Escudero, and J. A. García-Naya, "An overview of iot architectures, technologies, and existing open-source projects," *Internet of Things*, p. 100626, 2022.

[136] T. Wiangwiset, C. Surawanitkun, W. Wongsinlatam, T. Remsungnen, A. Siritaratiwat, C. Srichan, P. Thepparat, W. Bunsuk, A. Kaewchan, and A. Namvong, "Design and implementation of a real-time crowd monitoring system based on public wi-fi infrastructure: A case study on the sri chiang mai smart city," *Smart Cities*, vol. 6, no. 2, pp. 987–1008, 2023.

[137] A. Medina-Pérez, D. Sánchez-Rodríguez, and I. Alonso-González, "An internet of thing architecture based on message queuing telemetry transport protocol and node-red: A case study for monitoring radon gas," *Smart Cities*, vol. 4, no. 2, pp. 803–818, 2021.

[138] A. Bader, O. Kopp, and M. Falkenthal, "Survey and comparison of open source time series databases," *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband*, 2017.

[139] R. Kumar Kaliyar, "Graph databases: A survey," in *International Conference on Computing, Communication & Automation*. IEEE, 2015, pp. 785–790.

[140] M. A. Rodriguez and P. Neubauer, "The graph traversal pattern," in *Graph data management: Techniques and applications*. IGI global, 2012, pp. 29–46.

[141] A. Aggoune and Z. Benratem, "Ecg data visualization: Combining the power of grafana and influxdb," in *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*. IEEE, 2023, pp. 1–6.

[142] T. Schranz, Q. Alfalouji, T. Hirsch, and G. Schweiger, "An open iot platform: Lessons learned from a district energy system," in *2022 Second International Conference on Sustainable Mobility Applications, Renewables and Technology (SMART)*. IEEE, 2022, pp. 1–9.

[143] L. Girish and S. K. Rao, "Anomaly detection in cloud environment using artificial intelligence techniques," *Computing*, vol. 105, no. 3, pp. 675–688, 2023.

[144] P. John, J. Hynek, T. Hruška, and M. Valný, "Application of time series database for iot smart city platform," in *2023 Smart City Symposium Prague (SCSP)*. IEEE, 2023, pp. 1–6.

[145] Y. Tian, "The world of graph databases from an industry perspective," *ACM SIGMOD Record*, vol. 51, no. 4, pp. 60–67, 2023.

[146] T. Sasaki, Y. Morita, and A. Jada, "Access control architecture for smart city iot platform," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2019, pp. 717–722.

[147] R. Angles and C. Gutierrez, "Survey of graph database models," *ACM Computing Surveys (CSUR)*, vol. 40, no. 1, pp. 1–39, 2008.

[148] R. H. Güting, "Graphdb: Modeling and querying graphs in databases," in *VLDB*, vol. 94. Citeseer, 1994, pp. 12–15.

[149] P. Bellini, P. Nesi, and G. Pantaleo, "Benchmarking rdf stores for smart city services," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*. IEEE, 2015, pp. 46–49.

[150] M. Gavrilescu, "Chapter 5 - Theoretical Predictive Air Quality Models," in *The Quality of Air*, ser. Comprehensive Analytical Chemistry, M. d. l. Guardia and S. Armenta, Eds. Elsevier, 2016, vol. 73, pp. 97–115, iSSN: 0166-526X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166526X16300502

[151] Y. Horie, "Air quality management plan 1988 revision, appendix vp: Ozone episode representativeness study for the south coast air basin," *Report prepared for the South Coast Air Quality Management District, El Monte, CA by Valley Research Corporation, Van Nuys, CA, VRC Project*, no. 057, 1988.

[152] J. C. Cassmassi, "Development of an objective ozone forecast model for the south coast air basin," in *Proceedings, Annual Meeting, Air Pollution Control Association;(USA)*, vol. 4, no. CONF-870695-, 1987.

[153] M. C. Hubbard, *Development of a regression model to forecast ground-level ozone concentration in Jefferson County, Kentucky*. University of Louisville, 1997.

[154] W. F. Ryan, "Forecasting severe ozone episodes in the baltimore metropolitan area," *Atmospheric Environment*, vol. 29, no. 17, pp. 2387–2398, 1995.

[155] V. Mallet and B. Sportisse, "Air quality modeling: From deterministic to stochastic approaches," *Computers & Mathematics with Applications*, vol. 55, no. 10, pp. 2329–2337, 2008.

[156] B. Sportisse, "A review of current issues in air pollution modeling and simulation," *Computational Geosciences*, vol. 11, no. 2, pp. 159–181, 2007.

[157] S. E. Peckham, "Wrf/chem version 3.3 user's guide," 2012.

[158] P. Tuccella, G. Curci, G. Visconti, B. Bessagnet, L. Menut, and R. J. Park, "Modeling of gas and aerosol with wrf/chem over europe: Evaluation and sensitivity study," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D3, 2012.

[159] P. C. Campbell, Y. Tang, P. Lee, B. Baker, D. Tong, R. Saylor, A. Stein, J. Huang, H.-C. Huang, E. Strobach *et al.*, "Development and evaluation of an advanced national air quality forecasting capability using the noaa global forecast system version 16," *Geoscientific Model Development*, vol. 15, no. 8, pp. 3281–3313, 2022.

[160] L. Menut, B. Bessagnet, R. Briant, A. Cholakian, F. Couvidat, S. Mailler, R. Pennel, G. Siour, P. Tuccella, S. Turquety *et al.*, "The chimere v2020r1 online chemistry-transport model," *Geoscientific Model Development*, vol. 14, no. 11, pp. 6781–6811, 2021.

[161] L. Menut, B. Bessagnet, R. Briant, A. Cholakian, F. Couvidat, S. Mailler, R. Pennel, G. Siour, P. Tuccella, S. Turquety, and M. Valari, "The CHIMERE v2020r1 online chemistry-transport model," *Geoscientific Model Development*, vol. 14, no. 11, pp. 6781–6811, Nov. 2021, publisher: Copernicus GmbH. [Online]. Available: https://gmd.copernicus.org/articles/14/6781/2021/

[162] S. Gulia, A. Kumar, and M. Khare, "Performance evaluation of calpuff and aermod dispersion models for air quality assessment of an industrial complex," 2015.

[163] M. Rzeszutek and A. Szulecka, "Assessment of the aermod dispersion model in complex terrain with different types of digital elevation data," in *IOP Conference Series: Earth and Environmental Science*, vol. 642, no. 1. IOP Publishing, 2021, p. 012014.

[164] S. Freitag, A. Clarke, S. Howell, V. Kapustin, T. Campos, V. Brekhovskikh, and J. Zhou, "Combining airborne gas and aerosol measurements with hysplit: a visualization tool for simultaneous evaluation of air mass history and back trajectory consistency," *Atmospheric Measurement Techniques*, vol. 7, no. 1, p. 107, 2014.

[165] M. Ibrahim, "Air quality analyses for photochemical smog associated with atmospheric aerosol particles and ozone precursors using cmaq and camx modeling systems," *International Journal of Scientific Research in Science and Technology*, vol. 224, p. 235, 2019.

[166] I. Bey, D. J. Jacob, R. M. Yantosca, J. A. Logan, B. D. Field, A. M. Fiore, Q. Li, H. Y. Liu, L. J. Mickley, and M. G. Schultz, "Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation," *Journal of Geophysical Research: Atmospheres*, vol. 106, no. D19, pp. 23 073–23 095, 2001.

[167] A. M. Manders, P. J. Builtjes, L. Curier, H. A. Denier van der Gon, C. Hendriks, S. Jonkers, R. Kranenburg, J. J. Kuenen, A. J. Segers, R. M. Timmermans *et al.*, "Curriculum vitae of the lotos–euros (v2. 0) chemistry transport model," *Geoscientific Model Development*, vol. 10, no. 11, pp. 4145–4173, 2017.

[168] L. Gidhagen, C. Johansson, J. Langner, and V. Foltescu, "Urban scale modeling of particle number concentration in stockholm," *Atmospheric Environment*, vol. 39, no. 9, pp. 1711–1725, 2005.

[169] L. W. Horowitz, S. Walters, D. L. Mauzerall, L. K. Emmons, P. J. Rasch, C. Granier, X. Tie, J.-F. Lamarque, M. G. Schultz, G. S. Tyndall *et al.*, "A global simulation of tropospheric ozone and related tracers: Description and evaluation of mozart, version 2," *Journal of geophysical research: Atmospheres*, vol. 108, no. D24, 2003.

[170] D. Spracklen, K. Pringle, K. Carslaw, M. Chipperfield, and G. Mann, "A global off-line model of size-resolved aerosol microphysics: I. model development and prediction of aerosol properties," *Atmospheric Chemistry and Physics*, vol. 5, no. 8, pp. 2227–2252, 2005.

[171] V. Mallet, D. Quélo, B. Sportisse, M. Ahmed de Biasi, É. Debry, I. Korsakissok, L. Wu, Y. Roustan, K. Sartelet, M. Tombette, and H. Foudhil, "Technical Note: The air quality modeling system Polyphemus," *Atmos. Chem. Phys.*, vol. 7, no. 20, pp. 5,479–5,487, 2007.

[172] M. E. Emetere and E. T. Akinlabi, *Introduction to environmental data analysis and modeling.* Springer Nature, 2020, vol. 58.

[173] D. S. McKenna, P. Konopka, J.-U. Grooß, G. Günther, R. Müller, R. Spang, D. Offermann, and Y. Orsolini, "A new chemical lagrangian model of the stratosphere (clams) 1. formulation of advection and mixing," *Journal of Geophysical Research: Atmospheres*, vol. 107, no. D16, pp. ACH–15, 2002.

[174] A. Stohl, C. Forster, A. Frank, P. Seibert, and G. Wotawa, "The lagrangian particle dispersion model flexpart version 6.2," *Atmospheric Chemistry and Physics*, vol. 5, no. 9, pp. 2461–2474, 2005.

[175] I. Pisso, E. Sollum, H. Grythe, N. I. Kristiansen, M. Cassiani, S. Eckhardt, D. Arnold, D. Morton, R. L. Thompson, C. D. Groot Zwaaftink *et al.*, "The lagrangian particle dispersion model flexpart version 10.4," *Geoscientific Model Development*, vol. 12, no. 12, pp. 4955–4997, 2019.

[176] N. Bousserez, J.-L. Attié, V. Peuch, M. Michou, G. Pfister, D. Edwards, L. Emmons, C. Mari, B. Barret, S. Arnold *et al.*, "Evaluation of the mocage chemistry transport model during the icartt/itop experiment," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D10, 2007.

[177] P. A. Makar, W. Gong, J. Zhang, J. Milbrandt, S. Gravel, B. Pabla, and P. Cheung, "A process analysis of the impact of air-quality/weather feedbacks using gem-mach," in *Air Pollution Modeling and its Application XXIII.* Springer, 2014, pp. 229–235.

[178] A. Chu, R. C.-W. Kwok, and K. Yu, "Study of pollution dispersion in urban areas using computational fluid dynamics (cfd) and geographic information system (gis)," *Environmental Modelling & Software*, vol. 20, no. 3, pp. 273–277, 2005.

[179] L. Thouron, Y. Kim, B. Carissimo, C. Seigneur, and B. Bruge, "Intercomparison of two modeling approaches for traffic air pollution in street canyons," *Urban Climate*, vol. 27, pp. 163–178, 2019.

[180] L. Carvalho, "Smart cities from scratch? a socio-technical perspective," *Cambridge Journal of Regions, Economy and Society*, vol. 8, no. 1, pp. 43–60, 2015.

[181] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira, *Smart cities and the future internet: Towards cooperation frameworks for open innovation.* Springer Berlin Heidelberg, 2011.

[182] H. Rahman and M. I. Hussain, "A comprehensive survey on semantic interoperability for internet of things: State-of-the-art and research challenges," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, p. e3902, 2020.

[183] P. Antonios, K. Konstantinos, and G. Christos, "A systematic review on semantic interoperability in the ioe-enabled smart cities," *Internet of Things*, p. 100754, 2023.

[184] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[185] M. Ruta, F. Scioscia, I. Bilenchi, F. Gramegna, G. Loseto, S. Ieva, and A. Pinto, "A multiplatform reasoning engine for the semantic web of everything," *Journal of Web Semantics*, vol. 73, p. 100709, 2022.

[186] O. Novo and M. D. Francesco, "Semantic interoperability in the iot: extending the web of things architecture," *ACM Transactions on Internet of Things*, vol. 1, no. 1, pp. 1–25, 2020.

[187] U. Straccia, *Foundations of fuzzy logic and semantic web languages.* Taylor & Francis, 2013.

[188] O. Lassila, F. van Harmelen, I. Horrocks, J. Hendler, and D. L. McGuinness, "The semantic web and its languages," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 6, pp. 67–73, 2000.

[189] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, "Owl 2: The next step for owl," *Journal of Web Semantics*, vol. 6, no. 4, pp. 309–322, 2008.

[190] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[191] W. W. W. Consortium *et al.*, "Owl 2 web ontology language document overview," 2012.

[192] C. H. Papadimitriou, "Computational complexity," in *Encyclopedia of computer science*, 2003, pp. 260–265.

[193] H. Boley, S. Tabet, and G. Wagner, "Design rationale for ruleml: A markup language for semantic web rules." in *SWWS*, vol. 1, 2001, pp. 381–401.

[194] J. Pérez, M. Arenas, and C. Gutierrez, "Semantics and complexity of sparql," in *International semantic web conference*. Springer, 2006, pp. 30–43.

[195] J. Galgonek and J. Vondrášek, "A comparison of approaches to accessing existing biological and chemical relational databases via sparql," *Journal of Cheminformatics*, vol. 15, no. 1, pp. 1–14, 2023.

[196] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[197] B. Eine, M. Jurisch, and W. Quint, "Ontology-based big data management," *Systems*, vol. 5, no. 3, p. 45, 2017.

[198] E. Ruijer, F. Détienne, M. Baker, J. Groff, and A. J. Meijer, "The politics of open government data: Understanding organizational responses to pressure for more transparency," *The American review of public administration*, vol. 50, no. 3, pp. 260–274, 2020.

[199] M. d'Aquin, J. Davies, and E. Motta, "Smart cities' data: Challenges and opportunities for semantic technologies," *IEEE Internet Computing*, vol. 19, no. 6, pp. 66–70, 2015.

[200] H. Patni, C. Henson, and A. Sheth, "Linked sensor data," in *2010 International Symposium on Collaborative Technologies and Systems*. IEEE, 2010, pp. 362–370.

[201] B. Galitsky, D. Ilvovsky, and E. Goncharova, "Relying on discourse trees to extract medical ontologies from text," in *Russian Conference on Artificial Intelligence*. Springer, 2021, pp. 215–231.

[202] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done," *Queue*, vol. 17, no. 2, pp. 48–75, 2019.

[203] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[204] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[205] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys (Csur)*, vol. 54, no. 4, pp. 1–37, 2021.

[206] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "Linkclimate: An interoperable knowledge graph platform for climate data," *Computers & Geosciences*, vol. 169, p. 105215, 2022.

[207] R. Petrolo, V. Loscri, and N. Mitton, "Towards a smart city based on cloud of things, a survey on the smart city vision and paradigms," *Transactions on emerging telecommunications technologies*, vol. 28, no. 1, p. e2931, 2017.

[208] A. A. Chikkamannur *et al.*, "Semantic annotation of iot resource with ontology orchestration," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*. IEEE, 2020, pp. 1–7.

[209] F. Shi, Q. Li, T. Zhu, and H. Ning, "A survey of data semantization in internet of things," *Sensors*, vol. 18, no. 1, p. 313, 2018.

[210] C. Bizer and A. Seaborne, "D2rq-treating non-rdf databases as virtual rdf graphs," in *Proceedings of the 3rd international semantic web conference (ISWC2004)*, vol. 2004. Springer Hiroshima, 2004.

[211] M. Hert, G. Reif, and H. C. Gall, "A comparison of rdb-to-rdf mapping languages," in *Proceedings of the 7th international conference on semantic systems*, 2011, pp. 25–32.

[212] O. Corcho, F. Priyatna, and D. Chaves-Fraga, "Towards a new generation of ontology based data access," *Semantic Web*, vol. 11, no. 1, pp. 153–160, 2020.

[213] M. Lefrançois, A. Zimmermann, and N. Bakerally, "Flexible rdf generation from rdf and heterogeneous data sources with sparql-generate," in *European Knowledge Acquisition Workshop*. Springer, 2016, pp. 131–135.

[214] G. M. Santipantakis, K. I. Kotis, G. A. Vouros, and C. Doulkeridis, "Rdf-gen: Generating rdf from streaming and archival data," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 2018, pp. 1–10.

[215] M. Palmblad, E. Asein, N. P. Bergman, A. Ivanova, L. Ramasauskas, H. M. Reyes, S. Ruchti, L. Soto-Jácome, and J. Bergquist, "Semantic annotation of experimental methods in analytical chemistry," *Analytical Chemistry*, vol. 94, no. 44, pp. 15 464–15 471, 2022.

[216] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[217] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D344–D350, 2007.

[218] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "Chebi in 2016: Improved services and an expanding collection of metabolites," *Nucleic acids research*, vol. 44, no. D1, pp. D1214–D1219, 2016.

[219] G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton, "Pubchemrdf: towards the semantic annotation of pubchem compound and substance databases," *Journal of cheminformatics*, vol. 7, no. 1, pp. 1–15, 2015.

[220] D. Liang, Z. Li, J. Vlaanderen, Z. Tang, D. P. Jones, R. Vermeulen, and J. A. Sarnat, "A state-of-the-science review on high-resolution metabolomics application in air pollution health research: Current progress, analytical challenges, and recommendations for future direction," *Environmental Health Perspectives*, vol. 131, no. 5, p. 056002, 2023.

[221] C.-H. S. Chen, T.-H. Yuan, R.-H. Shie, K.-Y. Wu, and C.-C. Chan, "Linking sources to early effects by profiling urine metabolome of residents living near oil refineries and coal-fired power plants," *Environment international*, vol. 102, pp. 87–96, 2017.

[222] J. F. Rousseau, E. Oliveira, W. M. Tierney, and A. Khurshid, "Methods for development and application of data standards in an ontology-driven information model for measuring, managing, and computing social determinants of health for individuals, households, and communities evaluated through an example of asthma," *Journal of Biomedical Informatics*, vol. 136, p. 104241, 2022.

[223] P. Moreno, S. Beisken, B. Harsha, V. Muthukrishnan, I. Tudose, A. Dekker, S. Dornfeldt, F. Taruttis, I. Grosse, J. Hastings *et al.*, "Binche: a web tool and library for chemical enrichment analysis based on the chebi ontology," *BMC bioinformatics*, vol. 16, pp. 1–7, 2015.

[224] I. Tudose, J. Hastings, V. Muthukrishnan, G. Owen, S. Turner, A. Dekker, N. Kale, M. Ennis, and C. Steinbeck, "Ontoquery: easy-to-use web-based owl querying," *Bioinformatics*, vol. 29, no. 22, pp. 2955–2957, 2013.

[225] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, and E. Consortium, "The environment ontology: contextualising biological and biomedical entities," *Journal of biomedical semantics*, vol. 4, pp. 1–9, 2013.

[226] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall, "The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation," *Journal of biomedical semantics*, vol. 7, pp. 1–12, 2016.

[227] U. Ahle and J. J. Hierro, "Fiware for data spaces," *Designing Data Spaces*, p. 395, 2022.

[228] Z. Wang, Q. Shen, P. Hua, S. Jiang, R. Li, Y. Li, G. Fan, J. Zhang, and P. Krebs, "Characterizing the anthropogenic-induced trace elements in an urban aquatic environment: A source apportionment and risk assessment with uncertainty consideration," *Journal of Environmental Management*, vol. 275, p. 111288, Dec. 2020, place: London Publisher: Academic Press Ltd- Elsevier Science Ltd WOS:000582474500049.

[229] A. Di Antonio, O. A. M. Popoola, B. Ouyang, J. Saffell, and R. L. Jones, "Developing a Relative Humidity Correction for Low-Cost Sensors Measuring Ambient Particulate Matter," *Sensors*, vol. 18, no. 9, 2018. [Online]. Available: ⟨GotoISI⟩://WOS:000446940600049

[230] B. Bessagnet, M. Beauchamp, L. Menut, R. Fablet, E. Pisoni, and P. Thunis, "Deep learning techniques applied to super-resolution chemistry transport modeling for operational uses," *Environmental Research Communications*, vol. 3, no. 8, p. 085001, 2021.

[231] R. Borge, D. Jung, I. Lejarraga, D. de la Paz, and J. M. Cordero, "Assessment of the madrid region air quality zoning based on mesoscale modelling and k-means clustering," *Atmospheric Environment*, vol. 287, p. 119258, 2022.

[232] M. Valari, K. Markakis, E. Powaga, B. Collignan, and O. Perrussel, "Explume v1. 0: a model for personal exposure to ambient o 3 and pm 2.5," *Geoscientific Model Development*, vol. 13, no. 3, pp. 1075–1094, 2020.

[233] K. C. Tsimillis, "Measurement uncertainty: requirements set in the accreditation standards," *Accreditation and Quality Assurance*, vol. 23, no. 2, pp. 109–114, 2018.

[234] D. Ghernaout, M. Aichouni, and A. Alghamdi, "Overlapping iso/iec 17025: 2017 into big data: A review and perspectives," *International Journal of Science and Qualitative Analysis*, vol. 4, no. 3, pp. 83–92, 2018.

[235] L. Menut, B. Bessagnet, D. Khvorostyanov, M. Beekmann, N. Blond, A. Colette, I. Coll, G. Curci, G. Foret, A. Hodzic, S. Mailler, F. Meleux, J.-L. Monge, I. Pison, G. Siour, S. Turquety, M. Valari, R. Vautard, and M. G. Vivanco, "CHIMERE 2013: a model for regional atmospheric composition modelling," *Geoscientific Model Development*, vol. 6, no. 4, pp. 981–1028, Jul. 2013, publisher: Copernicus GmbH. [Online]. Available: https://gmd.copernicus.org/articles/6/981/2013/

[236] J. G. Verwer, "Gauss–seidel iteration for stiff odes from chemical kinetics," *SIAM Journal on Scientific Computing*, vol. 15, no. 5, pp. 1243–1250, 1994.

[237] R. Atkinson, D. L. Baulch, R. A. Cox, R. F. Hampson, J. A. Kerr, M. J. Rossi, and J. Troe, "Evaluated Kinetic, Photochemical and Heterogeneous Data for Atmospheric Chemistry: Supplement V. IUPAC Subcommittee on Gas Kinetic Data Evaluation for Atmospheric Chemistry," *Journal of Physical and Chemical Reference Data*, vol. 26, no. 3, pp. 521–1011, May 1997, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.556011

[238] B. G. Loh and G. H. Choi, "Calibration of portable particulate matter–monitoring device using web query and machine learning," *Safety and health at work*, vol. 10, no. 4, pp. 452–460, 2019.

[239] L. Parri, M. Tani, D. Baldo, S. Parrino, E. Landi, M. Mugnaini, and A. Fort, "A distributed iot air quality measurement system for high-risk workplace safety enhancement," *Sensors*, vol. 23, no. 11, p. 5060, 2023.

[240] J. A. Bernabé-Díaz, M. del Carmen Legaz-García, J. M. García, and J. T. Fernández-Breis, "Efficient, semantics-rich transformation and integration of large datasets," *Expert Systems with Applications*, vol. 133, pp. 198–214, 2019.

[241] B. Svenningsson, J. Rissler, E. Swietlicki, M. Mircea, M. Bilde, M. C. Facchini, S. Decesari, S. Fuzzi, J. Zhou, J. Mønster, and T. Rosenørn, "Hygroscopic growth and critical supersaturations for mixed aerosol particles of inorganic and organic compounds of atmospheric relevance," *Atmospheric Chemistry and Physics*, vol. 6, no. 7, pp. 1937–1952, Jun. 2006, publisher: Copernicus GmbH. [Online]. Available: https://acp.copernicus.org/articles/6/1937/2006/

[242] M. R. Giordano, C. Malings, S. N. Pandis, A. A. Presto, V. McNeill, D. M. Westervelt, M. Beekmann, and R. Subramanian, "From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors," *Journal of Aerosol Science*, vol. 158, p. 105833, 2021.

[243] B. Dessimond, I. Annesi-Maesano, J.-L. Pepin, S. Srairi, and G. Pau, "Academically produced air pollution sensors for personal exposure assessment: The canarin project," *Sensors*, vol. 21, no. 5, p. 1876, 2021.

[244] T. Rönkkö and H. Timonen, "Overview of sources and characteristics of nanoparticles in urban traffic-influenced areas," *Journal of Alzheimer's disease*, vol. 72, no. 1, pp. 15–28, 2019.

[245] N. A. Ramli, N. F. F. Md Yusof, S. Shith, and A. Suroto, "Chemical and Biological Compositions Associated with Ambient Respirable Particulate Matter: a Review," *Water, Air, & Soil Pollution*, vol. 231, no. 3, p. 120, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s11270-020-04490-5

[246] A. Santurtún, M. L. Colom, P. Fdez-Arroyabe, A. Del Real, I. Fernández-Olmo, and M. T. Zarrabeitia, "Exposure to particulate matter: direct and indirect role in the covid-19 pandemic," *Environmental Research*, vol. 206, p. 112261, 2022.

[247] M. Godin, A. K. Bryan, T. P. Burg, K. Babcock, and S. R. Manalis, "Measuring the mass, density, and size of particles and cells using a suspended microchannel resonator," *Applied physics letters*, vol. 91, no. 12, 2007.

[248] M. M. Daryani, T. Manzaneque, J. Wei, and M. K. Ghatkesar, "Measuring nanoparticles in liquid with attogram resolution using a microfabricated glass suspended microchannel resonator," *Microsystems & Nanoengineering*, vol. 8, no. 1, p. 92, 2022.

[249] S. Olcum, N. Cermak, S. C. Wasserman, K. S. Christine, H. Atsumi, K. R. Payer, W. Shen, J. Lee, A. M. Belcher, S. N. Bhatia *et al.*, "Weighing nanoparticles in solution at the attogram scale," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1310–1315, 2014.

[250] J. P. Specht, S. Esfahani, Y. Xing, A. Köck, M. Cole, and J. W. Gardner, "Thermally modulated cmos-compatible particle sensor for air quality monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[251] P. W. Oluwasanya, G. Rughoobur, and L. G. Occhipinti, "Design, modeling and simulation of a capacitive size-discriminating particulate matter sensor for personal air quality monitoring," *IEEE Sensors Journal*, vol. 20, no. 4, pp. 1971–1979, 2019.

[252] S. Ezrre, M. A. Reyna, C. Anguiano, R. L. Avitia, and H. Márquez, "Lab-on-a-chip platforms for airborne particulate matter applications: a review of current perspectives," *Biosensors*, vol. 12, no. 4, p. 191, 2022.

[253] T. R. Kandukuri, I. Prattis, P. Oluwasanya, and L. G. Occhipinti, "Pathogen detection via impedance spectroscopy-based biosensor," *Sensors*, vol. 24, no. 3, p. 856, 2024.

[254] M. Schaap, C. Cuvelier, C. Hendriks, B. Bessagnet, J. M. Baldasano, A. Colette, P. Thunis, D. Karam, H. Fagerli, A. Graff, R. Kranenburg, A. Nyiri, M. T. Pay, L. Rouïl, M. Schulz, D. Simpson, R. Stern, E. Terrenoire, and P. Wind, "Performance of European chemistry transport models as function of horizontal resolution," *Atmospheric Environment*, vol. 112, pp. 90 – 105, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231015300066

[255] M. Lachatre, S. Mailler, L. Menut, S. Turquety, P. Sellitto, H. Guermazi, G. Salerno, T. Caltabiano, and E. Carboni, "New strategies for vertical transport in chemistry transport models: application to the case of the Mount Etna eruption on 18 March 2012 with CHIMERE v2017r4," *Geoscientific Model Development*, vol. 13, no. 11, pp. 5707–5723, Nov. 2020, publisher: Copernicus GmbH. [Online]. Available: https://gmd.copernicus.org/articles/13/5707/2020/

[256] L. Wu, J. Hang, X. Wang, M. Shao, and C. Gong, "Apfoam 1.0: integrated computational fluid dynamics simulation of o 3–no x–volatile organic compound chemistry and pollutant dispersion in a typical street canyon," *Geoscientific Model Development*, vol. 14, no. 7, pp. 4655–4681, 2021.

[257] J.-L. Santiago, E. Rivas, B. Sanchez, R. Buccolieri, and F. Martin, "The impact of planting trees on nox concentrations: The case of the plaza de la cruz neighborhood in pamplona (spain)," *Atmosphere*, vol. 8, no. 7, p. 131, 2017.

[258] Y. Zhang, X. Ye, S. Wang, X. He, L. Dong, N. Zhang, H. Wang, Z. Wang, Y. Ma, L. Wang *et al.*, "Large-eddy simulation of traffic-related air pollution at a very high resolution in a mega-city: evaluation against mobile sensors and insights for influencing factors," *Atmospheric Chemistry and Physics*, vol. 21, no. 4, pp. 2917–2929, 2021.

[259] R. Briant, C. Seigneur, M. Gadrat, and C. Bugajny, "Evaluation of roadway gaussian plume models with large-scale measurement campaigns," *Geoscientific Model Development*, vol. 6, no. 2, pp. 445–456, 2013.

[260] P. E. Benson, "A review of the development and application of the caline3 and 4 models," *Atmospheric Environment. Part B. Urban Atmosphere*, vol. 26, no. 3, pp. 379–390, 1992.

[261] L. Soulhac, P. Salizzoni, F.-X. Cierco, and R. Perkins, "The model sirane for atmospheric urban pollutant dispersion; part i, presentation of the model," *Atmospheric environment*, vol. 45, no. 39, pp. 7379–7395, 2011.

[262] Y. Kim, Y. Wu, C. Seigneur, and Y. Roustan, "Multi-scale modeling of urban air pollution: Development and application of a street-in-grid model (v1. 0) by coupling munich (v1. 0) and polair3d (v1. 8.1)," *Geoscientific Model Development*, vol. 11, no. 2, pp. 611–629, 2018.

[263] J. Wang, X. Zhang, Z. Guo, and H. Lu, "Developing an early-warning system for air quality prediction and assessment of cities in china," *Expert systems with applications*, vol. 84, pp. 102–116, 2017.

[264] S.-A. Sansone, P. Rocca-Serra, M. Wilkinson, and L. Harland, "Fair: Making data ai-ready," in *Artificial Intelligence for Science: A Deep Learning Revolution.* World Scientific, 2023, pp. 627–640.

[265] N. Shukla, S. Gulia, and S. K. Goyal, "Regional scale air quality modelling system in india: issues, challenges and suggestive framework," *Arabian Journal of Geosciences*, vol. 16, no. 6, p. 387, 2023.

[266] J. Oliveira, C. Delgado, and A. C. Assaife, "A recommendation approach for consuming linked open data," *Expert Systems with Applications*, vol. 72, pp. 407–420, 2017.

[267] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE computer graphics and applications*, vol. 10, no. 4, pp. 76–82, 1990.

[268] M. D. Legaz-Garcia, J. A. Minarro-Gimenez, M. Menarguez-Tortosa, and J. T. Fernandez-Breis, "Generation of open biomedical datasets through ontology-driven transformation and integration processes," *Journal of Biomedical Semantics*, vol. 7, 2016. [Online]. Available: ⟨GotoISI⟩://WOS:000377175600001

[269] K. Velasquez, "Mechanisms for latency reduction in fog environments," Ph.D. dissertation, Coimbra, Portugal, 5 2021.

[270] Y. Tang, N. Cheng, W. Wu, M. Wang, Y. Dai, and X. Shen, "Delay-minimization routing for heterogeneous vanets with machine learning based mobility prediction," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3967–3979, 2019.

[271] D. Rahbari and M. Nickray, "Task offloading in mobile fog computing by classification and regression tree," *Peer-to-Peer Networking and Applications*, vol. 13, pp. 104–122, 2020.

[272] J. Park, H. Ryu, E. Kim, Y. Choe, J. Heo, J. Lee, S.-H. Cho, K. Sung, M. Cho, and W. Yang, "Assessment of pm2. 5 population exposure of a community using sensor-based air monitoring instruments and similar time-activity groups," *Atmospheric Pollution Research*, vol. 11, no. 11, pp. 1971–1981, 2020.

[273] G. S. Evans, H. Kloke, and S. Jahn, "Review of ethics for occupational hygiene hazard monitoring surveys using sensors," *Journal of Occupational and Environmental Hygiene*, vol. 20, no. 10, pp. 439–451, 2023.

[274] S. Masri, K. Cox, L. Flores, J. Rea, and J. Wu, "Community-engaged use of low-cost sensors to assess the spatial distribution of pm2. 5 concentrations across disadvantaged communities: Results from a pilot study in santa ana, ca," *Atmosphere*, vol. 13, no. 2, p. 304, 2022.

[275] D. Karanassos, C. Kyfonidis, G. Angelis, A. Emvoliadis, T.-I. Theodorou, A. Zamichos, A. Drosou, and D. Tzovaras, "Socio-bee: A next-generation citizen science platform for citizens' engagement to air pollution measuring," in *2023 IEEE International Smart Cities Conference (ISC2)*. IEEE, 2023, pp. 01–05.