XVI Jornadas de Ingeniería Telemática

JITEL 2023

La Salle – Universitat Ramon Llull

Actas de las XVI Jornadas de Ingeniería Telemática (JITEL 2023), Barcelona (España), 8-10 de noviembre de 2023.

Desarrollo y evaluación de técnicas para el análisis de la calidad de la información proveniente de infraestructuras IoT

Laura Martín, Luis Sánchez, Jorge Lanza, Pablo Sotres
Departamento Ingeniería de Comunicaciones,
Universidad de Cantabria
Plaza de la Ciencia s/n, Santander, 3005, España.
lmartin@tlmat.unican.es, lsanchez@tlmat.unican.es, jlanza@tlmat.unican.es, psotres@tlmat.unican.es

Hoy en día, los datos son el motor de la prosperidad económica y la base fundamental de sistemas de toma de decisiones, apoyados por la Internet de las Cosas como una de las tecnologías más representativas. Por ello, la calidad de estos datos se convierte en un aspecto crítico. En este trabajo se presenta una solución para evaluar varias dimensiones de calidad de los flujos de datos IoT a medida que se generan. Además, el enfoque seguido se centra en añadir la información sobre la calidad de los datos como metadatos vinculados a cada elemento del flujo de datos, aprovechando los principios de los datos enlazados y el estándar NGSI-LD para armonizar y enriquecer fuentes de datos heterogéneas. Por último, se evalúa el diseño propuesto, logrando un compromiso sólido entre funcionalidad y sobrecarga, con un rendimiento estable y escalable.

***Palabras Clave*—calidad de los datos, curado de los datos, Inteligencia Artificial, Internet de las Cosas**

I. INTRODUCCIÓN

El creciente número de fuentes de datos asociadas al despliegue de la Internet de las Cosas (IoT, *Internet of Things*), así como aquellas vinculadas a portales de datos abiertos y plataformas de redes sociales están generando una inmensa cantidad de información. Estos datos son sumamente útiles y beneficiosos, tanto para el sector público como privado, gracias al desarrollo de servicios de valor añadido, el aumento de la transparencia y la disponibilidad de las administraciones o el fomento de la eficiencia de los servicios públicos. En febrero de 2020, la Comisión Europea anunció la Estrategia Europea de Datos [1], cuyo objetivo es crear un mercado único para que los datos se compartan e intercambien entre diferentes sectores de manera eficiente y segura dentro de la UE.

Entre las tecnologías que van a desempeñar un papel clave en la futura Economía de los Datos, la IoT es reconocida como una tecnología que cambia las reglas del juego y amplía su aplicabilidad a una enorme variedad de dominios [2]–[5]. Precisamente, los datos generados constantemente por la miríada de sensores incrustados en el entorno pueden transformarse en conocimiento de valor añadido si se procesan de manera inteligente.

Sin embargo, cuanto mayor es la infraestructura IoT desplegada, más probables son los fallos del sistema y de la red [6]. Los errores se deben, principalmente, al uso generalizado de dispositivos de bajo coste, aspecto fundamental para conseguir tener un mayor alcance y adaptación de la IoT en la sociedad. Esto también genera inquietudes en cuanto al uso de estos dispositivos, debido a sus niveles de precisión, fiabilidad, aplicabilidad sobre el terreno y rendimiento. Los errores y fallos introducidos por el uso de dispositivos de coste reducido pueden dar lugar a una mala calidad de los datos (DQ, *Data Quality*), que, a su vez, terminan conduciendo a resultados indeseados en sistemas de toma de decisiones.

Dentro de la literatura se pueden encontrar múltiples definiciones del concepto de DQ. La definición propuesta por [7] es una de las más extendidas, acuñando la filosofía "*adecuado para su uso*". Respecto a las clasificaciones de las métricas para la medición de la calidad de los datos, se identifican numerosas compilaciones y definiciones de las dimensiones de calidad [7]–[10], destacando el conjunto inicialmente formulado por [7] (Categorías Intrínseca, Contextual, Accesibilidad y Representación).

La mayoría de trabajos enfocados al concepto de DQ se han concentrado en el aspecto Big Data de la IoT, investigando, clasificando y discutiendo su gestión para grandes conjuntos de datos [11]. Sin embargo, la IoT posee una naturaleza dinámica fundamental y una parte importante de su potencial proviene de su capacidad para generar datos en tiempo real organizados en los llamados flujos de datos (es decir, series temporales de observaciones sucesivas tomadas por dispositivos IoT). Por lo tanto, es de

gran importancia que la calidad digital se evalúe y mejore a medida que se generan y recogen los datos. Centrando el estudio en la calidad de la información en entornos IoT y flujos de datos, en [12]–[14] se realizan diferentes selecciones de dimensiones de calidad de datos dedicadas a este tipo de escenarios.

Dados los análisis realizados por estas tres publicaciones sobre las dimensiones de calidad enfocadas a entornos IoT, se han priorizado aquellas que proveen información sobre los datos tanto de manera individual como en conjunto, escogiendo finalmente cinco dimensiones: Exactitud —*Accuracy*, Precisión —*Precision*, Puntualidad —*Timeliness*, Integridad —*Completeness* y Usabilidad —*Usabillity*. Esta última propiedad, Usabilidad, se aporta en este trabajo, promoviendo la interoperabilidad de la información y su fácil consumo, de acuerdo con la representación homogénea y estadarizada de los flujos de datos.

Junto con la definición teórica de las dimensiones DQ, se deben especificar los métodos para evaluar la calidad de los flujos de datos en entornos IoT. En [15], los autores se centran en el cálculo de un conjunto de dimensiones DQ y en la aplicación de este tipo de técnicas de mejora y evaluación de la calidad, pero sólo una vez se ha obtenido un conjunto de datos lo suficientemente extenso como para llevar a cabo estas acciones. Por el contrario, en este artículo se propone el modelado y cálculo de las dimensiones DQ seleccionadas, así como varias técnicas de enriquecimiento de la calidad de los datos. Ambos procesos se van a poder aplicar sobre las observaciones recogidas dentro de despliegues IoT de forma continua, es decir, a medida que se van generando.

Asimismo, esta información de calidad obtenida como metadatos DQ se incluirán junto con los flujos de datos evaluados. De este modo, se consigue el máximo potencial del curado de datos al poder acceder directamente a estas características DQ para cada elemento del flujo de datos. Este enfoque no está muy extendido en la literatura, y es más típico evaluar las dimensiones de calidad de manera periódica y enviar alertas en caso de que se superen ciertos límites, tal y como propone [16].

Por otro lado, existen diferentes enfoques sobre la integración del módulo de evaluación de la calidad en las arquitecturas de adquisición y consumo de datos. En [16], se propone una arquitectura y metodología para la evaluación y monitorización de la DQ de forma externa a una plataforma IoT. Sin embargo, en [17] se discute la incorporación de un conjunto de actividades que pueden ser realizadas en paralelo o integradas con el resto de actividades en curso. En este trabajo se propone la integración del Módulo de Curado de Datos IoT en una arquitectura DET (*Data Enrichment Toolchain*) para evaluar la viabilidad e idoneidad de incluir dimensiones DQ en plataformas IoT existentes para mejorar aún más la dimensión Usabilidad que se ha definido.

Así, las contribuciones clave de este trabajo son: (i) identificar y definir las dimensiones DQ más relevantes en los flujos de datos IoT y los mecanismos para evaluar cada una de ellas; (ii) especificar e implementar soluciones de curado de datos que empleen algoritmos de IA (Inteligencia Artificial) para enriquecer los flujos de datos IoT incrementando sus características de calidad; (iii) integrar estos dos tipos de mecanismos (es decir, la evaluación y el enriquecimiento de la DQ de los flujos de datos IoT) en una DET operativa que aproveche los principios de datos enlazados y el estándar NGSI-LD para proporcionar datos enriquecidos semánticamente; y (iv) llevar a cabo una síntesis y reevaluación de los resultados presentados en [18] con el objetivo de resaltar los beneficios conseguidos por el Módulo de Curado de Datos IoT en las diferentes dimensiones de calidad de la información evaluadas.

II. DIMENSIONES DE CALIDAD DE LA INFORMACIÓN

El concepto DQ es crucial para los procesos de minería de datos y análisis de la información, siendo crítico en el uso de la tecnología IoT debido a su gran impacto en los productos finales [14], [19]. DQ define el grado de cumplimiento de los requisitos impuestos por los consumidores de datos, y las dimensiones DQ se refieren a los criterios que deben cumplirse para que los resultados del análisis y el consumo de información sean óptimos y no se vean comprometidos.

Es importante destacar que el planteamiento seguido se enfoca en proporcionar conocimiento sobre las dimensiones DQ que sea útil para que los consumidores sean capaces de comprender los datos que van a recibir. Es decir, no se trata de obtener un valor único que indique si un ítem de un flujo de datos es de alta o baja calidad, sino de incluir metadatos que permitan decidir si ese elemento tiene suficiente calidad o no, enriqueciendo así los criterios de selección del flujo de datos.

A. Definiciones

Siguiendo la clasificación en categorías de las dimensiones de calidad ofrecida por [7], se han escogido las siguientes: Exactitud (*Accuracy*) como dimensión Intrínseca clave; Integridad (*Completeness*), Puntualidad (*Timeliness*) y Precisión (*Precision*) como dimensiones Contextuales; y, una quinta dimensión aportada en este trabajo, Usabilidad (*Usability*), que de alguna manera aúna los aspectos de las categorías DQ de Representación y Accesibilidad.

En los siguientes apartados se exponen las definiciones utilizadas y sus correspondientes métodos de cálculo para cada una de las dimensiones DQ seleccionadas.

1) Exactitud: (*Accuracy*) indica lo cerca que está el valor medido por el dispositivo IoT del valor considerado como verdad absoluta. Esto se representa en la Ec. 1, tomando las unidades del valor de observación y representado por el símbolo ± a su lado.

$$exactitud = |valorObservado - valorReferencia| \quad (1)$$

2) Integridad: (*Completeness*) cuantifica el número de observaciones perdidas en una ventana temporal determinada. El método seguido para el cálculo de este parámetro se muestra en la Ec. 2 y se representa en parte por unidad (ppu). Los parámetros que intervienen en este

método son: $ventana\_temporal$, ventana temporal; n, número de observaciones consideradas perdidas en esa ventana temporal; y $tasa$, frecuencia de llegada media de las observaciones (también conocida como Puntualidad o *Timeliness* en este trabajo).

$$integridad = \frac{ventana\_temporal - n \cdot tasa}{ventana\_temporal} \quad (2)$$

3) Puntualidad: (*Timeliness* o tn) es una dimensión con una gran cantidad de definiciones dentro de la literatura [8], [13], [14], [19]. El enfoque seguido en este trabajo define esta propiedad como la tasa de actualización en el sistema del valor observado y su cálculo se encuentra representado en la Ec. 3 como una media ponderada a través del parámetro α del valor de tn de la iteración anterior (tn_{i-1}) y el calculado con la llegada de la observación que se está evaluando (tn_{bruto}).

$$tn_i = \alpha \cdot tn_{i-1} + (1 - \alpha) \cdot tn_{bruto} \quad (3)$$

4) Precisión: Pese a que la propiedad de Precisión (*Precision*) no pertenece al conjunto original de dimensiones de calidad para IoT encontrado en la literatura, sí que suele aparecer en segundo plano [7], [8], [10], [12]. Tomando como base la definición ISO 5725 de Precisión [20], el objetivo es evaluar la cercanía entre los valores del flujo de datos. El método seguido para su cálculo se muestra en la Ec. 4 y su resultado, incluido con el símbolo ±, toma las unidades de los valores del conjunto de datos. Los parámetros involucrados en este cálculo son: x_i, elemento i del conjunto de datos; μ, media aritmética del conjunto de datos o, en este caso, valor de la observación que se está evaluando; n, número de elementos en el flujo de datos.

$$precisión = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \quad (4)$$

5) Usabilidad: A diferencia de las anteriores y dentro del alcance del estudio de este trabajo, la dimensión Usabilidad no está definida, como tal, en la literatura [7], [10], [21]. El concepto más similar podría ser el de "*adecuación para su uso*" descrito en [7]. Ambos conceptos comparten el enfoque de definir una dimensión de calidad polifacética que esté centrada en el consumidor. Sin embargo, se orienta a dimensiones intrínsecas objetivas, sin prestar suficiente atención a las dimensiones subjetivas y contextuales relacionadas con el procedimiento real del consumo de datos.

Usabilidad, al tratarse de una dimensión en la perspectiva subjetiva de DQ, no incluye una expresión cerrada para evaluarla, pero para que los datos puedan calificarse como de alta calidad en sus términos deben cumplir tres aspectos básicos. En primer lugar, los datos deben estar representados de manera uniforme, siguiendo modelos estándar y consiguiendo así coherencia en su representación. En segundo lugar, es importante comprender cuál es la procedencia y linaje de los datos, conociendo cuál ha sido el tratamiento previo al que han sido sometidos, con el objetivo de proveer la confianza necesaria en los consumidores para inyectar estos datos en sus aplicaciones. Por último, la representación de los datos de debe realizarse de manera semánticamente rica, de modo que los datos no sean sólo valores, sino que estén vinculados a cualquier característica asociada a ese valor, en particular las relacionadas con sus dimensiones de calidad.

B. Mecanismos basados en IA para el curado de la información

En [19] se proponen técnicas de mejora de la calidad para datos proporcionados en un entorno IoT: detección de valores atípicos, interpolación, integración de datos, eliminación de duplicados y limpieza de datos. De entre todas estas técnicas, en este trabajo se hace hincapié sobre la imputación de valores, tanto en conjuntos de datos estáticos como en flujos de datos IoT en tiempo real.

La ausencia de valores en conjuntos de datos es un asunto crucial que debe gestionarse, ya que estos valores perdidos pueden sesgar los resultados de la aplicación de otras técnicas o, directamente, reducir la calidad de la información [19]. Dos de los métodos más conocidos para abordar la ausencia de valores son la interpolación y la aplicación de Inteligencia Artificial [19], [21]–[23]. Las técnicas derivadas de este último método se basan, principalmente, en el aprendizaje supervisado, estimando los valores ausentes en función de la información disponible por los valores existentes. El algoritmo kNN (k-Nearest Neighbour) es uno de los más significativos que, utilizado en este enfoque de imputación de valores, implica la estimación del hueco en base a una métrica de distancia entre sus k vecinos más cercanos.

En un escenario IoT, donde la información se distribuye en series temporales y debe procesarse en tiempo real, algunas de las técnicas y dimensiones DQ comentadas anteriormente no se pueden aplicar. Las técnicas que utilizan métodos de Inteligencia Artificial se ven obligadas a reentrenar sus modelos por cada nueva observación recibida de los dispositivos IoT desplegados, ya que la marca temporal o timestamp es crucial. Por tanto, en un entorno IoT con una gestión de los datos en tiempo real se puede determinar que esta situación de reentreno continuo no es escalable. El planteamiento propuesto para abordar este tipo de situaciones se basa en la estimación de valores futuros, aumentando sintéticamente el conjunto de datos y así poder aplicar las técnicas de enriquecimiento sobre este intervalo de tiempo adicional. La estructura de datos más común en un entorno IoT, como ya se ha comentado, es la serie temporal, donde se pueden manifestar características como la estacionalidad. Debido a esto, se ha considerado que el modelo más adecuado para su análisis y estimación es el algoritmo ARIMA Estacional (SARIMA) [24].

III. CURADO AUTÓNOMO DE DATOS IOT

Haciendo referencia a la dimensión de calidad propuesta como Usabilidad, esta propiedad impone requisitos sobre la representación y el uso de los datos tratados. En este trabajo se defiende la idea de que los resultados, tanto de la aplicación de técnicas de enriquecimiento como de la evaluación de las dimensiones de calidad, no tienen valor

por sí mismos, sino que este valor se potencia una vez que se integran en los flujos de datos procesados. De esta manera, se consigue que el consumidor tenga acceso a información de mayor calidad (resultante de las técnicas de enriquecimiento y curado de datos) cuyas características de calidad (producto de las técnicas de evaluación de calidad y de las dimensiones DQ) sean conocidas.

El enfoque seguido para incorporar los resultados de los mecanismos de curación de datos desarrollados e introducir los metadatos asociados a la información proporcionada por los entornos IoT es a través de la integración de estos mecanismos dentro de una DET.

En los siguientes apartados se presenta la arquitectura general de la cadena de enriquecimiento (DET) y los aspectos clave que se han tenido que cumplir para integrar los mecanismos necesarios.

A. *Arquitectura de la cadena de enriquecimiento*

En este trabajo, la DET se define como una cadena de microservicios heterogéneos que tiene como resultado la mejora progresiva de la calidad y valor de la información original. Para ello, se han identificado cinco funciones clave: descubrimiento, con la capacidad de descubrir y solicitar recopilaciones de conjuntos y flujos de datos; formateo, con la transformación de los datos en bruto en conjuntos bien formados y estructurados de acuerdo a los modelos de datos descritos en el estándar NGSI-LD [25]; curado, con la identificación y/o corrección de datos que no reflejan la calidad esperada; enlazado, con la capacidad de relacionar diferentes conjuntos de datos; y, finalmente, enriquecimiento, con la capacidad de comprender y enmarcar las estructuras de datos en situaciones y contextos.

Como se puede ver en la Fig. 1, las fuentes de datos heterogéneas (como por ejemplo conjuntos y flujos de datos, CSV, JSON) se descubren e identifican en el primer paso descrito. A continuación, se aplica la transformación a NGSI-LD con el objetivo de armonizar los datos entrantes y permitir que los procesos posteriores se aprovechen de la semántica y los principios de los datos enlazados. Justo antes de que todos estos datos se almacenen en el Context Broker (pieza central de la API de NGSI-LD [25]), se aplican las técnicas de curado de los datos con el fin de maximizar la DQ, de modo que su resultado sirva como base para las aplicaciones y el resto de bloques funcionales de enlazado y enriquecimiento de la arquitectura.

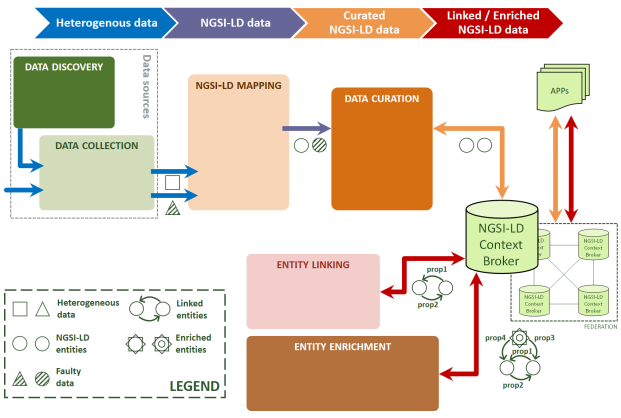

Fig. 1. Arquitectura de la cadena de enriquecimiento (DET) [26].

B. *Modelado de la información de calidad*

Tal y como se ha ido comentado a lo largo del artículo, los resultados de la aplicación de técnicas de enriquecimiento de DQ y el análisis de las dimensiones DQ sólo son valiosos si se incorporan en los flujos de datos como metadatos. De esta manera, los consumidores tienen la capacidad de comprender totalmente el significado de la información, no sólo a partir de su valor en bruto, sino también de sus características DQ.

Por lo tanto, se ha determinado que no sólo las mediciones generadas por los dispositivos IoT, sino también sus dimensiones y propiedades DQ obtenidas por las técnicas de mejora de la calidad, se representarán en el formato definido por el estándar NGSI-LD [25]. La aplicación de este estándar para la representación de los flujos de datos y sus características DQ se alinea con la materialización de los objetivos fijados en la dimensión Usabilidad. Se consigue la representación uniforme de la información de acuerdo con la normativa, la trazabilidad basada en los principios de los datos enlazados y la provisión de valor extra con metadatos a través de los campos de las propiedades incrustadas.

Con anterioridad a la realización de este trabajo, no existía ningún modelo de datos específico para la representación de la información de evaluación de la calidad en el catálogo de la iniciativa Smart Data Models [27]. Es por ello que, como contribución, se ha diseñado y proporcionado un modelo de datos que recoge todas las dimensiones DQ abordadas, así como diferentes propiedades adquiridas con las técnicas de enriquecimiento de datos para la trazabilidad de su aplicación. La documentación relevante se encuentra disponible en [28].

El objetivo principal de este formato de modelado de la información es vincular la observación recibida por los dispositivos IoT (entidad valor) y las propiedades de calidad obtenidas (modelo de datos propuesto).

C. *Módulo de curado de información IoT*

El componente responsable del curado de la información dentro de la arquitectura de la cadena de enriquecimiento (DET) se denomina Módulo de Curado de Datos IoT. Su principal objetivo es aplicar las técnicas de enriquecimiento de DQ y obtener las dimensiones de calidad para poder almacenar la información con la mayor calidad posible en el Context Broker.

El elemento de entrada de este componente se basa en la observación recibida por las fuentes heterogéneas representada ya en el formato definido por el estándar NGSI-LD en el modelo de datos de la iniciativa Smart Data Models que más se ajuste. Como ya se ha comentado, el objetivo de este componente es recopilar las características de calidad de la observación obtenida a la entrada, consiguiendo a su salida dos entidades. La primera de ellas se corresponde con la entidad procesada y, la segunda, con el modelo de

datos propuesto en el apartado anterior que reúne todas las propiedades que representan la calidad de la observación procesada. Es importante comentar que a la entidad valor (o entidad procesada) se le debe añadir un nuevo campo para crear la relación entre ambas entidades de salida, consiguiendo que estén explícitamente vinculadas y así alcanzar los requisitos impuestos por la dimensión de Usabilidad propuesta en este trabajo.

IV. EVALUACIÓN DE LAS TÉCNICAS DE ENRIQUECIMIENTO DE LA CALIDAD DE LA INFORMACIÓN

En esta sección se describe la evaluación llevada a cabo para caracterizar el comportamiento y el rendimiento de las técnicas de mejora de la calidad de la información propuestas en apartados anteriores.

A. Evaluación de las dimensiones de calidad

Tras la definición de las dimensiones DQ seleccionadas y propuestas anteriormente, se caracteriza su comportamiento y rendimiento en un entorno de pruebas.

Se ha simulado la existencia de 100 sensores desplegados que generan observaciones cada 2 minutos, hasta un total de 100 observaciones por dispositivo IoT. Con esto, se consigue un flujo de datos de 10000 elementos de datos como entrada para la cadena de evaluación DQ, por cada una de las 60 iteraciones del método de Monte Carlo [29] realizadas, con el fin de extraer conclusiones de validez general. Todas estas pruebas se han realizado sobre una máquina Ubuntu 20.04.5 LTS (2 núcleos CPU, 2.40 GHz de reloj, 16 GB de RAM) en la que se ha empleado un Broker Scorpio NGSI-LD como Context Broker. Al terminar cada una de las 60 iteraciones, el Context Broker se reinicia para comenzar desde su estado inicial.

La fórmula utilizada para calcular la sobrecarga (*sobrecarga*), en términos de tamaño, introducida por la inclusión de los metadatos generados a través del procedimiento de evaluación de la calidad se presenta en la Ec. 5. En ella, *enriquecida* se refiere al tamaño de la entidad incluyendo información sobre las dimensiones DQ, y *en bruto* se refiere al tamaño original de la entidad (es decir, sólo su valor sin elementos adicionales).

$$sobrecarga(\%) = \frac{enriquecida - en\,bruto}{en\,bruto} \cdot 100 \quad (5)$$

De manera complementaria a la sobrecarga en términos de tamaño, también se introduce cierto retardo en el tiempo de cálculo. Este retardo total se puede dividir en dos fases: el tiempo empleado en obtener toda la información necesaria para realizar el cálculo de la dimensión DQ por cada elemento de datos (denominado, *Retardo de petición*), y el tiempo necesario para efectuar realmente ese cálculo (denominado, *Retardo de procesado*).

En los siguientes apartados, se detallan los procesos de cálculo de cada una de las dimensiones de calidad.

1) Exactitud: Haciendo referencia a la definición propuesta para la dimensión de Exactitud, su valor se obtiene a partir de un valor de referencia. Por tanto, es evidente que se debe realizar una petición a una fuente externa de confianza que provea este valor. A efectos de esta investigación, la fuente externa se corresponde con la Agencia Estatal de Meteorología (AEMET) [30], ya que las observaciones que se están analizando son de tipo Temperatura dentro de la zona de Santander, Cantabria. Una vez que el submódulo recibe esta información de referencia, realiza el cálculo correspondiente (Ec. 1).

2) Integridad: Para realizar el cálculo de la dimensión de Integridad, es necesario que el Context Broker soporte el almacenamiento de valores temporales. De esta manera, el primer paso consiste en solicitar los valores almacenados en una ventana temporal predefinida para cada flujo recibido. Esta ventana temporal predefinida refuerza la correlación temporal para la evaluación de la dimensión de Integridad, lo que significa que se trata de una dimensión que debe tener en cuenta un historial limitado y no el completo del flujo de datos. El submódulo tiene que consultar estos valores históricos tanto para el tipo de entidad evaluado (por ejemplo, Temperatura) como para la entidad DataQualityAssessment [28] vinculada a ella. Tras obtener estos datos, el submódulo es capaz de evaluar la expresión en la Ec. 2. Es importante señalar que *n* sería el número de valores etiquetados como sintéticos en el modelo de datos propuesto y *rate* sería el valor de Puntualidad (*Timeliness*) de la observación actual.

3) Puntualidad: El cálculo de la dimensión de Puntualidad, al igual que las propiedades anteriores, es un procedimiento que se divide en dos fases: solicitud de la información necesaria y procesado para la obtención del valor de Puntualidad. En la fase de solicitud, el submódulo consulta al Context Broker el último valor de Puntualidad registrado de la entidad de evaluación de la calidad vinculada al *id* de la observación recibida. Una vez obtenido este valor, el submódulo continua con la fase de procesado, ejecutando el cálculo definido en la Ec. 3.

4) Precisión: La última dimensión que se procesa es Precisión. En este caso, para cada observación o entidad recibida en el submódulo, se debe realizar una solicitud al Context Broker sobre todas las entidades registradas que se encuentren dentro de un área determinado, tanto para el tipo de entidad evaluado como para la información de calidad vinculada a esta. Una vez el submódulo obtiene toda esta información, es capaz de efectuar la evaluación de la dimensión de Precisión que se describía en la Ec. 4.

5) Análisis de la sobrecarga introducida: Una vez descritos los procesos de cálculo de las dimensiones de calidad, conociendo cuáles son los requisitos de información previa de cada uno de ellos, se presentan los resultados obtenidos en términos de retardo y sobrecarga en las Tablas I y II.

En la Tabla I se muestran los valores medios, tras las 60 simulaciones de Monte Carlo, de los retardos en cada una de las fases necesarias para el cálculo de las dimensiones. Se puede ver que la dimensión Exactitud es la que mayor retardo implica, principalmente ocasionado por esta necesidad de petición a una fuente externa de referencia. En la Tabla II se incluye la sobrecarga introducida en términos de tamaño de cada una de las

Tabla I
RETARDO POR EL CÁLCULO DE LAS DIMENSIONES DE CALIDAD.

	Retardo de petición (ms)	Retardo de procesado (ms)	Retardo total (ms)
Dim. Exactitud	185.1	0.004	185.1
Dim. Integridad	40.7	0.06	40.8
Dim. Puntualidad	12.2	0.3	12.5
Dim. Precisión	64.7	14.8	79.8

Tabla II
SOBRECARGA INTRODUCIDA POR LAS DIMENSIONES DE CALIDAD A LA ENTIDAD BÁSICA.

	Tamaño (bytes)	Sobrecarga (%)
Entidad básica	1205	–
Dim. Exactitud	134	11.1
Dim. Integridad	137	11.4
Dim. Puntualidad	140	11.6
Dim. Precisión	134	11.1

dimensiones de calidad en el momento en el que se integran como metadato a la entidad observada. Este valor de sobrecarga es prácticamente constante en todas las dimensiones, añadiendo alrededor de 136 bytes cada una de ellas, frente a los 1205 bytes de la entidad original.

A la vista de estos resultados, se puede concluir que la inclusión de los metadatos relacionados con DQ impone una sobrecarga no despreciable que podría considerarse un inconveniente para la solución propuesta. Sin embargo, a partir de este análisis, queda claro también que tanto el aumento de los requisitos de almacenamiento como el retraso en el procesamiento, debido a esta integración de las propiedades adicionales, no es un precio tan alto a pagar a cambio de comprender el verdadero significado de los datos disponibles.

El acceso a funcionalidades de valor añadido (i.e. información valiosa sobre la calidad de los datos) siempre conlleva ciertos compromisos. La solución propuesta ofrece un equilibrio entre funcionalidad y sobrecarga.

B. Evaluación de las técnicas basadas en IA

Tras la revisión realizada sobre las técnicas elegidas para incrementar la calidad de la información que se hizo en la Sección II, en este apartado se evalúan algunos de los algoritmos y métodos para llevar a cabo estas técnicas.

En primer lugar, es importante destacar cuál es el conjunto de datos utilizado para la evaluación de estas técnicas. Este conjunto se compone de los registros históricos de temperatura, desde el 01 de enero de 2021 hasta el 13 de junio de 2022, de los sensores pertenecientes a SmartSantander [31]. Como proyecto de Ciudad Inteligente, SmartSantander cuenta con un gran número de sensores que informan de manera periódica cada 1, 2 o 5 minutos, alcanzando un volumen de alrededor de 2 GB de datos, con aproximadamente 16 millones de observaciones, durante el periodo de tiempo mencionado. En línea con la premisa de ser un proyecto de *Smart City*, la calidad de los dispositivos desplegados alrededor de la ciudad de Santander no es elevada, lo que provoca inexactitudes en las observaciones e incluso recurrencia de valores absurdos. Además de estos problemas, también es importante considerar la degradación temporal que sufren los dispositivos, ya que llevan desplegados desde el inicio del proyecto en 2011. En consecuencia, es evidente que el conjunto de datos inicial debe ser procesado antes de la aplicación de las técnicas de mejora de la calidad. De acuerdo con esto, se han aplicado tres métodos relacionados con el conocimiento del entorno: eliminación de absurdos, correlación espacial y correlación temporal.

El primer método, la eliminación de absurdos, se basa en establecer rangos lógicos a los valores de las observaciones recogidas por los sensores. En este caso, al tratarse del fenómeno atmosférico de temperatura en la ciudad de Santander, se han consultado los registros de AEMET [30] en el mismo rango temporal que el conjunto de datos.

El segundo método se corresponde con la correlación espacial. Debido a que en el proyecto SmartSantander se colocaron algunos dispositivos en autobuses y taxis, se han registrado medidas fuera de Santander, como por ejemplo, en Bilbao o Madrid. Por lo tanto, al centrar el estudio de todo el conjunto de datos dentro del área geográfica de Santander, se han eliminado aquellas observaciones tomadas fuera de estas coordenadas.

El último método se centra en la correlación temporal. Para ello, se han agrupado las observaciones en una frecuencia periódica (horaria) y se han promediado sus valores. De este modo, se elimina la posibilidad de que se dupliquen las marcas temporales de las observaciones de distintos sensores y se reduce considerablemente el volumen de datos, sin llegar a perder información.

Una vez preprocesados los datos, el conjunto ya se encuentra listo para ser evaluado bajo las técnicas de enriquecimiento de la calidad propuestas en este trabajo.

Comenzando por la imputación de valores, el objetivo de la evaluación de esta técnica es comparar los diferentes métodos propuestos a través de errores absolutos (MAPE – *Mean Absolute Percentage Error* y MAE – *Mean Absolute Error*). Para ello, se ha dividido el conjunto de los datos en dos grupos: conjunto de entrenamiento y conjunto de validación. El primer grupo se compone de un subconjunto de datos que incluye huecos entre sus observaciones. Estas observaciones "perdidas" son las que componen el subconjunto de validación. Por tanto, los métodos de interpolación y el algoritmo kNN elegidos se aplican sobre el primer subconjunto de datos, obteniendo como resultado los valores de las observaciones ausentes. Tras ello, se

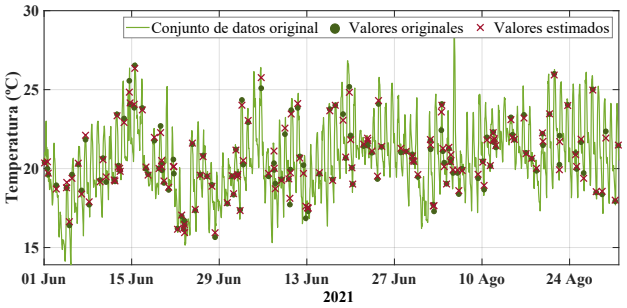

Fig. 2. Imputación de valores utilizando la interpolación polinomial de grado 2. Zoom temporal sobre un periodo de 3 meses.

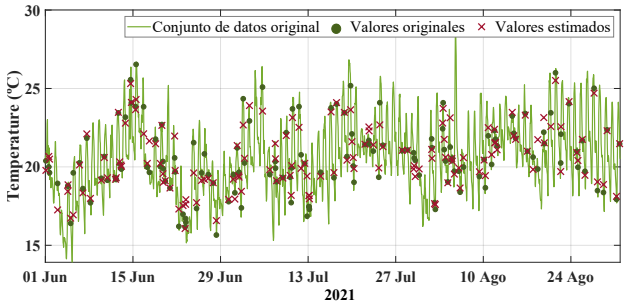

Fig. 3. Imputación de valores utilizando el algoritmo kNN con 5 vecinos. Zoom temporal sobre un periodo de 3 meses.

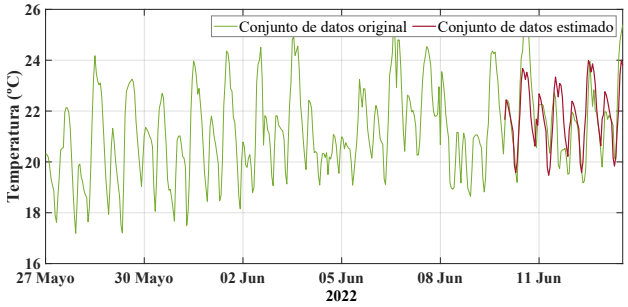

Fig. 4. Estimación del futuro inmediato con el algoritmo SARIMA (configuración (0,1,1)(2,1,0)[24]).

comparan estos valores imputados con los que se habían extraído en el segundo subconjunto de datos, validando y evaluando la exactitud de estas técnicas.

Con el objetivo de facilitar la comprensión sobre los resultados expuestos en las figuras, se han aplicado las técnicas sobre un periodo temporal de 3 meses (junio, julio y agosto de 2021). De esta forma, se obtienen los resultados mostrados en la Fig. 2 y la Fig. 3 de manera más visual, y los resultados numéricos mostrados en las dos primeras filas de la Tabla III de las técnicas aplicadas, interpolación polinomial de grado 2 y algoritmo kNN con 5 vecinos, respectivamente. Las figuras muestran la diferencia mínima en el rendimiento de ambas técnicas, representando los valores originales (conjunto de validación) con puntos verde oscuro y los valores estimados con cruces rojas (resultados de la aplicación de las técnicas). En cuanto a los datos de exactitud mostrados en la Tabla III, se observa que la interpolación polinomial de grado 2 obtiene mejores resultados que el algoritmo kNN, alcanzando valores de exactitud de 99.941% y 99.825% (1-MAPE), respectivamente. La diferencia es insignificante, concluyendo que ambas técnicas son buenas opciones para la imputación de valores.

La otra técnica de enriquecimiento mencionada es la ampliación sintética del conjunto de datos base con la intención de poder aplicar otras técnicas sobre los flujos de datos en tiempo real y así disponer de este rango temporal adicional como apoyo. Dada la naturaleza del conjunto de datos (series temporales con estacionalidad), se ha utilizado el algoritmo SARIMA. En este caso, debido a este aspecto de estacionalidad, se debe acortar el conjunto de datos (a 3 meses) para eliminar la periodicidad anual

Tabla III
EVALUACIÓN DE LAS TÉCNICAS DE IMPUTACIÓN DE VALORES Y ESTIMACIÓN DEL FUTURO INMEDIATO.

	Error porcentual absoluto medio (MAPE)	Error absoluto medio (MAE)
Interpolación pol. de grado 2	0.000592	0.0121
kNN con 5 vecinos	0.001740	0.03541
SARIMA (0,1,1)(2,1,0)[24]	0.040075	0.87221

Tabla IV
MEJORA TRAS LA APLICACIÓN DE LAS TÉCNICAS DE CURADO.

	Pre-procesado (mín-máx)	Post-procesado (mín-máx)	Ganancia (%)
Exactitud ($\pm^oC$)	$0 - 183.1$	$0.0 - 6.7$	95.078
Integridad (ppu)	$0.455 - 1$	$1 - 1$	0.864
Puntualidad (min)	$0.997 - 13.26$	$0.997 - 10.94$	1.124
Precisión ($\pm^oC$)	$0 - 193.07$	$0 - 4.49$	97.642

y mantener la diaria. Teniendo esto en cuenta, se aplica este algoritmo con la configuración más óptima de sus parámetros y el resultado se muestra en la Fig. 4. Además, en la última fila de la Tabla III se muestran las métricas utilizadas para la evaluación numérica del método. Se puede ver que se obtiene una exactitud del 96% (1-MAPE), lo que puede considerarse un buen rendimiento.

Por último, la Tabla IV muestra la mejora conseguida mediante la aplicación de las técnicas de curado ya descritas. Para ello, se han evaluado las dimensiones de calidad en cada uno de los 65 datasets (series temporales correspondientes a cada uno de los sensores utilizadas en la evaluación) y se han comparado los valores de éstas, antes y después de haber sido procesadas en el Módulo de Curado de Datos IoT implementado. Como se puede ver, se consiguen mejoras significativas en todas y cada una de las dimensiones evaluadas. Exactitud y Precisión son en las que se alcanzan mayores ganancias debido a la existencia de una gran cantidad de valores anómalos, los cuales han sido eliminados en el proceso de curado. Integridad y Puntualidad presentan valores de ganancia residuales, ya que dependen del número de observaciones perdidas (marginal en el flujo de datos que se ha empleado en la validación). Aún así, el efecto introducido por el proceso de curado es siempre de mejora. Los parámetros evaluados que se muestran en la Tabla IV son: para la Exactitud y la Precisión, la amplitud del rango de temperaturas (en unidades de variación de grados Celsius, ±ºC); para la Integridad, la cantidad de observaciones no perdidas (en partes por unidad, ppu); y para la Puntualidad, la tasa de actualización (en minutos).

V. CONCLUSIONES

Dado el incremento de la importancia de los datos en la actualidad, garantizar su calidad y ser capaces de comprender todas las dimensiones que tiene el ámbito de la calidad de la información (DQ), se convierten en condiciones ineludibles para cualquier plataforma de gestión de datos. En particular, en las plataformas IoT esto es especialmente importante debido a las particularidades de este tipo de

infraestructuras, que hacen que en ocasiones sea imposible imponer unos requisitos mínimos de DQ.

En este artículo se presenta el trabajo llevado a cabo para evaluar dimensiones específicas de DQ para flujos de datos IoT y compensar, mediante mecanismos de curado de datos habilitados por IA, los valores deficientes en estas dimensiones. La sobrecarga en términos de retardo y de tamaño, debido al cálculo y evaluación de las dimensiones de calidad, ha presentado valores adecuados dentro de un equilibrio entre funcionalidad y rendimiento dado el servicio de valor añadido presentado. En este sentido, se ha evaluado la mejora introducida por el módulo de curado que se ha implementado para este trabajo. Por otro lado, las técnicas de mejora y enriquecimiento de la calidad del conjunto de datos han demostrado un comportamiento apropiado, medido a través de las métricas MAPE y MAE, además de haber expuesto el presente beneficio tras su uso.

Los próximos pasos de esta investigación se basan en el estudio de la posible ampliación del número de dimensiones DQ evaluadas y su elección dinámica por parte del consumidor en la cadena de enriquecimiento (DET). Además, también se plantea investigar diferentes técnicas de enriquecimiento no contempladas hasta el momento e incorporarlas en esta cadena de curado de la información.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Programa CEF de la Comisión Europea a través del proyecto SALTED "Situation-Aware Linked heTerogeneous Enriched Data" bajo el Número de Acción 2020-EU-IA-0274 y por la Agencia Estatal de Investigación (AEI) mediante el proyecto THROTTLE "Mercado de Datos de Movilidad Urbana Confiable" bajo el Acuerdo de Subvención nº TED2021-131988B-I00 (en el marco del Plan de Recuperación, Transformación y Resiliencia) y el proyecto SITED "Semantically-enabled Interoperable Trustworthy Enriched Data-spaces" bajo el Acuerdo de Subvención nº PID2021-125725OB-I00.

REFERENCIAS

[1] European Commission, "A European strategy for data," 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

[2] P. Bellini, P. Nesi, and G. Pantaleo, "Iot-enabled smart cities: A review of concepts, frameworks and key technologies," *Applied Sciences*, vol. 12, no. 3, p. 1607, 2022.

[3] B. B. Sinha and R. Dhanalakshmi, "Recent advancements and challenges of internet of things in smart agriculture: A survey," *Future Generation Computer Systems*, vol. 126, pp. 169–184, 2022.

[4] Y. Yang, H. Wang, R. Jiang, X. Guo, J. Cheng, and Y. Chen, "A review of iot-enabled mobile healthcare: technologies, challenges, and future trends," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9478–9502, 2022.

[5] K. Garg, C. Goswami, R. Chhatrawat, S. K. Dhakar, and G. Kumar, "Internet of things in manufacturing: A review," *Materials Today: Proceedings*, vol. 51, pp. 286–288, 2022.

[6] P. Sotres, J. R. Santana, L. Sánchez, J. Lanza, and L. Muñoz, "Practical lessons from the deployment and management of a smart city internet-of-things infrastructure: The smartsantander testbed case," *IEEE Access*, vol. 5, pp. 14 309–14 322, 2017.

[7] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5—33, Mar. 1996.

[8] Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86—-95, Nov. 1996.

[9] T. C. Redman, *Data Quality for the Information Age*, 1st ed. Artech House, Inc., 1997.

[10] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 16, 7 2009.

[11] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*, 2018, pp. 166–173.

[12] L. Zhang, D. Jeong, and S. Lee, "Data Quality Management in the Internet of Things," *Sensors*, vol. 21, no. 17, p. 5834, Aug. 2021.

[13] S. Geisler, S. Weber, and C. Quix, "An Ontology-based Data Quality Framework for Data Stream Applications," in *ICIQ 2011 - 16th International Conference on Information Quality*, 01 2011.

[14] A. Klein and W. Lehner, "Representing Data Quality in Sensor Data Streaming Environments," *Journal of Data and Information Quality (JDIQ)*, vol. 1, no. 2, Sep. 2009.

[15] M. Gomez-Omella, B. Sierra, and S. Ferreiro, "On the Evaluation, Management and Improvement of Data Quality in Streaming Time Series," *IEEE Access*, vol. 10, pp. 81 458–81 475, 2022.

[16] C. Batini, D. Barone, M. Mastrella, A. Maurino, and C. Ruffini, "A Framework and a methodology for data quality assessment and monitoring," in *ICIQ 2007 - 12th International Conference on Information Quality*, Sep. 2007, pp. 333–346.

[17] F. de Haro Olmo, A. Valencia, A. Varela Vaca, and J. Alvarez-Bermejo, "Data Curation in the Internet of Things: a Decision Model approach," *Computational and Mathematical Methods*, vol. 3, Sep. 2021.

[18] L. Martín, L. Sánchez, J. Lanza, and P. Sotres, "Development and evaluation of artificial intelligence techniques for iot data quality assessment and curation," *Internet of Things*, vol. 22, p. 100779, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660523001026

[19] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016.

[20] "Technical Committee ISO/TC 69, "ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions"." [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en

[21] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: a systematic review," *Journal of Big Data*, vol. 7, no. 1, Dec. 2020.

[22] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–37,, Dec. 2021.

[23] N. Y. Yen, J.-W. Chang, J.-Y. Liao, and Y.-M. Yong, "Analysis of interpolation algorithms for the missing values in IoT time series: a case of air quality in Taiwan," *The Journal of Supercomputing*, vol. 76, no. 8, pp. 6475–6500, Aug. 2020.

[24] R. Adhikari and R. K. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," Feb. 2013.

[25] "Context Information Management (CIM) ETSI Industry Specification Group (ISG), "NGSI-LD API," 2021. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.04.01_60/gs_cim009v010401p.pdf

[26] Situation-Aware Linked heTerogeneous Enriched Data (SALTED), "D2.1: Report on Data Linking and Enrichment Architecture," Project Deliverable, 2022.

[27] "Smart Data Models – A global program led by FIWARE Foundation, TMForum, IUDX, and OASC." [Online]. Available: https://smartdatamodels.org/

[28] "DataQualityAssessment, Smart Data Model, GitHub Repository." [Online]. Available: https://github.com/smart-data-models/dataModel.DataQuality

[29] J. Von Neumann and S. Ulam, "Monte carlo method," *National Bureau of Standards Applied Mathematics Series*, vol. 12, no. 1951, p. 36, 1951.

[30] "Agencia Estatal de Meteorología - AEMET. Gobierno de España." [Online]. Available: https://www.aemet.es/es/portada

[31] L. Sanchez, L. Muñoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, and D. Pfisterer, "SmartSantander: IoT experimentation over a smart city testbed," *Computer Networks*, vol. 61, pp. 217–238, 2014, Special issue on Future Internet Testbeds -- Part I.