Abstract

Air pollution is now well known to be one of the major causes of human and climate health issues. The global crisis related to COVID-19 pandemic has brought to the fore such theme. The importance of air quality has been rediscovered and counted among the main positive effects of lockdown. The spread of low-cost electrochemical sensors, joined with diffusion of the Internet of Things (IoT) technologies, will allow in the near future, the birth of a generation of air quality monitoring networks, characterized by the integration of regulatory grade analyzers and such IoT smart electrochemical and particulate multisensory devices. The former will provide a backbone of sparse but high reliable, high quality, measurements at a significant procurement and operational costs, while smart multisensory devices will provide high resolution and possibly redundant measurements with affordable costs and with reduced precision and accuracy. Consequently, high-resolution pollution maps will be provided, constituting an advanced informative support tool for institutional decision makers.

However, this paradigm shift in air quality monitoring, is currently hampered by a series of problems concerning the low-cost sensors, high fabrication variance and the dynamic nonstationary nature of the working environment where these devices have to operate; but the primer concern is related to the measurements data quality.

Field calibration, relying on statistical or machine learning models more generally, seems the only viable and feasible method to guarantee the short-term accuracy and precision of these systems. Although its robustness to long term deployment and so different environmental and pollution composition is criticized. Field calibration allows to expose, rapidly and cheaply, the sensors grabbing their response to a variety of (uncontrollable) conditions that are similar to the ones that will be encountered during operational life, in opposition to laboratory-based calibration that would need significant time and human efforts to achieve similar variability in controlled settings.

Addressing the long-awaited achievement of data quality objective (DQO), in our opinion, could be a turning point for the rapid large-scale diffusion of this technology, especially in smart city applications.

With this objective in mind, the present PhD research has been focused in the first part, in the assessment of the machine learning techniques for the calibration of low-cost air quality monitoring systems (LCAQMSs), comparing multivariate linear regression and neural networks. The purpose of

this analysis was aimed at understanding whether a simpler technique is equally able to carry out acceptable performances in terms of data quality with respect to advanced but much more complex techniques. A mid-term experimental co-location campaign as well as a citizen science company have been performed for such kind of investigation, evidencing the effectiveness of the multivariate approach, both in fixed and mobile applications.

The extensive literature analysis executed has shown that most of the efforts of the scientific community operating in this research area was given to the inspection and assessment of the calibration models able to provide the best performances, while less emphasis is found looking for the answer to a simple question: *When does the sensor node need to be recalibrated?*

After the calibration phase, the LCAQMS will be subjected to performance degradation and forced to operate in conditions never seen before during the training phase. The outcomes will be bad quality measurements both in accuracy and precision. One of the phenomena that most influences this trend is the so-called Concept Drift. The awareness that the used model is no longer able to provide reliable data implies the risk to invalidate the model and to request a model update. Consequently, an original methodology based on the two-sample Kolmogorov–Smirnov test (TSKS test) is proposed to automatically detect the presence of the concept drift and a scheme of an add-on block based on the proposed approach is designed for the continuous monitoring of the metrological performance exhibited by the calibration model. As disposed by European directive, the relative expanded uncertainty (REU) is the paramount metric we will refer to. This functional block, in addition to monitoring the calibration model performance, is able to provide an alert to the user when a proper threshold is exceeded. Consequently, retraining or updating the calibration model ensuring compliance of the DQOs, is possible.

In the last part, different strategies have been analyzed to update the calibration model, trying to mitigate the effects of the concept drift in an air quality network operational scenario. Specifically, two alternative calibration models are taken into consideration: the general calibration model and the importance weighting calibration model. In some cases, both models have shown improvement of performances or matching those of the ad-hoc model, bringing the REU back to values in compliance with DQOs without requiring reference data. These models have also been used as the first layer in a stacking ensemble approach with the outcome of a further improving performance by requiring only the reference labels in the training process. The proposed approach guarantees the continuity of the data quality and extends the validity of the calibrations.

Sintesi

È ormai noto come l'inquinamento atmosferico sia una delle principali cause dei problemi sulla salute umana e dei cambiamenti climatici. La crisi globale legata alla pandemia COVID-19 ha portato alla ribalta questo tema. L'importanza della qualità dell'aria è stata riscoperta e annoverata tra i principali effetti positivi del lockdown. La diffusione di sensori elettrochimici a basso costo, unita alla diffusione delle tecnologie dell'Internet of Things (IoT), consentirà nel prossimo futuro la nascita di una generazione di reti di monitoraggio della qualità dell'aria, caratterizzate dall'integrazione di analizzatori di livello normativo e di dispositivi multisensoriali sia elettrochimici che per il particolato intelligenti. I primi forniranno una spina dorsale di misurazioni sparse, altamente affidabili e di alta qualità ma a costi significativi, mentre i dispositivi multisensoriali intelligenti (a basso costo) forniranno misurazioni ad alta risoluzione e possibilmente ridondanti, a costi accessibili ma con precisione e accuratezza ridotte. Di conseguenza, saranno fornite mappe dell'inquinamento ad alta risoluzione, che costituiranno uno strumento di supporto informativo avanzato per i decisori istituzionali.

Tuttavia, questo cambiamento di paradigma nel monitoraggio della qualità dell'aria è attualmente ostacolato da una serie di problemi riguardanti i sensori a basso costo ossia, l'elevata varianza di fabbricazione e la natura dinamica e non stazionaria dell'ambiente di lavoro in cui questi dispositivi devono operare; ma la preoccupazione principale è legata alla qualità dei dati di misurazione.

La calibrazione in campo, basata su modelli statistici o più in generale su modelli di machine learning, sembra l'unico metodo praticabile e fattibile per garantire l'accuratezza e la precisione a breve termine di questi sistemi. La calibrazione in campo allo stesso tempo viene anche criticata per la sua robustezza a lungo termine poiché il sistema si troverà ad operare successivamente in una diversa composizione dell'ambiente e dell'inquinamento nella sua vita operativa. Tale procedura però consente di esporre, in modo rapido ed economico, i sensori a una varietà di condizioni (incontrollabili) simili a quelle che si incontreranno durante la vita operativa, a differenza della calibrazione in laboratorio che richiederebbe tempi e sforzi umani significativi per ottenere una variabilità simile in ambienti controllati.

Il raggiungimento dei tanto attesi obiettivi di qualità dei dati, potrebbe rappresentare un punto di svolta per la rapida diffusione su larga scala di

questa tecnologia, soprattutto nelle applicazioni per le cosiddette “città intelligenti”.

La presente ricerca di dottorato si è concentrata su questo obiettivo: la qualità del dato, inteso come la misurazione rilasciata dal nodo sensoriale. Nella prima parte è stata affrontata la valutazione delle tecniche di apprendimento automatico per la calibrazione dei sistemi di monitoraggio della qualità dell'aria a basso costo, confrontando i modelli della regressione lineare multivariata e delle reti neurali. Lo scopo di questa analisi è stato quello di capire se una tecnica più semplice (come la regressione lineare multivariata) sia ugualmente in grado di realizzare prestazioni accettabili in termini di qualità dei dati rispetto a tecniche avanzate, ma molto più complesse come le reti neurali. Per questo tipo di indagine sono state realizzate una campagna sperimentale di co-locazione a medio termine e una di citizen science. Entrambe hanno dimostrato l'efficacia dell'approccio multivariato, sia in applicazioni fisse che mobili.

L'ampia analisi della letteratura scientifica sull’argomento ha evidenziato come la maggior parte degli sforzi della comunità scientifica operante in quest'area di ricerca sia stata dedicata alla ricerca e alla valutazione di modelli di calibrazione in grado di fornire le migliori prestazioni (alla ricerca del “modello ottimo”), mentre minore enfasi è stata posta nella ricerca della risposta a una semplice domanda: Quando è necessario ricalibrare il nodo sensore?

Dopo la fase di calibrazione, il nodo sarà soggetto a un degrado nelle prestazioni del modello poiché sarà costretto a operare in condizioni mai viste prima durante la fase di addestramento. Il risultato sarà una “cattiva” qualità delle misure, sia in termini di accuratezza che di precisione. Uno dei fenomeni che più influenzano questa tendenza è la cosiddetta *deriva concettuale* (o concept drift). La consapevolezza che il modello utilizzato non è più in grado di fornire dati affidabili comporta il rischio di invalidare il modello e di richiederne l'aggiornamento. Viene proposta una metodologia originale basata sul test statistico di Kolmogorov-Smirnov a due campioni per rilevare automaticamente la presenza della deriva concettuale e viene progettato uno schema di principio di un blocco aggiuntivo basato sull'approccio proposto, per il monitoraggio continuo delle prestazioni metrologiche esibite dal modello di calibrazione. Come previsto dalla direttiva europea, l'incertezza estesa relativa (REU) è la metrica principale a cui si fa riferimento. Inoltre, questo blocco funzionale, oltre a monitorare le prestazioni del modello di calibrazione, è in grado di fornire un avviso all'utente quando viene superata una soglia critica. Di conseguenza, è possibile riaddestrare o aggiornare il modello di calibrazione garantendo la conformità agli obiettivi di qualità dei dati definiti nella direttiva europea per le misurazioni indicative.

Nell'ultima parte, sono state analizzate diverse strategie per aggiornare il modello di calibrazione, cercando di mitigare gli effetti del concept drift in uno scenario operativo di rete di qualità dell'aria. In particolare, sono stati presi

in considerazione due modelli di calibrazione alternativi: il modello di calibrazione generale (general calibration model) e il modello di calibrazione con ponderazione dell'importanza (importance weighting calibration model). In alcuni casi, entrambi i modelli hanno dimostrato di migliorare le prestazioni o di eguagliare quelle del modello ad hoc, riportando l'incertezza estesa relativa a valori conformi agli obiettivi di qualità del dato senza bisogno di dati di riferimento. Questi modelli sono stati utilizzati anche come primo livello in un approccio che utilizza un modello di machine learning chiamato stacking ensemble. I risultati mostrano un ulteriore miglioramento delle prestazioni, richiedendo solo le etichette di riferimento per processo di addestramento. L'approccio proposto garantisce la continuità della qualità dei dati ed estende la validità delle calibrazioni.