The PANORAMA knowledge graph based prediction of outdoor air quality impact on health

Nareesa Karmali[1] , **Cedric Gil-Jardiné**[1] , **Gayo Diallo**[1]
[1]Team AHeaD, Bordeaux Population Health Inserm 1219, Univ. Bordeaux, F-33000, Bordeaux, France
{first, last}@u-bordeaux.fr

Abstract

Current air quality surveillance ecosystems cannot monitor air pollution levels across time and space continuously, and lack the ability to integrate historical data from heterogeneous sources into artificial intelligence algorithms capable of providing more insights on the impact of severe air pollution. Integrating in-situ sensor and satellite data has attracted recent attention, but the characteristics of graph-based learning with heterogeneous air quality data have not been explored in depth.

In this paper, we present the PANORAMA knowledge graph based prediction approach of the links between pollutants exposure and health outcomes. PANORAMA aims at integrating heterogeneous air quality and health data in order to better predict and explain their relationships. An Extract, Transform, and Load process is used to design and incorporate discrete data into the knowledge graph, and a knowledge graph embedding model is trained and tested to assess inferential capabilities of the graph. A use case with a dataset related to the Gironde department in France is showcased. The promising preliminary results of a 31.376 MR, 0.312 MRR, and 0.254, 0.335, and 0.524 Hits@ 1, 3, and 10 respectively on a TransE model suggest the potential of leveraging semantic structuring and knowledge graph technology for environmental public health.

1 Introduction

There has been a correlation between climate and air quality and the incidence of disease over the past few decades. Despite this, real-time information regarding air pollution is limited due to the fragmentation of data sources. According to the World Health Organization (WHO), air pollution causes over 7 million premature deaths annually, the majority of which occur in low- and middle-income countries[1]. By 2060, the number of deaths attributable to outdoor air pollution is expected to double, yet many countries lack public air quality information. Recent reports indicate that 141 countries do not regularly monitor particulate matter, and only 23 countries have more than three monitors per million citizens [Holloway *et al.*, 2021].

Gaseous pollutants and particulate matter contribute to the development and progression of disease[Kampa and Castanas, 2008]. Globally, ground monitors are the most popular source of air quality data, but their limited observation range means they may not accurately reflect the air quality where populations live and work [Ma *et al.*, 2019]. Due to its expansive spatial and geographic coverage and consistent data quality, satellite monitoring has gained popularity in the past decade. Satellite data cannot directly assess particulate matter concentration; aerosol optical depth data must be post-processed into particulate matter concentrations. IoT and other portable sensor options have emerged, enabling low-cost, wearable technologies to monitor real-time air quality. Integrating in-situ sensor and satellite data has become increasingly crucial for providing comprehensive air quality data for public health planning, policy development, and research.

There is gap regarding the automated prediction or forecasting of disease caused by air quality factors while integrating the various available data sources for air quality. This is particularly the case in the context of Low and Middle Income Countries (LMICs) where in-situ sensors are lacking [Raheja *et al.*, 2022]. Integrating information from multiple sources, such as ground and satellite sensors, results in a multimodal learning approach for air quality data analysis. Rather than relying solely on a single modality for understanding the effects of air pollutants, data integration enables a more comprehensive knowledge base and thorough analysis. This area has not been thoroughly explored.

In this paper, we described the PANORAMA approach (Public heAlth iNdicatOrs generation from Air quality data) for integrating heterogeneous air quality data and health related outcomes into a triple-based knowledge graph [Somé *et al.*, 2019; Some *et al.*, 2019; Hogan *et al.*, 2022] in order to enable predicting the relationship between outdoor air quality and diseases and later explaining this relationship. We report the design and development of the first version of the PANORAMA knowledge graph, as well as the preliminary results obtained by applying a missing links prediction process over the graph to detect evidenced association between pollutants and health outcomes (diseases). In contrast to previous studies, which employs traditional machine learning

[1]https://www.who.int/westernpacific/health-topics/air-pollution

techniques for such a task, we use symbolic AI and knowledge graph-based techniques.

2 Related Work

In this section, we highlight some relevant approaches on air quality monitoring and the prediction of the impact of poor air quality on health outcomes. They are categorized in two categories: traditional Machine Learning (ML) based approaches and graph based approaches.

2.1 Traditional ML approaches

Statistical and machine learning prediction methods have received increasing attention in recent years for the advantages they provide in describing complex linear and non-linear relationships [Wen *et al.*, 2019]. Prediction methods involving air quality have been categorized into two main classes: physical and data-driven methods. Physical-based approaches have been widely used for air quality modelling in the environmental community, and includes strategies such as chemical transport models, operational street canyon models, Gaussian plume models, and reduced-scale models. An advantage of physical models is their ability to provide insights about physical and chemical processes that govern atmospheric pollutants and thus explain the variability in the chemical and dynamical mechanisms of pollutants [Zhao *et al.*, 2019; Zhou *et al.*, 2019]. However, physical methods often require a great deal of empirical assumptions and expert-based adjustments to adapt to different regions and contexts. Multiple linear regression has been used quite extensively for understanding the relationships between air pollution observations, health, and other meteorological variables. This technique was built upon when creating land-use regression, which extends the linear regression model to make use of land characteristics, road use, traffic, and other physical environmental variables as predictors. This spatial interpolation modelling method involves the inclusion of Euclidean buffers around pollution monitoring stations, while conventional linear regression only reflects the coefficients within a study area [Beckerman *et al.*, 2013; Shaddick *et al.*, 2018]. With this method, land use information can be used to interpolate how pollution may be distributed at locations that do not have direct monitors [Beckerman *et al.*, 2013]. However, land-use regression has been criticized for its inflexibility and limitations when dealing with time-series or temporal data, predictor interaction, or collinearity [Beckerman *et al.*, 2013; Kerckhoffs *et al.*, 2019]. Following the use of multiple linear regression and Poisson regression for time-series data, generalized additive models (GAM) became popular in the late 1990s for dealing with complex and non-linear time-series data both widely and in the air quality field [Ma *et al.*, 2020]. The GAM is more flexible for seasonality and time trends, as well as nonlinear relationships, and defines the model using a non-parametric smoothing function [Khojasteh *et al.*, 2021].

The study in [Khojasteh *et al.*, 2021] used the Dickey-Fuller test to determine the long-term effects of air pollution on respiratory morbidity and mortality and developed nonlinear autoregressive and artificial neural network models to accurately predict it. They found that NO2, SO2, O3, and PM10 did not affect respiratory morbidity and mortality in the sensitivity analysis.

2.2 Graphs based approaches

There have been a variety of use cases of graph learning in healthcare. Patient records for instance, particularly electronic health records (EHRs), have been commonly represented as networks and used to personalize patient predictions for precision medicine [Li *et al.*, 2022]. However, the field of graph learning in environmental public health has not been thoroughly investigated, and the characteristics of graph representation of heterogeneous air quality sensor data are not fully understood. Due to the complex topographic structure and diverse types of interactions between nodes, graph learning can be challenging to develop and implement [Li *et al.*, 2022]. [Ferrer-Cid *et al.*, 2021] described the topology of an air pollution monitoring network using graph learning techniques. The primary nodes of air pollution monitoring networks are ground sensors, but the limited coverage of these sensors demonstrates the critical need for modeling techniques that can increase spatial resolution and utilize the knowledge of neighboring sensors to reconstruct the signal in regions without sensor data [Ferrer-Cid *et al.*, 2021]. The study in [Lin *et al.*, 2020] integrates theories of spatial statistics into neural networks for the prediction of geographically-based spatiotemporal phenomena on a fine spatial scale. Using observed air quality measurements from a network of low-cost sensors as well as contextual data describing environmental characteristics, the objective is to estimate air quality values at locations lacking sensors across a fine spatial grid.

The current study uses a knowledge graph, a heterogeneous graph that captures knowledge from different datasets, to integrate a variety of air quality and health outcome data [Li *et al.*, 2022]. Knowledge graphs are networks of data expressed through nodes and edges corresponding to important entities of information and the relations that link them together, respectively. Using the principles of Linked Data, knowledge graphs provide a tool for combining, exploring, and traversing data in context with other connected resources.

3 The PANORAMA approach

Our objective is to develop a versatile method that can integrate diverse data types and modalities to predict the effects of exposure to poor air quality. To achieve this, we aim to rely on readily available Open Data, especially in LMICs, where reliable statistics and monitoring tools could be scarce. Initially, we integrated satellite data on air quality monitoring, ground sensor data, geographical subdivision data, and hospital admissions data to validate the approach's fundamental principle. Our primary focus was on the Gironde department in the southwest of France, where we were able to retrieve the necessary multimodal data to apply the approach. By leveraging these data types, we aim to develop a generic method that can be used in the future to predict the effects of air pollution exposure.

Figure 2 exhibits the overall workflow of the PANORAMA approach, which consists of the steps described in the following sections.

3.1 Design of the knowledge model

A model was designed based on nodes of interest to be integrated regarding air quality data monitoring and the relationships between these nodes. Therefore, the Knowledge graph construction begins with modelling the nodes and relations that are set to be included in the graph. The knowledge graph is a graph-based information represented in the form of triplets, to represent entities and their relationships [Li *et al.*, 2020] that are a set of symbols. Nodes were selected based on discrete air quality data available for the Gironde department. The links and their corresponding labels were more particular, and thus chosen depending on the semantics available within each discrete dataset and the intersecting information between the datasets. Custom names for relations were assigned to links between the nodes. The high level view of the graph is depicted in figure 1.

3.2 Knowledge sources selection

The resulting conceptual model in the previous step was used to determine the information to be retrieved from the knowledge sources. Several knowledge sources were assessed for location, reliability and frequency of measurements, range of pollutants, and availability of health outcome data. Such evaluation of knowledge sources was necessary to understand how fragmented, unavailable, and unreliable data may affect the ability to integrate and produce downstream findings. While this is an expected challenge considering the data of interest, it is essential that the PANORAMA knowledge graph is developed and tested on consistent and valid data to ensure its applicability in settings where data may be more disparate, difficult to obtain, or unreliable. Two air quality data sources and one health outcome data source were selected for their frequency of measurements, spatial and temporal coverage, reliability, and ability to align location.

Copernicus Atmospheric Monitoring Service (CAMS)

The Copernicus Atmospheric Monitoring Service (CAMS) is the European Union's Earth Observation Programme, monitoring and forecasting European air quality and pollutants through satellite and non-satellite observations[2]. CAMS provides consistent and quality-controlled information through daily forecasts of air composition and forecasts of global long-range transport of pollutants. CAMS focuses on five main areas: Air quality and atmospheric composition, ozone layer and ultraviolet radiation, emissions and surface fluxes, solar radiation, and climate forcing. Free and openly accessible, CAMS provides data from two groups of missions: (i) the Sentinels, which have been developed specifically for CAMS and provided dedicated observations, and (ii) the Contributing Missions operated by National, European, or International organizations and siphon data to CAMS. For satellite data obtained from CAMS, which includes measurements with a date, value, the pollutant that is measured, and the geographical coordinates of the measurement, the following custom links were established: *hasValue, hasDate, hasPollutant, fromLocation*.

[2]Copernicus Services — Copernicus. Accessed April 25, 2023. https://www.copernicus.eu/en/copernicus- services

Atmo Nouvelle-Aquitaine pollutants data

Atmo Nouvelle-Aquitaine is the regional air observatory in the Nouvelle-Aquitaine region of France[3]. According to the law on air and rational use of energy, Atmo has several goals, including the monitoring of air 24 hours a day and forecasting the air quality in the region, predicting pollution episodes and altering the authorities. As of December 2020, Atmo Nouvelle-Aquitaine has a fixed sensor network of 44 pollution measurement stations and over 100 analyzers, which operation continuously days a week, 24 hours a day. For data obtained from Atmo service, which includes measurements with a unit, date, value, and name, the pollutant that is measured, and the location of the sensor, the following custom links were established: *hasValue, hasDate, hasUnit, hasPollutant, fromLocation*.

SAMU Urgences de France

SAMU Urgences de France (SUDF) is the French organization of prehospital emergency medicine and ensure appropriate medical attention, from which emergency calls receive the most appropriate response as soon as possible according to the urgency and severity. From the emergency call, a variety of services are available depending on the nature of the call. These options include medical advice, a private ambulance, a general practitioner, etc. There is one SAMU per French department, approximately one for every 500,000 inhabitants, and nearly 100 in total. SAMU data are organized at the level of entries. Specifically, entry-level data are collected for each contact with SAMU and include the entry identification number, date, main complaint, INSEE area code, age and sex of the individual, and possible diagnostic information including an International Classification of Diseases (ICD) code. For data obtained from SAMU Urgences de France, which includes entries with an identification number, date, main complaint, sex, age, possible diagnosis, and the location of the entry, the following custom links were established: *hasReason, hasDate, isSex, isAge, hasCondition, hasICD, hasDiagnosis, fromLocation*.

The National Institute of Statistics and Economic Studies

The National Institute of Statistics and Economic Studies (INSEE) collects, analyses and disseminates information on the French economy and society[4]. Created in 1946, INSEE is a Directorate-General of the Ministries for the Economy and for Finances and is the official body for the design, production and dissemination of official statistics. Pertaining to the present study, INSEE describes and analyzes regions and territories in France according to codes. Within France, INSEE provides codes for each region, department, arrondissement, canton, and municipality, and thus provides granular geographic and regional information across the country. For specificity, INSEE provides codes, descriptors, and the relationship between geographical levels. All data provided by

[3]Atmo Nouvelle-Aquitaine, l'observatoire régional de l'air. Atmo Nouvelle-Aquitaine. Published December 5, 2016. Accessed April 25, 2023. https://www.atmo-nouvelleaquitaine.org/article/atmo-nouvelle-aquitaine-lobservatoire-regional-de-lair

[4]Getting to know INSEE — Insee. Accessed April 25, 2023. https://www.insee.fr/en/information/2381925

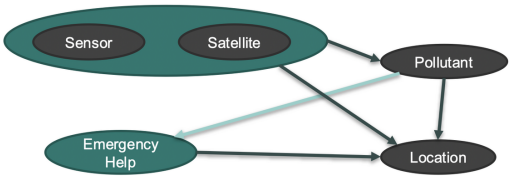

Figure 1: High level view of the conceptual model of the knowledge graph. The link to be predicted is the one between the health outcomes node (Emergency Help) and the Pollutants node.

INSEE is open-access and free. The department of Gironde, which is of interest in this study, is associated with code 33, and the codes for each arrondissement, canton, and municipality in this department begin with this code and add subsequent digits as the geographic level of interest becomes more granular.

International Classification of Diseases, Version 10

The WHO International Classification of Diseases (ICD) Version 10 Vocabulary was obtained from BioPortal, an online biomedical ontology and terminology repository [Noy *et al.*, 2009]. Medical and clinical terminology are described in the ICD as codes, and these codes are used as the main foundation for most disease and health records and statistics in medical care, research, and public health.

3.3 Extraction and transformation of knowledge

Information was extracted from the knowledge sources, cleaned, and made available as triples in Turtle format with PANORAMA identifiers and the name of each relationship between entities. CAMS data was delivered in netCDF format for all global geographical coordinates and required the initial step of conversion to tabular Comma-Separated Value (CSV) format prior to cleaning. Post-conversion, data was restricted to the geographical coordinates of Gironde, France according to Google coordinates. Data obtained from Atmo and CAMS were subject to the same preparation: (i) Dates between 01-01-2017 and 31-12-2019; (ii) Only pollutants of interest included: PM2.5, PM10, CO, NO2, O3, SO2. Data obtained from SUDF was similarly prepared: (i) Dates between 01-01-2017 and 31-12-2019; (ii) Only entries which include a possible diagnosis (ICD code). Then, post-cleaning was necessary in order to normalize the entities to ensure that links could be created between the disparate knowledge sources. Indeed, in air quality and health data, different terms or descriptions exist for the same entity. Entity normalization is required to map the terms, ensuring that linking entities are represented analogously between resources. In the PANORAMA knowledge graph, all nodes must be harmonized according to their location (which acts as a proxy) to create links between the selected knowledge sources. Entity normalization was achieved using INSEE geographical data, which describes different levels of regions in Gironde as codes. SUDF data was previously mapped to INSEE geographical codes prior to data being obtained. For instance, the Atmo Sensor Location $<< Bordeaux - Bastide >>$ was normalized to the INSEE Commune $<< Bordeaux >>$ with the INSEE Code "33063.0", while $<< Ambes2 >>$ was normalized to "Ambes" with INSEE Code $<< 33528.0 >>$. CAMS data described location through geographical coordinates to the tenth place. Location was mapped to INSEE codes accordingly. For instance, CAMS Coordinate $<< 44.8"N, -0,6"W >>$ was mapped to the INSEE Commune $<< Bordeaux >>$ with the code $<< 33063.0 >>$.

3.4 Loading into the triplestore

RDF statements are comprised of a three-part structure linking resources, which can be identified with a URI, together depicting two entities and the relation between them. This semantic data is represented as subject-predicate-object triples for implementation into a knowledge graph. A set of triples makes an RDF graph, and in a graph, nodes represent entities and edges represents relationships between entities. RDF triples are well suited for representing highly interconnected data. The set of PANORAMA triples are loaded and interconnected in a knowledge graph database. In the present study, 11 custom relations are defined in this study: hasValue, hasDate, hasUnit, hasPollutant, fromLocation, hasReason, isSex, isAge, hasCondition, hasICD, hasDiagnosis.

Triples were extracted from unstructured tabular data using RDFLib 6.1.1, a Python package for working with RDF. For each data source, a tabular data frame was loaded and parsed with RDFLib. A namespace of 'http://panorama.org/' was set. A namespace is a mapping that connects URIs to a set prefix. For instance, a URI concept was created for each measurement accordingly for Atmo Nouvelle-Aquitaine and CAMS data: $measureConcept = URIRef("http://panorama.org/Measure")$. Finally, all the triples have been loaded into a semantic graph database, Blazegraph[5]. The latter is an open source Triplestore and graph database for storing and querying linked data and support RDF and SPARQL Protocol and RDF Query Language (SPARQL).

3.5 Predicting the links between pollutants and health outcomes

As already mentioned, we turn the identification of links between pollutants and health outcomes to a knowledge graph completion problem. Generally, knowledge graphs are incomplete sets with many missing facts. Thus, Knowledge Graph Completion aims to fill in missing triples to support graph completeness and improve prediction and the performance of other downstream applications, such as link prediction [Guan *et al.*, 2018]. Knowledge graph embeddings are

[5]https://blazegraph.com/

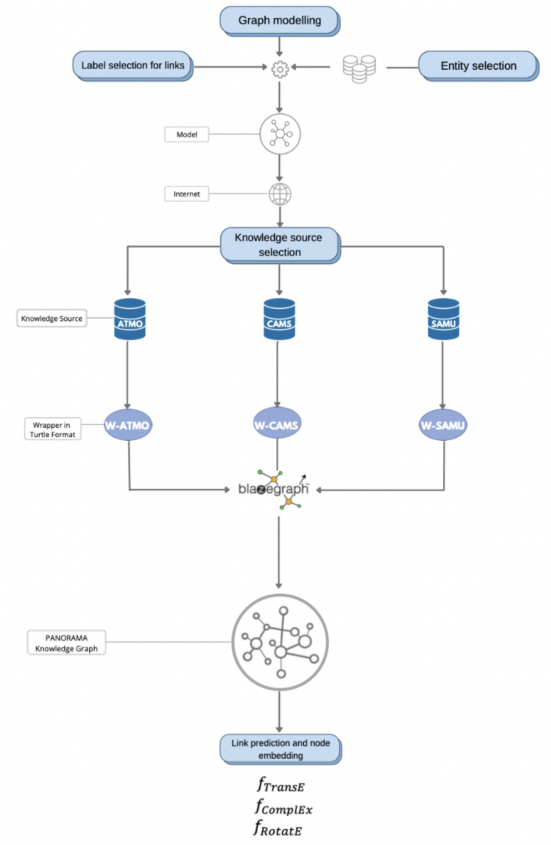

Figure 2: Overall workflow of PANORAMA. Nodes are selected based on interest for integration and connected with edges to form the conceptual knowledge graph. After loading in a triplestore, the embedding model is developed and link prediction is conducted to predict the possible association between pollutants and health outcomes.

supervised learning models that learn vector representations of nodes and edges of labeled, directed multigraphs. The aim is to map and project entities and relations into a low- dimensional and continuous vector space, on which other operations can be conducted. In our case, a Knowledge Graph Embedding Model (KGEM) was trained and evaluated on the PANORAMA knowledge graph using several model algorithms and a set of hyperparameters (see figure 4. A training knowledge graph is subjected to the architecture of a model with a couple layers: (i) lookup layer (simply assigns an embedding to each node and edge type of the graph); (ii) scoring layer, interacting with a loss function; (iii) negatives generation process; (iv) optimization (to train embeddings, which are used by the above scoring function to predict new links, the downstream task).

In this study, we have tested three different embeddings models, respectively Translating Embeddings (TransE)[Bordes *et al.*, 2013], RotatE[Huang *et al.*, 2021] and ComplEx[Trouillon *et al.*, 2016] which is a semantic matching model that serves as an alternative to translational distance models. RotatE is another translational distance model inspired by TransE, but instead switches to complex vector

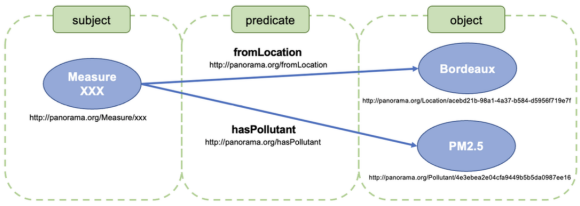

Figure 3: Examples of triples extracted in the study. A measure node, acquired from tabular Atmo or CAMS data, is assigned to the http://panorama.org/ namespace.

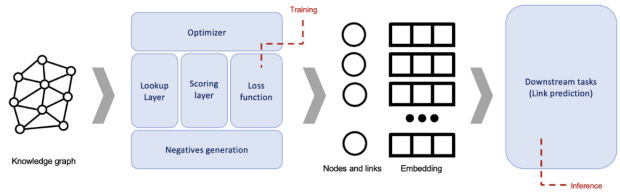

Figure 4: A KGEM comprises several layers: a lookup layer, a scoring layer, a negatives generation process, and optimization. Embedding of nodes and links of the original knowledge graph allows for downstream tasks such as link prediction.

space C to define relations as rotational from head to tail of a triple.

4 Results

4.1 The PANORAMA Knowledge graph

The graph is depicted with all conceptual nodes and edges in figure 6. It contains 1,922,998 triples, consisting of 443,332 entities and 13 relation types (Table 1). The graph contains entities from CAMS, Atmo, SUDF, ICD10, and INSEE. There are two pre-defined relations in the graph acquired from RDF Vocabulary: $rdf : type$, which is an instance of $rdf : Property$ that is used to declare that a resource is an instance of a class, and $rdfs : label$, which is another instance of $rdf : Property$ that provides a label or readable version of a resource's name. An example of Turtle format rendering for a pollutant measure is depicted in figure 5.

There are a total of 2,145 unique ICD10 codes included in the graph, identified from SUDF entries and 6 pollutants have been included: PM2.5, PM10, NO_2, O_3, SO_2 and CO. The different measurements were acquired from both Atmo and CAMS data, but CO was only available in CAMS data. There are a total of 6 locations included in the knowledge graph. They are assigned to the following INSEE communes: "Le Temple", "Ambes", "Merignac", "Bordeaux", "Bassens", and "Talence". These communes are assigned to INSEE URIs

Triples	$\|E\|$	$\|R\|$	Training	Testing
1,922,998	443,332	13	1,634,548	288,450

Table 1: The number of triples, entities, and relations along with the training and testing split of triples for the first PANORAMA knowledge graph.

```
<http://panorama.org/Measure/0007d0e2-62d9-4b54-a45b-7fedae671fe6> a pano:Measure ;
    pano:fromLocation <http://panorama.org/Location/aa63f8a3-49a2-41e3-a039-36dc639be4a1> ;
    pano:hasDate "2018-02-07" ;
    pano:hasPollutant <http://panorama.org/Pollutant/f75aa1bd232b495c19fa9ab7b3bf2f73> ;
    pano:hasValue "2.374648e-08" .
```

Figure 5: Triples from the CAMS dataset. Relations, in green, are assigned to nodes according to the measurement of interest

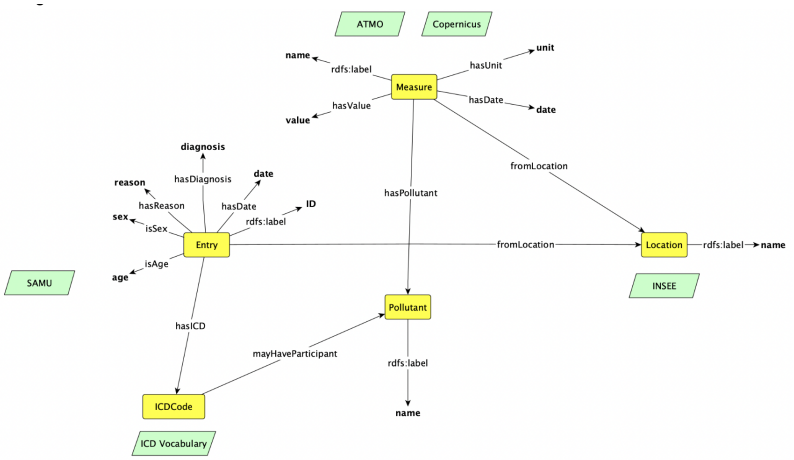

Figure 6: The conceptual knowledge graph obtained after the integration of the different knowledge sources. The green banners indicate the knowledge sources. Nodes are indicated with yellow squares and Literal nodes are indicated with bolded text. Links are indicated with directed black arrows and are labelled according to the relation.

identified from the INSEE RDF space according to the mapping exemplified earlier.

4.2 Predicting the link between pollutants and diseases

Link prediction is the main downstream task used to evaluate knowledge graph embedding models. There are two main types of link prediction that can be conducted: (i) Transductive link prediction, in which all links that are predicted are relations already seen in the training graph; (ii) Inductive link prediction, in which the goal is to predict links between unseen entities in the training graph. For the PANORAMA knowledge graph, transductive link prediction has been used to predict the relation of mayHaveParticipant, which is already seen in the training graph between ICD10 and Pollutant entities. In this work, the PyKEEN framework[6] was used to perform the analysis over the knowledge graph. PyKEEN is a Python package for reproducible, facile knowledge graph embeddings.

There exist 2,150 triples in the dataset including the mayHaveParticipant link between ICD10 and Pollutant entities, which serves as the future link of interest in the KGEM. For a proof of concept, the KGEM was trained on the following hyperparameters: model algorithms TransE, ComplEx, RotatE, epoch size 2, 5, embedding size 50, batch size 128. Similar to previous knowledge graph work, test triples are ranked against all candidate triples not appearing in the training set. The knowledge graph dataset was divided into an 85-15 split for training and testing, respectively. This evaluation strategy employs a learning to rank problems, in which we can see how well each positive triple in a test set ranks against synthetic negatives, built under the same procedure used in the training. This uses information retrieval metrics such as Mean Rank (MR), Mean Reciprocal Rank (MRR), and Hits@K respectively. MR computes the arithmetic mean over all individual ranks. MRR takes the reciprocal of the rank to guarantee robustness against outliers. It is more affected by changes of low ranked triples than high ranked triples. Hits@K calculates how many positive triples in a test set show up in the top k position against synthetic negative triples. This measure describes the fraction of positive facts that appear in the first k triples of the sorted rank list, and thus does not differentiate between triples that rank higher than k. Table 2 exhibits the model losses and the three main metrics of evaluation for the KGEM: MR, MRR, and Hits@K at three levels – 1, 3, and 10 hits. From this table, we can see that the TransE scoring function performs better than other model algorithms. An epoch size of 5 also outperformed the competing epoch size of 2 for all metrics in this model. The TransE model with 5 epochs achieved the best results with a mean rank of 31.376, mean reciprocal rank of 0.312, and 0.254, 0.335, and 0.524 Hits@1, 3, and 10, respectively.

5 Discussion and Perspectives

5.1 Contributions of the study

The present study demonstrated the ability to apply knowledge graph technology, which is extremely novel in public health, to sparse air quality data sources and health data. The PANORAMA knowledge graph provides data coherency and a holistic presentation of several knowledge sources. Further, the graph is able to fill gaps that may exist in discrete

[6] https://pykeen.github.io/

	Epochs	Losses	MR	MRR	Hit@ 10	Hit@ 3	Hit@ 1
TransE	2	0.00437,0.00206	928.582	0.142	0.217	0.182	0.158
	5	0.00138,0.00106	**31.376**	**0.312**	**0.524**	**0.335**	**0.254**
ComplEx	2	0.01091,0.00580	170058.469	0.0197	0.078	0.0026	0.0038
	5	0.00580,0.00576	165203.078	0.0249	0.0182	0.0078	0.00396
RotatE	2	0.00782,0.00781	157281.492	0.0358	0.00468	0.00156	0.00156
	5	0.00614,0.00568	26690.462	0.0313	0.000377	0.000223	0.000108

Table 2: Metrics of the knowledge graph embedding model with the following parameters: Scoring functions TransE, ComplEx, RotatE, epoch size 2, 5, embedding size 50, and batch size 128. The TransE epoch 5 model produced optimal results for the embedding (highlighted) – a MR of 31.376, MRR of 0.312, and Hits@1, 3, and 10 of 0.254, 0.335, and 0.524 respectively.

knowledge sources, such as missing measurements on particular dates or in particular locations, that may be achieved with the integration of other air quality data. This graph can recognize ambiguous entities in context, connect entities to similar data sources, leverage related information in a dataset, and provides a foundation for use in other settings and contexts. Moreover, promising results of the embedding model demonstrate inferential reasoning abilities about health from air quality and pollution. A mean rank of 31.376 and mean reciprocal rank of 0.312 for an initial embedding model using a TransE scoring function show that future renditions of the model will likely be able to produce insightful prediction results. A Hits@10 result of 0.524 indicates that there is a proportion of 0.524 of correct positive facts (plausible association between pollutants and health outcomes) appearing in the top 10 entries of a sorted ranked list again synthetic negatives in the model, demonstrating further aptitude of this knowledge graph's embeddings and link prediction. These hits indicate that over half of results in a sorted ranked list are relevant, and as the model matures with the fine-tuning of hyperparameters, inferential reasoning is probable.

5.2 Limitations of the study

The study presents some limitations and challenges. Firstly, the ability to create a semantic knowledge graph from a variety of heterogeneous sources is entirely reliant on the ability to align and harmonize the discrete data. In this study, it was possible to merge the data sources on location (which acts as a proxy), simply due to the availability of these different sources in the same region. However, this will be a continued challenge in applying this knowledge graph without adaptation in other settings, as the presence of harmonizing variables is fully dependent on data availability in the context of interest. In a global health setting, in which the PANORAMA knowledge graph may be of important use, such harmonizing variables may not be always present. In the context of LMICs, relying on Open Data could be a relevant alternative for prediction tasks at the macro level. A major challenge with using complex and heterogeneous data is the varying granularity (particularly spatial coverage of satellite Vs in-situ sensors based data) and frequency of data variables, particularly those which will be used for harmonization. When only a limited number of measurement locations are available, or when the measurement locations are sparse and not spatially representative of the target region, taking into account only neighborhood-level spatial or temporal dependencies may result in overfitting [Lin *et al.*, 2020]. Air quality data tends to be available at different specificities, and in the current study, proved to be challenging for integrating Atmo, CAMS, and SUDF data. CAMS data is geolocated according to geographical coordinates, but only to the tenths place for each latitude and longitude value. When examining a small region such as Gironde for application of PANORAMA, it proved difficult to align such high-level coordinates from CAMS with Atmo in-situ sensors, which record data at the sensor's precise location. Thus, it is expected that there will be some variation in the locations, which may impede the process of harmonization. This challenge will likely disappear when applied to larger areas or with data sources that provide more precise location records, such as GPS coordinates. However, it is unrealistic to assume that such perfect data will be obtained for other applications; this represents a limitation to be considered moving forward. Lastly, ambiguity among data sources for identical variable records presents a challenge. This study was fortunate to have access to health data containing symptom and disease identifiers based on the internationally recognized ICD vocabulary. However, health data is indicated differently across the globe, and this lexical gap may make harmonization difficult.

5.3 Future works

The PANORAMA knowledge graph was initially created with the primary purpose of being used in global health contexts where air quality data is sparse and heterogeneous. However, the graph's potential extends beyond this initial application, and there is an interest in testing its capabilities in other settings such as the Global South. Doing so will provide insights into global health contexts and help us understand the graph's limitations with various data sources. One way to enhance the graph's potential is by integrating data collected by citizen scientists or the Internet of Things (IoT) using low-cost sensors. It will be relevant to investigate similarly to [Lin *et al.*, 2020] how to exploit both local autocorrelation patterns and global autocorrelation trends of predictions and labeled data in the learned spatio-temporal embedding space for accurate fine-scale prediction tasks with limited labeled data. Additionally, it would be interesting to explore the dynamic modality of measures of exposure as discussed in Letellier et al. [Letellier *et al.*, 2022], while utilizing geolocated metadata provided by mobile operators' Call Data Records.

References

[Beckerman *et al.*, 2013] Bernardo S. Beckerman, Michael Jerrett, Randall V. Martin, Aaron Van Donkelaar, Zev Ross, and Richard T. Burnett. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. *Atmospheric Environment*, 77:172–177, October 2013.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usinier, Alberto Garcia-Dur\'{a}n, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 2787–2795. Curran Associates Inc., 2013.

[Ferrer-Cid *et al.*, 2021] Pau Ferrer-Cid, Jose M. Barcelo-Ordinas, and Jorge Garcia-Vidal. Graph Learning Techniques Using Structured Data for IoT Air Pollution Monitoring Platforms. *IEEE Internet of Things Journal*, 8(17):13652–13663, September 2021.

[Guan *et al.*, 2018] Saiping Guan, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. Shared Embedding Based Neural Networks for Knowledge Graph Completion. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 247–256, Torino Italy, October 2018. ACM.

[Hogan *et al.*, 2022] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs. *ACM Computing Surveys*, 54(4):1–37, May 2022.

[Holloway *et al.*, 2021] Tracey Holloway, Daegan Miller, Susan Anenberg, Minghui Diao, Bryan Duncan, Arlene M. Fiore, Daven K. Henze, Jeremy Hess, Patrick L. Kinney, Yang Liu, Jessica L. Neu, Susan M. O'Neill, M. Talat Odman, R. Bradley Pierce, Armistead G. Russell, Daniel Tong, J. Jason West, and Mark A. Zondlo. Satellite Monitoring for Air Quality and Health. *Annual Review of Biomedical Data Science*, 4(1):417–447, July 2021.

[Huang *et al.*, 2021] Xuqian Huang, Jiuyang Tang, Zhen Tan, Weixin Zeng, Ji Wang, and Xiang Zhao. Knowledge graph embedding by relational and entity rotation. *Knowledge-Based Systems*, 229:107310, October 2021.

[Kampa and Castanas, 2008] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367, January 2008.

[Kerckhoffs *et al.*, 2019] Jules Kerckhoffs, Gerard Hoek, Lützen Portengen, Bert Brunekreef, and Roel C. H. Vermeulen. Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environmental Science & Technology*, 53(3):1413–1421, February 2019.

[Khojasteh *et al.*, 2021] Davood Namdar Khojasteh, Gholamreza Goudarzi, Ruhollah Taghizadeh-Mehrjardi, Akwasi Bonsu Asumadu-Sakyi, and Masoud Fehresti-Sani. Long-term effects of outdoor air pollution on mortality and morbidity–prediction using nonlinear autoregressive and artificial neural networks models. *Atmospheric Pollution Research*, 12(2):46–56, February 2021.

[Letellier *et al.*, 2022] Noémie Letellier, Steven Zamora, Chad Spoon, Jiue-An Yang, Marion Mortamais, Gabriel Carrasco Escobar, Dorothy D. Sears, Marta M. Jankowska, and Tarik Benmarhnia. Air pollution and metabolic disorders: Dynamic versus static measures of exposure among Hispanics/Latinos and non-Hispanics. *Environmental Research*, 209:112846, June 2022.

[Li *et al.*, 2020] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, and Yuting Liu. Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine*, 103:101817, March 2020.

[Li *et al.*, 2022] Michelle M. Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369, October 2022.

[Lin *et al.*, 2020] Yijun Lin, Yao-Yi Chiang, Meredith Franklin, Sandrah P. Eckel, and Jose Luis Ambite. Building Autocorrelation-Aware Representations for Fine-Scale Spatiotemporal Prediction. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 352–361, Sorrento, Italy, November 2020. IEEE.

[Ma *et al.*, 2019] Jun Ma, Yuexiong Ding, Jack C.P. Cheng, Feifeng Jiang, and Zhiwei Wan. A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5. *Journal of Cleaner Production*, 237:117729, November 2019.

[Ma *et al.*, 2020] Ling Ma, Cheng-An Zhao, Tingxian Wu, Yingying Guo, and Ziru Zhao. The Effects of AIr Quality and Weather Conditions on Weather Sensitive Diseases. In *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, pages 60–65, Guiyang, China, April 2020. IEEE.

[Noy *et al.*, 2009] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web Server issue):W170–173, July 2009.

[Raheja *et al.*, 2022] Garima Raheja, Kokou Sabi, Hèzouwè Sonla, Eric Kokou Gbedjangni, Celeste M. McFarlane, Collins Gameli Hodoli, and Daniel M. Westervelt. A Network of Field-Calibrated Low-Cost Sensor Measurements of PM $_{2.5}$ in Lomé, Togo, Over One to Two Years. *ACS Earth and Space Chemistry*, 6(4):1011–1021, April 2022.

[Shaddick *et al.*, 2018] Gavin Shaddick, Matthew L. Thomas, Heresh Amini, David Broday, Aaron Cohen, Joseph Frostad, Amelia Green, Sophie Gumy, Yang Liu, Randall V. Martin, Annette Pruss-Ustun, Daniel

Simpson, Aaron van Donkelaar, and Michael Brauer. Data Integration for the Assessment of Population Exposure to Ambient Air Pollution for Global Burden of Disease Assessment. *Environmental Science & Technology*, 52(16):9069–9078, August 2018.

[Some *et al.*, 2019] Borlli Michel Jonas Some, Georgeta Bordea, Frantz Thiessard, Stefan Schulz, and Gayo Diallo. Design Considerations for a Knowledge Graph: The WATRIMed Use Case. *Studies in Health Technology and Informatics*, 259:59–64, 2019.

[Somé *et al.*, 2019] Borlli Michel Jonas Somé, Georgeta Bordea, Frantz Thiessard, and Gayo Diallo. Enabling West African Herbal-Based Traditional Medicine Digitizing: The WATRIMed Knowledge Graph. *Studies in Health Technology and Informatics*, 264:1548–1549, August 2019.

[Trouillon *et al.*, 2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. In *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, pages 2071–2080, 2016.

[Wen *et al.*, 2019] Congcong Wen, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu, and Tianhe Chi. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of The Total Environment*, 654:1091–1099, March 2019.

[Zhao *et al.*, 2019] Jiachen Zhao, Fang Deng, Yeyun Cai, and Jie Chen. Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere*, 220:486–492, April 2019.

[Zhou *et al.*, 2019] Yanlai Zhou, Fi-John Chang, Li-Chiu Chang, I-Feng Kao, and Yi-Shin Wang. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of Cleaner Production*, 209:134–145, February 2019.