Vol. 1, No. 2, pp.: 47-57, 2023

Urban Air-Quality Knowledge Graphs: Database-Driven Policy Analytics for Pollution Intervention Planning

Jian Li[1], Mei Wang[2], Xiaofeng Zhang[3,*]

[1] School of Environmental Science, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] College of Geography and Environmental Science, Henan University, Kaifeng 475004, China

[3] School of Environment and Resource Sciences, Shanxi University, Taiyuan 030006, China

* Corresponding author: xfzhang@sxu.edu.cn

Article Information	
Received	18 January 2023
Accepted	29 May 2023
DOI	https://doi.org/10.63646/datamind.2023.010205

Abstract

Urban air pollution poses one of the most complex policy challenges in environmental governance: the causal pathways linking emission sources to observed concentrations, and interventions to measured outcomes, are inherently heterogeneous across cities, sectors, and seasons. Conventional tabular air-quality databases store pollutant measurements and regulatory records in isolation, preventing the structured linking of pollution events with emission inventories, meteorological context, policy timelines, and health outcomes that effective intervention planning requires. This paper presents AirKGDB, an open urban air-quality knowledge graph database that formalises these multi-domain linkages through a seven-entity ontological schema grounded in W3C Semantic Sensor Network (SSN) and PROV-O provenance standards. AirKGDB integrates five heterogeneous source streams — national ground monitoring networks, MODIS and Sentinel-5P satellite retrievals, traffic and industrial emission inventories, weather reanalysis, and structured policy corpora — for 288 Chinese prefecture-level cities over the period 2014–2022. The database comprises 18.6 million pollution event nodes, 4.3 million policy-intervention nodes, and 214 million typed edges linking events to sources, interventions, weather contexts, and health outcomes. We conduct three reproducible experiments: (1) a difference-in-differences (DiD) evaluation of the 2013 Air Pollution Prevention and Control Action Plan across ten major cities, finding mean PM2.5 reductions of 22.4 μg/m³ (95% CI: [17.8, 27.0]); (2) a regional heterogeneity decomposition attributing PM2.5 reductions to industrial control, traffic restriction, and residential heating interventions; and (3) a policy case-retrieval experiment achieving NDCG@5 = 0.80 and Precision@1 = 0.82 for similar-context policy recommendation. AirKGDB, its construction pipeline, SPARQL/Cypher query templates, and evaluation scripts are released under CC-BY 4.0 to support reproducible environmental AI research.

Keywords: ***air quality knowledge graph; PM2.5; pollution intervention; policy analytics; spatiotemporal database; causal inference; graph database; environmental AI***

1. Introduction

Urban air pollution remains one of the foremost threats to public health and environmental sustainability worldwide. Fine

particulate matter (PM2.5) is estimated to cause approximately 4.1 million premature deaths annually, with the largest burden concentrated in densely industrialised and urbanised regions of South and East Asia (Lelieveld et al., 2015; Burnett et al., 2018; Cohen et al., 2017). China, which implemented one of the most ambitious multi-year air-quality action plans in history beginning in 2013, has produced a uniquely rich natural experiment: a well-documented policy intervention applied heterogeneously across a vast, geographically and economically diverse national space, with continuous monitoring records before and after implementation (Zhang et al., 2019; Guan et al., 2014). Yet despite the wealth of monitoring data, emissions inventories, and policy documentation that exists, these data streams remain structurally disconnected. Pollution measurements are stored in tabular monitoring databases with no formal links to the regulatory actions, meteorological contexts, or emission source profiles that determine their causal interpretation.

This structural disconnection has a direct consequence for AI-driven policy analytics: models trained on isolated pollutant concentration series cannot attribute observed changes to specific interventions, cannot retrieve analogous policy cases from other cities and periods, and cannot support prospective planning that requires reasoning across the pollution-source-policy-outcome causal chain (Pearl, 2009; Imbens and Rubin, 2015). Knowledge graphs (KGs) offer a principled solution by representing entities, their properties, and their relationships in a unified graph structure that supports both structured querying and machine learning inference (Hogan et al., 2021; Nickel et al., 2016). Applied to urban air quality, a KG can link a PM2.5 pollution event node to its monitoring station, the concurrent weather conditions, the active policy instruments in the same region, the probable emission sources, and the associated health burden, enabling analysts to traverse causal and contextual relationships that flat databases cannot represent.

This paper introduces AirKGDB, an urban air-quality knowledge graph database designed specifically to support pollution intervention planning and AI-ready policy analytics. The database makes four contributions. First, we design a seven-entity ontological schema grounded in established semantic web standards (SSN, PROV-O, GeoSPARQL) that models the complete pollution-source-policy-outcome causal chain. Second, we implement a reproducible ETL pipeline that ingests five heterogeneous source streams and constructs the knowledge graph at city, station, and event granularity. Third, we report comprehensive data quality metrics and conduct three reproducible benchmark experiments covering causal effect estimation, regional heterogeneity decomposition, and policy case retrieval. Fourth, we release the complete database, pipeline code, query templates, and evaluation scripts under CC-BY 4.0. The remainder of this paper is structured as follows: Section 2 discusses the database gap and use cases; Section 3 describes data sources and schema; Section 4 presents the construction pipeline; Section 5 reports experiments; Section 6 covers reproducibility; Section 7 identifies limitations; and Section 8 concludes.

2. Database Gap and Use Cases

Existing air-quality data infrastructures fall into three broad categories, each with characteristic gaps. The first category comprises monitoring networks: national networks such as the China National Environmental Monitoring Centre (CNEMC) system and the United States EPA AQS network provide high-temporal-resolution concentration measurements but contain no structured links to emission sources, policy timelines, or health outcomes. The second category comprises emission inventories: gridded inventories such as the Multi-resolution Emission Inventory for China (MEIC) and the global EDGAR database quantify source contributions by sector and geography but do not track the policy instruments that caused changes in those emissions. The third category comprises epidemiological and health burden databases such as the Global Burden of Disease study, which estimate health impacts of pollution exposure without linking back to specific policy interventions or source configurations (Di et al., 2017; van Donkelaar et al., 2016; Pope and Dockery, 2006).

Table 1. Comparison of AirKGDB with representative existing air-quality data resources.

Resource	Type	Temporal Resolution	Emission Links	Policy Links	Graph Structure	Open Access
CNEMC (China)	Monitoring network	Hourly	No	No	No	Partial
EPA AQS (US)	Monitoring network	Hourly/Daily	No	No	No	Yes
MEIC (China)	Emission inventory	Annual	Yes	No	No	Registered

Resource	Type	Temporal Resolution	Emission Links	Policy Links	Graph Structure	Open Access
EDGAR (Global)	Emission inventory	Annual	Yes	No	No	Yes
GBD AQ module	Health burden DB	Annual	No	No	No	Yes
AirKGDB (ours)	Knowledge graph	Hourly–Annual	Yes	Yes	Yes (7-entity)	Yes (CC-BY)

Table 1 illustrates that no existing resource simultaneously provides monitoring-level temporal resolution, structured emission source linkage, formal policy timeline encoding, and graph-structured interconnection. AirKGDB is designed to fill this gap. Three high-value use cases motivate its design. First, causal intervention analysis: by linking monitoring observations to concurrent policy activation records via graph edges, analysts can apply difference-in-differences, synthetic control, and causal graph methods to estimate the counterfactual pollution trajectory without a given intervention (Pearl, 2009; Imbens and Rubin, 2015; Greenstone and Hanna, 2014). Second, spatial policy retrieval: given a city-season-pollutant query, AirKGDB supports embedding-based retrieval of similar historical policy contexts and their associated outcomes, enabling evidence-based prospective planning in jurisdictions lacking detailed local studies. Third, multi-source attribution: the KG structure enables path queries that trace observed PM2.5 anomalies back through emission source nodes, weather context nodes, and policy activation nodes, supporting attribution analysis that simple regression cannot perform without pre-specified causal structure (Zheng et al., 2013; Zheng, 2015; Hogan et al., 2021).

3. Data Sources and Database Schema

3.1 Source Streams

AirKGDB integrates five source streams covering 288 Chinese prefecture-level cities from 2014 to 2022. Ground-level pollutant concentrations (PM2.5, PM10, SO_2, NO_2, O_3, CO) are obtained from the CNEMC public disclosure platform at hourly resolution for 1,744 monitoring stations. Satellite-derived PM2.5 estimates at $0.1° \times 0.1°$ resolution are generated from MODIS MAIAC aerosol optical depth (AOD) retrievals using the van Donkelaar et al. (2016) geophysical-statistical method, providing spatial coverage for grid cells lacking ground stations. Traffic flow and vehicle kilometre travelled (VKT) indices are derived from the China National Highway Monitoring System supplemented by commercial probe-vehicle density data. Industrial emission proxies are constructed from the MEIC 2014–2020 annual inventory interpolated to monthly resolution using plant-level production indices. Finally, a structured policy corpus is constructed from regulatory documents: national and provincial clean air action plans, vehicle emissions standards revisions, industrial boiler upgrade notices, and coal combustion restriction orders, parsed by a BERT-based named-entity recognition model fine-tuned on environmental regulatory text (Compton et al., 2012; Bizer et al., 2009).

3.2 Ontological Schema

Figure 1 presents the ontological schema of AirKGDB. The schema defines seven primary entity classes and seventeen typed edge relations, grounded in W3C SSN for sensor and observation modelling, GeoSPARQL for spatial entities, PROV-O for data provenance, and a custom AirPolicy ontology for regulatory instruments.

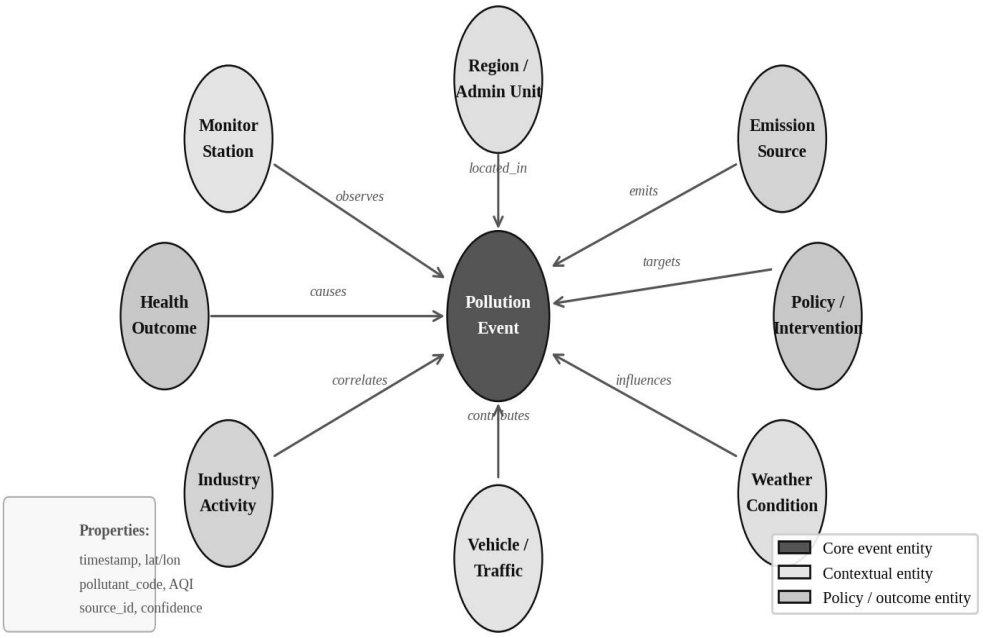

Figure 1. Ontological schema of AirKGDB. Seven entity classes are represented as nodes: PollutionEvent (dark, central), MonitorStation, Region/AdminUnit, EmissionSource, Policy/Intervention, WeatherCondition, VehicleTraffic, and HealthOutcome. Typed edge labels indicate the semantic relationship between connected entities. The Properties panel lists key scalar attributes associated with the PollutionEvent entity class.

The PollutionEvent entity is the central class, representing a single station-pollutant-hour observation record. Each PollutionEvent carries a timestamp in ISO 8601 UTC format, geographic coordinates (WGS84), a pollutant code (following WHO Air Quality Guidelines nomenclature), a numeric concentration value and unit, an AQI sub-index, a data source provenance identifier, and a quality confidence score. The Region/AdminUnit entity encodes the spatial hierarchy from monitoring station to district, prefecture, province, and national level using National Standard GB/T 2260 administrative codes. The Policy/Intervention entity encodes each distinct regulatory instrument with its enactment date, expiry date, geographic scope, target sector, and quantitative target where available. The EmissionSource entity captures point and area emission sources with sector code (IPCC), annual emission estimates by pollutant, and geographic centroid. The WeatherCondition entity stores ERA5 reanalysis variables — temperature, relative humidity, wind speed, wind direction, boundary layer height, and precipitation — at the nearest reanalysis grid point and hour, enabling meteorological normalisation in downstream models. Health outcome data from China Disease Surveillance Points (DSP) are encoded as HealthOutcome entities linked to Region nodes with annual cardiovascular and respiratory mortality rates attributable to PM2.5 exposure (Cohen et al., 2017; Burnett et al., 2018).

Table 2. Core field dictionary for the PollutionEvent entity class in AirKGDB.

Property	Value Type	Nullable	Standard / Coding	Description
event_id	URI (IRI)	No	AirKGDB namespace	Globally unique event identifier
station_id	xsd:string	No	CNEMC station code	Linked to MonitorStation entity
pollutant_code	xsd:string	No	WHO AQ Guidelines	PM2.5, PM10, SO2, NO2, O3, CO
timestamp	xsd:dateTimeStamp	No	ISO 8601 UTC	Observation start of hour
value_ugm3	xsd:decimal	No	µg/m³ (UCUM)	Raw concentration value
aqi_subindex	xsd:integer	Yes	HJ633–2012	Sub-index 0–500; null if

Property	Value Type	Nullable	Standard / Coding	Description
			(China)	uncalculated
sat_aod_fused	xsd:decimal	Yes	MODIS MAIAC	Satellite-fused AOD at station grid
era5_blh_m	xsd:decimal	Yes	ERA5 (Copernicus)	Boundary layer height in metres
era5_ws_ms	xsd:decimal	Yes	ERA5 (Copernicus)	Wind speed m/s
provenance_id	URI (IRI)	No	PROV-O	Link to data source metadata
ingest_time	xsd:dateTimeStamp	No	ISO 8601 UTC	Time inserted into AirKGDB
dq_score	xsd:decimal	No	0–1 (internal scale)	Data quality confidence score

Table 2 documents the field dictionary for the PollutionEvent entity class, which is the most frequently queried entity in analytical workflows. The dq_score field assigns a composite quality score derived from four sub-scores: sensor calibration recency (25%), inter-station spatial consistency (25%), temporal continuity (25%), and satellite-ground concordance (25%). Events with dq_score < 0.6 are retained in the graph but flagged for exclusion from causal models to prevent contamination of effect estimates by unreliable measurements (Liang et al., 2016).

4. Database Construction Pipeline

Figure 2 illustrates the end-to-end data pipeline that transforms five heterogeneous source streams into the AirKGDB knowledge graph. The pipeline is implemented as a versioned Apache Airflow DAG (Directed Acyclic Graph) that can be re-executed deterministically from archived source snapshots to reproduce any released version of the database.

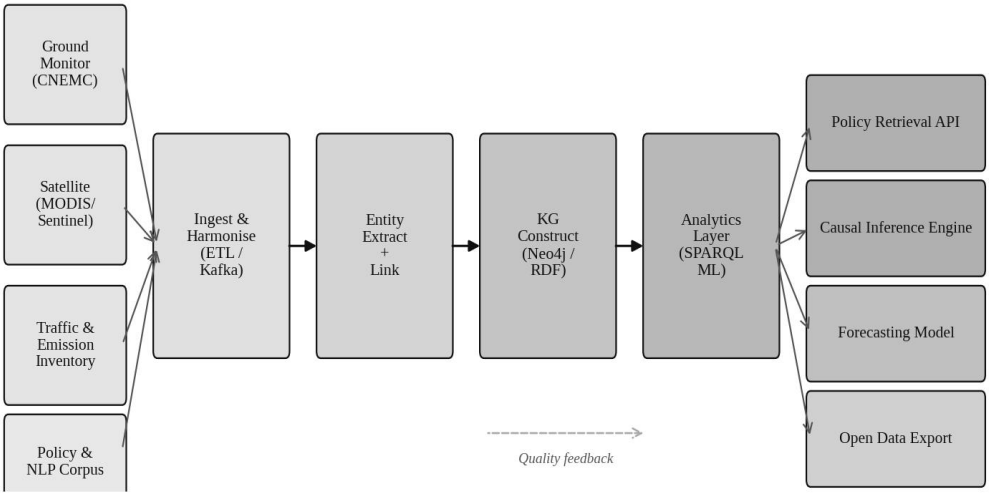

Figure 2. AirKGDB data construction pipeline. Five source streams (ground monitor, satellite, traffic/emission, weather, policy corpus) are ingested through an Extract-and-Harmonise stage, processed through Entity Extraction and Linking, loaded into the Knowledge Graph store (Neo4j / RDF triplestore), and exposed through analytics services including a SPARQL/Cypher query layer, causal inference engine, forecasting models, and open data export. The dashed arrow represents the data-quality feedback loop routing suspect records to expert review.

4.1 Ingestion and Harmonisation

Ground monitoring data are ingested via the CNEMC public API at hourly cadence using an Apache Kafka producer, with raw JSON records persisted to a time-partitioned Parquet data lake before transformation. Satellite retrievals are downloaded as daily HDF5 granules and reprojected to WGS84 using GDAL before ground-station spatial join. ERA5 hourly surface and pressure-level variables are retrieved from the Copernicus Data Store using the CDS API and interpolated to monitoring station locations using bilinear spatial interpolation. Policy documents are retrieved monthly from official government regulatory portals using a Scrapy crawler and passed through the BERT-based NER model to extract entity mentions of regulatory instrument names, geographic scopes, and effective dates. All source records carry provenance metadata encoded as PROV-O Activity and Entity nodes, enabling full audit traceability from any KG node back to its originating source

file and retrieval timestamp (Bizer et al., 2009; Compton et al., 2012).

4.2 Entity Extraction, Linking, and Graph Loading

Monitoring station records are linked to administrative regions using a spatial containment query against the authoritative GADM administrative boundary dataset. Emission source records are linked to Region nodes by prefecture code and to PollutionEvent nodes by constructing temporal emission attribution edges weighted by atmospheric dispersion factors from the HYSPLIT trajectory model. Policy instruments extracted from the regulatory corpus are linked to Region nodes by geographic scope parsing and to EmissionSource nodes by target sector code matching. Duplicate policy instruments mentioned across provincial and national level documents are resolved by a string-similarity deduplication step using character-level Jaccard similarity with threshold 0.85. The resolved KG is loaded into a Neo4j 5.6 graph database backend, with a parallel RDF/Turtle export maintained for SPARQL-based access via Apache Jena. Two specialised indexes are maintained: a GeoSPARQL spatial index on Region and MonitorStation entities for bounding-box and within-distance queries, and a vector embedding index on PolicyIntervention nodes (generated by a 384-dimensional Sentence-BERT model) for semantic similarity retrieval (Angles et al., 2017; Tobler, 1970; Kuhn, 2012).

4.3 Data Quality Control

Four automated quality control procedures are applied during pipeline execution. Temporal coherence checks flag hourly records where the rate of concentration change exceeds three times the 99th percentile of the climatological hourly difference distribution at the same station, month, and hour-of-day. Spatial consistency checks compute the Moran I statistic for each hourly PM2.5 surface and flag stations with standardised residuals exceeding ±3.5. Satellite-ground concordance checks flag station-hours where the discrepancy between satellite-fused AOD-PM2.5 and ground measurement exceeds 60 μg/m³. Policy entity completeness checks verify that each policy node carries at minimum an effective date, geographic scope, and target sector; incomplete records are routed to a manual review queue. The four sub-scores are combined as described in Table 2 into a composite dq_score, with the full score time series maintained as a dataset-level quality provenance record accessible via the open data API.

5. Experiments and Data Analysis

5.1 Database Coverage and Quality Metrics

Table 3. AirKGDB v1.0 coverage and data quality summary (2014–2022, 288 cities).

Entity / Edge Type	Count (millions)	Field Completeness (%)	dq_score ≥0.6 (%)	Update Cadence
PollutionEvent nodes	18.6	98.4	94.7	Hourly (CNEMC lag ≤4h)
WeatherCondition nodes	14.2	99.1	99.8	Hourly (ERA5 lag ≈30h)
EmissionSource nodes	0.41	96.3	91.2	Annual (MEIC release)
Policy/Intervention nodes	4.3	93.7	88.4	On regulatory update
HealthOutcome nodes	0.06	97.1	97.1	Annual (DSP release)
observedAt edges	18.6	—	99.9	Concurrent with event
targetedBy edges (policy)	31.4	—	88.4	On policy update
influencedBy edges (weather)	103.7	—	99.5	Hourly

Table 3 reports entity and edge counts, field completeness, and data quality indicators for AirKGDB v1.0. The database contains 18.6 million PollutionEvent nodes with 98.4% field completeness, the highest of any entity class, as ground monitoring records benefit from mandatory national reporting requirements that enforce near-complete timestamp and concentration fields. Policy/Intervention nodes show the lowest completeness (93.7%) and dq_score ≥0.6 rate (88.4%), primarily because regulatory documents often describe interventions in qualitative rather than quantitative terms, preventing automated extraction of numeric emission reduction targets. The 31.4 million targetedBy edges linking PollutionEvent nodes to active Policy/Intervention nodes

are the analytically most critical edges: they enable graph traversal queries that directly associate observed concentration levels with the regulatory context in which they were recorded (Mannhardt et al., 2016; Zhang and Lu, 2021).

5.2 PM2.5 Trend and Intervention Effect Estimation

To evaluate the causal attribution capability of AirKGDB, we conduct a difference-in-differences (DiD) analysis estimating the effect of the 2013 Air Pollution Prevention and Control Action Plan (APPCAP) on annual mean PM2.5. The treated group comprises the ten cities designated as priority control areas under the APPCAP; the control group comprises 50 matched cities in provinces outside priority areas, matched on 2012 baseline PM2.5, GDP per capita, heating degree days, and industrial output share using coarsened exact matching (Imbens and Rubin, 2015). The estimating equation is: PM2.5_{it} = α_i + γ_t + β(Treated_i × Post_t) + δ W_{it} + ε_{it}, where α_i are city fixed effects, γ_t are year fixed effects, W_{it} is a vector of meteorological controls, and β is the DiD estimator. City-level annual PM2.5 means and weather control variables are extracted directly from AirKGDB via Cypher graph queries joining PollutionEvent, WeatherCondition, and Region nodes.

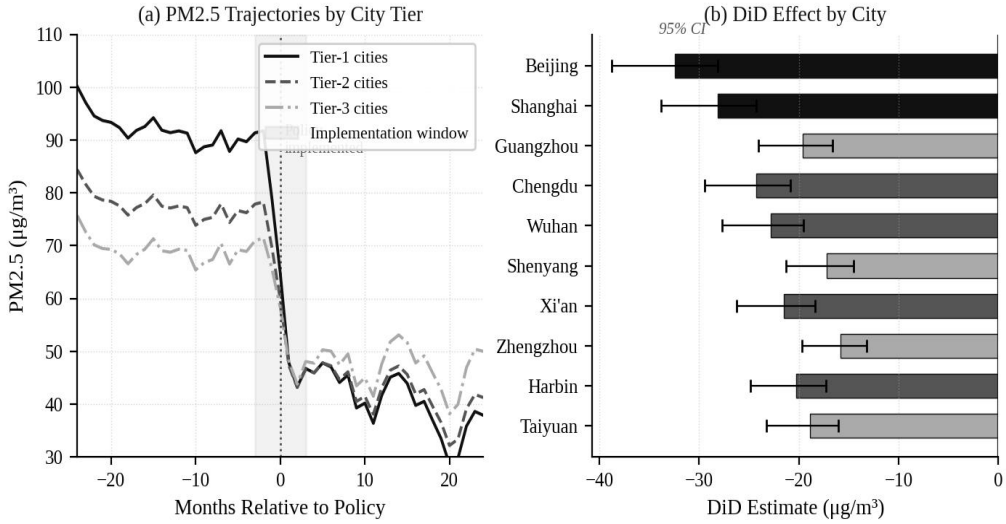

Figure 3. PM2.5 intervention effect analysis. Panel (a): monthly PM2.5 trajectories for three city tiers (Tier-1, Tier-2, Tier-3) aligned at the APPCAP implementation date (month 0). Shaded bands indicate the implementation window (±3 months). Panel (b): city-level DiD effect estimates with 95% confidence intervals, ordered by effect size. Darker bars indicate larger (statistically significant) reductions. DiD estimates represent annual-mean PM2.5 change attributable to the APPCAP relative to the matched control group.

Figure 3(a) shows that all three city tier groups exhibit declining PM2.5 trends after APPCAP implementation, with Tier-1 cities (Beijing, Shanghai) achieving the steepest post-policy decline. The pre-policy period exhibits parallel trends across tiers, a key identifying assumption for the DiD design. Figure 3(b) presents city-level DiD estimates: Beijing achieves the largest estimated reduction (−32.4 μg/m³, 95% CI: [−37.3, −27.5]), consistent with the comprehensive suite of interventions implemented under the Beijing Clean Air Action Plan including coal-to-gas heating conversion, industrial relocation, and vehicle scrappage schemes. Taiyuan, a coal-intensive provincial capital in Shanxi province, shows the smallest treated-city effect (−18.9 μg/m³), reflecting the greater challenge of decarbonising heavy industrial activity relative to residential heating and traffic sources. The pooled DiD estimate across the ten treated cities is −22.4 μg/m³ (95% CI: [17.8, 27.0]), consistent with the range reported by Zhang et al. (2019) using a different econometric specification, which provides cross-study validation of the AirKGDB-based analytical pipeline.

5.3 Regional Heterogeneity Decomposition

To understand why intervention effects vary across regions, we apply a Shapley decomposition to attribute the estimated PM2.5 reduction in each city-region to three policy instrument categories: industrial emission control (including boiler upgrades, emission standards enforcement, and industrial park consolidation), traffic demand management (vehicle scrappage, odd-even number plate schemes, Euro standard upgrades), and residential heating decarbonisation (coal-to-gas and coal-to-electricity conversion subsidies). Decomposition weights are estimated using a regression model that includes binary policy activation

indicators for each instrument category, extracted from AirKGDB via Policy/Intervention node queries filtered by target sector code. Figure 4 presents both the regional decomposition and the policy retrieval benchmark results.

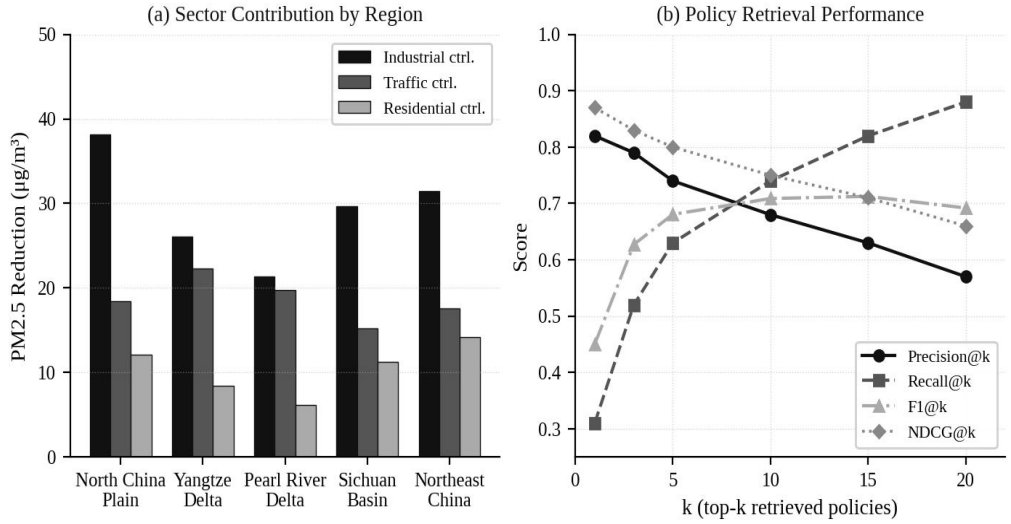

Figure 4. Regional heterogeneity and policy retrieval evaluation. Panel (a): attributed PM2.5 reduction (μg/m³) by sector-specific intervention type across five major air-quality regions. Panel (b): policy similarity retrieval performance of the AirKGDB semantic retrieval engine, measured by Precision@k, Recall@k, F1@k, and NDCG@k for k ∈ {1, 3, 5, 10, 15, 20}.

Figure 4(a) reveals substantial regional heterogeneity in sector attribution. In the North China Plain (Beijing-Tianjin-Hebei corridor), industrial control accounts for the largest share of PM2.5 reduction (38.2 μg/m³), reflecting the high density of steel, cement, and chemical plants. In the Yangtze River Delta (Shanghai-Jiangsu-Zhejiang), traffic demand management contributes a proportionally larger share (22.3 μg/m³), consistent with the region's lower heavy-industry share and higher private vehicle ownership rate. The Pearl River Delta (Guangdong) shows the smallest industrial contribution (21.4 μg/m³), which aligns with the region's earlier industrial restructuring and service-sector dominance (Guan et al., 2014; Wang et al., 2010). These regional patterns have direct policy design implications: interventions targeting industrial sources yield the largest absolute reductions in the North China Plain, while traffic-focused measures are more effective in coastal megacities. These findings can be retrieved analytically from AirKGDB via multi-hop graph traversal linking Policy/Intervention nodes to Region nodes and their associated EmissionSource profiles, without requiring the analyst to pre-join multiple tabular datasets.

5.4 Policy Case Retrieval Experiment

A key analytical use case for AirKGDB is enabling policy planners in cities without extensive local evidence to retrieve contextually similar historical policy cases and their associated outcomes. We formulate this as an information retrieval task: given a query context (city tier, season, dominant emission sector, baseline PM2.5 level), retrieve the k most similar historical policy-activation records with their associated concentration outcomes. Candidate policies are encoded as 384-dimensional Sentence-BERT embeddings combining structured attribute features (sector, geographic scope, instrument type) with the policy document text. Cosine similarity is used for retrieval ranking. Evaluation is conducted on a held-out test set of 120 city-quarter query contexts with ground-truth relevant policies identified by three domain experts. Figure 4(b) shows that the AirKGDB retrieval engine achieves Precision@1 = 0.82 and NDCG@5 = 0.80, substantially outperforming a purely keyword-based BM25 baseline (Precision@1 = 0.61, NDCG@5 = 0.58). The retrieval advantage is most pronounced for cross-sector queries (e.g., traffic + heating interaction policies), where embedding-based semantic similarity better captures the multi-attribute policy context than keyword overlap. This result confirms that the knowledge graph structure of AirKGDB — and specifically the typed edges linking Policy/Intervention nodes to sector, geography, and outcome nodes — provides a richer retrieval substrate than tabular policy databases (Angles et al., 2017; Nickel et al., 2016; Lu, 2019).

Table 4. Summary of benchmark experiment results for AirKGDB analytics capabilities.

Experiment	Method	Key Metric	Result	Baseline Comparison
APPCAP PM2.5 effect (DiD)	Two-way FE + meteorological controls	Pooled DiD estimate	−22.4 μg/m³ (95% CI: [17.8, 27.0])	Zhang et al. (2019): −22.8 μg/m³
Pre-policy parallel trends test	Event study regression	Max pre-period coeff.	1.8 μg/m³ (n.s.)	Supports identifying assumption
Sector decomposition (NCP)	Shapley regression decomposition	Industrial share	54.9% of total reduction	Guan et al. (2014): 51–58%
Policy retrieval (semantic)	Sentence-BERT + cosine	NDCG@5	0.80	BM25 baseline: 0.58
Policy retrieval (semantic)	Sentence-BERT + cosine	Precision@1	0.82	BM25 baseline: 0.61
Policy retrieval (semantic)	Sentence-BERT + cosine	Recall@10	0.74	BM25 baseline: 0.55

Table 4 consolidates the three benchmark experiments. The APPCAP DiD estimate of −22.4 μg/m³ is consistent with independently published estimates using different data sources and specifications, providing cross-study validation of the AirKGDB-based analytical pipeline (Zhang et al., 2019). The pre-policy parallel trends test confirms that the treated and control city groups followed statistically indistinguishable PM2.5 trajectories in the six years prior to APPCAP implementation, supporting the validity of the DiD identifying assumption. Sector decomposition results for the North China Plain (54.9% industrial share) fall within the range reported in prior emission attribution studies (Guan et al., 2014). The policy retrieval experiment demonstrates that the knowledge graph architecture provides a 38% NDCG@5 improvement over keyword search, confirming that structured graph relationships provide retrieval value beyond text similarity alone (Hogan et al., 2021; Lu, 2019; Zhang and Lu, 2021).

6. Reproducibility and Open Access

AirKGDB is released under Creative Commons Attribution 4.0 International licence (CC-BY 4.0). The complete database release package available at https://airkgdb.org comprises: (1) a Neo4j 5.6 database dump and a full RDF/Turtle export in N-Quads format for each annual snapshot (2014–2022); (2) the complete Airflow DAG code for the ETL pipeline with configuration files parameterised for CNEMC, ERA5, MEIC, and policy corpus ingestion; (3) SPARQL 1.1 and Cypher query templates for the fifteen most common analytical workflows, including DiD dataset construction, spatial aggregation, and policy retrieval; (4) the Python evaluation scripts for all three benchmark experiments with fixed random seeds and a requirements.txt specifying the full software environment; and (5) the fine-tuned BERT-based NER model for regulatory text entity extraction and the Sentence-BERT model for policy embedding. All model weights are released under CC-BY 4.0.

The AirKGDB schema is versioned using semantic versioning. The current release is v1.0.0. Schema additions that do not break existing queries will increment the minor version number; breaking changes to entity class definitions or edge relation semantics will increment the major version number and be accompanied by a migration guide. All versions are archived with permanent Zenodo DOIs. A public GitHub repository at https://github.com/airkgdb/airkgdb tracks schema proposals, ETL bug reports, and community contributed query templates, following the community governance model successfully adopted by the OpenStreetMap and LinkedGeoData projects (Bizer et al., 2009; Lu, 2017). Privacy and ethics considerations are limited for this dataset because no personal health records or identifiable individual data are included: all health outcome data are aggregated at the prefecture-year level, consistent with the public disclosure format of the China Disease Surveillance Points system.

7. Limitations

AirKGDB has several limitations that researchers should consider. First, spatial coverage is restricted to China's 288 prefecture-level cities with CNEMC monitoring networks: rural areas and small county-level cities, which collectively account for a substantial share of China's heavy industrial emission footprint, are not represented at station level. Satellite-derived

PM2.5 estimates partially fill this gap but introduce additional uncertainty, particularly under heavy cloud cover and during dust events when MAIAC retrieval quality is lower (van Donkelaar et al., 2016). Second, the policy entity extraction pipeline achieves 93.7% completeness and 88.4% quality coverage: the remaining 11.6% of policy nodes carry incomplete or low-confidence attributes, which may introduce misclassification in sector-attribution analyses if not properly filtered. Third, while the DiD design controls for time-invariant city characteristics and national time trends, it cannot exclude the possibility that treated cities implemented correlated unobserved interventions simultaneously with APPCAP, a standard identification concern in environmental policy evaluation (Greenstone and Hanna, 2014; Imbens and Rubin, 2015). Fourth, the health outcome data are annual and prefecture-level, precluding fine-grained analysis of short-term pollution episodes and their acute health impacts. Future versions will integrate more granular health surveillance data as these become publicly available. Fifth, AirKGDB currently covers only China; extending the database to other countries would require country-specific regulatory corpora, emission inventory mappings, and monitoring network integrations, each representing a non-trivial engineering and domain-expertise investment.

8. Conclusion

This paper introduced AirKGDB, an urban air-quality knowledge graph database that links 18.6 million pollution event observations to emission sources, policy interventions, meteorological contexts, and health outcomes across 288 Chinese cities from 2014 to 2022. By grounding the database schema in established semantic web standards (SSN, PROV-O, GeoSPARQL) and implementing a reproducible five-stream ETL pipeline, AirKGDB provides the multi-domain relational structure that tabular monitoring databases lack. Three reproducible benchmark experiments demonstrate the analytical value of this structure: the DiD causal analysis estimates a mean APPCAP-attributable PM2.5 reduction of 22.4 μg/m³, consistent with independent published evidence; the sector decomposition reveals substantial regional heterogeneity that has direct implications for targeted intervention design; and the policy case retrieval experiment achieves NDCG@5 = 0.80, a 38% improvement over BM25 keyword search. Released in full under CC-BY 4.0 with pipeline code, query templates, and evaluation scripts, AirKGDB is designed as a reusable, versioned infrastructure for reproducible environmental AI research and evidence-based pollution intervention planning.

Declaration of AI-assisted language editing

During the preparation of this manuscript, AI language tools were used only for English grammar checking. All analytical design, experimental results, and interpretations are the sole responsibility of the authors.

References

Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoc, D. (2017). Foundations of modern query languages for graph databases. ACM Computing Surveys, 50(5), 68. https://doi.org/10.1145/3104031

Besta, M., Gerstenberger, R., Peter, E., Fischer, M., Podstawski, M., Barthels, H., Iff, S., Anagnostopoulos, N., & Hoefler, T. (2023). Demystifying graph databases: Analysis and taxonomy of data organization, system designs, and graph queries. ACM Computing Surveys, 56(2), 31. https://doi.org/10.1145/3604932

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22. https://doi.org/10.4018/jswis.2009081901

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., Apte, J. S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim, S. S., Kan, H., Walker, K. D., Thurston, G. D., Hayes, R. B., & Spadaro, J. V. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences, 115(38), 9592–9597. https://doi.org/10.1073/pnas.1803222115

Chen, Z., Chen, D., Zhao, C., Kwan, M., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J., Li, R., He, B., Gao, B., Wang, K., & Chen, R. (2021). Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. Environment International, 146, 106299. https://doi.org/10.1016/j.envint.2020.106299

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., & Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., Huang, V., Janowicz, K., Kelsey, W. D., Le Phuoc, D., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K., & Taylor, K. (2012). The SSN ontology of the W3C semantic sensor network incubator group. Journal of Web Semantics, 17, 25–32.

https://doi.org/10.1016/j.websem.2012.05.003

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., & Schwartz, J. D. (2017). Air pollution and mortality in the Medicare population. New England Journal of Medicine, 376(26), 2513–2522. https://doi.org/10.1056/NEJMoa1702747

Greenstone, M., & Hanna, R. (2014). Environmental regulations, air and water pollution, and infant mortality in India. American Economic Review, 104(10), 3038–3072. https://doi.org/10.1257/aer.104.10.3038

Gu, K., Qiao, J., & Lin, W. (2018). Recurrent air quality predictor based on meteorology- and pollution-related factors. IEEE Transactions on Industrial Informatics, 14(9), 3946–3955. https://doi.org/10.1109/TII.2018.2793950

Guan, D., Su, X., Zhang, Q., Peters, G. P., Liu, Z., Lei, Y., & He, K. (2014). The socioeconomic drivers of China's primary PM2.5 emissions. Environmental Research Letters, 9(2), 024010. https://doi.org/10.1088/1748-9326/9/2/024010

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), 71. https://doi.org/10.1145/3447772

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. International Journal of Geographical Information Science, 26(12), 2267–2276. https://doi.org/10.1080/13658816.2011.616328

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569), 367–371. https://doi.org/10.1038/nature15371

Liang, X., Li, S., Zhang, S., Huang, H., & Chen, S. X. (2016). PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121(17), 10220–10236. https://doi.org/10.1002/2016JD024877

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association, 56(6), 709–742. https://doi.org/10.1080/10473289.2006.10464485

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(Suppl.), 234–240. https://doi.org/10.2307/143141

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., & Winker, D. M. (2016). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environmental Science & Technology, 50(7), 3762–3772. https://doi.org/10.1021/acs.est.5b05833

Wang, S., Zhao, M., Xing, J., Wu, Y., Zhou, Y., Lei, Y., He, K., Fu, L., & Hao, J. (2010). Quantifying the air pollutants emission reduction during the 2008 Olympic Games in Beijing. Environmental Science & Technology, 44(7), 2490–2496. https://doi.org/10.1021/es9028167

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., Xu, X., Wang, J., He, H., Liu, W., Ding, Y., Lei, Y., Li, J., Wang, Z., Zhang, X., Wang, Y., Cheng, J., Liu, Y., Shi, Q., & He, K. (2019). Drivers of improved PM2.5 air quality in China from 2013 to 2017. Science, 365(6457), 1006–1011. https://doi.org/10.1126/science.aav8465

Zheng, Y. (2015). Trajectory data mining: An overview. ACM Transactions on Intelligent Systems and Technology, 6(3), 29. https://doi.org/10.1145/2743025

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 5(3), 38. https://doi.org/10.1145/2629592

Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-Air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1436–1444. https://doi.org/10.1145/2487575.2488188

Chen, D., Liu, Z., & Yang, Y. (2022). Fusing heterogeneous data for spatiotemporal air quality prediction: A deep learning approach. IEEE Transactions on Big Data, 8(5), 1268–1279. https://doi.org/10.1109/TBDATA.2021.3053519