

Semantic specification of urban environmental exposure calculations within The World Avatar: a Singapore case study

Kok Foong Lee¹, Jielin Chen¹, Xu Tang^{1,2}, Yuan Wang^{1,3},
Sebastian Mosbach^{1,4}, Mei Qi Lim¹, Rudi Stouffs^{1,3}, Jethro Akroyd^{1,4},
Ronita Bardhan^{1,2}, Markus Kraft^{1,4}

released: September 29, 2026

¹ CARES
Cambridge Centre for Advanced
Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

² Sustainable Design Group
Department of Architecture
University of Cambridge
Cambridge, United Kingdom

³ Department of Architecture
College of Design and Engineering
National University of Singapore
4 Architecture Drive
Singapore 117566

⁴ Department of Chemical Engineering
and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

Preprint No. 348



Keywords: environmental exposure, epidemiology, data provenance, knowledge graph

Edited by

Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

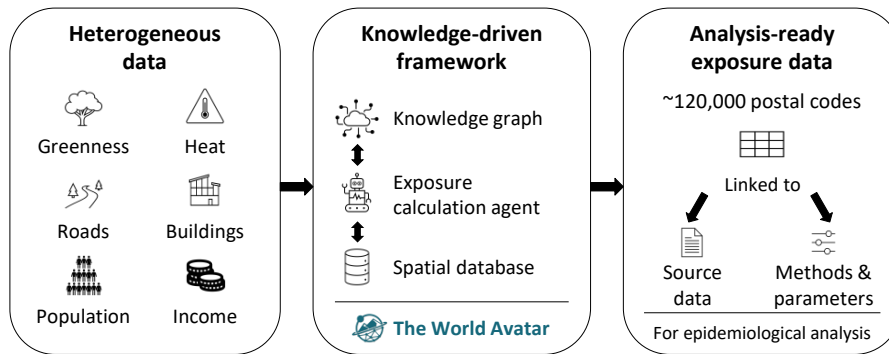
E-Mail: mk306@cam.ac.uk

World Wide Web: <https://como.ceb.cam.ac.uk/>



Abstract

Health-data-driven urban design requires coherent environmental exposure measures for defined study locations at scale. However, assembling measures from datasets that differ in environmental domain, spatial form and temporal coverage often relies on scripts whose assumptions and provenance are difficult to recover. This study develops a framework for HD4, a programme funded by Singapore's National Research Foundation to link environmental exposures with population health outcomes. The framework uses a semantic model to represent calculation methods and parameters separately from source datasets, while a spatial database performs raster and vector operations. Each result remains linked to its study location, source data and calculation specification. Applied across approximately 120,000 Singapore postal-code locations, the framework generated physical and visual greenness, humid heat, urban morphology and socioeconomic measures. The case study demonstrates how these heterogeneous measures can be retrieved by exposure characteristics and temporal context, with each value traceable to its source data, calculation method and parameters.



Highlights

- Represents calculation methods and parameters separately from source datasets.
- Uses a common semantic model for raster, vector and attribute-based exposures.
- Links semantic calculation specifications to spatial execution in PostGIS.
- Retrieves exposures with their source data, temporal context and calculation choices.
- Generates traceable exposures for approximately 120,000 Singapore postal codes.

Contents

1	Introduction	4
2	Background	5
2.1	Study setting and design requirements	5
2.2	The World Avatar	5
2.3	Integration of GIS and Semantic Web technologies	6
3	Knowledge-driven exposure calculation framework	7
3.1	Overview	7
3.2	Core ontology	8
3.3	Hybrid system architecture	9
4	Data	11
4.1	Raster datasets	12
4.1.1	NDVI	12
4.1.2	UTCI	12
4.1.3	Population	12
4.2	Vector datasets	13
4.3	Attribute datasets	13
4.3.1	BGVI	13
4.3.2	Income distribution	14
5	Exposure calculations	14
5.1	Neighbourhood-based exposure calculations	16
5.1.1	Raster-based exposures	16
5.1.2	Vector-based exposures	17
5.2	Direct assignment	17
6	Results	17
6.1	Semantic representation of exposure results	17
6.2	Generation of analysis-ready exposure results	18
6.3	Relationships between exposure results	21
6.4	Semantic retrieval and provenance	22

7	Limitations and future work	25
8	Conclusions	26
	Nomenclature	27
A	Supplementary Information	29
A.1	Income PCA	29
A.2	Ontology representation and example queries	29
A.2.1	Notation and namespaces	30
A.2.2	Subject	31
A.2.3	Exposure dataset	32
A.2.4	Calculation	33
A.2.5	Exposure result	35
	References	38

1 Introduction

This work was conducted as part of the HD4 programme [7] in Singapore, which aims to develop a city-scale data pipeline linking heterogeneous urban environmental datasets for investigations into how the built environment influences public health outcomes. A central component of the programme is a collaboration with SG100K [50], a large-scale longitudinal cohort study designed to integrate genomic, lifestyle, and clinical information from 100,000 Singaporeans, with the aim of supporting precision medicine research tailored to Asian populations. Within this broader programme, the present study focuses on constructing and analysing urban environmental exposure data that can subsequently support epidemiological research.

Supporting such epidemiological studies requires exposure measures covering several aspects of the environment. Exposure calculation is often handled through fragmented scripts, manually assigned variable names, and limited provenance. This becomes increasingly difficult to maintain as datasets are updated, new exposure domains are introduced, or exposure definitions need to be recalculated at different buffer sizes, time periods, or spatial units.

Recent environmental epidemiology studies increasingly combine diverse datasets to quantify exposures such as greenness, heat, urban morphology, transportation infrastructure, and socioeconomic characteristics [46, 51, 61], highlighting the need to integrate heterogeneous environmental data. This allows several aspects of the environment to be characterised for the same study locations, but also increases the number of datasets and calculation choices that must be managed. Interpreting and updating the resulting exposure measures requires clear records of their source datasets, temporal context, calculation methods, and parameters.

Semantic technologies have been increasingly explored within epidemiology to improve the integration and interoperability of heterogeneous health data. Existing work has largely focused on ontology development for epidemiological concepts [48], exposure science and environmental exposure modelling [37, 65], as well as the automated construction of knowledge graphs from epidemiological literature and surveillance data [17, 36]. Broader reviews have also highlighted the increasing adoption of Semantic Web technologies within healthcare and epidemiology [73].

Among systems that employ semantic technologies for epidemiological applications, the multi-agent decision support system proposed by Li and Mackaness [35] is one of the more closely related to the present work. Their framework integrates epidemiological datasets, analytical services, and spatio-temporal information through domain ontologies, enabling users to select datasets and computational services for epidemic management.

Semantic Web technologies have also been applied to geospatial processing workflows, where workflow components, datasets, and geoprocessing services are represented semantically to improve workflow discovery, interoperability, and reuse [59, 72]. While these approaches focus on general-purpose geoprocessing workflows, the present work instead models the domain knowledge required for environmental exposure assessment, capturing the semantics of exposure calculations.

This work builds upon the framework presented by Chen et al. [9] for environmental

exposure assessment. Here, each exposure estimate is explicitly linked to its source data, calculation method and parameters. These links allow researchers to retrieve estimates calculated using specified methods or buffer sizes, together with the information needed to interpret them. Retrieval of exposure estimates together with their linked calculation specifications was not demonstrated in the earlier work. Within the HD4 programme, the present study addresses the generation of exposure measures for fixed locations at city scale, complementing the earlier work on individual movement trajectories.

We demonstrate the framework by generating environmental, urban-morphology and socioeconomic exposure measures for approximately 120,000 postal-code locations in Singapore. The case study shows how heterogeneous exposure measures can be generated at city scale and retrieved together with their source data and calculation specifications, providing traceable, analysis-ready datasets for subsequent epidemiological research.

2 Background

2.1 Study setting and design requirements

Within the broader HD4 programme, this paper focuses exclusively on the construction and analysis of urban environmental exposure data using a knowledge-driven framework. The subsequent linkage of these environmental data to individual-level health records is outside the scope of this paper and is handled separately within HD4.

Supporting environmental exposure assessment at city scale requires the integration of a wide range of geospatial datasets describing different aspects of the urban environment. Geographic Information Systems (GIS) provide the primary framework for storing, managing, and analysing these data, which are commonly represented as vector or raster datasets. These datasets often differ in their spatial and temporal resolution, may be updated over time, and can represent observations from different periods. Such heterogeneity makes them difficult to organise and integrate consistently. In this work, we demonstrate how a knowledge-driven approach based on the ideas underlying The World Avatar [1] addresses these challenges by providing a semantic framework for organising urban environmental datasets while preserving their metadata, provenance, and relationships.

2.2 The World Avatar

The proposed framework is implemented within The World Avatar (TWA), a dynamic knowledge graph based on Semantic Web technologies [1]. One of the distinguishing features of TWA is its use of autonomous software agents that interact with the knowledge graph. These range from lightweight agents that update information, such as weather observations, to agents that perform complex calculations. TWA has been applied to urban and environmental problems that draw on data and models from several domains. Applications include district heating optimisation and municipal heat planning [28, 64], the integration of national-scale energy and climate data [58], analyses of how heat-pump

adoption and energy prices affect social and regional inequalities [57, 70], building retrofit assessment [8] and question answering for urban planning [11]. These applications show how TWA can link information about buildings, energy systems, socioeconomic conditions and the wider urban environment, making it available for computation and retrieval. Applications in molecular engineering [6, 32], scientific question answering [45] and digital laboratory monitoring [52, 53] demonstrate the wider scope of the approach.

TWA represents knowledge using the Resource Description Framework (RDF) [60], in which information is expressed as triples consisting of a subject, predicate, and object. These triples describe entities and the relationships between them and are queried using SPARQL. Within TWA, RDF data can be managed using two complementary approaches. The first employs dedicated triple stores, such as Blazegraph [5] and RDF4J [13], where RDF triples are explicitly stored and queried directly. The second uses Ontop [69], a virtual knowledge graph, in which the underlying data remain in relational databases while RDF triples are generated dynamically at query time according to declarative mapping files. This virtualisation approach enables existing relational datasets to be queried using SPARQL without duplicating the data as RDF triples.

Although RDF provides a flexible representation of heterogeneous knowledge, practical urban environmental analysis also requires the efficient processing of large geospatial datasets. The design requirements of this work therefore extend beyond semantic knowledge representation to include the management of datasets with diverse spatial and temporal characteristics. The following section reviews the integration of GIS and Semantic Web technologies, motivating the hybrid architecture presented later in Section 3.3.

2.3 Integration of GIS and Semantic Web technologies

SPARQL has been extended to support spatial querying through GeoSPARQL, the OGC-recommended standard for representing and querying geospatial data on the Semantic Web [42]. Other research extensions, such as SPARQL-ST, incorporate both spatial and temporal predicates but are not part of the official OGC GeoSPARQL standard [47].

GeoSPARQL defines a standard vocabulary and set of spatial predicates for RDF data. While GeoSPARQL-compliant triples can be stored in any triple store, the availability of GeoSPARQL functions depends on the implementation, with most triple stores supporting only a subset of the standard. Consequently, both functionality and performance differ considerably between vendors [31]. Furthermore, geospatial triple stores generally remain less mature than established spatial database systems such as PostGIS, with reported limitations in scalability and performance when handling large datasets and complex spatial joins [30].

Another shortcoming of GeoSPARQL is its lack of support for raster operations. GeoSPARQL was primarily designed for vector geometries represented using Well-Known Text (WKT) literals and does not define a standard representation or query functions for raster data. Unlike vector geometries, raster datasets represent continuous spatial fields and often require computationally intensive, tile-based operations that are difficult to express within the GeoSPARQL standard. Although several studies have proposed extensions to incorporate raster support into GeoSPARQL [18, 26], these approaches remain

largely experimental and have not been adopted as part of the official standard.

Ontop [3, 4] provides ontology-based access to relational databases through a SPARQL-to-SQL translation layer, enabling SPARQL queries over relational data. It also implements the GeoSPARQL standard, allowing users to perform spatial queries using GeoSPARQL predicates and functions. This approach is well suited to applications that primarily require simple geospatial querying and semantic access to existing spatial data. However, it is less suitable for workflows that rely on the flexible use of database-specific spatial functions for computationally intensive geospatial processing. The optimisation of SPARQL-to-SQL translation has also been recognised as a challenge [29, 55], with issues such as redundant joins potentially affecting query performance. For geometry-intensive workloads, it can be difficult to ensure that the generated SQL consistently exploits the spatial indexing and query optimisation capabilities provided by spatial database systems such as PostGIS. These considerations motivate a hybrid architecture in which semantic technologies are used for knowledge representation and metadata management, while computationally intensive geospatial operations are executed directly within a spatial database.

Executing geospatial analyses directly within a spatial database also provides greater procedural flexibility than ontology-based query engines. Intermediate results can be materialised, temporary tables created, database sessions maintained, and iterative computations incorporated into controlled analytical workflows. Such procedural operations are difficult to express efficiently through declarative SPARQL queries alone, reinforcing the complementary roles of Semantic Web technologies for knowledge representation and spatial databases for computationally intensive geospatial analysis within a hybrid architecture.

3 Knowledge-driven exposure calculation framework

3.1 Overview

Environmental exposure assessment relies on heterogeneous geospatial datasets that differ in their data structures, temporal characteristics, and calculation requirements. To manage this complexity, this work proposes a knowledge-driven framework that enables exposure results to be retrieved together with the information used to generate them.

In the framework presented by Chen et al. [9], exposure calculations depend on the geometry type of the environmental feature. The present work represents the method and its parameters through a separate calculation specification, linking each derived exposure value to that specification, its source dataset and the study subject. This supports different calculations for the same spatial data representation while retaining the methodological choices associated with each result.

Figure 1 provides a conceptual overview of the proposed framework for generating traceable environmental exposure results that are ready for epidemiological analyses. The exposure specification is represented as computable knowledge, so the same description of the dataset, method and parameters can guide the calculation and be retrieved alongside its result. The framework is illustrated using a representative query that an epidemiologist

From a research question to a traceable exposure value

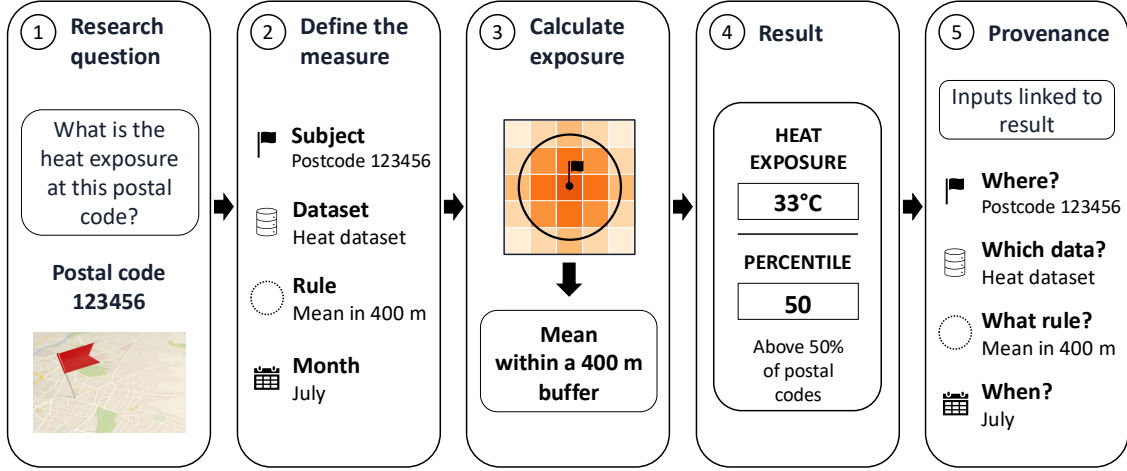


Figure 1: Overview of the proposed framework for traceable environmental exposure assessment.

may wish to answer. For example, in this study, we may be interested in determining the heat exposure at a particular postal code. From here, the information required to perform the calculation is identified, including the subject of interest (a particular postal code), environmental dataset, calculation method, and associated parameters. The exposure is then computed through spatial analysis to produce an analysis-ready exposure value that can subsequently be linked with health data. By explicitly representing these components within the knowledge base, every exposure result remains linked to the postal code for which it was calculated, the dataset used, the calculation method, and parameters such as the buffer size. These links allow each exposure value to be retrieved together with its source dataset and calculation specification.

3.2 Core ontology

To support semantic representation of the exposure calculation workflow, we developed an ontology describing the core entities and relationships involved. An ontology provides a formal specification of concepts and relationships within a domain, enabling information to be represented consistently and interpreted by both humans and machines.

One of the key advantages of the proposed framework is that it explicitly separates a dataset's spatial representation from the method used to derive an exposure. These aspects are modelled as separate concepts within the ontology, allowing the same type of spatial dataset to support different exposure calculations. In this study, for example, raster datasets are used to characterise heat, population and physical greenness.

The ontology centres on four concepts:

- Subject

- `ExposureDataset`
- `Calculation`
- `ExposureResult`

Together, these concepts capture the information required to define an exposure calculation and represent its result. The exposure calculation agent receives instances of `Subject`, `ExposureDataset`, and `Calculation`. It calculates the exposure value, stores it as an `ExposureResult`, and links the result to the three input instances.

The `Subject` represents the entity for which an environmental exposure is calculated. In this work, subjects correspond to fixed postal-code locations represented as vector points, although the framework is designed to support dynamic subjects in future applications. The `ExposureDataset` represents the environmental dataset from which the exposure is derived, while the `Calculation` specifies the computational method together with its associated parameters, such as the buffer radius. Finally, the `ExposureResult` links the derived exposure value to the corresponding subject, source dataset, and calculation.

3.3 Hybrid system architecture

This work adopts a hybrid architecture that combines Semantic Web technologies with a conventional GIS database, PostGIS [49], to support urban environmental exposure calculations. In the framework of Chen et al. [9], environmental data are accessed through semantic queries and spatial exposure calculations are performed within the agent. Here, the agent retrieves the calculation specification from the knowledge base and executes the spatial operations directly in PostGIS. For the city-scale calculations in this study, processing exposure datasets directly in PostGIS limits data transfer and makes use of its spatial indexing and geospatial functions. The semantic layer is represented using RDF and queried using SPARQL, whereas the spatial layer is managed in PostGIS using SQL.

Throughout this paper, specific entities represented by the ontology are referred to as ontology instances. Each instance belongs to an ontology class and represents a particular subject, environmental dataset, calculation method or exposure result. Representative instances are used throughout the figures to illustrate how these concepts are represented within the proposed framework.

The conceptual workflow shown in Figure 1 is implemented by the hybrid architecture illustrated in Figure 2. Each exposure calculation is initiated by one instance of each ontology class, representing the subject (`Subject`), exposure dataset (`ExposureDataset`), and calculation method (`Calculation`). The knowledge base stores the semantic descriptions of these instances. The exposure calculation agent retrieves the corresponding metadata through SPARQL, translates it into the corresponding SQL query, and executes the spatial operations directly in PostGIS. After the calculation is completed, the resulting exposure value is stored as an `ExposureResult` together with explicit links to the semantic inputs used to derive it, preserving its calculation provenance. The exposure results can subsequently be mapped to a data dictionary, providing standardised variable definitions for downstream statistical analysis and data sharing. For illustrative pur-

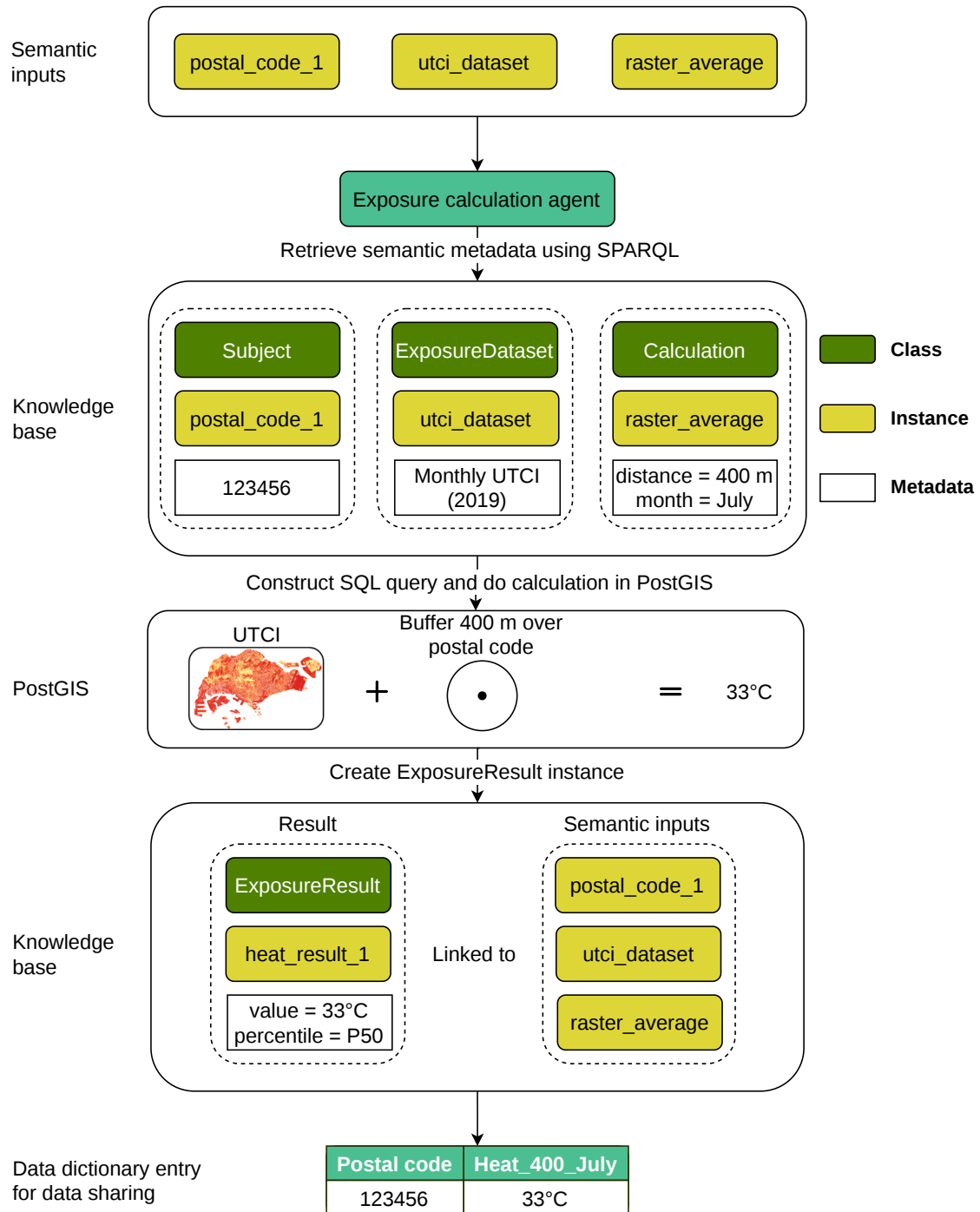


Figure 2: Workflow for calculating exposures. Demonstrated for heat exposure at a selected postal code.

poses, Figure 2 uses the representative instances `postal_code_1`, `utci_dataset`, and `raster_average`.

4 Data

Our Singapore case study uses the data sources shown in Figure 3 to characterise three aspects of the urban environment: environmental conditions, urban morphology and socio-economic conditions. The datasets are supplied in three common forms used in geospatial analysis: raster data, vector data and attributes associated with spatial entities. These different native structures provide a useful test of the framework’s ability to represent and manage heterogeneous environmental data consistently.

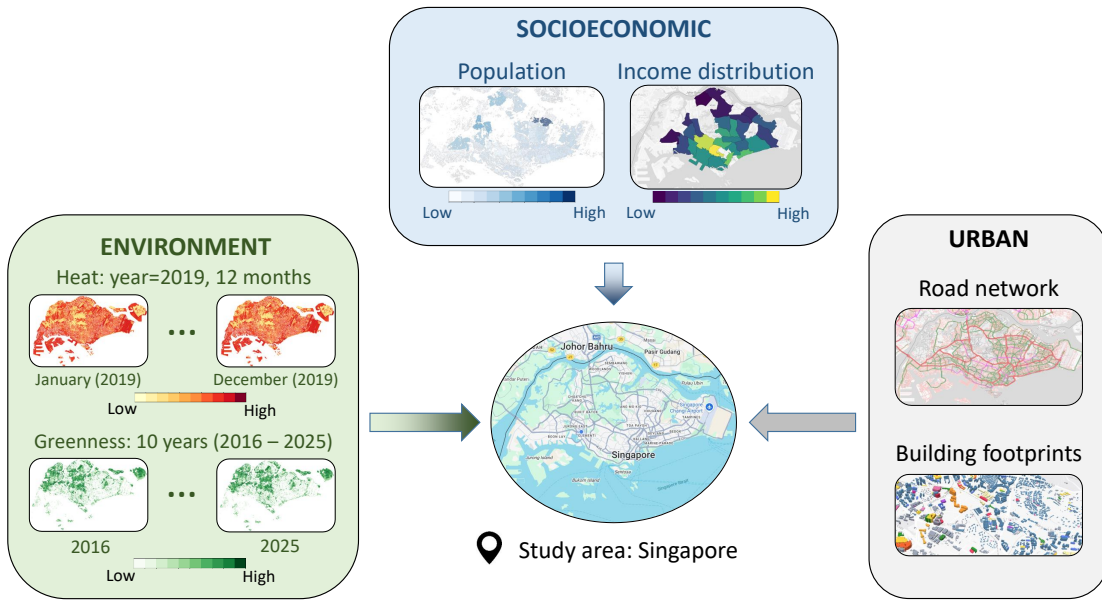


Figure 3: *Datasets used to characterise the urban environment of Singapore. Map snapshot obtained from Google Maps [20].*

To demonstrate the temporal representation supported by the proposed framework, multiple years of greenness datasets (2016–2025) and monthly heat datasets for a representative year (2019) were compiled. Different temporal resolutions were selected to reflect the assumed temporal characteristics of the underlying environmental processes. As urban greenness is assumed to change relatively slowly, annual datasets using a consistent processing and compositing workflow were considered sufficient to represent long-term vegetation conditions for each calendar year. By contrast, thermal conditions exhibit substantial seasonal variability, so monthly datasets were used to capture temporal changes in the urban thermal environment within a single year. The remaining urban morphology and socioeconomic datasets were represented using a single snapshot, as comparable time-series datasets were not available.

The 2016–2025 period was selected to align with the study period of the SG100K [50]

cohort, supporting future downstream epidemiological studies. The year 2019 was chosen for the heat datasets because it was among Singapore’s warmest years on record, providing a representative example of elevated thermal conditions [41].

4.1 Raster datasets

4.1.1 NDVI

Urban greenness has been widely associated with physical and mental health, thermal comfort, and quality of life, and is commonly quantified using satellite-derived vegetation indices such as the Normalised Difference Vegetation Index (NDVI) [56] and Enhanced Vegetation Index (EVI) [38, 39].

In this work, NDVI was used to quantify physical greenness exposure. Annual NDVI datasets with 10 m resolution were generated from Sentinel-2 Level-1C imagery [10, 12] for the period 2016–2025 using Google Earth Engine [21]. Cloud-contaminated observations were identified using the s2cloudless Sentinel-2 Cloud Probability product, with pixels having cloud probability $\geq 40\%$ excluded [62]. Explicit cloud-shadow projection was not applied, as comparative testing showed that it could remove genuine vegetated areas and introduce inconsistent artefacts across years. NDVI was calculated according to [40]:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \quad (1)$$

where NIR and Red denote the near-infrared and red spectral bands, respectively. A pixel-wise median annual composite was then produced to generate a representative greenness surface for each calendar year while reducing the influence of transient artefacts and temporal variability [25].

4.1.2 UTCI

The Universal Thermal Climate Index (UTCI) was used to represent heat exposure in this study. High-resolution UTCI datasets at a 4 m spatial resolution were generated using the methodology of Wang et al. [66], which integrates urban morphology, land-surface information, meteorological conditions, and human angular radiation factors to estimate pedestrian-level outdoor thermal stress across Singapore. Monthly UTCI rasters for 2019 were incorporated into TWA, with each raster representing the pixel-wise maximum UTCI for that month. Using the monthly maximum captures the highest estimated thermal stress at each location, providing a measure of potential peak heat exposure.

4.1.3 Population

Population data were obtained from the 2020 population distribution dataset provided by AI for Good at Meta [14, 63]. This 30 m resolution raster estimates the spatial distribution of the resident population by combining census information with satellite imagery and ancillary geospatial data. Population was included as a socioeconomic indicator to

investigate whether environmental exposures, including greenness, heat, and urban infrastructure, vary with the concentration of the resident population.

4.2 Vector datasets

Building footprints and road network data were obtained from OpenStreetMap [44]. These vector datasets were used to characterise urban morphology by deriving building coverage and total road length within the surrounding neighbourhood. Although OpenStreetMap provides extensive spatial coverage and frequent updates, its data quality varies across locations because of its crowdsourced nature.

4.3 Attribute datasets

4.3.1 BGVI

Building Green View Index (BGVI) was used to quantify visual greenness exposure. Unlike NDVI, which measures the physical presence of vegetation, BGVI measures the extent to which surrounding vegetation is visible from a building's perspective, reflecting the greenness experienced by building occupants [71].

BGVI was calculated using the ray-tracing approach developed by Feng and Yang [15]. Rays are cast from observer locations distributed along the building boundary, and the proportion of rays whose first intersection is vegetation is used to derive the index. Figure 4 illustrates the ray-tracing process, with rays whose first intersection is vegetation highlighted in green. BGVI ranges from 0 to 1, where 0 indicates that no vegetation is visible and 1 indicates that vegetation is visible in all directions. The calculation integrates building geometries, elevation, NDVI, and tree canopy data.



Figure 4: Ray tracing to calculate visual greenness.

BGVI estimates were generated for five representative floor groups in Singapore (floors 1–2, 3–6, 7–12, 13–22, and 23+). The case study in this paper used estimates for the first floor group (floors 1–2) only, since the choice of floor group does not affect the design of the framework.

4.3.2 Income distribution

Income distribution data were obtained from the 2020 population census published through Singapore’s open data portal [22]. The dataset reports the number of resident households in each planning area by monthly household income from work, using discrete income brackets (*e.g.*, below S\$1,000; S\$1,000–1,999; S\$2,000–2,999; etc.). Principal component analysis (PCA) was applied to reduce the dimensionality of these correlated income distributions, summarising them into a small number of components for subsequent analyses.

Household counts were converted to proportions within each planning area so that the PCA captured differences in income composition rather than differences in planning area population size. The proportion of households in each income category was standardised across planning areas to zero mean and unit variance before applying PCA so that categories with greater variance did not dominate the analysis solely because of their scale. The resulting principal components were used as indicators of neighbourhood socioeconomic characteristics.

Figure 5 presents the results of the PCA. The first two rows show the income distributions of four planning areas selected to represent distinct socioeconomic profiles: Bukit Timah (high affluence), Yishun (lower affluence), Outram (larger proportion of lower-income households), and Pasir Ris (stronger upper-middle- and high-income profile). The third row presents the PCA results. The first two principal components (PC1 and PC2) explain 83.3% of the total variance, supporting the use of a two-component representation. PC1 primarily summarises neighbourhood affluence, while PC2 contrasts lower-income with upper-middle- and high-income household profiles. The component loadings supporting these interpretations are presented in the Supplementary Information. The PCA score plot projects all planning areas onto the PC1–PC2 space, with the four example planning areas highlighted.

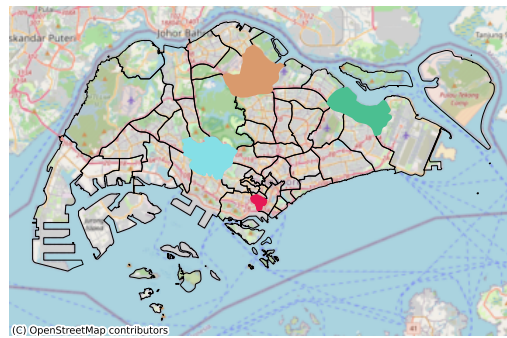
5 Exposure calculations

In the Singapore case study, the framework was applied to calculate urban environmental exposure measures for approximately 120,000 postal-code locations. In Singapore, six-digit postal codes identify individual delivery points at the level of a house or building. Postal-code locations were obtained from [19]. Within the framework, each postal-code location was represented as a `Subject` for which environmental exposure measures were calculated.

For each postal code, the following exposure metrics were derived and subsequently used in the later analyses:

- **Environment:** physical greenness, visual greenness, and heat exposure.
- **Urban:** neighbourhood building coverage and total road length.
- **Socioeconomic:** population and income PC1/PC2.

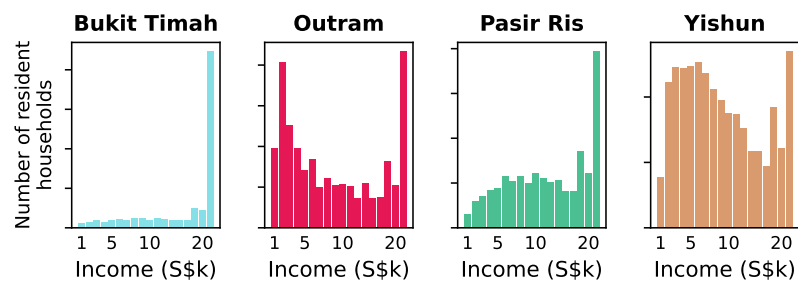
Planning areas



Selected areas

- Bukit Timah
- Outram
- Pasir Ris
- Yishun

Selected income distributions



PCA

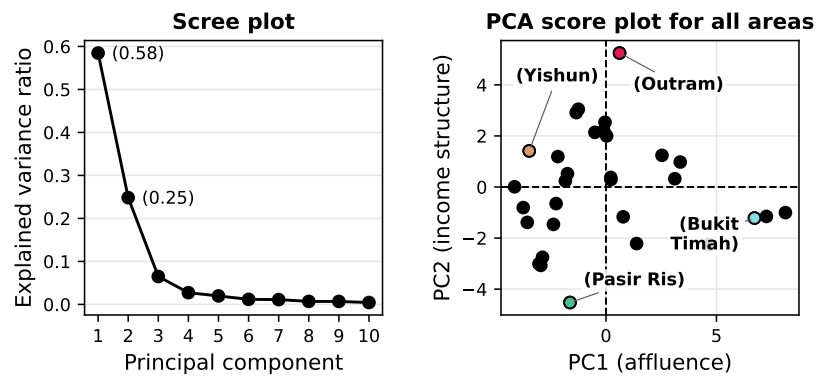


Figure 5: PCA of income distribution.

The methodology used to derive each exposure depends on the characteristics of the underlying dataset, including its spatial representation (*e.g.*, raster, vector, or attribute data) and the intended interpretation of the exposure metric. Table 1 summarises the datasets used in this study together with the corresponding calculation methods, which are described in detail in the following subsections.

Table 1: *Summary of exposure calculations.*

Exposure	Dataset	Dataset representation	Exposure calculation
Physical greenness	NDVI	Raster	Area proportion within buffer
Heat	UTCI	Raster	Mean value within buffer
Population	Population count	Raster	Sum within buffer
Building coverage	Building footprints	Vector	Area proportion within buffer
Road length	Road network	Vector	Total length within buffer
Visual greenness	BGVI	Attribute	Direct assignment
Income PC1/PC2	Income distribution	Attribute	Direct assignment

5.1 Neighbourhood-based exposure calculations

Unless otherwise stated, neighbourhood-based measures were calculated using a 400 m circular buffer centred on each postal-code point. The distance was selected to represent the walkable neighbourhood commonly used in environmental epidemiology studies [24].

5.1.1 Raster-based exposures

Physical greenness. Following annual compositing, pixels with $\text{NDVI} \geq 0.2$ were classified as vegetated before ingestion into the framework, consistent with thresholds commonly applied in tropical and Asian urban vegetation mapping studies [23, 27]. The exposure calculation then derived the proportion of vegetated area within the neighbourhood buffer surrounding each postal code. While the framework supports alternative physical greenness measures, this proportion was used as a representative NDVI-derived indicator for demonstrating the proposed semantic framework.

Heat. Heat exposure was quantified using the UTCI raster dataset. Heat exposure for each postal code was calculated as the spatial mean of the pixel-wise monthly maximum UTCI values within its 400 m buffer. Averaging is appropriate because UTCI is a continuous variable with physically meaningful values, allowing the pixel values within the buffer to be aggregated into a single exposure metric.

Population. Neighbourhood population was estimated by summing population values from raster cells within the buffer around each postal code. This measure approximates

the number of residents in the surrounding area, subject to the resolution and accuracy of the underlying population data.

5.1.2 Vector-based exposures

Building coverage. Building coverage was quantified as the proportion of the buffer occupied by building footprints. This was calculated by intersecting building polygons with the buffer, summing the areas of the resulting intersections, and normalising by the buffer area.

Road length. Road exposure was quantified as the total road length within the buffer. This was calculated by clipping road segments to the buffer boundary and summing the lengths of the resulting segments.

5.2 Direct assignment

Exposure values derived from datasets associated with individual spatial features were assigned directly to postal codes using a simple spatial intersection.

Visual greenness. Visual greenness was estimated at the building level. As the building and postal-code datasets were obtained from different sources and did not share a common identifier, each postal-code point was associated with its corresponding building through spatial intersection. The building-level BGVI value was then assigned to the postal code.

Income distribution. Income distribution data were available at the planning area level rather than the postal-code level. Each postal code was therefore assigned the corresponding PC1 and PC2 values (Section 4.3.2) of the planning area in which it was located.

6 Results

The results address two questions: first, whether the proposed representation can accommodate heterogeneous exposure calculations at city scale; and second, whether the resulting measures can be retrieved with their calculation specifications and traced to their source data. We begin by illustrating the common semantic representation of exposure results. We then examine the spatial distributions and relationships among measures generated for approximately 120,000 Singapore postal-code locations. The final examples demonstrate retrieval using specified exposure criteria, tracing individual values to their inputs, and comparing estimates calculated using alternative buffer sizes.

6.1 Semantic representation of exposure results

Figure 6 shows the semantic representation of three exposure calculations from the Singapore case study, using raster, vector and attribute data. Although the underlying datasets and calculation methods differ, each exposure is represented consistently through linked

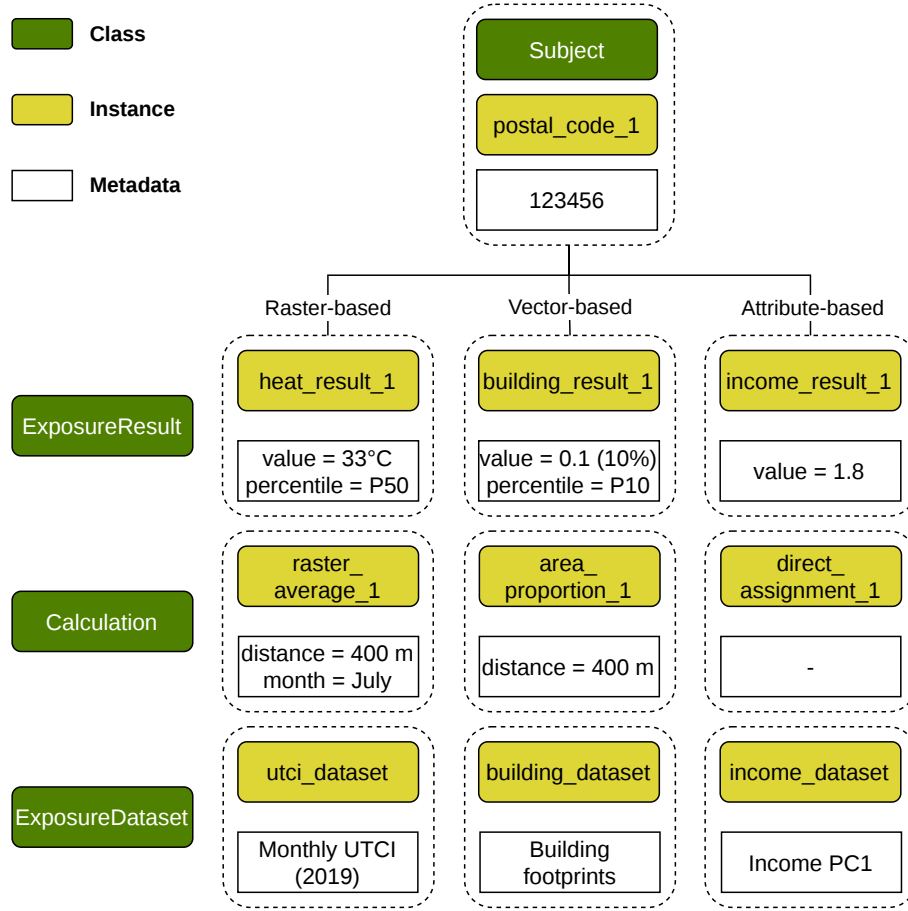


Figure 6: *Illustrative example showing how multiple environmental exposures associated with a single subject are represented using a common semantic structure.*

Subject, ExposureDataset, Calculation and ExposureResult instances. Where needed, calculation-specific metadata, such as buffer distance and temporal parameters, are associated with the corresponding Calculation instance. These examples demonstrate that exposure measures derived from different data structures and calculation methods can share the same subject–dataset–calculation–result representation while retaining calculation-specific information.

More detailed representations and additional explanations of the proposed framework are provided in the Supplementary Information.

6.2 Generation of analysis-ready exposure results

The purpose of the following analysis is to demonstrate that the framework generates spatially resolved, epidemiologically interpretable exposure variables across multiple domains, rather than to establish new environmental associations.

Figures 7 and 8 summarise the exposure values generated for approximately 120,000 postal codes across Singapore, showing their distributions and spatial patterns. The frame-

work can manage datasets covering multiple time points. The results presented here use the 2019 NDVI and BGVI datasets and the July 2019 UTCI dataset as representative examples.

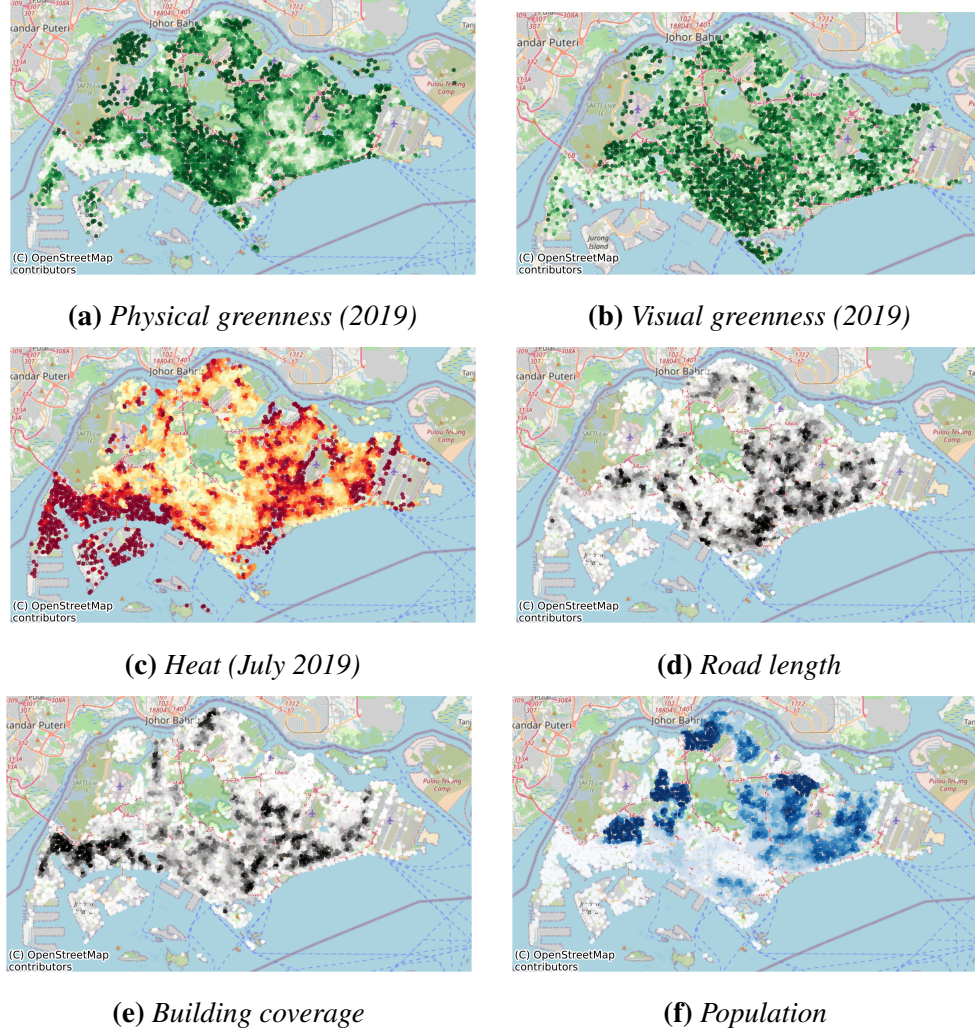


Figure 7: Spatial distribution of the environmental, urban morphology, and socioeconomic exposure metrics across Singapore. Each map uses the same colour scale as the corresponding histogram in Figure 8.

Figure 7 shows the spatial distribution of the exposure values, with each postal code represented by a point coloured according to its value. Figure 8 shows the distribution of these values across all postal codes, with P10, P50 and P90 denoting the 10th, 50th (median) and 90th percentiles. A consistent colour scheme is used in both figures to facilitate comparison. Income distribution is omitted because the socioeconomic indicators (PC1 and PC2) have already been characterised through the PCA analysis presented in Figure 5. Plotting their distributions across postal codes would largely reflect the repeated assignment of planning area values rather than provide additional insight.

The generated exposure metrics exhibit different distributions and spatial patterns, re-

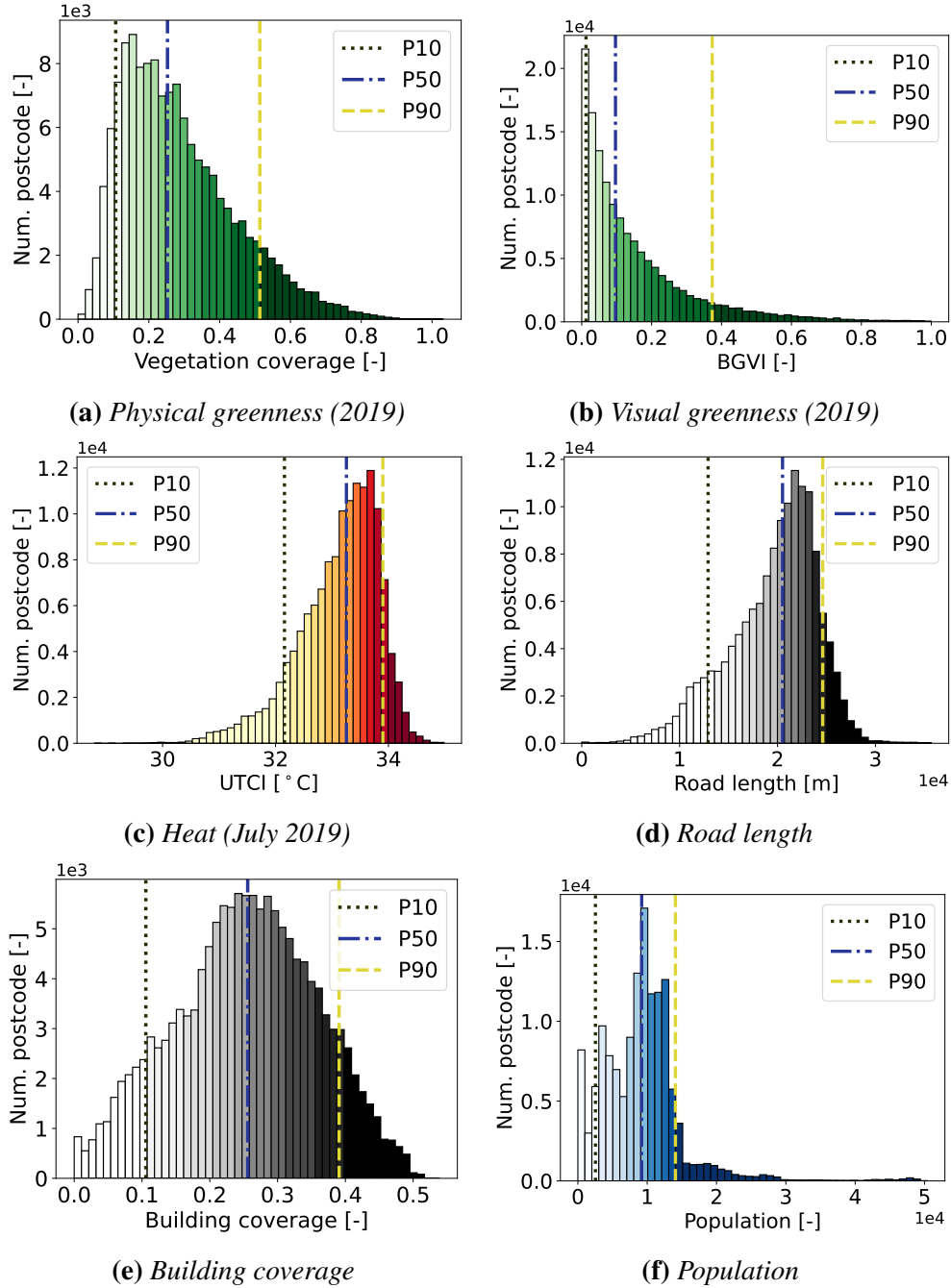
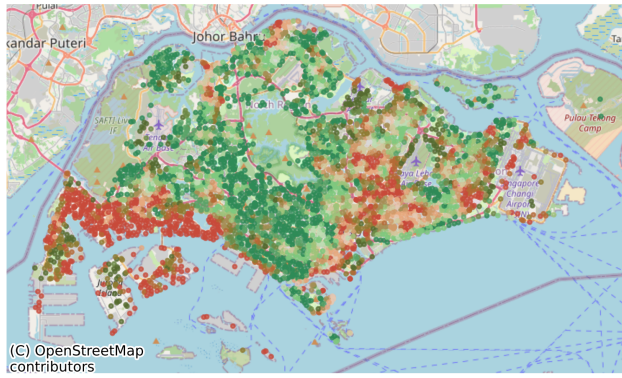


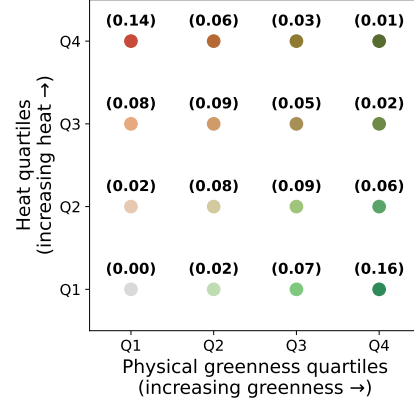
Figure 8: Histograms showing the distribution of environmental, urban morphology, and socioeconomic exposure metrics used in this study.

flecting the diversity of the underlying environmental characteristics. For example, by comparing Figures 7a and 7c, areas with high physical greenness generally correspond to lower heat exposure, suggesting an inverse relationship between these two variables. While Figures 7 and 8 present each exposure independently, environmental exposures are often interrelated and should therefore be considered jointly.

Figure 9 therefore illustrates the joint spatial distribution of heat exposure and physical



(a) Spatial distribution of postal codes classified according to the combined quartiles of heat exposure and physical greenness.



(b) Colour scheme used in Figure 9a. Numbers in parentheses indicate the proportion of postal codes within each heat–greenness quartile combination.

Figure 9: Joint classification of postal codes according to heat exposure and physical greenness.

greenness by classifying postal codes according to the quartiles of both variables. Nearly half of all postal codes fall along the diagonal corresponding to higher greenness with lower heat and lower greenness with higher heat, illustrating a strong inverse relationship between these two exposure metrics.

Together, these results demonstrate that the framework generates exposure measures with distinct spatial distributions and supports their joint examination across Singapore. The following section examines relationships among the generated measures.

6.3 Relationships between exposure results

While Figure 9 qualitatively illustrates the relationship between heat and physical greenness through their joint spatial distribution, the following analyses examine relationships across all exposure metrics simultaneously. These analyses provide a descriptive check of whether the generated measures retain interpretable urban–environmental relationships and illustrate their use in joint exposure analysis.

Figure 10 presents a correlation matrix based on Spearman’s rank correlation coefficient, providing an overview of pairwise associations between the generated exposure datasets. Correlation coefficients range from -1 to 1 , where positive values indicate that two variables tend to increase together, negative values indicate an inverse relationship, and values closer to ± 1 represent stronger associations. Among all pairwise comparisons, the largest absolute Spearman correlation coefficients are observed between the greenness metrics (physical and visual) and heat exposure, between the greenness metrics and building coverage, and between population and Income PC1.

Physical and visual greenness are strongly correlated but have different distributions (Fig-

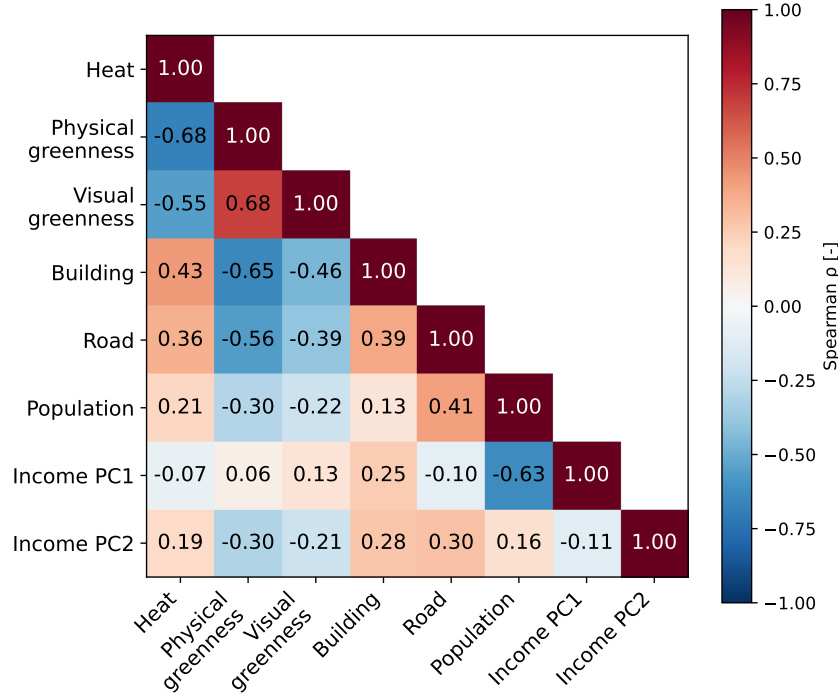


Figure 10: Correlation heatmap between different exposures.

ures 8a and 8b), with visual greenness more concentrated near zero. They also represent different aspects of exposure: physical greenness measures neighbourhood vegetation coverage, whereas visual greenness depends on vegetation arrangement and occlusion by the built environment. Previous comparisons similarly show that visible-greenery measures capture information not represented by NDVI [34].

The negative association between greenness and heat is consistent with the cooling effects of urban vegetation reported in previous studies [16, 67]. Greenness is also negatively correlated with building coverage, reflecting the competing spatial requirements of vegetation and the built environment. The negative correlation between population and Income PC1 indicates that more affluent planning areas tend to contain postal-code locations with lower neighbourhood population counts.

These relationships illustrate how the generated datasets can be analysed jointly while preserving the distinct interpretation of each exposure measure.

6.4 Semantic retrieval and provenance

The principal methodological benefit of the framework is not merely that it calculates exposure values, but that the resulting values remain computationally connected to their source datasets, calculation methods and parameters. The following examples demonstrate how these links support the selection of locations meeting specified exposure criteria, the tracing of individual results to their inputs, and the comparison of estimates calculated using different buffer sizes.

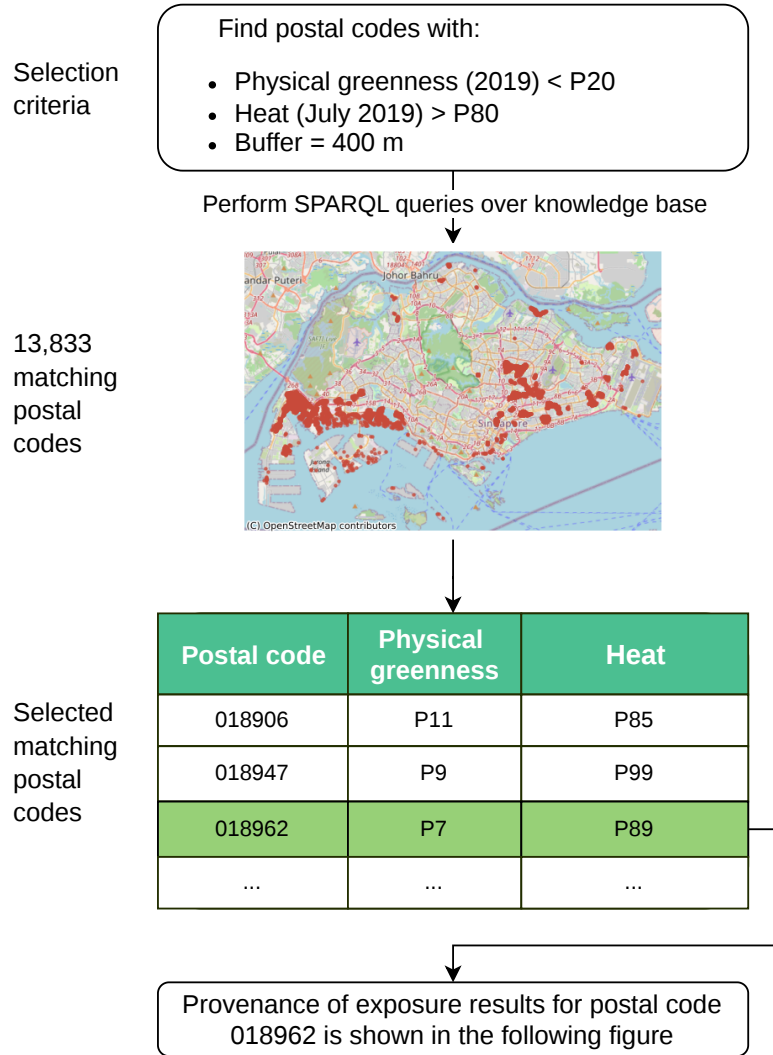


Figure 11: Retrieving postal codes satisfying the specified exposure criteria. The highlighted postal code, 018962, is examined in Figure 12.

Figure 11 demonstrates how an epidemiological query is evaluated against the knowledge base to identify postal codes satisfying the specified exposure criteria. In this example, the query retrieves postal codes with physical greenness in 2019 below the 20th percentile and heat exposure in July 2019 above the 80th percentile, using a 400 m buffer. The query returns 13,833 matching postal codes, with selected records also shown.

Figure 12 traces the exposure results for postal code 018962, highlighted in Figure 11. The postal code is represented as a Subject, whose `ExposureResult` instances are linked to the corresponding `ExposureDataset` and `Calculation` instances. The `Calculation` instance labelled `raster_average_400` computes the mean UTCI within the 400 m buffer, whereas `raster_area_400` computes the proportion of the 400 m buffer area classified as vegetated in the NDVI dataset. These links retain the source datasets, temporal context and calculation parameters, allowing researchers to retrieve each value together with the information needed to interpret it.

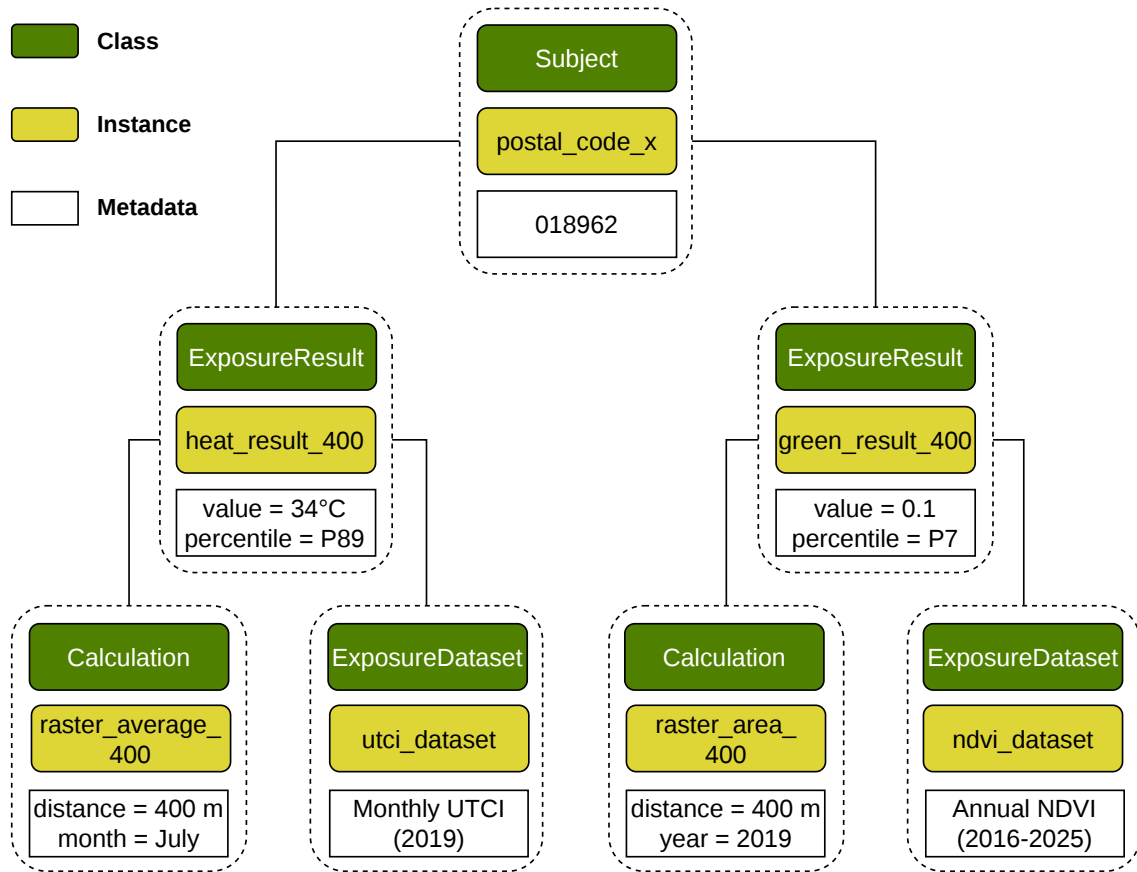


Figure 12: *Tracing the provenance of exposure values for a retrieved subject.*

Figure 13 demonstrates how alternative exposure specifications can be compared while retaining their provenance. For postal code 018962, heat exposure was recalculated by increasing the buffer radius from 400 m to 800 m, using the same 2019 UTCI dataset and July selection. This changes the estimated UTCI by less than 0.02 °C, with both values rounding to 34 °C, but raises the percentile rank from approximately P89 to P96 within the respective buffer-specific distributions. The two `Calculation` instances specify the same raster-averaging method but different buffer radii. Each `ExposureResult` instance retains its value and percentile rank and is linked to the same subject and dataset and to its respective calculation instance. These links allow the estimates to be retrieved and compared while preserving the calculation specification associated with each result.

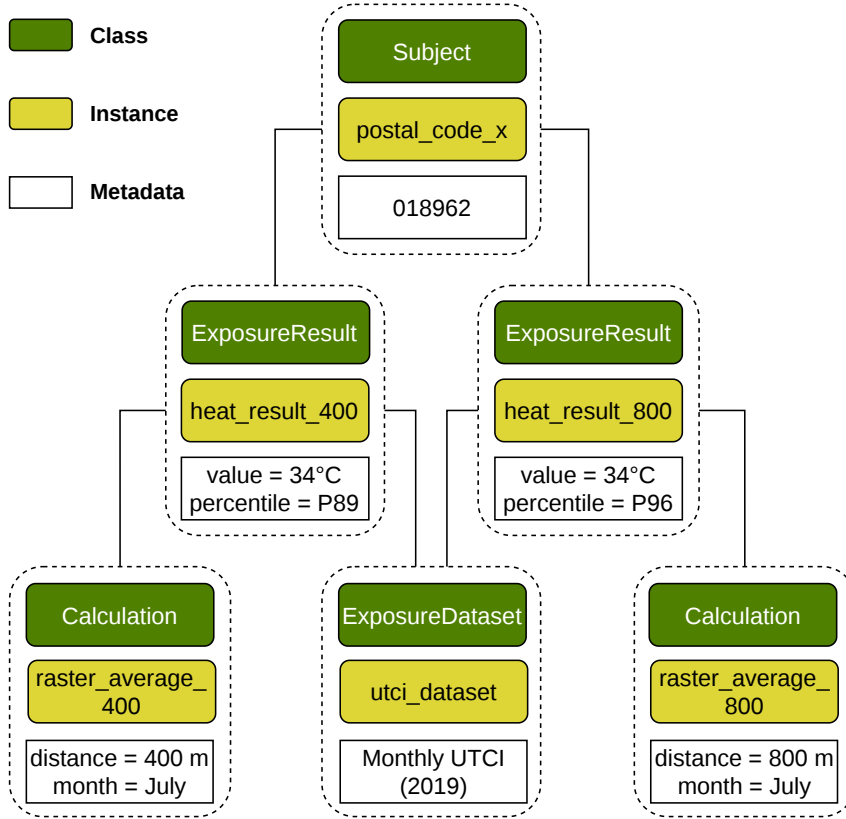


Figure 13: Comparison of heat exposure calculated using 400 m and 800 m buffers for postal code 018962. Percentiles were calculated separately for each buffer radius using the same set of postal codes.

7 Limitations and future work

The framework standardises the semantic representation, management and retrieval of environmental exposure data, while preserving links between derived exposure metrics and their source datasets. The quality, accuracy and uncertainty of these metrics therefore depend on the underlying data, including any limitations or biases they contain. Evaluating dataset suitability remains a domain-specific task for data providers and subject-matter experts, rather than a function of the framework itself.

A related consideration is the temporal coverage of the source data. The framework records the temporal context of each dataset and exposure result, allowing observations to be queried by time and compared across periods. This supports the selection of exposure data for specific periods and for longitudinal analyses. In the Singapore case study, the source datasets cover different years and time periods, so each exposure metric should be interpreted with reference to its recorded temporal coverage rather than as part of a temporally synchronised snapshot.

In this paper, provenance covers the source dataset, temporal context, calculation method and parameters associated with each result. The proposed framework captures the semantic relationships between environmental datasets, exposure calculations and exposure

results, but does not currently record software versions, library dependencies, execution environments or workflow execution histories. Existing tools could capture aspects of this execution-level provenance. For example, Apache Hamilton [33] could define and execute the calculation workflows, while OpenLineage [43] could record individual executions. Integrating these tools is beyond the scope of this work. General computational scalability, interoperability with external semantic systems and the use of the generated datasets in downstream epidemiological analyses are also beyond the scope of this work.

A practical consideration in adopting semantic knowledge graphs is the initial effort required to construct and maintain them. Ontology development, the mapping of heterogeneous datasets and the instantiation of knowledge graph entities continue to require domain expertise. However, tools that reduce this effort are an active area of research. Recent work within TWA has demonstrated how existing ontologies can be compiled into executable tools that guide large language model agents in extracting, validating and instantiating structured knowledge [74]. Natural-language interfaces can similarly translate user requests into SPARQL queries, making semantic data more accessible to non-expert users [11, 54]. Together, these developments offer promising routes to reducing the effort required to populate and access semantic knowledge graphs.

The present study considers fixed postal-code locations. Future work will extend the explicit representation of calculation specifications and their links to exposure results to subjects whose locations change over time. Building on the trajectory-based assessment demonstrated by Chen et al. [9], this would allow exposures along individual movement trajectories to be retrieved and compared together with the data, methods and parameters used to derive them.

8 Conclusions

We developed a knowledge-driven framework that represents environmental exposure calculation methods and parameters separately from their source datasets and links each result to the subject, data and calculation specification used to derive it. This common semantic structure accommodates raster, vector and attribute-based exposure measures while preserving their methodological differences. Spatial operations are performed directly in PostGIS, and the resulting values remain linked to their semantic inputs.

The Singapore case study demonstrated this approach across environmental, urban-morphology and socioeconomic measures for approximately 120,000 postal-code locations. Exposure profiles could be retrieved using specified exposure criteria and time periods, individual values traced to their source datasets and calculation specifications, and estimates compared under alternative buffer sizes. These results demonstrate the generation of analysis-ready exposure datasets at city scale with explicit records of the source data, temporal context, methods and parameters needed to interpret each value.

For epidemiological research, these links make it possible to select exposure measures with the required temporal coverage and calculation settings, and to distinguish estimates derived under different assumptions. Researchers can therefore assemble analysis-ready datasets while retaining access to the methodological information needed to interpret and

compare the selected measures. Future work will link these exposure measures to longitudinal health data from participants in the SG100K study [50].

Nomenclature

BGVI Building green view index

GIS Geographic Information Systems

NDVI Normalised difference vegetation index

PC1 First principal component of PCA

PC2 Second principal component of PCA

PCA Principal component analysis

RDF Resource description framework

SPARQL SPARQL Protocol and RDF Query Language

SQL Structured query language

TWA The World Avatar

UTCI Universal Thermal Climate Index

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Code availability

The software repositories used to calculate the exposure measures are available at the following links:

- Exposure calculation agent: <https://github.com/TheWorldAvatar/exposure-calculation-agent>
- TWA stack: <https://github.com/TheWorldAvatar/stack>
- Configuration to host HD4 data on TWA stack: <https://github.com/TheWorldAvatar/hd4-stack>

A Supplementary Information

A.1 Income PCA

Figure 14 shows the contribution of each income category to the two principal components. Larger absolute loadings indicate a greater contribution. The highest income categories contribute most strongly to PC1, supporting its interpretation as a neighbourhood affluence score. PC2 contrasts lower-income categories with upper-middle- and high-income categories, so it captures differences in household income patterns across planning areas.

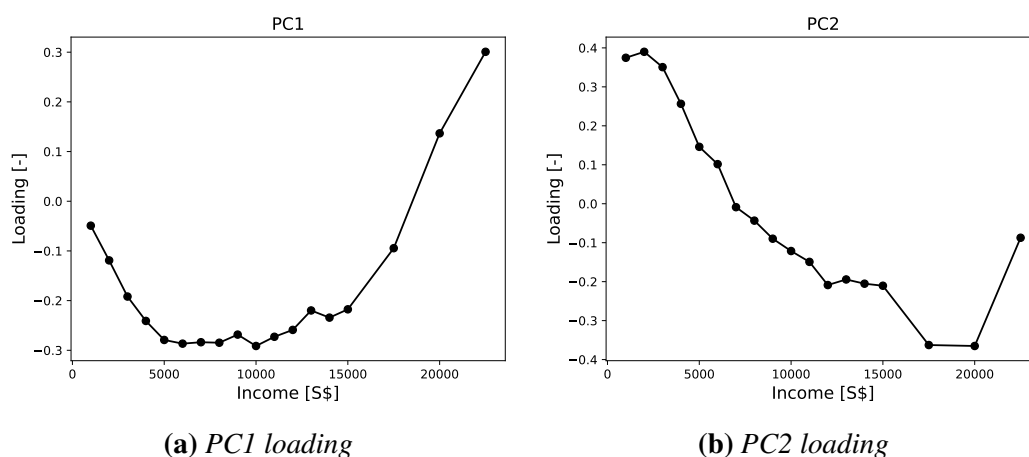


Figure 14: Component loadings from the PCA.

A.2 Ontology representation and example queries

The remainder of this section explains how the knowledge graph represents the main parts of an exposure assessment. It begins with subjects (Section A.2.2) and exposure datasets (Section A.2.3), then describes calculations (Section A.2.4) and exposure results (Section A.2.5). The corresponding ontology classes are `exp:Subject`, `exp:ExposureDataset`, `exp:Calculation`, and `exp:ExposureResult`. Together, these classes represent the location being assessed, the environmental data, the operation applied to those data and the resulting exposure value.

The figures distinguish between two parts of the ontology: the TBox and the ABox. The TBox defines the general structure, including the classes, properties and relationships. The ABox records specific instances of those classes and the relationships between them. The TBox figures therefore show the general model, while the ABox figures show how the model represents particular subjects, datasets, calculations and results.

Section A.2.1 introduces the namespace prefixes used in the figures and example queries. Each later section explains the role of one component, shows how the ontology represents it and, where relevant, provides an example SPARQL query for retrieving the information.

A.2.1 Notation and namespaces

The ontology diagrams and SPARQL queries use the namespace prefixes listed below. A prefix is a short form of a URI and identifies the vocabulary that defines a term. For example, `exp:Subject` refers to the `Subject` class in the exposure ontology, while `geo:Feature` refers to the `Feature` class in GeoSPARQL. Throughout this section, typewriter text marks formal ontology identifiers, including classes, properties and instances.

```
dcat: <http://www.w3.org/ns/dcat#>
dcterms: <http://purl.org/dc/terms/>
der: <https://www.theworldavatar.com/kg/ontoderivation/>
exp: <https://www.theworldavatar.com/kg/ontoexposure/>
geo: <http://www.opengis.net/ont/geosparql#>
owl: <http://www.w3.org/2002/07/owl#>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
xsd: <http://www.w3.org/2001/XMLSchema#>
```

A.2.2 Subject

An exposure assessment begins by identifying the subject: the person, place or other entity whose exposure will be calculated. The subject provides the location used in the spatial calculation. In the present study, the subjects are fixed postal-code locations represented as instances of `exp:Subject`.

Figure 15 shows how the ABox represents `exp:postal_code_1`, an instance of `exp:Subject` for the postal code ‘123456’. The postal code is represented by a separate entity with properties such as the postal code value and address. The figure omits these relationships because the exposure calculation does not use them. `exp:Subject` is a subclass of `geo:Geometry`, so it inherits the properties of `geo:Geometry`. The exposure calculation agent only needs the geometry as a Well-Known Text (WKT) literal, which it accesses through the `geo:asWKT` property.

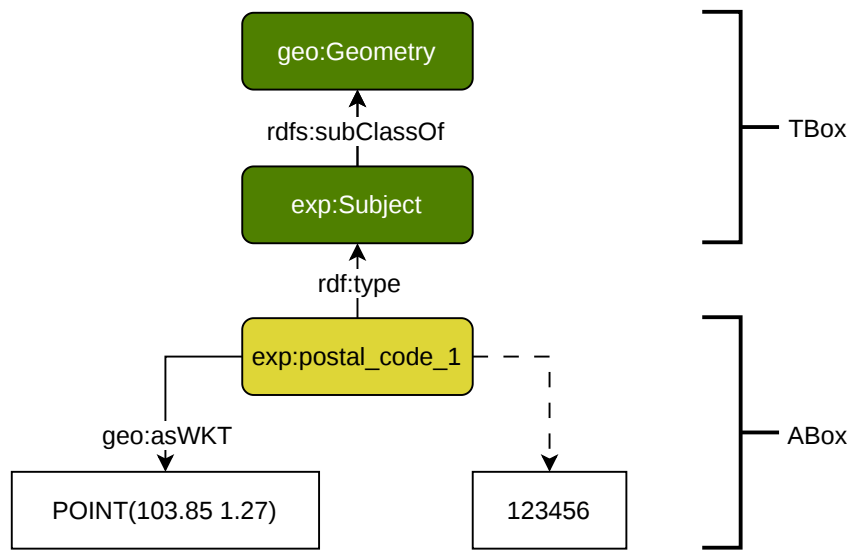


Figure 15: ABox representation of an `exp:Subject` instance.

Figure 16 shows an example query used by the exposure calculation agent to retrieve the WKT literal for `exp:postal_code_1`:

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX exp: <https://www.theworldavatar.com/kg/ontoexposure/>
SELECT ?wkt
WHERE {
  exp:postal_code_1 geo:asWKT ?wkt .
}
```

Figure 16: Example query to retrieve a subject’s location.

A.2.3 Exposure dataset

Once the subject has been identified, the framework specifies the environmental data used to calculate the exposure. An exposure dataset identifies the source data and records the metadata needed to find and interpret them. Exposure datasets are represented as instances of `exp:ExposureDataset`.

Figure 17 shows how the ABox represents `exp:utci_dataset`, an instance of `exp:ExposureDataset` for the UTCI data used in the case study. `exp:ExposureDataset` is a subclass of `dcat:Dataset`. It can therefore use the metadata properties defined by DCAT [68], including dataset descriptions and access URLs. This makes it possible to represent datasets stored in different locations or accessed through different services. The current exposure calculation agent only uses the `dcterms:title` property, which identifies the corresponding dataset in PostGIS.

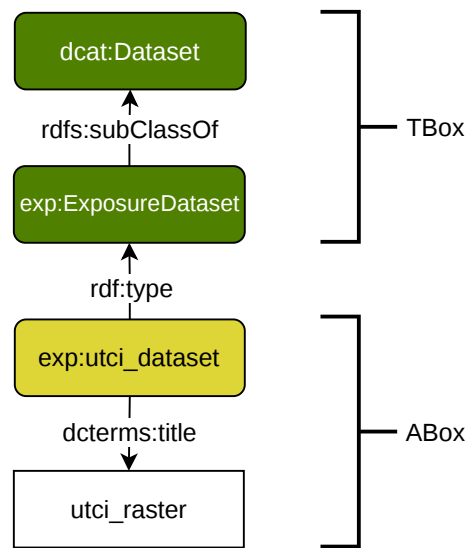


Figure 17: ABox representation of an `exp:ExposureDataset` instance.

Figure 18 shows an example query used by the exposure calculation agent to identify the dataset stored in PostGIS:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX exp: <https://www.theworldavatar.com/kg/ontoexposure/>
SELECT ?dataset
WHERE {
    exp:utci_dataset dcterms:title ?dataset .
}
```

Figure 18: Example query to identify an exposure dataset in PostGIS.

A.2.4 Calculation

With the subject and exposure dataset identified, the calculation specifies how to derive the exposure value. It records the spatial operation and its parameters separately from the data on which the operation is performed. The same calculation method can therefore be described consistently and reused with different subjects or datasets.

Figure 19 shows how the TBox represents `exp:Calculation`, including its method and parameters. `exp:Subject` and `exp:ExposureDataset` extend classes from established ontology standards: GeoSPARQL [42] and DCAT [68], respectively. By contrast, `exp:Calculation` is defined specifically for environmental exposure assessment. Subclasses of `exp:Calculation`, such as `exp:RasterAverage` and `exp:RasterSum`, tell the exposure calculation agent which spatial operation to perform. Other properties define the calculation parameters. For example, `exp:hasDistance` gives the buffer distance, while `exp:hasDatasetFilter` links a calculation to an `exp:DatasetFilter` instance. The `exp:DatasetFilter` identifies the attribute to filter, such as `month`, and the required value, such as `July`. Multiple time-specific datasets, including monthly UTCI rasters, can therefore be stored in one PostGIS dataset. The agent uses the semantic metadata to select the required subset.

Under the current design, each new calculation method requires a corresponding subclass of `exp:Calculation` and an implementation in the exposure calculation agent. However, the set of spatial operations used in environmental exposure assessment is expected to remain manageable. Defining a subclass for each method makes the method explicit, while shared ontology properties provide a consistent way to record common parameters.

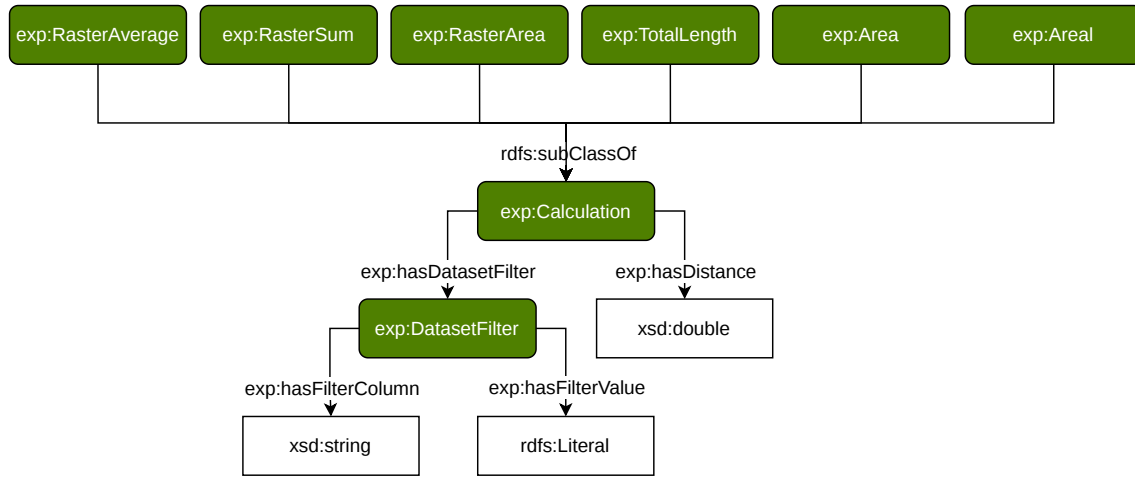


Figure 19: TBox representation of `exp:Calculation`.

Figure 20 shows how the ABox uses this structure to represent `exp:raster_average_calculation_1`. The `rdf:type` statement `exp:RasterAverage` tells the exposure calculation agent to perform a raster average operation. The calculation is linked to `exp:dataset_filter_1`, which tells the agent to select rows where the ‘month’ attribute is ‘July’. Table 2 shows example rows from the corresponding UTCI raster table. The table stores several monthly rasters and uses the ‘month’ column to distinguish between them. The `exp:hasDistance` property gives the buffer radius used in the calculation.

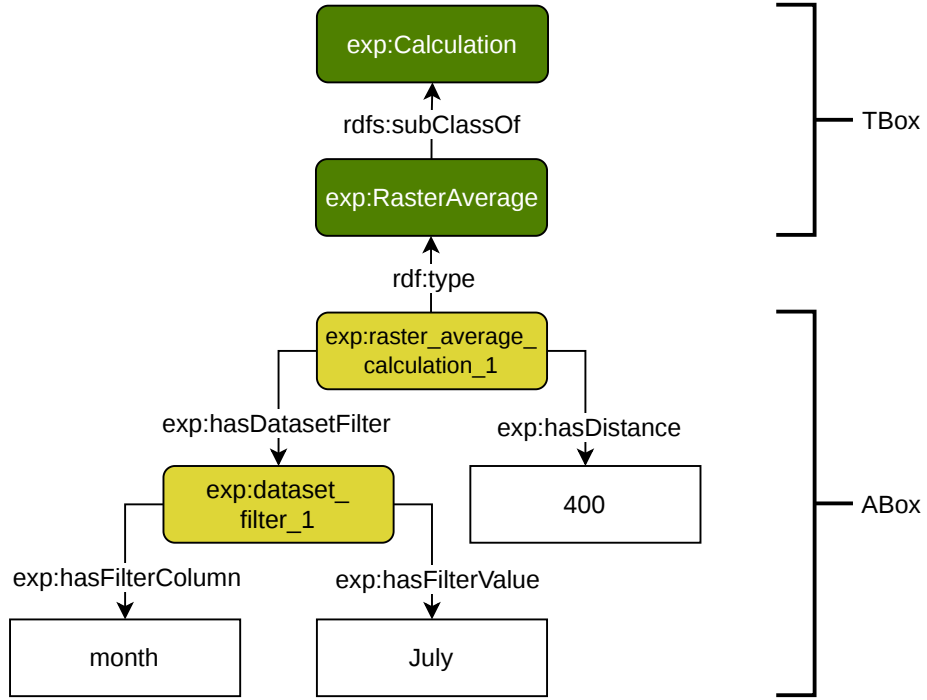


Figure 20: ABox representation of an *exp:Calculation* instance.

Table 2: Illustrative rows from the UTCI raster table.

rid	rast	month
1	0100000100268C1EE87D8C17...	July
2	0100000100268C1EE87D8C17...	January

Figure 21 shows an example query used by the exposure calculation agent to retrieve the metadata for *exp:raster_average_calculation_1*:

```

PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX exp: <https://www.theworldavatar.com/kg/ontoexposure/>
SELECT ?rdf_type ?buffer_distance ?filter_column ?filter_value
WHERE {
  exp:raster_average_calculation_1 a ?rdf_type ;
  exp:hasDistance ?buffer_distance;
  exp:hasDatasetFilter ?dataset_filter.
  ?dataset_filter exp:hasFilterColumn ?filter_column;
  exp:hasFilterValue ?filter_value.
}
  
```

Figure 21: Example query to retrieve the parameters of a calculation.

A.2.5 Exposure result

An exposure result links the subject, exposure dataset and calculation described in the preceding sections. It records the exposure value and the information needed to determine where it applies, which data were used and how it was calculated.

Figure 22 shows how the TBox represents `exp:ExposureResult` and links it to the other parts of the exposure assessment. `exp:ExposureResult` uses the `der:Derivation` class proposed by Bai et al. [2], making it compatible with the derived information framework. An `exp:ExposureResult` records the exposure value, its unit and its percentile. It also links to the relevant `exp:Calculation`, `exp:Subject` and `exp:ExposureDataset`. The raw value is retained, while the percentile provides a normalised measure that may be more useful in later epidemiological analyses.

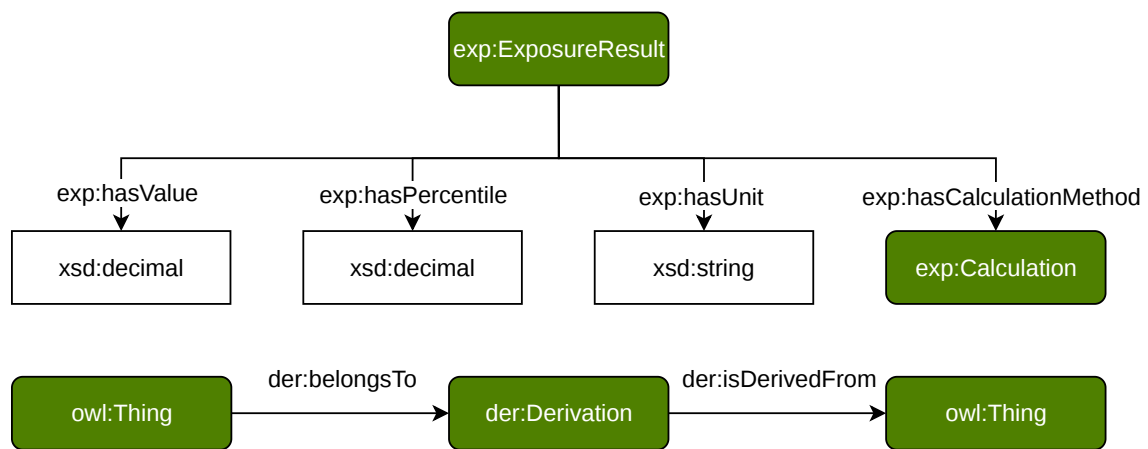


Figure 22: TBox representation of `exp:ExposureResult`.

Figure 23 shows how the ABox represents `exp:heat_result_1`, an instance of `exp:ExposureResult`. The links to the `exp:Calculation`, `exp:Subject` and `exp:ExposureDataset` instances preserve the provenance of the exposure value. Users can follow these links with SPARQL queries to determine how the value was produced. Storing the value, unit, percentile and provenance together makes environmental exposure datasets easier to trace, interpret and compare for later epidemiological analyses.

Figure 24 shows an example query that retrieves subjects (postal codes) above the 90th percentile for heat exposure, using `exp:raster_average_calculation_1` and `exp:utci_dataset`:

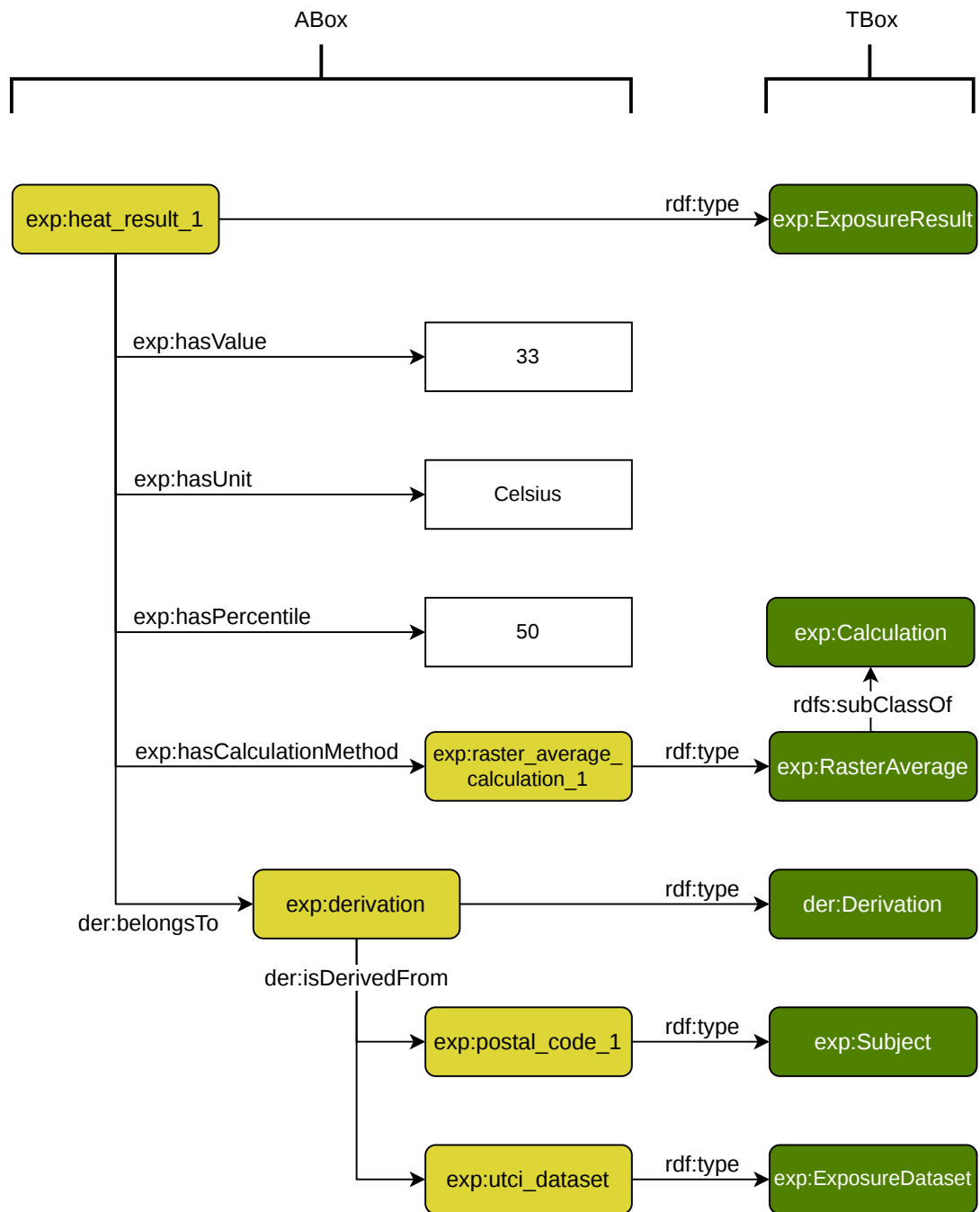


Figure 23: ABox representation of an `exp:ExposureResult` instance.

```

PREFIX der: <https://www.theworldavatar.com/kg/ontoderivation/>
PREFIX exp: <https://www.theworldavatar.com/kg/ontoexposure/>
SELECT ?subject
WHERE {
  ?derivation der:isDerivedFrom exp:utci_dataset;
  der:isDerivedFrom ?subject.
  ?result exp:hasCalculationMethod exp:raster_average_calculation_1;
  der:belongsTo ?derivation;
  exp:hasPercentile ?percentile.
  FILTER (?subject != exp:utci_dataset && ?percentile > 90)
}

```

Figure 24: *Example query to retrieve postal codes above the 90th percentile for heat exposure.*

References

- [1] J. Akroyd, S. Mosbach, A. Bhawe, and M. Kraft. Universal Digital Twin - A Dynamic Knowledge Graph. *Data-Centric Engineering*, 2:e14, 2021. doi:10.1017/dce.2021.10.
- [2] J. Bai, K. F. Lee, M. Hofmeister, S. Mosbach, J. Akroyd, and M. Kraft. A derived information framework for a dynamic knowledge graph and its application to smart cities. *Future Generation Computer Systems*, 152:112–126, 2024. doi:10.1016/j.future.2023.10.008.
- [3] K. Bereta, G. Stamoulis, and M. Koubarakis. Ontology-based data access and visualization of big vector and raster data. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 407–410. IEEE, 2018.
- [4] K. Bereta, G. Xiao, and M. Koubarakis. Ontop-spatial: Ontop of geospatial databases. *Journal of Web Semantics*, 58:100514, 2019.
- [5] Blazegraph. Blazegraph, 2020. URL <https://blazegraph.com>. Accessed: 1 July 2024.
- [6] P. W. Butler, S. D. Rihm, S. Mosbach, J. Akroyd, and M. Kraft. Leveraging general molecular fragments to expand the design space of metal–organic polyhedra. *Physical Chemistry Chemical Physics*, 2026. doi:10.1039/d6cp02069k.
- [7] Cambridge CARES. HD4 Project, 2026. URL <https://www.cares.cam.ac.uk/research/hd4-project/>. Accessed: June 2026.
- [8] J. Chen, J. Bai, J. Xu, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Transforming building retrofits: Linking energy, equity, and health insights from The World Avatar. *Advances in Applied Energy*, page 100230, 2025.
- [9] J. Chen, J. Akroyd, S. Mosbach, F. Farazi, X. Deng, K. F. Lee, M. Q. Lim, G. Brownbridge, T. Bishop, W. Cunningham, et al. Assessment of individualised dynamic environmental exposures within The World Avatar. *Computers, Environment and Urban Systems*, 129:102476, 2026. doi:10.1016/j.compenvurbsys.2026.102476.
- [10] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, and C. Justice. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219:145–161, 2018. doi:10.1016/j.rse.2018.09.002.
- [11] X. Deng, Y.-K. Tsai, S. Ganguly, D. Tran, H. Y. Quek, W. Ang, S. Z. Phua, S. Mosbach, J. Akroyd, and M. Kraft. Question answering system for smart cities and urban planning with The World Avatar. *IET Smart Cities*, 7(1):e70009, 2025. doi:10.1049/smc2.70009.
- [12] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: ESA’s Optical High-Resolution Mission for

- GMES Operational Services. *Remote Sensing of Environment*, 120:25–36, 2012. doi:10.1016/j.rse.2011.11.026. The Sentinel Missions - New Opportunities for Science.
- [13] Eclipse Foundation. Eclipse RDF4J, 2024. URL <https://rdf4j.org/>. Accessed: 1 July 2024.
- [14] Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. High resolution settlement layer (HRSL), 2016. URL <https://data.humdata.org/dataset/singapore-high-resolution-population-density-maps-demographic-estimates>. Source imagery for HRSL © 2016 DigitalGlobe. Accessed: July 2026.
- [15] Y. Feng and F. Yang. Developing a voxel-based sightline sampling algorithm for calculating panoramic visible green index in high-density urban environment. *Journal of Building Design and Environment*, 1(2):15556–15556, 2023. doi:10.37155/2811-0730-0201-5.
- [16] L. S. Ferreira and D. H. S. Duarte. Exploring the relationship between urban form, land surface temperature and vegetation indices in a subtropical megacity. *Urban Climate*, 27:105–123, 2019. doi:10.1016/j.uclim.2018.11.002.
- [17] Z. Gao, M. Gao, C.-h. Chen, Y. Zhou, Z.-H. Zhan, and Y. Ren. Knowledge graph of wastewater-based epidemiology development: A data-driven analysis based on research topics and trends. *Environmental Science and Pollution Research*, 30(11):28373–28382, 2023. doi:10.1007/s11356-023-25237-9.
- [18] A. Ghosh, A. Pano, G. Xiao, and D. Calvanese. OntoRaster: Extending VKGs with raster data. In *International Joint Conference on Rules and Reasoning*, pages 108–123. Springer, 2024.
- [19] Github. singapore-postal-codes-1, 2020. URL <https://github.com/isen-ng/singapore-postal-codes-1/tree/master>. Accessed: 1 July 2024.
- [20] Google. Google Maps, 2026. URL <https://maps.google.com/>. Accessed: 1 September 2026.
- [21] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017. doi:10.1016/j.rse.2017.06.031.
- [22] Government Technology Agency of Singapore. Singapore’s Open Data Portal, 2024. URL <https://data.gov.sg/>. Accessed: 1 July 2024.
- [23] S. Guha. Dynamic seasonal analysis on LST-NDVI relationship and ecological health of Raipur City, India. *Ecosystem Health and Sustainability*, 7(1):1927852, 2021. doi:10.1080/20964129.2021.1927852.
- [24] L. D. Gunn, T. L. King, S. Mavoa, K. E. Lamb, B. Giles-Corti, and A. Kavanagh. Identifying destination distances that support walking trips in local neighborhoods. *Journal of transport & health*, 5:133–141, 2017. doi:10.1016/j.jth.2016.08.009.

- [25] O. Gutiérrez-Hernández and L. V. García. Uncovering true significant trends in global greening. *Remote Sensing Applications: Society and Environment*, 37: 101377, 2025. doi:[10.1016/j.rsase.2024.101377](https://doi.org/10.1016/j.rsase.2024.101377).
- [26] Y. Hamdani, G. Xiao, L. Ding, and D. Calvanese. An ontology-based framework for geospatial integration and querying of raster data cube using virtual knowledge graphs. *ISPRS International Journal of Geo-Information*, 12(9):375, 2023.
- [27] H. Hashim, Z. Abd Latif, and N. A. Adnan. Urban vegetation classification with NDVI threshold value method with very high resolution (VHR) Pleiades imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:237–240, 2019. doi:[10.5194/isprs-archives-XLII-4-W16-237-2019](https://doi.org/10.5194/isprs-archives-XLII-4-W16-237-2019).
- [28] M. Hofmeister, K. F. Lee, Y.-K. Tsai, M. Müller, K. Nagarajan, S. Mosbach, J. Akroyd, and M. Kraft. Dynamic control of district heating networks with integrated emission modelling: A dynamic knowledge graph approach. *Energy and AI*, 17:100376, 9 2024. doi:[10.1016/j.egyai.2024.100376](https://doi.org/10.1016/j.egyai.2024.100376).
- [29] D. Hovland, D. Lanti, M. Rezk, and G. Xiao. OBDA constraints for effective query answering. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, pages 269–286. Springer, 2016.
- [30] T. Ioannidis, G. Garbis, K. Kyzirakos, K. Bereta, and M. Koubarakis. Evaluating geospatial RDF stores using the benchmark Geographica 2. *Journal on Data Semantics*, 10(3):189–228, 2021. doi:[10.1007/s13740-021-00118-x](https://doi.org/10.1007/s13740-021-00118-x).
- [31] M. Jovanovik, T. Homburg, and M. Spasić. A GeoSPARQL compliance benchmark. *ISPRS International Journal of Geo-Information*, 10(7):487, 2021.
- [32] A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Automated Rational Design of Metal–Organic Polyhedra. *Journal of the American Chemical Society*, 144(26):11713–11728, 2022. doi:[10.1021/jacs.2c03402](https://doi.org/10.1021/jacs.2c03402).
- [33] S. Krawczyk, E. ben Izzy, and D. Quinn. Hamilton: enabling software engineering best practices for data transformations via generalized dataflow graphs. In *DEco@VLDB*, pages 41–50, 2022.
- [34] A. Larkin and P. Hystad. Evaluating street view exposure measures of visible green space for health research. *Journal of Exposure Science & Environmental Epidemiology*, 29(4):447–456, 2019. doi:[10.1038/s41370-018-0017-1](https://doi.org/10.1038/s41370-018-0017-1).
- [35] S. Li and W. A. Mackaness. A multi-agent-based, semantic-driven system for decision support in epidemic management. *Health informatics journal*, 21(3):195–208, 2015. doi:[10.1177/1460458213517704](https://doi.org/10.1177/1460458213517704).
- [36] Z. Lin, X. Lin, and X. Yang. An Automated Analysis Framework for Epidemiological Survey on COVID-19. *IEEE Journal of Biomedical and Health Informatics*, 28(5):3186–3199, 2024. doi:[10.1109/JBHI.2024.3370253](https://doi.org/10.1109/JBHI.2024.3370253).

- [37] C. J. Mattingly, T. E. McKone, M. A. Callahan, J. A. Blake, and E. A. C. Hubal. Providing the Missing Link: the Exposure Science Ontology ExO. *Environmental Science & Technology*, 46(6):3046–3053, 2012. doi:10.1021/es2033857.
- [38] R. Mitchell, T. Astell-Burt, and E. A. Richardson. A comparison of green space indicators for epidemiological research. *J Epidemiol Community Health*, 65(10): 853–858, 2011. doi:10.1136/jech.2010.119172.
- [39] A. Mizen, D. A. Thompson, A. Watkins, A. Akbari, J. K. Garrett, R. Geary, R. Lovell, R. A. Lyons, M. Nieuwenhuijsen, S. C. Parker, et al. The use of Enhanced Vegetation Index for assessing access to different types of green space in epidemiological studies. *Journal of exposure science & environmental epidemiology*, 34(5): 753–760, 2024. doi:10.1038/s41370-024-00650-5.
- [40] M. A. Naser, R. Khosla, L. Longchamps, and S. Dahal. Using NDVI to differentiate wheat genotypes productivity under dryland and irrigated conditions. *Remote Sensing*, 12(5):824, 2020. doi:10.3390/rs12050824.
- [41] NEA. 2024 Is Singapore’s Warmest Year On Record, 2026. URL <https://www.nea.gov.sg/media/news/news/index/2024-is-singapore-s-warmest-year-on-record>. Accessed: June 2026.
- [42] Open Geospatial Consortium. OGC GeoSPARQL - A Geographic Query Language for RDF Data, 2024. URL <https://www.ogc.org/standards/geosparql/>. Accessed: 2 March 2026.
- [43] OpenLineage Project. OpenLineage: An Open Framework for Data Lineage Collection and Analysis, 2026. URL <https://openlineage.io/docs/>. Accessed: 17 September 2026.
- [44] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org> . <https://www.openstreetmap.org>, 2023.
- [45] L. Pascazio, D. Tran, S. D. Rihm, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Question-answering system for combustion kinetics. *Proceedings of the Combustion Institute*, 40(1-4):105428, 2024. doi:10.1016/j.proci.2024.105428.
- [46] A. Peluso, D. Rastogi, H. B. Klasky, J. Logan, D. Maguire, J. Grant, B. Christian, and H. A. Hanson. Environmental determinants of health: Measuring multiple physical environmental exposures at the United States census tract level. *Health & place*, 89:103303, 2024. doi:10.1016/j.healthplace.2024.103303.
- [47] M. Perry, P. Jain, and A. P. Sheth. *SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries*, pages 61–86. Springer US, Boston, MA, 2011. doi:10.1007/978-1-4419-9446-2_3.
- [48] C. Pesquita, J. D. Ferreira, F. M. Couto, and M. J. Silva. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. *Journal of biomedical semantics*, 5(1):4, 2014. doi:10.1186/2041-1480-5-4.

- [49] PostGIS. PostGIS. <https://postgis.net/>, 2025. Accessed: 2026-07-17.
- [50] PRECISE. Phase II: PRECISE SG100K, 2026. URL <https://www.npm.sg/phase-ii-precise-sg100k/>. Accessed: June 2026.
- [51] A. I. Ribeiro, M. d. F. d. Pina, and R. Mitchell. Development of a measure of multiple physical environmental deprivation. After United Kingdom and New Zealand, Portugal. *The European Journal of Public Health*, 25(4):610–617, 2015. doi:10.1093/eurpub/cku242.
- [52] S. D. Rihm, Y. R. Tan, W. Ang, M. Hofmeister, X. Deng, M. T. Laksana, H. Y. Quek, J. Bai, L. Pascazio, S. C. Siong, et al. The digital lab manager: Automating research support. *SLAS technology*, 29(3):100135, 2024. doi:10.1016/j.slast.2024.100135.
- [53] S. D. Rihm, Y. R. Tan, W. Ang, H. Y. Quek, X. Deng, M. T. Laksana, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. The digital lab facility manager: automating operations of research laboratories through “The World Avatar”. *Nexus*, 1(3), 2024. doi:10.1016/j.ynexs.2024.100031.
- [54] S. D. Rihm, D. N. Tran, A. Kondinski, L. Pascazio, F. Saluz, X. Deng, S. Mosbach, J. Akroyd, and M. Kraft. Natural language access point to digital metal–organic polyhedra chemistry in The World Avatar. *Data-Centric Engineering*, 6:e22, 2025. doi:10.1017/dce.2025.12.
- [55] M. Rodriguez-Muro and M. Rezk. Efficient SPARQL-to-SQL with R2RML mappings. *Journal of Web Semantics*, 33:141–169, 2015.
- [56] J. W. Rouse Jr, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains with ERTS. In *NASA. Goddard Space Flight Center 3d ERTS-1 Symp., Vol. 1, Sect. A*, number PAPER-A20, 1974.
- [57] T. Savage, J. Akroyd, S. Mosbach, M. Hillman, F. Sielker, and M. Kraft. Universal Digital Twin – the impact of heat pumps on social inequality. *Advances in Applied Energy*, 5:100079, 2022. doi:10.1016/j.adapen.2021.100079.
- [58] T. Savage, J. Akroyd, S. Mosbach, N. Krdzavac, M. Hillman, and M. Kraft. Universal Digital Twin: Integration of national-scale energy systems and climate data. *Data-Centric Engineering*, 3, 2022. doi:10.1017/dce.2022.22.
- [59] S. Scheider and A. Ballatore. Semantic typing of linked geoprocessing workflows. *International Journal of Digital Earth*, 11(1):113–138, 2018. doi:10.1080/17538947.2017.1305457.
- [60] Semantic Web Standards. Resource Description Framework (RDF). <https://www.w3.org/RDF/>, 2026. Accessed: 30 June 2026.
- [61] R. Shrestha, J. Flacke, J. Martinez, and M. Van Maarseveen. Environmental health related socio-spatial inequalities: identifying “hotspots” of environmental burdens and social vulnerability. *International journal of environmental research and public health*, 13(7):691, 2016. doi:10.3390/ijerph13070691.

- [62] S. Skakun, J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, et al. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274:112990, 2022. doi:10.1016/j.rse.2022.112990.
- [63] T. G. Tiecke, X. Liu, A. Zhang, A. Gros, N. Li, G. Yetman, T. Kilic, S. Murray, B. Blankespoor, E. B. Prydz, et al. Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839*, 2017. doi:10.48550/arXiv.1712.05839.
- [64] Y.-K. Tsai, M. Hofmeister, S. Ganguly, K. Rustagi, Y. R. Tan, S. Mosbach, J. Akroyd, and M. Kraft. Municipal heat planning within The World Avatar. *Energy and AI*, 20:100479, 2025. doi:10.1016/j.egyai.2025.100479.
- [65] C. Vámos, S. Scheider, T. Sonnenschein, and R. Vermeulen. Ontology of active and passive environmental exposure. *Semantic Web*, 15(5):1733–1761, 2024. doi:10.3233/SW-243546.
- [66] Y. Wang, S. Yi, X. Li, P. Liu, Z. Yang, R. Bardhan, and R. Stouffs. From physical surfaces to human-centric heat stress: LST and UTCI heat mapping reveals nonlinear effects of urban morphology. *Sustainable Cities and Society*, page 107659, 2026. doi:10.1016/j.scs.2026.107659.
- [67] Q. Weng, D. Lu, and J. Schubring. Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. *Remote sensing of Environment*, 89(4):467–483, 2004. doi:10.1016/j.rse.2003.11.005.
- [68] World Wide Web Consortium (W3C). Data Catalog Vocabulary (DCAT) – Version 3. <https://www.w3.org/TR/vocab-dcat-3/>, 2024. W3C Recommendation, Accessed: 2026-07-20.
- [69] G. Xiao, D. Lanti, R. Kontchakov, S. Komla-Ebri, E. Güzel-Kalaycı, L. Ding, J. Corman, B. Cogrel, D. Calvanese, and E. Botoeva. The virtual knowledge graph system ontop. In *International Semantic Web Conference*, pages 259–277. Springer, 2020. doi:10.1007/978-3-030-62466-8_17.
- [70] J. Xu, S. Mosbach, J. Akroyd, and M. Kraft. Impact of heat pumps and future energy prices on regional inequalities. *Advances in Applied Energy*, 17:100201, 2025. doi:10.1016/j.adapen.2024.100201.
- [71] S. Yu, B. Yu, W. Song, B. Wu, J. Zhou, Y. Huang, J. Wu, F. Zhao, and W. Mao. View-based greenery: A three-dimensional assessment of city buildings’ green visibility using Floor Green View Index. *Landscape and Urban Planning*, 152:13–26, 2016. doi:10.1016/j.landurbplan.2016.04.004.
- [72] P. Yue, X. Guo, M. Zhang, L. Jiang, and X. Zhai. Linked Data and SDI: The case on Web geoprocessing workflows. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:245–257, 2016. doi:10.1016/j.isprsjprs.2015.11.009.

- [73] X. Zenuni, B. Raufi, F. Ismaili, and J. Ajdari. State of the art of semantic web for healthcare. *Procedia-Social and Behavioral Sciences*, 195:1990–1998, 2015. doi:10.1016/j.sbspro.2015.06.213.
- [74] X. Zhou, P. Butler, C. Yang, S. D. Rihm, T. Angkanaporn, J. Akroyd, S. Mosbach, and M. Kraft. Ontology-to-tools compilation for executable semantic control of large language models. c4e-Preprint Series Technical Report 343, University of Cambridge, 2026. URL <https://como.ceb.cam.ac.uk/preprints/343/>. Accessed: 9 August 2026.