Check for updates

DATA NOTE

REVISED Identifying “cities” of the world: a methodological approach, lexicon, and database for urban administrative units

[version 2; peer review: 1 approved, 2 approved with reservations]
Previously titled ; "Identifying “cities” of the world: a methodological approach, lexicon, and database for human settlements"

Hiba Karam, Lalita Phatthanachaisuksiri, Vijay Palliyil, Meng Cai, Niklas Suhre, Eva Kassens-Noor

Institute of Transport Planning and Traffic Engineering, Technische Universitat Darmstadt Fachbereich Bau- und Umweltingenieurwissenschaften, Darmstadt, Hesse, Germany

v2 **First published:** 15 Aug 2025, **5**:241
https://doi.org/10.12688/openreseurope.21078.1
Latest published: 13 Apr 2026, **5**:241
https://doi.org/10.12688/openreseurope.21078.2

Abstract

Since 2015, there has not been a comprehensive database of cities across the world. While debates about what defines a “city” remain open in urban science, we center our efforts toward developing a replicable and transparent methodology to construct a globally valid dataset of cities, agglomerations, capitals, municipalities, and urban centers. Such an endeavor is both timely and necessary given that comparative research, governmental collaborations, and cross-national comparisons among urban areas yield valuable knowledge about human settlements to inform climate resilience and sustainability planning. Our approach is bottom-up and anchored in accessibility. We begin with Wikipedia, a dynamic, crowd-sourced platform, and cross-check its entries with an expert database based on census data. When discrepancies arise, we turn to official governmental websites to confirm information. We integrate these three sources into a harmonized database that includes every urban settlement worldwide with at least 50,000 residents (7,468 in total). For China, Brazil, France, Japan, India, Pakistan, Somalia and North Korea we apply a higher inclusion threshold of 100,000. Each entry lists country and city names and is complemented by a terminology lexicon explaining how different nations define and classify cities. This approach enables us to not only catalogue urban settlements but also to make visible how cities are constructed and recognized within

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 13 Apr 2026	✓ view	? view	? view
version 1 15 Aug 2025	? view	? view	

1. **Milad Malekzadeh**, University of Helsinki, Helsinki, Finland
2. **Johannes H. Uhl**, Joint Research Centre (JRC), Ispra, Italy
3. **Yile Chen**, Macau University of Science and Technology, Taipa, China

Any reports and responses or comments on the article can be found at the end of the article.

diverse global national governance frameworks.

Keywords
Cities, Human Settlements, Database



This article is included in the Horizon Europe gateway.



This article is included in the European Research Council (ERC) gateway.

Corresponding author: Eva Kassens-Noor (ivv@verkehr.tu-darmstadt.de)

Author roles: Karam H: Data Curation, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Phatthanachaisuksiri L**: Data Curation, Visualization, Writing – Review & Editing; **Palliyil V**: Data Curation, Writing – Review & Editing; **Cai M**: Writing – Review & Editing; **Suhre N**: Writing – Review & Editing; **Kassens-Noor E**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the [European Union's Horizon 2020 research and innovation programme] [European Union's Horizon Europe research and innovation programme] (Grant agreement No. [101087218]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Karam H *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Karam H, Phatthanachaisuksiri L, Palliyil V *et al.* **Identifying "cities" of the world: a methodological approach, lexicon, and database for urban administrative units [version 2; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2026, **5**:241 https://doi.org/10.12688/openreseurope.21078.2

First published: 15 Aug 2025, **5**:241 https://doi.org/10.12688/openreseurope.21078.1

REVISED **Amendments from Version 1**

In the revised version (V2), we clarified key terminology and refined how existing city lists are described, avoiding any confusion between methodological frameworks and the datasets that implement them. We also streamlined the limitations discussion and added a dedicated Limitations and Future Work section to clearly state the scope, constraints, and next steps. Overall, these revisions improve conceptual precision and transparency without changing the study's aims or conclusions.

Any further responses from the reviewers can be found at the end of the article

Introduction

The question of "What is a city?" dates back to ancient philosophers grappling to define its form and function and the corresponding rights of its citizens (Shahidipak, 2022). From a contemporary perspective, cities function more like dynamic ecosystems—interconnected spaces where residents, institutions, and economic actors interact and co-evolve (Batty, 2013). Yet despite their centrality to everyday life and to research across disciplines, one fact remains constant: cities continually reshape themselves in response to changing agglomeration forces. Rapid urban population growth, compounded by nationally varying definitions (Allen *et al.*, 2005; Gregory *et al.*, 2009; Parr, 2007) and divergent disciplinary classifications (Caves, 2005; Chant & Goodman, 1999; Doxiadis, 1970; Paddison, 2001), has vastly outpaced the capacity of policymakers, industries, and researchers to maintain a timely, reliable, and verifiable global list of cities.

Even global entities, like the United Nations, have been struggling to answer that question since 1967 (United Nations, 1969). Over the past five decades, the United Nations Population Division has been compiling a comprehensive demographic account of the global urbanization process, now recognized as the biennial "World Urbanization Prospects". Yet despite these efforts, persistent inconsistencies in national definitions have prevented the creation of a unified global list. For example, the 2018 edition surveyed 233 countries and areas and found that 104 differed in both the number and criteria used to define their own cities (United Nations, 2019).

Part of the challenge on defining what a city is lies in the conceptual instability of the term itself. The term city can be applied to "anything and everything" (Caves, 2005) and is often used interchangeably with terms like urban area, metropolitan region, city proper, urban agglomeration, and capital (UN-Habitat, 2020). This semantic fluidity complicates the task of producing a singular, globally valid definition (UN-Habitat, 2020). A recent study brilliantly summarizes the various definitions of cities and their evolution (Dong *et al.*, 2024).

Rather than defining what a city is, we focus instead on how to create a flexible methodology and verifiable list of cities. We will create a lexicon of settlements based on a new methodology, whereby users can flexibly identify cities as it suits their definition. It will fulfil two urgent needs: an up-to-date verifiable list of cities with their corresponding information (with populations of 50,000 and above[1]), and an adaptable methodology that allows this lexicon to be expanded with lower population thresholds and continuously updated by anyone with access to the internet. This approach directly responds to UN-Habitat's need for timely data and metrics (UN-Habitat, 2020). to operationalize any local and/or global policy.

To the best of our knowledge (2024), five datasets (Table 1) exist that list cities across the world. Two of the most comprehensive and academically grounded global datasets are the Atlas of Urban Expansion (Angel *et al.*, 2016) and the latest public release of the GHSL Urban Centre Database (Melchiorri *et al.*, 2024) New York University and UN-Habitat compiled the first dataset in 2010 as part of the *Atlas of Urban Expansion* project. This dataset identified 4,231 free-standing cities in 172 countries or territories with populations exceeding 100,000, using spatial criteria based on built-up areas and open spaces, rather than administrative boundaries (Angel *et al.*, 2016). The second is the GHSL Urban Centre Database (GHS-UCDB R2024A/ "UCDB 2025"), a harmonised global inventory of urban centres derived from GHSL's implementation of the Degree of Urbanisation (DoU) methodology (Melchiorri *et al.*, 2024; Dijkstra *et al.*, 2021). Using consistent gridded population and built-up inputs (e.g., GHS-POP and GHS-SMOD), the latest release reports 11,422 quality-controlled urban centres worldwide (Melchiorri *et al.*, 2024).

Both approaches produce harmonised urban footprints that may not fully coincide with administrative or statistical units. Satellite-derived Urban Expansion methods can capture urban form that extends across jurisdictional boundaries (Angel *et al.*, 2016)., and grid-based delineations using population-density and contiguity thresholds may also differ from national administrative definitions (Dijkstra *et al.*, 2021). The first method identifies cities based solely on physical

[1]Except for China, Brazil, France, Japan, India, Pakistan, Somalia and North Korea we apply a higher inclusion threshold of 100,000 residents due to resource constraints.

Table 1. Existing city datasets and their limitations. Source: The authors.

Name of the dataset	Organisation	Date of the used data	Criteria	Number of cities	Shortcomings
1. Cities in the World	The OECD[2]	2000–2015	Densely populated areas with at least 50,000 people and 1,500 inhabitants per square kilometre.	9,028	The dataset doesn't include all countries in the world.
2. UN data Database	United Nations Statistics Division[3]	1970–2023	The government decided on the list of cities since the data was collected based on questionnaires dispatched annually to national statistical offices.	4,968	The dataset is outdated because recent government lists have not been updated, e.g., Lebanon.
3. World Cities Database - simple maps	Free and paid versions[4]	March 31, 2023	Based on authoritative sources from the USA.	44,691	It is non-governmental origin and solely determined by U.S. government agencies, potentially limiting its representativeness and inclusivity.
4. Geonames	Open source[5]	November 10, 2023	All cities with a population > 1000 or seats of administrative divisions (ca 80.000).	141,184	It is a crowdsourcing scheme with an interface that allows everyone to sign in and edit the database, hence false information can be entered, and such information can remain undetected.
5. Urban Centre Database (GHS-UCDB R2024A/ "UCDB 2025" public release)	European Commission (JRC/GHSL)[6]	Urban-centre boundaries reference year: 2025 (fixed boundary option); attributes reported 1975–2030	Derived from GHSL's implementation of the Degree of Urbanisation (DoU) using gridded population and built-up layers. Urban centres are delineated on a 1 km grid based on population density and contiguity thresholds, independent of administrative boundaries.	11,422 (quality-controlled urban centres)	Because this method uses global grid density/contiguity thresholds, it delineates morphological/statistical urban centres that may differ from administrative city limits or functional urban areas.
6. The 2010 Universe of Cities	UN Habitat, New York University, Lincoln Institute[7]	2010	Urban Extent, Built-Up, and Urbanized Open Space.	19,289	The dataset is outdated since it relies on data from 2010.

[2]Source: Cities in the World | OECD
[3]Source: UNdata | record view | City population by sex, city and city type
[4]Source: World Cities Database | Simplemaps.com
[5]Source: Geonames - All Cities with a population > 1000 — Opendatasoft
[6]Source: GHS-UCDB R2024A - GHS Urban Centre Database 2025
[7]Source: The 2010 Universe of Cities - Lincoln Institute of Land Policy

features, while the second relies on population data. Recognizing that cities also reflect other factors, several authors have proposed alternative ways to define their boundaries. For example, an exciting new approach to identifying cities has been developed in theory. Dong *et al.* (2024) proposed using cell-phone datasets, which capture the movement of people and serve as a proxy for the temporal and spatial connections that define urban life. The authors estimate that 7 billion cell-phone users create geolocated data across the world. This data can indicate people's presence in specific areas and their movement between locations. They suggest that this method has the potential to become widely accepted for delineating cities. However, they do not address the challenge of individuals owning multiple phones or relying exclusively on wireless technology.

In summary, all three methods use an expert-driven, top-down methodological approach: the classification of cities from a top-down perspective, meaning they work with global governmental entities like the United Nations to decide what a city is. They require access to significant amounts of private and/or expensive data and need significant manpower, expertise, and resources to make a list of cities based on the three methodologies. In addition to these three, several other researchers have compiled similar lists (see Table 1), applying varying criteria for what constitutes a city and often capturing only a partial subset of cities worldwide (Yap & Biljecki, 2023).

Despite numerous calls by geographers, demographers, and statisticians for internationally comparable definitions of urban areas and cities, the many different national definitions have not shown any tendency to converge to a common one. International comparability, especially in measuring urbanisation in a particular country, is likely not a major concern for national authorities (Buettner, 2015).

In light of the persistent lack of consensus on what defines a city, we shift the focus from enforcing a universal definition to building a dataset grounded in how cities are defined by individual governments. We argue that the question is no longer *what* a city is, but *how* it evolves within its national context. Given the resource constraints and temporal limitations inherent in data collection, pursuing a single, fixed definition has become less relevant than developing a methodology that is adaptable over time. Our approach acknowledges that each government operates with its own criteria—criteria that inevitably shift in response to political, demographic, and spatial change. To accommodate this complexity, we offer a methodology, lexicon, and dataset that are responsive, transparent, and flexible tools designed to support researchers and policymakers alike.

Method

To build our lexicon of cities, we draw on three foundational perspectives. Altshuler emphasizes the governance dimension of cities, highlighting the role of expert planners (Warren, 1967); Jacobs portrays cities as dynamic, evolving ecosystems shaped by everyday human interaction (Jacobs, 1961); and Corburn underscores the importance of local knowledge and participatory processes. Together, these views support a bottom-up understanding of urban classification (Corburn, 2005).

In our lexicon, we identify each self-declared settlement alongside its official website, adopting the definitions and lists provided by local governments. This aligns with the principle of local governance (Ndreu, 2016), where urban status is determined by municipal authorities rather than global institutions. This approach ensures contextual accuracy, official recognition, inclusivity, and currency. Our method triangulates three data sources: dynamic and crowd-sourced platform (i.e., Wikipedia (Wikipedia, 2024), expert-curated databases City Population (Brinkhoff, 2024a), and localized information (using websites of the human settlements) to create a final list.

Creating the lexicon and database contained six steps:

Step 1: To compile data on cities for the Lexicon, we simultaneously examined City Population (Brinkhoff, 2024a) and Wikipedia (Wikipedia, 2024). City Population, managed by Thomas Brinkhoff in Germany, is, to the best of our knowledge, the only website providing comprehensive global city information based on census results and official estimates (Brinkhoff, 2024b). Wikipedia, being crowdsourced, allows for frequent updates by anyone. We compiled the data into an Excel file, grouping countries listed by the United Nation (United Nations, 2025) into world regions as per Angel *et al.* (2016).

Step 2: We validated the list of cities in each country by initially referring to the region index pages in the City Population database and cross-referencing it with the country's city list in Wikipedia. Subsequently, we conducted further validation by comparing the final list of country cities with the original data sources listed on both websites.

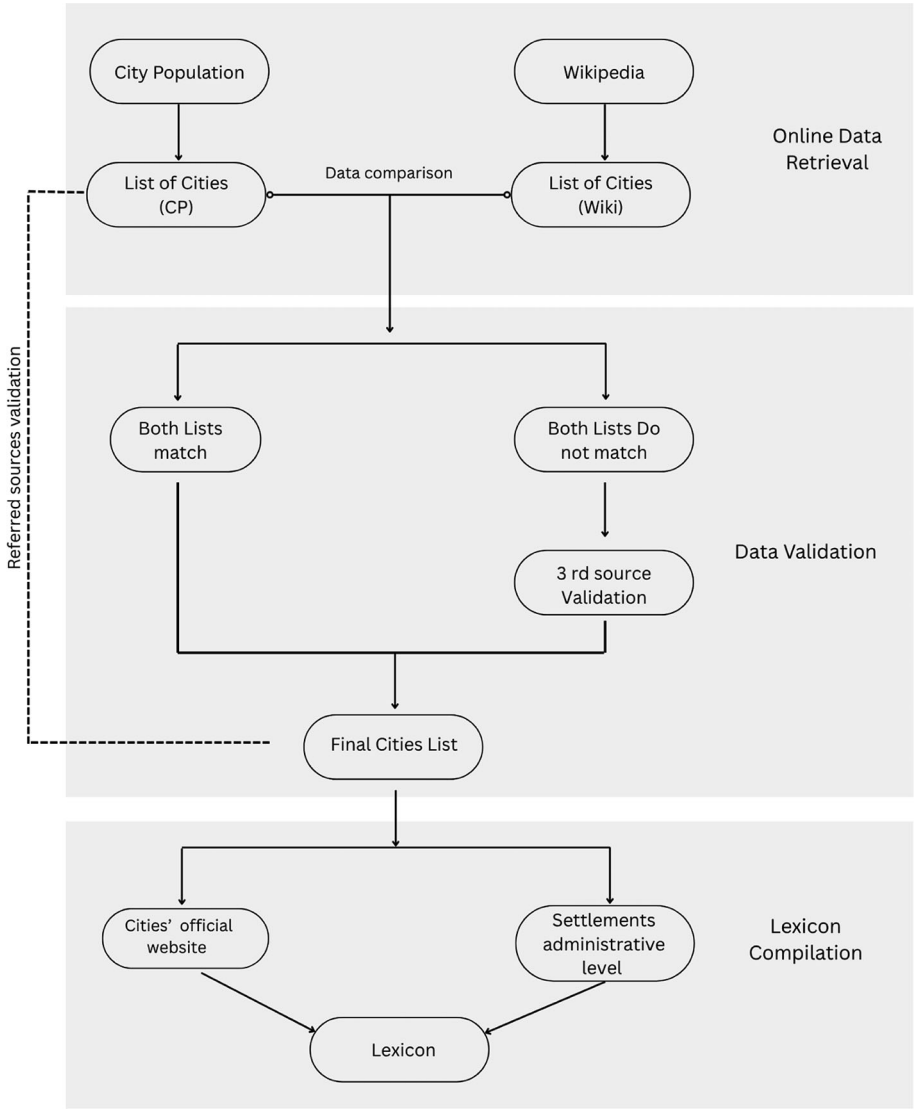
Step 3: If the lists from the two websites do not match, we check the government's official documents for verification. Additional scrutiny involved checking the specific Wikipedia page of any extra city identified or other sources such as

Google Maps. This process ensures the existence of the city. Moreover, if the source of information could not be verified on the City Population or Wikipedia websites, we only considered the list from the website with a verified reference source, and that source was further validated (e.g., Cambodia).

Step 4: If the years of the population numbers differ between the two websites, we select the latest available census or estimated population list. Moreover, during our data verification process, we may discover that the country has updated its city list on its official government website. However, this updated list may not have been reflected on the two websites we referenced. In such cases, we prioritized the list sourced directly from the government website, ensuring that we incorporated the most accurate and up-to-date information available (e.g., Thailand).

Step 5, for each city included in the final list, we documented its administrative level, identifying whether it was classified as a municipality, district capital, special-status city, or another type of administrative unit. This information was obtained either directly from the country's official government list or from the city's official website, depending on availability.

Finally, we used Google search to find the official website, alongside cross-referencing with the City Population and Wikipedia pages, ensuring that the website is indeed the city's official site (See Workflow diagram 1). The websites of US cities were adopted from Cai *et al.* (2023).



Workflow diagram 1. Illustrates the logic, operations, and decisions involved in the compilation and validation processes of compiling the lexicon. Source: The authors.

Limitation

A core challenge in developing a global dataset through open-source platforms lies in their heavy reliance on predominantly Western-oriented sources, especially the English-language version of *Wikipedia.* This dependence raises valid concerns about representativeness and inclusivity—particularly in regions where *Wikipedia* content is censored or inconsistently developed across languages (see: Censorship of Wikipedia). To address this, we mitigated the bias by triangulating data from government websites and expert-curated databases.

Moreover, a further limitation is that, in countries where governments do not maintain up-to-date or accessible city lists, triangulation against official sources is constrained, and coverage may depend more on open platforms with uneven completeness. In such contexts, population-based settlement products could provide a complementary basis for future extensions.

Future work

Integrating spatial data (e.g., from OpenStreetMap) is a promising avenue for future work to support mapping, geocoding, and spatial validation. In addition, the dataset is intended to enable iterative improvement by other researchers. We intentionally designed the workflow to be simple and transparent, so that anyone with internet access can reproduce the procedure and update the dataset.

Data availability

Zenodo repository: "World Cities Lexicon – Urban Administrative Units and Population Data" https://zenodo.org/records/16419786 (Karam *et al.*, 2025).

This project contains following underlying data:

- LexiconDataset_25072025.xlsx Contains major cities across world regions with their administrative classification, population data, official digital presence, and relevant geographic identifiers.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Contribution

We do not claim that our dataset is exhaustive, nor do we offer a silver bullet for the persistent challenge of defining what constitutes a city. Instead, we provide a replicable method to construct and validate a flexible, globally relevant dataset. We contribute to global comparability by clarifying how cities are defined and designated in different national contexts. Through the compilation of a lexicon that documents country-specific criteria, thresholds, and classification systems, the study equips users to better interpret the variation in how cities are recognized worldwide. This transparency enables more informed, purposeful sampling in cross-national research and policy development. In doing so, the lexicon promotes comparability not by enforcing sameness, but by rendering visible the diversity embedded in urban definitions making it accessible to researchers, practitioners, and policymakers alike.

A natural question arises: if an official list of cities, such as those from national census data, is available, why not use it directly as the foundation for the lexicon, given its presumed authority? However, locating and interpreting official government data is often challenging and time-consuming, especially when sources are available only in local languages unfamiliar to the researchers. In our dataset, when the lists from Wikipedia and City Population align, compiling the information becomes straightforward and efficient. But when discrepancies arise, cross-referencing the citations and references in Wikipedia or City Population can provide useful entry points, as either source may link to a more accurate or accessible government source.

Acknowledgement

We would like to extend our sincere gratitude to the following individuals for their invaluable contributions to compile the lexicon: Ánh Minh Nguyen, Dacheng Lu, Dinah Theresia Buhr, Ezra Garcia, Ines Wich, Kristian Mayer, Marie Kühn, Ngoc Trinh Nguyen, and Pardis Safarafard.

References

Allen J, Massey D, Pile S: *City worlds.* Routledge; 2005.
Publisher Full Text

Angel S, Blei A, Parent J, *et al.*: *Atlas of urban expansion: the 2016 edition, volume 1: areas and densities*. NYU Urban Expansion Program, UN-Habitat, & Lincoln Institute of Land Policy; 2016.
Reference Source

Batty M: *The new science of cities*. The MIT Press; 2013.
Publisher Full Text

Brinkhoff T: *City population.* 2024a.
Reference Source

Brinkhoff T: *City population—help & information.* City Population Help & Information, 2024b.
Reference Source

Buettner T: **Urban estimates and projections at the United Nations: the strengths, weaknesses, and underpinnings of the world urbanization prospects.** *Spat Demogr.* 2015; **3**(2): 91–108.
Publisher Full Text

Cai M, Huang H, Decaminada T: **Local data at a national scale: introducing a dataset of official municipal websites in the United States for text-based analytics.** *Environ Plan B: Urban Anal City Sci.* 2023; **50**(7): 1988–1993.
Publisher Full Text

Caves RW: *Encyclopedia of the city.* Taylor & Francis; 2005.
Reference Source

Chant C, Goodman D: *European cities and technology: industrial to post-industrial cities*. Routledge; 1999.
Reference Source

Corburn J: *Street science: community knowledge and environmental health justice*. The MIT Press; 2005.
Publisher Full Text

Dijkstra L, Florczyk AJ, Freire S, *et al.*: **Applying the degree of urbanisation to the globe: a new harmonised definition reveals a different picture of global urbanisation.** *J Urban Econ.* 2021; **125**: 103312.
Publisher Full Text

Dong L, Duarte F, Duranton G, *et al.*: **Defining a city—Delineating urban areas using cell-phone data.** *Nat Cities.* 2024; **1**(2): 117–125.
Publisher Full Text

Doxiadis CA: *Emergence and growth of an urban region the developing urban detroit area volume 3: a concept for future development.* Detriot Edison Co.; 1970.
Reference Source

Gregory D, Johnston R, Pratt G, *et al.*: *The dictionary of human geography.* 6th edn. Chichester: Wiley-Blackwell; 2009.
Reference Source

Jacobs J: *Death and life of Great American cities.* Random House; 1961.
Reference Source

Karam H, Phatthanachaisuksiri L, Palliyil V: **World cities Lexicon – urban administrative units and population data.** *Zenodo.* 2025.
Publisher Full Text

Melchiorri M, Mari Rivero I, Florio P, *et al.*: Stats in the City – the GHSL Urban Centre Database 2025: Public release GHS-UCDB R2024. Luxembourg: Publications Office of the European Union; 2024.
Publisher Full Text | **PubMed Abstract**

Ndreu A: **The definition and importance of local governance.** *Social and Natural Sciences Journal.* 2016; **10**(1).
Publisher Full Text

Paddison R, editor: *Handbook of urban studies.* SAGE; 2001.
Reference Source

Parr JB: **Spatial definitions of the city: four perspectives.** *Urban Stud.* 2007; **44**(2): 381–392.
Publisher Full Text

Shahidipak DM: **Philosophy of democracy, shaping city in ancient Greece by Plato and Aristotle.** 2022; **1**(1): 25–29.
Reference Source

UN-Habitat: **What is a city?** *UN-Habitat*, 2020.
Reference Source

United Nations: **Growth of the world's urban and rural population, 1920–2000 (44; POPULATION STUDIES)**. *Department of Economic and Social Affairs*, 1969.
Reference Source

United Nations: **World urbanization prospects the 2018 revision.** 2019.
Reference Source

United Nations: **Member states**. United Nations; 2025.
Reference Source

Warren R: **Altshuler, Alan A., The city planning process: a political analysis.** *Nat Res J.* 1967; **7**(4): 666.
Reference Source

Wikipedia: *Lists of cities by country.* February 29, 2024.
Reference Source

Yap W, Biljecki F: **A global feature-rich network dataset of cities and dashboard for comprehensive urban analyses.** *Sci Data.* 2023; **10**(1): 667.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 2

Reviewer Report 24 April 2026

https://doi.org/10.21956/openreseurope.25458.r72204

© **2026 Uhl J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Johannes H. Uhl
Joint Research Centre (JRC), Ispra, Italy

Dear authors, thank you for addressing my previous suggestions. However, there are still some things that could be improved.

1) The first sentence of the abstract states "Since 2015, there has not been a comprehensive database of cities across the world." - I am struggling with this statement. That implies there was such a dataset before 2015, but not after? Why 2015? This is very unclear. In the introduction, the authors mention the Atlas of Urban Expansion (Angel et al. 2016) and the Urban Centre Database (Melchiorri et al. 2024) - this contradicts that statement. I strongly suggest to make this statement clearer and less ambiguous, for example you could say "Creating a comprehensive database of cities across the world is a challenging task" or similar.

2) In the intro, you mention the Atlas of Urban Expansion, but this dataset is then not included in Table 1. Is there a reason for that?

3) There are a few further statements that seem a bit problematic to me:

"Despite numerous calls by geographers, demographers, and statisticians for internationally comparable definitions of urban areas and cities, the many different national definitions have not shown any tendency to converge to a common one."

--> The Degree of Urbanisation aims to tackle exactly this issue. There is no need that national definitions "converge to a common one". But there is a need that countries are able to disaggregate their statistical data into internationally comparable rural-urban classes. This process is ongoing and the number of countries that implement the Degree of Urbanisation as an additional classification (besides the national rural-urban definition) is steadily increasing (see https://doi.org/10.2785/3471430 for a recent summary)

--> Moreover, there should be a reflection about national definitions of "cities" vs. "urban" / "rural" - many countries classify their territory into "urban" and "rural" (using statistical or non-statistical

methods). The administrative delineation of a “city” may or may not coincide with “urban” definitions.

--> I think it would be important to make clear that the administrative term "city" is often referring to an arbitrary geographic entity - in many cases, only representing the "core" of a city, that may or may not be the area where population concentrates. There are many examples where the "city" from an admin. point of view covers mainly the business district and surroundings, while the majority of the urban population resides in the outskirts where urban expansion occurred, and may not be included in the population counts used in this work. This is an important shortcoming of the method used herein.

When discussing the approach of Dong et al. (2024), using cellphone data for urban delineations, the authors state “However, they do not address the challenge of individuals owning multiple phones or relying exclusively on wireless technology.”

- It is unclear what is meant by “relying exclusively on wireless technology”
- The problem of population overcounting by “individuals owning multiple phones” is likely a minor problem. The main issues with this approach are (a) that cellphone data are typically not open nor authoritative data, making the applicability of the approach dependent on data policies of individual countries or individual companies, and (b) data are not available for historical points in time. Measuring urban change in a consistent way is crucial to assess sustainable development and planning.

Lastly, I am still a bit concerned about the arbitrary change in population threshold for some countries. You state
“Rather, due to resource constraints related to processing and validating very large numbers of settlements, we applied a higher inclusion threshold for China, Brazil, France, Japan, India, Pakistan, Somalia, and North Korea, limiting entries to settlements with at least 100,000 residents rather than the 50,000-threshold used elsewhere.”

- This arbitrary choice causes the settlement structure of these countries to be represented in a biased way and makes the dataset very difficult to use. I understand that part of this methodology consists of manual website checks, which may not be feasible in some cases – however, it would be a much better strategy to keep the threshold fixed for all countries, and omit the manual quality check in these countries, and flag these records as “not quality-checked”.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Urbanization, geospatial analysis, human settlement modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 22 April 2026

https://doi.org/10.21956/openreseurope.25458.r72203

© 2026 Malekzadeh M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Milad Malekzadeh
University of Helsinki, Helsinki, Finland

Thank you for responsing to my concerns. The authors have adequately addressed the issues I raised.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Geography, GIS, Spatial Analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 April 2026

https://doi.org/10.21956/openreseurope.25458.r72391

© 2026 Chen Y. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yile Chen
Department of Architecture, Faculty of Humanities and Arts, Macau University of Science and Technology, Taipa, Macau, China

Dear author, I can see that the article has been revised.
Overall, I think it's very good, but there's one thing I don't quite understand: the article claims, "For China, Brazil, France, Japan, India, Pakistan, Somalia, and North Korea, we adopted a higher inclusion threshold, namely 100,000 cities," but the reasons for the different countries are not explained.

Also, some hypothetical additions could be made to the applicable scenarios.
In other words, besides supporting geographic mapping, are there any other application scenarios?
China does not seem to include Hong Kong, Macau, and Taiwan. This is actually quite common, and I understand it. However, its limitations should be explained.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: urban planning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 05 November 2025

https://doi.org/10.21956/openreseurope.22800.r59250

© 2025 Uhl J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Johannes H. Uhl
Joint Research Centre (JRC), Ispra, Italy

Dear Authors, thank you for your hard work.
Your effort is laudable, but this report needs further clarifications before it should be indexed

1) I was unable to find a rationale for excluding China, Brazil, France, Japan, India, Pakistan, Somalia and North Korea from the lexicon. Could you clarify why you excluded these countries? It is particularly surprising e.g. for Brazil, which has high-quality geospatial data and classifications of their settlements.

2) There are some misunderstandings regarding the Degree of Urbanisation, the Urban Centre Database, and the GHSL-SMOD datasets:

- The Degree of Urbanisation (DoU) is a methodology, not a dataset. EC, UN-Habitat and others are working towards the implementation of this method in individual countries, which means that national statistical offices adopt the method to classify their territory acording to the Degree of Urbanisation method, enabling them to produce statistics for globally harmonised rural-urban classes using their own census data. The Degree of Urbanisation does not require that countries abandon their national definition of urban / rural, rather that they have an additional territorial

classification allowing them to report internationally comparable urban/rural statistics. The "implementation" of the DoU means that individual countries adopt this method, and this implementation is an ongoing process, but this is not very relevant for your work, because of the following:

- The European Commission Joint Research Centre (Global Human Settlement Layer project, GHSL) uses globally available census and population data to produce global gridded population data (GHS-POP R2023A). GHS-POP R2023A is available globally from 1975-2030. Based on GHS-POP, GHSL produced global gridded data of the Degree or Urbanisation (GHS-SMOD; "Settlement model") for the same spatio-temporal scope, also publicly available.
Based on GHS-SMOD, GHSL extracted urban centres, which were enriched with a wide range of attributes, and are published as a global-scope dataset, called the Urban Centre Database (UCDB R2024A). In your report, you cite an old version of the UCDB. Please update this to the recent version of the UCDB.
Moreover, GHSL classifies global administrative data according to the DoU method, published as the GHS-DUC (Degree of Urbanisation Classification).

Hence, the fact that countries have not implemented the DoU method to their own census data does not hamper the applicability of the DoU method to the globe, since global gridded population data are available (GHS-POP) and the DoU method has been applied to that. The completeness / exhaustive property of the UCDB is not affected by the fact that the DoU implementation in countries is still ongoing.

This needs to be clarified in the report.

Moreover, the authors should clarify how exactly they searched for "cities" in Wikipedia. Often, the term "city" is used for an administrative unit (eg municipality), but in many cases, the "city" as defined by administrative institutions only comprises the center part of an urban settlement, while urban growth / sprawl / conurbation processes have caused the urban object (or: settlement) to extend far beyond the administrative borders of the "city".

I see the dataset is called "Urban Administrative Units" - This is actually a much more precise name. Your dataset is about urban administrative units, according to an arbitrary definition of "urban". It is not about "cities", and not about "settlements", it is about "Urban Administrative Units". I strongly suggest replacing "human settlements" by "Urban Administrative Units" in the title of this report.

Another issue is the temporal heterogeneity of the population estimates reported in the proposed dataset, ranging from or beyond 2014 to 2023. Alongside with the lack of information for several important countries, I do not see how these population data could be useful for analytical purposes.

I suggest to cross-compare the population numbers obtained in this database with the UCDB 2024A for the respective reference years.

References
1. M, Melchiorri I, Mari Rivero: Stats in the City – the GHSL Urban Centre Database 2025. *https://publications.jrc.ec.europa.eu/repository/handle/JRC139768*. 2024.
2. GHSL data package 2023. *https://op.europa.eu/en/publication-detail/-/publication/a1483a27-ed5d-11ed-a05c-01aa75ed71a1/language-en*. 2023.

Is the rationale for creating the dataset(s) clearly described?
Partly

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Urbanization, geospatial analysis, human settlement modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Mar 2026

Hiba Karam

Dear Johannes H. Uhl, Thank you very much for your careful reading of our manuscript and for your thoughtful and constructive comments. We greatly appreciate the time and effort you invested in reviewing our work. Your feedback has been very helpful in improving the clarity and quality of the paper. We have revised the manuscript accordingly, and below we respond to each comment in detail.

1) I was unable to find a rationale for excluding China, Brazil, France, Japan, India, Pakistan, Somalia and North Korea from the lexicon. Could you clarify why you excluded these countries? It is particularly surprising e.g. for Brazil, which has high-quality geospatial data and classifications of their settlements.

Thank you for raising this point, and we apologize for the confusion. We did not exclude these countries from the lexicon. Rather, due to resource constraints related to processing and validating very large numbers of settlements, we applied a higher inclusion threshold for China, Brazil, France, Japan, India, Pakistan, Somalia, and North Korea, limiting entries to settlements with at least 100,000 residents rather than the 50,000-threshold used elsewhere. We have revised the manuscript to explain this rationale more explicitly and to

avoid any ambiguity that might suggest these countries were excluded.

2) There are some misunderstandings regarding the Degree of Urbanisation, the Urban Centre Database, and the GHSL-SMOD datasets:

- The Degree of Urbanisation (DoU) is a methodology, not a dataset. EC, UN-Habitat and others are working towards the implementation of this method in individual countries, which means that national statistical offices adopt the method to classify their territory acording to the Degree of Urbanisation method, enabling them to produce statistics for globally harmonised rural-urban classes using their own census data. The Degree of Urbanisation does not require that countries abandon their national definition of urban / rural, rather that they have an additional territorial classification allowing them to report internationally comparable urban/rural statistics. The "implementation" of the DoU means that individual countries adopt this method, and this implementation is an ongoing process, but this is not very relevant for your work, because of the following:
- The European Commission Joint Research Centre (Global Human Settlement Layer project, GHSL) uses globally available census and population data to produce global gridded population data (GHS-POP R2023A). GHS-POP R2023A is available globally from 1975-2030. Based on GHS-POP, GHSL produced global gridded data of the Degree or Urbanisation (GHS-SMOD; "Settlement model") for the same spatio-temporal scope, also publicly available.
Based on GHS-SMOD, GHSL extracted urban centres, which were enriched with a wide range of attributes, and are published as a global-scope dataset, called the Urban Centre Database (UCDB R2024A). In your report, you cite an old version of the UCDB. Please update this to the recent version of the UCDB.
Moreover, GHSL classifies global administrative data according to the DoU method, published as the GHS-DUC (Degree of Urbanisation Classification). Hence, the fact that countries have not implemented the DoU method to their own census data does not hamper the applicability of the DoU method to the globe, since global gridded population data are available (GHS-POP) and the DoU method has been applied to that. The completeness / exhaustive property of the UCDB is not affected by the fact that the DoU implementation in countries is still ongoing.
This needs to be clarified in the report.

Thank you for this detailed information. We have revised our manuscript accordingly in the introduction section and updated Table 1 with the indicated new dataset. We understand that the Degree of Urbanisation is a method, and we rewrote our sentences to make that clearer.

3) Moreover, the authors should clarify how exactly they searched for "cities" in Wikipedia. Often, the term "city" is used for an administrative unit (eg municipality), but in many cases, the "city" as defined by administrative institutions only comprises the center part of an urban settlement, while urban growth / sprawl / conurbation processes have caused the urban object (or: settlement) to extend far beyond the administrative borders of the "city".

Thank you for raising this important clarification. We used Wikipedia’s curated “Lists of cities https://en.wikipedia.org/wiki/Lists_of_cities” pages as a structured entry point to identify

settlement across countries. These Wikipedia pages were used to retrieve and standardize city/settlement names and to provide reference links for user-led follow-up checks. We revised the wording in the methodology to make this clearer. Crucially, Wikipedia was not the primary determinant of inclusion in the final lexicon. The final list is derived from our core sources and harmonization procedure, with Wikipedia serving only a supplementary role (i.e., to support initial identification and facilitate verification).

4) I see the dataset is called "Urban Administrative Units" - This is actually a much more precise name. Your dataset is about urban administrative units, according to an arbitrary definition of "urban". It is not about "cities", and not about "settlements", it is about "Urban Administrative Units". I strongly suggest replacing "human settlements" by "Urban Administrative Units" in the title of this report.

Thank you for the comment. This is a great point. We agree and have changed the name accordingly.

5) Another issue is the temporal heterogeneity of the population estimates reported in the proposed dataset, ranging from or beyond 2014 to 2023. Alongside with the lack of information for several important countries, I do not see how these population data could be useful for analytical purposes. I suggest to cross-compare the population numbers obtained in this database with the UCDB 2024A for the respective reference years.

Thank you for this comment. We agree that the population figures are temporally heterogeneous and not suitable for direct analytical comparison without harmonization. We have clarified in the manuscript that they are provided mainly as contextual metadata rather than as a standardized analytical variable. We also added a limitation noting that, where official city lists are unavailable or outdated, coverage may depend more on open sources with uneven completeness, and that population-based settlement products could support future extensions.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 September 2025

https://doi.org/10.21956/openreseurope.22800.r60429

© 2025 Malekzadeh M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Milad Malekzadeh
University of Helsinki, Helsinki, Finland

The rationale for creating this dataset is clearly articulated, and the need for such a resource is evident. However, while the motivation is well presented, there are existing datasets that cover

similar ground, as listed by the authors in Table 1. Unfortunately, some of the information provided about these datasets is inaccurate. For example, I quickly reviewed the dataset from *simplemaps.com*. The basic version of their dataset is freely available, whereas the pro version contains substantially more information and is therefore not open access, though it is highly comprehensive. Additionally, the authors state that the city lists in previous datasets are outdated. However, according to the *simplemaps* website, their dataset was most recently updated in 2025. These discrepancies should be addressed and corrected in the revised manuscript.

Regarding the contribution, this new dataset has clear value, and I commend the authors for their efforts in compiling and sharing it. That said, I have some concerns about the methodology. A widely used and authoritative source of spatial data for geographers, spatial analysts, and related fields is OpenStreetMap (OSM). Extracting a list of cities from OSM is straightforward by querying features labeled as "city." While Wikipedia can certainly be a useful resource, it is a crowdsourced platform and, therefore, has limited reliability as a primary source for scientific data. Although the authors have attempted to ensure reliability by cross-checking information across multiple sources, it would strengthen the paper to include a comparison between the city list derived from OSM (while still crowdsourced, but the validity of its data has been studied over and over) and the dataset they have compiled. OSM has the additional advantage of providing spatial coordinates, which would make the dataset more readily applicable for quantitative spatial analyses.

Furthermore, *simplemaps* states on their website that their dataset was "built from the ground up using authoritative sources such as the NGIA, US Geological Survey, US Census Bureau, and NASA." It would be important to know whether the authors have investigated *simplemaps*' methodology or contacted them directly to clarify their approach. If *simplemaps* also uses a bottom-up, authoritative method, then the novelty of the present dataset may be less clear. In such a case, the primary contribution of the authors' work might shift toward providing an open and accessible version of similar data.

Finally, for countries where governments do not maintain up-to-date city lists, or where Wikipedia lacks adequate information or cannot be fully trusted, approaches based on population data may offer a more comprehensive and reliable alternative. It would be helpful for the authors to discuss these issues explicitly and to justify their methodological choices in light of these considerations.

Is the rationale for creating the dataset(s) clearly described?

Partly

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Geography, GIS, Spatial Analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Mar 2026

Hiba Karam

Dear Milad Malekzadeh, We sincerely thank you for your thorough review of our manuscript and for your insightful comments and suggestions. We appreciate the opportunity to revise our work in response to your feedback.

1. The rationale for creating this dataset is clearly articulated, and the need for such a resource is evident. However, while the motivation is well presented, there are existing datasets that cover similar ground, as listed by the authors in Table 1. Unfortunately, some of the information provided about these datasets is inaccurate. For example, I quickly reviewed the dataset from *simplemaps.com*. The basic version of their dataset is freely available, whereas the pro version contains substantially more information and is therefore not open access, though it is highly comprehensive.

Thank you for highlighting this point. We agree that Simplemaps.com offers both a freely available basic version and a more comprehensive paid (“Pro”) version. We have revised Table 1 accordingly to reflect this distinction more accurately. Importantly, this clarification does not affect our broader argument regarding the dataset’s limitations for our purposes, as our concern relates primarily to the underlying data source and its characteristics (as discussed in point 6), rather than to the access model itself.

1. Additionally, the authors state that the city lists in previous datasets are outdated. However, according to the *simplemaps* website, their dataset was most recently updated in 2025. These discrepancies should be addressed and corrected in the revised manuscript.

Thank you for raising this point. We have changed the name from World Cities to World Cities Database – simplemaps to resolve any confusions as we did not characterize the simplemaps dataset as outdated (see Table 1 for our revised manuscript). Rather, our statement regarding outdated information refers specifically to (i) the UN Data Database (1970–2023) and (ii) the 2010 Universe of Cities, which is based on 2010 data, as reflected in Table 1.

1. Regarding the contribution, this new dataset has clear value, and I commend the authors for their efforts in compiling and sharing it. That said, I have some concerns about the methodology. A widely used and authoritative source of spatial data for geographers, spatial analysts, and related fields is OpenStreetMap (OSM). Extracting a list of cities from OSM is straightforward by querying features labeled as “city.”

Thank you for this helpful suggestion. We understand that OpenStreetMap (OSM) is a community-driven resource that draws on local knowledge, with contributors using aerial imagery, GPS devices, and field mapping to support accuracy and regular updates. Our dataset, however, is not intended to provide spatial geometries; rather, it focuses on compiling and harmonizing city lists and associated attributes. That said, we recognize the

value of extending the dataset with spatial information, and we have incorporated this point into the manuscript's "Future Work" section as a clear avenue for further development and the potential integration of spatial data.

1. While Wikipedia can certainly be a useful resource, it is a crowdsourced platform and, therefore, has limited reliability as a primary source for scientific data.

Thank you for this important point. We acknowledge that Wikipedia is a crowdsourced platform and therefore has limitations as a primary scientific source. In our dataset, however, it is not used as a principal data source, but rather as a practical starting point for users seeking to further update the city list. We have explicitly addressed this concern in the manuscript's Limitations section, where we discuss the uneven geographic and linguistic coverage of open-source platforms, including the disproportionate reliance on English-language content, and the resulting implications for representativeness, particularly in contexts of censorship or uneven coverage across language editions. As noted there, we mitigate this limitation by triangulating entries against government websites and expert-curated databases. In addition, we clarify that when discrepancies arise between lists, cross-referencing the citations and references provided in Wikipedia or City Population may offer useful entry points, as either source can lead users to a more accurate or accessible government source.

1. Although the authors have attempted to ensure reliability by cross-checking information across multiple sources, it would strengthen the paper to include a comparison between the city list derived from OSM (while still crowdsourced, but the validity of its data has been studied over and over) and the dataset they have compiled. OSM has the additional advantage of providing spatial coordinates, which would make the dataset more readily applicable for quantitative spatial analyses.

Thank you for your suggestion. We aim to make differences across city lists transparent in order to support more informed comparative case selection, rather than to provide a ready-to-use spatial dataset. However, we have added the integration of spatial data as a clear avenue for future work.

1. Furthermore, *simplemaps* states on their website that their dataset was "built from the ground up using authoritative sources such as the NGIA, US Geological Survey, US Census Bureau, and NASA." It would be important to know whether the authors have investigated *simplemaps*' methodology or contacted them directly to clarify their approach. If *simplemaps* also uses a bottom-up, authoritative method, then the novelty of the present dataset may be less clear. In such a case, the primary contribution of the authors' work might shift toward providing an open and accessible version of similar data.

Thank you for pointing this out. Yes, we reviewed Simplemaps' methodology. Simplemaps describes its World Cities Database as being compiled from "authoritative sources" such as the National Geospatial-Intelligence Agency, the U.S. Geological Survey, and the U.S. Census Bureau. Its FAQ further states that, in response to the question "What counts as a city/town?", the database includes "any populated place in the world as determined by U.S. government agencies." This suggests a top-down inclusion criterion rooted in institutional classification, which is conceptually different from our bottom-up procedure. Accordingly, we do not position our contribution as providing "better" data. Rather, we present a dataset designed for a different research purpose: to make differences across city lists transparent and thereby support more representative and reproducible comparative case selection, based on explicit inclusion criteria and traceable, ground-level evidence rather than a single

authoritative determination of what qualifies as a “populated place.”

1. Finally, for countries where governments do not maintain up-to-date city lists, or where Wikipedia lacks adequate information or cannot be fully trusted, approaches based on population data may offer a more comprehensive and reliable alternative. It would be helpful for the authors to discuss these issues explicitly and to justify their methodological choices in light of these considerations.

Thank you for your comment. This is clearly one of our work’s limitations. We have included it in the limitation section in our revised manuscript.

Competing Interests: No competing interests were disclosed.