Spatiotemporal Data Management for subnational Census Data on global Scale

Thomas Brinkhoff [1]

[1] Institute for Applied Photogrammetry and Geoinformatics, Jade University of Applied Sciences, Oldenburg, Germany
- thomas.brinkhoff@jade-hs.de

Keywords: Spatiotemporal Data Management, Global Mapping, Population Census, Demography, SDGs.

Abstract

Knowledge of the regional distribution of the world's population is essential for political and social decisions not only but especially for achieving the 17 Sustainable Development Goals (SDGs). Census and other population data at the subnational level are important for this purpose. However, current population data management platforms largely ignore the spatiotemporal nature of census data. Here, we outline the requirements for a spatiotemporal population data management system and present its general architecture, data model and state of implementation. The developed system currently stores population data from approximately 200 countries, nearly 11 million spatial units and around 770 million individual population figures.

1. Introduction

In 2015, the UN launched the SDGs. All of the SDGs that are based on ensuring that a certain percentage of the population has access to specific resources, or achieves a sufficient level of social, economic, or physical health. Quantifying them requires reliable demographic data (Tatem, 2017). Population censuses are an essential source of such information. The most recent 2020 census round was completed in 2024, though some results were not released until 2026. The 2030 census round began in 2025 (United Nations Statistics Division, 2026).

For global analyses and applications, we need unobstructed, interoperable access to census data. Additionally, census data are geospatial referring to spatial areas such as administrative or census units. These units often change between two surveys. Therefore, census data are inherently spatiotemporal. However, current population data management systems largely ignore this fact.

As part of the "WorldPop Global Demographic Data" project (WorldPop, 2025), we processed global census data for higher administrative levels as well as for subnational units. To accomplish this, we have designed and implemented a spatiotemporal data management system and populated it primarily with data from the 2020 census round.

The next section reviews the state of the art regarding population data management systems, their interoperability, the incorporation of geospatial features, and spatiotemporal data management in general. In the third section, the demands and challenges of managing spatiotemporal population data are elaborated. Section 4 presents the requirements, architecture, data model and implementation of the spatiotemporal data management system. The paper concludes with a brief summary and an outlook on future work.

2. State of the Art

2.1 Population Data Management Systems

Many data management systems that organize and disseminate census data, do not support spatial data. *Redatam*, e.g., is a program that facilitates the processing of microdata from population and housing censuses, as well as similar surveys. It was developed by the Population Division of the United Nations Economic Commission for Latin America and the Caribbean (CELADE, 2026) and allows the creation of "tables that can be exported to create thematic maps". *PxWin* and *PxWeb* are programs of Statistics Sweden (2026) for the management and web-based dissemination of statistical data. They are particularly popular in Scandinavia. Data are represented by so-called PX files. However, they are not able to store spatial attributes. *STATcube* is the statistical database system of Statistics Austria (2026). It also lacks support for spatial data.

Some systems offer a limited geospatial support. *Nomis* is a statistical service provided by the University of Durham on behalf of the UK Office for National Statistics (2026a). It is able to combine statistical data with about 90 different spatial datasets (down to output areas of the 2001, 2011 and 2021 census). A Leaflet-based web map application can display a maximum of 10,000 units. An export to a geospatial format is not possible. *Data.census.gov* is the platform for accessing demographic and economic data from the U.S. Census Bureau (2026). It allows users to select geographies (down to the census block level) in web maps and to display query results in maps. Again, an export into a geospatial data format is not supported.

For population estimates and projections, most systems allow data from different years to be combined. The basic assumption is that the subnational units have not changed their boundaries during this period. Otherwise, no data will be provided for such units or another time series covering a different period will have to be selected. For census data, an integrated access to several dates is typically not provided. They often need to be accessed separately.

2.2 Dissemination of Geospatial Population Data

There are some organizations that disseminate geospatial population data for subnational units on a global scale: The U.S. Census Bureau (2025) currently provides such datasets for about 60 countries. In general, a dataset consists of an Excel file and a matching shapefile. The population figures are mostly projections and not census results. For most countries, data are provided only for the first and second administrative level. The geospatial datasets often origin from HDX.

HDX is the humanitarian data exchange platform of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA, 2026a). For many countries, geospatial data for subnational administrative areas are available on HDX. However, timeliness and accuracy vary and are difficult to assess. In several cases, population figures are provided that match the spatial dataset. Figures are mostly population estimates or projections; in some cases, census results are available.

Eurostat (2026a) operates a census data hub for the countries in the European Union and a few others. It provides data from the 2010 and 2020 census round, but is not global. Geospatial datasets are provided separately by *GISCO* (Eurostat, 2026b).

City Population (2026) presents population data for subnational administrative units of all countries of the world in both HTML tables and web maps. The website often integrates population data from several censuses and recent estimates. However, it is not a dissemination platform, as there are no download options.

2.3 Interoperability of Population Data

Most statistical agencies and data platforms allow population data to be downloaded as Excel, CSV, or JSON files. Associated geospatial data (if available) are typically provided as a shapefile, GeoJSON, or GeoPackage. It is not uncommon for census data to be available only as PDF report.

In 2001, seven international organizations launched an initiative to develop a global standard for an API-based exchange of statistical data and metadata (Willekens, 2023). This work resulted in the *SDMX* standard (SDMX Community, 2026), which was published in 2004. In 2005, the International Organization for Standardization (ISO) approved SDMX as an international standard. The current version is ISO 17369:2013. It was last reviewed and endorsed in 2023 (ISO, 2023). The "place of residence" can only be represented by a code that is unique within a country. Only some portals allow the export of SDMX, such as the data warehouse of the Italian National Institute of Statistics.

The European Commission's *INSPIRE* initiative specified a data specification for population distribution (European Commission, 2013). Annex C specifies demography-related codes. Spatial units are represented by the Geography Markup Language (GML). This specification is actually geospatial, but is rarely used.

A more lightweight approach to an interoperable exchange of population data is the use of *HLX* annotations. The Humanitarian Exchange Language allows a user to specify the semantics of a column in a spreadsheet by adding a row for an additional header. The cells in this header store predefined hashtags. They allow indicating the level, name, code and boundary coordinates of administrative units as well as thematic information like sex, age group, and date. However, HDX stopped supporting HLX in January 2026 (OCHA, 2026b).

Unambiguous coding of subnational units is important for the interoperability of population data. Section 3.1 discusses this topic in more detail.

2.4 Spatiotemporal Data Management

Spatiotemporal data modeling is a well-known topic in geospatial information science (Langran, 1989) (Worboys, 1994) (Peuquet, 1999). Important aspects are data types, the nature of changes, the evolution in time and space, space-time topology, and object identities (Pelekis et al., 2004). Several spatiotemporal data models have been proposed. According to the taxonomy presented by Shekhar et al. (2015), a polygonal object model that accounts for temporal changes (e.g., split/merge events) is adequate for managing census data. Subnational units can be considered analogous to the "socioeconomic units" described by Worboys (2001).

Spatial database systems are also an established technology (Brinkhoff & Kresse, 2022). They are an obvious choice for building a geospatial data management system for subnational population census data. However, spatial database systems do not currently provide support for specific spatiotemporal data models. Most NoSQL systems with support for spatiotemporal data deal only with trajectories (Alam et al., 2022). But the modeling capabilities of graph databases are appropriate for the representation of spatiotemporal events (Zheng et al., 2017) (Wu et al., 2020).

Statistical offices also strive to handle spatiotemporal data appropriately. For example, Sindoni et al. (2002) presented experiences with the development of a spatiotemporal information system by the Italian National Institute of Statistics. Martin et al. (2015) use a gridded population model instead.

3. Spatiotemporal Data Management

3.1 Requirements and Challenges

For the "WorldPop Global Demographic Data" project, we were especially interested in demographic data structured by sex, 5-year age group and their combination. For system design, we had to consider what data are available from statistical offices. Regarding sex (or gender), there are up to four categories: male, female, third gender (e.g., Hijra), and unknown. The maximum 5-year age group varies. However, sometimes other or no age groups are available. Additionally, data may be structured differently for different administrative or statistical levels.

In many cases, a privacy-maintaining post-processing hinders consistency checks and makes aggregation difficult or impossible. Two methods are typically applied. One approach is to skip values equal or below a given threshold (e.g., 3, 5 or 9). With the second method, records may be swapped between different geographical areas, which perturbs the counts by small amounts. For example, the UK Office for National Statistics (2023) uses this method. Both approaches hinder consistency checks. Sometimes, both approaches are combined: The 2022 census reports a total population of 12 for the Dierfeld commune in Germany, consisting of 13 males and an omitted number of females (Statistisches Bundesamt Deutschland, 2024). The threshold method makes aggregation difficult or even impossible.

Boundaries of administrative and statistical areas often change between censuses. In most cases, no information about these changes is provided. For example, the number of subprefectures of Guinea increased from 343 during the 2014 census to 378 reported for the 2025 census (Institut Nationale de la Statistique de Guinée, 2026). No descriptions, maps or even geospatial data indicate the boundary changes. Thus, a geospatial storage of this dataset as well as comparisons with the former census data are not possible on this level. On prefecture level, the number and names of the entities remained the same. However, there were obviously (at least) between the prefectures of Conakry, Coyah and Dubréka significant boundary changes with considerable impacts on population numbers.

Another issue with existing datasets is the frequent absence of a stable identity code for administrative units. In many cases, these codes are entirely absent. In other cases, the same code may be used in two censuses but not represent the same boundaries. Figure 1 shows the boundaries of U.S. places in the 2010 and 2020 censuses, along with their identifiers. The brown areas are covered by both the 2010 and 2020 polygons. The green areas belong only to the 2020 places, while the red areas belong only to the 2010 places.

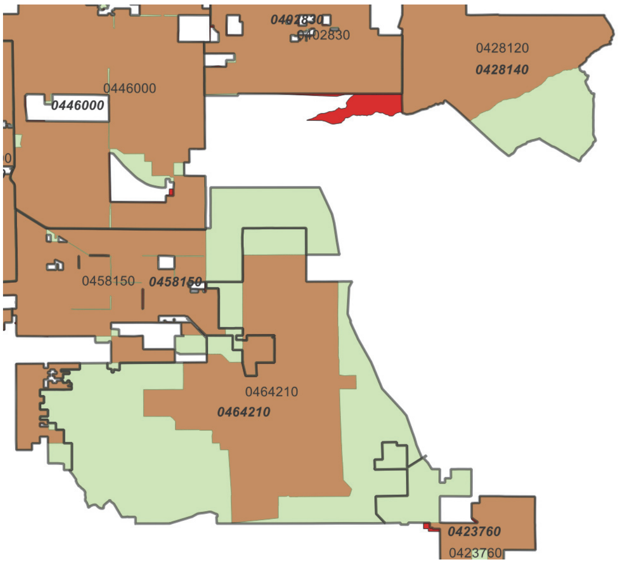

Figure 1. Boundary and identifier comparison of U.S. places according to the 2010 and 2020 population census.

Regardless of the source of the geospatial dataset, it is not unusual for population and spatial records to differ. This can occur if the datasets have different validity periods, if unpopulated areas are missing from population datasets, or if records exist for unlocated populations. Furthermore, timeliness and accuracy are difficult to assess and vary.

In summary, a data management system for spatio-temporal population data should

- support the import of multiple input formats and models,
- handle population and geospatial data and their internal and external relationships,
- provide thorough consistency and quality checks that account for privacy-related changes and omissions,
- represent spatiotemporal changes to ensure comparability between different censuses,
- store appropriate metadata, and
- export data using accepted and easy-to-use data formats.

3.2 Architecture

Figure 2 shows the general architecture of our spatiotemporal census data management system. The core of the architecture is a spatial database system that stores population data, geospatial data, and metadata. Data are mainly organized by country, administrative level, and survey date. Population data from different censuses may refer to the same or different geospatial layers. In the second case, additional spatiotemporal relationships are stored. Depending on the changes and data availability, these relationships allow the computation of the previous, the subsequent, an intermediate and/or upper administrative or census division.

As discussed above, census data are not currently very interoperable. Due to the wide variety of data formats, it is not possible to provide specific import mechanisms for every country or census in the world. Instead, we first transform the original data into a suitable intermediate format. There are several types of templates for this intermediate format in order to support typical data structures. Then, importing can be done automatically.

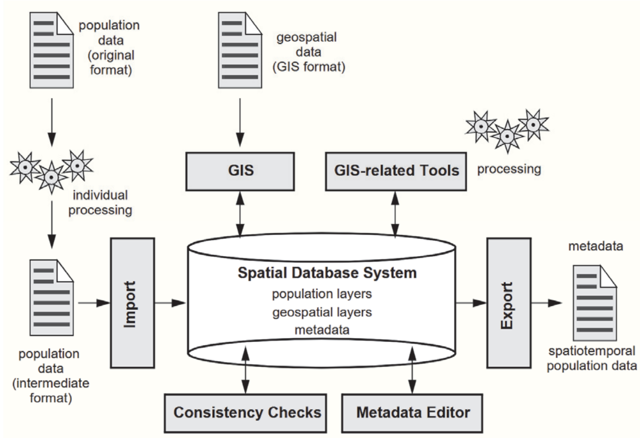

Figure 2. System architecture.

The second step in the processing is a consistency check. This step is primarily performed by software programmed in Python. Data cleansing is often necessary. These checks can also identify spatiotemporal changes that are not described in census reports or other metadata.

GIS tools are used to import, visualize, and process geospatial data. The metadata editor allows maintaining the metadata. Another component exports data and metadata by using accepted, interoperable, and easy-to-use data formats and models.

3.3 Data Model

Figure 3 shows the basic data model developed as UML class diagram. The countries represented are listed in the table “countries” and identified by their ISO 3 code. If a subnational statistical office is responsible for censuses or population estimates, that subnational area is considered as a country. Examples include the emirates of the United Arab Emirates and the UK countries Scotland and Northern Ireland. In these cases, a modified ISO 3 code is used (e.g., “ARE-DUB” for Dubai).

For each census or estimate and administrative or statistical division, an entry is stored in the metadata table “pop_metadata”. Its primary key “tablename” refers to the table that stores the corresponding population figures (denoted as “iso3_level_year” in the UML diagram). The attributes “division” and “level” provide the name of the represented units (e.g., “municipalities” or “census blocks”) and their position in the division hierarchy. The derived attributes “/upper_tab” and “/lower_tab” refer to the metadata entries above and below the current entry, respectively. The census day (or other reference date) is stored in “popdate”. For population projections, a period of several years can be specified. Then, single dates are stored in an additional column of the population table. The exact structure of the population table is defined by four attributes: “t”, “sex”, “t_agegroup”, and “sex_agegroup”. The “round_diff” attribute indicates the maximum deviation of a population figure from its exact value due to privacy or accuracy concerns. It is applied to check aggregations.

The “geo_table” attribute refers to one entry in the table “geo_metadata” that represents the metadata about the geospatial layers. Its attributes have analogous meanings. The primary key “tablename” refers to the table that stores the geometries (denoted as “iso3_level_v_geo” in the UML diagram). An entry in “geo_metadata” can be referenced by several entries in “pop_metadata” if there has been no boundary change. The spatial hierarchy is represented by the “upper_adm” attribute, which stores the foreign key to an entry in an upper geospatial layer.

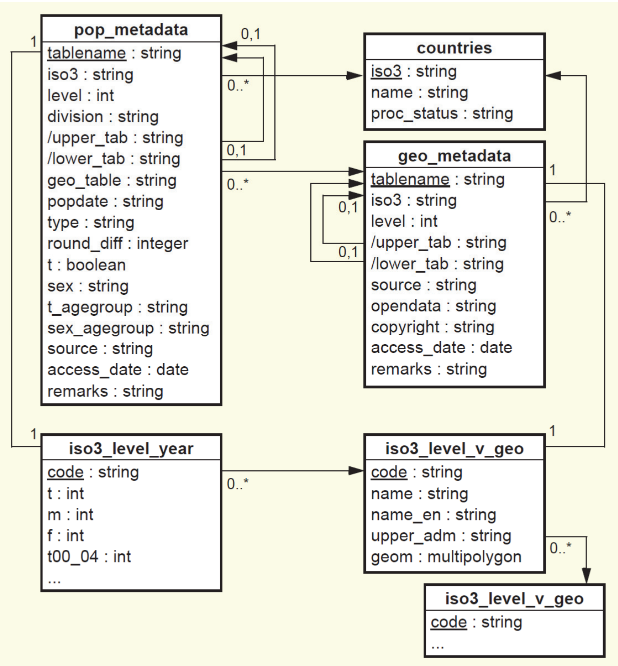

Figure 3. Basic data model.

Corresponding entries in a geospatial layer and in one or several population tables have the same identifier. The flexible support of population tables with different table schemas and different geometry layers for time-dependent divisions results in a large number of tables and hinders the database management system to check foreign keys. Instead, this task is performed by the aforementioned component for consistency checks.

The data model of Figure 3 does not explicitly handle censuses or other population surveys with changing boundaries. The “v” in the “iso3_level_v_geo” notation indicates that several versions of a division may exist. Then, we need an additional mapping between the entries of two different versions of a geometry layer. The previous discussion has made it clear that the level of information can vary greatly. Depending on the situation, different solutions should be applied.

Ideally, a lower division exists that allows us to build both versions of the modified division. Then, we can add a table describing the corresponding n-to-m relationship. As depicted in Figure 4, it relates two versions x and y of a layer of level i with a layer of level j and version x with $j > i$. In this case, we can compare the entries of versions x and y exactly. If the population layers of levels i and j have the same attributes, it is possible to compute all population attributes.

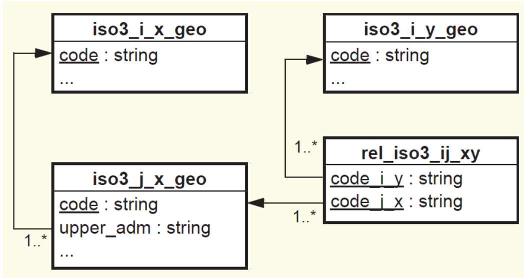

Figure 4. Modelling to relationship between two versions x and y of a layer of level i by referencing a layer of lower level j.

Lookup tables can represent relationships between subnational units that have remained unchanged or undergone modification between two subsequent censuses. Change indicators describe the possible cases. The UK Office for National Statistics (2025), for example, provides a lookup table between the output areas of the 2021 and 2011 census. It contains 189,567 entries and four types of change indicators. Of the 188,880 output areas in 2021,

- 175,469 are unchanged from 2011 (U),
- 12,654 are the result of splitting into two or more areas (S),
- 479 are the result of mergers of two or more areas (M), and
- 278 are the result of changes that are irregular and fragmented (X).

The identifiers of the units and the change indicator can be stored directly in the relationship table “rel_iso3_i_xy” depicted in Figure 5. As in the previous case, it refers to versions x and y of a layer of level i. The “frac_i_x” and “frac_i_y” attributes store the population share (the so-called split relation) of the referenced area compared to the other area. In case of unchanged units, both attributes have a value of 100%. In the event of a split, “frac_i_y” is 100%, while “frac_i_x” is less than 100%. The sum of all “frac_i_x” values stored for one area of version x should equal 100%. In the merge case, the behaviour is reversed: “frac_i_x” is 100%, while “frac_i_y” is less than 100%. The sum of all “frac_i_y” values stored for one area of version y is 100%. In the case of an irregular or fragmented split, the values of both attributes are less than 100%. The quality of comparisons and population computations depends on the accuracy of the split relation. In many cases, the split relation is estimated. Thus, we can only make approximate comparisons between versions x and y. Zoraghein et al. (2016) argue that the split relation should be supplemented with a quality indicator and metadata about the estimation technique used.

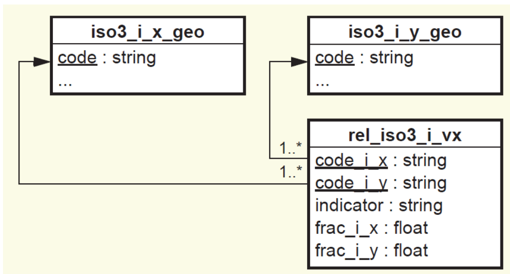

Figure 5. Modelling to relationship between two versions x and y of a layer of level i by using a lookup table.

3.4 Implementation

A first version of the spatiotemporal census data management system has been implemented. It uses PostgreSQL/PostGIS as spatial database system. Available data from the 2020 census round have been imported, verified for accuracy, and georeferenced. For countries with reliable population estimates, these estimates have been used in the absence of recent census data. Additionally, some data from the 2010 census round are included. The general export component has not yet been implemented. For now, QGIS is used to export GeoPackages containing population data and the corresponding geospatial layers.

As of mid-November 2025, population data from approximately 200 countries had been processed. These data are organized into 900 population tables and 670 geospatial layers, comprising approximately 10.8 million spatial units. This results in about 770 million individual population figures. Figure 6 illustrates statistical areas in New Zealand along with the corresponding 2018 census data.

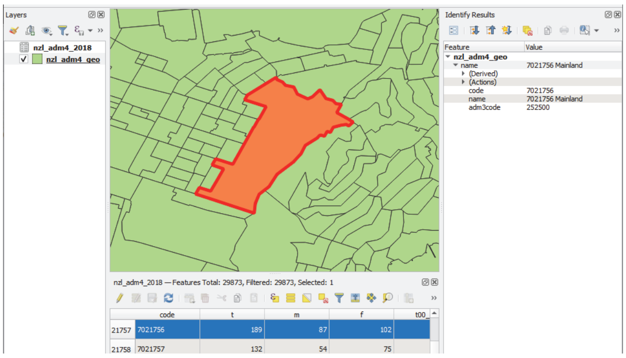

Figure 6. Statistical areas in New Zealand with corresponding 2018 census data depicted by QGIS.

The census data were incorporated into WorldPop’s gridded population datasets, which were published in September 2025 (Tatem, 2025).

4. Conclusions

We presented a spatiotemporal data management system for subnational population census data. The need for its implementation is driven by the poor interoperability of existing datasets and the lack of links to current, valid geospatial information. We derived the requirements for this system from our experiences with the “WorldPop Global Demographic Data” project. A first version of the data management system has been implemented and populated with data particularly from the 2020 census.

A complete set of spatiotemporal relationships and procedures cannot be elaborated using the current state of the database. Therefore, we need to add more datasets from past and future surveys.

The current data model assumes that the type of division and resolution of census data remains consistent throughout the country or at least within an administrative region. However, some countries, such as Chile and Mexico, provide differently structured datasets for urban and rural areas.

Additionally, most of the implemented tools are still in the prototype stage and require further development for more productive and convenient use.

5. Acknowledgements

This work was supported by funding from the Bill & Melinda Gates Foundation (INV-045237).

References

Alam, M., Torgo, L., Bifet, A, 2022. A Survey on Spatio-temporal Data Analytics Systems. *ACM Comput. Surv.*, 54, 10s, 219, doi.org/10.1145/3507904.

Brinkhoff, T., Kresse, W., 2022. Databases. In: Kresse, W., Danko, D. (eds.), *Springer Handbook of Geographic Information*, Springer, 53-93, doi.org/10.1007/978-3-030-53125-6_3.

CELADE, 2026. Redatam: Processing, analysis and dissemination of statistics and census around the world. redatam.org/en (20 March 2026).

City Population, 2026. Population Statistics in Maps and Charts. www.citypopulation.de (20 March 2026).

European Commission, 2013. INSPIRE Data Specification on Population Distribution – Technical Guidelines English. inspire-mif.github.io/technical-guidelines/data/pd/dataspecification_pd.pdf (20 March 2026).

Eurostat, 2026a. European Statistical System – Data Browser. ec.europa.eu/CensusHub2/selectHyperCube (20 March 2026).

Eurostat, 2026b. GISCO – the Geographic Information System of the Commission. ec.europa.eu/eurostat/web/gisco/overview (20 March 2026).

Institut Nationale de la Statistique de Guinée, 2026. Quatrième Recensement Général de la Population et de l’Habitation (RGPH-4) 2025 Résultats Préliminaires. www.stat-guinee.org/images/Documents/Publications/INS/rapports_enquetes/RGPH-4/Resultats_preliminaires_rgph_4.pdf (20 March 2026).

ISO, 2023. ISO 17369:2013 – Statistical data and metadata exchange (SDMX). www.iso.org/standard/52500.html (20 March 2026).

Langran, G., 1989. A review of temporal database research and its use in GIS applications. *International Journal of Geographical Information System*, 3, 3, 215-232, doi.org/10.1080/02693798908941509.

Martin, D., Cockings, S., Leung, S., 2015. Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers*, 105, 4, 754-772, doi.org/10.1080/00045608.2015.1022089.

OCHA, 2026a. HDX – The Humanitarian Data Exchange. data.humdata.org (20 March 2026).

OCHA, 2026b. Retiring HXL Services. centre.humdata.org/retiring-hxl-services/ (20 March 2026).

Pelekis, N., Theodoulidis, B., Kopanakis, I., Theodoridis, Y., 2004. Literature review of spatio-temporal database models. *The Knowledge Engineering Review*, 19, 3, 235–274, doi.org/10.1017/S026988890400013X, 2004.

Peuquet, D. J., 1999. Time in GIS and geographical databases. *Geographical Information Systems*, 1, 91-103.

SDMX Community, 2026. Statistical Data and Metadata eXchange. sdmx.org (20 March 2026).

Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V. M., Zhou, X., 2015. Spatiotemporal data mining: A computational perspective. *ISPRS International Journal of Geo-Information*, 4, 4, 2306-2338, doi.org/10.3390/ijgi4042306.

Sindoni, G., De Francisci, S., Paolucci, M., Tininini, L., 2002. Experiences in developing a spatio-temporal information system. *Research in Official Statistics*, 5, 1, 45-57.

Statistics Austria, 2026. STATcube – statistical database. www.statistik.at/en/databases/statcube-statistical-database (20 March 2026).

Statistics Sweden, 2026. Statistical programs for px files. www.scb.se/en/services/statistical-programs-for-px-files/ (20 March 2026).

Statistisches Bundesamt Deutschland, 2024. Ergebnisse des Zensus 2022 – Demografie. https://www.destatis.de/static/DE/zensus/gitterdaten/Regionaltabelle_Demografie.xlsx (20 March 2026).

Tatem, A., 2017. WorldPop, open data for spatial demography. Sci Data, 4, 170004. doi.org/10.1038/sdata.2017.4.

Tatem. A., 2025. WorldPop ‘Global2’: Global high-resolution population estimates for 2015-2030. www.worldpop.org/blog/worldpop-global2-global-high-resolution-population-estimates-for-2015-2030/ (20 March 2026).

UK Office for National Statistics, 2023. Protecting personal data in Census 2021 results. www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/protectingpersonaldataincensus2021results (20 March 2026).

UK Office for National Statistics, 2025. Open Geography Portal, OA (2011) to OA (2021) to Local Authority District (2022) Exact Fit Lookup in EW (V3). geoportal.statistics.gov.uk/datasets/93ffd0c524db494aa11914d44023c730_0/explore (20 March 2026).

UK Office for National Statistics, 2026a. Nomis – official census and labour market statistics. www.nomisweb.co.uk (20 March 2026).

UK Office for National Statistics, 2026b. Open Geography Portal. geoportal.statistics.gov.uk (20 March 2026).

United Nations Statistics Division, 2026. World Population and Housing Census Programme. unstats.un.org/unsd/demographic-social/census/ (20 March 2026).

U.S. Census Bureau, 2025. Subnational Population Data by Geographic Area. www.census.gov/geographies/mapping-files/time-series/demo/international-programs/subnationalpopulation.html (20 March 2026).

U.S. Census Bureau, 2026. Explore Census Data. data.census.gov (20 March 2026).

Willekens, F., 2023. Programmatic access to open statistical data for population studies: The SDMX standard. *Demographic Research* 49, 40, 1117-1162, doi.org/10.4054/DemRes.2023.49.40.

Worboys, M. F., 1994. A Unified Model for Spatial and Temporal Information. *The computer journal*, 37, 1, 26-34, doi.org/10.1093/comjnl/37.1.26.

Worboys, M. F., 2001. Modelling changes and events in dynamic spatial systems with reference to socio-economic units. In: Frank, A., Raper, J., Cheylan, J.-P. (eds.), Life and Motion of Socio-Economic Units, pp. 115-123, Taylor & Francis.

WorldPop, 2025. WorldPop Global Demographic Data. www.worldpop.org/current-projects/completed-projects/worldpop-global-demographic-data/ (20 March 2026).

Wu, B., Yu, B., Shu, S., Wu, Q., Zhao, Y., Wu, J., 2020. A spatiotemporal structural graph for characterizing land cover changes. *International Journal of Geographical Information Science*, 35, 2, 397-425, doi.org/10.1080/13658816.2020.1778706.

Zheng, L., Zhou, L., Zhao, X., Liao, L., Liu, W., 2017. The Spatio-Temporal Data Modeling and Application Based on Graph Database. *Proceedings 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 741-746, doi.org/10.1109/ICISCE.2017.159.

Zoraghein, H., Leyk, S., Ruther, M., Buttenfield, B. P., 2016. Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, 58, 19-28, doi.org/10.1016/j.compenvurbsys.2016.03.004.