https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

HISDAC-ES: Historical Settlement Data Compilation for Spain (1900 - 2020)

Johannes H. Uhl[1,2], Dominic Royé[3], Keith Burghardt[4], José A. Aldrey Vázquez[5], Manuel Borobio Sanchiz[5], Stefan Leyk[1,6]

[1]Institute of Behavioral Science, University of Colorado Boulder, Boulder, CO 80309, USA
[2]Earth Lab, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO 80309, USA
[3]Climate Research Foundation (FIC), 28013 Madrid, Spain
[4]Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
[5]Department of Geography, Universidad de Santiago de Compostela, 15703 Santiago de Compostela, Spain
[6]Department of Geography, University of Colorado Boulder, Boulder, CO 80309, USA

Correspondence to: Johannes H. Uhl (Johannes.uhl@colorado.edu)

Abstract. Multi-temporal measurements quantifying the changes to the Earth's surface are critical for understanding many natural, anthropogenic, and social processes. Researchers typically use remotely sensed earth observation data to quantify and characterize such changes in land use and land cover (LULC). However, such data sources are limited in their availability prior to the 1980s. While an observational window of 40 to 50 years is sufficient to study most recent LULC changes, processes such as urbanization, land development, and the evolution of urban, and coupled nature-human systems often operate over longer time periods covering several decades or even centuries. Thus, to quantify and better understand such processes, alternative historical-geospatial data sources are required that extend farther back in time. However, such data are rare and processing is labor-intensive, often involving manual work. To overcome the resulting lack in quantitative knowledge of urban systems and the built environment prior to the 1980s, we leverage cadastral data with rich thematic property attribution, such as building usage and construction year. We scraped, harmonized, and processed over 12,000,000 building footprints including construction years to create a multi-faceted series of gridded surfaces, describing the evolution of human settlements in Spain from 1900 to 2020, at 100 m spatial and 5 years temporal resolution. These surfaces include measures of building density, built-up intensity, and built-up land use. We evaluated our data against a variety of data sources including remotely sensed human settlement data and land cover data, model-based historical land use depictions, as well as historical maps and historical aerial imagery, and find high levels of agreement. This new data product, the Historical Settlement Data Compilation for Spain (HISDAC-ES), is publicly available (https://doi.org/10.6084/m9.figshare.22009643; Uhl et al., 2023a) and represents a rich source for quantitative, long-term analyses of the built environment and related processes over large spatial and temporal extents and at fine resolutions.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

1 Introduction

The built environment, encompassing cities, towns, villages, and transportation infrastructure connecting them, represents a fundamental component of our civilization. It determines the social, environmental, economic, identity-related, perceptual, safety- and health-related aspects of human settlements in urban and rural settings. The built environment interacts with human life and society in various ways. For example, the morphological structure and dimension of the built environment and of cities in general affects the efficiency and sustainability of cities and urban ecosystems, human health, economic development and social inequality (Alonso 1964; Ewing and Rong 2008; Saiz 2010; Seto et al., 2011).

Measuring the physical, functional, and socio-economic characteristics of the built environment, as well as their evolutionary trajectories, helps researchers to understand the development of complex urban systems, and enables informed decision making in urban and regional planning. Information about different dimensions of the built environment and the building stock can be obtained from a variety of data sources, such as remote sensing data or volunteered geographic information (VGI). In particular, detailed building data are critical for the development of long-term urban sustainability strategies (Hudson 2018). Data sources for contemporary studies or analyses covering the last 30 to 40 years include gridded data on impervious surfaces (e.g., Gong et al., 2020), built-up areas, building functions, building height and volume (Marconcini et al., 2020; Haberl et al., 2021; Pesaresi et al., 2021; Esch et al., 2022; Li et al., 2022; Schiavina et al., 2022), urban fabric classification (Demuzere et al., 2019), mass and material of the building stock (Haberl et al., 2021). Moreover, building-level data is available from industry-generated data sources, such as Google (Sirko et al., 2021), Microsoft[1], from VGI (OpenCityModel[2], Atwal et al., 2022), as well as increasingly from cadastral data sources (Uhl and Leyk, 2022a). In addition, (commercial) property / real estate data can be obtained through large-scale, data harmonization and dissemination efforts (e.g., ZTRAX[3], Regrid[4], ParcelAtlas[5], EuroGeographics[6]). Such efforts have catalyzed the data-driven study of environmental processes in general (Nolte et al., 2021) and opened up new avenues to increase our knowledge of the human habitat and its interactions from a multi-dimensional, quantitative perspective.

However, such multi-source, multi-modal data often suffer from spatial, temporal, or semantic inconsistencies, incompatibilities and impede the direct, quantitative analyses of the built environment from a multi-dimensional perspective. Moreover, while data on the contemporary state and recent history of the built environment are available for many places in the world, analysis-ready geospatial vector or raster data, or systematically georeferenced historical information for cities, towns and villages prior to the 1980s is generally scarce (Uhl and Leyk, 2022a).

[1] https://github.com/microsoft/GlobalMLBuildingFootprints
[2] https://github.com/opencitymodel/opencitymodel
[3] https://www.zillow.com/research/ztrax/
[4] https://regrid.com/
[5] https://www.boundarysolutions.com/BSI/ParcelAtlas/page1.html
[6] https://www.mapsforeurope.org/datasets/cadastral-all

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

We argue that cadastral data sources (i.e., parcel and building data including construction dates and other thematic information on building size, material, or function) allow to mitigate these two shortcomings and complement the traditional data sources (e.g., remote sensing data). Cadastral data are increasingly available as open data (von Meyer and Jones, 2013) and have been used in a variety of geographic, demographic, and economic studies (e.g., Tapp 2010; Leyk et al., 2014; Zoraghein et al., 2016; Nolte 2020; Domingo et al., 2023). In previous work, for example, (Uhl and Leyk, 2022a) integrated cadastral parcel data and building footprint data to generate multi-temporal building footprint data for some regions within the conterminous United States (CONUS), which constitutes a valuable data source for accuracy assessments of remote sensing derived built-up surface data (Leyk et al., 2018; Uhl et al., 2018; Uhl and Leyk, 2022b; Uhl and Leyk, 2022c), and remote sensing based construction year estimation (Uhl and Leyk, 2017; Uhl and Leyk, 2020). We also used cadastral and property data sources to create accessible, geohistorical data infrastructure on the built environment in the United States (Leyk and Uhl, 2018; Uhl et al., 2021; McShane et al., 2022), and demonstrated the value of such data for quantitative analyses of long-term urbanization and land development (Leyk et al., 2020; Uhl et al., 2021), road network evolution (Burghardt et al., 2022), long-term urban scaling analyses (Burghardt et al., 2023), long-term settlement trends in the context of natural hazards (Braswell et al., 2022; Iglesias et al., 2021), and for assessments of historical neighborhood changes (Connor et al., 2020).

Specifically, in our past work, we employed the Zillow Transaction and Assessment Dataset (ZTRAX), an industry-generated property dataset on over 150,000,000 properties in the US, resulting from a large cadastral data harmonization effort, to generate the Historical Settlement Data Compilation for the US (HISDAC-US). HISDAC-US consists of gridded datasets that measure built-up intensity and settlement age (Leyk and Uhl, 2018), building density (Uhl et al., 2021), and building function (McShane et al., 2022), at 250m spatial resolution from 1810 to 2015, and from 1940 to 2015, respectively. These datasets have widely been used by researchers for various scientific studies (e.g., Millhouser 2019; Balch et al., 2020; Mietkiewicz et al., 2020; McDonald et al., 2021; Ferrara et al., 2021; Boeing 2021; Li et al., 2021; Dornbierer et al., 2021; Millard-Ball 2022; Salazar-Miranda 2022; Wan et al., 2022).

A lack of comparable data outside of the US has impeded similar efforts for other regions of the world. However, the INSPIRE directive (“Infrastructure for Spatial Information in Europe”) has paved the way for the availability of open cadastral and building data for the member countries of the European Union (EU). INSPIRE is the legal framework to implement a European Spatial Data Infrastructure (SDI; Bernard et al., 2005; Minghini et al., 2021), enabling harmonized and searchable data resources across the European Union. One of the core components of INSPIRE are standardized metadata specifications (Cetl et al., 2019). Moreover, INSPIRE provides a taxonomy of geospatial data into 34 data themes, encompassing cadastral parcels, buildings, land cover, as well as population, environmental, and infrastructure-related topics (Minghini et al., 2021). One of the data themes defined in the INSPIRE data scheme is the building theme, representing a set of data models for spatial vector data on the geometric and thematic attributes of buildings, with a rich set of thematic

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

attributes describing the physical, functional, and temporal characteristics of buildings. The INSPIRE building model can be implemented at different geometric and thematic levels of detail, specified in the "core 2d", "extended 2d", "core 3d", and "extended 3d" profiles, accounting for different levels of data availability across EU member countries (Gröger and Plümer 2014).

Many member countries of the European Union have made such data publicly accessible (as regulated for instance in the EU PSI directive (European Union 2019), typically derived from cadastral data records, at varying levels of geometric detail and attribute completeness. We compared the building footprint data available for European countries and found that data from Spain has high levels of data coverage and attribute completeness. However, these building data are maintained by different institutions within Spain, i.e., the chartered communities ("diputaciones forales") of Navarra and the provinces of the Basque country, as well as the national cadastral agency ("Dirección General del Catastro") for the remaining autonomous communities[7], and are available as large, distributed datasets in slightly different data models and data formats, impeding direct and wide usage of these data for country-level analyses.

To increase the accessibility of such data, we obtained and harmonized INSPIRE-conforming building data from Spain to create an accessible and consistent geospatial data resource enabling the analysis of the built environment in Spain from a physical, functional, and temporal perspective. More specifically, we generated a set of fine-grained gridded surfaces describing physical, functional, and temporal dimensions of the built environment in Spain. These surfaces encompass, for example, the building area, the number of housing units, predominant land use type, and building age statistics at a fine spatial resolution of 100m × 100 m. Moreover, we used building age information available in these INSPIRE-conforming building databases to estimate and map historical building densities and built-up land from 1900 to 2020.

These gridded surfaces are intended to enable researchers from various disciplines to carry out fine-scale, multi-dimensional analyses of the built environment in Spain, consistently enumerated in a common spatial grid, and to facilitate long-term studies of the evolution of the built environment within an observational window of up to 120 years. We call this dataset the ***Historical Settlement Data Compilation for Spain (HISDAC-ES)*** and make all data publicly available (https://doi.org/10.6084/m9.figshare.22009643; Uhl et al., 2023a). This data descriptor presents our data curation effort (Section 2), highlights the resulting gridded surfaces (Section 3), and includes a thorough evaluation of the created data, encompassing several comparative analyses against a variety of independent data sources on land use and built-up land across both, space and time (Section 4), and concludes with some final remarks and an outlook (Section 5).

[7] Herein, we refer to the autonomous communities as "regions".

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

2 Data and methods

INSPIRE-conforming building footprint data for Spain is hosted by national, regional, or provincial authorities. The data processing workflow consisted of the following steps: 1) We acquired around 12,000,000 building footprints plus attributes as polygonal, geospatial vector data in Geographic Markup Language (GML) format, from official web resources using automated and manual downloads (Section 2.1.1). 2) We harmonized the data (Section 2.1.2), and 3) We aggregated the data into gridded surfaces, and computed zonal statistics at the municipality level (Section 2.1.3). Furthermore, we evaluated the resulting gridded surfaces through comprehensive comparisons with a wide variety of independent spatial datasets (Sections 2.2 and 2.3).

2.1 INSPIRE building data processing

2.1.1 Data collection

For most parts of Spain, INSPIRE-conforming building data is available through an ATOM interface. The ATOM feed format is an XML-based language that allows ~~the~~ automated, web-based content retrieval by providing a machine-readable web interface (IBM 2023). The ATOM XML files are organized in a hierarchical manner (see Table 1 for examples) and allow accessing the building data in Geographic Markup Language (GML) format, at the municipality level. We created a Python script to automatically download these GML files (https://github.com/johannesuhl/hisdac-es). In some cases (e.g., Basque country, Navarra region, which have their own cadastral systems) we manually downloaded the data available as Web Feature Service (WFS) (Table 1). All data was projected in UTM Zone 30N (EPSG:25830).

2.1.2 Data preprocessing

After downloading and gathering the building data for 8,131 municipalities, covering all regions of Spain, we first converted the polygonal building footprint data to centroids, to reduce the computational effort for the subsequent data processing. Despite the common INSPIRE framework, attribute names and building function classes differed slightly between the different data sources. Thus, we harmonized the data by renaming columns, and by applying a common building function classification scheme.

We decided to provide gridded surfaces in three different spatial reference systems: (a) ETRS89 UTM Zone 30N (EPSG:25830) for the Spanish mainland, Balearic Islands, as well as the exclaves Ceuta and Melilla located in Northern Africa; (b) REGCAN-95 (EPSG:4083) for the Canary Islands; and (c) ~~for all Spanish territory in~~ the reference grid of the European Environmental Agency (EEA), which is based on the ETRS89 Lambert Azimuthal Equal Area projection (LAEA; EPSG:3035) and is commonly used for pan-European statistical mapping. This way, users can refer to the data in UTM / REGCAN for mapping purposes (North-oriented, angle-preserving); while the datasets in area-preserving LAEA projection

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

can be used for statistical modelling and integration with other datasets (e.g., gridded statistical data from Eurostat[8] or gridded Spanish census data (INE grid[9]). Thus, we reprojected the building centroids into these reference systems, yielding two sets of harmonized building centroid data: (a) in UTM / REGCAN and (b) in LAEA projection.

After an initial examination of the attribute coverage and completeness, we decided to focus on six well-covered attributes, measuring different aspects of the built environment, including the construction year, building function, number of dwellings, number of building units, building indoor area, and building footprint area (Fig. 1). For clarity, the number of dwellings describes the number of housing units in residential buildings, whereas the number of building units counts the number of units within non-residential buildings (e.g., number of commercial businesses within a building complex, etc.). Furthermore, the building indoor area represents the attribute "official area" and measures the gross indoor area (across all stories) within a building.

2.1.3 Data aggregation

Based on the preprocessed, harmonized building data, and the six selected thematic attributes, we created a range of different aggregated representations of the data. These aggregations include (a) spatial aggregation into grid cells within regular spatial grids of 100m × 100m, (b) aggregation to the municipality level by calculating zonal statistics per municipality polygon, and (c) temporal aggregation by stratification into different temporal classes. The combination of spatial aggregation and temporal stratification applied to the different thematic attributes yields a range of different sets of gridded surfaces. For example, we calculated the sum and the mean of the building units (BUNITS) ~~per~~ dwelling (DWEL) over all buildings within a given grid cell, as well as the sum and mean building indoor area (BIA) and building footprint area (BUFA), respectively, based on the building centroids located within a grid cell. The resulting gridded surfaces represent ***physical features*** of the built environment. Similarly, we calculated the minimum, maximum, mean, median, mode, and the variety of construction years (COY) per grid cell, which measures settlement age (heterogeneity), and quantifies construction / remodeling activity within each grid cell. Thus, COY statistics describe the ***age-related features*** of the built environment.

We stratified the building records by their construction year into temporal classes (epochs) based on 5-year intervals (e.g., built-up before 1900. before 1905, etc.) and calculated the number (or density) of buildings (BUDENS), and the total building footprint area (BUFA) per grid cell, in each of these epochs from 1900 to 2020. We further binarized these grid cells to measure developed area (DEVA) (i.e., grid cells containing at least one building) and undeveloped areas in each epoch. These surfaces measure the ***physical evolution*** of the built environment in Spain. Similarly, we thematically disaggregated the building count surfaces per epoch based on the building function attribute of the buildings in each grid cell and epoch, yielding time series of function-specific building density surfaces, for six types of building functions (i.e., residential, commercial, industrial, agricultural, public services, and offices) as a proxy measure for ***built-up land use***

[8] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids
[9] https://www.ine.es/censos2011_datos/cen11_datos_resultados_rejillas.htm

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

evolution from 1900 to 2020. Table 2 provides an overview of the gridded surfaces and spatial variables generated by these data processing steps.

As mentioned before, all gridded surfaces are available in UTM Zone 30N for the Iberic peninsula, in REGCAN-95 for the Canaries, as well as in Lambert Azimuthal Equal Area (LAEA) projection for the whole extent. For selected variables, we also provide a time series of zonal statistics, aggregated to the municipality boundaries[10] (see Section 3.5). All data processing, as well as the evaluation experiments and data visualizations were implemented in Python 3.8, using libraries such as Numpy, Scipy, GDAL, Geopandas, Matplotlib, PIL, and ESRI ArcPy. The core component of our grid cell aggregation procedure is the "binned statistic 2d" function in Scipy[11]. The overall processing workflow from the INSPIRE-conforming building data to the spatial layers of the HISDAC-ES is shown in Fig. 2.

2.2 Evaluation data and agreement assessments

As historical spatial data is generally scarce, the evaluation of the produced historical data is difficult. In order to evaluate the quality of the produced spatial layers in the HISDAC-ES as thoroughly as possible, we employed a range of independent datasets that exhibit coherence to the spatio-temporal processes measured by HISDAC-ES and carried out different evaluations and cross-comparisons. Specifically, we used three types of spatial data for these experiments: (a) recent, remote sensing derived datasets (i.e., the Global Human Settlement Layer, Corine Land Cover), (b) spatial-historical land use models (i.e., History Database of the Global Environment), and (c) historical cartographic data (i.e., historical maps and urban atlases) and orthoimagery.

2.2.1 Global Human Settlement Layer

To evaluate the plausibility and reliability of the developed area (DEVA) layers from 1975 to 2015, we used built-up areas from the Global Human Settlement Layer (GHS-BUILT R2018, Florczyk et al., 2019) for comparison. The GHS-BUILT surfaces are derived from multispectral remote sensing data (Landsat sensors, Sentinel-2) and map built-up areas globally from 1975 to 2014, at a spatial resolution of 30m (Fig. 3a,b). They are accompanied by a rural-urban classification (settlement model GHS-SMOD, Fig. 3c,d). GHS-SMOD is available at a resolution of 1km and classifies each location on Earth into one of seven classes of urbanness, ranging from sparse rural settlements to high-density urban centers (Florczyk et al., 2019). While more recent versions of the GHS-BUILT are available at the time of writing, we decided to use the GHS-BUILT R2018A due to its fine spatial resolution, and because a lot of work has been done and published to quantify the accuracy of the GHS-BUILT R2018A across the rural-urban continuum, and over time (e.g., Liu et al., 2020; Uhl and Leyk 2022b,c), whereas little information is available on the accuracy of newer, multitemporal GHS-BUILT datasets.

[10] https://doi.org/10.7419/162.09.2020
[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binned_statistic_2d.html

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

We resampled the GHS-BUILT to the HISDAC-ES grid (i.e., upsampling from 30m to 100m spatial resolution) for the epochs 1975, 1990, 2000, and 2015 and thus, obtained binary grid cells (i.e., built-up vs. not built-up) for each epoch. To reduce spatial misalignment effects, we first upsampled the 30m GHS-BUILT data to a 10m x 10m grid, nesting within the 100m × 100m HISDAC-ES grid, and then downsampled to the target grid, assigning 1 (built-up) if at least one 10m grid cell within the target cell was labeled as built-up. We then quantified the agreement between the resampled, binary GHS-BUILT and the DEVA layers using Precision, Recall, and F1 score for each epoch and for each municipality. Moreover, we expected the agreement to vary across the rural-urban gradient (cf. Leyk et al., 2018; Uhl and Leyk, 2022a,b). Hence, we calculated these agreement metrics for each year, within each of the seven GHS-SMOD rural-urban classes. Moreover, we generated a continuous rural-urban index for each municipality based on the GHS-SMOD layers, constructed from the weighted histogram of SMOD class instances within each municipality polygon (see Uhl et al., 2022 for details), and assessed the municipality-level agreement trends across this rural-urban continuum (Section 4.1). Here, it is worth noting that we qualitatively compared the built-up areas from GHS-BUILT to the World Settlement Footprint (WSF) Evolution dataset (Marconcini et al., 2020), and found high levels of agreement between these two datasets (Fig. A1). Thus, herein, we compared HISDAC-ES to GHS-BUILT only.

2.2.2 Corine land cover data

While the comparison of DEVA and GHS-BUILT evaluated the presence / absence of buildings in the HISDAC-ES, we also used Corine Land Cover data (CLC, Büttner 2014) and compared it to the land use / building function layers in HISDAC-ES. Specifically, we obtained CLC data for the earliest (1990) and most recent (2018) available epoch (Fig. 3e,f), and resampled it from the original resolution of 30m to the HISDAC-ES grid of 100m spatial resolution, using a majority resampling rule. We then overlaid these resampled CLC surfaces with the land use-specific building count surfaces of the respective years, and cross-tabulated the building counts for each combination of INSPIRE building function class and CLC land cover class (Section 4.2).

2.2.3 History database of the global environment (HYDE)

While the remotely sensed data from the GHSL and CLC allows for assessing the plausibility of the HISDAC-ES since 1975 and 1990, respectively, it does not provide any insight into the plausibility of the long-term trends (1900-2020) measured in the HISDAC-ES. To account for this, we employed the History Database of the Global Environment (HYDE V3.2, Klein-Goldewijk et al., 2017), consisting of a set of global, gridded land use layers from 10000 BC to 2015, which are model-based and available at a spatial resolution of 5' × 5' (approx. 6km × 9km in Spain). Specifically, we used the layer "urban area fraction" from HYDE for each decade from 1900 to 2015 (Fig. 3g,h) and aggregated both the building footprint area (BUFA) and developed area (DEVA) from the HISDAC-ES to the HYDE grid cells. We then conducted a correlation and regression analysis to quantify the agreement between BUFA, DEVA, and the total urban area as reported in HYDE, per grid

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

235 cell. To account for potential regional differences in the agreement, we stratified our analyses into regions obtained from the NUTS (Nomenclature of territorial units for statistics) administrative dataset[12] (Section 4.3).

2.2.4 Historical maps and orthoimagery

While the evaluation approaches described in the previous sections are based on measured or modeled data, they suffer from measurement error, resampling errors, and other incompatibilities that may bias the comparative analyses. Thus, we decided
240 to include alternative, historical data sources into our evaluation analyses, allowing for a more unbiased evaluation of HISDAC-ES layers in early years. These data sources include (a) historical "planimetría" maps (shown for the city of Alicante in Fig. 4a). (b) aerial photographs from 1956 (Fig. 4b), and (c) an urban atlas (Remírez et al., 1988) depicting different urban development phases (Fig. 4c). Specifically, we manually digitized the areas developed in different time periods for the city of Alicante and Madrid, and obtained a similar vector dataset, depicting different historical urban
245 development phases for the city of Valencia (courtesy of Zornoza-Gallego 2022a). We quantitatively assessed the agreement between these historical urban extents and the MINCOY / DEVA layers (Section 4.4). Moreover, we visually compared the built-up areas depicted in "planimetría" topographic maps from around 1900[13] to the HISDAC-ES DEVA layer for several urban and rural places (Section 4.4). Finally, we manually delineated the approximate urban boundaries for the cities of Santiago de Compostela, Madrid, and Alicante based on aerial imagery from 1956[14] and compared them qualitatively to the
250 developed areas in HISDAC-ES in the same year (Section 4.5).

2.2.5 Attribute completeness

In addition to the comparison to external datasets, we also aimed to quantify internal uncertainties in the HISDAC-ES, or the underlying INSPIRE building data, respectively, by assessing the completeness of relevant building attributes at the municipality level (Section 4.6).

255 ## 3 Results

In this section, we present the different spatial layers contained in HISDAC-ES, resulting from the spatial aggregation and temporal stratification of the INSPIRE-conforming building data. This includes the gridded surfaces related to the four thematic components of HISDAC-ES (physical, temporal, physical and land use evolution, Sections 3.1 to 3.4), as well as the municipality-level statistics (Section 3.5).

[12] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts
[13] Maps from 1870-1950, predecessors of the minutas catastrones (MTN50), scale 1:50,000, available at http://www.ign.es/wms/minutas-cartograficas?request=GetCapabilities&service=WMS
[14] Ortofotos AMS (B) 1956-1957 (IGN, Instituto Geográfico Nacional), available at http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=FPNOA

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

3.1 Physical characteristics

Fig. 5 displays the gridded surfaces measuring selected contemporary, physical features (in the year 2020) of the Spanish built environment, exemplarily shown for the city of Valencia. ~~These sum surfaces~~ exhibit interesting spatial patterns of the density of building indoor area (BIA) and footprint area (BUFA), decreasing from the city center towards the outskirts, whereas the density of dwellings (i.e., housing units) is higher in the outskirts than in the center part. The grid cell means of these variables are measures of (vertical / horizontal) building size, and exhibit different patterns, illustrating the presence of large multi-apartment complexes in the outskirts and small, historical buildings in the center part of the city.

3.2 Temporal characteristics

Temporal, or age-related statistics of the Spanish building stock are measured by different statistics calculated using the construction year attribute of the buildings within each grid cell. For example, the minimum construction year (MINCOY) impressively shows the settlement age patterns in the metropolitan area of Valencia (Fig. 6), depicting the historical city core, as well as recently developed suburban areas in the urban fringe, and older settlements in the surrounding villages. Similar patterns can be observed in the mean (MEANCOY), median (MEDCOY), and the most frequent (i.e., mode) construction year (MODECOY). The maximum construction year (MAXCOY) per grid cell measures the year of the last modification of the building stock, and alongside with the construction year variety (i.e., the number of unique construction years per grid cell, VARCOY) is a measure of construction activity, highlighting areas characterized by heavy urban renewal processes. Here it is worth noting that the construction year on record may also represent the year of the last building reformation, which introduces certain bias in the created surfaces (see Discussion section).

3.3 Evolutionary characteristics

Grid-cell level statistics (i.e., sums of BUFA, counts / densities of buildings BUDENS) stratified by the construction year attribute yield a time series of gridded surfaces measuring the long-term evolution of cities, towns, and villages. For example, the BUFA and BUDENS surfaces show how the built-up intensity (as measured by built-up surface density and building density) has increased from 1900 to 2020 (Fig. 7). These surfaces uniquely document the long-term urban growth processes, measured at fine spatial granularity and over long time periods. The derived DEVA surfaces show developed / undeveloped land for each point in time, facilitating quantitative, multitemporal analysis of urban form, e.g., using landscape metrics (Uhl et al., 2021). The high temporal resolution (i.e., 5 years) of these multi-temporal layers enables measuring the evolution of urban extents and building density at fine spatial and temporal detail, as illustrated in a complete time series for the city of Valencia (Appendix Fig. B1). The additional stratification of the BUDENS surfaces by building function disaggregates the building stock spatially, by age, and by function. As an example shown in Fig. 7 (right column), industrial land use has heavily increased in suburban areas, in particular in the Northern suburbs (between 1900 and 1960) and later in the Southern suburbs (after 1960).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

While the examples in Figs. 5-7 show the city of Valencia, we would like to emphasize the country-wide coverage of HISDAC-ES. For example, the minimum construction year surface (MINCOY) reveals commonalities and differences between settlement age patterns in different cities in Spain (Fig. 8), including polycentric development (e.g., Barcelona, Sevilla) and monocentric development (most other cities shown). Moreover, HISDAC-ES not only measures urban development across different cities, e.g., by means of the MINCOY surface (Fig. 8), but also long-term land development processes in rural areas, including towns, villages, and scattered, unincorporated settlements, as exemplified by the DEVA surfaces (Fig. 9). The DEVA surfaces reveal further detail on the spatial configuration of cities in early years (Appendix Fig. B2). We also provide several supplementary animated data visualizations illustrating the value of HISDAC-ES for quantifying long-term urbanization processes (see Video Supplement).

3.4 Built-up land use surfaces

Lastly, we show the building density surfaces stratified by building function, measuring the spatial distribution of different built-up land use classes (Fig. 10). These surfaces not only highlight the dominance of residential land use, but also illustrate peri-urban clusters of industrial land use, as well as spatial patterns of commercial land use, which has a mixed clustered and scattered spatial pattern, or agricultural land use, mostly occurring in peri-urban areas to the Northeast of Valencia. These surfaces, along with the corresponding multi-temporal land use surfaces (Fig. 7d) enable the quantitative assessment of land use specific evolution of the built environment and add a unique thematic component to the HISDAC-ES data layers.

3.5 Municipality-level statistics

We provide zonal statistics of INSPIRE-conforming building footprint data for all 8,131 municipalities in Spain as tabular data and geospatial vector data. These datasets contain the zonal sums of grid-cell level variables (i.e., building counts, as well as BUFA, BIA, DWEL, BUNITS) as well as corresponding densities (per municipality area) and allow for coarse-scale analyses, and for the joint analysis with historical population data, available at the municipality level. The visualizations in Fig. 11 illustrate the usefulness of such aggregated statistics to observe and quantitatively assess broad-scale settlement and building stock age patterns. These patterns can be interpreted in the context of historical settlement development, but also provide insight into the contemporary building stock age and its spatial variation. As the absolute counts per municipality may be affected by regional trends of municipality area (Fig. 11, top row), we also provide these statistics as densities normalized by the municipality area, which show a different picture (Fig. 11, bottom row), see Video Supplement.

4 Evaluation

We compared the layers from HISDAC-ES to a variety of related but independent datasets to evaluate spatial, temporal, and thematic components of our data (Sections 4.1-4.5). These efforts are summarized in Table 3. Moreover, we assessed the attribute completeness of the building data underlying HISDAC-ES (Section 4.6).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

4.1 Multi-temporal built-up area evaluation (1975-2014)

The comparison of the developed area (DEVA) to built-up areas from the GHS-BUILT reveals several trends: 1) The agreement between DEVA and GHS-BUILT changes from rural to urban areas. Specifically, the precision of DEVA is high in urban areas, and low in rural areas (Fig. 12a). A low precision in an agreement assessment would indicate high commission errors, which in this case implies that DEVA labels much more grid cells as built-up than GHS-BUILT. This is encouraging as previous work has shown that the GHS-BUILT tends to underreport built-up areas in rural regions (Leyk et al., 2018; Uhl and Leyk, 2022), and thus, the DEVA layers appear to account for this shortcoming. Similarly, recall of DEVA is slightly lower in urban areas such as the Madrid region (Fig. 12a). As previous work revealed, GHS-BUILT tends to overestimate built-up areas in urban settings (i.e., roads are often classified as built-up). The DEVA layers are not affected by this type of misclassification, resulting in lower recall values. Hence, both low precision (rural settings) and low recall (urban settings) imply high accuracy in the DEVA layers, as the reference data (i.e., the GHS-BUILT) suffer from the described shortcomings. 2) We observe an increase of precision over time (Fig. 12b), which is likely due to increasing completeness of built-up areas in the GHS-BUILT, particularly in rural areas. Recall, however, shows a different trend over time (Fig. 12b), maximizing, on average across all municipalities, in 1990 and decreasing towards recent epochs. This is likely a ~~superposed~~ effect of (a) increasing incompleteness in GHS-BUILT as we go back in time due to poorer quality of underlying Landsat data, and (b) increasing incompleteness in DEVA as we go back in time due to a survivorship bias in the INSPIRE building footprint data. Specifically, new buildings that replace an existing (old) building will be attributed with the construction year of the replacement, and the building that existed prior to the replacement is not contained in our data. Thus, urban renewal causes this bias in our data, and this bias manifests in lower recall values towards early points in time. Looking at the agreement trends over time and across the GHS-SMOD rural-urban classes (Fig. 12c), we observe a sharp increase of agreement from rural to urban areas, and a slight increase over time, implying that the reliability of the DEVA layers is highest in urban centers. Lastly, the distributions of municipality-level agreement metrics across rural-urban strata confirm this trend (Fig. 12d). The peaks in recall in the low-density and rural cluster strata, across all years, indicate that the effects of incompleteness in the reference data (caused by omitting rural settlements) and in the DEVA (caused by survivorship bias) are of similar magnitude and thus, cause higher levels of agreement. Here it is noteworthy that the more recent GHS-BUILT v2022 is likely to perform better in rural areas, and thus, precision of the HISDAC-ES is expected to increase in such areas, and the agreement gradients across the rural-urban continuum are expected to be less steep (cf. Uhl et al., *forthcoming*). However, as mentioned above, we prefer to use the GHS-BUILT R2018A because its accuracy has been well-studied and makes our interpretations more robust.

4.2 Land use evaluation 1990-2020

We cross-tabulated the land use-specific building counts from HISDAC-ES in 1990 and 2020, within the land cover classes from CLC for the years 1990 and 2018 (Fig. 13). The absolute building counts per land cover class in Fig. 13a indicated that

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

most buildings underlying the HISDAC-ES fall into urban fabric, industrial or commercial areas, or in agriculturally used areas, and to a lesser extent, in areas characterized by forest (Fig. 13a). When plotting the proportions of buildings per HISDAC-ES land use class (Fig. 13b) or per CLC land cover class (Fig. 13c), we observe more interesting patterns. For example, most buildings of any land use (except agriculture) are located in areas of continuous urban fabric. Agriculturally used buildings are mostly located in areas classified as "complex cultivation patterns" in CLC. This indicates that the agricultural land use as reported in the INSPIRE building data is highly accurate. Moreover, in 2020, the proportion of buildings in "discontinuous urban fabric" has increased, as compared to 1990, which may be an effect of suburbanization, and increasing low-density built-up areas. Finally, the cross-tabulation relative to the CLC classes show that the residential land use is the most dominant across all land cover classes, with a few interesting exceptions: Industrial land use has also high proportions in CLC "industrial or commercial units", and INSPIRE "public services" buildings have a peak in "port areas" and "airport" CLC classes. Agricultural buildings peak in CLC classes "rice fields", "annual crops", and "agro-forestry areas", which confirms the high levels of coherence between the two datasets. The peak of agricultural buildings in the "inland marshes" class may indicate higher levels of confusion between CLC classes "inland marshes" and "rice fields" which may be difficult to differentiate. All these observations confirm that the land use information from the INSPIRE building data and the derived HISDAC-ES land use layers seem highly plausible. Note that there is a slight temporal gap between the two datasets, as the most recent CLC data is from 2018. However, we expect this discrepancy to be of minor importance.

4.3 Long-term trajectory evaluation (1900-2015)

While the comparisons of HISDAC-ES to the GHS-BUILT and Corine Land Cover data focus on recent decades, the comparison to HYDE's urban fraction estimates examines the long-term agreement with the BUFA and DEVA evolution layers. We observe that the correlation between urban area fraction and BUFA is high after 1980 (i.e., >0.8 for most regions), and decreases as we go back in time, but is still at around 0.6 in 1900, in most regions (Fig. 14a). This drop in correlation could be explained by the previously mentioned survivorship bias in HISDAC-ES, but could also be attributed to uncertainties in the model-based urban area fractions in HYDE. The correlation to DEVA (Fig. 14b) shows a similar trend, but slightly lower levels of correlations. Interestingly, correlations are highest in the Southern region (i.e., Andalucia), possibly due to low levels of building stock renewal and thus, a weaker effect of the survivorship bias in the HISDAC-ES data. Besides this comparison of quantitative measures per grid cell, we also compared the agreement between developed and undeveloped grid cells in DEVA and HYDE, as measured by the time series of F1 scores in Fig. 14c. We observe extremely high agreement (>0.95) in recent decades, and just slightly lower F1 scores in the beginning of the 20th century. These high levels of agreement of the HISDAC-ES and model-based urban area estimations from HYDE underline the high quality of the HISDAC-ES evolutionary layers. It is noteworthy that the high F-1 scores may be an effect of the relatively low spatial resolution of HYDE (5' × 5' grid cells).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023


Open Access Earth System Science Data Discussions

385 **4.4 Comparison to historical maps**

The previous evaluations are based on either remotely sensed data, or model-based reference data. Thus, those datasets are limited in their temporal coverage, or suffer from uncertainty themselves. Hence, we used multi-temporal urban areas manually digitized from historical maps, covering the time period from approximately 1900 to present, for the cities of Alicante, Madrid, and Valencia (Fig. 15). These extents were manually digitized for Alicante and Madrid from an urban atlas
390 (Remírez et al., 1988; Valencia data are courtesy of C. Zornoza-Gallego), resulting in increments of urban area newly added in a given time period (see Fig. 4a,b). We rasterized the resulting vector data in the HSDAC-ES grid, encoding the earlier year of each time period (Fig. 15a), ~~using the sme grid like the HISDAC-ES MINCOY settlement age surface (Fig. 15b)~~. For each city, we reclassified the MINCOY surface to match the settlement age categories from the digitized urban areas, and calculated agreement measures for each city, based on both the cumulative urban area per point in time, and on the newly
395 added built-up area in each time period (Table 4). As the MINCOY surface only encodes the year of earliest settlement per grid cell, disregarding the settlement density in that year, we also used the BUDENS surfaces for the respective years (shown for 1900 and 2015 in Fig. 15c,d), to weight each grid cell by its building density. We did this because we assumed that misclassifications are more likely in sparsely populated areas, likely not contained in the urban extents from the historical urban atlases due to generalization. Thus, agreement metrics based on building counts rather than grid cell counts, would be
400 more representative and realistic for a comparison between these two datasets. Based on this evaluation strategy, we observe the following:

1) **Agreement levels are generally relatively low (F1 between 0.28 and 0.74).** This may be attributed to the survivorship bias in HISDAC-ES, but also due to definitional differences (i.e., HISDAC-ES measures built-up area, while the urban extents derived from historical atlases measure urban area, and thus, are already a generalized
405 representation of the developed land in a given point in time . They are likely to omit low-density, scattered, peri-urban settlements, but include roads, impervious surfaces, intra-urban greenspaces (e.g., parks, cemeteries, etc.) which are not directly measured in HISDAC-ES which is based on the presence of built structures only.
2) **Recall is higher than precision.** Low precision indicates high commission error (i.e., overestimation), likely because peri-urban, rural settlements are contained in HISDAC-ES, but not in the urban extents due to
410 generalization and the (arbitrary) definition of the urban boundaries in the historical urban atlases, similar to what we observed in the comparison to the GHSL (Section 4.1).
3) **Agreement for cumulative urban extents is higher than for incremental time slices**. This effect is to be expected, as the confusion between historical increments is irrelevant when comparing the total built-up / urban area in a given point in time. As the urban areas (and the increments) in early time periods can be small,
415 misclassification is more likely, also due to higher levels of survivorship bias in HISDAC-ES for early time periods.
4) **Precision based on the number of buildings is higher than precision based on the number of "occupied" grid cells.** This indicated that grid cells label ed as "built-up" in HISDAC-ES but not in the historical urban areas tend to

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

have low building density, confirming our aforementioned assumption that low-density settlements are not mapped in the historical urban extents.

5) **For cumulative urban extents, precision and recall increase over time.** This is a direct effect of the survivorship bias, manifesting in higher omission errors (and thus lower recall) as we go back in time. Moreover, the proportion of scattered low-density settlements (which are not contained in the historical urban atlases) in relation to dense, urban settlements was higher in early than in recent epochs, resulting in an increase of precision over time.

Moreover, the agreement levels are relatively similar across the three cities under study, implying that these observations are likely to be generalizable at least across the larger cities in Spain. Despite the low absolute numbers of the agreement metrics reported in Table 4, which are likely due to definitional differences, the observed trends are in line with theoretical expectations (e.g., survivorship bias decreases over time), and with evaluation results of the HISDAC-US (Uhl et al., 2021), indicating similar characteristics of the historical settlement layers derived from cadastral / property data in the U.S. and in Spain. Confusion matrices underlying the agreement metrics shown in Table 4 can be found in Appendix Fig. E1.

The qualitative comparison of the DEVA layers and "planimetría" historical maps (predecessors of the "Minutas catastrones" from 1870-1950, at scale 1:50,000) confirms the previously made observations. As shown in Fig. 16, the DEVA (1900) layer mimics quite well the urban areas as depicted in the maps, with some omission errors mainly in peri-urban areas, e.g. for Madrid, Sevilla, and Tarrasa. This could indicate that building replacements (causing survivorship bias in the HISDAC-ES) tend to occur least in the medieval city centers, which are subject to monument protection (assuming concentric growth patterns). Moreover, these disagreement patterns may also be due to temporal inconsistencies between DEVA (1900) and the "planimetría" maps, which may have slightly different temporal references. In the small villages around Hornillos del Camino (Burgos) (Fig. 16 right column) we rather observe over- than underestimation. This could indicate that in rural, economically less prosperous areas, where less building remodeling occurs, the original building stock is still dominating and thus, HISDAC-ES is less affected by survivorship bias. This observation may imply higher levels of data quality in small, rural places, a promising perspective for long-term settlement modeling in the often understudied rural settings. The bottom row (Fig. 16c) shows the contemporary (i.e., 2020) building densities, illustrating a positive association between building density and settlement age. On a side note, the building density in the small, rural communities around Hornillos del Camino is similar to the densities in the center parts of the large cities.

4.5 Comparison to historical orthophotos from 1956

The urban extents manually digitized from historical orthophotos acquired in 1956 show the urban boundary in those years in Alicante, Madrid, and Santiago de Compostela (Fig. 17 a,b,c). When we overlay these boundaries to the built-up areas from HISDAC-ES in the same year, we observe varying levels of agreement: In Alicante the agreement is relatively high (Fig. 17d), as well as in Santiago (Fig. 17f), whereas we observe higher levels of disagreement in the Madrid data (Fig. 17e), mainly occurring in suburban areas. While some of the disagreement may be attributed to differences in definitions (i.e. , the

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

urban boundaries drawn in the orthophotos only include dense, urban settlements), in the Madrid case there are additional issues, related to notable activities of urban renewal in previously informal settlements (e.g., the Entrevías neighborhood (example 4 in Fig. 17e). See Fig. F1 for a more detailed discussion of historical reasons for these discrepancies.

4.6 Attribute completeness

Lastly, we report the INSPIRE building attribute completeness at the municipality level, ~~as referred~~ to the total number of buildings available in each municipality (Fig. 18). We observe very high levels of completeness of the construction year attribute. The building function attribute has a high coverage, except in the Gipuzkoa and Bizkaia provinces. Moreover, there is a building function attribute "other" that is only available in the Navarra region, which we excluded from the HISDAC-ES dataset. Thus, the uncertainty in the (historical) land use layers in HISDAC-ES in Navarra is slightly higher, as it is unknown what building function the "other" class encompasses. The indoor area, number of dwellings, and number of building units attributes have also lower levels of completeness in some areas of the Basque country. Conversely, information on the number of floors is highly complete in the Navarra region, but otherwise not covered in the remaining provinces, and thus, has not been used in this version of HISDAC-ES. Generally, these high levels of attribute completeness are encouraging and indicate that the layers derived from these attributes are expected to be highly reliable at least for recent points in time. We made these municipality-level attribute completeness statistics available in the data repository.

6 Data availability

All datasets are available at https://doi.org/10.6084/m9.figshare.22009643 (Uhl et al., 2023a). All raster datasets are available in LZW-compressed GeoTIFF format and have a spatial resolution of 100m × 100m. All raster datasets are available in EPSG:3035 (LAEA, all Spanish territory), EPSG:4083 (REGCAN, Canary Islands), and EPSG:25830 (UTM30N, Iberic peninsula). The raster datasets are organized in subfolders as follows: They are grouped by geographic coverage (all, can, ibe) and reference system (laea, regcan, utm), and by theme (evolution, landuse, physical, temporal). For example, "ibe_utm_temporal" contains the surfaces measuring temporal characteristics of the built environment, covering the Iberic peninsula, referenced to the UTM grid. In total, there are 717 GeoTIFF files. Municipality-level aggregates and uncertainty measures are available in CSV format. In addition to that, municipality-level uncertainty metrics are available as GeoPackage (.gpkg) format (EPSG:25830) including municipality boundaries. The CSV files can be joined to the municipality GPKG using the "lau_code" attribute field. The total uncompressed data volume is around 3GB. The data in the repository is partitioned in four ZIP-compressed archives, one for the raster data in each of the three spatial reference systems, and one for the municipality-level aggregates.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

7 Code availability

The Python code used to create HISDAC-ES is publicly available at https://github.com/johannesuhl/hisdac-es. R users can access the Spanish cadastral data underlying HISDAC-ES using the CatastRo package, which is available at (https://ropenspain.github.io/CatastRo/index.html), and a comprehensive instructions for building age visualization in R can be found at https://dominicroye.github.io/en/2019/visualize-urban-growth/.

8 Conclusions

In this data descriptor, we presented the creation and characteristics of HISDAC-ES, a set of geospatial raster and vector layers measuring the built environment in Spain from different perspectives, including physical, temporal, evolutionary, and functional aspects. HISDAC-ES aims to (a) facilitate the access to and use of information derived from cadastral building data by spatial, temporal, and semantic aggregation, (b) provide empirically measured, historical geospatial data, enabling contemporary, but also long-term, historical analyses of urban growth, sprawl, and change; and (c) demonstrate the usefulness of cadastral data for geographic applications in general and domains of social and environmental sciences, more specifically. HISDAC-ES represents an extension of related work, recently conducted on U.S. property data (HISDAC-US; Leyk and Uhl, 2018; Uhl et al., 2021; McShane et al., 2022), and demonstrates the benefit of open data policies and large-scale data harmonization efforts for the scientific community and beyond.

HISDAC-ES provides a valuable data source for urban analysts, regional planners, and policy makers, enabling or upscaling the quantitative measurement and interpretation of long-term urbanization and land development processes (e.g., Arribas-Bel et al., 2011; Alvarez-Palau et al., 2019; Sapena and Ruiz, 2019; Zornoza-Gallego 2022b; Domingo et al., 2023). Together with the sister product HISDAC-US, it will enable the comparative study of urban size, shape, and morphology over long time periods, across different continents and across historical as well as cultural settings.

We evaluated the agreement of HISDAC-ES with a range of related datasets obtained from remote sensing data and historical maps, identifying varying levels of agreement. While the associations between datasets imply some level of coherence and are generally promising, it is important to point out that a rigorous quality assessment of historical geospatial data is difficult. The main reasons are the general lack of reliable, historical reference data, but also differences in definitions and semantic discrepancies (ambiguity) between the evaluation datasets and HISDAC-ES, as well as vagueness in the evaluation datasets (e.g., arbitrarily defined urban boundaries).

Nonetheless, there are a few shortcomings of HISDAC-ES that need to be addressed in the future. The main issue is the survivorship bias in the data: We infer settlement age based on the construction year on record in the cadastral building data. It remains unknown whether a construction date represents the first establishment of a building at a given location, or whether there was a built-up structure existing prior to that. Similarly, buildings that existed in the past, but have been demolished, are not contained in HISDAC-ES. Thus, HISDAC-ES can only measure urban growth, but not urban shrinkage or urban renewal). Fortunately, the latter process is rare. It is also unknown if the different attributes in the cadastral building

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

data underlying the HISDAC-ES were measured at the same time. For example, the building function, indoor area, etc. reflects the contemporary state of a building (i.e., as of the year 2020), but these properties may have changed since the construction year on record, which may introduce additional uncertainty in the evolutionary layers in HISDAC-ES. Finally, the gridded surfaces in HISDAC-ES are based on discrete point locations, rather than the actual building footprint geometries, in order to reduce computational processing effort. Thus, large buildings extending across two or more grid cells may not be represented correctly in HISDAC-US, introducing some positional uncertainty in the data.

Future work should focus on validation of the HISDAC-ES dataset, for example by employing large-scale historical map collections (cf. Olazabal et al., 2019), or other historical records. The integration of HISDAC-ES with historical population data in a dasymetric modeling framework could be useful to create fine-grained, historical population estimates (cf. Burghard et al., 2023). Moreover, other components of Spanish INSPIRE-conforming building data could be used, such as sub-building level information (e.g., building parts), to create fine-grained data on building function at the sub-building level. Lastly, with the prospect of increasing availability of INSPIRE-conforming cadastral building data, HISDAC-related efforts will be expanded to other European countries where cadastral building data is of similarly high completeness, quality, and thematic richness.

Video supplement

We provide eight animated data visualizations (available at http://doi.org/10.6084/m9.figshare.22064798, Uhl et al., 2023b), showcasing several data layers of the HISDAC-ES, at the grid cell level and at the municipality level:

1) Evolution of developed areas (DEVA) in 30 urban centers in Spain (01_hisdac_es_developed_area_evolution.gif)
2) Evolution of building density (BUDENS) in 30 urban centers in Spain (02_hisdac_es_building_density_evolution.gif)
3) Evolution of building footprint area (BUFA) in 30 urban centers in Spain (03_hisdac_es_building_footprint_area_evolution.gif)
4) Evolution of residential land use in 30 urban centers in Spain (04_hisdac_es_residential_landuse_evolution.gif)
5) Evolution of commercial land use in 30 urban centers in Spain (05_hisdac_es_commercial_landuse_evolution.gif)
6) Evolution of industrial land use in 30 urban centers in Spain (06_hisdac_es_industrial_landuse_evolution.gif)
7) Evolution of building footprint area (BUFA) per municipality (07_hisdac_es_municipality_bufa_density.gif)
8) Evolution of building density (BUDENS) per municipality (08_hisdac_es_municipality_building_density.gif)

Animations 1-3 are shown from 1920 to 2020. Animations 4-6 are shown from 1950 to 2020. Note that for the land-use related animations (4-6), we binarized the building density layers stratified by land use category (i.e., highlighting grid cells where at least one building of the respective land use class exists). Animations 7 and 8 show the municipality-level

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

aggregates, converted into densities (i.e., built-up area per km², buildings per km²), shown in percentiles based on the data distributions across all years.

Author contribution

JU, DR and SL designed the data model. JU and DR acquired the source data. DR, MBS, JAV and JU gathered evaluation data. JU implemented the code for data production. JU, DR, SL and KB designed the evaluation experiments. JU implemented and carried out the data production process and the evaluation analyses. JU and DR visualized the results. JU prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

Thanks to Dr. Carmen Zornoza-Gallego for providing data on historical urban extents for the city of Valencia, used for the evaluation studies presented herein. Partial funding for this work was provided through the Humans, Disasters, and the Built Environment and the Human Networks and Data Science – Infrastructure programs of the US National Science Foundation (Award Numbers 1924670 and 2121976, respectively) to the University of Colorado Boulder. Moreover, this research benefited from support provided to the University of Colorado Population Center (CUPC, Project 2P2CHD066613-06) from the Eunice Kennedy Shriver Institute of Child Health Human and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH or CUPC. Finally, this study was supported partially by the Open Philanthropy Project.

References

Alonso, W. (1964). The historic and the structural theories of urban form: Their implications for urban renewal. *Land Economics, 40(2), 227-231.*

Alvarez-Palau, Eduard J., Jordi Martí-Henneberg, and Jorge Solanas-Jiménez. 2019. "Urban Growth and Long-Term Transformations in Spanish Cities Since the Mid-Nineteenth Century: A Methodology to Determine Changes in Urban Density" *Sustainability* 11, no. 24: 6948.

Arribas-Bel, D., Nijkamp, P., and Scholten, H. (2011). Multidimensional urban sprawl in Europe: A self-organizing map approach. *Computers, environment and urban systems, 35*(4), 263-275.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Atwal, K. S., Anderson, T., Pfoser, D., and Züfle, A. (2022). Predicting building types using OpenStreetMap. *Scientific Reports, 12(1), 19976.*

Balch, J. K., Iglesias, V., Braswell, A. E., Rossi, M. W., Joseph, M. B., Mahood, A. L., ... and Travis, W. R. (2020). Social-environmental extremes: Rethinking extraordinary events as outcomes of interacting biophysical and social systems. *Earth's Future, 8(7), e2019EF001319.*

Bernard, L., Kanellopoulos, I., Annoni, A., and Smits, P. (2005). The European geoportal—one step towards the establishment of a European Spatial Data Infrastructure. Computers, environment and urban systems, 29(1), 15-31.

Boeing, G. (2021). Off the grid… and back again? The recent evolution of American street network planning and design. *Journal of the American Planning Association, 87(1), 123-137.*

Braswell, A. E., Leyk, S., Connor, D. S., and Uhl, J. H. (2022). Creeping disaster along the US coastline: Understanding exposure to sea level rise and hurricanes through historical development. *PLoS one, 17(8), e0269741.*

Burghardt, K., Uhl, J. H., Lerman, K., and Leyk, S. (2022). Road network evolution in the urban and rural United States since 1900. *Computers, Environment and Urban Systems, 95, 101803.*

Burghardt, K., Uhl, J. H., Lerman, K., and Leyk, S. (2023). Analyzing urban scaling laws in the United States over 115 years. *arXiv preprint arXiv:2209.10852.*

Büttner, G. (2014). CORINE land cover and land cover change products. *Land use and land cover mapping in Europe: practices & trends, 55-74.*

Cetl V, de Lima VN, Tomas R, Lutz M, D'Eugenio J, Nagy A, Robbrecht J (2017) Summary report on status of implementation of the INSPIRE directive in EU. JRC Technical Report. https://doi.org/10.2760/143502. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC109035/jrc109035_jrc109035_jrc_inspire_eu_summaryreport_online.pdf. Accessed: 2020-02-06

Connor, D. S., Gutmann, M. P., Cunningham, A. R., Clement, K. K., and Leyk, S. (2020). How entrenched is the spatial structure of inequality in cities? Evidence from the integration of census and housing data for Denver from 1940 to 2016. *Annals of the American Association of Geographers, 110(4), 1022-1039.*

Demuzere, M., Bechtel, B., Middel, A., and Mills, G. (2019). Mapping Europe into local climate zones. *PloS one, 14(4), e0214474.*

Domingo, D., Van Vliet, J., and Hersperger, A. M. (2023). Long-term changes in 3D urban form in four Spanish cities. *Landscape and Urban Planning, 230, 104624.*

Dornbierer, J., Wika, S., Robison, C., Rouze, G., and Sohl, T. (2021). Prototyping a methodology for long-term (1680–2100) historical-to-future landscape modeling for the conterminous United States. *Land, 10(5), 536.*

Esch, T., Brzoska, E., Dech, S., Leutner, B., Palacios-Lopez, D., Metz-Marconcini, A., ... and Zeidler, J. (2022). World Settlement Footprint 3D-A first three-dimensional survey of the global building stock. *Remote sensing of environment, 270, 112877.*

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

European Union (2019). Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019. on open data and the re-use of public sector information, in OJ L 172, 26.6.2019, 56–83, ELI http://data.europa.eu/eli/dir/2019/1024/oj

Ewing, R., and Rong, F. (2008). The impact of urban form on US residential energy use. *Housing policy debate, 19(1), 1-30.*

Ferrara, A., Testa, P., and Zhou, L. (2021). New area-and population-based geographic crosswalks for US counties and congressional districts, 1790-2020. *Available at SSRN 4019521.*

Florczyk, A.J., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., Melchiorri, M., Pesaresi, M., Politis, P., Schiavina, M. and Sabo, F. (2019). GHSL data package 2019. *Luxembourg, EUR, 29788(10.2760), p.290498.*

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., ... and Zhou, Y. (2020). Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment, 236, 111510.*

Gröger, G., and Plümer, L. (2014). The interoperable building model of the European Union. In Geoinformation for Informed Decisions (pp. 1-17). Springer, Cham.

Haberl, H., Wiedenhofer, D., Schug, F., Frantz, D., Virág, D., Plutzar, C., ... and Hostert, P. (2021). High-resolution maps of material stocks in buildings and infrastructures in Austria and Germany. *Environmental science & technology, 55(5), 3368-3379.*

Hudson, P. (2018). Urban Characterisation; Expanding Applications for, and New Approaches to Building Attribute Data Capture. *The Historic Environment: Policy & Practice, 9(3-4), 306-327.*

IBM 2023. Online resource, available at https://www.ibm.com/docs/en/baw/19.x?topic=formats-atom-feed-format

Instituto Nacional de Geografía (2022). Planimetría historical maps Web Map Service. Online resource. http://www.ign.es/wms/minutas-cartograficas?request=GetCapabilities&service=WMS&

Iglesias, V., Braswell, A. E., Rossi, M. W., Joseph, M. B., McShane, C., Cattau, M., ... and Travis, W. R. (2021). Risky development: Increasing exposure to natural hazards in the United States. *Earth's future, 9(7), e2020EF001795.*

Klein Goldewijk, K., Beusen, A., Doelman, J., and Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth System Science Data, 9(2), 927-953.*

Leyk, S., and Uhl, J. H. (2018). HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Scientific data, 5(1), 1-14.*

Leyk, S., Ruther, M., Buttenfield, B. P., Nagle, N. N., and Stum, A. K. (2014). Modeling residential developed land in rural areas: A size-restricted approach using parcel data. *Applied Geography, 47, 33-45.*

Leyk, S., Uhl, J. H., Balk, D., and Jones, B. (2018). Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States. *Remote sensing of environment, 204, 898-917.*

Leyk, S., Uhl, J. H., Connor, D. S., Braswell, A. E., Mietkiewicz, N., Balch, J. K., and Gutmann, M. (2020). Two centuries of settlement and urban development in the United States. *Science Advances, 6(23), eaba2937.*

Li, M., Wang, Y., Rosier, J. F., Verburg, P. H., and van Vliet, J. (2022). Global maps of 3D built-up patterns for urban morphological analysis. *International Journal of Applied Earth Observation and Geoinformation, 114,* 103048.

Li, X., Zhou, Y., Hejazi, M., Wise, M., Vernon, C., Iyer, G., and Chen, W. (2021). Global urban growth between 1870 and 2100 from integrated high resolution mapped data and urban dynamic modeling. *Communications Earth & Environment, 2(1), 201.*

Liu, F., Wang, S., Xu, Y., Ying, Q., Yang, F., and Qin, Y. (2020). Accuracy assessment of Global Human Settlement Layer (GHSL) built-up products over China. *Plos one, 15*(5), e0233164.

Marconcini, M., Gorelick, N., Metz-Marconcini, A., and Esch, T. (2020, June). Accurately monitoring urbanization at global scale–the world settlement footprint. In *IOP Conference Series: Earth and Environmental Science (Vol. 509, No. 1, p. 012036). IOP Publishing.*

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., ... and Strano, E. (2020). Outlining where humans live, the World Settlement Footprint 2015. *Scientific Data, 7(1), 242.*

Mc Shane, C., Uhl, J. H., and Leyk, S. (2022). Gridded land use data for the conterminous United States 1940–2015. *Scientific Data, 9(1), 493.*

McDonald, R. I., Biswas, T., Sachar, C., Housman, I., Boucher, T. M., Balk, D., ... and Leyk, S. (2021). The tree cover and temperature disparity in US urbanized areas: Quantifying the association with income across 5,723 communities. *PloS one, 16(4), e0249715.*

Mietkiewicz, N., Balch, J. K., Schoennagel, T., Leyk, S., St. Denis, L. A., and Bradley, B. A. (2020). In the line of fire: consequences of human-ignited wildfires to homes in the US (1992–2015). *Fire, 3(3), 50.*

Millard-Ball, A. (2022). The width and value of residential streets. *Journal of the American Planning Association, 88(1), 30-43.*

Millhouser, P. (2019). Evaluating Landscape Connectivity and Habitat Fragmentation Effects on Elk in the Roaring Fork and Eagle Valleys. *Master's Thesis, Pennsylvania State University, University Park, Pennsylvania.*

Minghini, M., Cetl, V., Kotsev, A., Tomas, R., and Lutz, M. (2021). INSPIRE: The entry point to Europe's big geospatial data infrastructure. In *Handbook of Big Geospatial Data* (pp. 619-641). Cham: Springer International Publishing.

Miranda, A. S. (2022). The micro persistence of layouts and design: quasi-experimental evidence from the United States housing corporation. *Regional Science and Urban Economics, 95, 103755.*

Nolte, C. (2020). High-resolution land value maps reveal underestimation of conservation costs in the United States. *Proceedings of the National Academy of Sciences, 117(47), 29577-29583.*

Nolte, C., Boyle, K. J., Chaudhry, A. M., Clapp, C. M., Guignet, D., Hennighausen, H., ... and Uhl, J. H. (2021). Studying the impacts of environmental amenities and hazards with nationwide property data: best data practices for interpretable and reproducible analyses. *WVU College of Law Research Paper, (2021-013).*

Olazabal, Eduardo, Eduard Alvarez Palau, Rafael Barquín, María Olga Macías Muñoz, Jordi Martí Henneberg, Mateu Morillas-Torné, Pedro Pablo Ortúñez, and Jorge Solanas. "Catálogo digital de cartografía urbana contemporánea en España (1800-1950)." Biblio 3W. Revista Electrónica Bibliográfica de Geografía y Ciencias Sociales Universidad de Barcelona, 2019, vol. XXIV, núm. 1267 (2019). https://revistes.ub.edu/index.php/b3w/article/view/27944

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Pesaresi, M., Corbane, C., Ren, C., and Edward, N. (2021). Generalized Vertical Components of built-up areas from global Digital Elevation Models by multi-scale linear regression modelling. *Plos one, 16(2), e0244478.*

Remírez, P., Martínez de Marigorta, M. Á., Fragua, R., Izquierdo, M., LLinàs, J., Lorenzo, E., Más, R. M., GradualMap, SL, Rugoma SA, RevisaAtlas SA, Técnica Cartográfica Andaluza SA, Digraf SL, and Tracasa SA. (1998). Gran atlas de España Planeta (Segunda). Editorial Planeta, S.A.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics, 125(3), 1253-1296.*

Sapena, M., and Ruiz, L. Á. (2019). Analysis of land use/land cover spatio-temporal metrics and population dynamics for urban growth characterization. *Computers, environment and urban systems, 73*, 27-39.

Schiavina M., Melchiorri M., Pesaresi M., Politis P., Freire S., Maffenini L., Florio P., Ehrlich D., Goch K., Tommasi P., Kemper T., GHSL Data Package 2022, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-53071-8, doi:10.2760/19817, JRC 129516

Seto, K. C., Fragkias, M., Güneralp, B., and Reilly, M. K. (2011). A meta-analysis of global urban land expansion. *PloS one, 6(8), e23777.*

Silveira, L. E. D., Alves, D., Painho, M., Costa, A. C., and Alcântara, A. (2013). The evolution of population distribution on the Iberian Peninsula: a transnational approach (1877–2001). *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 46(3), 157-174.*

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., ... and Quinn, J. (2021). Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283.*

Tapp, A. F. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science, 37(3), 215-228.*

Uhl, J. H., and Leyk, S. (2017, July). A framework for radiometric sensitivity evaluation of medium resolution remote sensing time series data to built-up land cover change. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1908-1911). IEEE.*

Uhl, J. H., and Leyk, S. (2020). Towards a novel backdating strategy for creating built-up land time series data using contemporary spatial constraints. *Remote Sensing of Environment, 238, 111197.*

Uhl, J. H., and Leyk, S. (2022a). MTBF-33: A multi-temporal building footprint dataset for 33 counties in the United States (1900–2015). *Data in Brief, 43, 108369.*

Uhl, J. H., and Leyk, S. (2022b). A scale-sensitive framework for the spatially explicit accuracy assessment of binary built-up surface layers. *Remote Sensing of Environment, 279, 113117.*

Uhl, J. H., and Leyk, S. (2022c). Assessing the relationship between morphology and mapping accuracy of built-up areas derived from global human settlement data. *GIScience & remote sensing, 59(1), 1722-1748.*

Uhl, J. H., Connor, D. S., Leyk, S., and Braswell, A. E. (2021). A century of decoupling size and structure of urban spaces in the United States. *Communications earth & environment, 2(1), 20.*

Uhl, J. H., Hunter, L. M., Leyk, S., Connor, D. S., Nieves, J. J., Hester, C., ... and Gutmann, M. (2022). Place-level urban-rural indices for the United States from 1930 to 2018. *arXiv preprint arXiv:2202.13767*.

Uhl, J. H., Leyk, S., McShane, C. M., Braswell, A. E., Connor, D. S., and Balk, D. (2021). Fine-grained, spatiotemporal datasets measuring 200 years of land development in the United States. *Earth system science data, 13*(1), 119-153.

Uhl, J. H., Leyk, S., and Balk, D. (*forthcoming*). Exploring local accuracy patterns of high-resolution built-up surface indicators from the Global Human Settlement Layer in the United States.

Uhl, J. H., Royé, D., Burghardt, K., Aldrey Vázquez, J. A., Borobio Sanchiz, M., and Leyk, S. (2023a). HISDAC-ES: Historical Settlement Data Compilation for Spain (1900-2020). figshare. Dataset. https://doi.org/10.6084/m9.figshare.22009643

Uhl, J. H., Royé, D., Burghardt, K., Aldrey Vázquez, J. A., Borobio Sanchiz, M., and Leyk, S. (2023b). Visualizing long-term urbanization and land development in Spain (1900-2020). figshare. Media. https://doi.org/10.6084/m9.figshare.22064798

Uhl, J. H., Zoraghein, H., Leyk, S., Balk, D., Corbane, C., Syrris, V., and Florczyk, A. J. (2018). Exposing the urban continuum: Implications and cross-comparison from an interdisciplinary perspective. *International journal of digital earth 13(*1), 22-44.

Von Meyer, N., and Jones, B. (2013). Building National Parcel Data in the United States: One State at a Time. *International Association of Assessing Officers Fair and Equitable, 3-10.*

Wan, H., Yoon, J., Srikrishnan, V., Daniel, B., and Judi, D. (2022). Population downscaling using high-resolution, temporally-rich US property data. *Cartography and Geographic Information Science, 49(1), 18-31.*

Zoraghein, H., Leyk, S., Ruther, M., and Buttenfield, B. P. (2016). Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems, 58, 19-28.*

Zornoza Gallego, C. (2022a). Estructura urbana y movilidad en el área metropolitana de Valencia. Análisis de la evolución entre los años 1902-1942-1982-2017. València: Publicacions de la Universitat de València. Https://roderic.uv.ess/handle/10550/81983

Zornoza-Gallego, C. (2022b). Means of Transport and Population Distribution in Metropolitan Areas: An Evolutionary Analysis of the Valencia Metropolitan Area. *Land*, 11(5), 657.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Figures

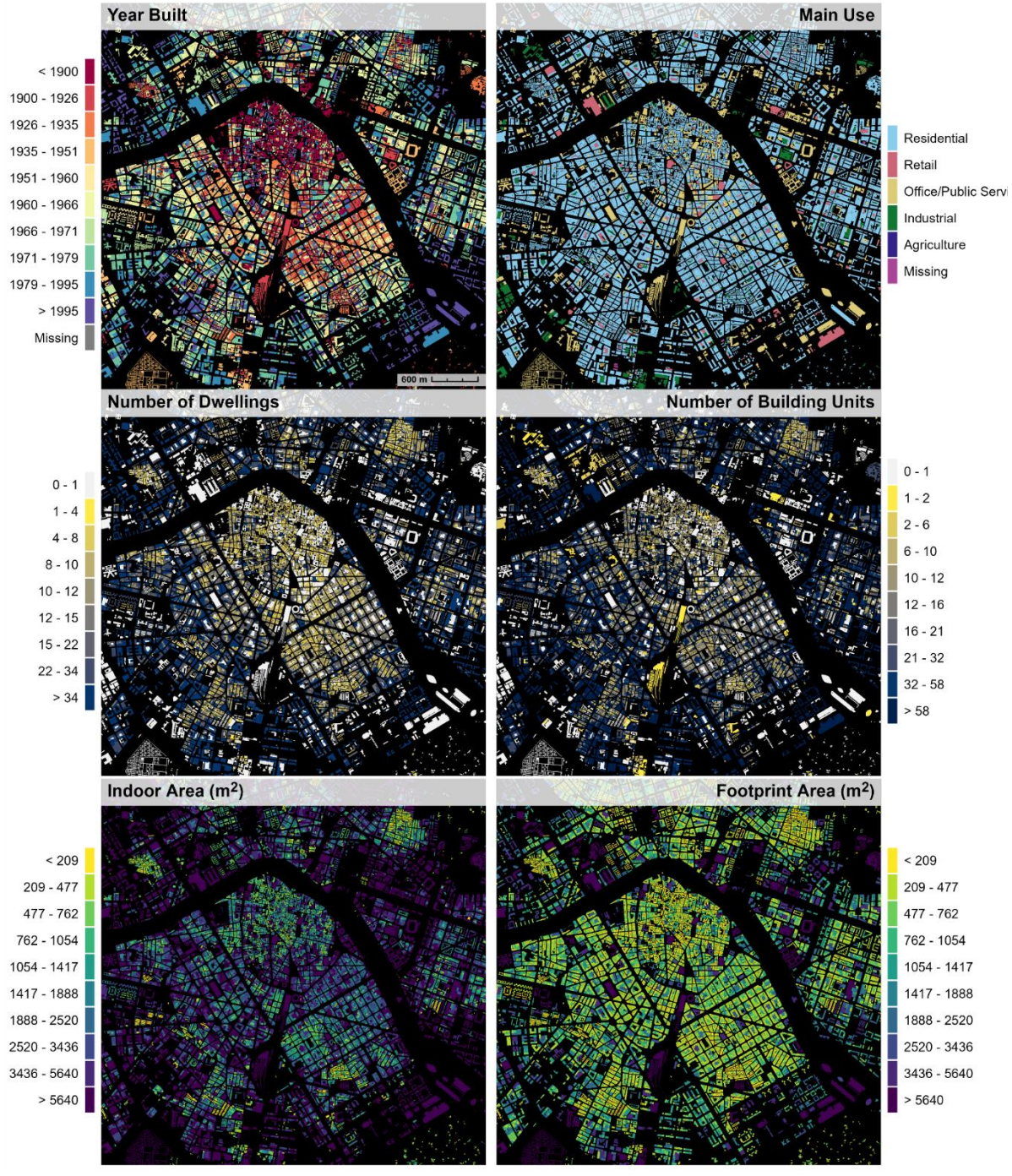


Figure 1: Input data: INSPIRE-conforming building footprint data provided by the Spanish authorities, including several attributes. (a) Year built, (b) building use, (c) number of dwellings, (d) number of building units, (e) building indoor area, and (f) building footprint area. Data shown for the city of Valencia.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

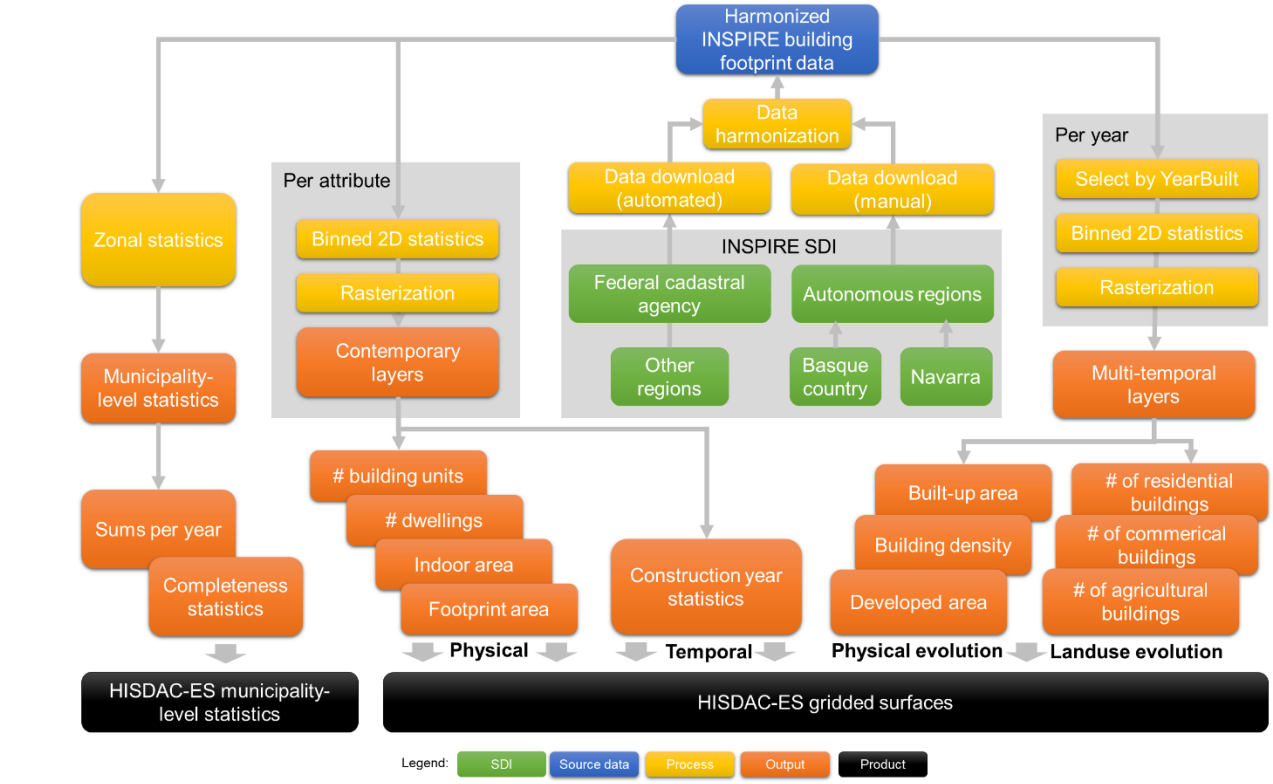
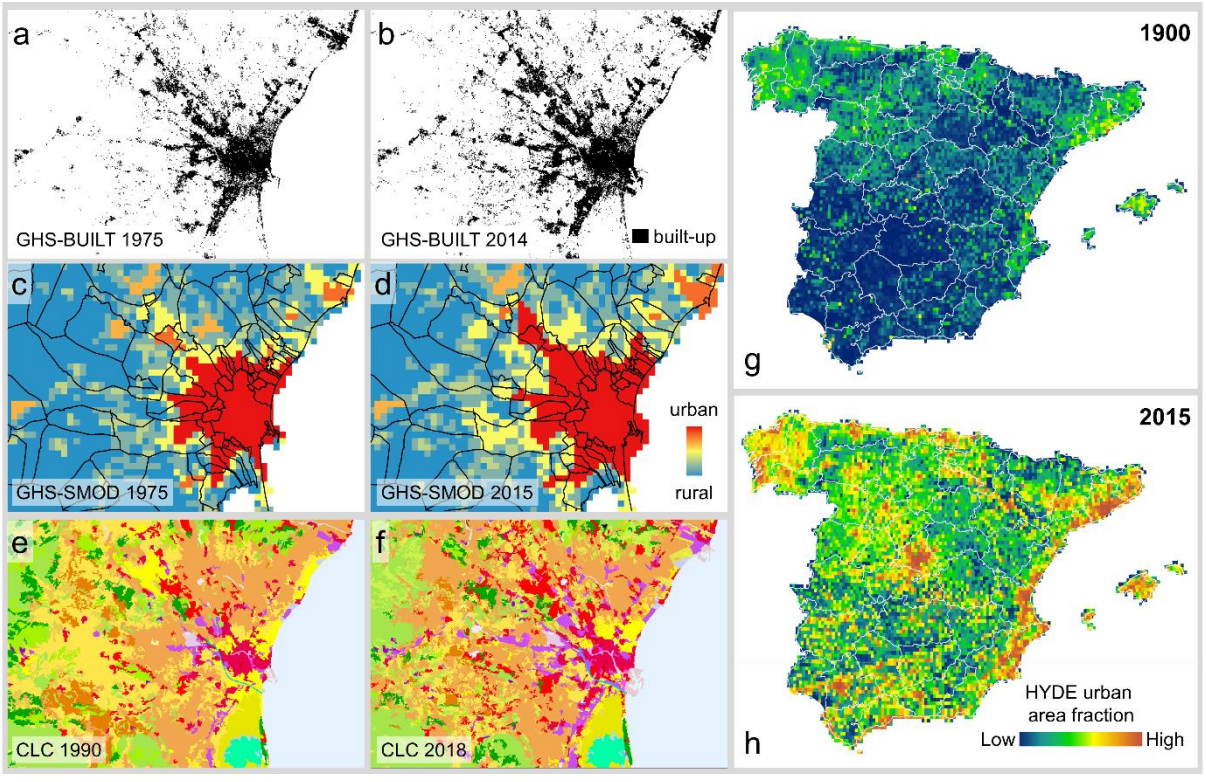


Figure 2: Data processing workflow to create the spatial data layers of the HISDAC-ES, measuring multiple dimensions of the built environment in Spain from 1900 to 2020, obtained from cadastral building footprint data available via the INSPIRE spatial data infrastructure (SDI).

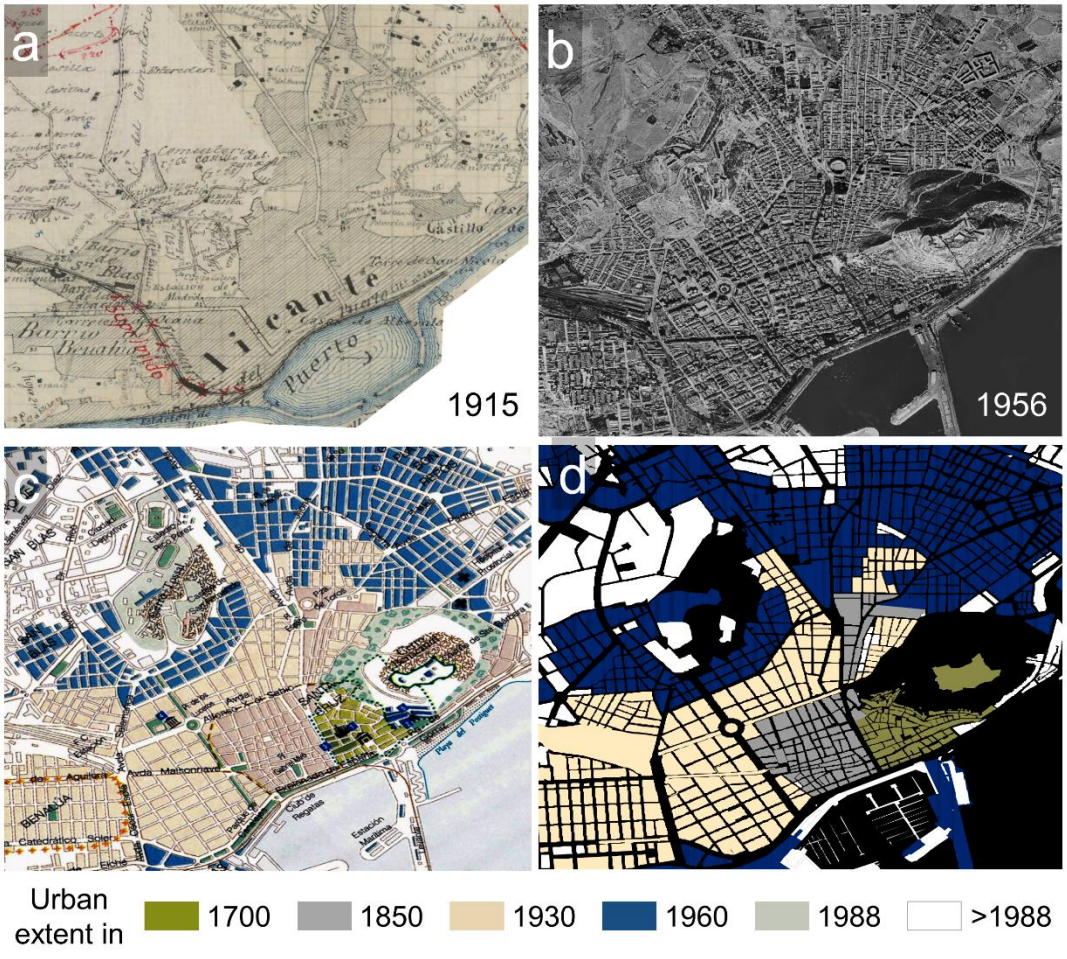
https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



735

Figure 3: Remote-sensing and model-based evaluation data: built-up areas from the GHS-BUILT R2018A in (a) 1975 and (b) 2014; GHS-SMOD rural-urban classes in (c) 1975 and (d) 2014; Corine Land Cover data in (e) 1990 and (f) 2018; Urban area fraction from the HYDE v3.2 dataset in (g) 1900 and (h) 2015. (a) - (f) show the city of Valencia, (g) and (h) show mainland Spain and the Balearic Islands.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



740

Figure 4: Evaluation data based on historical cartography and orthoimagery: (a) "Planimetría" historical topographic map from 1910 (Minutas catastrones at scale 1:50,000; Instituto Nacional de Geografía 2022), (b) historical aerial image from 1956 (Ortofotos AMS(B) 1956-1957), (c) Historical urban atlas illustrating different urban development phases (Remírez et al., 1988), and (d) urban areas for different time periods, manually digitized from c). All datasets shown for the city of Alicante.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

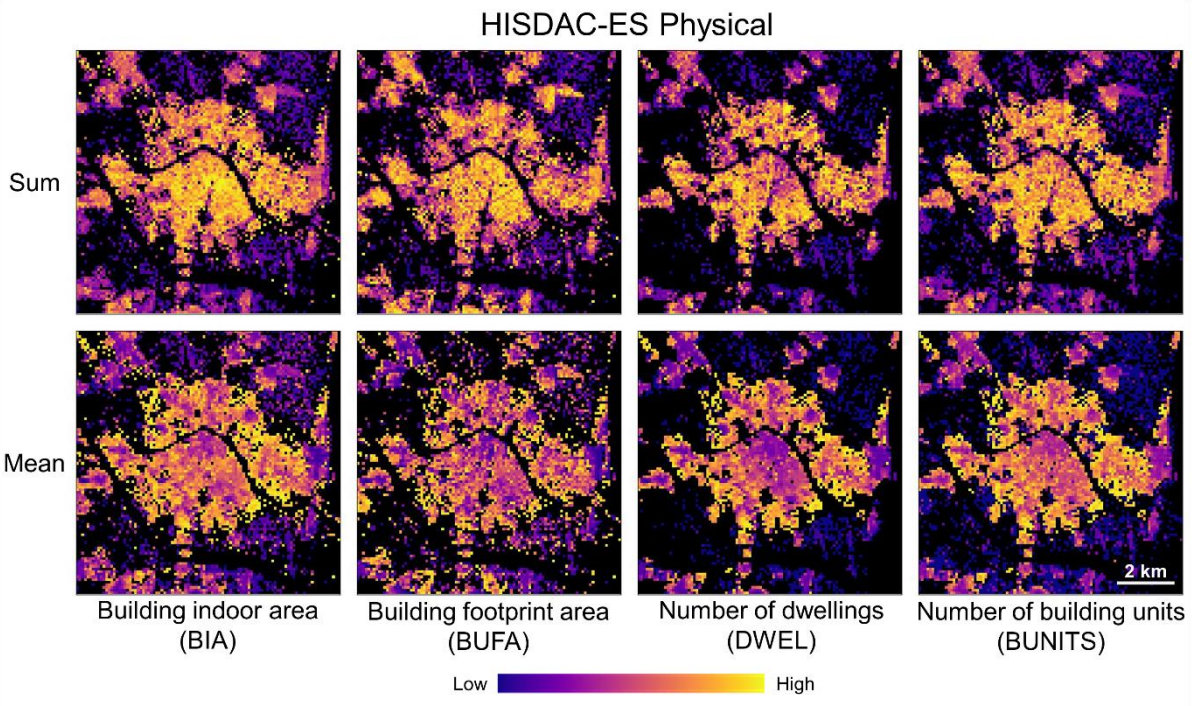
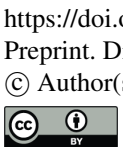
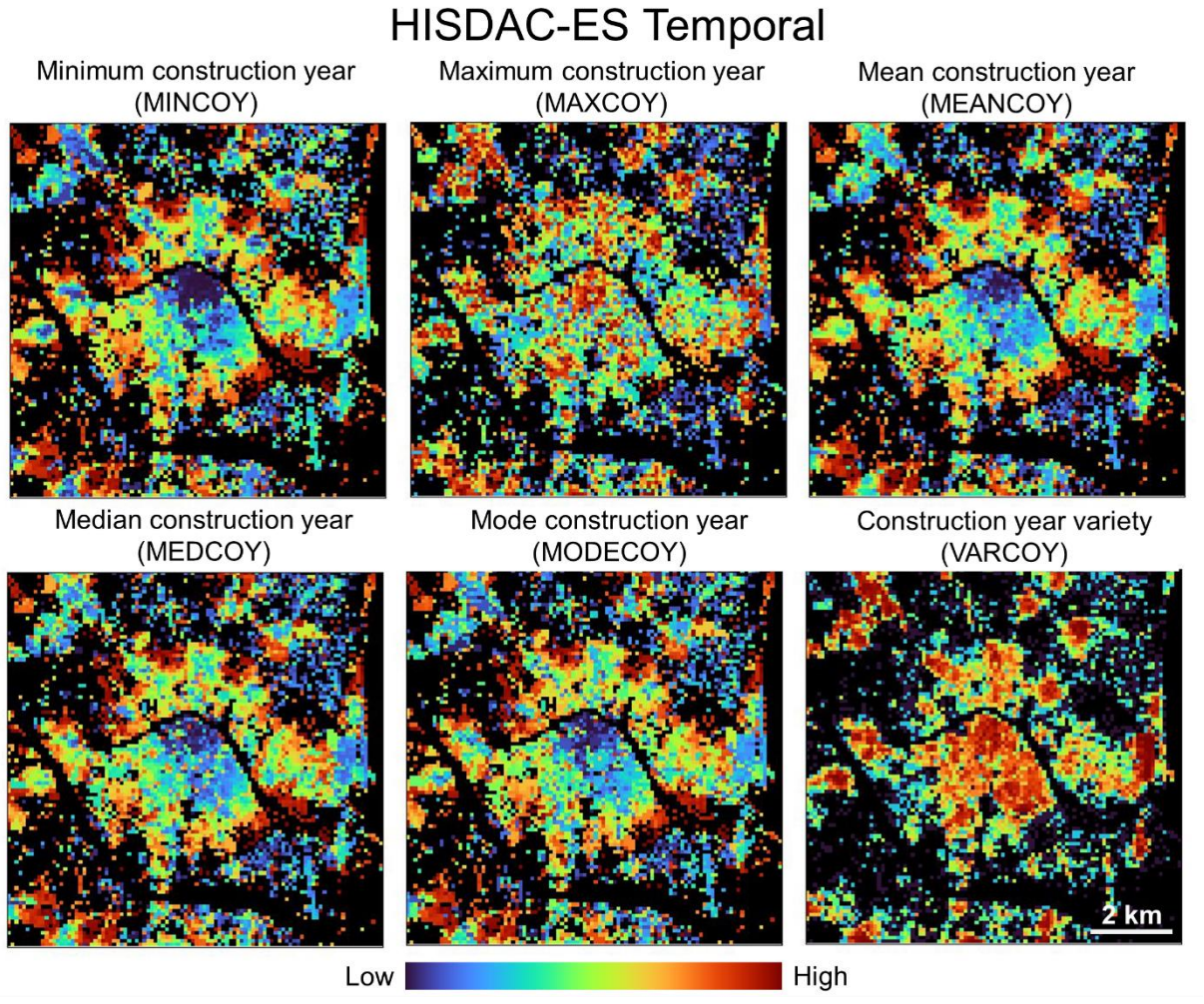


Figure 5: Examples of the created gridded surfaces of the HISDAC-ES, quantifying physical characteristics of the built environment: building indoor area (BIA), building footprint area (BUFA), number of dwellings (DWEL), and number of building units (BUNITS). HISDAC-ES contains the sum (top row) and the mean (bottom row) of these variables per grid cell. All datasets shown for the city of Valencia.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



Open Access Earth System Science Data Discussions



750

Figure 6: Examples of the created gridded surfaces of the HISDAC-ES, describing building-age related characteristics of the built environment: minimum construction year, maximum construction year, mean construction year, median construction year, mode (i.e., most frequent) construction year, and the variety (i.e., number of unique construction years) per grid cell. All datasets shown for the city of Valencia.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

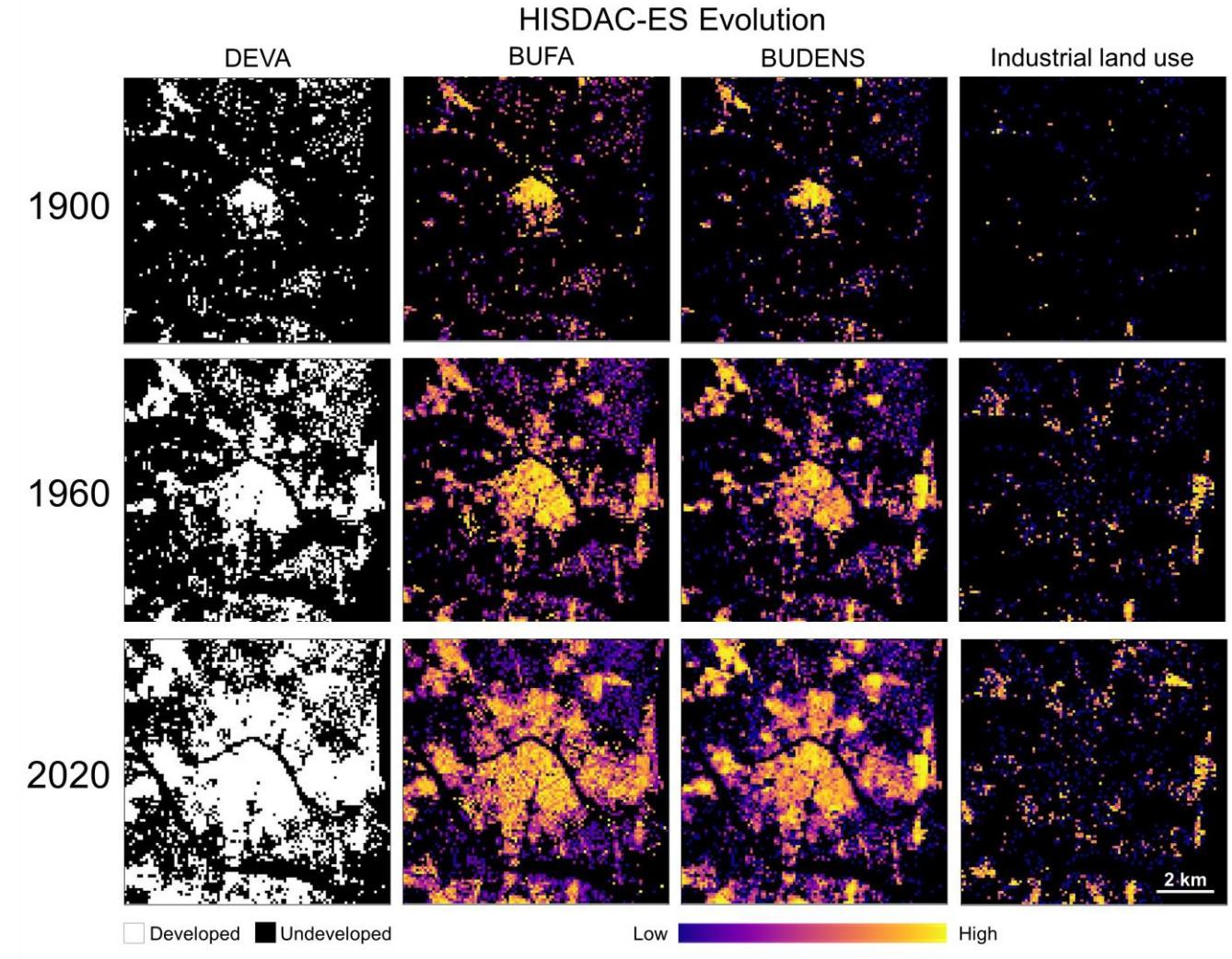
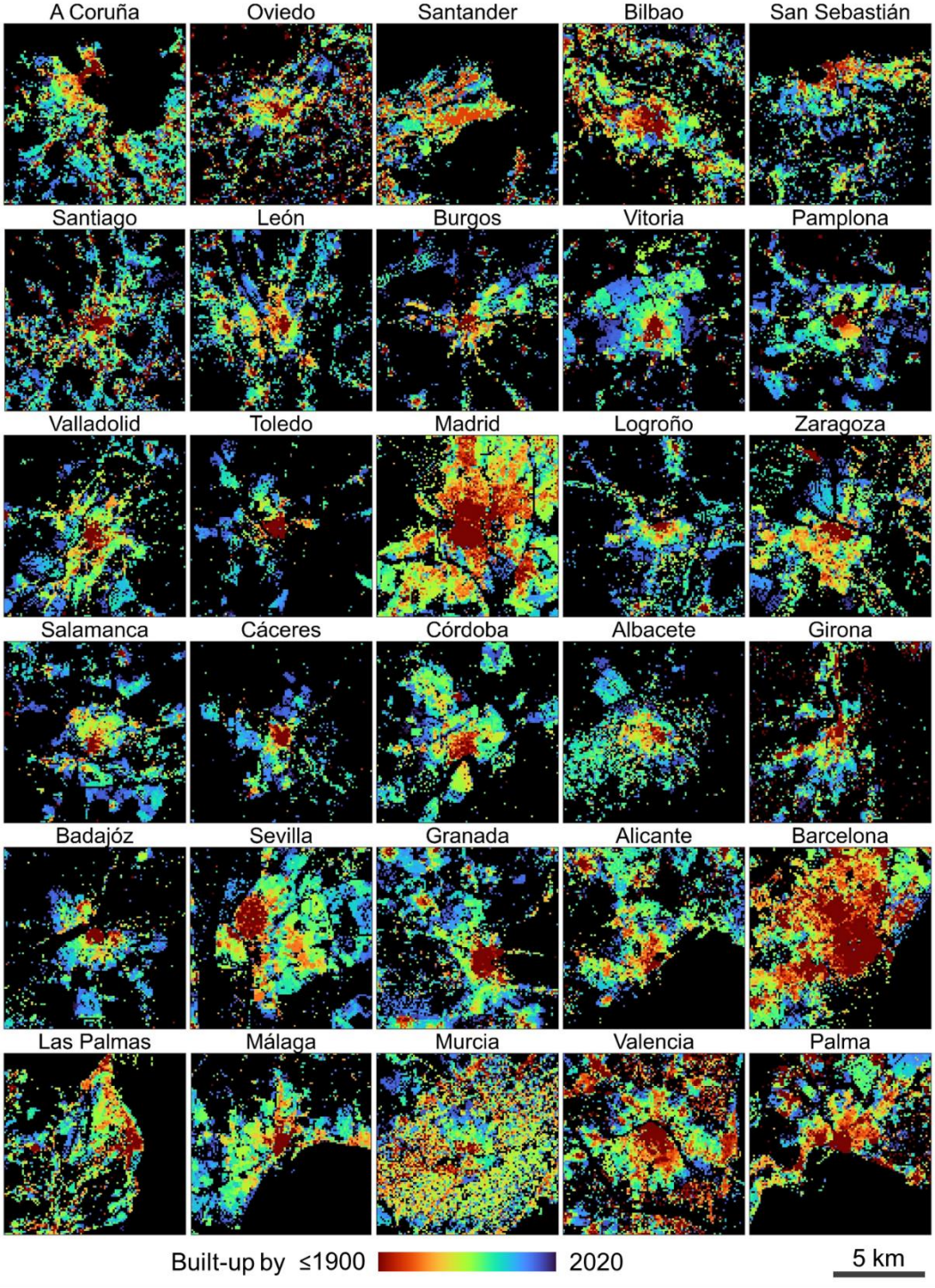


Figure 7: Examples of the created gridded surfaces of the HISDAC-ES, quantifying evolutionary characteristics of the built environment: Multi-temporal layers of developed area (DEVA), building footprint area (BUFA), and building density (BUDENS), as well as industrial land use in 1900, 1960, and 2020 (as one exemplary land use class). All datasets shown for the city of Valencia.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



760

Fig. 8. HISDAC-ES settlement age surfaces for 30 cities in Spain: Minimum construction year (MINCOY) shown for a selection of 30 cities, arranged in an approximate, quasi-geographic space (upper right = Northeast, lower left = Southwest)

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

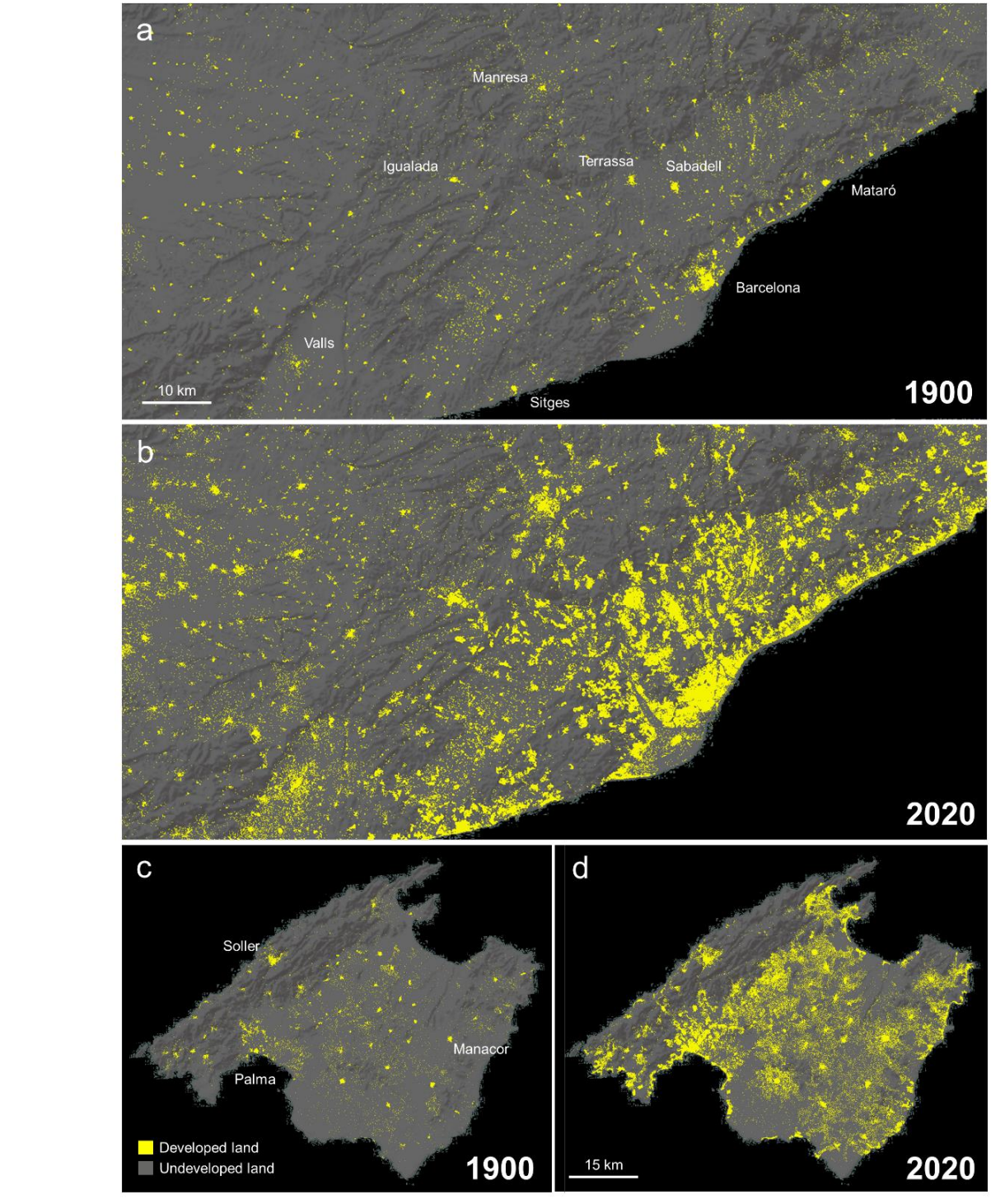


Figure 9: HISDAC-ES multitemporal surfaces of developed areas: DEVA layers shown for Greater Barcelona in (a) 1900, and (b) 2020, and for the Island of Mallorca in (c) 1900 and (d) in 2020. Basemap: Esri, USGS, NOAA.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

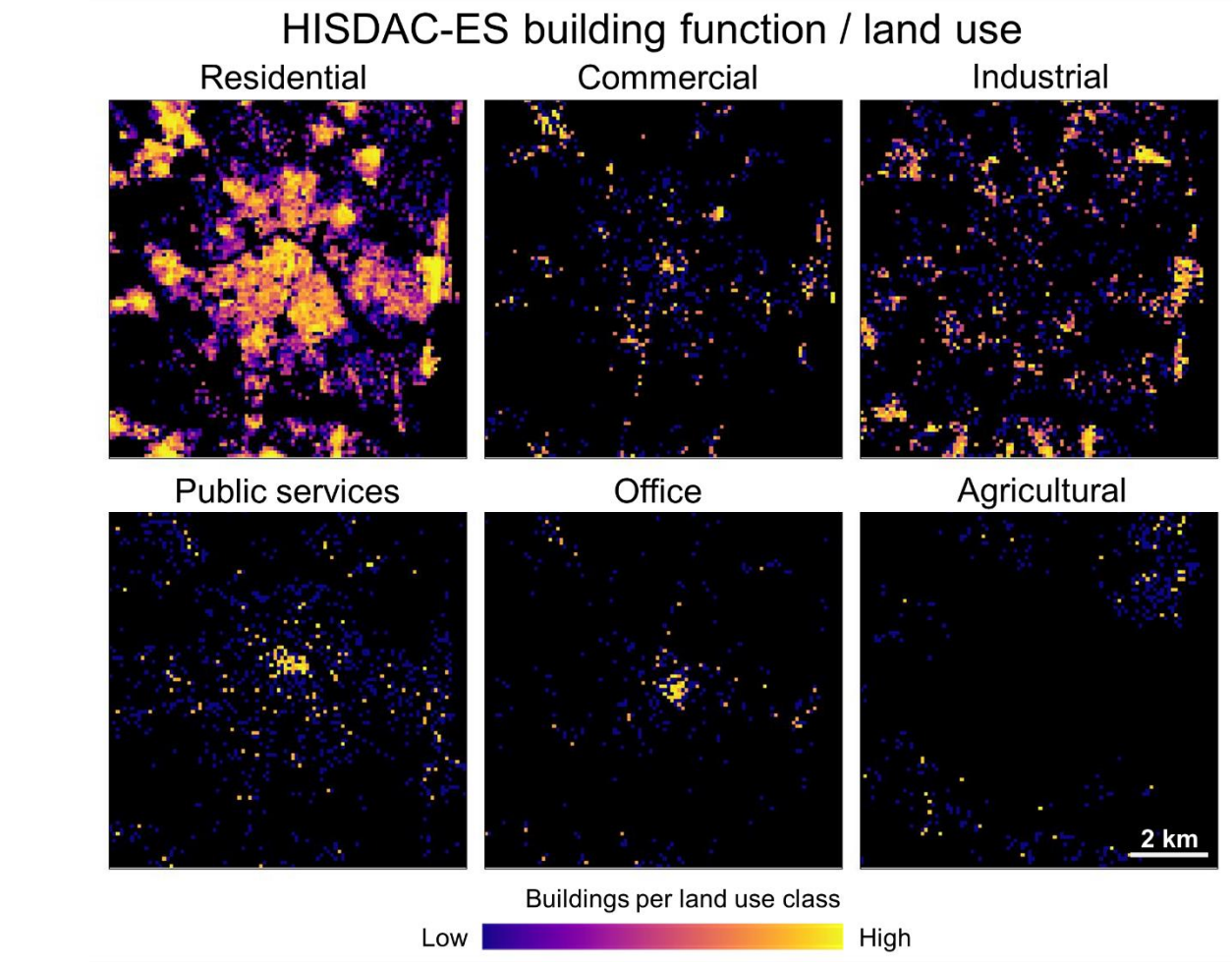


Fig. 10. Gridded surfaces describing the spatial distributions of different building function classes or building-related land use classes, shown for the city of Valencia in 2020.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

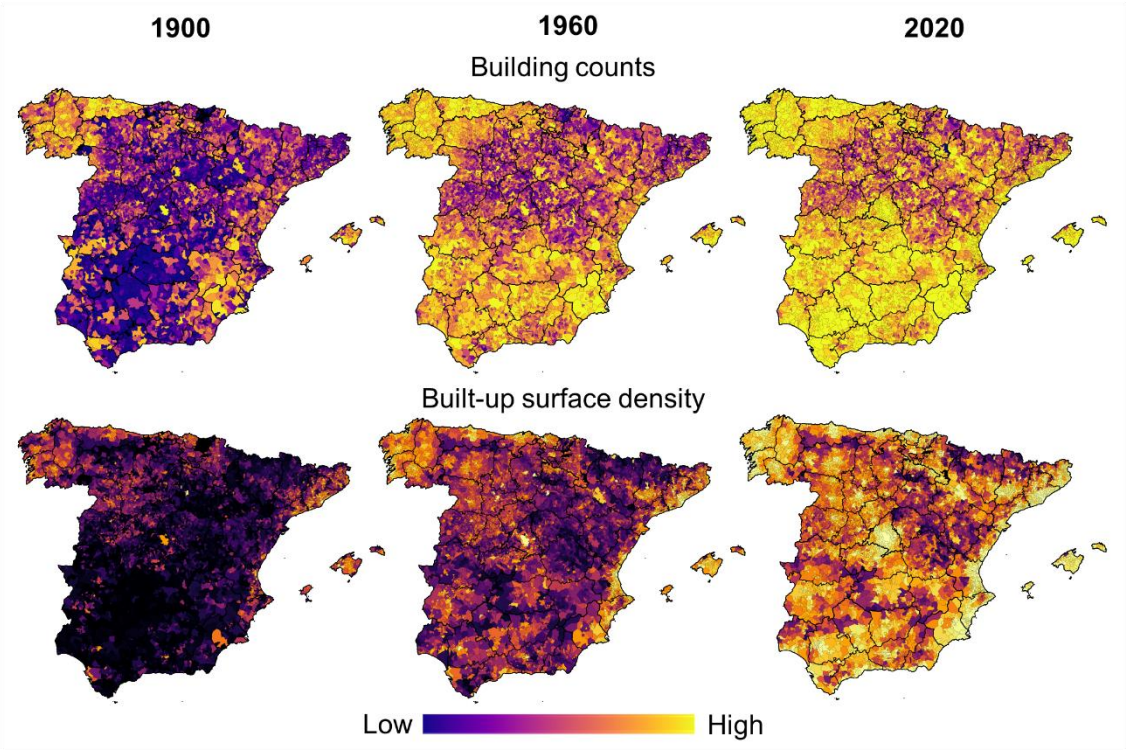


Figure 11: Selected multi-temporal municipality-level statistics: Building counts and built-up surface density per municipality in 1900, 1960, and 2020. See Appendix Fig. C1 for corresponding maps of the Canary Islands.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

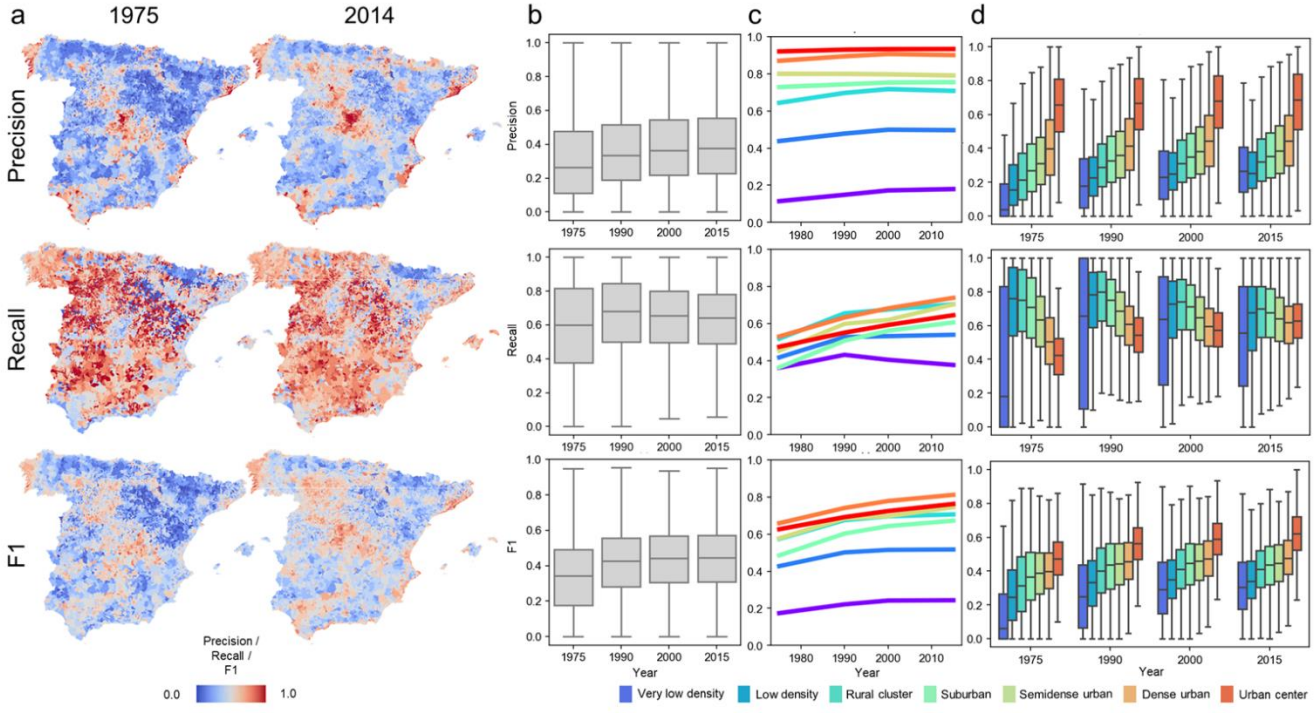
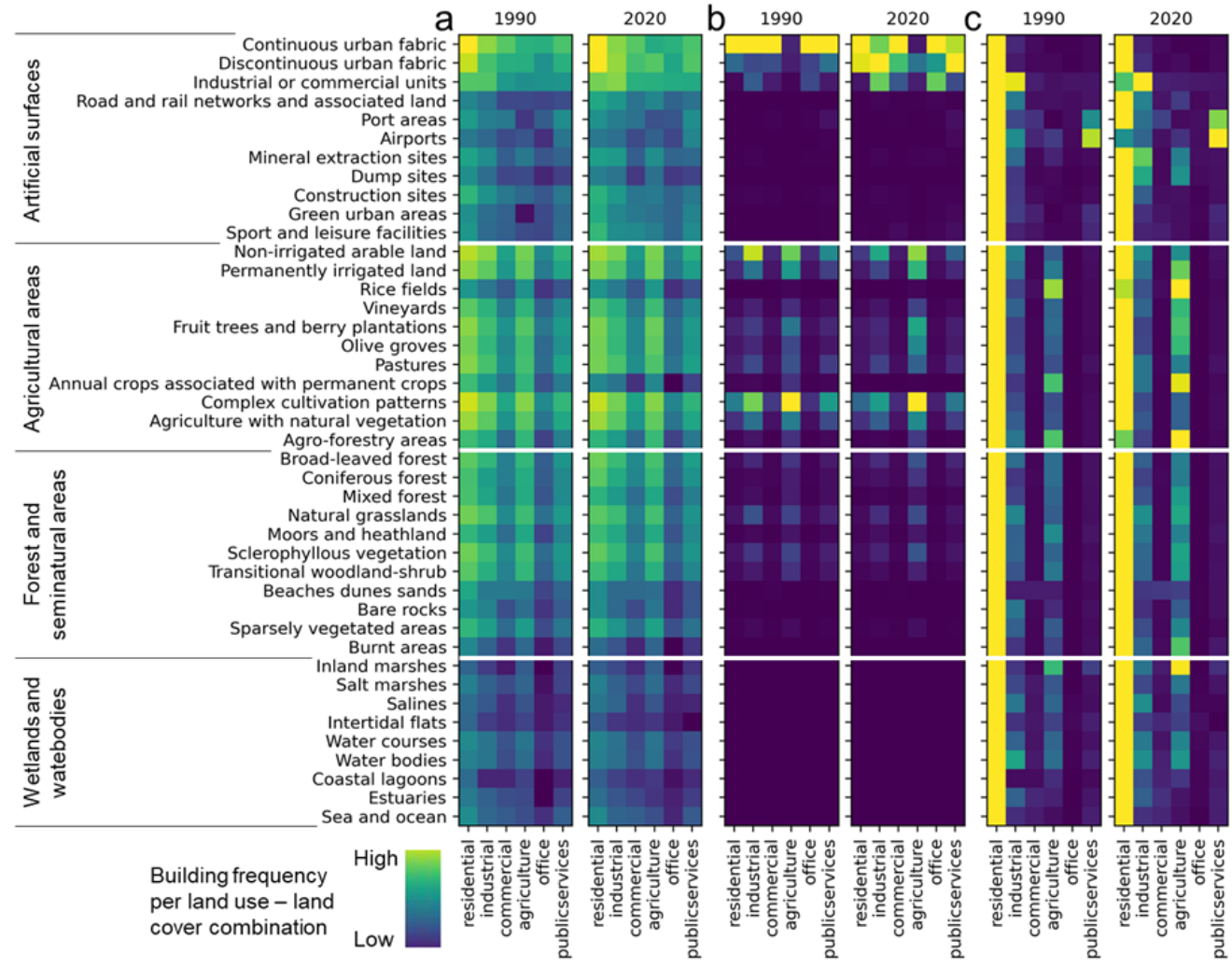


Figure 12: Evaluation of the HISDAC-ES developed area (DEVA) in comparison to built-up areas from the Global Human Settlement Layer per municipality, over time, and across the rural-urban continuum, by means of map comparison. (a) Maps of municipality-level precision, recall, and F1 score in 1975 and 2014, (b) overall temporal trends of municipality-level precision, recall and F1 score, (c) temporal trends of precision, recall and F-1 score calculated within strata of the GHS-SMOD rural-urban classes, and (d) distributions of municipality-level agreement metrics within strata of a GHS-SMOD based rural-urban index calculated per municipality. See Appendix Fig. D1 for corresponding maps of the Canary Islands.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



780

Figure 13: Comparison of the HISDAC-ES land use data (columns) to land cover classes from Corine Land Cover (rows) for the years 1990 and 2018. The heatmaps show the number of buildings (BUDENS) assigned to grid cells within each combination of HISDAC-ES land use and CLC land cover classes. These bidimensional frequency maps are shown in three variants (a) showing absolute building counts, (b) proportions of buildings per HISDAC-ES land use class, and (c) proportion of buildings per CLC land cover class. Note that the 2020 HISDAC-ES data was compared to the 2018 CLC data.

785

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

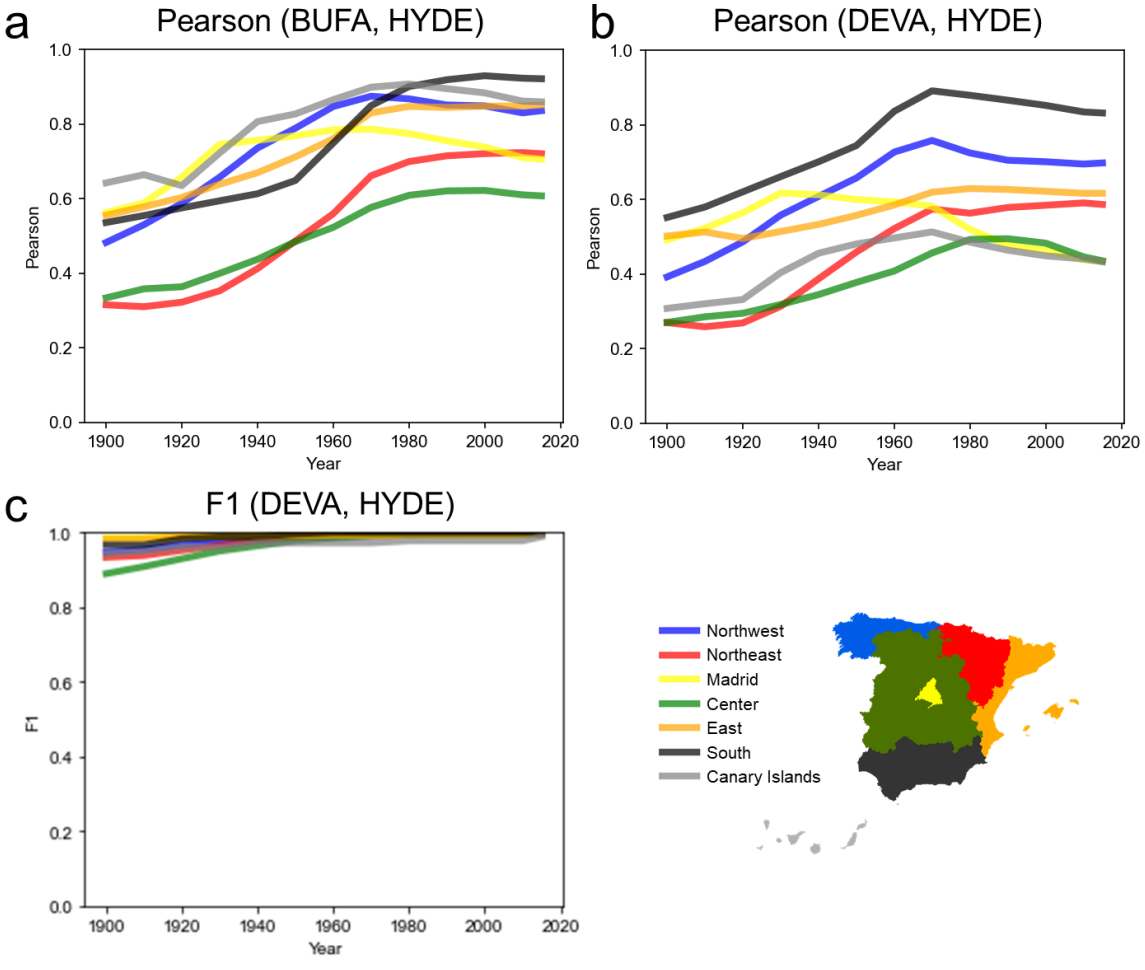


Figure 14: Regionally stratified agreement analysis of HISDAC-ES developed area (DEVA) and building footprint area (BUFA) variables against HYDE urban area estimates. (a) time series of Pearson's correlation coefficient of BUFA and HYDE, (b) time series of Pearson's correlation coefficient of DEVA and HYDE, and (c) time series of F1 scores based on developed vs. non-developed grid cells in HISDAC-ES and HYDE urban area fraction (5'x5' grid cells). Regional stratification was done based on the seven NUTS-1 regions (https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

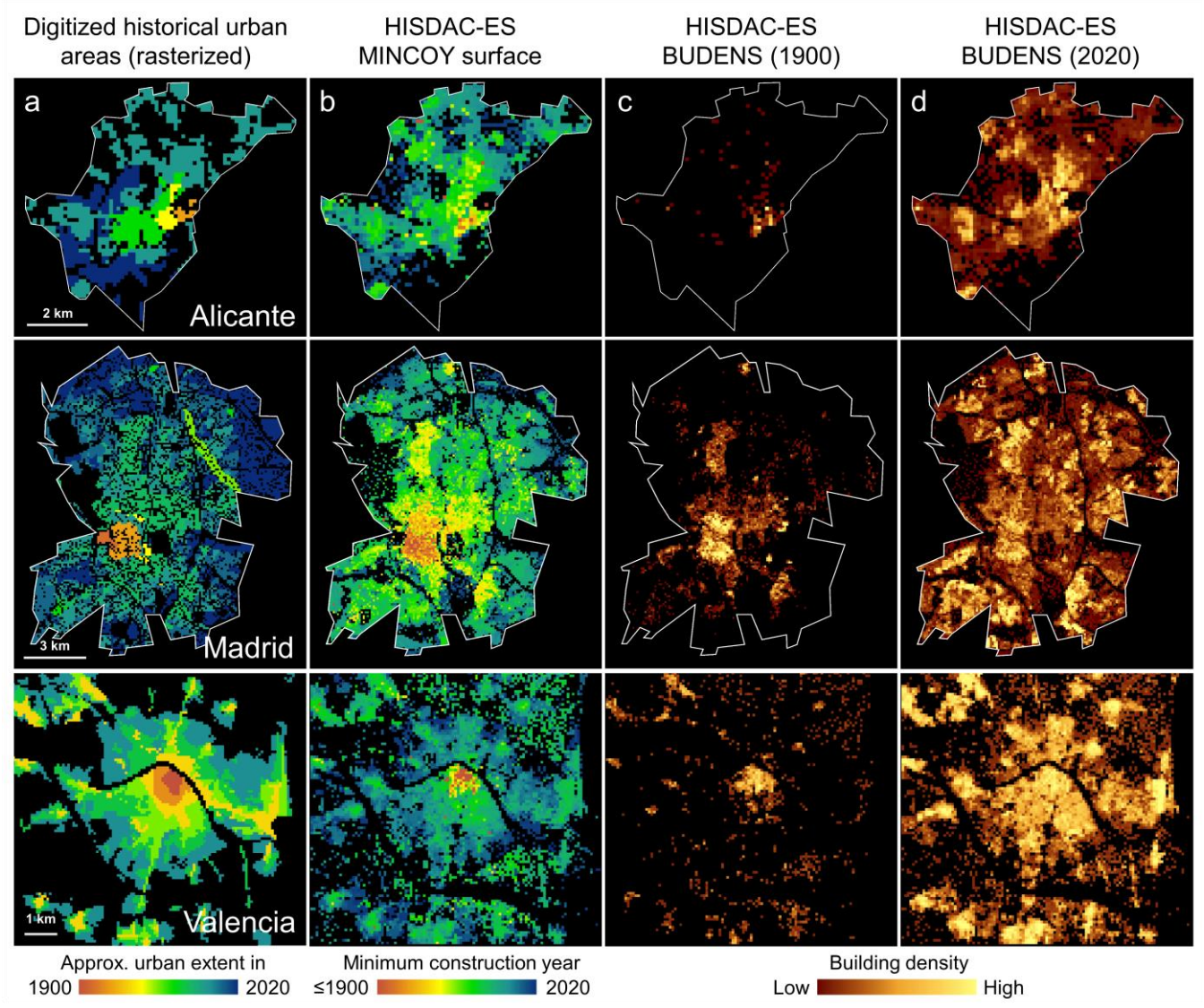


Figure 15: Data used for the quantitative comparison of the HISDAC-ES to historical urban extents, derived from historical maps for the cities of Alicante (top row), Madrid (middle row), and Valencia (bottom row): (a) Historical urban areas digitized from historical maps after rasterization, (b) the HISDAC-ES MINBUY layer, and the HISDAC-ES building density layers for (c) 1900 and (d) 2020. Data source of the historical urban extents for Valencia: courtesy C. Zornoza-Gallego, Figure in the lower left corner adapted from Zornoza-Gallego (2022a).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

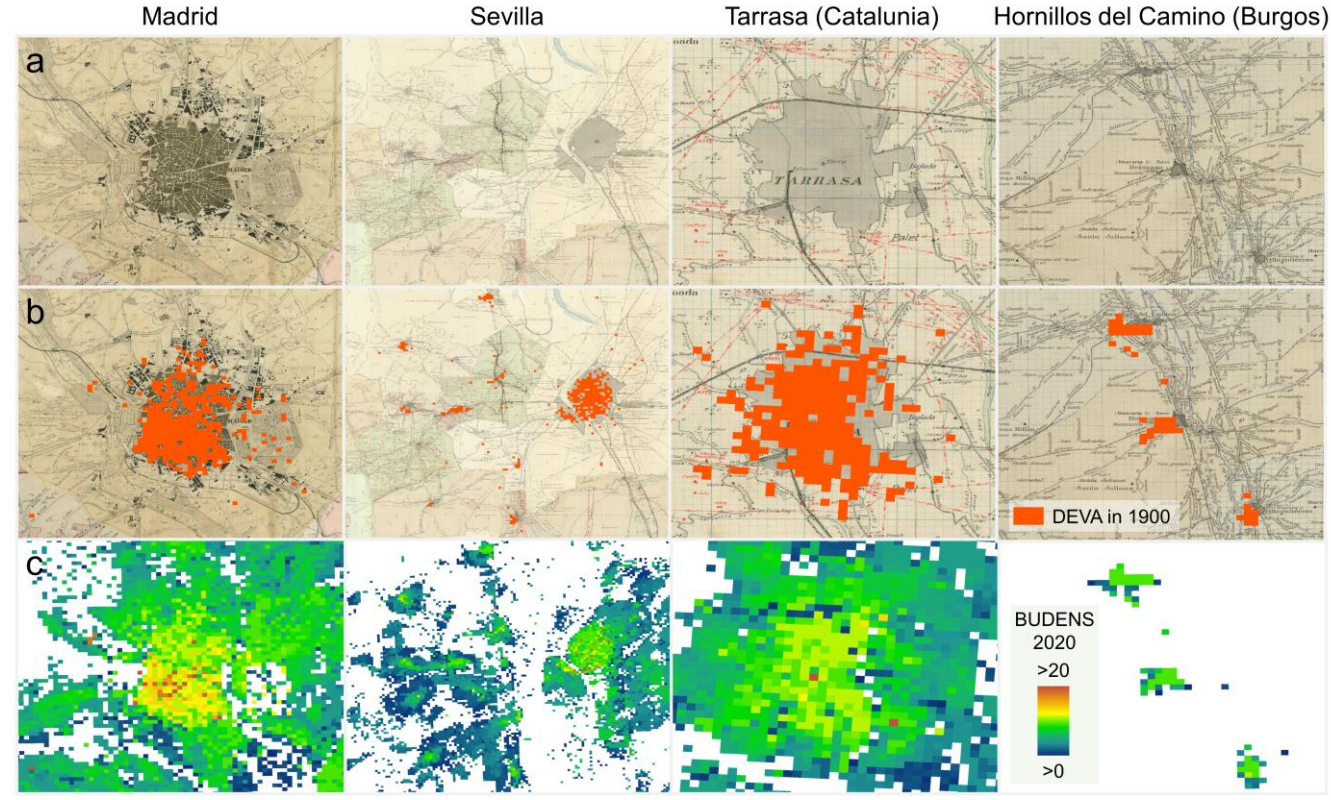
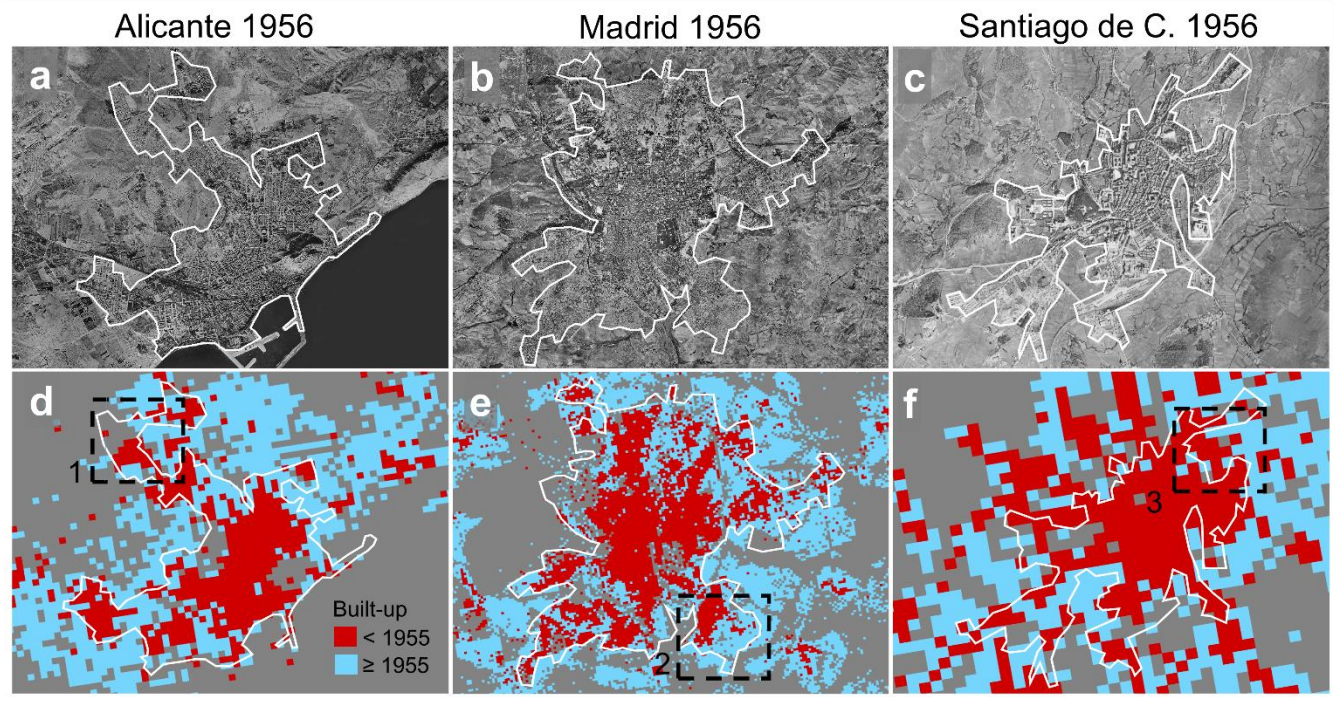


Figure 16: Visual comparison of HISDAC-ES developed areas (i.e., DEVA layer in 1900) with historical maps. (a) "Planimetría" maps (b) overlaid with DEVA in 1900, and (c) contemporary building density layer (BUDENS 2020) illustrating the change between 1900 and 2020. Each layer is shown for Madrid, Sevilla, Tarrasa (Catalunia), and Hornillos del Camino (Burgos). Historical map source: Instituto Nacional de Geografía 2022.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023



805 **Figure 17: Visual comparison of HISDAC-ES to aerial imagery from 1956. Aerial image overlaid with manually digitized urban extents for (a) Alicante, (b) Madrid, and (c) Santiago de Compostela; (d) – (f) overlaid with the HISDAC-ES built-up grid cells after and prior to 1955. Dashed black boxes are selected areas of discrepancies (i.e., turquoise areas within the digitized urban boundaries, or red areas without the urban boundary), which are enlarged and discussed in Appendix Fig. A6.**

810

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

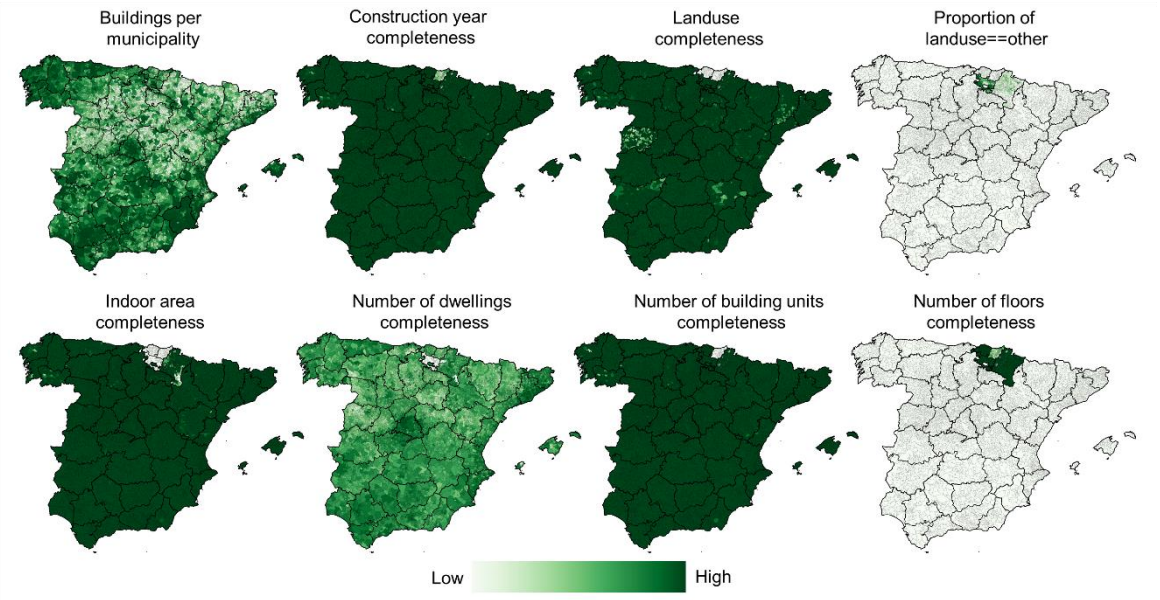


Figure 18: Attribute completeness at the municipality level. See Appendix Fig. G1 for corresponding maps of the Canary Islands.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

Tables

Table 1: Source data overview.

Province / Region	Data format	Online resource
Araba province (Basque Country)	WFS	https://geo.araba.eus/WFS_Katastroa?SERVICE=WFS&VERSION=1.1.0&REQUEST=GetCapabilities
Bizkaia province (Basque Country)	ATOM/GML	https://web.bizkaia.eus/es/inspirebizkaia
Gipuzkoa province (Basque Country)	ATOM/GML	https://b5m.gipuzkoa.eus/web5000/es/utilidades/inspire/edificios/
Navarra region	WFS	https://inspire.navarra.es/services/BU/wfs
Other regions (country level ATOM file)	ATOM	http://www.catastro.minhap.es/INSPIRE/buildings/ES.SDGC.bu.atom.xml
Other regions (regional level ATOM example)	ATOM	http://www.catastro.minhap.es/INSPIRE/buildings/03/ES.SDGC.bu.atom_03.xml
Other regions (municipality level example)	GML	http://www.catastro.minhap.es/INSPIRE/Buildings/03/03004-AIGUES/A.ES.SDGC.BU.03004.zip

Table 2: Overview of the gridded surfaces of the HISDAC-ES, measuring different components of the built environment, including the underlying variables and statistics.

Component	Variable	Statistics per grid cell	Measurement of	Surface name
Physical	Building Indoor area	Sum, mean across buildings per grid cell	Settlement intensity, built-up surface density, building density	BIA
	Building footprint area			BUFA
	Building units			BUNIT
	Dwellings			DWEL
Age	Construction year	Minimum, Maximum, Mean, Median, Mode, Variety per grid cell	Settlement age, building renewal, construction activity, building age heterogeneity	COY
Physical evolution	Building footprint area	Cumulative sum per 5-year time step ***T***, per grid cell	Land development, urban growth, settlement expansion, (sub)urbanization, urban sprawl, densification, infilling	$BUFA_T$
	Building density			$BUDENS_T$
	Developed area			$DEVA_T$
Land use evolution	# Residential buildings	Cumulative sum per 5-year time step ***T***, per grid cell	Land use dynamics, land development stratified by building function	LU_RES_T
	# Commercial buildings			LU_COM_T
	# Industrial buildings			LU_IND_T
	# Agricultural buildings			LU_AGR_T
	# Public buildings			LU_PUB_T
	# Office buildings			LU_OFF_T

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

Table 3: Overview of the comparative evaluation efforts for HISDAC-ES.

HISDAC-ES variable	Evaluation data product (+URL)	Evaluated time period	Evaluated area	Section
DEVA	GHS-BUILT, GHS-SMOD (https://ghsl.jrc.ec.europa.eu/)	1975-2014	Spain	4.1
Land use	Corine Land Cover (https://land.copernicus.eu/pan-european/corine-land-cover)	1990-2018	Spain	4.2
BUFA, DEVA	HYDE v3.2 (https://doi.org/10.17026/dans-25g-gez)	1900-2015	Spain	4.3
MINCOY	Historical urban extents (Zornoza-Gallego 2022a, Remírez et al., 1988)	1900-2020	Alicante, Madrid, Valencia	4.4
DEVA (qualitative)	Historical maps (Minutas catastrales, 1:50,000) (http://www.ign.es/wms/minutas-cartograficas?request=GetCapabilities&service=WMS)	1910	Several cities / villages	4.4
DEVA (qualitative)	Historical orthophotos 1956-1957 (https://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=FPNOA)	1955	Alicante, Madrid, Santiago de Compostela	4.5

Table 4: Map comparison results of digitized historical urban areas digitized from historical maps and HISDAC-ES settlement age surface (MINCOY), for both increments (i.e., newly developed grid cells between two points in time), and cumulatively (i.e., for the total developed land at a given point in time. Note that there is a slight temporal gap in the earliest Madrid epoch, where MINCOY from 1900 is compared to urban extents from 1892.

		Incremental				Cumulative				
		Cell-based			Building-weighted		Cell-based			Building-weighted
Study area	Time period	Precision	Recall	F1	Precision	Year	Precision	Recall	F1	Precision
Alicante urban atlas	1900-1928	0.32	0.42	0.36	0.44	≤1928	0.32	0.42	0.36	0.44
	1928-1970	0.47	0.52	0.49	0.56	≤1970	0.64	0.73	0.68	0.73
	1970-1990	0.26	0.20	0.23	0.27	≤1990	0.66	0.70	0.68	0.73
Madrid urban atlas	≤1900	0.44	0.40	0.42	0.53	≤1900	0.44	0.40	0.42	0.53
	1900-1932	0.40	0.54	0.46	0.44	≤1932	0.44	0.56	0.50	0.51
	1932-1962	0.25	0.40	0.31	0.33	≤1962	0.50	0.71	0.59	0.53
	1962-1970	0.35	0.23	0.28	0.39	≤1970	0.59	0.68	0.64	0.61
	1970-2020	0.31	0.22	0.26	0.33	≤2020	0.65	0.67	0.66	0.66
Valencia urban extents	≤1902	0.34	0.46	0.39	0.59	≤1902	0.34	0.46	0.39	0.59
	1902-1944	0.26	0.54	0.35	0.38	≤1944	0.44	0.74	0.55	0.65
	1944-1980	0.32	0.51	0.4	0.50	≤1980	0.53	0.87	0.66	0.74
	1980-2011	0.46	0.25	0.33	0.52	≤2011	0.68	0.80	0.74	0.88

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

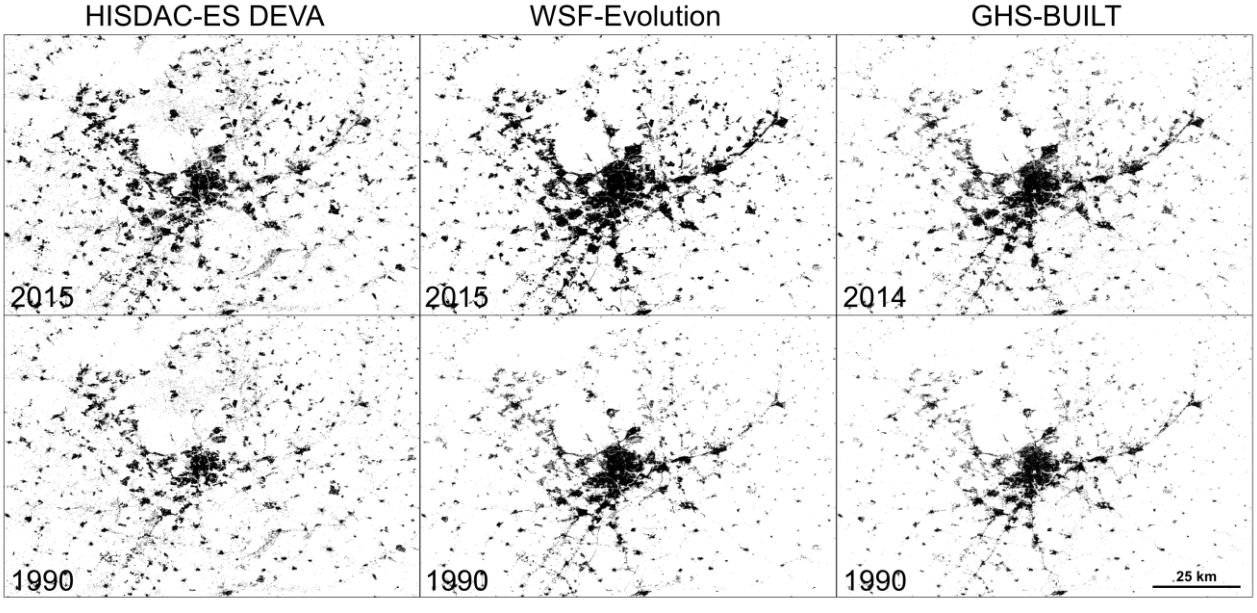
Open Access Earth System Science Data Discussions

Table 5: HISDAC-ES dataset overview.

Component	Variable	Surface name	Data format	File name (pattern)
Physical	Building Indoor area	BIA	GeoTIFF	hisdac_es_phys_bia_<*mean/sum*>_v1_100.tif
	Building footprint area	BUFA	GeoTIFF	hisdac_es_phys_bufa_<*mean/sum*>_v1_100.tif
	Building units	BUNIT	GeoTIFF	hisdac_es_phys_bunits_<*mean/sum*>_v1_100.tif
	Dwellings	DWEL	GeoTIFF	hisdac_es_phys_dwel_<*mean/sum*>_v1_100.tif
Age	Construction year	COY	GeoTIFF	hisdac_es_temp_<*statistic*>coy_v1_100.tif
Physical evolution	Building footprint area	$BUFA_T$	GeoTIFF	hisdac_es_evol_bufa _v1_100_<*year*>.tif
	Building density	$BUDENS_T$	GeoTIFF	hisdac_es_evol_budens _v1_100_<*year*>.tif
	Developed area	$DEVA_T$	GeoTIFF	hisdac_es_evol_deva _v1_100_<*year*>.tif
Land use evolution	# Residential buildings	LU_RES_T	GeoTIFF	hisdac_es_landuse_residential_v1_100_<*year*>.tif
	# Commercial buildings	LU_COM_T	GeoTIFF	hisdac_es_landuse_commercial_v1_100_<*year*>.tif
	# Industrial buildings	LU_IND_T	GeoTIFF	hisdac_es_landuse_industrial_v1_100_<*year*>.tif
	# Agricultural buildings	LU_AGR_T	GeoTIFF	hisdac_es_landuse_agriculture_v1_100_<*year*>.tif
	# Public buildings	LU_PUB_T	GeoTIFF	hisdac_es_landuse_publicservices_v1_100_<*year*>.tif
	# Office buildings	LU_OFF_T	GeoTIFF	hisdac_es_landuse_office_v1_100_<*year*>.tif
Municipality-level aggregates	various	Municipality-level variables	.csv	hisdac_es_municipality_stats_multitemporal_v1.csv
	% complete	Municipality-level attribute completeness	GeoPackage, .csv	hisdac_es_municipality_stats_completeness_v1.csv/gpkg

Appendices

Appendix A



Appendix Figure A1: Visual comparison of HISDAC-ES DEVA, WSF-Evolution, and GHS-BUILT in 1990 and 2015.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

Appendix B.

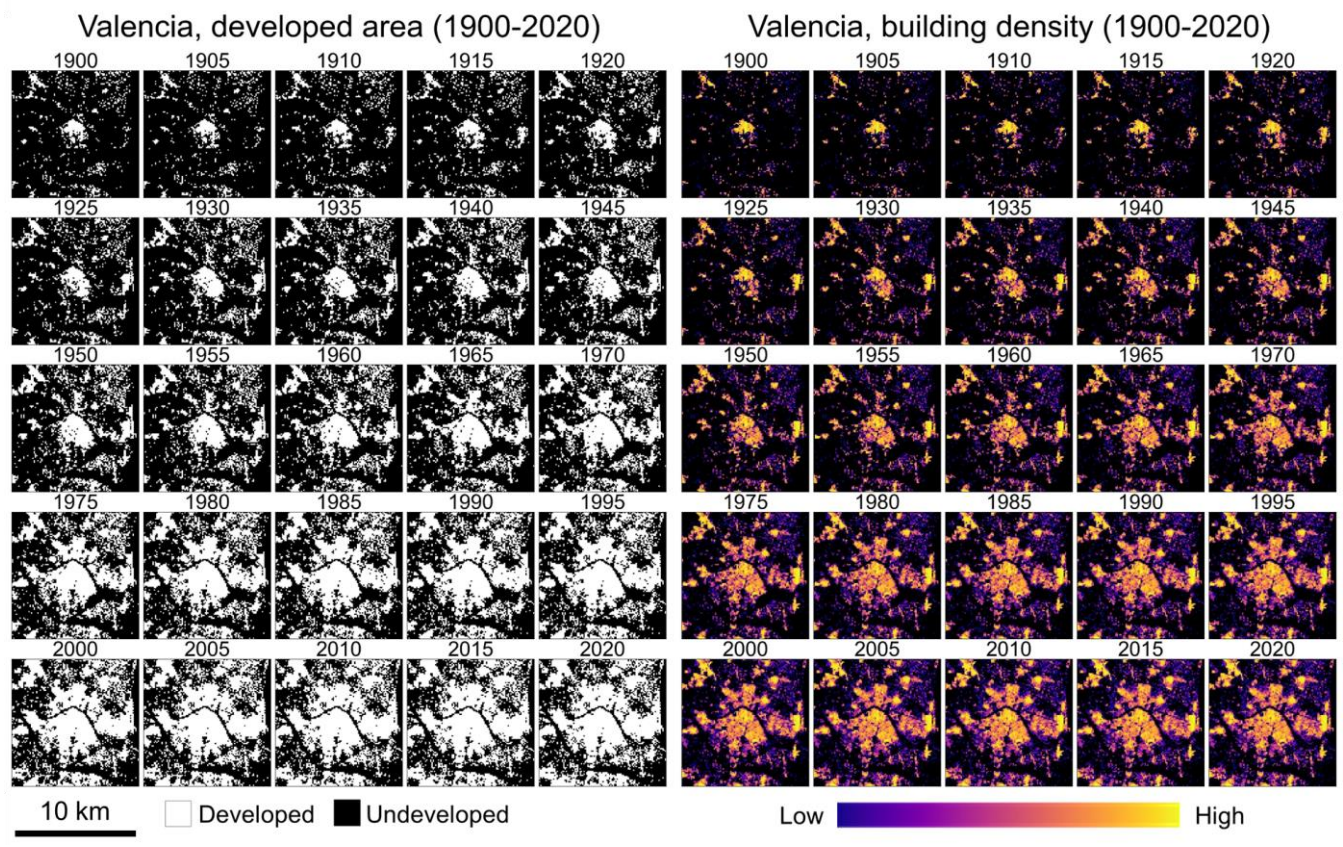
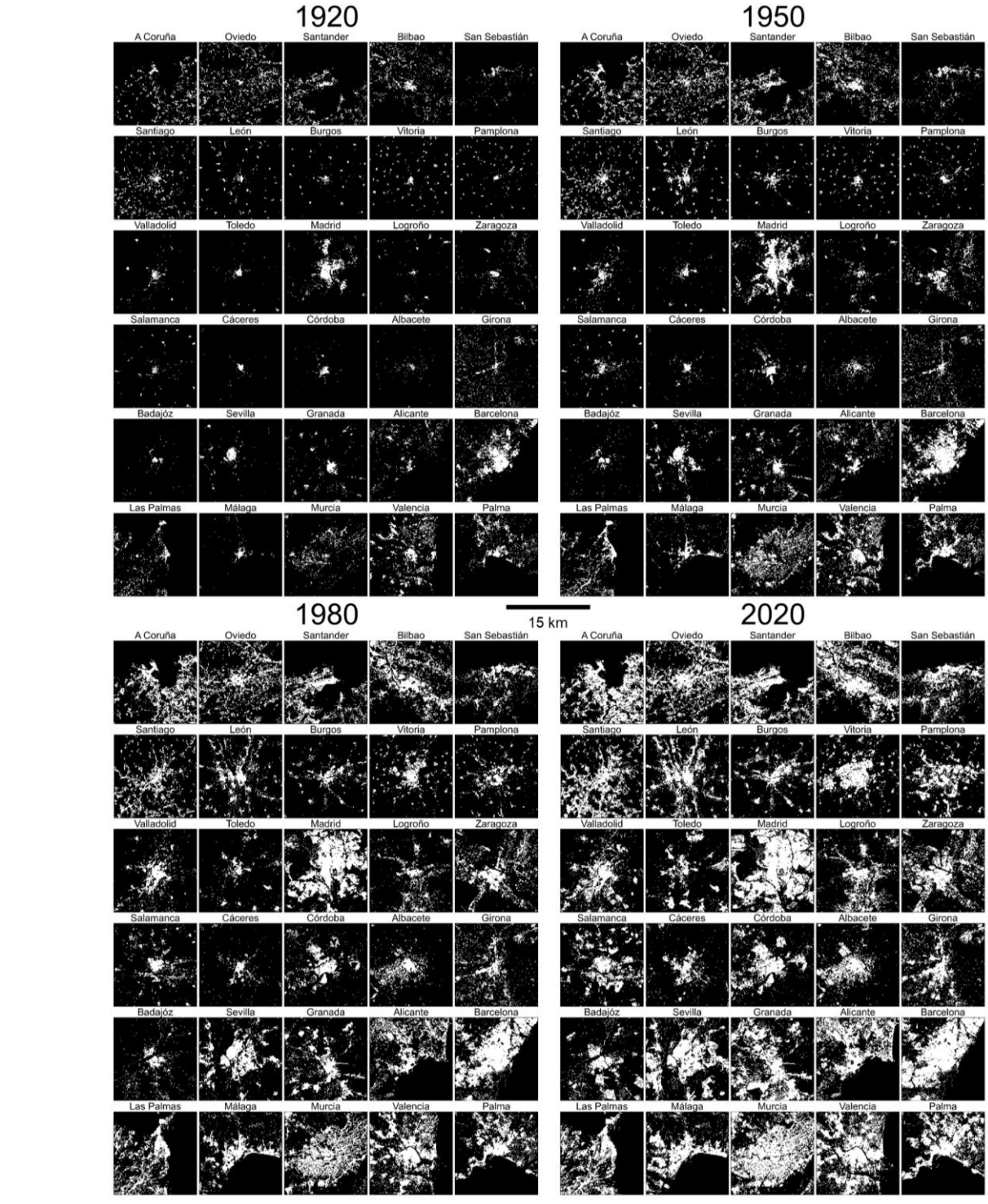


Figure B1: Complete time series of the DEVA (developed area, left) and BUDENS (building density, right) raster time series in 5-year intervals from 1900 to 2020. Data shown for the city of Valencia.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

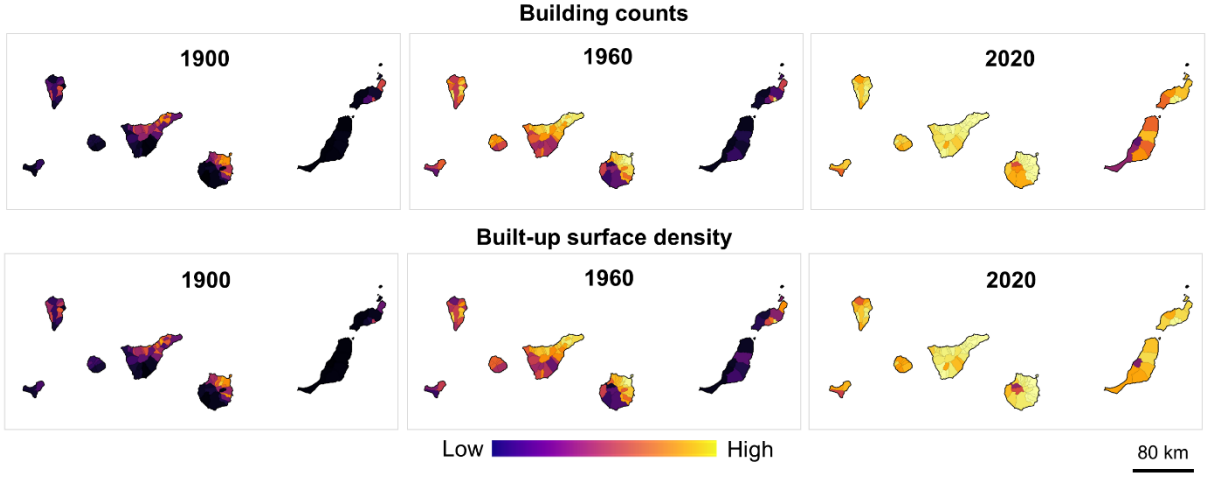


Appendix Figure B2: HISDAC-ES developed area layers (DEVA) for 1920, 1950, 1980, and 2020 for a selection of 30 cities in Spain. Cities are arranged in an approximate, quasi-geographic space (upper right = Northeast, lower left = Southwest).

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

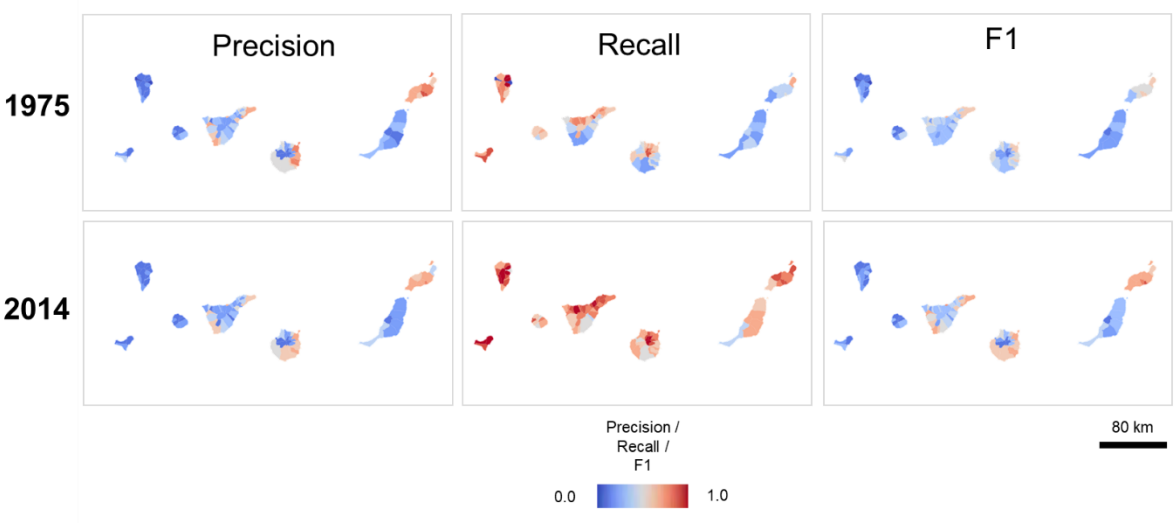
Open Access Earth System Science Data Discussions

Appendix C.



Appendix Figure C1: Exemplary municipality-level aggregates for the Canary Islands in 1900, 1960, and 2020.

Appendix D.

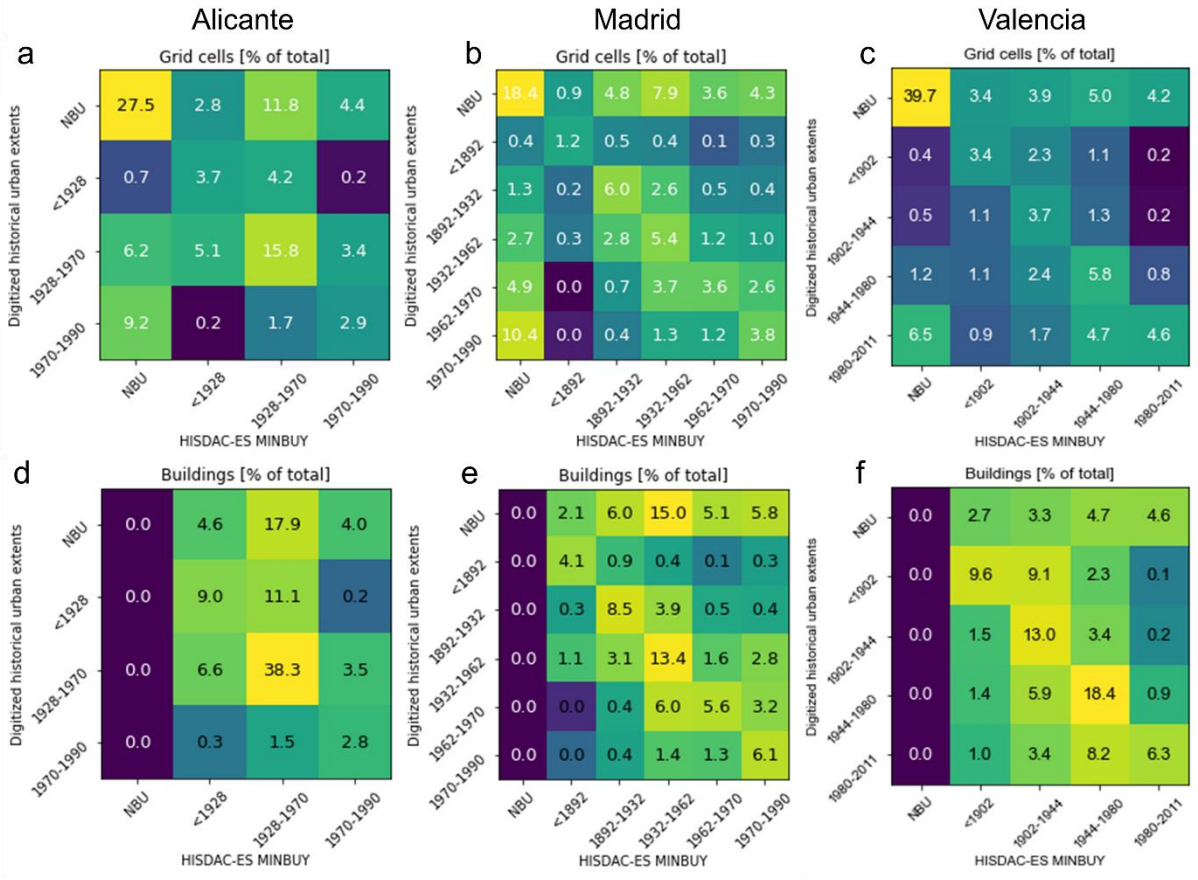


Appendix Figure D1: Municipality-level agreement of HISDAC-ES DEVA and GHS-BUILT R2018A for the Canary Islands.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Open Access Earth System Science Data Discussions

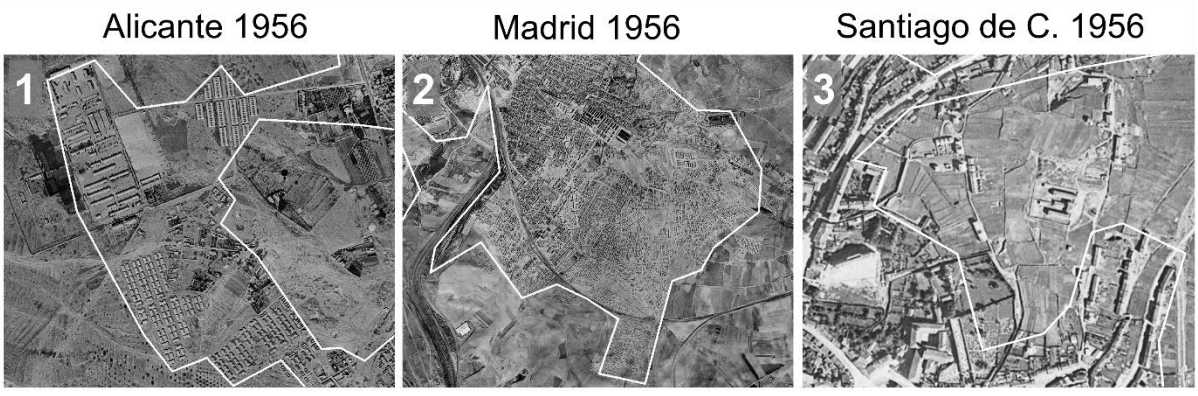
Appendix E.



Appendix Figure E1: Confusion matrices underlying Table 4 from map comparison of the MINBUY and urban extents digitized from historical maps for (a) Alicante, (b) Madrid, and (c) Valencia. While panels (a) – (c) cross-tabulate the grid cell counts in each category, Panels (d) – (f) are based on the number of buildings reported in HISDAC-ES within each category.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

Appendix F.



Appendix Figure F1: Examples of discrepancies between HISDAC-ES and the 1955 aerial imagery. Shown are three regions highlighted in Fig. 17.

1. Alicante: Alférez Rojas Navarrete Barracks. Even though this military facility existed in 1955, it is not well covered in the cadastral data underlying the HISDAC-ES. Despite covering a large area, it is only represented by one grid cell in the HISDAC-ES layers for 1955. It is likely that military facilities were not mapped the same way like regular residential neighborhoods, possibly for security reasons.

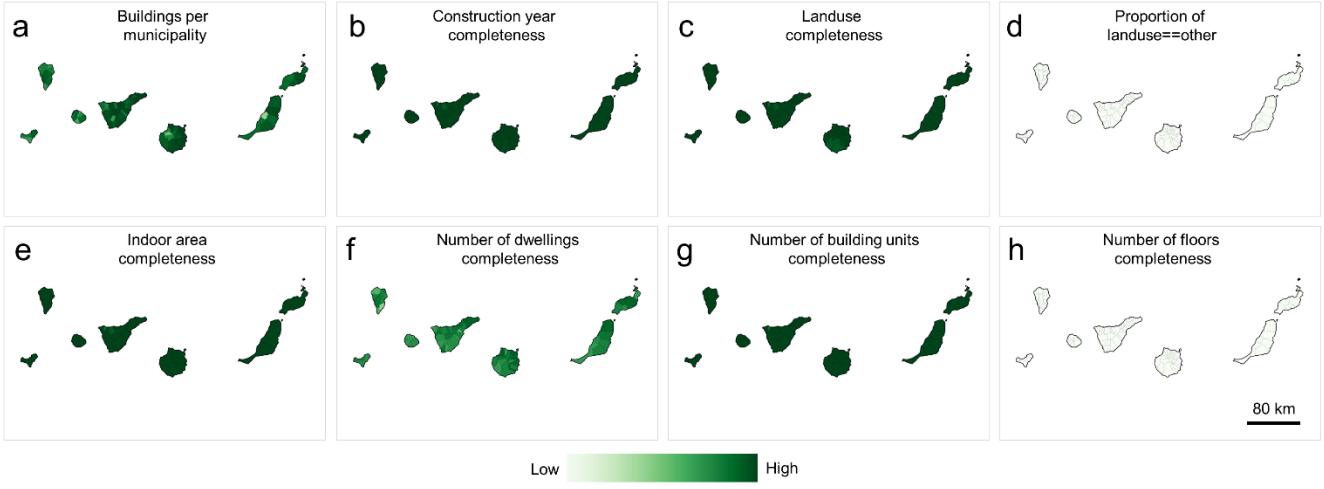
2. Madrid: Housing colonies Entrevías (Vallecas II). A large social housing colony that was completely renewed at a later point in time. This is an extreme example of urban renewal, which cannot be measured by the HISDAC-ES data.

3. Santiago de Compostela: This discrepancy between the HISDAC-ES data and the manually drawn urban boundary is due to a definitional problem. In the aerial image of the Monte da Almáciga region, the large building complex is a college, surrounded by scattered settlements, possibly of agricultural usage. In this case, we excluded these areas from the "urban extent" due to the low settlement density, even though these areas should be considered as "urban" due to their functional importance. This definitional mismatch causes this discrepancy.

https://doi.org/10.5194/essd-2023-53
Preprint. Discussion started: 2 March 2023

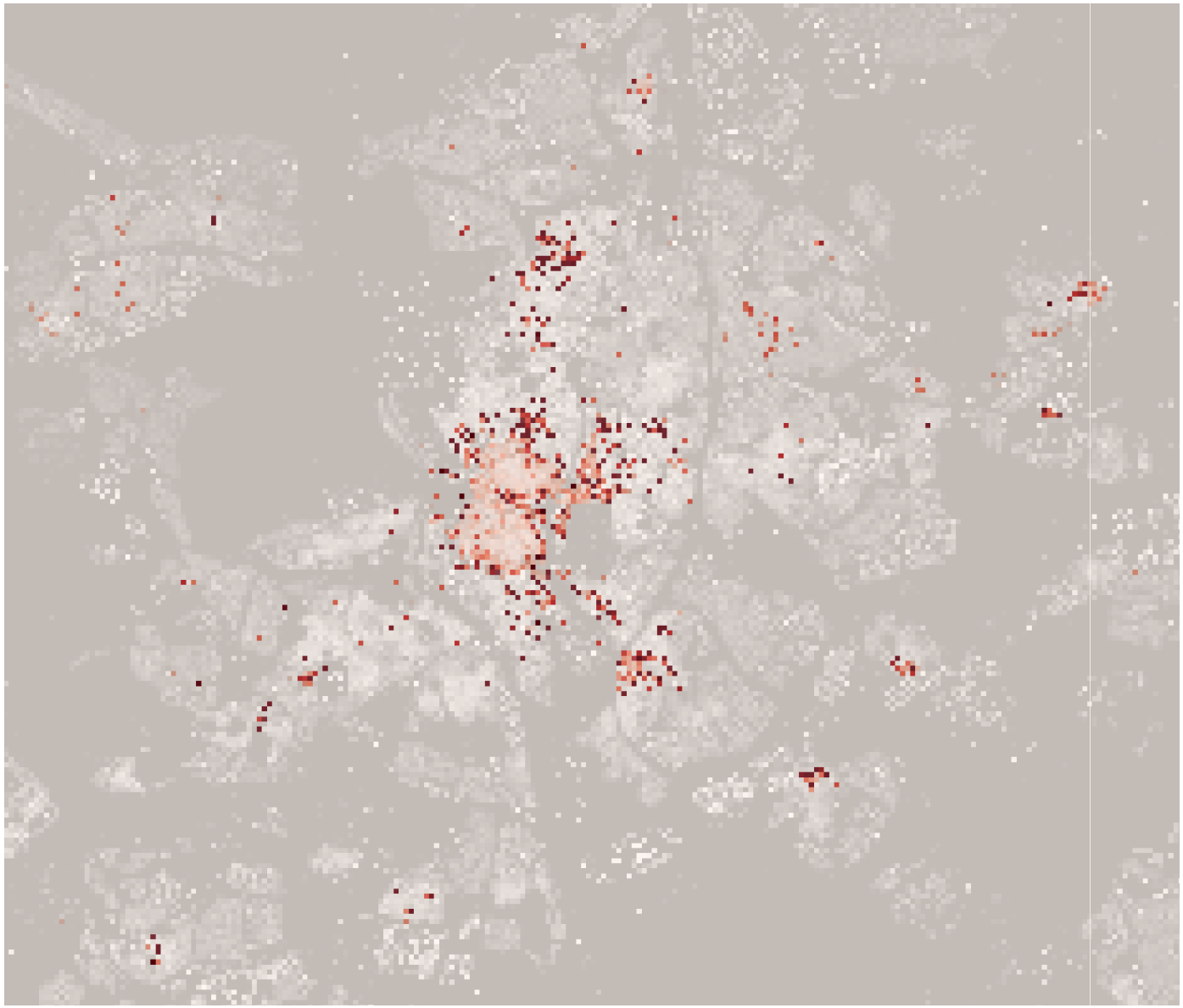
Open Access Earth System Science Data Discussions

Appendix G.



Appendix Figure G1: Attribute completeness at municipality level for the Canary Islands.

Area around Madrid



Apparent discrepancy between phys_bufa_mean and evol_bufa_2020

hisdac_es_phys_bufa_mean_v1_100
Band 1: Band_1 (Gray)
5,000
0

hisdac_es_evol_bufa_v1_100_2020
Band 1 (Gray)
10,000
0