Comparison of built-up area maps produced within the global human settlement framework

Filip Sabo | Christina Corbane | Aneta J. Florczyk | Stefano Ferri | Martino Pesaresi | Thomas Kemper

European Commission, Joint Research Centre, Ispra, Italy

Correspondence
Filip Sabo, European Commission Joint Research Centre Ispra, Disaster Risk Management Unit, Via Enrico Fermi 2749, I-21027 Ispra (VA), Italy.
Email: filipsabo@yahoo.com

Abstract
The validation of built-up areas derived from different sensors is crucial for gaining a deeper understanding of the consistency and interoperability between them. This article presents the methodology and results of an inter-sensor comparison of built-up area data derived from Landsat, Sentinel-1, Sentinel-2, and SPOT5/SPOT6. The assessment was performed for 13 cities across the world for which cartographic reference building footprints were available. Several validation approaches were used: cumulative built-up curve analysis, pixel-by-pixel performance metrics, and regression analysis. The results indicate that Sentinel-1 and Sentinel-2 contribute greatly to improved built-up area detection compared to Landsat, within the global human settlement framework. However, Sentinel-2 tends to show high omission errors while Landsat tends to have the lowest omission error. The built-up area obtained from SPOT5/SPOT6 shows high consistency with the reference data for all European cities, and hence can potentially be considered as a reference dataset for wall-to-wall validation in Europe.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2018 The Authors. Transactions in GIS published by John Wiley & Sons Ltd.
Knowledge of the spatial distribution of human settlements and monitoring of urban expansion are crucial for a large number of applications, such as exposure mapping and risk assessment, infrastructure planning, biodiversity conservation, climate change, and urban development (Chrysoulakis et al., 2014; Triantakonstantis et al., 2015; Florczyk et al., 2016). One key for understanding worldwide urbanization processes and developing actions toward sustainable urban and rural development is the availability of detailed, up-to-date, accurate, and consistent—in time and space—information on human settlements. Human settlements are closely related to population distribution, as well as to social and economic development which, from the spatial perspective, is not directly measurable. Built-up areas, which refer to the physical space used for human habitation, represent the physical dimension of human settlements and the most appropriate surrogate for urbanization processes. The spatial aspect involved in the definition of built-up areas is what makes remote sensing appealing for extracting global and up-to-date information on the extent of the built-up environment that can be related directly to human settlements.

The availability of open and free, high-spatial-resolution earth observation data (Copernicus Sentinel-1 and Sentinel-2 data, Landsat imagery), combined with robust algorithms for automatic information extraction and high-performance computing systems, is opening the path to a new generation of high-resolution (HR) products (10–30 m) describing built-up areas. Over the past decade, several initiatives have successfully produced pan-European and global maps of built-up areas (Aune-Lundberg & Geir-Harald, 2010; Gong et al., 2013; Ferri et al., 2014; Montero, Van Wolvelaer, & Garzón, 2014). The most exhaustive and widely used maps of built-up areas are those produced with TerraSAR-X data, namely the Global Urban Footprint (GUF) (Esch et al., 2013) or with free Landsat data, such as GlobeLand30 (Chen et al., 2015), and the Global Human Settlement Layer (GHSL) (Pesaresi, Ehrlich et al., 2016).

In the framework of the GHSL project, a generic methodology has been developed for automatic extraction of built-up surfaces from large volumes of remote-sensing images (Pesaresi et al., 2013). The main characteristics of the GHSL methodology are scene-based processing and a multi-scale learning paradigm that combines auxiliary datasets with the extraction of textural and morphological image features. This methodological framework proved to be robust and scalable, with the possibility of adapting it for processing imagery in different spatial resolutions from a variety of sensors, including optical and radar imagery. It was applied successfully to the production of continental and global-scale built-up products. At the continental scale, the GHSL methodology was used to produce the European Settlement Map (ESM) from high-resolution SPOT imagery (Ferri et al., 2014; Florczyk et al., 2016) and the South African settlement map (Kemper et al., 2015). The ESM is the first high-resolution built-up layer for Europe and currently the most detailed map expressing the proportion of the pixel area covered by buildings at a spatial resolution of 10 m. At the global scale, a step toward the production of a global human settlements monitoring system was achieved with the first multi-temporal GHSL data derived from Landsat imagery (Pesaresi, Ehrlich et al., 2016), hereinafter referred to as GHSL-Landsat. Compared to concurrent maps of human settlements, GHSL-Landsat currently represents the most up-to-date, multi-temporal, and open data on the physical characteristics and dynamics of human settlements. Drawing on 40 years of Landsat data, the multi-temporal grids describing built-up areas have been produced for the periods 1975, 1990, 2000, and 2015 as part of the global human settlement (GHS) framework.

The open data policy of the GHSL is currently driving the future development of global built-up products. The plan is to maximize the use of Copernicus earth observation missions (i.e. Sentinel-1 and Sentinel-2) in conjunction with Landsat data. In that context, the technology used for producing the first global layers from Landsat data was adapted to the processing of a worldwide coverage of mono-temporal Copernicus Sentinel-1 data (https://ghsl.jrc.ec.europa.eu/s1_2017.php). The output of this experiment was a new global built-up layer at a spatial resolution of 20 m. The prospect of deploying the system for the analysis of Sentinel-2 imagery is a cutting-edge technological issue which requires a series of tests for analyzing the suitability of this new sensor data for mapping
human settlements. In this regard, a prototype has been tested on the first Sentinel-2 images released during the commissioning phase (Pesaresi, Corbane et al., 2016). This study confirmed (i) the noticeable improvement of Sentinel-2 in comparison to Landsat for built-up classification; and (ii) the added value of combining Sentinel-1 and Sentinel-2 data for improving the discriminatory ability when comparing to results obtained using one-sensor data. The prototype for Sentinel-2 is currently in its final development phase and has also been tested on a set of cities across the world.

Despite the successful adaptation of the GHSL methodology to the analysis of several types of satellite data, currently there is a lack of assessment of the consistency among the derived products and insufficient information on their accuracy. Users of these products should get minimum guidance on which dataset to use and for what purpose. Besides, to fully exploit the synergies between the different sensors and possibly integrate those products in the future, it is essential to understand the potential and limitations of each of them and identify complementarities. In this respect, the validation of built-up data derived from different sensors is crucial for gaining a deeper understanding of the consistency and interoperability between the different products. In the specific case of the Sentinel-2-derived built-up data, validation is necessary for improving the workflow currently under development by providing insights into the main issues that need further research and experimentation.

The aim of this article is to provide validation and comparison of the built-up maps, derived with the GHSL framework, from different input sensors and with different spatial coverages:

- At the local scale, with Sentinel-2 for which the workflow for built-up extraction is currently in the prototyping and testing phase.
- At the regional scale, with SPOT5/SPOT6 used for producing the ESM.
- At the global scale, with Landsat and Sentinel-1 used as input data for the production of GHSL-Landsat and GHSL-Sentinel-1, respectively.

As a benchmark, fine-scale building footprints were used as reference data. The analysis was performed over 13 cities selected from different continents, with diverse landscapes, settlement structures, and densities. To paint a complete picture, we developed and applied a validation framework that incorporates conventional methods of pixel-by-pixel accuracy assessment and analysis of grid-based differences in built-up data in relation to settlement densities. Through this detailed validation, this article attempts to identify the main differences in the detection of built-up data across sensors and identify potential areas of synergy between the different maps.

In order to provide an overview of the agreement/disagreement between the different products, an inter-comparison of several globally available layers was also performed.

## 2 Previous Assessments of Built-Up Area Maps

The challenge of providing a consistent and practical definition of “built-up areas” places a limit on the accuracy of the generated products, which is perhaps as significant as the sensor-based errors, such as the spectral data quality or geolocation. In the GHSL paradigm, the fundamental link between earth observation sensor data and human presence is the observable presence of built-up structures or buildings. Specifically, we define built-up areas as the aerial units recording the full or partial presence of buildings and the space in-between buildings (Tenerelli & Ehrlich, 2011). Practically speaking, for the purpose of validation, the definition of a built-up area is translated as “the area where the intersection between the reference building footprint and the raster grid is greater than 0.” Figure 1 shows an illustration of the concept of “built-up area” in a given grid of regularized pixels. The size of the cell is determined by the spatial resolution of the remote sensor.
This definition of built-up areas determines also the technical specifications of the reference datasets required for validating the built-up area maps. The reference datasets should contain information based on building footprints or density of footprints (Pesaresi et al., 2013). At present, there are no global validation datasets with these properties, and not even a statistically representative sample of reference building footprints. The existing global reference datasets have been produced to validate land cover classes, with the built-up areas implicitly considered under the “urban” or “artificial surfaces” land cover classes (Potere, Schneider, Angel, & Civco, 2009; Schneider, Friedl, & Potere, 2009; Gong et al., 2013; Chen et al., 2015). As a rule, given the need to assess the quality of multiple classes, the number of samples representing spatially rare land cover classes (such as the built-up area) may be very small, therefore limiting their utility for the validation of a thematic map (here, a single class built-up area map).

A large majority of the global validation efforts are also based on reference point datasets collected through visual interpretation of high and very-high-resolution satellite imagery. These point-based reference datasets are inadequate for the purpose of our analysis, which aims at preserving the geometric and thematic reliability of the different maps. Table 1 provides a summary of the validation efforts, with a regional or global scope, aimed at assessing the quality of land cover maps, including the classes related to the built-up environment. Some of these initiatives correspond to mono-class validation targeting only the “built-up areas.”

3 | OVERVIEW OF THE FOUR BUILT-UP PRODUCTS AND THEIR CLASSIFICATION METHODS

We present here a brief description of the four built-up maps under assessment, together with an overview of the information extraction workflows which were developed for producing them using imagery from different satellite sensors (Landsat, Sentinel-1, Sentinel-2, and SPOT5/SPOT6).
**TABLE 1** An overview of validation studies of global and regional layers: the “built-up specific” column corresponds to a mono-class validation experiment (targeting the “built-up area” class)

| Dataset(s) | Built-up specific | Validation extent | Validation methodology | Source |
|------------|-------------------|-------------------|------------------------|--------|
| VMAP0, GLC2000, GLOBC, HYDE3, IMPISA, MOD500, MOD1k, GRUMP | ✓ | Global | Tier accuracy assessment, confusion matrix | Potere et al. (2009) |
| GHSL, GUF, MOD500, GLOBC | ✓ | Site-specific | Multi-scale comparison | Klotz, Kemper, Geiß, Esch, and Taubenböck (2016) |
| GHSL | ✓ | Site-specific | Statistical learning | Ouzounis, Syrris, and Pesaresi (2013) |
| GLCC, UMD, GLC2000, MCD12Q1 2001, GLOBC | ✓ | Country | Spatial agreement, consistency analysis | Bai et al. (2014) |
| GLOBC | ✓ | Global | Confusion matrix | Bicheron et al. (2008) |
| IGBP DISCover, UMD, MCD12Q1 2001, GLC2000 | ✓ | Global | Confusion matrix, spatial agreement | Herold, Mayaux, Woodcock, Baccini, and Schmullius (2008) |
| FROM-GLC, MCD12Q1 2010, GLOBC | ✓ | Global | Confusion matrix | Zhao et al. (2014) |
| IGBP DISCover, UMD, GLC2000, GLCNMO 2008, CCI-LC 2000, CCI-LC 2010, MCD12Q1 2001, MCD12Q 2010, GlobeLand30 | ✓ | Country | Area and spatial comparison | Yang, Xiao, Feng, and Li (2017) |
| ESM | ✓ | Site-specific | Confusion matrix, building densities | Florczyk et al. (2016) |

Abbreviations used: vector map level 0 (VMAP0); Global Land Cover 2000 (GLC2000); GlobCover (GLOBC); history database of the global environment (HYDE); global land cover characterization (GLCC); impervious surface area (IMPSA); MODIS urban land cover 500 m (MOD500); MODIS urban land cover 1 km (MOD1k); Global Rural–Urban Mapping Project (GRUMP); University of Maryland land cover product (UMD); finer resolution observation and monitoring of global land cover (FROM-GLC); MODIS land cover (MOD12Q1); International Geosphere–Biosphere Program Data and Information System (IGBP); Climate Change Initiative land cover (CCI LC).
The technology at the core of the Landsat, Sentinel-1, and Sentinel-2 built-up products relies on the symbolic machine learning (SML) supervised classifier (Pesaresi, Syrris, & Julea, 2016). The basic concepts of the SML methodology are presented briefly in the next subsection. Then, the application of the SML to the classification of Landsat data for the generation of GHSL-Landsat is recalled, followed by a concise introduction of the adapted methods for the analysis of Sentinel-1 and Sentinel-2 data. Finally, a brief overview of the methodology used for the generation of the ESM from SPOT imagery is also given here for information purposes.

3.1 | The SML classifier

SML is a new generic supervised classification framework developed at the JRC (https://ghsl.jrc.ec.europa.eu/tools/php), which provides a scalable solution for complex and large multi-scene satellite data processing (Pesaresi, Vasileios, & Julea, 2016).

The SML schema is based on two relatively independent steps:

1. Reducing the data instances to their symbolic representation (i.e. a set of unique discrete data sequences).
2. Evaluating the association between the unique data sequences subdivided into two parts: $X$ (input features) and $Y$ (known class abstraction).

The association is measured as a confidence index, called the Evidence-based Normalized Differential Index (ENDI), which provides a continuum of positive and negative values ranging from $-1$ to $1$. The ENDI expresses the strength of association between the image data layers and the reference data. Values close to $1$ indicate that the data sequence is strongly associated with the image class of interest (the built-up area in our case), while values close to $-1$ indicate that the feature is strongly associated with classes other than the built-up area.

3.1.1 | Built-up extraction from Landsat (GHSL-Landsat)

The GHSL-Landsat product was generated using an information extraction technique based on the SML classifier. The input data consists of four Landsat data collections for 1975, 1990, 2000, and 2014. The classification of Landsat data in the GHSL framework uses reference data for supervised learning composed of coarse-scale [i.e. MODIS global urban extents (Mertes, Schneider, Sulla-Menashe, Tatem, & Tan, 2015), MERIS GLOBCover 2009 (Bontemps et al., 2011), and LandScan population density grid (Dobson, Bright, Coleman, Durfee, & Worley, 2000)] and fine-scale data [e.g. Open Street Map (OSM) and settlement polygons extracted from the urban classes of CORINE Land Cover (Aune-Lundberg & Geir-Harald, 2010)].

The GHSL-Landsat product that was validated in this study corresponds to the 2015 epoch (i.e. Landsat images acquired during 2013–2014), hereinafter referred to as Landsat.

Details on the information extraction procedure are given in Pesaresi, Ehrlich et al. (2016) and Corbane, Pesaresi et al. (2017a).

3.1.2 | Built-up extraction from Sentinel-1

In 2016, the availability of a global coverage of high-resolution SAR data collected from the European Sentinel-1 mission was a motivation for testing the applicability of the SML classifier to this new imagery in view of improving and updating GHSL-Landsat (Corbane, Lemoine et al., 2017).

The SML workflow was adapted to exploit the key features of the Sentinel-1 ground range detected (GRD) data, which are: (i) spatial resolution of 20 m with a pixel spacing of 10 m; and (ii) availability of dual polarization acquisitions (VV and VH), widely used for monitoring urban areas since different polarizations have different sensitivities and different backscattering coefficients for the same target (Matsuoka & Yamazaki, 2004).
The input features to the SML classifier consisted of: (i) dual polarized backscatter intensities processed at a resolution of 20 m; (ii) mean and standard deviations of backscatter intensities calculated for different window sizes (i.e. 3 × 3, 5 × 5, 7 × 7, and 9 × 9); and (iii) topographic features (slope, aspect, and crest lines) derived from a digital elevation model (SRTM 90 m Digital Elevation Database v4.1) with the purpose of attenuating the confusion between built-up and topographic structures.

The learning data at the global level consisted of the union of the built-up data obtained from Landsat and the Global Land Cover map at 30 m resolution (GLC-30). The latter has also been derived from Landsat imagery through operational visual analysis techniques (Chen et al., 2015).

A simplification of the adapted SML workflow for the classification of Sentinel-1 data ($S_1$) is shown in Figure 2, with a total of 21 input features, and a reference built-up layer (derived from the built-up data from Landsat and GLC-30) used for learning in the association analysis.

### Built-up extraction from Sentinel-2

At the time of writing this article, the prototype for built-up extraction from Sentinel-2 imagery was still under development. However, an almost stable version of the algorithm was being tested in different landscapes with satisfactory results. The algorithm builds on the SML workflow with adjustments designed to exploit the key features of Sentinel-2 data: (i) availability of four 10 m spatial resolution bands (B2, blue; B3, green; B4, red; B8, near infrared); (ii) the availability of six bands at 20 m resolution, especially in the near infrared and shortwave infrared (B5, B6, B7, B8a in near infrared; B11, B12 in shortwave infrared).

The following features derived from Sentinel-2 are used for classification of the Sentinel-2 image with the SML approach.

---

**Figure 2** Simplified workflow showing the adaptation of the SML to the classification of Sentinel-1 images at the global level. The input features comprise 18 features derived from dual-polarization Sentinel-1 intensity data and 3 topographic features derived from a global digital elevation model (i.e. SRTM).
• Spectral features: four 10 m resolution and six 20 m resolution bands.
• Textural features: four textural features derived from the four input 10 m bands by applying a texture-derived built-up presence index, pantex (Pesaresi, Gerhardinger, & Kayitakire, 2008). These four features were combined into a single feature using the minimum operator. The textural feature is used to refine the output confidence layer by eliminating overdetections (mainly roads and open spaces classified as built-up areas).

The learning set is based on Landsat. The Sentinel-2 images used in the classification are first atmosphere and terrain corrected [level 2A outputs are referred to bottom of atmosphere (BOA)] with the SEN2COR toolbox (Louis et al., 2016) prior to classification. This is relevant especially when deriving other thematic layers from Sentinel-2, such as the vegetation layer, and for monitoring large areas at continental scale. Both layers are seen to be part of the global GHSL product derived from Sentinel-2 during 2017. One of the outputs of the atmospheric correction in SEN2COR is scene classification at 20 m resolution with 10 different classes. This rough classification is used in the SML classifier for stratifying the learning set based on Landsat. This allows the training set to be tailored to the image under processing, especially in the presence of clouds or cloud shadows, and hence allows reducing commission and omission errors.

Figure 3 presents a simplified version of the workflow for the classification of Sentinel-2 data (S2).

**FIGURE 3** Simplified workflow showing the adaptation of the SML to the classification of built-up area from Sentinel-2 BOA images (excluding the extraction of additional thematic layers, e.g. vegetation, water, etc.). The input features comprise reflectance values of the four 10 m bands resampled at 20 m and the reflectance values of the 20 m bands. The original 10 bands are used to derive a textural feature that is used for postprocessing the ENDI confidence and refining the detection of built-up pixels.
3.2 | ESM based on SPOT5 and SPOT6

The ESM is the first high-resolution built-up layer for Europe at 2.5 m. The ESM classification workflow used more than 3,500 satellite images from SPOT5 and SPOT6 sensors as input data with spatial resolutions of 2.5 and 1.5 m (Florczyk et al., 2016). The ESM production workflow was based strongly on the GHSL methodology introduced by Pesaresi et al. (2013); however, the GHSL approach was adapted in order to allow the 2.5 m input data and use of the vegetated surface detector. The main elements of the GHSL methodology were still the core of the applied method: (i) scene-based processing; and (ii) multi-scale learning which combines auxiliary datasets with the extraction of textural and morphological image features. The validation of the ESM layer against the Land Use/Cover Area frame Survey (LUCAS) data (Eurostat, 2015) gave an overall accuracy of 96% with omission and commission errors lower than 4% and 1%, respectively. Full details on the ESM workflow and validation can be found in Florczyk et al. (2016).

In this work, the validation is performed using the recent release of ESM at 2.5 m spatial resolution (Ferri, Siragusa, Sabo, Pafi, & Halkia, 2017). Table 2 provides a summary of the layers analyzed in this study. It should be noted that the S2 layer is still under production, and the results presented here were derived from sensor images acquired in 2016.

3.3 | Ancillary datasets

The SML framework makes large use of already available spatial information describing human settlements from various sources at different scales, thematic definition, and completeness or accuracy conditions. SML is a very robust classifier. This has been tested in three typical scenarios of altered learning sets:

- scale generalization noise in which the learning set was degraded to 10 different spatial resolutions;
- random thematic noise, in which a salt-and-pepper noise was introduced to the learning set;
- spatial displacement noise, in which the learning set was systematically shifted by 36 different displacement vectors with a worst displacement of 1,080 m.

The results showed that the SML outperforms parametric classifiers in terms of robustness. It is largely agnostic, both with respect to the statistical distribution of the input data and to the relations between the image data and the information to be detected.

TABLE 2  Characteristics of the built-up products under validation

| Product (imagery) | Spatial resolution (m) | Imagery timestamp | Coverage | Data source |
|-------------------|------------------------|-------------------|----------|-------------|
| **SPOT5/SPOT6 (SPOT5/SPOT6)** | 1.5, 2.5 | 2010–2013 | Europe | https://land.copernicus.eu/pan-european/GHSL |
| **S1 (Sentinel-1)** | 10, 20 | 2016 | Global | https://ghsl.jrc.ec.europa.eu/ghs_bu_s1.php |
| **S2 (Sentinel-2)** | 10, 20 | 2016 | Selected cities | Under production |
| **Landsat (Landsat-8)** | 30, 38 | 2013–2014 | Global | https://ghsl.jrc.ec.europa.eu/ghs_bu.php |
This fact largely facilitates the generalization of the identical classification process to image data collected by different sensors and with different local data collection conditions (season, illumination, building practices, materials, settlement patterns, natural background) (Pesaresi, Syrris et al., 2016).

All the datasets that were used here as a training set or in the postprocessing step are referred to as ancillary datasets.

- MODIS global urban extents (MODIS) was produced with a supervised decision tree classification algorithm for 1-year MODIS data input (2001–2002). Validation of MODIS with fine-scale building footprints showed an overall accuracy of 0.86 and a kappa coefficient of 0.09 (Pesaresi, Ehrlich et al., 2016).
- The GlobCover 2009 land cover product is the second 300 m global land cover map produced from an automated classification of MERIS time series. Validation of GlobCover 2009 class “artificial surfaces” with fine-scale building footprints yielded an overall accuracy of 0.88 with a kappa coefficient of 0.08 (Pesaresi, Ehrlich et al., 2016).
- GLC-30 was produced with a hybrid pixel–object–knowledge-based classification. Reported overall accuracy is 80%, with a kappa coefficient of 0.75 and a user accuracy of 86.70% for artificial surfaces.
- Both the Soil Sealing Layer from the European Environment Agency (EEA) and Corine Land Cover (CLC) 2006 rasters with 100 m pixel size were used as a training set for the production of the SPOT5/SPOT6 product. The training set used to rescale the pantex and saliency features was defined as an intersection of sealed surface greater than 0 and urban fabrics from the CLC.

All the datasets used, either as a learning set or in the postprocessing step, are summarized in Table 3.

**TABLE 3** Ancillary datasets used for training the built-up products or in the postprocessing step

| Product                | Learning set  | Reference for learning set | Learning set timestamp | Pixel size (m) | Product auxiliary set     |
|------------------------|---------------|----------------------------|------------------------|----------------|---------------------------|
| MODIS                  | Schneider et al. (2009) | 2002 | ~500 | Open Street Map          |
| Landsat                | MERIS Globe Cover | Arino et al. (2012) | 2009 | ~300 | Geonames                  |
|                        | LandScan population grids | Dobson et al. (2000) | 2013 | ~1,000 |                          |
| S1 Landsat             | Pesaresi, Ehrlich et al. (2016) | 2014 | 38 | /                  |
| GLC-30                 | Chen et al. (2015) | 2013 | 30 |                     |
| S2 Landsat             | Pesaresi, Ehrlich et al. (2016) | 2014 | 38 | /                  |
| SPOT5/SPOT6 Soil Sealing Layer | EEA* | 2010 | 100 | Urban Atlas            |
| Corine Land Cover      | Aune-Lundberg and Geir-Harald (2010) | 2006 | 100 | Open Street Map          |

*https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-imperviousness-2
4 | VALIDATION METHODOLOGY

In this section we propose a validation framework which aims to explore a series of hypotheses related to the sensors under assessment and the features used for built-up extraction from the different sensors. In terms of sensor performance, we test the following:

- The SPOT5/SPOT6-derived product outperforms the other products due to its fine spatial resolution.
- The Landsat-derived product provides the least accurate results due to its lower spatial resolution in comparison to the other sensors.
- The products derived from Sentinel-2 and Sentinel-1 outperform the Landsat-derived product due to their higher spatial resolution and, in the case of Sentinel-1, we expect underdetection in dense built-up areas due to shadow effects that occur on those sides of vertical structures that are facing away from the incoming radar beams.
- Textural features, which allow discrimination of built-up structures from other image features (edges and lines), have the drawback of producing a higher omission rate (Pesaresi et al., 2008).

The validation framework developed for testing these hypotheses is described in this section. It incorporates conventional methods of pixel-by-pixel accuracy assessment and analysis of grid-based differences in built-up data in relation to settlement densities. The purpose of this framework is to achieve a comprehensive and systematic description of the accuracy and validity of a layer under comparison.

First, we determine the absolute accuracies and performances on the basis of a pixel-by-pixel accuracy matrix. Then, the dependencies between the built-up and the physical settlement structures are explored using a grid-based analysis. This is performed following two approaches: (i) analysis of the relationship between built-up densities derived from the reference data and the cumulative built-up area from the different layers; and (ii) correlation analysis between the sums of built-up pixels per cell derived from the different layers and the observed reference built-up cell sums. The rationale and details of these three validation approaches are described in the following subsections.

The methodology and results of the global inter-comparison follow, after the methodology and results of the three previously mentioned validation approaches.

4.1 | Performance metrics

This subsection is dedicated to the standard pixel-based accuracy assessment, which is a de facto standard in the remote sensing community. It is based on the metrics derived from the confusion matrix (i.e. error matrix) (Congalton, 1991). For the analysis of absolute classification accuracies, we follow the recommendations of Foody (2008) and base the interpretation of results on a combination of meaningful accuracy metrics beyond the use of a single statistic. As standard descriptive measures, we report the overall accuracy (OA), kappa statistics, commission error (CE), and omission error (OE) according to Equations (1–4):

\[
\text{OA} = \frac{TP + TN}{TP + FP + FN + TN}
\]  \hspace{1cm} (1)

\[
\text{kappa} = \frac{p_o - p_e}{1 - p_e}
\]  \hspace{1cm} (2)

\[
\text{CE} = \frac{FP}{TP + FP}
\]  \hspace{1cm} (3)
where the standard labels used in the binary confusion matrix are: TP = true positive; TN = true negative; FP = false positive; FN = false negative and $p_o = OA$; $p_e = \frac{(TP + FN) \cdot (TP + FP) + (FP + TN) \cdot (FN + TN)}{TP + FP + FN + TN}$.

The overall accuracy informs us about the correct classifications for all the pixels in a specific site (Equation 1). However, this metric is known for being vulnerable to bias from skew due to imbalanced data, as in the case of built-up areas (Jeni, Cohn, & De La Torre, 2013). In this work, we focus on the relative comparison of the different built-up layers that we analyze in combination with additional performance metrics. The kappa coefficient (Equation 2), coined by Cohen (1960), highlights the differences between the actual agreement in the error matrix (i.e. the correctly classified sample units presented by the major diagonal) and the chance agreement presented by the column and row totals. Landis and Koch (1977) proposed a categorization of kappa in which values of 0.00 to 0.20 are regarded as poor, 0.21 to 0.4 as fair, 0.41 to 0.6 as moderate, 0.61 to 0.8 as substantial, and 0.81 to 1.00 as almost perfect agreement in order to ease the comparison of multiple classification outputs. CE (Equation 3) is defined as the fraction of values that were classified as built-up area but do not belong to that class. OE (Equation 4) is defined as the fraction of values that belong to the built-up class but were not classified as such.

### 4.2 Built-up density analysis

The analysis of built-up density introduced by Ferri et al. (2014) and Florczyk et al. (2016) is a grid-based method with a predefined cell size, which allows assessing the relationship between the built-up density and the cumulative sum of built-up area derived per layer under study. This allows us to establish a stronger understanding of the mapping capabilities of each sensor in relation to built-up density and structural characteristics. In this validation experiment, the cell size is set to $500 \times 500$ m. This cell size is a good compromise between the areas of the cities under analysis, the necessity to capture commission and omission errors, and computational constraints. Several study sites are not big enough to use a 1 km or bigger cell size (see Table 4). Also, an experiment with smaller cell sizes, less than 50 m, is computationally demanding. For each cell, the sum of the built-up area for each layer is calculated. The maximum cell sum is $250,000$ m$^2$, because the common pixel size used for the resampled layers is 1 m. In order to derive the built-up density of reference cells, the total sum of pixels classified as built-up area in a cell is divided by the total area of the cell. The formula is as follows:

$$dens_i = \frac{\sum_{k=1}^{N} bu_k}{w_i \times h_i}$$

where $dens_i$ is the density for cell $i$, $bu_k$ is the $k$th built-up pixel in the specific cell $i$, $N$ is the maximum number of built-up pixels in one cell, $w_i$ and $h_i$ are the width and height of the cell, respectively.

The reference densities are ordered from the highest to the lowest. The cumulative built-up area sums are calculated for layers under validation and plotted against the reference built-up area. Figure 4 displays the built-up density analysis for the city of Amsterdam with different cell sizes ($250$, 500, and 1,000 m). It shows that this analysis is robust to changes in cell size.

The GHSL approach measures the percentage of built-up area in a given spatial unit and this kind of analysis can explain whether the specific layer overestimates or underestimates the built-up area and for which given densities of built-up area.
4.3 | Regression analysis of built-up area

The regression analysis uses similar input as the cumulative built-up density analysis by exploring the correlation between the sums of pixels classified as built-up area per cell derived from the different layers, and the reference sums of built-up area per cell. We analyze the scatterplot matrices displaying pairwise correlations. To support the analysis of scatterplots, we also compute the Pearson coefficient of correlation $r$, the slope and intercept from first-order linear regression:

$$y = a \cdot x + b$$

where $a$ is the slope coefficient and $b$ is the intercept, $y$ corresponds to the specific layer under analysis (SPOT5/SPOT6, S1, S2, or Landsat), and $x$ to the reference layer derived from building footprints.

Through this analysis, we explore the possibility of estimating the “actual” built-up area given an input layer derived from one of the satellite sensors assessed in this work. The outputs of the correlation analysis can also give an indication of the degree of interoperability between the different built-up layers, which may facilitate the task of sharing data between different agencies and organizations working on human settlements.
In addition to the selected cities validation, a global inter-comparison experiment was performed covering the following datasets: S1 and Landsat from the GHSL framework, GLC-30, and MODIS (Table 3). The SPOT5/SPOT6 product is available only in Europe, so it was excluded from this analysis as well as the S2 product, which is still in the production phase. All the products were warped to a common pixel size of 30 m. The projection used for the global comparison was Google Mercator.

The built-up area in pixels (30 m) was derived and compared per continent. A cross-correlation of built-up densities was performed by calculating the sums of built-up pixels in a grid with cell size 150 × 150 km. The total number of grid cells was 23,000.

5 | STUDY SITES AND REFERENCE DATA

The proposed validation framework was applied to 13 study sites located in Europe, North America, Australia, Africa, and Asia. The study sites were selected based on availability of reliable reference data (with preference for cadastre sources) and the need to cover different types of built-up structures, settlement densities, and landscapes.

Reference data consist of fine-scale building footprints obtained from different sources: national mapping agencies, national geoportals, and OSM official data. Building footprints downloaded from the OSM were first visually inspected with the help of very-high-resolution (VHR) images to ensure completeness of the geographic coverage. A very conservative approach was adopted in which the areas with missing building structures (as visually assessed using VHR) were excluded. This approach ensured that gaps in the OSM reference data are avoided. The reference building footprints were first rasterized to 1 m pixel size to preserve maximum possible
detail. All other layers were resampled to the same resolution as the reference data using nearest-neighbor interpolation in order to match the pixel size of reference building footprints of 1 m. A common Universal Transverse Mercator projection was used for all layers. All the layers under assessment were up-sampled in order to not favor a particular layer. The 13 selected cities for which the fine-scale reference data were available are shown in Figure 5.

Table 4 gathers, for each city, the source, timestamp (year), and total area of the reference dataset. The second column refers to the extent of the validated area (approximately) and not to the specific city boundary area. This extent was determined according to the reference footprints coverage grid. The total validated area is approximately 12,588 km$^2$.

Reference built-up densities for European cities are plotted in Figure 6, with the median value also shown in the corresponding color. The curves are similar in shape for the cities of Milan, Torino, and Novara, where the building footprints cover only dense city areas. For Montpellier, Warsaw, Amsterdam, and Oslo, the building footprints cover a much bigger area, which includes smaller settlements, villages, and other built-up areas. Here we can notice that the curves still increase significantly when the reference density is converging to 0%. The median value for the Warsaw study site is higher than for all the other sites where the reference data does not cover only dense urban zones, which means that there is a significant amount of built-up area outside the Warsaw urbanized zone.

### 6 | RESULTS AND DISCUSSION

In this section the proposed validation framework is applied to the selected 13 study areas (cities) followed by several visual examples of the built-up classification results. The results and discussion for the global inter-comparison are also given in this section.
6.1 | Results of performance metrics

Figure 7 displays the comparative performance metrics for each city. The analysis of the four different performance metrics shows that the SPOT5/SPOT6 layer resulted in the highest values of overall accuracy and kappa and the lowest value of CE. The performances of S1 and S2 in European cities are very similar, with the average kappa value being slightly higher for S1, while the OE for S2 is significantly higher than the OE of the other layers. This may be explained by the fact that the S2 workflow includes a textural analysis (pantex) aimed at refining the derived built-up layers by removing roads and other non-built-up areas with high confidence values. The textural refinement seems to exclude a lot more than non-built-up areas by also excluding large buildings (e.g. industrial) and some groups of buildings in very dense cities (Figure 8). Landsat has the highest value of CE. This is because Landsat tends to classify some road networks, runways, bare soil in arable land and river beds, and so on as built-up areas (Figure 8). In general, all four layers show high average commission errors (i.e. greater than 0.5). In terms of OE, Landsat shows the best results with very low values, even lower than SPOT5/SPOT6. S1 also shows good overall performance between those of S2 and Landsat.

6.2 | Results of built-up density analysis

The results of the built-up density analysis are presented in Figure 9, showing the behavior of the cumulative built-up area of each layer, including the reference building footprints (in black) in relation to the reference built-up area densities. The densities calculated per grid cells of 500 × 500 m are ordered from the highest to the lowest density. The cumulative sums of built-up areas are calculated for each validated layer and plotted respecting the
order of reference densities. The analysis is made by visual comparison of the curves derived from SPOT5/SPOT6, S1, S2, and Landsat to the “reference curve.”

Since these are cumulative sums of built-up area, the maximum value is the total built-up area reported by the layers. Evidently, the curves calculated from SPOT5/SPOT6 are closest to the reference curves in terms of both shape and distance. It is notable that for Milan and Torino, the cumulative built-up curves for SPOT5/SPOT6 almost coincide with the reference layer. Hence, the final estimated area is close to the reference data. However, SPOT5/SPOT6 still tends to overestimate the built-up area in general terms, except in the case of Torino, where it slightly underestimates the total built-up area. The almost perfect match between the SPOT5/SPOT6 curve and the reference curve, especially in the case of Torino and Milan, can be explained by the fact that the SPOT5/SPOT6 workflow uses the same reference building footprints as an auxiliary dataset for filling the gaps in the final SPOT5/SPOT6 product. This certainly introduces a bias in the presented validation.

Landsat overestimates the built-up area significantly, and this overestimation is reached very fast. S2 shows good results, with a curve very close and parallel to the SPOT5/SPOT6 curve (e.g. in the case of Montpellier, Oslo, Novara, and Milan). S1 performed better than Landsat, but also has a general tendency to overestimate the built-up area in sparse urban zones.
For the cities of Warsaw, Oslo, Amsterdam, and Montpellier, the curves start to diverge when the reference built-up density is around 20%. At that point, the cumulative values for all layers start to differ greatly, and that is when we begin to observe the overestimation. The reference data in those cities covers not only the dense city centers but also sparse built-up areas and some rural zones. In those cases, we can observe the tendency of the different layers to overdetect the built-up area in zones characterized by sparse and scattered settlement patterns. In the Italian cities (Novara, Torino, and Milan) the reference data covers essentially the urban core areas, where the density of the built-up area is very high. In that situation, all the layers show almost the same behavior and start overestimating around 50% of the built-up density.

Overestimation of built-up areas for all three layers (S1, S2, and Landsat) is expected due to the nature of the built-up area observed from the satellite sensors, and due to the inherent semantic definition of settlement areas that does not comply with individual building outlines.

It should be noted that the total built-up area estimated by S2 for Aizuwakamatsu (Japan) is higher than that estimated from Landsat (see Figure 11 later). This site is characterized by many rural areas which surround the city. Unlike the other sites, the OE of S2 was 0.27, as opposed to that of Landsat (0.33) and S1 (0.36) (Figure 6), due to the predominance of small buildings. Besides this, the Landsat built-up layer rarely (if ever) detects small...
FIGURE 9  Cumulative built-up area curve analysis. The x-axis corresponds to decreasing reference built-up densities calculated for grid cells of 500 × 500 m. The y-axis is the cumulative sum of built-up areas for all layers, including the reference layer.
scattered settlements such as those detected by S2, and usually underestimates them. A visual example is provided in Figure 10.

Figure 11 provides a summary of the estimated built-up area for each city and for each layer, as well as the reference built-up area. The results highlight some similarity in the total built-up area, as derived from SPOT5/SPOT6 and S2, especially for the cities of Montpellier, Novara, and Oslo. S1 and Landsat also provided high similarity of the total built-up area estimated for several cities: Glenorchy, Novara, Dar es Salaam, and Warsaw.

### 6.3 Results of built-up regression analysis

The results of the regression analysis are shown in Figures 12 (for European cities, without Novara) and 13 (for non-European cities). The values of the scatterplot matrix on the x-axis correspond to the reference built-up area,
whilst the values on the y-axis are the values of the built-up area for the layers under validation (SPOT5/SPOT6, S1, S2, and Landsat).

**FIGURE 11** Estimated built-up area per layer and the reference built-up area
FIGURE 12  Scatterplot matrices with correlation coefficients for European cities excluding Novara. The x-axis corresponds to the reference built-up area extracted from 500 × 500 m grid cells, while the y-axis corresponds to the built-up area obtained from grid cells for specific layers (SPOT5/SPOT6, S1, S2, and Landsat). Trendlines are shown in red.
FIGURE 13 Scatterplot matrices with correlation coefficient in non-European cities
The results confirm the cumulative built-up curve analysis, implying a more accurate representation of built-up structural variability with SPOT5/SPOT6 in comparison to the other layers. This is manifest by the high r values, which range between 0.91 and 0.96. Also, the linear relationship is very prominent between the SPOT5/SPOT6 and reference datasets, especially in the case of Milan and Torino.

From the scatterplot (Figure 12), it can also be seen that both Landsat and S1 overdetect built-up areas, with Landsat showing a second-order polynomial relation with the reference data: the values in the Landsat vs. reference scatterplot (first column, fifth row) are mostly concentrated in the left and upper-left corner of the plot. That is, for low values of reference built-up densities, Landsat gives much higher values than the other layers. A significant number of values are sealed at the top. This relationship is similar also in the case of Landsat vs. SPOT5/SPOT6, S1, and S2 plots (fifth row, columns two, three, and four), where values are concentrated in the upper parts of the scatterplot for most of the cities. This is a consequence of the fact that many Landsat cells have values of built-up density equal to 100%, while the ground-truth data report lower values for those cells.

Overall, when looking at the results obtained for all study sites, we notice that the r values are the lowest in the case of Landsat for 9 out of the 13 validated sites. However, for the sites characterized by significant presence of non-built-up area and sparse built-up area cells (e.g. Oslo, Glenorchy, Surrey, and AizuwaKamatsu), Landsat performs very well in terms of r values. For cities with very high building densities, Landsat gave the smallest coefficients of correlation, only 0.72 and 0.69 (New York, Milan, and Washington).

Both S1 and S2 are strongly correlated with the reference layer and with SPOT5/SPOT6. Similar to Landsat, they tend to overestimate the built-up areas, especially S1. It is important to note that the shape of the relation between S2 and the reference data is different from that observed for S1 and Landsat: there is less overdetection in S2, especially in fragmented built landscapes, as in the case of Montpellier and Amsterdam. These results confirm the results obtained from the analysis of cumulative built-up area curves.

Figure 14 displays the scatterplot matrix when using all observations together, independently of the city. S1 and S2 gave the same R value followed by a small decrease in R for the Landsat product. The slope values

![Figure 14: Scatterplot matrix with correlation coefficients when using all the observations](wileyonlinelibrary.com)
from the linear regression model for S1, S2, and Landsat when using merged observations are 2.8, 2.1, and 3.1, respectively.

The slopes of the regression analysis are displayed in Figure 15. For SPOT5/SPOT6, the regression slopes are stable and fall in the range [1, 2]. Given the estimated built-up area obtained with SPOT5/SPOT6 (Figure 11) in comparison to the actual reference built-up area, and considering the average slope value of the fitted linear models, it is reasonable to state that SPOT5/SPOT6 overestimates the built-up area by a factor of 1.5.

It is interesting to note that the regression slopes from S2 for four European cities are also in the range [1, 2], except for Warsaw. In the case of Warsaw, the built-up area is more concentrated in low-density areas, which is not common for other sites (Figure 6). The overestimation tends to be high for all products when the built-up area is more concentrated in areas with low built-up density. The slope value from S2 in the case of all cities is 2.1. Considering also a linear regression model, it is possible to quantify the overestimation of the built-up area for S1, S2, and Landsat. S2 overestimates the built-up area by a factor of 2.1. In the case of S1 and Landsat, the overestimations show a much larger variability per city and the slope values from the regression model when using all observations together are 2.8 and 3.1.

Figure 16 shows an example of the level of precision of the built-up area as described in the SPOT5/SPOT6 layer. The similarity and consistency between SPOT5/SPOT6 and building footprints is very striking. In addition, the SPOT5/SPOT6 layer detects some additional buildings, which are missing in the reference data.

The empirical results confirm most of our hypotheses. The SPOT5/SPOT6 product outperformed the other products because of its fine spatial resolution. The Landsat product was the least accurate since its spatial resolution is the lowest. Also, S1 and S2 outperformed the Landsat product due to their higher spatial resolution. We refuse the hypothesis that S1 was underdetecting the built-up area in dense urban zones due to shadow effects. The textural feature introduced more omission rates, since the pantex feature used in the S2 workflow contributed to the increased omission error.
6.4 | Global inter-comparison

The total built-up area for each product and for each continent is shown in Figure 17. The built-up area is expressed in number of pixels, where the pixel size is 30 m. As can be seen, there are significant variations between the products, especially in Europe. GLC-30 detects more built-up areas than all the other products for almost all continents. However, it should be noted that the GLC-30 "artificial class" considers all man-made structures as built-up (urban areas, roads, mines, rural cottages, concrete, etc.), which explains the systematic detection of more built-up pixels compared to other products, except for South America and Africa. S1 detection of built-up pixels varied throughout the continents. In Europe and Asia it detected more than Landsat, but not in other continents. It was shown by Corbane, Pesaresi et al. (2017b) that S1 contributed to fewer omission errors in Africa.
and Asia compared to the Landsat product. MODIS detected the least built-up pixels for almost all continents. Smaller settlements and smaller cities are not detected because of the coarse spatial resolution of the sensor. Furthermore, the MODIS methodology is designed to capture dense built-up land (>50% built-up) (Schneider, Friedl, & Potere, 2010). The variations in built-up area detection between the products are related to the type of
sensor (radar or optical), training sets, sensor resolution, and the methodology used for the production of layers. The correlation matrix is shown in Figure 18.

S1 and GLC-30 products provided very good correlation, with R value of 0.96, followed by MODIS and Landsat with 0.92. S1 used GLC-30 and Landsat as training data, and therefore the high R values were expected. A similar interdependency is found in the case of MODIS and Landsat, where MODIS was used as a training set for the Landsat product. MODIS and GLC-30 provided the lowest value of R.

7 | CONCLUSIONS

An inter-sensor comparison of built-up areas derived from different sensors was presented for 13 selected cities and four built-up layers obtained in the framework of the GHSL project. The quantitative validation was performed using detailed building footprints available from OSM or from national mapping agencies and local authorities. We first determined absolute accuracies and sensor performances based on a pixel-by-pixel accuracy matrix, considering the entire population of reference pixels. Then, we explored the dependencies between the built-up density and the physical settlement structures using a grid-based analysis. This was accomplished using two approaches: (i) analysis of the relationship between built-up densities derived from the reference data and the cumulative built-up area from the different layers; and (ii) correlation analysis between the sums of pixels classified as built-up area per cell derived from the different layers and the observed reference layer. Finally, a visual comparison of built-up classification results was performed, illustrating areas of agreement/disagreement between the layers.

An additional experiment focusing on the global built-up area products comparison was also performed. This analysis allows us to derive the following main observations:

1. ESM (SPOT5/SPOT6) shows the most accurate results relative to the reference data.
2. Both S1 and S2 showed an improved extraction of the built-up layer compared to that derived from Landsat.
3. All the layers overestimate the built-up area due to their inherent semantic definition of settlement areas that does not comply with individual building outlines. In particular, Landsat, with the lowest spatial resolution, was shown to be the product with the highest overestimation of built-up areas.
4. The results varied significantly across the different type of cities, suggesting the need to group the analysis per types of landscape and settlement patterns.
5. The total built-up area per continent varies significantly between the global products.

Based on all the validation experiments, SPOT5/SPOT6 gave the most accurate representation of the built-up area, independent of the settlement patterns and their densities. Given that reference building footprints are rarely available for large-scale validation, the outputs of this study suggest that the SPOT5/SPOT6 layer represents a good alternative for validation at the European level. Besides, previous validation work with SPOT5/SPOT6 (Florczyk et al., 2016) showed that this layer—generated by automatic image information extraction—achieves 96% agreement with the LUCAS dataset. The omission and commission errors are less than 4% and 1%, respectively. This also confirms its suitability as a good reference dataset for the validation of future built-up products within Europe.

Compared to Landsat, S2 and S1 showed significant improvements associated with the exclusion of agricultural fields, parking lots, sand beaches, and roads from the detected built-up area. S2 was good at detecting small, scattered settlements but failed to detect large buildings in dense urban zones. Conversely, S1 did not succeed in identifying scattered settlements, but correctly classified large industrial buildings. These results outline the complementarity between S1 and S2 sensors for increased accuracy in the detection of built-up areas. Deeper insights into the characterization of human settlements can certainly be gained from the integration of S1 and S2 results.
The results of the global built-up area comparison showed significant variation. GLC-30 detected the most built-up pixels in the global comparison for almost all continents, while MODIS detected the least built-up pixels (excluding South America). Interdependencies exist between the global products because of the training sets used for production. S1 was correlated with Landsat and GLC-30, while Landsat was correlated with MODIS.

The large variability in the results of the validation between the study sites suggests a need to expand the sample size to cover cities in different landscapes and with different settlement patterns. The results, which would incorporate a wide scattered sample size, could be used to develop robust cross-sensor built-up models. A validation campaign using image interpretation and crowdsourcing is currently ongoing, with the purpose of validating the GHSL products and other available global products describing human settlements (https://www.geo-wiki.org/branches/urban/). The progressive development of a global database of building reference footprints, which currently covers 300 cities spread across the world and is used for accuracy assessment of the different global built-up maps derived in the context of the GHSL, but also from other data providers, is also ongoing.

Future work will focus on a systematic and exhaustive consistency check of the different products and extensive validation in view of the development of sensor-specific models for deriving “actual” built-up areas from the different available built-up products.

This article showed that different validation experiments, combined with visual examples, can contribute to a better understanding of the similarities and disparities between the different built-up information layers. It also sheds light on the potential for exploiting the synergies between the sensors for consistent mapping of human settlements at regional and global scales.

ACKNOWLEDGMENTS

The authors wish to thank the GHSL team who contributed in several ways to this work mainly: Panagiotis Politis, Luca Maffenini, Sergio Freire, Daniele Ehrlich, Matina Halkia, Luigi Zanchetta and Donato Airaghi. We thank Kenji Osé from Irstea for providing us with the necessary data for the experiments. We would also like to extend our gratitude to Dr. Dušan Jovanović, from the Faculty of Technical Sciences in Serbia, for reviewing this paper.

ORCID

Filip Sabo http://orcid.org/0000-0003-0669-3939
Christina Corbane https://orcid.org/0000-0002-2670-1302
Thomas Kemper https://orcid.org/0000-0002-3446-8301

REFERENCES

Arino, O., Ramos Perez, J. J., Kalogirou, V., Bontemps, S., Defourny, P., & Van Bogaert, E. (2012). Global land cover map for 2009 (GlobCover 2009). Retrieved from https://doi.pangaea.de/10.1594/PANGAEA.787668
Aune-Lundberg, L., & Geir-Harald, S. (2010). CORINE Land Cover 2006: The Norwegian CLC2006 project. Retrieved from https://www.skogoglandskap.no/publikasjon/corine_land_cover_2006_the_norwegian_clc2006_project/publication_view
Bai, Y., Feng, M., Jiang, H., Wang, J., Zhi, Y., & Liu, Y. (2014). Assessing consistency of five global land cover data sets in China. Remote Sensing, 6(9), 8739–8759.
Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., ... Arino, O. (2008). GlobCover: Products description and validation report. Retrieved from https://www.theia-land.fr/sites/default/files/globcover_products_description_validation_report_i2.1.pdf
Bontemps, S., Defourny, P., Bogaert, E., Arino, O., Kalogirou, V., & Perez, J. (2011). GLOBCOVER 2009: Products description and validation report. Retrieved from https://mfkp.org/INRMM/article/12770349
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 3*(3), 980–1002.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20*(1), 37–46.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment, 37*(1), 35–46.

Corbane, C., Lemoine, G., Pesaresi, M., Kemper, T., Sabo, F., Ferri, S., & Syrris, V. (2017). Enhanced automatic detection of human settlements using Sentinel-1 interferometric coherence. *International Journal of Remote Sensing, 39*(3), 842–853.

Corbane, C., Pesaresi, M., Politis, P., Syrris, V., Florczyk, A. J., Soille, P., ... Kemper, T. (2017a). Mass processing of Sentinel-1 and Landsat data for mapping human settlements at global level. In *Proceedings of the 2017 Conference on Big Data from Space* (pp. 52–55). Toulouse, France.

Corbane, C., Pesaresi, M., Politis, P., Syrris, V., Florczyk, A. J., Soille, P., ... Kemper, T. (2017b). Big earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping. *Big Earth Data, 1*(1&2), 118–144.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering & Remote Sensing, 66*(7), 849–857.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., ... Dech, S. (2013). Urban footprint processor: Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience & Remote Sensing Letters, 10*(6), 1617–1621.

Eurostat. (2015). LUCAS: Land use and land cover survey. Retrieved from https://ec.europa.eu/eurostat/web/lucas.

Ferri, S., Siragusa, A., Sabo, F., Pafi, M., & Halkia, S. (2017). The European Settlement Map 2017 release, methodology and output of the European Settlement Map (ESM2p5m). Retrieved from https://ec.europa.eu/jrc/en/publication/european-settlement-map-2017-release-methodology-and-output-european-settlement-map-esm2p5m.

Ferri, S., Syrris, V., Florczyk, A., Scavazzon, M., Halkia, M., & Pesaresi, M. (2014). A new map of the European settlements by automatic classification of 2.5m resolution SPOT data. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium* (pp. 1160–1163). Quebec City, Canada: IEEE.

Florczyk, A. J., Ferri, S., Syrris, V., Kemper, T., Halkia, M., Soille, P., & Pesaresi, M. (2016). A new European remote settlement map from optical remotely sensed data. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 9*(5), 1978–1992.

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing, 29*(11), 3137–3158.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., ... Chen, J. (2013). Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing, 34*(7), 2607–2654.

Herold, M., Mayaux, P., Woodcock, C. E., Baccini, A., & Schmullius, C. (2008). Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment, 112*(5), 2538–2556.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data recommendations for the use of performance metrics. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 245–251). Geneva, Switzerland: IEEE.

Kemper, T., Mudau, N., Mangara, P., Pesaresi, M., European Commission, Joint Research Centre, & Institute for the Protection and the Security of the Citizen. (2015). *Towards a country-wide mapping & monitoring of formal and informal settlements in South Africa: Pilot-study in cooperation with the South African National Space Agency (SANSA)*. Luxembourg: European Commission Publications Office.

Klotz, M., Kemper, T., Geiß, C., Esch, T., & Taubenböck, H. (2016). How good is the map? A multiscale cross-comparison framework for global settlement layers: Evidence from Central Europe. *Remote Sensing of Environment, 178*, 191–212.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., ... Gascon, F. (2016). Sentinel-2 SEN2COR: L2A processor for users. In L. Ouwehand (Ed.), *Proceedings of the Living Planet Symposium 2016* (Vol. SP-740, pp. 1–8). Prague, Czech Republic: Spacebooks Online.

Matsuoka, M., & Yamazaki, F. (2004). Use of satellite SAR intensity imagery for detecting building areas damaged due to earthquakes. *Earthquake Spectra, 20*(3), 975–994.

Mertes, C. M., Schneider, A., Sulla-Menashe, D., Tatem, A. J., & Tan, B. (2015). Detecting change in urban areas at continental scale with MODIS data. *Remote Sensing of Environment, 158*, 331–347.
Montero, E., Van Wolvelaer, J., & Garzón, A. (2014). The European urban atlas. In I. Manakos & M. Braun (Eds.), Land use and land cover mapping in Europe (Vol. 18, pp. 115–124). Dordrecht, the Netherlands: Springer.

Ouzounis, G. K., Syrris, V., & Pesaresi, M. (2013). Multi-scale evaluation of global human settlement scenes against reference data using statistical learning. Pattern Recognition Letters, 34(14), 1636–1647.

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A., Syrris, V., & Soille, P. (2016). Assessment of the added value of Sentinel-2 for detecting built-up areas. Remote Sensing, 8(4), 299.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Carneiro Freire Sergio, M., Halkia, S., ... Syrris, V. (2016). Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. Ispra, Italy: Joint Research Centre.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. IEEE Journal of Selected Topics in Applied Earth Observation & Remote Sensing, 1(3), 180–192.

Pesaresi, M., Guo, H., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., ... Zanchetta, L. (2013). A global human settlement layer from optical HR/VHR RS data: Concept and first results. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 6(5), 2102–2131.

Pesaresi, M., Syrris, V., & Julea, A. (2016). A new method for Earth observation data analytics based on symbolic machine learning. Remote Sensing, 8(5), 399.

Pesaresi, M., Vasileios, S., & Julea, A. (2016). Analyzing big remote sensing data via symbolic machine learning (pp. 156–159). In Proceedings of the 2016 Conference on Big Data from Space. Santa Cruz de Tenerife, Spain.

Potere, D., Schneider, A., Angel, S., & Civco, D. L. (2009). Mapping urban areas on a global scale: Which of the eight maps now available is more accurate? International Journal of Remote Sensing, 30, 6531–6558.

Schneider, A., Friedl, M. A., & Potere, D. (2009). A new map of global urban extent from MODIS satellite data. Environmental Research Letters, 4, 044003.

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping global urban areas using MODIS 500 m data: New methods and datasets based on "urban ecoregions". Remote Sensing of Environment, 114, 1733–1746.

Tenerelli, P., & Ehrlich, D. (2011). Analysis of built-up spatial pattern at different scales: Can scattering affect map accuracy? International Journal of Digital Earth, 4(Suppl. 1), 107–116.

Triantakonstantis, D., Chrysoulakis, N., Sazonova, A., Esch, T., Feigenwinter, C., Düzgün, S., ... Tal, A. (2015). On-line evaluation of Earth observation derived indicators for urban planning and management. Urban Planning & Design Research, 3, 17.

Yang, Y., Xiao, P., Feng, X., & Li, H. (2017). Accuracy assessment of seven global land cover datasets over China. ISPRS Journal of Photogrammetry & Remote Sensing, 125(Suppl. C), 156–173.

Zhao, Y., Gong, P., Yu, L., Hu, L., Li, X., Li, C., ... Wang, X. (2014). Towards a common validation sample set for global landcover mapping. International Journal of Remote Sensing, 35(13), 4795–4814.

How to cite this article: Sabo F, Corban C, Florczyk A, Ferri S, Pesaresi M, Kemper T. Comparison of built-up area maps produced within the global human settlement framework. Transactions in GIS. 2018;22:1406–1436. https://doi.org/10.1111/tgis.12480