Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States

Stefan Leyk[a,*], Johannes H. Uhl[a], Deborah Balk[b], Bryan Jones[b]

[a] University of Colorado Boulder, Department of Geography, Boulder, CO 80309, USA
[b] City University of New York, Institute for Demographic Research and Baruch College, New York, NY 10010, USA

ARTICLE INFO

Keywords:
Spatiotemporal uncertainty
Global Human Settlement Layer
Open data
Data integration
Accuracy assessment
Urban land

ABSTRACT

Global data on settlements, built-up land and population distributions are becoming increasingly available and represent important inputs to a better understanding of key demographic processes such as urbanization and interactions between human and natural systems over time. One persistent drawback that prevents user communities from effectively and objectively using these data products more broadly, is the absence of thorough and transparent validation studies. This study develops a validation framework for accuracy assessment of multi-temporal built-up land layers using integrated public parcel and building records as validation data. The framework is based on measures derived from confusion matrices and incorporates a sensitivity analysis for potential spatial offsets between validation and test data as well as tests for the effects of varying criteria of the abstract term built-up land on accuracy measures. Furthermore, the framework allows for accuracy assessments by strata of built-up density, which provides important insights on the relationship between classification accuracy and development intensity to better instruct and educate user communities on quality aspects that might be relevant to different purposes. We use data from the newly-released Global Human Settlement Layer (GHSL), for four epochs since 1975 and at fine spatial resolution (38 m), in the United States for a demonstration of the framework. The results show very encouraging accuracy measures that vary across study areas, generally improve over time but show very distinct patterns across the rural-urban trajectories. Areas of higher development intensity are very accurately classified and highly reliable. Rural areas show low degrees of accuracy, which could be affected by misalignment between the reference data and the data under test in areas where built-up land is scattered and rare. However, a regression analysis, which examines how well GHSL can estimate built-up land using spatially aggregated analytical units, indicates that classification error is mainly of thematic nature. Thus, caution should be taken in using the data product in rural regions. The results can be useful in further improving classification procedures to create measures of the built environment. The validation framework can be extended to data-poor regions of the world using map data and Volunteered Geographic Information.

1. Introduction

How much we know about processes of urbanization and land conversion over the past decades depends heavily on the data that are available for analysis. In developed regions, there is usually an abundance of demographic and land use related data resulting in some confidence in analytical results at national levels for limited periods of time. However, such data are not available for most regions and countries of the world and thus global (and globally consistent) data products on population, land cover and built-up land play an important role for developing a better understanding of global dimensions of demographic key processes.

To date, existing large-scale datasets show significant limitations (e.g., Gong et al., 2013) and important differences in basic parameters such as spatial and temporal resolution, temporal coverage or thematic and regional consistency (for an overview see Grekousis et al., 2015). The most common drawback among all these data products is the persistent lack of knowledge of their spatial and thematic (Strahler et al., 2006) as well as spatio-temporal (Tsutsumida and Comber, 2015) accuracy and thus their fitness for different uses by the research user community. In addition, to date, there is a lack of consistent collection of global information across time impeding reporting efforts related to activities as required by the post-2015 Development Agenda (United Nations, 2012). It is well-known that classifications of built-up land or developed land suffer from lower levels of accuracy in less developed regions and rural settings (Smith et al., 2002; Wickham et al., 2013).

[*] Corresponding author.
E-mail addresses: stefan.leyk@colorado.edu (S. Leyk), johannes.uhl@colorado.edu (J.H. Uhl), deborah.balk@baruch.cuny.edu (D. Balk), bryan.jones@baruch.cuny.edu (B. Jones).

http://dx.doi.org/10.1016/j.rse.2017.08.035
Received 9 February 2017; Received in revised form 10 August 2017; Accepted 25 August 2017

The main reason for this inherent property is the poor performance of remote sensing based classifiers if the data are of rather coarse spatial and spectral resolution in relation to the object of interest, which can result in mixed pixel effects. Such effects are amplified by the use of natural construction material and smaller rooftop surfaces in less developed and rural regions as compared to more developed regions. Population data that is often used as an ancillary variable to improve such classifications is usually available at relatively coarse resolution, and if population distributions are to be created the demographic data is usually allocated to grid cells classified as developed or populated. Because of the lack of validation data for global land cover classifications in general (Zhao et al., 2014) such procedures introduce and propagate uncertainty, and this inherent uncertainty has rarely been quantified, modelled or evaluated, thoroughly. This shortcoming is even more impactful for large geographic areas and in less developed regions, as well as for earlier time periods since uncertainty is expected to be higher under all these conditions. This is considered a serious limitation of such databases because there is very limited understanding of the reliability of estimates and statistics derived from such data and this evokes potential misuse of the data. Some examples of datasets exposed to these issues are GlobCover (Bontemps et al., 2011), GlobLand30 (Chen et al., 2015) the Gridded Population of the World (GPW) and the related Global Rural Urban Mapping Project (GRUMP), which are snap-shots in time on population counts (Balk et al., 2005, 2006; Deichmann et al., 2001; CIESIN, 2005), Landscan (Dobson et al., 2000), which represents the ambient population per grid cell, the WorldPop project (Sorichetta et al., 2015) and the recent Global Human Settlement Layer (GHSL) (Pesaresi et al., 2015) which represents a first attempt to account for global built-up areas from decametric-resolution satellite data using a consistent multi-temporal data classification approach. Klotz et al. (2016) propose a cross-comparison framework to assess the accuracy of some of the above mentioned data products including the GHSL and the Global Urban Footprint dataset (GUF) (Esch et al., 2013) within European settings and for one point in time.

It is thus of great importance to derive further in-depth knowledge of the accuracy of such global datasets to better understand how confident researchers can be in using them in different regions of the world for demographic analysis, land cover change modeling or the characterization of relevant demographic processes such as urbanization. Focusing on built-up land layers, this urgent research need resonates with three main goals of the current study. *First*, exemplified for the United States, a validation dataset will be constructed through data integration procedures incorporating public parcel records and building footprints from several regions that can be used as an extensive sample for a national assessment at different points in time. *Second*, an analytical framework for accuracy assessment will be developed that allows evaluating multi-temporal spatial datasets on built-up or developed land at the national, continental and potentially global scale using such validation datasets. This framework will include a sensitivity analysis to address potential spatial offsets and differences in class definition. Most importantly, this framework will also support accuracy assessment within strata defined by different development intensities (i.e., loosely related to rural, peri-urban and urban land) to better understand relationships between development intensity and classification accuracy. *Third*, we will demonstrate the operationality of the developed analytical framework using the validation data layers constructed and testing the GHSL for the selected U.S. counties as our target dataset. This experiment will shed light on the efficacy of the framework to assess classification accuracy and its variation across the study regions, at different points in time, and across regions of varying development intensities.

Validation of multi-temporal data such as the GHSL is complex and challenging due to difficulties in creating historical versions of the test data at fine spatial resolution, issues of spatial and thematic inconsistency, and temporal mismatches between the validation data and target data. The GHSL is exemplified as the target data in this study because it represents a promising new public data product at fine spatial resolution with extensive temporal coverage of > 40 years (Pesaresi et al., 2015). This data product is already in high demand for multiple target applications in different disciplines including population projections (Linard et al., 2017), disaster management and risk assessment (Freire et al., 2015, 2016), as well as land use change modeling (Small and Sousa, 2016). However, the analytical framework is intended to be applicable as a general protocol for accuracy assessment to other datasets with properties and embedded class abstraction comparable to the GHSL and other regions. Furthermore, accuracy assessment results derived for one region may also provide first insights on expected accuracies in other countries in which validation data may be difficult to obtain or are non-existent. Such an accuracy assessment framework will not only enable the analyst to carry out more reliable analyses but will also present a pathway to future improvements of such data products through the detection of problematic regions or contexts which may need particular attention, different classification procedures or the inclusion of additional ancillary data.

2. Data and preprocessing

2.1. Global Human Settlement Layer (GHSL)

The Global Human Settlement Layer (GHSL) project aims to assess the human presence on the planet through analysis of evidences as collected from earth observation data, census data and volunteered geographic information. The GHSL information is shaped by a scalable abstraction schema overcoming the traditional land cover paradigm (Pesaresi et al., 2009) and is currently structured in three basic geo-information layers of increasing abstraction and decreasing spatial resolution: A) built-up areas, B) population grids, and C) the settlement classification model. In Pesaresi et al. (2013) the GHSL information production workflow targeting the "built-up area" class abstraction was defined and the automatic recognition was tested for a large set of sensors in the spatial resolution range of 0.5–10 m. These sensors may perform very well in detecting built-up areas using textural and morphological image-derived descriptors as input to the automatic classification but they are typically constrained regarding data access and processing and redistribution rights, which makes the scientific use of the derived products difficult or unsustainable. Moreover, they are typically available only for more recent years, and acquired in rather scattered ways for arbitrary points in time, which makes these data difficult to use for uniform and systematic information extraction and analysis of global, regional or even national trends.

In order to mitigate some of these issues, the GHSL built-up area recognition system was ported in the decametric resolution, open remote sensing data domain and tested with global collections of image data records collected by the Landsat satellite platform in the past 40 years (Pesaresi et al., 2016a). In the first edition, the GHSL was made available as seamless global mosaic at high spatial resolution (approx. 38 m) and for various epochs (1975, 1990, 2000, 2014, see Fig. 1). The GHSL satellite-derived information production technology is based on symbolic machine learning (SML), a supervised data classification inspired by methods for DNA microarrays data analysis used in biomedical informatics for the clustering of gene expressions (Pesaresi et al., 2016b) in large datasets. Current efforts at the Joint Research Center focus on the production of GHSL versions with improved temporal and spatial resolution using Sentinel1 and Sentinel2 satellite imagery.

2.2. Parcel data

Open cadastral and tax assessment data have become increasingly available to the public – often in GIS-compatible format – for several regions in the U.S. (von Meyer and Jones, 2013) and other countries. For this study, all publicly available parcel data that could be obtained through open-sources were used. Cadastral parcel boundaries are

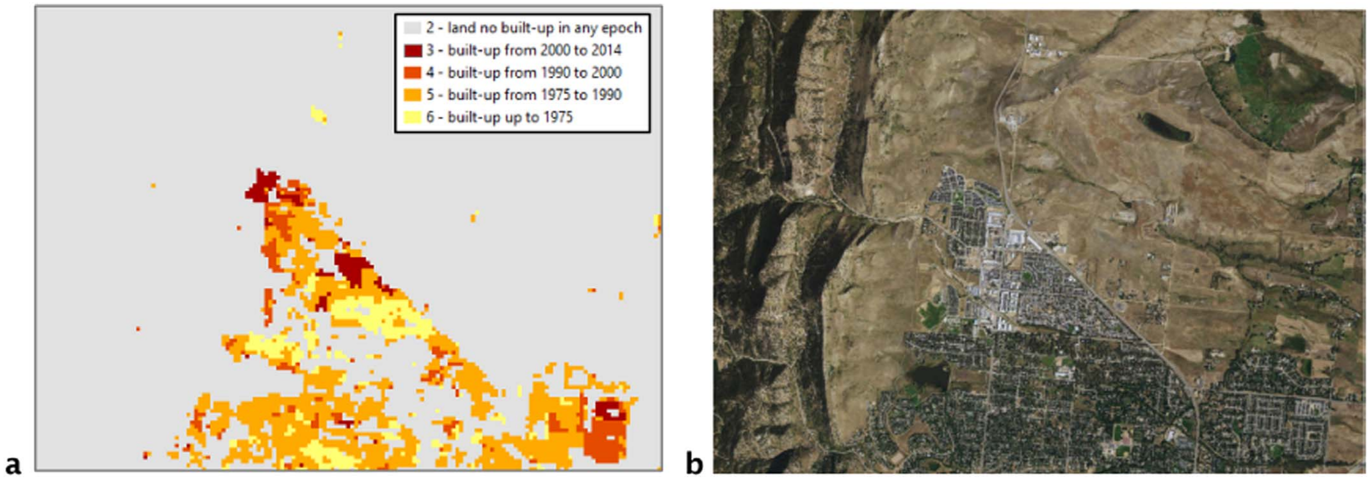

2 - land no built-up in any epoch
3 - built-up from 2000 to 2014
4 - built-up from 1990 to 2000
5 - built-up from 1975 to 1990
6 - built-up up to 1975
a
b

Fig. 1. (a) GHSL built-up land identified for four time periods from 1975 to 2014 and (b) corresponding satellite image acquired in 2015 (Source: ESRI) for a subset of Boulder County (CO). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

typically acquired using terrestrial or navigation technology-based land surveying methods. Parcel data usually contain rich attribute information related to the type of land use, which can be inconsistent across counties, characteristics of the structure and the year when a structure in a parcel has been established (built year). This allows the creation of snapshots of built-up parceled land for any point in time with a temporal resolution of one year. While a valuable data source for modeling areas of residential development (Leyk et al., 2014) and population estimates for small areas (Tapp, 2010) and different points in time (Zoraghein et al., 2016), the use of parcel data can be challenging. Such data are still difficult to obtain for most areas, can suffer from attribute inconsistencies and processing them can be computationally expensive (Manson et al., 2009). Furthermore, there is some significant size variation in land parcel units, especially across urban-rural trajectories, which needs to be taken into account if parcels are used as analytical units (Leyk et al., 2013).

2.3. Building data

Data representing building footprints are typically derived from LiDAR measurements or digitized based on aerial imagery. With increasing investments in detecting and mapping existing structures, building data are becoming increasingly available as open data and such open datasets are used in this study to spatially refine the snapshots of built-up parceled land. This refinement is expected to be especially effective in rural areas where parcel units can have large areal extents and are expected to drastically overestimate built-up land if they remain unrefined. For some administrative regions in the U.S. (31 counties) in which parcel data are available, also building data could be accessed (Fig. 2 for some examples). Since this study uses spatial data that are publicly available, the described database can be expanded in the near future to incorporate additional study areas in the U.S. and other countries.

2.4. Study areas for demonstrating the assessment framework

As described above, selection of the study areas is driven by the availability of the reference data. We systematically collected data from areas in which both parcel and building data were available to create a large validation dataset that can be used to quantify the accuracy of built-up land layers in different regions of the U.S. and thus under different landscape conditions, vegetation types, settlement histories, etc., and for different points in time. This sample represents an extensive basis to assess data accuracy, objectively, across the United States and hence at the national to continental scale. In total, we collected data for 31 counties (Fig. 2; see Table A1) summing to an area of 45,014 km^2, and containing 6,362,429 parcel features across the country. These 31 counties show great variation in areal extent and urban character as well as in the rate of development over time (Fig. 3). While it can be assumed that these official cadastral and building data have high quality standards, detailed accuracy information is not available for most of these datasets.

3. Methods

In this section, *first*, the creation of the validation dataset will be described. The parcel and building data have to be integrated in order to produce fine resolution representations of built-up land (i.e., at the structure level) that, at the same time, carry the temporal information (i.e., the built-year attribute found in parcel records). This spatial join process can be challenging due to data volume, inconsistency and inaccuracy or incompleteness of parcel records and building features. This integrated data product has to be converted to raster format and encoded, properly, in order to be compatible with the built-up land layer to be evaluated.

Second, the validation data and built-up land layers for different points in time are spatially overlaid and compared. We create confusion matrices (see Fig. A1) as is common in classification accuracy assessments and compute various accuracy measures for each study area as well as across all areas some of which allow adjusting for imbalanced data distributions. This accuracy assessment will also include a sensitivity assessment in order to examine potential effects due to possible spatial offsets between data layers and varying definitions of the abstract concept of "built-up land" in the case of the GHSL. This accuracy assessment will also be done for different strata of development intensity in order to examine the accuracy of built-up land layers such as GHSL within urban, peri-urban and rural settings, which will provide important insights on the usability, limitations and suitability of such data products under different conditions related to landscape development and population density. Finally, regression models are created to examine the relationship between spatially aggregated GHSL-derived built-up quantities and those computed from the reference data. This will shed light on the effects of residual spatial misalignment on the assessment and provide a basis for more robust quality assessments in regions not covered by validation data.

3.1. Creating multi-temporal validation data layers

3.1.1. Integration of parcel and building data and consistency tests

In order to create spatially refined parcel information, parcel data

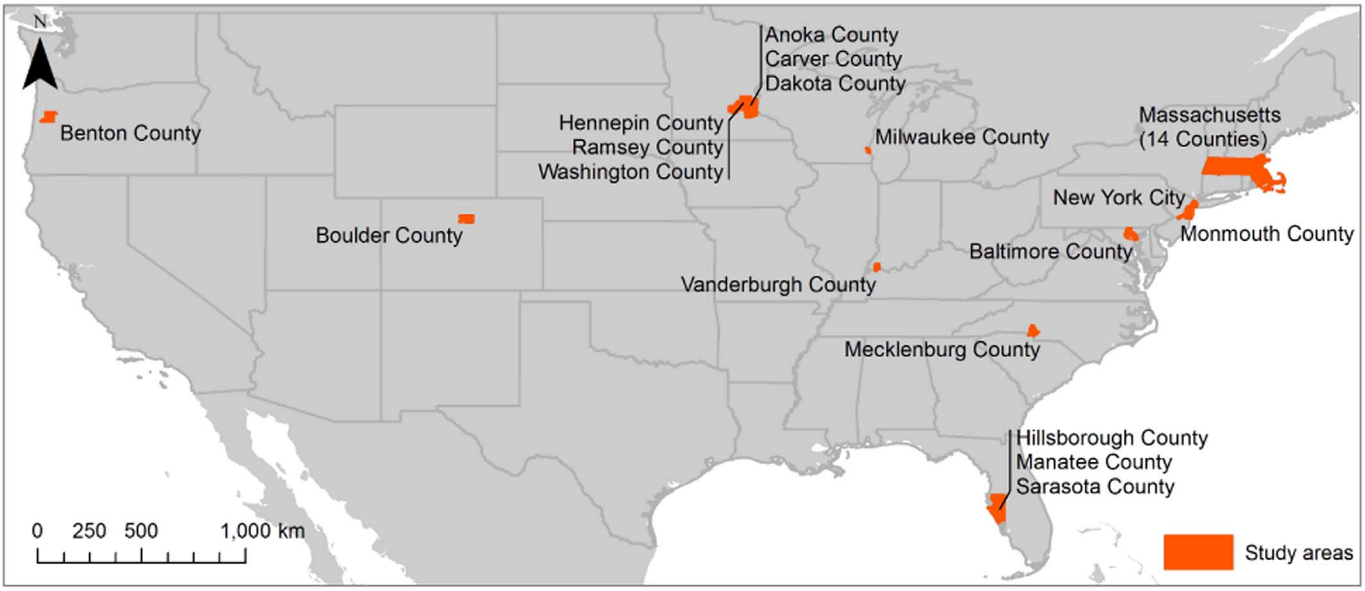


Fig. 2. Study areas in the U.S. where parcel records including built year information and building footprint data are publicly available. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

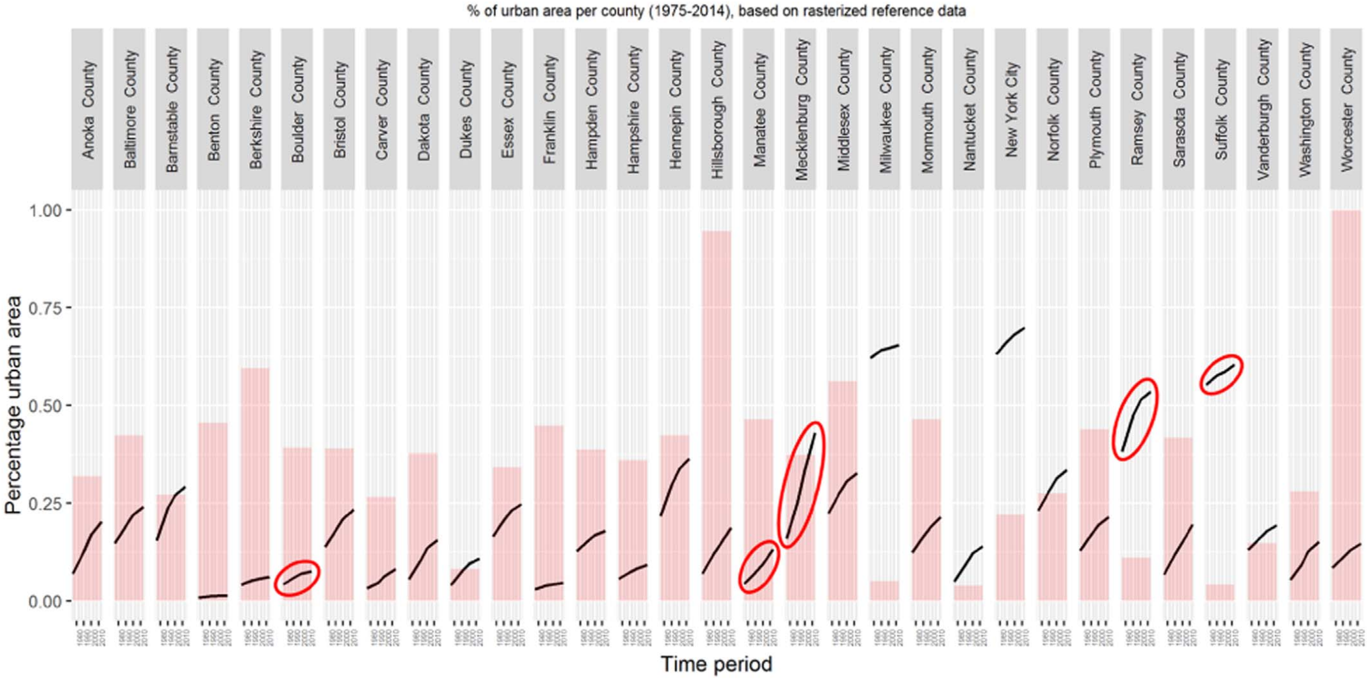


Fig. 3. Characterization of all counties used as study areas. Light red columns indicate the relative areal extents of the counties [0,1] in relation to Worcester County. The black lines illustrate the proportion of built-up land and its change for each county between 1975 and 2014. Red ovals indicate the five counties for which Table 2 presents detailed results; they cover different scenarios of development intensity and its change over time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and building footprints have to be spatially integrated. For this process to be objective, topological relationships – such as rules of containment – between parcel and building objects have to be defined, computed and examined. This process is impeded due to different data acquisition methods applied and specific geometric and topological characteristics of the different datasets as well as possible n:m relationships between building and parcel objects.

Different types of bidirectional spatial join methods were tested to carry out the data integration step across rural and urban parts of some of the study areas in which parcel size and building densities are expected to vary significantly. This includes spatial joins based on: (a) tests of *containment of building centroids* in parcel polygons, (b) the *largest overlapping area* of a building polygon with adjacent parcel polygons, and (c) tests of *complete containment* of building features within parcel polygons.

The spatial join methods append parcel attributes (e.g., presence of a structure, built year) to the joined building feature(s). Typically, such semantic relationships between spatial objects are established in two steps: 1) data matching and 2) data linking. Data matching is performed by spatially joining parcel and corresponding building objects based on geometrical (e.g., minimum size) and topological (e.g., containment) criteria. Data linking implicitly transfers the unique identifiers and

Table 1
Evaluation tests of selected spatial join methods (shown for a subset of Boulder County, CO).

Spatial join method	Maximum number of parcels to be joined	Spatially joined parcels	Correctly joined parcels ($r \le 1$)	Maximum omission error [%]	Correctly joined rate [%]
A: building centroid-based	1585	1367	1290	13.8	94.4
B: area majority	1585	1362	1292	14.1	94.9
C: complete containment	1585	1085	1083	31.5	99.8

other attributes of the matched (or source) objects (parcel features) to the attribute table of the target objects (building features) and thus allows you to retrieve building features associated with parcel-level attribute information.

In order to evaluate the performance of the different spatial join methods, the summarized area of the buildings joined to a parcel was computed and appended to the parcel and building objects. The ratio of summarized building areas in relation to the parcel area r was calculated and used as a test measure to examine plausibility of the join. A join is considered plausible if the aggregated area of k buildings associated with a parcel does not exceed the area A_{Parcel} of the parcel itself:

$$r = \frac{\sum_{1}^{k} A_{Building\ i}}{A_{Parcel}} \le 1$$

The centroid-based method and area majority-based method show similar robustness and relatively low maximum omission errors (parcels not joined out of the maximum number of parcels that potentially can be joined to a building; Table 1). The similarity in performance indicates that problems due to buildings that may be split between more than one parcel can be neglected. However, the centroid-based method is preferred as it is based on a simple point-in-polygon query for object-matching and thus computationally more efficient. This is highly relevant given the data volume to be processed across all 31 counties. A visual check of the data ensured there was no systematic misalignment between parcels and buildings that could cause dramatic uncertainties. The complete containment-method does not consider buildings that overlap more than one parcel polygon which leads to a high rate of correct joins but results in a high maximum omission error.

Beside the plausibility test, an additional consistency check was carried out to filter out locations that could not successfully be evaluated. Parcels without built-year information that were joined to buildings as well as parcels with built-year information that were not joined to any building (no existing building footprint) were removed from the validation dataset and these areas were excluded from the subsequent accuracy assessment. This cross-comparison of the two independent datasets increases reliability of the integrated data product (Fig. 4). We created a binary raster layer (exclusion layer) in which grid cells were labeled 1 if they were located in parcel units that did not pass the plausibility test or failed the consistency check above and 0 otherwise. The cells with value 1 were excluded from the accuracy assessment (Figs. 4b and 5).

3.1.2. Creation of reference surfaces

Since GHSL data is available in raster format at a spatial resolution of approximately 38 m, a raster-based accuracy assessment is reasonable and straight-forward. Therefore, the building objects containing built year information are used to generate GHSL-compatible raster data. Here, compatibility includes three aspects: temporal, thematic, and spatial.

Temporal compatibility: The built year information (i.e., the year the current structure has been established) stemming from the parcel data is converted and encoded to match the temporal categories of the dataset to be validated, here the GHSL (class 2: land not built-up, classes 3–6: land built-up from 2000 to 2014, 1990–2000, 1975–1990 and before 1975, respectively). For simplification purposes, years are used as temporal cut-offs. However, it has to be kept in mind that the built-up area for a given epoch in GHSL is derived from an image collection acquired within a certain time frame (for example, the built-up area for epoch 1990 is based on images gathered between 1985 and 1994).

Thematic compatibility: The definition of the abstract class "built-up land" as used in the GHSL is simulated for pixel value assignment during the raster conversion process. According to Pesaresi et al. (2016a), a pixel is considered built-up in the classification process if at least one structure overlaps the pixel area. Consequently, any overlaps between vector building objects and GHSL raster cells have to be identified and labeled as built-up land. To do so, the building objects are first rasterized to very fine spatial resolution of 2 m using

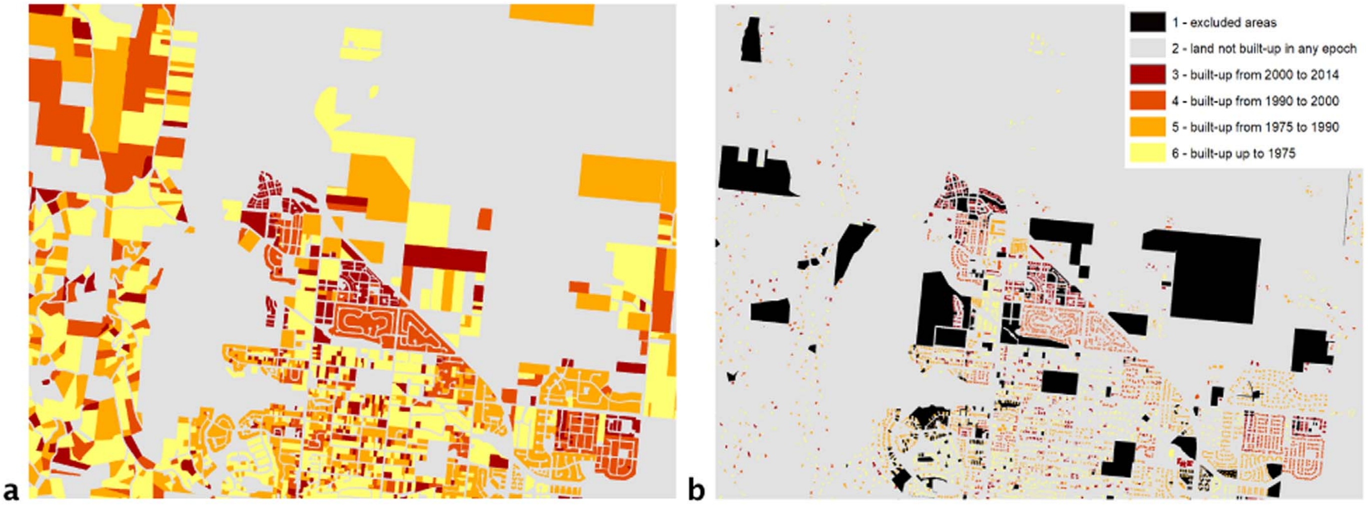

a
b
1 - excluded areas
2 - land not built-up in any epoch
3 - built-up from 2000 to 2014
4 - built-up from 1990 to 2000
5 - built-up from 1975 to 1990
6 - built-up up to 1975

Fig. 4. (a) Parcel records (area features) showing built-year, (b) parcel-based reference data refined by building footprints after the plausibility and consistency test for a subset of Boulder County (Colorado). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

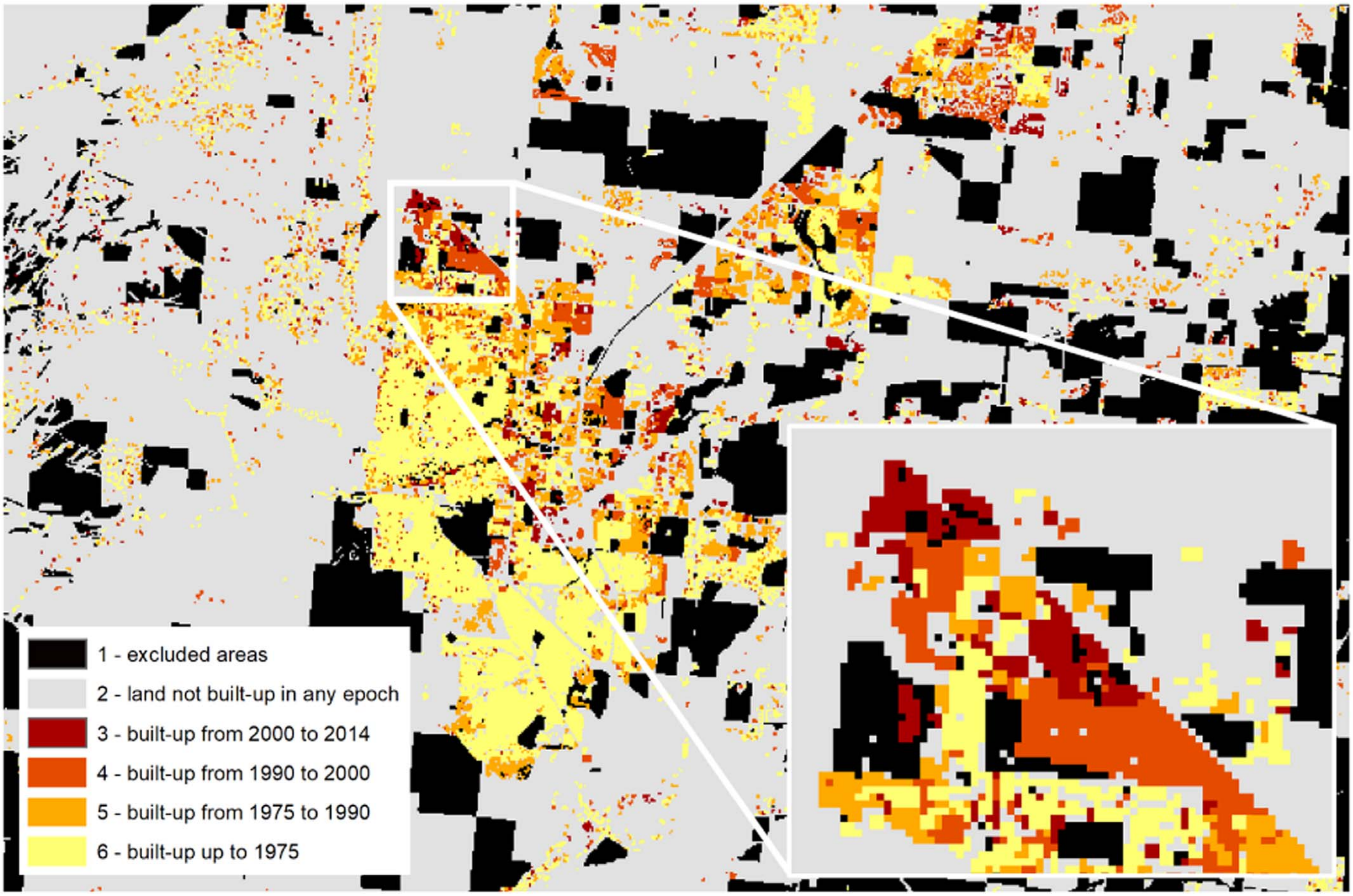

1 - excluded areas
2 - land not built-up in any epoch
3 - built-up from 2000 to 2014
4 - built-up from 1990 to 2000
5 - built-up from 1975 to 1990
6 - built-up up to 1975

Fig. 5. Spatially refined reference dataset rasterized and resampled to GHSL resolution for a subset of Boulder County (CO). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the GHSL-matched built-year class as raster value. If any of these 2 m resolution cells intersect with the extent of a GHSL raster cell, this cell is labeled the same built-year class. The resolution of 2 m is considered a good trade-off between maintaining characteristic building outline features and computational efficiency.

Spatial compatibility: The intermediate raster dataset is then aggregated and resampled to the GHSL cell extents, creating a GHSL compatible reference surface called $GHSL_{ref}$ thus maintaining the spatial resolution and registration properties of GHSL. If a raster cell extent in $GHSL_{ref}$ contains at least one 2 m resolution cell encoded as built-up land, the $GHSL_{ref}$ cell will be classified as built-up. If there are 2 m cells with different temporal labels within the $GHSL_{ref}$ cell extent, the oldest built-up category will be used for $GHSL_{ref}$ cell assignment. If there are no 2 m cells with classes 3–6 within the $GHSL_{ref}$ cell extent it will be assigned the not built-up class (class 2). Fig. 5 shows an example of the converted 38 m resolution reference surface $GHSL_{ref}$ based on spatially integrated parcels and building footprints encoded with GHSL-matched temporal categories.

3.2. *Multi-temporal accuracy assessment & sensitivity analysis*

3.2.1. *Confusion matrices for computing accuracy measures*

The GHSL raster and the created reference surfaces are compared pixel-wise to build confusion matrices, which allow the derivation of various accuracy measures (Fielding and Bell, 1997; Nguyen et al., 2009) for different time periods and each of the study areas. The derived accuracy measures include Kappa Coefficient of Agreement (Kappa) (Cohen, 1960), which does not adjust for imbalanced data distributions, and three class-independent measures that do adjust for such imbalances. These are the Normalized Mutual Information Criterion (NMI) (Forbes, 1995), the F-measure, which is the harmonic mean of recall and precision (Fawcett, 2006), and the geometric mean (G-mean), which takes into account sensitivity and specificity (Kubat and Matwin, 1997). Evaluating accuracy using different error measures allows the analyst to understand trends, class-specific deviations as well as the nature of inherent agreement or disagreement and provides rich opportunities for in-depth interpretation of the results. Each GHSL epoch is evaluated cumulatively (pre-1975, pre-1990, pre-2000 and pre-2014) in order to characterize the behavior of accuracy measures over time. The workflow for the data integration, conversion and accuracy assessment is illustrated in Fig. 6.

3.2.2. *Sensitivity analysis to account for positional and thematic uncertainty*

It is important to examine possible spatial mismatches between reference and test data as such discrepancies may bias the assessment results. Positional uncertainty could occur in the reference data itself due to data acquisition methods (digitization from maps) or distortions in aerial photographs used as source data, or may be introduced during the rasterization and aggregation process. Positional uncertainty can also be present in the satellite imagery used to create the built-up land layers due to possible shifts of raster cells during resampling, registering and reprojection processes. Such misalignments between reference and test data can have dramatic impacts in areas of low built-up density. To examine the effect of such discrepancies a sensitivity analysis of the assessment results will be performed by incorporating systematic offsets between reference data and the built-up layer of 19 m and 38 m (dimensions of one half and one GHSL pixel) in each of the eight main directions (Fig. 7d). The accuracy assessment is carried out for each of these scenarios. Note, that due to complex geocoding and ortho-

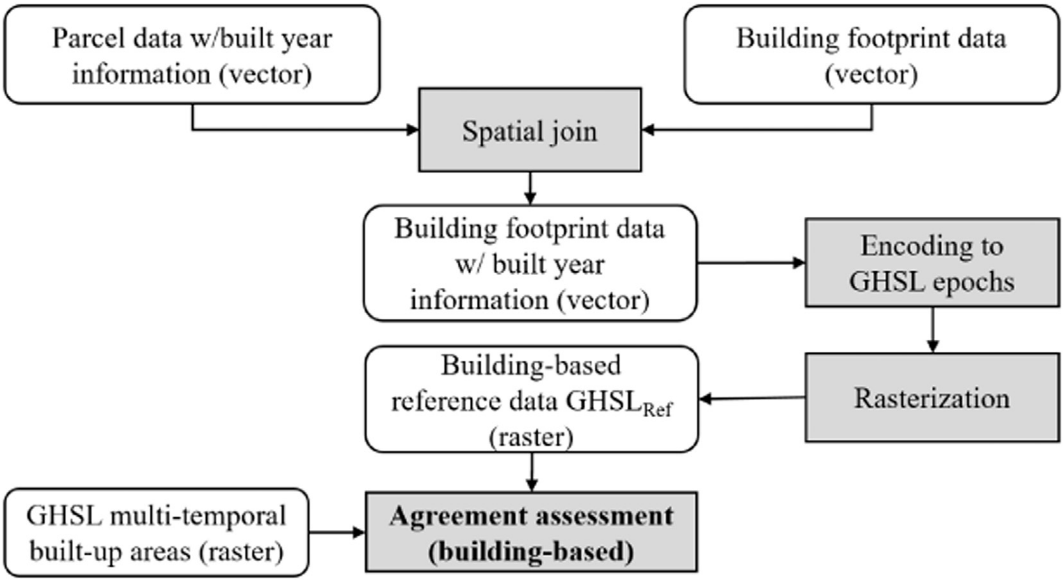

Parcel data w/built year information (vector)
Building footprint data (vector)
Spatial join
Building footprint data w/ built year information (vector)
Encoding to GHSL epochs
Rasterization
Building-based reference data GHSL_{Ref} (raster)
GHSL multi-temporal built-up areas (raster)
Agreement assessment (building-based)

Fig. 6. Workflow for accuracy assessment of GHSL built-up areas at different points in time.

rectification processes there could be non-stationary or even random positional mismatches between test and validation data, which can influence accuracy measures in fine resolution data. This sensitivity analysis assumes stationarity in positional mismatches within one county due to the considerable size of the study areas.

One additional crucial point in this context is the definition of built-up land which relates to the inherent thematic uncertainty in the GHSL impacted by possible mixed pixel and aggregation effects (Horwitz et al., 1971; Detchmendy and Pace, 1972) during classification and interlaces with positional uncertainty. As described above, to comply with GHSL specifications, the baseline assessment assumed that a $GHSL_{ref}$ pixel is considered and labeled built-up if it overlaps with a building object (i.e., overlap threshold > 0%, see Fig. 7a). To test for more conservative rules of overlap (i.e., larger proportions of the building have to overlap the GHSL pixel extent), this threshold is systematically increased to up to 40% of the $GHSL_{ref}$ pixel area. This means that $GHSL_{ref}$ pixels, which have an overlap with building area less than that threshold, are not considered as built-up (Fig. 7b, c). These

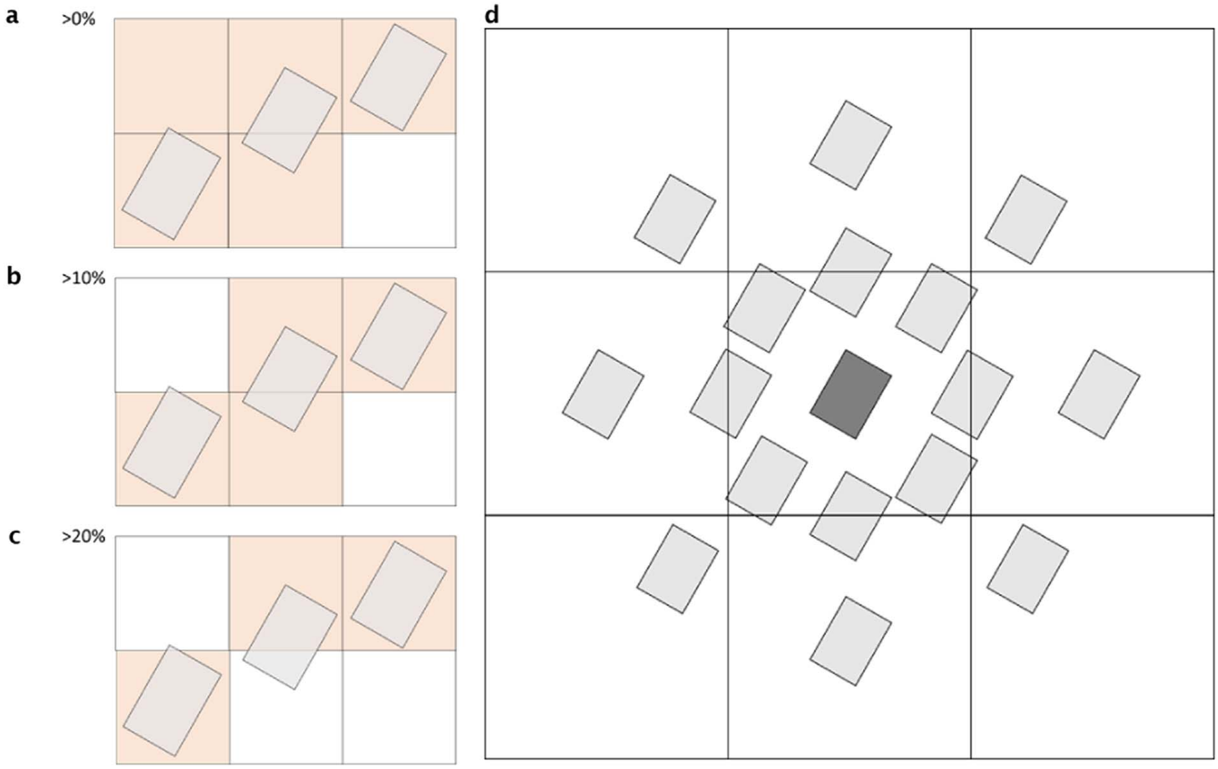


Fig. 7. Schematic illustration of the sensitivity analysis including changing overlap thresholds between building footprint and GHSL cell extent used to label the cell "built-up" in $GHSL_{ref}$. (a: > 0% overlap means 5 cells will be labeled built-up, b: > 10% overlap results in 4 cells built-up; c: > 20% overlap results in only 3 cells labeled built-up), and spatial offset between building footprints and GHSL cell (by 19 m and 38 m) in each of the main directions (d). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracy measures based on the comparison between GHSL and the reference data for five different counties (see the remaining 26 counties in Table A1) and across all test regions.

Test region	GHSL class	F-measure	G-mean	Kappa	NMI	Sensitivity	Specificity
Boulder County, Colorado	Not built-up	0.972	0.614	0.495	0.318	0.993	0.379
	< 2015	0.519	0.614	0.495	0.318	0.379	0.993
	≤ 2000	0.519	0.619	0.495	0.309	0.387	0.992
	≤ 1990	0.507	0.621	0.486	0.291	0.389	0.991
	< 1975	0.380	0.508	0.366	0.231	0.259	0.996
	All epochs	0.579	0.595	0.467	0.293	0.481	0.870
Manatee County, Florida	Not built-up	0.930	0.865	0.589	0.338	0.889	0.842
	< 2015	0.655	0.865	0.589	0.338	0.842	0.889
	≤ 2000	0.631	0.876	0.582	0.347	0.838	0.915
	≤ 1990	0.629	0.889	0.591	0.370	0.847	0.933
	< 1975	0.583	0.855	0.559	0.344	0.761	0.961
	All epochs	0.686	0.870	0.582	0.347	0.836	0.908
Mecklenburg County, North Carolina	Not built-up	0.735	0.640	0.317	0.079	0.794	0.516
	< 2015	0.576	0.640	0.317	0.079	0.516	0.794
	≤ 2000	0.484	0.597	0.255	0.052	0.446	0.799
	≤ 1990	0.438	0.581	0.282	0.071	0.386	0.874
	< 1975	0.398	0.571	0.300	0.089	0.355	0.919
	All epochs	0.526	0.606	0.294	0.074	0.499	0.780
Ramsey County, Minnesota	Not built-up	0.684	0.715	0.443	0.151	0.640	0.800
	< 2015	0.757	0.715	0.443	0.151	0.800	0.640
	≤ 2000	0.748	0.720	0.448	0.154	0.794	0.653
	≤ 1990	0.731	0.722	0.446	0.155	0.795	0.655
	< 1975	0.605	0.673	0.345	0.089	0.629	0.721
	All epochs	0.705	0.709	0.425	0.140	0.732	0.694
Suffolk County, Massachusetts	Not built-up	0.613	0.668	0.480	0.283	0.453	0.983
	< 2015	0.840	0.668	0.480	0.283	0.983	0.453
	≤ 2000	0.827	0.678	0.476	0.253	0.967	0.476
	≤ 1990	0.820	0.676	0.467	0.242	0.963	0.474
	< 1975	0.796	0.675	0.440	0.197	0.934	0.488
	All epochs	0.779	0.673	0.469	0.251	0.860	0.575
All test regions (counties)	Not built-up	0.912	0.767	0.562	0.267	0.914	0.644
	< 2015	0.650	0.767	0.562	0.267	0.644	0.914
	≤ 2000	0.622	0.749	0.542	0.253	0.607	0.924
	≤ 1990	0.617	0.753	0.547	0.263	0.608	0.933
	< 1975	0.558	0.709	0.503	0.239	0.528	0.952
	All epochs	0.672	0.749	0.543	0.258	0.660	0.873

different proportional overlaps are built-in to the above sensitivity analysis to develop a full understanding of how sensitive the accuracy assessment will be to different spatial offsets in different directions as well as varying “criteria” to identify built-up land. We test all these scenarios for three counties and evaluate the sensitivities of the resulting accuracy measures to establish a well-defined protocol for objective assessment of cumulative built-up land for all test counties in each epoch. Similar simulative approaches have been proposed by Glick et al. (2016) and Foody (2010).

3.2.3. *Accuracy assessment in strata of different development intensities*

Remote sensing derived classes of developed or impervious land tend to be underestimated in rural settings (Wickham et al., 2013; Leyk et al., 2014), especially in earlier time periods due to issues of spatial resolution and mixed pixel effects which impede detection of individual structures in remote, isolated locations. One important question in this study is how accurate data on multi-temporal built-up land are in different parts of the landscape that are shaped by different settlement histories and development intensities. Therefore, the study areas will be stratified based on the level of built-up density and in each of these strata the accuracy assessment will be carried out, separately. Knowing how classification accuracy varies with changing development intensity at different points in time will be useful for users to understand data quality of the built-up land layer and propose appropriate ways to apply the data in different settings as well as to account for inherent uncertainty. Stratified accuracy assessment has been proposed for example by Congalton (1991), Wulder et al. (2006) and recently by Olofsson et al. (2013).

Based on building centroids derived from the vector reference layers and the building area, continuous density surfaces of built-up area per km^2 are computed for each epoch using a point density function for different radii between 200 m and 500 m implying different analytical scales that are providing reasonable local density estimates. Different thresholds will be applied on this continuous density surface to identify three strata of built-up density in all test counties (e.g., > 5%, 0.5%–5% and 0–0.5%) that may be loosely linked to urban, peri-urban and rural settings, respectively. For each test county and each epoch a confusion matrix will be built and accuracy measures will be computed for each of these strata. There is no theoretical guidance on the choice of appropriate density thresholds to reflect the level of urbanness. Therefore, threshold values will be systematically shifted to identify sets of density cutoffs that allow for optimal differentiation of accuracy measures related to these three strata across all test counties. Such thresholds can be further investigated to understand their meaningfulness for data quality assessment.

3.2.4. *Regression analysis to evaluate statistical GHSL$_{ref}$-GHSL relationships*

A simple linear regression analysis is carried out in order to examine how well GHSL can estimate built-up land as derived from the reference data. Spatially aggregated built-up land proportions are computed for whole blocks of 10 × 10 pixels (approximately, 380 × 380 m) for both data layers. Based on these aggregated built-up land quantities, linear regression models are created for each point in time and each built-up density stratum within each county as well as across all counties. This regression analysis is useful because it will: (1) mitigate potential spatial misalignment issues and help separate between thematic and spatial uncertainty; (2) provide a robust statistical framework to estimate this relationship across different built-up density strata; and (3) provide a basis to quantify systematic bias and derive parameters that could be typical for the intensity stratum considered or the sensor or extraction method applied. Thus the results could potentially be useful for applying the GHSL in other regions to compensate for uncertainty or underestimation.

4. Results & discussion

The results presented in this section are used to demonstrate the assessment framework described and are reported in relation to the 31 counties in the United States, embracing a variety of landscape and development conditions. Table 2 presents the results of the baseline accuracy assessment over all 31 tested counties as well as for five individual counties for each GHSL epoch. These five counties were selected because they represent different typical combined characteristics with regard to areal extent, proportion of urban development, and rate of development growth over time (Fig. 3). The results for the remaining 26 counties can be found in Table A1.

4.1. Aggregate multi-temporal accuracy measures over all test regions

Kappa, NMI, F-measure and G-mean of the built-up land layer summarized over all epochs (cumulative built-up land) and including the non-built-class have values of 0.543, 0.258, 0.672 and 0.749, respectively. When breaking down these measures to different GHSL epochs it can be seen that there is some variability in these measures (e.g., F-measure varies from 0.56 for the epoch ≤ 1975 to 0.65 for ≤ 2014) illustrating that different time periods have different levels of accuracy, possibly, a direct result of improving data quality in remote sensing products over time. The MSS sensor onboard of early Landsat generations used for the 1975 epoch in the GHSL data provided a source resolution of 68 × 83 meters and had 4 spectral bands coded in 64 levels (4 bit data) as compared to 15 m/30 m resolution, 11 spectral bands and 12 bit data provided by the Landsat 8 OLI sensor used for the epoch 2014. Thus, the GHSL for the epoch 1975 was produced through oversampling the coarser resolution source data, which may explain the drop in classification accuracy for the earliest epoch. Furthermore, specific conditions at the time of image acquisition (e.g., haziness or cloud cover) can influence the accuracy measures. Fig. 8 shows the temporal trends of computed accuracy measures for built-up land across all test regions (blue line) and illustrates that most accuracy measures have an increasing trend over time.

4.2. Region-specific multi-temporal accuracy measures

As shown in Table 2 (which includes 5 example counties) and Table A1, and schematically illustrated in Fig. 8, there is considerable variation in each of the accuracy measures across test regions, regardless if these account for class-specific characteristics, chance agreement or imbalanced class proportions. For example, in land built-up until 1990, Kappa ranges from 0.28 in Mecklenburg County to 0.59 in Manatee County. NMI ranges from 0.07 in Mecklenburg County to 0.37 in Manatee County.

In general, lower accuracy appears to be loosely linked to lower proportions of developed areas within the corresponding county (more rurality) but also to the sensor technology available at the time as described above. This can be seen in Fig. 9, which shows the temporal trends of Kappa, NMI, F-measure and G-mean for the five counties included in Table 2. The grey line in each of these plots illustrates the level of (urban) development in each epoch derived from proportions of built-up land in the corresponding reference data. These plots illustrate a direct link between changes in development intensity and classification accuracy in different epochs. Increasing development is clearly related to increasing classification accuracy. This observation may be explained by the fact that remote sensing classification using medium resolution imagery performs better in regions that are more developed i.e., proportionally less isolated development and larger contiguous patches of developed area which reduces the well-known mixed pixel problem. This might also relate to spatial mismatches between reference and test data that would affect the accuracy assessment particularly in areas of low development density, which is dominated by small patches of built-up land. Overall, as more development can be observed over time, the more gain in classification accuracy can be expected.

4.3. Sensitivity of accuracy measures

For each study region (i.e., county) 136 different scenarios were run to examine two spatial offsets for each of the main directions as well as 8 different thresholds of overlap between reference buildings and $GHSL_{ref}$ pixel extents. As a general outcome, it was found that the main driver for sensitivity in the results is the proportion of overlap used to label a $GHSL_{ref}$ pixel as built-up. As can be seen in Fig. 10, exemplified for Kappa computed for one county, the spatial offset and changes in offset direction both result in minor deviation in accuracy, while changes in overlap thresholds show significant effects and in this case result in decreasing accuracy when compared to the benchmark scenario (i.e., no offset and overlap threshold > 0% to identify built-up land).

The level of sensitivity varies between accuracy measures, across counties and between epochs. However, the main pattern remains the same across the study areas: changing overlap thresholds to label built-up land dominates and results in decreasing accuracy for global accuracy measures though very few exceptions were found in which measures would first increase and then decrease. G-mean shows a reverse trend, which indicates some increasing imbalance effects due to such changes in overlaps. Thus overall, the baseline scenario for accuracy assessment (built-up if > 0% overlap, no offset in any direction) can be seen as an operational setting, indicating that data co-registration between reference data and GHSL is acceptable and an overlap of > 0% between building footprints and $GHSL_{ref}$ cell extents is indeed the most sensitive criterion for defining built-up land (Fig. 11). Therefore, the following stratified accuracy assessment was done using these baseline parameters. However, to fully reflect the limitations of this test, the reader should be cautioned that, first, even this baseline scenario might suffer from spatial mismatches, which can affect the accuracy assessment, particularly in low-density development areas. Second, stationarity of existing mismatches is assumed within each county, which may hide some potential non-stationarity effects.

4.4. Accuracy assessment results stratified by development intensity

The results of the accuracy assessments carried out separately in areas that can be related to low, medium and high intensity developed land (referred to as rural, peri-urban and urban strata below) (Fig. 12) are shown in Figs. 13 and 14 for two different threshold sets (low-threshold and high-threshold scenarios), respectively. All results shown are based on a 200 m radius used in the point density function. Using radii of 400 m or greater resulted in lower levels of distinction between strata as illustrated in Fig. 12, possibly an indication that development intensity can be better understood as a highly localized process.

Using different thresholds to distinguish between levels of development intensity results in deviating patterns across all counties but uncovers a general trend. Recall that there is no theoretical basis or definition related to meaningful thresholds for built-up density that would indicate certain levels of urbanization that could be used in this assessment. Furthermore, each county is expected to have different

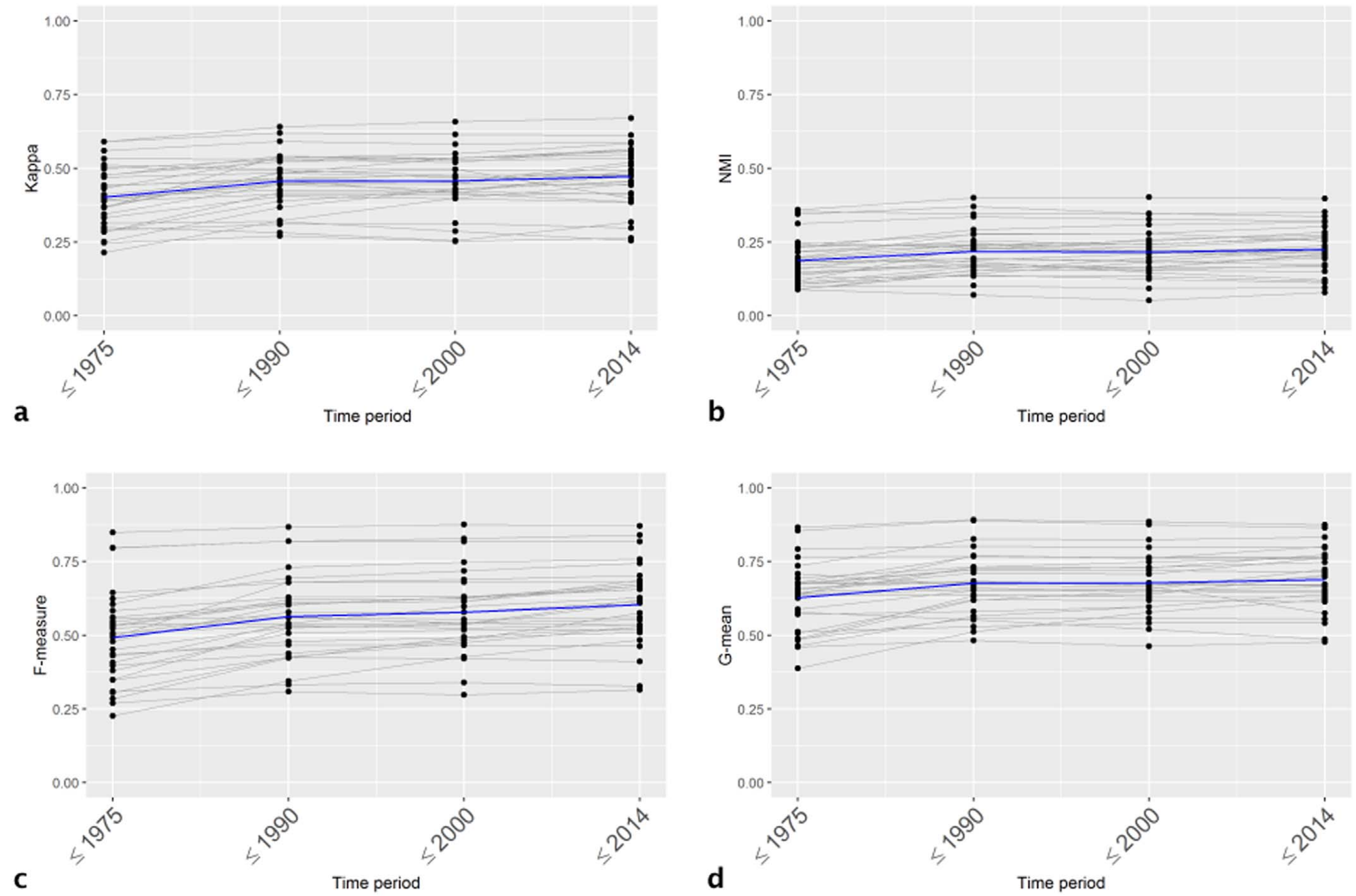


Fig. 8. Temporal trends of the accuracy measures computed for cumulative built-up land (≤ 1975 to ≤ 2014) plotted for each study region (grey thin lines) and as average measures over all study regions (blue thick line): (a) Kappa, (b) NMI, (c) F-measure, (d) G-mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

optimal thresholds depending on individual settlement patterns and development history. Nevertheless, the iterative evaluations described here can be seen as a starting point to establish relationships between classification accuracy of built-up land layers and development intensity across all counties. Fig. 13 shows different accuracy measures over time for all 31 counties and as average for the low-thresholds scenario (0–0.5% for rural, 0.5–5% for peri-urban, > 5% for urban land) and illustrates a relatively clear separability between the different strata. Fig. 14 shows the results for the high-thresholds scenario (0–2% for rural, 2–9% for peri-urban, > 9% for urban land) for comparison and to illustrate changes in separability. Below the individual accuracy measures are described in more detail.

4.4.1. F-measure

F-measure levels in different strata in the *low-thresholds scenario* (Fig. 13) are highly distinct with high values in the urban stratum (> 5% density). F-measure focuses on the positive minority class (built-up land) in imbalanced data distributions; since both Precision (positive predictions that are correct = User's Accuracy) and Recall (positives that are correctly detected = Producer's Accuracy) are defined with respect to that positive class and result in high values, F-measure shows high classification accuracy in relatively dense built-up areas. In the low-intensity developed stratum F-measures are extremely low documenting one of the major problems in remote-sensing based classification, namely, the mixed pixel problem. Whether these results are also affected by spatial misalignment between reference and test data will be evaluated in the regression analysis described below. The peri-urban stratum shows higher spread of values but remains well separated from the rural and urban strata. In the *high-thresholds scenario* (Fig. 14) the separability of F-measures between strata is still clear but values in the rural and urban strata are more spread.

4.4.2. Kappa

Kappa values show a combined pattern of trends of traditional Producer's Accuracy (PA) and User's Accuracy (UA). Average Kappa values in the *low-thresholds scenario* (Fig. 13) appear to be well distinct between strata but there is considerable overlap between county-level measures, particularly between peri-urban and urban strata. A possible explanation for these overlaps is that the mixed pixel problem but also the spatial offset effects as described above are not only related to the calculated built-up density but also to the local pattern or clustering of structures, which is not reflected in the density measures, and therefore PA and thus Kappa values could regionally vary, considerably. Furthermore, it can be expected that the large proportions of non-built-up land result in imbalanced class distributions and may affect Kappa calculations. Separability of Kappa values between strata further decreases in the high-thresholds scenario (Fig. 14).

4.4.3. G-mean

G-mean is described as a balanced accuracy measure that is commonly used if classification performances for both minority and majority classes in imbalanced data distributions are of interest. The

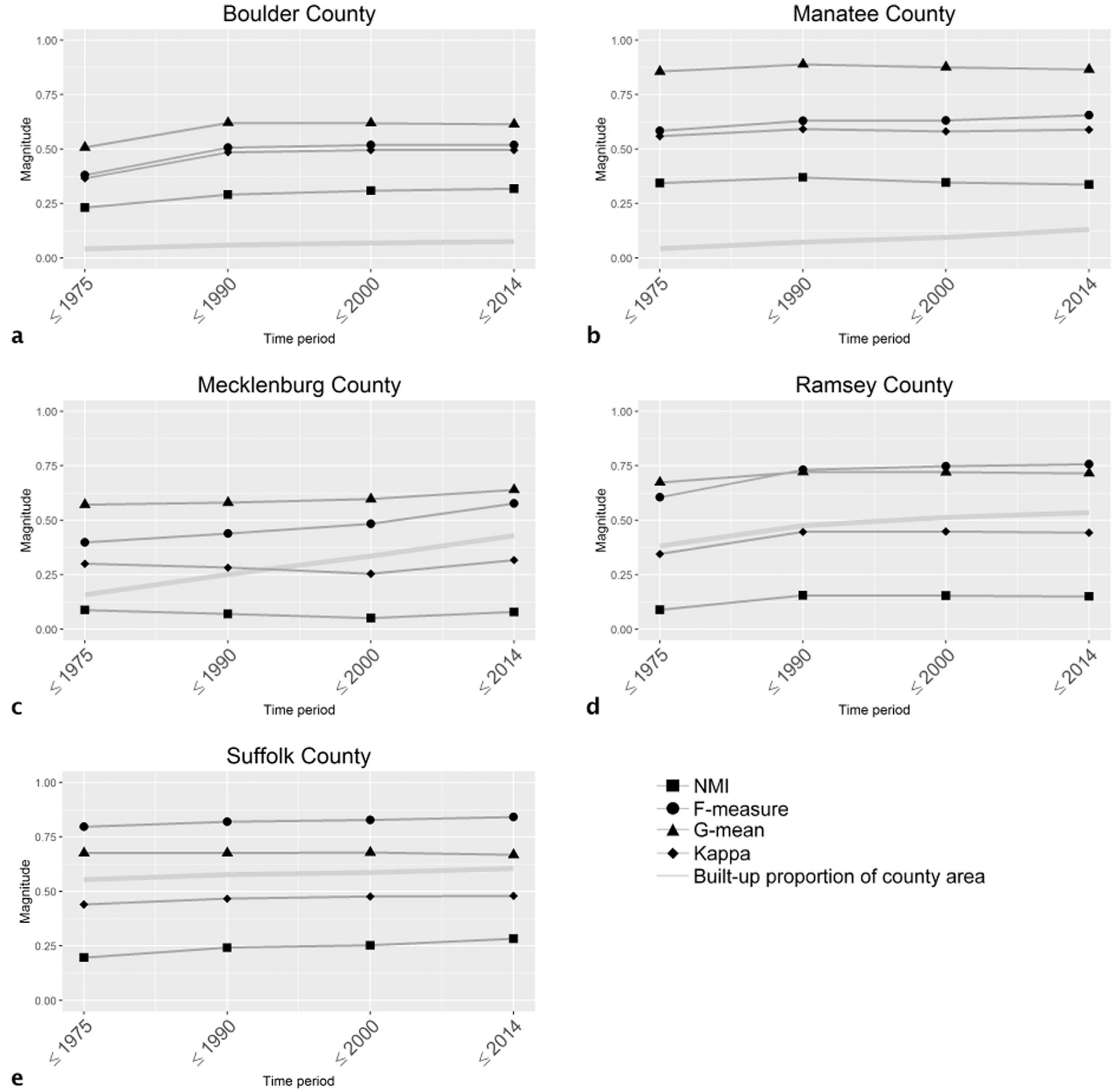


Fig. 9. Temporal trends (1975–2014) of selected county-specific accuracy measures (Kappa, NMI, F-measure and G-mean, see Table 2) computed for cumulative built-up land classes: (a) Boulder County, (b) Manatee County, (c) Mecklenburg County, (d) Ramsey County, (e) Suffolk County. Grey lines in each plot indicate the proportion of developed land in each individual county.

metric takes Sensitivity (recall) and Specificity (accuracy of the majority class) into account. Because of the balancing effect (one class performance can balance the other one) the measures in the different strata appear to be more similar when compared to Kappa or F-measure and show higher levels of overlap. The mean values are well distinguished in the low-threshold scenario (Fig. 13) but very similar between medium and high density strata in the high-threshold scenario (Fig. 14). Furthermore, the high level of spread in the rural stratum in both threshold scenarios (Figs. 13 and 14) raises some concerns about the usefulness of balancing performance in majority and minority classes.

4.4.4. NMI

NMI represents the most conservative measure with some adjustment for the presence of a dominating class. The trend of the mean values is similar between scenarios but there is some overlap between individual strata basically reflecting the trends observed for Kappa but in a more condensed way.

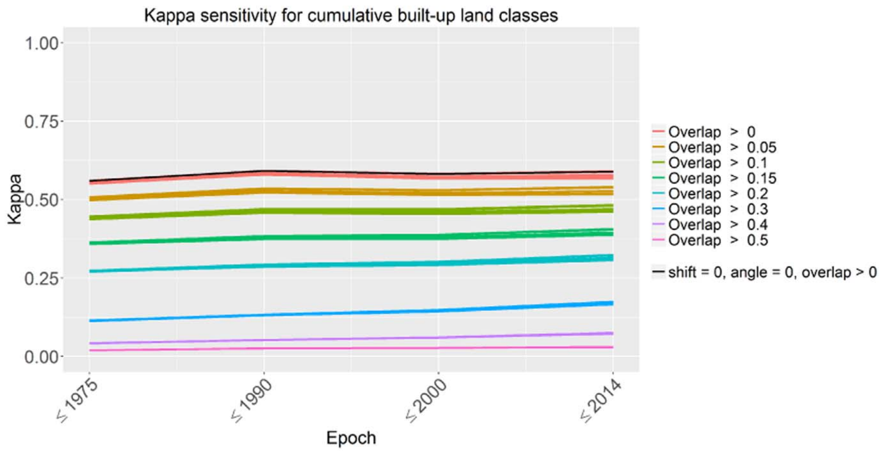


Fig. 10. Detailed results of the sensitivity analysis exemplified for Manatee County in relation to the baseline accuracy assessment (black line) exemplified for Kappa. Spatial offset (here 19 m) and changing directions result in "bundled" plot lines for each overlap threshold at each point in time. On the other side, changes in overlap have significant effects on the accuracy measures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4.5. *Accuracies at individual points in time*

While the above results refer to the general trends in each stratum, Figs. 13 and 14 also illustrate that the separability of county-level accuracy measures is higher for individual points in time. Average accuracy generally improves over time for all measures but shows steeper rates of improvement in the urban stratum for Kappa and NMI in the *low-thresholds scenario*. This observation may be related to the improvement of satellite technology and imagery over time. However, again, there is a high level of variation among counties which can be related to unexpected data quality measures such as cloud cover or environmental variability, which could be responsible for lower accuracy measures also in later time epochs compared to earlier ones.

4.5. *Regression analysis results*

Table 3 presents the regression analysis results. As can be seen, the linear regression coefficients (slope in Table 3) for estimating built-up land across all counties are highly significant and very similar for all

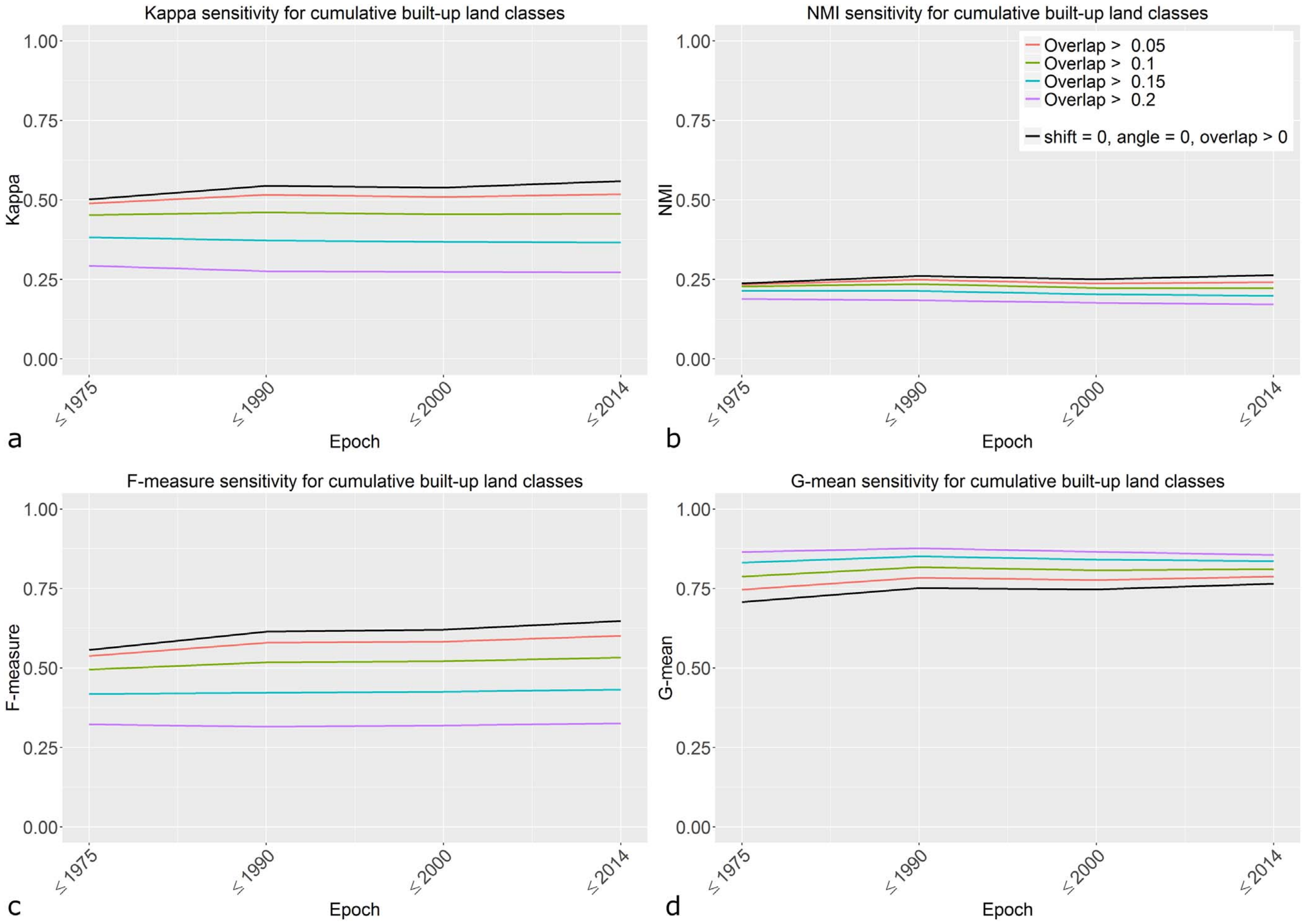


Fig. 11. Results of the sensitivity analysis over all counties over time in relation to the baseline accuracy assessment; the plots show the resulting accuracy measures for increasing overlap thresholds between reference data (fine-resolution building footprints) and GHSL cell extents to label $GHSL_{ref}$ cells as built-up: (a) Kappa, (b) NMI, (c) F-measure, (d) G-mean. The results for spatial offsets in different directions show minor effects and are not included. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

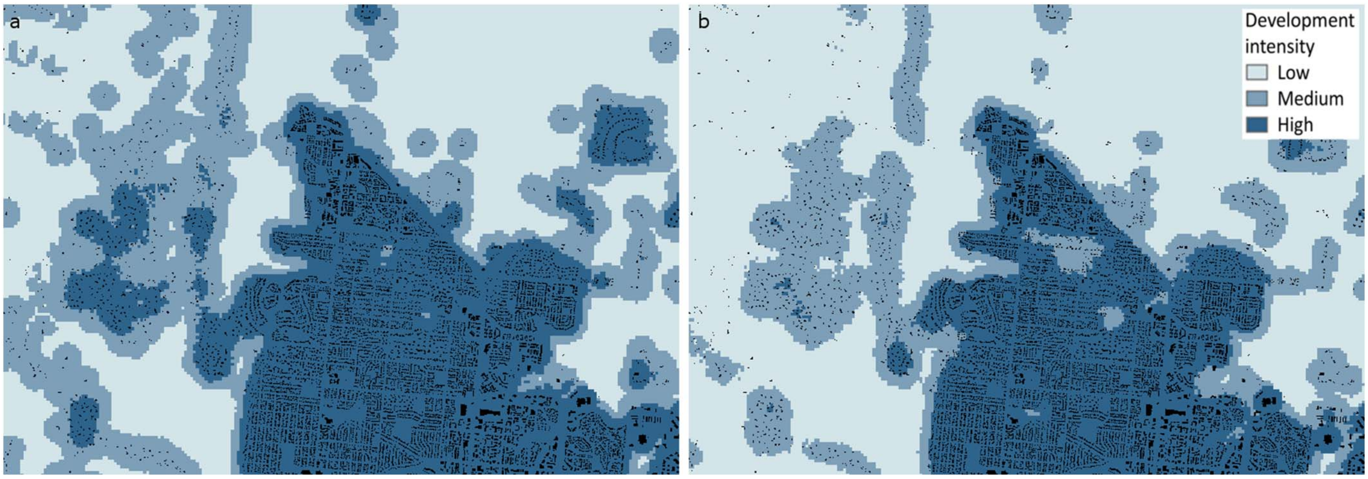


Fig. 12. Development intensity raster and corresponding vector reference data for 2014 using a 200 m radius with applied thresholds for low, medium, and high development intensity (a) for the low threshold scenario (0.5% and 5.0%) and (b) for the high threshold scenario (2.0% and 9.0%). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

epochs (0.93–0.95) indicating that overall GHSL seems to be a highly reliable predictor for built-up land derived from the reference database. However, as suspected there are high levels of non-stationarity hidden by such a global model. When this regression analysis is carried out for the different development strata a different picture is revealed. For the medium and high density strata the coefficients increase from epoch < 1975 to epoch < 2014 and indicate relatively robust relationships (0.63–0.84 for medium density; 0.78–0.84 for high density). For the low density stratum regression coefficients are very small positives (0.03–0.06) and occasionally switch sign for different counties indicating a weak statistical relationship and thus low explanatory power of GHSL for built-up land.

Overall, these results confirm the outcomes from the classification accuracy assessment but add some important new insights. First, since these models are based on aggregated raster cell blocks (10 × 10 pixels), the low classification accuracy in the low-density

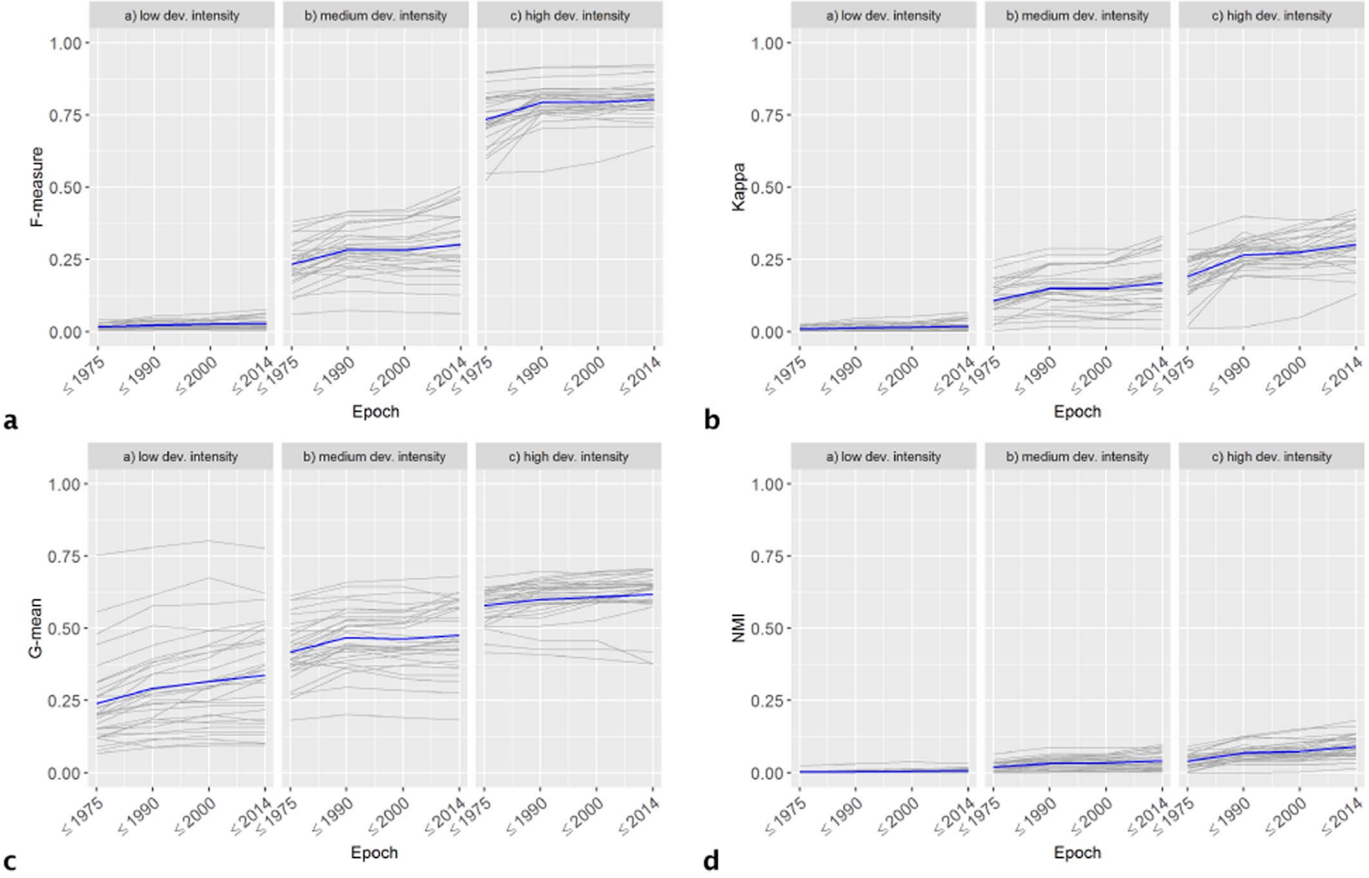


Fig. 13. Multi-temporal accuracy assessment across all 31 counties for different strata (defined by development intensity) computed using point density functions with 200 m radius and building area density as attribute. The thresholds used here are: 0–0.5% for low intensity (rural), 0.5–5% for medium intensity (peri-urban) and > 5% for high intensity (urban) developed land. The blue thick line is the average measure in the corresponding stratum. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

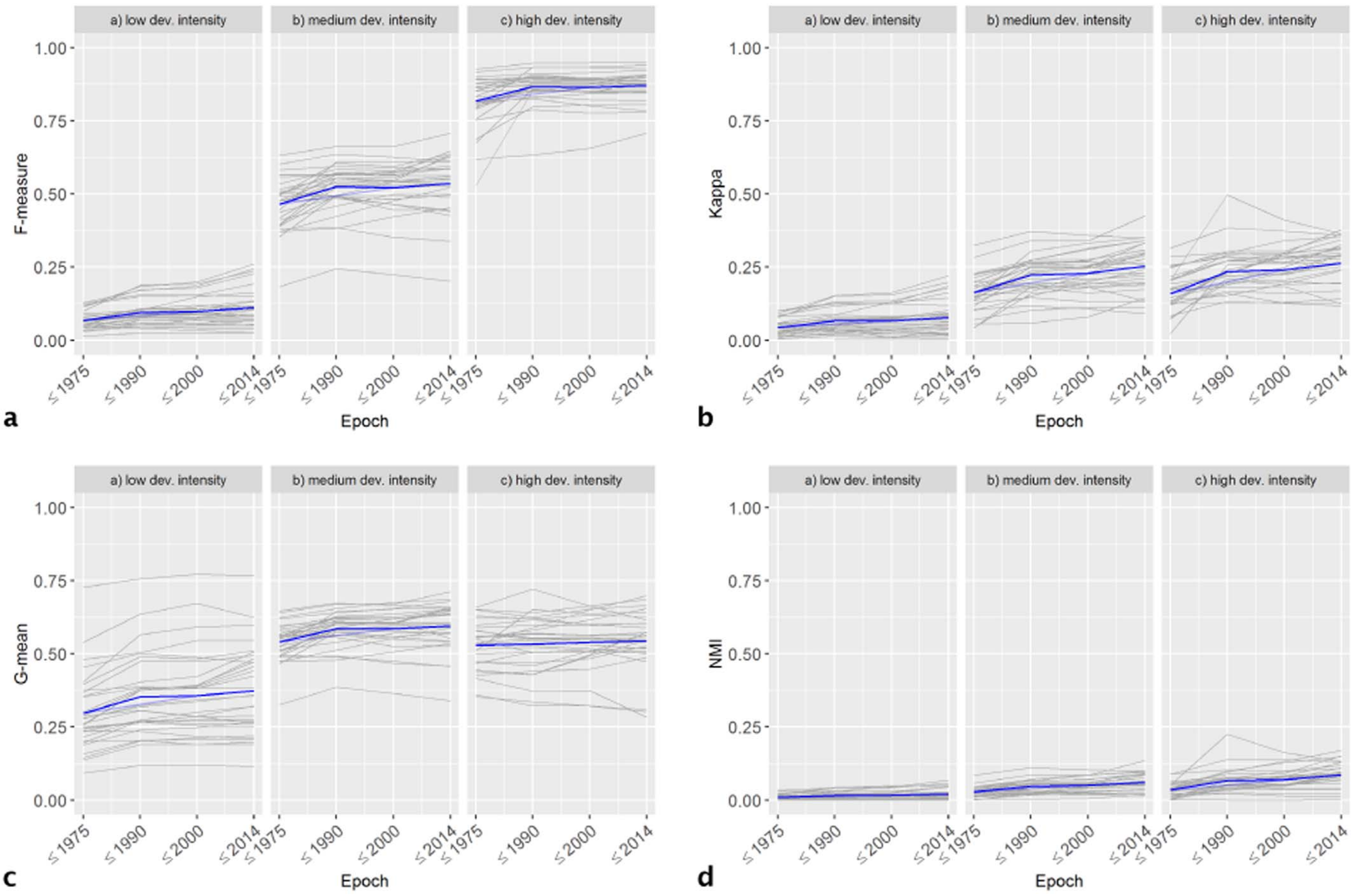


Fig. 14. Multi-temporal accuracy assessment across all 31 counties for different strata (defined by development intensity) computed using point density functions with 200 m radius and building area density as attribute. The thresholds used here are: 0–2% for low intensity (rural), 2–9% for medium intensity (peri-urban) and > 9% for high intensity (urban) developed land. The blue thick line is the average measure in the corresponding stratum. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Results of the regression analysis; all coefficients and intercepts show are statistically significant (alpha = 0.001).

Epoch	Coefficient	All counties, all strata	All counties, rural	All counties, periurban	All counties, urban
< 2014	Intercept	0.04	0.01	0.08	0.05
	Slope	0.93	0.03	0.84	0.84
< 2000	Intercept	0.04	0.01	0.07	0.05
	Slope	0.94	0.04	0.79	0.83
< 1990	Intercept	0.03	0.01	0.06	0.04
	Slope	0.95	0.06	0.75	0.83
< 1975	Intercept	0.03	0.01	0.07	0.06
	Slope	0.95	0.04	0.63	0.78

development stratum, described above, appears to be relatively unaffected by spatial misalignment between reference data and test data. Thus, the classification error appears to be mainly of thematic nature indicating that the target data layer has to be used with care in rural settings. Second, the regression coefficients for medium-density and particularly high-density settings are very robust and may provide a basis for defining sensor- and extraction method specific parameters to estimate built-up land reliably using GHSL in regions for which no validation data exist. Furthermore, it can be seen that all models have positive intercepts which indicates some systematic underestimation of the target data source for built-up land which complements above interpretations.

5. Final evaluation & concluding remarks

This study had three main objectives. First, an analytical framework for multi-temporal accuracy assessment of built-up land layers for different points in time and for different development-based strata was created using publicly available integrated cadastral and building data (see Section 5.1). Second, in order to demonstrate this framework, the "built-up area" class abstraction as included in the first release of the GHSL has been evaluated for selected counties in the United States (see Section 5.2). Third, the framework has been presented as a unique way to establish relationships between the intensity of development (built-up land density) and the accuracy in the target dataset (Section 5.2).

5.1. A validation framework

Creating a *validation database* that can be used for multi-temporal accuracy assessment of a built-up land data product is a challenging task as high spatial resolution and temporal information are required to be able to carry out the analysis. In the case of the U.S., parcels, which in rural areas are often very large in relation to the built-up portion, often carry temporal information indicating the year an existing structure has been built. The integration of parcels with building footprints results in an effective spatial refinement and allows for the creation of unique fine-resolution reference data for different points in time. It is important to cross-compare parcel records and building footprints to identify gaps in either dataset which have to be excluded from any validation effort. Comparable validation data can be created where similar data are available and accessible. The built-year attribute is subject to some uncertainty related to structures torn down and rebuilt or the most current extension to a structure, which is difficult to assess. One very time-consuming way to do such an assessment would be the cross-comparison with historical map sheets to verify the existence of a building at a certain point in time.

Converting vector data to fine-resolution (2 m) raster data and *resampling* them to coarser resolution such as 38 m to match the GHSL resolution change the data structure and introduces aggregation effects. This can result in systematic mismatches and offsets between the validation data and the target data, which is often underestimated in such data processing efforts. The *sensitivity analysis* described in this study examines such possible effects by incorporating spatial offsets in different directions, which also addresses recently reported problems with some accuracy measures (Pontius and Millones, 2011). This analysis also included tests of changing criteria (% overlap between building footprints and GHSL cell extents) to identify built-up land. In this study, the baseline scenario (no offset, overlap > 0%) showed the best performance. While this procedure cannot fully assess residual spatial disagreement in the baseline scenario the regression results provided evidence that the main factor for classification error is of thematic nature. However, this may be different for other regions, countries or datasets and therefore this sensitivity analysis should be an essential component of an accuracy assessment. The resulting reference surfaces ($GHSL_{ref}$) can be created for any point in time and any spatial target resolution, which enables the analyst to carry out unique multi-temporal validation experiments to better understand the data quality of the dataset of interest at different points in time and for different regions.

5.2. Demonstrating the validation framework for different development-derived strata using GHSL as target data

Overall, the GHSL data product shows high levels of accuracy across the selected U.S. counties compared to accuracies published for developed or impervious land layers in national (e.g., Wickham et al., 2013; Homer et al., 2015) or global datasets (e.g., Gong et al., 2013; Esch et al., 2013) though the authors caution that these outcomes are valid only for the selected counties, and comparisons are difficult since the target classes are defined and abstracted in different ways. Comparisons between global datasets that represent compatible abstractions will be carried out in the near future. As expected, accuracies in the GHSL increase over time as a direct result of improved sensor technology and imagery in more recent points in time reducing prominent issues such as limited spatial or spectral resolution that relate to the mixed pixel effect, especially in isolated rural regions. Deviations from this trend can be caused by environmental variations (Maclaurin and Leyk, 2016a), unexpected cloud cover or haziness. The assessment results for the 31 individual counties used in this study reveal a picture of varying data quality depending on above conditions as well as the proportion of developed land and development rate (i.e., increase of built-up land over time) at the county level. Fast growing counties appear to have higher accuracies in later points in time. In contrast, counties with small proportions developed land (rural settings) tend to have lower classification accuracies because of higher proportions of isolated structures that are difficult to detect using data with resolutions between 15 × 15 m and 68 × 83 m. While these associations were expected, the described analytical framework allows the analyst to put them in numbers. These results are encouraging for using GHSL for estimates of settlement activity in other regions of the world where no validation data exist and at the same time allow better communication of the expected data quality in different settings. However, it has to be noted that, similar to any other existing dataset, the pixel-level use of GHSL does not provide reliable estimates of built-up area for rural regions. The low accuracy measures in these regions are mainly due to misclassification errors as could be confirmed by the regression analysis, but might also be affected by spatial misalignment between reference and test data to some minor degree. Therefore, it is recommended to use the data in more aggregate forms (e.g., census units) to mitigate such quality issues in rural areas.

The *stratified accuracy assessment* results provide unique insights into associations between development intensity (or built-up density) and classification accuracy. These results are valuable as they not only allow one to quantify uncertainty in different strata that can loosely be related to rural, peri-urban and urban land, but they also can be used to communicate expected over- or underestimation under different conditions, which is essential for correct use of the data by the user community and can be useful in future efforts to improve the GHSL in different regions.

Both underestimation and overestimation in built-up land are very high in rural settings. However, as mentioned before, GHSL has not been designed for mapping purposes at large cartographic scales but rather for deriving statistical estimates at higher aggregation levels. In settings of higher development intensity GHSL shows high accuracies, even at the pixel level, and can be seen as one of the most reliable global, open and free data available to estimate built-up area (similar to the GUF data product for one point in time) and identify changes in urban land over time. Nevertheless, the results for lower intensity developed land indicate that issues of misclassification, imbalances in spatial distributions as well as spatial misalignments between test and reference data need to be addressed, before this data product can be used for the study of key processes such as urbanization. For this reason, future efforts will focus on improvements of the current GHSL version in rural and peri-urban settings through information extraction approaches (e.g., Maclaurin and Leyk, 2016b; Uhl and Leyk, 2017) that will increase the detection rate of isolated built-up entities.

5.3. Final remarks

The presented analytical framework for accuracy assessment of multi-temporal built-up land layers such as the GHSL provides a strong foundation for evaluation and improved quantification of expected data quality over large geographic extents but also for different regions and underlying conditions. These results can be used to instruct the user community on how to use the data and what the limitations will be. Future research will also tap into questions of critical scale for accuracy assessments to further improve the use such data layers.

Future analysis would benefit from closer examination of the spatial patterns of the relationships observed here, for example, to determine whether the patterns of low-density development and the measures of accuracy of peri-urban and rural areas are dependent on the spatial proximity to higher density land or major urban centers.

Integrated parcel data and building footprints represent a valuable validation database. However, there are uncertainties inherent in these data (Zoraghein et al., 2016) and such data layers are not available everywhere. Different data sources may be useful to attempt similar data integration procedures including extracted data from topographic maps (Uhl et al., 2017), Volunteered Geographical Information (VGI) that could be combined with administrative or remote sensing data or map-derived records as well as already existing validation datasets (Tsendbazar et al., 2015; Zhao et al., 2014). Furthermore, road networks could be incorporated into such assessments to evaluate the target data with and without road information. This will be the focus of future research within the population studies research community.

Acknowledgements

Research reported in this publication was, in part, supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD066613. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was also funded, in part, by the US National Science Foundation award #1416860 to the City University of New York, the Population Council, the National Center for Atmospheric Research and the University of Colorado at Boulder as well as award IIS #1563933 to the University of Colorado at Boulder and IIS #1564164 to the University of Southern California. Finally, Innovative Seed Grant funding from the University of Colorado as well as a development grant received from its Population Center (CUPC) at the Institute of Behavioral Science, are acknowledged. We are grateful for the advice and critiques from Martino Pesaresi, director of the GHSL program at the Joint Research Center, European Commission, Ispra, Italy.

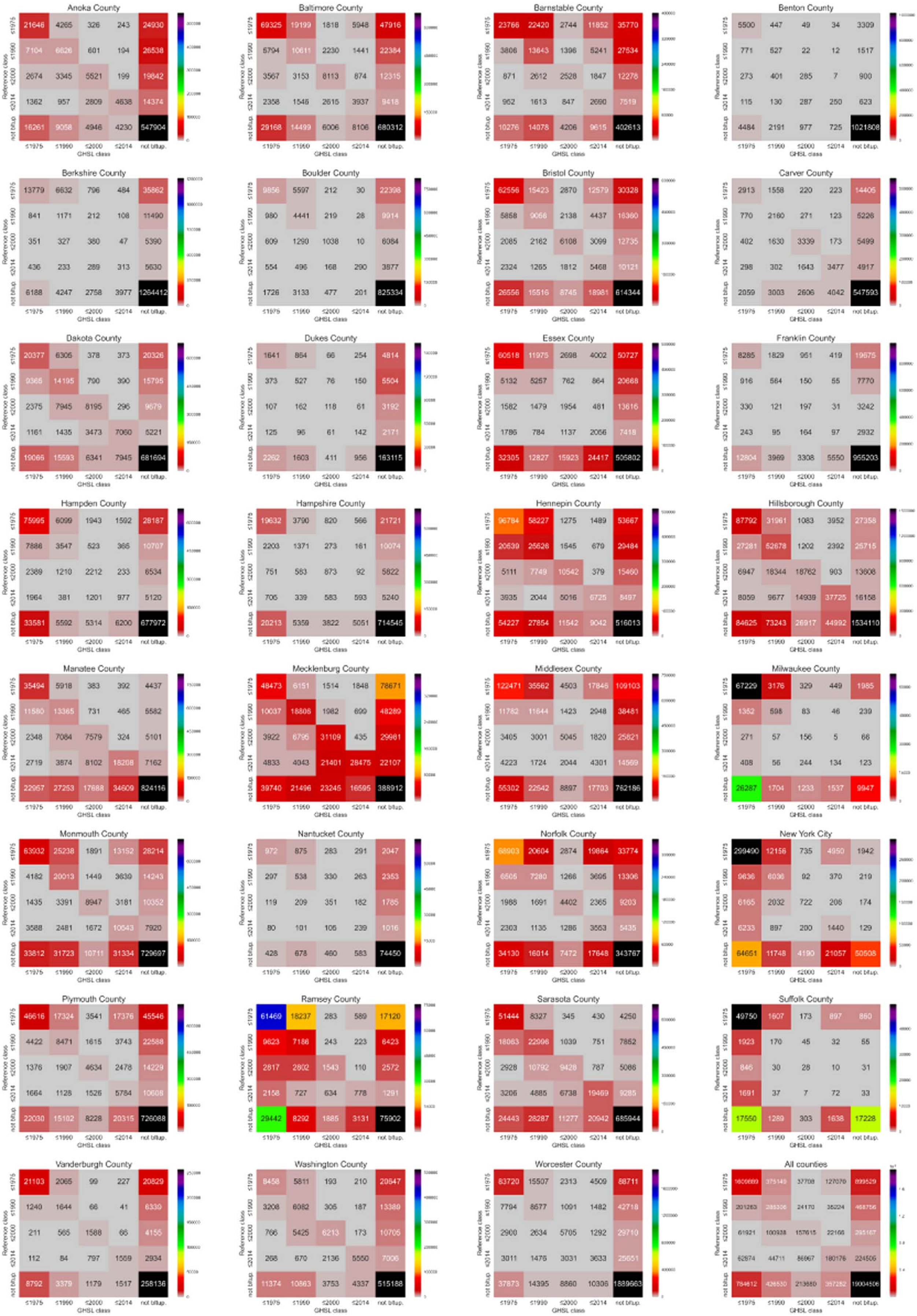
Appendix A

Table A1
Accuracy measures of the remaining 26 counties.

Test region	GHSL class	F-measure	G-mean	Kappa	NMI	Sensitivity	Specificity
Anoka County	Not built-up	0.901	0.630	0.416	0.167	0.941	0.422
	< 2015	0.510	0.630	0.416	0.167	0.422	0.941
	≤ 2000	0.493	0.629	0.410	0.162	0.420	0.942
	≤ 1990	0.478	0.637	0.413	0.168	0.429	0.947
	< 1975	0.431	0.636	0.389	0.164	0.421	0.960
	All epochs	0.563	0.632	0.409	0.166	0.526	0.842
Baltimore County	Not built-up	0.901	0.748	0.557	0.262	0.922	0.608
	< 2015	0.656	0.748	0.557	0.262	0.608	0.922
	≤ 2000	0.627	0.731	0.533	0.244	0.577	0.926
	≤ 1990	0.607	0.723	0.522	0.238	0.562	0.931
	< 1975	0.545	0.676	0.478	0.216	0.481	0.951
	All epochs	0.667	0.725	0.530	0.244	0.630	0.867
Barnstable County	Not built-up	0.869	0.704	0.492	0.207	0.913	0.543
	< 2015	0.620	0.704	0.492	0.207	0.543	0.913
	≤ 2000	0.538	0.638	0.417	0.164	0.438	0.930
	≤ 1990	0.525	0.634	0.418	0.167	0.429	0.936
	< 1975	0.349	0.489	0.284	0.113	0.246	0.970
	All epochs	0.580	0.634	0.421	0.172	0.514	0.858
Benton County	Not built-up	0.993	0.765	0.546	0.352	0.992	0.589
	< 2015	0.553	0.765	0.546	0.352	0.589	0.992
	≤ 2000	0.542	0.764	0.536	0.345	0.589	0.992
	≤ 1990	0.536	0.768	0.530	0.344	0.594	0.993
	< 1975	0.537	0.765	0.533	0.354	0.589	0.995
	All epochs	0.632	0.765	0.538	0.349	0.671	0.912
Berkshire County	Not built-up	0.971	0.554	0.386	0.199	0.987	0.311
	< 2015	0.411	0.554	0.386	0.199	0.311	0.987
	≤ 2000	0.420	0.558	0.398	0.216	0.314	0.989
	≤ 1990	0.425	0.558	0.405	0.229	0.314	0.991
	< 1975	0.348	0.488	0.333	0.194	0.239	0.994
	All epochs	0.515	0.542	0.381	0.207	0.433	0.854
Bristol County	Not built-up	0.898	0.774	0.565	0.263	0.898	0.667
	< 2015	0.667	0.774	0.565	0.263	0.667	0.898
	≤ 2000	0.615	0.728	0.520	0.232	0.576	0.920
	≤ 1990	0.610	0.732	0.531	0.246	0.575	0.932
	< 1975	0.561	0.694	0.499	0.234	0.505	0.952
	All epochs	0.670	0.740	0.536	0.247	0.644	0.874
Carver County	Not built-up	0.963	0.621	0.448	0.227	0.979	0.394
	< 2015	0.483	0.621	0.448	0.227	0.394	0.979
	≤ 2000	0.427	0.579	0.399	0.197	0.341	0.983
	≤ 1990	0.345	0.512	0.323	0.153	0.266	0.987
	< 1975	0.226	0.387	0.214	0.107	0.151	0.994
	All epochs	0.489	0.544	0.366	0.182	0.426	0.867
Dakota County	Not built-up	0.932	0.762	0.559	0.274	0.933	0.622
	< 2015	0.627	0.762	0.559	0.274	0.622	0.933

	≤ 2000	0.598	0.749	0.536	0.258	0.599	0.937
	≤ 1990	0.540	0.731	0.485	0.224	0.569	0.939
	< 1975	0.407	0.640	0.371	0.157	0.427	0.961
	All epochs	0.621	0.729	0.502	0.237	0.630	0.878
Dukes County	Not built-up	0.940	0.477	0.263	0.095	0.969	0.235
	< 2015	0.316	0.477	0.263	0.095	0.235	0.969
	≤ 2000	0.298	0.462	0.252	0.093	0.220	0.973
	≤ 1990	0.309	0.482	0.270	0.103	0.239	0.975
	< 1975	0.270	0.460	0.248	0.099	0.215	0.984
	All epochs	0.427	0.472	0.259	0.097	0.375	0.827
Essex County	Not built-up	0.850	0.671	0.386	0.122	0.855	0.526
	< 2015	0.535	0.671	0.386	0.122	0.526	0.855
	≤ 2000	0.541	0.670	0.416	0.147	0.503	0.893
	≤ 1990	0.560	0.684	0.458	0.184	0.510	0.919
	< 1975	0.523	0.661	0.443	0.182	0.466	0.938
	All epochs	0.602	0.671	0.418	0.152	0.572	0.826
Franklin County	Not built-up	0.970	0.541	0.298	0.117	0.974	0.301
	< 2015	0.328	0.541	0.298	0.117	0.301	0.974
	≤ 2000	0.340	0.542	0.314	0.133	0.300	0.979
	≤ 1990	0.332	0.530	0.310	0.134	0.285	0.982
	< 1975	0.308	0.512	0.290	0.127	0.266	0.986
	All epochs	0.456	0.533	0.302	0.126	0.425	0.844
Hampden County	Not built-up	0.931	0.797	0.612	0.321	0.930	0.682
	< 2015	0.682	0.797	0.612	0.321	0.682	0.930
	≤ 2000	0.680	0.798	0.616	0.327	0.681	0.935
	≤ 1990	0.679	0.801	0.620	0.336	0.683	0.940
	< 1975	0.645	0.793	0.591	0.313	0.668	0.941
	All epochs	0.723	0.797	0.610	0.324	0.729	0.886
Hampshire County	Not built-up	0.949	0.646	0.412	0.174	0.954	0.438
	< 2015	0.463	0.646	0.412	0.174	0.438	0.954
	≤ 2000	0.466	0.650	0.420	0.184	0.441	0.959
	≤ 1990	0.467	0.655	0.427	0.193	0.445	0.963
	< 1975	0.436	0.640	0.404	0.183	0.422	0.969
	All epochs	0.556	0.647	0.415	0.181	0.540	0.857
Hennepin County	Not built-up	0.831	0.763	0.533	0.221	0.834	0.698
	< 2015	0.702	0.763	0.533	0.221	0.698	0.834
	≤ 2000	0.688	0.761	0.528	0.218	0.692	0.838
	≤ 1990	0.680	0.770	0.541	0.232	0.695	0.852
	< 1975	0.494	0.638	0.367	0.117	0.458	0.890
	All epochs	0.679	0.739	0.501	0.202	0.675	0.822
Hillsborough County	Not built-up	0.908	0.832	0.585	0.303	0.870	0.796
	< 2015	0.674	0.832	0.585	0.303	0.796	0.870
	≤ 2000	0.628	0.824	0.550	0.280	0.769	0.882
	≤ 1990	0.603	0.827	0.536	0.278	0.764	0.895
	< 1975	0.479	0.735	0.432	0.198	0.577	0.937
	All epochs	0.658	0.810	0.537	0.272	0.755	0.876
Middlesex County	Not built-up	0.839	0.698	0.457	0.171	0.879	0.554
	< 2015	0.615	0.698	0.457	0.171	0.554	0.879
	≤ 2000	0.578	0.671	0.428	0.156	0.504	0.894
	≤ 1990	0.578	0.679	0.446	0.170	0.510	0.903
	< 1975	0.503	0.626	0.393	0.144	0.423	0.925
	All epochs	0.623	0.674	0.436	0.162	0.574	0.831
Milwaukee County	Not built-up	0.375	0.487	0.255	0.113	0.244	0.969
	< 2015	0.818	0.487	0.255	0.113	0.969	0.244
	≤ 2000	0.817	0.521	0.287	0.125	0.963	0.282
	≤ 1990	0.819	0.552	0.320	0.138	0.959	0.318
	< 1975	0.797	0.579	0.314	0.107	0.919	0.364
	All epochs	0.725	0.525	0.286	0.119	0.811	0.435
Monmouth County	Not built-up	0.897	0.801	0.565	0.269	0.872	0.735
	< 2015	0.667	0.801	0.565	0.269	0.735	0.872
	≤ 2000	0.625	0.761	0.533	0.241	0.642	0.903
	≤ 1990	0.620	0.767	0.541	0.254	0.644	0.914
	< 1975	0.534	0.679	0.476	0.218	0.483	0.954

	All epochs	0.669	0.762	0.536	0.250	0.675	0.876
Nantucket County	Not built-up	0.941	0.640	0.473	0.238	0.972	0.421
	< 2015	0.528	0.640	0.473	0.238	0.421	0.972
	≤ 2000	0.475	0.597	0.426	0.210	0.365	0.976
	≤ 1990	0.428	0.564	0.389	0.188	0.325	0.980
	< 1975	0.305	0.464	0.284	0.137	0.218	0.989
	All epochs	0.536	0.581	0.409	0.202	0.460	0.868
Norfolk County	Not built-up	0.834	0.762	0.520	0.211	0.820	0.708
	< 2015	0.686	0.762	0.520	0.211	0.708	0.820
	≤ 2000	0.615	0.707	0.453	0.163	0.584	0.856
	≤ 1990	0.610	0.712	0.467	0.177	0.580	0.873
	< 1975	0.530	0.654	0.411	0.148	0.472	0.907
	All epochs	0.655	0.720	0.474	0.182	0.633	0.833
New York City	Not built-up	0.492	0.574	0.399	0.254	0.332	0.993
	< 2015	0.871	0.574	0.399	0.254	0.993	0.332
	≤ 2000	0.876	0.666	0.497	0.279	0.977	0.454
	≤ 1990	0.867	0.671	0.497	0.276	0.975	0.462
	< 1975	0.849	0.709	0.512	0.243	0.938	0.536
	All epochs	0.791	0.639	0.461	0.261	0.843	0.555
Plymouth County	Not built-up	0.902	0.723	0.511	0.223	0.917	0.571
	< 2015	0.609	0.723	0.511	0.223	0.571	0.917
	≤ 2000	0.536	0.656	0.447	0.186	0.459	0.939
	≤ 1990	0.528	0.652	0.451	0.195	0.449	0.948
	< 1975	0.451	0.588	0.394	0.172	0.357	0.966
	All epochs	0.605	0.669	0.463	0.200	0.551	0.868
Sarasota County	Not built-up	0.925	0.874	0.670	0.398	0.890	0.859
	< 2015	0.744	0.874	0.670	0.398	0.859	0.890
	≤ 2000	0.719	0.885	0.659	0.403	0.867	0.903
	≤ 1990	0.693	0.892	0.641	0.399	0.873	0.912
	< 1975	0.624	0.866	0.590	0.359	0.794	0.946
	All epochs	0.741	0.879	0.646	0.391	0.857	0.902
Vanderburgh County	Not built-up	0.913	0.673	0.478	0.213	0.946	0.479
	< 2015	0.562	0.673	0.478	0.213	0.479	0.946
	≤ 2000	0.554	0.671	0.477	0.214	0.474	0.948
	≤ 1990	0.561	0.681	0.493	0.231	0.486	0.954
	< 1975	0.557	0.678	0.503	0.249	0.476	0.965
	All epochs	0.629	0.675	0.486	0.224	0.572	0.858
Washington County	Not built-up	0.926	0.665	0.454	0.194	0.944	0.469
	< 2015	0.527	0.665	0.454	0.194	0.469	0.944
	≤ 2000	0.495	0.650	0.431	0.181	0.446	0.948
	≤ 1990	0.423	0.619	0.368	0.142	0.403	0.950
	< 1975	0.285	0.483	0.251	0.090	0.239	0.974
	All epochs	0.531	0.616	0.392	0.160	0.500	0.857
Worcester County	Not built-up	0.936	0.653	0.474	0.226	0.964	0.443
	< 2015	0.535	0.653	0.474	0.226	0.443	0.964
	≤ 2000	0.524	0.649	0.468	0.223	0.436	0.966
	≤ 1990	0.532	0.661	0.485	0.241	0.451	0.969
	< 1975	0.507	0.648	0.470	0.238	0.430	0.975
	All epochs	0.607	0.653	0.475	0.231	0.545	0.863



Anoka County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	21646	4265	326	243	24930
≤1990	7104	6626	601	194	26538
≤2000	2674	3345	5521	199	19842
≤2014	1362	957	2809	4638	14374
not bltup.	16261	9058	4946	4230	547904

Baltimore County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	69325	19199	1818	5948	47916
≤1990	5794	10611	2230	1441	22384
≤2000	3567	3153	8113	874	12315
≤2014	2358	1546	2615	3937	9418
not bltup.	29168	14499	6006	8106	680312

Barnstable County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	23766	22420	2744	11852	35770
≤1990	3806	13643	1396	5241	27534
≤2000	871	2612	2528	1847	12278
≤2014	962	1613	847	2690	7519
not bltup.	10276	14078	4206	9615	402613

Benton County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	5500	447	49	34	3309
≤1990	771	527	22	12	1517
≤2000	273	401	285	7	900
≤2014	115	130	287	250	623
not bltup.	4484	2191	977	725	1021808

Berkshire County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	13779	6632	796	484	35862
≤1990	841	1171	212	108	11490
≤2000	351	327	380	47	5390
≤2014	436	233	289	313	5630
not bltup.	6188	4247	2758	3977	1264412

Boulder County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	9856	5597	212	30	22398
≤1990	980	4441	219	28	9914
≤2000	609	1290	1038	10	6084
≤2014	554	496	168	290	3877
not bltup.	1726	3133	477	201	825334

Bristol County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	62556	15423	2870	12579	30328
≤1990	5858	9066	2138	4437	16360
≤2000	2085	2162	6108	3099	12735
≤2014	2324	1265	1812	5468	10121
not bltup.	26556	15516	8745	18981	614344

Carver County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	2913	1558	220	223	14405
≤1990	770	2160	271	123	5226
≤2000	402	1630	3339	173	5499
≤2014	298	302	1643	3477	4917
not bltup.	2059	3003	2606	4042	547593

Dakota County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	20377	6305	378	373	20326
≤1990	9365	14195	790	390	15795
≤2000	2375	7945	8195	296	9679
≤2014	1161	1435	3473	7060	5221
not bltup.	19066	15593	6341	7945	681694

Dukes County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	1641	854	66	254	4814
≤1990	373	527	76	150	5504
≤2000	107	162	118	61	3192
≤2014	125	96	61	142	2171
not bltup.	2262	1603	411	956	163115

Essex County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	60518	11975	2698	4002	50727
≤1990	5132	5257	762	864	20668
≤2000	1582	1479	1954	481	13616
≤2014	1786	784	1137	2066	7418
not bltup.	32305	12827	15923	24417	505802

Franklin County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	8285	1829	951	419	19675
≤1990	916	564	150	55	7770
≤2000	330	121	197	31	3242
≤2014	243	95	164	97	2932
not bltup.	12804	3969	3308	5550	955203

Hampden County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	75995	6099	1943	1592	28187
≤1990	7886	3547	523	366	10707
≤2000	2389	1210	2212	233	6534
≤2014	1964	381	1201	977	5120
not bltup.	33581	5592	5314	6200	677972

Hampshire County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	19632	3790	820	566	21721
≤1990	2203	1371	273	161	10074
≤2000	751	583	873	92	5822
≤2014	706	339	583	593	5240
not bltup.	20213	5359	3822	5051	714545

Hennepin County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	96784	58227	1275	1489	53667
≤1990	20639	25626	1545	679	29484
≤2000	5111	7749	10542	379	15460
≤2014	3935	2044	5016	6725	8497
not bltup.	54227	27854	11542	9042	516013

Hillsborough County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	87792	31961	1083	3952	27358
≤1990	27281	52678	1202	2392	25715
≤2000	6947	18344	18762	903	13608
≤2014	8059	9677	14939	37725	16158
not bltup.	84625	73243	26917	44992	1534110

Manatee County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	35494	5918	383	392	4437
≤1990	11580	13365	731	465	5582
≤2000	2348	7084	7579	324	5101
≤2014	2719	3874	8102	18208	7162
not bltup.	22957	27253	17688	34609	824116

Mecklenburg County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	48473	6151	1514	1848	78671
≤1990	10037	18806	1982	699	48289
≤2000	3922	6795	31109	435	29981
≤2014	4833	4043	21401	28475	22107
not bltup.	39740	21496	23245	16595	388912

Middlesex County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	122471	35562	4503	17846	109103
≤1990	11782	11644	1423	2948	38481
≤2000	3405	3001	5045	1820	25821
≤2014	4223	1724	2044	4301	14569
not bltup.	55302	22542	8897	17703	762186

Milwaukee County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	67229	3176	329	449	1985
≤1990	1352	598	83	46	239
≤2000	271	57	156	5	66
≤2014	408	56	244	134	123
not bltup.	26287	1704	1233	1537	9947

Monmouth County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	63932	25238	1891	13152	28214
≤1990	4182	20013	1449	3639	14243
≤2000	1435	3391	8947	3181	10352
≤2014	3588	2481	1672	10643	7920
not bltup.	33812	31723	10711	31334	729697

Nantucket County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	972	875	283	291	2047
≤1990	297	538	330	263	2353
≤2000	119	209	351	182	1785
≤2014	80	101	106	239	1016
not bltup.	428	678	460	583	74450

Norfolk County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	68903	20604	2874	19864	33774
≤1990	6505	7280	1266	3696	13306
≤2000	1988	1691	4402	2365	9203
≤2014	2303	1135	1286	3563	5435
not bltup.	34130	16014	7472	17648	343767

New York City

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	299490	12156	735	4950	1942
≤1990	9636	6036	92	370	219
≤2000	6165	2032	722	206	174
≤2014	6233	897	200	1440	129
not bltup.	64651	11748	4190	21057	50508

Plymouth County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	46616	17324	3541	17376	45546
≤1990	4422	8471	1615	3743	22588
≤2000	1376	1907	4634	2478	14229
≤2014	1664	1128	1526	5784	10608
not bltup.	22030	15102	8228	20315	726088

Ramsey County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	61469	18237	283	589	17120
≤1990	9623	7188	243	223	6423
≤2000	2817	2802	1543	110	2572
≤2014	2158	727	634	778	1291
not bltup.	29442	8292	1885	3131	75902

Sarasota County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	51444	8327	345	430	4250
≤1990	18063	22996	1039	751	7862
≤2000	2928	10792	9428	787	5086
≤2014	3206	4885	6738	19469	9285
not bltup.	24443	28287	11277	20942	685944

Suffolk County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	49750	1607	173	897	860
≤1990	1923	170	45	32	55
≤2000	846	30	28	10	31
≤2014	1691	37	7	72	33
not bltup.	17550	1289	303	1638	17228

Vanderburgh County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	21103	2065	99	227	20829
≤1990	1240	1644	66	41	6330
≤2000	211	565	1588	66	4155
≤2014	112	84	797	1559	2934
not bltup.	8792	3379	1179	1517	258136

Washington County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	8458	5811	193	210	20647
≤1990	3208	6082	306	187	13389
≤2000	766	5425	6213	173	10705
≤2014	268	670	2136	5550	7006
not bltup.	11374	10863	3753	4337	515188

Worcester County

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	83720	15507	2313	4509	88711
≤1990	7794	8677	1091	1482	42718
≤2000	2900	2634	5705	1292	29710
≤2014	3011	1476	3031	3633	25651
not bltup.	37873	14395	8860	10306	1889663

All counties

Reference class	GHSL class				
	≤1975	≤1990	≤2000	≤2014	not bltup.
≤1975	1606689	375149	37708	127070	899529
≤1990	201263	285306	24170	38224	468756
≤2000	61921	100938	157615	22166	295107
≤2014	62874	44711	86967	180176	224506
not bltup.	784612	426530	213880	357282	19004506

Fig. A1. Confusion matrices of the 31 individual counties and across all study areas.

References

Balk, D., Pozzi, F., Yetman, G., Deichmann, U., Nelson, A., 2005. The distribution of people and the dimension of place: methodologies to improve the global estimation of urban extents. In: International Society for Photogrammetry and Remote Sensing, Proceedings of the Urban Remote Sensing Conference, March 2005, Tempe, AZ.

Balk, D.L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S.I., Nelson, A., 2006. Determining global population distribution: methods, applications and data. Adv. Parasitol. 62, 119–156.

Bontemps, S., Defourny, P., Bogaert, E.V., Arino, O., Kalogirou, V., Perez, J.R., 2011. GLOBCOVER 2009-products description and validation report. http://due.esrin.esa.int/files/GLOBCOVER2009_Validation_Report_2.2.pdf, Accessed date: 24 January 2017.

Center for International Earth Science Information Network/Columbia University, 2005. Global Rural-Urban Mapping Project (GRUMP). Socieconomic Data and Applications Center (SEDAC), Columbia University.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30 m resolution: a pok-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

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

Deichmann, U., Balk, D., Yetman, G., 2001. Transforming population data for interdisciplinary usages: from census to grid. Available at: http://sedac.ciesin.columbia.edu/gpw-v2/GPWdocumentation.pdf.

Detchmendy, D.M., Pace, W.H., 1972. A model for spectral signature variability for mixtures. In: Proceedings of the Conference on Earth Resources Observations and Information Analysis, pp. 596–620 13–14 March, Tullahoma, Tennessee.

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. Photogramm. Eng. Remote. Sens. 66 (7), 849–857.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenböck, H., Müller, A., Dech, S., 2013. Urban footprint processor—fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. IEEE Geosci. Remote Sens. Lett. 10 (6), 1617–1621.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24 (01), 38–49.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sens. Environ. 114 (10), 2271–2285.

Forbes, A.D., 1995. Classification algorithm evaluation: five performance measures based on confusion matrices. J. Clin. Monit. Comput. 11, 189–206.

Freire, S., Florczyk, A., Ehrlich, D., Pesaresi, M., 2015. Remote sensing derived continental high resolution built-up and population geoinformation for crisis management. In: Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, pp. 2677–2679.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., Mills, J., 2016. Development of new open and free multi-temporal global population grids at 250 m resolution. In: Proceedings of the 19th AGILE Conference on Geographic Information Science, (Helsinki, Finland, June 14–17, 2016).

Glick, H.B., Routh, D., Bettigole, C., Oliver, C.D., Seegmiller, L., Kuhn, C., 2016. Modeling the effects of horizontal positional error on classification accuracy statistics. Photogramm. Eng. Remote. Sens. 82 (10), 789–802.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, H., Zhu, P., Zhao, Z., Zhang, H., Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, L., Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, J., Chen, J., 2013. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM + data. Int. J. Remote Sens. 34, 2607–2654.

Grekousis, G., Mountrakis, G., Kavouras, M., 2015. An overview of 21 global and 43 regional land-cover mapping products. Int. J. Remote Sens. 36 (21), 5309–5335.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., Megown, K., 2015. Completion of the 2011 national land cover database for the conterminous united states - representing a decade of land cover change information. Photogramm. Eng. Remote. Sens. 81 (5), 345–354.

Horwitz, H.M., Nalepka, R.F., Hyde, P.D., Morgenstern, J.P., 1971. Estimating the Proportions of Objects Within a Single Resolution Element of a Multispectral Scanner. University of Michigan, Ann Arbor (NASA Contract NAS-9-9784).

Klotz, M., Kemper, T., Geiß, C., Esch, T., Taubenböck, H., 2016. How good is the map? A multi-scale cross-comparison framework for global settlement layers: evidence from Central Europe. Remote Sens. Environ. 178, 191–212.

Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the 14th International Conference on Machine Learning (ICML). vol. 97. pp. 179–186.

Leyk, S., Buttenfield, B.P., Nagle, N.N., Stum, A.K., 2013. Establishing relationships between parcel data and landcover for demographic small area estimation. Cartogr. Geogr. Inf. Sci. 40 (4), 305–315.

Leyk, S., Ruther, M., Buttenfield, B.P., Nagle, N.N., Stum, A.K., 2014. Modeling residential developed and in rural areas: a size-restricted approach using parcel data. Appl. Geogr. 47 (1), 33–45.

Linard, C., Kabaria, C.W., Gilbert, M., Tatem, A.J., Gaughan, A.E., Stevens, F.R., Sorichetta, A., Noor, A.M., Snow, R.W., 2017. Modelling changing population distributions: an example of the Kenyan Coast, 1979–2009. Int. J. Digital Earth 1–13. http://dx.doi.org/10.1080/17538947.2016.1275829.

Maclaurin, G.J., Leyk, S., 2016a. Geographic extension of existing land cover data in an active machine learning and corrective sampling framework. Int. J. Remote Sens. 37 (21), 5213–5233.

Maclaurin, G.J., Leyk, S., 2016b. Temporal replication of the national land cover database using active machine learning. GISci. Remote. Sens. 53 (6), 759–777.

Manson, S.M., Sander, H.A., Ghosh, D., Oakes, J.M., Orfield Jr., M.W., Craig, W.J., Luce Jr., T.F., Myott, E., Sun, S., 2009. Parcel data for research and policy. Geogr. Compass 3 (2), 698–726.

Nguyen, G. Hoang, Bouzerdoum, A., Phung, S., 2009. Learning pattern classification tasks with imbalanced data sets. In: Yin, P. (Ed.), Pattern Recognition. In-Teh, Vukovar, Croatia, pp. 193–208.

Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sens. Environ. 129, 122–131.

Pesaresi, M., Ehrlich, D., Gamba, P., Herold, M., 2009. A methodology to quantify built-up structures from optical VHR imagery. In: Global Mapping of Human Settlement: Experience, Datasets and Prospects. Taylor and Francis, New York, pp. 27–59.

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M.A., Ouzounis, G.K., Scavazzon, M., Soille, P., Syrris, V., Zanchetta, L., 2013. A global human settlement layer from optical HR/VHR RS data: concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 6 (5), 2102–2131.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Haag, F., Halkia, M., Julea, A.M., Kemper, T., Soille, P., 2015. Global human settlement analysis for disaster risk reduction. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 40 (7), 837–843.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, S., Julea, A., Kemper, T., Soille, P., Syrris, V., 2016a. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. In: JRC Technical Report EUR 27741 EN, http://dx.doi.org/10.2788/253582 (online).

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

Pontius Jr., R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 32 (15), 4407–4429.

Small, C., Sousa, D., 2016. Humans on Earth: global extents of anthropogenic land cover from remote sensing. Anthropocene 14, 1–33.

Smith, J.H., Wickham, J.D., Stehman, S.V., Yang, L., 2002. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. Photogramm. Eng. Remote. Sens. 68, 65–70.

Sorichetta, A., Hornby, G.M., Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., 2015. Americas Datasets. vol. 1 Harvard Dataverse.

Strahler, A., Boschetti, G., Foody, M., Fiedl, M., Hansen, M., Herold, P., Mayaux, P., Morisette, J., Stehman, S., Woodcock, C., 2006. Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps. Report of Committee of Earth Observation Satellites (CEOS)–Working Group on Calibration and Validation (WGCV). Office for Official Publications of the European Communities, Luxembourg.

Tapp, A.F., 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. Cartogr. Geogr. Inf. Sci. 37 (3), 215–228.

Tsendbazar, N., Bruin, S., Herold, M., 2015. Assessing global land cover reference datasets for different user communities. ISPRS J. Photogramm. Remote Sens. 103, 93–114.

Tsutsumida, N., Comber, A.J., 2015. Measures of spatio-temporal accuracy for time series land cover data. Int. J. Appl. Earth Obs. Geoinf. 41, 46–55.

Uhl, J.H., Leyk, S., 2017. A framework for radiometric sensitivity evaluation of medium resolution remote sensing time series data to built-up land cover change. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2017, (Fort Worth, TX, USA, July 23–28, 2017).

Uhl, J.H., Leyk, S., Chiang, Y.-Y., Duan, W.-W., Knoblock, C.A., 2017. Extracting human settlement footprint from historical topographic map series using context-based machine learning. In: Proceedings of the 8th International Conference on Pattern Recognition Systems, ICPRS-2017, (Madrid, Spain, July 11–13, 2017).

United Nations, 2012. Realizing the Future We Want for All. Report to the Secretary-General UN System Task Team on the Post-2015 UN Development Agenda, New York (2012).

von Meyer, N., Jones, B., 2013. Building National Parcel Data in the United States: One State at a Time. Fair and Equitable. International Association of Assessing Officers, pp. 3–10.

Wickham, J.D., Stehman, S.V., Gass, L., Dewitz, J., Fry, J.A., Wade, T.G., 2013. Accuracy assessment of NLCD 2006 land cover and impervious surface. Remote Sens. Environ. 130 (15), 294–304.

Wulder, M.A., Franklin, S.E., White, J.C., Linke, J., Magnussen, S., 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. Int. J. Remote Sens. 27 (4), 663–683.

Zhao, Y.Y., Gong, P., Yu, L., Hu, L., Li, X.Y., Li, C.C., Zhang, H., Zheng, Y., Wang, J., Zhao, Y., Cheng, Q., Liu, C., Liu, S., Wang, X., 2014. Towards a common validation sample set for global land-cover mapping. Int. J. Remote Sens. 35 (13), 4795–4814.

Zoraghein, H., Leyk, S., Ruther, M., Buttenfield, B.P., 2016. Exploiting temporal information in parcel data to refine small area population estimates. Comput. Environ. Urban. Syst. 58, 19–28.