Exploring the Potential of Integrated Cadastral and Building Data for Evaluation of Remote-Sensing based Multi-Temporal Built-up Land Layers

Johannes H. Uhl, Stefan Leyk, Aneta J. Florczyk, Martino Pesaresi, and Deborah Balk

ABSTRACT: An increasing amount of multi-temporal land use and built-up land datasets will be made available in the near future. However, little research has been done regarding the spatiotemporal uncertainty of these datasets. Publicly available cadastral parcel data including temporal information about construction dates of structures may be a useful source of reference data for spatiotemporal accuracy assessment, especially when parcel records are integrated with building footprints to create a spatially refined reference dataset. In this work we discuss the suitability of such an approach for establishing protocols for future spatiotemporal validation of multi-temporal built-up land data, exemplified by the novel Global Human Settlement Layer (GHSL), which assesses human presence on the planet on a global scale based on automatic classification of multi-temporal remote sensing data for a temporal extent from 1975 to 2014.

KEYWORDS: Spatiotemporal uncertainty, Global Human Settlement Layer, Open data, Data integration, Accuracy assessment

Introduction

The Global Human Settlement Layer (GHSL) project aims to map built-up land based on the integration of remotely sensed image data and census data (Pesaresi et al., 2015). In Pesaresi et al., (2013) the GHSL information production workflow was tested for a large set of sensors in the spatial resolution range of 0.5-10m. These sensors may perform very well in detection of built-up areas but are typically constrained regarding data access, 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 extrapolation of global, regional or even national trends.

In order to mitigate some of these issues, the GHSL system was ported in the 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., 2016). The GHSL is available as seamless global mosaic at high spatial resolution (approx. 38m) and for various epochs (1975, 1990, 2000, 2014, see Figure 1). This new dataset offers promising opportunities for population projections, disaster management and risk assessment (Freire et al., 2015; Freire et al., 2016), as well as for analysing and modelling urban dynamics and land use change.

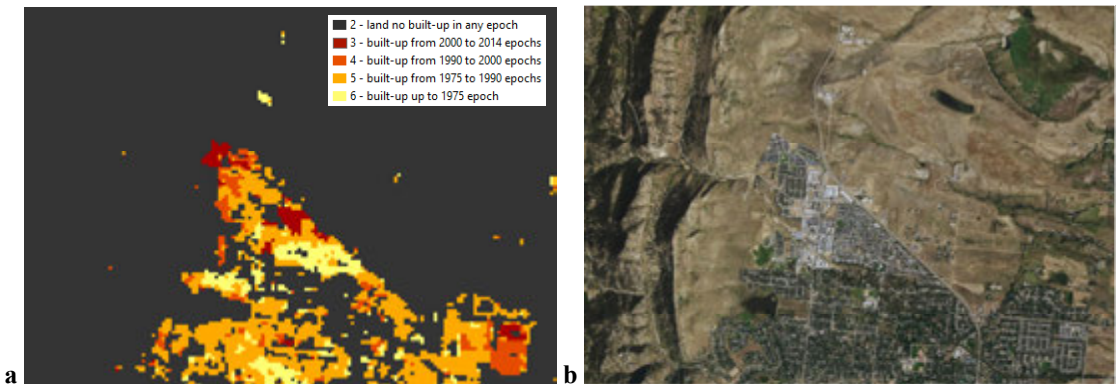

Figure 1: (a) GHSL built-up land identified for four time periods from 1975 to 2014 and (b) corresponding satellite image acquired in 2015 (Source: ESRI) for a subset of Boulder County (Colorado).

However, before such novel data products can be made available to the research community, an extensive quality assessment is required. Such assessments are difficult to realize due to the lack of reliable reference data particularly for earlier time periods and in less developed regions. In this paper we present and discuss a novel approach that can be applied to develop protocols for consistent future evaluation of multi-temporal spatial data on built-up land such as GHSL or developed land cover classes (e.g., in the National Landcover Database in the U.S.) using publicly available parcel (cadastral) data integrated with building footprints.

The aim of this study is to examine the suitability of such integrated data as reference data in accuracy assessments of fine-scale multi-temporal built-up land layers derived from automatic classification of satellite data. Possible spatial mismatches between data sources, which can be caused by positional uncertainty in the reference data, geometric inaccuracy of the imagery used to create the built-up land layers, aggregation and mixed pixel effects in the raster data or shifts of raster cells during resampling and projection processes, are addressed through a sensitivity analysis of the assessment results. In this sensitivity analysis, systematic offsets between reference data and the built-up data are incorporated and accuracy metrics are computed for each of these shifts. In the case of GHSL, a quality assessment protocol should be capable to separate the thematic and spatial components of overall errors, since the objective of GHSL is to report on the amount of built-up area within an administrative unit in order to support the monitoring of implementation of international frameworks and not to produce topographic maps.

This study focuses on feasibility of the proposed approach; consequently, the outcomes are not statistically representative quality measures for GHSL products but merely are intended to demonstrate the potential use of the integrated data products to establish effective evaluation frameworks and to show the sensitivity involved in such a comparison. Selection of the study areas is driven by the availability of the reference data. Even though these study areas are in the U.S., such an evaluation process can be extremely useful to shed light on the quality of such products in different contexts and represents a basis for further improvement of the multi-temporal built-up land layer.

Method

Data and study area

Open data policy makes cadastral, tax assessment and occasionally building data increasingly available to the public – often in GIS-compatible format – for many regions in the U.S. and other countries. Parcel data usually contain rich attribute information related to the type of land use, 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. Building data are becoming increasingly available and 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 overestimate built-up land if they remain unrefined. Some administrative regions in the U.S. provide these valuable data publicly and are used as study areas (Figure 2 for some examples).

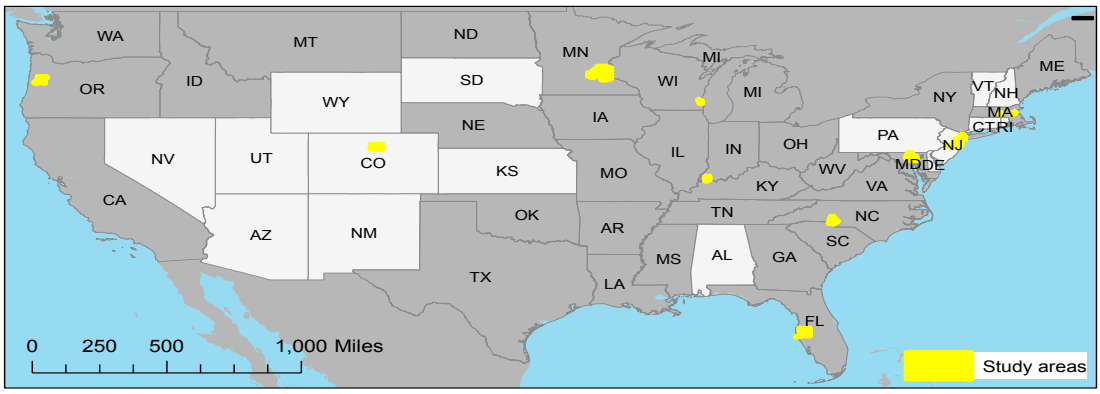

Figure 2: Study areas in the U.S. where parcel records including built year information and building footprint data are publicly available.

Integration of parcel records and building footprint data

Cadastral parcel boundaries are typically acquired using terrestrial or GNSS-based land surveying methods. Building footprints are often derived from LiDAR measurements or digitized based on aerial imagery. In order to create spatially refined parcel information, parcel data and building footprints have to be spatially integrated by establishing topological relationships between parcel and building objects such as containment. This is a challenging task due to different data acquisition methods and specific geometric and topological characteristics of the data as well as possible n:m relationships between building and parcel objects. In this work different vector data integration methods based on spatial joins are evaluated. During the data integration, parcel information (e.g., built year) is appended to the building feature(s) contained in the parcel, and for parcels that contain one or more building objects, area statistics are appended to the parcel object. According to Butenuth et al., (2007), the establishment of semantic relationships between spatial objects consists of two steps: a) data matching and b) data linking. Here, data matching is performed by spatially joining parcel and corresponding building objects based on geometrical and topological criteria. Data linking is performed implicitly during the spatial join: The unique identifier and other attributes of the matched object are

transferred to the attribute table of the target object and allows to retrieve associated objects within a GIS or other database environments. When the relationships between parcels and contained buildings are modeled, the choice of an efficient, accurate, and robust spatial join method is crucial. Three promising bidirectional spatial join methods were implemented and compared for a subset of Boulder County (Colorado) that contains both urban and rural areas, where parcel sizes and building densities are expected to vary, significantly. These methods are:

a. Spatially join buildings and parcels based on containment of building centroids in parcel polygons.
b. Spatially join buildings and parcels based on the majority of the overlapping area between the parcel and building polygons.
c. Spatially join buildings and parcels based on "complete containment" and "completely within" criterion.

Based on the established relationships between parcels and building objects, the built year information from the parcel is transferred to the building. In addition to that, the summarized area of the buildings joined to a parcel is computed and appended to the parcel and building objects. The ratio of summarized building areas in relation to the parcel area $\boldsymbol{r}$ is used to evaluate the performance of the spatial join method. A join is considered correct if the aggregated area of $\boldsymbol{k}$ buildings associated with a parcel does not exceed the area A_{Parcel} of the parcel itself:

$$r = \frac{\sum_1^k A_{Building\ i}}{A_{Parcel}} \leq 1$$

The centroid-based method and area majority-based method show similar robustness to geometric and topologic inconsistencies between the two datasets and relatively low maximum omission errors (Table 1). The maximum omission errors are estimated using the number of parcels that overlap with a building object which represents the maximum number of parcels that potentially can be joined to a building. The complete containment-method does not consider buildings that also overlap with adjacent parcels which leads to a high rate of correct joins but results in a high maximum omission error.

Table 1: Evaluation results of selected methods to spatially join parcels and building objects for a subset of Boulder County (CO).

Spatial join method	*Maximum number of parcels potentially to be joined*	*Spatially joined parcels*	*Correctly joined parcels (r≤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

The centroid-based joining method requires less computational power than the area majority method, since object matching is accomplished by a simple point-in-polygon query. For this reason and given the large amount of data to be integrated if several counties are included, the centroid-based joining method is chosen to transfer built-year information from parcel objects (Figure 3a) to building objects. The building objects with built-year information transferred from the parcel objects represent spatially refined temporal snapshots of built-up land (Figure 3b). In addition to the statistics in Table 1, further evaluation of the spatial join is performed by obtaining the number of buildings that are joined to parcels that do not have built-year information, and parcels with built-year information that were not joined to any building. This cross-comparison of the two independent data products allows to assess the completeness and reliability of the integrated data product, and provides important information about thematic uncertainty in the reference data introduced by the data integration process.

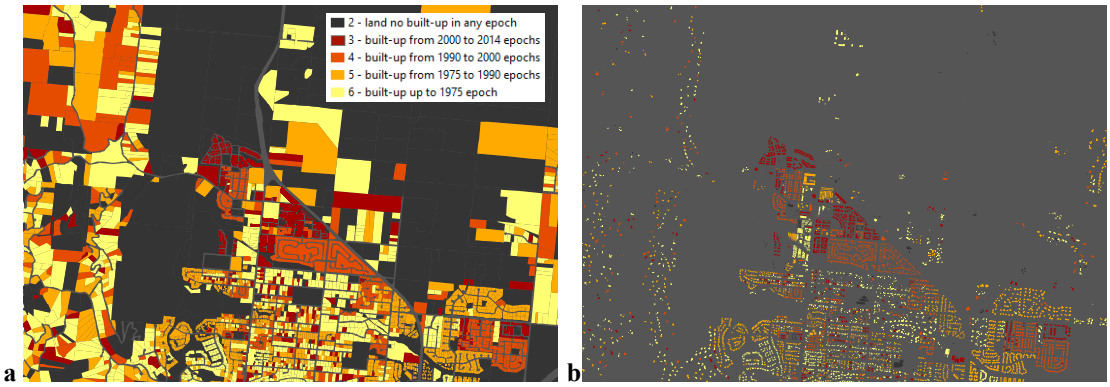

Figure 3: (a) Parcel-based reference data (vector), (b) parcel-based reference data refined by building footprints f or a subset of Boulder County (Colorado).

Creation of reference surfaces

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

The **built year information is converted and encoded** to match the temporal categories in GHSL (2: land not built-up, 3: land built-up from 2000-2014, 4: 1990-2000, 5: 1975-1990, and 6: <1975). For simplification purposes, GHSL epochs are treated as global temporal thresholds. In fact, 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 1985-1995).

The **same definition of the abstract class *built-up land*** as used in the GHSL is applied here for pixel value assignment. According to Pesaresi et al. (2016), a pixel is considered to be built-up if at least one structure detected by the GHSL classification method overlaps the pixel area. To identify overlaps between building objects and GHSL raster

cells, the building objects are rasterized to an intermediate spatial resolution of 2m using the GHSL-matched built-year class as raster value and it will be examined whether any of these cells intersect with the GHSL raster cell extents. The resolution of 2m is considered a good trade-off between maintaining characteristic building outline features and a feasible computational effort.

The intermediate raster dataset is then aggregated (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 in $GHSL_{ref}$ contains at least one 2m resolution cell encoded as built-up land, the $GHSL_{ref}$ cell will be classified as built-up and the oldest built-up category that occurs inside will be used for cell assignment, otherwise it will be assigned as not built-up. Figure 4 shows the converted 38m resolution reference surfaces based on parcel data only (Figure 4a) and spatially integrated parcels and building footprints (Figure 4b) both encoded with GHSL temporal categories.

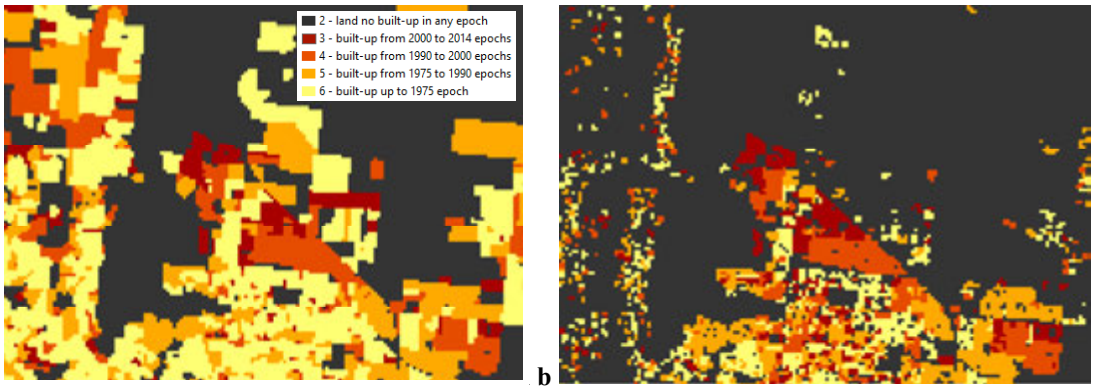

Figure 4: (a) parcel-based reference dataset and (b) spatially refined reference dataset both rasterized and resampled to GHSL resolution for a subset of Boulder County (Colorado).

Assessment of agreement between reference data and built-up land layers

The agreement between GHSL built-up land and the created reference surfaces is evaluated based on created confusion matrices which allow the derivation of various accuracy metrics (Fielding and Bell, 1997) for different time periods and each of the study areas. Each GHSL time span is evaluated cumulatively in order to characterize agreement behavior over time. These metrics can be used to quantify the classification accuracy of GHSL data for each time span assuming the reference data reflect built-up land with high accuracy, and provide rich material for discussion and interpretation. To evaluate the effect of the spatial refinement of parcel units using building footprints, both parcel-based and building-based reference surfaces are used (as shown in Figure 4).

The described method to create parcel-based and building-based reference surfaces is illustrated in Figure 5.

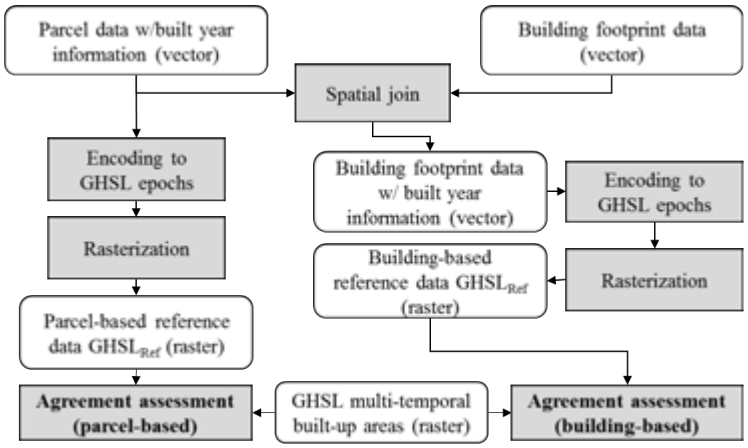

Figure 5: Workflow for parcel-based and building-based agreement assessment for GHSL built-up areas.

Results

Parcel-based versus building-based assessment of agreement

In a first comparative assessment the accuracy metrics for (overall) cumulative built-up areas in each epoch are computed using both parcel-based and building-based reference surfaces for Boulder County (Colorado), which has a rural area coverage of 88.25% according to U.S. Census 2010 Urban/Rural Classification. As can be seen in Table 2, using the spatially refined, building-based reference surface, Kappa, NMI, PCC and Producer's Accuracy, and Omission error indicate increased agreement compared to parcel-based reference data. One main reason can be found in the overestimation of built-up land in rural areas where parcel units can be very large but only small portions of the parcel area are actually built-up. Thus the integration of parcel and building data appears to create more realistic reference surfaces, especially for rural regions.

Table 2: Agreement metrics for comparing GHSL with the two reference surfaces for Boulder County (Colorado). Highlighted are accuracy metrics that improved by using spatially refined reference data.

Reference data type	GHSL class	Producer's Accuracy	User's Accuracy	Kappa	NMI	PCC	Omission error (%)	Comission error (%)
Parcel-based	**not built-up**	0.996	0.686	0.190	0.079	0.701	0.415	31.434
Parcel-based	**<2015**	0.158	0.954	0.190	0.079	0.701	84.205	4.619
Parcel-based	**≤2000**	0.150	0.846	0.178	0.058	0.727	85.024	15.366
Parcel-based	**≤1990**	0.132	0.766	0.163	0.047	0.772	86.837	23.404
Parcel-based	**<1975**	0.069	0.742	0.099	0.027	0.826	93.055	25.767
Refined by buildings	**not built-up**	0.980	***0.965***	***0.400***	***0.170***	***0.947***	2.025	***3.480***
Refined by buildings	**<2015**	***0.369***	0.505	***0.400***	***0.170***	***0.947***	***63.058***	49.452
Refined by buildings	**≤2000**	***0.375***	0.493	***0.401***	***0.174***	***0.951***	***62.456***	50.660
Refined by buildings	**≤1990**	***0.378***	0.515	***0.415***	***0.188***	***0.959***	***62.236***	48.537
Refined by buildings	**<1975**	***0.249***	0.586	***0.338***	***0.139***	***0.973***	***75.111***	41.431

Assessing agreement in rural and urban regions

Using the spatially refined reference data, agreement assessment is conducted for all administrative unit areas separately for urban and rural regions which are defined based on the U.S. census 2010 percentage of urban and rural land area per county estimate. Each county is divided into regions of mostly urban and rural character using a threshold of 50% coverage of urban land area per county. The preliminary agreement metrics in Figure 7 and Figure 7 demonstrate how agreement between built-up land layers and reference data could be assessed for different points in time, and show interesting differences for urban and rural regions regarding ranges and tendencies of the agreement metrics over time in the study areas as a totality covering different geographic settings.

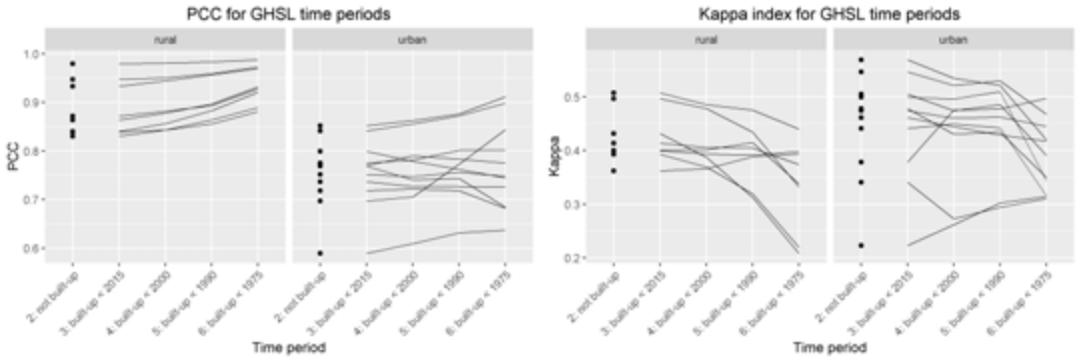

Figure 6: Temporal behavior of PCC and Kappa index for GHSL total built-up areas (using building-based reference data) in urban and rural regions of the different study areas.

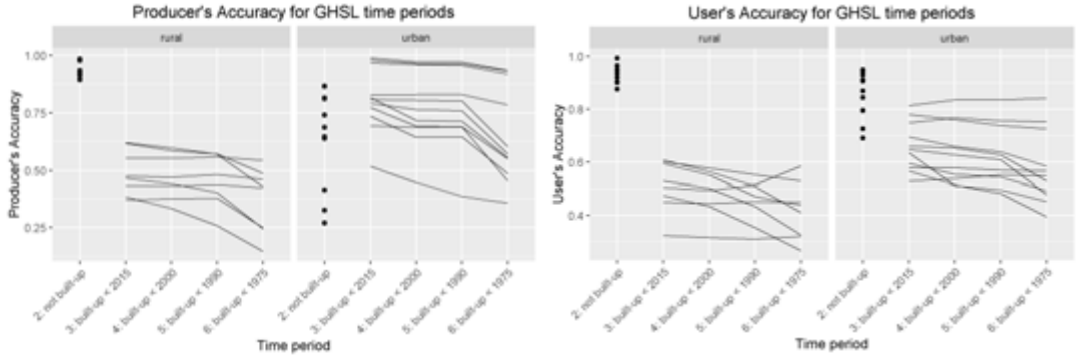

Figure 7: Temporal behavior of Producer's and User's accuracy for GHSL total built-up areas (using building-based reference data) in urban and rural regions of the different study areas.

Sensitivity Analysis

Important components of uncertainty in remotely sensed data and derived products include positional and thematic uncertainty. Furthermore, uncertainty in the reference data itself needs to be taken into account. Positional inaccuracy in the reference data may be introduced by the data acquisition method, such as digitization of building footprints from scanned maps or geometric distortions in aerial photography used for digitization. The satellite imagery used to create GHSL built-up land layers may suffer from displacements due to inaccurate image registration and may introduce positional inaccuracy. These aspects can cause positional discrepancies between reference data and

target data which may bias the assessment of agreement. To quantify the sensitivity of agreement metrics to potential positional discrepancies, in this study systematic offsets between the two datasets are simulated and the behavior of the agreement metrics will be observed.

One crucial point is the aggregation of high-resolution building footprints (2m resolution; Figure 8a) to create $GHSL_{ref}$ (approx. 38m resolution). To comply with GHSL specifications, the baseline agreement assessment described above assumed that a $GHSL_{ref}$ pixel is considered as built-up if it overlaps with a building object (i.e., overlap threshold >0%, see Figure 8b). To test for stricter rules of overlap (i.e., larger proportions of the building have to overlap the pixel area), this threshold is systematically increased to up to 20% 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 (Figure 8c). On the other hand, buildings outside of the pixel area may in reality overlap the pixel if the datasets are spatially off-set. To also account for this potential discrepancy, building footprints are systematically buffered by up to 40m (approx. one GHSL pixel size). $GHSL_{ref}$ pixels that overlap with these buffer areas (>0%) are considered built-up (Figure 8d) in the different scenarios.

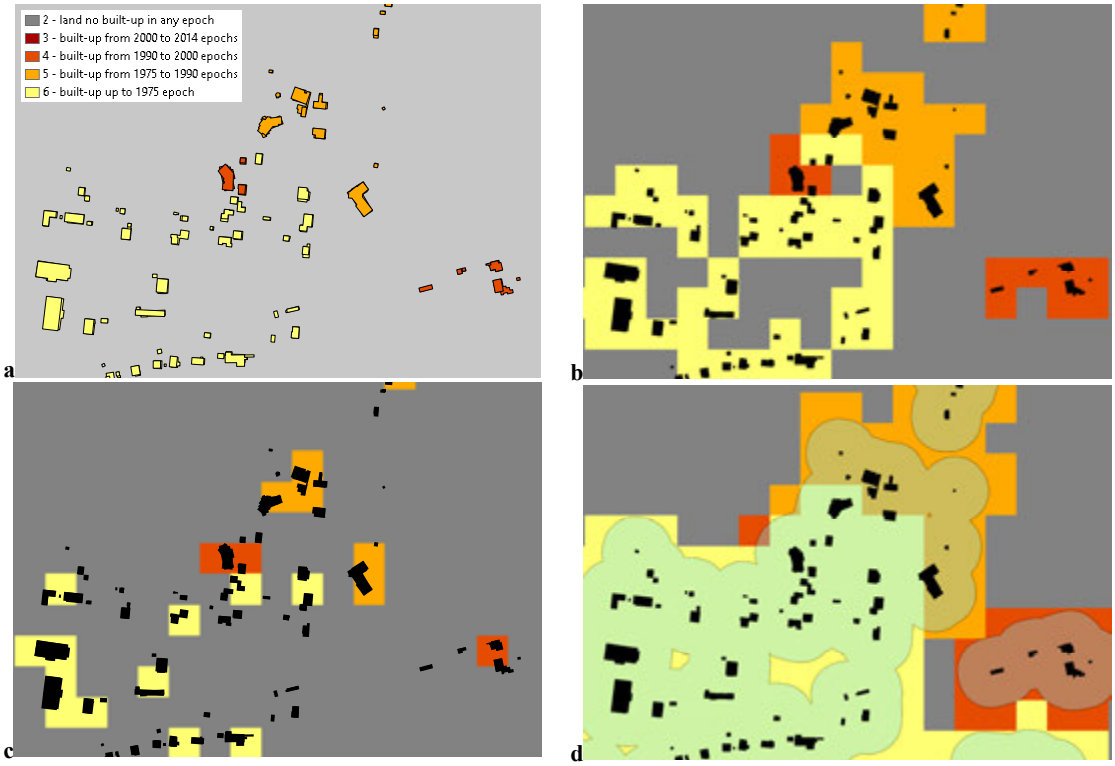

Figure 8: (a) Building footprints with built year information from parcel data integration, and reference surface $GHSL_{ref}$ based on (b) building footprints and >0% overlap threshold, (c) building footprints and >20% overlap, and (d) building footprints buffered by 40m and >0% overlap threshold with buffered area.

For each of these scenarios a reference surface $GHSL_{ref}$ is created for Boulder County (Colorado) and for each resulting surface the assessment of agreement between $GHSL_{ref}$ and GHSL built-up land is performed for cumulative built-up land in each time span.

For each scenario and time span the agreement metrics are shown in Figure 9 and Figure 10. It can be noted that some agreement metrics show higher degrees of sensitivity than others: While Producer's Accuracy and User's Accuracy ranges vary from 0.01 to 0.44 and 0.10 to 0.31, respectively, (Figure 9), Kappa and PCC show lower degrees of variation (differences of up to 0.23 and 0.08 between maximum and minimum, respectively) (Figure 10). PCC seems to be less sensitive to positional discrepancies between reference and test data. Furthermore, Producer's Accuracy and PCC seem to be less sensitive for older time periods, which might be due to decreasing area evaluated in older epochs. Sensitivity of Producer's Accuracy and Kappa seems to be nearly constant over time except for the earliest time epoch that show lower values.

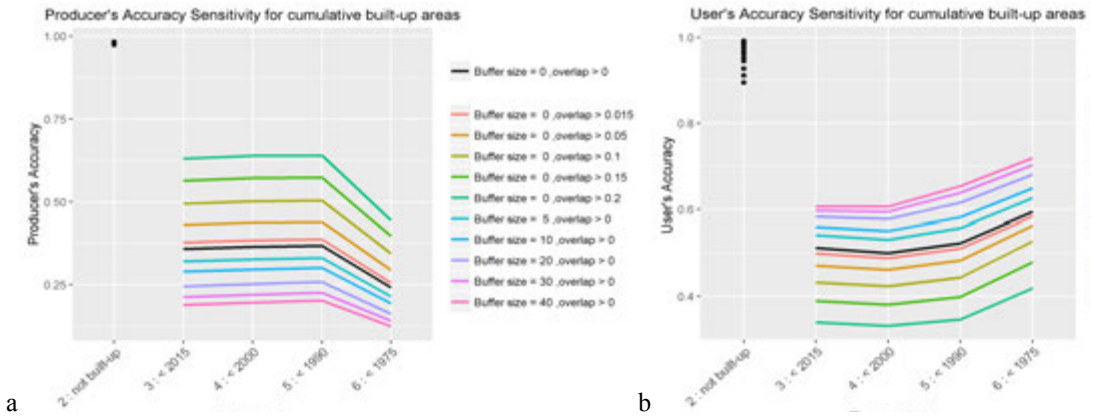

Figure 9: (a) Results of sensitivity analysis for Producer's Accuracy, and (b) User's Accuracy for cumulative built-up areas in Boulder County (Colorado).

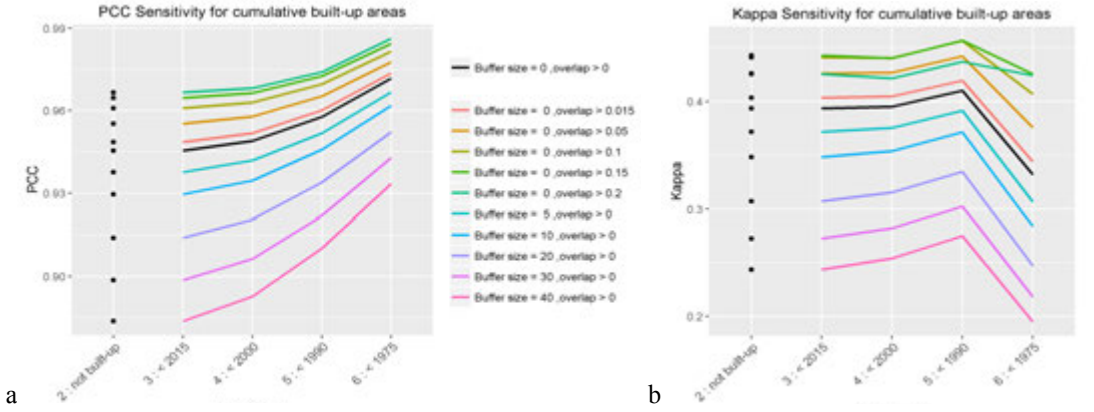

Figure 10: (a) Results of sensitivity analysis for PCC and (b) Kappa for cumulative built-up areas in Boulder County (Colorado).

Discussion

This study aims to show how publicly available cadastral and building data can be used for effective future spatiotemporal accuracy assessment of built-up land data to shed light on the methodological challenges and potential sensitivities in such assessments. In the case of the U.S., parcels in rural areas are often very large in relation to the built-up

portion and therefore bias the validation results if used as reference data. However, the integration with building footprints and the inherent spatial refinement allows to create more realistic reference data. Uncertainty in the reference data can be explored by cross-comparing parcel and building data. The conducted sensitivity analysis allows to quantify the sensitivity of the agreement metrics to possible positional discrepancies between reference and test data but also to variations of how to define the abstract class "built-up land" in remote sensing based classification procedures which also relates to the general concept of scale inherent in different data. Future work will focus on these sensitivities and more meaningful thresholds that will be employed in future large-scale validation analyses.

The presented preliminary results for the GHSL data show that the integrated parcel/building-based agreement assessment approach allows to reveal different degrees and tendencies of agreement over time for urban and rural areas. Furthermore, these results aim to demonstrate how agreement behaviour and possible ranges of agreement metrics over time can be graphically presented. However, they are not representative for the global GHSL data product. This approach can be applied to other multi-temporal land use / land cover products, such as the National Land Cover Database (NLCD) in the U.S.

Since the required reference data for this approach (parcel records including built year information and building footprints) are increasingly available to the public, the study areas can be extended in the near future which allows to assess the accuracy of built-up land layers across different geographic settings across the U.S. or internationally, and thus makes it possible to associate these places with varying degrees of underlying data quality. If applied to a wider set of study areas, this approach could provide a broader understanding of the spatial displacement in GHSL built-up area labels. However, further work is needed to separate thematic and spatial accuracy components.

Rural development and features associated with rural and small urban settlements are hard to discern from administrative data that tend to be coarse. A fuller understanding on the association between finely resolved parcel data or the like and globally-available GHSL data hold much promise for data poor regions of the world.

References

Butenuth, M. Gösseln, G. V. Tiedge, M. Heipke, C. Lipeck, U. and Sester, M. (2007) Integration of heterogeneous geospatial data in a federated database. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(5), 328-346

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

Freire, S. Florczyk, A. Ehrlich, D. and Pesaresi, M. (2015) Remote sensing derived continental high resolution built-up and population geoinformation for crisis

management. *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, 2677-2679.

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

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. and Zanchetta, L. (2013) A global human settlement layer from optical HR/VHR RS data: concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), 2102-2131.

Pesaresi, M. Ehrlich, D. Ferri, S. Florczyk, A. Freire, S. Haag, F. Halkia, M. Julea, A. M. Kemper, T. and Soille, P. (2015) Global Human Settlement Analysis for Disaster Risk Reduction. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 40, no. 7: 837.

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

Johannes H. Uhl, University of Colorado Boulder, Department of Geography, Boulder, CO 80309, United States of America

Stefan Leyk, University of Colorado Boulder, Department of Geography, Boulder, CO 80309, United States of America

Aneta J. Florczyk, European Commission – Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, 21027 Ispra, Italy

Martino Pesaresi, European Commission – Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, 21027 Ispra, Italy

Deborah Balk, City University of New York, Institute for Demographic Research and Baruch College, New York, NY 10010, United States of America