Remote Sensing of Environment 178 (2016) 191–212

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe

M. Klotz [a,*], T. Kemper [b], C. Geiß [a], T. Esch [a], H. Taubenböck [a]

[a] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen, 82234 Wessling, Germany
[b] European Commission, Joint Research Center (JRC), Institute for the Protection and Security of the Citizen, 21027 Ispra, Italy

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 18 February 2016
Accepted 1 March 2016
Available online xxxx

Keywords:
Global settlement mapping
Remote Sensing
Cross-comparison
Land cover validation
Accuracy assessment
Global Urban Footprint
Global Human Settlement Layer
Globcover
MODIS
Urbanization

ABSTRACT

Mapping of settlement areas from space is entering a new era. With the recently developed Global Urban Footprint (based on radar data from TanDEM-X) and the Global Human Settlement Layer (based on optical data), two new initiatives that promise to map complex settlement patterns at global scales and unprecedented spatial resolutions are about to enter the scientific and map user community. However, comparative studies on these layers' strengths and weaknesses, especially in terms of their potential added value with regard to existing lower resolution maps, as well as their assessed accuracy are still absent. In this regard, we introduce a multi-scale cross-comparison framework that uses the best existing urban maps as a benchmark. To paint a complete picture, we simultaneously address several components of map accuracy including relative inter-map agreement, absolute accuracies and pattern-based classification differences. This framework is applied to present regionally representative results from two Central European test sites. In this, we find that the new base maps bring decisive advancements in preserving the small-scale complexity of global human settlement patterns beyond urban core areas. Relative inter-map comparison exposes low density settlement regions traditionally under-represented by lower resolution maps that are now recognized. Absolute metrics such as the Kappa coefficient of agreement (K) show that accuracies of the new high resolution layers ($\overline{K} = 0.56$–0.58) nearly double those of existing products. Beyond, they feature substantial consistency between urban ($\overline{K} = 0.46$–0.50) and rural landscapes ($\overline{K} = 0.41$–0.45). Results from pattern-based exploration further reveal significant correlation of accuracies with physical pattern variations such as settlement density and mark a clear shift of accuracies from large to medium and small patch sizes. This differentiated view on classification accuracies shows that the new generation of urban maps constitutes a significantly enhanced spatial representation of large-scale settlement patterns.

1. Introduction

Global urbanization may well be the most important transformation that our planet will undergo in the 21st century. Even today, more than half or the world's population – approximately 54% – is living in urban areas, marking the dawn of the "urban century" (UN, 2014a). According to the United Nations' population projections, this share is expected to further increase to two-thirds in 2050 making cities the focal places of worldwide demographic growth. As our world progresses to demographically urbanize, the upcoming decades will bring along substantial changes with regard to size and spatial patterns of human settlements on our planet. New dimensions of urban landscapes such as mega-regions are increasingly being recognized (e.g., Florida, Gulden, & Mellander, 2008; UN-HABITAT, 2013; Taubenböck et al., 2014). Beyond, spatial complexity of urban transformation through e.g., peri-urbanization (e.g., Simon, 2008, Taubenböck, 2015), growth of urban villages and edge cities (Garreau, 1991; Anthrop, 2000), or the infrastructural delinking of rural areas located in the urban shadow (Main, 1993, Taubenböck & Wiesner, 2015) is constantly increasing. From a spatial perspective, the social, economic and environmental implications of global urbanization are not directly tangible. Nonetheless, the requirement of detailed, up-to-date, accurate and consistent information on the spatial patterns and dynamics of global settlements is today widely acknowledged (Potere & Schneider, 2007; Taubenböck et al., 2012; Esch, Marconcini, Felbier, Heldens, & Roth, 2014; Esch et al., 2012; UN 2014a, 2014b; GEO, 2014). In fact, it presents one key to understanding worldwide urbanization processes, and prerequisite to developing and supporting actions towards sustainable urban and rural development goals.

In this regard, satellite-based earth observation (EO) from space has long been recognized as an independent tool for the provision of area-wide spatial information on the location of settlement features and their spatial distribution from global (i.e., large-scale urban areas) to local scales (i.e., individual buildings). In the past decades, several initiatives coming from both government and academia have produced

* Corresponding author.

http://dx.doi.org/10.1016/j.rse.2016.03.001
0034-4257/

profound global maps on the size and spatial distribution of human settlements or related spatial attributes. This first generation of urban maps heavily relied on satellite sensors of relatively low geometric resolution (LR; ≥300 m acc. to EC-Copernicus, 2014). However, with the Global Urban Footprint (GUF) (Esch et al., 2013) and the Global Human Settlement Layer (GHSL) (Pesaresi et al., 2013), two new initiatives that promise to be capable of mapping fine-scale and complex human settlement patterns at unprecedented spatial resolutions and global scales are now becoming available. Knowledge on these layer's strengths and weaknesses in terms of their assessed accuracy, quality and overall agreement is however yet few and far.

In this regard, we present the first comprehensive cross-comparison that integrates these recent advancements in high resolution (HR; 4–30 m) settlement mapping into the portfolio of existing coarse resolution urban maps. To answer the call for a degree of confidence associated with the results from remote sensing-based land cover classifications (e.g., Richards, 1986; Congalton, 1991; Foody, 2002), our focus is on the capabilities of recently produced HR settlement maps respecting the best existing LR maps as a benchmark. To paint a complete picture, we develop and apply a comprehensive, multi-layered comparison framework that incorporates techniques of absolute accuracy assessment, analysis of relative inter-map agreement and exploration of pattern-based classification differences. We apply this framework for two large-scale test sites of varying landscape character in Central Europe. Within this setting, we present quantitative regional evidence on the mapping capabilities of the latest efforts in HR settlement mapping. In this, we address several specific research questions on the accuracy and validity of the respective layers under test:

(1) How and to which degree do new high resolution settlement layers correspond to existing global products of lower geometric resolution in a Central European setting?
(2) How accurate are different – high and low resolution – global geo-information layers in absolute terms regarding the representation of complex settlement features and their spatial configuration in Central Europe?
(3) How does the accuracy of these layers vary for structurally different areas, i.e., urban versus rural landscapes, in Central Europe?
(4) Does the accuracy of global settlement layers show spatial variation with regard to the physical configuration of human settlements, i.e., size or density, in Central Europe?

Building upon the presented framework, we aim at fostering the user-oriented assessment and definition of the novel products on the way to a global inventory of high resolution settlement information. The remainder of this work is organized as follows. The subsequent section presents relevant background information on past and present mapping and validation efforts followed by a review of techniques for meaningful accuracy assessment techniques. Section 3 depicts the layers under study from a technical perspective in combination with a brief description of the selected test sites, reference and ancillary data employed. The key methodological framework is summarized in Section 4 along with the scale-dependent steps of analysis taken. Section 5 presents the main results that are summarized and discussed in Section 6. Section 7 concludes with a final perspective and future directions.

2. Background & rationale

2.1. Overview of past and present global settlement mapping initiatives and their validation

Until the year 2000, only one dataset existed that aimed at representing the extent of the Earth's urban areas. The Digital Chart of the World (Danko, 1992; also known as Vector Map Level 0 (VMAP0)) was the predecessor to several global human settlement mapping initiatives since the millennium. These initiatives have produced an extended portfolio of ten global urban maps. Among these, six present urban areas as distinct human settlement outlines. In addition, four more layers model continuous physical features related to human settlement activity such as the degree of imperviousness of the land surface, the intensity of stable night-time illumination or the ambient local population. Satellite remote sensing data employed were mainly imagery from coarse resolution optical sensors such as the Moderate Resolution Imaging Spectroradiometer (MODIS), Satellite Pour l'Observation de la Terre (SPOT) or the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS). Table 1 gives a comprehensive overview of these layers including particular thematic and geometric specifications, data employed for map generation and their assessed accuracy according to Potere, Schneider, Angel, and Civco (2009).

Although these layers' usefulness and applicability for global analysis of larger urban areas are widely recognized, there are some problematic issues associated with their use: Heterogeneity in terms of geometric resolution (300 m–10 km), thematic representation (multi-category/binary/continuous information), and employed input data (EO/census/maps/data fusion) demands a high degree of expert knowledge by the map user. Beyond, the issue of a missing universally accepted, consistent and unambiguous definition of urban areas across these datasets is one of the major drawbacks with regard to their application (Schneider, Friedl, & Potere, 2010). Consequently, there is a large disagreement between the maps' total estimated shares of urban land at the global scale (Fig. 1). Beyond, inconsistencies between different scales of map representation (i.e., global vs. regional) are evident as indicated by the regional numbers for the city of Cologne, Germany. Further issues arise from low update frequencies, often data-dependent representation of human settlements (Schneider et al., 2010) as well as limited accuracy of the maps due to the spectral and spatial heterogeneity of built environments (Forster, 1983, 1985; Small, 2001, 2005). Ultimately, the coarse geometric resolution of EO data exploited so far does not embrace the full spatial complexity of large scale settlement patterns (Welch, 1982) and calls for novel HR layers that enable an enhanced spatial representation.

To answer this call, JRC and DLR have initiated the development of two new global products that promise to be a major leap forward regarding the derivation of spatially highly resolved settlement information on the global scale. The Global Urban Footprint (GUF) (Esch et al., 2012, 2013) builds upon the known capabilities of radar imagery for classification, monitoring and analysis of urban agglomerations at supranational levels (Henderson & Xia, 1997). It employs satellite imagery that is independent from weather, time-of-day and environmental conditions (Lewis, 1968). In contrast, the Global Human Settlement Layer (GHSL) (Pesaresi et al., 2013) initiative proposes a novel approach to map, analyze and monitor human settlements and ongoing urbanization processes in the 21st century. Exploiting high and very high resolution (HR/VHR) optical satellite imagery, GHSL is – although not globally available yet – up-to-date the largest and most complete known experiment based on optical EO data. Another promising approach that uses multi-spectral satellite imagery in combination with existing urban area maps is presented by Miyazaki, Shao, Koki, and Shibasaki (2013). It is, however, not subject to analysis in this work as the respective settlement layer only covers larger cities (>100.000 inhabitants) while disregarding other, lower density, settlement landscapes. Similarly, GUF and GHSL define urban areas based on distinct physical settlement features: Man-made vertical structures (GUF) or buildings (GHSL), respectively, mark the structuring elements for the derivation of generalized aerial representations of built-up areas (JRC, 2012; Esch et al., 2012). This eases the simultaneous assessment of these new high resolution geo-information products in the remainder of this work.

Despite these extensive efforts in global human settlement mapping now and in the past, comparative studies on these layers' strengths and weaknesses in terms of their assessed accuracy are still limited. Fig. 2 presents a comprehensive but non-exhaustive categorization of the published literature in this regard. While most studies relating to

Table 1
Overview of new (bottom section) and existing (top section) global urban maps; datasets employed in this study are printed in bold (adopted and updated from Schneider et al., 2010; accuracy statistics adopted from Potere et al., 2009).

Abbr.	Map (Reference)	Producer	Time-stamp	Definition of urban areas and map representation	Resolution	Primary data sources	Urban extent (km^2)	Relative City scale agreement– *Adj. R^2*	Absolute Accuracy– *Overall acc./ Kappa*
VMAP0	Vector Map Level 0 / Digital Chart of the World (5th ed.) (Danko, 1992)	US National Imagery and Mapping Agency (US-NIMA)	1992	Class: Populated places (thematic multi-category)	Scale 1:1.000.000	Operational navigation charts, maps	276.000	0.56	0.977/0.49
GLC00	Global Land Cover 2000 (Bartholme & Belward, 2005)	European Commission Joint Research Center (EC-JRC)	1999/2000	Class: Artificial surfaces and associated areas (thematic multi-category)	~ 1.000m	EO (SPOT-Vegetation 2000), LITES (Africa)	308.000	0.36	0.970/0.45
GLOBC	GlobCover v2 (Arino et al., 2007; ESA, 2011)	European Commission Joint Research Center (EC-JRC)	2009	Class: Artificial surfaces and associated areas (urban areas > 50%) (thematic multi-category)	~ 300m	EO (MERIS), GLC00	336.000	0.30	0.968/0.46
HYDE	History Database of the Global Environment (Goldewijk, 2011)	Netherlands Environmental Assessment Agency (PBL)	2000	Percentage of urban areas (built-up, cities) (continuous, %)	~ 10.000m	LSCAN, GLC2000, national / sub-national census & land use statistics, administrative city gazetteers	532.000	0.73	0.969/0.44
IMPSA	Global Impervious Surface Area (Elvidge et al., 2007)	US National Geophysical Data Center (US-NGDC)	2000/2001	Density of constructed impervious surfaces (continuous, %)	~ 1.000m	LSCAN, LITES	572.000	0.60	0.975/0.61
MOD500	MODIS 500m Map of Global Urban Extent (Schneider et al., 2009)	University of Wisconsin and Boston (US-NASA)	2001/2002	Built environment including non-vegetated, human constructed elements (> 50%) (thematic binary)	~ 500m	EO (MODIS 500m)	657.000	0.90	0.972/0.63
MOD1K	MODIS 1km Map of Global Urban Extent (Schneider et al., 2003)	Boston University (US-NASA)	2000/2001	Urban and built-up areas (thematic binary)	~ 1.000m	EO (MODIS 1km), LITES	727.000	0.67	0.960/0.50
GRUMP	Global Rural-Urban Mapping Project (CIESIEN, 2004)	Earth Institute at Columbia University (CIESIN)	1995	Urban extent (thematic binary)	~ 1.000m	VMAP, census data, LITES	3.532.000	0.75	0.839/0.22
LITES	DMSP-OLS Nighttime Lights (Elvidge et al., 2001)	US National Geophysical Data Center (US-NGDC)	1992-2015 (ongoing)	Nighttime illumination intensity (continuous, %)	~ 1.000m	EO (DMSP-OLS)	NA	-	-
LSCAN	Landscan (Bhaduri et al., 2002)	Oak Ridge National Laboratory (US-ORNL)	1998-2014 (ongoing)	Ambient (average per 24h) global population distribution (continuous, counts)	~ 1.000m	VMAP0, LITES, MOD1K, maps, census data, HR imagery	NA	-	-
GUF	Global Urban Footprint (Esch et al. , 2013)	German Aerospace Center (DLR)	2011-2013	Built-up areas marked by the presence of vertical structures (e.g., buildings)	12m	EO (TerraSAR-X/TanDEM-X)	NA	-	-
GHSL (PANTEX)	Global Human Settlement Layer (Pesaresi et al., 2013)	European Commission Joint Research Center (EC-JRC)	2014 (ongoing)	Built-up areas marked by the presence of buildings	10m	HR EO (multi-sensor optical data; 0.5-10m)	NA	-	-

the first generation of global urban maps focused on single product accuracies, more comprehensive studies comparing multiple products are scarce. The majority of these relate to the assessment of multi-category land over datasets that do not allow urban-specific conclusions. The most comprehensive, urban-specific review and comparison have been given by Potere and Schneider (2009) as well as Potere et al. (2009). In this, they present a quantitative non site-specific comparison of eight coarse resolution urban maps on the global, continental and city

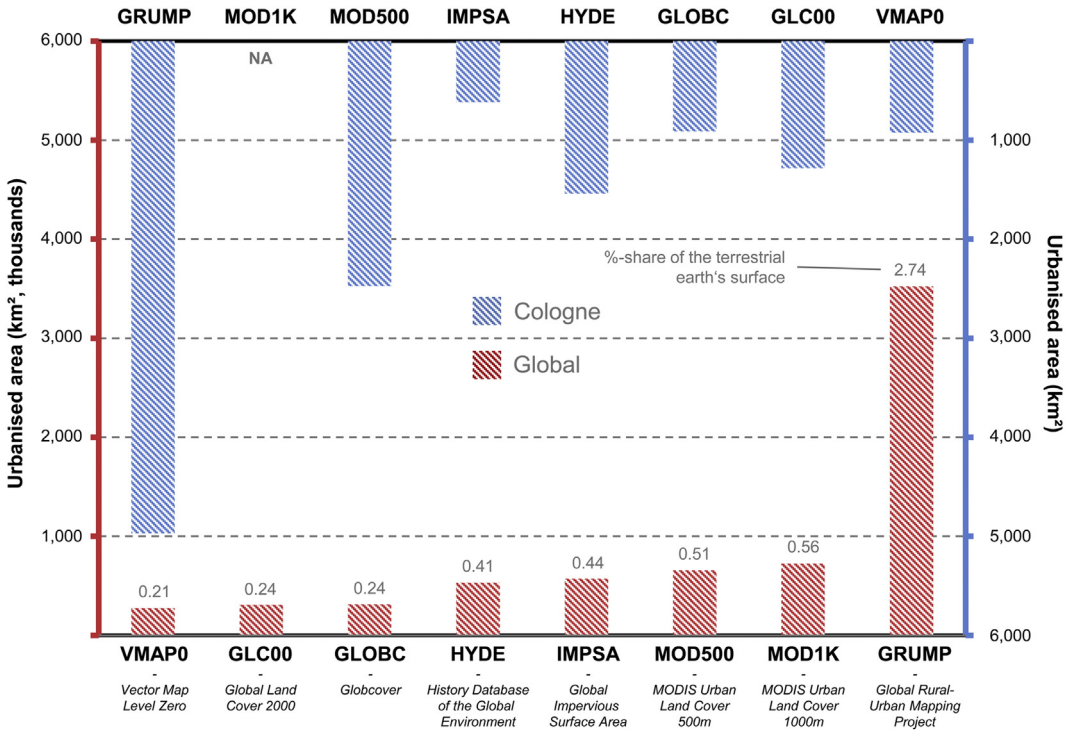

Fig. 1. Comparison of total urban area estimates of eight global urban maps on a global (red; thsds. km^2) and a regional scale for the city of Cologne, Germany (blue; km^2) (adopted from Klotz, Wurm, & Taubenböck, 2015; global estimates adopted from Schneider et al., 2010). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

level. Beyond, they determine absolute accuracies with regard to a stratified sample of 140 cities as well as relative inter-map agreement on the city-scale.

Accuracy investigations relating to recent initiatives in HR settlement mapping such as GUF and GHSL are yet relatively limited and focused on single-product accuracies. In fact, until today no study has aimed at an integrative and comparative validation of HR settlement layers with regard to existing lower resolution products. The rationale of our research intends to address this gap.

2.2. Land cover classification accuracy assessment

The exploration of classification accuracy methods to retrieve information about the quality of the thematic maps derived from remotely sensed data has been the focus of many studies (e.g., Congalton, 1991; Congalton & Green, 2008; Stehman & Czaplewski, 1998; Foody, 2002). Validation concepts have evolved considerably from early first-level visual confidence tests of the derived maps and non-site specific comparisons of gross classification rates to more sophisticated

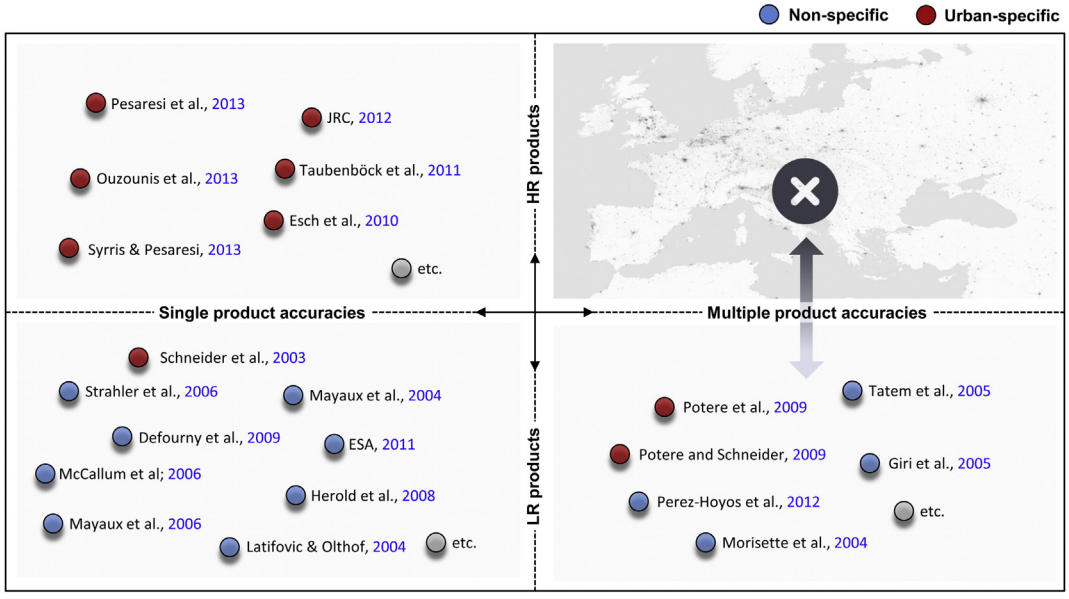

Fig. 2. Schematic categorization of a literature review on past validation efforts of multi-category and urban-specific land cover datasets (Defourny et al., 2009; Giri et al., 2005; Herold et al., 2008; Latifovic and Olthof, 2004; Mayaux et al., 2006; Mayaux et al., 2004; McCallum et al., 2006; Morisette et al., 2004; Stehmann, 2000; Strahler et al., 2006; Tatem et al., 2005).

pixel- (Congalton, 1994; Jensen, 1999) and object-based (Rutzinger, Rottensteiner, & Pfeifer, 2009) approaches. Nowadays, site-specific comparisons of class labels derived by classification with appropriate validation data have been widely accepted as a standard for reporting quantitative accuracy metrics (Congalton, 1994; Congalton & Green, 2008; Foody, 2006). At the core of most studies, the error or confusion matrix between the actual (columns) and the predicted (rows) class labels establishes a statistical basis for the description of classification accuracies (Congalton, 1991). Several descriptive measures and analytical techniques have been proposed to quantify accuracy and respective error: The producer's accuracy (commonly called sensitivity or recall for binary classification scenarios) is the counterpart of the error of omission. In contrast, the user's accuracy (precision) is complementary to the error of commission (Story & Congalton, 1986). Beyond, as the share of all correctly assigned sample units of the error matrix, the overall accuracy is commonly reported as a standard descriptive measure. However, many authors (e.g., Foody, 2002) have criticized its sole reporting due to the negligence of class-specific errors and thus, overestimation of the overall thematic map value.

As a consequence, several multivariate analytical measures have been proposed that rely on the entire error matrix and consider both types of errors: Kappa (Galton, 1892) presents a discrete multivariate statistic to test if binary (Cohen, 1960) or multi-categorical (Fleiss, 1971) classifications are significantly different from one another or respective reference data. Kappa highlights the differences between the actual agreement in the error matrix (i.e., the correctly classified sample units presented by the major diagonal) and the chance agreement presented by the column and row totals. Landis and Koch (1977) proposed a categorization of Kappa in which values of 0.00 to 0.20 are regarded as poor, 0.21 to 0.4 as fair, 0.41 to 0.6 as moderate, 0.61 to 0.8 as substantial, and 0.81 to 1.00 as almost perfect agreement to ease the comparison of multiple classification outputs. Although commonly applied as a standard, several authors have criticized the use of Kappa for its uni-modal response to prevalence (Allouche, Tsoar, & Kadmon, 2006), i.e., imbalanced class distributions in the reference data, and its over-estimation of chance agreement (Foody, 1992). Consequently, Congalton and Green (2008) have introduced several ways for the modification of Kappa such as weighting of errors and calculation of confidence limits whereas others have proposed alternative measures insensitive to prevalence such as the True-Skill-Statistic (Allouche et al., 2006).

Beyond single descriptive and analytical measures, other multivariate techniques have been applied to the error matrix (Congalton & Green, 2008): Normalization has been proposed to establish direct comparability between matrices of different-sized sample populations (Congalton, Oderwald, & Mea, 1983; Stehman, 2004). In addition, fuzzification of the error matrix is meant to account for semantic uncertainties and ambiguities induced by geolocation errors, differences in geometric resolution or fuzziness of thematic class descriptions between different land cover classification schemes (e.g., Gopal & Woodcock, 1994; Powers, 2007; Perez-Hoyos, García-Haro, & San-Miguel-Ayanz, 2012). In this, fuzzy sets establish a variable degree of class membership to rate the appropriateness of class allocation, thus increasing the significance of the accuracy assessment.

Although many of the presented techniques are today widely accepted as standards for reporting quantitative accuracy estimates, there are still many critical considerations associated with the design of systematic validation frameworks. Foody (2002, 2008) gives a comprehensive review in this regard: Predominantly, the selection of appropriate accuracy measures plays a key role for the conceptualization of a meaningful approach. Until today, there is no single universally accepted measure of agreement that is insensitive to all different features of the error matrix. Instead, a reasonable selection of measures must always consider the different components of accuracy to be evaluated and should be extended beyond the use of a single metric. Furthermore, poor quality of collected reference data may be transferred to the map resulting in reduced accuracy values of the thematic information content. A considerate selection of the sampling scheme, i.e., the sampling design and size, is thus one of the most important *a priori* considerations (Dicks & Lo, 1990; Stehman, 1999). Non-thematic misregistration and geolocation errors can further decrease accuracy of results accompanied by negative effects of mis-interpretation and mixed pixels. The latter is problematic due to the rigidness of the error matrix which assumes pure pixels and neglects the possibility of mixed land cover types within single pixels. Finally, to account for the structural and physical peculiarities of settlement areas, there has been a recent call for techniques respecting the variability of landscape patterns within the accuracy assessment process (Foody, 2006; Taubenböck, Esch, Felbier, Roth, & Dech, 2011).

Based on this review, we introduce a systematic framework incorporating different comparative, descriptive, analytical and pattern-based techniques beyond standard accuracy assessment protocols. To overcome differences in spatial resolution, thematic representation and landscape pattern, we follow a *multi-layered comparison concept*: As a *first step*, we conduct a relative comparison between recent initiatives in HR settlement mapping and existing urban maps of coarser geometric resolution to identify the potential added value of the new base maps. Based on these results, we *secondly* evaluate absolute overall and landscape-specific accuracies of all layers under test in terms of mapping spatially detailed and complex settlement features. In this, we build upon a considerate selection and combination of meaningful accuracy measures. To complete the picture, we explore the influence of the physical pattern variations on the classification results using pattern-based validation techniques.

3. Study sites and data

3.1. Test sites

We apply the proposed comparison framework to two large-scale test sites of Central Europe comprising square regions of 100 by 100 km. These have been selected for their varying settlement character and the large-scale availability of appropriate reference data for map validation. The test site of Cologne in the Western parts of Germany features a strong polarity between urban and rural landscape character. The eight larger cities (Cologne, Düsseldorf, Essen, Dortmund, Bochum, Duisburg, Wuppertal and Bonn) located in the Northern and Western parts of the test site are home to more than 4 million people (UN, 2014a). Contrasting this highly urbanized area, the Sauerland region located in the Eastern and South-eastern parts of the test site comprises several medium-sized towns (<300,000 inhabitants) of peri-urban character. Beyond, distinctly rural areas with <150 inhabitants per km^2 (according to the national census definition listed by UN (2014b)) exhibit a more fragmented settlement pattern (Fig. 3). The second test site of Tuscany, Italy, features a somewhat different spatial and demographic picture: As the only larger city, Florence, located in the Northern part of the test site, has around 700,000 inhabitants. Beyond, very few medium-sized towns describe the region's polycentric settlement structure (Burgalassi, 2010). However, the largest part of the test site is marked by rural livelihoods, i.e. communes with <10.000 inhabitants according to the national census definition. Continuously low densities reflect the dispersed and fragmented settlement structure. This makes the test sites specifically relevant for accuracy investigations with regard to the presumably enhanced mapping capabilities of new HR settlement layers in such landscapes.

3.2. Reference data

As an agreed standard for meaningful accuracy assessment (e.g. U.S. Bureau of the Budget, 1947; ASPRS, 1990; FGDC, 1998), the comparison of any given map product should be conducted against reference data that are independent from the map and preserves better

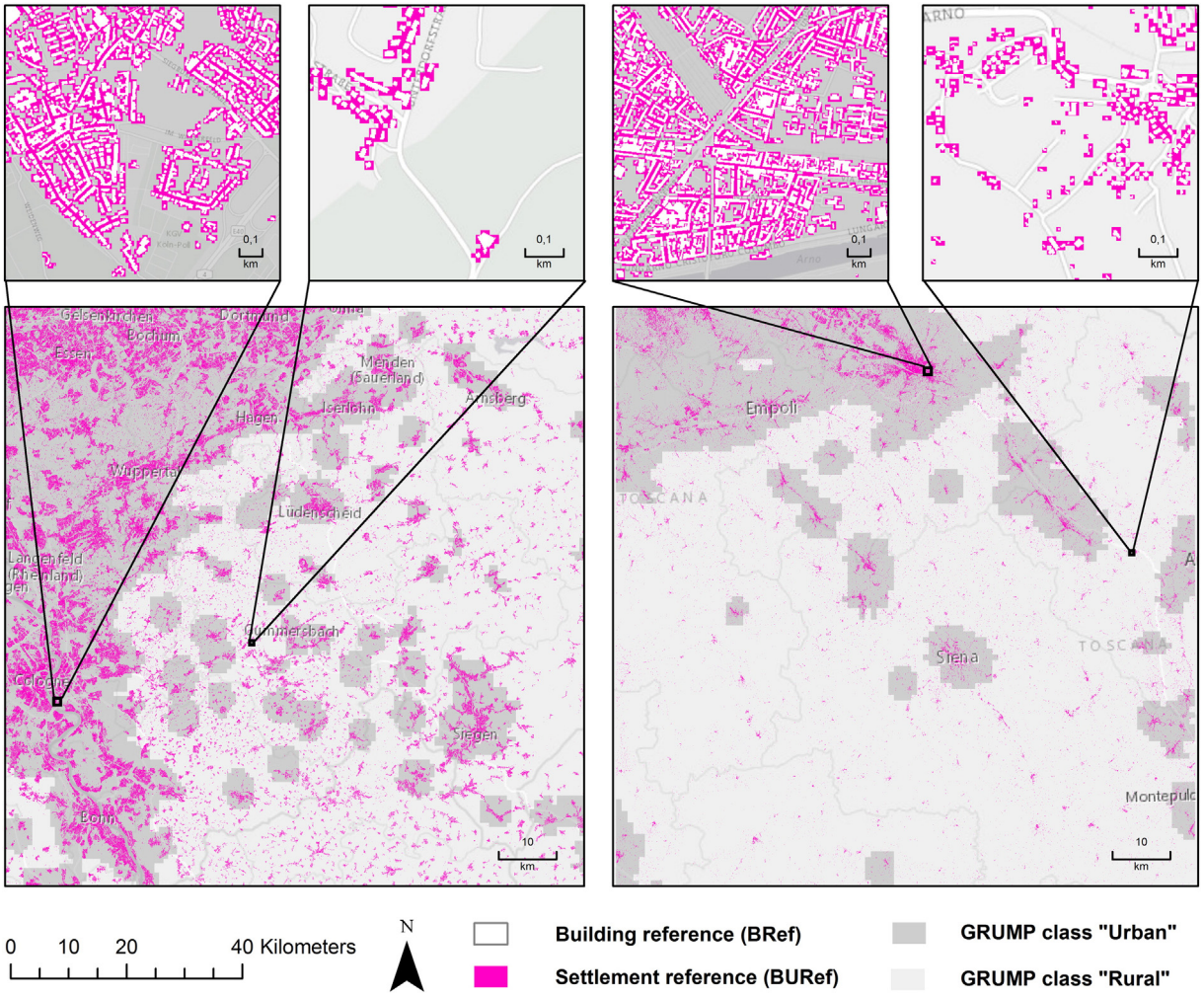

Fig. 3. Aerial views and subsets of the test sites Cologne, Germany (left), and Tuscany, Italy (right), including the reference layers, *BRef* (subsets) and *BURef*, as well as ancillary spatial information from GRUMP employed in this study. Source background map: Esri, HERE, DeLorme, MapmyIndia, OpenStreetMap contributors, and the GIS user community. Building outlines from TK25 (BKG, 2014) and the Carta Technica Regionale (Regione Toscana, 2015a).

geometric or thematic reliability. Since the main focus of our study is on the accuracy in mapping fine-scale settlement patterns, we employ two reference features of varying spatial complexity at the core of our reference database: Buildings (*BRef*) and settlement areas (*BURef*). As sound ground truth information, we opt for building outlines as the core element of settlement areas which we derive from consistent and reliable sources. As cadastral data is rarely existent or accessible for large aerial extents such as the selected test sites (10,000 km^2), we employ alternative sources. For the test site of Cologne, we derive footprints of individual buildings from the German topographic map 1:25,000 (BKG, 2014) from 2008 at a spatial resolution of 2.5 m as described by Wurm, d'Angelo, Reinartz, and Taubenböck (2014). For Tuscany, building outlines with equal geometric specifications were derived from the Carta Technica Regionale of 2010 (Regione Toscana, 2015a). Subsequently, all data have been transformed to Universal Transverse Mercator (UTM) projection (zone 32) with ellipsoid World Geodetic System 84 as displayed in the subsets of Fig. 3.

Quality considerations regarding the reference data relate to their geometric, temporal and semantic specifications: The layers under test exhibit maximum temporal shifts with regard to the reference of 6 and 8 years, respectively. To rate the influence of these time gaps, visual confidence checks have been performed by backdating a one percent stratified random sample of the reference (>3,000 buildings for both sites) to built-up areas identified from Landsat imagery of the year of the respective map. Change of less than three percent for both sites show that temporal shifts can be widely disregarded due to low urban growth dynamics. Beyond, positional accuracies of the reference data are around 10 m for the city of Cologne (GeoBasisNRW, 2013) and 6 m for the Tuscany test site (Regione Toscana, 2015b). Although these inaccuracies must be accepted with regard to the building mask, we can partly compensate these by the spatial generalization of settlement areas. In this regard, we employ the widely accepted semantic definition of built-up areas that describe aerial units recording the full or partial presence of buildings and the space-in-between buildings (Tenerelli & Ehrlich, 2011). Thus, we derive a generalized settlement mask, *BURef*, from the building mask, *BRef*, using a grid cell size of 12 m (Fig. 3) based on the native geometric resolution of the HR settlement maps investigated in this study. Beyond the compensation of inaccuracies due to mis-registration, the value of this additional HR reference layer lies in the more comprehensive representation of settlements encompassing other human-constructed elements such as roads and or impervious surfaces in close proximity to buildings (JRC, 2013). As an equivalent semantic descriptor, we use the term settlement area synonymous to built-up area from this point onward.

3.3. Settlement layers under test

Although this paper focuses on the validation of new advancements in HR global settlement mapping, i.e., GUF and GHSL, with the MODIS 500 m Map of Global Urban Extent (MOD500) and Globcover

(GLOBC), we integrate two more layers into the analysis. This is as these products yet set the benchmark of state-of-the art global settlement mapping in terms of thematic accuracy and spatial resolution according to Potere et al. (2009). Fig. 4 depicts the aerial representations of all layers subject to analysis for the selected test sites. The following subsections briefly summarize their inherent methodologies, underlying source data and semantic definitions of settlement areas.

3.3.1. MODIS map of global urban extent (MOD500)

MOD500 (Schneider et al., 2009) has been widely applied for global analysis in past academic research. The higher-ranking goal of this initiative was to produce an up-to-date, seamless and spatially consistent map of urban areas from a global MODIS Collection 5 coverage of the years 2001 and 2002. The data featuring a spatial resolution of ca.

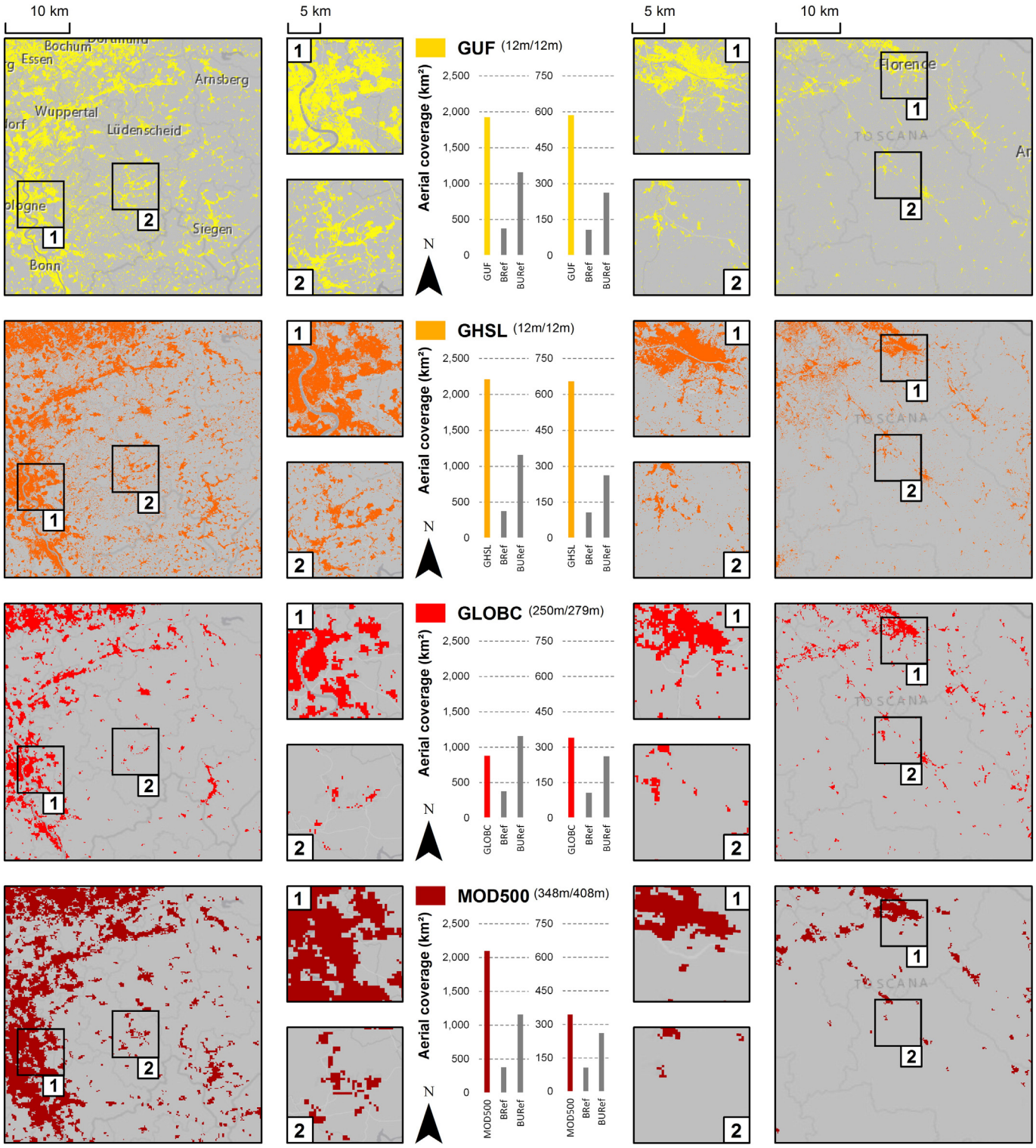

Fig. 4. Aerial views and subsets of the test sites Cologne, Germany (left), and Tuscany, Italy (right), displaying the layers under test in this study, namely GUF, GHSL, GLOBC and MOD500; the center columns represents total estimates of settlement areas (km^2) by these layers for both test sites compared to *BRef* and *BURef*, respectively. Source background map: Esri, HERE, DeLorme, MapmyIndia, OpenStreetMap contributors, and the GIS User Community.

500 m has been processed through a stratified supervised classification approach based on training data visually collected from higher resolution optical imagery. In addition, *posteriori* class membership functions were exploited for classification optimization. Due to the sole reliance on optical EO data, MOD500 defines urban areas based on physical attributes, i.e., places that are dominated by the built environment. These include a mix of human-constructed elements and impervious surfaces. In this context, the term 'dominated' implies aerial coverage >50% of a pixel. Reviewing Potere et al. (2009), MOD500 has been selected for this study as it is up-to-date the best known global urban map in terms of thematic accuracy (Table 1). For our study, MOD500 has been re-projected from its native geographic projection to UTM resulting in geometric resolutions of 347.9 m (Cologne) and 405.7 m (Tuscany), respectively.

3.3.2. Globcover (GLOBC)

Compared to MOD500, GLOBC is a multi-category global land cover product that has been first published in 2005 and updated in 2009 (Arino et al., 2007). GLOBC employs an automated land cover classification scheme based on spectro-temporal clustering of stratified production regions using a full year of observations from Medium Resolution Imaging Spectrometer (MERIS) on-board ENVISAT (ESA, 2011). It comprises 22 thematic land cover classes — one dedicated to artificial surfaces and associated areas. Similar to MOD500, this category is defined as pixels having an urban area percentage of >50%. Although Potere and Schneider (2009) found only moderate accuracies for GLOBC (Table 1), it is yet considered in this study as it is up-to-date the geometrically highest resolved dataset. Re-projected from its native geographic projection the dataset features an output resolution of 249.9 m and 278.5 m, respectively.

3.3.3. Global Urban Footprint (GUF)

Based on the German satellite constellation of TerraSAR-X (TSX) and TanDEM-X (TDX) two global coverages of the Earth's entire land-mass have been acquired between 2011 and 2013. The GUF processor exploits the local speckle information, i.e., the local coefficient of variation and the fading texture of the radar imagery which is acquired in single-polarized StripMap mode at 3 m spatial resolution. This texture information highlights heterogeneous built-up areas featuring strong backscattering signals due to direct or double bounce reflection in the proximity of vertical structures such as buildings (Esch et al., 2013). This information feeds into a fully automatic unsupervised classification that spatially generalizes these seeds to derive a binary built-up/non built-up classification in a so far unique spatial resolution of 12 m. The GUF output is mainly related to built-up regions that feature vertical structures (e.g., houses, walls, traffic signs, etc.) but excludes impervious surfaces without a vertical component. Initial validation efforts carried out for the GUF report region-specific accuracies ranging between 60% and 95% (Esch et al., 2010, 2013; Taubenböck et al., 2011, 2012).

3.3.4. Global Human Settlement Layer (GHSL)

GHSL is a similar product aiming at the derivation of a globally consistent spatial representation of human settlements. It exploits multi-resolution (0.5–10 m), multi-platform, multi-sensor (panchromatic, multispectral) and multi-temporal optical image sources. Up until today, the dataset covers more than 24 million square kilometers of the earth's land surface. The information extraction is heavily based on PANTEX (Pesaresi, Gerhardinger, & Kayitakire, 2008) — a rotation-invariant, anisotropic textural measure based on the grey level co-occurrence matrix of the input imagery. The strong correlation between this textural information and the local density of buildings produces a continuous built-up index [0, 1] that can be semantically translated to a dichotomic built-up mask (Pesaresi et al., 2013; Ouzounis, Syrris, & Pesaresi, 2013). Thus, the term 'built-up' of the GHSL definition refers explicitly to pixels that coincide with buildings, but ideally excludes non-building vertical structures as compared to GUF. First validation efforts by JRC (2012) used a visual validation protocol and reported overall accuracies for PANTEX between 80% and 90% with region-specific accuracy variations. For this study, the 10 m GHSL output, i.e. PANTEX, has been derived from pan-sharpened SPOT-5 imagery. Based on the assumption that all positive values greater zero present pixels that coincide with buildings, we threshold PANTEX accordingly to derive a binary built-up mask. This step complies with the conceptual definition of built-up areas by Ehrlich and Tenerelli (2013) and is thus analogous to the spatial generalization of *BURef*. Eventually, nearest neighbor resampling to the GUF resolution is applied to create a comparable pair of HR layers.

3.4. Ancillary data

In order to assess and compare classification accuracies of the layers under test in areas of varying landscape character, i.e., urban vs. rural areas, a transferrable rule for spatial zoning is required. Since an accepted and semantically consistent global definition of urban and rural areas is yet non-existent, we employ a data-driven approach. By the Global Rural-Urban Mapping Project's (GRUMP) urban extent layer, we use an ancillary data source that allows for a general distinction of urban and rural landscapes. GRUMP is not subject to validation in this study as it features by far the lowest accuracies among all urban maps reviewed by Potere et al. (2009) (Table 1). This is mainly due to its strong reliance on buffered census data and LITES, thus corresponding more closely to population than built-up areas. However, its very generalized and clumpy layout presents a particular strength that allows for a consistent coarse-level separation of urban and rural extents (Fig. 3). As we consider the density of the built environment as one of the main distinguishing features of urban and rural areas (Fina, Krehl, Siedentop, Taubenböck, & Wurm, 2014), consistency and plausibility of the structural divergence of the GRUMP classes are verified in Table 2. Thus, spatial zoning via GRUMP enables us to examine landscape-specific accuracies of the layers under test.

In terms of the spatial transferability of this approach, it must be acknowledged that while the original GRUMP dataset dating back to 1995 presents an eligible choice for regions of Central Europe, it is deemed outdated for parts of the world that feature higher urban growth dynamics. Nevertheless, its inherent methodology described by Balk, Pozzi, Yetman, Deichmann, and Nelson (2005) could be reasonably applied to update and transfer spatial zoning based on up-to-date data sources to other cultural areas of the world.

4. Methodological framework

We introduce a *multi-scale comparison framework* (Fig. 5) to achieve a comprehensive and systematic description of the accuracy and validity of the layers under test. In this, map accuracy is deconstructed to various components for both new HR and existing LR settlement layers in an integrative manner. Based on the review of techniques for meaningful accuracy assessment in Section 2.2, the analytical framework incorporates different pixel-, object- and pattern-based validation

Table 2
Mean building and settlement densities of the GRUMP classes derived from *BRef* and *BURef*, respectively.

	Building density (*BRef*, %)		Settlement density (*BURef*, %)	
	Cologne	Tuscany	Cologne	Tuscany
GRUMP class "Urban"	5.76	3.22	17.04	2.43
GRUMP class "Rural"	1.40	0.37	5.10	1.04

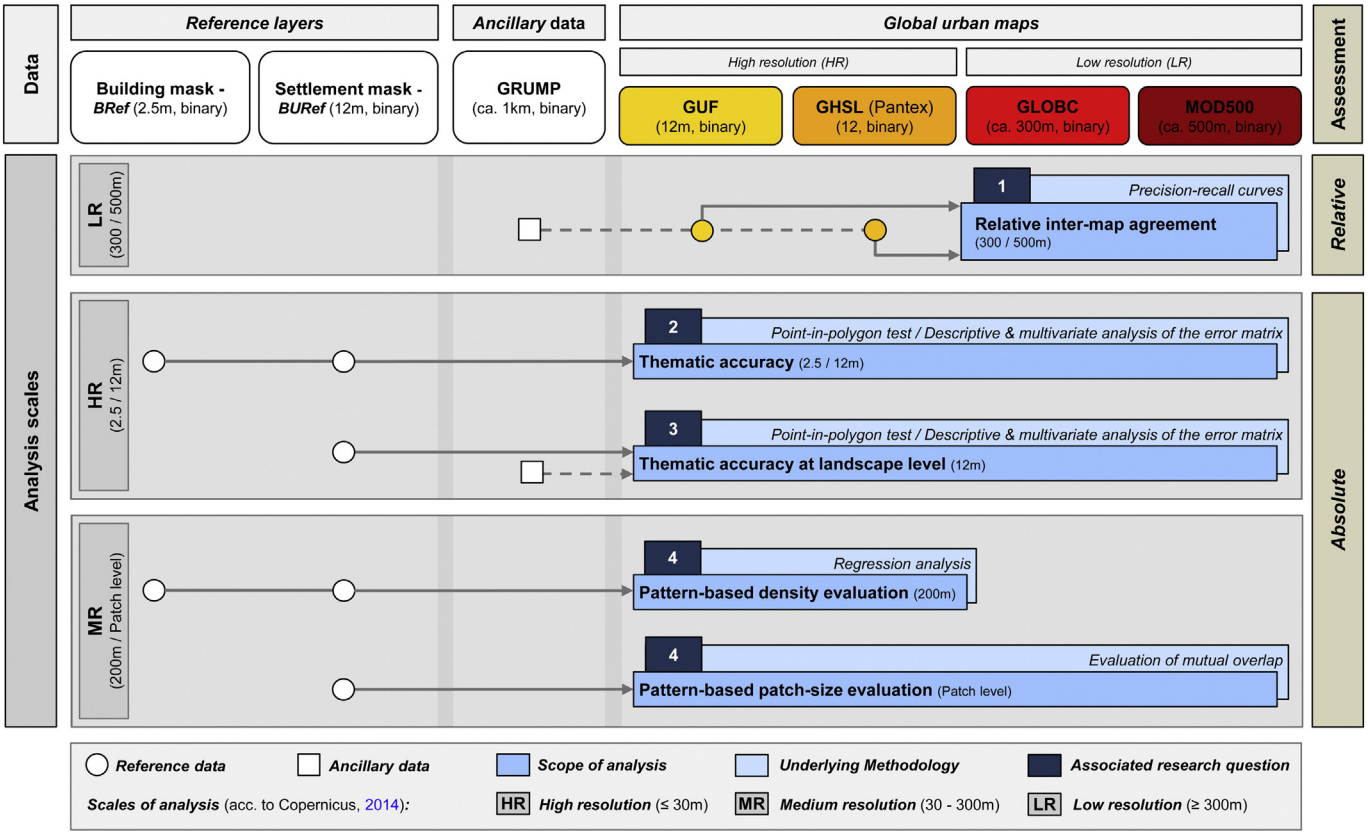

Fig. 5. Multi-scale cross-comparison framework listing the scale-specific steps of analysis, associated research questions as well as the respectively employed test, reference and ancillary data layers.

techniques to answer the research questions addressed in the introduction.

As a *first step*, we explore the potential added value of the novel HR maps with regard to existing products of coarser geometric resolution. This analysis is conducted on the LR scale that relates to the native spatial resolution of GLOBC and MOD500. Using techniques of performance evaluation (Section 4.1), we evaluate the trade-off between over- and under-representation of these benchmark datasets by novel HR layers. This relative assessment establishes an understanding on the degree and nature of inter-map agreement. As a proof of concept, regions of map deviation and thus, potential HR map evolution, are quantitatively explored.

Based on these results, *secondly*, we determine absolute accuracies on the *high resolution scale of analysis* (Section 4.2). Building upon on a considerate selection of accuracy metrics, we assess the performance of the respective layers under test with regard to their accuracy in mapping explicit spatial settlement features. In this context, we asses all maps with regard to both, the building (*BRef*) and the settlement mask (*BURef*), to explore and specify each layer's semantic definition. Beyond the perspective on the overall test site extent, we integrate landscape-specific statistics for urban and rural areas in this assessment.

Ultimately, in a *third step*, we conduct a differentiated assessment of absolute classification accuracies that respects spatially altering structural characteristics of the built environment (Section 4.3). This assessment on the *medium resolution scale* (MR; 30–300 m) of analysis helps us to refine our understanding of layer-specific accuracies. We exploit pattern-based evaluation techniques to examine accuracy variations as a function of settlement density and size. In this, we present an advanced object-based approach evaluating the spatial overlap between mapped and referenced settlement patches. This establishes an understanding of the scale-dependent accuracy variations.

The following subsections describe the applied methods on each specific analysis scale. As a central conceptualization for all methodological steps taken, the error matrix (Table 3) formally compares spatial units (i.e., pixels or objects) of the binary classifications under test against the reference data as the basis for the calculation of specific accuracy measures. Given an arbitrary binary classifier, there are four possible outcomes for n elements of an error matrix: True positives (*TP*) and true negatives (*TN*) describe correctly detected presences and absences, respectively. In contrast, false positives (*FP*) and false negatives (*FN*) are incorrectly classified absences (commission) and incorrectly rejected presences (omission), respectively, of the reference. This common conceptualization is naturally adopted to the urban/non-urban categorization in the remainder of this work. With regard to the review of accuracy metrics presented in Section 2.2, we follow the terminology commonly used for the performance evaluation of binary classifiers from this point onward as compared to remote sensing-specific notations such as e.g., producer's and user's accuracy.

Table 3
Conceptualization of the error matrix for a binary classification scenario; $n = TP + FP + FN + TN$.

		Reference data	
		Presence	Absence
Layer under test	Presence	*TP*	*FP*
	Absence	*FN*	*TN*

4.1. Low resolution analysis

On the LR scale of analysis, we employ precision-recall curves (PR curves) and related error statistics to quantify the *relative inter-map agreement* between each pair of HR and LR settlement layer (Fig. 5). As a conceptual foundation of the cross-comparison framework, this

step is meant to establish an understanding on the degree and nature of map correspondence and disagreement, respectively, between multi-resolution layers. Beyond, the motivation of this step inheres in the exploration of the potential added value of the novel HR products with regard to existing LR maps. PR analysis allows us to identify particular regions of disagreement via the HR maps' density domain. PR curves are strongly related to the receiver operating characteristics (ROC; Kullback, 1968) and have long been used in evaluation of information retrieval systems (e.g. Raghavan, Bollmann, & Jung, 1989; Manning & Schutze, 1999; Fawcett, 2006). In this, they have been proven advantageous compared to ROC curves as they paint a more informative picture when dealing with highly skewed class distributions (Davis & Goadrich, 2006).

As a prerequisite for PR analysis, spatial aggregation is applied to HR layers to produce a continuous density derivate at lower geometric resolution that bridges the resolution gap between layers. Let $X \in \{0,1\}$ be an image with a binary domain presenting non built-up and built-up areas at a high geometric resolution, i.e., GUF or GHSL, Eq. (1) produces the built-up density D for each pixel $x \in X$ (Ouzounis et al., 2013):

$$D(x) = \frac{\sum_{x \in N} \{x \mid \text{class}(x) = 1\}}{s^2} \tag{1}$$

In this, $D(x)$ denotes the fraction of pixels labelled as 'built-up' (coded as '1') within the square structuring element N. The edge length s of N corresponds to the native spatial resolution of lower geometric resolution layers, i.e., GLOBC or MODIS.

Following this preliminary, PR curves reflect the trade-off of recall and precision between the pairs of thresholded density aggregates from HR layers and binary LR layers. In this regard, recall presents the fraction of *TP* out of the positives of the LR layer (*TP rate*; also called sensitivity or completeness), whereas precision is the fraction of *FP* out of the negatives of the corresponding HR map (positive predictive value (*PPV*)). This ratio is calculated and visualized in PR space for all possible threshold values T applied to the continuous density representation of D of HR layers. Thus, the lowering of the density threshold from 100% to 0% is conceptually equivalent to the discretization (or masking) of the density layers from GUF and GHSL, respectively, at a particular threshold value $D(X) > T$. This is followed by the calculation of respective performance measures – recall and precision – against the respective LR counterpart. The PR curve is thus, complementary to the gradual course of over- (p_{FP}) and under-representation (p_{FN}) of a LR layer by a respective HR layer:

$$p_{FN}(T) = \frac{\sum_i 1(d_i < T) 1(y_i = 1)}{\sum_i 1(y_i = 1)} \tag{2}$$

$$p_{FP}(T) = \frac{\sum_i 1(d_i \geq T) 1(y_i = 0)}{\sum_i 1(d_i \geq T)} \tag{3}$$

whereby, $d_i \in D$ presents the continuous density measurements derived from HR layers and its respective dichotomic class representation at a particular threshold T, and $y_i \in \{0,1\}$ reflects the binary class representation of LR layers. In this manner, we employ PR curves to rate and examine inter-map agreement between layers as a function of the built-up density measured by HR layers. In order to compare agreement between map combinations that produce different curves in PR space, the minimum error rate ER_{min} presents a consistent quality metric. It describes the minimum total of over- (p_{FP}) and under-representation (p_{FN}) identified from PR space:

$$\text{ER}_{\min} = \min[\text{p}_{\text{FN}}(\text{T}) + \text{p}_{\text{FP}}(\text{T})], \exists \text{T}. \tag{4}$$

ER_{min} theoretically ranges between 0% and 200% whereby small values testify good correspondence of the respective HR and LR layer in their way of discriminating settlement areas at a given threshold. Beyond, we use the capabilities of PR analysis for threshold optimization to empirically identify the optimal cut-off value at which the error rate is minimal ($T(ER_{min})$). This allows us to quantitatively characterize density margins of map correspondence and disagreement, respectively. Following this protocol, we examine PR curves in combination with the statistics of ER_{min} and $T(ER_{min})$ to investigate overall and landscape-specific agreement between each pair of HR and LR layers across sites.

4.2. High resolution analysis

Based on the results of the previous comparison, the focus of the HR analysis is on assessing *site-specific absolute accuracies* of all layers in terms of mapping explicit settlement features. To give a first quantitative estimate of the accuracy of the maps under study, we follow a straight forward object-based approach. The point-in-polygon test (PIP; Rutzinger et al., 2009) employs vector representations of *BRef* to count the buildings spatially covered by an aerial classification. We overlay the settlement extent of each dataset with the centroid of each building footprint from *BRef* to calculate shares of buildings covered and omitted, respectively. Although, this type of assessment does not embrace a complete thematic description of absolute map accuracy, the analysis gives a first indication of the completeness of classifications in terms of capturing the core elements of human settlements.

Subsequently, we conduct a pixel-based evaluation based on the error matrix and the respective reference data described in Section 3.2. As the tabulation of the error matrix postulates equal geometric resolutions of the map and the reference, when necessary, we resample the datasets under investigation to the geometric resolutions of *BURef* (2.5 m) and *BRef* (12 m), respectively. In this scenario, positional inaccuracies due to up-sampling amount to relatively small maximum errors of 1.25 m and 6 m, respectively. For the analysis of absolute classification accuracies, we follow the recommendations by Foody (2006, 2008) and base the interpretation of results on a combination of meaningful accuracy metrics beyond the use of a single statistic. As standard descriptive measures, we report the overall accuracy (A), the *TP* rate (*TPR*), the *TN* rate (*TNR*) and the precision (positive predictive value (*PPV*)) according to Eqs. (5), (6), (7) and (8).

$$A = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{n} \tag{5}$$

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$TNR = \frac{TN}{FP + TN} \tag{7}$$

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

These measures enable a descriptive qualification of classification accuracy: Sensitivity and specificity are complementary to errors of omission and measure the completeness of the built-up and non built-up categories, respectively. On the contrary, precision addresses the correctness of the classification result and is intrinsically related to errors of commission.

Beyond these descriptive measures, we employ further multivariate analytical measures that consider both types of errors based on the entire error matrix. These are Kappa (K, Eq. (9); Congalton et al., 1983), F-score (F, Eq. (10); Rijsbergen, 1979) and True-Skill-Statistic (*TSS*, Eq. (11); Allouche et al., 2006):

$$K = \frac{\left(\frac{TP + TN}{n}\right) - \frac{(TP + FP)(TP + FN) + (FN + TN)(TN + FP)}{n^2}}{1 - \frac{(TP + FP)(TP + FN) + (FN + TN)(TN + FP)}{n^2}} \tag{9}$$

$$F = \frac{2TP}{2TP + FP + FN} = \frac{2 \times TPR \times PPV}{TPR + PPV} \tag{10}$$

$$TSS = \frac{TP \times TN - FP \times FN}{(TP + TN)(FP + FN)} = TPR + TNR - 1 \quad (11)$$

K has been extensively used in research as it is specifically designed to correct the overall accuracy of the classification by the accuracy induced by chance. This is quantified by the first and second term in the enumerator of Eq. (9). As an equally conservative measure of classification performance, *F* measures the trade-off between sensitivity (*TP* rate) and precision (*PPV*) by the harmonic mean of these descriptive accuracy terms. Thus, *F* is essentially a class-specific quality measure adding to the results of *K* by penalizing both missed detection and false alarm within the urban domain (Labatut & Cherifi, 2011). Although both of these comprehensive measures are commonly applied in the published literature, they are critically discussed as being sensitive to imbalanced datasets (Jeni, Cohn, & Torre, 2013). This is commonly the case for settlement areas as they cover – especially for large-scale perspectives on urban, peri-urban and rural landscapes – only a small portion of the entire test site extent. To encounter this potential weakness, we additionally employ *TSS* which corrects for the dependence on prevalence while still keeping all advantages of *K* (Allouche et al., 2006).

The proposed set of measures enables a systematic and comprehensive description of absolute classification performance with regard to the described reference data base. In this context, an added value arises from the simultaneous assessment of the layers under test with regard to both the building (*BRef*) and the settlement mask (*BURef*). This allows for a reverse exploration of each layer's thematic definition. Beyond, we integrate the spatial zoning via GRUMP into both PIP-tests and the pixel-based assessments to further determine landscape-specific accuracy estimates.

4.3. Medium resolution analysis

On the MR scale of analysis, we conduct a final pattern-based assessment to give a structured insight into the mapping capabilities of GUF and GHSL (Fig. 5). We study absolute accuracy statistics as a function of selected physical features of the built environment. Thus, we are able to account for the physical heterogeneity of the settlement fabric in the assessment and refine our understanding of layer-inherent accuracy variations. Spatial aggregation is used to increase the level of abstraction of the thematic information content. This allows us to assess and compare spatial functions derived from the built-up representations in the map and the reference.

We follow this conception in two ways: (1) We investigate *density* functions that describe the urban fabric represented by each HR layer using linear regression techniques. This approach has been proposed in previous work (e.g., Taubenböck et al., 2011; Ouzounis et al., 2013) and is originally motivated in the exploration of the influence of physical density variations on the classification output. (2) Beyond, we study classification differences with regard to the *size of settlement patches* to establish a stronger understanding on the scale-dependent mapping capabilities of each layer. In doing so, we present a novel object-based approach that exploits information of the mutual overlap between mapped and referenced settlement patches.

(1) The objective of the first pattern-based assessment lies in the quantification how well density metrics derived from the independent reference are described by the automatically-derived settlement layers GUF and GHSL. We derive continuous measurements representing building (*BRef*) and settlement (*BURef*) densities from the reference. Using Eq. (1), both reference and HR settlement layers are aggregated at a spatial resolution of 200 m. This scale has been found to feature a reasonable trade-off between generalization and fitting quality by Pesaresi, Halkia, and Ouzounis (2011) and Ouzounis et al. (2013), as well as independent investigations. The authors showed a strong increase of fitting quality for the density function through stepwise aggregation up to a scale of 200 m (R^2 ~ 80%) and distinct saturation at coarser scales.
Given the density aggregates derived from GUF and GHSL, we explore their correlation with both the building (*BRef*) and the settlement (*BURef*) density. To support the analysis of the scatterplots of these bivariate distributions, we compute the Pearson coefficient of correlation *r* from first order linear regression (Everitt, 2002):

$$\mathrm{r} = \sqrt{1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}} \quad (12)$$

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(d_i - \hat{d}\right)^2 \quad (13)$$

$$TSS = \sum_{i=1}^{n} \left(\mathrm{d}_i - \overline{d}\right)^2 \quad (14)$$

where *RSS* and *TSS* are the residual and total sums of squared errors, respectively, *n* is the total number of observations, $\overline{d}$ is the average value of the generalized reference *D* and *d* is the response of the regression model.

(2) In a second pattern-based assessment, we further analyze classification variation with regard to the size of adjacent built-up areas. On this patch level, we evaluate the mutual overlap between mapped and referenced settlement patches (*BURef*). This approach has previously been applied for the evaluation of building extraction protocols from HR airborne or satellite sensors (e.g., Rottensteiner, Trinder, Clode, & Kubik, 2005; Rutzinger et al., 2009; Wurm et al., 2014). However, compared to the analysis of building footprints, an unambiguous one-to-one allocation between settlement patches of the reference and the map is not feasible by GIS-based procedures. Thus, we establish a many-to-many relationship, initially merging all adjacent settlement patches of each layer. Subsequently, the overlap between a reference patch $p_r \in P_r$ and its intersecting mapped patches $p_m \in P_m$ is computed in the way that $o_{rm} = a_{r \cap m}/a_r$. In turn, the overlap between a mapped patch and its corresponding reference patches is described as $o_{mr} = a_{m \cap r}/a_m$. From this, we proceed by classifying the percent mutual overlap between patches according to Rottensteiner et al. (2005):

$$o = \begin{cases} \text{none} & & o_{\mathrm{rm}}/o_{\mathrm{mr}} \leq 10\% \\ \text{weak} & 10\% & < o_{\mathrm{rm}}/o_{\mathrm{mr}} \leq 50\% \\ \text{partial} & 50\% & < o_{\mathrm{rm}}/o_{\mathrm{mr}} \leq 80\% \\ \text{strong} & 80\% & < o_{\mathrm{rm}}/o_{\mathrm{mr}} \leq 100\%. \end{cases} \quad (15)$$

To eventually obtain a measure of the accuracy with regard to patch size variation, we compute the completeness (*TP* rate, Eq. (6)), correctness (*PPV*, Eq. (8)) and overall quality (*F*, Eq. (10)) of each layer for pre-defined patch size bins. To do so, we count the respective numbers of *TP*, *FN* and *FP* for each bin based on a threshold $T_o = 50\%$ to account for both completely or partly corresponding patches of the maps and the reference:

- TP_{comp} ($o_{rm} > T_o$): number of reference patches that are either partly or completely mapped;
- TP_{corr} ($o_{mr} > T_o$): number of the mapped patches that are either partly or completely referenced;
- *FN* ($o_{rm} \leq T_o$): number of reference patches that are not (partly) mapped;
- *FP* ($o_{mr} \leq T_o$): number of mapped patches that are not (partly) referenced.

Fig. 6 gives a schematic exemplification of various cases of patch relationships and their resulting classification applying a threshold of $T_o = 50\%$. The calculation of consistent quality metrics for each patch size bin allows us to asses each product's capabilities in capturing the scale-dependent complexity of the settlement pattern.

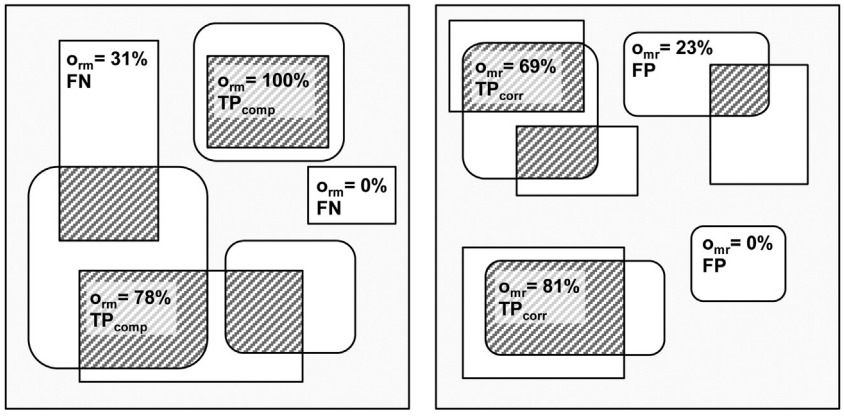

Fig. 6. Schematic display of possible patch relationships (sharp corners: referenced patches, round corners: mapped patches), calculation of the percent mutual overlap and the resulting binary classification TP_{comp}/FN and TP_{corr}/FP, respectively, using a threshold value T_o of 50%.

5. Results

Fig. 4 depicts the map representations of GUF, GHSL, GLOBC and MOD500 across the selected test sites allowing for a first visual comparison of all maps under study. The map representations of GUF and GHSL seem quite similar in overall pattern and extent, but are somewhat more expansive compared to MOD500 and expecially, GLOBC. More specifically, high spatial detail and fragmentation of the settlement pattern, especially beyond the urban core areas, is contrasted by limited numbers of patches and spatial complexity of LR layers. Spatial statistics displayed further enable a quantitative comparison of gross settlement area estimates with the reference. It can be seen that all maps significantly over-represent the building mask (*BRef*). Over-estimation ranges from 231% and 318% for the most conservative estimate by GLOBC, to maximum values of 587% and 612% by GHSL. Lower disagreement is found with regard to the settlement mask (*BURef*). While GUF and GHSL roughly double the extent of the referenced settlement areas of both sites, the estimates by the LR layers approximate the reference more closely, especially for the rural Tuscany region.

Although these numbers are assumed to be related to layer-inherent errors of commission or omission, respectively, the non site-specific nature of this assessment cannot ascertain locational agreement between classifications and the reference. The following sections present the results from multi-scale cross-comparison that allow for site-specific conclusions on each thematic map's value.

5.1. Low resolution analysis — relative inter-map comparison

We first investigate if the significantly increased spatial resolution of recent developments in global settlement mapping translates to enhanced mapping capabilities. The inter-map agreement between pairs of HR and LR layers serves as valid indicator marking areas of potential map evolution in the HR maps' density domain. PR curves displayed in Figs. 7 and 8 depict the trade-off between precision (*TP rate*) and recall (*PPV*) of the aggregated HR layers with regard to GLOBC and MOD500, respectively. Complementary plots within these figures show the course of *ER* across the entire threshold domain as well as the empirically derived optimal threshold values $T(ER_{min})$.

The PR curves reveal that only moderate correspondence exists between all pairs of HR and LR layers on the test site level. None of the curves approaches the top right-hand corner of the PR space that marks optimum recall and precision. Based on the area under the PR curve and the visual impression from Fig. 4, GUF and GHSL exhibit only minor differences compared to a particular LR layer and across test sites. This indicates a substantial degree of agreement between HR layers. Both correspond slightly better to GLOBC which presents, especially for Cologne, the more conservative estimate of urban land. Beyond, significantly stronger inter-map agreement for all map combinations is evident for the Cologne site as the respective curves dominate Tuscany's for the largest part of the threshold domain. This is presumably due to Cologne's significantly higher

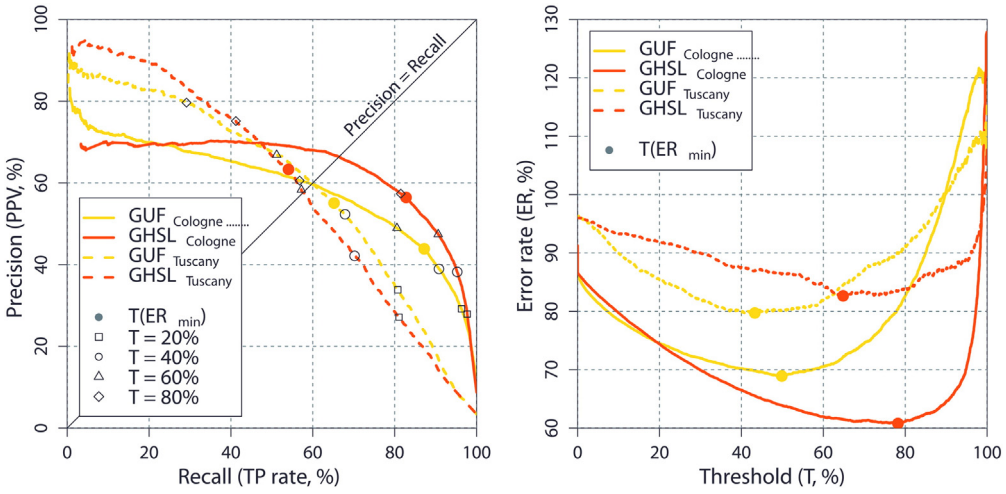

Fig. 7. PR curve (left) and threshold plot (right) displaying the overall correspondence between HR settlement information by GUF/GHSL against GLOBC; dots mark empirically derived threshold values $T(ER_{min})$.

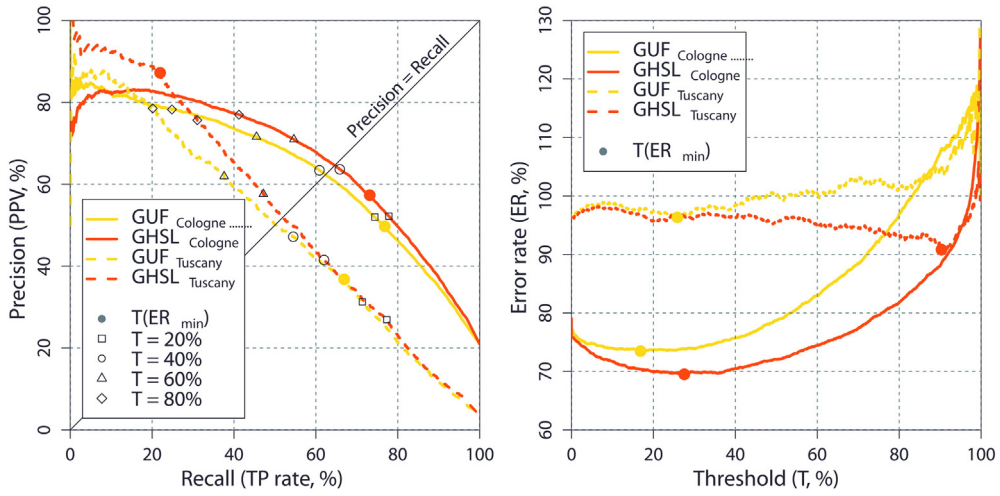

Fig. 8. PR curve (left) and threshold plot (right) displaying the overall correspondence between HR settlement information by GUF/GHSL against MOD500; dots mark empirically derived optimal threshold values $T(ER_{min})$.

settlement density (Table 2) and total share of urban land (Fig. 4) that are reflected more extensively by LR layers on the test site level.

Despite the limited overall agreement, a distinct density-dependent relationship between HR and LR layers is visible. This is revealed by the examination of particular regions of the PR space. As *T* decreases all curves move from conservative areas (low recall, high precision) of the PR space to the liberal ones (high recall, low precision). In this, PR curves of Cologne feature a distinct convex profile. With the lowering of *T* to high and subsequently medium cut-off values, a strong increase of *TP* marks a substantial rise of recall at the cost of an only moderate loss of precision. With regard to the density domain, this indicates substantial agreement between HR and LR layers for medium to high magnitudes, presumably larger urban areas featuring continuous urban fabric. With the continued shifting of *T* to lower densities, increasing *FP* occurrences are observed in favour of an only moderate further gain in recall. Consequently, strong degradation of precision and thus, overall agreement, reflect decisive over-representation of the LR settlement extents by HR layers in the low density range. Although the density margin of maximum agreement is less pronounced for the Tuscany site due to limited overall agreement, a similar tendency is evident by the slight convex form of the respective PR curves. In line with these observations, complementary threshold plots show a visible decline of *ER* with increasing densities for all map combinations reaching distinct minima in the medium and high density range. Due to the spatially more conservative delineation of urban areas by GLOBC, these minima are found within higher density margins as compared to MOD500. Nevertheless, all observations strongly indicate that the spatial reduction of GUF and GHSL to medium and high density areas, respectively, supports stronger inter-map agreement. In contrast, correspondence in low density areas seems generally limited.

To quantitatively confirm these observations, we study respective error statistics – ER_{min} and $T(ER_{min})$ – derived from PR space in more detail (Tables 4 & 5). In line with the limited agreement observed in PR space, we find substantial error rates exceeding 60% for all pairs of layers. It is noteworthy that error rates for both GUF and GHSL yield very similar magnitudes with regard to each LR layer and test site — a further indication of the marked consistency between these layers. For Cologne, ER_{min} between HR and LR layers ranges from 60% to 73%, whereas substantially higher magnitudes of error between 80% and 96% are observed in Tuscany in line with the course of the respective PR curves. Across sites, $T(ER_{min})$ exceeds 43% (GUF) and 65% (GHSL) against GLOBC, as well as 17% and 28% against MOD500. This proofs that lower density built-up areas of GUF and GHSL below these empirically derived thresholds must be largely disregarded to preserve good inter-map agreement. PR statistics for urban and rural areas (Tables 4 and 5) by incorporating GRUMP for spatial zoning further add to these results by landscape-specific qualification of map agreement. Lower values of ER_{min} and thus, stronger agreement, in high density urban areas support the previous findings. The respective optimum thresholds are naturally lower as compared to the overall test site perspective. In contrast, maximum correspondence in rural areas is rather poor and associated with very high threshold values. Thus, best agreement between HR and LR layers in these regions is theoretically established by masking out the largest parts of settlement areas mapped by GUF and GHSL.

Overall, these findings establish a clear proof of concept and understanding of map evolution with regard to structural and landscape-specific image regions. Although map combinations exhibit moderate agreement within medium to high density margins of HR layers, it is evident that HR layers identify built-up areas in regions where LR do not provide this capability. These regions are identified to be primarily low to medium density rural areas beyond the extents of LR products. As these landscape characteristics however mark the general nature of disagreement between layers, map evolution can also be expected in terms of precision for other image regions of lower densities: E.g., peri-urban areas, at the urban fringe, in low density inner-urban districts or close to inner-urban spaces of unoccupied land. As the spatial generalization of HR layer however, reduces the degree of site-specificity in this analysis, it remains to be seen if these disparities

Table 4
Overall and landscape-specific statistics of inter-map agreement between HR layers and GLOBC.

		GUF			GHSL		
		Overall	Urban	Rural	Overall	Urban	Rural
Cologne	ER_{min} (%)	68.92	62.15	88.28	60.78	55.23	96.93
	$T(ER_{min})$ (%)	49.87	48.74	69.12	78.22	67.43	78.85
Tuscany	ER_{min} (%)	79.71	71.16	98.52	82.65	74.57	98.89
	$T(ER_{min})$ (%)	43.24	42.95	97.82	64.82	64.79	98.44

Table 5
Overall and landscape-specific statistics of inter-map agreement between HR layers and MOD500.

		GUF			GHSL		
		Overall	Urban	Rural	Overall	Urban	Rural
Cologne	ER_{min} (%)	73.47	54.52	95.56	69.52	54.45	94.15
	$T(ER_{min})$ (%)	16.85	8.56	83.61	27.54	14.00	95.05
Tuscany	ER_{min} (%)	96.33	82.05	96.20	90.05	80.97	98.68
	$T(ER_{min})$ (%)	25.84	5.32	90.72	29.61	8.51	96.70

translate to absolute accuracies with regard to reliable reference data in subsequent analysis.

5.2. High resolution analysis — absolute accuracy assessment

The previous analysis has proven significant overall and landscape-specific disparities between HR and LR settlement information. However, the relative comparison does not reveal to what extent the various products detect settlement features correctly. To obtain knowledge about the absolute accuracy, especially for HR layers, we analyze statistics computed from PIP-test. This is followed by the analysis of the selected set of pixel-based accuracy metrics derived from overlay with the building (*BRef*) and the settlement mask (*BURef*).

PIP-statistics displayed in Fig. 9 show that the HR settlement maps, GUF and GHSL, paint a very complete picture of Cologne's building inventory (900,871 buildings) capturing the lion's shares of 87% and 90%, respectively. In contrast, they omit roughly one-third of the buildings in Tuscany (364,544). These increased omission rates are most probably related to the smaller sizes of settlements and stronger fragmentation in this region. Nevertheless, these numbers are opposed by significantly larger shares of unmapped buildings by GLOBC and MOD500. While MOD500 captures at least close to half of the building centroids for Cologne, more conservative GLOBC reaches only 25%. For the more scattered settlement patterns of Tuscany both layers omit an even larger share of buildings (~80%). These basic descriptive measures establish a first impression regarding the completeness of the mapped settlement pattern with regard to the core features of human settlements.

More robust and differentiated information about the accuracy of the classification results can be retrieved from the extended set of pixel-based accuracy measures presented as bar charts in Fig. 10. Overall accuracies displayed in Fig. 10a imply high accuracies and only insignificant differences between all layers. With regard to *BRef*, layer-specific accuracies lie well above 80% for Cologne and 90% for Tuscany despite large differences in spatial resolution. As *A*, however, does not respect class-specific errors, further descriptive statistics enable a more differentiated assessment of the thematic maps' value. Sensitivity (*TP* rate) and specificity (*TN* rate) give information about the completeness, whereas precision (*PPV*) exposes the correctness of classifications. Building footprints of the reference amount to very small shares of the test sites' aerial extents, i.e., five percent for Cologne and one percent for Tuscany. Considering these low prevalence rates, overall high degrees of specificity across sites constitute the general ability of each approach to identify non-urban areas. In contrast, significant differences arise between HR and LR layers from the examination of sensitivity. While GUF and GHSL detect more than 77% of building pixels across sites, GLOBC and MODIS paint a less complete picture of the building stock with a *TP rate* in the range of 35% and 40%. This is in line with the results from the previous PIP-analysis. Contrasting these still moderate magnitudes of omission, it is obvious that all layers essentially do not feature the capabilities to map individual building outlines correctly. Precision barely reaches 15% in Tuscany and just exceeds 20% in Cologne testifying a high share of mapped pixels not belonging to the building mask. This is due to the fact that all layers essentially define built-up or settlement areas as aerial extents including spaces in between buildings. These findings translate to only poor to moderate classification accuracies in terms of *K* and *F* across test sites, whereby HR layers perform slightly better. It is worth noticing that the improved spatial resolution of HR layers barely reflects in *K* and *F*. Only *TSS* indicates decisive differences in classification accuracy of approximately 36% between HR ($\overline{TSS} = 0.75$) and LR ($\overline{TSS} = 0.39$) layers averaged across test sites. This is due to the fact that *K* and *F* respond to low prevalence by maximizing values for GLOBC and MOD500 when specificity exceeds sensitivity (Allouche et al., 2006; Jeni et al., 2013). In contrast, *TSS* is widely insensitive to prevalence.

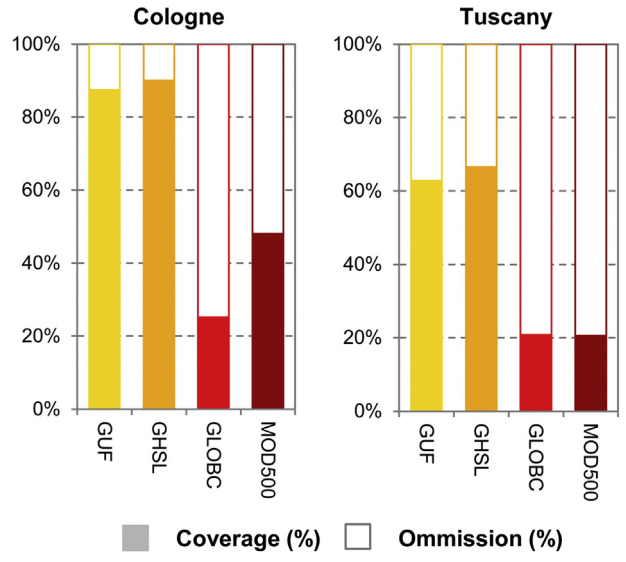

Fig. 9. Overall PIP-statistics for all layers under study.

With regard to the settlement mask (Fig. 10b), quantitative metrics exhibit very similar orders and magnitudes in terms of sensitivity, specificity and overall accuracy as compared to the building mask. The spatial generalization of the building reference to a settlement mask (*BURef*), however, results in improved precision of the classifications, especially for GUF and GHSL. In this regard, remaining over-classification by GLOBC and MOD500 is naturally related to these layers' coarse spatial resolution. For the GUF, errors of commission have been found to mainly relate to false alarm in image regions featuring texture characteristics similar to that of built-up areas such as rugged terrain (Esch, Marconcini, Marmanis et al., 2014) or to the horizontal displacement of strong backscattering signals detected as urban seeds next to vertical structures (Taubenböck et al., 2011). In contrast, GHSL responds in this to the spatial generalization of PANTEX accompanied by incorrect commission of objects resembling built-up textures (e.g., excavations, construction sites, etc.) (Wania, Kemper, & Tiede, 2014). Due to reduced prevalence rates – ca. 12% for Cologne and 3% for Tuscany – differences of accuracies between HR ($\overline{K} = 0.58$; $\overline{F} = 0.42$) and LR ($\overline{K} = 0.31$; $\overline{F} = 0.24$) layers now reflect also in *K* ($\Delta\overline{K} = 0.17$) and *F* ($\Delta\overline{F} = 0.18$) across test sites. With regard to the HR settlement reference, GUF ($\overline{F} = 0.55$; $\overline{K} = 0.51$; $\overline{TSS} = 0.72$) and GHSL ($\overline{F} = 0.53$; $\overline{K} = 0.48$; $\overline{TSS} = 0.75$) reach moderate to substantial absolute accuracies. In contrast, GLOBC ($\overline{F} = 0.37$; $\overline{K} = 0.28$; $\overline{TSS} = 0.28$) and MOD500 ($\overline{F} = 0.33$; $\overline{K} = 0.27$; $\overline{TSS} = 0.34$) do not exceed fair accuracies. This is mainly due to their coarse geometric resolution resulting in an average 42%-difference with regard to HR layers in terms of *TSS*. These results give clear evidence that HR settlement layers feature significantly improved completeness and correctness. Beyond, the simultaneous assessment of the layers under test with regard to both *BRef* and *BURef* enables a reverse exploration of each HR layer's semantic definition. In this regard, GUF and GHSL draw a generalized outline of the spatial building distribution due to the stronger correspondence to the settlement mask.

From previous analysis, we can assume that layer-specific accuracies are very consistent across test sites. From these results, we move on to the analysis of landscape-specific statistics incorporating GRUMP's urban and rural classes into the analysis for spatial zoning (Fig. 11). PIP-tests on the landscape level clearly reveal higher degrees of completeness for GUF and GHSL as compared to GLOBC and MOD500, particularly in the rural parts of the selected test sites. Although both HR layers feature a slightly higher building share in urban areas, they still capture more than 80% of all buildings in rural areas of Cologne (249,910 buildings) and almost 50% in Tuscany (133,436). In contrast, GLOBC and MOD500 show significantly lower shares in urban areas

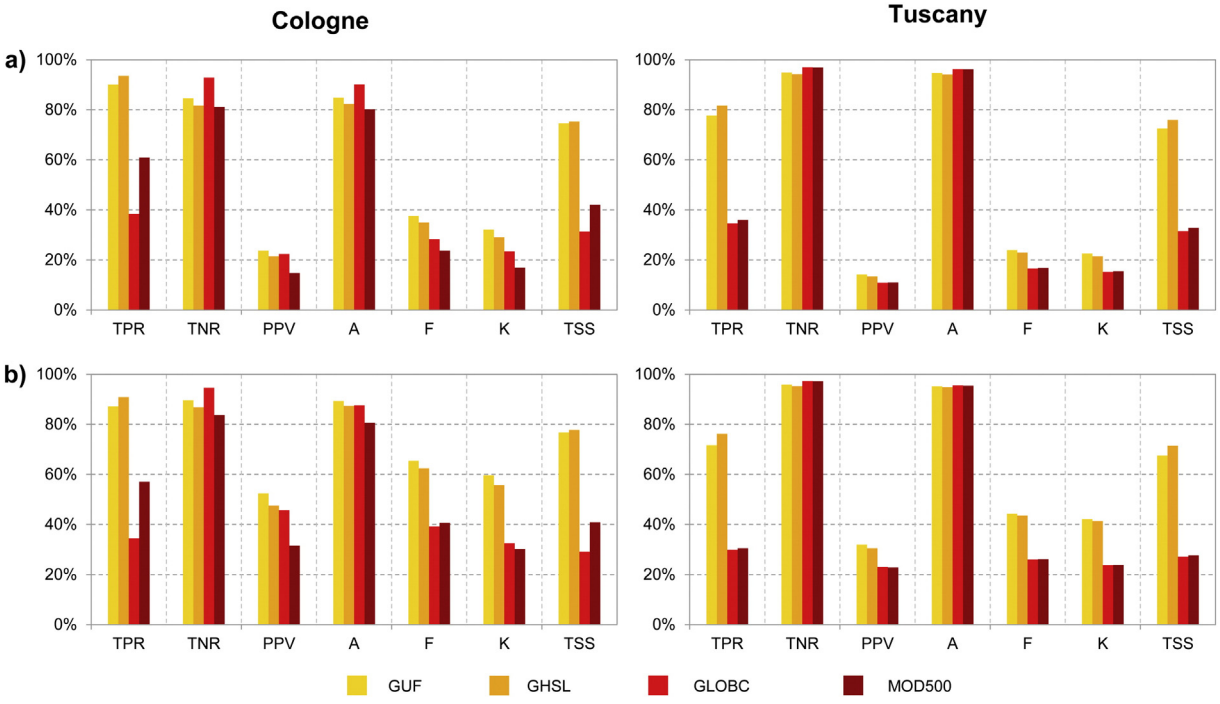

Fig. 10. Overall pixel-based accuracy measures for all layers under study compared to a) *BRef* and b) *BURef* (*F*, *K* and *TSS* are rescaled to 100%).

and almost entirely neglect buildings in rural areas. Their omission rates constantly exceed 88% across sites. While, GUF and GHSL again show high consistency, GLOBC presents more conservative estimates in terms of building coverage as compared to MOD500. These results – both analogies and differences – are well in line with our findings from PR analysis.

We subsequently analyze pixel-based map accuracies at the landscape level for both test sites (Fig. 11). As the thematic and spatial representations of the maps have been proven to correspond more closely to the semantic definition of the generalized settlement mask (*BURef*), we disregard the building mask (*BRef*) at this point. Naturally, we again find moderate to substantial agreement for both GUF ($\overline{F} = 0.59$; $\overline{K} = 0.50$; $\overline{TSS} = 0.68$) and GHSL ($\overline{F} = 0.56$; $\overline{K} = 0.46$; $\overline{TSS} = 0.71$) in the urban domain. In this, they outperform the respective LR layers as their advantageous spatial resolution allows more accurately tracing the spatially detailed outlines of built-up areas as marked by higher sensitivity and precision. This indicates decisive map evolution in areas located at the urban fringe or close to inner-urban open spaces. Nevertheless, both GLOBC ($\overline{F} = 0.38$; $\overline{K} = 0.28$; $\overline{TSS} = 0.30$) and MOD500 ($\overline{F} = 0.38$; $\overline{K} = 0.23$; $\overline{TSS} = 0.33$) perform significantly better than chance manifesting their potential for global analysis of larger urban areas. Maximum discrepancies, however, exist between HR and LR layers

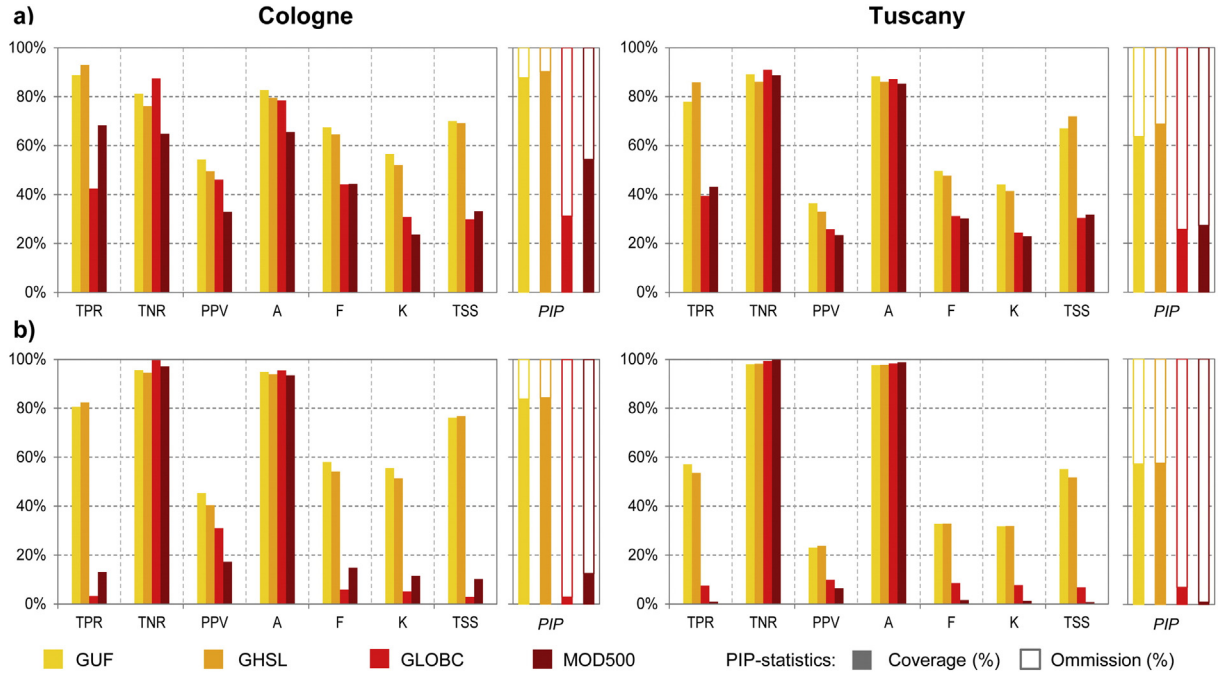

Fig. 11. Landscape-specific PIP-statistics and pixel-based accuracy measures for all layers under study for a) urban b) rural areas (*F*, *K* and *TSS* are rescaled to 100%).

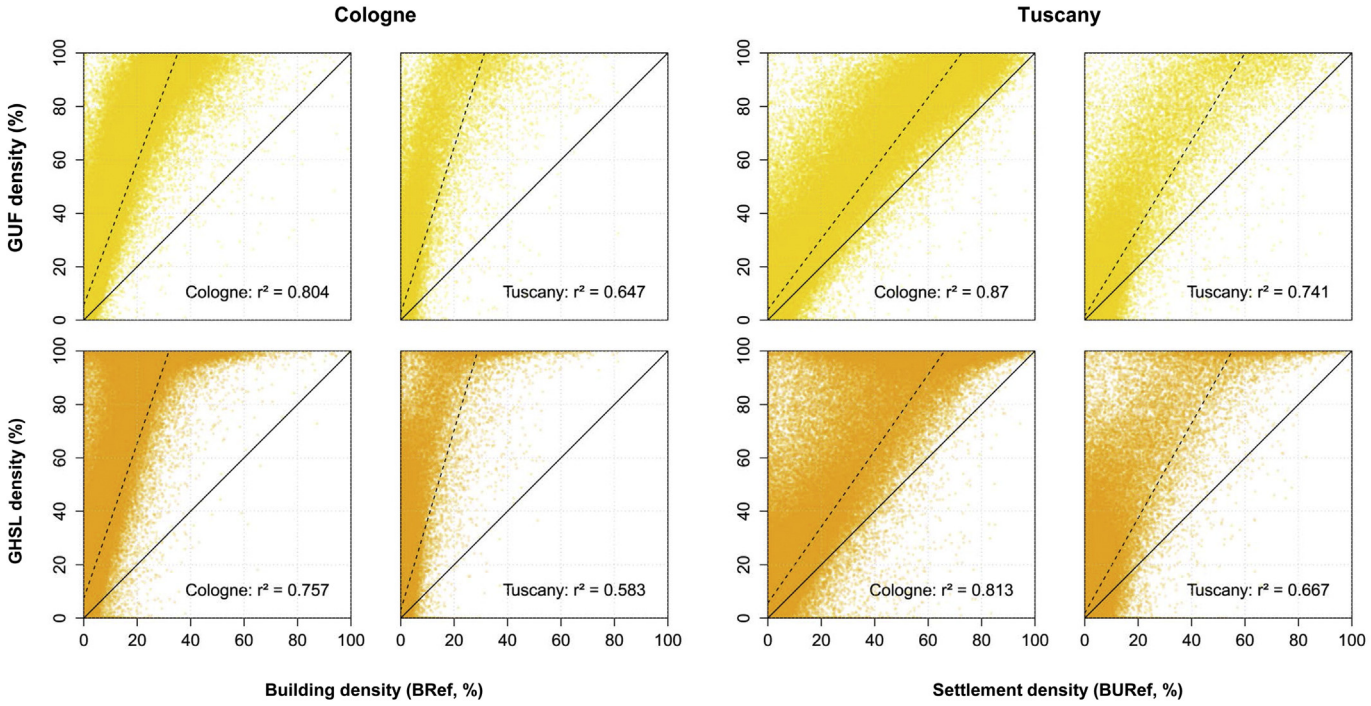

Fig. 12. GUF (top) and GHSL (bottom) as a function of building (*BRef*; left) and settlement density (*BURef*, right) displayed for both test sites at an aggregation scale of 200 m.

in the rural setting. Here, HR layers significantly exceed the geometric capabilities of existing LR products which are too coarse to detect, and too generalized to delineate, small-scale fragmented settlement patches. This is testified by significantly reduced sensitivity and precision of LR layers. In contrast, both GUF ($\overline{F} = 0.45; \overline{K} = 0.44; \overline{TSS} = 0.66$) and GHSL ($\overline{F} = 0.44; \overline{K} = 0.41; \overline{TSS} = 0.64$) exhibit only relatively little performance loss compared to urban areas. Moderate to substantial agreement found with regard to the reference underpin their extended applicability in rural areas. This verifies our findings from Section 5.1 in the way that regions of disagreement identified in PR analysis largely correspond to correctly detected settlement areas by GUF and GHSL. These areas have been traditionally neglected by GLOBC and MOD500. Consequently, average values of *F*, *K* and *TSS* are below 10% for both LR layers.

5.3. Medium resolution analysis – pattern-based accuracy assessment

The previous results have underlined the enhanced capabilities of new developments in global settlement mapping. However, these standard accuracy assessments did not consider spatially altering structures of urban, peri-urban and rural environments. To explore pattern-based dependencies between the classifications and the physical settlement structure, we present the results from two final absolute assessments of GUF and GHSL that respect the physical variability of the landscape pattern by the notations of (1) *urban density* and (2) *settlement patch size*. When applicable, GLOBC and MOD500 again function as benchmarks.

(1) The first pattern-based assessment focuses on the relationship between density metrics derived from the independent reference and the automatically-derived HR settlement layers. Scatterplots in Fig. 12 display the correlations of density measurements from GUF and GHSL against the observed building (*BRef*) and settlement (*BURef*) densities at the empirically found aggregation scale. As a functional descriptor, a linear regression model is fitted to the point clouds. The quality of each model and thus, the degree of correlation, is described by the coefficient of correlation *r* that quantifies the share of the real world structural variability explained by the respective map.

From the scatterplots on the left side of Fig. 12 relating to *BRef*, it can be generally retained that both layers significantly overclassify real-world building densities. The approximation by the linear regression lines shows increasing over-estimation with increasing building densities when compared to the ideal trend line that presents a theoretical maximum of 100% explained variability. This pattern-based dependency is due to the layer-inherent semantic definition of settlement areas that does not comply with individual building outlines (see Section 5.2). Once again, both layers perform considerably better for Cologne, where more than 80% of the structural variability is explained by GUF and 76% by GHSL. This is due to locational, morphological and structural characteristics. The Cologne site features a highly structured building pattern in its large-scale urban agglomeration located in a flat region. In contrast, Tuscany exhibits a more diverse and fragmented arrangement with fuzzy transitions in a more rugged terrain facilitating the classification process. Although both layers show significant over-classification and quantitative inconsistencies, especially in high density categories (Table 6), *r* still reveals that both GUF and GHSL present at least systematic first-level proxies for a two-class distinction of high and low building densities.

The correlation with built-up or settlement densities derived from *BURef* represented on the right-hand side of Fig. 12 implies a more accurate representation of the structural variability by GUF and GHSL. This is manifested by higher *r* values ranging between 0.67 and 0.87 across sites. Maximum values are again found for Cologne. Mean densities in Table 6 endorse this finding by more consistent orders of magnitudes of the mapped densities. Nevertheless, again reduced but explicit over-estimation is found, particularly in the medium density range between 20% and 80%. These are due to limitations in terms of precision identified for both GHSL and GUF in Section 5.2. This corresponds to the findings of Taubenböck et al. (2011) who evaluated pattern-based accuracies for the GUF for a city region in Indonesia. In contrast, both classification approaches work more accurately in areas with extremely high or extremely low densities. Thus, a clear pattern-based dependency is proven in

Table 6
Trends of over-classification with regard to categorized building (*BRef*; top) and settlement (*BURef*; bottom) density classes by GUF and GHSL.

Reference	Density (%)	Area (km²)		GUF density (%)			GHSL density (%)		
		Cologne	Tuscany	Cologne	Tuscany	Mean	Cologne	Tuscany	Mean
Building density (*BRef*)	0	5939	6902	1.67	0.89	1.28	2.91	0.88	1.90
	>0–20	2996	2974	30.69	14.20	22.44	34.99	16.23	25.62
	20–40	938	105	84.45	83.10	83.77	92.58	88.92	90.75
	40–60	118	16	90.13	84.99	87.55	97.22	93.96	95.58
	60–80	8	2	86.32	87.35	86.84	96.24	98.93	97.58
	80–100	1	1	86.22	72.48	79.35	97.10	98.44	97.76
Settlement density (*BURef*)	0	5788	6734	1.57	0.85	1.21	2.78	0.82	1.80
	>0–20	2166	2913	17.10	9.69	13.39	20.17	11.43	15.80
	20–40	758	208	52.62	60.37	56.49	59.37	65.44	62.40
	40–60	632	93	75.72	80.99	78.35	84.46	86.84	85.66
	60–80	548	43	88.68	89.81	89.25	95.97	95.13	95.55
	80–100	108	9	93.68	91.20	92.44	98.41	98.76	98.59

the way that accuracies of GUF and GHSL are sensitive to built-up density variations. Consequently, we can retain that both GUF and GHSL hold – on a spatial aggregation level – the potential for structural qualification of the built environment beyond binary formats.

(2) In addition to the analysis of building and built-up densities, we analyze the dependence of the classification results with respect to the varying size of settlements at the object level. Beyond the exploration of pattern-dependent characteristics of the classifications, we aim at a user-oriented quantification of accuracies that can be expected with regard to a particular scale of analysis that could be defined by a minimum settlement size. Fig. 13 a displays patch size frequencies of all maps under study and the reference normalized to each layer's total number of patches. The abscissa uses an exponential scaling of patch size bins to a base of two. In this, the minimum patch size of one pixel (144 m^2) corresponds to the geometric resolution of *BURef* (12 m). For the ease of comparison the patch sizes are further grouped into three patch size ranges using equal intervals of the exponential domain. It can be seen that both GUF and GHSL show frequency-size distributions very similar to the reference with maxima in the lower patch size range between 2^2 and 2^5 pixels ($\approx$500 m^2–5000 m^2) and exponential decay approaching the abscissa with increasing patch sizes. Only GUF features a second maximum of patches consisting of single pixels, presumably isolated buildings or false positives. In contrast, LR layers cover only the medium and large patch size range due to their coarse geometric resolution. Beyond, they show a trend of log-linear decay in line with the recognized rank-size rule of city sizes (Potere & Schneider, 2009).
Results of the evaluation on the patch level are presented in Fig. 13b to d. The line charts show completeness (*TP* rate), correctness (*PPV*) and overall quality of the classification (*F*) plotted against the range of patch size bins. In terms of completeness (Fig. 13b), we find strong analogies between GUF and GHSL as well as between GLOBC and MOD500. HR layers exhibit a steady gain in completeness, i.e., the share of reference patches at least partially detected, from small- to medium-sized patches. They reach substantial agreement (*TP* rate > 60%) at 2^4 pixels ($\approx$2,300 m^2) and perfect agreement at the transition from medium to large patches (2^{11} pixels $\approx$ 0.3 km^2). In contrast, LR source data of GLOBC and MOD500 do not enable the detection of most small- and medium-scale settlements. Thus, they exhibit only low completeness rates (<40%) up to a patch size that corresponds to their native spatial resolution (2^9 to 2^{10} pixels). For larger patches, completeness values slowly increase reaching substantial agreement not until a size of 2^{12} pixels ($\approx$0.6 km^2). Consequently, perfect completeness is only reached for maximum patch sizes.
Simultaneously, the analysis of correctness (Fig. 13c) reveals information about the patch size-dependent precision of classifications. Again, both HR layers show very similar trends. 70% or more of all mapped patches consisting of only very few pixels ($\leq 2^4 \approx$ 2300 m^2) present, however, false positives. This fact can be mainly attributed to misclassification and partly to spatial mis-match between mapped and the referenced patches. Consequently, low correctness is evident for small settlements consisting of only a few buildings. Nevertheless, a strong gain in correctness is observed in the medium patch size range approximating, and ultimately reaching, 100% for medium to large settlement sizes. In correspondence with the frequency-size distributions (Fig. 13a), GLOBC and MOD500 are naturally too generalized to correctly represent small- to medium-sized settlements up to a size of 2^9 to 2^{10} pixels ($\approx$70,000 m^2–0.15 km^2). Nevertheless, they still possess capabilities in mapping settlements larger than their squared native spatial resolution in good correctness according to the applied mutual overlap threshold.
The overall quality of each map with regard to patch size is depicted in Fig. 13d by the course of *F* which is computed as the harmonic mean of correctness and completeness. Thus, the F-score presents a conservative measure of map quality. Reflecting the previous findings, for all layers a distinct pattern-based quality gain is observed with increasing patch size. While HR maps feature distinct limitations in terms of correctness for the range of smaller patches up to 2^4 pixels ($\approx$2,300 m^2), they exhibit a decisive increase in map quality in the medium patch size range, that GLOBC and MOD500 naturally neglect. Substantial agreement (*F* > 60%) with referenced patches from *BURef* is reached at a size of 2^6 pixels ($\approx$10,000 m^2). This size corresponds to an edge length 100 m of an idealized square patch. As GUF and GHSL promote both higher degrees of sensitivity and precision over the entire patch size range, this results in an improved scale-dependent representation of spatially detailed settlement patterns. This is particularly evident fore medium to large patch sizes for which HR layers approach map qualities of 100% at smaller magnitudes than LR products.

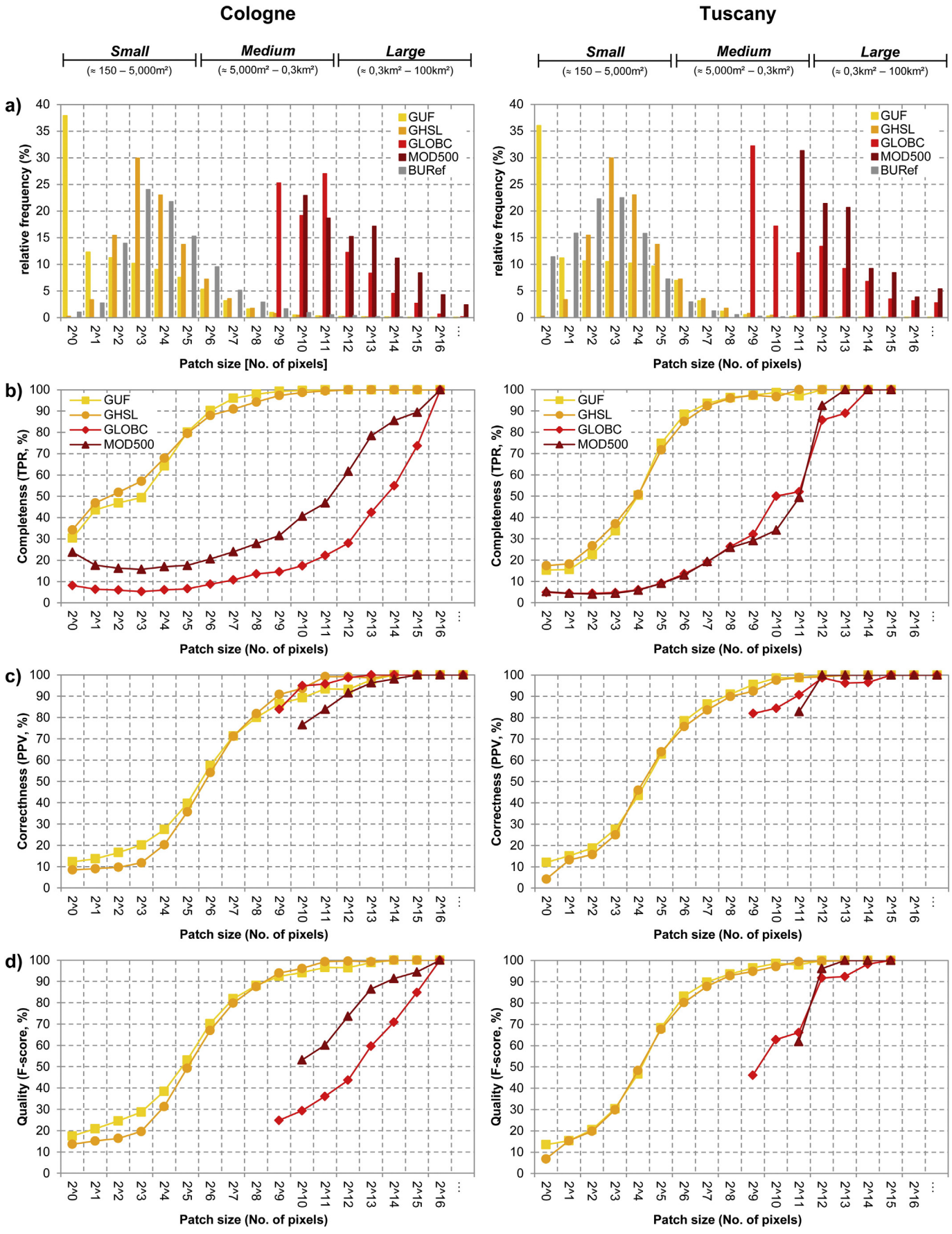

Fig. 13. Analysis of patch size-dependent b) completeness (*TP rate*), c) correctness (*PPV*) and overall quality (*F*) values for all layers under test; a) relative frequency-size distributions display the range and occurrences of patch sizes for each layer.

6. Summary & discussion

For the presentation and discussion of the main findings from *multi-scale cross-comparison* it is referred to the research questions addressed in Section 1 of this work:

(1). How and to which degree do new high resolution settlement layers correspond to existing global products of lower geometric resolution in a Central European setting?

Using PR curves to investigate the inter-map agreement between new and existing urban maps, we find a distinct potential added value in recent settlement mapping efforts. While pairs of HR and LR layers moderately correspond in core, high density urban areas, we demonstrate that GUF and GHSL clearly detect urban areas beyond the extents of GLOBC and MOD500. Respecting landscape-specific differences of the selected test sites we localize regions of strong map disagreement predominantly in low and medium density, rural and peri-urban areas. These have so far been neglected by global maps due to their coarse geometric resolution. Based on this finding, GUF and GHSL can further be expected as potentially more accurate in other low density spaces of urban environments (e.g., areas at the urban fringe, inner-urban low density spaces, etc.) — both in terms of precision and completeness. Results from this relative map comparison give a profound understanding of the disparities between new HR and existing LR settlement maps.

(2). How accurate are different – high and low resolution – global geo-information layers in absolute terms regarding the representation of complex settlement features and their spatial configuration in Central Europe?

We define settlement areas by spatially two explicit features of these environments, i.e., buildings and built-up areas. From absolute pixel-based accuracy assessment and PIP-tests with regard to the building mask, we find that GUF and GHSL feature significantly improved completeness in terms of mapping the spatial buildings distribution. Nevertheless, they generally lack the capabilities to map outlines of individual buildings due limitations in terms of precision. Although highly resolved, they rather paint a generalized picture of the spatial building configuration that encompasses structured built-up areas with enhanced spatial detail. This strengthens our understanding of these novel products' semantic definition. In contrast, GLOBC and MOD500 feature a far more generalized spatial representation incorporating further aerial features such as intra-urban spaces and two-dimensional impervious surfaces. In consequence, when assessed against the settlement mask, both HR layers naturally greatly outperform LR layers marked by substantial agreement with the reference. Although both GUF and GHSL are prone to quantifiable layer-specific errors of commission, improved accuracies relate to both increased sensitivity and precision in terms of mapping the small-scale spatial complexity of the settlement pattern. In contrast, LR layers are too generalized to correctly delineate the complex, irregular outlines of larger urban areas, and too coarse to completely detect smaller scale human settlements. In this context, measures insensitive to low prevalence such as *TSS* have been found to give a more robust indication of accuracy differences.

(3). How does the accuracy of these layers vary for structurally different areas, i.e., urban versus rural landscapes, in Central Europe?

By the application of landscape-specific PIP-tests and pixel-based accuracy assessments with regard to the settlement reference, we spatially differentiate results for urban and rural landscapes. In this, the advanced capabilities of HR settlement maps become more evident than before. While existing LR products generally allow for the delineation of larger cities in their correct dimension and a spatially generalized form, they are too coarse for applications in rural areas. Confronting these limitations, also GUF and GHSL exhibit quantifiable weaknesses in terms of sensitivity and precision in these regions. These, however, amount to an only slight reduction of overall accuracies as compared to urban areas manifested in moderate to substantial agreement in terms of *K*, *F* and *TSS*. Taking into account that GLOBC and MOD500 almost entirely neglect settlements in rural areas, high consistency between urban and rural areas of the new HR base maps impressively underline their extended spatial applicability. Beyond their increased completeness in rural areas, they were also found advantageous as they promote higher precision in delineating complex outlines of larger urban areas.

(4). Does the accuracy of global settlement layers show spatial variation with regard to the physical configuration of human settlements, i.e., size or density, in Central Europe?

Ultimately, we explore the influence of physical pattern variations on the classification results of both GUF and GHSL. First, we find that both layers significantly over-estimate true building densities and thus, only allow for general binary density separation of the settlement fabric. This observation corresponds to our findings from pixel-based accuracy assessment and originates from the semantic definitions of GUF and GHSL. In contrast, built-up or settlement densities are represented quite consistently by both layers which is manifested in strong correlations between the maps and the reference. Although we find significant over-estimation for medium magnitudes, a clear pattern-based dependency with regard to built-up densities is proven. Thus, a significant added value of GUF and GHSL lies in the structural qualification of settlement configuration beyond binary formats at a spatial aggregation scale. These results may provide useful information with regard to continued efforts in classification optimization on a regional scale.

Secondly, we explore pattern-dependent classification differences of GUF and GHSL with regard to the size of settlement patches. The scale-dependent evaluation of the thematic quality clearly reflects the perspectives on completeness and correctness with regard to patch sizes. All layers depict a pattern-based accuracy gain with increasing patch size. For smaller patch sizes that are exclusively existent in HR maps, these layers still adhere distinct limitations in terms of precision. Nevertheless, they clearly improve the scale-dependent representation of the settlement pattern marked by a clear shift of accuracies from large to medium- and small-scale patch sizes due to significantly higher degrees of completeness as compared to GLOBC and MOD500. This is manifested in a strong gain in map quality exceeding substantial agreement to the reference at a patch size of ca. 10,000 m^2.

The presented findings clearly testify the decisive advancements of recent efforts in global human settlement mapping. In general terms, GUF and GHSL exhibit significant improvements in terms of completeness, precision and accuracy with regard to existing lower resolution products. Increased sensitivity of the new base maps promotes a more complete representation of the settlement pattern, especially in rural areas. On the other hand, improved precision adds to a more correct delineation of the complex form of small-scale settlement characteristics. In this, the high consistency of accuracies between urban and rural areas of the new base maps is especially noteworthy. Beyond, we find quantifiable accuracy variations with regard to spatially altering structural characteristics such as density and settlement size. It becomes clear that HR layers possess additional structural information on an aggregated spatial scale that may even allow for a differentiated qualification of the built environment beyond binary formats. Although relying on naturally very different source data (radar vs. optical), it is especially important to note that both GUF and GHSL feature only insignificant disparities throughout the analysis. In fact, HR layers show the highest degree of correspondence among all pairs of layers in terms of Kappa

($K = 0.63$). This consistency additionally strengthens our confidence associated with these new maps resulting from independent initiatives.

In the context of this paper, it must be underlined that the presented results only give representative evidence for highly structured urbanized landscapes that can be found e.g., in Central Europe. Although these sites feature considerable physical variability, settlement characteristics vary to a much greater degree around the globe. As a consequence, accuracies may differ considerably in correspondence to differences in building material, construction type, settlement structure and physical surrounding. Some independent studies focusing on site-specific settings of other cultural areas have already shown that classification accuracies of HR layers can be significantly lower. For the GHSL, these are e.g., arid regions in Africa that feature bright open soil surfaces and scattered vegetation resulting in higher probabilities of false alarm (JRC, 2012). In turn, the GUF information extraction can feature weaknesses when applied to areas of sparse and scattered settlement structures with a weak vertical expression, particularly in terms of confusion with other vertical elements such as trees or high river banks (Esch et al., 2013, 2010). An independent investigation by DLR revealed that although Kappa statistics did not exceed agreement of 20% in a rural setting of sub-Saharan Africa, GUF and GHSL exclusively detect shares of the small-scale, fragmented settlement structures among the pool of global settlement maps. Further region-specific validation studies are essential to confirm these findings and present comprehensive knowledge on the accuracy variation of HR settlement information around the globe. The increasing availability of large inventories of HR validation data from open digital sources will ease the way to a stronger understanding of each map's strengths and weaknesses in both space and time.

Ultimately, it should be acknowledged, that within this work, we promote the comparison of datasets that rely on different specifications in terms of their geometric resolution and their inherent semantic definitions of settlement areas. One might come to the conclusion that these conceptual differences hamper comparison of the respective maps, especially the assessment of LR layers with regard to high resolution reference data. In contrast, we consider these base conditions as the main determining factor for exploring disparities of past and present mapping efforts and reasonably respect them during the interpretation of the obtained results.

7. Conclusion

It has been seen shown that the design of meaningful accuracy assessment framework needs to consider various components of map accuracy beyond traditional pixel-based approaches to paint a complete picture. In this regard, we present a comprehensive and systematic cross-comparison framework that integrates both recently initiated high, and the best known low resolution settlement products of global coverage. With regard to appropriate reference data available for two large-area test sites of varying landscape character in Central Europe, we explore multiple aspects of map accuracy. These include relative inter-map agreement between HR and LR layers, absolute overall and landscape-specific accuracies as well as pattern-dependent classification differences. In general, we find significantly improved mapping capabilities of the new base maps in terms of spatial completeness and precision, particularly in areas naturally neglected by LR products.

With regard to the immense dynamics of global urban transformation and the evolution of new forms and patterns of human settlements, an understanding of the strengths and weaknesses of global settlement information is of crucial importance when applying these datasets. The presented work gives strong evidence that the development and application of HR datasets will decisively add to our understanding and managing of the manifold aspects of worldwide urbanization on our planet. In this, products such as GUF and GHSL will extend their applicability way beyond global analysis of core urban areas. In fact, they lay the foundation for monitoring the growth of cities as well as the regional evolution of peri-urban and rural settlement patterns in high spatial detail, independently and at global scales.

Acknowledgments

The authors would like to acknowledge the support by the European Commission's Seventh Framework Program [FP7/2007–2013], under grant agreement no. 312972 “Framework to integrate Space-based and in-situ sENSing for dynamic vUlnerability and recovery Monitoring (SENSUM)” as well as by the German Federal Ministry for Economic Affairs and Energy, under grant agreement no. 01MD15008D “Smart Data — Disaster Management (sd-kama)” in the framework of the technological innovation program “Smart data — Innovations from Data”.

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. http://dx.doi.org/10.1111/j.1365-2664.2006.01214.x.

Anthrop, M. (2000). Changing patterns in the urbanized countryside of Western Europe. *Landscape Ecology*, *15*, 257–270. http://dx.doi.org/10.1023/A:1008151109252.

Arino, O., Gross, D., Ranera, F., Leroy, M., Bicheron, P., Brockman, C., ... Weber, J. L. (2007). GlobCover: ESA service for global land cover from MERIS. *Proceedings of the Geoscience and Remote Sensing Symposium (IEEE International, IGARSS 2007), Barcelona, Spain, 23–28 July 2007*. http://dx.doi.org/10.1109/IGARSS.2007.4423328.

ASPRS (1990). APSPRS accuracy standards for large scale-maps (American Society for Photogrammetry and Remote Sensing). *Photogrammetric Engineering and Remote Sensing*, *56*(7), 1068–1070.

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. *Proceedings of the ISPRS Urban Remote Sensing Conference, Tempe, AZ, March 2005*.

Bartholome, E., & Belward, S. (2005). GLC2000: A new approach to global land cover mapping from earth observation data. *International Journal of Remote Sensing*, *26*, 2005. http://dx.doi.org/10.1080/01431160412331291297.

Bhaduri, B. L., Bright, E. A., Coleman, P. R., & Dobson, J. E. (2002). LandScan: Locating people is what matters. *Geoinformatics*, *5*(2), 34–37.

BKG (2014). Digital topographic map 1:25 000, DTK25-V. (preliminary edition). Online: http://www.geodatenzentrum.de/docpdf/dtk25-veng.pdf (Accessed 14 May 2014)

Burgalassi, D. (2010). Defining and measuring polycentric regions: The case of Tuscany. MPRA discussion papers, 101, No. 25880. Online: http://mpra.ub.uni-muenchen.de/25880/ (Accessed: 4 April 2015).

CIESIN (2004). Global Rural-Urban Mapping Project (GRUMP) — Urban extents. Center for International Earth Science Information Network. Online: http://sedac.ciesin.columbia.edu/data/set/grump-v1-urban-extents (Accessed 21 November 2014).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. http://dx.doi.org/10.1177/001316446002000104.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46. http://dx.doi.org/10.1016/0034-4257(91)90048-B.

Congalton, R. G. (1994). Accuracy assessment of remotely sensed data: future needs and directions. *Proceedings of Pecora 12 Land Information from Space-based Systems*. Beteshda, MD: ASPRS.

Congalton, R. G., & Green, K. (2008). *Assessing the accuracy of remotely sensed data — Principles and practices*. Boca Raton, FL: CRC Press.

Congalton, R. G., Oderwald, R. G., & Mea, R. A. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, *49*, 1671–1678.

Danko, D. M. (1992). The digital chart of the world project. *Photogrammetric Engineering and Remote Sensing*, *58*(8), 1125–1128.

Davis, J., & Goadrich, M. (June 2006). The relationship between precision-recall and ROC curves. *Proceedings of the International Conference on Machine Learning, Pittsburgh, PA, U.S.* (pp. 25–29).

Defourny, P., Schouten, L., Bartalev, S., Bontemps, S., Caccetta, P., de Witt, A., ... Arino, O. (2009). Accuracy assessment of a 300-m Global Land Cover map: The GlobCover experience. *Proceedings of the 33rd International Symposium on Remote Sensing of Environment (ISRSE), Stresa, Italy, May 2009*.

Dicks, S. E., & Lo, T. H. (1990). Evaluation of thematic map accuracy in a land-use and land-cover mapping programs. *Photogrammetric Engineering and Remote Sensing*, *56*(9), 1247–1252.

EC-Copernicus — European Commission Copernicus Program (2014). *Date warehouse requirements — Version 2.0: Specifications on the space-based Earth Observation needs for the period 2014–2020*. European Commission., Online: http://www.copernicus.eu/sites/default/files/library/Data_Warehouse_V2_0.pdf (Accessed 29 September 2015).

Ehrlich, D., & Tenerelli, P. (2013). Optical satellite imagery for quantifying spatio-temporal dimensions of physical exposure in disaster risk assessments. *Natural Hazards*, *68*(3), 1271–1289. http://dx.doi.org/10.1007/s11069-012-0372-5.

Elvidge, C., Imhoff, M. L., Baugh, K. E., Hobson, V. R., Nelson, I., Safran, J., ... Tuttle, B. T. (2001). Nighttime lights of the world: 1994–95. *ISPRS Journal of Photogrammetry and Remote Sensing*, *56*, 81–99.

Elvidge, C., Tuttle, B. T., Sutton, P. C., Baugh, K. E., Howard, A. T., Milesi, C., ... Nemani, R. (2007). Global distribution and density of constructed impervious surfaces. *Sensor*, *7*(9), 1962–1979.

ESA (2011). *Globcover 2009 — Product description and validation report.* European Space Agency (ESA) Online: https://globcover.s3.amazonaws.com/LandCover2009/GLOBCOVER2009_Validation_Report_1.0.pdf (Accessed 6 August 2014).

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., ... Dech, S. (2013). Urban Footprint Processor — Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters*, *10*, 1617–1621. http://dx.doi.org/10.1109/LGRS.2013.2272953.

Esch, T., Marconcini, M., Felbier, A., Heldens, W., & Roth, A. (2014a). Adding a new dimension to global urban observations: Inventory of human settlements pattern and urban morphology using VHR SAR data of the TanDEM-X mission. *3rd International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Changsha, China, 11–14 June 2014.* http://dx.doi.org/10.1109/EORSA.2014.6927873.

Esch, T., Marconcini, M., Marmanis, D., Zeidler, J., Elsayed, S., Metz, A., ... Dech, S. (2014b). Dimensioning urbanization — An advanced procedure for characterizing human settlement properties and patterns using spatial network analysis. *Applied Geography*, *55*, 212–228. http://dx.doi.org/10.1016/j.apgeog.2014.09.009.

Esch, T., Taubenböck, H., Roth, A., Heldens, W., Felbier, A., Thiel, M., ... Dech, S. (2012). TanDEM-X mission—New perspectives for the inventory and monitoring of global settlement patterns. *Journal of Applied Remote Sensing*, *6*, 1–21. http://dx.doi.org/10.1117/1.JRS.6.061702.

Esch, T., Thiel, M., Schenk, A., Roth, A., Muller, A., & Dech, S. (2010). Delineation of urban footprints from TerraSAR-X data by analyzing speckle characteristics and intensity information. *IEEE Transactions on Geoscience and Remote Sensing*, *48*, 905–916. http://dx.doi.org/10.1109/TGRS.2009.2037144.

Everitt (2002). *The Cambridge dictionary of statistics* (2nd ed.). Cambridge: Cambridge University Pres.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.

FGDC (1998). *Geospatial positioning accuracy standards — Part 3: National standard for spatial data accuracy.* Washington, D.C.: Federal Geographic Data Committee.

Fina, S., Krehl, A., Siedentop, S., Taubenböck, H., & Wurm, M. (2014). Dichter dran! Neue Möglichkeiten der Vernetzung von Geobasis-,Statistik- und Erdbeobachtungsdaten zur räumlichen Analyse und Visualisierung von Stadtstrukturen mit Dichteoberflächen und –profilen. *Raumforschung & Raumordnung*, *72*, 1–14. http://dx.doi.org/10.1007/s13147-014-0279-6.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. http://dx.doi.org/10.1037/h0031619.

Florida, R., Gulden, T., & Mellander, C. (2008). The rise of mega-region. *Cambridge Journal of Regions, Economy and Society*, *1*(3), 459–476. http://dx.doi.org/10.1093/cjres/rsn018.

Foody, G. M. (1992). On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, *58*(10), 1459–1460.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*(1), 185–201. http://dx.doi.org/10.1016/S0034-4257(01)00295-4.

Foody, G. M. (2006). What is the difference between two maps? A remote senser's view. *Journal of Geographical Systemts*, *8*(2), 119–130. http://dx.doi.org/10.1007/s10109-006-0023-z.

Foody, G. M. (2008). Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, *29*(11), 3137–3158. http://dx.doi.org/10.1080/01431160701442120.

Forster, B. C. (1983). Some urban measurements from Landsat data. *Photogrammetric Engineering and Remote Sensing*, *49*(1983), 1693–1707.

Forster, B. C. (1985). An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing*, *6*(1), 139–151. http://dx.doi.org/10.1080/01431168508948430.

Galton, F. (1892). *Finger prints.* London: McMillan.

Garreau, J. (1991). *Edge city: Life on the new frontier.* New York: Doublebay.

GEO (2014). Manifesto for a global human settlement partnership. Online: http://www.earthobservations.org/documents/ghs/ghs_brochure.pdf (Accessed 5 May 2015)

GeoBasisNRW (2013). ATKIS—Digitale Topographische Karte 1:25.000 (DTK25), Bezirksregierung Köln. Online: http://www.bezreg-koeln.nrw.de/brkinternet/presse/publikationen/geobasis/faltblattgeobasisatkis01.pdf (Accessed 4 May 2014)

Giri, C., Zhu, Z. L., & Reed, B. (2005). A comparative analysis of the global land cover 2000 and MODIS land cover data sets. *Remote Sensing of Environment*, *94*, 123–132. http://dx.doi.org/10.1016/j.rse.2004.09.005.

Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, *60*(2), 181–188.

Henderson, F. M., & Xia, Z. G. (1997). SAR applications in human settlement detection, population estimation and urban land use pattern analysis: A status report. *Geoscience and Remote Sensing*, *35*(1), 79–85. http://dx.doi.org/10.1109/36.551936.

Herold, M., Mayaux, P., Woodcock, C. E., Baccinie, A., & Schmullius, C. (2008). Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment*, *112*(5), 2538–2556. http://dx.doi.org/10.1016/j.rse.2007.11.013.

Jeni, L. A., Cohn, J. F., & Torre, F. D. L. (2013). Facing imbalanced datasets: Recommendations for the use of performance metrics. *Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013.*

Jensen, J. R. (1999). *Introductory digital image processing — A remote sensing perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1999.

JRC (2012). A global human settlement layer from optical high resolution imagery concept and first results. *European Commission Joint Research Center — Institute for the Protection and Security of the Citizen*. Luxembourg: Publications Office of the European Union.

JRC (2013). Visual collection of built up information from High Resolution satellite imagery. *European Commission Joint Research Center - Institute for the Protection and Security of the Citizen*. Luxembourg: Publications Office of the European Union.

Klein Goldewijk, K., Beusen, A., de Vos, M., & van Drecht, G. (2011). The HYDE 3.1 spatially explicit database of human induced land use change over the past 12,000 years. *Global Ecology and Biogeography*, *20*, 73–86. http://dx.doi.org/10.1111/j.1466-8238.2010.00587.x.

Klotz, M., Wurm, M., & Taubenböck, H. (2015). Der Werkzugkasten der urbanen Fernerkundung - Daten und Produkte. In H. Taubenböck, M. Wurm, T. Esch, & S. Dech (Eds.), *Globale Urbanisierung - Perspektive aus dem All* (pp. 29–38). Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-662-44841-0_5.

Kullback, S. (1968). *Information theory and statistics.* New York: Dover.

Labatut, V., & Cherifi, H. (2011). Evaluation of performance measures for classifiers comparison. *Ubiquitous Computing and Communication Journal*, *6*, 21–31.

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

Latifovic, R., & Olthof, I. (2004). Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data. *Remote Sensing of Environment*, *90*(2), 153–165. http://dx.doi.org/10.1016/j.rse.2003.11.016.

Lewis, A. J. (1968). Evaluation of multiple-polarized radar imagery for the detection of selected cultural features. *NASA technical letter 130*. Washington, D.C.: NASA Online: http://pubs.usgs.gov/of/1968/0168/report.pdf (Accessed 1 October 2015).

Main, H. (1993). Urbanisation, rural environmental degradation and resilience in Africa. In U. Agnihotri (Ed.), *Environment and development* (pp. 199). New Delhi: Concept Publishing.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing.* Massachusetts: MIT Press.

Mayaux, P., Hugh, E., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., ... Roy, P. S. (2006). Validation of the Global Land Cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, *44*, 1728–1739. http://dx.doi.org/10.1109/TGRS.2006.864370.

McCallum, I., Obersteiner, M., Nilsson, S., & Shvidenko, A. (2006). A spatial comparison of four satellite derived 1 km global land cover datasets. *International Journal of Applied Earth Observation and Geoinformation*, *8*(4), 246–255. http://dx.doi.org/10.1016/j.jag.2005.12.002.

Mayaux, P., Bartholome, E., Fritz, S., & Belward, A. (2004). A new land-cover map of Africa for the year 2000. *Journal of Biogeography*, *31*, 861–877. http://dx.doi.org/10.1111/j.1365-2699.2004.01073.x.

Miyazaki, H., Shao, X., Koki, I., & Shibasaki, R. (2013). An automated method for urban area mapping by integrating ASTER satellite images and GIS data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(2), 1004–1019.

Morisette, T. J., Privette, J. L., Strahler, A., Mayaux, P., & Justice, C. O. (2004). Validation of land-cover products by the committee on Earth observing satellites. In R. S. Lunetta, & J. G. Lyon (Eds.), *Remote sensing and GIS accuracy assessment*. Boca Raton, FL: CRC Press.

Ouzounis, G. K., Syrris, V., & Pesaresi, M. (2013). Multiscale quality assessment of global human settlement layer scenes against reference data using statistical learning. *Pattern Recognition Letters*, *34*(14), 1636–1647. http://dx.doi.org/10.1016/j.patrec.2013.04.004.

Perez-Hoyos, A., García-Haro, F. J., & San-Miguel-Ayanz, J. (2012). Conventional and fuzzy comparisons of large scale land cover products: Application to CORINE, GLC2000, MODIS and GlobCover in Europe. *ISPRS Journal of Photogrammetry and Remote Sensing*, *74*, 185–201. http://dx.doi.org/10.1016/j.isprsjprs.2012.09.006.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(3), 180–192. http://dx.doi.org/10.1109/JSTARS.2008.2002869.

Pesaresi, M., Halkia, M., & Ouzounis, G. (2011). Quantitative estimation of settlement density and limts based on textural measurements. *Proceedings of the Joint Urban Remote Sensing Event (JURSE), Munich, Germany, 11–13 April 2011*. http://dx.doi.org/10.1109/JURSE.2011.5764726.

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., ... Zanchetta, L. (2013). A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(6), 2102–2131. http://dx.doi.org/10.1109/JSTARS.2013.2271445.

Potere, D., & Schneider, A. (2007). A critical look at representations of urban areas in global maps. *GeoJournal*, *69*(1–2), 55–80. http://dx.doi.org/10.1007/s10708-007-9102-z.

Potere, D., & Schneider, A. (2009). Comparison of global urban maps. In P. Gamba, & M. Herold (Eds.), *Global mapping of human settlement. Experiences, datasets, and prospects* (pp. 269–308). Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781420083408-c13.

Potere, D., Schneider, A., Angel, S., & Civco, D. (2009). Mapping urban areas on a global scale: Which of the eight maps now available is more accurate? *International Journal of Remote Sensing*, *30*, 6531–6558. http://dx.doi.org/10.1080/01431160903121134.

Powers, M. W. (2007). Evaluation: From precision, recall and f-factor to ROC, informedness, markedness & correlation. *Technical report SIE-07-001, 2007*. Adelaide, Australia: School of Informatics and Engineering, Flinders University Online: http://david.wardpowers.info/BM/Evaluation_SIETR.pdf (Accessed 8 February 2015).

Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval systems performance. *ACM Transactions on Information Systems*, *7*, 205–229.

Regione Toscana (2015a). Cartografica tecnica regionale e scarico dati geografici. Online: http://www.regione.toscana.it/-/cartografia-tecnica-regionale-e-scarico-dati-geografici (Accessed: 4 April 2015).

Regione Toscana (2015b). Prescrizioni Techniche per la cartografica fotogrammetrica numerica in scala 1:10.000. Livello 4.0. Online http://www.regione.toscana.it/documents/10180/12431710/SpecificheTecniche_CTR10K_liv40.pdf/b2aca347-e08a-4e69-8395-62ddaade2ee9 (Accessed: 4 April 2015).

Richards, J. A. (1986). *Remote sensing digital image analysis — An introduction.* Berlin/Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-662-02462-1.

Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Newton, MA: Butterworth-Heinemann Ltd.

Rottensteiner, F., Trinder, J., Clode, S., & Kubik, K. (2005). Using the Dempster–Shafer method for the fusion of LIDAR data and multi-spectral images for building detection. *Information Fusion*, *6*(4), 283–300.

Rutzinger, M., Rottensteiner, F., & Pfeifer, N. (2009). A comparison of evaluation techniques for building extraction from airborne laser scanning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*(1), 11–20. http://dx.doi.org/10.1109/JSTARS.2009.2012488.

Schneider, A., Friedl, M., McIver, D., & Woodcock, C. (2003). Mapping urban areas by fusing multiple sources of coarse resolution remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *69*, 1377–1386. http://dx.doi.org/10.1109/IGARSS.2003.1294530.

Schneider, A., Friedl, M. A., & Potere, D. (2009). A new map of global urban extent from MODIS data. *Environmental Research Letters*, *4*, 044003. http://dx.doi.org/10.1088/1748–9326/4/4/044003 (10.1201/9781420083408-c13).

Schneider, A., Friedl, M. A., & Potere, D. (2010). Monitoring global urban areas using MODIS 500 m data: New methods and datasets based on urban ecoregions. *Remote Sensing of Environment*, *114*, 1733–1746. http://dx.doi.org/10.1016/j.rse.2010.03.003.

Simon, D. (2008). Urban environments: Issues on the peri-urban fringe. *Annual Review of Environment and Ressources*, *33*, 167–185. http://dx.doi.org/10.1146/annurev.environ.33.021407.093240.

Small, C. (2001). Estimation of urban vegetation abundance by spectral mixture analysis. *International Journal of Remote Sensing*, *21*(7), 1305–1334. http://dx.doi.org/10.1080/01431160151144369.

Small, C. (2005). A global analysis of urban reflectance. *International Journal of Remote Sensing*, *26*(4), 661–681. http://dx.doi.org/10.1080/01431160310001654950.

Stehman, S. V. (1999). Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, *20*(12), 2426–2441. http://dx.doi.org/10.1080/014311699212100.

Stehman, S. V. (2004). A critical evaluation of the normalized error matrix in map accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, *70*, 743–756. http://dx.doi.org/10.14358/PERS.70.6.743.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, *64*, 331–344. http://dx.doi.org/10.1016/S0034-4257(98)00010-8.

Stehmann, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, *72*(1), 35–45. http://dx.doi.org/10.1016/S0034-4257(99)00090-5.

Story, M., & Congalton, R. G. (1986). Accuracy assessment — A user's perspective. *Photogrammetric Engineering and Remote Sensing*, *52*(3), 397–399.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., ... Woodcock, C. E. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps.* Luxembourg: Office for Official Publications of the European Commission.

Syrris, V., Pesaresi, M., Syrris, V., & Pesaresi, M. (2013). On the assessment of automatically processing HR/VHR imagery using low-resolution global reference data. *Proceedings of the Joint Urban Remote Sensing Event (JURSE)* (pp. 21–23). http://dx.doi.org/10.1109/JURSE#.2013.6550671 Sao Paulo, Brazil, April 2013.

Tatem, A. J., Noor, A. M., & Hay, S. I. (2005). Assessing the accuracy of satellite derived global and national urban maps in Kenya. *Remote Sensing of Environment*, *96*(1), 87–97. http://dx.doi.org/10.1016/j.rse.2005.02.001.

Taubenböck, H., Esch, T., Felbier, A., Roth, A., & Dech, S. (2011). Pattern-based accuracy assessment of an urban footprint classification using TerraSAR-X data. *IEEE Geoscience and Remote Sensing Letters*, *8*, 278–282. http://dx.doi.org/10.1109/LGRS.2010.2069083.

Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., & Dech, S. (2012). Monitoring urbanization in mega cities from space. *Remote Sensing of the Environment*, *117*, 162–176. http://dx.doi.org/10.1016/j.rse.2011.09.015.

Taubenböck, H., Wiesner, M., Felbier, A., Marconcini, M., Esch, T., & Dech, S. (2014). New dimensions of urban landscapes: The spatio-temporal evolution from a polynuclei area to a mega-region based on remote sensing data. *Applied Geography*, *47*, 137–153. http://dx.doi.org/10.1016/j.apgeog.2013.12.002.

Taubenböck, H. (2015). Ohne Limit? Das Flächenwachstum der Megacities. In H. Taubenböck, M. Wurm M, T. Esch, & S. Dech (Eds.), *Globale Urbanisierung - Perspektive aus dem All* (pp. 49–58). Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-662-44841-0_7.

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

U.S. Bureau of the Budget (1947). *National map accuracy standards.* Washington, D.C.: Federal Geographic Data Committee.

UN-Habitat (2013). State of the world's cities 2012/2013 — Prosperity of cities. *United Nations Human Settlements Programme (UN-Habitat)*. New York: Routledge. http://dx.doi.org/10.1080/07293682.2013.861498.

United Nations (2014a). *World urbanization prospects: The 2014 revision, highlights (ST/ESA/SER.A/352).* Population Division, Department of Economic and Social Affairs: United Nations Publications.

United Nations (2014b). *World urbanization prospects: The 2014 revision, sources for urban population.* Population Division, Department of Economic and Social Affairs: United Nations Publications, Online: http://esa.un.org/unpd/wup/CD-ROM/WUP2014_DOCUMENTATION/WUP2014-DataSource-UrbanPopulation.xls (Accessed: 07 October 2015).

Wania, A., Kemper, T., & Tiede, D. (2014). Mapping recent built-up area changes in the city of Harare with high resolution satellite imagery. *Applied Geography*, *46*, 35–44. http://dx.doi.org/10.1016/j.apgeog.2013.10.005.

Welch, R. (1982). Spatial resolution requirements for urban studies. *International Journal of Remote Sensing*, *3*(2), 139–146. http://dx.doi.org/10.1080/01431168208948387.

Wurm, M., d'Angelo, P., Reinartz, P., & Taubenböck, H. (2014). Investigating the applicability of Cartosat-1 DEMs and topographic maps to localize large-area urban mass concentrations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*, 4138–4152. http://dx.doi.org/10.1109/JSTARS.2014.2346655.