Systematic assessment of human settlement datasets in the context of humanitarian mapping

MSc GISc Dissertation

Acknowledgements

I would like to give thanks to the following people who supported me in completing this project:

- Dr Andrea Ballatore for the conversations that guided my thinking in this work. His knowledge and insight assisted the development of my own ideas.
- Dr Shino Shiode and Dr Paul Elsner for their useful advice and feedback on this project.
- Gabor Bakos and Andrew Braye of Missing Maps for their assistance in helping define an interesting and useful research project.
- My family and friends for their ongoing support and interest in the project.
- Not least, my girlfriend Rebekah Neal who offered patience and understanding while I spent hours working on this project, in what would have been our shared time.

Abstract

The spatial detail of automated human settlement mapping is advancing. New datasets utilise developments in computing power, high resolution satellite imagery, and novel detection approaches to classify settlements. These advancements create opportunities to apply the datasets to humanitarian settlement mapping.

In response to humanitarian crises agencies require geospatial data to assist with response decision making. Many vulnerable countries lack geospatial data and so data is collected through Volunteered Geographic Information campaigns (the Missing Maps Project). This manual digitisation workflow is time consuming and could be enhanced by integrating new automated human settlement datasets. However, studies on the capabilities of these products in a humanitarian mapping context are limited.

We introduce a comparison framework to assess the relative capabilities of datasets for a site previously mapped by volunteers (Leyte and Samar, Philippines). The framework incorporates techniques to quantify inter-map agreement in total, by settlement character, and by settlement pattern. Inter-map differences are assessed to understand their source and potential causes. The framework is applied to three high spatial detail datasets, the Global Human Settlement Layer, Global Urban Footprint, and High Resolution Settlement Layer.

We find disagreement between human settlement datasets, particularly in areas with fragmented rural settlements due to increased detection complexity. The Global Human Settlement Layer is identified as unsuitable for humanitarian mapping due to weak classification capabilities for rural settlements. Global Urban Footprint and High Resolution Settlement Layer show satisfactory capability in these settings with a higher proportion of classifications for smaller settlement sizes, and errors of commission indicating capability to detect dispersed rural settlements missed by comparison datasets.

Overall, we find that no one product offers complete settlement mapping capabilities. Future efforts should focus on incorporating current products for the prioritisation of volunteer mapping, and on fine tuning classifiers with training data generated by volunteer observations.

Contents page

1 Introduction 1

2 Background 4

2.1 Overview of humanitarian mapping 4

2.2 Overview of earth observation human settlement datasets 8

3 Study rationale 16

4 Study site and data 18

4.1 Study site 18

4.2 Datasets under test 20

4.3 Datasets for comparison 20

4.4 Dataset independence 20

4.5 Ancillary data 22

5 Methodology 24

5.1 Data pre-processing 24

5.2 RQ1: To what extent do high spatial detail human settlement datasets vary in a Global South setting? 24

5.3 RQ2: To what extent do these datasets vary for different settlement landscape character? 29

5.4 RQ3: What factors may contribute to inter-map disagreement for these datasets? 30

6 Results and Discussion 33

6.1 RQ1: To what extent do high spatial detail human settlement datasets vary in a Global South setting? 33

6.2 RQ2: How do the datasets vary for different settlement landscape character? 41

6.3 RQ3: What factors may contribute to inter-map disagreement for these datasets? 50

6.4 Limitations 60

7 Conclusion 61

8 References 63

Figures

Figure 1: Missing Maps workflow (based on (Albuquerque, Herfort and Eckle, 2016)) 5
Figure 2: MapSwipe mobile application for classification. Green tiles have been labelled by a volunteered as containing buildings. 6
Figure 3: OSM Tasking Manager for digitisation. Buildings and roads are traced by volunteers with drawing and labelling tools. 6
Figure 4: Human settlement satellite imagery examples (Bing Maps) 7
Figure 5: Study site location: Islands of Leyte & Samar, Philippines 19
Figure 6: VIIRS NTL Luminosity data landscape character zone threshold categorisation 23
Figure 7: Resampling of human settlement datasets for PIP test comparison 26
Figure 8: Total area classified as human settlement (km2), and total pixel/polygon count for human settlements datasets included in this study (Leyte and Samar, Philippines). 34
Figure 9: Map view of human settlement datasets under test, with inset map for area around Tacloban city (Leyte and Samar, Philippines). 35
Figure 10: Map view of human settlement datasets for comparison, with inset map for area around Tacloban city (Leyte and Samar, Philippines). 36
Figure 11: Inter-map agreement illustrations, A) HRSL FP relative to GUF B) MOD500 high FN relative to GHSL 39
Figure 12: GHSL inter-map agreement error matrix results 40
Figure 13: GUF inter-map agreement error matrix results 40
Figure 14: HRSL inter-map agreement error matrix results 40
Figure 15: Change in PPV by NTL spatial zones for datasets under test 43
Figure 16: Change in TPR by NTL spatial zones for datasets under test 43
Figure 17: Change in F by NTL spatial zones for datasets under test 44
Figure 18: Patch size distribution for human settlement datasets. Normalised to total number of patches per dataset. 45
Figure 19: Change in PPV by patch size 48
Figure 20: Change in TPR by patch size 48
Figure 21: Change in quality (F) by patch size 49
Figure 22: Sample of HRSL and GUF FP and FN classifications 51
Figure 23: Inter-map disagreement by proximity to human settlements 53
Figure 24: FP and FN error rates for zones of good and bad imagery 53
Figure 25: Inter-map agreement GUF:HRSL 55
Figure 26: Inter-map agreement GUF:OSM 56

Figure 27: Inter-map agreement HRSL:GUF 57
Figure 28: Inter-map agreement HRSL:OSM 58
Figure 29: Inter-map agreement HRSL:OSM, with OSM bad imagery mask 59

Tables

Table 1: Summary of global human settlement mapping initiatives, datasets under test by this study are shown in red, and datasets for comparison in blue 13
Table 2: Summary of sensors and satellite missions commonly used for human settlement extraction 14
Table 3: Summary of datasets under test 21
Table 4: Summary of datasets for comparison 21
Table 5: Summary of ancillary datasets 21
Table 6: Inferred landscape character spatial zones from VIIRS NTL data 22
Table 7: Dataset comparison matrix (X indicates inter-map agreement assessment conducted) 25
Table 8: Inter-map agreement bitwise logic for raster summing 26
Table 9: Error statistics 28
Table 10: Bitwise binary logic for raster sums by landscape character (HS = human settlement) 29
Table 11: Inter-map disagreement scenarios 31
Table 12: Dataset comparison matrix for RQ3 50
Table 13: Comparison pair error sample size (95% confidence level, and 5% margin of error) 50
Table 14: Map errors descriptions 52

Appendices

Appendix A: Research question two Night-time Light spatial zone error matrix results 71

1 Introduction

Humanitarian crises are emergency events arising from natural and anthropogenic disasters. They are often sudden events which threaten people's health, safety, and wellbeing. Poor populations with minimal infrastructure and resources are more vulnerable, due to their inability to endure the negative impacts (Humanitarian Coalition, 2017). Humanitarian crises have occurred throughout history, and are likely to continue with the rising impact of climate change, population growth, disease, and other factors (Zook et al., 2010).

In response to humanitarian crises national and international agencies require reliable geospatial data and tools to assist decision making (Goodchild and Glennon, 2010). However, the areas of the world where humanitarian crises are more common appear to be poorly represented in geospatial and other data (Graham et al., 2014). Inadequate building, road, and other geospatial data can hinder response effectiveness (Goodchild and Glennon, 2010).

The collection and processing of this data is a major resource challenge for humanitarian relief agencies with limited resources during an emergency (Goodchild and Glennon, 2010). Volunteered Geographic Information (VGI) is the creation, assembly, and dissemination of geographic data voluntarily by individuals (also called user-generated content, crowdsourcing or crowdmapping) (Goodchild, 2007; Haklay et al., 2014). VGI methodology can assist humanitarian agencies in the supply of geospatial data for crisis relief. This utilisation of volunteer labour allows agencies to focus their resources on humanitarian assistance (Haklay et al., 2014).

The humanitarian crisis following the 2010 Haiti earthquake illustrates the usefulness of volunteer mapping for disaster relief. Volunteers from around the world used satellite imagery in OpenStreetMap (OSM) to trace buildings, roads, and other features for the Port-au-Prince region, providing in a few days what would have taken years to map (Zook et al., 2010; Herfort, Eckle and Albuquerque, 2016).

To facilitate the collection of VGI for humanitarian work the Missing Maps project was founded in November 2014. Missing Maps is a collaboration between the Humanitarian OSM Team (HOT), American Red Cross, British Red Cross, and Médecins Sans Frontières (MSF) and aims to "map the most vulnerable places in the developing world, so that international and local NGOs and individuals can use the maps and data to better respond

to crises affecting the areas". To achieve this, Missing Maps supports technological development of crowdmapping tools and connects volunteers with these tools (Missing Maps, 2017; Herfort, Eckle and Albuquerque, 2016).

Missing Maps volunteers use VGI tools to classify and digitise their observations to generate spatial data for humanitarian projects. The contribution of these volunteers is enormous, with over 30,000 volunteers digitising 12 million buildings and 1.3 million kilometres of road (between November 2014 and July 2017) (Missing Maps, 2017). However, volunteer time is limited, with half of contributors spending less than 70 minutes, and one third less than 30 minutes mapping in total (Pete Masters and Benjamin Herfort, 2016). Of 2,600 projects in the OSM Tasking Manager, only 595 are estimated as being more than 90% complete, and more than half are less than 50% complete (Giraud, 2017). The current workflow is time consuming, with volunteers manually scanning all imagery, whilst human settlements occur on <1% of the Earth's surface (Schneider, Friedl and Potere, 2010). This intensive approach could benefit from automation using earth observation data and techniques (Kunce, 2016; Pete Masters and Benjamin Herfort, 2016).

Use of earth observation data for human settlement extraction has been widely studied (Li and Gong, 2016; Chen et al., 2015). Techniques have been applied in other fields including urban planning, environmental change, population growth, and economic development (Weng, 2014). While accurate mapping of dense urban centres has improved over recent years, extracting dispersed rural settlements remains a challenge. Improvements in image resolution and computing power have strengthened rural settlement detection capabilities (Giri et al., 2013). Medium and high to very high resolution global products classifying human settlement extents are now available and are relevant to humanitarian mapping, where targets are often small settlements in scattered rural areas (Chen and Zipf, 2017).

Missing Maps representatives have identified a need to better understand the fitness-for-use of these new products for humanitarian mapping. Whilst product accuracy is assessed by the institutions who develop them, it is best measured for the requirements of a particular user (Foody, 2002). The Missing Maps use case is the classification of human settlements across rural to urban character within vulnerable countries of the Global South (Chen and Zipf, 2017).

The aim of this research is to assess fitness-for-use of high and very high spatial resolution (high spatial detail) human settlement datasets in the context of humanitarian mapping. This will allow Missing Maps to understand the strengths and weakness of each product. A

comparison framework is introduced to study relative inter-map agreement, and the relative capabilities of each dataset under test (Klotz et al., 2016). The comparison framework incorporates techniques to quantify inter-map agreement in total, by settlement character, and by settlement pattern.

An absolute accuracy assessment has not been completed, as reference data for the Global South study site subject to human settlement VGI campaigns is unavailable (Congalton and Green, 2008). The framework is applied to three datasets that represent the latest advancements in human settlement detection (described in section 1.1). Three additional human settlement datasets are included for comparison (section 4.3). Through the application of the framework, three research questions are addressed for the datasets under test:

1) To what extent do high spatial detail human settlement datasets vary in a Global South setting?
2) How do the datasets vary for different settlement landscape character?
3) What factors may contribute to inter-map disagreement for these datasets?

The study is organised as follow; section 2 provides background on both humanitarian mapping requirements and practises and earth observation human settlement detection techniques. Section 3 depicts the rationale for this study and context in existing literature. Section 4 introduces the study site and datasets. Section 5 outlines the comparison framework. Results and discussion are presented in section 6. Section 7 outlines study conclusions and recommendations.

2 Background

2.1 Overview of humanitarian mapping

This section describes the Missing Maps target and VGI workflow so that the applicability of an equivalent automated solution can be understood.

2.1.1 Humanitarian mapping target

The Missing Maps project maps areas where humanitarian organisations are working to meet the needs of vulnerable people (Missing Maps, 2017). Buildings are vital human infrastructures and therefore offer a indicator of human presence for mapping (Pesaresi et al., 2016b). The Missing Maps project aims to map all buildings in its project areas (Chen and Zipf, 2017).

Human settlements occur in a range of forms from scattered rural settlements to dense urban areas. Settlements are urban when they have high population density and human land use (built-up). Settlements are rural when they have low population density (Kemper et al., 2016; Deuskar, 2015).

The Global South extends across continents and climatic zones, each with diverse cultures and levels of development. Building form can vary greatly due to these factors (Jensen, 2013). Missing Maps works across the Global South and informs on local building form for classification and digitisation (Missing Maps, 2017).

Considering these requirements, and for the purposes of this study, the Missing Maps human settlement mapping target (‘humanitarian mapping target’) is defined as aerial units containing a building or partial-building (Tenerelli and Ehrlich, 2011)).

2.1.2 Human settlement volunteered geographic information

Crowdsourcing geographic information has become popular in the last decade (Sui, Elwood and Goodchild, 2012). Its application to crisis mapping has allowed geospatial data to be produced for relief efforts following emergency events (Goodchild and Glennon, 2010). The value of this approach was demonstrated in responses to the Haitian earthquake (Zook et al., 2010), and the Santa Barbara wildfires (Goodchild and Glennon, 2010).

The Missing Maps project facilitates collection of VGI for crisis preparedness and response by developing necessary workflows and tools (Missing Maps, 2017). The current Missing Maps workflow consists of three tasks, and two VGI tools (shown in

Figure 1). First, satellite imagery is classified by volunteers using the MapSwipe mobile application (launched by MSF July 2016). Map tiles are classified as either: obscured by cloud cover, containing buildings and/or roads, or unknown (Figure 2).

Next, classified images are passed to the OSM Tasking Manager for digitisation. This web platform allows volunteers to digitise roads, residential areas, and building footprints observed in satellite imagery as vector point, line, or polygon features (Figure 3). Pre-processing in MapSwipe ensures volunteer digitisation tasks always contain features to trace, removing the need to scan for human features or bad imagery.

Finally, conflation of data is completed by merging data and adding attributes. This conflation task is completed by experienced mappers in the OSM community (Albuquerque, Herfort and Eckle, 2016; Herfort, Reinmuth and Zipf, 2017; Bakos and Ballatore, 2017).

The quality of geospatial data produced by volunteers is often questioned (Fan, Yang and Zipf, 2016). The data is unlikely to have been exposed to the quality control procedures authoritative mapping agencies use to validate outputs. However, it is thought the quality of VGI can match that of authoritative sources for three reasons. Firstly, information collected by many observers can be more accurate than that collected by one (Haklay et al., 2010). Secondly, geospatial data is rich in context and seemingly inconsistent features may stand out (such as a building in a lake). Finally, VGI can be quickly edited so is more likely to be up to date (Goodchild and Glennon, 2010). Missing Maps VGI workflow reflects these mechanisms (Albuquerque, Herfort and Eckle, 2016).

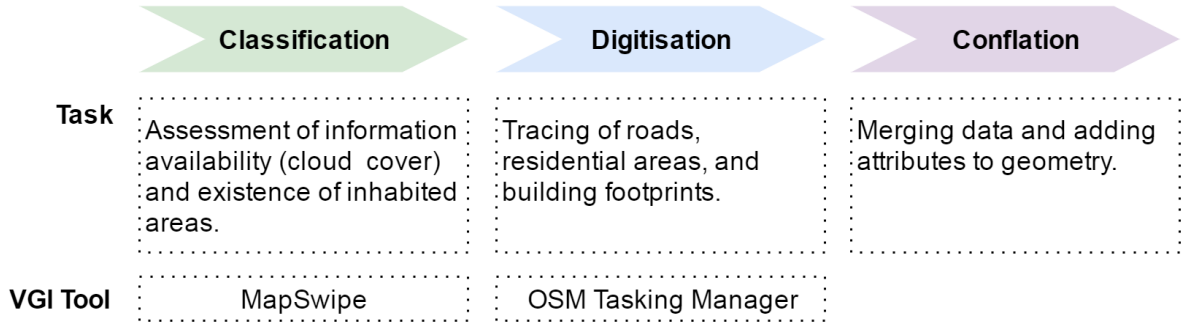


Figure 1: Missing Maps workflow (based on (Albuquerque, Herfort and Eckle, 2016))

Figure 2: MapSwipe mobile application for classification. Green tiles have been labelled by a volunteered as containing buildings.

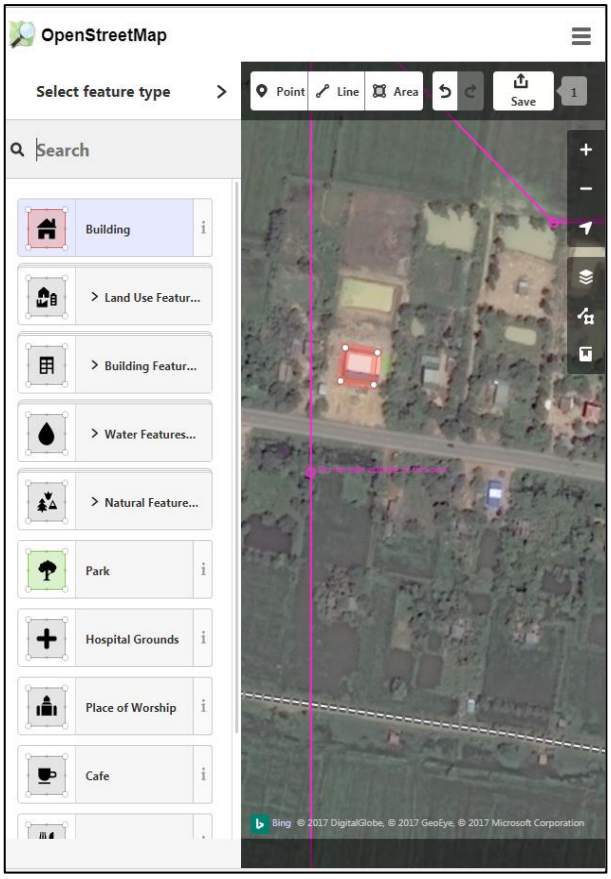


Figure 3: OSM Tasking Manager for digitisation. Buildings and roads are traced by volunteers with drawing and labelling tools.

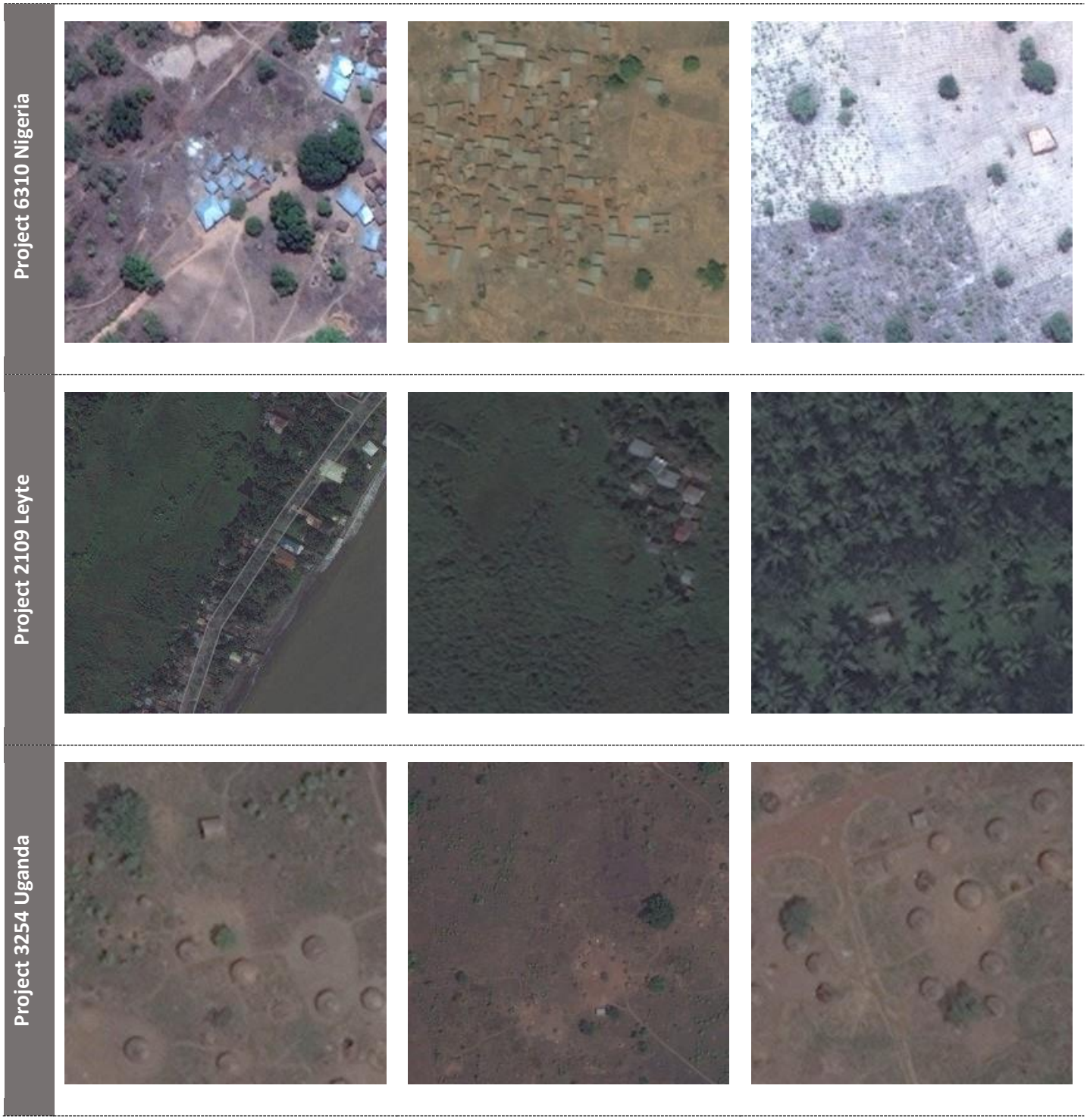


Figure 4: Human settlement satellite imagery examples (Bing Maps)

2.2 Overview of earth observation human settlement datasets

This section reviews global settlement mapping initiatives relevant to the humanitarian mapping target. It is limited to products with input imagery after year 2000, and spatial resolution <= 1000m. Techniques are described by spatial resolution, low (300-1000m), medium (30-300m), and high to very high (<30m) (Hoersch and Amans, 2015).

2.2.1 Methodological issues

Extraction of human settlements from satellite imagery is challenging as they are often characterised by a spectral mix of materials (Jensen, 2013). Settlements are defined by function (land use), rather than form (land cover). This land use can consist of many land cover entities with inhomogeneous reflectance values such as trees, grass, concrete, metal, plastic, water, and soil (Weng, 2014). Challenges can also arise in rural settlements of the Global South where buildings are made of local materials which are spectrally indistinguishable from the landscape (Small, 2009).

On average the scale of spectrally homogenous human settlement features is 10 to 20m in cities, down to as low as 2m for rural settlements and refugee camps (Small, 2003, 2005). The relationship between the spatial resolution of sensors and the scale of settlement features determines the mixing within a pixel. When the pixel is larger than the feature (generally medium to low resolution imagery) spectral confusion may occur as multiple features are mixed within a pixel (Lu et al., 2008). When the pixel is smaller than the feature (generally high to very high-resolution imagery) individual spectrally homogenous features can be detected. Dispersed rural settlements often have a tiny footprint that can only be detected in high resolution imagery (Small and Sousa, 2016). The Urban Industrial Land-use Classification guidance indicates that sensor data with a spatial resolution of 1 to 5 m is required to detect residential, commercial, and industrial buildings (Jensen, 2013).

2.2.2 Low resolution (LR) (300 to 1000m)

Several low spatial resolution global land cover (GLC) datasets are available for different time periods. Products include the MODIS Map of Global Urban extent (MOD500) at 500m spatial resolution (Schneider, Friedl and Potere, 2010; Potere et al., 2009), GlobCover v2 at 300m (Arino, Kalogirou and Ramos Perez, 2011), Landscan at 1000m (Dobson et al., 2000), Gridded Population of the World at ~1000m (Doxsey-Whitfield et al., 2015), and the Global Rural-Urban Mapping Project (GRUMP) at ~1000m (Balk et al., 2005). Due to detection limits these products classify highly urbanised environments only (Jokar Arsanjani, See and Tayyebi, 2016). For example, the MODIS human settlement class is defined as the built

environment < 50% of 500m pixel, and GlobCover v2 as artificial surfaces <50% of 300 m pixel (Arino, Kalogirou and Ramos Perez, 2011; Schneider, Friedl and Potere, 2010). These products fail to detect small settlements (Klotz et al., 2016), and are less effective in most of Africa and Asia, than in Europe and North America, where highly built-up land use (artificial surfaces) clearly delineates urban settlements (Potere et al., 2009). Additional shortfalls identified in low resolution GLC datasets include low agreement between products (for example 20% greater cropland in GlobCover than MODIS) (Fritz et al., 2011), and low overall accuracies (GlobCover v2 Kappa of 0.44, MODIS Kappa 0.63) (Klotz et al., 2016) .

In addition to categorical GLC datasets, night-time light (NTL) sensors detect anthropogenic lighting on the earth's surface as a continuous measurement of a spectral property (Weng, 2014; Elvidge et al., 2013). NTL's have been shown to correlate with human settlement features (for example population (Zhang and Seto, 2011) and urbanisation (Zheng et al., 2017)). Two global cloud free composites are available which represent stable NTL (fires, gas flares, and other non-stable lights removed), DMSP-OLS and VIIRS (described in Table 1) (Weng, 2014). NTL data detects light emitting human settlements only. In the Global South NTL less accurately identifies settlements where there is limited or no access to electricity (Zhang and Seto, 2013; Brenner and Schmid, 2014). Doll et al. (2010) determined that in Europe numbers of DMSP-OLS dark pixels (no NTL detected) declined to 10% at a population density of 50 people/km2, whereas in Africa, 75% of pixels remained unlit at a population density of 250 people/km2 (Doll and Pachauri, 2010).

These low-resolution products have been widely used for analysis of large urban areas. However limited accuracy due to spectral and spatial heterogeneity of human settlements has led to the development of higher resolution products (Klotz et al., 2016; Small, 2005).

2.2.3 Medium resolution (MR) (30 to 300m)

Medium resolution products include GlobeLand30 at 30m spatial resolution (Chen et al., 2015), and the Global Human Settlement Layer at 38m spatial resolution (Pesaresi et al., 2016b).

GlobeLand30

The National Geomatics Centre of China released GlobeLand30 (GL30) in 2014 following the opening of the Landsat archive (Giri et al., 2013). This medium resolution product provides a global land cover dataset of 10 land use classes for 2000 and 2010 (Chen et al., 2015). Human settlements are represented as 'artificial surfaces' and defined as 'lands modified

by human activities, including all kinds of habitation, industrial, mining, transport, and interior urban green areas and water bodies'. Over 10,000 Landsat Thematic Mapper and Enhanced Thematic Mapper scenes were classified through an integrated approach of pixel classification, object-based classification, and human review (known as pixel-object-knowledge-based (POK-based)) (Chen et al., 2015).

Higher resolution of GL30 may address some limitations of the low-resolution products described. Its overall accuracy is reported to be 80% with a kappa statistic of 0.676 to 0.950 (Chen et al., 2015). Jokar Arsanjani et al. (2016) compared GL30 to 821 ground truth samples in Iran with overall accuracy (proportion of correctly classified pixels) of 78% (Jokar Arsanjani, Tayyebi and Vaz, 2016). Jokar Arsanjani et al. (2016) study in Germany in comparison with OSM reference data indicated overall accuracy and Kappa index between the two datasets of 85% and 77% respectively (Jokar Arsanjani, See and Tayyebi, 2016). Brovelli et al. (2015) calculated an overall accuracy of 80% in comparison with authoritative national datasets in Italy (Brovelli et al., 2015).

Global Human Settlement Layer

The Global Human Settlement Layer (GHSL) funded by the European Commission Joint Research Centre (JRC) is a global dataset estimating built up areas for 1975, 1990, 2000 and 2014. The product consists of a series of layers including a built-up grid at 38m spatial resolution, population grid at 250m spatial resolution, and settlement model (settlement size classification) at 1km resolution (Pesaresi et al., 2016c; Freire et al., 2016). Frameworks to incorporate high and very high resolution imagery are under development but not yet available (Pesaresi et al., 2013; Ferri et al., 2014; Pesaresi et al., 2016a).

Extraction of built-up areas is completed from Landsat imagery using supervised and unsupervised classification methods. Scene supervised classification through a Symbolic Machine Learning classifier is completed in two processing loops. The first loop is trained with coarse resolution global land cover data (including GlobCover, and WorldPop) and produces a conservative dataset which would under-represent small settlements. The second loop is trained with higher resolution training data (OSM and Geonames) and refines the first. The 2014 dataset utilises textural image features in Landsat 8 data to enhance the classification. Areas labelled as built-up by the supervised classification are reclassified by an unsupervised classification approach. This discriminates areas labelled as built-up by vegetation levels (NDVI), and 3D roughness (DSM and DEM data). Degree of

built-up per cell is calculated with a fuzzy information unmixing model to estimate the proportion covered by buildings (Pesaresi et al., 2016b).

The GHSL base product is the built-up grid representing 'built-up areas' (constructed man-made objects) with a value of building footprint area per grid cell. Validation against a European Union land use survey (270,000 points), and building footprints from America and Europe indicated overall accuracy of 96% (Pesaresi et al., 2016c). Validation on comparison to MODIS 500m and LandScan indicated 91.5% agreement globally (Pesaresi et al., 2013). There appear to be few studies into the usefulness of this dataset, apart from Klotz et al (2016) who compared the GHSL to the low resolution GlobCover and MODIS in Cologne and Tuscany. They concluded that the GHSL offers improved completeness, precision, and accuracy in both high and low density landscapes (Klotz et al., 2016).

Integrated methodologies

Integrated methodologies for human settlement detection using medium resolution imagery have been proposed but are not currently available globally. Vegetation (widely detected using the normalised difference vegetation index (NDVI) (Jensen, 2013)) is negatively correlated with urban settlements (Weng, 2014) and has been used to mask non-settlement pixels (Patel et al., 2015). Zhang et al. integrated Landsat NDVI masking with DMSP-OLS NTL to detect urban areas at 30m resolution (Zhang et al., 2015). Li et al. (2016) combined NDVI masking with NDVI mean (bareland temporal changes mask), modified normalised difference water index (water surface mask), and short wave infrared (bareland mask) before extracting settlements through supervised classification (Li and Gong, 2016).

2.2.4 High to very high-resolution (HR) (< 30m) initiatives

High to very high resolution products include the Global Urban Footprint published at 12m spatial resolution (from 3m resolution input data) (Esch et al., 2017), and the High Resolution Settlement Layer published at 30m spatial resolution (from 0.5m resolution input data) (Tiecke, 2016).

Global Urban Footprint

The Global Urban Footprint (GUF) was developed by Germany's Aerospace Center (Deutsche Zentrum für Luft- und Raumfahrt (DLR)) from 2012/13 TerraSAR-X/TanDEM-X imagery. It provides a binary classification of 'built-up areas' ('a region featuring man-made building structures with a vertical component'). Roads and other flat impervious surfaces are not detected (Esch et al., 2011, 2017). Built-up areas are extracted using X band

microwave radar data from the TanDEM-X satellite (3m resolution). Speckle divergence (a texture feature) is extracted to highlight diverse sensor backscattering associated with heterogeneous entities, common in urban areas. This is passed to a unsupervised classification algorithm to produce a binary settlement layer at 12m spatial resolution (Esch et al., 2013)

GUF validation was completed by manual comparison with HR imagery for Buenos Aries, Nairobi, New Delhi, Munich and Padang. Overall accuracy was stated at between 94.8% and 96.4% (Esch et al., 2013). Klotz et al. (2016) found the accuracy of the GUF to be nearly twice that of low resolution products in both rural and urban landscape of Cologne and Tuscany (Klotz et al., 2016). Muck et al. (2017) validated GUF in Burkina Faso, a predominantly rural Global South setting, against national settlement data and found urban centres had high accuracy, but rural landscapes were under classified (Mück, Klotz and Taubenböck, 2017). Taubenböck et al. (2011) reviewed the accuracy of GUF for Padang (Indonesia) with high overall accuracies (Taubenbock et al., 2011).

High Resolution Settlement Layer

The High Resolution Settlement Layer (HRSL) was collaboratively developed by the Facebook Connectivity Lab, the Centre for International Earth Science Information Network at Columbia University and the World Bank (Tiecke, 2016). It provides population estimates (number of people per pixel) from 2015 HR DigitalGlobe imagery and the Gridded Population of the World (GPW) dataset (Center for International Earth Science, 2017). HRSL aims to facilitate settlement pattern analysis in internet infrastructure planning (Gros and Tiecke, 2016). Data is made available for thirteen countries (Center for International Earth Science, 2017). Documentation is not yet available on methodology or accuracy of results.

The product uses a novel approach to human settlement detection. HRSL has been developed using computer-vision techniques with HR satellite imagery. The Facebook image recognition deep convolutional neural network (CNN) was trained with 8,000 binary labelled human settlement images from one country. The trained algorithm processed millions of HR satellite imagery tiles to recognise settlement features. Analysis has now been completed for twenty countries (Tiecke, 2016).

Table 1: Summary of global human settlement mapping initiatives, datasets under test by this study are shown in red, and datasets for comparison in blue

Spatial Resolutio	Product name	Producer	Time stamp	Spatial resolution	Primary imagery	Definition of human settlement representation	Reference
Low	MODIS 500 m Map of Global Urban Extent (MOD500)	Universities of Wisconsin and Boston	2001/02	~ 500m	MODIS	'Built environment'	(Schneider, Friedl and Potere, 2010, 2009)
	GlobCover V2	European Commission Joint Research Centre	2009	~ 300m	MERIS	'Artificial surfaces and associated areas'	(Arino, Kalogirou and Ramos Perez, 2011)
	Global Rural-Urban Mapping Project (GRUMP)	Columbia University	2000	~ 1000m	Census, DMSP-OLS	People per grid square	(Balk et al., 2005)
	Gridded Population of the World	Columbia University	2010	~ 1000m	Census, administrative boundaries	People per grid square	(Doxsey-Whitfield et al., 2015)
	Global Impervious Surface Area	US National Geophysical Data Centre	2000/ 20001	~ 1000m	DMSP-OLS	'Constructed impervious surface area'	(Elvidge et al., 2007)
	Defence Meteorological Satellite Program Operational Linescan System (DMSP-OLS) Stable Night-time Lights	US National Geophysical Data Centre	1992 to 2015	~ 1000m at equator	DMSP-OLS	Nocturnal radiance with intermittent light sources (e.g. fires) removed.	(Elvidge et al., 2001)
	Visible Infrared Imaging Radiometer Suite (VIIRS) Stable Night-time Lights	US National Geophysical Data Centre	2011 to present	~ 500m at equator	VIIRS	Nocturnal radiance with intermittent light sources (e.g. fires) removed.	(Elvidge et al., 2013)
	LandScan™	Oak Ridge National Laboratory	2012	~ 1000m	Census and landcover data	n/a - gridded population	(Dobson et al., 2000)

Medium	GlobeLand30 (GL30)	National Geomatics Centre for China	2010	30m	Landsat TM/ETM+, and HJ1	'Artificial surfaces' (lands modified by human activities)	(Chen et al., 2015)
	Global Human Settlement Layer (GHSL)	European Commission, Joint Research Centre	2013/14	38m	Landsat 8	'Built-up land cover'	(Kemper et al., 2016)
	WorldPop, AsiaPop, AfriPop, AmeriPop	GeoData Institute, University of Southampton	By location	100m	Census and landcover data	People per grid square	(GeoData Institute, 2017)
High	Global Urban Footprint (GUF)	Deutsche Zentrum für Luft- und Raumfahrt (DLR)	2011 to 2014	12m	TanDEM-X	'built-up area with vertical structuring'	(Esch et al., 2011)
	High Resolution Settlement Layer (HRSL)	Facebook Connectivity Lab, Centre for International Earth Science Information Network at Columbia University, World Bank	2010 to 2016	~30m	DigitalGlobe high (0.5m) resolution	'Urban and rural human settlements'	(Tiecke, 2016; Gros and Tiecke, 2016)

Table 2: Summary of sensors and satellite missions commonly used for human settlement extraction

Spatial Resolutio	Instrument / mission name	Agencies	Missions	Mission dates	Spatial resolution	Spectral resolution	Revisit time	Data availability	Reference
Low	MODIS 500 m Map of Global Urban Extent (MOD500)	National Aeronautics and Space Administration (NASA)	Terra, Aqua	1999, 2002 to present	250m	36 bands	1 to 2 days	Public	(Weng, 2014)
	Moderate Resolution Imaging Spectrometer (MERIS)	European Space Agency (ESA)	Envisat	2002 to 2012	300m	15 bands	3 days	Public	(Weng, 2014)
	Defence Meteorological Satellite Program Operational Linescan System (DMSP-OLS)	US National Geophysical Data Centre	5 DMSP missions	1992 to 2015	~ 1000m at equator	1 band	0.5 days	Public	(Elvidge et al., 2001)

	Visible Infrared Imaging Radiometer Suite (VIIRS)	US National Geophysical Data Centre	Suomi	2011 to present	~ 500m at equator	22 bands	1 day	Public	(Elvidge et al., 2013)
Medium	Landsat 1 to 8	NASA and National Oceanic and Atmospheric Administration (NOAA)	Landsat 1 to 8	1972 to present	30m	Land surface data primarily 6 band.	16 days	Public	(Weng, 2014; Loveland and Dwyer, 2012)
High	Sentinel 1 to 3	European Space Agency (ESA)	Sentinel 1 to 3	2014 to present	S1 50m to 9m, S2 10	Multispectral Instrument 13 bands	6 days	Public	(European Space Agency, 2017)
	Satellite for observation of Earth (SPOT)	Spot image	SPOT 1 to 7	1986 to present	SPOT 6 & 7 panchromatic 1.5m, multispectral 6m	5 bands	Daily	Commercial	(Airbus, 2017)
	DigitalGlobe	DigitalGlobe	EarlyBird-1, IKONOS, QuickBird, GeoEye-1, Worldview1 to 3	1997 to present	Worldview-3 panchromatic 0.31m, multispectral 1.24m	29 bands	Daily	Commercial	(DigitalGlobe, 2016)
	TerraSAR-X add-on for Digital Elevation Measurement (TanDEM-X)	German Aerospace centre (DLR) and Airbus Defence and Space	TanDEM-X	2010 to present	12m	n/a	11 days	Restricted	(Esch et al., 2011)

3 Study rationale

The background to this study highlights extensive efforts in human settlement mapping. Recent developments of novel detection approaches with high resolution imagery has resulted in major advancements in the spatial detail of human settlement products (Klotz et al., 2016). These products could be used to optimise humanitarian mapping workflows in inadequately mapped areas of the Global South (Kunce, 2016). However, existing studies on the capabilities of these products in a humanitarian mapping context are limited.

With advancements in the spatial detail of human settlement datasets comes the opportunity to assess the fitness-for-use of these products in a new context. The applicability of human settlement datasets to humanitarian mapping is linked to the spatial resolution of the product. HR imagery can capture smaller spectrally homogenous human settlement features (such as individual buildings) than MR and LR imagery (Small, 2003, 2005). Datasets developed using HR imagery may therefore better detect the small-scale complexity of settlement patterns and may offer greater completeness in rural to urban landscapes (Klotz et al., 2016; Small and Sousa, 2016). These products could help improve the current resource intensive VGI approach to humanitarian mapping (Kunce, 2016; Pete Masters and Benjamin Herfort, 2016). As such, this study compares three high spatial detail human settlement datasets for a Global South study site, GHSL, GUF, and HRSL (summarised in Table 1 and Table 3).

A multi-dataset review of high spatial detail products has not been previously completed in a Global South setting. Product accuracy is assessed by the institutions who develop them prior to release, and by independent researchers. These assessments are described by product in section 2.2. Comparative studies of the strengths and weaknesses of multiple high spatial detail products are however few. Small et al. (2016) described anthropogenic land cover datasets of all scales and compared low resolution products by patch (contiguous pixels) size distribution. The study found consistency of distribution between products (Small and Sousa, 2016) . Klotz et al. (2016) completed a comparative analysis of the GHSL and GUF for two test sites in Europe and found they provide advancements in preservation of small-scale settlement complexity (Klotz et al., 2016). Potere et al (2009) compared 8 MR and LR products globally against ground truth data and found MOD500 to have the highest accuracy (Potere et al., 2009). Of these multi product assessments only Klotz et al (2016) focussed on high spatial detail products. Their case study sites were however in Central Europe, a highly urbanised and built up area (Klotz et al., 2016). This

study looks to add value by comparing high spatial detail datasets in a Global South context to understand their strengths and weaknesses in a humanitarian mapping context as elaborated below.

4 Study site and data

In this section, the study site and human settlement datasets are introduced.

4.1 Study site

The assessment will be completed for a study site in the Philippines comprising the islands of Leyte and Samar (Figure 5). This site (43,600km^2) was selected for its varying settlement landscape, data availability, and status as a vulnerable region previously mapped by a Missing Maps VGI campaign.

The human settlement landscape includes 6 cities, 136 towns, and small rural settlements (OpenStreetMap contributors, 2017). The most populous cities include Tacloban (242,000 people), and Calbayog (183,000 people) (Philippines National Statistics Office, 2010). Towns and small fragmented rural settlements dot the islands (OpenStreetMap contributors, 2017). The polarity of settlement types allows human settlement datasets to be tested across varying landscape character (Klotz et al., 2016).

The study site has complete data availability of human settlement datasets in this study, including VGI settlement data from a Missing Maps campaign in 2016/17 (projects 2033, 2109). The campaign was completed in support of Project NOAH (Nationwide Operational Assessment of Hazards). NOAH generates hazard maps and risk maps to support the Philippine government (Reyes, 2016). The campaign over Leyte and Samar highlights the need for VGI human settlement data in the area. This confirms the test sites relevance to the study aim of assessing human settlement datasets in a humanitarian mapping context.

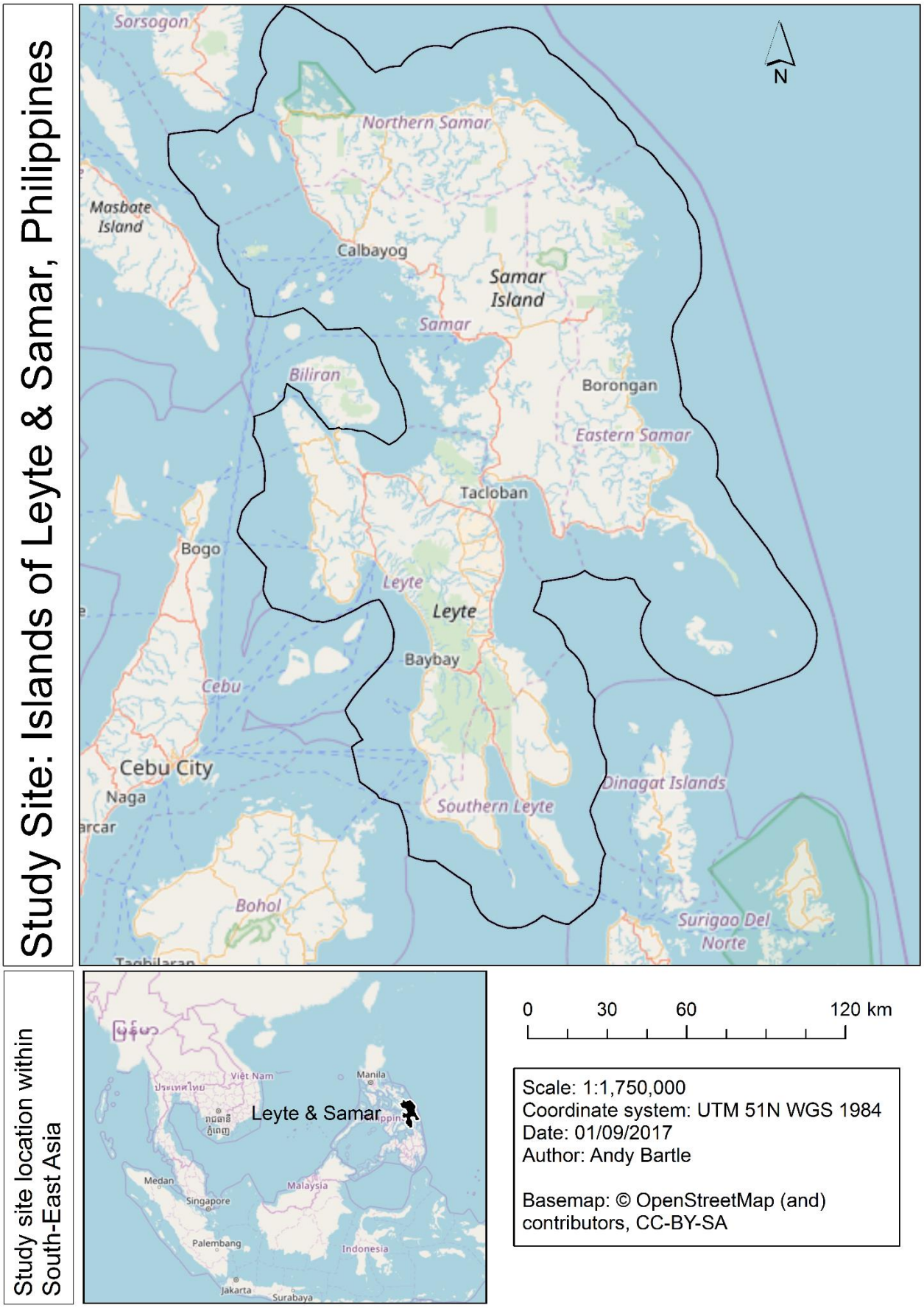


Figure 5: Study site location: Islands of Leyte & Samar, Philippines

4.2 Datasets under test

Datasets under test include GHSL, GUF, and HRSL. These products represent the latest advancements in the detection of human settlements due to the use of HR input imagery and novel detection approaches (Klotz et al., 2016; Weng, 2014). Metadata and acquisition details presented in Table 3.

Human settlement datasets are described as 'high spatial detail' rather than 'high spatial resolution' to reflect the HR input satellite imagery, rather than spatial resolution of final product. For example, HRSL is a MR dataset (30m pixels), developed from VHR imagery (0.5m pixels). It is described as 'high spatial detail' to reflect the HR detection capability.

4.3 Datasets for comparison

Three datasets are incorporated in this study so capabilities can be examined relative to benchmark products. These include OSM, GL30, and MOD500. Metadata and acquisition details in Table 4.

The OSM VGI is a high spatial detail vector building footprint dataset (section 2.1.2). Inter-map comparison will provide understanding of capabilities relative to an existing humanitarian mapping dataset. MOD500 and GL30 are incorporated so capabilities relative to lower spatial resolution products can be assessed. It was suggested that MOD500 offers the highest thematic accuracy of LR products (Potere et al., 2009). GL30 has been used previously as a comparison dataset (Mück, Klotz and Taubenböck, 2017; Chen et al., 2016), and offers high thematic accuracy (section 2.2.3).

4.4 Dataset independence

Integration of independent datasets in remote sensing classification is common (Jensen, 2013; Congalton and Green, 2008). For example, ancillary data is used to provide information for classification, or as reference data for post-processing (Jensen, 2013). Data used to produce a classification is not independent and should not be used as reference data for accuracy assessments (Lillesand and Kiefer, 2000).

This study compares capabilities of products but doesn't calculate absolute accuracies. To qualify inter-map agreement it is noted that GHSL integrated OSM as ancillary data (Joint Research Centre, 2016), and GUF integrated OSM and GL30 as reference data during post processing (DLR-DFD Oberpfaffenhofen, 2016). OSM campaigns for Leyte & Samar were completed following the release of GHSL and GUF datasets (Giraud, 2017) .

Table 3: Summary of datasets under test

Dataset name	Dataset version	Input imagery	Dataset acquisition	Classification	Reference
GHSL	Epoch 2013/14	Landsat (2013/14)	European Commission FTP repository	Binary	(Pesaresi et al., 2016c)
GUF	GUD_DLR_v01	TanDEM-X (2011 to 2012 (93%) 2013/14 (7%))	On request from the German Aerospace Earth Observation Centre	Binary	(Esch et al., 2017),
HRSL	n/a	DigitalGlobe (2010 to 2016)	Center for International Earth Science Information Network website	Continuous (people per pixel)	(Tiecke, 2016)

Table 4: Summary of datasets for comparison

Dataset name	Dataset version	Input imagery	Dataset acquisition	Classification	Reference
OSM VGI	as of 05th August 2017	DigitalGlobe WorldView1, 2 and 3 (18/05/2015, 21/03/16, 01/06/16, 10/06/2016, 24/06/16, 31/07/2016)	Geofabrik (www.geofabrik.de) OSM service provider. OSM Map data is copyrighted OpenStreetMap contributors and available from www.openstreetmap.org (OpenStreetMap contributors, 2017)	Binary (vector building footprints)	(Giraud, 2017)
GL30	2010	Landsat (2010)	on request through the global land cover website (www.globallandcover.com)	Binary	(National Geomatics Center of China, 2014)
MOD500		MODIS (2001-2002)	NASA Land Processes Distributed Active Archive Center	Binary	(Schneider, Friedl and Potere, 2009, 2010)

Table 5: Summary of ancillary datasets

Dataset name	Dataset version	Input imagery	Dataset acquisition	Reference
VIIRS	vcm-orm-ntl	VIIRS 2015 annual composite with cloud mask and intermittent light sources (e.g. fires) removed	NOAA National Centers for Environmental Information Earth Observation Group website	(National Oceanic and Atmospheric Administration, 2017)

4.5 Ancillary data

To compare human settlement datasets across varying landscape character (rural to urban) a spatial zoning dataset is required. In the absence of an authoritative dataset a data driven approach is used. VIIRS NTL luminosity data (described in section 2.2.2) has been widely shown to correlate with population density and urbanisation for areas with electricity access (Elvidge et al., 2013; Zhang and Seto, 2011). In the Philippines 89% of the population has electricity access, including 83% of the rural population (World Bank, 2016). VIIRS is therefore enlisted in this study as a proxy to estimate settlement character zones.

Spatial zones are defined using a simple NTL thresholding approach. By manually calibrating VIIRS NTL luminosity values against OSM city and town extents three zones are delineated (Henderson et al., 2003; L.Imhoff et al., 1997). The thresholds are presented in Table 6 and resulting classification in Figure 6. VIIRS metadata and acquisition details presented in Table 5.

Table 6: Inferred landscape character spatial zones from VIIRS NTL data

Spatial zone	VIIRS NTL threshold	Inferred landscape character description
High NTL	>= 0.4 nanoWatts/cm2/sr	Higher settlement density with a continuous urban fabric.
Medium NTL	< 0.4 and >= 0.1 nanoWatts/cm2/sr	Medium settlement density with a discontinuous urban fabric.
Low NTL	< 0.1 nanoWatts/cm2/sr	Lower settlement density with isolated buildings.

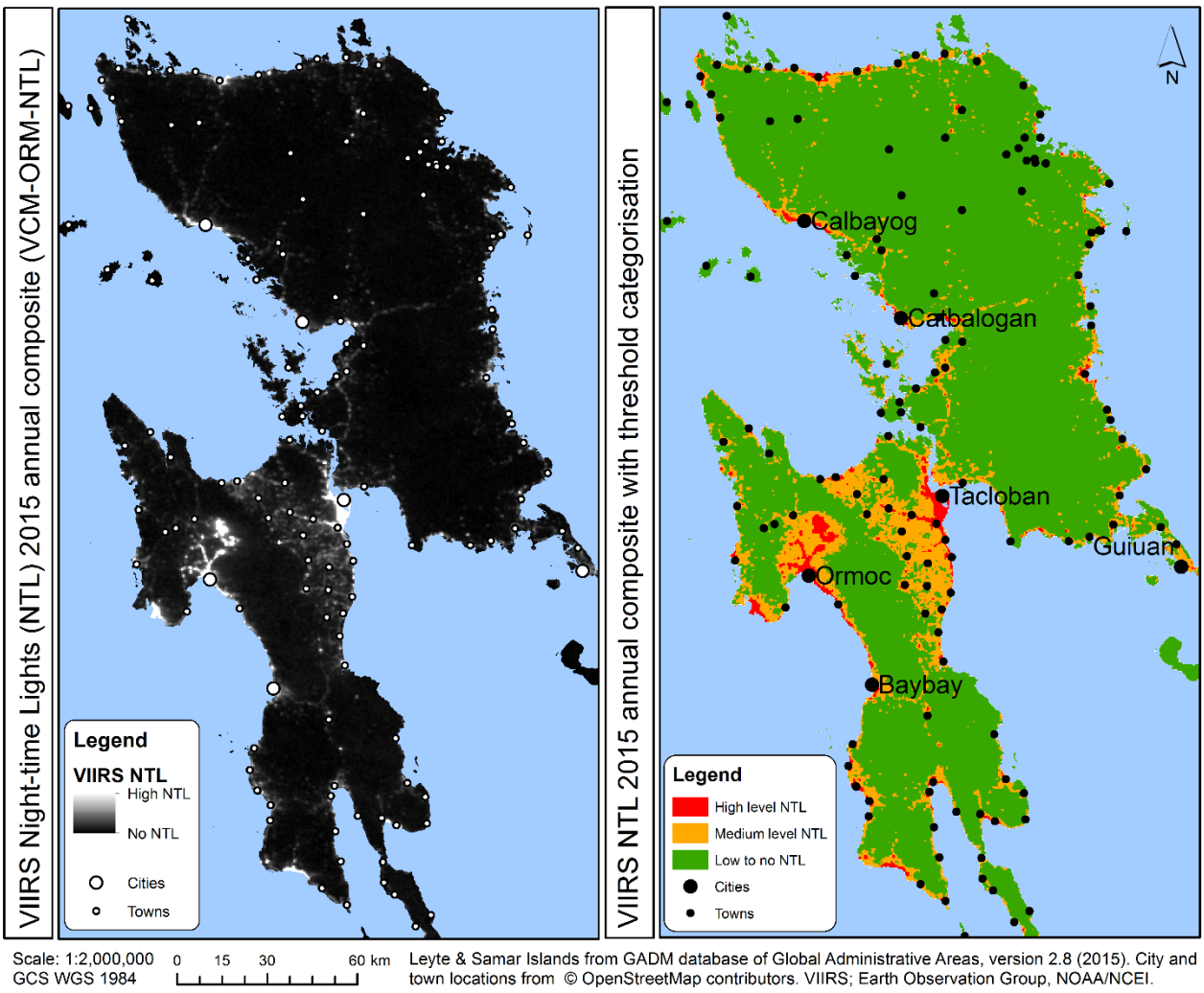


Figure 6: VIIRS NTL Luminosity data landscape character zone threshold categorisation

5 Methodology

A comparison framework is introduced to respond to the research questions (RQs) outlined in the introduction. The comparison framework incorporates techniques to quantify inter-map agreement in total, by settlement character, and by settlement pattern.

5.1 Data pre-processing

Ahead of analysis, all available datasets are processed to facilitate comparison. First, all datasets are transformed to Universal Transverse Mercator (UTM) projection local zones 51 North with ellipsoid World Geodetic System 84. This allows area calculations to be completed. Second, datasets are clipped to the study area extent (shown in Figure 5). Finally, all raster datasets are reclassified to a consistent binary classification, with 1 representing human settlements and 0 representing no human settlements.

5.2 RQ1: To what extent do high spatial detail human settlement datasets vary in a Global South setting?

To answer RQ1, non-site-specific and site-specific inter-map agreement is quantified for each human settlement dataset under test.

5.2.1 Non-site-specific assessment

Non-site-specific assessment of human settlement datasets is completed to understand variation across the study site. First, each human settlement dataset is mapped for visual comparison. Second, the pixel area positively classified by each product is summed, and the pixel quantity counted. The results are plotted for comparison. This non-site-specific approach is routinely conducted for simple multi-dataset comparisons (Congalton and Green, 2008; Potere and Schneider, 2007; Lillesand and Kiefer, 2000; Mück, Klotz and Taubenböck, 2017).

The analysis contributes to Missing Maps understanding of each datasets fitness-for-use by demonstrating variation in the extent of total classification. Comparison between datasets indicates relative agreement or disagreement of total human settlement area (Congalton and Green, 2008).

5.2.2 Site-specific assessment

Site-specific comparison extends the first approach by comparing map agreement for individual pixels. Site-specific comparison of classifications against reference data is an accepted approach for assessing dataset accuracy (Foody, 2006; Congalton, 1991;

Congalton and Green, 2008). Absolute accuracy is beyond the scope of this study as appropriate reference data is unavailable. Availability of human settlement reference data is an inherent problem in areas requiring humanitarian mapping VGI campaigns. In the absence of reference data, site-specific inter-map agreement is calculated to understand relative capabilities. This approach was used in studies to quantify map agreement and evolution (Potere and Schneider, 2007; Klotz et al., 2016).

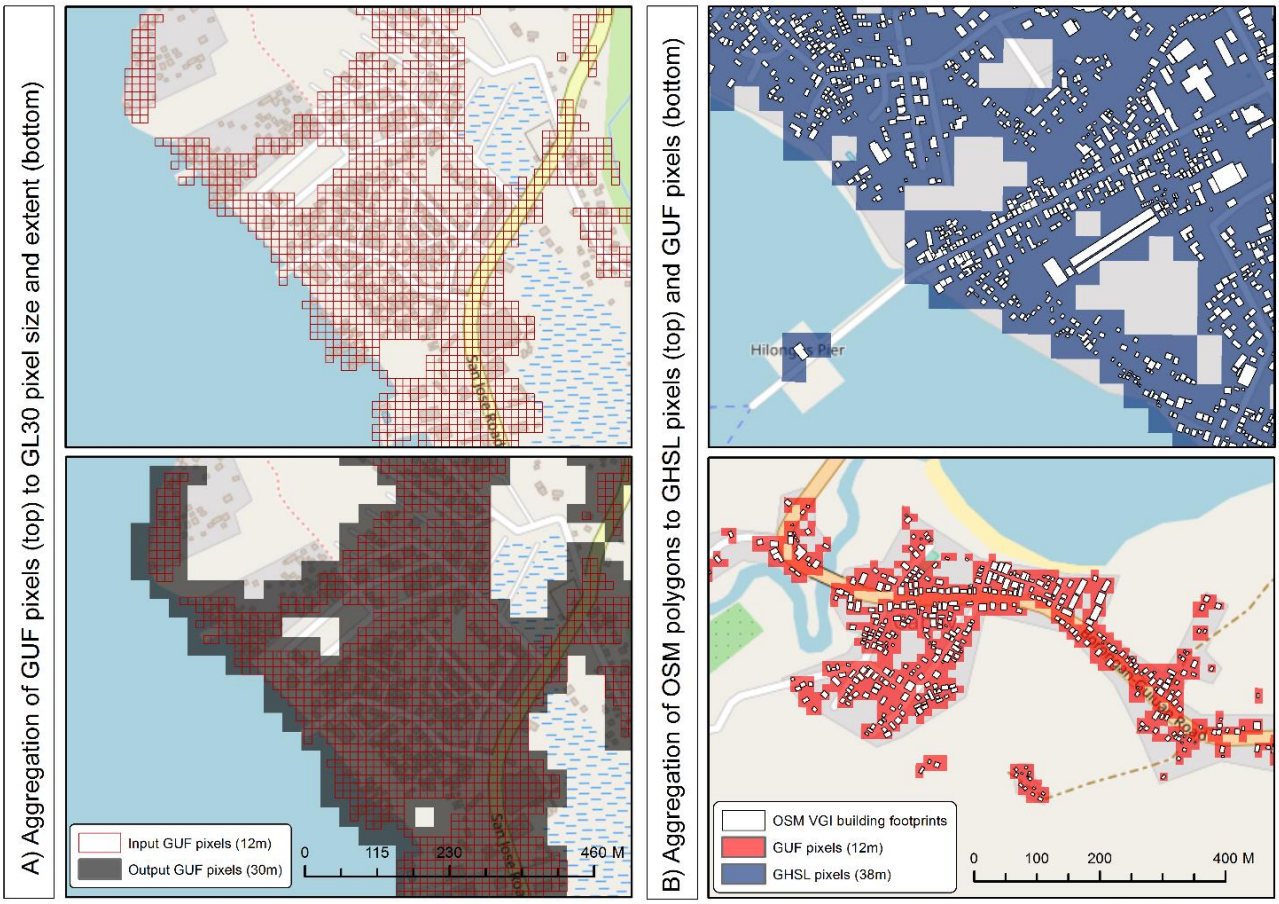
Inter-map agreement calculation follows an accuracy assessment methodology. Agreement is calculated for each dataset under test relative to all other products (described as 'comparison pairs') (comparison matrix in Table 7). An object-based point-in-polygon (PIP) test is applied to compare pixel sites, and error matrix used to quantify results.

Table 7: Dataset comparison matrix (X indicates inter-map agreement assessment conducted)

		Comparison dataset					
		GHSL	GUF	HRSL	OSM	GL30	MOD500
Dataset under test	GHSL		X	X	X	X	X
	GUF	X		X	X	X	X
	HRSL	X	X		X	X	X

A PIP test compares a classified pixel with a spatially overlaid comparison site (reference data point) to determine relative agreement (Foody, 2002). In this study, we compare each pixel with the full pixel extent of the comparison dataset (rather than a point). This allows the potential extent of a settlement within the pixel to be represented, thus acknowledging that the classifier could have detected a human settlement at any location within the pixel (Rutzinger, Rottensteiner and Pfeifer, 2009; Weng, 2014).

A direct comparison requires equal geometric resolution (Potere and Schneider, 2007). For each dataset under test, the comparison datasets are resampled to the same raster size and extent. This is completed by vectorising positively classified pixels, and rasterising these to the size of the dataset under test (Congalton and Green, 2008). This is illustrated in Figure 7. In (A) positively classified GUF pixels are resampled to GL30 extent. In (B) OSM VGI polygons are aggregated to GHSL and GUF pixels.



Basemap © OpenStreetMap (and) contributors, CC-BY-SA. Globeland30; National Geomatics Centre for China. Global Urban Footprint (GUF); DLR 2016.

Figure 7: Resampling of human settlement datasets for PIP test comparison

The dataset under test and resampled comparison dataset are reclassified with unique base two integers (designated in Table 8) and the two datasets are summed to identify locations where pixels agree and disagree (logic detailed in Table 8) (Beazley and Jones, 2013).

Table 8: Inter-map agreement bitwise logic for raster summing

Dataset under test	Value	Comparison dataset	Value	Total	Result
Human settlement	1	Human settlement	4	1 + 4 = 5	True positive (TP)
Human settlement	1	No settlement	8	1 + 8 = 9	False positive (FP)
No settlement	2	Human settlement	4	2 + 4 = 6	False negative (FN)
No settlement	2	No settlement	8	2 + 8 = 10	True negative (TN)

Results of the inter-map pixel comparison (Table 8) are presented by error matrix. An error matrix is a widely used method of quantifying accuracy assessments. It provides a cross-tabulation of each dataset under test with its comparison dataset, and descriptive statistics

to quantify differences (Foody, 2002; Jensen, 2013). Whilst these statistics are widely used in accuracy assessments, each has limitations from sensitivity to different components of the error matrix. Several statistics are therefore used to assess different components (Table 9) (Lillesand and Kiefer, 2000; Foody, 2002).

This site-specific quantification of inter-map agreement provides locational understanding of map differences (Congalton and Green, 2008) and allows Missing Maps to understand map agreement and relative classification capabilities for each dataset under test.

Table 9: Error statistics

Statistic name	Description and limitations	Equation
1. Positive predictive value (PPV)	PPV is the proportion of true positive results, or how often the classification will be correct relative to the comparison dataset. PPV is also known as user's accuracy and precision. The inverse is the error of commission (Congalton and Green, 2008).	$$PPV = \frac{TP}{TP + FP}$$ **Equation 1: Positive predictive value**
2. True positive rate (TPR)	TPR is the proportion of correctly identified positives relative to the comparison dataset. TRP is also known as producer's accuracy, recall and sensitivity. The inverse is the error of omission (Congalton and Green, 2008).	$$TPR = \frac{TP}{TP + FN}$$ **Equation 2: True positive rate**
3. True negative rate (TNR)	TNR is the proportion of correctly identified negatives relative to the comparison dataset TNR is also known as specificity (Congalton and Green, 2008).	$$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{FP} + \mathbf{TN}}$$ **Equation 3: True negative rate**
4. Overall accuracy (OA)	OA is the proportion of correctly classified pixels. This simple metric is highly vulnerable to bias from prevalence (Jeni, Cohn and De La Torre, 2013).	$$A = \frac{TP + TN}{TP + FP + TN + FN}$$ **Equation 4: Overall accuracy**
5. F-score (F)	F is the harmonic mean of PPV and TPR, and so considers precision and recall. Perfect F is reached at 1, and worst at 0. Again, by integrating PPV and TPR in one metric F is affected by prevalence (Lantz and Nebenzahl, 1996)	$$F = \frac{2TP}{2TP + FP + FN}$$ **Equation 5: F-score**
6. True-Skill-Statistic (TSS)	TSS considers PPV and TPR (as with K and F). The metric is corrected for chance (as with K), but is also corrected for dependence on prevalence. TSS therefore offers the advantages of K without dependence on prevalence (Allouche, Tsoar and Kadmon, 2006). Perfect TSS is reached at 1, and worst at -1 (Jeni, Cohn and De La Torre, 2013).	$$TSS = TPR + TNR - 1$$ **Equation 6: True-Skill-Statistic**
7. Kappa (K)	K measures the level of agreement relative to the comparison dataset corrected for chance. K extends on OA by adjusting for the estimated effect of chance. Perfect agreement is reached at 1, and complete randomness at 0 (or below) (Foody, 2002). By integrating PPV and TPR in one metric K is affected by prevalence (Lantz and Nebenzahl, 1996).	$$\boldsymbol{K} = \frac{\frac{(\boldsymbol{TP} + \boldsymbol{TN})}{\boldsymbol{n}} - \frac{(\boldsymbol{TP} + \boldsymbol{FP})(\boldsymbol{TP} + \boldsymbol{FN}) + (\boldsymbol{FN} + \boldsymbol{TN})(\boldsymbol{TN} + \boldsymbol{FP})}{\boldsymbol{n^2}}}{\mathbf{1} - \frac{(\boldsymbol{TP} + \boldsymbol{FP})(\boldsymbol{TP} + \boldsymbol{FN}) + (\boldsymbol{FN} + \boldsymbol{TN})(\boldsymbol{TN} + \boldsymbol{FP})}{\boldsymbol{n^2}}}$$ **Equation 7: Kappa**

5.3 RQ2: To what extent do these datasets vary for different settlement landscape character?

Urban settlement mapping has greatly improved however extraction of fragmented rural settlements remains a challenge (Weng, 2014). Assessment techniques which respect this variation in detection complexity are recommended (Taubenbock et al., 2011; Foody, 2006; Klotz et al., 2016). RQ2 therefore looks at inter-map agreement by settlement character (rural to urban) as defined by NTL spatial zones, and settlement size.

5.3.1 Inter-map agreement by NTL spatial zoning

Inter-map agreement is calculated by spatial zone using VIIRS NTL luminosity data. The spatial zones represent inferred settlement character (described in section 4.5). Inter-map agreement is calculated for each comparison pair using a PIP test and error matrix (as described in section 5.2.2).

The landscape character raster dataset (NTL spatial zones) is resampled to the size and extent of the dataset under test for direct comparison. The three raster datasets (dataset under test, comparison dataset, NTL spatial zones) are then reclassified with unique base 2 integers, and summed to identify locations where pixels agree and disagree (Table 10). An error matrix is used to quantify the results (Table 9).

Table 10: Bitwise binary logic for raster sums by landscape character (HS = human settlement)

Landscape character		Dataset under test		Comparison dataset		Total	Result
Low NTL	16	HS	1	HS	4	= 21	Low NTL - TP
	16	HS	1	No-HS	8	= 25	Low NTL - FP
	16	No-HS	2	HS	4	= 22	Low NTL – FN
	16	No-HS	2	No-HS	8	= 26	Low NTL - TN
Medium NTL	32	HS	1	HS	4	= 37	Med NTL – TP
	32	HS	1	No-HS	8	= 41	Med NTL – FP
	32	No-HS	2	HS	4	= 38	Med NTL – FN
	32	No-HS	2	No-HS	9	= 42	Med NTL – TN
High NTL	64	HS	1	HS	4	= 69	High NTL – TP
	64	HS	1	No-HS	8	= 73	High NTL – FP
	64	No-HS	2	HS	4	= 70	High NTL – FN
	64	No-HS	2	No-HS	8	= 74	High NTL - TN

5.3.2 Inter-map agreement by settlement size

As a second part in understanding classification by landscape patterns, inter-map agreement is quantified by settlement size. This provides a user-oriented measure of map agreement that considers this varying parameter. A similar approach was applied to quantify completeness and correctness of building footprint digital elevation models by building area and height (Wurm et al., 2014; Taubenbock et al., 2011; Rutzinger, Rottensteiner and Pfeifer, 2009)

First, human settlement size is compared between datasets using a frequency distribution chart. This is completed by merging contiguous human settlement pixels into patches (polygons representing human settlement areas), counting these by patch size bins, and plotting the results (Klotz et al., 2016).

Second, inter-map agreement is quantified by human settlement size using the raster images produced in section 5.2.2. Raster datasets for each comparison pair display the spatial distribution of TP, FP, FN, and TN pixels. Contiguous pixels of each type are merged into patches. This produces polygons representing areas of agreement (TP and TN), and disagreement (FP and FN). These are then counted by patch size bins, and error statistics (Table 9) calculated to describe accuracy by patch size (Wurm et al., 2014; Taubenbock et al., 2011).

This assessment informs understanding of dataset relative capabilities with respect to landscape character. The use of zones for settlement density, and contiguous pixels for settlement size provides varying representations of this phenomenon. The analysis allows Missing Maps to understand capabilities across varying character, including fragmented settlements which are complex to detect but critical to the humanitarian mapping target (Rutzinger, Rottensteiner and Pfeifer, 2009; Missing Maps, 2017).

5.4 RQ3: What factors may contribute to inter-map disagreement for these datasets?

Inter-map agreement analysis completed for RQ1 and RQ2 identified site-specific disagreement between comparison pairs. The disagreements include errors of commission and omission.

Errors of commission are represented by FPs, occurring when there is no inter-map match for a pixel positively classified by the dataset under test. Errors of omission are represented by FNs, occurring when a negatively classified pixel has a positively classified inter-map

match. The errors are identified through inter-map comparison (rather than reference data comparison) so the true source of the disagreement is unknown. Disagreement could be caused by an error in either the dataset under test or the comparison dataset. Possible scenarios are presented in Table 11.

RQ3 looks to identify potential factors contributing to inter-map disagreement, including the source of disagreement (which dataset), and potential causes of classification differences.

Table 11: Inter-map disagreement scenarios

Inter-map disagreement	Disagreement scenario	Dataset in error	Correct dataset
False positive	Human settlement present	Comparison dataset	Dataset under test
	No human settlement	Dataset under test	Comparison dataset
False negative	Human settlement present	Dataset under test	Comparison dataset
	No human settlement	Comparison dataset	Dataset under test

5.4.1 Inter-map disagreement dataset error source

To understand the source of inter-map disagreement the proportion of error results attributed to error by either the dataset under test, or its comparison dataset is quantified. This is completed through a manual assessment of randomly selected error pixels.

The sample size is calculated using a binomial probability theory (Equation 8). This approach is based on established statistical theory and has been used in remote sensing to determine sample size for remote sensing reference data (Foody, 2009a; Jensen, 2013). Three terms are specified, confidence level z-score, margin of error, and expected percentage accuracy. Typed used values have been adopted for a sample with 95% confidence level and 5% margin of error (Foody, 2009a).

$$sample\ size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$

N = population size
p = expected percent accuracy
e = margin of error
z = z-score for confidence level

Equation 8: Binomial probability sample size equation

The sample is selected using the Python random tool to return unique random pixels until the specified sample size is met (Python Software Foundation, 2017). The randomly selected sample is then vectorised and analysed in Google Earth to determine building

presence per pixel. Satellite imagery to match imagery year is used when cloud cover permits.

5.4.2 Causes of classification differences

Two potential causes of classification differences are assessed. These include, classification differences associated with proximity to human settlements, and errors of omission due to bad imagery (cloud cover or low resolution).

Classification differences associated with proximity to human settlements can be caused by technical problems or semantic definition differences (Weng, 2014). Technical problems such as misregistration, or quantization errors can result in misaligned datasets and therefore local classification differences (Rutzinger, Rottensteiner and Pfeifer, 2009). Semantic differences between datasets result in comparison of different targets (Potere et al., 2009). For example, GL30 classifies 'lands modified by human activities' which includes roads and parks. GUF classifies 'built-up area with vertical structuring' which encompasses buildings and structures only (definitions of each dataset in Table 1).

Spatial analysis techniques are used to quantify the proportion of FN and FP results adjacent to settlements. TP results are buffered by 30m (or one pixel space) to define a zone surrounding human settlements. The quantity of FN and FP pixels that spatially intersect this zone are counted to quantify the proportion of FN and FP errors adjacent to human settlements.

Classification differences relative to OSM due to bad imagery are quantified. Imagery tiles with digitiser comments noting bad imagery are collated to define a zone of bad imagery (Giraud, 2017). Error rates are then calculated (area of errors / area of zone) for the bad imagery zone, and good imagery zone. Equivalent analysis for GHSL, GUF, and HRSL is not completed as there are no bad imagery warnings for the study site.

By identifying the source and potential causes of inter-map disagreement Missing Maps can understand the strengths and weaknesses of each dataset and therefore fitness-for-use for humanitarian mapping.

6 Results and Discussion

The results and discussion is presented by research question (chapter 1) and methodological approach (chapter 5).

6.1 RQ1: To what extent do high spatial detail human settlement datasets vary in a Global South setting?

In this subsection, we address RQ1 by comparing non-site-specific and site-specific inter-map agreement of each dataset under test.

6.1.1 Map pattern and total area comparison for study-site

We first investigate variation by comparing map extent and the total area classified by each product.

Mapped representations of the datasets under test are shown in Figure 9, and the datasets for comparison in Figure 10. Visual comparison indicates that the three datasets under test, and the higher detail OSM VGI share a similar extent and pattern. These datasets are all more spatially expansive than GL30 and MOD500, which show only a few patches. Additionally, these higher detail datasets identify fragmentation of the settlement pattern around Tacloban city, where GL30 and MOD500 do not offer this spatial detail. These differences reflect the greater detail offered by the datasets under test in detecting human settlement patterns beyond urban cores. The extent and pattern similarities between OSM VGI and the datasets under test indicates similar detection abilities between these high spatial detail products.

Total land area classified as human settlement by each product is shown in Figure 8. This indicates significant disagreement between datasets. Variation between all datasets differs four-fold (415% increase) and variation between the datasets under test is three-fold (290%). HRSL has the greatest spatial share of human settlement area, possibly due to the aggregation of VHR input data (0.5m) to MR 30m pixels. This would result in spatial generalisation of high detail input data (for example a small building detected in 0.5m imagery will be represented by the spatial extent of a 30m pixel following aggregation). OSM has the lowest spatial share due to the high detail of the vector building footprints. The plotted pixel / polygon count is associated with spatial resolution and provides an insight into the spatial detail of each product. For example, human settlement classification by MOD500 is twice that of OSM but with a fraction of the pixel / polygon count. This indicates that the spatial detail for OSM is much higher than MOD500.

Lack of agreement between datasets is assumed to be related to errors of omission and commission. Non-site-specific analysis does not allow us to identify where difference occur. We therefore build on current results by completing a site-specific comparison of datasets.

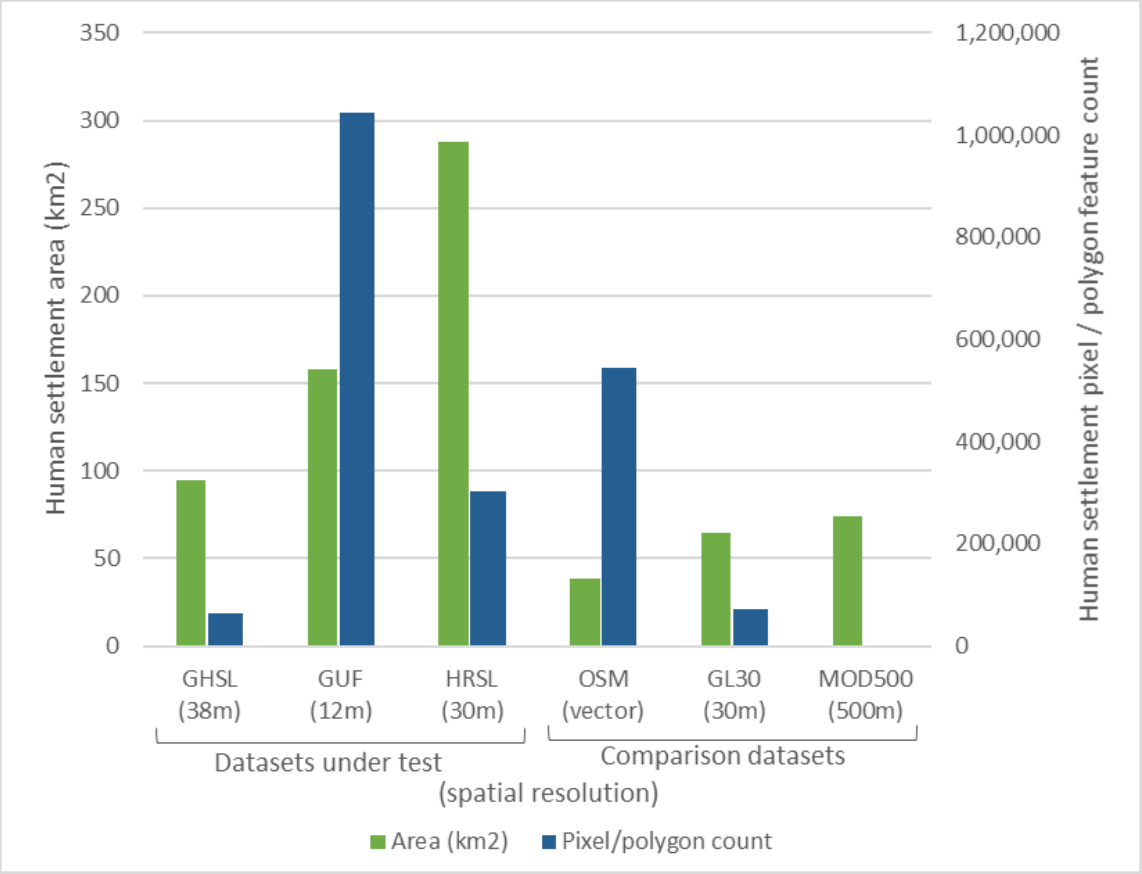


Figure 8: Total area classified as human settlement (km2), and total pixel/polygon count for human settlements datasets included in this study (Leyte and Samar, Philippines).

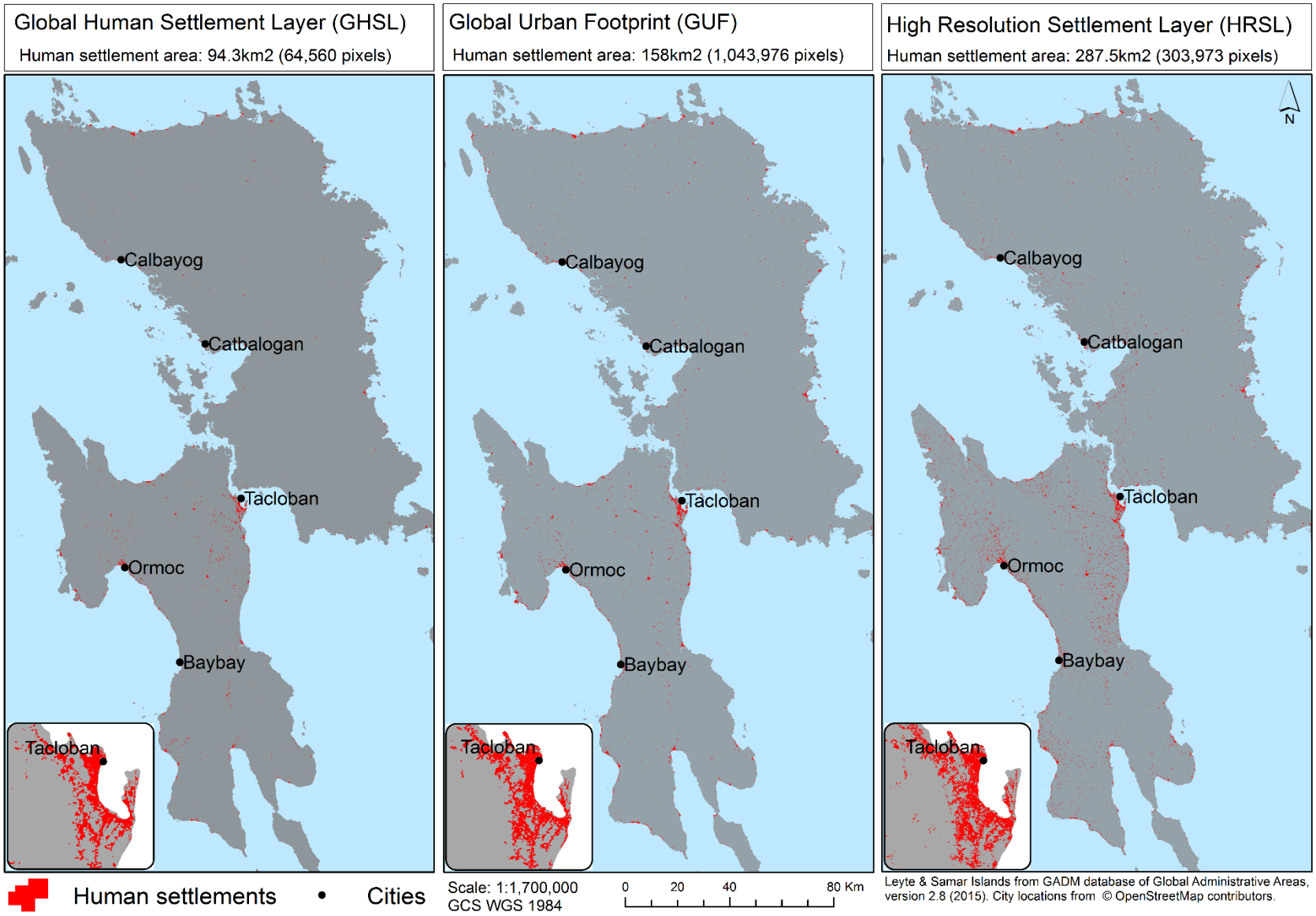


Figure 9: Map view of human settlement datasets under test, with inset map for area around Tacloban city (Leyte and Samar, Philippines).

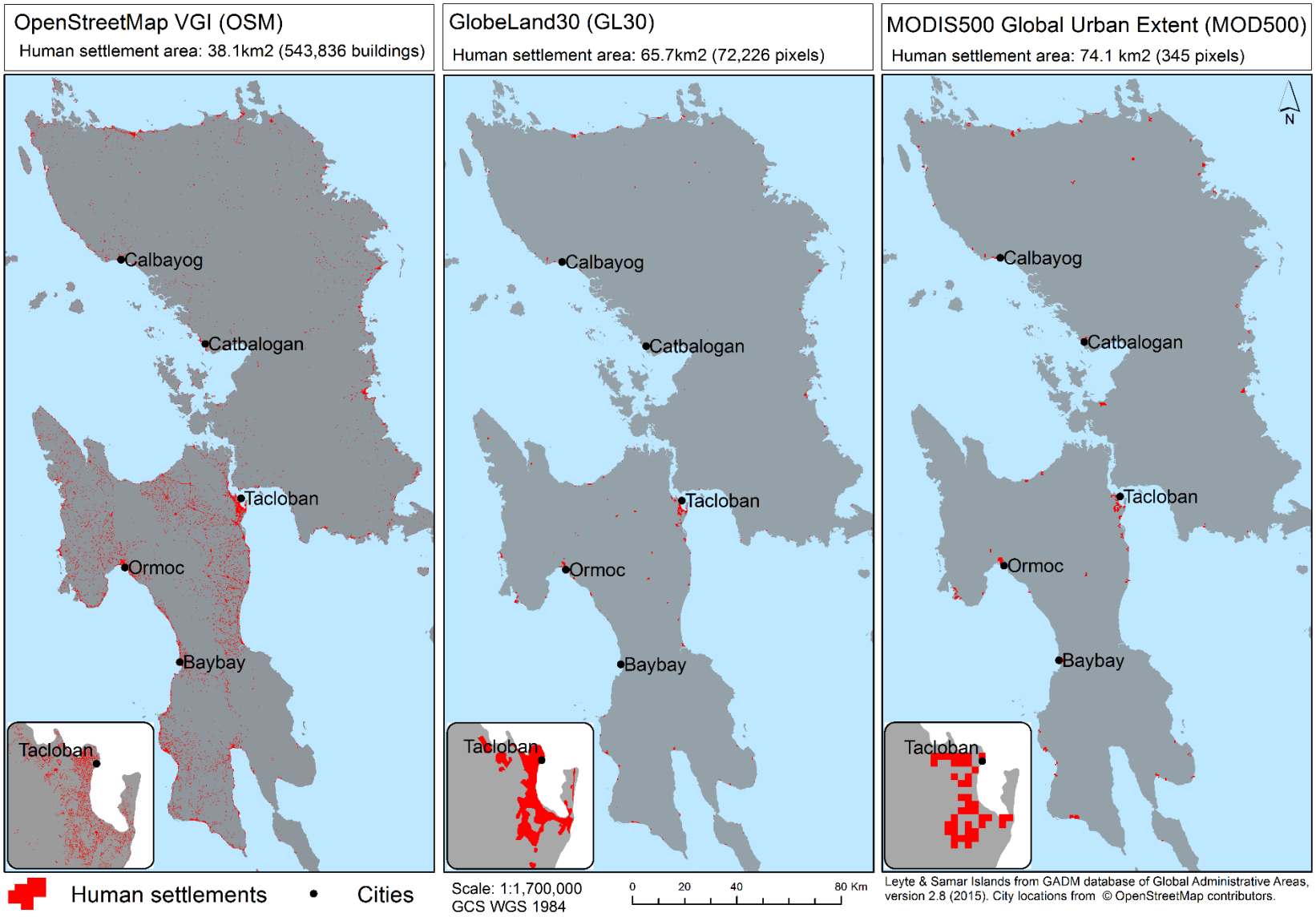


Figure 10: Map view of human settlement datasets for comparison, with inset map for area around Tacloban city (Leyte and Samar, Philippines).

6.1.2 Site-specific inter-map agreement and disagreement

Following non-site-specific analysis, we complete a site-specific assessment by calculating inter-map agreement for each comparison pairs (Table 7).

Error statistics for the quantification of inter-map agreement are presented as bars charts in Figure 12 (GHSL), Figure 13 (GUF), and Figure 14 (HRSL).

Inter-map agreement can be described overall by error statistics OA, F, K, and TSS. For all datasets (Figure 12 to Figure 14) OA implies high accuracy (consistently > 98%). However, the prevalence of negative classifications results in a highly skewed result and so the statistic is not meaningful (Jeni, Cohn and De La Torre, 2013; Li and Guo, 2014).

Multivariate error statistics F and K indicate poor to moderate classification accuracy. F and K show that inter-map agreement of GHSL (Figure 12) is strongest relative to GL30, and weakest relative to HRSL. Agreement relative to GL30 could be due to the shared source input imagery (Landsat 30m spatial resolution). Disagreement relative to HRSL may be caused by significantly finer resolution of HRSL VHR imagery (0.5m spatial resolution). F and K indicate minor variation in inter-map agreement of GUF (Figure 13) relative to all products, other than MOD500. Both GUF and HRSL show weakest inter-map agreement relative to MOD500 due its coarse spatial resolution with low patch count and spatial detail. HRSL agreement (Figure 14) is strongest relative to GUF and OSM.

TSS shows the greatest relative difference between products due it its low sensitivity to prevalence (Allouche, Tsoar and Kadmon, 2006). TSS indicates inter-map agreement is greatest for all datasets under test relative to GL30. This is due to the relative high TP and low FN of classifications relative to this MR dataset. The detection of urban areas from 30m Landsat data by GL30 is repeatable by datasets under test.

Class specific errors are measured by PPV and TPR. GHSL PPV results (Figure 12) show that GUF, HRSL, and OSM positively classify 77%, 82%, and 66% respectively of the human settlement pixels identified by GHSL. However, TPR results show that GHSL only classified 26% of GUF human settlement pixels, 11% of HRSL, and 15% of OSM. This indicates approximately 80% of GHSL positively classified pixels are also detected by the GUF and HRSL datasets, and that GHSL misses between 89% and 74% of pixels classified by GUF and HRSL. GUF PPV results (Figure 13) are lower for comparison datasets than GHSL results. Of the pixels positively classified by GUF, the GHSL, HRSL and OSM datasets provide a match for 45%, 78%, and 48% respectively. TRP shows that GUF also classifies 62% of the

positively classified GHSL pixels, 30% of HRSL pixels, and 45% of OSM. This indicates that approximately half of the pixels classified as human settlement by GUF are not detected by GHSL and OSM, whereas 80% are detected by HRSL. In addition, GUF fails to detect 70% of HRSL human settlement pixels. HRSL PPV results (Figure 14) for all comparison datasets but OSM are lower than that of GHSL and GUF. GHSL, GUF, and OSM positively classify 17%, 48%, and 56% of HRSL human settlement pixels respectively. TPR results show that HRSL classifies between 43% and 56% of GHSL, GUF and OSM. HRSL therefore classifies significantly more pixels as human settlements than other datasets under test, and misses less of the pixels classified by other datasets than GHSL and GUF.

For all datasets under test PPV relative to the low spatial detail GL30 and MOD500 products is low (between 8% and 43%). This is due to the capability of high spatial detail datasets in detecting a greater extent of fragmented settlements, and represents the evolution of these products relative to GL30 and MOD500. In addition, for all datasets under test TPR relative to MOD500 is low, meaning inter-map agreement of positively classified MOD500 pixels is poor. This is due to the coarse MOD500 pixels (500m spatial resolution) covering a large mix of land use. A human settlement may cover only 50% of a MOD500 pixel and so over classifies relative to high spatial detail datasets (illustrated in Figure 11-B) (Schneider, Friedl and Potere, 2009).

This site-specific assessment reveals poor inter-map agreement between comparison pairs. Relative to high spatial detail datasets GHSL shows low errors of commission, high errors of omission, and low total classified area (Figure 8). This indicates GHSL offers a conservative classification of human settlements, with high relative correctness, but low completeness. It may be useful for identification of urban settlements, but not fragmented settlements.

HRSL shows high errors of commission, moderate errors of omission, and very high total classified area. The dataset classifies areas as human settlements that are not identified in other datasets, and has the lowest relative errors of omission. This may represent high completeness, high rates of false-positives, or a combination of both. For example, Figure 11-A illustrates dispersed settlements on the fringes of Tacloban city that are positively classified by HRSL but with no GUF inter-map agreement. Each could be HRSL TP and GUF FN, or HRSL FP and GUF TN. This dataset, if proven to be accurate, may offer high completeness in its classification of human settlements, and could therefore be a suitable product for humanitarian mapping.

GUF shows moderate rates of commission and omission. GUF classifies areas as human settlement not found in comparison datasets. However, the recall of GUF relative to HRSL is low (70% error of omission). GUF results place between the relative conservative classification of GHSL and generous classification extent of HRSL.

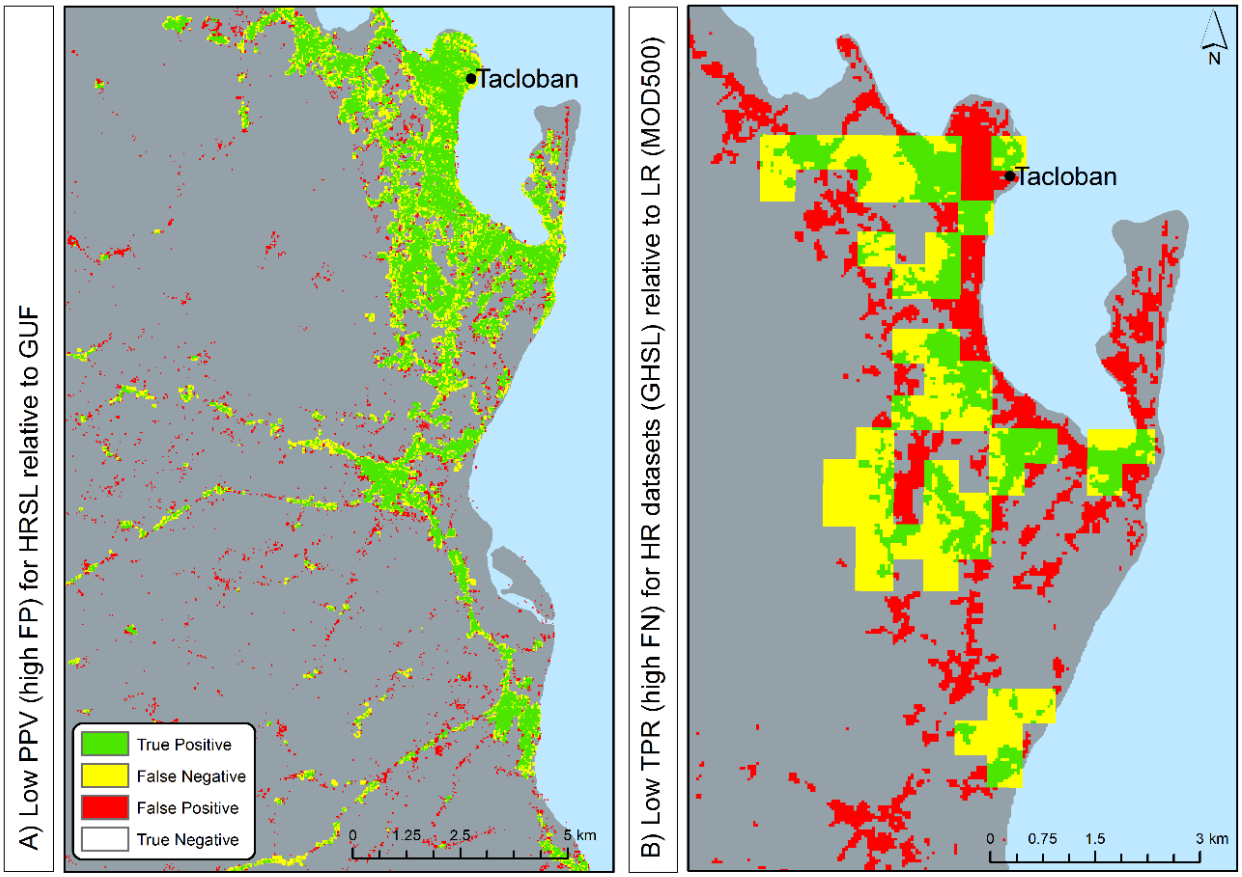


Figure 11: Inter-map agreement illustrations, A) HRSL FP relative to GUF B) MOD500 high FN relative to GHSL

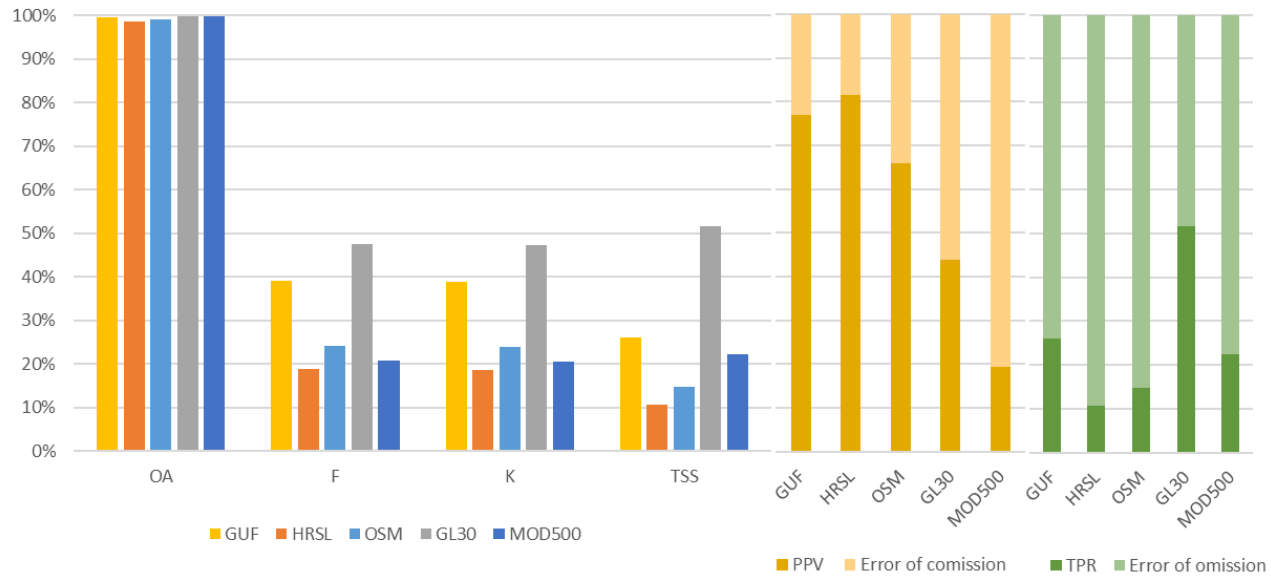


Figure 12: GHSL inter-map agreement error matrix results

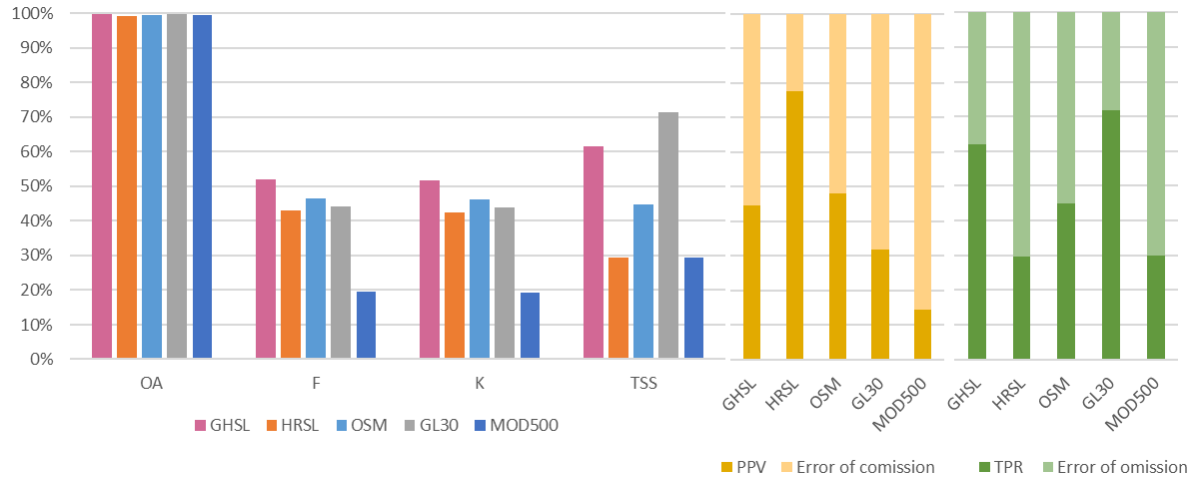


Figure 13: GUF inter-map agreement error matrix results

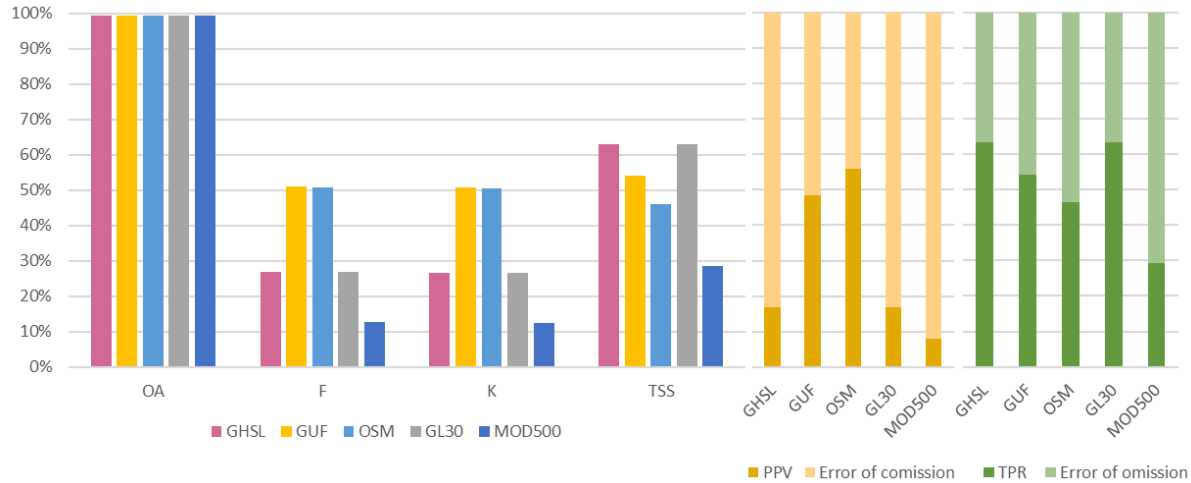


Figure 14: HRSL inter-map agreement error matrix results

6.2 RQ2: How do the datasets vary for different settlement landscape character?

We address RQ2 by quantifying inter-map agreement by settlement landscape character. Landscape character is spatially defined using NTL zones to infer settlement density, and contiguous cells to represent settlement size.

6.2.1 Inter-map agreement by NTL spatial zoning

Assessment of Inter-map agreement in RQ1 indicated significant disagreement between human settlement datasets. This approach quantifies inter-map agreement by VIIRS NTL spatial zones to understand differences in inter-map agreement by inferred settlement density.

Error statistics PPV, TPR, and F are presented by spatial zone in Figure 15, Figure 16, and Figure 17. Full error statistics are in Appendix A. PPV, TPR and F are used as summary statistics due to their applicability to highly skewed binary datasets. The statistics quantify the error matrix without considering high prevalence TN results (no human settlements) (Congalton and Green, 2008; Allouche, Tsoar and Kadmon, 2006).

The presented results show that PPV, TPR, and F increase with NTL luminosity zone (zones thresholds described in Table 6). This corresponds to greater inter-map agreement for zones inferred to represent higher settlement density with a continuous urban fabric, than zones of lower settlement density with isolated buildings. This trend reflects the greater complexity of classification for fragmented rural settlements over urban centres (Giri et al., 2013). Results relative to MOD500 for low and medium NTL zones are inconsistent due to sparse data in these areas, with coarse 500m pixels covering a large mix of land use resulting in high relative FN results.

As found in RQ1 GHSL shows high correctness (PPV), but low completeness (TRP) relative to high spatial detail datasets. This is more apparent in areas of low and medium NTL (PPV 30% to 65%, and TPR < 15%). GHSL fails recall > 85% of the settlements classified by high spatial detail pairs in areas of low and medium NTL. Quality (F) shows low inter-map agreement (F < 20%) across low and medium NTL zones.

GUF results relative to HRSL indicate high correctness across all zones (PPV 69% to 82%), but low completeness in areas of low and medium NTL (TPR 15% to 21%). HRSL classified a high proportion of areas detected by GUF, but GUF fails to recall many areas detected by

HRSL. Quality (F) shows GUF has low inter-map agreement (F 20% to 35%) for low and medium NTL zones.

HRSL results relative to GUF and OSM show moderate correctness (PPV 19% to 53%), and moderate completeness (TPR 30% to 48%). HRSL positively classifies many areas that have no inter-map agreement. Quality (F) shows HRSL has moderate inter-map agreement relative to GUF and OSM (F 27% to 46%) for low and medium NTL zones.

GHSL offers the lowest relative capability in zones of low and medium NTL, with low completeness relative to high spatial detail datasets. HRSL offers greatest relative capability, as demonstrated by optimal completeness (TPR) and quality (F) of the datasets under test relative to high spatial detail datasets pairs (OSM, GUF, HRSL).

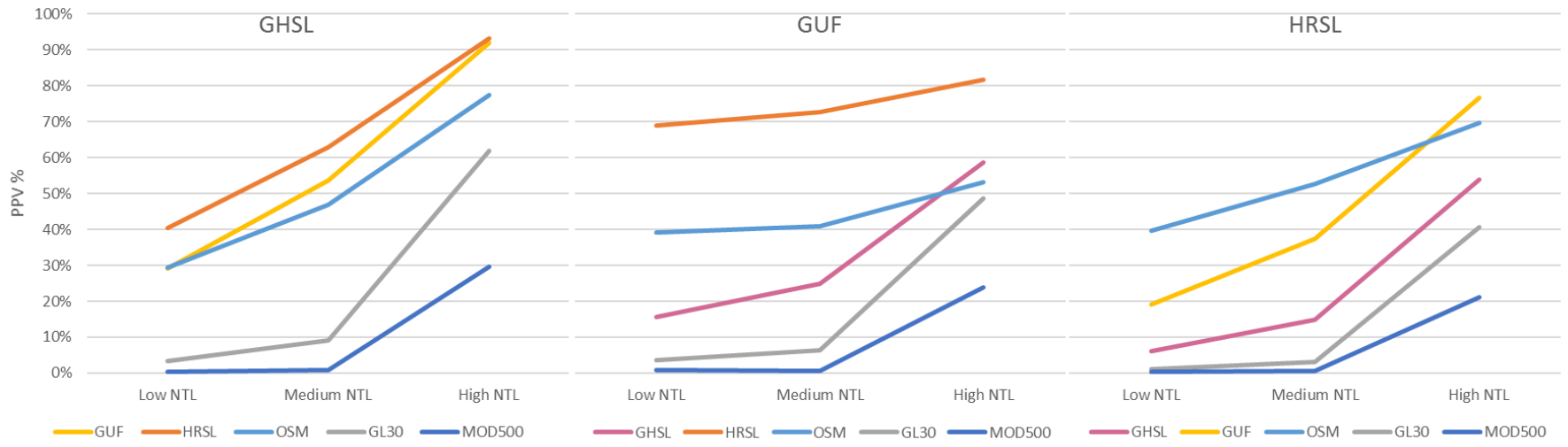


Figure 15: Change in PPV by NTL spatial zones for datasets under test

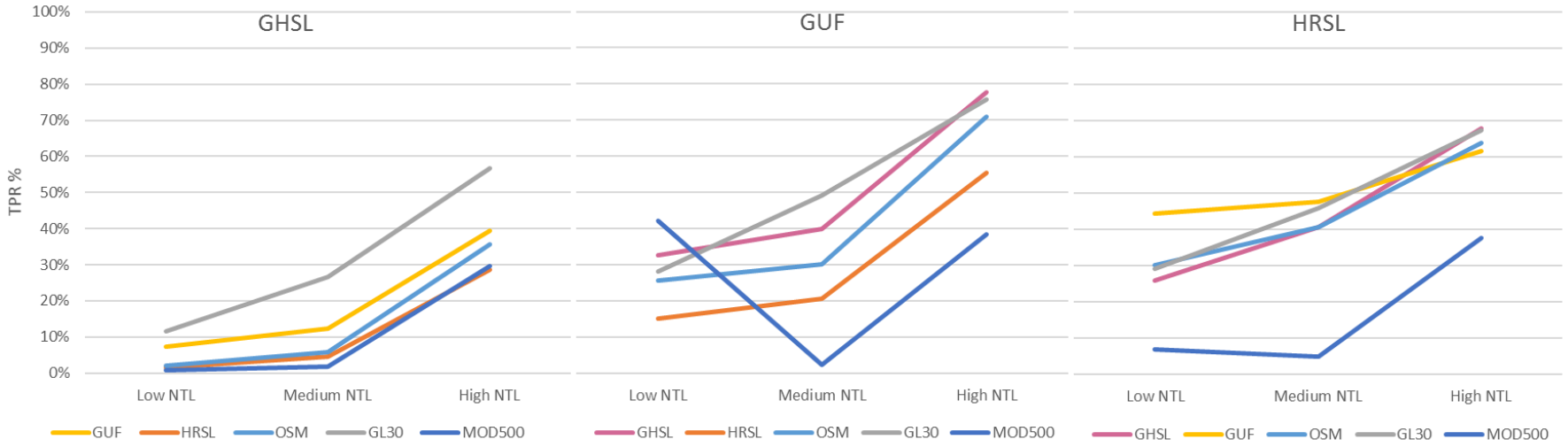


Figure 16: Change in TPR by NTL spatial zones for datasets under test

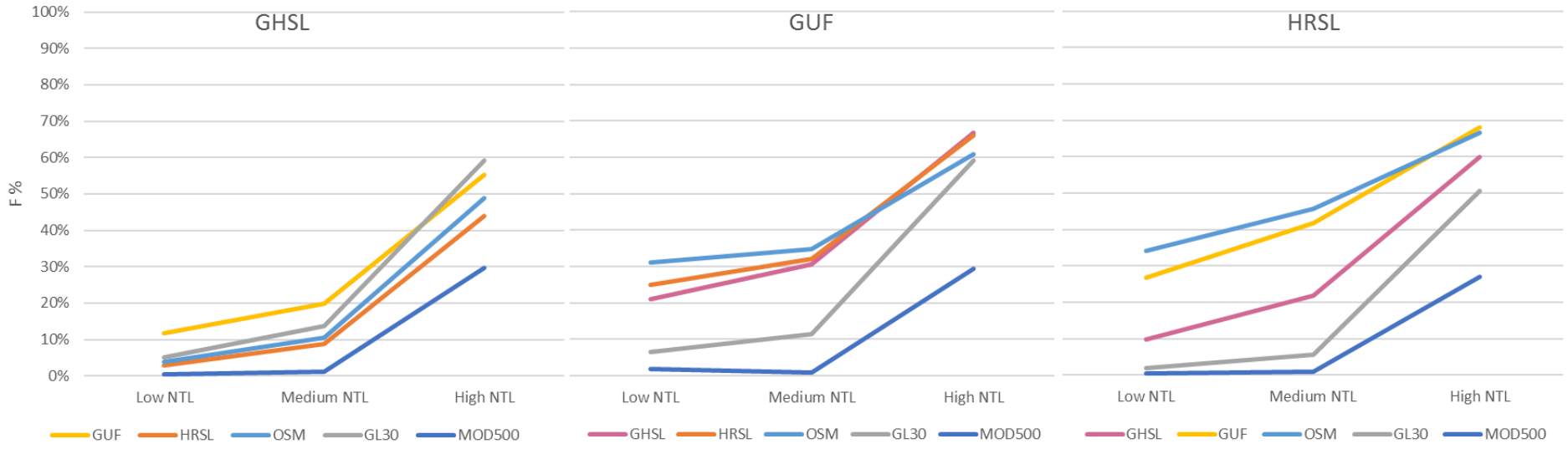


Figure 17: Change in F by NTL spatial zones for datasets under test

6.2.2 Inter-map agreement by settlement size

Variation of inter-map agreement by settlement size is quantified to understand relative capabilities of each dataset under test for different scales. Settlement size is defined using a pattern-based assessment where contiguous human settlement pixels are merged into patches (Wurm et al., 2014; Rutzinger, Rottensteiner and Pfeifer, 2009).

Patch size distribution is shown in Figure 18. Patch size bins use an exponential scale with a base of two. The minimum patch size is 10m. The chart illustrates the ability of each dataset to classify settlements of different patch size. The minimum patch size of each dataset is equal to its pixel size. Across the datasets under test, the minimum bin size for GHSL and HRSL (38m, and 30m pixels) is 2^4 (1600m^2), and the minimum bin size for GUF (12m pixels) is 2^1 (up to 200m^2). Across a common bin size of 2^4, 36% of the patches classified by GHSL fall within this bin, 68% of HRSL, and 51% of GUF (within or below the bin size). This indicates HRSL and GUF detect a greater dataset proportion of fragmented settlement patches than GHSL. The MOD500 dataset (500m pixels) frequencies only feature for the larger patch sizes. All datasets but GL30 show a decay in frequency with increased patch size. This conforms to the theory that city population size relative to its size rank follows a log-linear decay trend (Small and Sousa, 2016; Lotka and Thorndike, 1941). OSM has not been plotted as the vector building footprint data does not support patch size analysis.

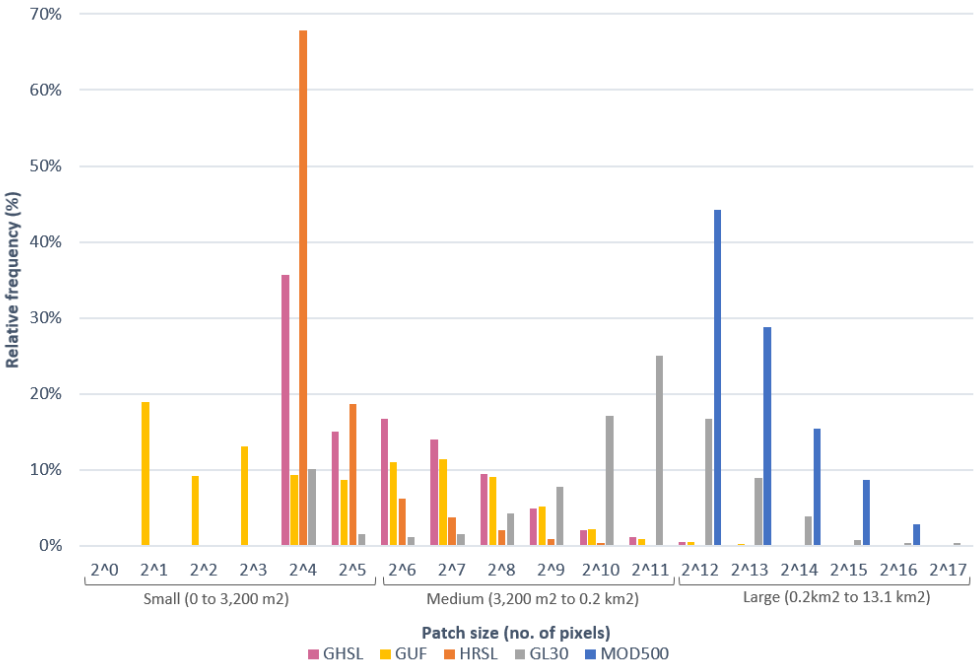


Figure 18: Patch size distribution for human settlement datasets. Normalised to total number of patches per dataset.

Error statistics PPV, TPR, and F are presented by patch size in Figure 19, Figure 20, and Figure 21. MOD500 has been excluded as patch sizes do not corroborate with high spatial detail datasets under test (Figure 18).

The results show PPV increases with patch size for all datasets under test. Gains in PPV relative to GL30 occur across medium patch sizes due to the MR product failing to detect small patch size settlements. Similar increases by patch size are observed in TPR and F statistics for GUF and HRSL. However, GHSL shows an initial decrease in TRP and F relative to GUF and HRSL for small patch sizes, before showing increases over medium to large patch sizes. This is the result of a high rate of single pixel classifications from MR imagery (Figure 18) over medium sized towns, with good inter-map agreement from GUF and HRSL. These results correspond with greater inter-map agreement with increases in settlement size, due to the complexity of classifying fragmented settlements.

The relative capabilities of each dataset under test is analysed to determine fitness-for-use for the detection of fragmented settlements. Results largely reflect the NTL spatial zone analysis.

As determined in 6.1.2 and 6.2.1, GHSL demonstrates high PPV, but low TPR relative to high spatial detail pairs. GHSL PPV is higher than GUF and HRSL for small and medium settlements, however TRP is low across these ranges. This indicates GHSL has good correctness but poor recall relative to high spatial detail datasets for all but large patch sizes. GUF and HRSL results are comparable, with low PPV and TPR for small patch sizes, and steady increases in PPV and TPR over medium and large patch size bins. Qualify (F) indicates agreement of GUF is strongest relative to HRSL for small patches, and OSM for medium patches. HRSL inter-map agreement (F) is strongest relative to OSM for small and medium patch sizes. Overall there is significant disagreement between datasets for small and medium patch sizes. Strong consensus between datasets is only reached for large patch sizes.

In this subsection, we have quantified inter-map agreement by inferred settlement density, and settlement size. Findings demonstrate poor inter-map agreement between comparison pairs for fragmented settlements. Inter-map agreement increases for comparison pairs with density and settlement size. This reflects the challenge of detecting fragmented settlements.

Results indicate fitness-for-use of each dataset for humanitarian mapping. In areas with fragmented settlements GHSL demonstrates low errors of commission, but high errors of omission. This corresponds with good relative correctness, but poor relative classification of fragmented settlements. GHSL is therefore unsuitable for detecting the humanitarian mapping target.

Fitness-for-use of GUF and HRSL is greater. GUF and HRSL show high errors of commission relative to high spatial detail comparison pairs, high total classified area (Figure 8), and a high proportion of patches in small frequency patch bin sizes (Figure 18). This demonstrates capability to positively classify areas as human settlements that are not identified in comparison products. It is not known if these errors of commission correctly represent a human settlement. Additionally, GUF and HRSL show high relative errors of omission for areas with fragmented settlements. This indicates low agreement between products. It is important to understand factors contributing to error results to further understand capabilities.

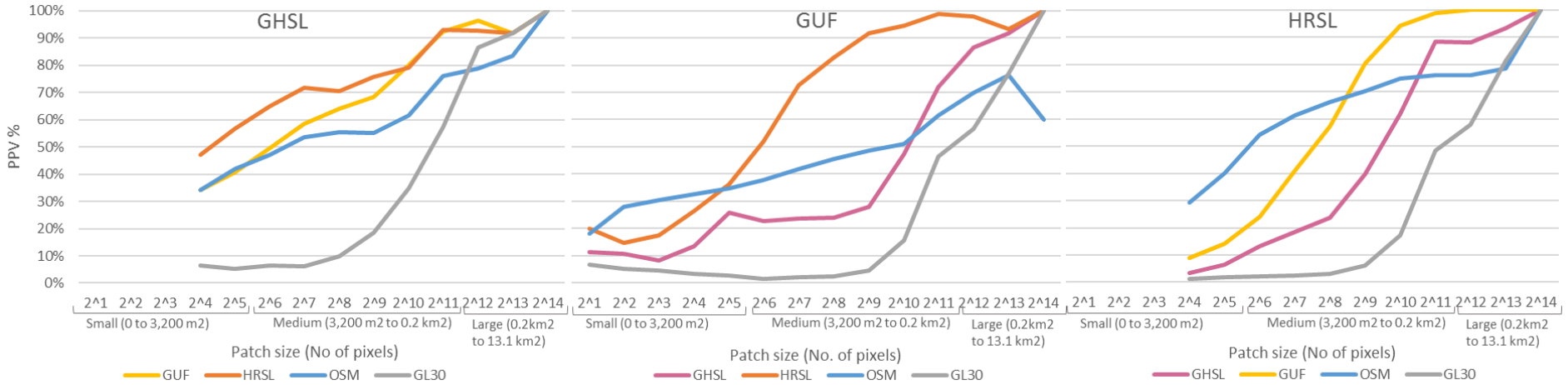


Figure 19: Change in PPV by patch size

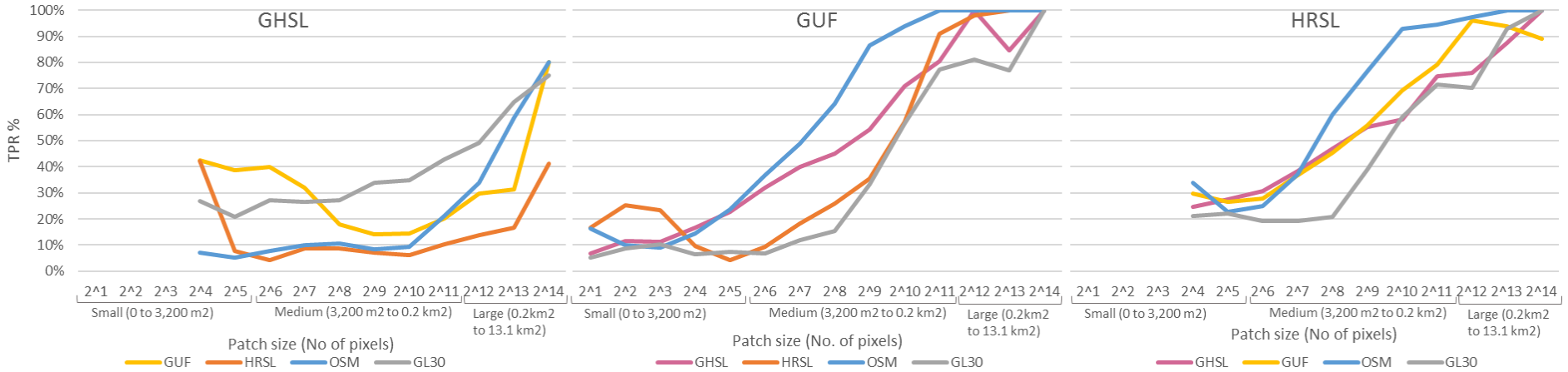


Figure 20: Change in TPR by patch size

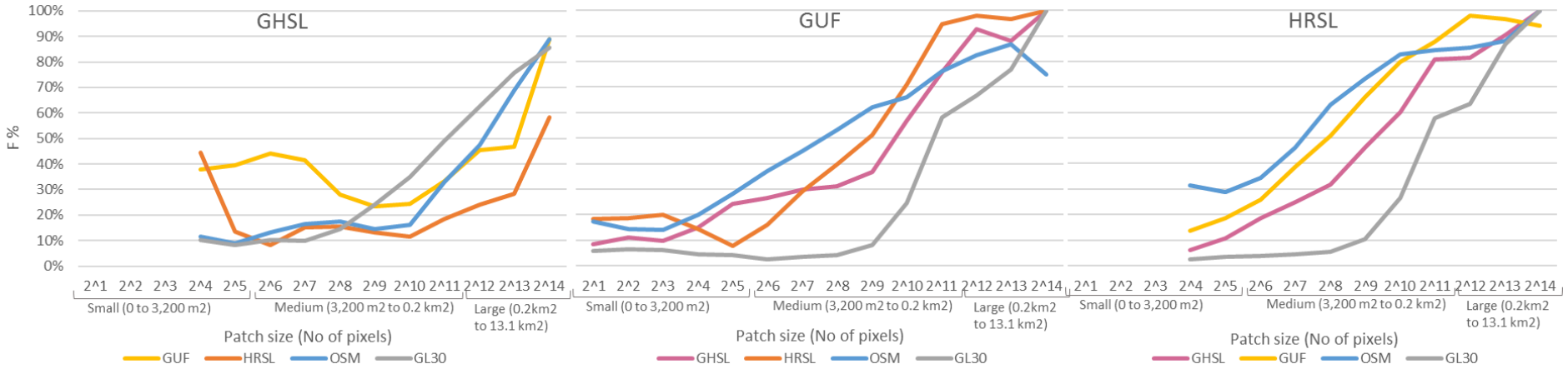


Figure 21: Change in quality (F) by patch size

6.3 RQ3: What factors may contribute to inter-map disagreement for these datasets?

To address RQ3 potential factors contributing to inter-map disagreement between high spatial detail datasets are assessed (FP and FN results). This includes the source of disagreement, and potential causes of classification differences.

This analysis is completed for GUF and HRSL relative to high spatial detail pairs (comparison matrix Table 12). GHSL is excluded as it was identified in RQ1 and RQ2 as not fit for use in a humanitarian mapping context.

Table 12: Dataset comparison matrix for RQ3

		Comparison dataset		
		GUF	**HRSL**	**OSM**
Dataset under test	GUF	-	GUF:HRSL FP and FN errors	GUF:OSM FP and FN errors
	HRSL	HRSL:GUF FP and FN errors	-	HRSL:OSM FP and FN errors

6.3.1 Inter-map disagreement: error source

Inter-map disagreement could be attributed to error in either the dataset under test or its comparison pair (Table 11). The proportion of errors by source dataset is quantified through manual assessment of 400 samples for each inter-map comparison error type (Table 13). The sample size represents 95% confidence level, and 5% margin of error.

Table 13: Comparison pair error sample size (95% confidence level, and 5% margin of error)

Inter-map comparison	Error	Pixel count	Minimum sample size
GUF:HRSL	FP	233,583	385
	FN	1,926,370	384
GUF:OSM	FP	543,003	384
	FN	612,894	384
HRSL:GUF	FP	157,139	384
	FN	123,336	383
HRSL:OSM	FP	133,886	384
	FN	196,341	384

Results are presented in N pattern reflects that of GUF.

Figure 22. The GUF:HRSL sample shows that the majority of FP results (68% true, no building) and FN results (86% true, visible building) are attributed to misclassification by GUF. HRSL:GUF FP results (17% true) attribute the majority of error to GUF, meaning HRSL correctly classifies many areas missed by GUF. However, HRSL:GUF FN results (58% true) show that GUF classifies sites missed by HRSL.

The GUF:OSM sample shows that 41% of FP errors are true, and 97% of FN errors are true. Most FP differences are attributed to errors in OSM. However, most FN errors are attributed to omission by GUF. This corresponds to VGI digitisers missing settlements, but rarely classifying buildings that are not present. HRSL:OSM sample results show that 51% of FPs are true, and 82% of FNs are true. HRSL and OSM each classify settlements not recalled by the other. The FN pattern reflects that of GUF.

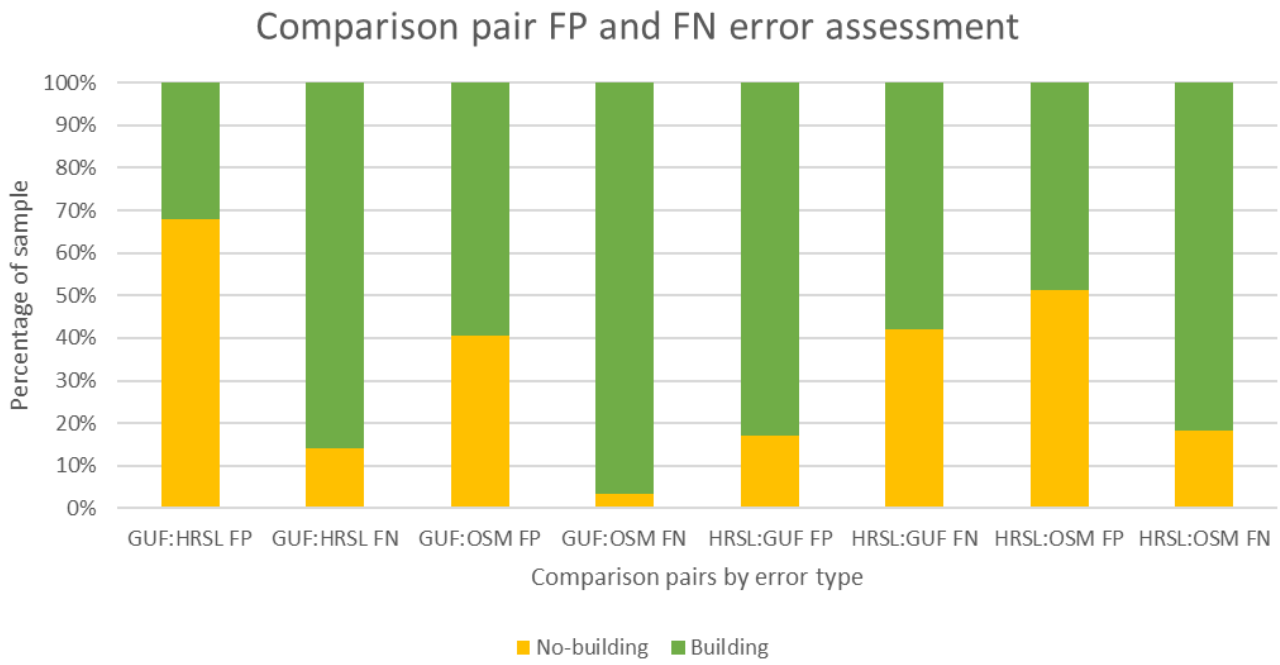


Figure 22: Sample of HRSL and GUF FP and FN classifications

6.3.2 Inter-map disagreement: error causes

Potential causes of classification differences between products provides insight into relative capabilities. Error result spatial distribution is investigated in relation to proximity to human settlements, and zones of bad imagery (cloud cover or low resolution).

Inter-map differences for a representative area in Northern Leyte around Tacloban city are shown in Figure 25 to Figure 28. This area was chosen for its mixed settlement landscape. Visual analysis of error spatial patterns is described in Table 14. FP results are frequently

clustered in and around settlements, potentially due to technical issues or semantic differences (described in 5.4.2). To understand the proportion of errors by proximity to human settlements we count errors within a 30m TP result buffer zone.

Table 14: Map errors descriptions

Inter-map comparison	Error	Error pixel count	Proportion of total errors	Error pattern description from visual assessment
GUF:HRSL (Figure 25)	FP	233,583	11%	Clustered around settlements
	FN	1,926,370	89%	Clustered and linear
GUF:OSM (Figure 26)	FP	543,003	47%	Clustered around settlements
	FN	612,894	53%	Dispersed and linear
HRSL:GUF (Figure 27)	FP	157,139	56%	Clustered around settlements
	FN	123,336	44%	Clustered and linear
HRSL:OSM (Figure 28)	FP	133,886	41%	Large clusters and dispersed
	FN	196,341	59%	Clustered and linear

Results are presented in Figure 23. GUF:HRSL FPs and HRSL:GUF FNs represent areas positively classified by GUF, but not detected by HRSL. Of this error type, more than 90% are within 30m of settlements. GUF:HRSL FNs and HRSL:GUF FPs represent areas positively classified by HRSL, but not GUF. Of this error type, more than 60% are not adjacent to settlements. HRSL under represents the extent of human settlements relative to GUF (missed adjacent pixels), and that the majority of GUF errors of omission are for fragmented settlements.

GUF:OSM FPs and HRSL:OSM FNs primarily occur adjacent to human settlements. This indicates that relative to OSM GUF over represents and HRSL underrepresents settlement extent. In contrast, GUF:OSM FNs, and HRSL:OSM FPs primarily occur away from settlements. This tells us GUF demonstrates higher errors of omission, and HRSL higher errors of commission relative to OSM in areas with fragmented settlements.

Visual analysis of GUF and HRSL relative to OSM (Figure 26 and Figure 28) shows a large cluster of FP results. This cluster (town of Burauen) falls within a zone of bad imagery and may represent an OSM error of omission (shown in Figure 29). Error rates for OSM good and bad image quality zones have been calculated to understand the effect of image quality on error distribution Results presented in Figure 29 indicate higher rates of error for zones of good imagery. This result is not logical and indicates additional variables impact the calculated error rates, such as settlement density, or land cover type.

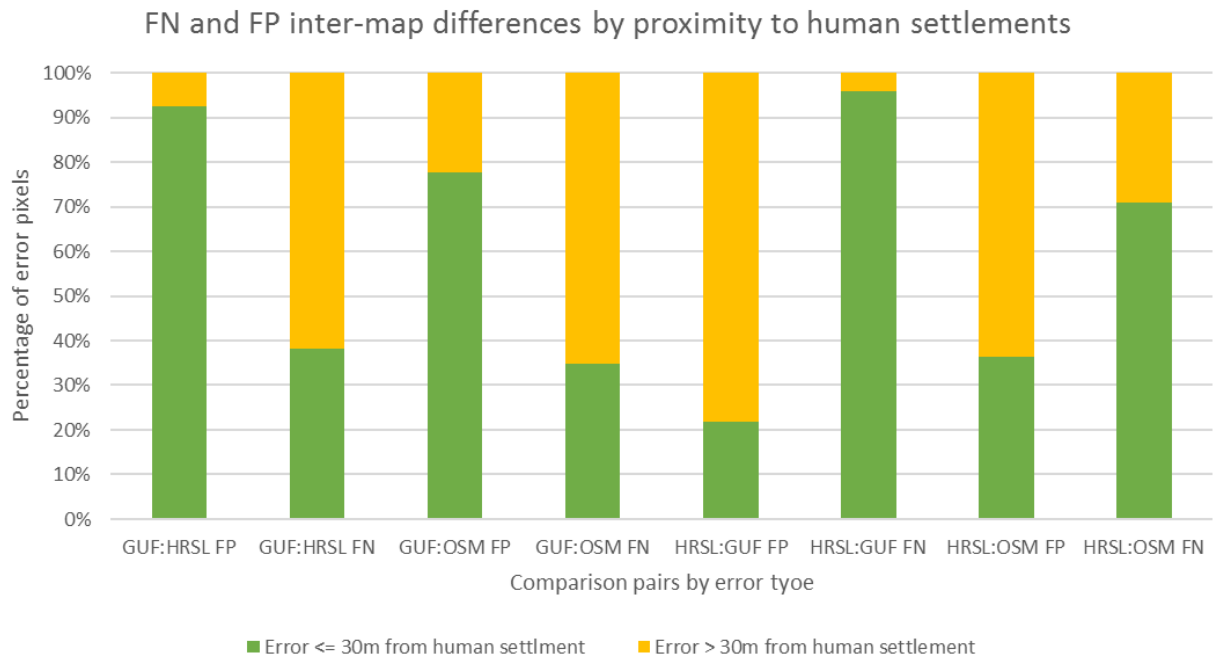


Figure 23: Inter-map disagreement by proximity to human settlements

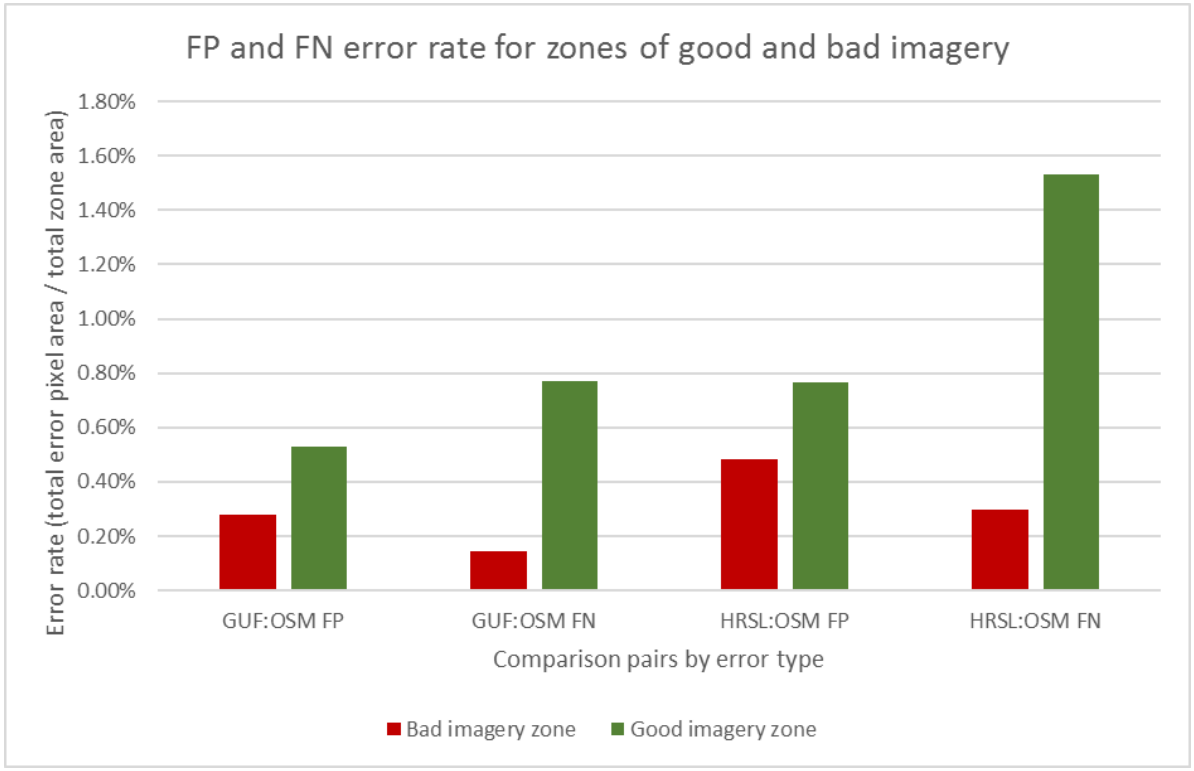


Figure 24: FP and FN error rates for zones of good and bad imagery

Findings demonstrate that inter-map disagreement of GUF and HRSL is primarily attributed to errors in GUF. HRSL correctly classifies pixels missed by GUF (GUF:HRSL FNs 86% true) the majority of which (60%) are for fragmented settlements. GUF successfully classifies settlements missed by HRSL (HRSL:GUF FN 58% true) more than 90% of which occur close to human settlements. HRSL therefore underrepresents the extent of settlements relative to GUF, and GUF fails to detect fragmented settlements relative to HRSL. These results reflect the findings of Muck et al. (2017) who reported that GUF under classifies in rural areas (Mück, Klotz and Taubenböck, 2017)

The samples of FP errors relative to OSM are more evenly attributed to both OSM and the dataset under test (GUF:OSM FP 59% true, HRLS:OSM FP 51% true). GUF FNs and HRSL FPs occur away from settlements, and GUF FPs, and HRSL FNs adjacent to settlements. This supports the finding that HRSL underrepresents the extent of settlements (adjacent pixels), and GUF underrepresents fragmented settlements.

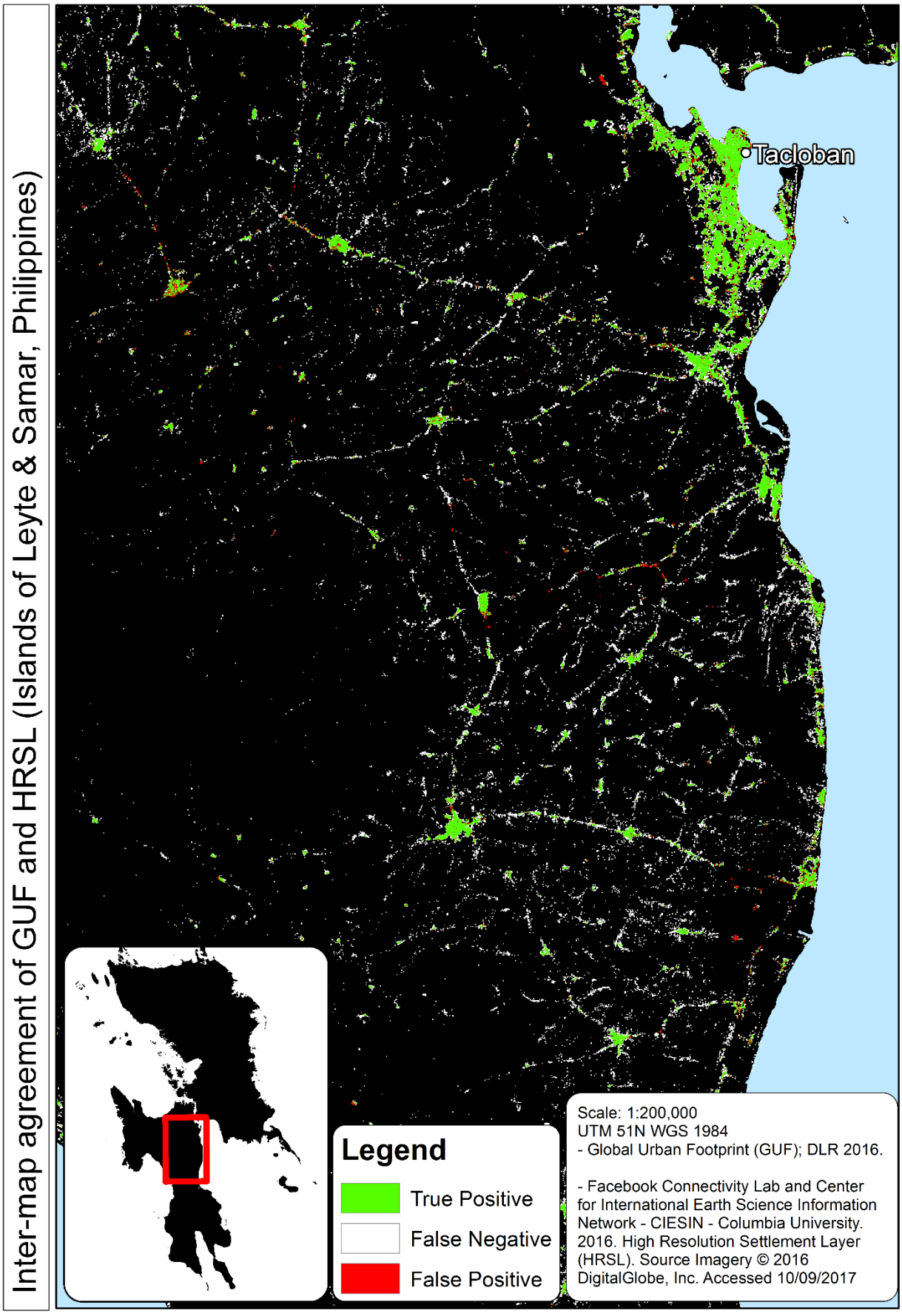


Figure 25: Inter-map agreement GUF:HRSL

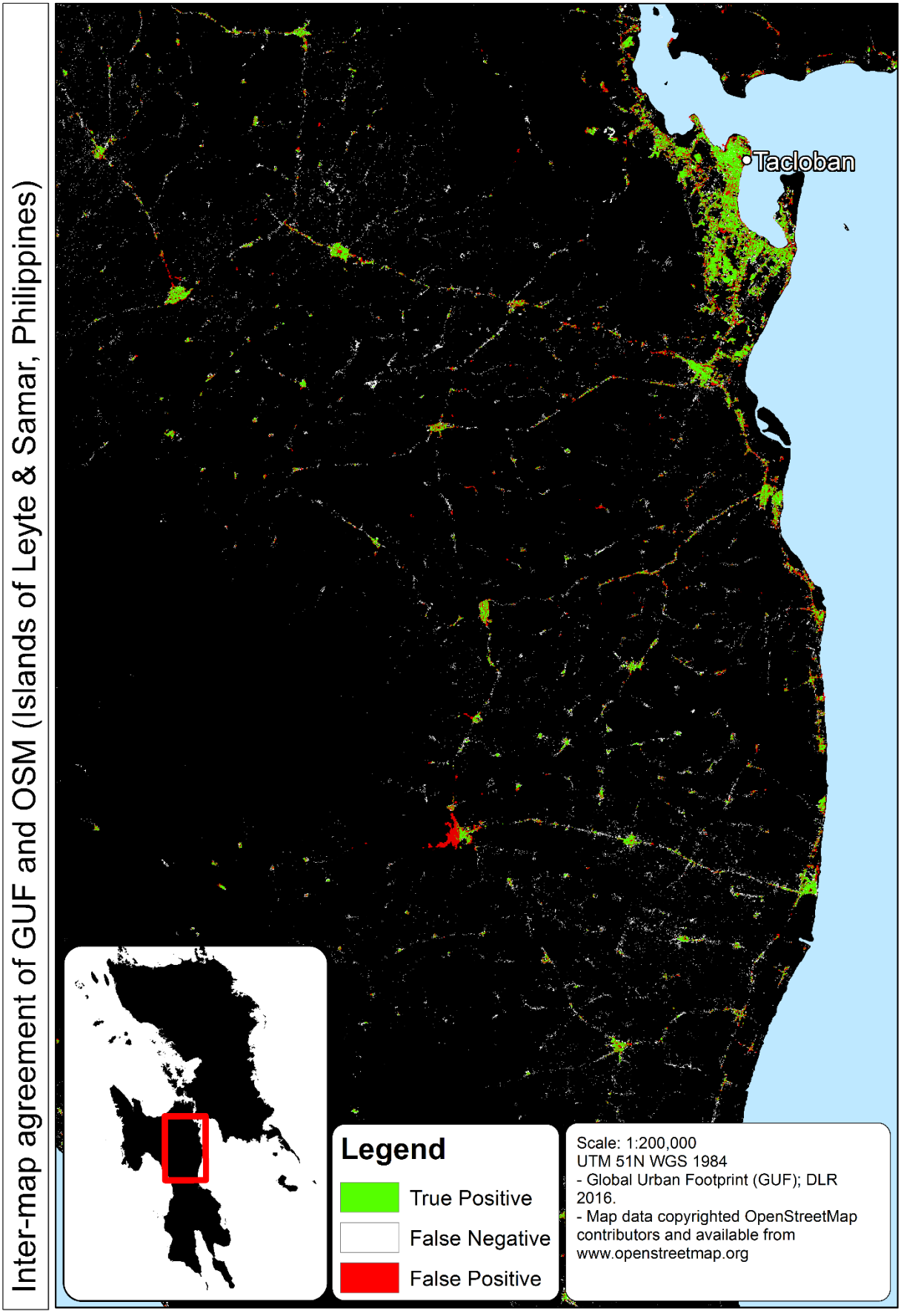


Figure 26: Inter-map agreement GUF:OSM

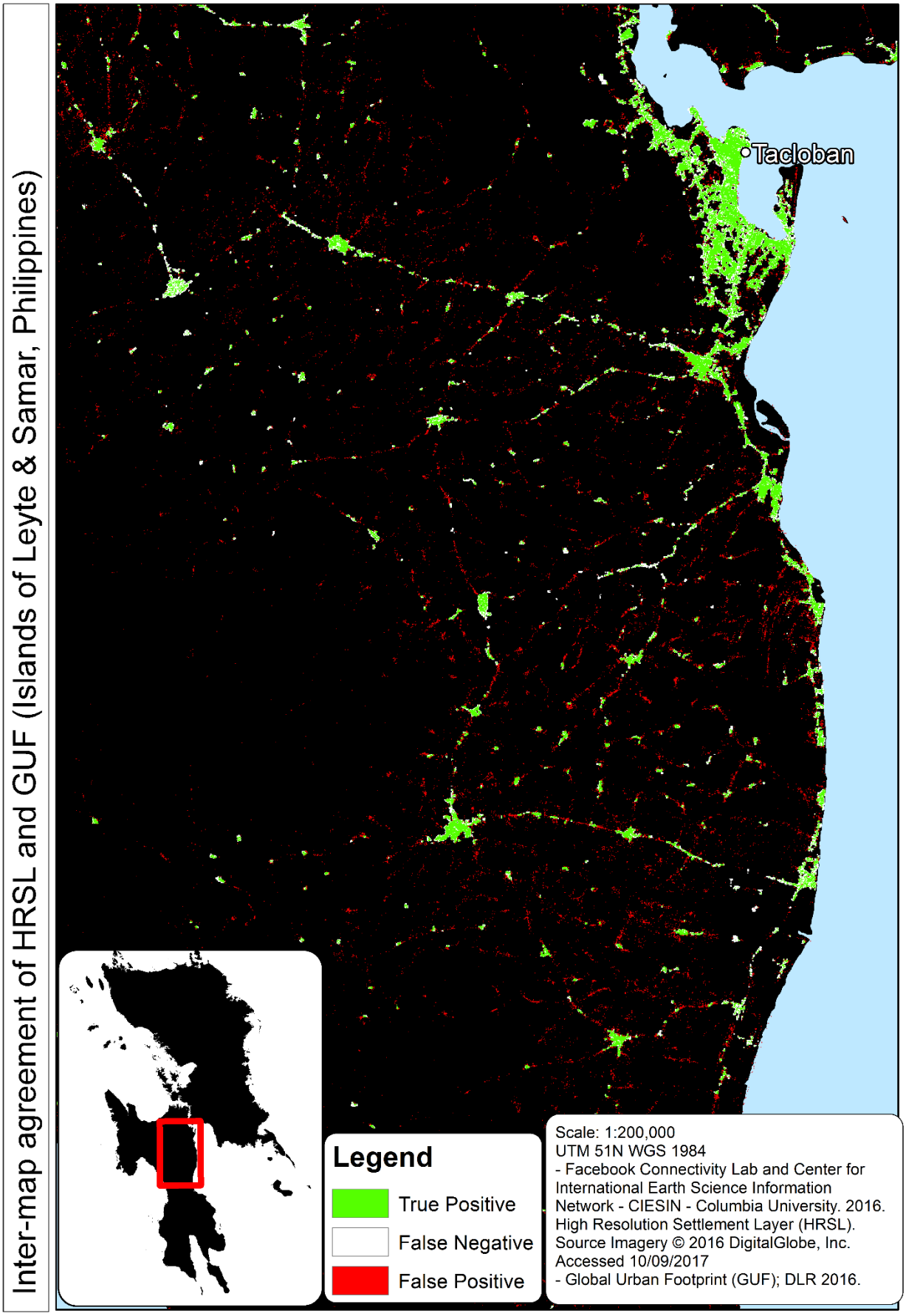


Figure 27: Inter-map agreement HRSL:GUF

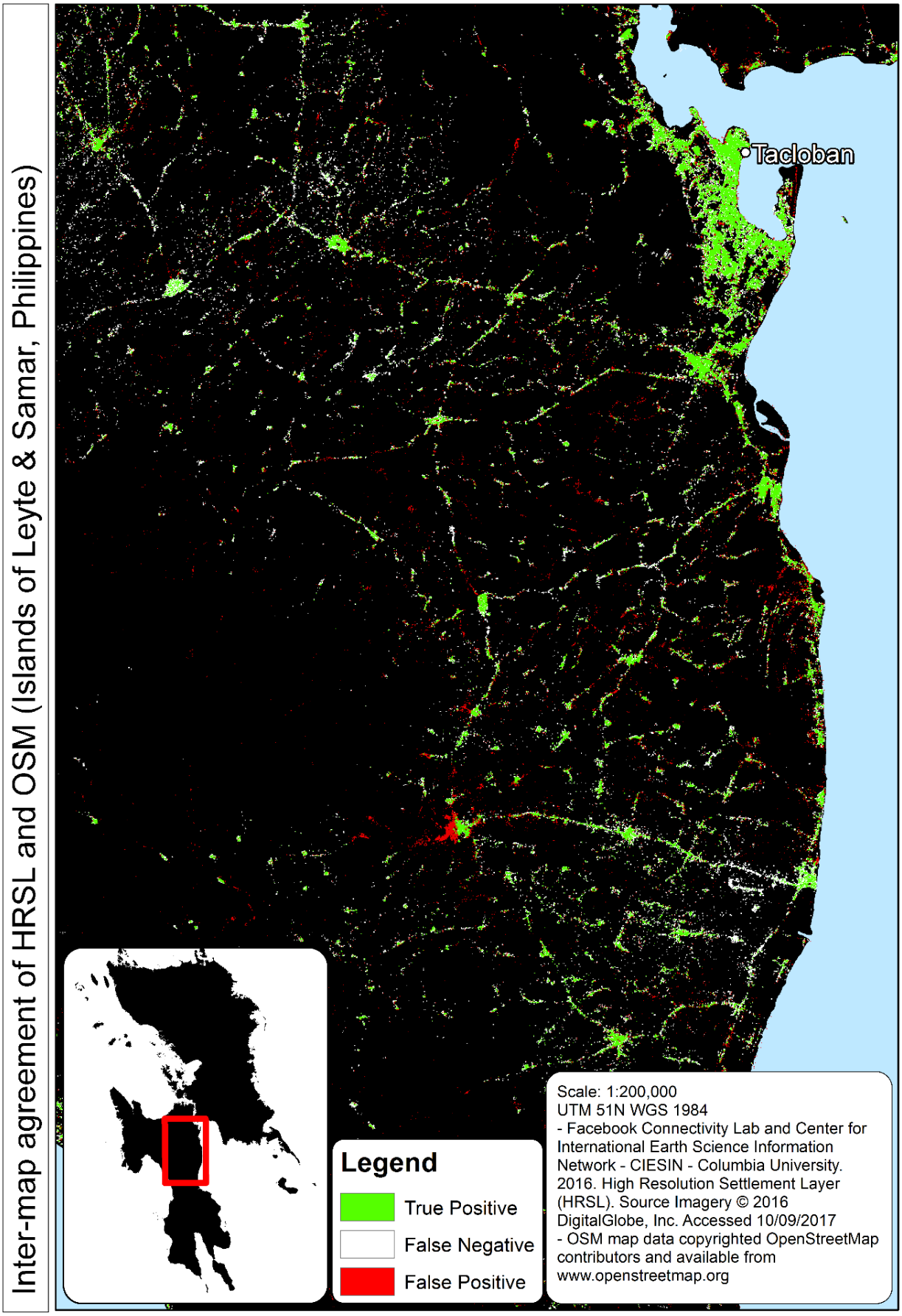


Figure 28: Inter-map agreement HRSL:OSM

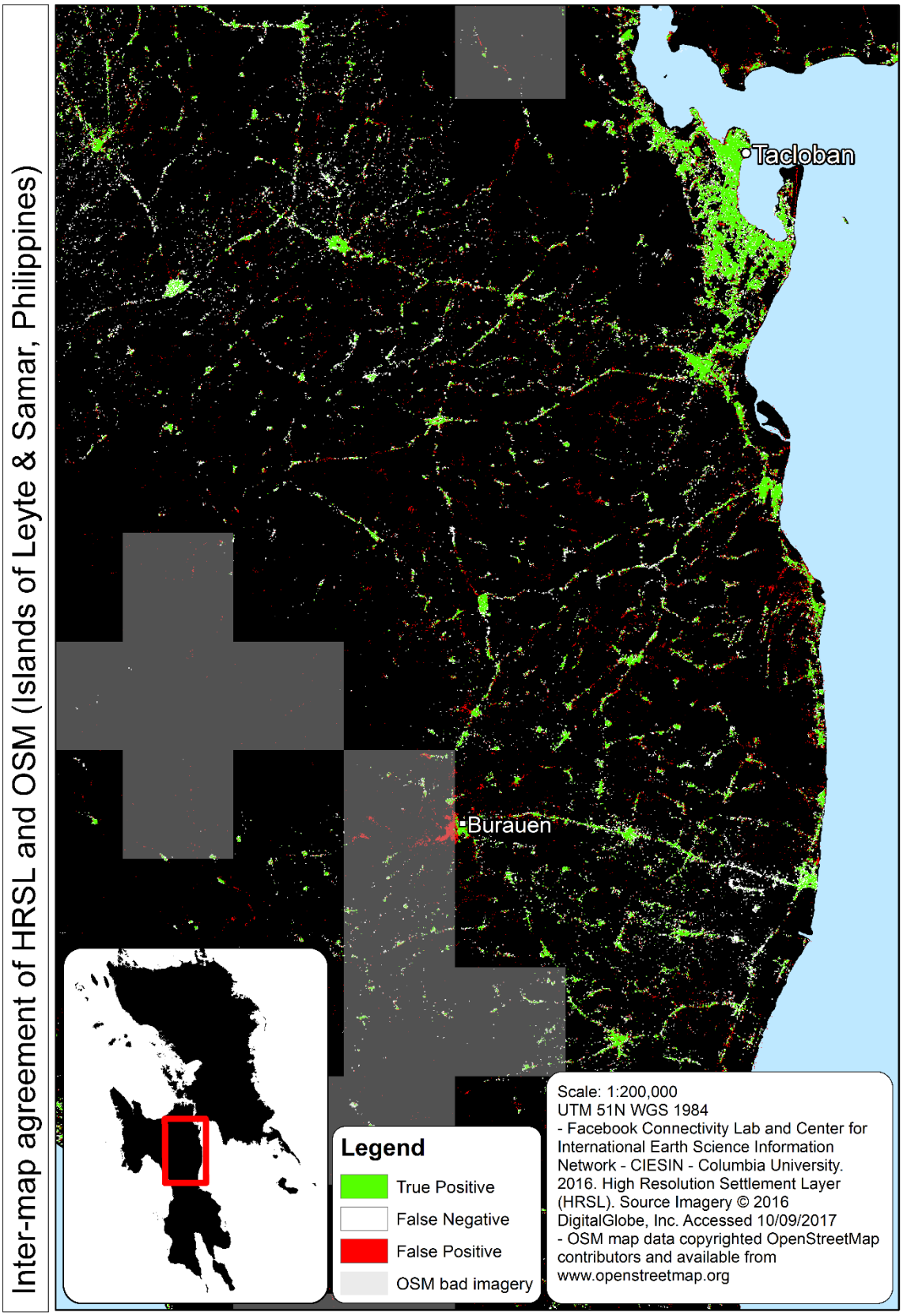


Figure 29: Inter-map agreement HRSL:OSM, with OSM bad imagery mask

6.4 Limitations

The applied methodology is limited by the availability of authoritative reference data. In its absence, comparisons between human settlement datasets have been conducted to understand relative capabilities. Results are therefore less definitive for actioning by Missing Maps than those that would be produced by calculating absolute accuracies (Lillesand and Kiefer, 2000).

Additionally, the compared datasets represent different time periods. There is maximum temporal shift for the datasets under test of 6 years (years in Table 3). The influence of the time gaps is unknown. For example, Typhoon Haiyan passed through Tacloban in 2013 (Carlowicz, 2013) between the imagery dates of GUF and GHSL. Differences between these products could be attributed to settlement destruction from the Typhoon.

The determination of landscape character using VIIRS NTL luminosity data (section 6.2.1) is limited in its thresholding approach. Thresholding introduces bias as the choice of threshold affects the results outcome (Henderson et al., 2003). The methodology could be improved by repeating the analysis for multiple threshold values to help define optimal values.

Finally, the inter-map error small sample size in section 6.3.1 is indicative only and not sufficient for determination of absolute accuracies (Foody, 2009a). The sample results can guide understanding of dataset strengths and weaknesses only.

7 Conclusion

This study aimed to assess the fitness-for-use of high spatial detail human settlement datasets in a humanitarian mapping context. The integration of an automated detection approach within the Missing Maps VGI workflow could optimise volunteer mapping for disaster relief (Pete Masters and Benjamin Herfort, 2016).

Accuracy assessments in Global South settings can be challenging due to lack of appropriate reference data (Foody, 2009b). In the absence of reference data, we introduced a comparison framework quantifying inter-map agreement through comparisons of human settlement products. The framework has been applied to three high spatial detail datasets (GHSL, GUF, HRSL) for a Global South study site to allow relative classification capabilities to be understood (Klotz et al., 2016).

It has been shown that there is significant disagreement between all human settlement datasets under test for the study site. Products show a three-fold variance in total area classified and low relative quality ($F < 60\%$) for all comparison pairs. Agreement between datasets is particularly poor in areas with fragmented rural settlements, but increases with settlement density and size. Inter-map disagreement demonstrates variation in classification capability particularly for complex fragmented settlements, and variation in the definition of human settlements.

Inter-map comparisons provide insight into relative capabilities of each product. The findings indicated that GHSL is not fit-for use for humanitarian mapping due to low relative detection capabilities for fragmented settlements (high correctness but low completeness). GUF and HRSL have shown greater capability in detecting fragmented settlements with a higher proportion of classifications for smaller settlement sizes, and high errors of commission in areas of fragmented settlement character. This corresponds with the capability to detect dispersed rural settlements missed by comparison datasets.

Analysis of inter-map disagreement (FP and FN error sample analysis) has demonstrated that differences in classifications between GUF and HRSL are primarily attributed to errors in GUF; however, each dataset correctly classifies sites missed by the other. Correct settlement classifications in GUF not recalled by HRSL largely occur (>90%) within 30m of other classified pixels. Correct classifications by HRSL not recalled by GUF frequently occur (62%) away from other classified pixels. This corresponds with HRSL underrepresenting the spatial extent of settlements relative to GUF (missed adjacent pixels), and GUF under

detecting fragmented settlements relative to HRSL. Results for GUF reflect findings of Muck et al. (2017) who reported that GUF under classifies in rural areas (Mück, Klotz and Taubenböck, 2017).

HRSL demonstrates greatest product fitness-for-use in the context of humanitarian mapping due to its greater ability in detecting fragmented settlements. However, results show that HRSL and all other high spatial detail datasets suffer errors of omission and commission. This means no one dataset offers complete settlement mapping capabilities. By understanding the relative classification abilities of existing datasets Missing Maps can focus future collaboration and research efforts. The following next steps are suggested;

Firstly, it is recommended that the automated settlement detection capabilities of GUF and HRSL be incorporated in the Missing Maps workflow to compliment VGI data collection. The workflow requires volunteers to manually scan all imagery tiles for human settlements (MapSwipe). This could be optimised by utilising GUF and HRSL to prioritise tiles for volunteer scanning and digitisation. Areas identified by GUF and HRSL as containing settlements would go to volunteers as a priority for review. This could generate useful data more quickly during a crisis.

Secondly, HRSL has been identified as fit-for use in humanitarian mapping. HRSL was developed using a CNN and HR imagery. It is recommended initiatives using similar technologies be pursued by Missing Maps. For example, performance of the HRSL classifier (or similar alternative) could be improved through fine-tuning with local training data. HRSL was trained with 8,000 images from one country. It has been demonstrated that existing classifiers can be optimised for remote sensing through fine-tuning with training imagery (Nogueira, Penatti and dos Santos, 2017). Developments in machine learning have been limited by availability of high quality training data (Alexander Wissner-Gross, 2016) and Missing Maps is in a unique position holding ample VGI data that is applicable as CNN training data to develop a CNN for humanitarian mapping.

8 References

Airbus, 2017. *SPOT 6/7 Satellite Imagery : Airbus Defence and Space*. [online] Available at: <http://www.intelligence-airbusds.com/en/147-spot-6-7-satellite-imagery> [Accessed 11 Sep. 2017].

Albuquerque, J., Herfort, B. and Eckle, M., 2016. The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. *Remote Sensing*, 8(10), p.859.

Alexander Wissner-Gross, 2016. *Datasets Over Algorithms*. [online] Available at: <https://www.edge.org/response-detail/26587> [Accessed 27 Sep. 2017].

Allouche, O., Tsoar, A. and Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), pp.1223–1232.

Arino, O., Kalogirou, V. and Ramos Perez, J., 2011. *GlobCover 2009 Products Description and Validation Report*. European Space Agency.

Bakos, G. and Ballatore, A., 2017. Estimating Population Distribution with Landsat Imagery and Volunteered Geographic Information.

Balk, D., Pozzi, F., Yetman, G., Deichmann, U. and Nelson, A., 2005. The distribution of people and the dimension of place: methodologies to improve the global estimation of urban extents. In: *International Society for Photogrammetry and Remote Sensing, Proceedings of the Urban Remote Sensing Conference*. [online] pp.14–16. Available at: <http://sedac.ciesin.columbia.edu/downloads/docs/grump-v1/ur_paper_webdraft1.pdf> [Accessed 11 Sep. 2017].

Beazley, D. and Jones, B.K., 2013. *Python Cookbook*. 3 edition ed. Sebastopol, CA: O'Reilly Media.

Brenner, N. and Schmid, C., 2014. The 'urban age' in question. *International Journal of Urban and Regional Research*, 38(3), pp.731–755.

Brovelli, M., Molinari, M., Hussein, E., Chen, J. and Li, R., 2015. The First Comprehensive Accuracy Assessment of GlobeLand30 at a National Level: Methodology and Results. *Remote Sensing*, 7(4), pp.4191–4212.

Carlowicz, M., 2013. Super Typhoon Haiyan Surges Across the Philippines : Natural Hazards. Available at: <https://earthobservatory.nasa.gov/NaturalHazards/view.php?id=82348> [Accessed 28 Sep. 2017].

Center for International Earth Science, 2017. *High Resolution Settlement Layer*. [online] Available at: <https://ciesin.columbia.edu/data/hrsl/> [Accessed 22 Jul. 2017].

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X. and Mills, J., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, pp.7–27.

Chen, J. and Zipf, A., 2017. DeepVGI: Deep Learning with Volunteered Geographic Information. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. [online] International World Wide Web Conferences Steering Committee, pp.771–772. Available at: <http://dl.acm.org/citation.cfm?id=3054250> [Accessed 31 Jul. 2017].

Chen, X., Cao, X., Liao, A., Chen, L., Peng, S., Lu, M., Chen, J., Zhang, W., Zhang, H., Han, G., Wu, H. and Li, R., 2016. Global mapping of artificial surfaces at 30-m resolution. *Science China Earth Sciences*, 59(12), pp.2295–2306.

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

Congalton, R.G. and Green, K., 2008. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Second Edition*. 2 edition ed. Boca Raton: CRC Press.

Deuskar, C., 2015. What does 'urban' mean? [Blog] *Sustainable Cities*. Available at: <http://blogs.worldbank.org/sustainablecities/what-does-urban-mean> [Accessed 14 Jul. 2017].

DigitalGlobe, 2016. *WorldView-3 Data Sheet*.

DLR-DFD Oberpfaffenhofen, 2016. *GUF Product Specifications (GUF_DLR_v01):* Available at: <http://www.dlr.de/eoc/en/Portaldata/60/Resources/dokumente/guf/GUF_Product_Specifications_GUF_DLR_v01.pdf> [Accessed 9 Feb. 2017].

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C. and Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7), pp.849–857.

Doll, C.N.H. and Pachauri, S., 2010. Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy*, 38(10), pp.5661–5670.

Doxsey-Whitfield, E., MacManus, K., Adamo, S.B., Pistolesi, L., Squires, J., Borkovska, O. and Baptista, S.R., 2015. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. *Papers in Applied Geography*, 1(3), pp.226–234.

Elvidge, C.D., Baugh, K.E., Zhizhin, M. and Hsu, F.-C., 2013. Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network*, 35(0), p.62.

Elvidge, C.D., Imhoff, M.L., Baugh, K.E., Hobson, V.R., Nelson, I., Safran, J., Dietz, J.B. and Tuttle, B.T., 2001. Night-time lights of the world: 1994-1995. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56, pp.81–99.

Elvidge, C.D., Tuttle, B.T., Sutton, P.S., Baugh, K.E., Howard, A.T., Milesi, C., Bhaduri, B.L. and Nemani, R., 2007. Global Distribution and Density of Constructed Impervious Surfaces. *Sensors (Basel, Switzerland)*, 7(9), pp.1962–1979.

Esch, T., Heldens, W., Hirne, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S. and Strano, E., 2017. Breaking new ground in mapping human settlements from space -The

Global Urban Footprint-. *arXiv:1706.04862 [physics]*. [online] Available at: <http://arxiv.org/abs/1706.04862>.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenbock, H., Muller, A. and 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(6), pp.1617–1621.

Esch, T., Schenk, A., Ullmann, T., Thiel, M., Roth, A. and Dech, S., 2011. Characterization of Land Cover Types in TerraSAR-X Images by Combined Analysis of Speckle Statistics and Intensity Information. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6), pp.1911–1925.

European Space Agency, 2017. *Sentinel Overview*. [online] Available at: <https://sentinel.esa.int/web/sentinel/missions> [Accessed 11 Sep. 2017].

Fan, H., Yang, A. and Zipf, A., 2016. The intrinsic quality assessment of building footprints data on OpenStreetMap in Baden-Württemberg. [online] Available at: <https://www.researchgate.net/profile/Hongchao_Fan/publication/311249851_Intrinsic_OSM_data_quality_assessment_The_intrinsic_quality_assessment_of_building_footprints_data_on_OpenStreetMap_in_Baden-_Wurttemberg/links/58401ccc08ae8e63e61f3081.pdf> [Accessed 24 Aug. 2017].

Ferri, S., Syrris, V., Florczyk, A., Scavazzon, M., Halkia, M. and Pesaresi, M., 2014. A new map of the European settlements by automatic classification of 2.5m resolution SPOT data. [online] IEEE, pp.1160–1163. Available at: <http://ieeexplore.ieee.org/document/6946636/> [Accessed 13 Aug. 2017].

Foody, G., 2009a. Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, 30.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote sensing of environment*, 80(1), pp.185–201.

Foody, G.M., 2006. What is the difference between two maps? A remote senser's view. *Journal of Geographical Systems*, 8(2), pp.119–130.

Foody, G.M., 2009b. The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, 30(12), pp.3275–3281.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E. and Mills, J., 2016. Development of new open and free multi-temporal global population grids at 250 m resolution - EU Science Hub - European Commission. *EU Science Hub*. [online] Available at: <https://ec.europa.eu/jrc/en/publication/development-new-open-and-free-multi-temporal-global-population-grids-250-m-resolution> [Accessed 13 Aug. 2017].

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., van der Velde, M., Boettcher, H., Havlík, P. and Achard, F., 2011. Highlighting continued uncertainty in global land cover maps for the user community. *Environmental Research Letters*, 6(4), p.044005.

GeoData Institute, 2017. *WorldPop*. [online] Available at: <http://www.worldpop.org.uk/> [Accessed 20 Apr. 2017].

Giraud, P., 2017. *OSM Tasking Manager*. [online] Available at: <http://tasks.hotosm.org/> [Accessed 31 Jul. 2017].

Giri, C., Pengra, B., Long, J. and Loveland, T.R., 2013. Next generation of global land cover characterization, mapping, and monitoring. *International Journal of Applied Earth Observation and Geoinformation*, 25, pp.30–37.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221.

Goodchild, M.F. and Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), pp.231–241.

Graham, M., Hogan, B., Straumann, R.K. and Medhat, A., 2014. Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers*, 104(4), pp.746–764.

Gros, A. and Tiecke, T., 2016. Connecting the world with better maps. *Facebook Code*. Available at: <https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/> [Accessed 15 Jul. 2017].

Haklay, M., Antoniou, V., Basiouka, S., Soden, R. and Mooney, P., 2014. *Crowdsourced geographic information use in government*. [online] World Bank Publications. Available at: <http://books.google.com/books?hl=en&lr=&id=boYLDAAAQBAJ&oi=fnd&pg=PA9&dq=%22Peter+mooney,+Department+of+Computer%22+%22202-473-1000%3B+Internet:%22+%22from+such+infringement+rests+solely+with%22+%22be+considered+an+official+World+Bank+translation.%22+%22upon+or+waiver+of+the+privileges+and%22+&ots=ObR4Pdey-2&sig=ouix5IUt9Rj3psVs6YmuiCN-V98> [Accessed 24 Aug. 2017].

Haklay, M. (Muki), Basiouka, S., Antoniou, V. and Ather, A., 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp.315–322.

Henderson, M., Yeh, E.T., Gong, P., Elvidge, C. and Baugh, K., 2003. Validation of urban boundaries derived from global night-time satellite imagery. *International Journal of Remote Sensing*, 24(3), pp.595–609.

Herfort, B., Eckle, M. and Albuquerque, J.P. de, 2016. Being specific about geographic information crowdsourcing: a typology and analysis of the Missing Maps project in South Kivu. In: *Proceedings of the ISCRAM 2016 Conference–Rio de Janeiro, Brazil, May 2016*. [online] Available at: <http://wrap.warwick.ac.uk/78703> [Accessed 30 Jun. 2017].

Herfort, B., Reinmuth, M. and Zipf, A., 2017. Towards evaluating crowdsourced image classification on mobile devices to generate geographic information about human settlements. *20th AGILE conference 2017*.

Hoersch, B. and Amans, V., 2015. *Copernicus Space Component Data Access Portfolio: Data Warehouse 2014 - 2020*. European Space Agency.

Humanitarian Coalition, 2017. *What is a Humanitarian Emergency?* [online] Available at: <http://humanitariancoalition.ca/sites/default/files/factsheet/fact_sheet-_what_is_a_humanitarian_emergency.pdf> [Accessed 30 Jun. 2017].

Jeni, L.A., Cohn, J.F. and De La Torre, F., 2013. Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. [online] IEEE, pp.245–251. Available at: <http://ieeexplore.ieee.org/document/6681438/> [Accessed 30 Aug. 2017].

Jensen, J.R., 2013. *Remote Sensing of the Environment: an Earth Resource Perspective 2nd Edition*. 2 edition ed. India: PEARSON.

Joint Research Centre, 2016. *GHSL Data Packages Instructions for data access. V1.0*. European Comission.

Jokar Arsanjani, J., See, L. and Tayyebi, A., 2016. Assessing the suitability of GlobeLand30 for mapping land cover in Germany. *International Journal of Digital Earth*, 9(9), pp.873–891.

Jokar Arsanjani, J., Tayyebi, A. and Vaz, E., 2016. GlobeLand30 as an alternative fine-scale global land cover map: Challenges, possibilities, and implications for developing countries. *Habitat International*, 55, pp.25–31.

Kemper, T., Pesaresi, M., Siragusa, A. and Melchiorri, M., 2016. *Atlas of the Human Planet 2016*. Luxembourg: Publications Office of the European Union.

Klotz, M., Kemper, T., Gei?, C., Esch, T. and Taubenb?ck, H., 2016. How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe. *Remote Sensing of Environment*, 178, pp.191–212.

Kunce, D., 2016. Facebook Collaboration. Available at: <http://www.missingmaps.org/blog/2016/11/16/facebook/> [Accessed 31 Jul. 2017].

Lantz, C.A. and Nebenzahl, E., 1996. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology*, 49(4), pp.431–434.

Li, W. and Guo, Q., 2014. A New Accuracy Assessment Method for One-Class Remote Sensing Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8), pp.4621–4632.

Li, X. and Gong, P., 2016. An 'exclusion-inclusion' framework for extracting human settlements in rapidly developing regions of China from Landsat images. *Remote Sensing of Environment*, 186, pp.286–296.

Lillesand, T. and Kiefer, R.W., 2000. *Remote Sensing and Image Interpretation*. 4th Revised edition edition ed. New York: John Wiley & Sons.

L.Imhoff, M., Lawrence, W.T., Stutzer, D.C. and Elvidge, C.D., 1997. A technique for using composite DMSP/OLS 'City Lights' satellite data to map urban area. *Remote Sensing of Environment*, 61(3), pp.361–370.

Lotka, A.J. and Thorndike, E.L., 1941. The law of urban concentration. *Science*, 94(2433), p.164.

Loveland, T.R. and Dwyer, J.L., 2012. Landsat: Building a strong future. *Remote Sensing of Environment*, 122, pp.22–29.

Lu, D., Tian, H., Zhou, G. and Ge, H., 2008. Regional mapping of human settlements in southeastern China with multisensor remotely sensed data. *Remote Sensing of Environment*, 112(9), pp.3668–3679.

Missing Maps, 2017. *Missing Maps*. [online] Available at: <http://www.missingmaps.org/about/> [Accessed 20 May 2017].

Mück, M., Klotz, M. and Taubenböck, H., 2017. Validation of the DLR Global Urban Footprint in rural areas: A case study for Burkina Faso. In: *Urban Remote Sensing Event (JURSE), 2017 Joint*. [online] IEEE, pp.1–4. Available at: <http://ieeexplore.ieee.org/abstract/document/7924618/> [Accessed 4 Aug. 2017].

National Geomatics Center of China, 2014. *GlobeLand30 Product Description*. National Geomatics Center of China.

National Oceanic and Atmospheric Administration, 2017. *VIIRS Day/Night Band Nighttime Lights*. [online] Available at: <https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html> [Accessed 2 Sep. 2017].

Nogueira, K., Penatti, O.A.B. and dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61, pp.539–556.

OpenStreetMap contributors, 2017. *Planet dump retrieved from https://planet.osm.org*.

Patel, N.N., Angiuli, E., Gamba, P., Gaughan, A., Lisini, G., Stevens, F.R., Tatem, A.J. and Trianni, G., 2015. Multitemporal settlement and population mapping from Landsat using Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, 35, pp.199–208.

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A., Syrris, V. and Soille, P., 2016a. Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas. *Remote Sensing*, 8(4), p.299.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P. and Syrris, V., 2016b. *Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014*. [online] JRC Technical Report. Available at: <https://www.researchgate.net/profile/Martino_Pesaresi/publication/299597485_Operating_procedure_for_the_production_of_the_Global_Human_Settlement_Layer_from_Landsat_data_of_the_epochs_1975_1990_2000_and_2014/links/573192c208aed286ca0e1831.pdf> [Accessed 23 Jun. 2017].

Pesaresi, M., Ehrlich, D., Florczyk, A.J., Freire, S., Julea, A., Kemper, T. and Syrris, V., 2016c. The global human settlement layer from landsat imagery. In: *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. [online] IEEE, pp.7276–7279. Available at: <http://ieeexplore.ieee.org/abstract/document/7730897/> [Accessed 23 Jun. 2017].

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M.A., Ouzounis, G.K., Scavazzon, M., Soille, P., Syrris, V. and Zanchetta, L., 2013. A Global Human Settlement Layer From Optical

HR/VHR RS Data: Concept and First Results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), pp.2102–2131.

Pete Masters and Benjamin Herfort, 2016. *MapSwipe and Pybossa: Further exploration in the power of crowds*. [online] Available at: <https://www.youtube.com/watch?v=pRZ_mWn0Lmc&feature=youtu.be> [Accessed 1 Jul. 2017].

Philippines National Statistics Office, 2010. *Philippine Yearbook 2010*. Manila: Philippines National Statistics Office.

Potere, D. and Schneider, A., 2007. A critical look at representations of urban areas in global maps. *GeoJournal*, 69(1–2), pp.55–80.

Potere, D., Schneider, A., Angel, S. and 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(24), pp.6531–6558.

Python Software Foundation, 2017. *9.6. random*. [online] Available at: <https://docs.python.org/3.5/library/random.html> [Accessed 20 Sep. 2017].

Reyes, R., 2016. *Nationwide Operational Assessment of Hazards*. [online] Available at: <https://drive.google.com/file/d/0B-I810Tobv9lU3c3MVd5LWJJR0k/view> [Accessed 1 Sep. 2017].

Rutzinger, M., Rottensteiner, F. and 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), pp.11–20.

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

Schneider, A., Friedl, M.A. and Potere, D., 2010. Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment*, 114(8), pp.1733–1746.

Small, C., 2003. High spatial resolution spectral mixture analysis of urban reflectance. *Remote Sensing of Environment*, 88(1–2), pp.170–186.

Small, C., 2005. A global analysis of urban reflectance. *International Journal of Remote Sensing*, 26(4), pp.661–681.

Small, C., 2009. The Color of Cities: An Overview of Urban Spectral Diversity. In: P. Gamba and M. Herold, eds., *Global Mapping of Human Settlement*. [online] CRC Press. Available at: <http://www.crcnetbase.com/doi/abs/10.1201/9781420083408-c4>.

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

Sui, D., Elwood, S. and Goodchild, M., 2012. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer Science & Business Media.

Taubenbock, H., Esch, T., Felbier, A., Roth, A. and Dech, S., 2011. Pattern-Based Accuracy Assessment of an Urban Footprint Classification Using TerraSAR-X Data. *IEEE Geoscience and Remote Sensing Letters*, 8(2), pp.278–282.

Tenerelli, P. and Ehrlich, D., 2011. Analysis of built-up spatial pattern at different scales: can scattering affect map accuracy? *International Journal of Digital Earth*, 4(sup1), pp.107–116.

Tiecke, T., 2016. Open population datasets and open challenges. *Facebook Code*. Available at: <https://code.facebook.com/posts/596471193873876> [Accessed 15 Jul. 2017].

Weng, Q. ed., 2014. *Global Urban Monitoring and Assessment through Earth Observation*. 1 edition ed. Boca Raton, FL: CRC Press.

World Bank, 2016. *Access to electricity*. [online] Available at: <https://data.worldbank.org/> [Accessed 2 Sep. 2017].

Wurm, M., dAngelo, P., Reinartz, P. and Taubenbock, 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(10), pp.4138–4152.

Zhang, Q., Li, B., Thau, D. and Moore, R., 2015. Building a Better Urban Picture: Combining Day and Night Remote Sensing Imagery. *Remote Sensing*, 7(9), pp.11887–11913.

Zhang, Q. and Seto, K., 2013. Can Night-Time Light Data Identify Typologies of Urbanization? A Global Assessment of Successes and Failures. *Remote Sensing*, 5(7), pp.3476–3494.

Zhang, Q. and Seto, K.C., 2011. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sensing of Environment*, 115(9), pp.2320–2329.

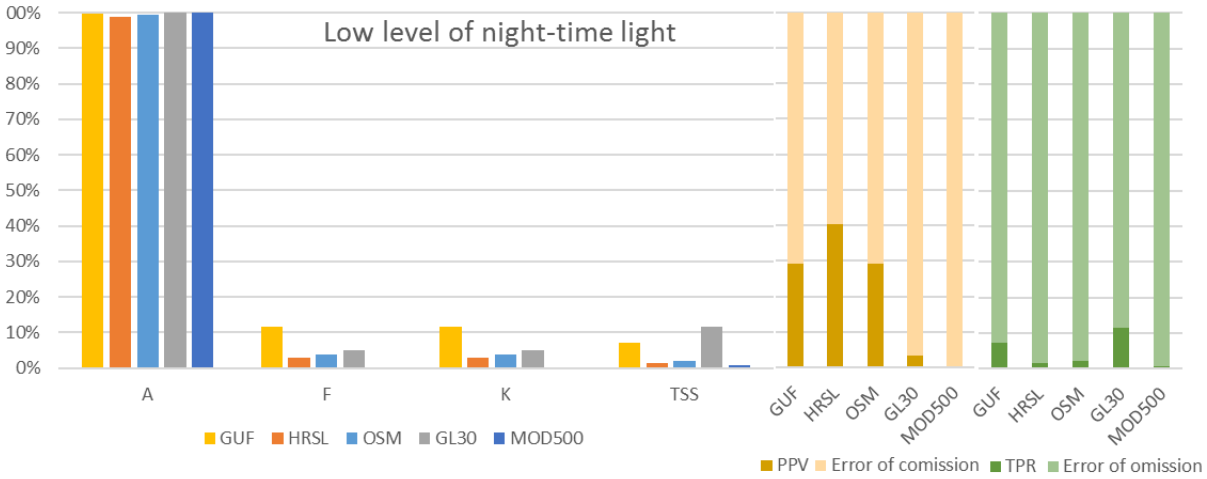
Zheng, Q., Zeng, Y., Deng, J., Wang, K., Jiang, R. and Ye, Z., 2017. 'Ghost cities' identification using multi-source remote sensing datasets: A case study in Yangtze River Delta. *Applied Geography*, 80, pp.112–121.

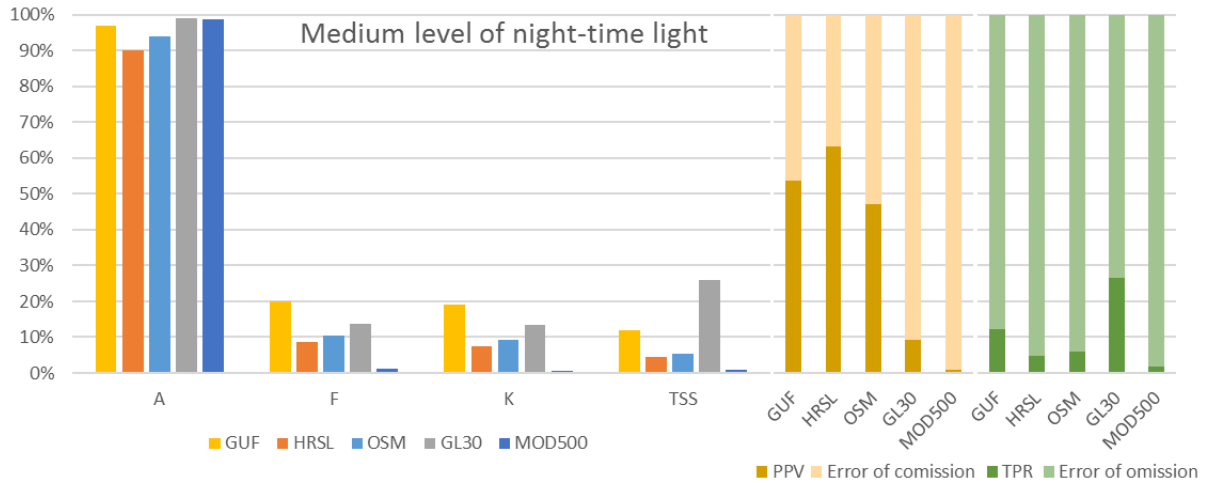
Zook, M., Graham, M., Shelton, T. and Gorman, S., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2), pp.6–32.

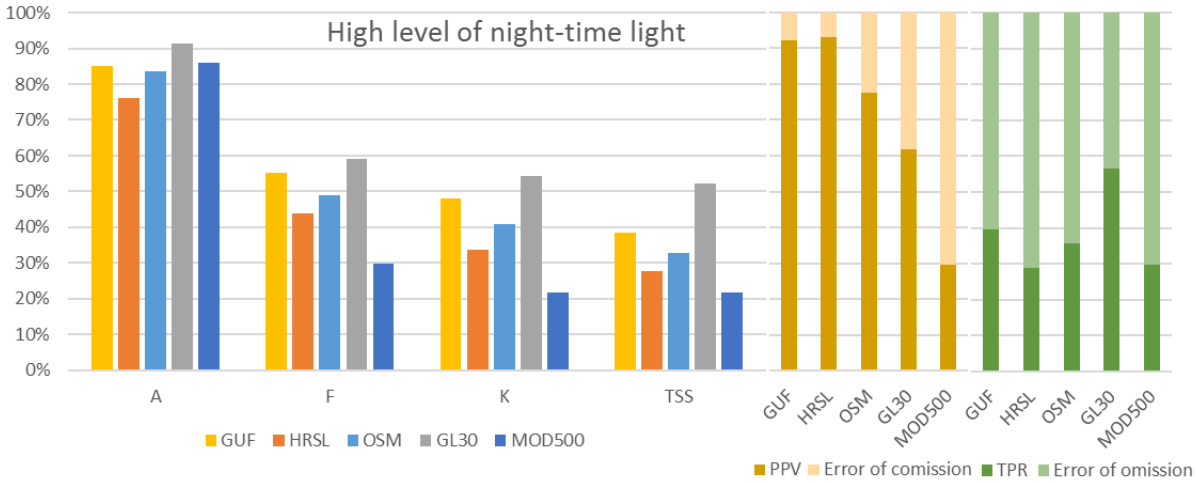
Appendices

Appendix A: Research question two Night-time Light spatial zone error matrix results

GHSL inter-map agreement by NTL spatial zone error matrix results







GUF inter-map agreement by NTL spatial zone error matrix results

HRSL inter-map agreement by NTL spatial zone error matrix results