Characterisation of Built-up Area using Artificial Intelligence and Open-Source Data for Assessment of Hazard Exposure

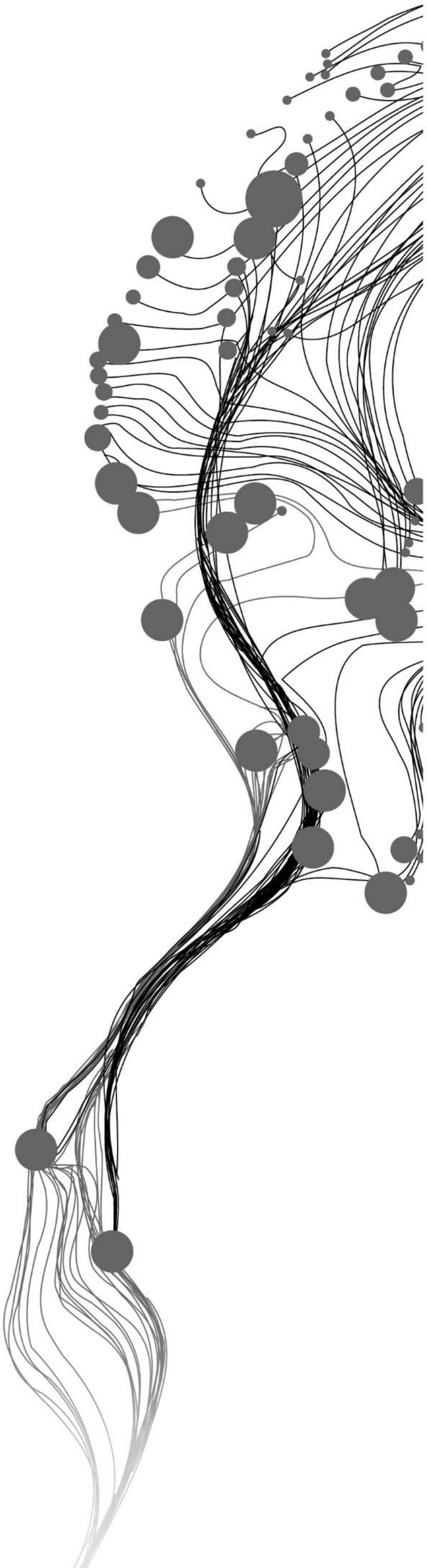
KUSHANAV BHUYAN

July 2021

SUPERVISORS:
Professor Dr C.J (Cees) van Westen
Asst. Professor Dr J (Jiong) Wang

ADVISOR:
Dr S.R (Sansar) Meena

THESIS ASSESSMENT BOARD:
Professor Dr V.G (Victor) Jetten (Chair)
Dr M (Mario) Floris, University of Padua, Italy (External Examiner)



Characterisation of Built-up Area using Artificial Intelligence and Open-Source Data for Assessment of Hazard Exposure

KUSHANAV BHUYAN

July 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.

Specialisation: Natural Hazards and Disaster Risk Reduction

SUPERVISORS:
Professor Dr C.J (Cees) van Westen
Asst. Professor Dr J (Jiong) Wang

ADVISOR:
Dr S.R (Sansar) Meena

THESIS ASSESSMENT BOARD:
Professor Dr V.G (Victor) Jetten (Chair)
Dr M (Mario) Floris, University of Padua, Italy (External Examiner)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Accurate elements-at-risk data (EaR) are one of the most important components to estimate the loss to both natural and anthropogenic hazards, particularly because of the potential increased exposure to these hazards, due to rapid urbanisation and poorly planned development strategies in hazardous regions. Therefore, it is important to not only map elements-at-risk but also to characterise them with attributes that are relevant for risk assessment. Mapping of building EaR includes the footprint information and their characteristics; however, acquiring them is difficult because of the following factors: lack of data accessibility, missing attribute data of buildings, data incompleteness and positional accuracy error, and many others. Major developments have taken place in the collaborative mapping of buildings, using platforms like OpenStreetMap. However, many areas in the world still lack this data. Therefore, the mapping of buildings footprints and their conversion into usable EaR maps is a challenge. To address these issues, we designed a semi-automated workflow that caters to the development of buildings EaR database by (1) detecting buildings footprints using a ResU-Net deep learning (DL) model and (2) characterising the footprints using building morphological metrics and open-source auxiliary data at a homogeneous block level. Based on our results, the building EaR footprints were detected with over 76% F1-score using the DL model and later classify them into building occupancy types like residential, commercial, industrial etc. Another major investigation that we examined is the transferability of the workflow in a different study area, which addresses the reproducibility of the method. After obtaining the final building EaR maps, we assessed the exposure of the building EaR by spatially overlaying the EaR maps over the flood susceptibility maps to understand how the building function and the occupants are affected. Our study has a huge significance, chiefly in (1) generating a building EaR database in data-scarce regions as a first approach (which were previously not explored), (2) transferring the methodology over a different test area and achieving good results, and for future applications in (3) linking the building occupancy types to hazard vulnerability and the subsequent hazard risk, and (4) serving projects and policy developments of regions for risk assessment, disaster risk mitigation and risk reduction.

Keywords: Building Detection, Building Characterisation, Building Morphology, Open-Source Data, Homogeneous Built-up Area, Exposure Assessment.

ACKNOWLEDGEMENTS

"Your struggle is just a part of your life, do not let it dictate you, rather let it guide you"

~ Shri Gokul Bhuyan (My Father)

I really do not know where or even how to begin this final section of my thesis. It has been a very surreal experience for me in the past nine months here in ITC with my research work. I think I will first start by thanking my caring, supportive, and remarkable supervisors at ITC, Prof. Dr Cees Van Westen and Asst. Prof. Dr Jiong Wang. I would have been lost without their constant cooperation and encouragement, who helped make this thesis come to a reality. A special thanks also to both Dr Jiong Wang and Dr Sansar Raj Meena, who helped me indulge and dive deeper into the world of Artificial Intelligence and Deep Learning. The three above were my research mentors, each playing a key role in different aspects of my thesis. This research has really been like a machine with different components where each of my mentors had their own role, and only because of them, these components functioned smoothly, thereby completing the thesis on time. Prof Cees has always been upfront and honest with my decisions and ideas, and the fact I have reached this position is surely due to his dedicated and energetic discussions with me. I owe all of my research skills and knowledge to him. I would not have reached this position of critical thinking and have the scientific vigour without his presence as my supervisor. I still have a lot to learn, but I take this as my first step to be a better researcher, learning from the path that Prof Cees has laid down for aspiring researchers like me.

I would also like to give special thanks to my friends Bharat Reddy, Ashok Dahal, Om Prasad Dhakal, Luo Eqi, Vasudha Chaturvedi and many others (sorry for not naming all of you) who helped me in some way or the other in my thesis journey. Of course, I cannot forget my family, Maa, Papa, and my brother, for their constant doses of blessings and love, which gave me the moral stamina to continue on this path toward finishing my master thesis. Honestly, I could not have been more blessed to have you all in my life. I wish you ALL the very best in your lives and endeavours in the future.

A great amount of thanks also goes to the Faculty of ITC. The professors here are some of the best that I have had the experience to be taught under, and I know for a fact that every bit of teaching from their side played some role in contributing and developing my research ideas. Professors like Dr Luigi Lombardo, Dr Bart Krol, Dr Norman Kerle, Dr Victor Jetten and many others have helped me shape my present and future mindset towards quality research. I thank you all for your time and energy in bestowing your knowledge to us students.

Finally, and not least, I cannot thank you enough, Miss Wu Zijing. You have been the backbone, the sustenance, and the drive that has made me not just learn but also love to appreciate the newer things in life (especially research). You have been there for me as a partner, friend, guide, critic, and most importantly, the best support I could have ever asked for. Thank you so much for your love, adoration, and patience in the past two years here with me. From sitting down to understand conceptual diagrams to addressing my quirks and drawbacks in my research writing, you have always been there. This journey could not have been possible without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI – Artificial Intelligence

BCE – Binary Cross Entropy

CBD – Central Business District

CNN – Convolutional Neural Network

DL – Deep Learning

DT – Decision Tree

EaR – Element-at-Risk

FCN – Fully Convolutional Network

GDB – Geographical Databases

GIS – Geographic Information System

GSV – Google Street View

GUF – Global Urban Footprint

ICFOSS - International Centre for Free and Open-Source Software

KSDMA – Kerala State Disaster Management Authority

LiDAR - Light Detection and Ranging

ML – Machine Learning

NN – Neural Network

NUIS – National Urban Information System

OBIA – Object-Based Image Analysis

RF – Random Forest

SAR – Synthetic Aperture Radar

SVM – Support Vector Machine

OSM – OpenStreetMap

UAV – Unmanned Aerial Vehicle

VGI – Volunteered Geographic Information

VHR – Very High Resolution

CHAPTER 1: INTRODUCTION

The research idea and the associated background that motivates the research in terms of the existing gaps are described in this chapter. This chapter includes the (1) background, (2) research problem and scientific significance, (3) research objectives, questions, (4) research design and conceptualisation, and (5) literature review.

1.1. Background

The fast urbanisation and poorly planned development strategies in hazardous regions have increased the potential of exposure to both natural as well as anthropogenic hazards. The impacts of hazards are manifold such as loss of life, property damage, and economic disruption, that need to be assessed for effective risk reduction planning (Eshrati, Mahmoudzadeh, & Taghvaei, 2015). A way of assessing the impacts of hazard events is by hazard risk assessment, which allows identifying expected loss caused by probable hazards and fosters the necessary information to make decisions on optimal risk mitigation and risk reduction measures (Gill & Malamud, 2014). A multi-hazard risk assessment also accounts for possible hazard interactions with multiple event probabilities for multiple types of elements-at-risk [1] (EaR) and multiple potential loss components. The risk associated with the hazard processes is quantified based on the hazard intensity, spatio-temporal probability, the exposed EaR and their respective physical vulnerability[2] (Chen et al., 2016). Elements-at-risk mapping is crucial for exposure analysis, vulnerability, and hazard risk assessment to identify who and what is at risk.

The identification of EaR includes the detection and characterisation of EaR, where detection refers to *the delineation of existing EaR footprints*, and characterisation refers to *the associated EaR typological attributes*. Typical EaR that is exposed to hazards are buildings, people, agricultural lands, vegetated areas, transportation networks etc. Implementing approaches for safeguarding EaR from hazard impacts is crucial and cannot be executed without proper datasets. Databases with updated information about elements exposed to hazards are fundamental for response activities and support crisis preparedness (Eshrati et al., 2015). Proper development of an *elements-at-risk database* is crucial as it takes into consideration the associated attributes or characteristics[3] of the EaR. Buildings are one of the most important EaR as it encompasses both population and material possessions that are of value. Information such as the building use, the structural type, the number of floors, content within the buildings, the replacement value, and the characteristics of the inhabitants are important. Furthermore, rapid mapping of EaR is also essential along with such contextual information as it has implications for vulnerability assessment of building EaR, disaster management, emergency planning, and formulation of mitigation measures (Papathoma et al. 2007).

Citizen based science and collaborative Geo-information Science are popular means of obtaining data on buildings EaR. Volunteered Geographic Information (VGI) (See et al. 2019) is one such example that has aided in mapping many activities and can be used to acquire information quickly and cheap over large areas.

[1] Elements-at-risk are population, properties, economic activities, or any other entity of value that may be affected by hazardous phenomena, either directly or indirectly, in a particular area.

[2] Physical vulnerability is expressed as the degree of loss or damage to a given element within the area affected by the hazard (Quan Luna et al., 2011).

[3] The words attributes, typology, and characteristics are used interchangeably in the context of the research.

With the recent advancements in Geo-information Science (GIS) technology and the progressive emergence of citizen science (Goodchild, 2007), collaborative approaches have contributed to many applications like land cover mapping (Ribeiro & Fonte, 2015), post-disaster mapping (Panek, 2015), landslide inventory mapping (Hao et al., 2020), mapping remote villages (Kanthi & Purwanto, 2016) and community development (Panek & Netek, 2019) in countries like South-Africa (Panek, 2015), Spain (López et al. 2014), and Malaysia (Husen, Idris, & Ishak, 2018). OpenStreetMap (OSM), which started in 2004, is now one of the best-known VGI projects that perform collaborative mapping and has been used for many applications. A study by Barrington and Millard-Ball (2017) estimated that OSM data had reached more than 80 per cent of completeness on a global scale. This completeness encourages the use of such datasets for developing solutions for practical applications in emergency planning, risk mitigation planning and many other fields.

However, the collection of attribute information for objects such as buildings has proven to be problematic. OSM data has also shown reasons for concern regarding positional accuracy and quality. Often there is no updating of the OSM database, and therefore, buildings on OSM might not be the actual buildings that are present in reality. For instance, buildings destroyed by a disaster still display the buildings in OSM that no longer exist physically (Foody et al., 2015). Accurate attribute data collection from OSM is also recognised as a major challenge as the OSM database has poor building characterisation since the building function cannot be seen by the voluntary mappers on the satellite images, and therefore are sometimes left blank (Zhang & Pfoser, 2019). The correct and complete attribute information of EaR is important to assess the vulnerability of EaR under different hazard scenarios and as input in risk assessment. Building typological attributes, for example, based on occupancy class (e.g., single-family dwelling), structure type (e.g., reinforced concrete) and the number of floors, is employed in the analysis of the vulnerability, loss estimation, and the subsequent risk. Current online products and tools like Mapillary, Google Street View (GSV) images, Google Maps, Bing Maps, global land cover maps, land use data, and other such auxiliary datasets can help provide contextual information (of occupancy type) about the attributes of the building EaR; however, the integration of such information with collaborative mapping can be challenging as the nature of these data are different across the board, and they cannot be linked directly. For example, the data of GSV, Google Maps, landuse are of three distinct data types: RGB photographic images, point and polygon vector data, and image raster data, respectively. Hence, an outlook towards a streamlined EaR identification framework is required that can help bridge this gap.

The advent of remote sensing has made ground-breaking contribution to the mapping of EaR. With the advances in satellite remote sensing technology, rapid progress in un-manned aerial vehicles (UAV), and substantial improvement in data acquisition, processing and interpretation have made it easier to detect land surface objects (Wu et al., 2020). Today, remote sensing techniques like Synthetic-Aperture Radar (SAR), multi-spectral imaging, hyperspectral imaging, Light Detection and Ranging (LiDAR), and UAVs enable the detection of many surface objects.

Traditional efforts in detecting building EaR from remote sensing imageries such as visual interpretation and manual digitisation approaches have witnessed significant drawbacks. OSM databases are an example that falls under such traditional efforts. As stated by Wu et al. (2020) and Ghorbanzadeh et al. (2020), such methods face difficulties, manifested mainly due to the following: (1) subjective visual interpretation; (2) discriminating closely located buildings as the same, (3) possibility of omission and misclassification, and (4) missing objects (such as buildings), attributes (such as occupancy type) and value (such as residential) in OSM data sets (Mobasheri, Zipf, & Francis, 2018), making it challenging to annotate different types of surface objects (e.g., buildings). Henceforth, it is essential to look at opportunities that can cater to a low-cost automated methodology to extract building EaR. There are several methods of classifying EaR like pixel-based and object-based classification techniques. While pixel-based methods only extract features from pixels by classifying each pixel accordingly, they do not take the spatial context into account. However, object-based methods (object-based image analysis) (Blaschke, 2010; Parker, 2013; Pesaresi, Gerhardinger, & Kayitakire, 2008) explicitly extract the spatial information of pixels from satellite images. In the past decade, methods such as artificial intelligence (AI), machine learning (ML) algorithms like neural networks (NN), support vector machine (SVM), decision trees (DT), Random Forest (RF) and deep learning (DL) have been widely employed for improved automation of surface object classification (Karpatne et al. 2016). The development of these classification methods has significantly increased the speed and amount in extracting information of surface objects. However, ML and DL-based classification are highly dependent on the number of training samples of the object of interest (Chen & Zipf, 2019). Obtaining training samples from visual interpretation (point/polygon digitisation) of satellite images can be time-consuming; however, it can be resolved partially through collaborative mapping (like OSM), as it contains footprint information of the building EaR. Thus, the intent for a low-cost-rapid approach for EaR detection through ML-DL approaches can have practical advantages in terms of the speed and accuracy of detecting the building EaR footprints by employing readily available building footprints as training labels.

The characterisation of building EaR is a challenge as well and needs further research. Remote sensing images alone cannot be used for characterising attributes of EaR based on visual interpretation. Characterising buildings to estimate typological attributes, for example, building occupancy types from just satellite images, is very difficult, especially by interpreting the (1) roof tops, (2) neighbours surrounding the buildings and (3) colour of buildings for example. Thus, there is a need to address the characterisation of building EaR for effective emergency monitoring, rapid response services, vulnerability assessment, and implementation of disaster risk reduction measures. Therefore, the thesis research intends to solve this problem.

1.2. Research Problem and Scientific Significance

1.2.1. Research Problem

Problem Statement: Accurate and rapid identification of elements-at-risk for hazard exposure, vulnerability, and risk assessment at low-costs is challenging due to the deficiency and incompleteness of existing EaR datasets.

The issue of (1) lack of data accessibility is common, where projects aimed at increasing resiliency of infrastructures often lack the required attribute information of EaR for assessing the exposure, vulnerability, loss, and the associated risk. Some of the common ways of obtaining EaR and its attribute data are through OpenStreetMap (OSM), Mapillary, Google Street View, field visits and others, but (2) the required attribute data are sometimes difficult to obtain, are absent or unavailable. Moreover, (3) data completeness and positional accuracy is a matter of concern as well. Existing remote sensing data and information are capable of interpreting many surface objects worldwide; however, (4) the detection and conversion of the detected object footprints into usable elements-at-risk maps is a challenge. (5) The characterisation of buildings at a footprint level is also a difficult task. Although there are auxiliary data such as population census, cadastral maps, human settlement and built-up area databases, label information from OSM, Mapillary and Google Street View, and other proxies that can help approximate the typological attributes of the EaR; however, (6) the integration/amalgamation of such data is difficult due to the differences in the type of data they inherently exhibit such as photographic images, geotags, and raster maps. The gaps mentioned above are realised, which is to be fulfilled by the MSc research thesis.

The question of "how to quickly produce quality building EaR database in a cost-effective way?" has become urgent. There is an absence of a streamlined-generalised workflow of acquiring EaR data of reasonable quality with their respective typological attributes, which can be reproducible in different areas.

1.2.2. Scientific Significance

The main goal of the research is to perform a first-approximated hazard exposure assessment of a study area by investigating the applicability of the generated EaR outputs from DL models and amalgamation with the open-source. Furthermore, the research novelty lays its foundation on (1) the development of a workflow of identifying EaR at homogeneous spatial units with open-source data, and (2) transferring the workflow to be reproducible inn different regions. The development of a generalised-reproducible workflow with lost-cost data is crucial. The research is envisioned at serving projects and policy developments inf regions that have a deficiency of quality EaR information required for risk assessment, disaster risk mitigation and risk reduction.

1.3. Research Objectives And Questions

1.3.1. General Objective

The main objective of the research is the *semi-automated detection* and *characterisation* of elements-at-risk into *homogeneous units of built-up area* from *open-source data* as input for hazard *exposure assessment*. In order to achieve this goal, a range of sub-objectives and research questions are identified as presented below:

Sub-objective 1 (S-O1): To detect building footprints and built-up areas from satellite image and OSM data.

1. Which of the many DL architectures would be suitable for detecting buildings?
2. To what extent can OSM data be employed as training data in DL? What are the constraints associated with the OSM data?
3. How well can the DL model be used to detect buildings and built-up areas in different study areas?

Sub-objective 2 (S-O2): To develop a methodology for the sub-division of the built-up areas based on the characteristics of the building footprints within them.

1. What are the morphological metrics that can be obtained from buildings footprints?
2. Can these metrics be used to divide the built-up area into homogeneous unit? How to measure the homogeneity level of the units?
3. How can information on roads, railways, and other linear features help to refine the sub-division of homogeneous built-up areas?

Sub-objective 3 (S-O3): To develop a methodology to use geotags and labels from different data sources for the characterisation of the occupancy types of the homogenous units.

1. Can label information from OSM and Google Maps be obtained in an automated manner to characterise homogenous units?
2. Could these label information help to determine the most likely land-use type and occupancy type of the homogeneous buildings?
3. How to determine the occupancy type for units with too few or conflicting label information?
4. Can the methodology be transferable to different study areas?

Sub-objective 4 (S-O4): To evaluate the applicability of the resulting homogeneous units for exposure and vulnerability assessment.

1. To what degree are the homogeneous built-up area suitable for exposure, vulnerability, and risk assessment?
2. To what extent can the homogeneous built-up area be used with existing vulnerability curves?
3. Can the building footprint information be used to quantify exposure and vulnerability better?

The four sub-objectives will be addressed in the succeeding chapters. Bear in mind that the exposure assessment itself is not the main goal of the thesis to address, rather an exploration of whether the EaR outputs can be used to assess the exposure of the EaR.

1.4. Research Design and Conceptualisation

Figure 1 shows the overall design of the research, which is broadly divided as: Literature Review, (S-O1) deep learning model training for building detection, (S-O2) building morphometrics and homogenisation, (S-O3) auxiliary data for characterisation, (S-O3) final homogeneous built-up area generation through characterisation, and (S-O4) exposure assessment to flooding.

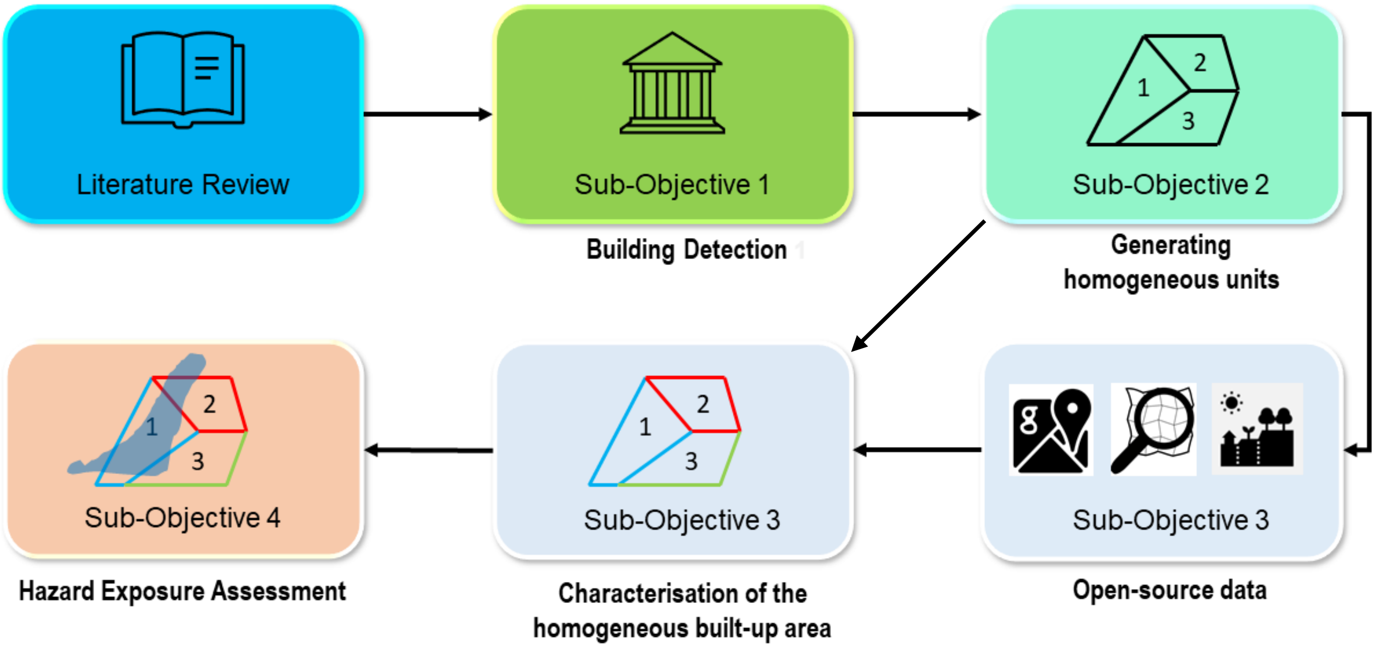


Figure 1: Research Design: Literature Review, (S-O1) deep learning model training for building detection, (S-O2) building morphometrics and homogenisation, (S-O3) auxiliary data for characterisation, (S-O3) final homogeneous built-up area generation, and (S-O4) exposure assessment to flooding.

1.5. Literature Review

1.5.1. Elements-at-Risk Detection

1.5.1.1. Object-Based Image Segmentation

The union of Geographic Information System (GIS) and image processing with Object-Based Image Analysis (OBIA) started to grow rapidly in the early 2000s, aiming to delineate readily available surface objects from satellite imageries by generating image objects that utilise the spectral and contextual information for classification of spatial properties through image segmentation. OBIA deals with the problem of pixel-based classification by grouping spectrally similar non-overlapping pixels in segments (Blaschke, 2010). By nesting pixels within the context of their discrete representations, OBIA mimics the human logic process (Parker, 2013). OBIA has been successfully used in many past applications for mapping population distribution, building and road footprint detection, and many other fields (Blaschke, 2010; Tavakkoli Piralilou et al., 2019). Prathiba et al. (2020) extracted building footprints from very high-resolution (VHR) images through the nearest neighbourhood classification after image object segmentation in Ahmedabad city in India. However, the problem of segmentation anomalies (over-segmentation and under-segmentation) that specifies the quality of *segmentation goodness* for footprint extraction of buildings still remained. Whenever the segmentation produces the objects that are used for the classification, the results may be influenced by the quality of this *segmentation goodness*, especially in highly heterogeneous urban areas where building and roads could frequently be misclassified (El-naggar, 2018). This issue makes it challenging to extract and categorise different types of buildings without prior contextual knowledge of the area (Angela, Norbert, & Jochen, 2013); thus, more robust techniques are required to improve building detection from satellite imagery. Furthermore, OBIA requires expert-based optimisation of segmentation parameters, and hence, the degree of automation is low compared to pixel-based methods(Sameen & Pradhan, 2019) like Convolutional Neural Networks discussed ahead.

1.5.1.2. Convolutional Neural Networks

During the past decade, DL methods, like Convolutional Neural Networks (CNNs), have achieved significant success in remote sensing image classification (Zhu et al., 2017). CNN is a DL algorithm under the umbrella of the machine learning family, which stems from the research on artificial neural networks and is based on the algorithm of back-propagation that allows feature learning (Zhou, 2018). Multiple hierarchical stacking and trainable layers enable CNNs to learn characteristic features and abstractions from satellite images (Fu et al., 2019). CNNs can extract hidden features, considering the common ones like colour, shape, and size, and deep features of ground objects such as spatial relationship features. CNNs consist of three layers: the convolutional layer, the pooling layer, and the fully connected layer. The convolutional layer defines a window or a filter that scans an entire image through this window and outputs a feature map. The pooling layer help compresses spatial information from the feature maps. Max pooling is one of the most popularly used examples; it returns the maximum value present inside the filter for each scanning location. Finally, the fully connected layer takes the convolution and pooling process results to classify the images. The output of this layer is flattened into a single vector of values, each representing a probability of features belonging to a specific label. Such characteristics have enabled CNN-based models to exhibit impressive accuracies in image classification (Xie et al. 2020), object detection (Ghorbanzadeh et al. 2019; Guirado et al. 2017; Sameen & Pradhan, 2019) and instance segmentation (Dai, He, & Sun, 2015; Iglovikov et al. 2018). The inherent characteristics of CNNs make it a plausible candidate for building footprint extraction (Alidoost & Arefi, 2018; Cohen et al. 2016; Stewart et al. 2020; Xie et al., 2020; Zhou et al. 2019).

As the research is interested in closely looking at the classification of buildings in satellite images, semantic image segmentation would be addressed from this point onward. *"Semantic image segmentation is a classic computer vision problem to mask out regions of interest"* (Pan et al. 2020). Essentially, it describes the association of image pixels to specific class labels such as buildings and non-buildings. In the venture for building segmentation, pixel-by-pixel manner semantic segmentation is performed using DL algorithms like Fully Convolutional Networks (FCN) (Wu et al., 2018). FCN is one of the most important networks in DL for semantic segmentation (Zhu et al. 2017). FCN introduced significant ideas like end-to-end learning of the upsampling algorithm via an *encoder-decoder* structure and skip connections to fuse information from different depths in the network. Some popular networks based on FCNs are U-Net (Ronneberger, Fischer, & Brox, 2015) and SegNet (Badrinarayanan, Kendall, & Cipolla, 2017). Many CNN models have been proposed in recent studies, such as DenseNet (Liu et al. 2020), U-Net (Yang et al. 2019), Mask R-CNN (Zhao et al. 2018), VGG-F (Ajami et al. 2019) and ResU-Net (Diakogiannis et al. 2020).

In recent years, many CNN architectures with excellent performance have been reported and used worldwide to classify and detect buildings. The growing development in remote sensing technologies with better spatial resolutions has laid the foundation for a whole new set of opportunities for urban risk planning, environmental monitoring, and other similar fields. Among many, U-Net proposed by Ronneberger et al. (2015) appears to be more adopted for remote sensing applications. Pan et al. (2020) highlight the integration of complex U-Net architecture with VHR satellite images to offer accurate building information in complex urban villages, which is frequently required for urban redevelopment in urban spaces. The feasibility, capability, accuracy, lesser training data-intensive, and overall lightweight nature of U-Net in semantic segmentation for high-density buildings was demonstrated in their research. Although the paper stated apparent issues with the separation of individual building polygons, the thesis research remains interested in a more homogeneous spatial unit based on built-up area rather than focusing on individual building footprints for the extraction of building characteristics in the later parts of the thesis. U-Net has also been widely used for road detection and road centerline extraction (X. Yang, Li, Ye, Zhang, et al., 2019) and thus, exhibits extensive usability of the architecture in many applications. However, recent studies indicate that very deep networks are associated with better performance when it came to semantic segmentation tasks. To experiment with this observation, Yi et al. (2019) made use of deep residual networks with the aforementioned U-Net model to understand how deeper networks really affect the performance and published results with an average 3.5% increase in overall F1-score accuracy in building segmentation.

Since CNNs like U-Net and ResU-Net uses the same feature maps that were used for the contraction and expansion of a vector (or matrix) to a segmented image in the network (during the encoding and decoding phases), this preserves the structural integrity of the image and thus, reduces distortion immensely (Ronneberger et al., 2015). Furthermore, the ResU-Net architecture excels at predicting with limited data (Qi et al. 2020). Recent research by Alidoost and Arefi (2018) have also reported improved building detection using the ResU-Net model, and therefore, the ResU-Net model was chosen for building detection.

1.5.2. Elements-at-Risk Characterisation

Research by Graff et al. (2019) employed the information of buildings EaR from multiple geographical databases[4] (GDB) produced by national institutes, VGI and archive documents to identify EaR at different scales. Their study shed light on adding information about the infrastructure like construction material, number of floors, building conditions etc., to the EaR to characterise them at different scales. Moreover, their study also emphasised the harmonisation of different GDBs to assess and characterise the EaR. However, the availability and accessibility of archive data as GDB can be challenging in some countries due to security and administrative reasons. Furthermore, information from VGI, like OSM, need auditing as rightly addressed by the author before using it for characterisation due to the possibility of erroneous EaR footprints and label information.

The first step towards characterising buildings is by looking at their physical morphology and how they relate to nearby buildings as well as the surrounding areas. Such morphological measurements or metrics can give meaningful insights about the types of buildings that potentially exist in certain places, linking to possible building functions like occupancy types. The current methods for calculating spatial metrics, such as FRAGSTAT (Grippa et al. 2018), offer a wide variety of landscape metrics for categorical map patterns. Unfortunately, it is limited by the size of the dataset (McGarigal, 2015) and offers limited automation. The Momepy urban morphology package (Fleischmann, 2019) is a Python library that was developed for quantitative analysis of urban form and morphometrics. The library allows calculation of building diversity, adjacency, area coverage and other structural parameters that can be key in clustering buildings into *homogeneous spatial units* (discussed more ahead; see chapter 4). Thus, the tool can be the bridge between the *detection* and *characterisation* of the building EaR. The intent behind semantic segmentation through DL and the focus on homogeneous spatial units revolve around developing the methodology (sub-objectives 1, 2 and 3) that enables the characterisation of homogeneous buildings (built-up area) required for exposure and vulnerability and risk assessment.

The idea of studying building morphologies is inspired by the works of Angela et al. (2013) and Blanco-Vogt and Schanze (2014) that laid the groundwork for semantically grouping buildings based on building characteristics like size, form, proximity to other buildings and building compactness; which can be leveraged from the Momepy library. Another crucial research work in employing urban morphometrics for characterising building was done by Fan, Zipf, and Fu (2014). While their work holds resemblance to the thesis research, the main difference comes from the fact that their study was conducted in five cities in Germany involving a complete OSM dataset with proper building data with over 2027 well-labelled buildings, thus possessing high data completeness. Such completeness of data can indeed help determine the building attributes with relative ease; however, in data-scarce regions like Palakkad and Kollam in Kerala, with only over 260 properly labelled buildings ranging over 26.6 km^2 of area, the task of identifying building tags with just the OSM becomes infeasible. Moreover, much of the building labels are tagged as *None* and *Yes*, creating confusion about which particular type of building it refers to in reality. Furthermore, the study of Fan, Zipf, and Fu (2014) focused on assessing the building types at an individual building level. However, due to the diversity and heterogeneity found in the buildings of Palakkad and Kollam and catering to a more general level of exposure and vulnerability assessment in terms of the building typology as occupancy type, the thesis research focused on determining the building types at a more homogeneous built-up area level.

[4] Geographic database is defined as a catalogue that stores spatially referenced data. Such databases are collections of data that are related either through location, data type, or a common underlying purpose.

Furthermore more, their studies did not include the usage of additional proxy data from online products and tools such as Facebook Mapping, Mapillary, Global Human Settlement Layer, Global Urban Footprint, WSF-3D[5], Google Maps, Google Street View, OSM labels, and land use/land cover maps for approximating building characteristics, probably as some of the former products were not available at the time of their research. Hence some of the products mentioned above will be employed for the purposes of the methodology in the succeeding chapters. The use of OSM for estimating building characteristics have been explored previously by Fan et al. (2014), Y. Sun, Shahzad, and Zhu (2017), and Cerri et al. (2021), where the latter recommended using OSM building information for flood vulnerability modelling stating that such openly accessible data makes it easier and cost-effective to study the effect of hazard to building EaR. They also discussed the employment of other proxies (or auxiliary) data for improved EaR exposure-vulnerability to flooding, which this thesis is partially addressing.

Building occupancy type is a very important attribute that is tightly connected to population activity patterns like shopping, residency living, recreation, and meetings. Stewart et al. (2016) developed a method to use Bayesian machine learning to estimate building occupancy type from population density tables that uses mined data of population statistics for a wide array of buildings for predicting/modelling occupancy of buildings. Hasan et al. (2018) used LiDAR data to extract building footprints and building heights automatically and estimated the building occupancy types for landslide exposure to EaR by manual interpretation efforts. The authors suggested developing a semi-automated process to detect EaR to reduce time and cost. The thesis research attempts at a different take on the estimation of the building occupancy type by developing a characterisation procedure with open-source data like OSM, Google Maps and available land use maps. The details of the methodology regarding the *characterisation* phase and the respective results will be discussed in chapters 5 and 6.

[5] World Settlement Footprint. Source.

CHAPTER 2: RESEARCH METHODOLOGY, TEST AREAS, AND DATA

This chapter includes (1) the research methodology, (2) the test area description and motivation, and (3) the data set acquisition and description.

2.1 Research Methodology

In order to integrate the detection and characterisation of buildings into a meaningful EaR identification, auxiliary data like Google Maps, OSM building tags, and land use information can be employed to develop a methodology of acquiring homogeneous spatial units with aggregated typological attributes. The characterisation of the building EaR will remain at a coarser level, where attributes of buildings will be largely estimated at a homogeneous block level.

Figure 2 illustrates the overall steps that are taken to accomplish the research sub-objectives. The steps include:

1. Preparation of remote sensing data from satellite images and ground truth data from OSM. (Chapter 3)
2. Sampling of data to generate training, validation and test sets, and model training to detect buildings. (Chapter 3)
3. Acquiring characteristics parameters from structural (morphological) and proxy (open-source data like OSM, land use data, and Google Maps) data. (Chapters 4 and 5)
4. Combining the characteristic parameters with the detected buildings to assert typological (occupancy type) attributes of the buildings at a homogeneous unit. (Chapter 5)
5. Exposure assessment with the derived output from the attributes of the buildings at a homogeneous unit as a means of exploring the opportunity to assess the exposure of the EaR. (Chapter 7)

Based on the steps mentioned above, the methodology essentially aims to *identify* building EaR. Meaning that using the state-of-the-art DL models and coupling the resultant outputs with openly available data, the amalgamation between them will be used to derive building occupancy types for the study areas (chapters 5 and 6).

The research sub-objectives 1, 2, 3 and 4 are addressed in chapters 3, 4, 5 and 7. These chapters will explain the respective research methods, results, discussions and, in the process, attempt to answer the associated research questions. Later on, to test the applicability and reproducibility of the workflow, the methodology will be applied to a second study area, where the results and discussion will be conferred in chapter 6. The codes for obtaining the different outputs can be found in the appendix section.

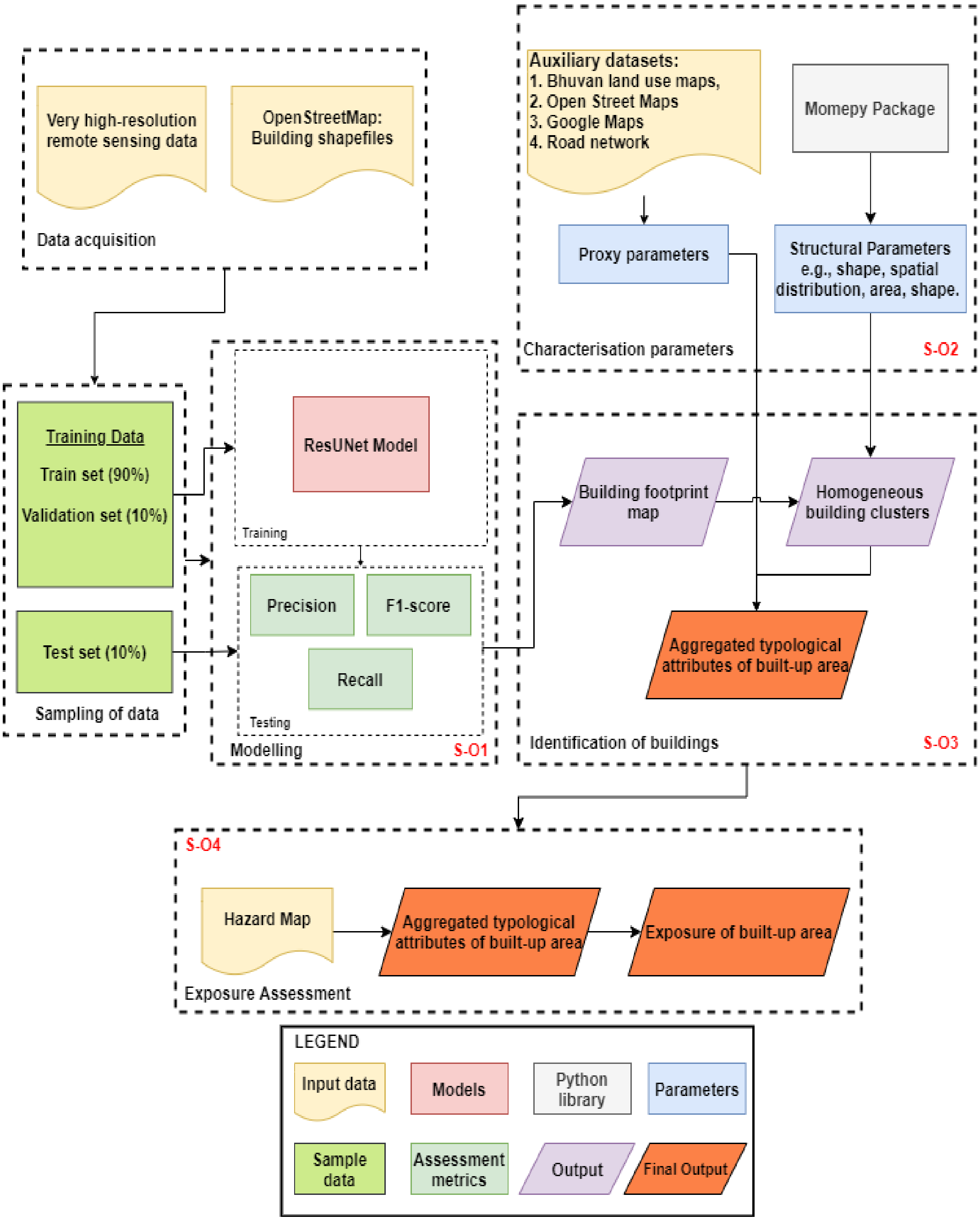


Figure 2: Research Methodology. Refer to section 1.3.1 for the sub-objectives (S-O1 to S-O4).

2.2 Test Areas

The study areas where the methodology will be tested can be seen below in figure 3. In developing countries like India, many regions suffer from data scarcity. In order to assess and evaluate hazard risk in terms of both monetary costs and the physical population that can be potentially affected, data of buildings EaR are quintessential. The year 2018 was a big year for monsoonal disasters in southwest India, particularly Kerala. Local and national news reported several flood and rainfall-induced landslides throughout many districts of Kerala, estimating the displacement of 85,000 people (figure 5) and the destruction of many properties (figure 6) where water had overrun riverbanks, submerged city buildings and left dozens of people dead (Dwyer, 2018). The Kerala State Disaster Management Authority (KSDMA) are portraying their roles in

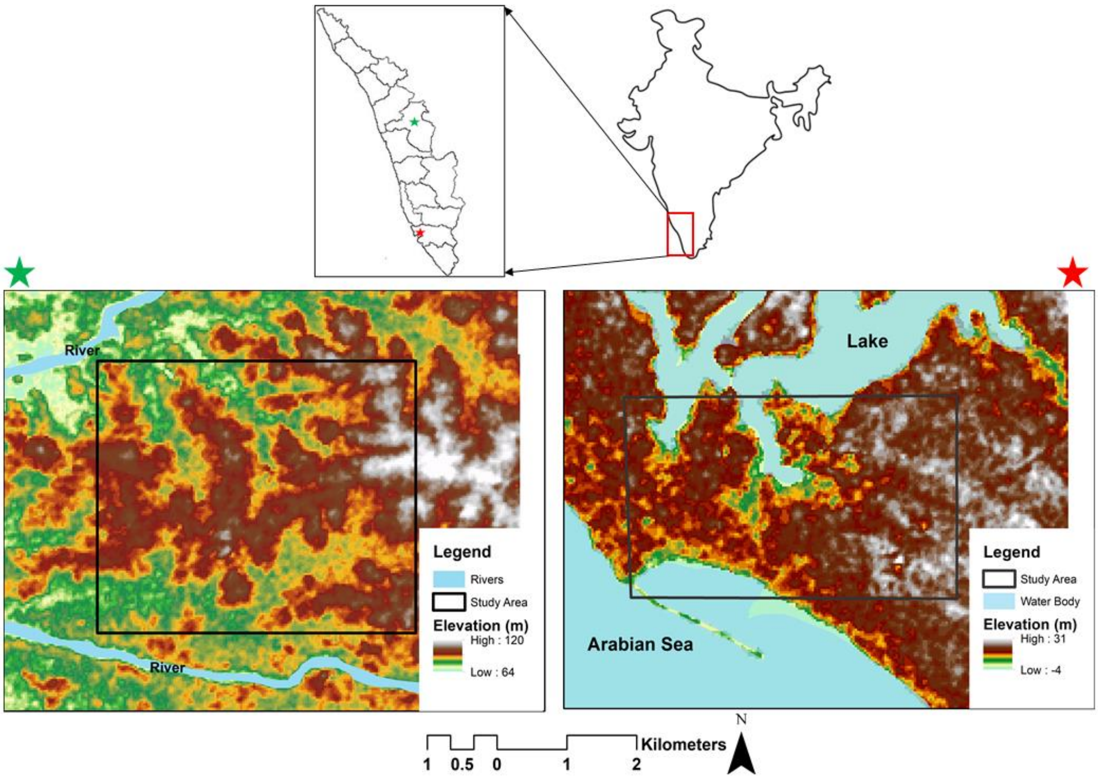
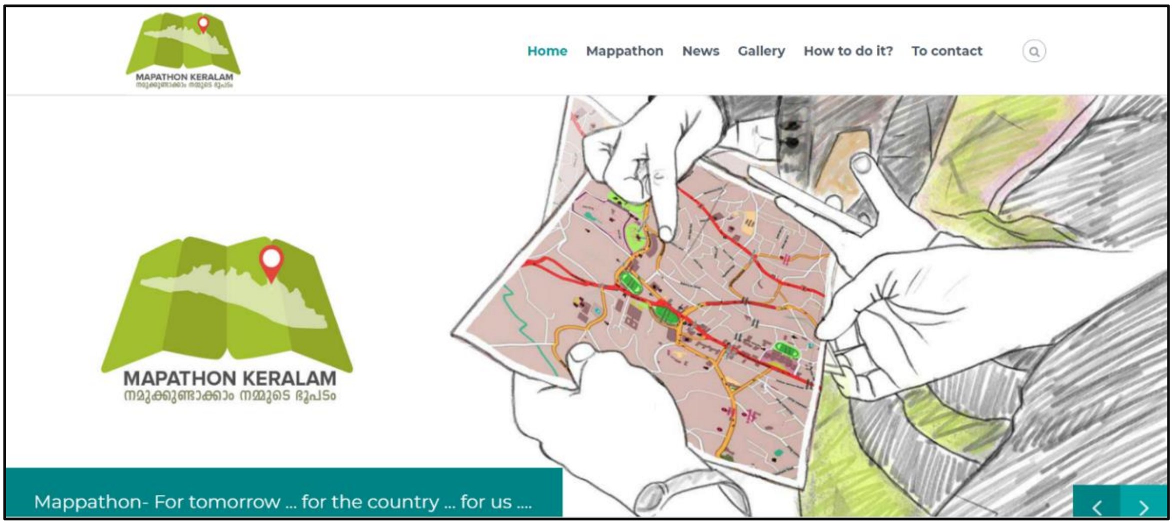


Figure 3: Study area of Palakkad (green star) and Kollam (red star).



Mappathon- For tomorrow ... for the country ... for us ....

Figure 4: Mapathon Kerala initiative. Image Source

improving the disaster risk management in Kerala in partnership with agencies like the International Centre for Free and Open-Source Software (ICFOSS) with collaborative mapping initiatives to develop building EaR database. The collaborative mapping initiative known as the *Mapathon Kerala* (figure 4) is a project that prepares maps of Kerala's public assets by public participation. The project realises the importance of mapping EaR and how it can be exceptionally supportive in recognizing individuals trapped in locating relief camps and recognizing how supplies can be utilized when buildings, bridges, and roads are flooded (Kerala State Spatial Data Infrastructure, 2021). It caters towards mapping footprints and adding relevant information about the EaR. However, the relevant open data are not available online yet due to the time that is required to generate them over different cities. Otherwise, it could have been a very good source of validation for the thesis methodology. Moreover, this also brings a challenge in rapid mapping of buildings EaR for emergencies. Unlike this project, which has been in the works for the past few months, the thesis wants to address the rapid mapping of buildings EaR by quickly developing buildings EaR database in data-scarce regions that can be used for emergency purposes like relief measures. Refer to appendix section A for further information on the time spent on each phase of the methodology of the thesis research.

Figure 5: People displaced and seeking refuge along the rivers of Palakkad. Image Source

Palakkad in Kerala, India, was chosen to answer the proposed research questions and achieve the research objectives. Palakkad (also known as Palghat) is a municipal city in the district of Palakkad in Kerala, India. Palakkad is one of the least urbanised cities in Kerala and is surrounded by tributaries of the Bharathapuzha River. Palakkad covers 26.6 km^2 of area with a population of 130,000 people (Census of India, 2011). Historically, the city was ruled by Rajas and fought off many invasions from the East India Company and its allies. Palakkad was a huge player in the two Anglo-Mysore wars against the British but ultimately ceded to the British (Shodhganga, 2019). The presence of a low mountain pass and the proximity to the major city Coimbatore made Palakkad economically very important, being one of the largest industrial hubs in Kerala. Therefore, many new projects are being set up in the city suburbs, witnessing rapid commercial and public development. Located on the western ghats of the Indian Peninsula and characterised by monsoonal rains with approximately 1216 mm of average annual rainfall, Palakkad often faces many hydrometeorological hazards like floods, landslides, and debris flow. As of August 2019, nearly 3000 people were shifted to relief camps because of rampant torrential rain-induced flood and landslides in the hilly regions around Palakkad (The Hindu, 2019).

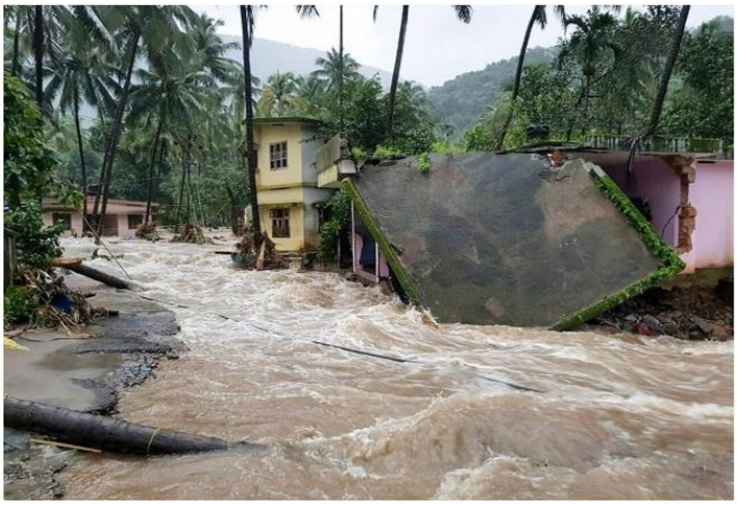


Figure 6: Destruction of buildings in Palakkad. Image Source

Figure 7: 2018 floods in Kollam, Kerala. Image Source

Moreover, to test the applicability of the proposed thesis methodology, a second study area in Kerala called Kollam will be chosen as a test site to assess the feasibility and transferability of the proposed framework. Kollam is an ancient seaport and has been a strong commercial city as early as the 9th century AD. Being an important port city, it was ruled by the Pandyas, Venads and later was influenced by the Portuguese, Dutch, and finally under the control of the British (Leela, 1986). Kollam is a fairly industrialised city known for its cashew trading and processing industry, encompassing over 34 factories and providing employment to around 26,000 workers (Raviz, 2018). Similar episodes of disasters were witnessed in the 2018 Kollam floods (figure 7), where flood waters rushed into many buildings and inundated many houses and shops, resulting in massive property damage and loss. Flood water inundated around 85.84 km^2 of the area during the 2018 floods (Lal et al., 2020). Fifty-six relief camps were set up to aid 3,600 displaced people (The Hindu, 2018). A massive lake surrounds Kollam towards the north and the Arabian Sea to the south, making it a very prone region to coastal and lake flooding during the monsoon season. Figure 8 depicts the inundation of the buildings during the 2018 flood event.

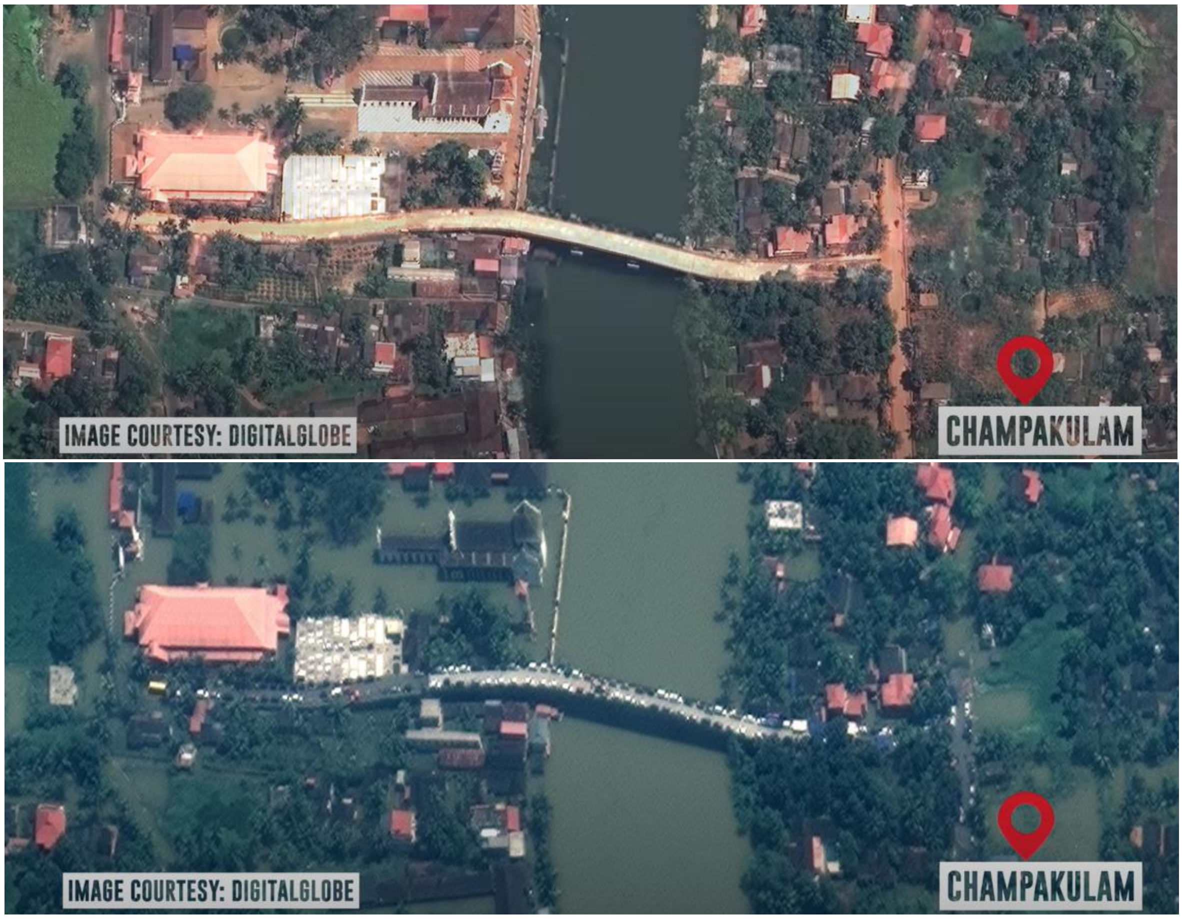
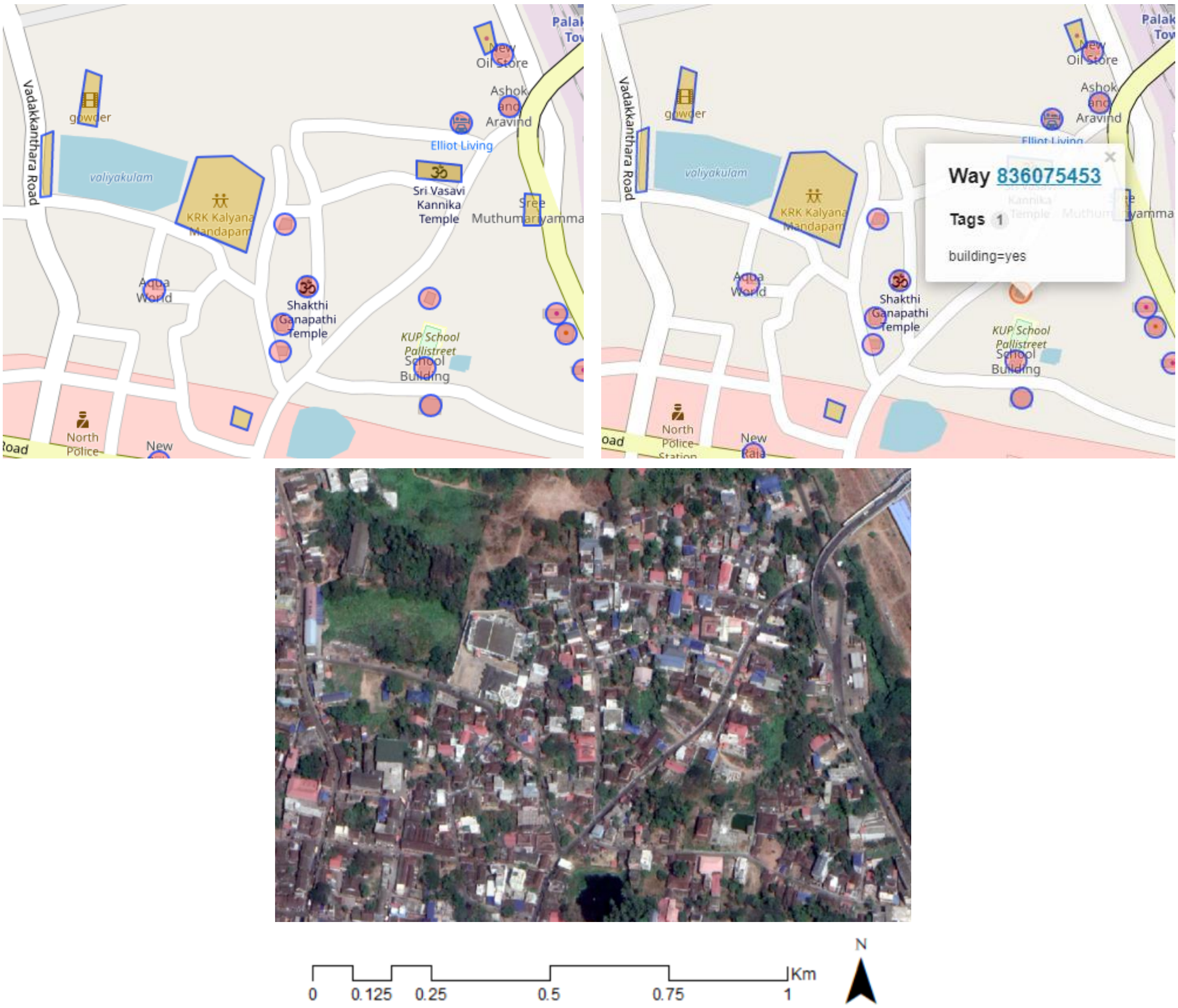


Figure 8: Digital Globe image depicting the inundation of city buildings pre (above) and post (below) 2018 flood event.

2.3 Dataset Acquisition

The dataset description covers the datasets and their respective sources in table 1. Urban land-use data was downloaded from the Indian geospatial website Bhuvan. The description of the software used in the thesis is presented in appendix section B.

The OSM dataset for Palakkad contains approximately 180 major district roads and more than a thousand building footprints exist. Such data is crucial for elements-at-risk information. However, most of the footprints lack attribute information, making it unfeasible for exposure, vulnerability, and risk assessment within the two cities, therefore emphasising and justifying the need for the research conducted.



Way 836075453

Tags 1

building=yes

Figure 9: (Top-left) OSM building footprint data, (Top-right) building attribute data, and (below) respective satellite image in Palakkad.

The figure above compares the image, the OSM building footprints, and the respective building tags to contextualise the lack of the attribute information. Also, as seen in the figure, only 20 building footprints can be seen in the top-left image, whereas in reality, there are more than 300 buildings in this area in Palakkad when seen in the respective satellite image. The figure, hence, depicts the lack of footprints and the lack of attribute information that exists in the OSM data of such regions, thus, reinforcing the need for the proposed methodology to address such issues.

Table 1: Data set description

Data	Type	Sources	Remarks	Purpose in the thesis
Satellite Imagery	Raster TIFF	SAS GIS	Google Earth Satellite image (3 bands) of November 2020.	Used for training the model and predicting on the two study areas.
Study site locations	Vector Polygon	MapCruzin	Administrative polygon of the two study areas.	Study area map generation and boundary delineation.
Building shapefiles	Vector Polygon	OpenStreetMap	Mapped building footprints.	Used as training data for the DL model.
Building labels/tags	Vector and Point Polygon	OpenStreetMap and Google Maps	Information about buildings (school, restaurants, offices etc.).	Used in the characterisation phase to aggregate the majority building functions from the building tags.
Global Urban Footprint [6]	Raster TIFF	GUF-DLR	Global urban settlement footprint 12 metre resolution.	Used to validate the predictions results of the prediction by spatially overlaying under the predicted footprints.
Landuse data	Raster TIFF and vector polygon	Bhuvan	Landuse data from the Indian geo-portal services under ISRO[7]. (Scale – 1:10000)	Used in the characterisation phase to aggregate the majority building functions from the land use information.
Susceptibility Map	Vector Polygon	KSDMA[8]	Flood Susceptibility Map of 2010. Consist of flood extent, no depth information.	Used to perform exposure assessment after characterising the buildings.

[6] The GUF data is freely available for non-commercial academic purposes.
[7] Indian Space Research Organisation
[8] Kerala State Disaster Management Authority

CHAPTER 3: BUILDING DETECTION USING DEEP LEARNING

This chapter aims to answer the first sub-objective and the respective research questions. The chapter is divided into sections of data preparation, setting up of the model, results, discussions, and the chapter summary. The chapter focuses on buildings as elements-at-risk and is associated to S-O1 in reference to figure 1. The challenges, limitations and possible suggestions of the methodology are discussed in chapter 8.

The workflow for this chapter consists of the following:

1. Obtaining and using the semi-manually labelled dataset from OSM for the study areas.
2. Establishing a ResU-Net architecture-based CNN model to predict building footprints on the satellite imageries of the study areas.
3. Transfer learning to use learnt weights from the first study area over to the second for seamless building detection.

3.1 Data Preparation

Existing building footprints were extracted from OSM using the Overpass API[9], which serves custom chosen parts of the OSM data. OSM data has been widely utilised for several applications, including land use and land cover classification studies, building and road footprint extraction (Grippa et al. 2018; Liu et al. 2020; Zhao et al. 2018) and thus, states the numerous prospects of being employed for future research projects. Although these data are sometimes not officially validated, they do provide contextual and spatial background about the buildings. With the help of the Overpass API, inputs as polygon shapefiles are derived corresponding to the buildings in the satellite images. A labelled building dataset is prepared (as vector polygons) and then used to create binary maps indicating the buildings and the rest as background. This binary mapping of the features behaves as annotations for the respective buildings, which are later used in the DL model.

3.1.1 OSM footprints and manual digitisation

The data set of buildings was downloaded from OSM for the city of Palakkad using the Overpass API. However, for DL models, large data are required to train the models to achieve higher accuracy properly. As a result, more buildings were digitised manually to increase the number of buildings to be used as training samples. The Palakkad data set contains approximately 6000 building polygons which were used for training the model. Additional 2000 training labels were manually digitised for improving the model accuracy (see section 3.2.3 for metric accuracy evaluation).

Figure 10 is an example of the manual digitisation of buildings in the city of Palakkad. Figure 10-C shows how some of the buildings are manually digitised. Table 2 refers to the number of tiles used for training and testing in Palakkad.

[9] Link: https://overpass-turbo.eu/

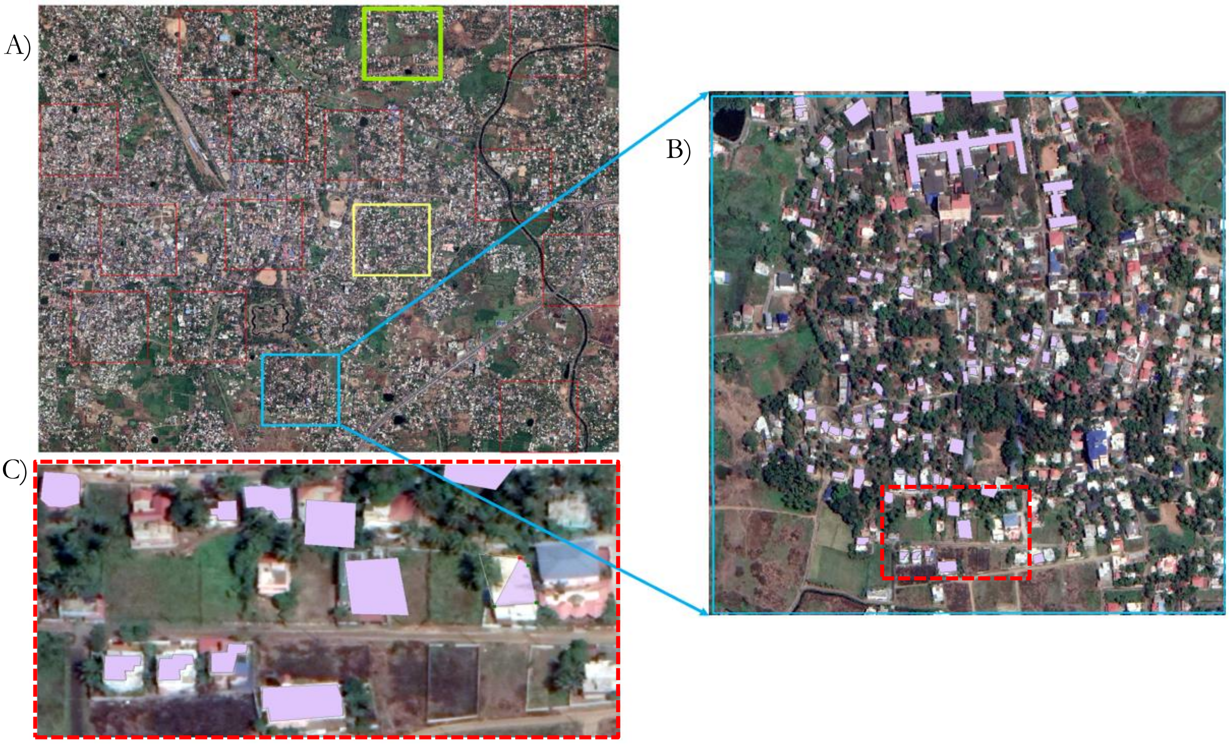


Figure 10: Example of manual building digitisation in one of the test sets in Palakkad. A) City of Palakkad with training and test sites, B) Testing tile, and C) Close-up of a few buildings manually digitised (in purple).

Table 2: Study site characteristics for training and testing sets at Palakkad.

Summary of training-testing sites	Size of tiles	Number of tiles	Number of patches
Training set	8000x8000	12	2700
Testing set	8000x8000	3	300
Total		15	3000

3.1.2 Data Preparation

A series of steps are taken into consideration to prepare the data set before training the model.

1. After manually digitising the buildings and obtaining the resultant training samples, the polygons were converted to raster images with the environment settings[10] of the satellite images (Figure 11). This step assured the spatial extent, coordinate system and cell size of the rasterised building footprints to adhere to that of the satellite images.

2. Following this, the rasterised building footprints were then reclassified as "0" and "255", where "0" indicates the non-building class and "255" indicates the building class.

3. This led to the generation of the labelled data that referred to the *building* and *non-building* classes.

[10] Settings: Maintain same coordinate reference system, mask buildings with satellite image extent, and maintain same pixel size.

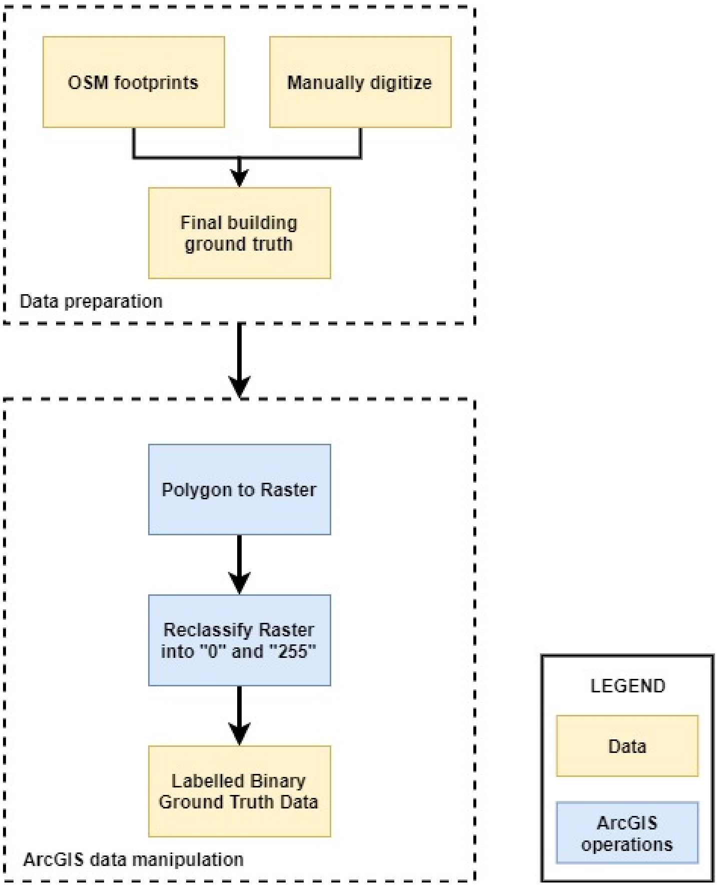


Figure 11: Data preparation steps using OSM and ArcGIS interface.

3.1.3 Data Splitting

The data set was split between training, validation, and test sets. The splitting was done over 15 image tiles for Palakkad spread strategically over the study area to encompass the complex environments where the buildings are located, which were then further patched into 512x512 sized image patches. The first 13 image tiles were further split in a 9:1 ratio, meaning 10 per cent of the image patches were used for validating the model. So, in total, 3000 image patches were used in training and validation. In figure 12, the tiling of the image into 15 tiles can be observed as an example. The remaining three tiles are used as the testing set where the model did not "see" the buildings in these three tiles, thus allowing evaluation of a truly un-seen building data assessed through the accuracy metrics described in section 3.2.3.

The red polygons are the training sites, while the blue, yellow, and green tiles are the test sites. The test sites will be used after model training to evaluate the metrics on these un-seen data before deploying the model for the entire study area of Palakkad. Similarly, the same was repeated for the second study area, Kollam. However, Kollam will also be using the learnt weights from Palakkad and re-train those weights to detect the buildings in the entirety of Kollam. More on this, in section 6.1.1.

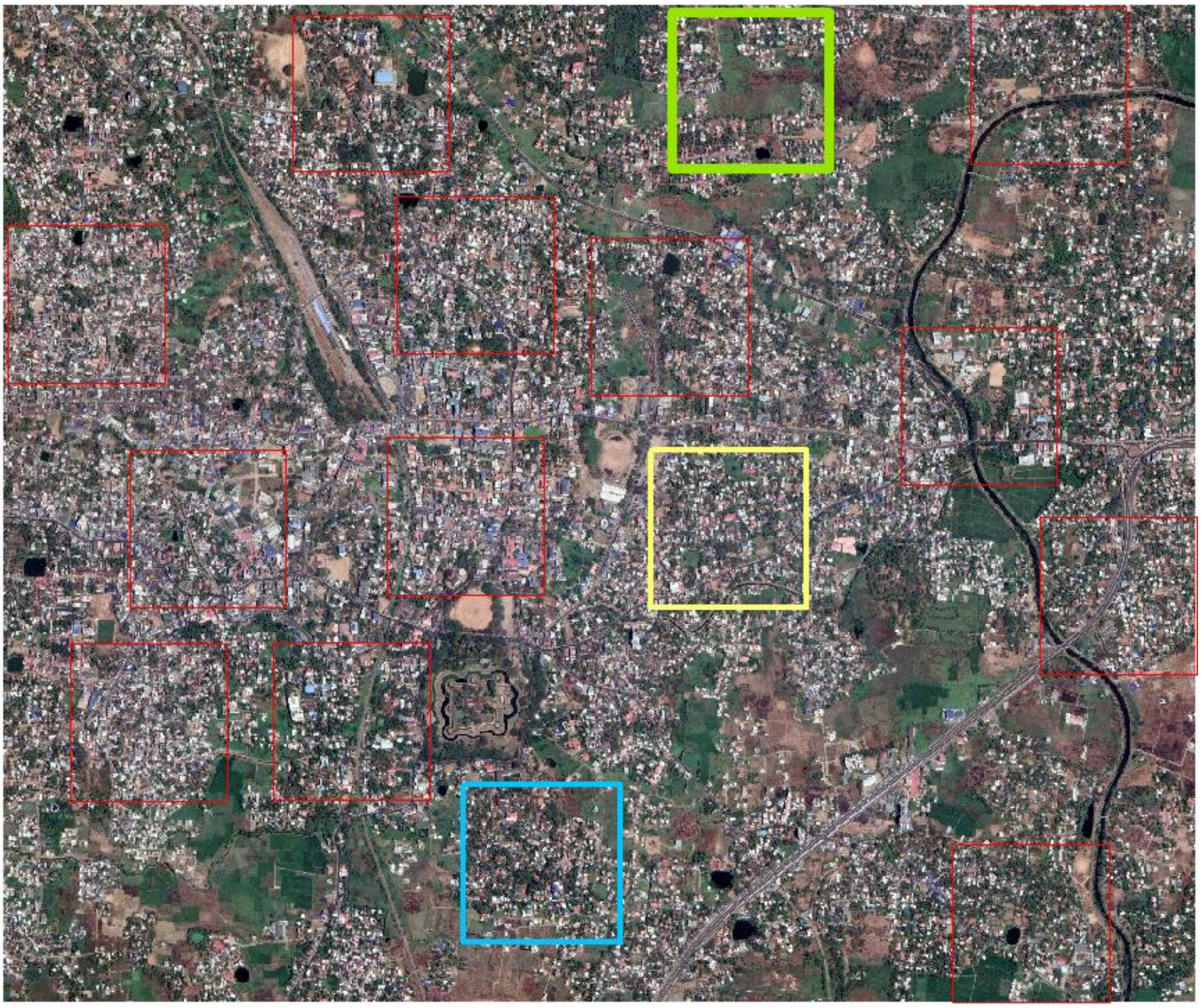


Figure 12: Example of training (red) and test (yellow, blue, and green) sites in Palakkad.

3.2 Deep Learning Model Set-Up

3.2.1 Model Architecture

The detection of building footprints in the study areas was carried out using the ResU-Net model (Diakogiannis et al. 2020) that specialises in detecting target objects with fewer training data or samples. The ResU-Net model is a semantic segmentation model inspired by the deep residual learning network (ResNet-50) (He et al. 2016) and U-Net (Ronneberger et al. 2015) that takes the advantage of both Residual network and U-Net models in achieving higher accuracies. The ResU-Net structure uses encoder-decoder parts with skip connections between them that effectively generate fine-grained segmentation results. These skip connections preserve the size of the original image and retain it in the feature maps, which makes them suitable for semantic segmentation applications. Figure 13 demonstrates the schematic structure of the ResU-Net adopted by Diakogiannis et al. (2020). Generally, the more training data is added, the better are the segmentation results (C. Sun, Shrivastava, Singh, & Gupta, 2017). Hence, from the training dataset, buildings were divided between training samples (90%) and validation samples (10%) for model training and fine-tuning, respectively. The model is then tested on three sites in the study area to test the accuracy and capability of the model.

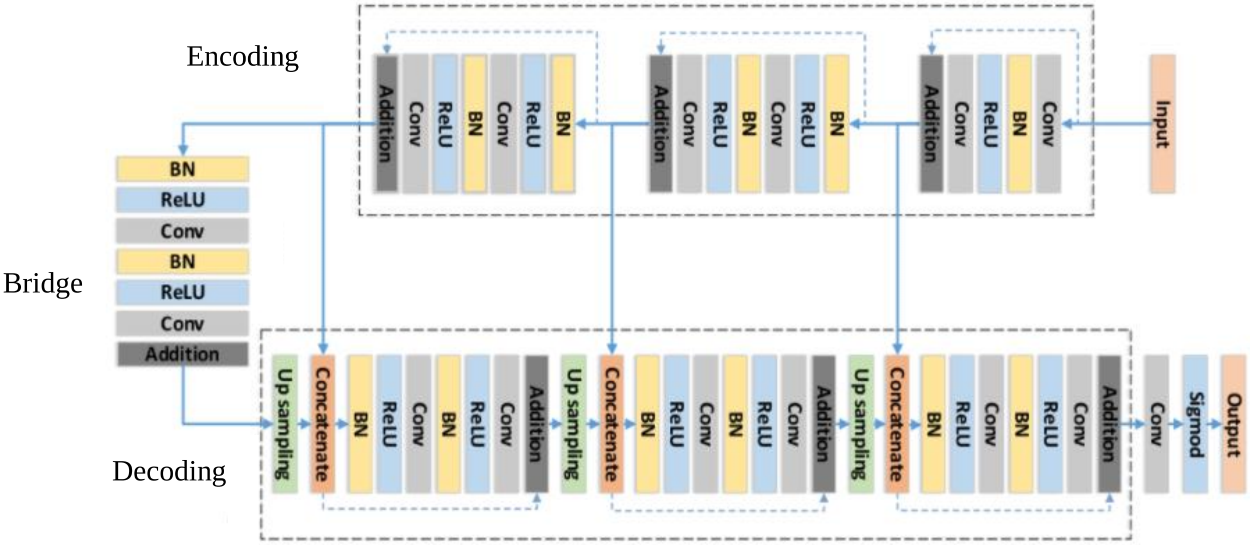


Figure 13: Schematic diagram of the ResU-Net model based on Diakogiannis et al. (2020)

The output after training is a binary classified image that distinguishes building pixels from non-building pixels. For training the model with the ResU-Net network, the whole process was run on an NVIDIA P100 GPU (16 GiB VRAM) and 25 GiB of RAM on Google Colab[11]. An example of the code to call the ResU-Net model is given in appendix section C. Please refer to that for further understanding and insight.

In the experiments, Adam optimiser was employed instead of the conventional Stochastic Gradient Descent optimiser as suggested by Bottou (2010) and Pan et al. (2020). The former is much faster because of its adaptive learning capability and converges faster to reduce the loss, ultimately increasing the overall accuracy. *Learning rate*[12] and *weight decay*[13] values were used to optimise the training speed and offset overfitting the network model. The output of this step would generate heat maps consisting of probability values belonging to classes: "buildings" and "non-buildings". The list of the so-called hyper-parameters that were used while training are as follows:

1. *Number of epochs*: The number of complete passes of training through a training dataset.
2. *Batch Size*: The number of training samples used in a pass of training.
3. *Optimisers*: Algorithms used to minimise the loss by updating parameters like weights.
4. *Learning Rate*: A hyper-parameter that controls the change in model in response to estimated error.

3.2.2 Model Tuning

Hyper-parameter tuning is one of the most important steps in controlling the overall behaviour of the model. The goal is to find an optimal combination of hyper-parameters that gives the best result by minimising the loss. In the research, two loss functions, namely the conventional *Binary Cross-Entropy* (Mannor, Peleg, & Rubinstein, 2005) and *Tversky Loss* (Abraham & Khan, 2019), were investigated. The Tversky loss has the added advantage of directly improving the Precision and Recall using the so-called *beta weights*. The Tversky loss function has two parameters, alpha and beta, that control the false positives and false negatives, respectively, ultimately influencing the Precision and Recall metrics. This parameter allows

[11] Colab allows to write and execute arbitrary Python code through the browser, and is especially well suited to machine learning, data analysis and education.

[12] Learning rate is a tuning parameter in an optimisation algorithm that determines the step size at each iteration while moving toward a minimum of a loss function.

[13] Weight decay is a hyperparameter that causes the weights of model to exponentially decay to zero to prevent the weights from growing too large.

balancing the imbalance between the data as "building" and "non-building" classes and helps reduce the loss of the model while training to achieve improvement in accuracy. Hence, differences between the Tversky and BCE loss will be also investigated. Other hyper-parameters like batch sizes, number of epochs and learning rate are also investigated to fine-tune the ResU-Net model in order to improve the segmentation/detection[14] of the buildings.

3.2.3 Model Accuracy Assessment

Using metrics of True Positives (TP), False Positives (FP) and False Negatives (FN), standard accuracy assessment like Precision, Recall, and F1 score are calculated for the results. Precision (1) indicates the proportion of buildings that are correctly identified by the proposed approach. Recall (2) is the proportion of the buildings in the labelled data that were correctly detected by the approach, the F1-score (3) is used to balance the Precision and Recall parameters. The Accuracy (4) represents all the predictions that the model got right with respect to the True Positives and True Negatives.

$$Precision = \frac{TP}{TP + FP} \qquad (1) \qquad Recall = \frac{TP}{TP+FN} \qquad (2)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (3) \qquad Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

3.2.4 Model Transferability

With the ability of automatic learning of feature representations (building features in this case) in the scenario of scarce training data, transfer learning can become very effective in transferring the learnt weights from previous studies (that is, the trained models) to newer data in different locations (Ravishankar et al., 2016). Therefore, to detect buildings in Kollam, transfer learning was used to address fewer training data. Moreover, since the building rooftop configurations (texture, shapes, colours) are similar to Palakkad, using transfer learning makes more sense than simply training from scratch with only label data from Kollam. Transfer learning also helps accomplish faster and seamless detection of buildings in new study areas with just a few training samples, thus allowing for effective transferability of the methodology in other regions. The results of transfer learning can be found in chapter 6, section 6.1.1.

3.2.5 Validation with existing settlement data from DLR

The Global Urban Footprint (GUF) data from the DLR[15] (Esch et al., 2013, 2012, 2011, 2010) is the mapped data of settlements with very high spatial resolutions (12-metre) developed by the DLR using TerraSAR-X and TanDEM-X scenes. This data is used for validation purposes after the detection of the buildings to substantiate the spatial distribution and extent of the detected buildings using the DL model.

3.3 Results and Experiments

3.3.1 Experiments with Hyper-Parameters for Fine-Tuning

3.3.1.1 The influence of the Binary Cross-Entropy and Tversky loss functions on the accuracy

As discussed in section 3.2.2, the two loss functions: Tversky loss and BCE, were investigated to see which of the two functions gave better results regarding the test set accuracies. Different weights (beta values) were experimented with for the Tversky loss to witness the varying differences in the accuracies. Batch sizes

[14] Segmentation and detection are used interchangeably in the context of the thesis.

[15] German Aerospace Centre (Deutsches Zentrum für Luft und Raumfahrt).

with the respective loss functions were also investigated to observe the best combination for optimal accuracy of the model for prediction.

Table 3: Table of BCE and Tversky loss against different batch sizes. Bold numbers are the best values.

BCE Loss					
	Batch Size	*Accuracy*	*Precision*	*Recall*	*F1-Score*
	8	0.948	0.732	0.733	0.729
	12	0.948	0.733	**0.735**	0.731
	16	**0.950**	**0.754**	0.725	**0.735**
Tversky Loss					
Beta weights	*Batch Size*	*Accuracy*	*Precision*	*Recall*	*F1-score*
0.6	8	0.95	0.73	0.74	0.73
0.6	12	0.95	0.73	0.76	0.74
0.6	16	0.95	0.74	0.75	0.74
0.7	8	**0.9536**	0.76	0.73	0.74
0.7	12	0.9535	0.77	0.74	**0.76**
0.7	16	0.95	0.78	0.72	0.74
0.8	8	0.94	0.70	**0.806**	0.74
0.8	12	0.95	0.73	0.75	0.74
0.8	16	0.95	0.75	0.74	0.74
0.85	8	0.95	0.72	0.78	0.74
0.85	12	0.95	0.75	0.74	0.74
0.85	16	0.9532	0.76	0.74	0.75
0.9	8	0.95	0.75	0.74	0.74
0.9	12	0.9531	0.76	0.74	0.74
0.9	16	0.9535	**0.784**	0.70	0.73

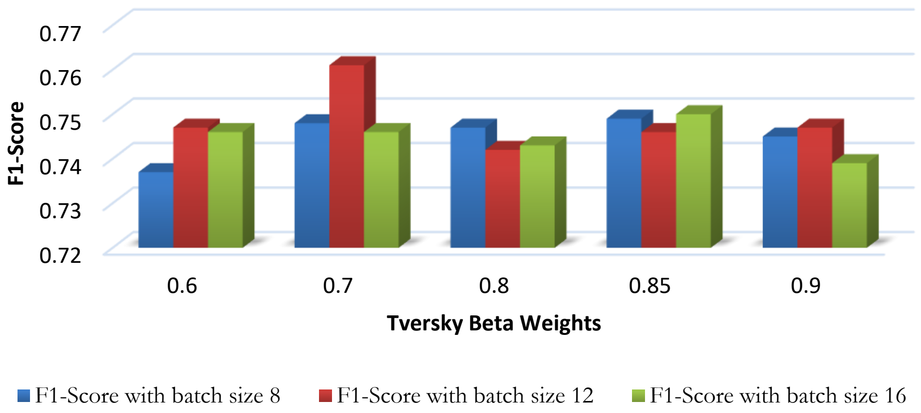


Figure 14: The effect of different batch sizes and Tversky beta weights on F1-scores.

The results from table 3 depict that the highest obtainable F1-score accuracies are 73.5% and 76.1% with the BCE and Tversky loss functions, respectively, in Palakkad. As illustrated in figure 14, a beta weight of 0.7 with a batch size of 12 generated the highest F1-score accuracy. Apart from this combination, the rest did not produce very different results. Most of the iterations with multiple hyper-parameter combinations resulted in similar results with over 72% F1 accuracy. Accordingly, based on the results observed in the table and figure 14, the best combination of a batch size of 12 and Tversky beta weight of 0.7 was considered in training the model. However, the influence of the learning rate was also important to be explored.

3.3.1.2 The influence of the learning rate on the Tversky loss

Table 4: Table of accuracies against different learning rates trained with Tversky beta weight of 0.7. Bold numbers are the best values.

Learning Rate	Batch Size	Loss	Accuracy	Precision	Recall	F1-score
1e-3						
	8	0.233354	0.952683	**0.808754**	0.682508	0.734267
	12	**0.231113**	**0.954026**	0.7877	**0.722834**	**0.748414**
	16	0.254153	0.952634	0.765052	0.720497	0.738014
1e-4						
	8	0.500679	0.94818	0.696967	**0.785013**	0.734116
	12	**0.433303**	**0.952269**	**0.759382**	0.730761	**0.740742**
	16	0.515036	0.948331	0.755611	0.684054	0.711797
1e-5						
	8	0.789249	0.94511	0.669366	**0.821011**	**0.734627**
	12	0.79551	0.940987	0.653644	0.789463	0.711991
	16	**0.724885**	**0.948133**	**0.73289**	0.715134	0.720403

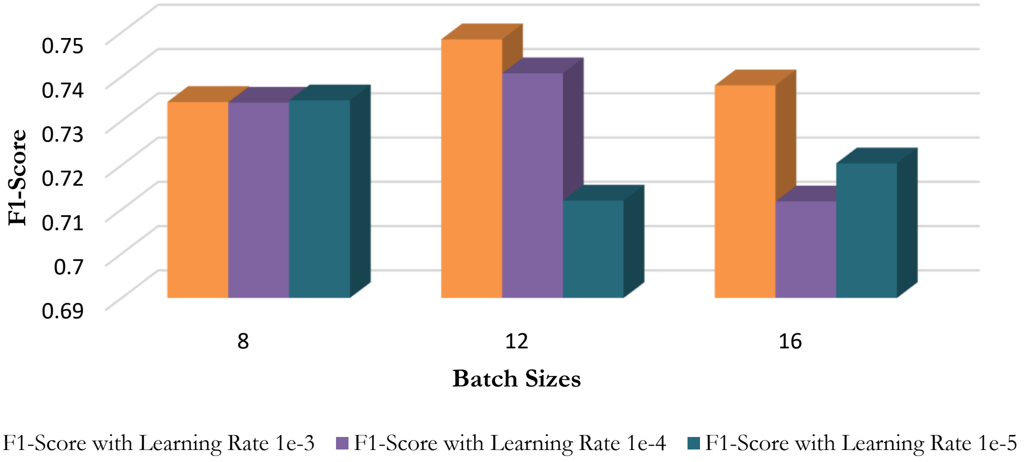


Figure 15: The effect of learning rate on F1-score with different batch sizes.

Table 4 shows the variety of results as an influence of the learning rate after being trained with a Tversky beta weight of 0.7. Based on the inference from table 3 and figure 14 discussed earlier, beta weight 0.7 was chosen for further experimentation and hence with this weight and batch sizes of 8, 12 and 16, the influence of the learning rate was inspected.

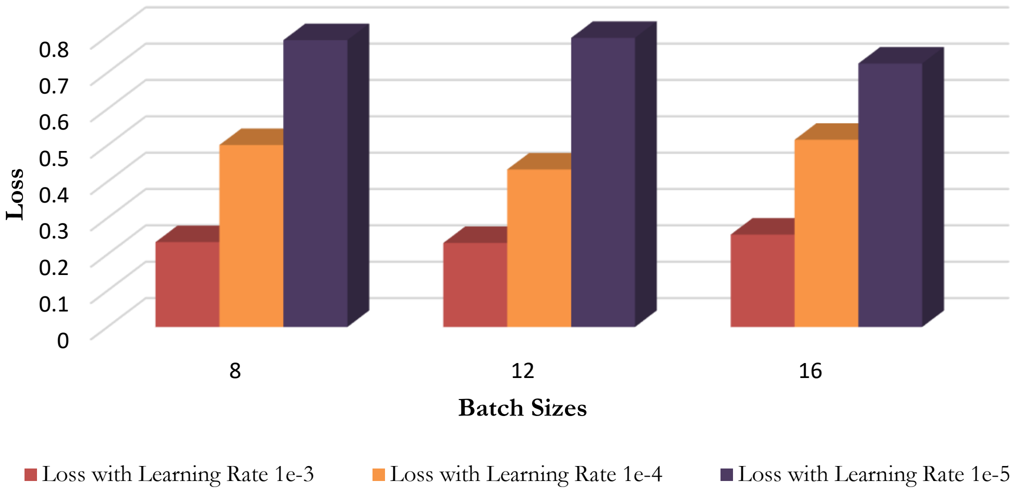


Figure 16: The effect of learning rate on the loss values with different batch sizes.

Figure 15 shows that with a batch size of 12 and a learning rate of 1e-3, the highest F1-score was achieved. The same can be seen in figure 16, where the lowest loss is with a learning rate of 1e-3 and batch sizes of both 8 and 12. The loss values are higher when trained with the lower learning rates of 1e-4 and 1e-5. The same can be seen in table 4, where the loss values are consistently above 0.5. Therefore, the learning rate of 1e-3 and batch size of 12 was chosen for final training in Palakkad (see table 5 for final hyper-parameter combinations).

Table 5: List of final hyper-parameter combination used for final training.

Hyper-parameters	Record	Remarks
Number of Epochs	25	The F1-score seems to remain relatively close to 76% even after training for more than 25 epochs. Also, training takes a long time, with 45-50 minutes per epoch.
Learning Rate	1e-3	The learning rates of 1e-4 and 1e-5 did not improve the F1-score metric significantly, while 1e-3 showed the best results.
Optimiser	Adam	See explanation in section 3.2.2.
Loss function	Tversky loss with a beta weight of 0.7	Tversky loss proves better in all metrics as it is better at addressing unbalanced classes.

3.3.2 Building Detection over Palakkad

After the model is trained on the first study area (Palakkad), the obtained weights were then used to detect buildings in the entire area of Palakkad using the TensorFlow API. In order to reduce the impact of boundary artefacts on the predictions, a sliding window method with a stride of 24 was applied to generate overlap images over each 512x512 sized patches, and the predictions of the said overlapped images were averaged to obtain the final segmentation results. Post-classification was performed to clean out multi-polygons and false-positive predictions. The final predictions over Palakkad can be seen in figure 17. As seen in figure 18, we observe that the predicted buildings of Palakkad overlay appropriately over the GUF map, also adhering to the spatial patterns and distribution of the buildings over the GUF. Moreover, as the learned weights contain the “knowledge” gained from the first study area, the same weights were used to predict buildings on the second study area (Kollam) (see chapter 6 - section 6.1.1).

Furthermore, there are predictions observed in the background of the GUF data, which are not false positives and are discussed more in the discussion section 3.4.2. The differences can be seen properly in figure 19.

Table 6: Summary table of final accuracies on the test set for Palakkad.

Metrics	Scores	Scores (in %)
ACCURACY	0.9535	95.35
PRECISION	0.779	77.9
RECALL	0.747	74.7
F1-SCORE	0.761	76.1

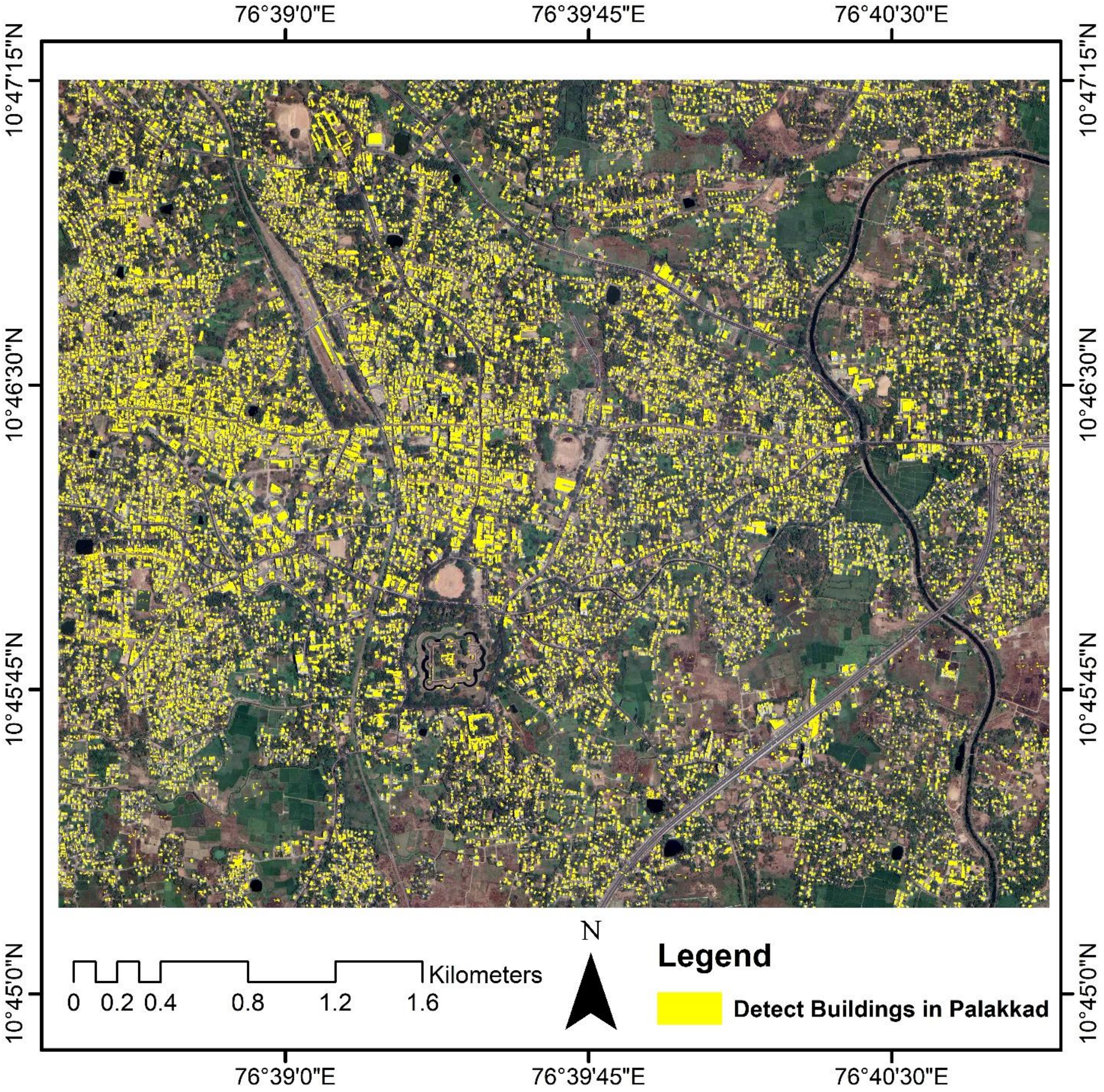


Figure 17: Detected buildings over Palakkad using the ResU-Net model.

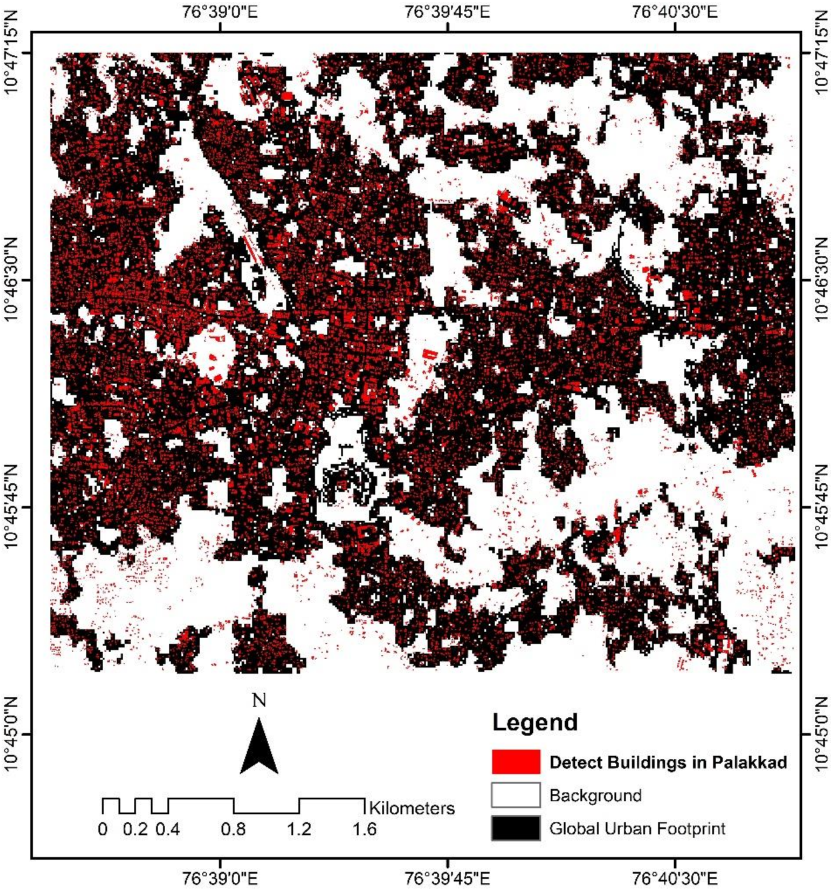


Figure 18: Overlay of the detected buildings with the Global Urban Footprint over Palakkad.

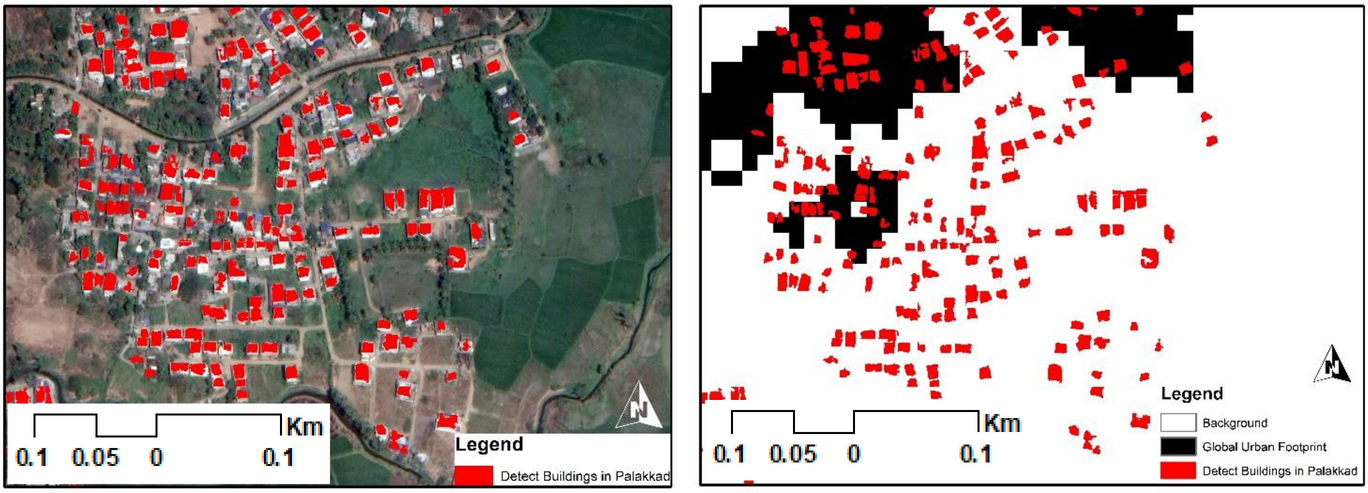


Figure 19: Difference in the predicted data from the GUF-DLR data with the recent satellite image as reference.

3.4 Discussion

3.4.1 Hyper-Parameters

Based on the hyper-parameter combinations, it was realised that the weights from the Tversky loss function influenced the model results. A key issue is the imbalance of the target objects and background. Using weighted loss functions, such as the Tversky loss, we can force the model to emphasise learning the target building pixels when in general, the target pixels occupy a very small proportion in the entire image (Lin et al. 2017). In the table in section 3.3.1.1, the accuracy metrics are almost very similar to each other ranging between 72 and 76 per cent F1-score. However, as we use the Tversky loss instead of the BCE loss, the F1-score increased by 2%. Also, as we increase the beta weights in the Tversky loss, the F1-score starts increasing, thus improving the overall accuracy slightly. But the accuracy peaked with the beta weight of 0.7 as increasing the beta weights further did not help achieve higher accuracies. The reason for this is because, with higher beta weights, the model could not properly learn from the training data, and as a result of adding higher weights on the false positive, the loss value plateaus while training, thus affecting the overall accuracy and hence, does not improve over time. Adding higher weights does not necessarily mean that the imbalance in class will be addressed linearly even when the Precision increases, as lower Recall brings the F1-score down, resulting in an overall lower F1-score over different batch sizes. Hence, a combination of the hyper-parameters that gave the best balance between Precision and Recall was considered for the final training.

Furthermore, the poor performance of loss values with lower learning rates of 1e-4 and 1e-5 shows that such lower learning rates are not good at updating the learnt weights and cannot optimise the training properly with the present data. Lower learning rates can cause the updating of the weights to get stuck as training will progress very slow due to tiny updates to the weights in the neural network. This process ultimately decreases the overall ability of the model to train optimally and hinder the model's potential performance in attaining higher accuracy.

3.4.2 Overall Prediction Remarks

There were many false positive predictions over Palakkad. This was because many buildings exhibit similar spectral information because of the similarities of many building rooftops with the road infrastructure. Thus, it explains the low Precision scores witnessed in the summary table (table 6). However, with careful post-classification clean-up of the false positives, these non-building artefacts can be easily removed. The near-perfect overlay of the predicted buildings over the GUF also gives a sense of confidence in the model and its output, and thus, from this point on, it was decided to move to the next part of the research, which is the characterisation of the detected building footprints. There were some detections observed in the background of the GUF, but these detections are not false positives; rather, newer buildings constructed recently as the GUF data was generated using TerraSAR-X and TanDEM-X scenes of the years 2013-2014 (see example in figure 19) and hence, did not include these new buildings. The transfer learning approach allowed gaining knowledge from Palakkad and using the learnt weights to recognise similar buildings present in Kollam (discussed more in chapter 6).

3.5 Chapter Summary

In this sub-chapter, we saw the review of many works of literature and research that delved into the investigation of various approaches to detect buildings for several purposes. By learning lessons from the literature, the ResU-Net model was chosen due to its sheer advantages compared to the traditional patch-based CNNs, emphasising using the state-of-the-art Tversky loss function that helps curb imbalance in data in model training. The sub-chapter also dealt with data collection and preparation for training, validation, and testing purposes. Later, the use of the ResU-Net DL model was discussed to detect buildings using openly available data from OSM as ground truth (with few manual digitisations). The results of the evaluated metrics on the test sets show that by fine-tuning, the best hyper-parameter combinations were chosen to finally detect the buildings over the entire study area of Palakkad. The summary table in section 3.3.1.2 shows the final accuracies of the model over Palakkad, after which the next step of characterisation of the buildings will be investigated. The results here show the possibility of a streamlined automatic process of building detection. With the completion of this chapter, the first sub-objectives are achieved.

CHAPTER 4: URBAN MORPHOLOGY METRICS AND HOMOGENISATION OF BUILT-UP AREA

This chapter aims to accomplish the second sub-objective of developing a methodology (S-O2 in figure 1) to sub-divide the built-up areas based on the characteristics of the building footprints within them. The chapter is divided into sections of data clean-up, building morphological metric evaluation, built-up area homogenisation, results, discussions, and the chapter summary.

4.1 Data clean-up

After the buildings are detected for Palakkad, the next crucial step is to clean the buildings that can cause issues with the methodology moving forward. For the morphological calculations, it is imperative not to have multi-and overlaying polygons that could potentially affect the calculation of the morphological metrics detrimentally. Hence, buildings under 25 sqm (square meters) were removed along with false positives that were detected on certain road infrastructures and overall cleaning of multi-and overlaying polygons were removed. Figure 20 shows the clean-up process of the data from the detection phase by using ArcGIS tools.

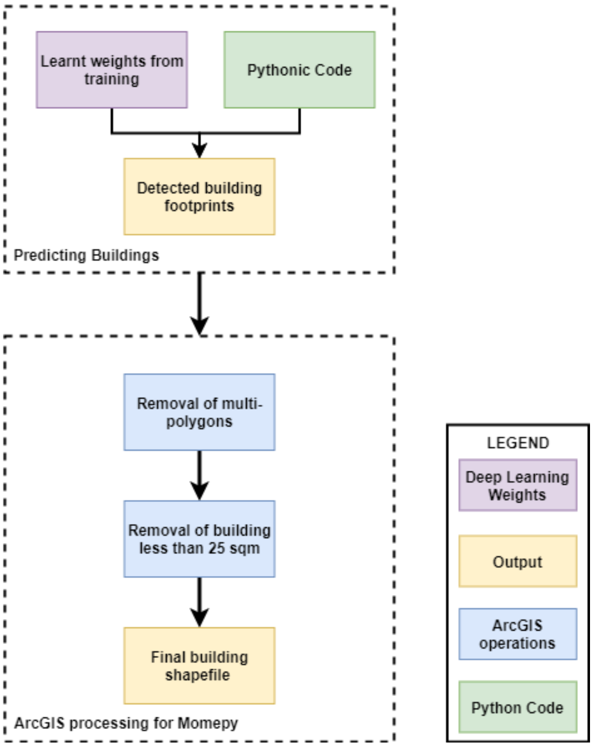

Learnt weights from training
Pythonic Code
Detected building footprints
Predicting Buildings
Removal of multi-polygons
Removal of building less than 25 sqm
Final building shapefile
ArcGIS processing for Momepy
LEGEND
Deep Learning Weights
Output
ArcGIS operations
Python Code

Figure 20: Flowchart for data clean-up using ArcGIS operations.

4.2 Building morphological metrics using Momepy

4.2.1 Measuring building morphology

After cleaning the building footprints, the next step is to employ spatial urban morphological metrics to understand the physical morphology of the buildings. The Momepy package is a library that allows quantitative analysis of urban form and morphometrics (Fleischmann, 2019). Among the many metrics available in the library[16], twenty-two morphological metrics were used to identify the spatial relationships and morphologies of the buildings to themselves and the adjacent surroundings. As the contextual configuration highly impacts building types among the building footprints (Fan et al., 2014), the spatial morphological analysis is conducted based on two important hypotheses:

a. Buildings would share attributes like building type, provided their footprints are of similar shape and size.
b. Buildings of similar footprints that are located close to each other would share attributes.

The Momepy library is installed using Python's Conda package management system. The core functionalities[17] of the library allow measuring the morphology of two major urban elements: buildings and streets that include building shapes, dimensions, density, diversity, and much more explained in table 7. Refer to appendix section D for the sample code on the selection of the morphological metrics.

[16] Link: http://docs.momepy.org/en/stable/api.html
[17] Link: http://docs.momepy.org/en/stable/index.html

In figure 21, we see two examples of the morphologies with the metrics *Cover Area Ration* and *Simpson's Diversity of Area*. Table 7 describes the twenty-two morphological metrics and their respective functions as definitions used in the methodology to categorise the buildings based on their physical and surrounding environmental conditions.

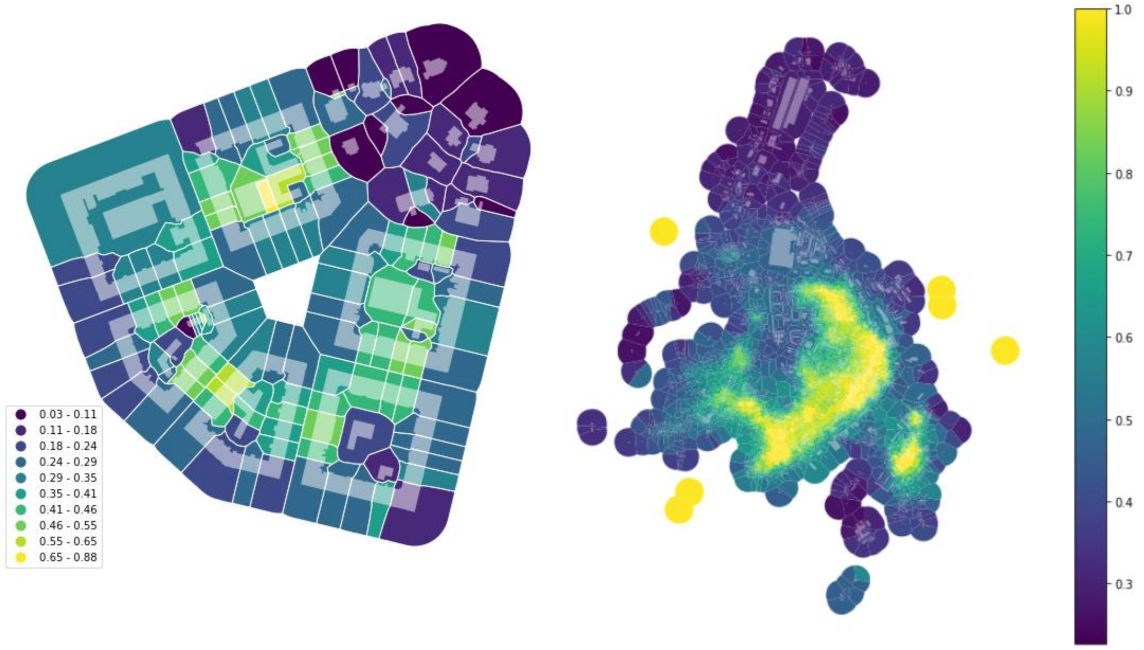


Figure 21: Examples of morphological metrics. (Left) Cover Area Ration and (Right) Simpson's diversity of area.

Table 7: List of urban morphological metrics used in the research. Refer website (Fleischmann, 2019).

Serial No.	Urban Morphological Metrics	Definition
	Dimensions	
1	Area	Calculates area of each building
2	Perimeter	Calculates perimeter of each building
	Shapes	
3	Corners	Calculates the number of corners of each building
4	Squareness	Calculates squareness of each building
5	Equivalent Rectangular Index	Calculates equivalent rectangular index of each building
6	Elongation	Calculates elongation of buildings seen as elongation of its minimum bounding rectangle.
7	Centroid Corners Mean	Calculates mean distance from the centroid
8	Centroid Corners Standard Deviation	Calculates the standard deviation of distance from centroid

	Spatial Distribution	
9	Orientation Blocks	Calculate the orientation of buildings as blocks
10	Orientation Tessellation	Calculate the orientation of buildings as tessellations
11	Cell Alignment	Calculate the difference between cell orientation and orientation of the building
	Tessellation Dimensions and Shapes	
12	Longest Axis Length of Tessellation	Calculates the length of the longest axis of the building
13	Area of Tessellation	Calculates area of each building tessellations
14	Circular Compactness of Tessellation	Calculates compactness index of each building tessellation
15	Equivalent Rectangular Index of Tessellation	Calculates equivalent rectangular index of each building
	Intensity using Queen Spatial Weights	
16	Area Ration of Tessellation	Calculate covered area ratio or floor area ratio of buildings
17	Alignment	Calculate the mean deviation of solar orientation of buildings on adjacent cells from other building
18	Neighbours	Calculate the number of neighbours captured by spatial weights
19	Neighbours Distance	Calculate the mean distance to adjacent buildings by spatial weights
20	Covered Area	Calculates the area covered by neighbours
21	Mean Inter-Building Distance	Calculate the mean interbuilding distance
22	Building Adjacency	Calculate the level of building adjacency

4.2.2 Homogeneous morphological patterns through clustering

After calculating the metrics of each individual buildings, the next step is to classify the buildings into *clusters*, where each cluster would contain information about the physical form of the buildings learnt from the morphological analysis. The clustering is achieved through K-Means unsupervised classification that allows the partitioning of the morphological metrics of the buildings into *k* clusters in which each observation belongs to a cluster with the nearest *mean* (or cluster centroid). The Silhouette score method[18] is then performed to determine the optimal number of clusters in the data set. However, the Silhouette score will only be considered when it is impossible to communicate with local experts or stakeholders for validation. The Silhouette scoring is taken as a reference in the absence of local validation or lack of communication. The selection of the number of clusters should be considered based on the input from the local stakeholders. An example of the code to perform the clustering is given in appendix section E.

[18] The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters (separation). (Marutho, Hendra Handaka, Wijaya, & Muljono, 2018).

4.3 Built-up area homogenisation

4.3.1 Homogenisation using road networks

The next step is to improve the homogenisation from the existing clusters by incorporating a refined unit of built-up area extent. This was done to address two issues:

(1) in the absence of administrative units like wards or census tracts, and
(2) choosing clusters as homogeneous units would mean under-approximating the morphological types of buildings when tackling smaller cluster numbers and vice-versa.

For example, considering four as the optimal number of clusters would mean that only four types of buildings (or built-up area) exist in the study area, which is practically impossible in a city like Palakkad in India. In such populous cities where there are many different types of buildings, such as long-elongated buildings, dense small buildings, dense big buildings, open-spaced buildings, and many others, choosing a small number of clusters does not make sense. Therefore, road networks in the city are used to split the study area into sixty-two *city blocks* or *homogeneous built-up area*[19] that would now inherit the majority urban morphological characteristics of the associated buildings within the blocks (figure 22) (discussed in depth in chapter 5). Using road networks as blocks for homogenising land parcels is not a new strategy and has been employed previously by Zeng et al. (2019) and Kuffer et al. (2020) for land use classification and obtained high accuracy of 83 per cent for the same. Furthermore, certain linear features like railway lines and river lines were also employed to overcome the lack of primary and secondary road networks in certain parts to derive homogeneous units.

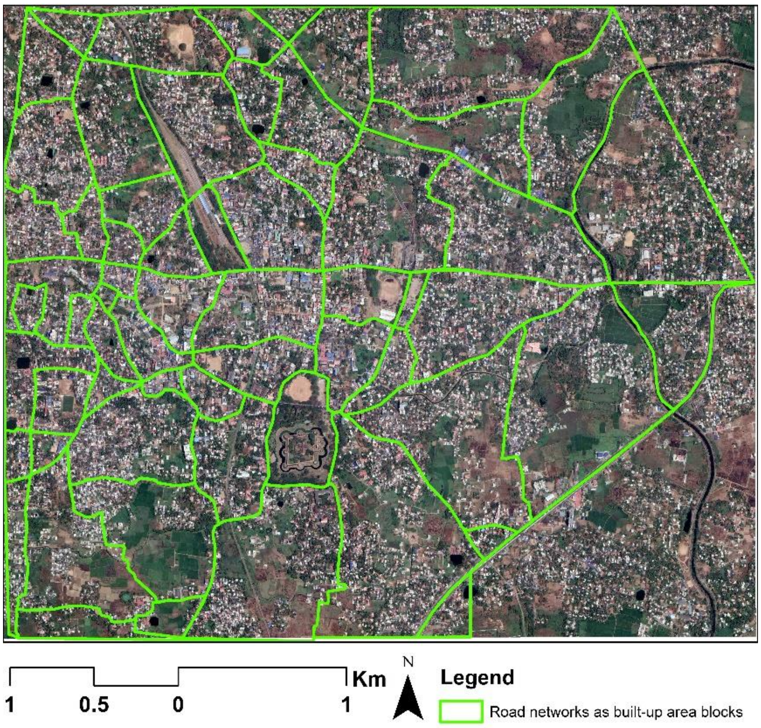


Figure 22: Road networks for built-up area blocks in Palakkad.

[19] In the thesis document, the words "homogeneous built-up area", "city blocks", and "blocks" are used interchangeably.

4.3.2 Homogeneity Score

In order to realise the homogeneity or similarity of the clusters within each block, it is essential to derive a metric that can help assess the true nature of the building morphology within each block. For example, buildings portraying morphologies of big buildings within a block, estimated to be rural, are implausible. Therefore, observing the homogeneity score can give more information about the similarity of the building morphologies present within each block. The formula of the homogeneity score is given below and is calculated at the block level. A lower percentage score can help understand what other types of buildings will possibly be inherent within that particular block.

$$\frac{\text{Frequency of first majority cluster buildings} + \text{Frequency of second majority cluster buildings}}{\text{Number of buildings per block}} \quad (4)$$

Fan et al. (2014) also used such a metric to assess the similarity between the buildings that presumably shared similar morphological attributes.

4.4 Results

4.4.1 Urban Morphological Metric (Momepy)

The study area is clusterised based on the K-Means un-supervised classification on the urban morphological metric values obtained from Momepy. Figure 23 shows the cluster classification based on k=8 clusters (chosen based on the suggestion of the local expert), thus attributing information about the types of buildings morphologically. The K-Means derived clusters for the buildings in Palakkad are described in tables 8.

Table 8: Cluster interpretation of the buildings in Palakkad after local expert validation.

Cluster Number	Interpretation
1	Large buildings associated with long corridors that are not densely located
2	Moderately sized buildings that are densely located
3	Small-sized buildings - Type 1 [20]
4	Densely packed moderate-sized buildings
5	Buildings with open surroundings
6	Small-sized buildings - Type 2 [21]
7	Moderately sized buildings sparsely located
8	Buildings with relatively fewer open surroundings

[20] Type 1 is more less regular in shape, meaning it is more circular.

[21] Type 2 is much regular in shape, having crisper boundaries.

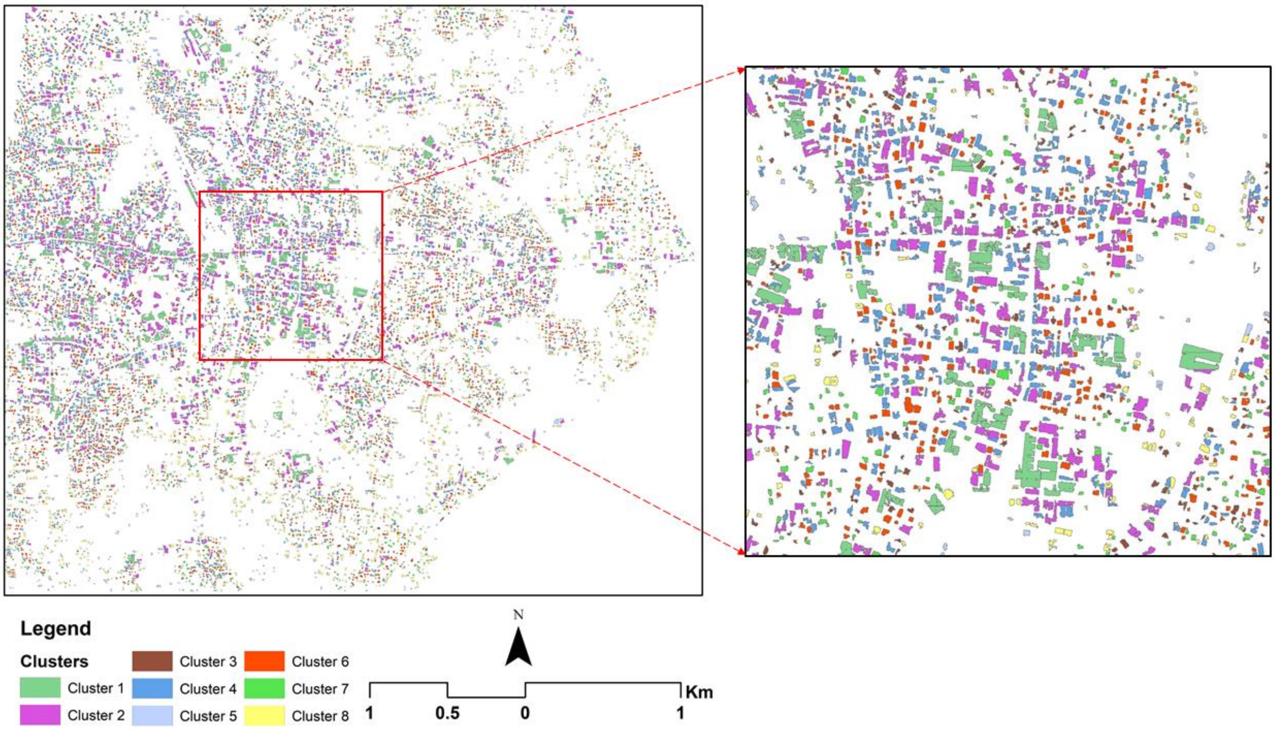


Figure 23: Morphological clusters of the buildings in Palakkad after performing K-Means classification.

In the figure above, we see the spatial distribution of the clusters of Palakkad. As an example, referring to the table above, we can visually notice that building cluster of green (cluster 1) are large buildings with long corridors and cluster 2 buildings of colour violet attribute morphologies of moderate-sized dense buildings.

4.4.2 Homogeneity Score

As discussed in section 4.3.2, the homogeneity score helps realise the homogeneity or similarity of the buildings as clusters within each block. It is essential to derive a metric that can help assess the true nature of the building morphology within each block. Based on the map from figure 24, we understand that in Palakkad, many homogeneous blocks exhibited scores of below 50%. Only 16 blocks showed homogeneity scores above 60%, meaning that buildings in these 16 blocks exhibited similar characteristics morphologically, whereas buildings in the remaining blocks are very dissimilar to each other.

4.5 Discussion

The clusters of Palakkad depict the morphological attributes of the buildings with respect to the physical characteristics of the buildings and their relationship to each other as well as to the surroundings. Many building clusters are dissimilar to each other, suggested by the low homogeneity scores in the blocks (figure 24). A reason for this can be the fact that many of the building footprints were a result of the detection from the DL model (chapter 3), and as the results are not always produced with crisp-straight boundaries, there are instances of some irregular polygonal building features that might affect the morphological metric calculation of Palakkad. Fan et al. (2014) and Qi and Li (2008) mention that the approach for using the homogeneity score help find approximately similar building footprints and are suitable for buildings that exhibit less detailed geometries. This reason explains why the scoring is less for Palakkad, where some buildings are shaped irregularly with many edges, confusing the model while calculating the morphological metrics.

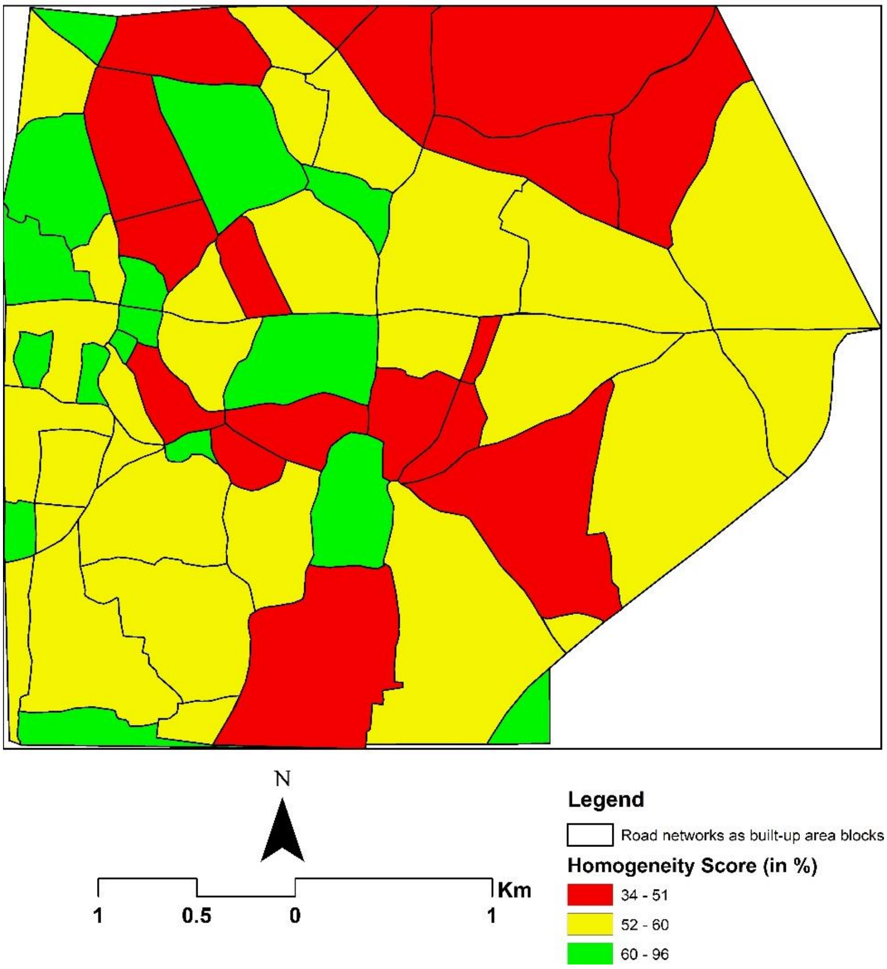


Figure 24: Homogeneity score of clusters in Palakkad.

Therefore, during the classification process for obtaining the occupancy types, care must be taken to consider the blocks with low homogeneity scores while estimating the occupancy type. Furthermore, these blocks would be subjected to further in-depth dialogue with the local knowledge experts for validation to ascertain other reasons for such low scores. Possible reasons can be due to the mixed types of buildings that are present in the blocks that are morphologically very different building types, which can effectively lower the scores. The use of other boundaries (or linear) features such as water bodies (canals, rivers), topographic differences, green zones, and others could further improve in determining the homogenous units (as done partially with railway and river lines, section 4.3.1). Another point of discussion is the size of the homogeneous units, where smaller units could be better suited for aggregating the building occupancy types. Administrative wards and even census tracts can be used (if/when available) as the smallest homogeneous units, which could probably separate the built-up and non-built-up area effectively (discussed more in section 8.1.2).

4.6 Chapter Summary

This chapter helped overcome the issues of characterising and estimating building morphology at the building level, inadvertently addressing the under-approximation of the estimated morphology. Therefore, road networks were established to homogenise the building clusters and categorise them into homogeneous built-up area blocks that would now inherit and exhibit the morphological characteristics of the buildings. Furthermore, the homogeneity of the blocks was assessed to witness which built-up areas manifested similar clusters of buildings, indicating towards the built-up areas that portray different building types morphologically. Local knowledge validation was also performed to help further refine the categorisation and interpretation of the building clusters of the homogeneous blocks.

CHAPTER 5: CHARACTERISATION OF HOMOGENEOUS UNITS WITH OPEN-SOURCE DATA

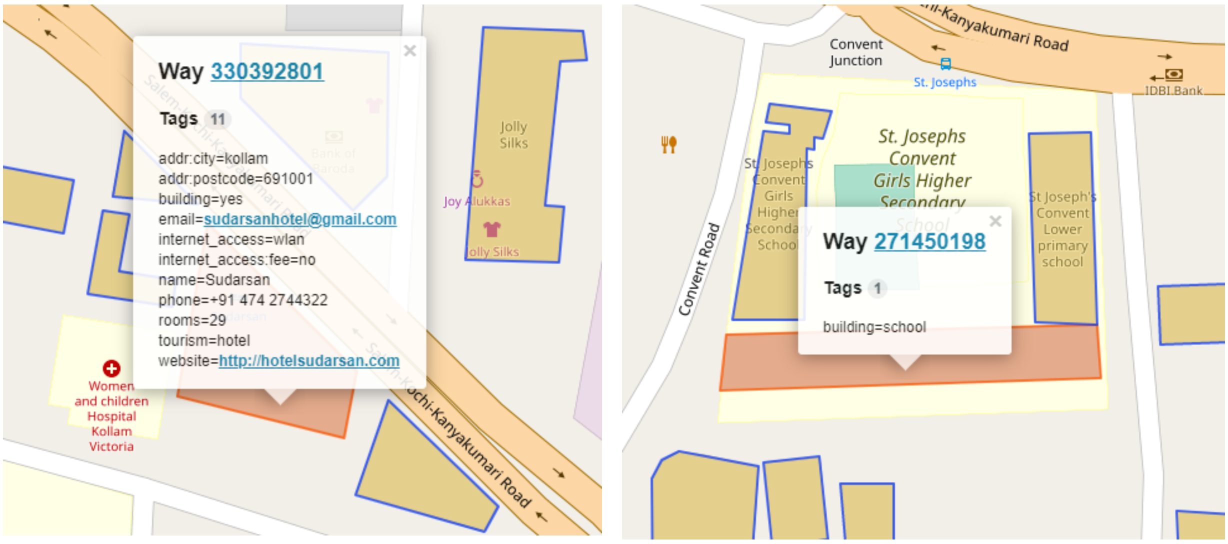
This chapter aims to achieve the third sub-objective and the respective research questions by amalgamating label information from open-source data with the building morphological metrics from the previous chapter. The open-source data are used for the *characterisation* of the built-up area (see figure 1: S-O3). Accordingly, the chapter is divided into their respective methodologies, results, discussions, and chapter summary.

5.1 Auxiliary data acquisition

In the research, mainly two main types of auxiliary data were chosen for the characterisation of the buildings. There was an intention of using free and open-source data that could be available at any point in time. OSM, Google Maps and land use maps are such data that are openly available and are of decent quality. Therefore, tags are used from OSM and Google Maps as *building tags* and *land use information* from the Bhuvan geospatial platform.

5.1.1 Building label information from OSM and Google Maps

The building labels (or tags) are used from OpenStreetMap and Google Maps. Using the Overpass API, the point and polygon data from OSM was downloaded that consisted of the building tags. Information such as shops, restaurants, offices, houses, residential apartments, commercial, recreational areas, schools, hotels, and many others (figure 25) was extracted from both point and polygon building data. Google Maps label data was manually digitised by using the Google Maps plug-in in QGIS to add more label data that were missing from the OSM data.



Way 330392801

Tags 11

addr:city=kollam
addr:postcode=691001
building=yes
email=sudarsanhotel@gmail.com
internet_access=wlan
internet_access:fee=no
name=Sudarsan
phone=+91 474 2744322
rooms=29
tourism=hotel
website=http://hotelsudarsan.com

Way 271450198

Tags 1

building=school

Jolly Silks
Joy Alukkas
Jolly Silks
Women and children Hospital Kollam Victoria
Salem-Kochi-Kanyakumari Road
Convent Junction
St. Josephs
-Kanyakumari Road
IDBI Bank
St. Josephs Convent Girls Higher Secondary School
St Josephs Convent Girls Higher Secondary School
St Joseph's Convent Lower primary school
Convent Road

Figure 25: Examples of building tag information from Palakkad.

5.1.2 Landuse data

The landuse map is a product of the urban land use survey conducted by the Ministry of Urban Development, India (MoUD) as a part of the National Urban Information System (NUIS) that generated the 1:10,000 scale urban geospatial database. The database can be seen in figure 26. The land use map is first downloaded using a Web Map Service using QGIS from the Bhuvan geospatial data platform and is then resampled to the required extent of the study area. As the data is in a GeoTIFF[22] image format, the GeoTIFF is converted to a polygon shapefile by manually digitising the image to obtain a shapefile of the landuse map.

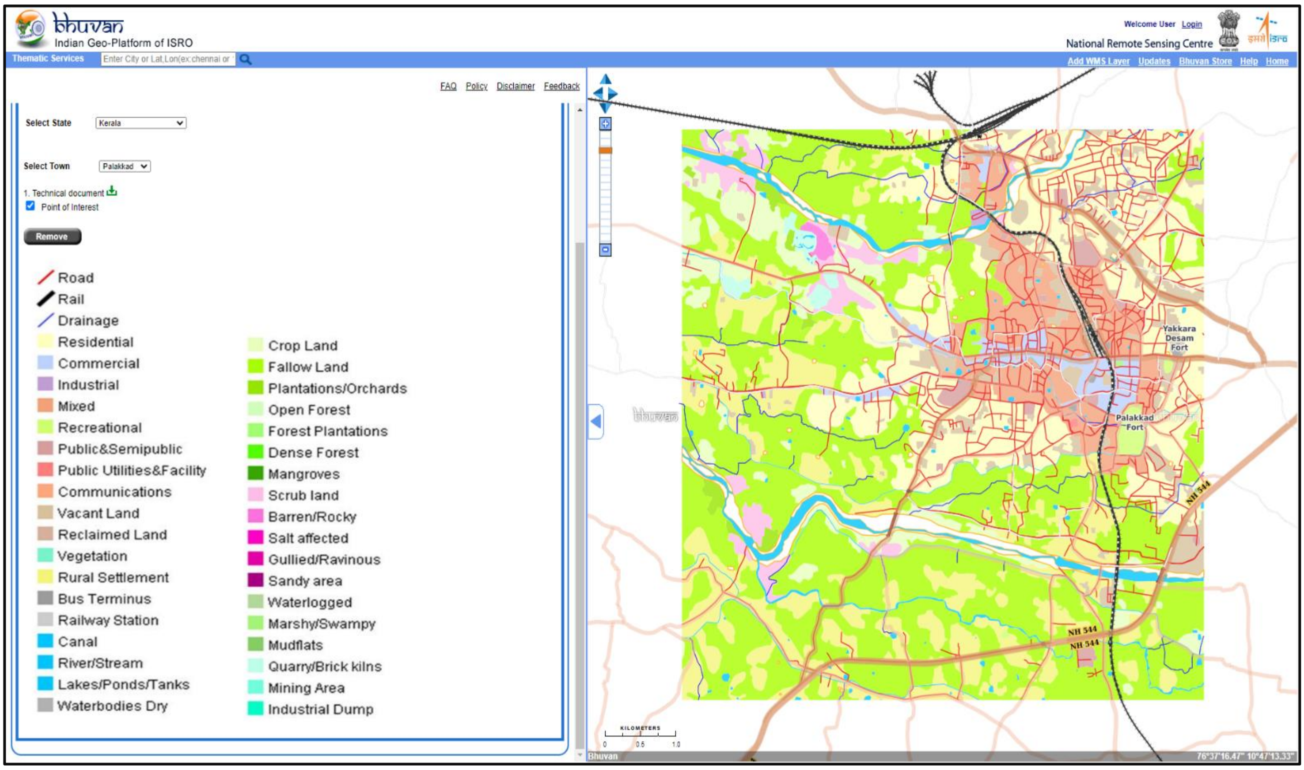


Figure 26: Bhuvan NUIS land use database for the region of Palakkad (Source).

5.2 Characterisation Strategy

The characterisation of the detected buildings with the auxiliary data is depicted in the flowchart in figure 27. Figure 27 is a schematic overview of achieving the characterisation of the buildings at a homogeneous built-up level with the auxiliary data from open sources as a proxy for estimating the *building occupancy/function type* as a typological attribute. As we reach a point of having acquired all the necessary data of *building tags*, *landuse*, *homogeneous built-up blocks,* and *morphological clusters*, we can combine them all to commence the characterisation procedure.

5.2.1 Amalgamating auxiliary data into the homogeneous built-up area

In this step, all auxiliary data are amalgamated or combined. This amalgamation is achieved in two steps:

1. Compile all auxiliary data of OSM, Google Maps and landuse, road network blocks and clustered morphology data and fuse them into one shapefile data using the Spatial Join tool in ArcGIS. This shapefile data is at a building footprint level, herewith called the *cluster-level* data. Next, the built-up area polygons made using the road networks herewith called the *block-level* data.

[22] GeoTIFF is a raster-based TIFF format image that is used as an interchange format for georeferenced raster imagery (Earthdata, 2019).

2. The data at the *cluster-level* are then used to calculate the majority cluster number, building tag, land use information, the homogeneity score (from section 3.2.5.2), and the number of buildings at the *block-level* using the shapefile key of the blocks.

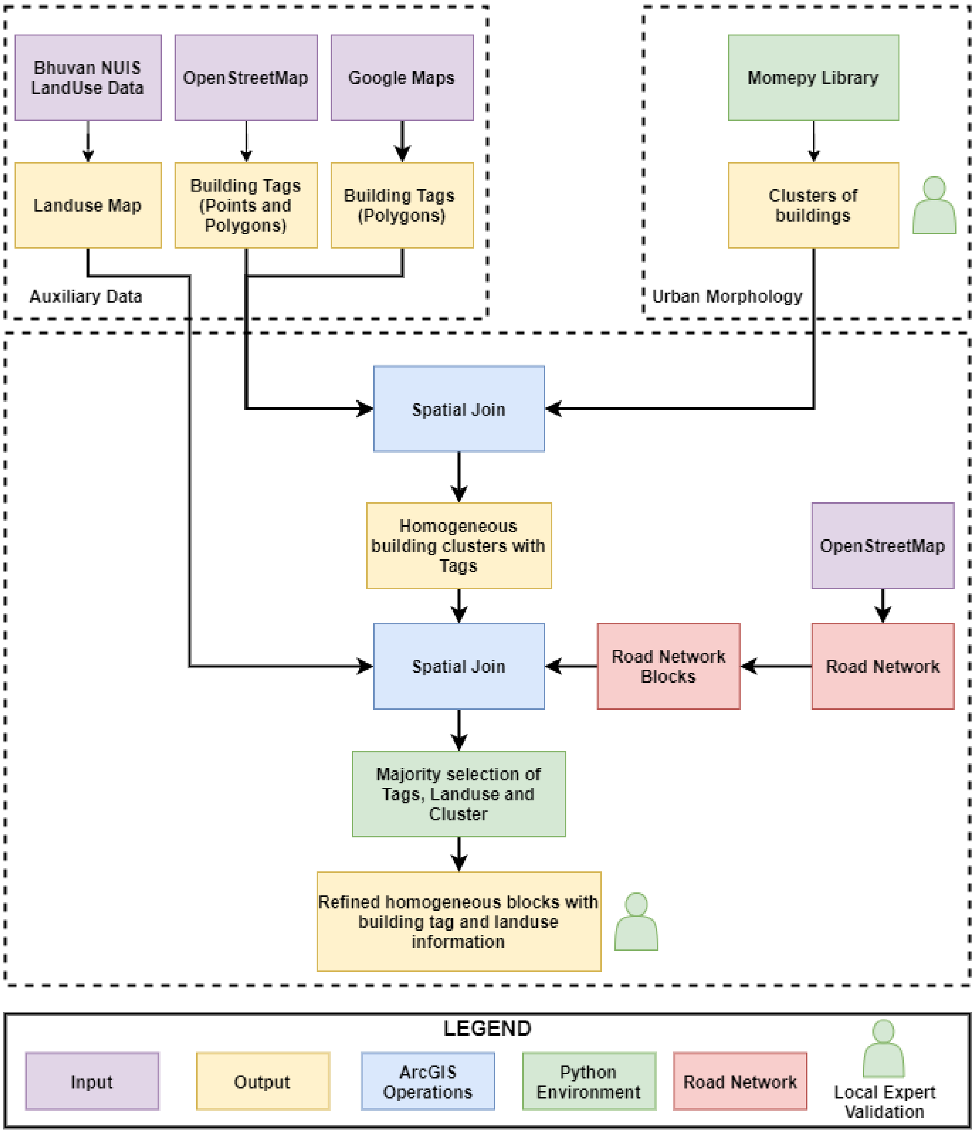


Figure 27: Flowchart of characterising buildings with data from OSM, land use maps, Google Maps, morphological metric information on the detected buildings and local expert validation.

This majority calculation is executed in a Python environment, and the code can be found in the appendix section (sections F and G).

5.2.2 Voting of the auxiliary data within homogeneous blocks for built-up area classification

A rule of *majority condition* is employed to assert the majority characteristics of the clusters to help homogenise according to the individual blocks. For example, if block number 2 has several buildings with different cluster numbers, but the majority of the buildings shared cluster number 7, then the block would be attributed and homogenised according to cluster number 7, meaning that block 2 now has the characteristics of cluster 7. Similarly, the majority information on the building tags and landuse will be attributed to the respective blocks accordingly.

After attaining the blocks with majority information of cluster number, building tags, landuse information, homogeneity score and the number of buildings, the information from these will be used to infer the classification of the blocks into semantic groups asserted to the building occupancy/function type. The voting system to achieve this can be observed in figure 28 and is described below:

1. Sort the data according to the homogeneity score.
2. Based on the scoring and the majority cluster value, the associated building morphology is interpreted with the building tags.
    a. If information from the building tags is not available, then skip step 2 and move to step 3 to use the land use information in lieu of the missing building tags.
    b. If information from the building tags is available, then use it and then move to step 3.
3. Next, the majority landuse information is used for further interpretation.
    a. If information from landuse is not available, then use the building tags from step 2 as the class label instead.
    b. If information from landuse is available, then use the information and move to step 4.
4. Classify the built-up area blocks using the inferred/interpreted building type from steps 2 to 3.
5. Sort the classified classes from Mixed-Built-up and then re-classify based on the distance from the Central Business District (CBD) or the city centre.

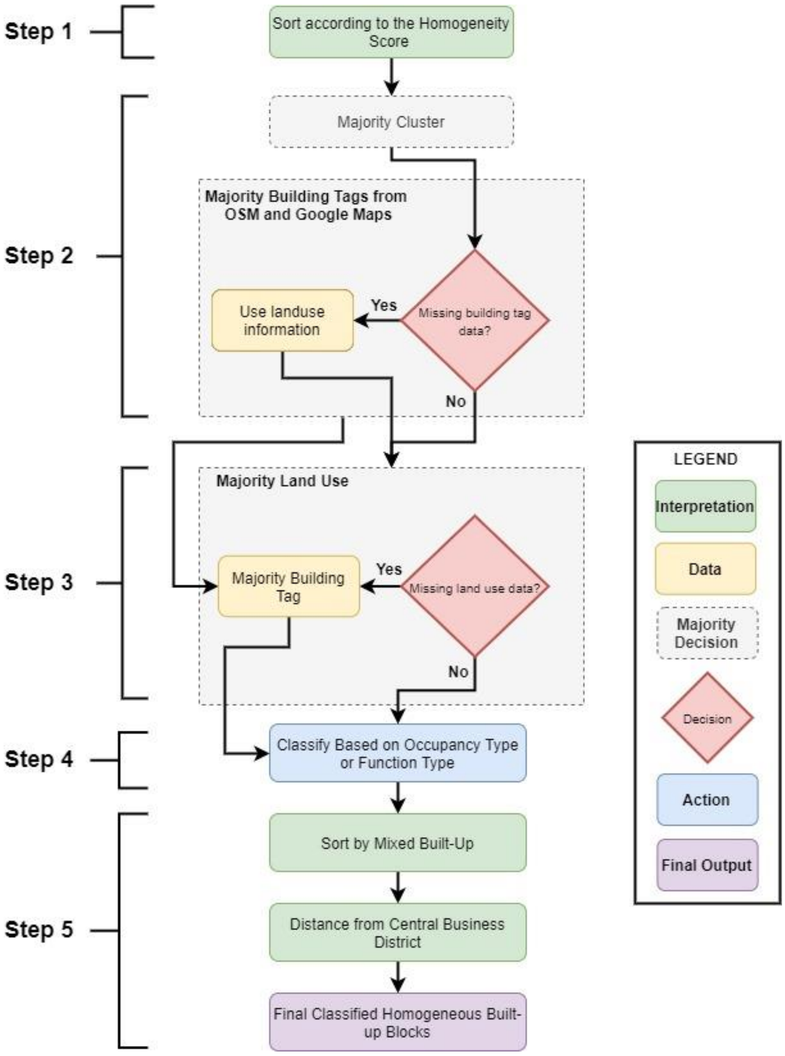

Step 1
Sort according to the Homogeneity Score
Step 2
Majority Cluster
Majority Building Tags from OSM and Google Maps
Use landuse information
Yes
Missing building tag data?
No
Step 3
Majority Land Use
Majority Building Tag
Yes
Missing land use data?
No
Step 4
Classify Based on Occupancy Type or Function Type
Step 5
Sort by Mixed Built-Up
Distance from Central Business District
Final Classified Homogeneous Built-up Blocks
LEGEND
Interpretation
Data
Majority Decision
Decision
Action
Final Output

Figure 28: Voting system for building classification based on the typology of the occupancy type.

One of the reasons as to why in step 5, the classified data is sorted and reclassified, as mentioned, is that there are no vulnerability curves for such mixed classes in the existing literature like Huizinga et al. (2017), where the vulnerability curves are generated empirically. Hence, in order to reclassify them into *Residential* or *Commercial* classes, distance from the CBD is considered. Refer to section 5.3.2 for further explanation.

5.2.3 Local expert validation

The next step is to figure out if the homogenisation with the classification essentially makes sense with what is present in reality in the study areas. For this purpose, local expert validation was performed by getting in touch with the members of the KSDMA and ICFOSS to validate the clusters and overall representation of the occupancy types of the buildings that were obtained from the methodology. It is imperative to have local knowledge to corroborate information from the clusters obtained programmatically against the reality in the study areas. Moreover, the experts can further validate and help improve the building classification into the proper building occupancy/function type. The local validation will be noted at two stages: (1) K-means cluster interpretation in the beginning and (2) final (re-) classification of the homogeneous built-up area at the end.

5.3 Results

5.3.1 Amalgamation with the open-source data

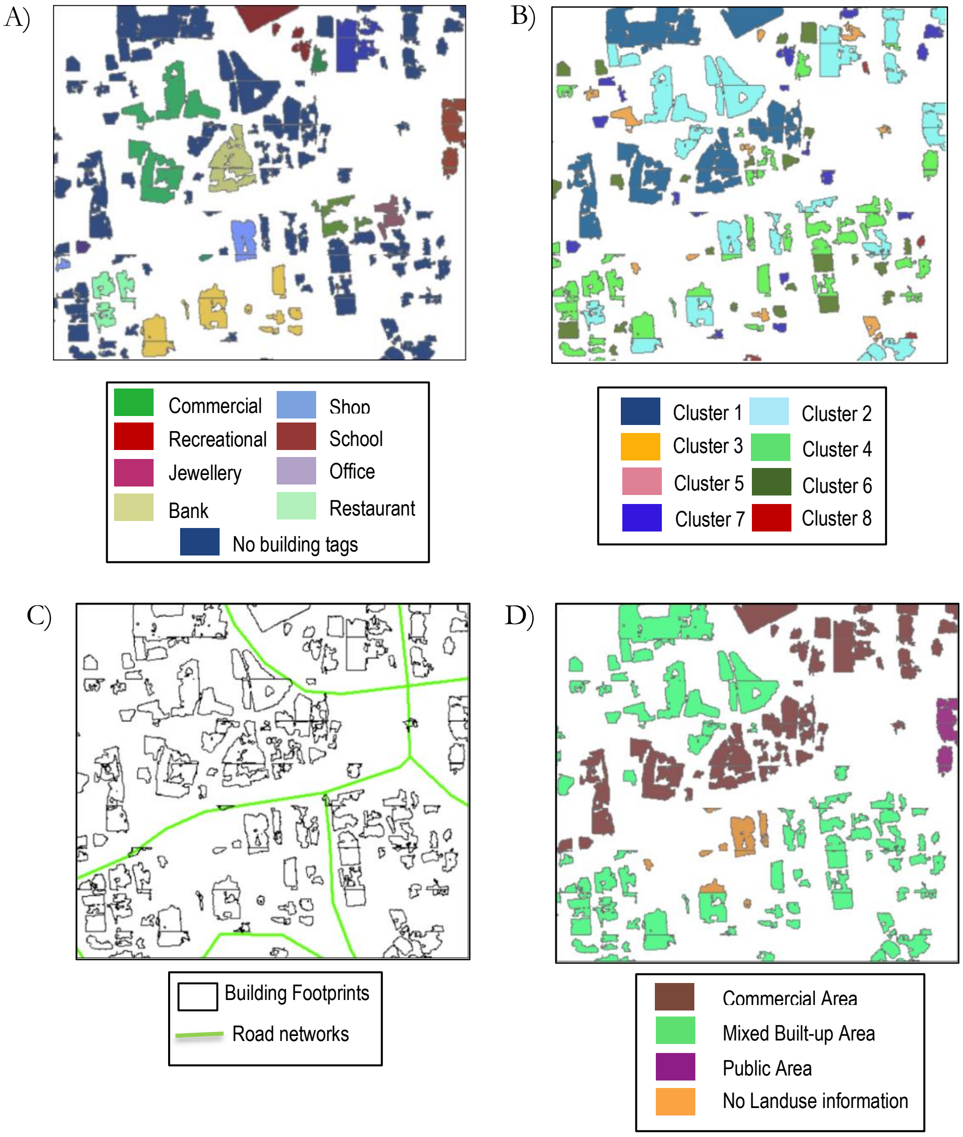
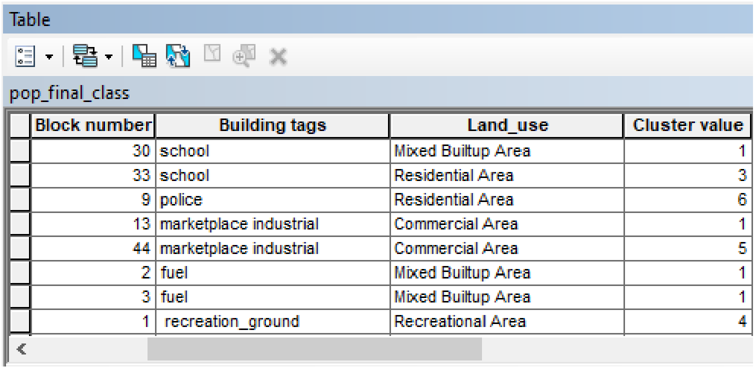


Figure 29: Example of the auxiliary data that are to be combined in the characterisation process. The data for (A) buildings tags from OSM and Google Map, (B) cluster values from Momepy, (C) landuse information, and (D) road network derived blocks.

The auxiliary data shown in figure 29 are then joined using the *Spatial Join* tool in ArcMap, which are then used to find the majority cluster values, buildings tags, and landuse information per block. Figure 30 depicts a snippet of the attribute table for the combined data.



Table

pop_final_class

Block number	Building tags	Land_use	Cluster value
30	school	Mixed Builtup Area	1
33	school	Residential Area	3
9	police	Residential Area	6
13	marketplace industrial	Commercial Area	1
44	marketplace industrial	Commercial Area	5
2	fuel	Mixed Builtup Area	1
3	fuel	Mixed Builtup Area	1
1	recreation_ground	Recreational Area	4

Figure 30: Snippet of the combined data at the block level after spatial join in ArcMap.

5.3.2 Local expert validation

Based on the recommendations and suggestions from the local expert, the cluster interpretation and final classification were adjusted to the authentic setting of the study area. Some example questions that were asked to the local experts during the clustering and final classification stages are shown in figure 31.

The initial number of road-derived homogeneous blocks was not enough to properly aggregate the amalgamated information properly, and therefore, the number of blocks was then increased based on the suggestions of the local experts. Moreover, to address the Mixed-Built-Up classes, the distance from CBD was considered for reclassification of the blocks. The classification would change from commercial,

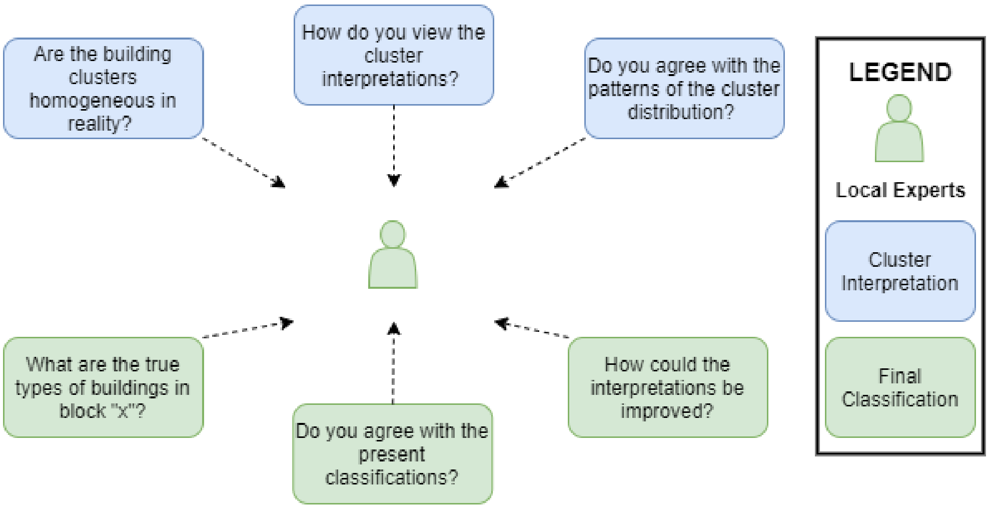

Are the building clusters homogeneous in reality?
How do you view the cluster interpretations?
Do you agree with the patterns of the cluster distribution?
LEGEND
Local Experts
Cluster Interpretation
Final Classification
What are the true types of buildings in block "x"?
Do you agree with the present classifications?
How could the interpretations be improved?

Figure 31: Schematic diagram of the local expert questioning and validation.

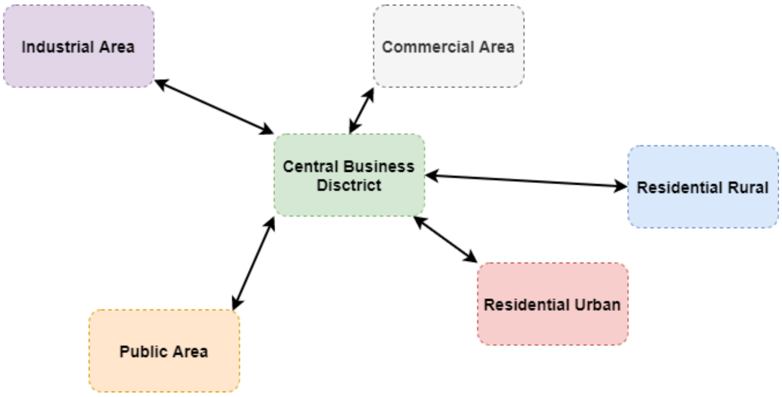


Figure 32: Distance from the CBD based classification approach.

residential urban, public, industrial, and residential rural with increasing distance (figure 32). This distance-based classification is a suggestion from the local expert who recommended using this approach in the context of the study areas.

5.3.3 Final classification of buildings after the voting system

After interpreting, validating, and finalising the results, the resulting attributes depict the building occupancy type (or function type) per block in the study area with additional information on the number of buildings and the area per block. The final classification of each block is derived based on expert-based knowledge, where the classification is based on the information from the majority building tags from OSM and Google Maps, majority land use information and majority cluster number. Figure 35 shows the final classification of the occupancy types of Palakkad. In this example in figure 34, with the help of the voting system (figure 28), the data is sorted according to the homogeneity score and then based on the majority cluster value, tags, and the landuse data, the classification as *Public Area Recreational* is decided upon in block 1. The final classification assigned as a *Public Area Recreational* is justified as buildings in such spaces are owned by the

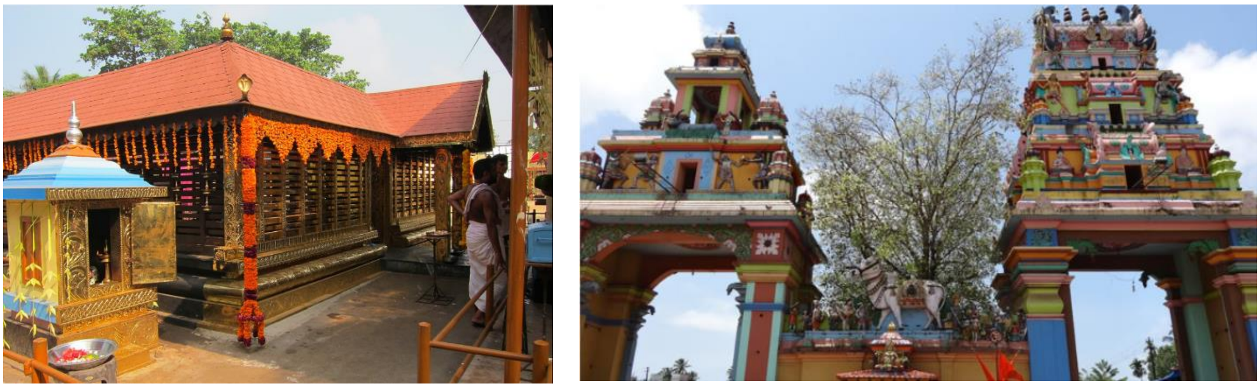
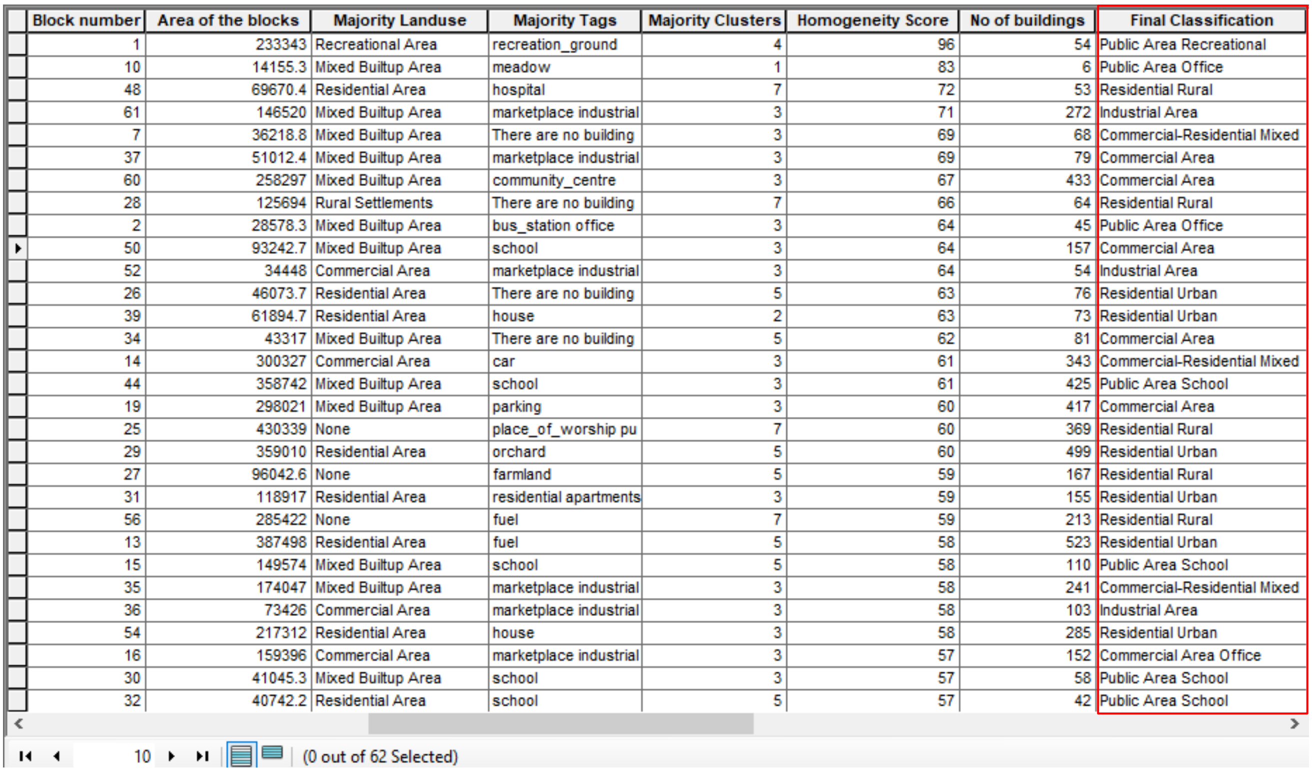


Figure 33: Reference of temple structures in Palakkad. Sources (left and right).



	Block number	Area of the blocks	Majority Landuse	Majority Tags	Majority Clusters	Homogeneity Score	No of buildings	Final Classification
	1	233343	Recreational Area	recreation_ground	4	96	54	Public Area Recreational
	10	14155.3	Mixed Builtup Area	meadow	1	83	6	Public Area Office
	48	69670.4	Residential Area	hospital	7	72	53	Residential Rural
	61	146520	Mixed Builtup Area	marketplace industrial	3	71	272	Industrial Area
	7	36218.8	Mixed Builtup Area	There are no building	3	69	68	Commercial-Residential Mixed
	37	51012.4	Mixed Builtup Area	marketplace industrial	3	69	79	Commercial Area
	60	258297	Mixed Builtup Area	community_centre	3	67	433	Commercial Area
	28	125694	Rural Settlements	There are no building	7	66	64	Residential Rural
	2	28578.3	Mixed Builtup Area	bus_station office	3	64	45	Public Area Office
▸	50	93242.7	Mixed Builtup Area	school	3	64	157	Commercial Area
	52	34448	Commercial Area	marketplace industrial	3	64	54	Industrial Area
	26	46073.7	Residential Area	There are no building	5	63	76	Residential Urban
	39	61894.7	Residential Area	house	2	63	73	Residential Urban
	34	43317	Mixed Builtup Area	There are no building	5	62	81	Commercial Area
	14	300327	Commercial Area	car	3	61	343	Commercial-Residential Mixed
	44	358742	Mixed Builtup Area	school	3	61	425	Public Area School
	19	298021	Mixed Builtup Area	parking	3	60	417	Commercial Area
	25	430339	None	place_of_worship pu	7	60	369	Residential Rural
	29	359010	Residential Area	orchard	5	60	499	Residential Urban
	27	96042.6	None	farmland	5	59	167	Residential Rural
	31	118917	Residential Area	residential apartments	3	59	155	Residential Urban
	56	285422	None	fuel	7	59	213	Residential Rural
	13	387498	Residential Area	fuel	5	58	523	Residential Urban
	15	149574	Mixed Builtup Area	school	5	58	110	Public Area School
	35	174047	Mixed Builtup Area	marketplace industrial	3	58	241	Commercial-Residential Mixed
	36	73426	Commercial Area	marketplace industrial	3	58	103	Industrial Area
	54	217312	Residential Area	house	3	58	285	Residential Urban
	16	159396	Commercial Area	marketplace industrial	3	57	152	Commercial Area Office
	30	41045.3	Mixed Builtup Area	school	3	57	58	Public Area School
	32	40742.2	Residential Area	school	5	57	42	Public Area School

Figure 34: Final classification (red) snippet with the majority information from the auxiliary data of Palakkad for building occupancy type.

public sector (or the government sector) and is maintained by the municipality of Palakkad. Furthermore, the majority cluster as 4(5)[23] is associated with *buildings with open surroundings*, which is true in the case of recreational buildings like those situated in parks or the Palakkad fort in this case. Figure 35 shows the final classified homogeneous built-up area (blocks) of Palakkad with the occupancy (or function) type after compiling, interpreting, and validating the results from the local expert. As we see, the outer-skirts of the city are classified appropriately as Residential Rural while the inner CBD is mainly classified as Commercial Areas with its own type of other occupancies like office, residential mix and so forth for the other occupancy types as well.

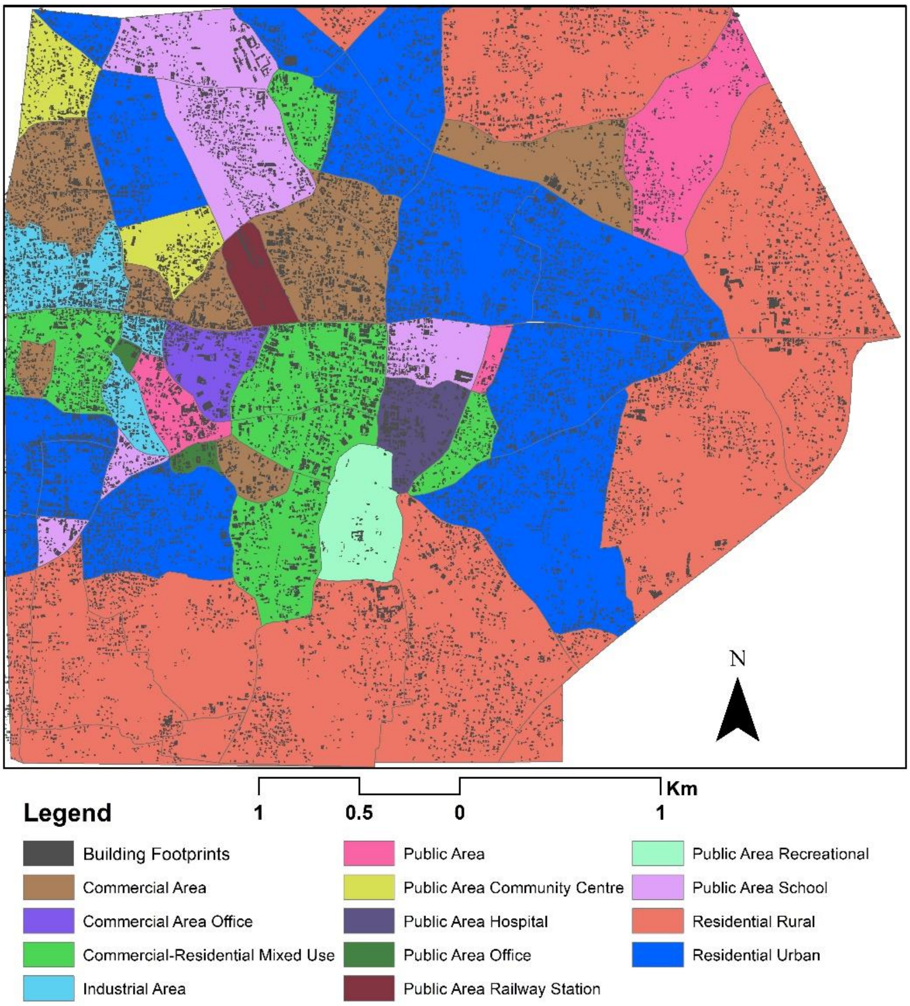


Figure 35: Occupancy types of the homogeneous built-up area in Palakkad.

5.4 Discussion

5.4.1 Amalgamation and local knowledge

The amalgamation of the auxiliary open-source data is one of the first attempts at approximating the building occupancy type at a homogeneous built-up level. The degree of completeness, quality, positional accuracy, and usability of the open-source data makes this methodology very limited to expand on. However, even with an area like Palakkad, where OSM tags and land use information are not abundantly available, it was still possible to reach a level of characterisation that can help understand the building function/occupancy types. Moreover, with additional input from the local experts and stakeholders, it was possible to refine the

[23] The clusters numbers in figures 34 and 41 are sorted from 0 to 7 instead of 1 to 8, hence, here 4 means 5.

classification to a great degree, which now represents the plausible building characteristics that exist in reality. However, due to the unavailability of certain other open-source data like the WSF-3D data, important attributes such as the building height could not be integrated for a more robust methodology to derive the number of floors per homogeneous block. The open-source data is the strength and the weakness of the methodology at the same time, whereby the availability or the deficit of open data will dictate the extent to which the methodology could be successfully applicable for the built-up area characterisation. Other obvious limitations that are associated with the open-source data, like the manufacturing date of the landuse maps where possible temporal uncertainties might be correlated and hamper the quality of the resulting output. Further limitations will be discussed in chapter 8.

5.4.2 Final Classification

Similar to the previous section, the final classification was possible due to manual interpretation of the building blocks based on the majority of building information like building tags, landuse information, clusters, and additional information from the distance from CBD. Apart from this, local expert suggestions helped shape the proper classification of the built-up area (or blocks) into building occupancy types. There were obvious challenges that were met which will be discussed more in-depth in the 8th chapter, but in short, had implications in the form of subjective classification of the buildings that is later resolved by local experts. The interpretation always requires prior knowledge to be combined with the information at hand, which is essentially external knowledge. Thus, different people holding different prior background knowledge would have a different interpretation. Such techniques make the methodology very subjective, especially at the building footprint level, and therefore, the research wanted to compensate for this by accomplishing the objectives at a homogenous level to counter the mentioned drawbacks. For each of the occupancy type, information such as the number of buildings, total floorspace area of buildings, percentage of built-up area per homogeneous unit, and the number of floors can also be calculated that can be later on used to estimate population data per homogeneous units, thus also giving us the population EaR data apart from the buildings.

Moreover, the time taken to refine the classification of the blocks with the local experts/stakeholders solely depends on the availability of said experts to spend time and hold numerous sessions to reach a point of agreement on the final classes of the occupancy types. However, with proper local knowledge and help from local stakeholders, it was possible to refine and improve on existing interpretations and reach a point of approximating the real potential classification of the built-up area. Furthermore, with compliance of the local stakeholders for the possible occupancy types in the study area, ruled out in favour of the applicability of the methodology overall in other data-scarce regions. Therefore, the reproducibility in terms of the methodology and the time taken, will be tested to investigate how fast the methodology could be transferred to a new area in chapter 6.

5.5 Chapter Summary

One of the most important questions that the research attempts to answer is "*how far it is possible to characterise the buildings in data-scarce regions with minimal auxiliary data that is available as open-source?*". The approach was to develop the database, starting from generating building footprints to characterising built-up area homogeneously. The thesis research practically delivered a workflow to *identify* building EaR that is useful for exposure, vulnerability, and hazard risk assessment in hazard-prone data-scarce areas. So far, based on the results and outcomes of the research, it has been possible to develop a semi-automated method of integrating aspects of DL and open-source auxiliary data to reach a point of characterising buildings, albeit at a homogeneous unit of level. The main focus of this chapter was to employ physical (morphological) characteristic metrics of buildings, coupled with information from auxiliary data, to be able to classify buildings based on certain typological attributes like occupancy type.

CHAPTER 6: APPLICATION OF THE METHOD IN A NEW TEST AREA

This chapter aims to test the developed workflow on the second study area of Kollam in Kerala, India, to investigate the reproducibility and time taken on newer-different regions.

6.1 Description and Results

6.1.1 Element-at-Risk Detection using Transfer Learning

With the ability of automatic learning of feature representations (building features in this case) within the scenario of scarce training data, transfer learning can become very effective in transferring the learnt weights from previous studies (that is, trained models) to newer data in different locations (Ravishankar et al., 2016). Therefore, to detect buildings in Kollam, transfer learning is used to address fewer training data. The Kollam data set contains 1100 building polygons in the training tiles, and the rest were manually digitised within the five training tiles (table 9 shows the number of tiles used for training and testing in Kollam). Moreover, since the building rooftop configurations (texture, shapes, colours) are similar to Palakkad's, using transfer learning makes more sense than simply training from scratch with label data from Kollam alone. Transfer learning also helps accomplish faster and seamless detection of buildings in new study areas with just a few training samples, thus allowing for effective transferability of the model in other similar regions. Such ability to detect buildings over a new and completely un-seen environment makes the use of such deep networks advantageous.

Using transfer learning from the weights learnt in Palakkad, the model trained over Kollam achieved over 74.6% F1-score accuracy. The predictions of buildings over Kollam can be seen in figure 36, and the respective overlay with the GUF for validation of the spatial distribution and patterns of the detection can

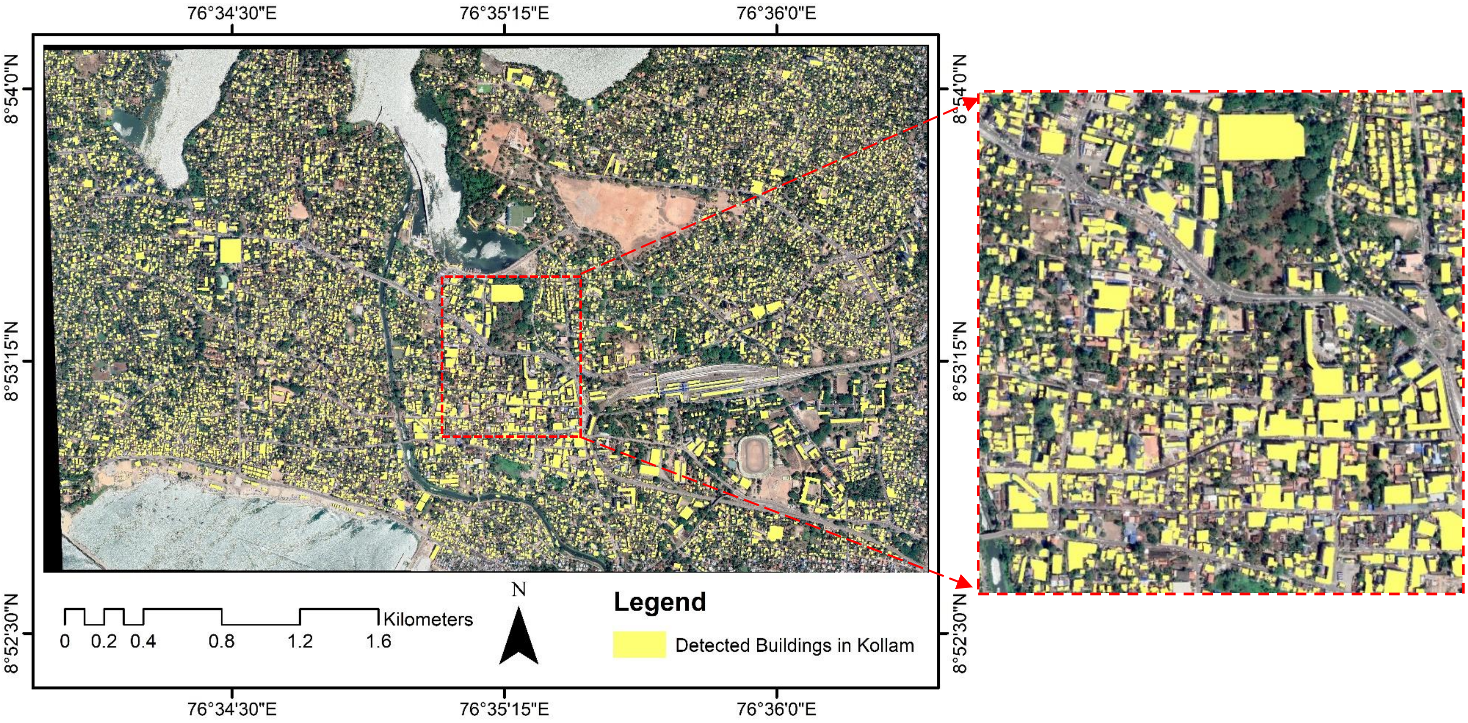


Figure 36: Detected buildings over Kollam using the ResU-Net model.

be seen in figure 37. Based on the figure, it can be seen that the detected building footprints overlay appropriately over the GUF data.

Table 10 shows the final scores after training the model, where the F1-score of 74% is relatively similar to that of Palakkad (76%). In this case with Kollam, the total time taken to prepare the building label data (manual digitisation) and detect the buildings with the ResU-Net model took around **8 hours**. Please refer to appendix section A for further information.

Table 9: Study site characteristics for training and testing sets at Kollam.

Summary of training-testing sites	Size of tiles	Number of tiles	Number of patches
Training set	8000x8000	5	1125
Testing set	8000x8000	3	300
Total		8	1300

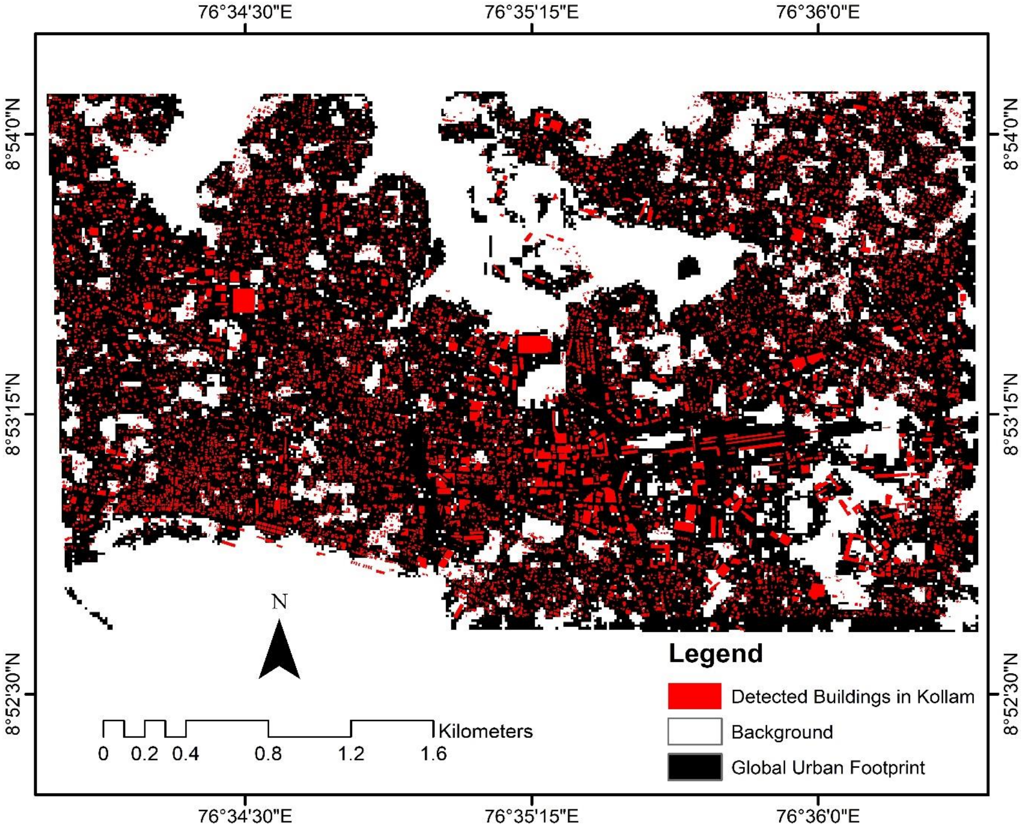


Figure 37: Overlay of the detected buildings with the Global Urban Footprint over Kollam.

Table 10: Summary table of final accuracies on the test set for Kollam.

Metrics	Scores
ACCURACY	0.8794
PRECISION	0.8018
RECALL	0.6975
F1-SCORE	0.7460

6.1.2 Urban Morphological Metrics

Figure 38 shows the cluster classification based on k=8 clusters (chosen based on the suggestion of the local expert), thus attributing information about the types of buildings morphologically. The K-Means derived clusters for the buildings in Kollam are described in table 11.

Table 11: Cluster interpretation of the buildings in Palakkad after local expert validation.

Cluster Number	Interpretation
1	Moderate-sized buildings
2	Densely located moderate-sized buildings
3	Sparsely located moderate-sized buildings - Type 1
4	Sparsely located moderate-sized buildings - Type 2
5	Very large and long buildings
6	Small-sized buildings
7	Large buildings with sharp corners
8	Very small-sized buildings

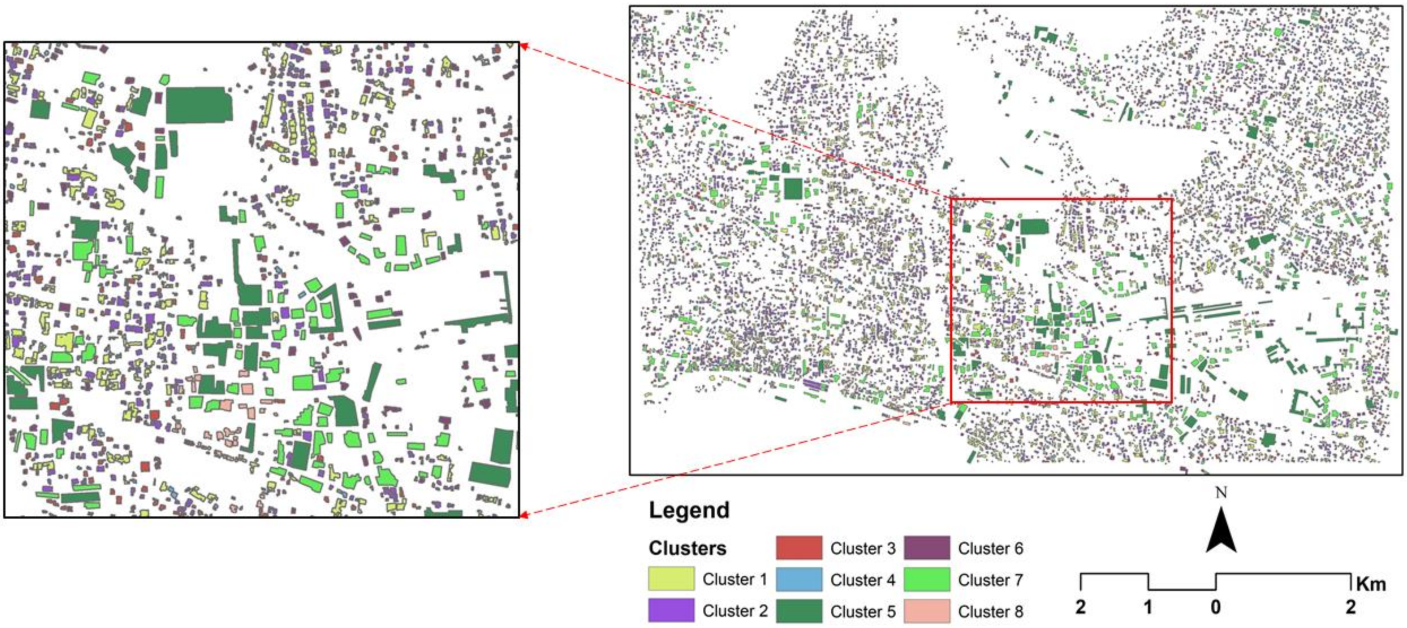


Figure 38: Morphological clusters of the buildings in Kollam after performing K-Means classification.

In Kollam (figure 38), it can be seen that very large buildings belong to cluster 5 and the interpretation from table 11 compliments very well. Similarly, cluster 3 interpreted as sparsely located moderate-sized buildings in the table can be visually observed in the map of Kollam as located sparsely in the zoomed part of the Kollam map. In this way, all the clusters are distributed geographically over the study areas, and the respective morphological metrics from Momepy helped cluster (or soft-classify) them into similar types of buildings.

The total time taken to clean the prediction outputs from the detection, followed by generating the morphological metrics of the buildings, took around **8 hours**.

6.1.3 Homogenisation with Road Networks and Homogeneity Score

Similar to Palakkad, the road networks[24] were used to improve the homogenisation from the existing clusters by incorporating a refined unit of built-up area extent (figure 39). Much of the buildings are similar to each other in Kollam in various blocks (figure 40). From the DL detection point of view, Kollam buildings were predicted with the learnt weights from Palakkad as well as trained with new building training samples, which has affected the predictions to be far better than that of Palakkad. Furthermore, post-processing after the detection can also explain why the buildings are much clearer and resemble the more regular polygonal shape. The Momepy library handles such errors or noise from irregular polygons fairly well; however, such situations are unavoidable and may affect the calculation of the morphological metrics.

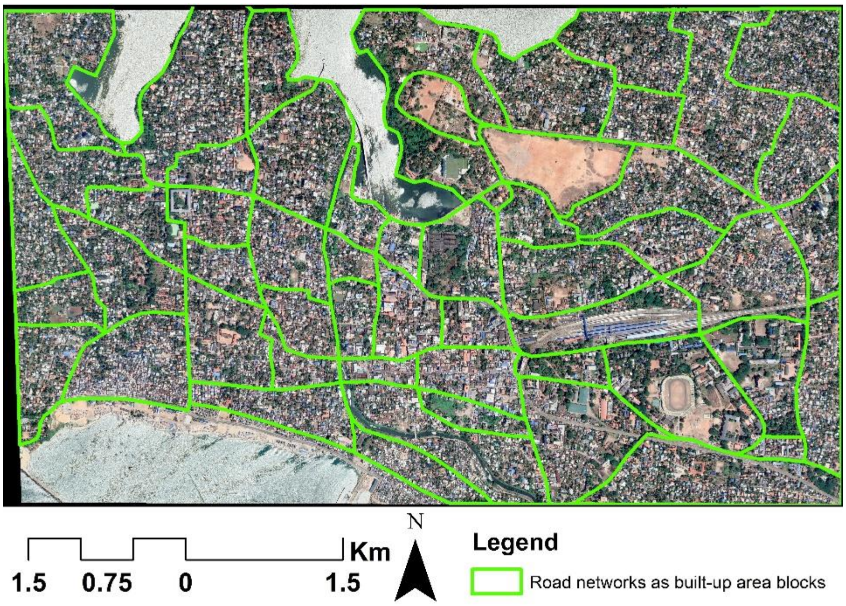


Figure 39: Road networks for built-up area blocks in Kollam.

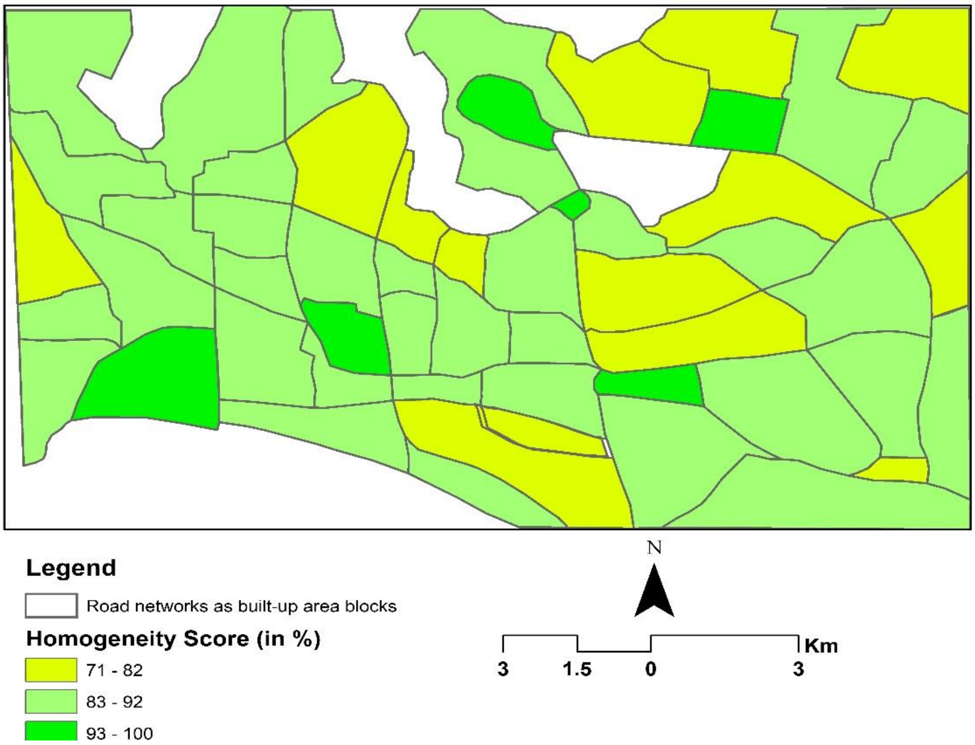


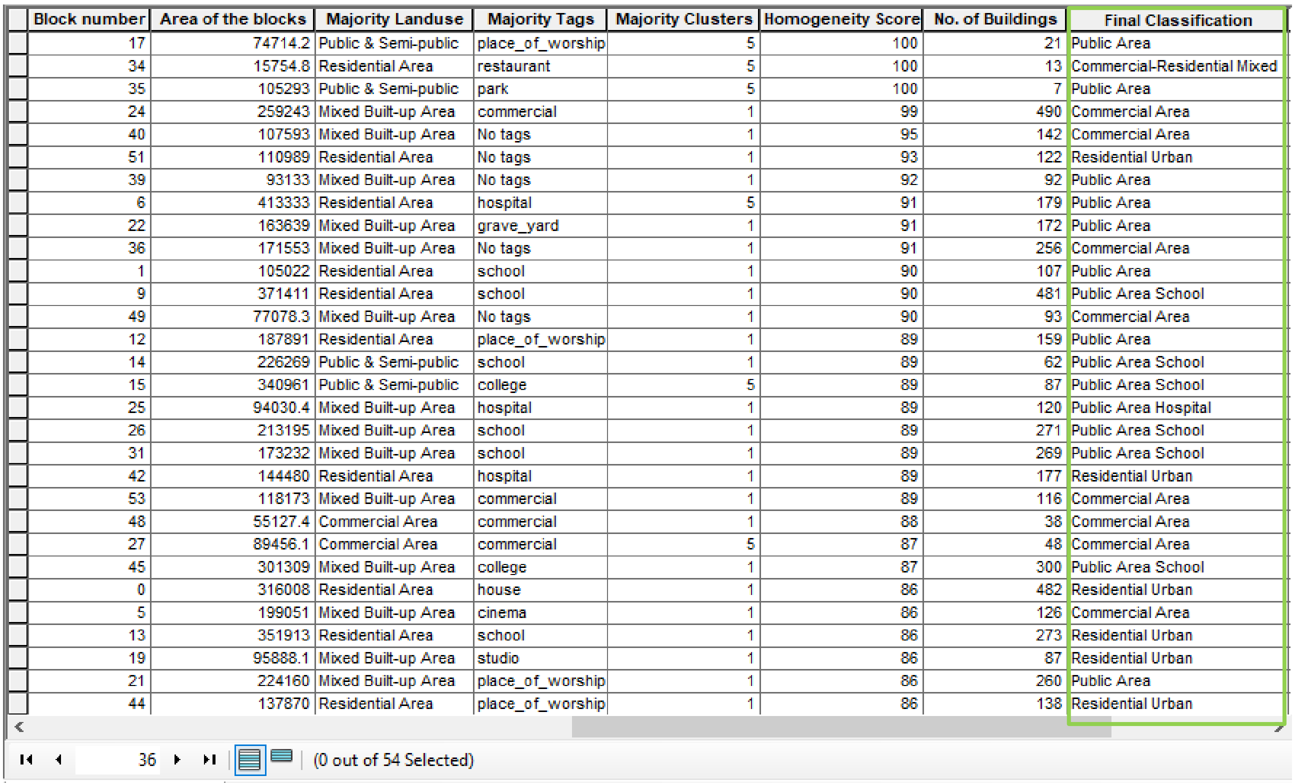
Figure 40: Homogeneity score of clusters in Kollam.

[24] In the instance of Kollam, railway lines were also used to help homogenise the study area along with road networks.

The total time taken to generate the road networks with other linear features like river and railway lines was about **1 hour**.

6.1.4 Final Classification

Figure 42 show the final classification of the occupancy types of Kollam. This result was obtained by employing the majority voting system (figure 28) similar to Palakkad in section 5.3.1 and 5.3.3. An example of the final tabular results (obtained from the voting system) is explained in figure 41 whereby in block 17, the majority building tag as *place_of_worship*, majority land use as *Public & Semi-public*, and majority cluster as 5(6) (which relates to regular buildings that are small in size), help classify the block as a *Public Area*. Places of worship like temples in such regions are located in places where fewer buildings surround it and are often small in size (small temples, archway gates, refer to figure 33); therefore, asserting a class of *Public Area* made sense. Therefore, in a similar manner, all the blocks were analysed based on such interpretations and were finally classified into occupancy types. Also similar to Palakkad, local knowledge validation was also performed for Kollam to investigate, improve, and refine the occupancy type classification.



Block number	Area of the blocks	Majority Landuse	Majority Tags	Majority Clusters	Homogeneity Score	No. of Buildings	Final Classification
17	74714.2	Public & Semi-public	place_of_worship	5	100	21	Public Area
34	15754.8	Residential Area	restaurant	5	100	13	Commercial-Residential Mixed
35	105293	Public & Semi-public	park	5	100	7	Public Area
24	259243	Mixed Built-up Area	commercial	1	99	490	Commercial Area
40	107593	Mixed Built-up Area	No tags	1	95	142	Commercial Area
51	110989	Residential Area	No tags	1	93	122	Residential Urban
39	93133	Mixed Built-up Area	No tags	1	92	92	Public Area
6	413333	Residential Area	hospital	5	91	179	Public Area
22	163639	Mixed Built-up Area	grave_yard	1	91	172	Public Area
36	171553	Mixed Built-up Area	No tags	1	91	256	Commercial Area
1	105022	Residential Area	school	1	90	107	Public Area
9	371411	Residential Area	school	1	90	481	Public Area School
49	77078.3	Mixed Built-up Area	No tags	1	90	93	Commercial Area
12	187891	Residential Area	place_of_worship	1	89	159	Public Area
14	226269	Public & Semi-public	school	1	89	62	Public Area School
15	340961	Public & Semi-public	college	5	89	87	Public Area School
25	94030.4	Mixed Built-up Area	hospital	1	89	120	Public Area Hospital
26	213195	Mixed Built-up Area	school	1	89	271	Public Area School
31	173232	Mixed Built-up Area	school	1	89	269	Public Area School
42	144480	Residential Area	hospital	1	89	177	Residential Urban
53	118173	Mixed Built-up Area	commercial	1	89	116	Commercial Area
48	55127.4	Commercial Area	commercial	1	88	38	Commercial Area
27	89456.1	Commercial Area	commercial	5	87	48	Commercial Area
45	301309	Mixed Built-up Area	college	1	87	300	Public Area School
0	316008	Residential Area	house	1	86	482	Residential Urban
5	199051	Mixed Built-up Area	cinema	1	86	126	Commercial Area
13	351913	Residential Area	school	1	86	273	Residential Urban
19	95888.1	Mixed Built-up Area	studio	1	86	87	Residential Urban
21	224160	Mixed Built-up Area	place_of_worship	1	86	260	Public Area
44	137870	Residential Area	place_of_worship	1	86	138	Residential Urban

Figure 41: Final classification (green) snippet with the majority information from the auxiliary data of Kollam for building occupancy type.

The steps to achieve the final classification of the homogeneous built-up area generally consisted of (1) downloading the auxiliary data, (2) generating land use maps, (3) digitisation of buildings tags manually from Google Maps, (4) amalgamating the auxiliary data and interpretation through the voting system, and (5) local expert validation, and took around **10 hours** to classify them as building occupancy types in Kollam.

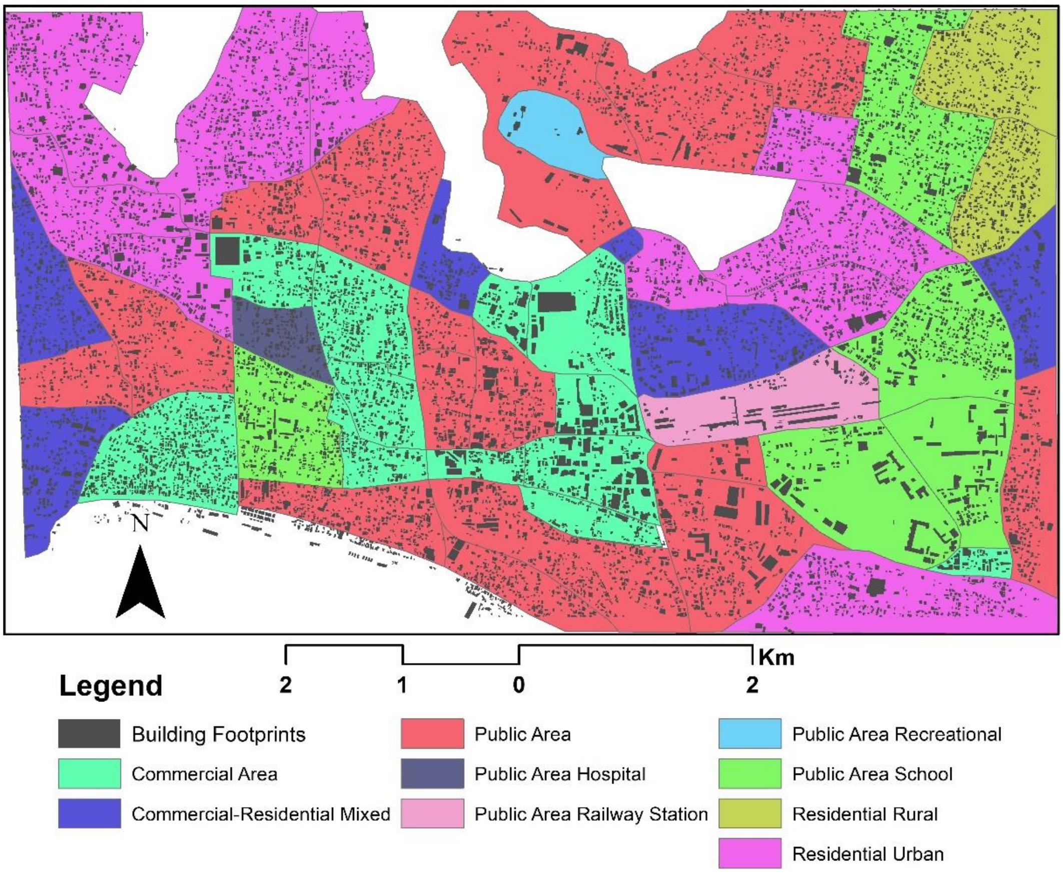


Figure 42: Occupancy types of the homogeneous built-up area in Kollam.

6.2 Overall Discussion

6.2.1 Transfer Learning

The transferability test results suggest the potential of the building detection model to be transferred to other areas. Normally, new regions require lots of training data to adapt to the new variations of the new study image; however, this is avoided due to the weights learnt from the previous study area. As discussed in the earlier sections, transfer learning was used in an attempt to detect buildings in Kollam with minimal training data added. Due to the complex learning nature of the model followed by retaining information from Palakkad as trained weights, the same model was able to perform optimally for predicting buildings with over 74% F1-score accuracy. Much of the work in terms of learning the weights with associated training data was done in the first training iteration with Palakkad that allowed the optimal detection of buildings in Kollam, thus alleviating the need to digitise more building labels in Kollam manually and saving time in the process.

6.2.2 Homogeneity Score

In contrast to Palakkad, Kollam exhibited high homogeneity scores, which can be understood as much of the buildings within the blocks were very similar to each other morphologically. The clusters of buildings in Kollam have three blocks with 100 per cent homogeneity scores. Figure 40 shows the disparity in the homogeneity or similarity of the building clusters in Kollam. The building clusters in Kollam do not have a single homogeneous block under 70 per cent in the homogeneity score. This good scoring depicts that the buildings with the respective clusters (and their probable interpretations) are comparable and similar to each other. Such scoring also gives a sense of confidence in the overall clusterisation of the buildings in Kollam, meaning that the interpretations in table 11 are reasonable interpretations and can be reliable during the classification routine for the occupancy type in the next steps moving forward.

6.2.3 Final Classification

Similar to what was seen in Palakkad, the amalgamation of the auxiliary open-source data allowed approximating the building occupancy type at a homogeneous built-up level. Many factors play a role in how well the auxiliary data can be used, especially in terms of the degree of completeness, quality, positional accuracy, and usability of the open-source. However, the final results were still conceivable to a great extent using the methodology in a different study area. Moreover, with additional input from the local experts and stakeholders, it was possible to refine the classification to a great degree, representing the plausible building characteristics that exist in reality.

6.2.4 Time required for reproducibility

This chapter attempted to address the reproducibility of the methodology over a new area by keeping in mind the time required to develop a buildings EaR database. With the collected information of the time at each phase, it took over **27 hours** at generating the database. This relatively short amount of time required to generate building occupancy types as typological attributes with just open-source data paves an interesting path towards producing buildings EaR database, which can be suitable for data-scarce regions in emergency situations.

Appendix section A highlights the time taken to generate a database from the very beginning, that is, without transfer learning from previous areas (like in this chapter), which otherwise saves much time. Therefore, differences in the overall time required to complete the methodology are witnessed between this chapter and appendix section A.

6.3 Chapter Summary

The point of this chapter and an important part of the MSc thesis research were to investigate whether the methodology can be feasible and reproducible in other areas. Based on the results seen in the sections above (performed over Kollam), related to:

(1) using transfer learning for seamless detection of buildings using prior weights,
(2) performing a morphological metric evaluation,
(3) homogenising buildings into built-up area using road networks,
(4) amalgamation and final classification, and
(5) local expert-knowledge validation,

gave confidence and assurance that the methodology can be applicable and reproducible on different test sites and quickly produce the required buildings EaR data. With the intent of employing this approach towards addressing the mapping of building EaR with attribute information, the methodology sets a good development in the direction of mapping building EaR in data-scarce regions. With this realisation, building EaR data can be generated in any other country or region, provided there is the availability of such open-source data which can be coupled with this research methodology.

CHAPTER 7: EXPOSURE ASSESSMENT AND THE LINK TO VULNERABILITY

This chapter aims to answer the last sub-objective and the respective research questions (Figure 1: S-O4). The chapter is an additional illustration of resulting homogeneous units in the context of hazard exposure and the possible link to the vulnerability of the buildings. However, it is to be kept in mind that this is not the main goal of the thesis, rather an exploration of the opportunity to assess the exposure of the EaR.

7.1 Flood Susceptibility Maps

The flood susceptibility maps were obtained from the KSDMA (website). The maps are binary, meaning information of only *flood* and *no flood* are present. Unfortunately, because of this reason, hazard intensities are not available and hence, vulnerability assessment cannot be investigated. In figure 43, the flood extent of the respective areas can be seen.

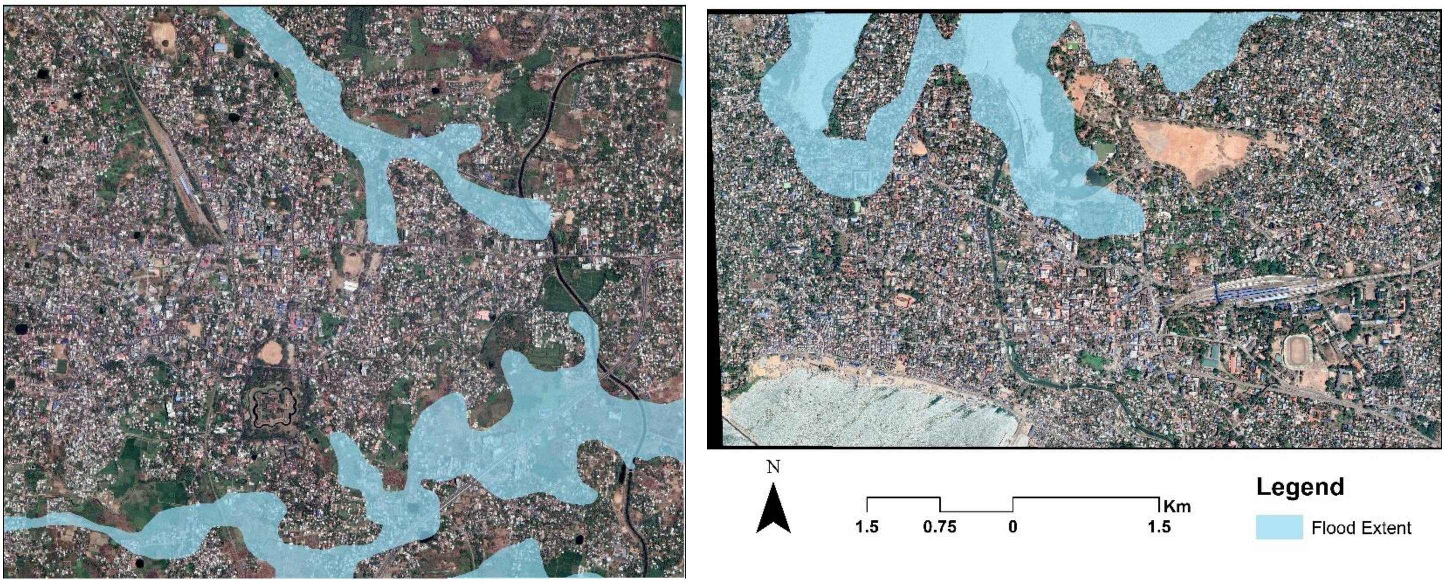


Figure 43: Flood susceptibility extent in Palakkad (left) and Kollam (right).

7.2 Flood Exposure Assessment

Flood exposure refers to the quantification of the EaR that are located in the flooded areas (De Moel, Aerts, & Koomen, 2011; Koks, Jongman, Husby, & Botzen, 2015). The exposure is calculated by performing a spatial overlay between the flood susceptibility map and the homogeneous built-up area blocks. Using the key ID (OBJECTID) of the blocks in the shapefile, the percentage of the area that is exposed by the flood extent is calculated. Moreover, the number of buildings affected are also reported. This way, the flood exposure is calculated using the *homogeneous built-up area blocks at the block-level*. The method is used in both areas of Palakkad and Kollam.

Another way of calculating the exposure is by using the individual building footprints and then aggregating them at the block level by using the key ID (OBJECTID) of the blocks, which relates the building footprints to their respective blocks, thereby generating the aggregated exposure. This way, the flood exposure is calculated using the *building footprints aggregated at the block-level*. Examples of the code that is used to calculate the exposure at both block-level and the aggregated block-level are shown in the appendix section H.

7.3 Results and Discussions

The crucial information that the thesis research provides is the classification of the built-up area. The exposure map provides information of the exposed buildings associated with their quantity, spatial distribution, and typological attributes based on the occupancy type. This information can be later used in scenarios where the hazard intensity is present (like flood depth) and can be used with vulnerability curves to associate the vulnerability of the buildings to their occupancy/functional type. Table 12 also relay the information of the number of exposed buildings per block in both study areas.

The number of blocks that were exposed in Palakkad are reported in the attribute table with the percentage affected by the flood extent can be seen in figure 44, and the exposure statistics are shown in table 12. Twenty-one homogeneous blocks are exposed to the flood in Palakkad both at the block-level and the aggregated block-level.

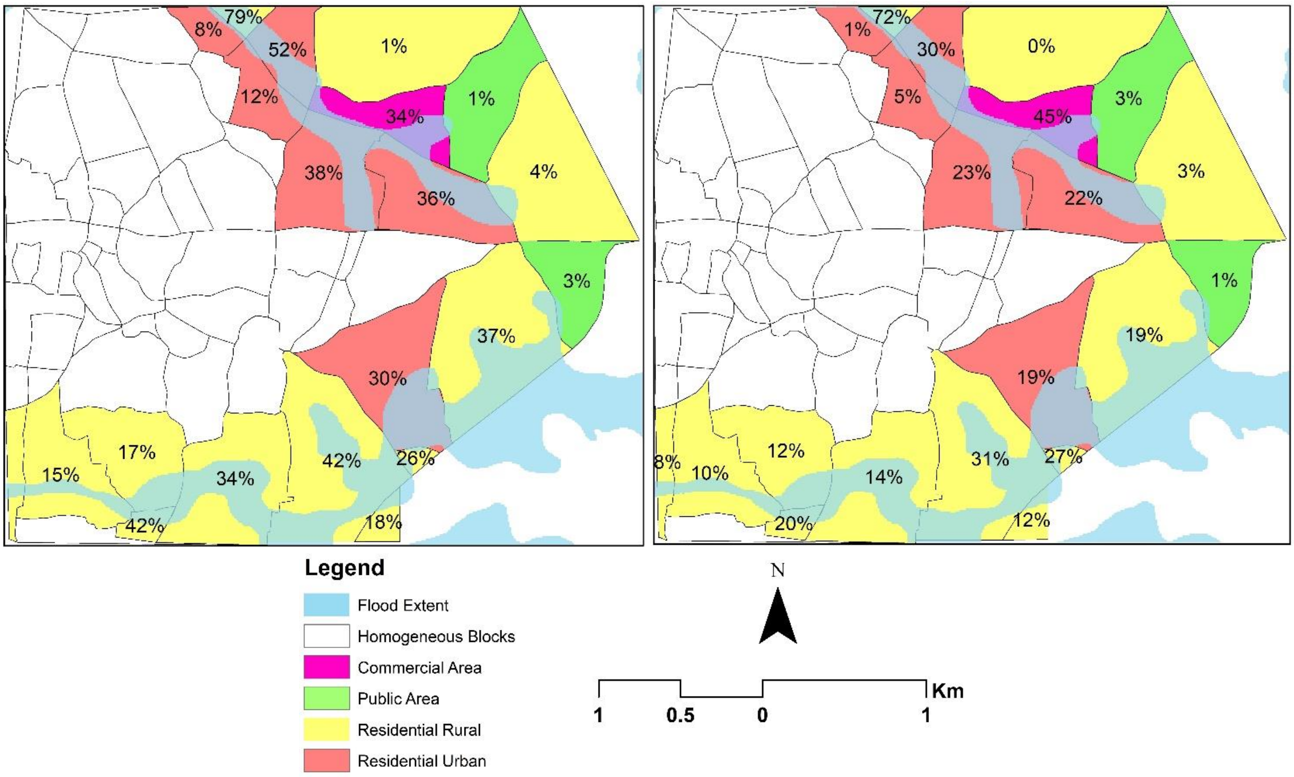


Figure 44: Flood exposure map of Palakkad with exposure as the percentage of the block exposed to flood (left) and the percentage of buildings within the blocks exposed to flood (right).

Similarly, the same was repeated for Kollam to report the exposure of the percentage area exposed to floods at both the block-level and the aggregated block-level, the number of buildings exposed, and the type (occupancy) of buildings that were exposed (figure 45 and table 12). Eight-teen homogeneous blocks are exposed to the flood in Kollam both at the block-level and the aggregated block-level.

Table 12: Information of the buildings exposed in terms of - number of buildings exposed, the percentage block exposure, and the percentage of buildings within the block exposure in Palakkad and Kollam.

	Homogeneous Blocks	No. of buildings exposed	Occupancy Type	Percentage of block exposed to flood (%)	Percentage of buildings within the block exposed to flood (%)
Palakkad	0	11	Residential Rural	26	27.5
	11	92	Residential Urban	30	19.2
	21	30	Residential Rural	79	72.8
	22	74	Residential Rural	34	14.6
	23	19	Residential Rural	17	12.4
	24	7	Residential Rural	42	20.6
	25	20	Residential Rural	15	10.4
	27	13	Residential Rural	11	8.03
	41	1	Residential Urban	8	0.96
	45	2	Residential Rural	0	0.19
	46	56	Commercial Area	34	45.7
	47	198	Residential Rural	42	31.1
	48	4	Residential Rural	18	12.4
	49	133	Residential Urban	38	23.6
	51	46	Residential Urban	52	30.5
	54	16	Residential Urban	12	5.4
	55	74	Residential Rural	37	19.1
	56	3	Public Area	3	2
	57	133	Residential Urban	36	22.38
	58	11	Public Area	1	3.8
	59	6	Residential Rural	4	3.8
Kollam	0	53	Residential Urban	23	20
	1	42	Public Area	45	40
	2	74	Public Area	25	18
	3	81	Residential Urban	48	47
	4	26	Commercial-Residential Mixed	39	24
	5	72	Commercial Area	49	69
	6	64	Public Area	48	45
	7	95	Public Area	37	30
	8	84	Public Area	34	29
	9	3	Public Area School	1	1
	19	3	Residential Urban	1	2
	20	41	Residential Urban	21	25
	33	18	Public Area	47	43
	35	3	Public Area	11	31
	38	130	Residential Urban	30	30
	41	1	Public Area Hospital	1	1
	44	36	Residential Urban	26	31

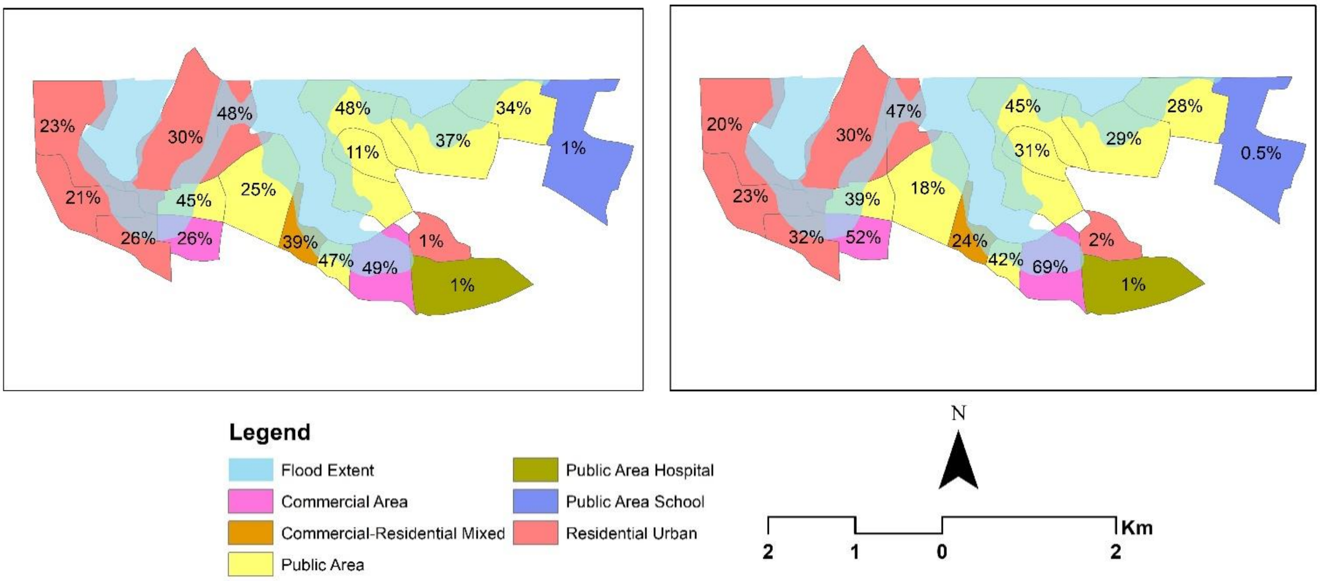


Figure 45: Flood exposure map of Kollam with exposure as the percentage of the block exposed to flood (left) and the percentage of buildings within the blocks exposed to flood (right).

The differences between the obtained results between the values of the exposure at block-level and the number of exposed buildings per block show some interesting results. The resulting outputs show differences in the percentage exposure of some blocks. Particularly blocks 11, 22, 23, 24, 41, 46, 47, 51, 54, and 57 in Palakkad, and blocks 4, 5, 8, 35, and 53 in Kollam have significant differences in the percentage of built-up area blocks exposed to the floods. The main reason for such differences is the non-uniform spatial distribution of the building footprints within the blocks. Hence, when compared against the percentage exposure at the block-level, the differences are apparent. A good example of this phenomenon is given in figure 46. As we see in the green box in the figure, the buildings are distributed non-uniformly in the entire block. Therefore, the percentage exposure of the buildings aggregated at the block-level is

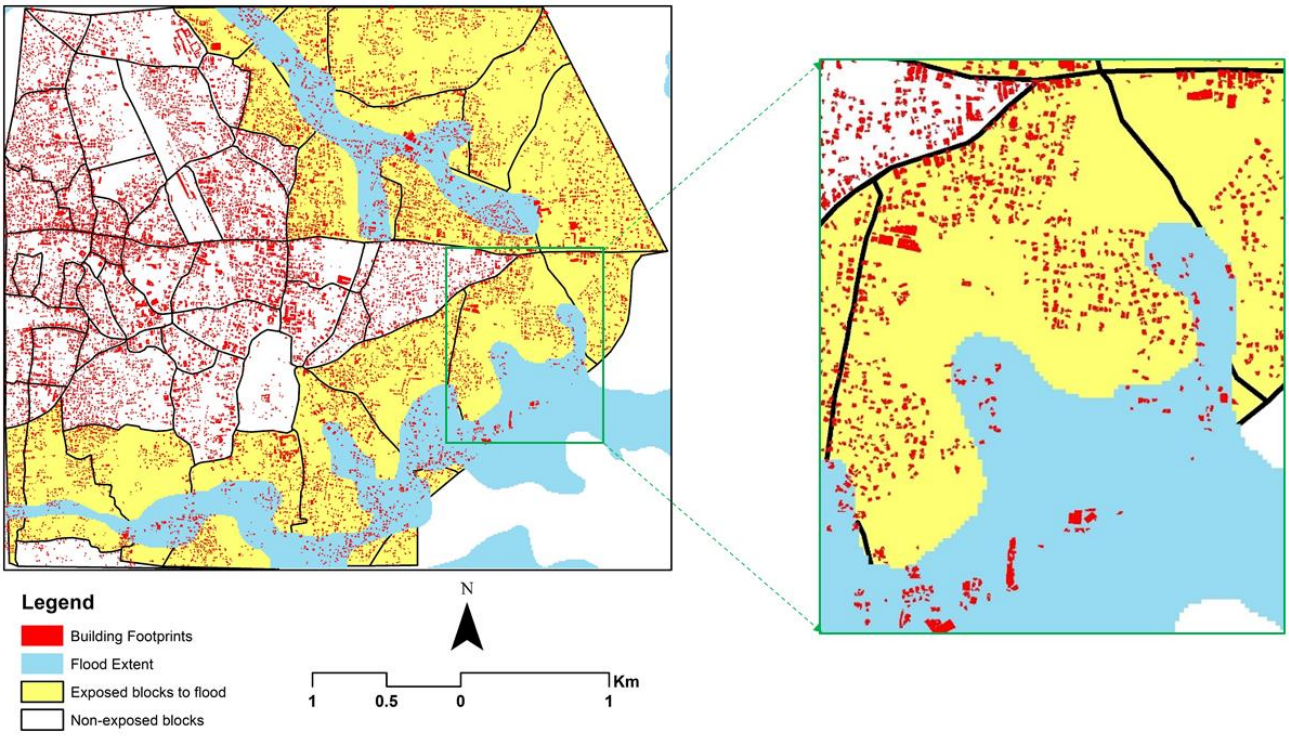


Figure 46: Flood exposure to blocks against the building footprints in Palakkad.

19.1%, whereas the overall block is exposed to 37% of the flood. Due to the distribution of most buildings towards the north, north-west and the west of the block, the percentage of buildings exposed is 19.1%.

Therefore, the final exposure assessment results are decided to be the values of the aggregated footprint level but reported at the block-level. Meaning that the values of exposed percentage will be that of the aggregated footprints but will be described on the built-up area blocks instead of the individual building footprints.

7.4 Link to vulnerability: the next journey

One of the crucial research questions to answer was how this methodology could link towards the physical vulnerability assessment. One of the main requirements of calculating the vulnerability is a physical vulnerability curve for buildings. The global flood-depth damage curves by Huizinga et al. (2017) are a useful resource to assess the physical vulnerability of buildings based on their occupancy/building function type. However, another crucial requirement is a hazard map that contains the hazard intensity. Since the available flood maps are susceptibility maps with values 0 (no-flood) and 1 (flood), it was not possible to perform a vulnerability assessment. However, provided there was a flood or any other hazard map, it would have been realistically possible to assess the physical vulnerability of the two study areas based on the research methodology. As the research output is homogeneous built-up areas with occupancy type as typological attributes, these attributes can be easily linked to the vulnerability curves by Huizinga et al. (2017) and can assess the vulnerability of said built-up areas (through the building footprints within the homogeneous units). The applications of this study can cater to not just large-scale areas but also smaller regional to city-scale areas. Studies by Q. Yang et al. (2020) have used these vulnerability curves in the city of Lishui in China for flood exposure and vulnerability assessment, where the city area (in sqm) is almost similar to that of both Palakkad and Kollam. Furthermore, the application of the vulnerability is subsequently used in loss estimation, risk assessment, and the ensuing measures and policies to reduce, mitigate, and avoid the risk of hazards. With the help of such curves for flood hazards but also for landslide hazards at a more general scale with landslide building vulnerability curves like Glade (2003), it is possible to estimate the vulnerability of the built-up area blocks based on the occupancy/function type and would be the direction for future research.

CHAPTER 8: LIMITATIONS, RECOMMENDATIONS AND FINAL CONCLUSION

In this chapter, the various limitations are discussed along with the possible suggestions to solve or counter these limitations, followed by the final conclusion of the thesis research.

8.1 Limitations

The limitations of the study are discussed in the sections below regarding each phase of the research.

8.1.1 Detection Phase

Prior to the training phase, it is commonly known that the DL model performs the best when huge sets of training data are available. In short, the more training data is available, the better would be the predicting capability of the model. Shortage of training samples from OSM and manual digitisation led to an average F1-score of 76% for Palakkad and 74% for Kollam. Additional training data could have helped in improving the overall F1-score and thus, allowing for better predictions by reducing the false-positive and refining the prediction of the building delineation.

Another core problem witnessed during the training phase was the lack of computing power despite using Google Colab. This is because the model was trained with a patch size of 512 over sample images of 8000x8000 pixels. Patch sizes under 512 cannot be used as the session crashes at every attempt due to the limited RAM allocation of 25GB. The investigation of the influence of patch size was, therefore, not possible because of this reason. Furthermore, data augmentation[25] techniques could not be investigated as well because of the deficit in RAM capacity in Google Colab. Augmenting the training samples with rotations, scaling, contrast changes, and cropping was not possible as a consequence.

Finally, the training of the ResU-Net model takes a very long time due to the deep network architecture of the model. Therefore, the training time of each epoch took around one hour. Coupled by the fact that Google Colab runs for only 24 hours before resetting the session, only 25 epochs of training were possible. Training with higher epochs was, hence, not possible.

8.1.2 Characterisation Phase

In the initial phases of using OSM data, some common issues witnessed concerned positional accuracy, data quality, lack of attribute information, and others, to name a few. As a result of this, building tags from Google Maps had to be incorporated to supplement the lack of such typological data of the buildings. Furthermore, the tags from Google Maps cannot be downloaded despite the fact that volunteers have placed them, thus posing a challenge in utilising all of the tags, thus requiring manual digitisation in Google Maps. Moreover, tags for residential buildings are not given, so there is a strong bias towards non-residential buildings, with an emphasis on commercial buildings. Lastly, even when there are building tags, due to the small number of individual tags present within a block, sometimes using the majority tags might not be the best way to represent the actual building occupancy.

In evaluating the morphological metrics, the limitations of the Momepy library stemmed from the fact that the clustering was based on the nature of the building footprints. Although the library is well-suited to

[25] Data augmentation are techniques utilised to increase amount of data or information by adding modified duplicates of existing data.

address noise from irregular-polygonal distortions, polygon overlays, and multi-polygons, such noisy errors are, unfortunately, still unavoidable. Therefore, the noise would affect the overall calculation of the morphological metrics at times, further affecting the clustering in the proceeding steps.

Another limitation was the homogeneity score of the building clusters in Palakkad, which were lower compared to Kollam due to the predictions as false positives after detection as irregular polygons and multi-polygons. These false-positive polygons can affect the morphological metric calculations and thus, affect the true understanding of the prevalent building morphologies in reality.

The road network was used in lieu of missing administrative units like wards or census tracts to counter under/over-approximation of the homogeneity within the clusters, provided the number of clusters was either too low or too high. However, due to the introduction of the roads as the foundation for the homogeneous blocks, there can be instances where due to the lack of secondary roads in certain regions, the road networks cannot appropriately be employed to aggregate the occupancy types. Sometimes, there are issues where vast areas or blocks are used for the aggregation of the building occupancy types, which could be counter-intuitive. Other linear features such as railway lines, river lines, topographic differences and others can be used, but it would very well vary for different topographies/regions in relation to finding the most suitable size of homogeneous units. For example, railway lines and river lines were used for homogenising units in Palakkad, but the absence of rivers in Kollam devoid the use of river lines for homogenising.

Moreover, uncertainties from the open-source data are one of the major concerns, and thus, attention must be given to carefully consider the quality, type, resolution, and source of the open-source data. The limitations of the landuse maps are also quite important to discuss. The year of the landuse maps from the Bhuvan portal was generated during 2015-2016, and therefore, possibly, there can be many uncertainties associated with it. The availability of landuse maps is also a point of discussion where normally, such maps might not be available in other areas. Therefore, in the voting system (refer back to figure 28), the building tags were set higher in the voting hierarchy while classifying the built-up area as the buildings tags were obtained from OSM and are more up-to-date.

Another limitation witnessed is that of the interpretation of the voting system, which can change accordingly at different places and hence would require local validation every time. Because of this reason, the methodology cannot be fully automated as there will always be a point where local knowledge validation would be necessary to authenticate the results.

8.1.3 Exposure and Vulnerability

The lack of hazard maps for the study areas did not allow the employment of the building occupancy types to assess the vulnerability. The requirement of the flood hazard intensities like flood height (in cm) is crucial to estimate the level of vulnerability and possible damage to buildings of certain occupancy types.

8.2 Suggestions and recommendations for future research

Some possible solutions and recommendations to counter the limitations faced in the research are listed below:

1. Using cloud machines like Microsoft Azure and other stronger machines with higher RAM capacities can help alleviate some of the time-constraint issues with training for more than 24-hours and training with smaller patch sizes to investigate the improvement of accuracies.

2. As mentioned previously about the issue of polygon irregularity, edge detection of buildings using OpenCV tools like Canny Edge can be incorporated to improve building boundary delineation.

3. Transfer learning can also be used temporarily in the same city to assess the changes in the building dynamics to study urban behavioural patterns, which can relay information about the construction of buildings towards hazard-prone areas. This approach can help prevent potential damage and suggest policies and measures to address the exposure to hazards and develop disaster mitigation plans.

4. The employment of more auxiliary data such as the WSF-3D data for building height can help refine the classification by estimating the number of floors per block. This information along with the average population at night and daytime from the CDMP[26] (Ara, 2013), can be used to estimate the population per block.

5. Another interesting suggestion can be the use of object detection DL algorithms like YOLO (Pham et al. 2020) with oblique photos from GSV for applications like (1) recognising the number of windows for estimating the number of building floors, (2) recognising occupancy types from signs outside the buildings, and many more.

6. The advent of smaller wards/homogeneous blocks can also further refine the information of built-up area by improving the aggregation within the blocks.

7. The development of a code to automate the voting system can greatly improve the time that is taken to classify the built-up area. Such a code can also streamline the process of the voting system and can accordingly change the hierarchy of the majority information that is used for the classification, depending on which auxiliary data to prioritise on.

8. The exposure results can also be broadcasted to the footprint level to address the issues of spatial non-uniformity, and also the loss can be calculated at the footprint level but then aggregated to the block level.

8.3 Final Conclusion

As discussed previously, one of the essential components for risk assessment is the elements-at-risk information which often lacks in certain developing nations. The research attempted to resolve this deficiency by developing a semi-automated *detection* and *characterisation* method to design an EaR database for buildings. The research sub-objectives were achieved by first (1) detecting buildings in Palakkad with an F1-score of 76%, followed by (2) homogenising the buildings into built-up areas with road networks and deriving the building morphological characteristics, and (3) obtaining the building EaR occupancy types like residential, commercial, industrial etc., by amalgamating data from open-sources. Moreover, we also tested the reproducibility of the methodology in a different region called Kollam. We achieved an F1-score of 74% in building detection and building occupancy type as the characterisation output. After obtaining the final building EaR maps, (4) we quantified the exposure of the buildings by spatially overlaying the EaR maps over the flood susceptibility maps, which can now also be used for vulnerability assessment with the building occupancy type output (provided hazard intensities are available).

The research addresses the needs to face hazards that can potentially damage and destroy EaR (buildings in this instance) and helps increase the resiliency of said EaR to tackle hazards more effectively and efficiently.

[26] Comprehensive Disaster Management Programme (https://www.preventionweb.net/events/view/33066?id=33066)

The research shows that it is possible to obtain the EaR information as building occupancy type using remote sensing image data in combination with freely available data on geotags and OSM by means of the state-of-the-art DL models, open-source remote sensing products, and validation with local expert/stakeholder. The research enabled the development of a building EaR database in data-scarce regions as the first step for estimating hazard vulnerability, risk assessment, rescue missions, and rehabilitation. This methodology also has implications for dasymetric mapping in developing nations or regions that lack building typological information.

Another important aspect of the research was to check how this methodology can be used for situations of emergency where rapid or timely mapping of EaR can be crucial for effective risk mitigation and disaster relief measures. Hence, section A in the appendix will shed light on the time taken for each step (or phase) to generate the required data whereby, it can be important in real-world crises to produce the buildings EaR database quickly.

LIST OF REFERENCES

Abraham, N., & Khan, N. M. (2019). A novel focal tversky loss function with improved attention u-net for lesion segmentation. *Proceedings - International Symposium on Biomedical Imaging*, *2019-April*, 683–687. https://doi.org/10.1109/ISBI.2019.8759329

Ajami, A., Kuffer, M., Persello, C., & Pfeffer, K. (2019). Identifying a slums' degree of deprivation from VHR images using convolutional neural networks. *Remote Sensing*, *11*(11), 1282. https://doi.org/10.3390/rs11111282

Alidoost, F., & Arefi, H. (2018). A CNN-Based Approach for Automatic Building Detection and Recognition of Roof Types Using a Single Aerial Image. *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, *86*(5–6), 235–248. https://doi.org/10.1007/s41064-018-0060-5

Angela, B. V., Norbert, H., & Jochen, S. (2013). Building extraction from remote sensing data for parameterising a building typology: A contribution to flood vulnerability assessment. *Joint Urban Remote Sensing Event 2013, JURSE 2013*, *856*, 147–150. https://doi.org/10.1109/JURSE.2013.6550687

Ara, S. (2013). *Analyzing Population Distribution and Its Effect on Earthquake Loss Estimation in Sylhet , Analyzing Population Distribution and Its Effect on Earthquake Loss Estimation in Sylhet ,*. 22.

Ariza-López, F. J., García-Balboa, J. L., Alba-Fernández, V., Rodríguez-Avi, J., & Ureña-Cámara, M. (2014). Quality assessment of the OSM data from the mapping party of Baeza (Spain). *Accuracy 2014 - Proceedings of the 11th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. International Spatial Accuracy Research Association (ISARA).

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLOS ONE*, *12*(8), e0180698. https://doi.org/10.1371/journal.pone.0180698

Blanco-Vogt, A., & Schanze, J. (2014). Assessment of the physical flood susceptibility of buildings on a large scale - Conceptual and methodological frameworks. *Natural Hazards and Earth System Sciences*, *14*(8), 2105–2117. https://doi.org/10.5194/nhess-14-2105-2014

Blaschke, T. (2010, January 1). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 65, pp. 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). https://doi.org/10.1007/978-3-7908-2604-3_16

Cerri, M., Steinhausen, M., Kreibich, H., & Schröter, K. (2021). Are OpenStreetMap building data useful for flood vulnerability modelling? *Natural Hazards and Earth System Sciences*, *21*(2), 643–662. https://doi.org/10.5194/nhess-21-643-2021

Chen, J., & Zipf, A. (2019). DeepVGI: Deep learning with volunteered geographic information. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 771–772. https://doi.org/10.1145/3041021.3054250

Chen, L., van Westen, C. J., Hussin, H., Ciurean, R. L., Turkington, T., Chavarro-Rincon, D., & Shrestha, D. P. (2016). Integrating expert opinion with modelling for quantitative multi-hazard risk assessment in the Eastern Italian Alps. *Geomorphology*, *273*, 150–167. https://doi.org/10.1016/j.geomorph.2016.07.041

Cohen, J. P., Ding, W., Kuhlman, C., Chen, A., & Di, L. (2016). Rapid building detection using machine learning. *Applied Intelligence*, *45*(2), 443–457. https://doi.org/10.1007/s10489-016-0762-6

COI. (2011). Census of India. *District Census Handbook, Coimbatore, Tamilnadu, Series-34*(Part XII-B), 232. Retrieved from https://censusindia.gov.in/2011census/dchb/KerlaA.html

Dai, J., He, K., & Sun, J. (2015). Instance-aware Semantic Segmentation via Multi-task Network Cascades. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 3150–3158. Retrieved from http://arxiv.org/abs/1512.04412

De Moel, H., Aerts, J. C. J. H., & Koomen, E. (2011). Development of flood exposure in the Netherlands during the 20th and 21st century. *Global Environmental Change*, *21*(2), 620–627. https://doi.org/10.1016/j.gloenvcha.2010.12.005

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *162*, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

Dwyer, C. (2018). Monsoon Hammers India With “Unprecedented Flood Havoc,” Killing Scores Of People : NPR. Retrieved June 12, 2021, from https://www.npr.org/2018/08/16/639224478/monsoon-hammers-india-with-unprecedented-flood-havoc-killing-scores-of-people?t=1623498005548

Earthdata. (2019). GeoTIFF | Earthdata. Retrieved May 23, 2021, from https://earthdata.nasa.gov/esdis/eso/standards-and-references/geotiff

El-naggar, A. M. (2018, December 1). Determination of optimum segmentation parameter values for extracting building from remote sensing images. *Alexandria Engineering Journal*, Vol. 57, pp. 3089–3097. https://doi.org/10.1016/j.aej.2018.10.001

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., … Dech, S. (2013). Urban footprint processor-Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters*, *10*(6), 1617–1621. https://doi.org/10.1109/LGRS.2013.2272953

Esch, T., Thiel, M., Schenk, A., Roth, A., Müller, A., & Dech, S. (2010). Delineation of Urban footprints from TerraSAR-X data by analyzing speckle characteristics and intensity information. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(2), 905–916. https://doi.org/10.1109/TGRS.2009.2037144

Esch, Thomas, Schenk, A., Ullmann, T., Thiel, M., Roth, A., & 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 PART 1), 1911–1925. https://doi.org/10.1109/TGRS.2010.2091644

Esch, Thomas, Taubenböck, H., Roth, A., Heldens, W., Felbier, A., Thiel, M., … Dech, S. (2012). TanDEM-X mission—new perspectives for the inventory and monitoring of global settlement patterns. *Journal of Applied Remote Sensing*, *6*(1), 061702–1. https://doi.org/10.1117/1.jrs.6.061702

Eshrati, L., Mahmoudzadeh, A., & Taghvaei, M. (2015). Multi hazards risk assessment , a new methodology. *International Journal of Health System and Disaster Management*, *3*(2), 79. https://doi.org/10.4103/2347-9019.151315

Fan, H., Zipf, A., & Fu, Q. (2014). Estimation of building types on openstreetmap based on urban morphology analysis. *Lecture Notes in Geoinformation and Cartography*, 19–35. https://doi.org/10.1007/978-3-319-03611-3_2

Fleischmann, M. (2019a). momepy: Urban Morphology Measuring Toolkit. *Journal of Open Source Software*, *4*(43), 1807. https://doi.org/10.21105/joss.01807

Fleischmann, M. (2019b). momepy: Urban Morphology Measuring Toolkit. *Journal of Open Source Software*, *4*(43), 1807. https://doi.org/10.21105/joss.01807

Foody, G. M., See, L., Fritz, S., Van Der Velde, M., Perger, C., Schill, C., … Comber, A. (2015). Accurate attribute mapping from volunteered geographic information: Issues of volunteer quantity and quality. *Cartographic Journal*, *52*(4), 336–344. https://doi.org/10.1080/00087041.2015.1108658

Fu, Y., Ye, Z., Deng, J., Zheng, X., Huang, Y., Yang, W., … Wang, K. (2019). Finer Resolution Mapping of Marine Aquaculture Areas Using WorldView-2 Imagery and a Hierarchical Cascade Convolutional Neural Network. *Remote Sensing*, *11*(14), 1678. https://doi.org/10.3390/rs11141678

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, *11*(2). https://doi.org/10.3390/rs11020196

Ghorbanzadeh, O., Tiede, D., Wendt, L., Sudmanns, M., & Lang, S. (2020). Transferable instance segmentation of dwellings in a refugee camp - integrating CNN and OBIA. *European Journal of Remote Sensing*. https://doi.org/10.1080/22797254.2020.1759456

Gill, J. C., & Malamud, B. D. (2014). Reviewing and visualizing the interactions of natural hazards. *Reviews of Geophysics*, Vol. 52, pp. 680–722. https://doi.org/10.1002/2013RG000445

Glade, T. (2003). Vulnerability assessment in landslide risk analysis. *Erde*, Vol. 134, pp. 123–146.

Goodchild, M. F. (2007, August 20). Citizens as sensors: The world of volunteered geography. *GeoJournal*, Vol. 69, pp. 211–221. https://doi.org/10.1007/s10708-007-9111-y

Graff, K., Lissak, C., Thiery, Y., Maquaire, O., Costa, S., Medjkane, M., & Laignel, B. (2019). Characterization of elements at risk in the multirisk coastal context and at different spatial scales: Multi-database integration (normandy, France). *Applied Geography*, *111*. https://doi.org/10.1016/j.apgeog.2019.102076

Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., … Wolff, E. (2018). Mapping urban land use at street block level using OpenStreetMap, remote sensing data, and spatial metrics. *ISPRS International Journal of Geo-Information*, *7*(7), 246. https://doi.org/10.3390/ijgi7070246

Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., & Herrera, F. (2017). *Deep-Learning Convolutional Neural Networks for scattered shrub detection with Google Earth Imagery*. Retrieved from http://arxiv.org/abs/1706.00917

Hao, L., van Westen, C., Ranjan Martha, T., Jaiswal, P., McAdoo, B., & Hao hao, L. (2020). *Constructing a complete landslide inventory dataset for the 2018 Monsoon disaster in Kerala, India, for land use change analysis Earth System Science Data Discussions*. *2*(June), 1–32. https://doi.org/10.17026/dans-x6c-y7x2

Hasan, R. C., A'Zad Rosle, Q., Asmadi, M. A., & Kamal, N. A. M. (2018). Extraction of element at risk for landslides using remote sensing method. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(4/W9), 181–188. https://doi.org/10.5194/isprs-archives-XLII-4-W9-181-2018

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hindu, T. (2018). Over 3,600 in relief camps in Kollam - The Hindu. Retrieved May 5, 2021, from https://www.thehindu.com/news/national/kerala/over-3600-in-relief-camps-in-kollam/article24709231.ece

Huizinga, J., de Moel, H., & Szewczyk, W. (2017). Global flood depth-damage functions : Methodology and the Database with Guidelines. In *Joint Research Centre (JRC)*. https://doi.org/10.2760/16510

Husen, S. N. R. M., Idris, N. H., & Ishak, M. H. I. (2018). The quality of OpenStreetMap in Malaysia: A preliminary assessment. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(4/W9), 291–298. https://doi.org/10.5194/isprs-archives-XLII-4-W9-291-2018

Iglovikov, V., Seferbekov, S., Buslaev, A., & Shvets, A. (2018). TernausNetV2: Fully convolutional network for instance segmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2018-June*, 228–232. https://doi.org/10.1109/CVPRW.2018.00042

Kanthi, N. S., & Purwanto, T. H. (2016). Application of OpenStreetMap (OSM) to Support the Mapping Village in Indonesia. *IOP Conference Series: Earth and Environmental Science*, *47*(1). https://doi.org/10.1088/1755-1315/47/1/012003

Karpatne, A., Jiang, Z., Vatsavai, R. R., Shekhar, S., & Kumar, V. (2016). Monitoring land-cover changes: A machine-learning perspective. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 8–21. https://doi.org/10.1109/MGRS.2016.2528038

Kerala State Spatial Data Infrastructure. (2021). Mapathon Keralam. Retrieved June 20, 2021, from https://mapathonkeralam.in/എന്തിനാണ്/

Koks, E. E., Jongman, B., Husby, T. G., & Botzen, W. J. W. (2015). Combining hazard, exposure and social vulnerability to provide lessons for flood risk management. *Environmental Science and Policy*, *47*, 42–52. https://doi.org/10.1016/j.envsci.2014.10.013

Kuffer, M., Thomson, D. R., Boo, G., Mahabir, R., Grippa, T., Vanhuysse, S., … Kabaria, C. (2020). The role of earth observation in an integrated deprived area mapping "system" for low-to-middle income countries. *Remote Sensing*, Vol. 12, p. 982. https://doi.org/10.3390/rs12060982

Lal, P., Prakash, A., Kumar, A., Srivastava, P. K., Saikia, P., Pandey, A. C., … Khan, M. L. (2020). Evaluating the 2018 extreme flood hazard events in Kerala, India. *Remote Sensing Letters*, *11*(5), 436–445. https://doi.org/10.1080/2150704X.2020.1730468

Leela, R. D. (1986). History of Kerala. Retrieved June 20, 2021, from https://books.google.nl/books?id=pXpuAAAAMAAJ&q=inauthor:%22R.+Leela+Devi%22&dq=inauthor:%22R.+Leela+Devi%22&hl=nl&sa=X&redir_esc=y

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. Retrieved from http://arxiv.org/abs/1708.02002

Liu, B., Du, S., Du, S., & Zhang, X. (2020). Incorporating Deep Features into GEOBIA Paradigm for Remote Sensing Imagery Classification: A Patch-Based Approach. *Remote Sensing*, *12*(18), 3007. https://doi.org/10.3390/rs12183007

Mannor, S., Peleg, B., & Rubinstein, R. (2005). The cross entropy method for classification. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, 561–568. https://doi.org/10.1145/1102351.1102422

Marutho, D., Hendra Handaka, S., Wijaya, E., & Muljono. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. *Proceedings - 2018*

International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, ISemantic 2018, 533–538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

McGarigal, K. (2015). *Fragstats help version 4.2* (pp. 1–182). pp. 1–182. Retrieved from http://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf

Mobasheri, A., Zipf, A., & Francis, L. (2018). OpenStreetMap data quality enrichment through awareness raising and collective action tools—experiences from a European project. *Geo-Spatial Information Science*, *21*(3), 234–246. https://doi.org/10.1080/10095020.2018.1493817

Pan, Z., Xu, J., Guo, Y., Hu, Y., & Wang, G. (2020). Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sensing*, *12*(10), 1574. https://doi.org/10.3390/rs12101574

Panek, J. (2015). How participatory mapping can drive community empowerment - A case study of Koffiekraal, South Africa. *South African Geographical Journal*, *97*(1), 18–30. https://doi.org/10.1080/03736245.2014.924866

Panek, J., & Netek, R. (2019). Collaborative mapping and digital participation: A tool for local empowerment in developing countries. *Information (Switzerland)*, *10*(8). https://doi.org/10.3390/info10080255

Papathoma-Köhle, M., Neuhäuser, B., Ratzinger, K., Wenzel, H., & Dominey-Howes, D. (2007). Elements at risk as a framework for assessing the vulnerability of communities to landslides. *Natural Hazards and Earth System Science*, *7*(6), 765–779. https://doi.org/10.5194/nhess-7-765-2007

Parker, O. P. (2013). *OBJECT-BASED SEGMENTATION AND MACHINE LEARNING CLASSIFICATION FOR LANDSLIDE DETECTION FROM MULTI-TEMPORAL WORLDVIEW-2 IMAGERY.*

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

Pham, M. T., Courtrai, L., Friguet, C., Lefèvre, S., & Baussard, A. (2020). YOLO-fine: One-stage detector of small objects under various backgrounds in remote sensing images. *Remote Sensing*, *12*(15), 2501. https://doi.org/10.3390/RS12152501

Prathiba, A. P., Rastogi, K., Jain, G. V., & Govind Kumar, V. V. (2020). Building Footprint Extraction from Very-High-Resolution Satellite Image Using Object-Based Image Analysis (OBIA) Technique. In *Lecture Notes in Civil Engineering* (Vol. 33, pp. 517–529). https://doi.org/10.1007/978-981-13-7067-0_41

Qi, H. B., & Li, Z. L. (2008). An Approach to Building Grouping based on hierarchical constraints. *ISPRS Archives -- Volume XXXVII Part B2, 2008*, *37*(B2), 449–454.

Qi, W., Wei, M., Yang, W., Xu, C., & Ma, C. (2020, August 1). Automatic mapping of landslides by the ResU-Net. *Remote Sensing*, Vol. 12, p. 2487. https://doi.org/10.3390/RS12152487

Quan Luna, B., Blahut, J., Van Westen, C. J., Sterlacchini, S., Van Asch, T. W. J., & Akbas, S. O. (2011). The application of numerical debris flow modelling for the generation of physical vulnerability curves. *Natural Hazards and Earth System Science*, *11*(7), 2047–2060. https://doi.org/10.5194/nhess-11-2047-2011

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., & Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10008 LNCS*, 188–196. https://doi.org/10.1007/978-3-319-46976-8_20

Raviz. (2018). Economy of the Ashtamudi Lake Area | Kollam. Retrieved June 20, 2021, from https://www.theraviz.com/blog/economy-of-the-ashtamudi-lake-area/

Ribeiro, A., & Fonte, C. C. (2015). A methodology for assessing openstreetmap degree of coverage for purposes of land cover mapping. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*(3W5), 297–303. https://doi.org/10.5194/isprsannals-II-3-W5-297-2015

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sameen, M. I., & Pradhan, B. (2019). Landslide Detection Using Residual Networks and the Fusion of Spectral and Topographic Information. *IEEE Access*, *7*, 114363–114373. https://doi.org/10.1109/ACCESS.2019.2935761

See, L., Fonte, C. C., Antoniou, V., & Minghini, M. (2019, March 11). Volunteered geographic information: looking towards the next 10 years. *Journal of Geographical Systems*, Vol. 21, pp. 1–3.

https://doi.org/10.1007/s10109-018-00291-x

Shodhganga. (2019). THE PALAKKAD DISTRICT: AN OVERVIEW. Retrieved June 20, 2021, from https://shodhganga.inflibnet.ac.in/bitstream/10603/1313/9/09_chapter 2.pdf

Stewart, C., Lazzarini, M., Luna, A., & Albani, S. (2020). Deep learning with open data for desert road mapping. *Remote Sensing*, *12*(14). https://doi.org/10.3390/rs12142274

Stewart, R., Urban, M., Duchscherer, S., Kaufman, J., Morton, A., Thakur, G., … Stewart stewartrn, R. (2016). *A Bayesian machine learning model for estimating building occupancy from open source data*. *81*, 1929–1956. https://doi.org/10.1007/s11069-016-2164-9

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*.

Sun, Y., Shahzad, M., & Zhu, X. X. (2017). Building height estimation in single SAR image using OSM building footprints. *2017 Joint Urban Remote Sensing Event, JURSE 2017*. https://doi.org/10.1109/JURSE.2017.7924549

Tavakkoli Piralilou, S., Shahabi, H., Jarihani, B., Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., … Aryal, J. (2019). Landslide Detection Using Multi-Scale Image Segmentation and Different Machine Learning Models in the Higher Himalayas. *Remote Sensing*, *11*(21), 2575. https://doi.org/10.3390/rs11212575

TheHindu. (2019). Palakkad continues to remain flooded - The Hindu. Retrieved May 3, 2021, from https://www.thehindu.com/news/national/kerala/palakkad-continues-to-remain-flooded/article28968475.ece

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., … Shibasaki, R. (2018). Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks. *Remote Sensing*, *10*(3). https://doi.org/10.3390/rs10030407

Wu, T., Luo, J., Zhou, Y., Wang, C., Xi, J., & Fang, J. (2020). Geo-Object-Based Land Cover Map Update for High-Spatial-Resolution Remote Sensing Images via Change Detection and Label Transfer. *Remote Sensing*, *12*(1), 174. https://doi.org/10.3390/rs12010174

Xie, Y., Cai, J., Bhojwani, R., Shekhar, S., & Knight, J. (2020). A locally-constrained YOLO framework for detecting small and densely-distributed building footprints. *International Journal of Geographical Information Science*, *34*(4), 777–801. https://doi.org/10.1080/13658816.2019.1624761

Yang, Q., Zhang, S., Dai, Q., & Yao, R. (2020). Improved framework for assessing vulnerability to different types of urban floods. *Sustainability (Switzerland)*, *12*(18). https://doi.org/10.3390/su12187668

Yang, X., Li, X., Ye, Y., Lau, R. Y. K., Zhang, X., & Huang, X. (2019). Road detection and centerline extraction via deep recurrent convolutional neural network U-Net. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(9), 7209–7220. https://doi.org/10.1109/TGRS.2019.2912301

Yang, X., Li, X., Ye, Y., Zhang, X., Zhang, H., Huang, X., & Zhang, B. (2019). Road Detection via Deep Residual Dense U-Net. *Proceedings of the International Joint Conference on Neural Networks*, *2019-July*. https://doi.org/10.1109/IJCNN.2019.8851728

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., & Zhao, T. (2019). Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sensing*, *11*(15), 1774. https://doi.org/10.3390/rs11151774

Zeng, J., Qian, Y., Ren, Z., Xu, D., & Wei, X. (2019). Road landscape morphology of valley city blocks under the concept of "open block"-taking lanzhou city as an example. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226258

Zhang, L., & Pfoser, D. (2019). Using openstreetmap point-of-interest data to model urban change—a feasibility study. *PLoS ONE*, *14*(2), e0212606. https://doi.org/10.1371/journal.pone.0212606

Zhao, K., Kang, J., Jung, J., & Sohn, G. (2018). Building extraction from satellite images using mask R-CNN with building boundary regularization. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2018-June*, 242–246. https://doi.org/10.1109/CVPRW.2018.00045

Zhou, K., Chen, Y., Smal, I., & Lindenbergh, R. (2019). Building segmentation from airborne vhr images using mask r-cnn. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(2/W13), 155–161. https://doi.org/10.5194/isprs-archives-XLII-2-W13-155-2019

Zhou, X. (2018). Understanding the Convolutional Neural Networks with Gradient Descent and Backpropagation. *Journal of Physics: Conference Series*, *1004*(1), 12028. https://doi.org/10.1088/1742-6596/1004/1/012028

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017, December 1). Deep

Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, Vol. 5, pp. 8–36. https://doi.org/10.1109/MGRS.2017.2762307

APPENDIX

A. TIME FRAME FOR EACH PHASE

A.1 Detection Phase

The time frame for the detection phase can be divided into three parts: (1) training data generation, (2) model training, and (3) model prediction.

A.1.1 Training data generation

The time required for the training data generation by manually digitising the buildings from the satellite image takes around 5 hours. Apart from that, downloading OSM building footprints and correcting them geometrically takes around 1 hour. In total **6 hours**.

A.1.2 Model Training

As discussed in the limitations, the model takes around **24 hours** to train completely with the various combinations of hyper-parameters. However, with transfer learning, the total time could be as low as 2 hours when using trained weights from previous studies.

A.1.3 Model Prediction

Luckily, model prediction takes around 5 mins to complete in the entire study area of Palakkad. Therefore, combining with Kollam, the final prediction of buildings in the two areas take about **10 minutes**.

Total time in the detection phase = **30 hours and 10 mins**.

A.2 Characterisation Phase

The time frame for the characterisation phase can be divided into four parts: (1) Cleaning the prediction output, (2) downloading and generating primary data from the open sources (3) generating results from the Momepy library, (4) amalgamation of the results to generate final classification of building occupancy, and (5) local expert/stakeholder validation.

A.2.1 Cleaning the prediction output

Post-processing to remove building overlaps, multi-polygons, and false-positives takes about 3 hours for each study area. In total, **6 hours**.

A.2.2. Downloading and generating primary data from the open sources

Downloading primary data from OSM, Google Maps and the Bhuvan landuse NUIS data takes around 30 minutes together. However, generating the resulting data by manual digitisation for the Google Maps, landuse data, and generating the road networks takes around 4 hours. In total **4 ½ hours**.

A.2.3 Generating results from the Momepy library

The generation of the morphologically distinct clusters of buildings with the Momepy library takes around **2 hours**. First, the generation of tessellation and calculating the morphological metrics take around 1 ½ hours followed by clustering the buildings based on the calculated morphological metrics (half hour).

A.2.4 Amalgamation to generate building occupancy

The entire process of the voting system (figure 28) takes around 2 hours to interpret the majority clusters, building tags, landuse and the other information like the number of buildings, distance from the CBD to classify the built-up are blocks into the respective occupancy types. Takes **4 hours** in total including the two study areas.

A.2.5 Local expert/stakeholder validation

Getting in touch with the local expert to interpret, re-evaluate and re-classify the cluster meaning and the occupancy types took about **3 hours** over four sessions of discussions.

Total time in the characterisation phase = **19 hours and 30 minutes**.

A.3 Exposure Phase

The time frame for the detection phase can be divided into two parts: (1) downloading and generating flood-susceptibility maps and (2) using Risk Changes Desktop library to perform exposure assessment.

A.3.1 Downloading and generating flood- susceptibility maps

The flood susceptibility maps were downloaded from the KSDMA. The conversion of the flood map from a polygon to a raster and then finally generating the flood raster maps take about **10 minutes**.

A.3.2 Risk Changes Desktop library to perform exposure assessment

The use of the flood raster map with the final outputs of the characterisation phase to generate the exposure maps takes around **5 minutes**.

Total time in the exposure phase = **15 minutes.**

A.4 Total Time

The total time taken from all the three phases to arrive at a point of characterising homogeneous built-up area with building typology as an important attribute for studies and research on exposure assessment is

= Detection Phase + Characterisation Phase + Exposure Phase

= (30 hours + 10 mins) + (19 hours + 30 mins) + (15 mins) = 49 hours + 55 minutes

≈ 50 hours

≈ **2 days and 2 hours**.

B. RESOURCES AND MATERIALS USED

For the research, the following resources were utilised:

1. *Overpass API*: A read-only API tool of OpenStreetMap that provides custom selected parts of the OSM map data.

2. *ArcGIS*: A geographic information system software developed by ESRI for working with maps, compiling geographic data, analysing mapped information, and managing geographic information in a database.

3. *QGIS*: QGIS is a free and open-source cross-platform desktop geographic information system application that supports viewing, editing, and analysing geospatial data. Overpass API can be used as a plug-in in QGIS.

4. *Jupyter Notebook*: Open-source web application that allows the creation of python codes and leverage machine learning libraries like Keras and TensorFlow to develop deep learning algorithms for the research.

5. *Google Colab*: A product from Google Research that allows writing and executing arbitrary code through the browser for machine learning, data analysis and education purposes.

6. *Momepy Toolkit*: A library for quantitative analysis of urban form and urban morphometrics.

7. *Bhuvan web-portal data*: Urban land use data in the scale of 1:10,000.

8. *Google Maps and Open Street Map*: Map data that consists of geocoded address and place names, road routes, building footprints and labels and many other information.

9. *Risk Changes Desktop*: A library developed by ITC and AIT, Thailand that can be used to calculate and estimate the exposure, vulnerability, loss, and risk of EaR against hazards such as floods, landslides, earthquakes, and others.

C. SAMPLE CODE OF THE DEEP LEARNING MODEL

```
#Residual network as an encoder for U-Net
from tensorflow.keras.applications import ResNet50

def ResUNet(loss, pretrained_weights=None, input_size=(PATCHSIZE, PATCHSIZE, NBANDS)
):
    inputs = Input(input_size, name="input_image")

    encoder = ResNet50(input_tensor=inputs, include_top=False, pooling=None)
    #encoder.trainable=False
    for l in encoder.layers:
        l.trainable = False

    skip_connection_names = ["input_image", "conv1_relu", "conv2_block3_out",
                             "conv3_block4_out", "conv4_block6_out"]
    encoder_output = encoder.get_layer("conv5_block3_out").output

    f = [3, 64, 256, 512, 1024] #[16, 32, 64, 128, 256, 512, 1024, 2048] * 32
    x = encoder_output
    for i in range(1, len(skip_connection_names)+1, 1):
        x_skip = encoder.get_layer(skip_connection_names[-i]).output
        x = UpSampling2D((2, 2))(x)
        x = Concatenate()([x, x_skip])

        x = Conv2D(f[-i], (3, 3), padding="same")(x)
        x = BatchNormalization()(x)
        x = Activation("relu")(x)

        x = Conv2D(f[-i], (3, 3), padding="same")(x)
        x = BatchNormalization()(x)
        x = Activation("relu")(x)

    x = Conv2D(1, (1, 1), padding="same")(x)
    x = Activation("sigmoid")(x)

    model = Model(inputs, x)
    print(model.summary())
    model.compile(optimizer=optimizer, loss=loss, metrics=metrics)

    if(pretrained_weights):
      model.load_weights(pretrained_weights)
    return model
```

D. SAMPLE CODE FOR MORPHOLOGICAL METRICS

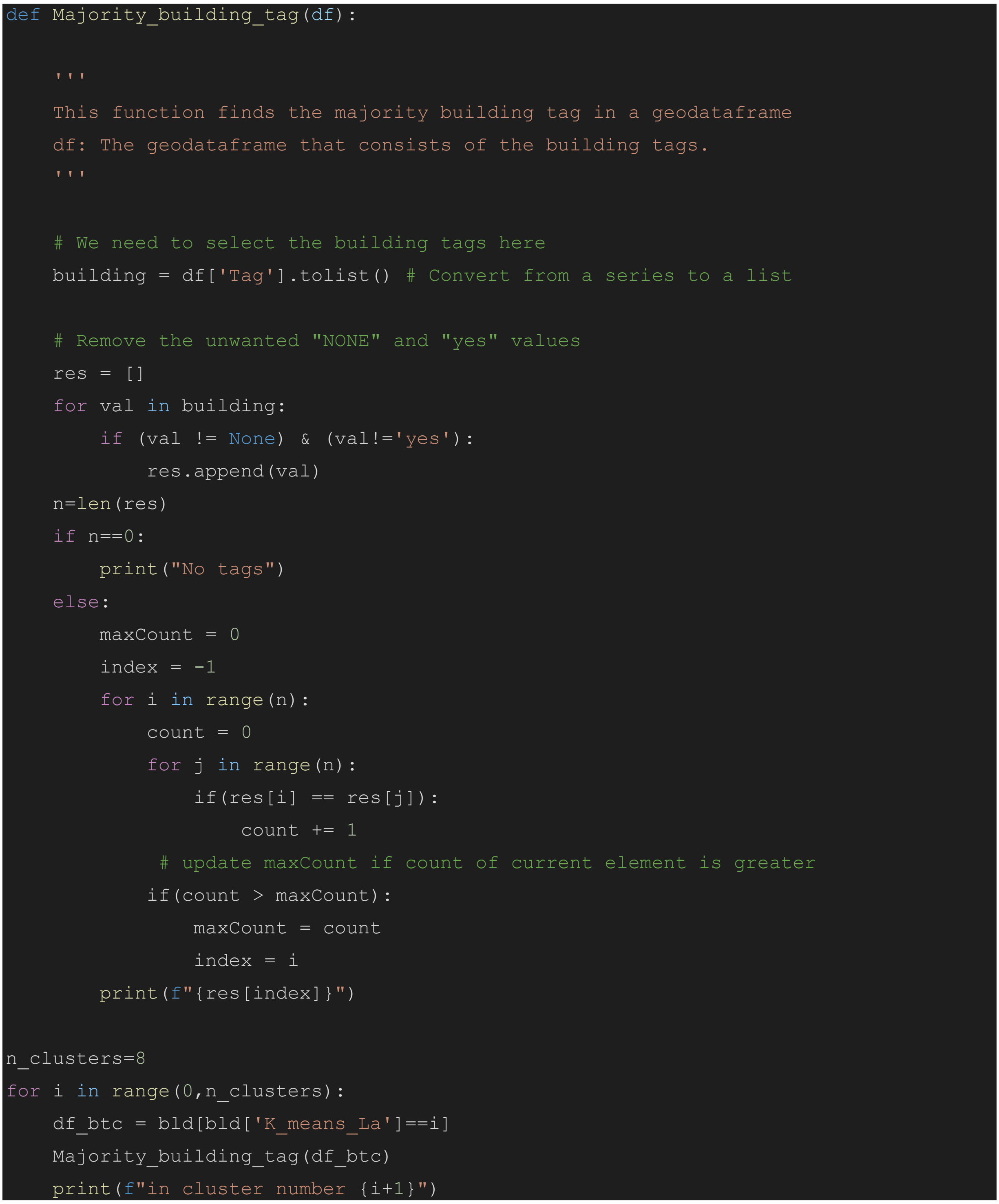
```
# Dimensions
clean_plyg["sdbAre"] = mm.Area(clean_plyg).series
clean_plyg["sdbPer"] = mm.Perimeter(clean_plyg).series

# Shapes
# clean_plyg["ssbCCo"] = mm.CircularCompactness(clean_plyg, "sdbAre").series
clean_plyg["ssbCor"] = mm.Corners(clean_plyg).series
clean_plyg["ssbSqu"] = mm.Squareness(clean_plyg).series
clean_plyg["ssbERI"] = mm.EquivalentRectangularIndex(clean_plyg, "sdbAre", "sdbPer")
.series
clean_plyg["ssbElo"] = mm.Elongation(clean_plyg).series
cencon = mm.CentroidCorners(clean_plyg)
clean_plyg["ssbCCM"] = cencon.mean
clean_plyg["ssbCCD"] = cencon.std

# Spatial distribution
clean_plyg["stbOri"] = mm.Orientation(clean_plyg).series
clean_plyg_tess["stcOri"] = mm.Orientation(clean_plyg_tess).series
clean_plyg["stbCeA"] = mm.CellAlignment(clean_plyg, clean_plyg_tess, "stbOri", "stcO
ri", "uID", "uID").series

# Tessellation dimensions and shapes
clean_plyg_tess["sdcLAL"] = mm.LongestAxisLength(clean_plyg_tess).series
clean_plyg_tess["sdcAre"] = mm.Area(clean_plyg_tess).series
clean_plyg_tess["sscCCo"] = mm.CircularCompactness(clean_plyg_tess, "sdcAre").series
clean_plyg_tess["sscERI"] = mm.EquivalentRectangularIndex(clean_plyg_tess, "sdcAre")
.series

# Intensity
clean_plyg_tess["sicCAR"] = mm.AreaRatio(clean_plyg_tess, clean_plyg, "sdcAre", "sdb
Are", "uID").series

queen_1 = Queen.from_dataframe(clean_plyg_tess, ids="uID")

clean_plyg["mtbAli"] = mm.Alignment(clean_plyg, queen_1, "uID", "stbOri").series
clean_plyg["mtbNDi"] = mm.NeighborDistance(clean_plyg, queen_1, "uID").series
clean_plyg_tess["mtcWNe"] = mm.Neighbors(clean_plyg_tess, queen_1, "uID", weighted=T
rue).series
clean_plyg_tess["mdcAre"] = mm.CoveredArea(clean_plyg_tess, queen_1, "uID").series

clean_plyg_queen = Queen.from_dataframe(clean_plyg, silence_warnings=True)
```

```
queen3 = mm.sw_high(k=3, weights=queen_1)

clean_plyg["ltbIBD"] = mm.MeanInterbuildingDistance(clean_plyg, queen_1, "uID", queen3).series
clean_plyg_tess["ltcBuA"] = mm.BuildingAdjacency(clean_plyg, queen3, "uID", clean_plyg_queen).series
```

E. SAMPLE CODE FOR CLUSTERING

```
s = {}
for n in range(2, 15):
    s[n] = []
    for r in range(5):
        kmeans = KMeans(n_clusters=n, random_state=r).fit(clean_norm3)
        labels = kmeans.labels_
        s[n].append(metrics.silhouette_score(clean_norm3, labels, metric='euclidean'))

vals = []
clus = []

for c, v in s.items():
    vals += v
    clus += [c] * 5

silhouettes3_clean = pd.DataFrame({'n_cluster': clus, 'silhouette': vals})
sns.lineplot(x="n_cluster", y="silhouette",
             data=silhouettes3_clean)

# Display
kmeans4 = KMeans(n_clusters=4, random_state=1).fit(clean_norm3)
clean_blg.plot(kmeans.labels_, figsize=(10,10), categorical=True, legend=True)
```

F. SAMPLE CODE FOR EVALUATING MAJORITY TAGS



```
def Majority_building_tag(df):

    '''
    This function finds the majority building tag in a geodataframe
    df: The geodataframe that consists of the building tags.
    '''

    # We need to select the building tags here
    building = df['Tag'].tolist() # Convert from a series to a list

    # Remove the unwanted "NONE" and "yes" values
    res = []
    for val in building:
        if (val != None) & (val!='yes'):
            res.append(val)
    n=len(res)
    if n==0:
        print("No tags")
    else:
        maxCount = 0
        index = -1
        for i in range(n):
            count = 0
            for j in range(n):
                if(res[i] == res[j]):
                    count += 1
             # update maxCount if count of current element is greater
            if(count > maxCount):
                maxCount = count
                index = i
        print(f"{res[index]}")

n_clusters=8
for i in range(0,n_clusters):
    df_btc = bld[bld['K_means_La']==i]
    Majority_building_tag(df_btc)
    print(f"in cluster number {i+1}")
```

G. SAMPLE CODE FOR EVALUATING MAJORITY LANDUSE

```
# Define a function to find the majority landuse
def Majority_landuse(df):
    '''
    This function finds the majority landuse in a geodataframe
    df: The geodataframe that consists of the landuse information.
    '''
    # We need to select the cluster number here
    landuse = df['Landuse']
    landuse = landuse.values.tolist()

    n=len(landuse)
    maxCount = 0
    index = -1
    for i in range(n):
        count = 0
        for j in range(n):
            if(landuse[i] == landuse[j]):
                count += 1
         # update maxCount if count of current element is greater
        if(count > maxCount):
            maxCount = count
            index = i
    print(landuse[index])

# Find majority landuse at block level
street_blocks=54

for blocks in range(0,street_blocks):
    df_landuse = bld[bld['Block_numb']==blocks]
    Majority_landuse(df_landuse)
#     print(f"in block number {blocks+1}")
```

H. SAMPLE CODE FOR THE EXPOSURE ASSESSMENT

```
# !pip install RiskChangesDesktop
import RiskChangesDesktop as RCD
from RiskChangesDesktop import DataManage
import Exposure
from RiskChangesDesktop import Loss
from RiskChangesDesktop import Risk

DataManage.CheckProjectionVector("data/footprint data/Kollam/footprints.shp")
DataManage.CheckProjectionVector("data/block data/Kollam/block.shp")
DataManage.MatchProjection("data/flood data/Kollam/flood_kollam.tif", "data/footprin
t data/Kollam/footprints.shp")

#Compute exposure at the block level
ear_block="data/block data/Kollam/block.shp"
susceptibility="data/flood data/Kollam/flood_kollam.tif"
ear_key_block="Block_numb"
outputname="data/exposure output/Kollam/exposure_block"
outputformat="shp"

Exposure.ComputeExposure(ear_block, susceptibility, ear_key_block, outputname, outpu
tformat)

#Compute exposure at the aggregated block level using the building footprints
ear_foot="data/footprint data/Kollam/footprints.shp"
susceptibility="data/flood data/Kollam/flood_kollam.tif"
ear_key_foot="OBJECTID"
admin_unit="data/block data/Kollam/block.shp"
agg_col="Block_numb"
outputname="data/exposure output/Kollam/exposure_assessed_foot_agg"
outputformat="shp"

Exposure.ComputeExposureAgg(ear=ear_foot, hazard=susceptibility, ear_key=ear_key_foo
t,
                            admin_unit=admin_unit, agg_col=agg_col, outputname=outpu
tname, outputformat=outputformat)
```

I. GITHUB LINK

The GitHub link contains the code and the data available to replicate the entire thesis. The data will be available till the end of July, after which interested parties would need to contact the author for the data. The GitHub link contains the codes for:

1. Building Detection
2. Building Characterisation
3. Exposure Assessment

LINK: https://github.com/kushanavbhuyan/Building-Identification-for-Exp-Vul-Risk-Assessment
Email: kushanavb@gmail.com