Cartography M.Sc.

Master thesis

Fine-scale Machine Learning Based Population Mapping

A Case Study of Munich

Zhuo Yang

Technical University of Munich

TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITY OF TWENTE. ITC

2022

Fine-scale Machine Learning Based Population Mapping

A Case Study of Munich

submitted for the academic degree of Master of Science (M.Sc.)
conducted at the Department of Aerospace and Geodesy
Technical University of Munich

Author:	Zhuo, Yang
Study course:	Cartography M.Sc.
Supervisor:	M.Sc Peng Luo (TUM)
Reviewer:	M.Eng Andrea Binn (TUW)

Chair of the Thesis
Assessment Board: Prof. Dr. Liqiu Meng

Date of submission: 13.09.2022

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Fine-scale Machine Learning Based Population Mapping: A Case Study of Munich"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 13.09.2022 Zhuo, Yang

Acknowledgement

With the completion of this thesis, my master's degree, which has been a life-changing experience for me, comes to an end. Everything I've accomplished over the last two years would not have been possible without the support of many people.

My first thanks go to Peng, my thesis supervisor. Without his close supervision throughout the thesis period, I would not have been able to complete the study successfully. I value his openness, wisdom, and patience at work, as well as his constant encouragement whenever I encountered a problem.

I would also like to express my profound thanks to Prof. Dr.-Ing. Liqiu Meng, my second supervisor. Her vast knowledge, guidance, and support provide me with a blueprint for my study, which has assisted me in determining what I should focus on when I am confused.

My sincerely gratitude also goes to Andrea, my thesis reviewer. During the proposal defense, mid-term presentation, and before thesis submission, I appreciated her advice, recognition, and encouragement.

I am also grateful for the opportunity to participate in the Cartography Master's program, which has been a truly unique experience for me. Juliane and the other four program coordinators deserve special recognition. We couldn't have survived in different universities in different places without their help. I would also like to thank my classmates with whom I had the pleasure to share this journey. I'll never forget the times we studied together, picnicked together, cooked together, and traveled together during the pandemic period.

Last but not least, I'd like to express my heartfelt gratitude to my family, boyfriend, and friends for always believing in me, supporting me, encouraging me, helping me. Love you!

Abstract

Fine-scale population mapping is of great importance to a wide spectrum of public interests. As census data is time-consuming and too expensive, many studies have been carried out to find other approaches as alternatives in the past decades. Especially in today's big data era, the conduction of high-resolution population mapping with the participation of remote sensing products and big geo data has become an active field of study. Compared with remote sensing data, big geo data has higher temporal and spatial resolution thus complementing the lack of semantic information which can indicate various human activities with a more precise location. Therefore, some scholars have utilized POI information, location-based social media data, together with remote sensing data to help conduct population mapping at a finer scale, like 100m grid-cell level and building level. However, there are three main problems: 1) collecting social media data is not always possible due to data privacy and ethical concerns; 2) optimal scales are rarely discussed, ignoring the fact that different features may have different acting ranges; and 3) at the building scale, population mapping results are frequently the disaggregation result of census data, rather than estimation. As a result, the investigation into integrating multiple freely available remote sensing products and semantic data using machine learning methods while considering optimal scales contributes to the enrichment of fine-scale population mapping products.

To accomplish this goal, a framework is proposed in Munich as the study area to incorporate collections of multi-source open remotely sensed products and semantic data to machine learning models. Following data preparation at the 100m and 50m gridded levels, the training process is carried out at the 100m gridded scale using five different machine learning models, followed by estimation at the 50m gridded scale using the best model Random Forest, and finally the results are allocated to building scale based on building volume. Additionally, Various visualization techniques were used on the final products, with the pros and cons of each method highlighted. Furthermore, the optimal scale for each POI category is identified, indicating that if POI is used in neighborhood environment investigation, the acting range difference should be considered. The significance of features and datasets has also been evaluated, with the result that OSM POI, OSM Building, and land use data help to

improve accuracy, implying that datasets with higher spatial resolution and more semantic information are more important in population mapping than other remotely sensed products, such as NTL and NDVI data. Furthermore, due to the high clustering of estimation errors, these variables do not perform equally well in different population density areas. However, it demonstrates the importance and characteristics of machine learning in population mapping by obtaining the spatial pattern of population distribution at a finer scale rather than the population value itself.

Keywords: population mapping, grid cell, building level, machine learning, Random Forest

Contents

Statement of Authorship i
Acknowledgement ii
Abstract iii
Contents v
List of Figures vii
List of Tables x
Abbreviations xi
1 Introduction 1
1.1 Motivation and Problem Statement 1
1.2 Research Identification 3
1.3 Thesis Outline 3
2 Related Work 5
2.1 Methods for Population Mapping 5
2.2 Integration of Ancillary Data for Population Mapping 8
2.3 Population Mapping at Building Level 11
2.4 Population Distribution Visualization 12
2.5 Summary 15
3 Methodology and Data Preparation 17
3.1 Workflow 17
3.2 Data Collection and Description 18
3.3 Data Preprocessing 23
3.4 Data Modeling 25
3.5 Model Evaluation 26
3.6 Visualization 26

4 Case Study and Results 28
4.1 Study Area 28
4.2 Data Modeling 28
4.3 Population Estimation 36
4.4 Population Visualization 40
5 Evaluation 47
5.1 Optimal Scale 47
5.2 Feature and Data Importance 49
5.3 Spatial Heterogeneity and Data Characteristics 51
5.4 Model Performance 55
5.5 Data Inconsistency and Quality Problem 56
6 Conclusion 58
6.1 Summary 58
6.2 Discussion 59
7 References 62
8 Appendix 69
8.1 Reclassification of OSM POI 69
8.2 Reclassification of Urban Atlas 2012 70

List of Figures

Figure 2.1 An illuminated choropleth map showing population density of counties in the conterminous United States.....14
Figure 2.2 A dasymetric map of London population density in year 2011 using residential area only 14
Figure 2.3 A proportional symbol map of global population distribution in year 2015 14
Figure 2.4 A cartogram map of global population distribution in year 2020.....14
Figure 2.5 A bivariate dot density map showing the relative concentrations of the Black and Hispanic populations in the United States in 2010 15
Figure 2.6 A heat map of population distribution in Guangzhou.....15
Figure 2.7 An pseudo-isarithmic map of population density in Germany.....15
Figure 2.8 A 3D mapping of population density in Europe.....15
Figure 3.1 Flowchart of this study.....17
Figure 3.2 Munich Stadtbezirke and 500m Grid Cells 18
Figure 3.3 Population Distribution of Munich 100m Grid Cells 18
Figure 3.4 Munich NDVI Data of 100m grid cells.....21
Figure 3.5 Munich NTL Data of 100m grid cells 21
Figure 3.6 CLC5 2018 of Munich.....21
Figure 3.7 Urban Atlas of Munich.....21
Figure 3.8 OSM POIs 22
Figure 3.9 OSM Building Sample.....22
Figure 3.10 Sample of CORINE dataset.....23
Figure 3.11 Sample of Urban Atlas dataset 23
Figure 4.1 Distribution of all features 29
Figure 4.2 Distribution of population density.....30
Figure 4.3 Information comparison from google map.....30
Figure 4.4 Correlation matrix of features 31

Figure 4.5 Connection of features with strong correlation 32
Figure 4.6 Scatter plot of true and predicted values 35
Figure 4.7 Histogram of Absolute Error between true and predicted population 36
Figure 4.8 Histogram of Population true values 37
Figure 4.9 P Histogram of Error between true and predicted population 37
Figure 4.10 Estimation ABSE and population distribution 37
Figure 4.11 Scatter plot of population true values and estimated errors 38
Figure 4.12 Error distribution in no population area 38
Figure 4.13 Histogram of error distribution in no population area 39
Figure 4.14 Land use of no population area in Munich 40
Figure 4.15 Building distribution in Munich 40
Figure 4.16 NTL index distribution in Munich 40
Figure 4.17 NDVI index distribution in Munich 40
Figure 4.18 Retail POI density distribution in Munich 40
Figure 4.19 Sport POI density distribution in Munich 40
Figure 4.20 Spatial distribution of predicted population 41
Figure 4.21 Spatial distribution of true population from Census 2011 41
Figure 4.22 Choropleth map of population distribution in 100m*100m grid cells of Munich 42
Figure 4.23 Dasymetric map of population distribution in 100m*100m grid cells of Munich 43
Figure 4.24 Dot map of population distribution of Munich 43
Figure 4.25 Proportional symbol map of population distribution of Munich 44
Figure 4.26 Heat map of population distribution of Munich 44
Figure 4.27 3D hexbin map of population distribution of Munich 45
Figure 4.28 Choropleth map of population distribution at building scale of Munich 45
Figure 4.29 3D hexbin map of population distribution at building level of Munich 46
Figure 4.30 3D hexbin map of population distribution at building level of Munich (part) 46

Figure 5.1 Retail POI density in six different ranges 48
Figure 5.2 Top 15 important features 50
Figure 5.3 Spatial auto correlation report 52
Figure 5.4 Overlay of ABSE and NTL dataset 54
Figure 5.5 Overlay of ABSE and NDVI dataset 54
Figure 5.6 Scatter plot of true and predicted population 55
Figure 5.7 Overlay of land use and no population area fall into residential area 56
Figure 5.8 Overlay of land use and grid cells with population fall into non-residential area 56
Figure 5.9 Overlay of land use and OSM Building 57
Figure 5.10 Overlay of land use and grid cells with population fall into non-residential area 57
Figure 8.1 OSM POI of Munich after reclassification 70
Figure 8.2 Urban Atlas 2012 of Munich after reclassification 71

List of Tables

Table 3.1 Sample data of German Census 2011 at 100m grid cell level.....19
Table 3.2 Overview of Data Used in This Study 22
Table 3.3 Sample result of data aggregation.....24
Table 4.1 VIF of features 33
Table 4.2 Evaluation of the applied five models.....35
Table 4.3 Pros and cons of the applied visualization methods at 100m*100m grid cell level.....41
Table 5.1 POI category and its optimal scale 49
Table 5.2 Results of different datasets 50
Table 8.1 Explanation of reclassification of OSM POI 69
Table 8.2 Explanation of reclassification of Urban Atlas 2012 70

Abbreviations

NTL	Night time Light
NDVI	Normalized difference vegetation index
POI	Point of interest
VGI	Volunteered Geographic Information
VIIRS	Visible infrared imaging radiometer suite
RF	Random Forest
GEE	Google Earth Engine
OSM	OpenStreetMap
SVR	Support Vector Regression
KNN	K Nearest Neighbors
RTUD	Realtime Tencent User Density
MODIS	Moderate Resolution Imaging Spectroradiometer

1 Introduction

1.1 Motivation and Problem Statement

Fine-scale population product is essential for many use cases. For example, it can largely support better resource allocation in urban-rural management, provides great help for natural disaster preparation, monitor human environment interactions and measure impactions of population growth (Stevens, Gaughan, Linard, & Tatem, 2015). Census data has always been the primary population product for fine-scale, however, the collection of census data is time-consuming and expensive. Most of the census data is conducted every ten years which makes it difficult to meet the increasing needs of more contemporary, and easily-updatable, as the use cases are far more complex than before. Moreover, census data is often based on administrative units which usually contains a large area thus hardly to allow a close look at detailed spatial distribution. Furthermore, many areas are not fully covered by census because of cost or privacy when comes to a finer scale. For example, the German Census 2011 stated that the reliability of some gridded-based data is limited due to statistical confidentiality under Article 16 of the Federal Statistics Law.

Given that, many scholars have made large efforts to search other approaches as alternatives. Among them, gridded population mapping (Leyk, et al., 2019) became extremely popular nowadays with the help of increasing accessibility of geographical big data with varying spatial and temporal resolution, as well as the blooming of novel techniques such as machine learning. Many researches have been carried out with the integration of multi-sourced geo data and population mapping models. It has been proven that machine learning methods like random forest model can bring out good results with the help of ancillary data for population mapping.

Most studies, however, use administrative unit level data as true values and assess modeling results at this level, making it difficult to examine real model performance because it often covers a much larger area than gridded cells. Alternatively, the evaluation is frequently carried out at the same gridded level by using the test dataset from the train-test-split method, which assesses performance on a very small subset of the data. Both methods have limitations in uncovering model performance, which may lead to incorrect conclusions. Because the German Census 2011 provides gridded results at the 1km and 100m levels, machine learning-based population mapping is possible on a very fine-scale of data training and evaluation, revealing more details about the machine learning method in population mapping.

On the other hand, variable optimal scales are rarely discussed when applying them to population mapping. Take point of interest data (POI) as example, they were widely

used as semantic ancillary information in population mapping as it has higher spatial resolution which can indicate human activities in a more precise manner (Ye, et al., 2019), however, most of previous studies chose the same scale when calculating its densities, ignoring the fact that POI has many types like accommodation, retail, hospital, and their neighbourhood characteristics vary in different acting ranges (Cheng, Zhang, & Huang, 2022). Therefore, the investigation of variable optimal scales need to be considered when integrating various data.

Other than that, spatial heterogeneity is another area that haven't been frequently talked in previous studies (Wang, Wang, Li, Cai, & Kang, 2022). Most of the existing studies are conducted within small areas which have less spatial heterogeneity, however, as different density population areas have different patterns, and different geospatial data captures different ground truth, especially when it comes to population distribution (Liu, et al., 2015), it is important to include it in the study thus increasing the credibility of the framework.

Lastly, considering that building level population mapping products are not rich enough due to the limitations in acquiring sufficient information. For instance, the 3D building data does not exist in every context (Pajares, Muñoz Nieto, Meng, & Wulfhorst, 2021). Nevertheless, the constantly growing volunteered geographic information (VGI) has enabled building level population mapping by providing various building attributes, like footprint and height. Therefore, it is possible to bring out a population mapping results at building level today.

In accordance with the proposed scientific perspectives, this study selects Munich as the research site for the following reasons:

- German Census 2011 published detailed population statistics at a 100-meter grid cell level, allowing for very fine-scale data modeling and results assessment rather than the administrative unit level;
- Despite Munich's dense population, there are many areas with no population and low population density that can be used to investigate spatial heterogeneity discussion.

As a result, the goal of this thesis is to map the population at fine-scale while optimizing the scale for corresponding variables and holding a discussion on spatial heterogeneity and data characteristics.

1.2 Research Identification

1.2.1 Research Objectives

The primary goal of this study is to propose a novel framework for estimating 50m*50m grid cell level populations using widely available free data, such as night-time light data, land use data, NDVI data, POI data, and building data, and then allocating the predictions to building scale. The final results will be visualizations of Munich's population at the building and 100m grid cell levels. Furthermore, it will investigate the significance and influence of various datasets in various areas, as well as machine learning methods for fine-scale population mapping.

1.2.2 Research Questions

It can be further detailed by following research questions:

- What is the approach for fine-scale population mapping?
- What is the optimal scale for each variable?
- Does certain data improve the accuracy in population mapping?
- Do they perform same level of accuracy in different areas?
- Is the result derived from machine learning methods has reasonable spatial details?

1.2.3 Innovation

This study aims to provide an approach for obtaining high-precision population maps at the building level by integrating remotely sensed products and big geo data into machine learning models.

Instead of disaggregating census data to gridded units and evaluating performance at the administrative unit level, the modeling process is based on 100m*100m grid cells, and the prediction is performed on 50m*50m grid cells, which are then aggregated to 100m gridded cell level to evaluate performance in a much larger area and at a much finer scale. Furthermore, the optimal scales for POI classes are also considered in the modeling process when using the OSM POI dataset, which is rarely discussed in population mapping. Finally, a wide range of result analysis is performed to gain insights on key features, datasets, and spatial heterogeneity.

1.3 Thesis Outline

The thesis is divided into six chapters that follow a logical structure that aims to establish a framework for fine-scale population mapping while discussing accuracies,

uncertainties, and spatial heterogeneity using an example to explain the entire process.

Chapter 1 introduces the study's motivations and addresses the research objectives and research questions, providing a general overview of the background of this thesis and what needs to be accomplished in the subsequent steps. The second chapter summarized the most recent population mapping research, including methods, popular datasets, fine-scales, and visualization methods. By presenting the related work, it provides a detailed impression of what has been accomplished thus far, what gaps must be filled, and what we can do to enrich the field. Chapter 3 introduces the data and methodology used in this study, as well as the workflow. The experiment in the study area is presented in Chapter 4, which includes data modeling, estimation evaluation, and population visualization. In chapter 5, the findings are categorized as optimal scale, feature importance, spatial heterogeneity, and data inconsistency. Finally, in Chapter 6, there is a follow-up summary and discussion.

2 Related Work

This chapter reviews the previous researches related to this study. It firstly introduced the methods in population mapping including conventional approaches and most recent popular procedures. Then the ancillary data widely used in population mapping is presented. Subsequently, the introduction of population mapping at building level is brought out which covered both of the data and methods in this field. Lastly, it summarized the fine-scale population visualization techniques.

2.1 Methods for Population Mapping

Spatial disaggregation methods are usually used to transform coarse scale spatial information to fine scales when facing a lack of high-resolution data (Qiu, Zhao, Fan, & Li, 2019). As census data is often available on administrative unit level, and normally municipality is the lowest level of official territorial division, these methods are also applied in fine-scale population mapping. Among them, areal interpolation and statistical modeling methods are the most widely used ones.

2.1.1 Areal Interpolation

Areal interpolation approach has developed for years, it means the transition of data from a source zone to many small subzones while intersecting with each other (Goodchild & Siu-Ngan Lam, 1980). It can be further classified into areal interpolation without and with ancillary data. Areal weighting approach without ancillary data assumes population is redistributed to all land surfaces cells evenly and does not use ancillary data (Doxsey-Whitfield et al., 2015). The Gridded Population of the World (GPWv4, https://sedac.ciesin.columbia.edu/data/collection/gpw-v4) released by the Center for International Earth Science Information Network (CIESIN) is the representative of this method while applying a water mask at the same time. The main advantage of this approach is the maintenance of the fidelity of the input data (Doxsey-Whitfield, et al., 2015) which helps to disaggregate census information in a large area especially at a global level. However, this approach does not perform as well when there are abrupt variations in value as it assumes homogeneity of population distribution whereas it is not in reality (Huang, Ottens, & Masser, 2007). Consequently, ancillary data such as land cover, road network data are integrated to this approach to improve the performance (Flowerdew & Green, 1994). Dasymetric mapping is one of most popular approaches of this category. It refines the weighting procedure by incorporating spatial ancillary data that is presumably related to population presence and density, therefore it involves redistribution weights in each cell (Eicher & Brewer, 2001).

More specifically, dasymetric mapping can be grouped to two classes, binary approach and weighted approach. Binary dasymetric approaches redistribute population using one or more ancillary layers that describe presence or absence of populated areas (Frantz, et al., 2021), the ancillary data is used to identify built-up cells that population can be redistributed to. In this way, on the basis of areal interpolation approach, it is done by discarding uninhabited areas reached from external data (Jain & Tiwari, 2017). Weighted dasymetric approaches assume that population density is unequally related to various variables (Nagle, Buttenfield, Leyk, & Spielman, 2014), such as road density, and population is then redistributed based on a cell-based weight. Large amount of literatures has proved that the spatial accuracy is improved by dasymetric techniques compared with conventional mapping techniques (Gallego, Batista, Rocha, & Mubareka, 2011; Langford, 2007)

The following is a typical formulation of many dasymetric population mapping applications:

$$P_i' = P_s \times \frac{A_i \times W_i}{\sum i\,(A_i \times W_i)}$$

where P_s is the known population of the source zone s and P_i' means the estimated population in the target zone. The source zones are discrete geographical units used to collect and/or report statistical data, whereas the target zones are finer zones within each source zone formed by the intersection of the source zone and ancillary spatial data (e.g. land use polygons). Finally, A_i denotes the target zone's area, and W_i is a weighting parameter related to the target zone's population density. It is important to note that the sum of the areas of the target zones within a source zone equals the area of the source zone, $\sum i\, A_i = A_s$ (Batista e Silva, Gallego, & Lavalle, A high-resolution population grid map for Europe, 2013).

In general, dasymetric mapping takes into consideration the spatial heterogeneity of population distribution to a cetain extent. However, dertermining population weights for each subdistrict remains difficult and does not capture the diversity of its subdistricts as the precision of the simple interpolation method is low and complex interpolation necessitates the use of multiple ancillary data sets. Therefore, although areal weighting method is simple and requires very few variables, it is hard to reflect the details at a grid cell location and actually population distribution is not homogeneous in reality (Luo, Zhang, Cheng, & Sun, 2019).

2.1.2 Statistical Modeling

Most recent studies propose combining several approaches to improve accuracy, commonly beginning with a dasymetric process, and then a statistical model. It is

founded on the notion that morphological characteristics such as the accessibility to the services, the transportation network influence the distribution in a region (Gallego, Batista, Rocha, & Mubareka, 2011). In short, this approach involves creating a regression model by using training data with the assumption that no population is assigned to places with no road network or built-up land cover (Langford, 2006). In this way, it simulates population spatial distribution by constructing fitted regression models with various variables, which then can both disaggregate population data to finer scales as well as estimate intercensal population or population of difficult-to-enumerate places (Wu, Qiu, & Wang, 2005).

The most common modeling technique is linear regression (Yao, et al., 2017). This approach is easy to model, and can easily identify meaningless negative population values in some rural areas due to this simple relationship between the variables which is impossible in reality (Chu, Yang, & Chou, 2019). However, it requires a linear relationship between the features and the population, and no covariance between the features can exist. As a result, it is usually applied to population disaggregation at large scales (Zeng, Zhou, Wang, Yan, & Zhao, 2011). Another popular model is logistical regression model (Gaughan, Stevens, Linard, Jia, & Tatem, 2013). However, if the predictors have a multivariate normal distribution, the solution may be more stable. Furthermore, as with other types of regression, even weak multicollinearity among predictors can result in biased estimates and inflated standard errors. (Nong & Du, 2011).

Then applying machine learning algorithms to population mapping became a hotspot in the light of increasing accessibility of big data (Šimbera, 2020). A common way is Random Forest (RF) model which evolved from decision trees and has strong generalization ability, allowing it to handle high-dimensional features and effectively improve population spatialization accuracy (Breiman, 2001). In population mapping, it uses population counts from census data and a weighting scheme where each weighting variable is predicted by using ancillary data such as land cover information, nighttime lights data, topographic information or vector-based features (Stevens, Gaughan, Linard, & Tatem, 2015). Currently, RF based population estimation is an important direction for population spatialization research (Sinha, et al., 2019; Ye, et al., 2019; He, Xu, & Li, 2020). As the accuracy of RF results is highly influenced by auxiliary data, most of these researches focus on how to effectively integrate multiple-sourced data and construct model features. Random Forest, on the other hand, typically models only numerical attribute information from the data and fails to take into account the potential influence of spatially data on each other. Furthermore, due to a lack of fine-scale training data, RF typically uses administrative unit data for modeling and then migrates the models to the grid for prediction, resulting in cross-scale problems between training and prediction (Mei, et al., 2022; Sinha, et al., 2019).

Deep learning has gradually been applied to the study of population spatialization in recent years (Robinson, Hohman, & Dilkina, 2017; Tiecke, et al., 2017; Zhao, Liu, Zhang, & Fu, 2020). It has been proven that deep neural networks outperform shallow machine learning models like Random Forest in national-level population spatialization with a higher estimation accuracy (Zhao, Liu, Zhang, & Fu, 2020). However, this method also involves many limitations. For example, Convolutional Neural Networks (CNN) can effectively capture the spatial autocorrelation properties of the learned grid population through convolutional operations, but this approach has not been widely used due to the difficulty of obtaining fine-scale population training samples (Robinson, Hohman, & Dilkina, 2017). The Facebook Connectivity Lab applied computer vision techniques to extract building outlines from high-precision remote sensing images and then assigned population to buildings, however, it is not easy to achieve accurate population-to-building assignment without better integration of more auxiliary information (Tiecke, et al., 2017).

2.2 Integration of Ancillary Data for Population Mapping

Without relevant geospatial data, fine-scale population mapping cannot be studied. Especially with the rapid advancement of remote sensing technology, volunteered geographic information, and global positioning technology over the past decades, many different types of ancillary data have been used in the above approaches (Murakami & Yamagata, 2019).

2.2.1 Remotely Sensed Products

As the increased availability and spatial granularity of remotely sensed information about vegetation and land cover (Frye, Wright, Nordstrand, Terborgh, & Foust, 2018), remote sensed products have been applied for years either as the main source for population estimation (Schug, Frantz, van der Linden, & Hostert, 2021; Sutton, Roberts, Elvidge, & Baugh, 2001) or alternatively used as a supplementary data source for use in spatially refining census population estimates (Chen, 2002).

The most popularly used remotely sensed product is land cover dataset. It was used to lessen the overglow effect and to differentiate between urban and rural environments (Sun, Zhang, Wang, & Cen, 2017). Furthermore, some fine scale products include precise information regarding land use that may be utilized to calculate population distribution, such as whether people would dwell in residential regions rather than sports areas. It is used as the primary mask in dasymetric mapping to filter out no-population regions.

The night-time light (NTL) imagery is another widely used product. Nocturnal light has arisen as a characteristic of contemporary civilization, providing a distinct feature for

detecting the presence of human activity (Sutton, Elvidge, & Obremski, 2003). Several satellite sensors for NTL combines the benefits of RS with the unique feature of nocturnal light which acts as an alternate manner to efficiently to map worldwide human habitation and human activities. The visible infrared imaging radiometer suite (VIIRS) carried by the National Polar-Orbiting Operational Environment Satellite System is a new kind of satellite sensor that captures NTL data (Sun, Zhang, Wang, & Cen, 2017) and is widely utilized in research.

Considering the limitations of some remote sensed products, like the underestimation of NTL in rural regions and saturation problem in high population density areas, some other products have been taken as supplements. For example, based on the concept that plant quantity is closely and inversely connected with impervious surface, vegetation index is used to combat the overglow impact of NTL images (Pozzi & Small, 2005). With its help, exaggerated illuminated regions like lakes and woods that were lighted by light reflection from urban area can be diminished (Sun, Zhang, Wang, & Cen, 2017). One of the popularly integrated dataset is Normalized Difference Vegetation Index (NDVI) derived from satellite images (Luo, Zhang, Cheng, & Sun, 2019).

However, because most free remotely sensed products have a limited spatial resolution when compared to big geo data such as POI, it has limitations in showing various human activities occurring in a certain geographic place. (Ye, et al., 2019).

2.2.2 Geospatial Big Data

The utilization of enormous volumes of multi-sourced geospatial big data has attracted significant interest in numerous sectors, including population mapping, in the big data age. Geospatial data is used to describe objects and things in relation to geographic space, often with the use of location coordinates in a spatial reference system. Ground surveying, hotogrammetry, and remote sensing are common methods for gathering geospatial data, but more recently, laser scanning, mobile mapping, volunteered geographic information (VGI), and global navigation satellite system (GNSS) have emerged as new methods for gathering large amounts of geospatial data (Evans, Oliver, Zhou, & Shekhar, 2019). Compared with conventional geospatial data, big geospatial data has a higher temporal and spatial resolution which can be a good complement in population mapping study (Ye, et al., 2019). Because VGI and mobile mapping are the most commonly utilized geospatial data in population mapping, we have only mentioned the applications of these two geospatial data sources.

1) Volunteered Geographic Information

Volunteers, rather than traditional data producers, create VGI through Web 2.0 applications (Goodchild M. F., 2007). Though it has been critiqued for potential data quality issues (Jackson, et al., 2013), it is nonetheless a useful addition to official statistics and commercial data (Goodchild M. F., 2007). Consequently, numerous researchers in the field of population mapping regard it as a valuable data source, and combine it with other remotely sensed products to improve the accuracy of population mapping.

Points of Interest (POI) is the most commonly used VGI. A point of interest (POI) is a map feature with precise point coordinates as well as rich spatial semantics, such as restaurants, hospitals, or hairdressers (Bakillah, Liang, Mobasheri, Jokar Arsanjani, & Zipf, 2014). POIs can be used to revise population estimates if a correlation can be established because certain types of places are associated with a higher population density (Zhang & Qiu, 2011), and high population density areas tend to have more POIs. For example, POI and multisensory remote sensing data are combined to assist in the disaggregation of population distribution from census to 250m*250m gird cells (Yang, et al., 2019). Additionally, temporal changes of population density in Europe is uncovered by applying POIs as well as land cover data in a multi-layered dasymetric approach at 1 km^2 resolution (Batista e Silva, et al., 2020).

Most studies, however, ignored the fact that different types of POIs have varying degrees of attractiveness to the population, and instead created quantity/density metrics based on POI data within a fixed same range (Cheng, Zhang, & Huang, 2022). For example, because a hospital is designed to cover a larger area than a restaurant, their densities cannot be calculated within the same acting range and then used to support population distribution estimation.

2) Location Based Data

Thanks to the global positioning system (GPS), Mobile location-based service (LBS) technology has advanced significantly in recent years which allows us to obtain a variety of position data, such as public transit trajectories, mobile communications, and check-in data. These data sets may be merged to uncover information about urban structures, economic vibrancy, and traffic congestion (Hu, Wang, & Li, 2014; Liu, et al., 2015) that are connected to numerous human activities at the microscale and accurately depict population distributions and human behaviors.

Location-based data has also been used in population mapping to obtain finer scale findings, such as 100m*100m grid cell level, building level population distribution. For example, the Realtime Tencent User Density (RTUD) provided by one of China's largest internet companies records the locations of smart phone users using Tencent

applications such as WeChat, allowing the calculation of down-scaling the street-level population distribution to the 25m*25m grid cell level. (Yao, et al., 2017). Furthermore, there are also many studies directly applying these location data to dynamic population mapping (Deville, et al., 2014), especially real-time population during disasters. For example, as the RTUD is updated on a regular basis, it has also been used to map the monthly population distribution and variance across China at 1-km resolution. (Cheng, Wang, & Ge, 2022). Additionally, by combining phone signal data with building data, an estimated real-time population distribution map is created that can be used to aid in earthquake emergency rescue work (Xia, Nie, Fan, & Zhou, 2020).

However, there are some clear drawbacks to mapping populations using cell phone data. First, a recent research found that the low correlation between call volume and underlying population demonstrates the insufficiency of considering mobile subscriber distribution as a representation of population distribution (Kang, Liu, Ma, & Wu, 2012). Furthermore, due to the government's personal privacy regulation, acquiring mobile phone data is difficult.

2.3 Population Mapping at Building Level

To date, most of population mapping studies focused on global population and upper level gridded population, few considered fine-scale population mapping at the local scale, largely because of a lack of reliable data and models (Yao, et al., 2017). A few studies conducted building level population mapping in small areas but using census data and detailed building information from city administration office (Lwin & Murayama, 2009) which has a lot limitations. With the improving of spatial resolution of remote sensing products, some scholars have extracted building footprints and heights in residential zones from aerial images thus conducting population mapping at building level (Ural, Hussain, & Shan, 2011).

Meanwhile, as geospatial big data booms, volunteered geographic information (VGI) can now be used to solve this problem. For example, a satisfactory mapping result has been obtained in Hamburg by applying POIs, building footprints, and fine land use/land cover data (LULC) to an areal interpolation approach (Bakillah, Liang, Mobasheri, Jokar Arsanjani, & Zipf, 2014). Another approach is made in the City of Munich, Freising and Fürstenfeldbruck in southern Germany by utilizing census data on the district level, land-use data, OpenStreetMap (OSM) (Pajares, Muñoz Nieto, Meng, & Wulfhorst, 2021).

However, most of the existing studies are conducted within small areas which is from disaggregating census data instead of estimations. Additionally, many are using location based social media data which is not always accessible and involves data

privacy. Therefore, we need to find another alternatives and can be applied to large areas as well.

2.4 Population Distribution Visualization

In a scientific setting, good visualizations help us comprehend the underlying data and capture the reader's attention without sacrificing truth for beauty (Riffe, Sander, & Klüsener, 2021) as the most commonly accepted argument says human brain processes visual information much faster than text (Healy, 2018). Visualizations are also frequently used to easily examine and uncover patterns in population data, as well as to assist viewers in gaining a better knowledge of the features to a large audience.

As a typical use case of thematic mapping, various visualization approaches are applied in population distribution visualizations, such as choropleth map, dasymetic map, isarithmic map, proportional symbol map, cartogram, dot map, heat map, and 3D mapping.

Choropleth mapping represents relative quantities in census or administrative areas using graded color or gray tone. A suitable color scheme should ideally represent the quantitative character of the data represented on choropleth maps (Slocum, McMaster, Kessler, & Howard, 2022). As a result, if the legend is improperly created, it may be rendered ineffective. To create an eye-catching image, an illuminated choropleth map is created by combining classic choropleth symbology with an illuminated 3-D prism view (Slocum, McMaster, Kessler, & Howard, 2022).

When the phenomena are not equally distributed between enumeration units, the dasymetric map is an alternative to the choropleth map. A dasymetirc map, like a choropleth map, presents standardized data using areal symbols, but the borders of the symbols do not always match the bounds of enumeration units. A dasymetric map is most commonly used to depict the attribute population (Slocum, McMaster, Kessler, & Howard, 2022). Additionally, a dasymetric map of population density requires supplementary information for the redistribution process, such as land cover. For example, the population density of a water land cover will be set to zero if no people live on the water.

The proportional symbol map is used to depict numerical data connected with point locations, whether those are genuine point locations or conceptual point locations such as the centroids of countries (Cabello, Haverkort, Van Kreveld, & Speckmann, 2010). Although a vast range of geometric and pictographic symbols are possible, circles are the most widely used proportional sign. Classed proportional symbol maps, also known as range-graded or graduated symbol maps, do not reflect actual data

relations, but rather employ a set of symbols that clearly distinguish each class from one another. As a result, classed maps can improve spatial pattern perception.

Cartographers may purposefully distort space based on attribute values, and the resulting map projection is known as a cartogram or value-by-area map. The area cartogram is the most commonly discussed type of cartogram in the geographic literature. It is generated by scaling enumeration units in response to the values of an attribute linked with the enumeration units (Sun & Li, 2010). For example, the size of the countries may be proportionate to their population. Cartograms have the unusual advantage of displaying small enumeration units with high attribute values that would be buried on a comparable projection. Population is a common attribute depicted on area cartograms.

Dot map is a method of discovering spatial patterns or the distribution of data over a geographical region by distributing equally sized points over the region (Kimerling, 2009). Dot maps are classified into two categories: one-to-one (one point indicates a single count or object) and one-to-many (one point represents a certain unit, such as 1 point equals 20 people). Dot maps are great for seeing how items are dispersed throughout a geographical region and can reveal patterns when the points cluster on the map, however, they are not ideal for demonstrating precise values.

Heat map depicts the magnitude of a phenomena in two dimensions with colors[1]. The color change may be via hue or intensity, providing the reader with apparent visual indications regarding how the occurrence is clustered or fluctuates across space. The cluster heat map and the spatial heat map are two fundamentally distinct types of heat maps. A spatial heat map depicts the magnitude of a geographical phenomenon as color on a map.

An isarithmic map depicts a continuous field by connecting points of equal value with line and/or region symbols. In general, these maps are used to aid in the visualization of continuous data sets through the use of color, particularly hue and value. There are several ways to portray isarithmic maps; regardless of the design style used, the phenomena being represented must be a quantitative continuous field. It was initially employed in population distribution by Dane Ravn in 1857, but it has been heavily criticized by geographers, who believe that particular population is not a point function but an area function, and that population distribution does not fluctuate continuously (Nordbeck & Rystedt, 1970). Other cartographers and geographers, however, challenged the criticisms of the isarithmic mapping approach, claiming that the critics reached wrong conclusions from a logical position when they accused the designers

[1] https://en.wikipedia.org/wiki/Heat_map

of isarithmic maps of logical mistakes. This debate over the existence of population isarithms continues to this day. Some researchers even refer to them as fake or pseudo-isarithms and advise geographers not to utilize them (Nordbeck & Rystedt, 1970).

Three-dimensional models offer a volumetric depiction of space, which is useful for a range of applications such as population density. Many map aficionados have been working hard in recent years to create attractive 3D visualizations of population distribution that emphasize demographic trends and physical limits. One of them is Alasdair Rae, who has created a series of 3D population density maps for various places across the world.

Figure 2.1 An illuminated choropleth map showing population density of counties in the conterminous United States.

Data Source: (Stewart & Kennelly, 2010)

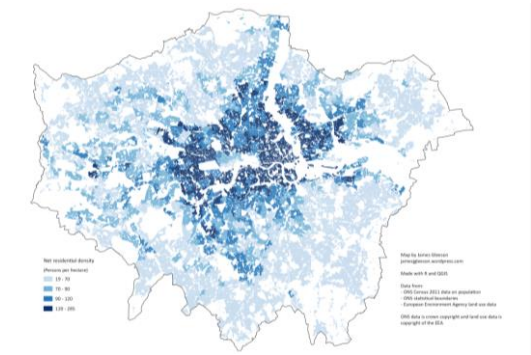


Figure 2.2 A dasymetric map of London population density in year 2011 using residential area only

Data Source: https://jamesjgleeson.wordpress.com/2013/01/23/dasymetric-map-of-londons-population-density-2011/

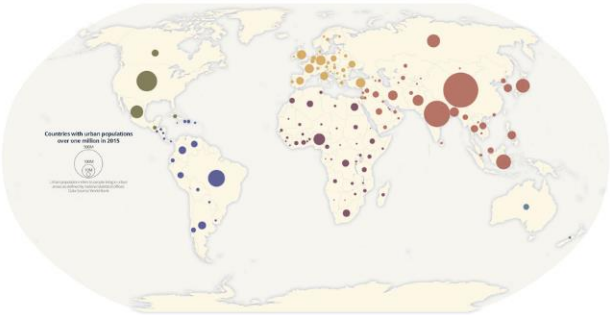


Figure 2.3 A proportional symbol map of global population distribution in year 2015

Data Source: https://carto.com/blog/proportional-symbol-maps/

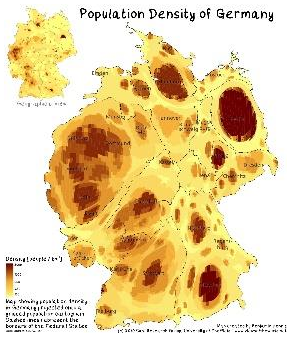


Figure 2.4 A cartogram map of global population distribution in year 2020

Data Source: https://worldmapper.org/maps/grid-population-2020/

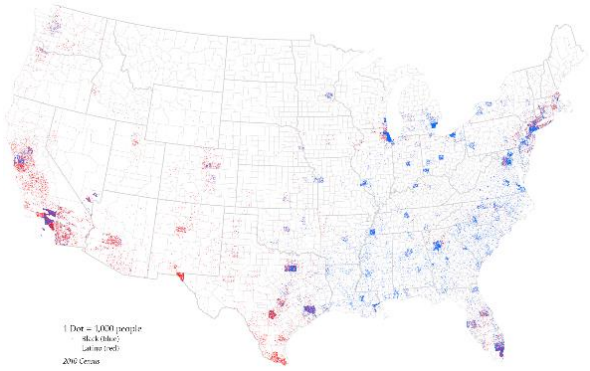


Figure 2.5 A bivariate dot density map showing the relative concentrations of the Black and Hispanic populations in the United States in 2010

Data Source: https://en.wikipedia.org/wiki/Dot_distribution_map

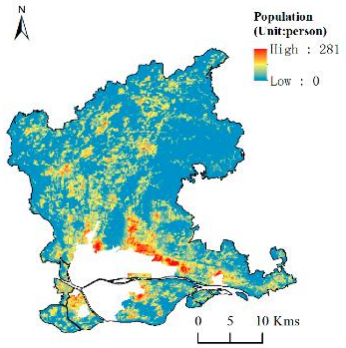


Figure 2.6 A heat map of population distribution in Guangzhou

Data Source: (Zhao & Yang, 2020)

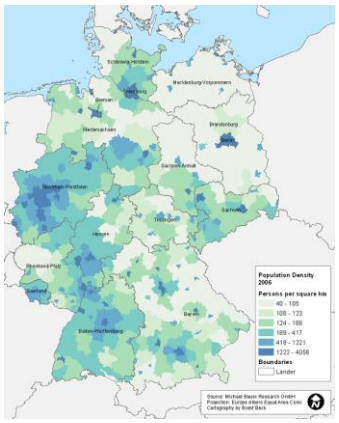


Figure 2.7 An pseudo-isarithmic map of population density in Germany

Data Source: Michael Bauer Research GmbH

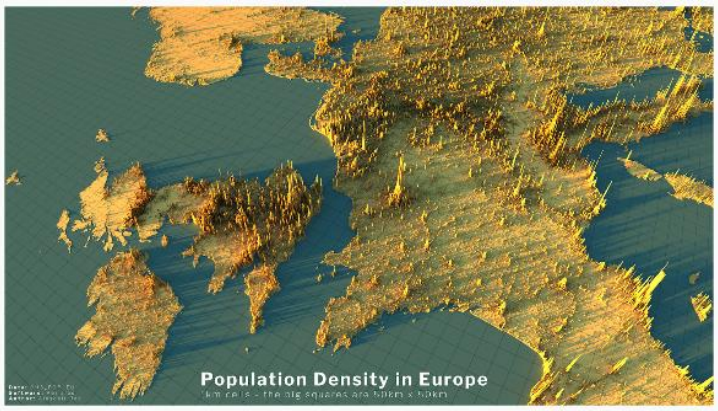


Figure 2.8 A 3D mapping of population density in Europe

Data Source: http://www.statsmapsnpix.com/2020/04/population-density-in-europe.html

2.5 Summary

In general, many studies have been carried out to find other approaches as alternatives for fine-scale population mapping in the past decades. Especially in today's big data era, the conduction of high-resolution population mapping with the participation of remote sensing products and big geo data has become an active field of study. Compared with remote sensing data, big geo data has higher temporal and spatial resolution thus complementing the lack of semantic information which can indicate various human activities with a more precise location, and widely used while in participation of remotely sensed products. However, there are three major problems: 1) collecting social media data is not always possible due to data privacy

and ethical concerns, another approach by using open data needs to be found; 2) optimal scales are rarely discussed, ignoring the fact that different features may have different acting ranges; and 3) at the building scale, population mapping results are frequently the disaggregation result of census data, rather than estimation. As a result, the investigation into integrating multiple freely available remote sensing products and semantic data using machine learning methods while considering optimal scales contributes to the enrichment of fine-scale population mapping products.

3 Methodology and Data Preparation

3.1 Workflow

Based on the previously mentioned motivations, research identification, and state of the art, the planned approach is to conduct data training at the 500m grid cell level, followed by population density estimation at the 50m grid cell level, and then evaluating the result at the 100m grid cell level after aggregating 50m grid cell level estimations to 100m grid cell level. However, it is hampered by a data-hungry problem because there are many no population grids in census data at the 100m grid cell level, making it difficult to obtain sufficient training data at the 500m grid cell level. As a result, the workflow was modified to perform the data modeling section at the 100m grid cell level rather than the 500m grid cell level. The final workflow of this study is as follows:

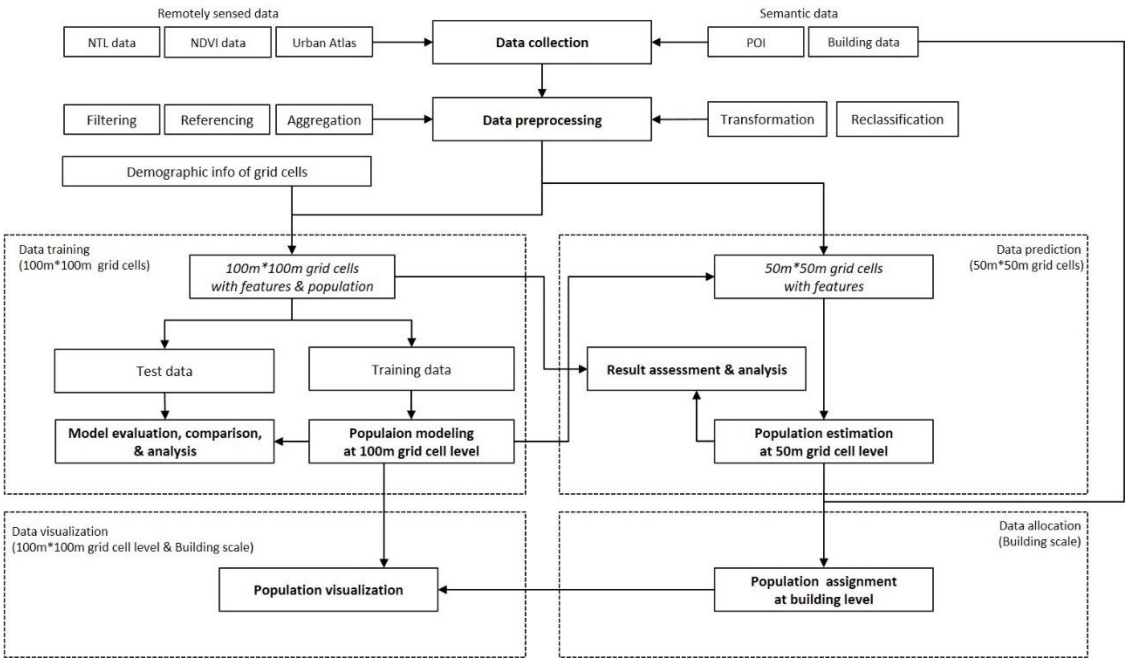


Figure 3.1 Flowchart of this study

To answer the research questions, the framework mainly contains eight sections, data collection, data preprocessing, data modeling at 100m*100m grid cell level, model evaluation and analysis, gridded population estimation at 50m*50m grid cell level, result assessment and analysis, population assignment at building level, and population visualization.

The general idea of the population estimation is to make a finer scale population prediction using machine learning models by integrating ancillary features. In this way,

following data preprocessing, two tables will be created, one of 100m*100m grid cells with features and population data, and one of 50m*50m grid cells with features. After being split into train and test datasets, the first table will be applied to the data modeling section to gain the weight layer of various features, which will then be used in the estimation process. The performance of the models is then assessed by comparing the prediction results of test data with actual data, as well as the difference between several machine learning models. This section will also go over the best feature scale and feature importance. Following that, population estimation is performed by applying the 50m*50m feature table to the model, followed by results assessment and analysis that takes into account spatial heterogeneity. Furthermore, population estimation at the building level is accomplished by assigning estimation results at the 50m grid cell level based on ancillary building information. Finally, all of the outcomes will be visualized and compared.

3.2 Data Collection and Description

This study takes Munich as the study area (Figure 3.2). Before collecting ancillary data, various levels of Munich administrative regions and grid cells, including Munich districts, 100m grid cells, 500m grid cells and 1km grid cells are gathered. All these data are stored as spatial vector data in Shapefile format or GeoPackage. Addtionally, 50m * 50m grid cells which are used for the scale of estimation is created by Grid Index Feature function in ArcGIS. All of these data serves as the foundation for subsequent data preprocessing work.

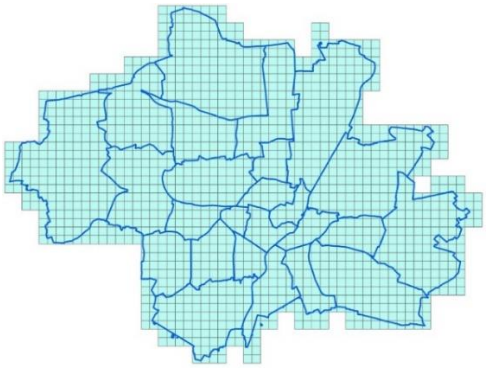


Figure 3.2 Munich Stadtbezirke and 500m Grid Cells

Data Source: Federal Agency for Cartography and Geodesy

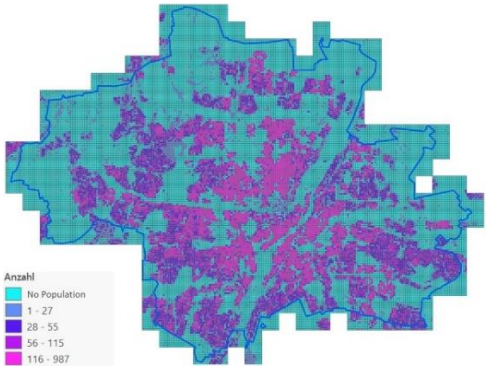


Figure 3.3 Population Distribution of Munich 100m Grid Cells

Data Source: Federal Agency for Cartography and Geodesy

In addition to that, this study incorporates the most commonly used free external data in population mapping, including three remotely sensed products and OpenStreetMap.

3.2.1 Census Data

The most recent public census data provided by the Federal Statistical Office and the statistical offices of the Länder is census 2011[2]. Its fine scales that can be downloaded are 1km and 100m grid-cell level. This is achieved by assigning geo-coordinates available address level statistics to grid cells meeting INSPIRE requirements. In accordance with the INSPIRE Directive, the Lambert Azimuthal Equal Area Projection was used (ETRS89-LAEA Europe - EPSG:3035). Geographic shape-files for this grid are the grid cells mentioned above. In this study, it is used as dependent values in modeling training and actual values in estimation assessment. The table below shows the sample of 100m grid cell census data.

Table 3.1 Sample data of German Census 2011 at 100m grid cell level

Gitter_ID_100m	Gitter_ID_100m_neu	Merkmal	Auspraegung_Code	Auspraegung_Text	Anzahl	Anzahl_q
100mN26891E43370	CRS3035RES100mN2689100E4337000	INSGESAMT	0	Einheiten insgesamt	8	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	ALTER_KURZ	1	Unter 18	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	ALTER_KURZ	3	30 - 49	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	ALTER_KURZ	5	65 und älter	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	FAMSTND_AUSF	1	Ledig	4	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	FAMSTND_AUSF	2	Verheiratet	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	GEBURTLAND_GRP	1	Deutschland	6	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	GEBURTLAND_GRP	21	EU27-Land	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	GESCHLECHT	1	Männlich	4	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	GESCHLECHT	2	Weiblich	4	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	RELIGION_KURZ	1	Römisch-katholische Kirche (öffentlich-rechtlich)	4	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	RELIGION_KURZ	3	Sonstige, keine, ohne Angabe	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_GRP	1	Deutschland	5	0

[2] https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html

100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_GRP	21	EU27-Land	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_HLND	1	Deutschland	5	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_HLND	8	Österreich	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_KURZ	1	Deutschland	5	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATSANGE_KURZ	2	Ausland	3	0
100mN26891E43370	CRS3035RES100mN2689100E4337000	STAATZHL	1	Eine Staatsangehörigkeit	7	0

Data Source: German Census 2011, https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html

Table 3.1 shows the census information of grid cell “100mN26891E43370”. Gitter_ID_100m is the id of 100m grid cell of the Lambert Azimuthal Equal Area Projection which can be used to match spatial vector data of these grid cells. Anzahl is the number of people in this grid according to Census 2011. Merkmal is the attribute of the population number, and Auspraegung_Code, Auspraegung_Text describe the corresponding attribute. As we can see from the table, there are a lot of attributes here, including the total value, age, family status, etc. This study only takes the total value of the grid, therefore, here we use the first line. Fig 3.3 shows the population distribution in 100m grid cells after census data processing.

3.2.2 Remotely Sensed Products

The remotely sensed products used in this study include Nighttime Light (NTL) Images, Normalized Difference Vegetation Index (NDVI), and land use data.

NDVI data is derived from the Terra satellite's MODerate Resolution Imaging Spectroradiometer (MODIS) at a spatial resolution of 250m. NTL data is from NPP-VIIRS DNB with a spatial resolution of 15 arc-seconds (about 500m) provided by the National Geophysical Data Center, NOAA, USA. The annual mean composition was applied to create a mean NDVI image and NTL image of year 2014 which could reduce the sensitivity of seasonal variations in the earliest year available. Both datasets were downloaded and preprocessed using Google Earth Engine (GEE). GEE is a cloud-based geospatial processing platform powered by Google's cloud architecture that offers free global scale geoscience data at the PB level (Gorelick, et al., 2017). The computation of GEE is based on automatic parallel processing of Google infrastructure, which dramatically improves data processing efficiency (Xiong, et al., 2017).

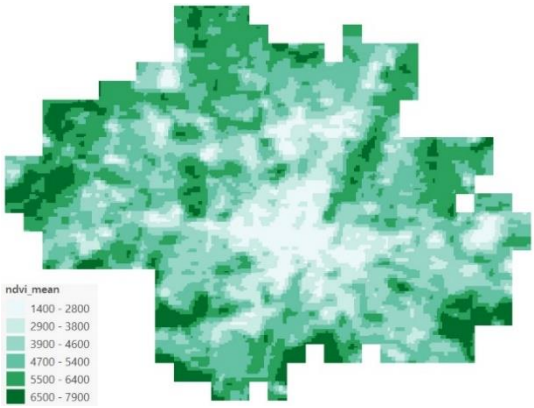


Figure 3.4 Munich NDVI Data of 100m grid cells

Data Source: Google Engine

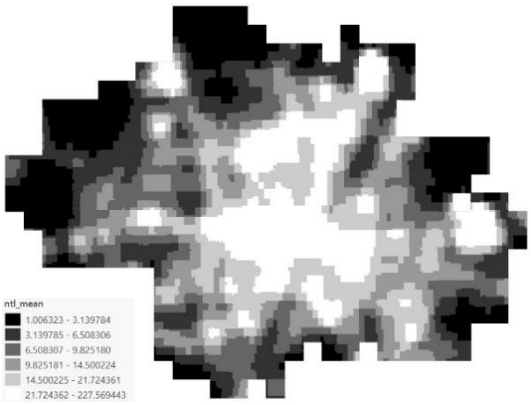


Figure 3.5 Munich NTL Data of 100m grid cells

Data Source: Google Engine

Two different sourced land use data were collected. The dataset CORINE Land Cover 5 ha (CLC5 2018) [3] represents a description of the landscape in vector format according to the nomenclature of CORINE Land Cover (CLC), which on the one hand reflects the land cover and on the other hand also includes aspects of land use. Urban Atlas [4] is part of Copernicus Land Monitoring Services, providing pan-European comparable land use and land cover (LULC) data derived from Very High Resolution (VHR) satellite imagery for Functional Urban Areas (FEA) with more than 50000 inhabitants. Figure 3.7 displays Urban Atlas LULC information related to the FUA DE003L1 – München for the reference year 2012. Its thematic information is derived by means of semiautomatic Sentinel–2 time series classification combined with visual interpretation using VHR satellite imagery made available through ESA CSCDA mechanism.

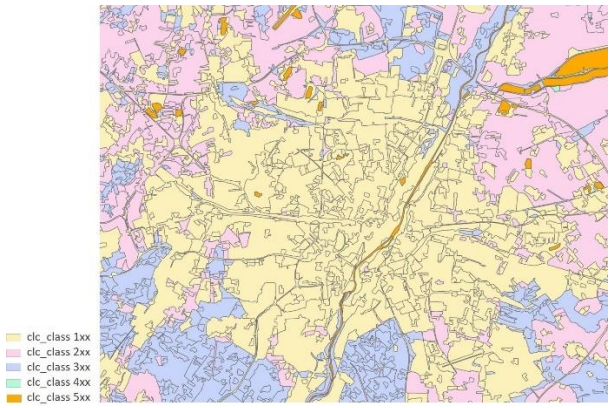


Figure 3.6 CLC5 2018 of Munich

Data Source: Federal Agency for Cartography and Geodesy

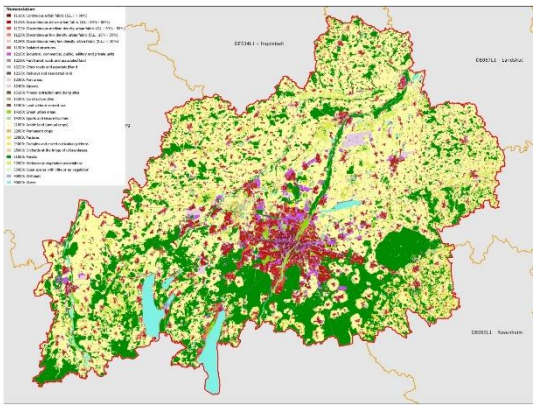


Figure 3.7 Urban Atlas of Munich

Data Source: Copernicus EU Land Monitoring Service

[3] https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/digitale-landschaftsmodelle/corine-land-cover-5-ha-stand-2018-clc5-2018.html
[4] https://land.copernicus.eu/local/urban-atlas

3.2.3 Semantic Products

The semantic data used in this study is POI and OSM Building provided by OpenStreetMap with spatial locations. OSM POIs are geo-located features that are described by point, line, or polygon with other attributes given by tags. An example of tag describing a POI could be postbox, shops, schools, and so on. When adding POIs to OSM, contributors are advised to utilize the terminology provided in the OSM Feature Wiki. Other contributors can also modify POIs generated by a contributor. OSM POIs (Figure 3.8) are achieved from Geofabrik[5] which is an open source website providing free data from OpenStreetMap. Building attributes (Figure 3.9) are collected from OSM Building[6] which is a Free and open source web viewer for 3D buildings supported by OpenStreetMap tagging schema. It is based on XYZ tiles, with a GeoJSON dataformat and geometry projection EPSG:4326 (WGS84). Attributes of building in the dataset include building name, building type, building height, building levels, etc.

Table 3.2 includes all data sources utilized in this analysis, along with the reference year for each. Because the proposed process works with data from various contexts representing the same objects, the mentioned data may be regarded a sample.

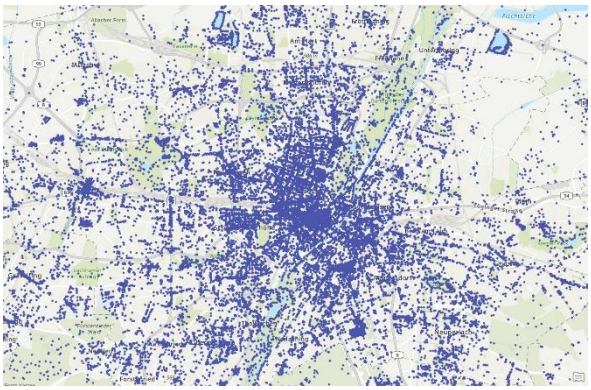


Figure 3.8 OSM POIs

Data Source: OpenStreetMap

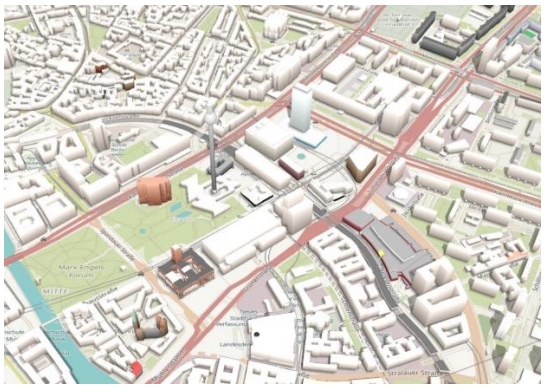


Figure 3.9 OSM Building Sample

Data Source: OSM Building

Table 3.2 Overview of Data Used in This Study

Type	Official Name	Data Provider	Reference Year
Administrative Units	Verwaltungsgebiete 1:250 000 (kompakt), Stand 31.12. (VG250 31.12.)	Federal Agency for Cartography and Geodesy	2011

[5] https://download.geofabrik.de/europe/germany/bayern/oberbayern.html
[6] https://osmbuildings.org/

Geographical Grids for Germany	DE_Grid_ETRS89-LAEA_100m		
	DE_Grid_ETRS89-LAEA_500m		2018
	DE_Grid_ETRS89-LAEA_1km		
NDVI data	MOD13Q1.061 Terra Vegetation Indices 16-Day Global 250m	NASA LP DAAC at the USGS EROS Center	2014
NTL data	VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1	Earth Observation Group, Payne Institute for Public Policy, Colorado School of Mines	2014
Land use	CORINE Land Cover 5 ha, as of 2018 (CLC5-2018)	Federal Agency for Cartography and Geodesy	2018
	Urban Atlas 2012 - München	Copernicus EU Land Monitoring Service	2012
Points of Interest	OpenStreetMap data	OpenStreetMap contributors	
Building attributes	OSM Building		2022

3.3 Data Preprocessing

Raw data typically contains a lot of information that is not needed in the study as well as some non-structured data. Therefore, before data modeling and visualization, data pre-selection, data georeferencing, data aggregation, and data reclassification are critical procedures.

3.3.1 Data Filtering

This step filters out unnecessary datasets, areas and features.

From the section above, we know that two land use datasets are collected for this study. As Urban Atlas has more detailed classification within a same area compared with CORINE Land Cover when display them in ArcGIS (Figure 3.10, 3.11), Urban Atlas is selected as the dataset for land use purpose.

Figure 3.10 Sample of CORINE dataset

Data Source: Federal Agency for Cartography and Geodesy

Figure 3.11 Sample of Urban Atlas dataset

Data Source: Copernicus EU Land Monitoring Service

Next, it is the area filtering as some of the data scope is not just within Munich but Germany. This is finished in ArcGIS by taking the Munich 100m grid cell Shapefile as the standard scope for the datasets of this study.

The last step is feature selection. In OSM Building dataset, more than 15 attributes are listed, like name, type, height, area, color, material, roof angle, etc. As height, and type are relevant to population mapping (Schug, Frantz, van der Linden, & Hostert, 2021), these three features are taken at first. However, more than 50% of the building types are missing in the dataset, this attribute was discarded in the end. As for the OSM POI dataset, more than 100 categories are found. Considering some of them are landscape infrastructures, 18 of them are dropped and the rest of them are reclassified in the next steps.

3.3.2 Data Referencing

As Grid Cell bases German Census 2011 takes the Lambert Azimuthal Equal Area Projection (ETRS89-LAEA Europe - EPSG:3035), all vector data are transformed to this projection in this study.

3.3.3 Data Aggregation

Since the base map is grid cells of different levels, each grid cell is a basic entity with census data as the dependent variable and ancillary data as independent variables. In this way, all the features need to be integrated to grid cells of different levels, therefore, data aggregation step is applied here which is achieved by using Pairwise Intersect function in ArcGIS. Finally, table format with grid cells as entities and features from other datasets as fields are reached, Table 3.3 shows the sample of building information after data aggregation process.

Table 3.3 Sample result of data aggregation

FID_buildi	Id	Height	Type	Shape_Area	id_100mmGrid
203638	279257061	5	fire_station	727.294684	100mN27860E44432
203639	279257060	4		28.8406134	100mN27860E44432
203640	288095895	4	storage	97.5579286	100mN27860E44432
203649	r57759	15		976.586378	100mN27860E44429
203650	r57760	15	office	903.004776	100mN27860E44429

3.3.4 Data Reclassification

As Urban Atlas 2012 and POI data has too many categories, category reclassification is applied before area proportion of land use category and multiscale POI density generation.

There are 21 categories of land use in Urban Atlas 2012, which contains several different kinds of urban fabric, transportation area, farmland, and forests. In order to simplify the land use information, those 21 categories are reclassified to 11 classes in the end.

According to the general attributes of POI information, the rest 108 categories are reclassified into 17 classes, including recycling services, education, leisure, sports, culture facilities, health services, green space, government, retail, food services, accommodation, life services, resorts, railway stops, public transport stops, airport and helipad.

3.3.5 Data Transformation

Data that has been properly transformed and verified enhances data quality and protects applications from potential landmines, such as incompatible formats, unexpected null values. Therefore, in order to better utilize the features, data transformation has been applied to OSM Building dataset, Urban Atlas dataset, and OSM POI dataset.

OSM Building dataset. Another feature Volume is created by multiply building footprint area and building height.

Urban Atlas dataset. As it only records the exact area of each category within a grid cell, the proportion of corresponding attribute is calculated by simply dividing its area with the total area of a grid cell.

OSM POI dataset. Since each POI has an acting range, POI density instead of POI amount is widely used to reflect urban functions and characteristics of a region. As each kind of POI has its own influencing area, the density of each POI type was calculated within predefined distances of 400 m, 800 m, 1200 m, 1600 m, 2000m, and 3000m which takes the neighbourhood scale and district scale, 5-15 minutes walking distance, 5-15 minutes riding distance into account. Finally, POI density of 17 types with 6 different scales are achieved by using ArcPy.

3.4 Data Modeling

Since machine learning has been shown to be effective for learning and discovering patterns in population mapping, several machine learning models are used in this study to model the complex and hidden relationships between a large number of factor variables and population data, and then to participate in a comparison. Some

most often used regression machine learning models are selected including linear and non-linear algorithms, such as linear regression, random forest, decision tree regression, support vector machine, and K Nearest Neighbors.

The independent variable in this study is 100m and 50m grid cell level data with features, and the dependent variable is its corresponding population density. The population mapping model is built in Python using data from 100m grid cells, with training samples accounting for 80% of the total dataset. The trained model is then applied to gridded population estimation at 100m and 50m grid cell levels, respectively.

3.5 Model Evaluation

To evaluate the prediction ability and transferability of population mapping models, the model evaluation session is divided into two parts: evaluation at the 100m grid cell level using test data representing 20% of the total dataset, and another evaluation process at the 100m grid cell level comparing the population of 100m grid cells and the aggregation results of 50m grid cell estimations.

Apart from R^2, the correctness of each model and dataset is measured using the mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMSE). The formulas are as follows:

$$MAE = \frac{1}{N}\sum_{i}^{N}|P_i - R_i|$$

$$MRE = \frac{1}{N}\sum_{i}^{N}\frac{|P_i - R_i|}{R_i}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(P_i - R_i)^2}$$

where P_i is the estimated population of grid i, R_i represents the census value of grid i, and N is the size of the grid cells.

3.6 Visualization

Some distinct visualization methods are used to illustrate trends in different aggregated levels to facilitate an intuitive and in-depth investigation of the properties of the population distribution. Choropleth maps, heat maps, and 3D extrusion, in particular, are widely utilized to aid in understanding population spatial trends. Different population density areas can be quickly and clearly identified using

choropleth maps and heat maps, however 3D extrusion perform better in presenting a direct idea of how significant the population disparity is.

4 Case Study and Results

In this section, the modeling and visualization methods are applied in a case study of Munich to test the proposed population mapping approach as well as the estimated population results visualization. More specifically, the accuracy assessment and the comparison of several machine learning models are analysed. Additionally, different visualization methods are used in the final output of the results.

4.1 Study Area

Munich is chosen as the study area of this study. It is located in the south-east of Germany and consists of 25 districts. It covers an area of 310.71 km^2 with around 1.56 million inhabitants in year 2021[7] ranking as the third-largest city in Germany after Berlin and Hamburg. Munich is the seat of the Bavarian administrative area of Upper Bavaria and the most densely inhabited and populous municipality in Germany, straddling the banks of the River Isar (a tributary of the Danube) north of the Bavarian Alps (4,800 people per km^2). As a result, urban spatial typologies have come to dominate many districts. However, the outlying districts also include low-density suburbs and high-rise housing complexes built in the 1960s and 1970s.

4.2 Data Modeling

The whole dataset contains 119 columns and 31783 rows in total after data preprocessing.

Before modeling the dataset at 100m Grid Level, feature distribution and multilinearity analysis are performed as these are necessary steps to better understand data attributes and take measures to process data before selecting suitable models.

In order to better visualize the result of the analysis, optimal scales of each POI category is chosen before visualizing them. In this way, the visuals contain 35 columns instead of 119 columns.

4.2.1 Feature Distribution Analysis

The mentioned 44 columns are comprised of two parts: the independent variables from NTL data (1 column), NDVI data (1 column), Urban Atlas 2012 (12 column), POI density in its optimal scale (17 column) and building information from OSM Building (3 column), and the dependent variable population data (1 column) from German Census 2011.

[7] https://en.wikipedia.org/wiki/Munich

NDVI and NTL data are stored in the ndvi mean and ntl mean columns, respectively. Volume, build area (the area of the building footprint), and MeanHeight are the three columns of building information. The Urban Atlas 2012 data is divided into 12 columns, with field titles beginning with "p." The remaining columns represent POI density values with their optimal acting range. As shown in the figure 4.1 below, the distribution of most features is highly skewed, particularly the urban atlas and poi data. Furthermore, we can see that there are many extreme values throughout the dataset, as well as many values close to 0.

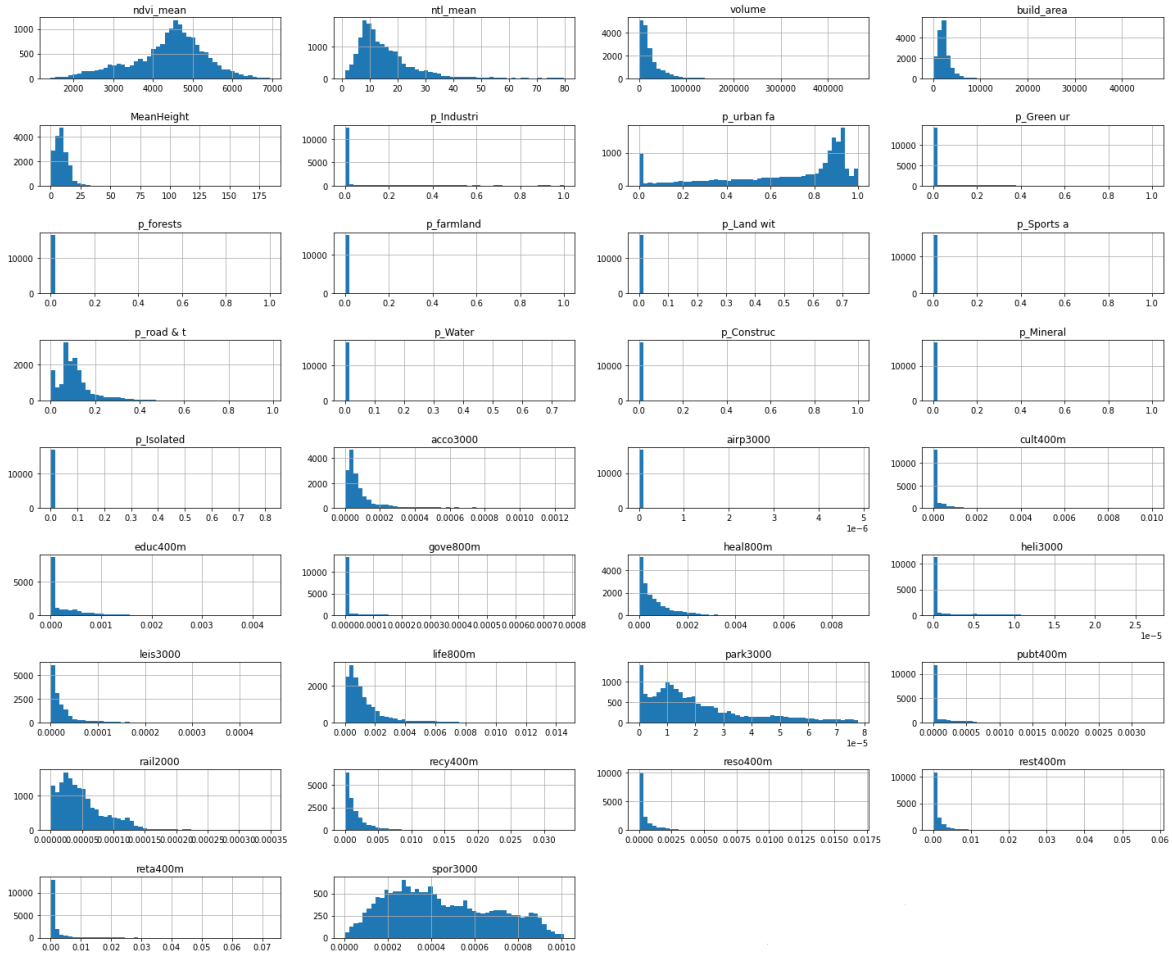


Figure 4.1 Distribution of all features

Because some machine learning algorithms, such as Linear Regression and Gaussian Naive Bayes, assume that numerical variables have a Gaussian probability distribution, and the current dataset are highly skewed and contains many values close to 0, the input data is transformed to have a Gaussian or more-Gaussian distribution to achieve better modeling performance. As a result, before utilizing the dataset in the regressor, a power transform is implemented. Power transforms, such as the Box-Cox transform and the Yeo-Johnson transform, are available in the Scikit-Learn Python machine

learning package and provide an automated manner of conducting these transformations on data.

As for the dependent variable, population density is calculated and acts as the y label in the modeling session instead of using the absolute amount of populations. Figure 4.2 shows the distribution of the population density. It is clearly to see that there are some extreme values which contains a grid cell with a maximum value up to 98700 /km^2, and three other values greater than 63000/km^2.

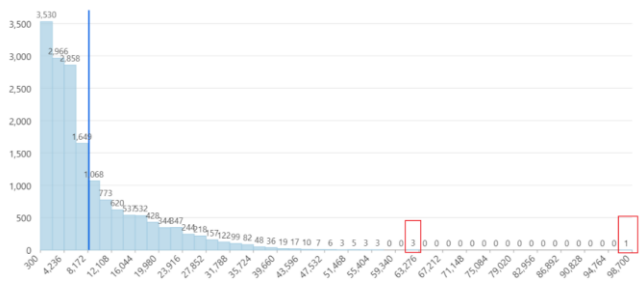


Figure 4.2 Distribution of population density

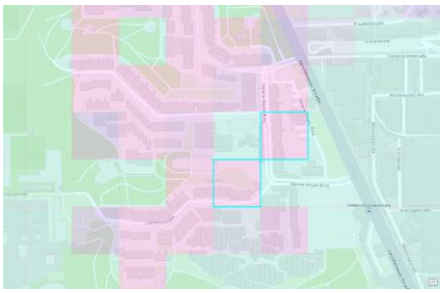


Figure 4.3 Information comparison from google map

The overlay analysis was performed to determine whether the extreme population density values are outliers. To be more specific, two of the grid cells with the aforementioned values were examined by comparing them to a Google map (Figure 4.3). The grid cell with the light blue color below has 987 people and the one above has 628 people, but they have far more residents than other grid cells even though they are in the same neighborhood. According to the documentation for the grid cell-based German Census 2011, the population allocation to grid cells is based on address, however, an address may spread out into several 100m*100m grid cells, and the population is assigned to one cell rather than several. As a result, we can call these extreme values outliers and remove them from our samples.

4.2.2 Feature Multilinearity Analysis

Because some regression algorithms, such as linear regression, cannot deal with multilinearity, feature multilinearity is also checked when selecting suitable models. Specifically, when there is a problem with multicollinearity, least-squares is unbiased and variances are huge, resulting in projected values that are far from the actual values. The variance inflation factor (VIF) and the correlation between all independent variables are computed during the check process.

1) Correlation matrix

Multilinearity is the condition in which many variables are highly connected and convey comparable information regarding variation within a particular dataset. In Python, this may be performed by using numpy's corrcoef function to obtain a confusion matrix of all variables. It may then be presented within a heat map. Figure 4.4 and 4.5 show the result of all the 34 columns after choosing optimal scales of POI.

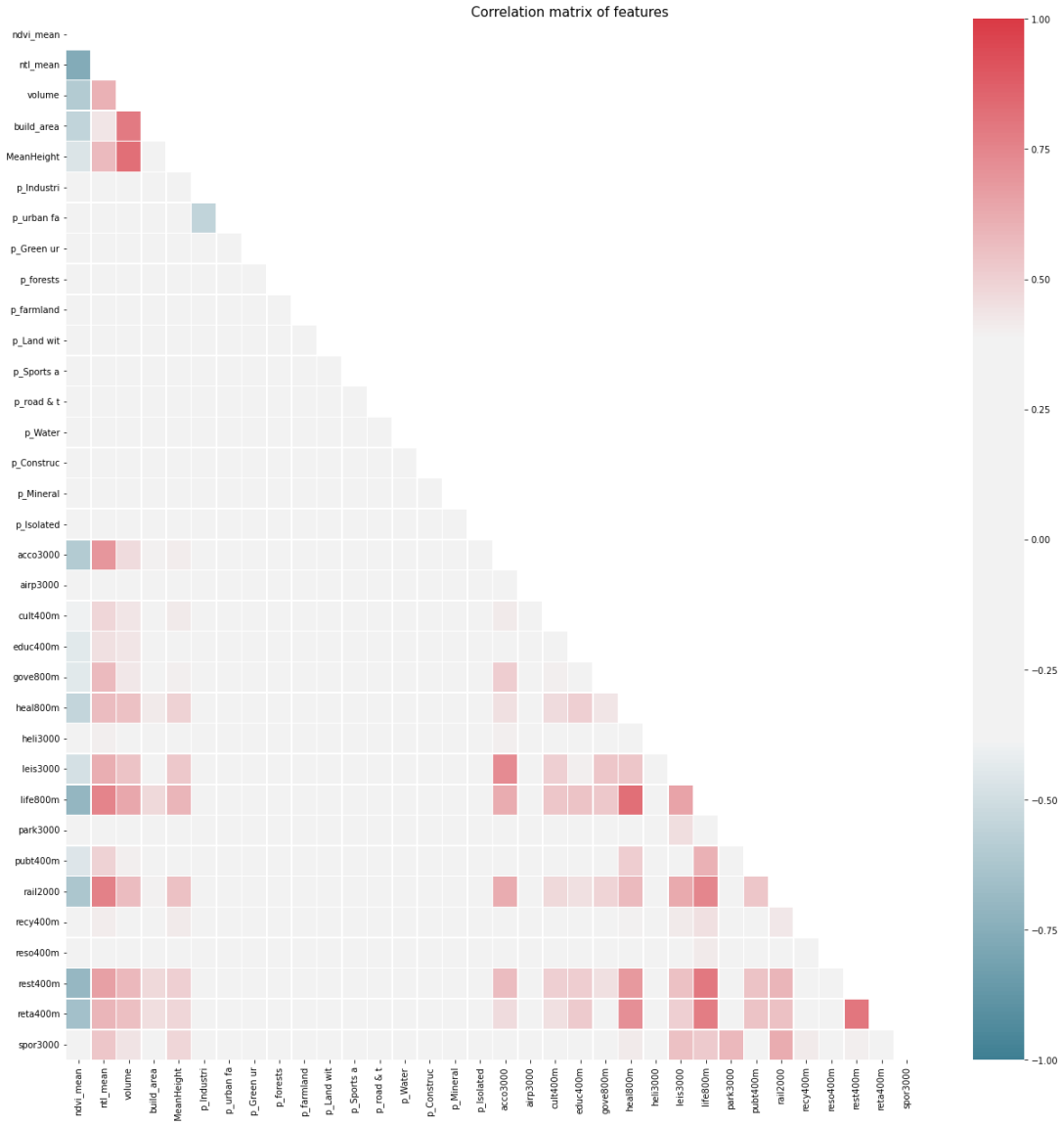


Figure 4.4 Correlation matrix of features

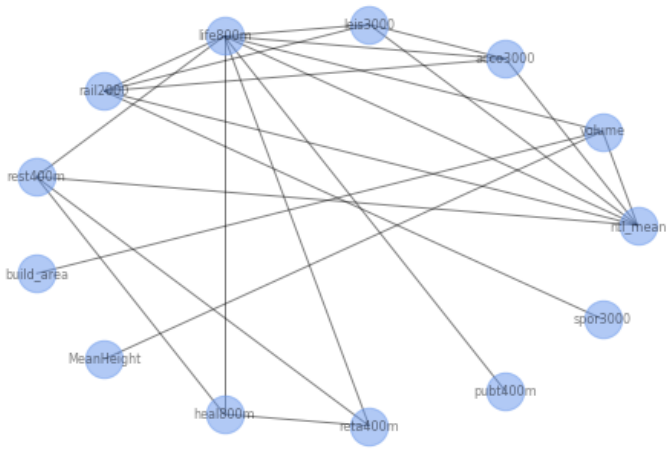


Figure 4.5 Connection of features with strong correlation

It is clear that NDVI has a strong negative collinearity with a variety of other features, including NTL, building information, restaurants, and retails. This corresponds to our perception, as these are highly urbanized areas with little vegetation. Another feature which has negative correlation with others is industrial, commercial area. Because census data only considers residents, it is understandable that there are no people living in those industrial and commercial areas. Other significant positive collinearities include NTL, building information, accommodation facilities, culture facilities, education, government, health care, leisure facilities, life services, railway stops, and restaurants, all of which are good representatives of urban areas.

2) VIF

VIF is a popular method for evaluating regression assumptions, especially multicolinearity, and unlike many statistical ideas, its formula is simple: It is equal to the ratio of the total model variance to the variance of a single independent variable model. Its formula is below:

$$VIF_{\ i} = \frac{1}{1 - R_i^2}$$

where R_i^2 is the regression equation's coefficient of determination in linear regression, with X_i on the left and all other predictor variables (all other X variables) on the right.

VIF ranges from 1 with no upper limit. The numerical value for VIF indicates the percentage by which the variance for each coefficient is inflated. For example, if the VIF of a feature is 1.8, it means that the variance of a certain coefficient is 80% more than what would be expected if there was no association with other variables. If VIF is equal to one, the variables are not correlated; if it is between one and five, the

correlation is moderate; and if it is greater than five, the variables are significantly connected.

Table 4.1 displays the VIF results for all 35 features. The top 11 features all have a value greater than 5, indicating that they are strongly related to others. Area proportions of urban fabric, roads, POI density of leisure facilities, accommodation, life service facilities, building volume, railway stops, restaurants, and retail, as well as NTL, which are indicators of densely populated areas, have extreme values. Furthermore, many features have a value between 1 and 5, indicating that they have a moderate correlation.

As a result, we can see that both the correlation matrix and the VIF show that there is strong multicollinearity between features that are closely connected within areas with high populations, and the number of them is large. Therefore, we must select machine learning algorithms that can produce good results even when there are collinerities.

Table 4.1 VIF of features

Id	Variable	Description	VIF
1	*p_urban fa*	Area proportion of Urban fabric	*91.18*
2	*leis3000*	Leisure facility density in 3km range	*24.25*
3	*acco3000*	Accommodation density in 3km range	*23.67*
4	*life800m*	Life service facility density in 800m range	*10.49*
5	*p_Industri*	Area proportion of Industrial, commercial, public, military and private units	*10.43*
6	*volume*	Building volume in a grid cell	*10.13*
7	*rail2000*	Railway stops density in 2km range	*7.77*
8	*rest400m*	Restaurant density in 400m range	*7.46*
9	*build_area*	Building footprint area in a grid cell	*6.14*
10	*ntl_mean*	Mean value of NTL in a grid cell	*5.04*
11	*reta400m*	Retail density in 400m range	*5.04*
12	p_road & t	Area proportion of Roads, port, airport, and associated land	4.87
13	heal800m	Health care facility density in 800m range	4.64
14	p_Green ur	Area proportion of Green urban areas	4.51
15	MeanHeight	Building mean height in a grid cell	4.04
16	ndvi_mean	Mean value of NDVI in a grid cell	3.85
17	p_farmland	Area proportion of Farmland	3.74
18	p_Sports a	Area proportion of Arable land	2.74
19	spor3000	Sport space density in 3km range	2.41
20	park3000	Green urban space density in 3km range	2.19
21	p_forests	Area proportion of Forests	1.90
22	pubt400m	Public transport stops density in 400m range	1.89
23	cult400m	Culture facility density in 400m range	1.76

24	gove800m	Government density in 800m range	1.68
25	educ400m	Education facility density in 400m range	1.49
26	heli3000	Helipad density in 3km range	1.42
27	reso400m	Resort density in 400m range	1.40
28	recy400m	Recycling facility density in 400m range	1.30
29	p_Construc	Area proportion of Construction sites	1.25
30	p_Land wit	Area proportion of Land without current use	1.22
31	airp3000	Airport density in 3km range	1.07
32	p_Isolated	Area proportion of Isolated structures	1.06
33	p_Water	Area proportion of Water	1.06
34	p_Mineral	Area proportion of Mineral extraction and dump sites	1.03

4.2.3 Modeling Results and Evaluation

Finally, in addition to Random Forest (RF) Regression, four other models are chosen: Ridge Regression, Decision Tree Regression, Support Vector Regression (SVR), and K Nearest Neighbors (KNN) Regression.

As mentioned in the literature review, Random Forest is the most often used model in statistical population mapping as it has a strong generalization ability, allowing it to handle high-dimensional features and effectively improve population spatialization accuracy (Breiman, 2001). Therefore, it is chosen in this study.

Four different models, both linear and non-linear, are chosen to test the performance of the RF model and ensure that it is the best among shallow machine learning approaches. Ridge Regression[8] is a type of linear regression that can analyze datasets with multicollinearity since it incorporates L2 regularization. Ridge Regression begins by normalizing the variables (both dependent and independent) by subtracting and dividing their means by their standard deviations, such that all calculations are based on strandardized variables (McDonald, 2009). Decision Tree Regression is the second model. To create regression models, Decision Tree employs a tree structure. It splits a dataset into smaller and smaller portions as it develops an associated decision tree. The resulting result is a tree with leaf and decision nodes[9]. Support Vector Regression is the third (SVR). SVR uses the same premise as Support Vector Machine techniques to predict discrete values. The primary principle of SVR is to locate the best fit line (hyperplane) with the most points[10]. Unlike other regression models, which seek to

[8] https://en.wikipedia.org/wiki/Ridge_regression
[9] https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/
[10]https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.

minimize the difference between the true and projected values, the SVR seeks to fit the best line within a certain range. The threshold value is the distance between the hyperplane and the boundary line, making it difficult to scale to datasets with more than a few tens of thousands of samples. The KNN regressor, which predicts the values of new data points based on 'feature similarity,' is the last one. This implies that the new point is assigned a value based on how closely it resembles the points in the training set[11].

All of the models shown above were created using the Scikit-Learn Python library. Table 4.2 shows the results. We can easily see that RF has the best overall performance, with the highest R^2 and the lowest mean absolute error and median absolute error. Ridge Regression and K Nearest Neighbors Regression outperform Decision Tree Regression and Support Vector Regression, while Support Vector Regression performs the worst in this case. Its R^2 indicates that it learned nothing during the modeling process.

Table 4.2 Evaluation of the applied five models

Performance	Random Forest	Ridge Regression	Decision Tree Regression	Support Vector Regression	K Nearest Neighbors Regression
R^2	0.77	0.58	0.49	-0.24	0.59
Mean absolute error (/km^2)	1520	2815	2145	3774	2029
Median absolute error (/km^2)	242	1647	0	452	200
Max error (/km^2)	42112	40086	43900	48947	51050

To be more specific in the evaluation of the RF model, in general, the RF regression model can explain the observed data well, revealing that the model explains 77% of the variability observed in the target variable. Figure 4.6 shows the scatter plot of the distribution of true values and predicted values, and its trend line. While the distribution of real population density has a mean value of 3790/km^2, the regression model's mean absolute error and median absolute error

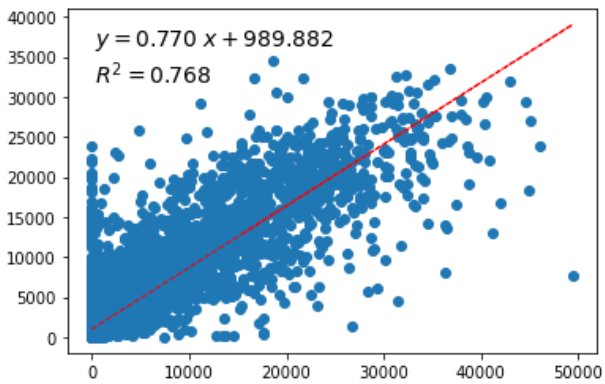


Figure 4.6 Scatter plot of true and predicted values

[11] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

are acceptable because they are around ½ of the true values. As for the max absolute error, since the value evaluated is population density, it means that the maximum gap of real and predicted value is 421 people in one 100m*100m grid cell.

As a result, RF is selected as the model to estimate population density at the 50m*50m grid cell level.

4.3 Population Estimation

The 50m*50m grid cells were created by dividing 100m*100m grid cells into 50m*50m grid cells, and they contain 127132 cells. All of the previous independent variables were aggregated to these cells and used as input data, which was then applied to the previously trained RF regression model. The estimations are then aggregated to 100m*100m grid cells and compared with the 31783 true values, providing a larger number of test cases to evaluate model performance and investigate the causes of prediction errors.

Figure 4.7 and 4.8 depicts the absolute error distribution of the true and predicted values, respectively. The mean absolute error is approximately 29, and the median absolute error is approximately 3. While the mean and median true values of the population are around 84 and 55, respectively, we can see that the distribution of absolute errors is acceptable because 76.4 percent of them are smaller than the mean population, and around 73% are smaller than 30, indicating that the estimation result is good.

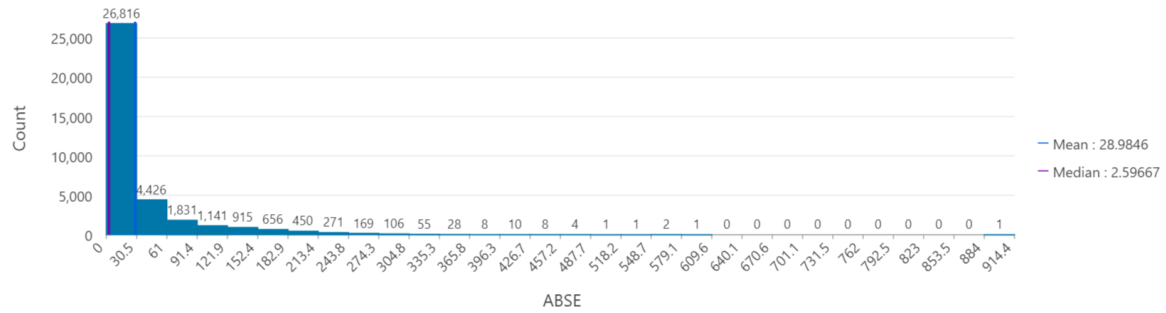


Figure 4.7 Histogram of Absolute Error between true and predicted population

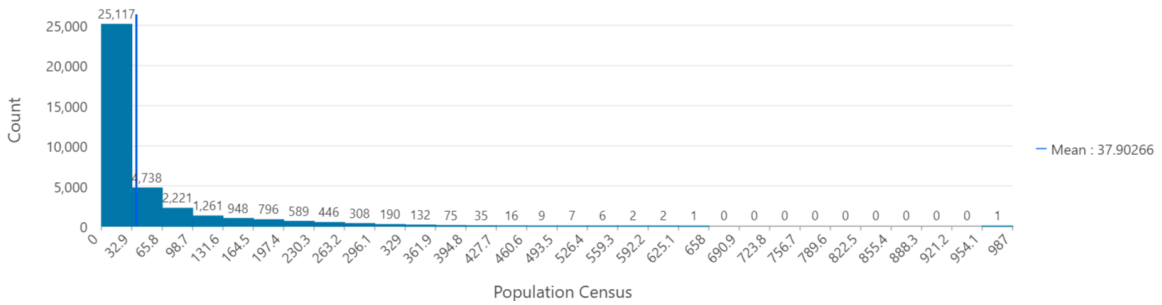


Figure 4.8 Histogram of Population true values

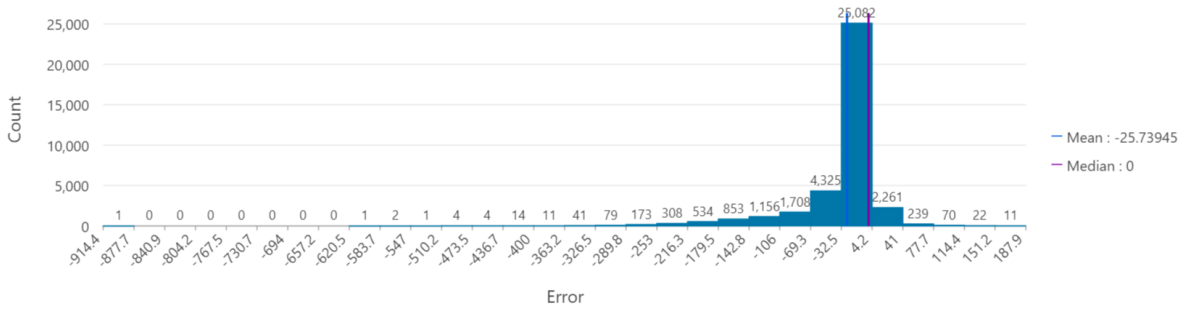


Figure 4.9 P Histogram of Error between true and predicted population

Figure 4.9 depicts the distribution of errors between predicted values and population true data. 68% grid cells contain an error between -32.5 and 4.2. Furthermore, the majority of grid cells are clearly underestimated, resulting in negative values. We can also see that the distribution of negative errors is wider than the distribution of positive errors. While 25 percent of grid cells have positive errors less than 40, 21 percent of their negative errors are less than -40. Furthermore, the maximum negative error is 904.8, which is much larger than the positive error of 165, indicating that the trained model makes more errors in underestimation as well.

When we look at the spatial distribution of the error values in figures below, we can get more details on how well the trained model predicted.

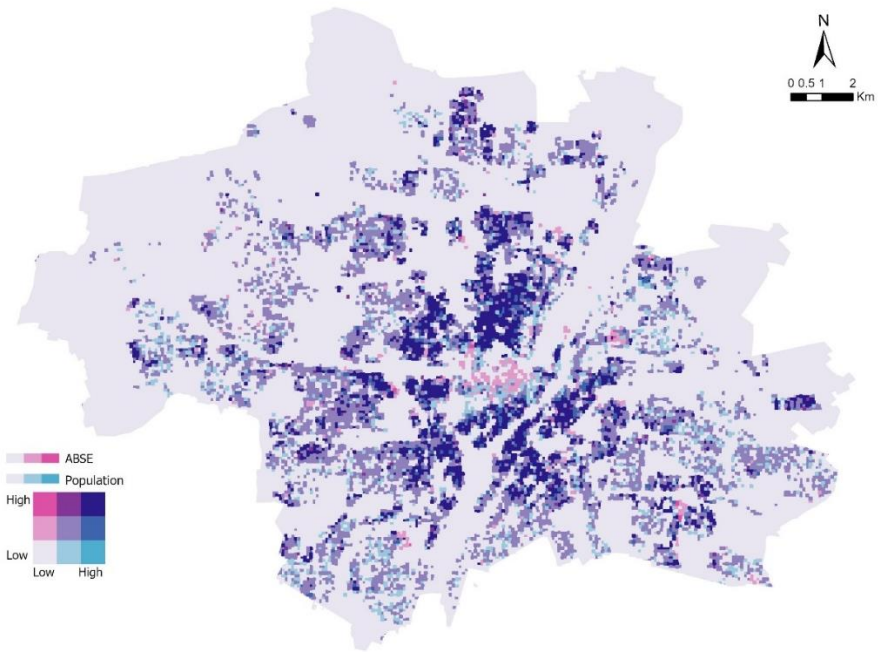


Figure 4.10 Estimation ABSE and population distribution

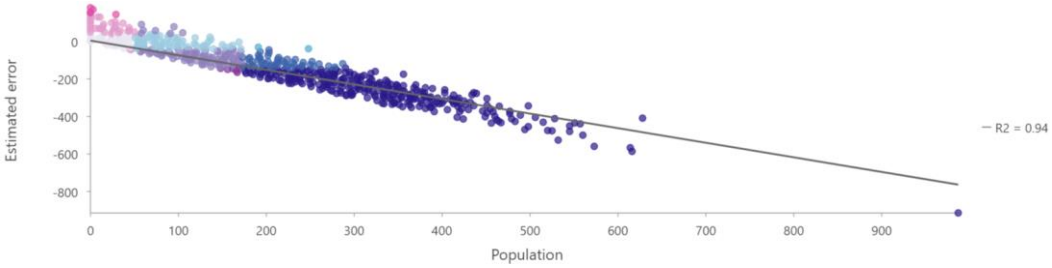


Figure 4.11 Scatter plot of population true values and estimated errors

As shown in figure 4.10, it appears that the higher the ABSE, the more people there are in a grid cell. Despite the fact that the legend has nine colors, the majority of the grid cells in the figure are light purple, medium purple, and dark purple, which correspond to low-low, medium-medium, and high-high in the legend. There are also some medium-low and low-medium grid cells that are light blue and light blue in color, but far fewer than dark and light colored grid cells. Furthermore, we can see that the central area of Munich has the most of the darkest grid cells while also having the highest population density, and the low-low grid cells are widely distributed in the outskirts. It gives us a hint that high density areas have higher estimated errors.

Looking at the scatter plot of estimated errors and the population (figure 4.11) confirms our previous conclusion. Positive error is represented by points above 0; negative error is represented by points below 0. We can clearly see that when the population approaches 100, the negative errors rapidly increase, implying that there are more underestimations in large population areas.

Figure 4.12 Error distribution in no population area

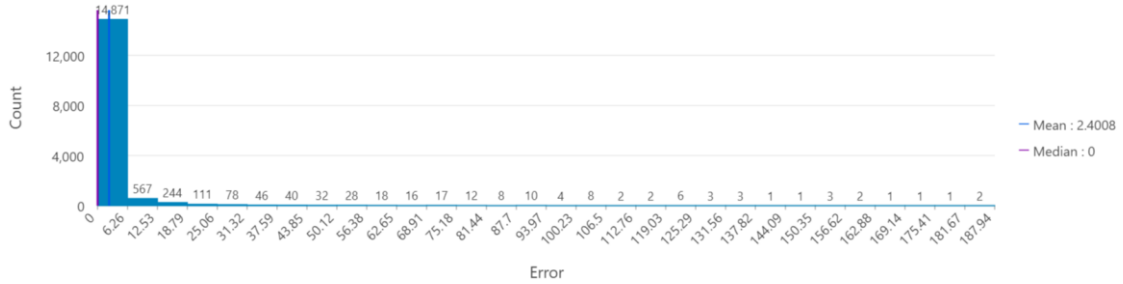


Figure 4.13 Histogram of error distribution in no population area

Another finding from the trained model is that many grid cells have an estimated population greater than one after being aggregated to 100m*100m grid cells, despite the fact that they have no population area in reality.

The distribution of estimated population errors and land use in these areas is depicted in Figure 4.12. While 76% of grid cells in no population areas have an estimated population error of less than one, the majority of cells with larger estimation errors after the overlay analysis are in industrial, commercial, and military areas, as well as farmland areas. The land use data is from Urban Atlas 2012. According to the definition of the German Census 2011, it records residents who have a registered address in the city hall, and the registered address typically refers to a residential area that corresponds to the urban fabric in Urban Atlas 2012. Therefore, the industrial, commercial, farmland, forest, and water areas should not have any residents. However, many estimated positive errors still fall into those areas, indicating that the trained model cannot perform as well in these areas as reality. Because there are constructions (industrial, commercial, public, military and private units in Figure 4.14), buildings (Figure 4.15), a high nighttime light index (Figure 4.16), a low NDVI index (Figure 4.17), and a high POI density (Figure 4.18 and 4.19), the model may become confused in these cases. Furthermore, building features are critical in the regression model, as demonstrated by feature importance, and they are widely distributed in grid cells with no census data, such as industrial and commercial areas. However, because there is no such information in the building dataset, we do not assign any penalty if the building is not residential, which may have caused some misunderstanding in the model. This corresponds to the distribution of the errors, as the majority of them fall under the orange colored land use, which is industrial and commercial. As a result, we can conclude that the model cannot recognize population areas when there are numerous artificial constructions especially commercial areas.

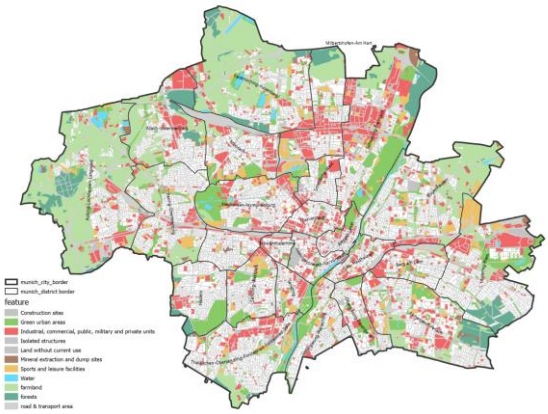


Figure 4.14 Land use of no population area in Munich

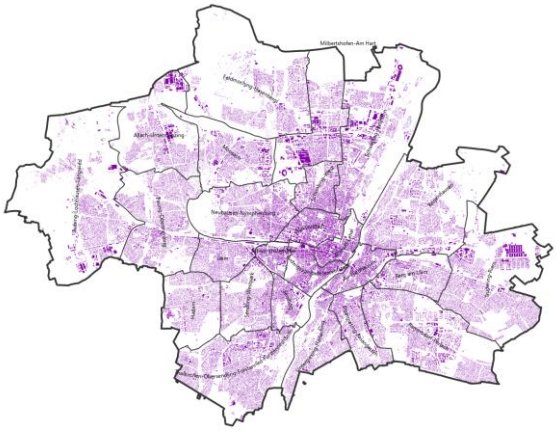


Figure 4.15 Building distribution in Munich

Figure 4.16 NTL index distribution in Munich

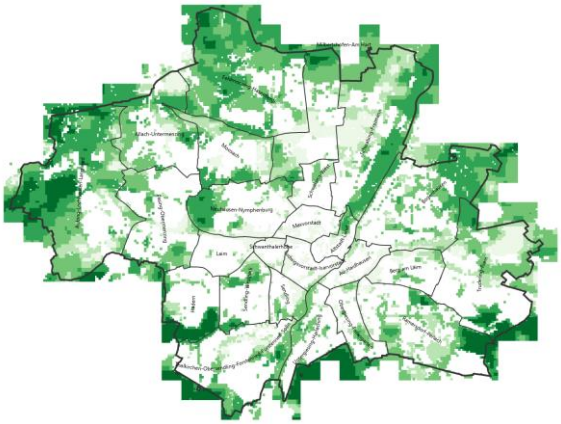


Figure 4.17 NDVI index distribution in Munich

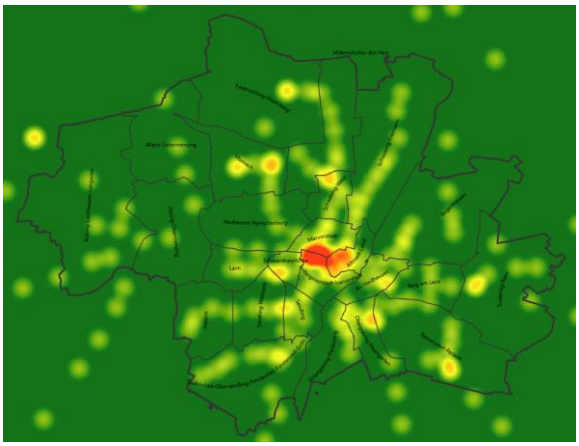


Figure 4.18 Retail POI density distribution in Munich

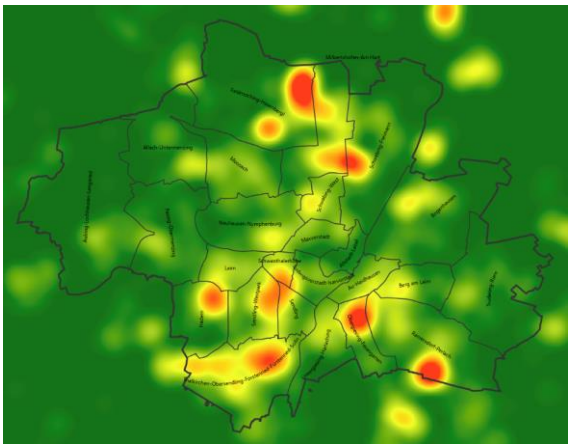


Figure 4.19 Sport POI density distribution in Munich

Furthermore, when the sum of the estimations is compared to the true population, there is a significant difference, indicating that the population is vastly underestimated. The spatial pattern of the estimations and the true values, however, are similar (Figures 4.20 and 4.21), indicating that the population values obtained by

the trained model are not reliable but their relationship is. As a result, to obtain the building population, first recalculate population values based on the estimated relationship and true data, and then allocate the results to the building level based on building volume.

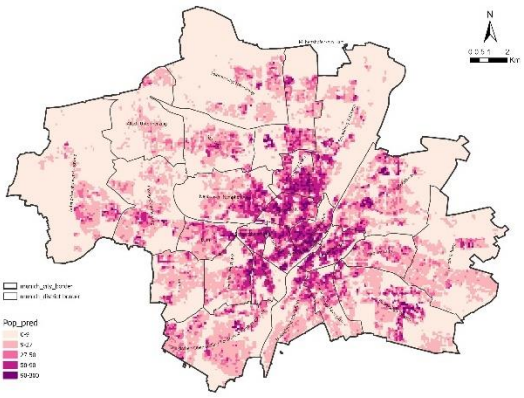


Figure 4.20 Spatial distribution of predicted population

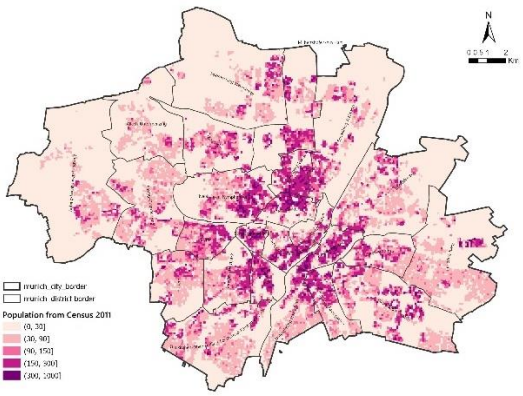


Figure 4.21 Spatial distribution of true population from Census 2011

4.4 Population Visualization

Following the estimation results, population visuals are created at the 100m*100m grid cell and building levels using ArcGIS and Kepler.gl (https://kepler.gl/), a free website for creating interactive maps.

4.4.1 100m Grid Cell Level

The 100m*100m grid cell level visuals are created based on real data, the German Census 2011. As there is lots of criticism on isarithmic map of population distribution, we just other six methods here for the visualization, including choropleth map, dasymetric map, dot map, proportional symbol map, heat map, and 3D hexbin map, from Figure 4.22 to Figure 4.27.

Table 4.3 below lists the advantages and disadvantages of each method used in this study to help you compare the visualization results more effectively. As we can see, each of them has advantages and disadvantages. As a result, selecting the best one is dependent.

*Table 4.3 Pros and cons of the applied visualization methods at 100m*100m grid cell level*

Method	Pros	Cons
Choropleth map (Figure 4.22)	• Use grid cells as the primary unit. • Color levels represent a range of values, giving the audience a quick	• Class sizes need to be carefully chosen.

	impression of how many people are present. • Groupings can be flexible in order to accommodate a wide range of values. • Effective for gaining insights into broad patterns.	• Assume that the entire unit has the same value. • It can be difficult to tell different shades apart.
Dasymetric map (Figure 4.23)	• Effective in conveying a large amount of information to the audience, such as the relationship between population distribution and nature. • Much closer to realistic population distribution than other methods.	• More difficult to create.
Dot map (Figure 4.24)	• It is easier to demonstrate variation in distribution. • Work well in black and white when color is not available.	• If the dot has been aggregated, it is difficult to determine the exact number. • Under the sample size, it is impossible to see the distribution of populations.
Proportional symbol map (Figure 4.25)	• Excellent for illustrating differences between locations. • Simple to recognize distribution patterns.	• It is difficult to read its value if the unit is too small because their symbols may overlap.
Heat map (Figure 4.26)	• Effective and efficient for gaining insights into spatial patterns such as clusters and population hotspot areas.	• It is impossible to obtain true value. • It is difficult to define an appropriate radius.
3D hexbin map (Figure 4.27)	• Useful for providing a direct impression of value variations to the audience.	• Greater efficiency in large areas with high variation. • More hardware support is required. • It is difficult to add additional information, such as city district.

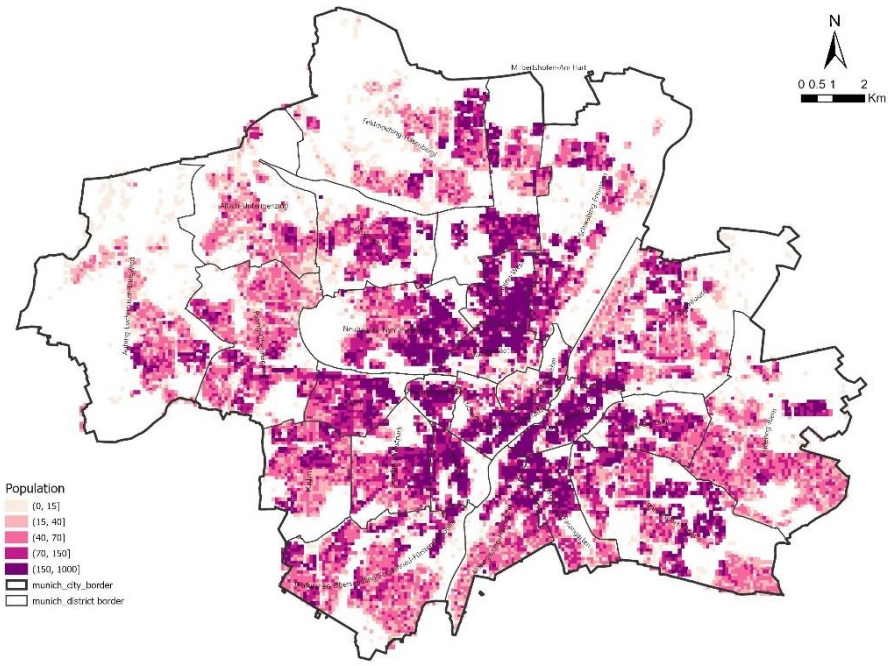


*Figure 4.22 Choropleth map of population distribution in 100m*100m grid cells of Munich*

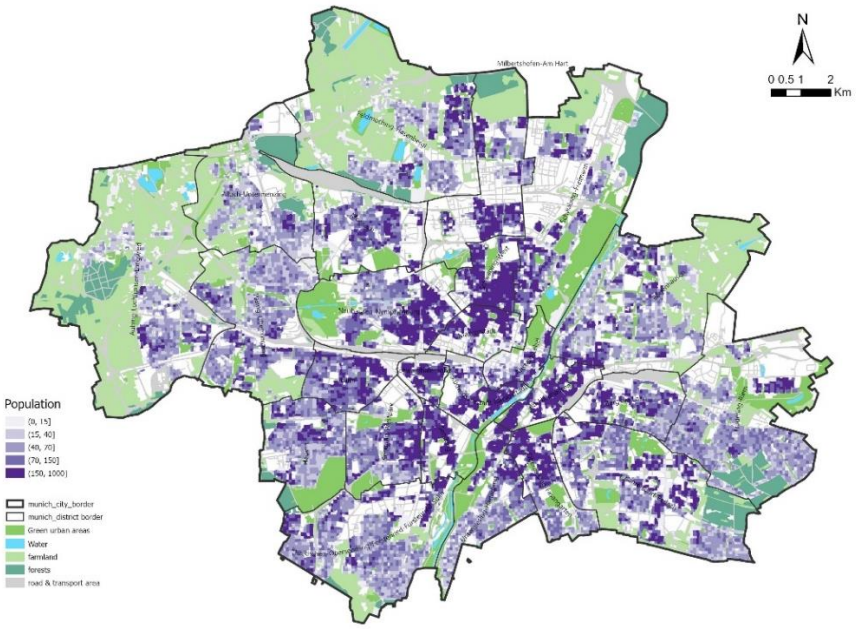


*Figure 4.23 Dasymetric map of population distribution in 100m*100m grid cells of Munich*

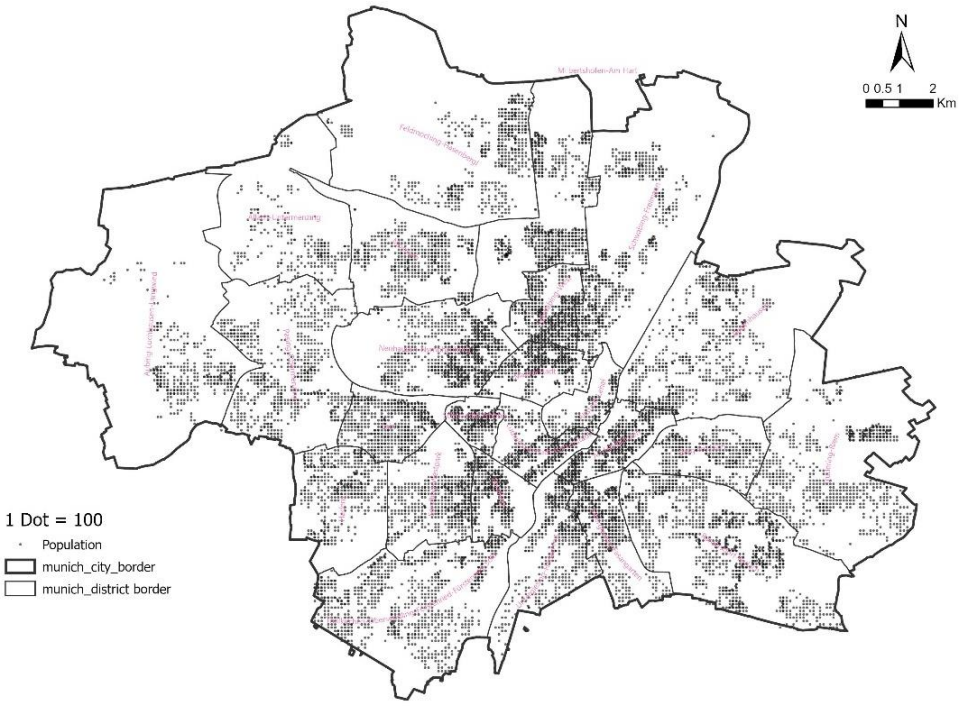


Figure 4.24 Dot map of population distribution of Munich

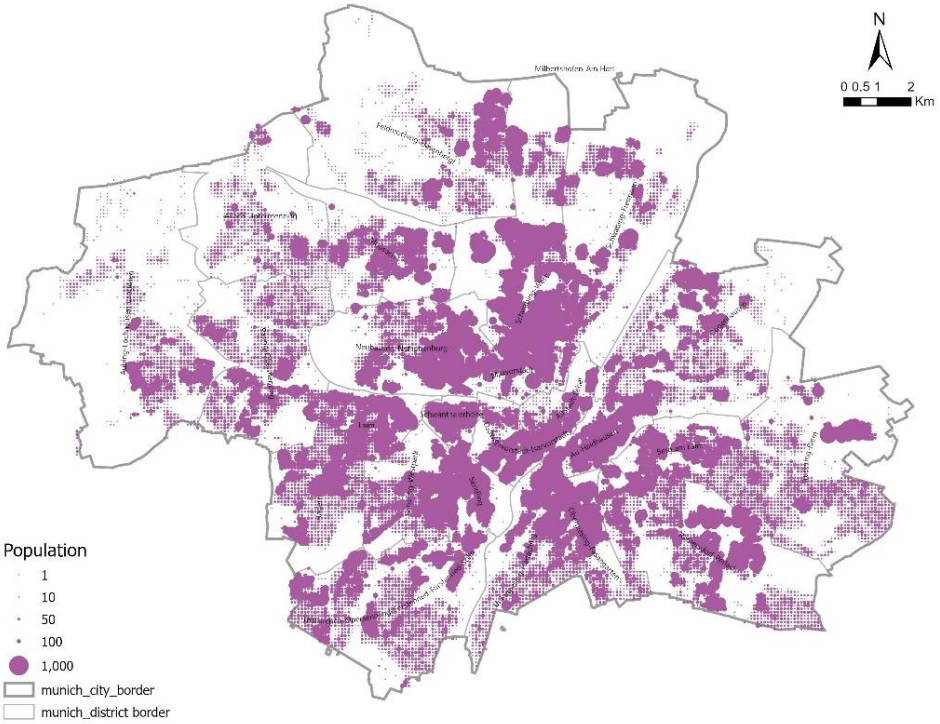


Figure 4.25 Proportional symbol map of population distribution of Munich

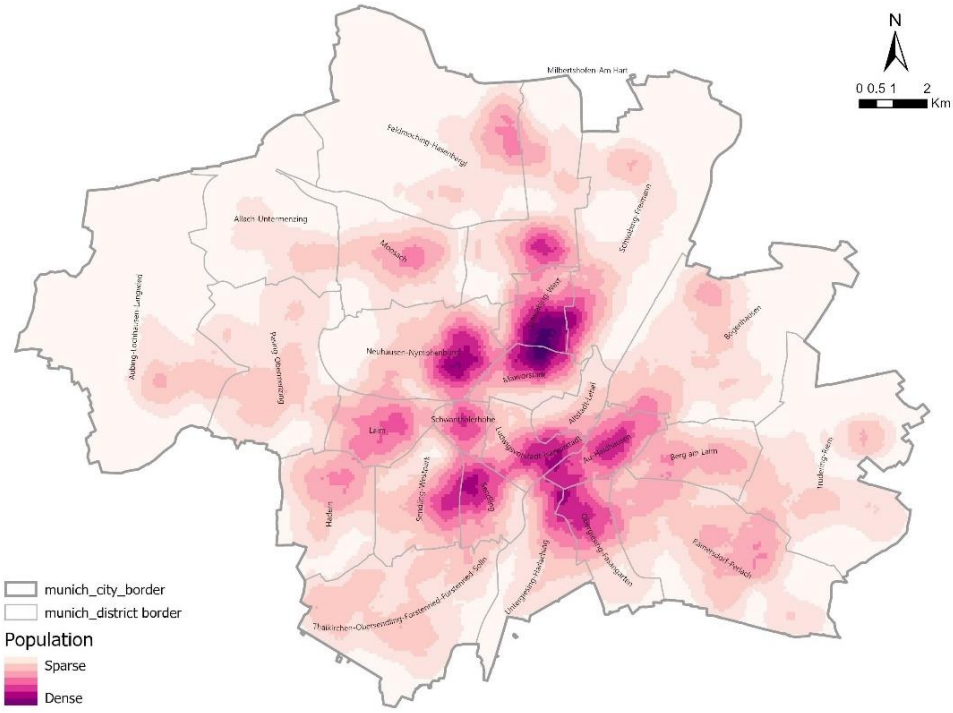


Figure 4.26 Heat map of population distribution of Munich

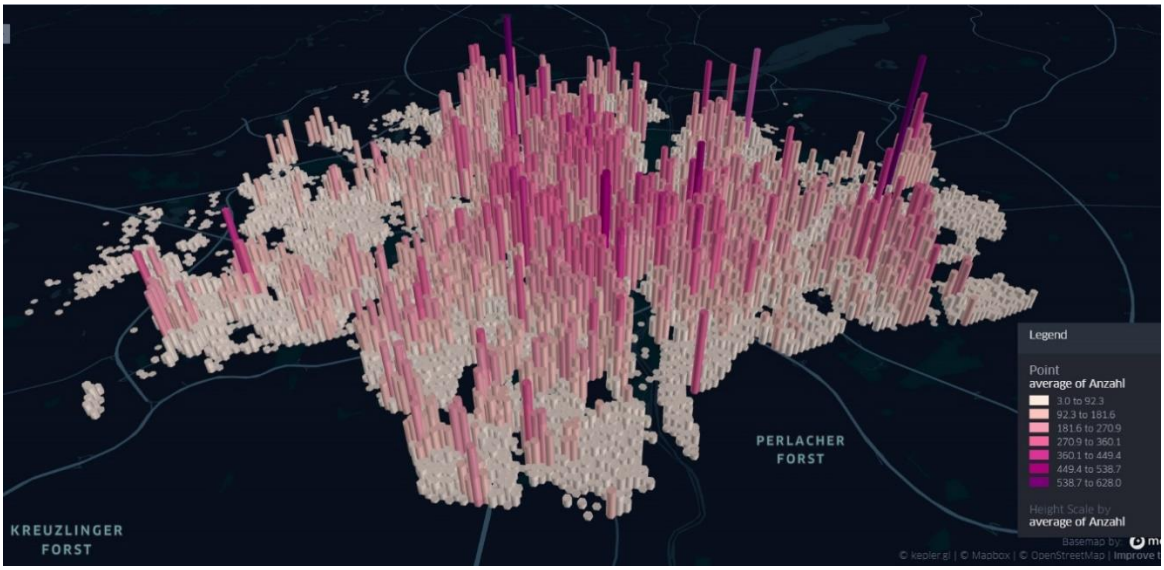


Figure 4.27 3D hexbin map of population distribution of Munich

4.4.2 Building Level

Building level estimation is obtained by allocating 50m*50m grid cell level estimation based on building volume. Choropleth maps and 3D mapping techniques are used to visualize the results. The results are depicted in the figures below.

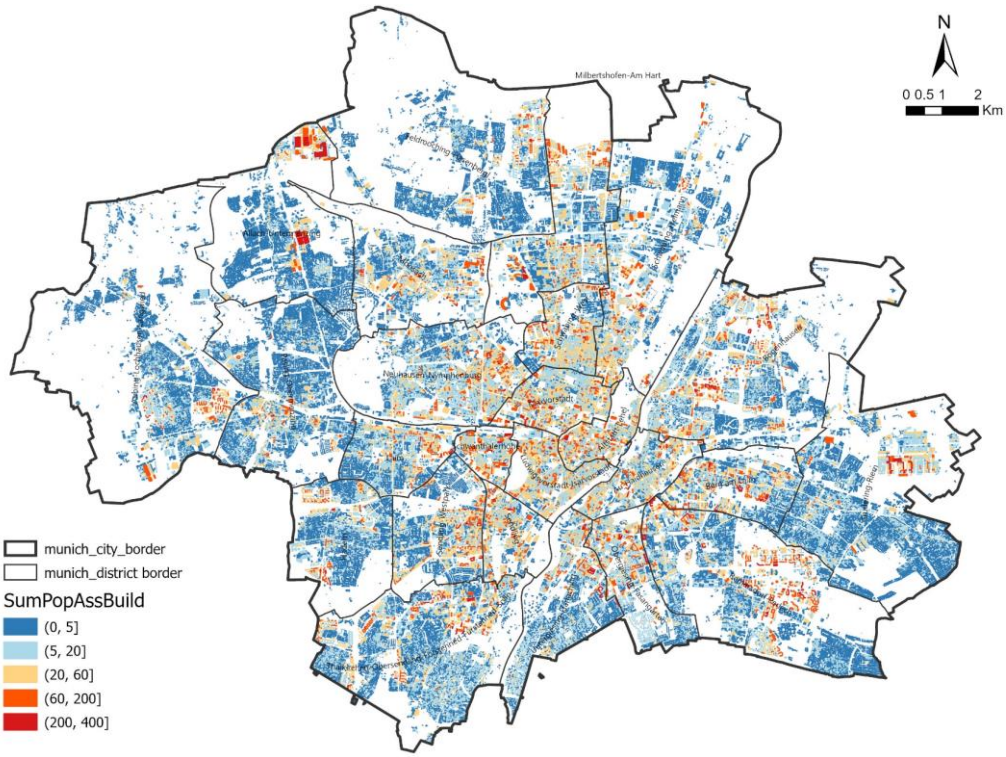


Figure 4.28 Choropleth map of population distribution at building scale of Munich

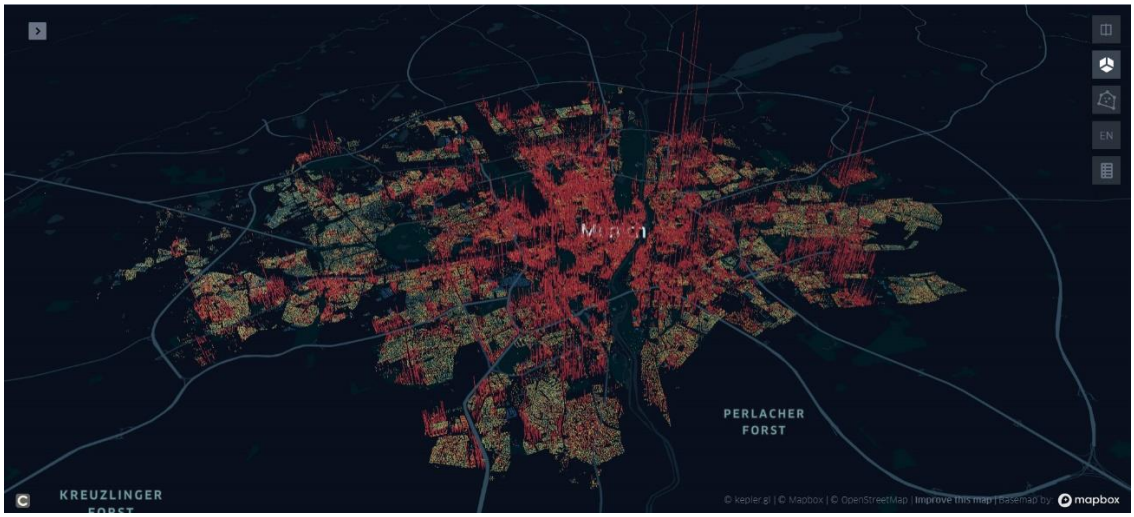


Figure 4.29 3D hexbin map of population distribution at building level of Munich

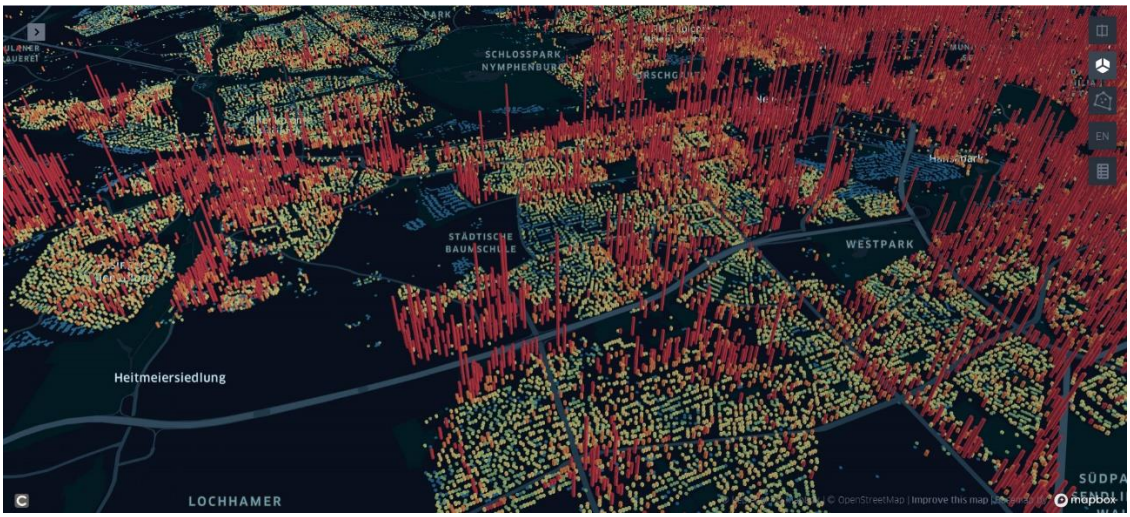


Figure 4.30 3D hexbin map of population distribution at building level of Munich (part)

As can be seen in the figures above, the choropleth map preserves the architectural texture of Munich and can easily display population values within a specific range while displaying the general pattern of population spatial distribution at the building scale. As for the 3D mapping method, we use 3D hexbin instead because it requires a lot of hardware support if we choose to visualize it by its original shapes. As previously stated in visualization at the 100m*100m grid cell level, 3D mapping provides a more direct impression of the spatial pattern than choropleth maps, which have a greater visual impact.

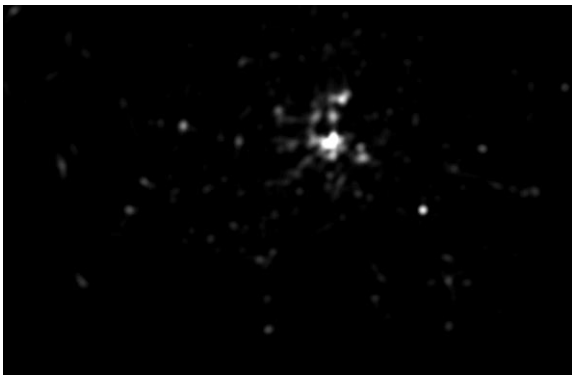
5 Evaluation

Because chapters 3 and 4 have already explained the approach used in this study, this chapter will only focus on the remaining research questions.

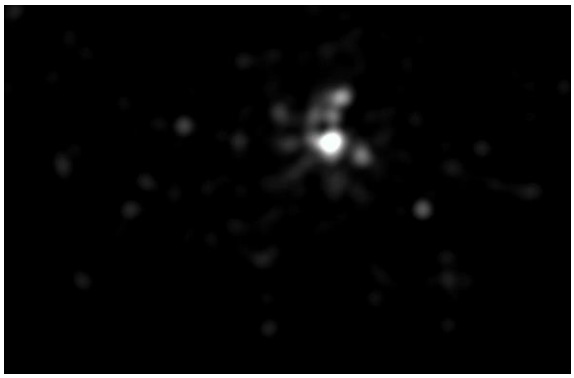
5.1 Optimal Scale

RQ2: What is the optimal scale for each variable?

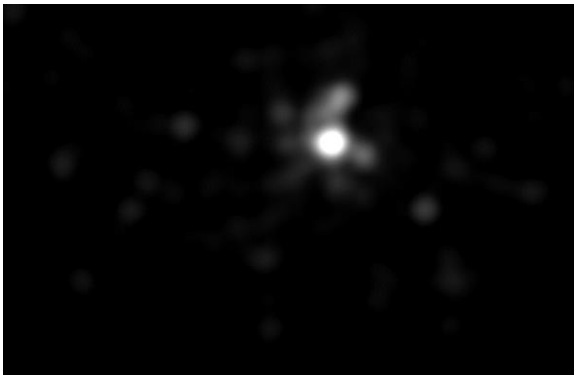
Because only POI data is calculated as density information, the optimal scale discussion is limited to the POI dataset. POI collected from OSM were reclassified into 17 categories during data preparation, including transportation, retail, restaurants, and lodging, among others. Later, 6 different scales were confirmed based on the neighborhood design principle in urban planning. These distances are as follows: 400m, 800m, 1200m, 1600m, 2000m, and 3000m. The density of retail points in those acting ranges is depicted in Figure 5.1. We can clearly see that there is a significant difference in various acting ranges, especially when the range exceeds 1200m, which increases the influencing area of the retail POI significantly. As a result, we can say the optimal scale discussion is critical when we use different types of POI information.

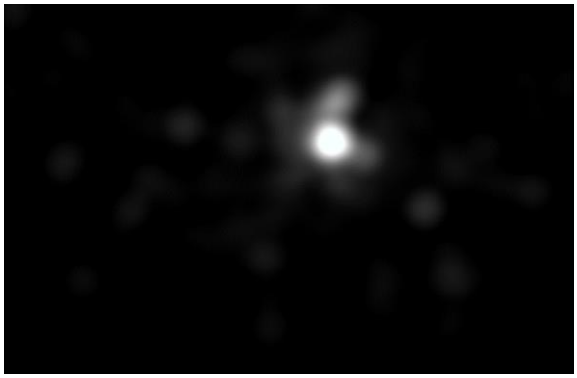


400m



800m





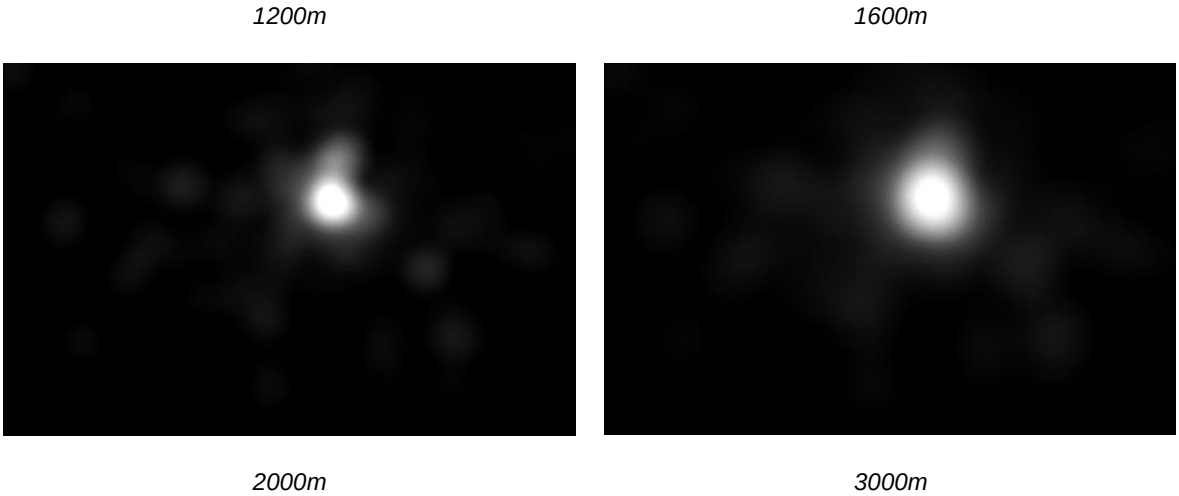


Figure 5.1 Retail POI density in six different ranges

The final input data used in the modeling process is the raster aggregation of these density values to previously obtained grid cells. The final input data contains 102 fields due to the 17 categories and 6 scales.

We use variable importance to evaluate the optical scale of the corresponding category in the POI data because variable importance can directly tell how relevant and important a feature is to the model.

In Random Forest algorithm, there is a built-in function to get the feature importance. Since Random Forest is tree based, it had a set of internal nodes and leaves. The selected feature is used to make decision on how to divide the data into separate sets. In regression model, the feature for internal nodes are selected based on how it decreases the impurity of the split. The average over all trees in the forest is the measure of the feature importance. This method can be directly achieved in calling rf.feature_importance_ function in the Scikit-Learn package.

Table 5.1 shows the result of choosing the scale with the highest importance from the six options. They differ significantly on the corresponding optimal scale, but there is still some regularity. For the majority of the categories, the optimal scale is within an 800m acting range, which is about a 10-minute walk. This means that these categories are more relevant at the neighborhood level than at the district level, and they serve as citizens' daily life infrastructure. Other categories, such as lodging, government, and leisure facilities, have a much larger influencing area that corresponds to our daily activities.

As a result, when applying POI data to population mapping, it is obviously critical to consider the difference in acting range.

Table 5.1 POI category and its optimal scale

Id	Category	Optimal Scale
1	Accommodation	3000m
2	Airport	3000m
3	Culture Facilities	800m
4	Education Facilities	400m
5	Government	3000m
6	Health Care Facilities	400m
7	Helipad	3000m
8	Leisure Facilities	3000m
9	Life Services	400m
10	Park	3000m
11	Public Transport Stops	400m
12	Railway Stops	1600m
13	Recycling Facilities	400m
14	Resorts	800m
15	Restaurant & Beverages	400m
16	Retail	400m
17	Sport Area	400m

5.2 Feature and Data Importance

RQ3: Does certain data improve the accuracy in population mapping?

1) Feature Importance

As previously stated, because feature importance can directly tell how relevant a feature is in the weight layer, it is discussed here to identify the dataset's important variables and determine which one improves population mapping accuracy.

Figure 5.2 depicts the top 15 features in the trained Random Forest model as determined by the built-in Random Forest algorithm in the Python library. We can easily see that the most important independent variable in the dataset, with a value of approximately 40%, is urban fabric. In this study, urban fabric is taken from the Urban Atlas 2012, and it refers to residential areas with densities ranging from 10% to 80%. Because our goal is to estimate residents, it is certain that citizens live in these residential areas. The next three significant variables are all about building information, implying that building data is critical for estimation. This is consistent with our assumption that residents live in rooms, and that the larger the volume of

the building, the more people it may house. Furthermore, sport areas, railway stops, recycling facilities, roads, education facilities, restaurants, resorts, and life servers are examples of POI data that are relevant to our daily lives. Since we are using Munich as the study area, these variables may differ in different use cases because people in different regions have different daily routines. NTL and NDVI are also important variables in this context, which corresponds to the meaning of these two indexes.

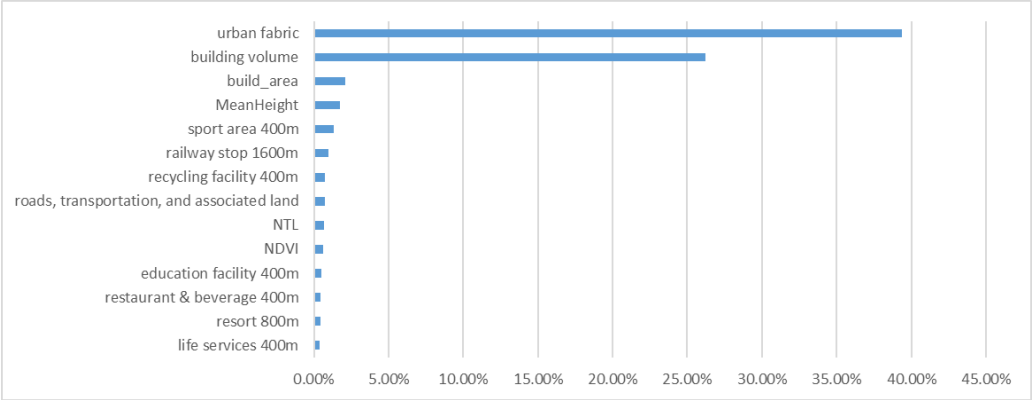


Figure 5.2 Top 15 important features

2) Data importance

Several datasets with different source combinations are applied in the Random Forest model to evaluate the importance of each dataset and see if any of them help to improve the accuracy of the final result. In particular, input data without building information (OSM Building), land use information (Urban Atlas 2012), and so on are tested, and the results are shown below in Table 5.2:

Table 5.2 Results of different datasets

Attribute	Datasets used	R^2	Mean absolute error (/km^2)	Median absolute error (/km^2)	Max error (/km^2)
---	NTL NDVI Urban Atlas 2012 OSM Building OSM POI	0.767	1520	242	42112
Without building information	NTL NDVI Urban Atlas 2012 OSM POI	0.720	1753	300	36382
Without land use data	NTL NDVI OSM Building	0.703	1854	393	41794

	OSM POI				
Without NTL data	NDVI Urban Atlas 2012 OSM Building OSM POI	0.767	1517	232	42165
Without NDVI data	NTL Urban Atlas 2012 OSM Building OSM POI	0.766	1518	240	41890
Without POI data	NTL NDVI Urban Atlas 2012 OSM Building	0.708	264	1706	39606

As shown in the table above, when the building information, land use data, and POI data are removed, the R^2 decreases significantly when compared to the original value of 0.768. The overall errors increased the most when there was no land use data, indicating that land use is the most important information in population mapping. It's understandable because land use provides basic information about where people are located, serving as a fundamental mask.

The second crucial dataset is OSM POI, which tells us which areas are more densely populated, assuming that dense areas have more infrastructure and facilities. The next important dataset is OSM Building, which differs from the previous result in that building information is the second most important feature after residential area proportion.

When there is building data, land use data, and POI information in the input dataset, the NTL and NDVI datasets do not contribute much. However, because the variables are multicollinear, this does not imply that they are useless. As a result, it may not be the perfect one to reflect which area has higher population density if we have a more precise dataset, such as OSM POI with a higher spatial resolution, but it is still a good one if a better dataset is not available.

5.3 Spatial Heterogeneity and Data Characteristics

RQ4: Do they perform same level of accuracy in different areas?

As previously stated in Chapter 4.3, our estimation tends to overestimate low-population density areas while underestimating high-population density areas, which indicates that the level of accuracy differs in different areas. This can be related to a lack of spatial heterogeneity information. The estimation is hampered by an averaging

effect imposed by the calculation of global parameters such as regression coefficients, which hides intrinsic heterogeneity in population distribution features, which expresses itself notably at population density extremes (Cockx & Canters, 2015).

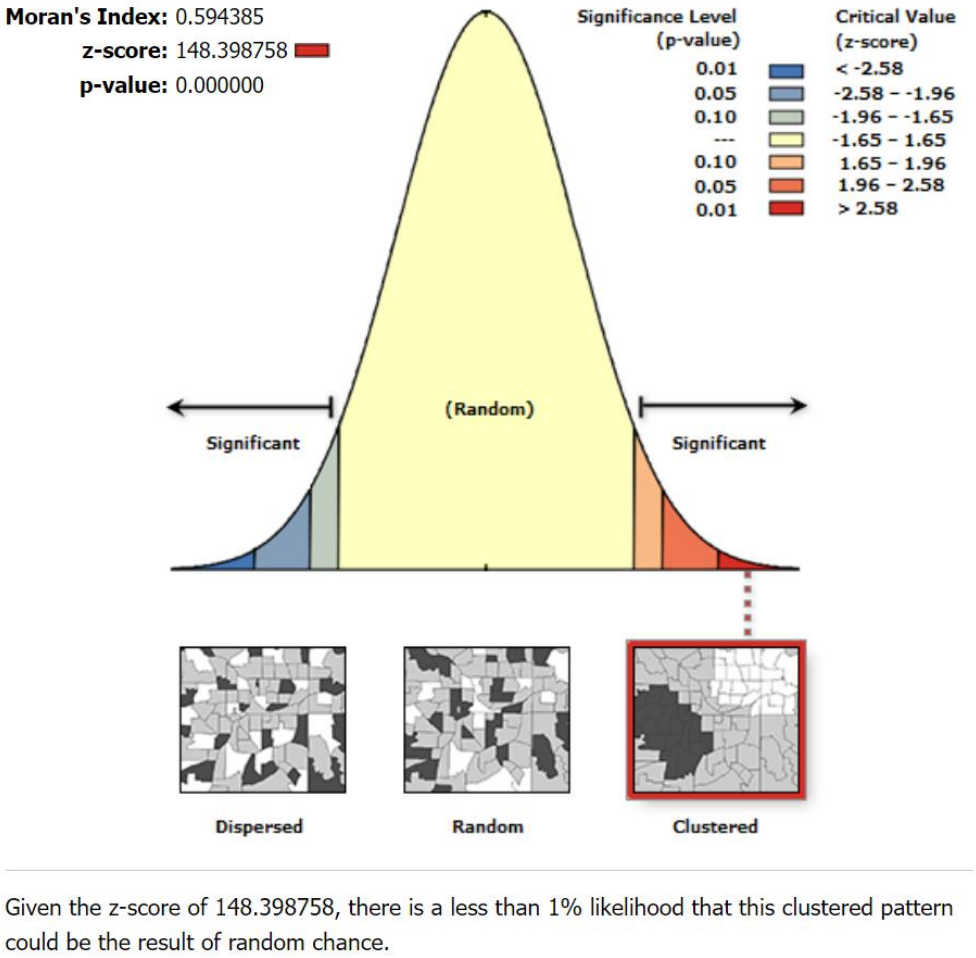


Figure 5.3 Spatial auto correlation report

The spatial autocorrelation report from ArcGIS can help with the further identification. ArcGIS Pro's spatial autocorrelation tool compares feature similarity based on both feature locations and feature values at the same time. It determines whether the pattern displayed is clustered, scattered, or random given a collection of characteristics and an associated attribute[12]. The tool computes the Moran's I index value as well as a Z score and p-value to assess the index's significance. A Moran's

12 https://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=How%20Spatial%20Autocorrelation:%20Moran%27s%20I%20(Spatial%20Statistics)%20works#:~:text=In%20general%2C%20a%20Moran's%20Index,many%20possible%20versions%20of%20random.

Index bigger than 0 refers to positive auto correlation while smaller than 0 means negative auto correlation. In our result (Figure 5.3), the Moran's index is around 0.6 with a p-value of 0 and Z score about 148, indicating that the clustering patter of these errors exist and there is a less than 1% likelihood that this clustered pattern could be the result of a random choice. As a result, it confirms our previous findings and statements, there are clear patterns of overestimations and underestimations in the neighbourhood area. Therefore, the dataset does not have the same level of accuracy in different areas.

Nevertheless, there are special problems lying in NTL data, and NDVI data. More information is presented when we overlay the significant features picked from the feature significance stage with the distribution of absolute error, providing us insights on the dataset's individual performance in different areas.

Aside from problems in high population density locations, NTL data does not operate well in the surrounding areas of densely populated regions, as seen in Figure 5.4, where the majority of the substantial absolute errors reside in the light grey area and the border of white areas. This is consistent with previous study, which found that the overglow or blooming effect of light is one of the key constraints in estimating population density due to light reflection from nearby locations and the atmosphere (Yang, Yue, & Gao, 2013). As a result, it is preferable to employ auxiliary data such as land use data in addition to NTL data.

As illustrated in Figure 5.5, the majority of the errors fall in the white area of the NDVI data, which refers to low vegetation regions, namely construction areas, indicating that the NDVI dataset works best in low population density areas.

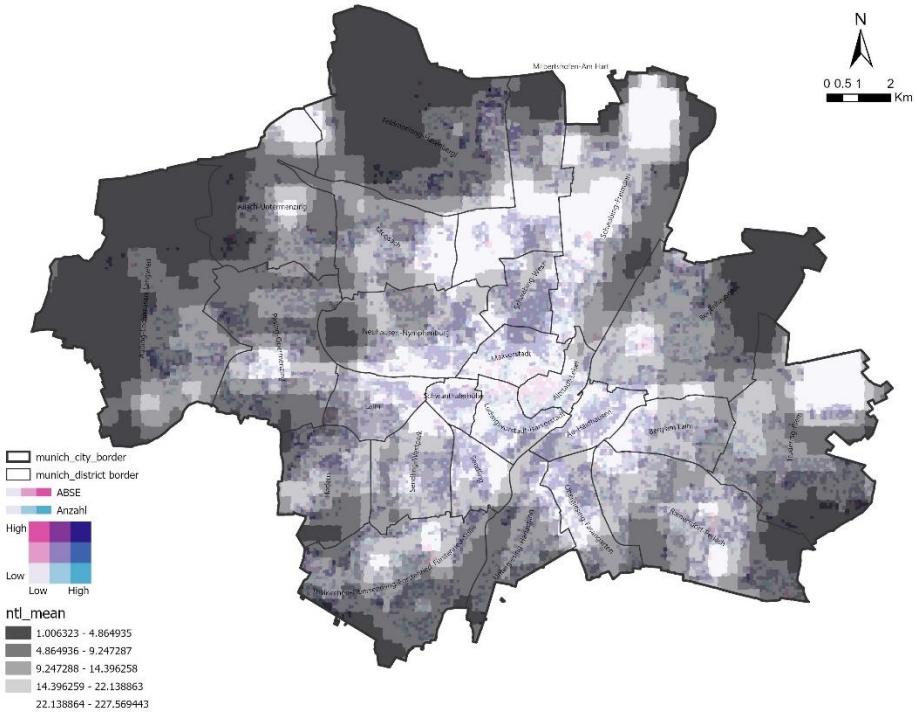


Figure 5.4 Overlay of ABSE and NTL dataset

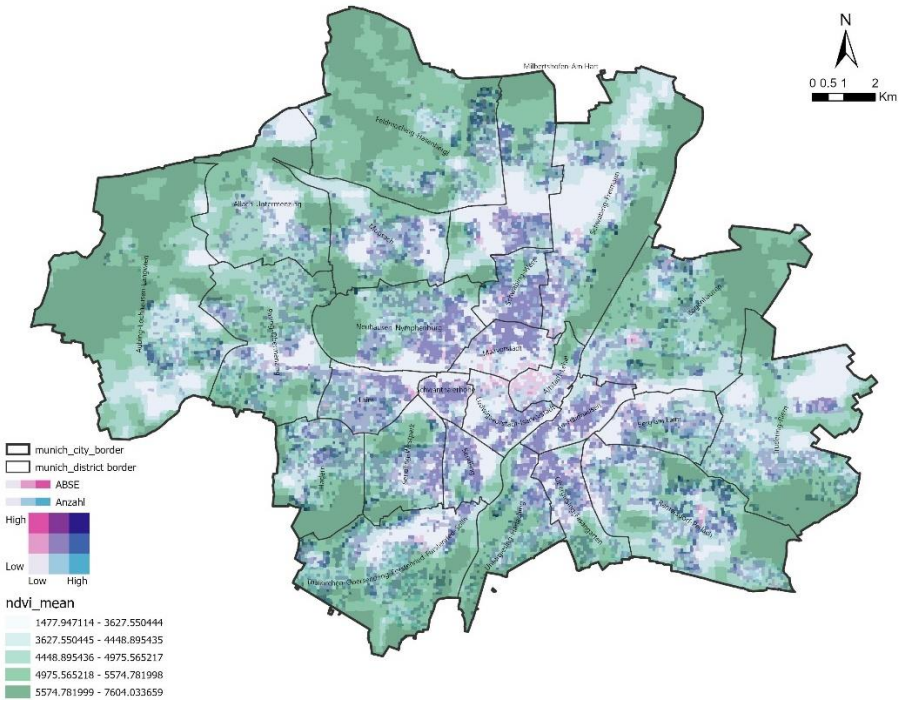


Figure 5.5 Overlay of ABSE and NDVI dataset

5.4 Model Performance

RQ5: Is the result derived from machine learning methods has reasonable spatial details?

Random Forest has a good performance at 100m*100m grid cell level gridded population mapping by evaluating its R square and mean absolute error, as stated in Chapters 4.2.4 and 4.3, as well as a good result at 50m*50m grid cell level. To learn more about the model's performance, we used a scatter plot and a choropleth map to display the predicted and true values.

The scatter plot in Figure 5.6 shows that the R square between predicted and true values is 0.57, indicating that the achieved regression model explains 57% of the variability observed in the target variable. Though not perfect, given that we are testing the model in the entire area using 50m*50m grid cells rather than a small portion of 100m*100m grid cells as in previous studies that used training and testing split at the same level, it is adequate in this context.

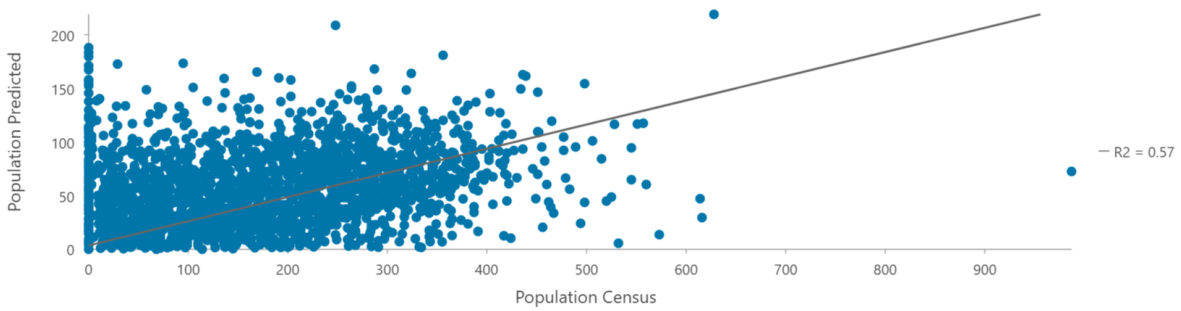


Figure 5.6 Scatter plot of true and predicted population

As shown in Figure 4.20 and Figure 4.21 previously, though the estimated population is far fewer than the true values, it is clear that the spatial pattern of the population distribution has been learned by the regression model as the two map almost show the same pattern though there are some differences in detail. We can see that the central part of Munich is largely overestimated by the regression model. Moreover, the predicted values have a tendency to be distributed in patches in medium-density areas, but these areas are actually densely packed with cells in low-density intervals.

In general, the result derived from machine learning methods has reasonable spatial details. Rather than population value itself, the population spatial distribution pattern can be well obtained, which helps us get a good population mapping result at a finer scale in another way while also recognizing feature characteristics.

5.5 Data Inconsistency and Quality Problem

1) Reference Year Difference

Six datasets were used in this study: The German Census 2011, MODIS NDVI data (2014), VIIRS NTL data (2014), Urban Atlas 2012, OSM POI (2022), and OSM Building (2022). They do not, however, have the same reference year due to availability. The MODIS NDVI and VIIRS NTL data were collected in 2014 as the earliest available year is 2014. Urban Atlas is available in three years: 2006, 2012, and 2018, so we chose 2012 as it is the closest one to 2011. Because OSM information can be updated at any time by any volunteer, the downloaded data is always current. As a result, these inconsistencies cannot be eliminated, and some errors may have occurred.

2) Inconsistency in Census Data and Land Use Data

The biggest errors coming after the referencing year difference are that more than 350 no population grid cells fall into residential area (shown in Figure 5.7 with light blue color), and around 100 grid cells with population data lie in non-residential area (shown in Figure 5.8 with light blue color), such as industrial, commercial, public and private units.

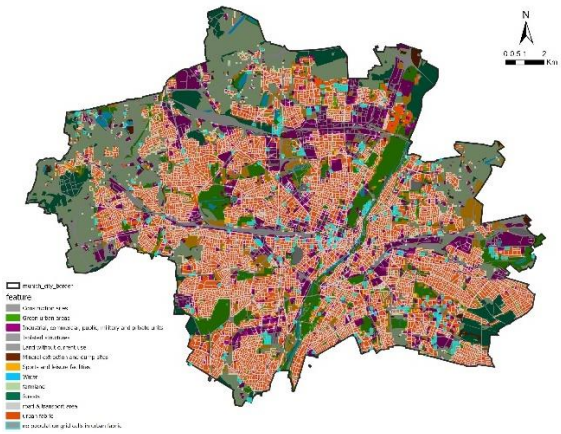


Figure 5.7 Overlay of land use and no population area fall into residential area

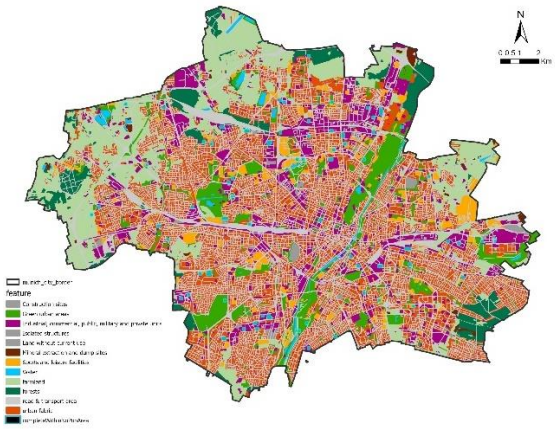


Figure 5.8 Overlay of land use and grid cells with population fall into non-residential area

3) Inconsistency in Land Use Data and Building Data

Figure 5.11 shows that there is inconsistency in both land use and building data. In Urban Atlas 2012, over 4000 buildings (in light blue color in Figure 5.9) are located in no-construction areas, which include forest, farmland, and water.

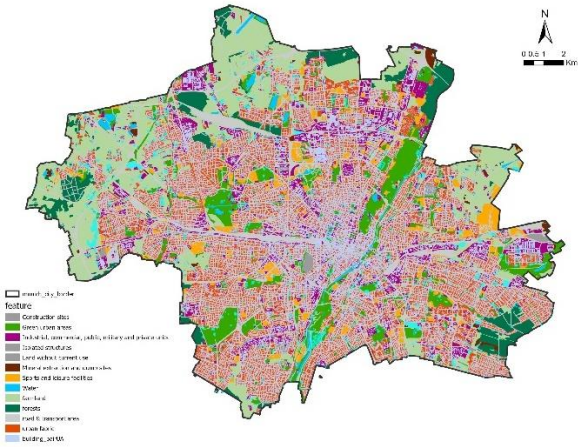


Figure 5.9 Overlay of land use and OSM Building

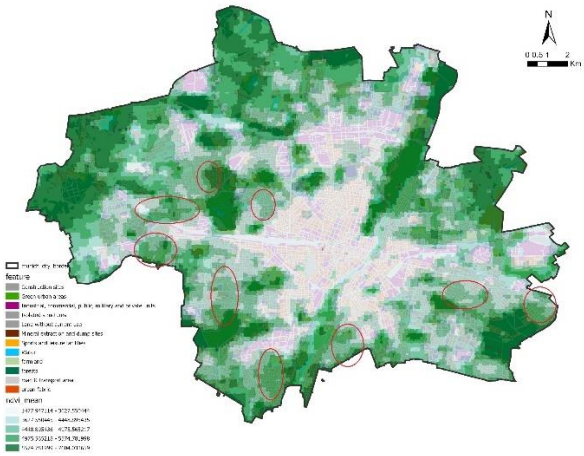


Figure 5.10 Overlay of land use and grid cells with population fall into non-residential area

4) Inconsistency in Land Use Data and NDVI Data

Many urban fabric areas in the Urban Atlas 2012 dataset are actually high vegetation regions in the NDVI dataset, as shown in the red circle in Figure 5.10, which is obviously inconsistent.

6 Conclusion

6.1 Summary

This study proposed a fine-scale population mapping framework based on machine learning by combining five different free geo datasets, including remotely sensed and semantic products. Rather than disaggregating census data, the final product at the building scale is based on the allocation of 50m*50m grid cell level estimation from five datasets and the weighted layer from machine learning. We began by aggregating and transforming all of the features from five datasets into 100m*100m grid cells, the smallest unit of German Census 2011. Then, after training the dataset with five different machine learning models, we chose the best one to get the weight layer. Following that, 500m*50m grid cells with all of the aggregated features are combined with the weight layer to generate population predictions, which are then subjected to error distribution analysis. Subsequently, buildings are assigned the 50m*50m grid cell level estimation based on building volume. Finally, various visualization techniques are used to visualize the census data at the 100m*100m grid cell level and the estimations at the building scale.

Furthermore, in the end, a wide range of results analysis is performed, including optimal scale for different POI categories, feature importance, dataset importance, spatial heterogeneity, data characteristics, machine learning performance, and data inconsistency.

The optimal scale for each POI class, as well as the top 15 relevant variables in the weight layer, are discovered using the initial function Feature Importance of Random Forest algorithm in Scikit-Learn Python library. According to the findings, most POI categories (such as culture facilities, education facilities, and retails) are more relevant at the neighborhood level than at the district level, with 400m and 800m being more relevant than 2000m and 3000m. However, some categories, such as accommodation, parks, and leisure facilities, have a much larger influencing area, with a radius of 3000m. As a result, when using POI data, we must consider various acting ranges. The top five important features in the weight layer are urban fabric, building volume, building footprint area, building height, and sporting area facility density in neighbourhood area. The first two has coefficients greater than 20%. Because there is much collinearity in all of the 119 features, especially in densely populated areas, land use, building, and POI datasets are more important in the weight layer than NTL and NDVI datasets, but the first three datasets have much higher spatial resolution than the latter two, thus contributing more to the weight layer. Moreover, estimation errors tend to cluster, with overestimations common in low population density areas and underestimations common in high population density areas, indicating that the

dataset performs differently in different areas. NTL data, in particular, does not perform well in high population density areas, whereas NDVI data performs better in low population density areas. In general, machine learning aids in the recognition of significant features in population distribution; however, because this approach relies on multi-sourced open data, it encountered many inconsistency issues, which may have resulted in significant estimation errors.

6.2 Discussion

6.2.1 Achievements

In general, the research objectives and research questions were met and widely discussed by evaluating the results from various perspectives.

A completed framework for population mapping at 50m*50m grid cell level and building scale has been proposed and evaluated by integrating multi-sourced open data with a machine learning method. Random Forest was found to be the best model in population estimation when compared to the other four shallow machine learning algorithms: Ridge Regression, Decision Tree Regression, Support Vector Regression (SVR), and K Nearest Neighbors (KNN) Regression. Following that, various thematic visualization techniques were applied to population visualization at the 100m*100m grid cell level and building scale, which was then compared to determine the advantages and disadvantages of various visualization methods.

Furthermore, optimal scales for each POI category are identified, indicating that the acting range difference between POI categories should be considered if it is used in neighborhood environment investigation. The importance of features and datasets has also been assessed, with the result that OSM POI, OSM Building, and land use data (Urban Atlas 2012) help to improve accuracy, implying that datasets with higher spatial resolution and more semantic information are more significant in population mapping than other remotely sensed products, such as NTL data and NDVI data. Moreover, because estimation errors are highly clustered, these variables do not perform at the same level of accuracy in different population density areas. However, as it shows the similar spatial pattern as true value, it indicates that machine learning can have a good result of pattern recognition instead of population value alone. Therefore, it demonstrates that the machine learning method is important in population mapping because it can help us get a good population distribution pattern at a finer scale while also recognizing feature characteristics.

6.2.2 Limitations

The whole study encounters many limitations which are easy to point out from a hindsight.

1) Limitations in Methodology

Deep learning has been shown to be more capable than shallow machine learning at acquiring and learning multisource data, achieving higher quality population spatialization (Zhao, Liu, Zhang, & Fu, 2020), and capturing the spatial autocorrelation properties of the learned grid population via convolutional operations (Robinson, Hohman, & Dilkina, 2017). However, in this study, only shallow machine learning algorithms were used, which ignored the spatial heterogeneity discussion during the modeling process. Otherwise, a clearer spatial heterogeneity analysis result can be obtained.

2) Limitations in Study area

Due to limited hardware support, this study only includes Munich city as a study area, with no rural or suburban areas. The majority of grid cells are either too densely populated or very sparsely populated. According to the German Census 2011, the overall population density of grid cells in population area is 8643/Km^2, while the rest of the grid cells have a density of 0, with a general population density in Munich of 4254/Km^2. As a result, the distribution of population density is not as broad as expected, preventing a good discussion of spatial heterogeneity as well as an examination of the performance of variables in different areas.

3) Limitations in Data Uncertainty Discussion

Although the rise of data-driven science and advanced computational capacity has enabled us to gain insights from various datasets and population mapping, we must acknowledge that much of the data explored is inherently uncertain due to limited knowledge, randomness and indeterminism, and vagueness (Chuprikova, 2019). A number of experiments in the population mapping field demonstrated that remote sensing data, such as land use and NTL data, cannot be used to conduct accurate population mapping at a fine scale, particularly in a complex urban environment (Bakillah, Liang, Mobasheri, Jokar Arsanjani, & Zipf, 2014). Therefore, it is critical to account for uncertainty when integrating big data into fine-scale population mapping.

4) Limitations in True Population Data Obtain

The majority of population mapping research uses census data, which refers to residents, as well as this study. Our modeling process predicted populations in areas

that did not appear to have residents based on census data documentation, such as commercial areas. However, due to data inconsistency in census and land use data, there are samples with populations that fall into non-residential areas, which could be true in most realistic cases. Because the spatial unit is so small in finer scale population mapping, we must have a complete understanding of true population data characteristics.

6.2.3 Future Research Prospects

Future research could go in the following directions, based on the limitations stated:

Deep learning methods and algorithms can be used to overcome the spatial auto correlation problem, or the study area can be expanded because spatial auto correlation is much stronger at the local scale and in smaller areas.

Conduct the research in a larger study area with a variety of urbanization patterns, such as urban, suburban, and rural areas.

Integrate data uncertainty analysis, which includes embedding statistical methods in the interactive environment of visual analytics to assist analysts in understanding and exploring data, as well as modeling uncertainty in the interim.

7 References

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science, 28*, 1940–1963.

Batista e Silva, F., Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M. A., Ziemba, L., . . . Lavalle, C. (2020). Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature communications, 11*, 1–11.

Batista e Silva, F., Gallego, J., & Lavalle, C. (2013). A high-resolution population grid map for Europe. *Journal of Maps, 9*, pp. 16-28.

Breiman, L. (2001). Random forests. *Machine learning, 45*, pp. 5-32.

Cabello, S., Haverkort, H., Van Kreveld, M., & Speckmann, B. (2010). Algorithmic aspects of proportional symbol maps. *Algorithmica, 58*, pp. 543-565.

Chen, K. (2002). An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing, 23*, 37–48.

Cheng, J., Zhang, X., & Huang, J. (2022). Optimizing the spatial scale for neighborhood environment characteristics using fine-grained data. *International Journal of Applied Earth Observation and Geoinformation, 106*, p. 102659.

Cheng, Z., Wang, J., & Ge, Y. (2022). Mapping monthly population distribution and variation at 1-km resolution across China. *International Journal of Geographical Information Science, 36*, pp. 1166-1184.

Chu, H.-J., Yang, C.-H., & Chou, C. C. (2019). Adaptive non-negative geographically weighted regression for population density estimation based on nighttime light. *ISPRS International Journal of Geo-Information, 8*, p. 26.

Chuprikova, E. (2019). *Visualizing Uncertainty in Reasoning.* Ph.D. dissertation, Technische Universität München.

Cockx, K., & Canters, F. (2015). Incorporating spatial non-stationarity to improve dasymetric mapping of population. *Applied Geography, 63*, pp. 220-230.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., . . . Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*, 15888–15893.

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

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science, 28*, 125–138.

Evans, R., Oliver, D., Zhou, X., & Shekhar, S. (2019). Spatial Big Data: Case Studies on Volume, Velocity, and Variety. *Big Data: Techniques and Technologies in Geoinformatics*, pp. 163-190.

Flowerdew, R., & Green, M. (1994). Areal interpolation and types of data. *Spatial analysis and GIS*, pp. 121-145.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., & Hostert, P. (2021). National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. *Remote Sensing of Environment, 252*, 112128.

Frye, C., Wright, D. J., Nordstrand, E., Terborgh, C., & Foust, J. (2018). Using classified and unclassified land cover data to estimate the footprint of human settlement. *Data Science Journal, 17*.

Gallego, F. J., Batista, F., Rocha, C., & Mubareka, S. (2011). Disaggregating population density of the European Union with CORINE Land Cover. *International Journal of Geographical Information Science, 25*, pp. 2051–2069.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PloS one, 8*, e55882.

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

Goodchild, M. F., & Siu-Ngan Lam, N. (1980). Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing, 1*, pp. 297--312.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment, 202*, pp. 18-27.

He, M., Xu, Y., & Li, N. (2020). Population spatialization in Beijing city based on machine learning and multisource remote sensing data. *Remote Sensing, 12*, p. 1910.

Healy, K. (2018). *Data visualization: a practical introduction.* Princeton University Press.

Hu, Q., Wang, M., & Li, Q. (2014). Urban hotspot and commercial area exploration with check-in data. *Acta Geodaetica et Cartographica Sinica, 43*, pp. 314-321.

Huang, Z., Ottens, H., & Masser, I. (2007). A Doubly Weighted Approach to Urban Data Disaggregation in GIS: A Case Study of Wuhan, China. *T. GIS, 11*, pp. 197-211.

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information, 2*, pp. 507-530.

Jain, D., & Tiwari, G. (2017). Population disaggregation to capture short trips--Vishakhapatnam, India. *Computers, Environment and Urban Systems, 62*, pp. 7-18.

Kang, C., Liu, Y., Ma, X., & Wu, L. (2012). Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology, 19*, pp. 3-21.

Kimerling, A. J. (2009). Dotting the dot map, revisited. *Cartography and Geographic Information Science, 36*, pp. 165-182.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, environment and urban systems, 30*, pp. 161-180.

Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems, 31*, pp. 19-32.

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., . . . others. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data, 11*, 1385–1409.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., . . . Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers, 105*, pp. 512-530.

Luo, P., Zhang, X., Cheng, J., & Sun, Q. (2019). Modeling population density using a new index derived from multi-sensor image data. *Remote Sensing, 11*, 2620.

Lwin, K., & Murayama, Y. (2009). A GIS approach to estimation of building population for micro-spatial analysis. *Transactions in GIS, 13*, 401–414.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*, pp. 93-100.

Mei, Y., Gui, Z., Wu, J., Peng, D., Li, R., Wu, H., & Wei, Z. (2022). Population spatialization with pixel-level attribute grading by considering scale mismatch issue in regression modeling. *Geo-spatial Information Science*, pp. 1-18.

Murakami, D., & Yamagata, Y. (2019). Estimation of gridded population and GDP scenarios with spatially explicit statistical downscaling. *Sustainability, 11*, 2106.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers, 104*, 80–95.

Nong, Y., & Du, Q. (2011). Urban growth pattern modeling using logistic regression. *Geo-spatial Information Science, 14*, pp. 62-67.

Nordbeck, S., & Rystedt, B. (1970). Isarithmic maps and the continuity of reference interval functions. *Geografiska Annaler: Series B, Human Geography, 52*, pp. 92-123.

Pajares, E., Muñoz Nieto, R., Meng, L., & Wulfhorst, G. (2021). Population Disaggregation on the Building Level Based on Outdated Census Data. *ISPRS International Journal of Geo-Information, 10*, 662.

Pozzi, F., & Small, C. (2005). Analysis of urban land cover and population density in the United States. *Photogrammetric Engineering & Remote Sensing, 71*, pp. 719-726.

Qiu, Y., Zhao, X., Fan, D., & Li, S. (2019). Geospatial Disaggregation of Population Data in Supporting SDG Assessments: A Case Study from Deqing County, China. *ISPRS International Journal of Geo-Information, 8*, p. 356.

Riffe, T., Sander, N., & Klüsener, S. (2021). Editorial to the Special Issue on Demographic Data Visualization. *Demographic Research, 44*, pp. 865-878.

Robinson, C., Hohman, F., & Dilkina, B. (2017). A deep learning approach for population estimation from satellite imagery. *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, (pp. 47-54).

Schug, F., Frantz, D., van der Linden, S., & Hostert, P. (2021). Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates. *Plos one, 16*, e0249044.

Šimbera, J. (2020). Neighborhood features in geospatial machine learning: The case of population disaggregation. *Cartography and Geographic Information Science, 47*, 79–94.

Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., & Tatem, A. J. (2019). Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Computers, Environment and Urban Systems, 75*, pp. 132-145.

Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2022). *Thematic cartography and geovisualization.* CRC Press.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE, 10*, pp. 1-22.

Stewart, J., & Kennelly, P. (2010). Illuminated Choropleth Maps. *Annals of the Association of American Geographers, 100*, pp. 513-534.

Sun, H., & Li, Z. (2010). Effectiveness of cartogram for the representation of spatial data. *The Cartographic Journal, 47*, pp. 12-21.

Sun, W., Zhang, X., Wang, N., & Cen, Y. (2017). Estimating population density using DMSP-OLS night-time imagery and land cover data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*, pp. 2674-2684.

Sutton, P. C., Elvidge, C., & Obremski, T. (2003). Building and evaluating models to estimate ambient population density. *Photogrammetric Engineering & Remote Sensing, 69*, pp. 545-553.

Sutton, P., Roberts, D., Elvidge, C., & Baugh, K. (2001). Census from Heaven: An estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing, 22*, 3061–3076.

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., . . . others. (2017). Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839*.

Ural, S., Hussain, E., & Shan, J. (2011). Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation, 13*, 841–852.

Wang, M., Wang, Y., Li, B., Cai, Z., & Kang, M. (2022). A Population Spatialization Model at the Building Scale Using Random Forest. *Remote Sensing, 14*, 1811.

Wu, S.-s., Qiu, X., & Wang, L. (2005). Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience & Remote Sensing, 42*, pp. 80-96.

Xia, C., Nie, G., Fan, X., & Zhou, J. (2020). Research on the estimation of the real-time population in an earthquake area based on phone signals: A case study of the Jiuzhaigou earthquake. *Earth Science Informatics, 13*, 83–96.

Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., . . . Thau, D. (2017). Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing, 126*, pp. 225-244.

Yang, X., Ye, T., Zhao, N., Chen, Q., Yue, W., Qi, J., . . . Jia, P. (2019). Population mapping with multisensor remote sensing images and point-of-interest data. *Remote Sensing, 11*, p. 574.

Yang, X., Yue, W., & Gao, D. (2013). Spatial improvement of human population distribution based on multi-sensor remote-sensing data: An input for exposure assessment. *International journal of remote sensing, 34*, pp. 5569-5583.

Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., & Zhang, Y. (2017). Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science, 31*, 1220–1244.

Ye, T., Zhao, N., Yang, X., Ouyang, Z., Liu, X., Chen, Q., . . . others. (2019). Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Science of the total environment, 658*, 936–946.

Zeng, C., Zhou, Y., Wang, S., Yan, F., & Zhao, Q. (2011). Population spatialization in China based on night-time imagery and land use data. *International Journal of Remote Sensing, 32*, pp. 9599-9620.

Zhang, C., & Qiu, F. (2011). A point-based intelligent approach to areal interpolation. *The Professional Geographer, 63*, pp. 262-276.

Zhao, G., & Yang, M. (2020). Urban Population Distribution Mapping with Multisource Geospatial Data Based on Zonal Strategy. *ISPRS International Journal of Geo-Information, 9*, p. 654.

Zhao, S., Liu, Y., Zhang, R., & Fu, B. (2020). China's population spatialization based on three machine learning models. *Journal of Cleaner Production, 256*, p. 120644.

8 Appendix

8.1 Reclassification of OSM POI

Table 8.1 Explanation of reclassification of OSM POI

Id	Classes after reclassification	Classes aggregated
1	recycling facility	recycling_glass, waste_basket, recycling_paper, recycling, recycling_clothes, recycling_metal
2	education facility	school, university, college, kindergarten
3	leisure facility	theme_park, doityourself, nightclub, cinema
4	sport area	playground, swimming_pool, sports_centre
5	culture facility	library, museum, theatre, artwork, arts_centre
6	health care facility	dentist, hospital, clinic, doctors, pharmacy, chemist, optician, veterinary, nursing_home
7	park	dog_park, garden_centre, park, camp_site, picnic_site
8	government	town_hall, police, embassy, courthouse
9	retail	Supermarket, gift_shop, market_place, sports_shop, beauty_shop, bicycle_shop, kiosk, shoe_shop, mobile_phone_shop, furniture_shop, computer_shop, outdoor_shop, bookshop, mall, butcher, clothes, toy_shop, jeweller, florist, greengrocer, caravan_site, video_shop, car_dealership, stationery, department_store
10	restaurant & beverages	restaurant, bakery, beverages, bar, fast_food, biergarten, cafe, food_court
11	accommodation	hotel, guesthouse, hostel
12	life services	hairdresser, car_wash, bank, laundry, post_box, post_office, fire_station, car_rental, community_centre, toilet, travel_agent, newsagent, atm, water_works car_sharing, bicycle_rental
13	resorts	castle, public_building, attraction, tower, ruins, observation_tower, monument, viewpoint, tourist_info, memorial
14	airport	airport
15	public transport stop	bus stop, tram stop
16	railway stop	railway station, railway stop
17	helipad	helipad

Note: these classes have been deleted as they are not useful in this study from personal understanding: camera_surveillance, vending_cigarette, comms_tower, bench vending_machine, shelter, pitch, vending_parking, convenience, vending_any, telephone, wayside_shrine, fountain, drinking_water, hunting_stand, wayside_cross, water_well, track.

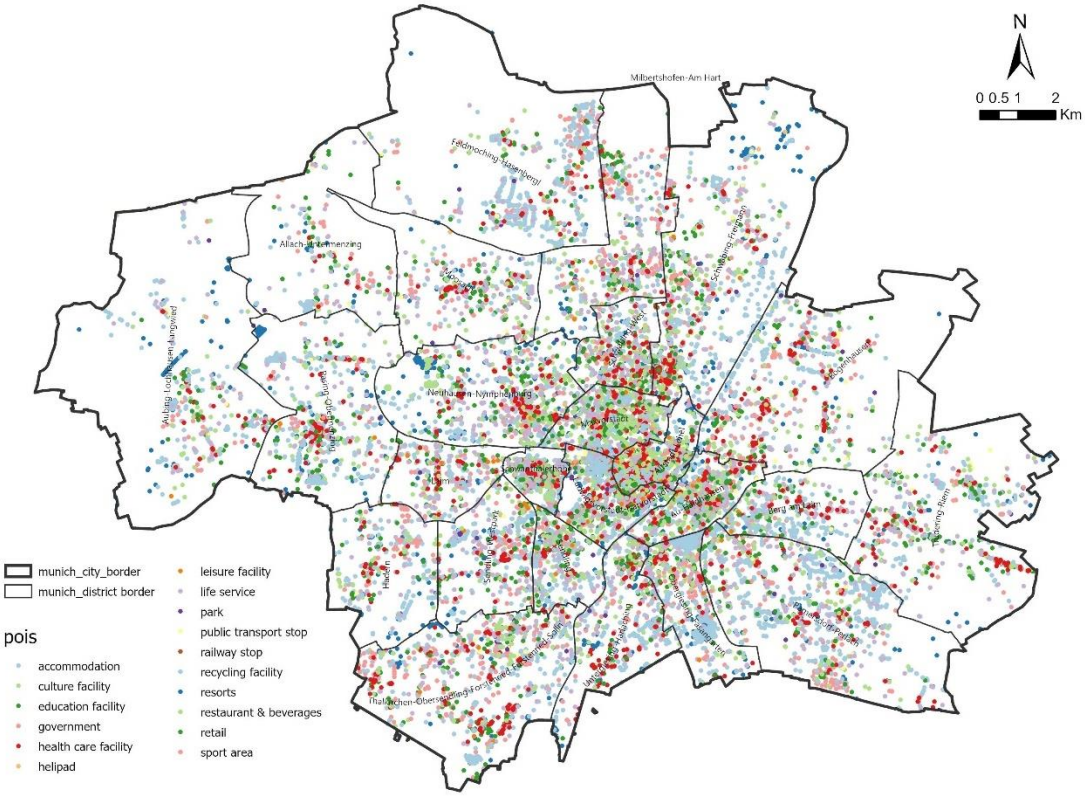


Figure 8.1 OSM POI of Munich after reclassification

Data source: OpenStreetMap

8.2 Reclassification of Urban Atlas 2012

Table 8.2 Explanation of reclassification of Urban Atlas 2012

Classes after reclassification	Classes aggregated
Construction sites	Construction sites
Farmland	Arable land (annual crops)
	Permanent crops
	Pastures
	Complex and mixed cultivation patterns
	Orchards at the fringe of urban classes
Forests	Forests
	Herbaceous vegetation associations
Green urban areas	Green urban areas
Industrial, commercial, public, military and private units	Industrial, commercial, public, military and private units
Isolated structures	Isolated structures
Land without current use	Land without current use

Mineral extraction and dump sites	Mineral extraction and dump sites
Sports and leisure facilities	Sports and leisure facilities
Urban fabric	Discontinuous very low density urban fabric (S.L.<10%) Discontinuous low density urban fabric (S.L.10-30%) Discontinuous medium density urban fabric (S.L.30-50%) Discontinuous dense urban fabric (S.L.50-80%) Continuous dense urban fabric (S.L.>80%)
Water	Water

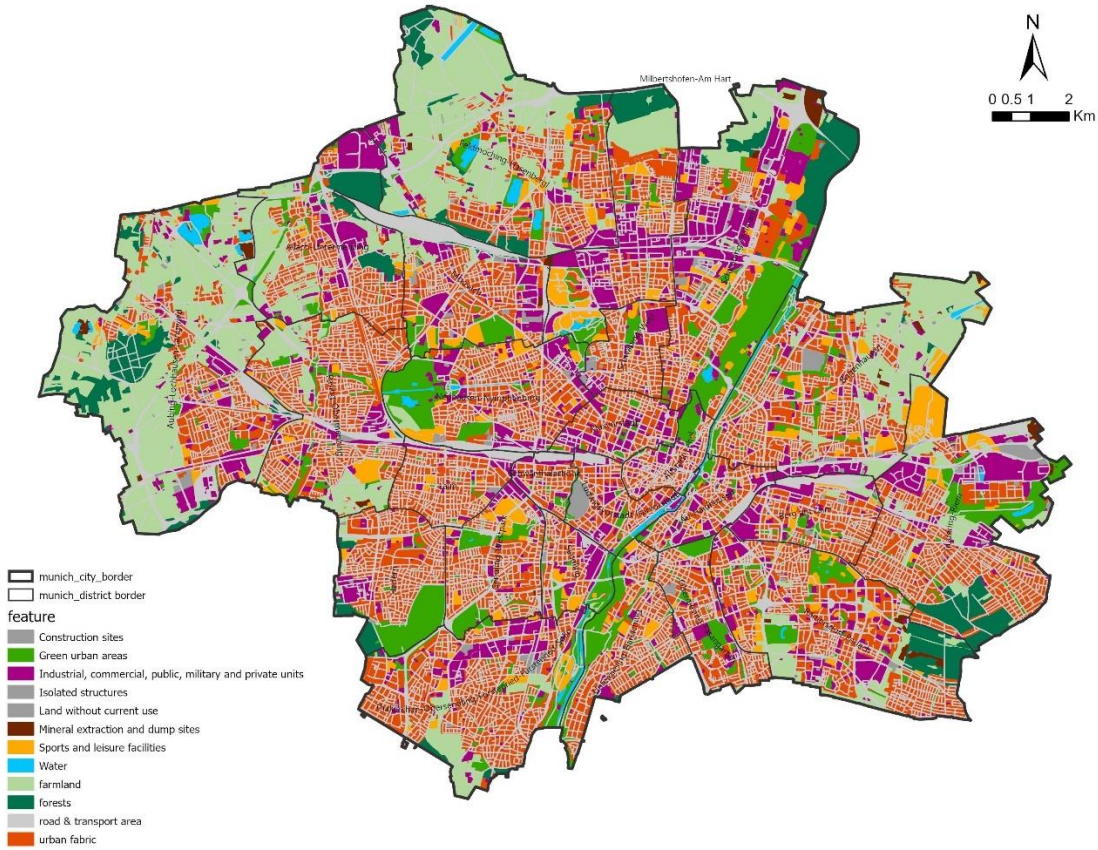


Figure 8.2 Urban Atlas 2012 of Munich after reclassification

Data source: https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012