Hochschule für Technik Stuttgart
University of Applied Sciences

DLR

Master of Science Programme
Photogrammetry and Geoinformatics
Master Thesis
Winter Term 2013/2014

Modeling Population Distribution Based on EO-Derived Data on the Built-Environment

by

Sergey Voinov

Supervisors: Prof. Dr.-Ing. Michael Hahn
Dr. Thomas Esch

Modeling Population Distribution Based on EO-Derived Data on the Built-Environment

by

Sergey Voinov

A dissertation presented in partial fulfillment of the requirements for the degree of Master of Science in the Department of Geomatics, Computer Science and Mathematics, Stuttgart University of Applied Sciences

Declaration

The following Master thesis was prepared in my own words without any additional help. All used sources of literature are listed at the end of the thesis.

I hereby grant to Stuttgart University of Applied Sciences permission to reproduce and to distribute publicly paper and electronic copies of this document in whole and in part.

Stuttgart, 28.02.2014 _____ (Your First and Family Name)

Approved by:

_____
(Name first supervisor)

Master Course Photogrammetry and Geoinformatics

Modeling Population Distribution Based on EO-Derived Data on the Built-Environment

Supervised by Prof. Dr.-Ing. Michael Hahn (HFT Stuttgart) and Dr. Thomas Esch (DLR)

Abstract

Last years, German Aerospace Center (DLR) made very big step forward in the context of global urban mapping. TanDEM-X mission, utilizing TerraSAR-X and TanDEM-X satellites, made it possible to derive very high resolution built-environment raster products – e.g., the Global Urban Footprint (GUF) settlement mask (Esch et al., 2013).

The goal of this master thesis project is to investigate the potential to model the human population distributions based on a combination of the above mentioned Global Urban Footprint product, statistical census data and – optionally - additional land cover maps (e.g., CORINE Land Cover). Resulting layers on the spatial distribution of population – once provided on global level - would be highly beneficial for sustainable spatial and environmental planning, land management as well as policy.

To conduct the study, algorithms were implemented as a tool for ArcGIS that were based on spatial/areal weighting and dasymetric mapping technics. In general, this includes the spatial disaggregating of information (Sleeter, 2004).

By comparing with officially available numbers, the results of evaluation showed good enough accuracy for the test area, namely Federal State of Bavaria. Further modifications and developments of this project are possible.

Keywords: GIS, population grids, dasymetric mapping, urban mapping

Acknowledgements

First of all, I would like to express my personal appreciation and thanks to:

To Prof. Dr. Hahn,

Thank you for supporting me in all my initiatives and your confidence. Your wise guidance and provided knowledge helped me during all time of the research. I could not wish for a better first supervisor.

To Dr. Esch,

Thank you for taking attention on my first e-mail and trust in me. Thank you for giving me such great chance, direction, support and friendship. This work was done because of you.

To Mr. Manfred Keil,

Thank you for your patience and advises. When it was necessary to concentrate on my master thesis, you did not pushing to work on my second project. Thank you for understanding.

To Mrs. Beate Baur,

Thank you for your patience, answering thousands of my questions and kind support.

To Prof. Dr. Schroeder,

Thank you for organizing this great MSc programme and your GIS lectures. I cannot imagine this work, without knowledge acquired from your lectures.

I would like to thank German Aerospace Center (DLR) administration for giving such opportunities to make master thesis internships on DLR base, use your equipment and software.

I thank HFT for the great studying conditions for international students. Your hospitality allows us to feel like home.

A special thanks goes to “Knoedler-Decker-Stiffung” foundation for giving me a scholarship for the master thesis period. This financial support made my life a bit easier for this six month.

Thanks to all authors listed in the reference section. Because of your research I was able to work.

My sincere thanks to Galina, for her patience and kind support. She is doing great job making my life better. Finally, I would like to thank my parents, for supporting me all the time and their hopes on my account.

Table of contents

List of figures

List of abbreviations

CLC - CORINE Land Cover

DEM - Digital Elevation Model

DLR - Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)

EEA - The European Environment Agency

EFTA - The European Free Trade Association

EO - Earth Observation

ESA - European Space Agency

GIS - Geographic Information System

GUF - The Global Urban Footprint

GUI – Graphical (guided) User Interface

HFT - Hochschule für Technik Stuttgart (University of Applied Sciences)

LAU - Local Administrative Unit

LC / LU - Land Cover / Land Use

MSS - Multispectral Scanner

NUTS - Nomenclature of Units for Territorial Statistics

ORNL - Oak Ridge National Laboratory

RMSE - Root Mean Square Error

RTAE - Relative Total Absolute Error

SAR - Synthetic-aperture radar

TDM - TanDEM-X satellite

TSX - TerraSAR-X satellite

1 Introduction

The United Nation reports that in the beginning of 21st century the global urbanization process passed the historical threshold, the proportion of people living in urban areas exceeded the share of living people in rural regions. It is expected that by 2050 the world population will be increased up to 9.3 billion and the population living in urban areas will grow up from 3.6 billion in 2011 to 6.3 billion in 2050 (United Nations, 2011).

Such urbanization trend raises high interest from decision makers in different fields of territories management. Impact of urbanization could cause consequences on socio-economic, political and mainly ecological nature. One of the tasks to be solved by means of Geographical Information Systems (GIS), in the framework of urbanization problem, is the calculation and analysis of the population distribution in a given territory.

The main obstacle to the solution of this task is a lack of spatially detailed information on the population distribution. Often, such information is available only in aggregate form (e.g., statistical numbers) at the administrative municipal or regional level. An established approach is dasymetric mapping, which disaggregates census data to a grid format with a spatial resolution corresponding to input built-up environment mapping products. Basic dasymetric mapping fundamentals are given in Theoretical background section of this thesis.

Some initiatives were done towards the modelling of global population distributions based on mentioned disaggregating technology, but still their spatial resolution in 1 km^2 is not enough to deal on a local level. Moreover, the updates of ancillary information (e.g. remote sensing products) are resource consuming activities. The most valuable research projects are “GEOSTAT Population Grid” (Eurostat, 2011), which mostly based on utilizing geo-referenced European statistical data in combination with Corine Land Cover Classification, and ORNL's LandScan (ORNL, 2012), based on utilizing remotely sensed data. The overview of GEOSTAT and LandScan are presented in Related Research section.

Populated places highly correlate to built-up areas. Availability of detailed urban masks could be beneficial in order to model population distributions. Difficulties in extraction of high resolution urbanized areas from EO (Earth Observation) as it was mentioned above, related to its high costs. Big step in this direction was done by DLR (German Aerospace Center) utilizing TanDEM-X mission (TDM). In a TDM framework were collected two global coverages of VHR

SAR (very high resolution synthetic aperture radar) imagery. This unique dataset has expanded the potential for built-environment analysis. Currently scientists from DLR working on technology for automatic deriving built-environment products from TDM imagery – the Global Urban Footprint (Esch et al., 2012).

The main goal of this master thesis research is to develop scale independent technology of modeling population distribution using DLR urban mapping products derived from SAR and optical EO-systems. In the section 2 the main objectives of the research are listed.

The developed methodology among the specifics of used data is described in the Principal chapter. In the same chapter evaluation and results are presented.

The last Conclusion part will summarize main aspects of the research.

2 Objectives

The aim of this master thesis project is to develop an approach to represent the pattern of population distributions for a given region in a raster/grid format. The calculations should be based on new DLR EO derived high resolution global products on built-environment, combined with statistical, vector information on administrative boundaries, and additional ancillary data.

First of all it is necessary to investigate the most established technics of deriving population distribution grids and analyze existing products in the matter of used data.

Based on acquired findings, the next step is to develop methodology that will handle specified DLR EO datasets in order to model population distributions.

As the next step, the software has to be developed in order to automate calculation routines and reduce user's interaction. The preferable software environment is ArcGIS.

The last objective of this study is to evaluate resulting grids by comparing with officially available numbers. Additionally, visual comparisons to existing products will be performed.

3 Modelling of Population Distribution

3.1 Theoretical background

The most common way of representation of such statistical data like population distribution is choropleth map, where one aggregated value is assigned to a specified zone (e.g. municipalities). But for better characterization of surface pattern a dasymetric map could be used, since it allows representing more detail coarse spatially aggregated information within zone boundaries.

According to ESRI GIS Dictionary, dasymetric mapping defined as "A technique in which attribute data that is organized by a large or arbitrary area unit is more accurately distributed within that unit by the overlay of geographic boundaries that exclude, restrict, or confine the attribute in question".

The dasymetric mapping method, in the form which is known nowadays, was developed and named in the beginning of 20th century by Russian cartographer Benjamin Semenov-Tyan-Shansky and popularized by American geographer Wright in 1920s and 1930s (Petrov, 2012). Although this method exists about 100 years, it did not become popular due to its laboriousness. The complication is in the fact that for each point (target map units) it is required to make complex calculations and depending on desirable degree of detalization it might be much resource consuming.

Even with the advent of the computers and GIS, dasymetric mapping did not gained popularity due to the large demands on computing power of computers. Only last 20 years (according to mass appearance of publications) the interest to the method grown up. This can be explained by growing performance of computers, functionality of GIS, and increased interest to environmental research. Generally, dasymetric mapping could be applied for modeling any kind of statistical information on a surface, but mostly it used for modeling population data.

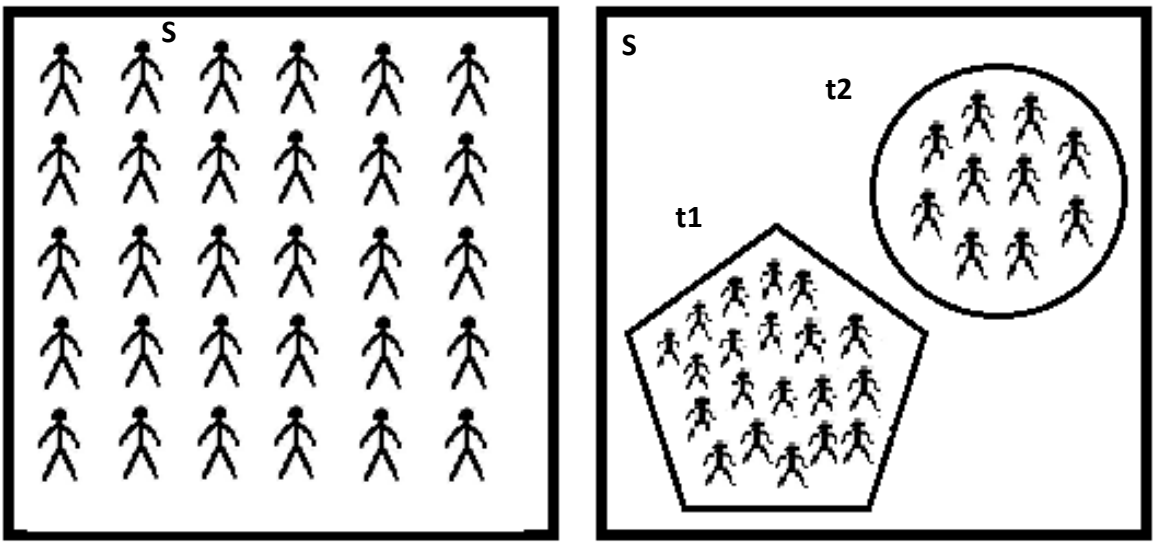
Technically, dasymetric mapping approach means transformation of data from one areal unit to another, from lower level - source zone, to the higher – target zone, e.g. from country boundary to municipalities. The process of such transformation could be interpreted as an areal interpolation (Mennis and Hultgren, 2006). The basic areal interpolation based on areal weighting, where disaggregation of source zone population information distributed between

target zones proportionally by areas of target zones. Langford (2006) defines this method as simple area weighting, and expresses it as:

$$P' = \frac{P_S A_t}{A_S} \quad (1)$$

where P_s is population of source zone, A_s is area of source zone, P' is estimated population of target source zone and A_t is area of target zone. Note that $A_s = \sum_t A_t$.

Figure 1 demonstrates an example how does simple area weighting work, where population of 30 inhabitants from the source zone (S) was distributed to a larger scale target zones (t1 and t2) proportionally to their areas – 20 and 10 inhabitants.



a) Choropleth representation of population of source zone

b) Dasymetric representation of population of source zone based on area weighting

Figure 1 Example of simple area weighting

Given the weakness of simple area weighting regarding the realistic representation of population distributions, a set of researches were done in order to improve the accuracy of the method (Langford et al., 1991; Langford, 2005; Langford, 2006; Mennis, 2003; Mennis and Hultgren, 2006; Batista, F. et al., 2013; M.-D Su et al., 2010 and many others). Mostly these improvements are based on including some additional ancillary information into a model.

Usually the ancillary information represents internal structure within a target zones, e.g. – built up masks (footprints), land cover classification, road networks, maps and etc. Using this information it is possible to divide source zones where one supposed to be populated and another one unpopulated - binary dasymetric (Langford et al., 1991). Another option, to define

priorities of distribution between target zones - multi-class dasymetric (M. -D. Su et al., 2010). A general formulation of enhanced areal weighing can be expressed as follows (Batista, F. et al. 2013):

$$P' = \frac{P_s A_t W_t}{\sum_t (A_t W_t)} \quad (2)$$

where P_s is population of source zone, P' is estimated population of target source zone, A_t is area of target zone and W_t is weighing coefficient for every target zone. As it was in formula 1, here follows the same rule $A_s = \sum_t A_t$, which means the sum of the areas of all target zones is equal to the area of the source zone.

Most of enhanced dasymetric mapping technics follows in some way the logic mentioned in formula 2, but may differ in the way of obtaining W parameter, which is directly related to ancillary data (Batista, F. et al., 2013).

3.2 Established Initiatives and Data Sets

This section presents an overview of the two most popular global-scale projects on the topic of population distribution modeling. Both approaches are based on dasymetric mapping technics, but utilized different datasets.

LandScan

The LandScan is a global population database, developed by Oak Ridge National Laboratory (ORNL), USA. It represents geographical average distribution of human population within 24h per day in a global scale using grid format with spatial resolution of 1 km^2 per cell. According to the developers, LandScan is the most accurate and reliable representation of global population, nevertheless, accuracy assessment information was not found.

alculations of grids are done by distribution model which is based on multi-class dasymetric mapping technics. The model based on following input and ancillary data: land cover / land use classification, road networks, elevation models, urban areas, village locations, and high resolution imagery. The algorithm calculates a composite probability coefficient for each grid cell, which used afterwards for proportional distribution of known population counts within source zones. In order to increase the accuracy, manual verification and improvement takes

place in the methodology. Most of the manual corrections are made to urbanized areas, due to not reveal urban properties in land cover/ land use classification errors, which can be eliminated by visual cross-checks of high resolution imagery. Example of LandScan map demonstrated on figure 2.

Specifications of LandScan are:

- Format: Raster GRID;
- Projection: Geographic; UTM;
- Datum: WGS84;
- Resolution: 0.00833 decimal degrees; 1x1km cell-size.

For US territory, LandScan provides higher spatial resolution data (up to 90m^2) and measure population for daytime and nighttime scenarios. This product widely used in US Federal Government Agencies.

Data availability: commercial licenses, by request. Free sample dataset available for Cyprus area.

The overview of LandScan based on official documentation of the product (ORNL, 2012).

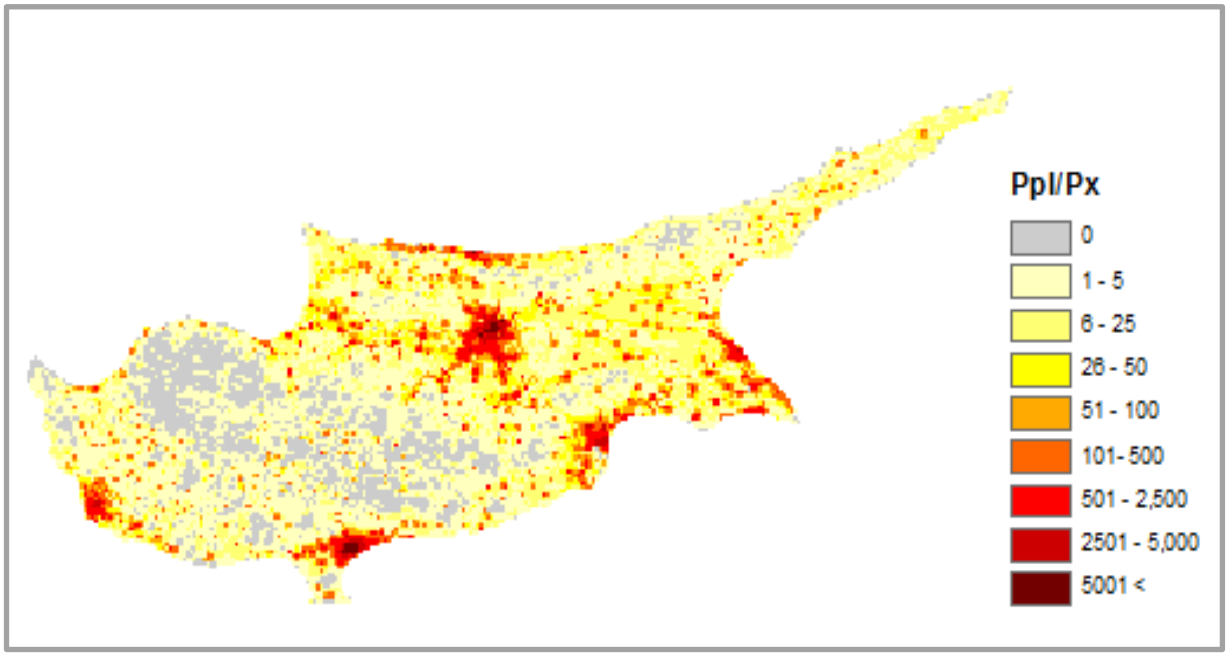


Figure 2 Sample of LandScan product (Cyprus Area)

GEOSTAT 2006 population grid

The GEOSTAT 2006 is a prototype of population grid with reference year of 2006 presented by EUROSTAT. Similarly to LandScan product, this grid represents the information on population distributions at 1 km^2 of spatial resolution for all EU countries, except Cyprus and for the four EFTA countries.

Production of GEOSTAT 2006 was separated between several institutions of European countries and then standardized and assembled into a single dataset. From country to country there were applied different technics of deriving population grid, but general methodology was similar. The calculations are based on disaggregating (dasymetric) model of population per local administrative unit (LAU), which corresponds to NUTS3, district or another minimum administrative unit, depending on country politics. In case of Germany, LAU corresponds to Amt. The principal utilized dataset were:

1) Degree of soil sealing 2006;
2) Administrative boundaries of LAU;
3) CORINE land cover classification 2006;
4) Open Street Map data.

The specifics of low resolution of 1 km^2 in GEOSTAT 2006 explained by many factors, starting from lack of reference of population information in some regions, difficulties of aqcuiring/processing of high resolution imagery and finishing with the most valuable reason – confidentiality, due to different security polices of membering countries.

Accuracy assessment. According to the official information, the grid information reliable approx. 90% in the matter if the single cell populated or not. Regarding the count of inhabitants per cell the comarisons to the register data shows sighnificant differences. The error ranges from 25 % (in Netherlands) to 70 % (in Norway). Mostly, the degree of error depends in the size of LAU. Another source of error comes from ancillary dataset, e.g. land cover classification layers. Unlike LandScan, GEOSTAT 2006 had no manual improvement stage. Example of GEOSTAT 2006 population grid shown on figure 3.

Data availability: free of charge download from official web-site.

Presented information based on official information from EUROSTAT, 2011.

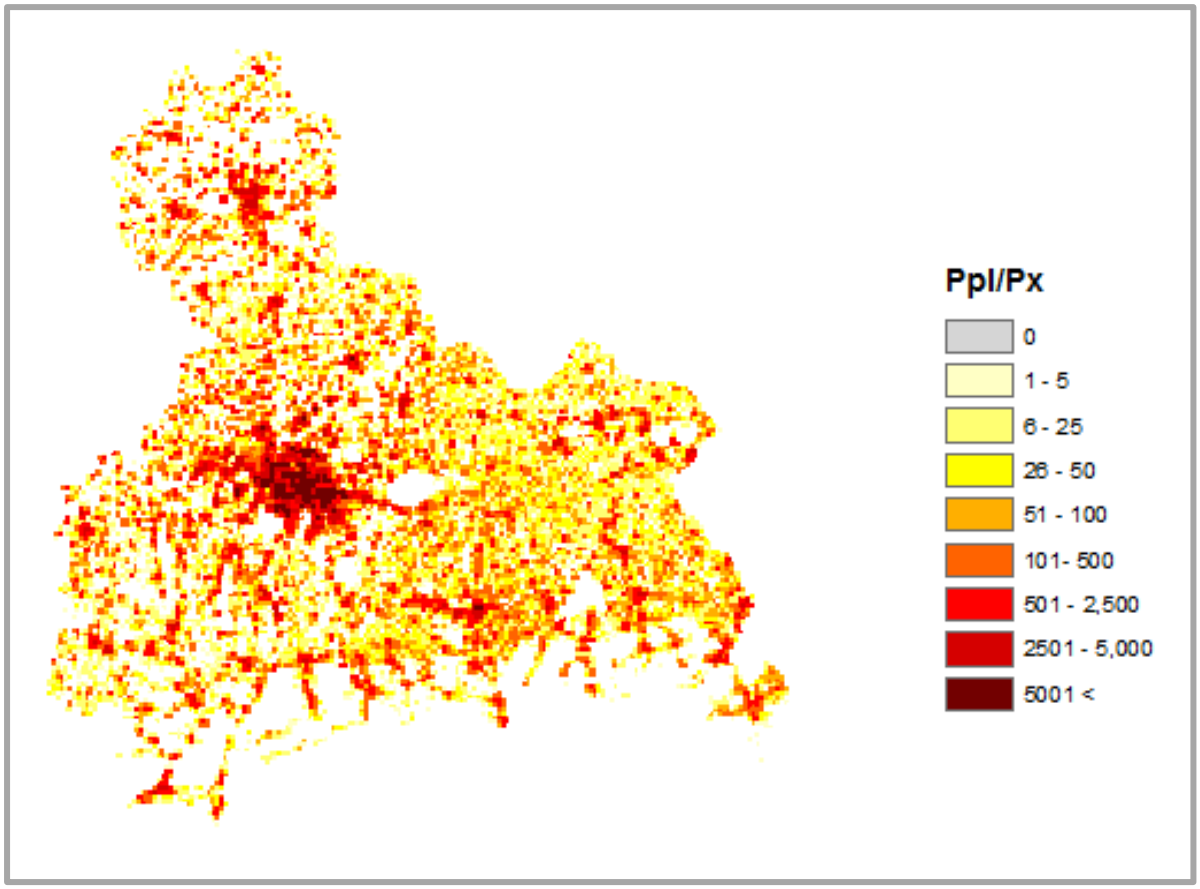


Figure 3 Sample of GEOSTAT 2006 product (Upper Bavaria)

4 Estimating Population Distribution using Novel Data on the Built-Environment

Fundamentals of the dasymetric mapping are given in the third section of the document. Generally, this approach has been applied many times in the context of modeling of human population distribution (Wright, 1936; Eicher C. L. and Brewer C. A. 2001; Langford, 2007; Mennis, 2003; Trusty 2004; ORNL, 2013; EUROSTAT, 2011) and demonstrated its effectiveness. Nevertheless there is still no special standard for mentioned approach. Most of the published methods were developed for specific territory and strict to a certain dataset. Some attempts were done in order to create a universal tool (e.g. Sleeter, 2007), but in the fact, it was not suitable for the available input data and defined scenario.

In the first part of this section, the description of used data is given. In the second and third parts the developed methodology and its implementation are explained. Afterwards the acquired results are demonstrated and evaluated. Finally, this section closes with discussions concerning this project and authors future view.

4.1 Data

As it was mentioned in Introduction, modeling should be based mainly on DLR EO products with minimum third-party datasets to make it as much independent as it possible. The minimum required and optional datasets (exactly these were used for this project) are listed below:

1) EO derived raster layers:
    a. Binary mask on built-up areas - the Global Urban Footprint (GUF), or
    b. Continuous raster layer, e.g. - Imperviousness Germany 2006 (% of impervious surface) or Building Volume layer (in the perspective).
2) Vector layers:
    a. Administrative borders with attributive information on census
3) (Optional) Additional ancillary data, namely Land Cover classification layer (raster)

The Global Urban Footprint (GUF) is a binary mask showing allocation of built-up areas. The spatial resolution of GUF is 8x8m/px. This product derived automatically from TDM SAR imagery and the mean overall accuracy is around 89%.

The general workflow of deriving GUF is follows. On the pre-processing step proceeds the analysis of local speckle statistics of SAR imagery. This is necessary to highlight textured image regions. Then derived texture image, in combination with the original intensity information, analyzed by pixel-based image classification method in order to produce a binary mask, showing built-up and non-built-up areas. The potential urban scatters identified by high amplitude and heterogeneous neighborhoods. Afterwards, coming through the set of thresholds and filters the true textures and then built-up areas extracted. Finally the built-up areas layer modified by focal maximum, to close small gaps, and focal minimum filtering to prevent the smoothing of settlements outlines. The output of the last procedure is GUF product, which is binary raster image, where a value of one used for built-up areas and a value of zero for non-built-up areas.

Described above routines are completely automated and implemented as UF processor at DLR DFD. TDM SAR images are automatically delivering to the processor's cache. And the final products are uploading to the product library at DLR.

More detailed description of deriving and characteristics of GUF could be found in paper work by Esch et al., 2012. Figure 4 shows an example of GUF, extracted from TSX image.

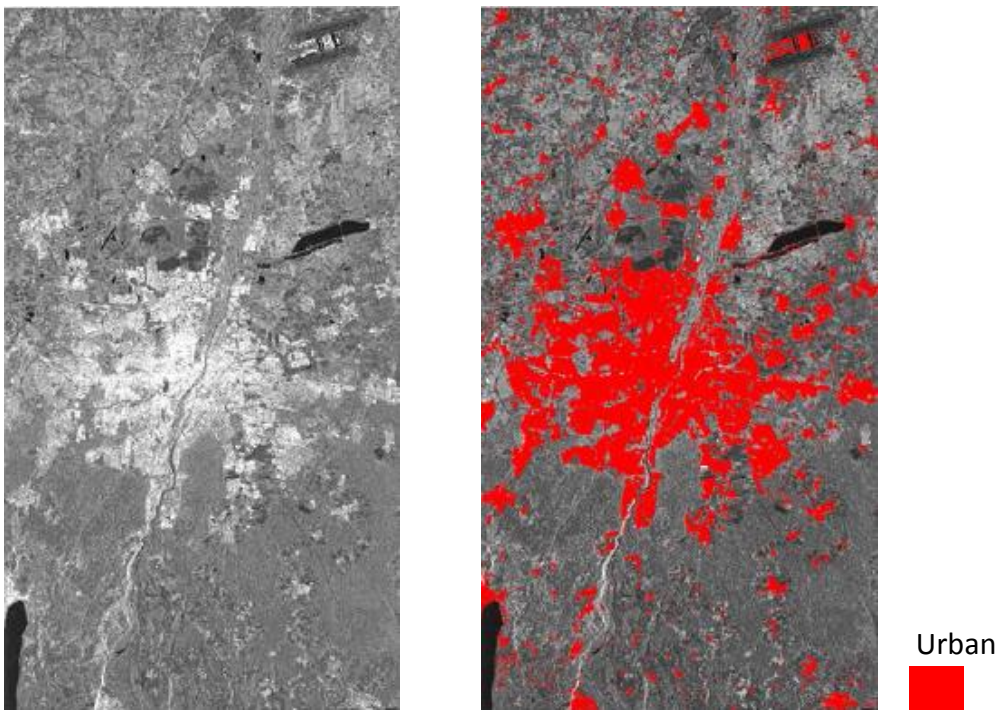
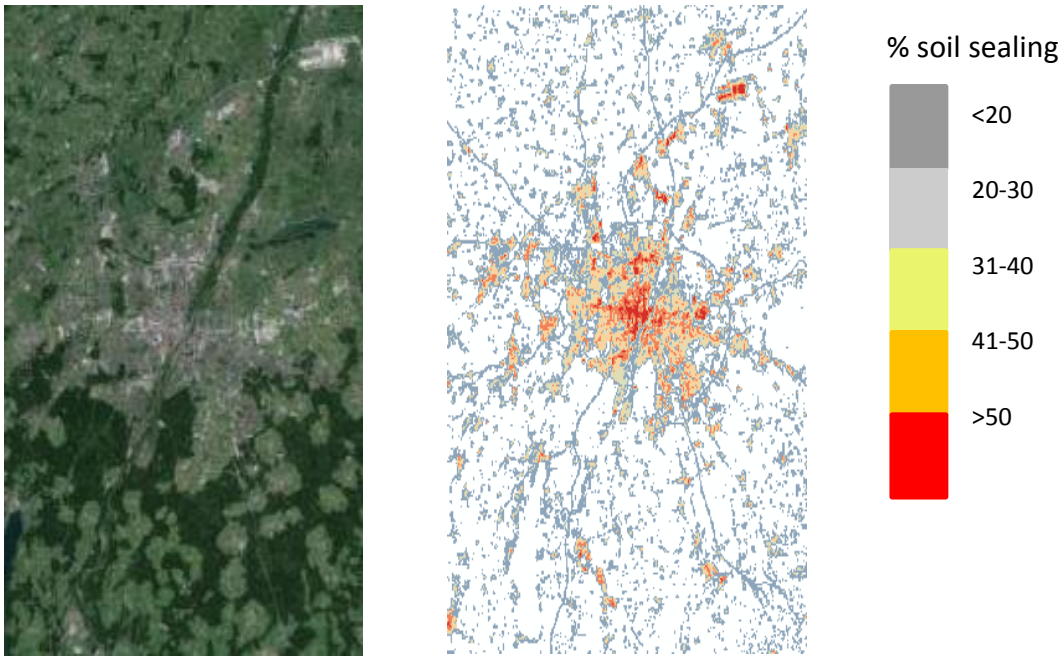


Figure 4 Original TSX image (left) and overlayed derived GUF (right)

Imperviousness Germany 2006 is a raster gray value layer each pixel of which represents the percentage of impervious surface. This layer derived from MSS imagery.

Impervious surface determines the degree of surface sealing by man-made objects, such like roads, buildings, etc. As a general rule, areas with high degree of imperviousness corresponds to high dense built-up areal. The methodology of estimating the imperviousness based on semi-automatic technique by means of Support Vector Machines utilizing remotely sensed data (Landsat) and cadastral information (vector). The first step of deriving imperviousness is creation of regression model by calculating of correlation between spectral bands of Landsat and reference data (cadastral information). The second step is estimation of imperviousness layer for the entire region using acquired regression model. More detailed explanation about characteristics and deriving of imperviousness layer could be found in paper works of Klein et al., 2009 and Esch et al., 2008. Figure 5 illustrates an example of imperviousness layer.



a) b)

Figure 5 Original MS optical image (a), derived imperviousness (b)

Building volume layer is a perspective product, which represents building volumes in m^3. The technology of deriving building volumes is still under developing stage, but for modeling of population distribution it might be ideal case since it is the most adequate description of probable human locations (EUROSTAT, 2011). This layer derived from TDM imagery as well as GUF. The concept of building volumes derived from TDM data introduced by Esch et al., 2012. Sample layer with building volumes demonstrated on Figure 6.

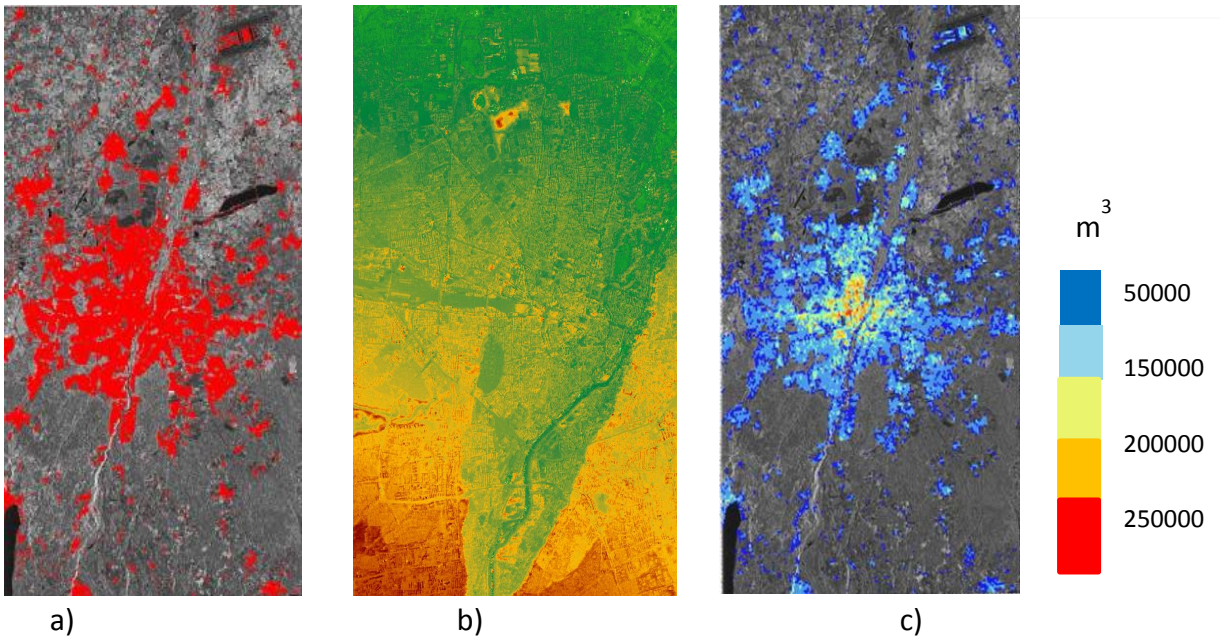


Figure 6 Derived GUF (a), DEM derived from TDM imagery (b), derived building volumes (c)

Administrative borders is a polygon vector layer describing area of interest (source zones) for modeling. This layer used to define the area within which should be done calculations.

Census information provides characterization of population distribution data on different level of administrative units. Normally this data could be acquired from statistical agencies in a table format, but not for every country such information is available, so there is no universal method of obtaining this data. Later on, this information should be integrated with vector data. In the framework of this master thesis, was utilized a set of different scale (NUTS1-NUTS3) datasets, combining census and administrative borders for Bavaria Federal state. NUTS division of Bavaria in different levels demonstrated on figure 7. Statistical (census) data for 2006-2012 year was acquired from EUROSTAT official web-site. These data are free of charge and open for everybody.

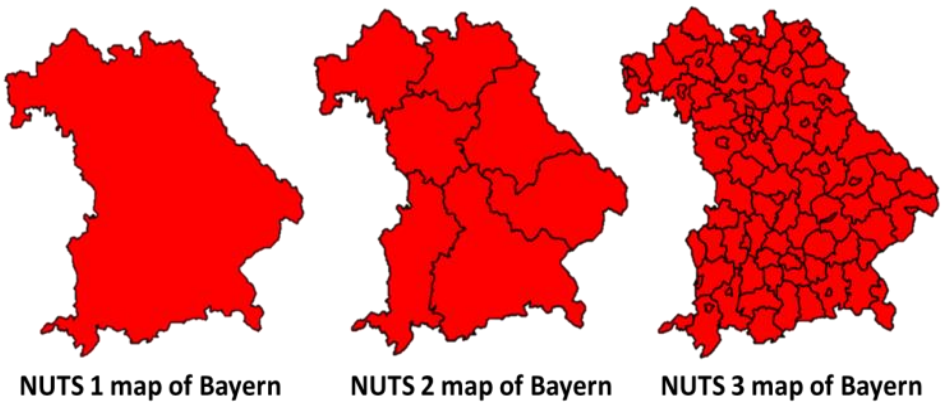


Figure 7 Maps of NUTS1-NUTS3 division of Bavaria

Land Cover and land use classification layer. In this project was utilized CORINE Land Cover (CLC) layer 2006, which consists of 44 classes. CLC 2006 was derived utilizing remote sensing data combined with previous CLC2000 dataset. Initially, CLC is a polygon vector dataset, where every polygon has a land cover class attribute. All classes of CLC are encoded with integer values and every value corresponds to the specific class. The minimum mapping unit size in CLC2006 is 25 ha. For convenient use, CLC2006 was converted to raster format using land cover class as a pixel value. The pixel size of CLC2006 raster is 20 m^2. The CLC dataset with the legend of classes demonstrated on figure 8.

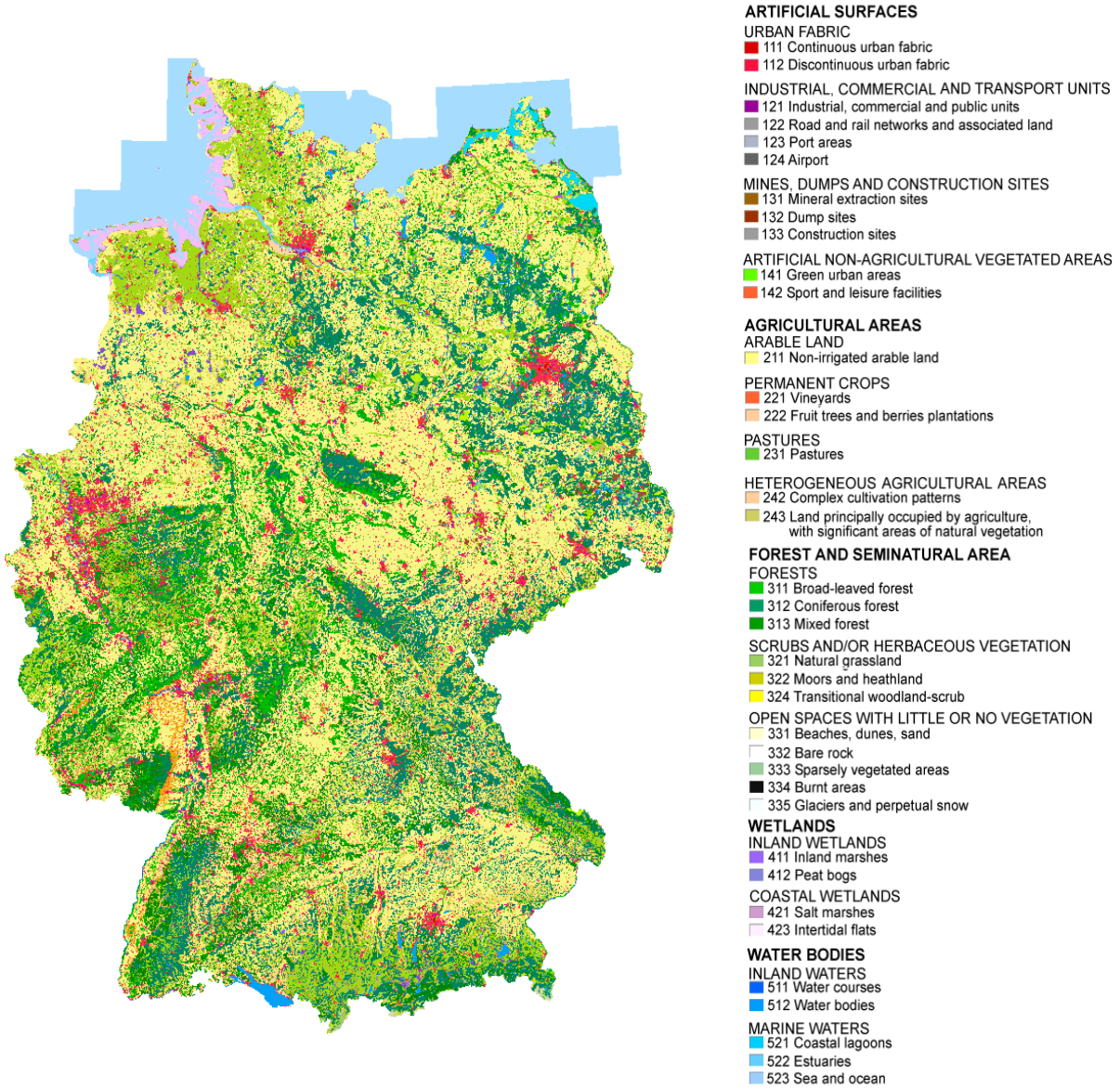


Figure 8 Raster representation of CLC dataset

4.2 Methodology

This subsection explained developed and applied methodology without any reference to the specific software. This means, that this approach might be implemented in many available open source and commercial GIS software.

Generally, proposed approach follows the concept expressed by formula 1 in the third section of this master thesis. Depending on combination of the input data there are four possible scenarios of calculations:

1) GUF
2) Imperviousness or (Volume)
3) GUF + CLC
4) Imperviousness or (Volume) + CLC

The main workflow is similar for any of suggested cases, but differs on a certain steps, according to the specific combination of input data. To perform the calculation "Three-step Supervised Dasymetric mapping" method was developed. The main idea is to divide the process into three segments:

1) Pre-processing of the input raster (filtering and aggregating in order to extract connected built-up segments), calculating of built-up segments areas;

2) Calculating of weights;

3) Deriving of final population grid. On this step takes place distribution of known population of the source zone to the pixels, which initially marked as built-up. The estimation of pixel population can be generally expressed as:

$$P' = \frac{SP \times W_t}{\sum_t W_t} \quad (3)$$

where P' is estimated pixel population, SP – total population of the source zone and W_t is weighing parameter for every target zone, which indirectly related to the population density. The way of calculating W_t is different for every combination of input data.

GUF and Imperviousness cases

For the modeling purpose, it was assumed that there should be a linear correlation between population and built-up areas. Of course, such assumption is the oversimplification of the real situation, therefore additional weighing factors are needed. As it generally known, the settlements is very heterogeneous substance and population density can be different within one town and in most cases the average density different between different cities, depending on its size. Based on this findings and assumptions it is possible to formulize W_t as:

$$W_t = f(A_t) \times W_p \tag{4}$$

where A_t is the area of the target zone and W_p is pixel weighing parameter of current pixel;

$$f(A_t) = a + \frac{(A_t - A_{min})}{(A_{max} - A_{min})} \times (b - a) \tag{5}$$

where A_t is the area of the target zone, A_{max} and A_{min} is maximum and minimum areas of built-up segments within a source zone, a and b is minimum and maximum possible weighing coefficients. The optimal setting of a and b for Bavaria is following: a = 1 and b = 1.6.

In case of GUF-based calculation W_p parameter acquired from the shortest distance between current pixel and the border of current zone (settlement). In situation with Imperviousness or Volume layer, the pixel value itself considered as W_p since it can describe good enough the structural heterogeneity of the settlement. The correlation between population density and such data like vegetation indices, Imperviousness (% of soil sealing) or building volumes are proved and discussed by Langford (2007), Taubenboeck (2008) and EUROSTAT (2011).

Accepting the assumption that in the day-time the population density is higher in the city center than in the peripheral areas, the optimum settings of W_p for GUF-based calculations are:

$$W_p = \begin{cases} 1.5, & D < 100\ m \\ 2, & D \geq 100\ m \\ 3, & D \geq 500\ m \\ 4, & D \geq 1500\ m \end{cases} \tag{6}$$

where D is a shortest distance between current pixel and built-up area border.

Bavarian settings for a, b and W_p parameters are acquired empirically and accuracy assessment shows that dimension of the results is agree with official NUTS census information, more detailed this will be discussed in Evaluation and Results subsections.

GUF+CLC and Imperviousness+CLC cases

Additional ancillary information like land cover classification makes it possible to exclude unpopulated areas and to define priorities between populated areas. For example, it is known where built-up areas are (from GUF or Imperviousness), but without classification it is not possible to distinguish them between commercial/public units or living houses. Assuming that the most people at the day time are located in center in the public units and having such classification it is possible to model human distribution more accurate. Approach for modeling of human population distribution, using land cover classification, well-proven many times (Langford, 2006; EOROSTAT, 2011; Batista, F. et al., 2013). The formulation of obtaining W_t parameter expressed as:

$$W_t = f(A_t) \times \mathrm{W_p} \times \mathrm{C_{tp}} \tag{7}$$

where C_{tp} is land cover class coefficient of current zone and pixel. This parameter is defined by user, e.g. - for class 111 (continuous urban fabric) C_{tp} = 2; for class 112 (discontinuous urban fabric) C_{tp} = 0; for class 512 (water bodies) C_{tp} = 0. This means that there will be distributed two times more people into class 111 than in 111 and completely was excluded water class (512). But these values are not absolute since there is influence of other weighing factors - A_t and W_p.

4.3 Implementation

As working environment for this this project was selected ArcGIS 10.1 with "Spatial Analyst" extension. ArcGIS is one of the most used GIS software in DFD-DLR and it can perform all necessary calculations for this project. In order to automate entire procedure of deriving population grids a tool for ArcGIS were developed. All scenarios, mentioned in the previous subsection (combinations of input data), were implemented in one Python (ver. 2.7) script utilizing ArcPy libraries.

The source code of the script could be found in Annex A, a screenshot of user interface demonstrated on figure 9. As well as single use with GUI, this tool could be integrated into

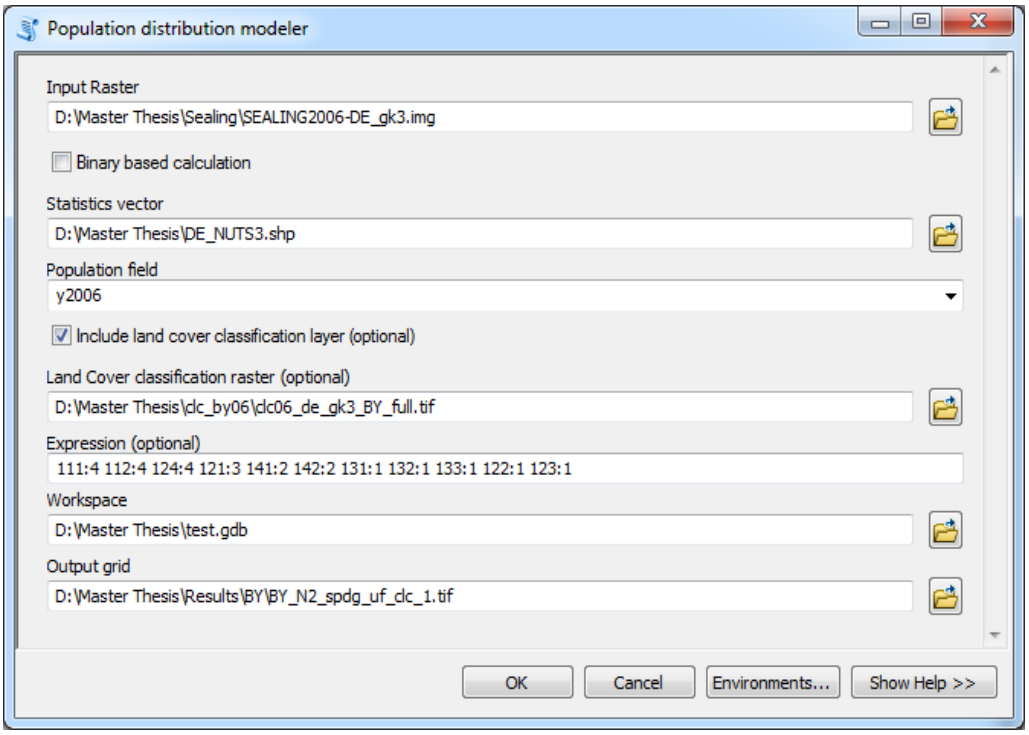


Figure 9 User interface of developed "Population distribution modeler" tool

ArcGIS model builder among the other tools to perform some complex operations.

As it can be seen from the picture, the software has simple and intuitive interface. User has to define following parameters:

1. **Input Raster.** This is the input raster on built-environment (binary or continious). All raster formats, which supported by ArcGIS can be used. Recommended pixel depth: 8 bit unsigned integer.

2. **Binary based calculation.** Should be checked in case of UF (GUF)-based calculation. If considered continuous-based calculation (imperviousness / volumes) this parameter should be unchecked (default).
3. **Statistics vector.** Polygonal vector file with boundaries and census information of source zones.
4. **Population field.** This parameter derived from the previous one. Here user has to select one of available fields with population data (numeric type).
5. **Include land cover classification layer.** This is optional parameter. Should be selected in order to include land cover / land use ancillary layer.
6. **Land cover classification raster.** Any supported by ArcGIS raster format, unsigned 8 bit integer.
7. **Expression.** String-type parameter. Used to define which classes are going to be used for modeling. A coefficient to each class has to be assigned using ":" symbol, separator between classes is space " ". E.g.: "111:4 112:2 …"
8. **Workspace.** ArcGIS File Geo Database. Used to store interim data during the calculation.
9. **Output Grid.** Path and file extension for output resulting Population Footprint. If extension is not defined, data will be stored in ArcGIS GRID format.

What is behind GUI?

To perform all necessary operations, a set of GIS operations were used. Logically these actions can be grouped into three main segments (steps), according to the mentioned in the previous subsection ideology. The generalized schematic view of the entire process presented on figure 10.

Block of inputs. Consists of two mandatory items (built-environment raster and vector file with boundaries and census information) and one optional (land cover / land use classification raster).

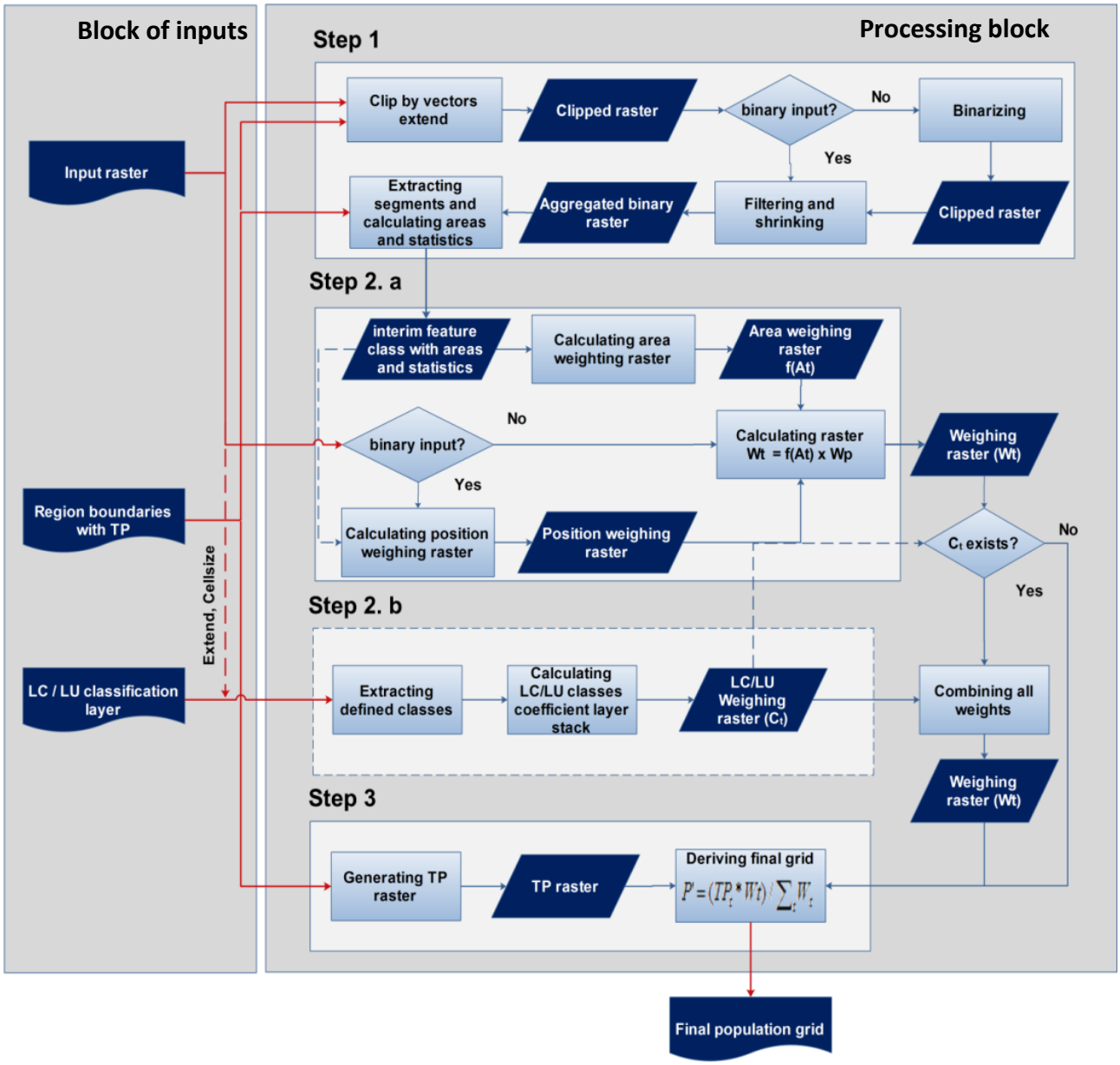


Figure 10 General schematics of population modeling workflow

Processing block

Step 1) This is the preprocessing stage of input built-environment raster. Firstly, operation "clip" were applied in order to have raster layer within provided vector file extend. This is needed to exclude areas that will not consider in calculations.

In case of using high resolution data, it is very common problem when "built-up" pixels of one settlement widely distributed and not connected to each other. Due to this fact, it is not possible to calculate the area of entire city correctly. In order to extract settlements areas as single undivided segments, various filtering and aggregation procedures are needed, therefore the next step is to aggregate and filter raster image. It is programmed to automatically detect

spatial resolution and generalize it to a pixel size of 1000x1000 m, then "shrinking" and "majority" filters are applied to remove isolated pixels.

Then derived layer were converted to a feature class. Resulting vector layer used to calculate areas and statistics (minimum and maximum areas within every source zone).

The last step is to calculate area weighing coefficients, using current, maximum and minimum areas within every source zone, the function were expressed in the formula 5.

Note, for final modeling of population distribution the original structure of input raster will be used instead of filtered one.

Step 2) The purpose of the second segment is to calculate weighing raster, which derived from combination of two (three) interim layers. First one is an areal weighing raster, which acquired by rasterizing polygons derived in the previous block. A field with weighing coefficients is used in rasterizing process.

The second weighing raster represents a pixel weights. In case of continuous-based calculation (imperviousness/volumes) the original pixel value will be used. If binary-based calculation selected, a pixel position relatively to the target zone boundary will be used. This concept presented in formulation 6. To perform such computations following GIS operations were used:

1) PolygonToLine – converts polygon feature class to lines. The input for this operation is polygon layer that were calculated for areas calculation;
2) Euclidean Distance – generates a raster layer, where pixel values represents a distances to the closest features. In this situation line features from the previous step were used;
3) Conditional Operation – to assign specific coefficients according to specified distances (see formulation 6);
4) Multiply – used to assign derived positional coefficients to original binary raster (to preserve original structure).

The third (optional) weighing raster derived from LC / LU classification layer. First of all, the software checks and read listed classes in separated raster layers, assigning appropriate coefficients to the corresponding cells. Afterwards all separated raster layers merged into one – LC / LU weighing raster.

The last step in this segment is to combine all weighing raster into one – final weighing raster. This is done by Multiply GIS operation.

Step 3) The last step is deriving of final product. Here are involved following operations:

1) Rasterizing of source population layer in order to perform raster-based calculations;
2) Complex map algebra operation, according to formulation 3.

The resulting grid – is a raster layer with the same structure as input built-environment layer with pixel values of number of inhabitants. The spatial resolution corresponds to the input data.

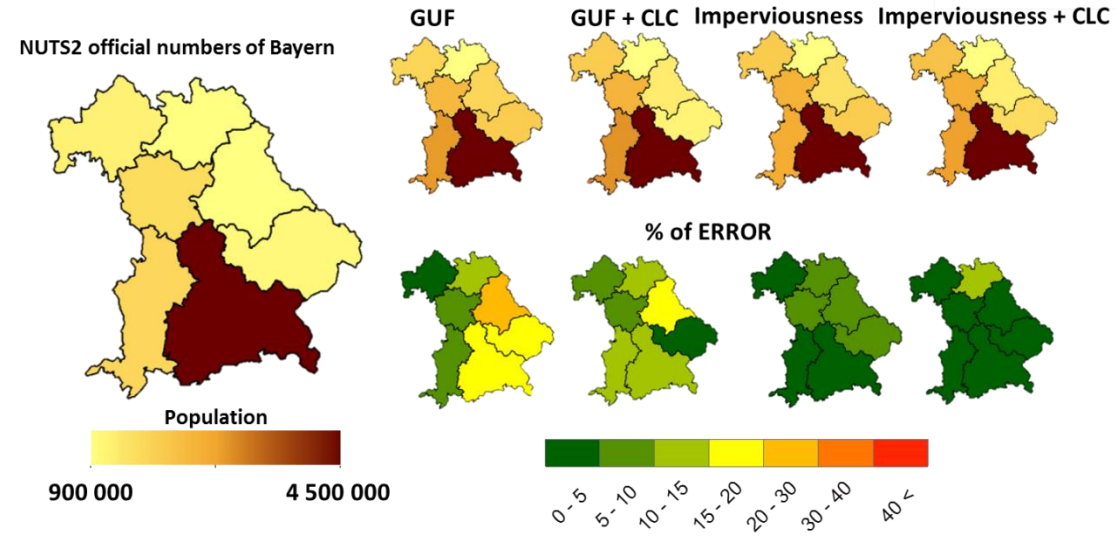
4.4 Evaluation

Evaluation is very important part of any research work. To get an idea how is accurate developed approach, following steps for evaluation were involved:

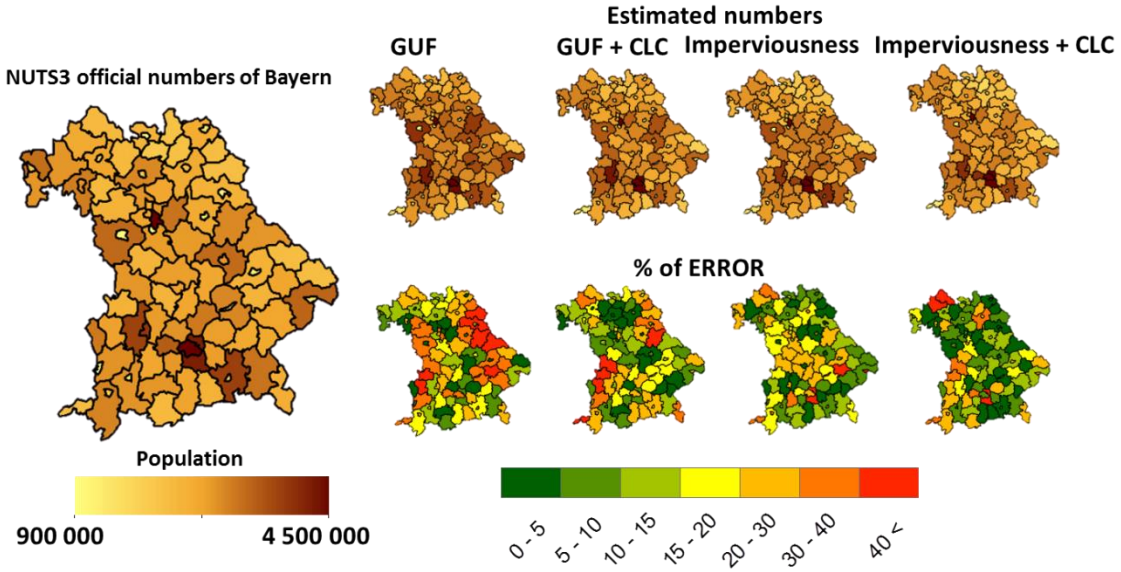
- Calculation of grids (for all possible input data calculations) for Bavaria using census numbers for NUTS 1, NUTS 2 and NUTS3 levels
- Aggregation of population values of NUTS 1 (Federal State) and NUTS 2 (Region) – based calculations to NUTS 3 (District) level (zonal statistics approach)
- Comparison of derived results with official NUTS 3 numbers
- Comparison of structural distribution of NUTS 3-based calculation with data from citypopulation.de resource and GEOSTAT 2006 on Munich area

The results of aggregated grids to NUTS2 and NUTS3 levels, generated utilizing NUTS1 and NUTS2 reported NUTS data shown in figure 12. Visual inspection shows that modeled distributions in all scenarios are agreed in the direction to the official numbers, and spatial configurations are similar. Nevertheless the errors are takes place, and distributions differ due to utilized input data.

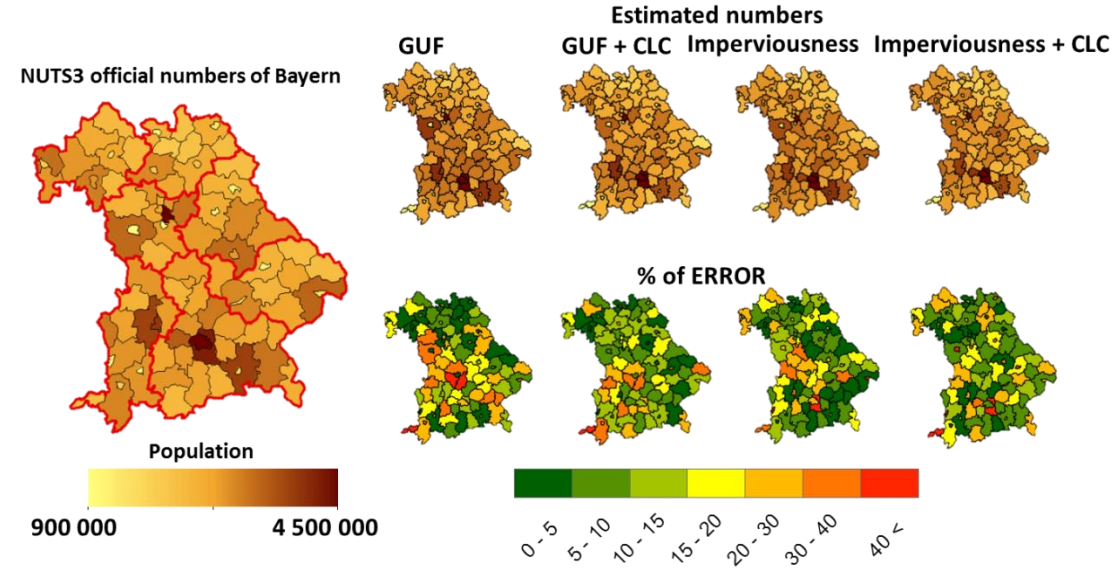
Comparing pure computations utilizing GUF or Imperviousness (sealing) layers without any ancillary data, on NUTS1 level continuous based scenarios demonstrates higher accuracy of distribution than binary based. In both cases the overall accuracy was increased by including land cover / land use classification layer, but at the same time some addition local errors were introduced.



a) NUTS1-based calculations, NUTS2 evaluation



b) NUTS1-based calculations, NUTS3 evaluation



c) NUTS2-based calculations, NUTS3 evaluation

Figure 11 Comparisons of aggregated grids with official numbers

In case of NUTS2 evaluation (See figure 12.a) there is no need to use any statistical indicators, since there are only seven numbers to compare. Degree of errors can be clearly distinguished between the regions. The most weak result demonstrated GUF-based calculation with the maximum error of 27,11%, and the most accurate shows the combination of input data with Imperviousness and CLC layers.

To evaluate the results on district level (NUTS3) it was decided to introduce additional statistical indicators. A variety of methods of measuring the accuracy exists. The choice was made in favor for two most used indicators in the field of measuring distributions - root mean square error (RMSE) and relative total absolute error (RTAE). RMSE, in one hand, applicable for count data (e.g. number of people) and might be interpreted as a value with the same units as measuring values (Eicher C. L. and Brewer C. A., 2001). On another hand, RTAE might be more robust in the matter of skewed distributions (Batista, F. et al., 2013). In authors opinion both indicators can supplement each other in order to assess the accuracy.

Derivation of RMSE and RTAE in this study can be expressed by following equations:

$$\text{RMSE} = \sqrt{\frac{\sum_1^n (P_i - P'_i)^2}{n}} \tag{8}$$

$$\text{RTAE} = \frac{\sum_1^n |P'_i - P_i|}{\sum_1^n P_i} \tag{9}$$

where n is the number of target zones within source zone, P_i is reported number of inhabitants of target zone i, P'_i is estimated number of inhabitants.

Note, RMSE shows the value unit, as measured – number of people. RTAE shows the value between 0 and 2 (RTAE $\in$ [0, 2]), where the value of 0 shows an excellent estimation and the value of 2 would indicate completely wrong modeling.

Table 1 shows the accuracy indicators for NUTS1-based (with only one number of residents for entire State) calculations, evaluated by comparison with NUTS3-reported numbers. Analyzing the results, it is clear that the most weak was GUF-based computation and the most accurate is combination of layers Imperviousness + CLC. The RMSE and RTAE are agreeing in their trends.

Statistics, demonstrated in table 2 for NUTS2-based (Regional level) calculations, evaluated by NUTS3 numbers. It shows significant improvements, which is unsurprisingly, since initially was

used more detailed information. Also here can be distinguished principal difference from NUTS1-based computations. If in the first case RMSE and RTAE everywhere were agreeing, in this case GUF+CLC combination shows better performance than Imperviousness+CLC in RTAE parameter and worse performance in RMSE indicator. Here additional parameters (minimum. maximum and mean errors) and visual inspection (figure 12) are helpful do identify, that the best performance gives Imperviousness+CLC combination, but difference from GUF+CLC is very slight in the matter of NUTS3 scale.

Table 1 Statistics for NUTS1-based calculations

Indicator / Input datasets	MIN Error (%)	MAX Error (%)	MEAN Error (%)	RMSE	RTAE
GUF	0,46	62,98	20,68	65560,91	0,2274
Imperviousness	0,02	52,99	15,95	38506,32	0,1661
GUF + CLC	0,05	57,98	15,98	44192,96	0,1695
Imperviousness + CLC	0,43	47,63	13,88	27423,96	0,1339

Table 2 Statistics for NUTS2-based calculations

Indicator / Input datasets	MIN Error (%)	MAX Error (%)	MEAN Error (%)	RMSE	RTAE
GUF	0,54	52,28	15,28	46378,57	0,1657
Imperviousness	0,32	57,06	14,52	38216,38	0,1466
GUF + CLC	0,38	50,46	13,32	28906,43	0,1326
Imperviousness + CLC	0,02	45,49	14,65	27209,18	0,1350

The second evaluation approach was to aggregate resulted grids to the cell size of 1km in order to compare it with Population Grid GEOSTAT 2006.

On figure 12 demonstrated visual comparisons of derived grids with GEOSTAT product. It can be seen, that results are very similar to GEOSTAT. More similarity demonstrates GUF+CLC and almost identical Imperviousness+CLC combinations. Anyway, GUF-derived grids should not be compared too serious, since here were utilized GUF and census data for the reference year of 2012 and CLC for 2006, when the GEOSTAT product completely based on the dataset of year 2006 (census, soil sealing and CLC). Hence, such assumption, that the real situation on the ground was changed in the period of 2006-2012 can take place.

On another hand, it can be seen from the illustration, that combination of Imperviousness+CLC gives almost identical result to GEOSTAT. This might be explained by similarity of input datasets for the same reference year.

Upper Bavaria and Munich area

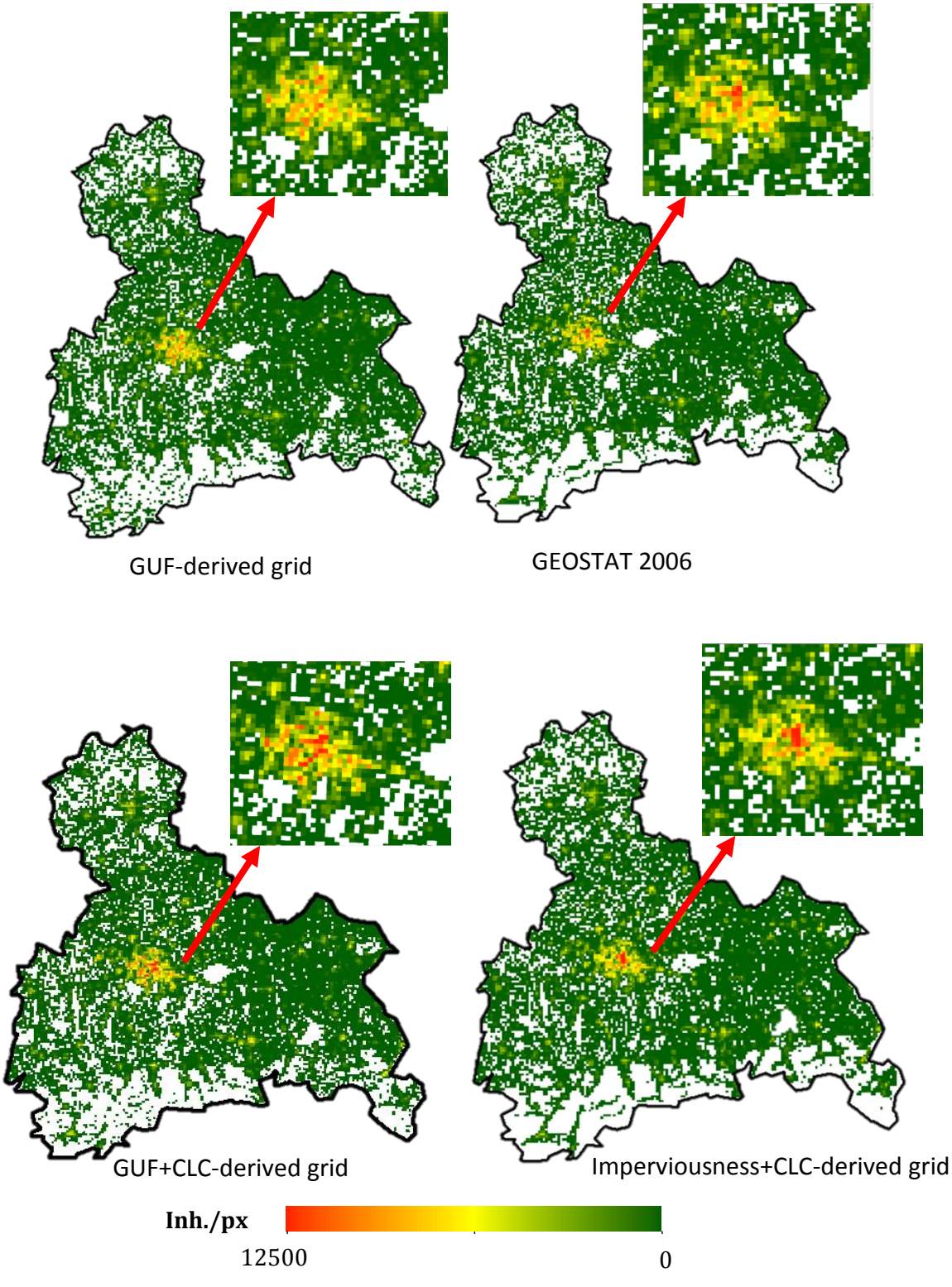


Figure 12 Comparison of derived grids with GEOSTAT

The last evaluation attempt was to compare internal settlement structure with high resolution reference data. Unfortunately, such data was not available in raster or vector format for this research. The only source was found is citypopulation.de portal. This resource provides the average population densities for certain cities districts. The population densities presented in units of inhabitants per square kilometer. It was possible to convert the scale to the pixel size, where 15 000 inh./km^2→ 6 inh./pixel with the pixel size of 20 m^2. Then, the same color scheme was applied to derived grids, shown in figure 13.

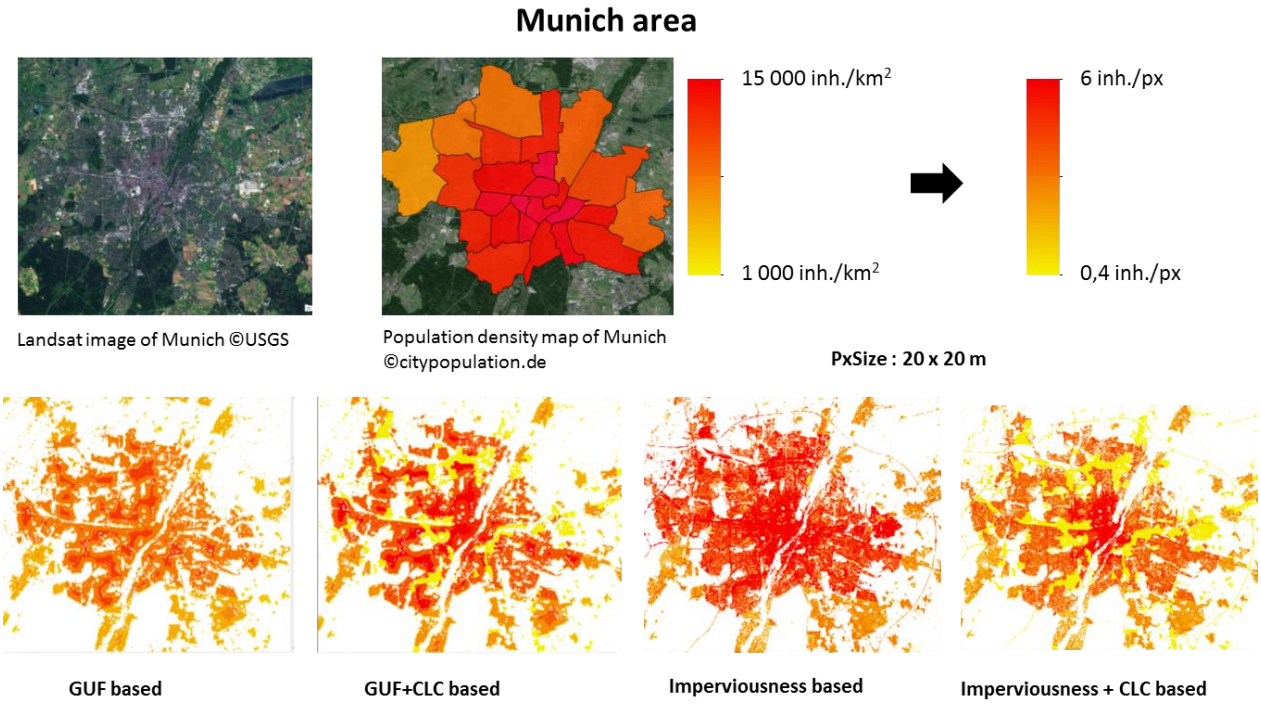


Figure 13 Comparison of derived grids with Citypopulation.de

The visual comparison shows that derived grids has more or less similar distribution shapes. More similarity can be seen in GUF+CLC and Imperviousness+CLC scenarios, when computations without land cover classification shows less realistic distributions. Again, as in case of comparisons with GEOSTAT 2006, this is only visual inspection. Data from citypopulation.de should not be considered as a reference data in this project, due to completely different representation format.

4.5 Results and discussion

The result of this master thesis work is approach of deriving population grids. Developed algorithm based on existing dasymetric mapping methods, which was adopted for utilizing GUF or Imperviousness layers as a base input. Additional ancillary information, such like LC/LU classification might be used as well.

Developed methodology, explained in 5.2 subsection, can be applied independently on specific software. Nevertheless, in order to automate the process, customizable python script was developed and introduced as an ArcGIS tool with user friendly interface. Term customizable in this matter means the possibility to adjust modeling parameters, such like LC/LU coefficients, position and area weighing functions for any specific territory.

Examples of the output grids, utilizing such input combinations, which shows the most weak and the most accurate results demonstrated in figure 14.

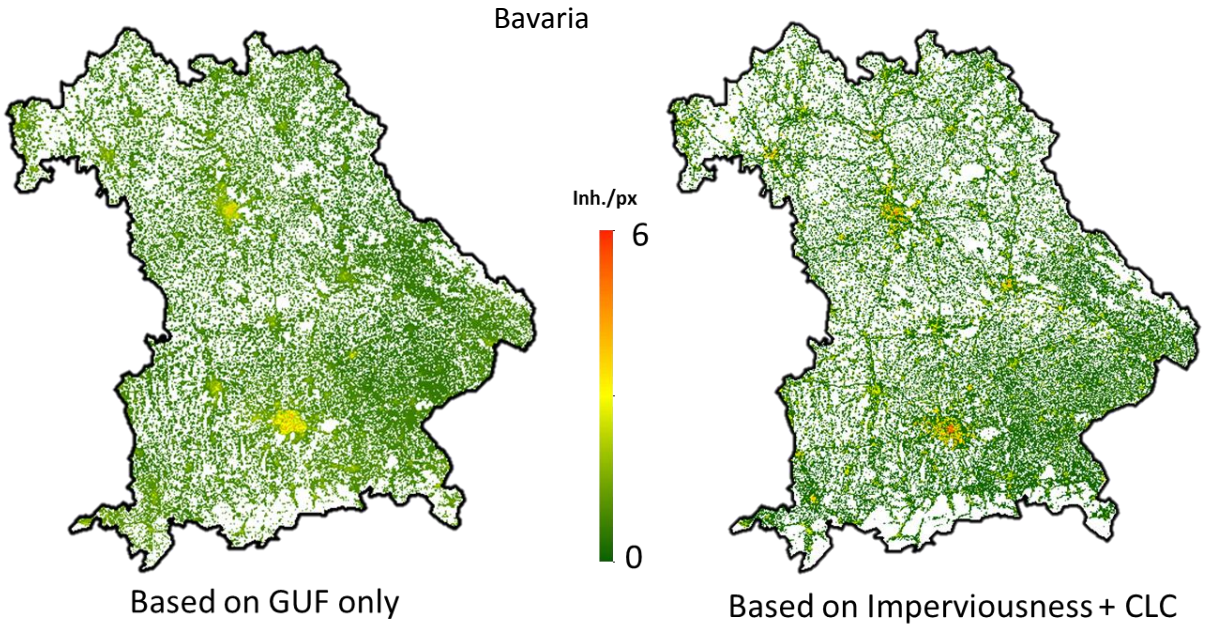


Figure 14 Example of derived population grids

Both examples are looks similar, but differences are exists. According to evaluation results, the most accurate output gives the combination of Imperviousness+CLC. Such combination is not acceptable for the global extend, since CLC dataset is available for the Europe extend only and Imperviousness layer only for Germany (alternative for Europe available as well, e.g. – Fast Track Service from EEA). Derivation of Imperviousness for the global extend is too resource consuming.

The compromise dataset for the global extend can be GUF in combination with global classification layer, e.g. – The GlobCover from European Space Agency (ESA), which represents similar to CLC land cover classification, but aggregates some classes into one. This combination will allow to generate population grid for global extend with comparable quality with GEOSTAT 2006.

Nevertheless, as it was shown in Evaluation subsection, the errors are exists and in some areas the degree of error is relatively high. The probable sources of errors are good subjects for discussion.

Obviously, the main cause of inaccuracy of population grid is the accuracy of input datasets. Figure 15 shows a good example, where Imperviousness-based generation of population grid gives only 4% of error and 41% with GUF-based for the same area.

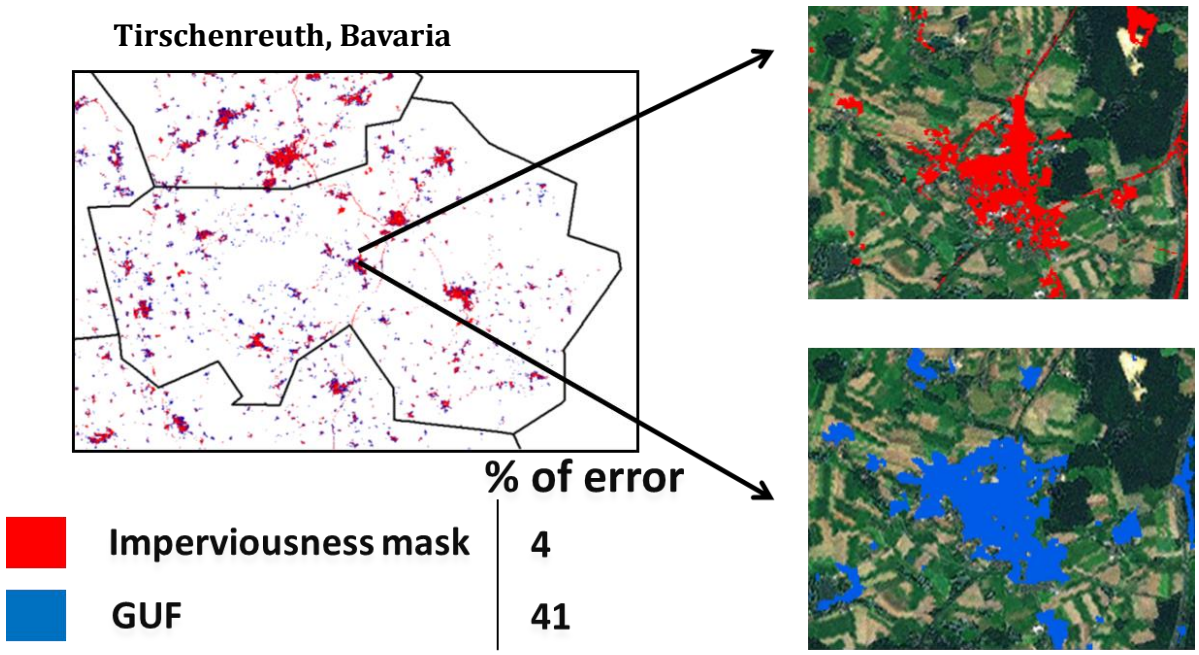


Figure 15 Differences between base input datasets

It can be clearly seen from the picture, that initial Imperviousness layer characterizes the surface better in the matter of built-up areas, when the GUF, evidently, is overestimated.

Elevation might be cause of errors as well. In this project modeling based on planar data and height information is not considered. In the fact, such factors as aspect, slopes and elevation heterogeneity might influence on people distributions.

Land cover / land use classification layers, increasing the overall accuracy, can introduce new local errors as well. Usually, classification is done automatically. It is almost impossible task to refine every location in global extend, so wrong classification can take place. Another reason is the spatial resolution of classification. For example, CLC2006 dataset has minimum mapping unit of 25ha, when utilized Imperviousness layer has pixel size of 20 sq. meter and GUF – 12 sq. meters. It means that some small residential units can be classified as completely different class.

Of course, other errors might arise due to cultural, economic and other local specifics (ORNL, 2012) as well as day time of people locations, which is not considered.

What are the ways to improve the accuracy? This question can be the matter for further project developments.

Depending on the target extend, the model parameters (area weighing coefficients, pixel weighing parameters and etc.) might be adjusted for specific territory. Additional ancillary data can be included as well, e.g. global DEM from DLR soon will be available. Such layers like road networks or socioeconomic layers can benefit for more accurate distributions as well.

Some principal improvements to the algorithm can include, but not limited to:

1) Network analysis. To investigate relationship between neighboring cities within and outside agglomerations. This will allow finding a region magnitude and cost of connections between cities, define priorities for people distributions.
2) Regression models. To find relationship between variables and adjust such parameters like land cover coefficients locally.

All these improvements can be included into one software product, but most likely there is necessity to switch software environment for that. ArcGIS provides all necessary means, but its performance leaves much to be desired. For example, calculation of GUF+CLC-based population grid for Bavaria with the pixel size of 20 m^2 takes approximately 50 minutes (Core 2 Duo 2.6 GHz, 8GB RAM, SSD). Such speed can be explained by specifics of 32 bit platform of ArcGIS, which is not able to use full amount of available RAM.

Unless, ESRI will not switch ArcGIS to x64 platform, the compromise option would be implementation of algorithm as module for SAGA GIS, which supports x64. This will allow reducing calculation time in 2-3 times and even more, depending on the hardware.

The uncompromised solution would be development of standalone application using available open source libraries, for example – GDAL.

5 Conclusion and Outlook

The population distributions are usually represented in aggregated form with the reference to administrative units – choropleth map. Such information might be too coarse for certain applications of spatial planning.

The goal of this work was to develop an approach of deriving raster layers representing distribution of human population based on the novel GUF settlement mask. Resulting grids are able to characterize demographic distributions within zones of choropleth maps in detail. Derived raster layers could be helpful in many GIS applications, where it is necessary to deal with counting of people within specified areas, for example – disaster management.

As main methodology, the dasymetric mapping technics was used. Dasymetric mapping is a technique that redistributes population data from coarse aggregated form (choropleth map zones) into the more detailed form that can describe internal zone structures, using ancillary information. In this project, mentioned approach was adopted to use the specified DLR inputs as main ancillary information for disaggregation.

The calculation includes a set of GIS and image processing operations for analyzing multiple datasets. Developed approach can be performed in most of modern GIS software, but for this study ESRI ArcGIS was utilized.

In order to automate calculation routines the software was developed and implemented as a tool for ArcGIS with guided user interface. This software performs entire procedure without any user interruption. If the calculation of population grid is needed as a part of more complex task, developed module can be called from ArcGIS modeler or Python window as well.

The accuracy assessment shows that resulting grids are agree with official reference data and comparable to existing products, for example GEOSTAT 2006.

Nevertheless the errors are takes place and ranges from 0,02% up to 57%, depending on the scale and combination of input data. The accuracy can be increased by including additional ancillary information and improvements of the algorithms.

The software performance tests show good enough computation speed, but it can be much less time consuming by switching to a 64 bit platform.

References

Batista, F. et al. (2013): A high-resolution population grid map for Europe. In: Journal of Maps, Vol. 9, Issue 1, pp. 16-28.

Eicher, C. L. and Brewer, C.A. (2001): Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. In: Cartography and Geographic Information Science, Vol. 28, Issue 2, pp. 125-138.

Esch, T. et al. (2009): Large-area Assessment of Impervious Surface based on integrated analysis of Single-date Landsat-7 Images and Geospatial vector Data. – In: Remote Sensing of Environment, Vol. 113, issue 8, pp. 1678 - 1690.

Esch, T. et al. (2012): TanDEM-X mission – new perspectives for the inventory and monitoring of global settlement patterns. In: Journal of Applied Remote Sensing, Vol. 6, Issue 1, pp. 1 – 21.

Eurostat (2011): Population grids
http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Population_grids
[accessed 10.10.2013]

Klein, D. et al. (2009): Assessment of urban extend and impervious of Cape Town using TerraSAR-X and Landsat images. In: Proceedings of IGARSS 2009 conference, Cape Town, South Africa.

Langford, M. (2006): Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. In: Computers, Environment and Urban Systems, Vol. 30, pp. 161-180.

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

Langford, M. et al. (2008): Urban population distribution models and service accessibility estimation. In: Computers, Environment and Urban Systems, Vol. 32, pp. 66-80.

Maantay, J. A. et al. (2007): Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). In: Cartography and Geographic Information Science, Vol. 34, Issue 2, pp. 77-102.

Mennis, J. (2003): Generating Surface Models of Population Using Dasymetric Mapping. In: The Professional Geographer, Vol. 55, Issue 1, pp. 31–42.

Mennis, J. and Hultgren, T. (2006): Intelligent dasymetric mapping and its application to areal interpolation. In: Cartography and Geographic Information Science, Vol. 33, Issue 3, pp. 179-194.

M. -D. Su et al. (2010): Multi-layer multi-class dasymetric mapping to estimate population distribution. In: Science of the Total Environment, Vol. 408, pp. 4807-4816.

ORNL (Oak Ridge National Laboratory) (2012): Landscan http://web.ornl.gov/sci/landscan/ [accessed 13.10.2013]

Petrov, A. (2012): One Hundred Years of Dasymetric Mapping: Back to the Origin. In: The Cartographic Journal, Vol. 49 Issue 3, pp. 256-264.

Sleeter, R. (2004): Dasymetric mapping techniques for the San Francisco Bay region, California. In: Urban and Regional Information Systems Association, Annual Conference, Proceedings, Reno, Nev., November 7–10.

United Nations (2011): World Urbanization Prospects, the 2011 Revision http://www.un.org/en/development/desa/population/theme/urbanization/index.shtml [accessed 26.02.2014]

Wright, J. K. (1936): A method of mapping densities of population. In: The Geographical Review, Vol. 26, Issue 1, pp. 103-110.

Annex A: Python script of Population Distribution Modeler

```
import arcpy
import os
from arcpy import env
from arcpy.sa import *
import numpy
from numpy import *
arcpy.env.overwriteOutput = True
arcpy.CheckOutExtension("Spatial")
input_raster = arcpy.arcpy.GetParameterAsText(0)
binary = arcpy.arcpy.GetParameterAsText(1)
stat_vector = arcpy.arcpy.GetParameterAsText(2)
stat_field = arcpy.arcpy.GetParameterAsText(3)
classified = arcpy.arcpy.GetParameterAsText(4)
LCLUC = arcpy.arcpy.GetParameterAsText(5)
LCLUC_exp = arcpy.arcpy.GetParameterAsText(6)
arcpy.env.workspace = arcpy.arcpy.GetParameterAsText(7)
output = arcpy.arcpy.GetParameterAsText(8)


exp= LCLUC_exp
items = exp.split(" ")
lcc={}
for item in items:
   key,value = item.split(':')
   lcc[key] = value
print lcc


description = arcpy.Describe(input_raster)
cellSize = round(description.children[0].meanCellHeight, 0)
arcpy.env.cellSize = cellSize
arcpy.env.extend = input_raster
arcpy.env.snapRaster = input_raster
arcpy.env.mask = input_raster


def lcc_coef(LCLUC):
   i=0
   for cl in lcc.keys():
      arcpy.AddMessage("Extraxtion of "+cl+" class...")
      if (i == 0):
         lcc_raster=arcpy.sa.Con(Raster(LCLUC) == int(cl), int(lcc[cl]), 0)
      else:
         lcc_raster+=arcpy.sa.Con(Raster(LCLUC) == int(cl), int(lcc[cl]), 0)
      i+=1
   return lcc_raster
arcpy.AddMessage("Processing input raster...")

if (binary == "true"):
   PIM = arcpy.sa.Con(Raster(input_raster) > 0, 1)
   BuiltBin_int = PIM
else:
   PIM = arcpy.sa.Con(Raster(input_raster) > 0, Raster(input_raster))
   BuiltBin_int = arcpy.sa.Con(PIM > 1, 1)
```

```
if cellSize<80:
   cellSizeFactor = round(100/cellSize, 0)
   BuiltBin = arcpy.sa.Shrink(arcpy.sa.MajorityFilter(arcpy.sa.Aggregate(BuiltBin_int, cellSizeFactor, "MAXIMUM", "TRUNCATE", "DATA"), "EIGHT", "HALF"), 2, [1])
else:
   BuiltBin = BuiltBin_int


if (classified == "true"):
   PIM = arcpy.sa.Times(PIM, lcc_coef(LCLUC))


arcpy.AddMessage("Calculating interim binary...")
Built_BIN_interim = arcpy.env.workspace+"/Built_BIN_interim"
arcpy.AddMessage("Extracting features...")
arcpy.RasterToPolygon_conversion(BuiltBin, Built_BIN_interim, "NO_SIMPLIFY", "VALUE")
arcpy.AddField_management(Built_BIN_interim, "ArCoef", "DOUBLE", 5, "", "", "ArCoef", "NULLABLE", "REQUIRED")
Built_BIN_interim_areas = arcpy.env.workspace+"/Built_BIN_interim_areas"
arcpy.AddMessage("Calculating builtup areas...")
arcpy.CalculateAreas_stats(Built_BIN_interim, Built_BIN_interim_areas)
BB = arcpy.env.workspace+"/BB"
arcpy.AddMessage("Calculating statistics...")
arcpy.SpatialJoin_analysis(Built_BIN_interim_areas, stat_vector, BB)
BB_stats = BB+"_stats"
arcpy.Statistics_analysis(BB, BB_stats, [["F_AREA", "SUM"], ["F_AREA", "MIN"], ["F_AREA", "MAX"]], "NUTS_ID")
arcpy.AddMessage("Processing statistical data...")
arcpy.JoinField_management (BB, "NUTS_ID", BB_stats, "NUTS_ID", ["SUM_F_AREA", "MIN_F_AREA", "MAX_F_AREA"])


def pos_weighing(Built_BIN_interim, cellSize):
   Built_BIN_lines = arcpy.env.workspace+"/Built_BIN_lines"
   arcpy.PolygonToLine_management(Built_BIN_interim, Built_BIN_lines)
   outEucDistance = arcpy.sa.EucDistance(Built_BIN_lines, "", cellSize, "")
   outExtractByMask = arcpy.sa.ExtractByMask(outEucDistance, Built_BIN_interim)
   pos_coef = arcpy.sa.Con(outEucDistance > 1500, 4, arcpy.sa.Con(outEucDistance > 500, 3, arcpy.sa.Con(outEucDistance > 100, 2, 1.5)))
   return pos_coef


arcpy.AddMessage("Calculating Spatial weights...")
rows = arcpy.UpdateCursor(BB)


def areainc(area, SumArea, MinArea, MaxArea):
 ymin=1
 ymax=1.65
 y=ymin+(area-MinArea)/(MaxArea-MinArea)*(ymax-ymin)
 return y


for row in rows:
   row.setValue("ArCoef", areainc(row.getValue("F_AREA"), row.getValue("SUM_F_AREA"), row.getValue("MIN_F_AREA"), row.getValue("MAX_F_AREA")))
   rows.updateRow(row)

BB_coefs = arcpy.env.workspace+"/BB_coefs"
arcpy.PolygonToRaster_conversion(BB, "ArCoef", BB_coefs, "CELL_CENTER", "", cellSize)
```

```
PIM2 = arcpy.sa.Times(PIM, BB_coefs)
#PIM2.save(arcpy.env.workspace+'/PIM2_test')


TP = arcpy.env.workspace+"/TP"
arcpy.PolygonToRaster_conversion(stat_vector, stat_field, TP, "CELL_CENTER", "", cellSize)


if (binary == "true"):
   PIM2 = arcpy.sa.Times(PIM2, pos_weighing(Built_BIN_interim, cellSize))
SIM_interim = arcpy.env.workspace+"/SIM_interim"
arcpy.AddMessage("Rasterizing features...")
arcpy.FeatureToRaster_conversion(Built_BIN_interim, "ID", SIM_interim, cellSize)
arcpy.AddMessage("Calculating attributive information...")
arcpy.BuildRasterAttributeTable_management(SIM_interim, "Overwrite")
arcpy.AddMessage("Calculating settlement zonal statistics...")
SIM = arcpy.sa.ZonalStatistics(SIM_interim, "VALUE", PIM2, "SUM", "DATA")
arcpy.AddMessage("Calculating total area zonal statistics...")
TIM = arcpy.sa.ZonalStatistics(stat_vector, "NUTS_ID", PIM2, "SUM", "DATA")
arcpy.AddMessage("Calculating settlement population...")
SP = arcpy.sa.Divide(arcpy.sa.Times(TP, SIM), TIM)
#SP.save(arcpy.env.workspace+'/SP_test')
arcpy.AddMessage("Generating final population grid...")
SPDG = arcpy.sa.Divide(arcpy.sa.Times(SP, PIM2), SIM)

SPDG.save(output)
```