Dasymetric estimation of population: A case study of the city of Enschede, The Netherlands

Mohammad Sharif
November, 2010

Dasymetric estimation of population: A case study of the city of Enschede, The Netherlands

by

Mohammad Sharif

Thesis submitted to the Faculty of Geo-information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: *Geoinformatics*

Thesis Assessment Board

Chair:	Dr. R.A. de By
External examiner:	Dr. Ir. R.J.A. van Lammeren
Supervisor:	Dr. R. Zurita-Milla
Second supervisor:	Ms. Ir. P.W.M. Augustijn

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION OF THE UNIVERSITY OF TWENTE
ENSCHEDE, THE NETHERLANDS

Disclaimer

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

To my mother, father and dear younger brothers

Abstract

This research is aimed at disaggregating census population data into the finest possible scale. Census data are usually issued aggregated, because of preventing the disclosure of information of people and also controlling the volume of data. In order to utilize this type of data for specific objectives such as disaster management, urban planning, and disease spread etc., aggregated census data need to become highly detailed and spatially localized by disaggregating them into the smallest unit of measurement.

Dasymetric methods, which are a type of areal interpolation techniques for transforming data from one set of spatial unit to another, are employed in this research. They use ancillary data in the process of disaggregating census data which help the internal distribution of populated regions to be inferred. Specifically, the chosen methods are Binary and 3-class dasymetric methods, where the former is quoted as a simple and easy to implement in GIS, while the later is severally cited as a robust method for disaggregating census tracts due to the use of a variety of ancillary classes (e.g. non-urban, low-density and high-density). These methods have been applied to different large study areas with various population density values. It is desired to examine the performance of dasymetric methods for a small and dense region; hence, the city of Enschede with 143 (km^2) area and average population of 1091 ($^{Person}/_{km^2}$) is chosen. The finest administrative units to be considered as the source and target zones are realized as district and neighbourhood levels, respectively.

The above methods are implemented on the census population data by making use of various types of ancillary datasets: first, the buildings' footprint land cover maps are utilized in order to assess the accuracy of 2D ancillary information, finding the right scale of these data and to justify previous studies regarding the performance of binary and 3-class dasymetric methods; second, the buildings' floor area map is used due to the evaluation of added value(s) of 3D data in the course of population estimation and finally, the postcode 6-digit information as an alternative to the area factor in current dasymetric methods.

The assessments of statistical results show the better performance of binary dasymetric method to the 3-class dasymetric method (23% vs. 45% error value) which contradicts previous studies regarding the high usability of 3-class variables. This matter can be referred to the assumption for classifying the study region to three class variables. Moreover, the evaluation of the ancillary datasets indicates the competency of TOP10 dataset in comparison to the other congener areal datasets. By utilizing 3-Dimentional data, which are the buildings heights and their corresponding floor areas, the performance of dasymetric methods have improved where the binary method generated 18% error value and 3-class method produces 23% uncertainty for disaggregating population. Despite previous efforts for finding a proper type of ancillary data, a new version of ancillary dataset –postcode 6-digit units– outweighs (with 17% coefficient variation) those datasets which make use of areal units in the course of population disintegration. Hence, as an answer to the main objective of this research, the census population data from a small study area with high population density can be disaggregated from a large region to various sub-regions by implementing binary dasymetric and utilizing either 3D TOP10 areal map or postcode 6-digit units as ancillary data with an approximate coefficient variation of 17% per person.

Keywords: Areal interpolation, dasymetric method, disaggregation census population data

Acknowledgements

All the praise and adoration to almighty God, the beneficent, the merciful as the everlasting leader and teacher in my entire life; without his divine grace, the prosperous achievement of this research was impossible.

My utmost grateful and appreciation goes to my supervisors, Dr. Raul Zurita-Milla and Ms. Ir. P.W.M. Ellen-Wien Augustijn for their precious guidance, critical advice and beneficial suggestions throughout the research period. All your encouragements helped me to focus on all parts of my research which I thank you.

I would like to extend my appreciation to Dr. Ali Sharifi, Dr. Richard Sliuzas and Ms. Monika Kuffer for sharing their invaluable knowledge and clarifying every question I posed anytime.

My sincere thanks go to all lecturers of K.N.Toosi University of technology, the program director of JKIP, Dr. Behzad Vosooghi, GFM coordinator Mr. Gerrit C. Huurneman and to all administrative staff of ITC for their support and effort during my M.Sc. period.

Thanks to the municipality of Enschede, I&O research organization and TNT post office for their consultation and data providing.

I thank my classmates, for all these eighteen month interval togetherness which yield to durable friendship, fellow-feeling and unity. Also, my special thanks to the friends whom voluntarily participate in my research, Babak Naimi, Davood Shojaei and Abdollah Narouie.

My heartfelt appreciation goes to my honourable parents and lovely younger brothers for their warm affection, encouragement and support; without them all these would not have been achievable.

Mohammad

Table of contents

List of figures

List of tables

1. Introduction

1.1. Overview

The world's growing population has produced great impacts on global resources, the environment and urban development. Timely and accurate population estimation and the spatial distribution of population and its dynamics become considerably significant in understanding the effects of population increase on social, economic, and environmental problems (Sutton et al., 1997). Moreover, population information at different levels, such as national, regional, and local, is very important for many purposes, such as urban planning, resource management and service allocation (Weng, 2009). In order to make suitable decisions in those matters, it is necessary to map the population distribution in an applicable and interpretable form.

Population mapping, in general, has two purposes: firstly, for cartographically depiction of population density through the area of interest, and secondly, to derive a quantitative estimation of population density for use in subsequent spatial analytical modelling tasks (Bielecka, 2005). A cartographic depiction of population has the form of a choropleth map. Dorling (1993) noted that choropleth maps of population by administrative areal unit give the notion that population is distributed homogeneously through the areal unit, even when the region in reality is uninhabited (see Figure 1.1a). This kind of mapping is very simple, but despite its simplicity, choropleth maps have limited utility for detailed spatial analysis of population data, especially where human populations are concentrated in relatively small numbers of villages, towns and cities (Bielecka, 2005). Moreover, it suffers from Modified Area Unit Problem (MAUP) where the aggregated data represent an arbitrary area which has no relation with disaggregated data (Li et al., 2007).

One approach for improving this shortcoming is to change the administrative units into smaller units through the process which is called dasymetric mapping. Eicher and Brewer (2001) stated that "Dasymetric mapping depicts quantitative areal data using boundaries that divide the mapped area into zones of relative homogeneity with the purpose of best portraying the underlying statistical surface." (see Figure 1.1b). This form of mapping has been quoted as an intelligent method to choropleth mapping regarding improving the area homogeneity (Sleeter, 2004). Based on Bielecka's (2005) definition, "Dasymetric mapping is a form of areal interpolation that uses ancillary data to transform population data from one set of spatial units to another".

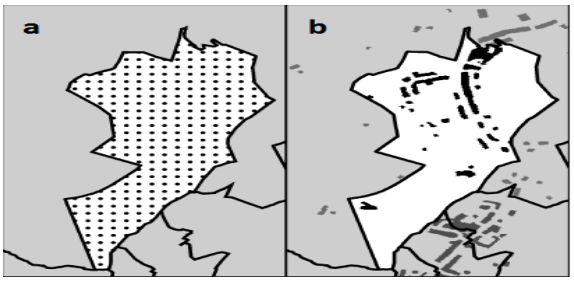


Figure 1.1. Two alternative representations of population: (a) a choropleth map-evenly distributed throughout the census zone, (b) a dasymetrically distributed map (Langford and Higgs, 2006).

1.2. Motivation and Problem statement

The aggregated census population data are using in a wide area of administrative projects. In order to utilize these types of data for specific objectives and intended goals, they need to become highly detailed and localized by disaggregating them into the smallest unit of measurement. For instance, in the case of disaster management, by knowing the population density of residential blocks/buildings, it is easier to relief people in natural disasters (i.e. earthquake, flood, storm and etc.) and plan for post-disaster reconstruction. In the context of urban planning, for establishing public constructions such as public transportation stations, schools, medical centres, etc., it is necessary to know the detailed distribution of population in the local area. Furthermore, in the case of spreading of diseases, the congestion of population in an area (neighbourhood, block of buildings or personal houses) shows the susceptibility of that part to epidemical diseases. As a matter of fact, it is important to know how to disaggregate these types of data to achieve the highest accuracy and less cost (time, complexity, effort etc). This disaggregation process would be performable by utilizing different dasymetric techniques.

Although the dasymetric mapping technique has been utilized since several years ago (Wright, 1936), it still suffers from an appropriate disaggregation method and suitable ancillary data. Two difficulties have been found regarding ancillary data. First one relates to the proper scale of raster/vector maps which are using in the process of population estimation. Various types of land use/land cover and topographic maps are produced in small/large scales. It should be found out, what type of map with which scale and what accuracy is suitable for implementing. Maybe a 1:50,000 or TOP10 map result the same efficiency as 1:500 scale maps. The second difficulty that relates to ancillary data has been finding the type of data that corresponds well with population distribution. Land use and/or land cover data are cited as the most frequent-used ancillary data for disaggregating population (i.e. Mennis, 2003; Holt et al., 2004). Other ancillary information includes remote sensing image with various spatial resolution, aerial photogrammetry images and Light Detection and Ranging (LIDAR) data (Harvey, 2002; Wu et al., 2006; Wu et al., 2008). Another difficulty which dasymetric mapping techniques deal with is finding the best method to be applied in the course of population disaggregation.

Frequent population estimation methods are usually designed for the scale of neighbourhood and upper than it (i.e. district, city and etc.). This is due to the limitation of data accuracy and/or errors associated with population estimation in small areas (Wu et al., 2008). The performance accuracy of these methods is largely related to the spatial distribution of spatial heterogeneity of population within the districts (Liao et al., 2010). New efforts can be done for modifying the previous methods or even designing a new approach by considering associated variables (i.e. ancillary class variables, residential density variables and etc.) for population estimation distribution in scales less than neighbourhood. The accuracy of such models can be compared by current population estimation models to evaluate the applicability of them. In addition, most of the studies in population estimation utilize a variety of 2-Dimensional land use/land cover maps with different scales and aerial/satellite images with different spatial resolution as ancillary data beside the dasymetric mapping techniques. Rarely in literature, can the role of 3-Dimensional ancillary data in population estimation accuracy be seen (Wu et al., 2008). The effects of such types of data on the final accuracy, compatibility and reliability of the result can also be evaluated and compared with validation data.

1.3. Research objectives

The main objective of this research is to downscale the census population data to the finest possible scale. This main objective can be divided into some categories of sub-objectives:

- Evaluation of existing dasymetric mapping methods (based on accuracy, complexity, efficiency and etc.)
- Assessing the "right scale" of the ancillary datasets
- Assessing the added value of 3D ancillary data
- Identifying the best way to evaluate the results (statistics, datasets, etc.)

1.4. Research questions

1. What are the existing population disaggregation methods?
    - Based on which criteria are they evaluated and chosen?
    - Which method acts properly (has better accuracy, less complexity, better efficiency and etc.) for the selected study area?
2. In what scales the ancillary datasets are available?
    - How are they assessed and chosen?
3. What are the extra values of 3D ancillary data? How are they going to be assessed?
4. What is a proper way for validating the results?

1.5. Methodology

The methodology section is divided to four steps (see Figure 1.2). It starts by reviewing similar works which have addressed the above problem from the literature; also, finding and assessing the current disaggregation methods and the ancillary data which have been used in the course of population disaggregation. In parallel, the initial and the required dataset for this research will be gathered. In the following, the implementation part will be begun in two phases. On the one hand, the disaggregation methods which have found as appropriate ones to the aim of this research will be applied on the selected 2D ancillary data. The output of this operation will be a disaggregated population map. On the other hand, the current methods will be modified by implementing 3D ancillary data. Again, a new disaggregated population map will be generated based on 3D data. The accuracy of the produced maps will be evaluated via a series of statistical analysis approaches. Finally, based on the calculated result, the best method for disaggregating census population data by incorporating the finest ancillary data will be introduced and the downscaled population map will be generated.

1.6. Innovation aimed at

Several studies have been done in the context of disaggregating population data by using current disaggregation methods and considering 2D ancillary data. This study aims at finding out the extra value of 3D ancillary data in population disaggregation results. Furthermore, a new type of ancillary data –postcode 6-digit– will be evaluated as a replacement to the existing dataset for contributing the process of population disaggregation.

1.7. Thesis outline

This thesis is organized in five chapters where: chapter one introduces the research problems and describes the research objectives and question with respect to the problem definitions. Chapter two is devoted to the review of existing population disaggregation methods including their definitions, usages in different stages and the advantages, limitation and the pitfall of the each method. Chapter three represents the study area, the corresponding dataset and an introduction to the applied methods consisting of the dasymetric methods and the statistical analysis approaches. In chapter four the results

which were obtained from chapter three will be discussed. Finally, chapter five concludes the research content and proposes some recommendations and possible studies for future.

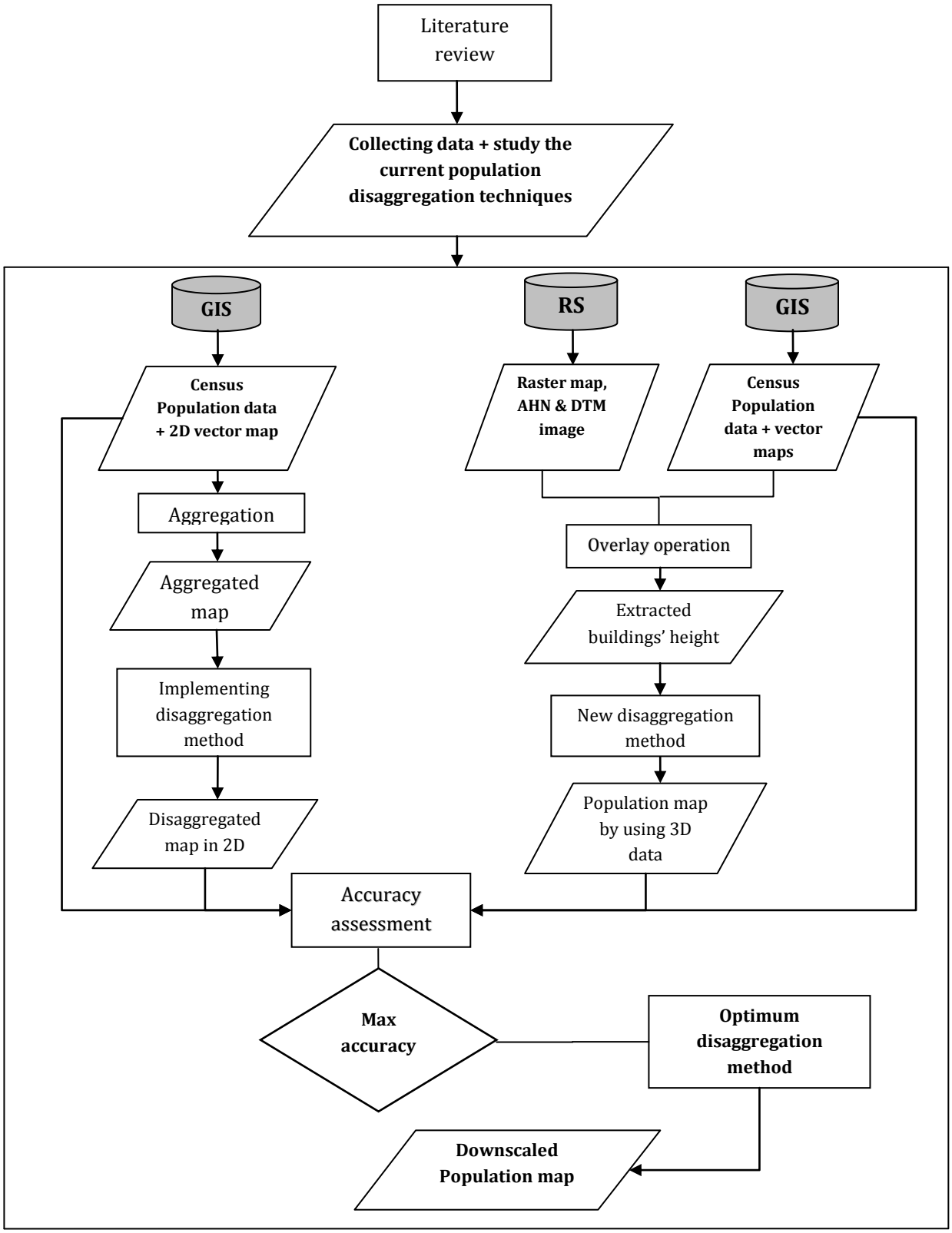


Figure 1.2. The processes of methodology

2. Literature Review

2.1. Overview

Census population are a type of socio-economic data which are usually published as aggregated due to the confidential protection and volume preservation of data. Some policies lead to gain detail internal variation of these types of data. Hence, there would be a need to break up the uniform source data to the smaller administrative units; where the concept is called spatial disaggregation. There is no definitive way for disaggregating population data due to the complexity of mismatch between boundaries in the larger zone and smaller sub-zones and also the heterogeneous density inside them. In this way, several disaggregation techniques have been introduced to mitigate these difficulties by making use of different underlying assumptions and various types of ancillary data. The primary purpose of this section is to introduce current population disaggregation methods, and to describe the datasets which have utilized as ancillary data corresponding to the methods.

2.2. Interpolation methods

Various types of interpolation methods have been introduced for population disaggregation in geographic information system and remote sensing literature. Based on the technique, required data and intended goals they can be categorized into two parts (Wu et al., 2005): statistical modelling approaches and areal interpolation methods. Table 2.1 is a categorization and summary of the current distribution methods by their corresponding techniques.

Table 2.1. Typology of methods for modelling population distribution (Freire, 2007)

<table>
<tr><th colspan="4">Methods</th><th>Techniques</th></tr>
<tr><td rowspan="4">Areal interpolation</td><td rowspan="3">Without ancillary data</td><td rowspan="2">Point-based</td><td>Exact</td><td>distance-weighting, kriging, spline, finite difference</td></tr>
<tr><td>Approximate</td><td>least squares, least squares fitting with splines, Fourier series models, power-series trend models</td></tr>
<tr><td colspan="2">Area-based</td><td>Areal weighting, Pycnophylactic interpolation</td></tr>
<tr><td rowspan="2">With ancillary data</td><td colspan="2">“Intelligent” interpolation</td><td>Control zones</td></tr>
<tr><td colspan="2">Dasymetric mapping</td><td>Limiting variables, Related variables</td></tr>
<tr><td colspan="4">Statistical modeling</td><td>Correlation with urban extent, land use, dwellings, image pixel, and several physical and socio-economic variables</td></tr>
</table>

Statistical modelling methods first introduced to minimize the limitations of the census population data, such as high cost, decennial interval of enumeration, effort intensity and etc. (Kraus et al., 1974) and further used to infer the relationship between population distribution and other variables for

aiming at estimating population for an area. In the other words, they use socioeconomic variables and theories in urban geography for estimating population density directly (Liu, 2003). The primary aims of statistical methods are to find the not enumerated census population data within the intercensal periods and estimating the population of areas difficult to enumerate, though it can be considered as a census population interpolating method (Wu et al., 2005). Checking the reliability of the census enumeration is another utility of these methods (Clayton and Estes, 1980). Generally, five statistical modelling methods (Table 2.1) which are based on the relationship between population and urban areas, land use, dwelling units, image pixels characteristics and other socioeconomic characteristics have been introduced in literature (Lo, 1986; Liu, 2003; Wu et al., 2005). But, because firstly this research is not going to model the population distribution and second, none of the above approaches use the census data directly for population estimation, they are found not suited to the goal of this research which is transforming census data from a large scale area to small sub-areas.

The term "Areal interpolation" (sometimes referred as cross-area estimation) is given to the process of transforming statistical data from one zonation of an area to another incompatible zonation of the same area (Goodchild and Lam, 1980). These types of methods usually use census population data as the basis and apply an interpolation or disaggregation technique to obtain population surface or refine population distribution (Liu, 2003). In general, two types of spatial units are required for transformation data from one set to another: source zones and target zones respectively (Lam, 1983). Source zones can be defined as a set of areal units for which the true value for enumeration is known, while target zones are a set of areal units for which an interpolated estimate is required (Langford, 2006). In the context of population distribution, areal interpolation techniques use census units as the source zones and apply a disaggregation method to obtain refined population surface (Liu, 2003). Although these techniques have been used for a long time by utilizing different areal units and data, they lack a standard performance methodology in using the method and also ancillary data. The areal interpolation can be categorized into two sets based on whether ancillary data are used which are going to be investigated in next section.

2.2.1. Areal interpolation methods without ancillary data

Areal interpolation methods without using ancillary data are divided into point-based and area-based methods (Lam, 1983).

- **Point-based methods**

In point-based interpolation, population counts are assigned to undetermined-value points based on a certain number of points with known location and values (Freire, 2007). With census data, point locations can represent the population in that area. Many point-based interpolation methods can be found in the literature where Lam (1983) categorized them into exact methods and approximate methods, based on whether they preserve the original sample point values or produce a general surface function, $f(x, y)$. The reason for this classification is concerned whether the interpolation methods keep the original values after the interpolation process done and a surface is created. Table 2.1 mentions the exact and approximate point-based methods. Each of these methods has their own merits and demerits and there is no superiority between them in all applications (Freire, 2007). Choosing the appropriate point-based interpolation method largely depends on the type of data, desired degree of accuracy and the amount of computational effort

afforded. Generally the exact methods are more trustworthy than the approximate ones because of their simplicity, flexibility and reliability (Lam, 1983; Wu et al., 2005).

Point-based areal interpolation methods deal with a few problems (Lam, 1983; Liu, 2003; Wu et al., 2005). First, it is regarding to the control points which are located in the centre of the source zone. The control points are demarked via a window, and the population of source zone is allocated to the grid cells covering this window. In this case, finding the centre of the area to be considered as source zone often generates errors, especially when the source zones are not symmetrical and relatively simple. Second problem of point-based method is transforming the total value of source zone to target zones. Volume preservation is an important task which gives the degree of reliability to the estimated values in target zones. In this context, another type of areal interpolation technique –area-based method– is introduced to cover the shortcomings of point-based methods.

- **Area-based methods**

Area-based interpolation methods in contrast to point-based methods have the ability for volume (total population) preserving. These methods use the source and target zones instead of control points as operation units (Liu, 2003). The simplest method in this category is Simple Areal Weighting which is an overlaying operation based on the geometric properties of the source and target zones (Fisher & Langford, 1995). By superimposing the target zone on source zone, the proportion of each source zone in each target zone will be obtained. The advantages of this method are its simplicity and also no need for ancillary data to guide the interpolation process (Langford, 2006). Furthermore, the functionality which is required is present in the almost GIS software. The major problem with this method is that, it assumes the density of population in the study area homogeneous. This might be generated for some study areas such as rainfall or agricultural but cannot be justified for population. This is the same conceptual problem which is raised when using choropleth map in cartography (Fisher & Langford, 1995). There have been numerous studies that shown the low accuracy of simple areal weighting method in the literature (Langford, 2006; Gregory and Paul, 2005).

2.2.2. Areal interpolation methods with ancillary data

Generally, some of the areal interpolation method use ancillary data in order to improve the population estimation accuracy. The purpose of the ancillary data is to allow the internal structure of population distribution within source zone to be inferred. These types of areal interpolation can be classified in two categories, the regression methods and dasymetric methods; based on the fact that, the regression methods need to generate a statistical best-fit relationship across the source zones while dasymetric models do not (Langford, 2006).

- **Regression method**

Based on the size of the study region (source zone), the regression model can be categorized into global and regional regression models.

Global regression: This method, first introduced by Langford and colleagues (1991) and due to the use of a full set of the study area (source zone), it is called global regression method (Donnay and Unwin, 2001). What discriminates the regression method from statistical methods is, the regression method is based on the assumption that there is a connection between the population count of the source zone and the presented land covers within it. For instance, it might be expected to see a higher population count if the majority of land cover is urban rather than agriculture (Langford, 2006). In other word, this method hypothesis that, the populated source zones may be a combination of various land classes with different density values. For example, different land classifications can be considered as independent variables for determining regression relationship with population counts. A disadvantage of this method is related to the density values which are based on a global context and they are fixed values in each land class throughout the source zone. So, the variation among different census source zones that have the same land class cannot be recognized easily (Li et al., 2007).

Regional regression: Yuan et al. (1997) found out that, there is no necessity to use the global regression source zone density to cover the study area. Hence, the regional regression method is introduced in order to enhance the reliability of the estimated density extracted from the regression method. Then, by utilizing the census counts, the globally estimated density is locally adapted for each land class inside each source zone (Li et al., 2007). Although the regional regression model performs better than the global method, still the spatial inconsistency between population and land cover exists. So, another areal interpolation technique that can better reveal the relationship between population and land cover is desired.

- **Dasymetric methods**

The method was first named *dasymetric* via a Russian cartographer Tian-Shansky, who developed the multi-sheet population density map of European Russia, scale 1:420000, published in the 1920s (Preobrazenski, 1954 in Bielecka, 2005). The first cartographer who made use of dasymetric mapping was Wright (1936) who believed that the word dasymetric concerns to *density measuring* and it can be employed for identifying the areas within zones which have different population densities via utilizing local knowledge (Fisher and Langford, 1995). In this way, the binary disaggregation method is utilized for partitioning the source zone to the small zones of population density by preserving the primary zone. In the past, it was difficult to execute dasymetric mapping; but nowadays by the development of digital data and GIS technology, the dasymetric methods can conveniently be implemented easily through different types of ancillary data. Two basic dasymetric methods which have been quoted in the literature will be introduced in the following paragraphs.

Binary dasymetric method: For population mapping the simplest approach is to divide the study area (source zone) into two sub-areas. One of them is considered as the populated area and the other one as unpopulated region; then the populated region is deemed to source zone (Langford, 2006). The concept of these limiting variable first introduced by Wright (1936), but has been widely used in literature where called *Binary Dasymetric Method* and described formulaically via Fisher and Langford (1996). This method is conceptually simple and relatively easy to perform where previous studies that have shown its robustness for error classification (Eicher & Brewer, 2001). However, this method is unable differentiating between more complex land uses that have a

variety of population concentrations; hence an alternative which have the ability to address the complexity inside the urban areas is desirable.

Three-class dasymetric method: The 3-class dasymetric mapping (Mennis, 2003; Langford, 2006) is cited as a better performance method in contrast to binary dasymetric method and regression model due to utilizing a limited number of ancillary class variables (e.g. high-density residential, low density residential and non-urban) for presenting the residential densities within source zones. This method assumes that the density for each land class is homogeneous inside the source zones where the relative density for each land class is required for implementing it. On the other hand, Mennis (2003) assigned a proportionate density value to each class in order to obtain a certain number of total populations from each source zone. This work is the modification of Eacher and Brewer (2001) dasymetric mapping method that utilized a constant ratio of population for each land use class.

2.3. Ancillary data types

In general, the areal interpolation methods which utilize ancillary data (particularly dasymetric methods) in the process of population estimation produce more reliable result than those which do not use such type of information. It supposes that, the ancillary data contribute the distribution of variables that should be mapped. Then, an example of the various types of ancillary data which have been used in the course of population estimation will be demonstrated.

Topographic maps are utilized in the basis of redistributing population data to the populated and non-populated regions as a primary effort for dasymetric mapping (Wright, 1936). The populated area is subdivided into smaller areas based on the settlement pattern information which is derived from the topographic maps. In continuous, these settlement patterns are weighted as density values, which were used for estimating population in inhabited areas.

Night-time light emission data from satellite imagery and land cover data are employed by Briggs (2007) to generate the population dissemination pattern at land parcel level across the European Union. In this way, regression methods implemented in order to make a connection between census population counts and other aforementioned ancillary data. The efforts resulted to a model which performs satisfactory for modelling the population distribution.

Landsat Thematic Mapper (TM) multispectral imagery and its derived land use map are contributed for generating a raster population surface (Langford et al. 1991).These ancillary data are utilized for creating a set of regression population density models with respect to each land class. The role these models are redistributing census data and producing a one-km-pixel-dimension population raster surface map for U.K census wards.

Ikonos high-resolution satellite imagery examined to see its correlation with census population density (Liu et al., 2006). Specifically, the image texture of these types of images in conjunction with census-block data is explored by adapting a linear regression. The result shown the presence between these dataset, but they varies regarding the type of image texture description methods. It is concluded that, these information have the potential for estimate the population distribution but not residential population forecasting.

The building volume and census housing statistics are used by Wu et al. (2008) for implementing in a deterministic model to estimate population in the sub-block unit. The model with the contribution of detail information such as, average space per housing unit, the housing unit occupancy rate and the average household size succeeded to disaggregate the census population data from building block-level to housing level with 0.15 error variation.

The above studies take advantage of various types of spatial ancillary data. As a general conclusion, they are assumed to be beneficial for mapping the population distribution with the association of areal interpolation methods and consistency to the study areas.

3. Data and Methods

3.1. Study area

The study area is the city of Enschede, which is located in the eastern Netherlands with total area of 143.1 km^2 (2009), the population of 156130 residents (Jan 2009) and the average density of 1091 ($^{person}/_{km^2}$). The city is divided into ten local districts and 70 neighbourhoods; and the residential settlement of Enschede as it can be seen in Figure 3.1, is mostly concentrated in the central districts.

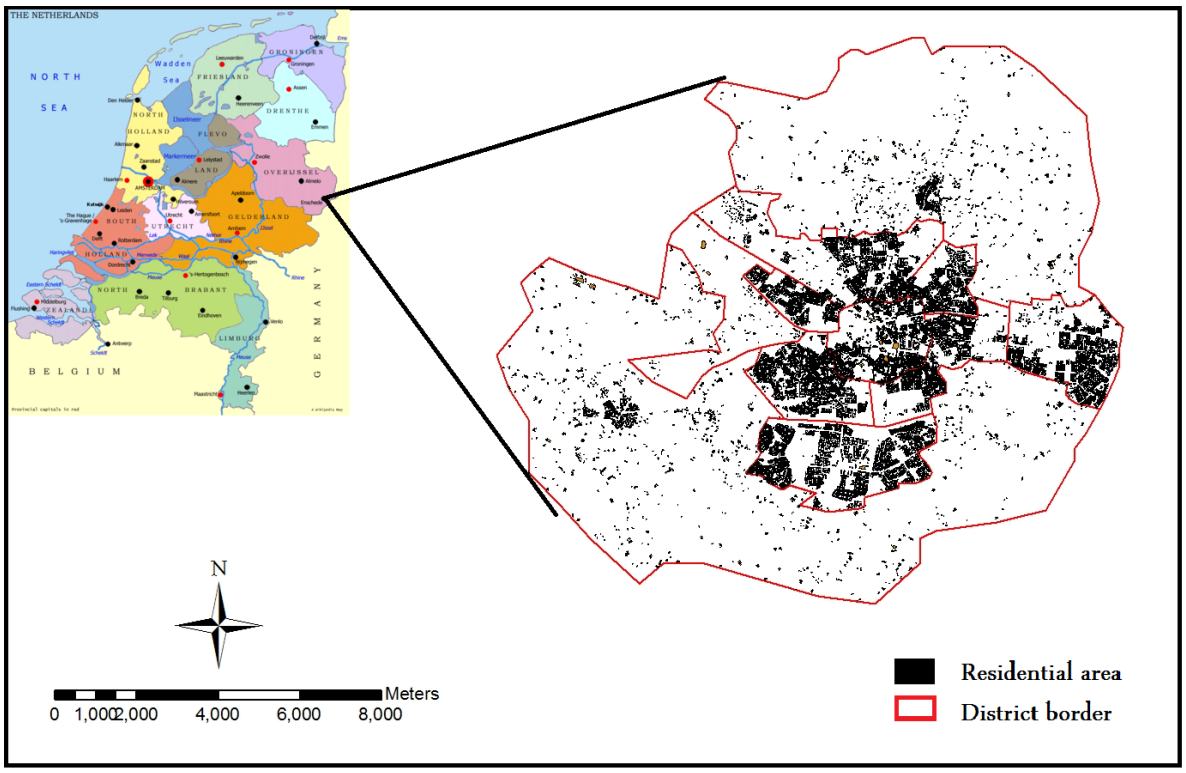


Figure 3.1. An illustration of the study area

The reason for choosing Enschede is that, previous studies for population estimation are mostly concerned in large study regions. For instance, Fisher and Langford (1996) and Langford (2006) employed the county of Leicestershire which covers around 820 km^2 in the centre of England; and Li et al. (2007) used a larger and more complex area namely called South East of Queensland (SEQ) in Australia with approximately 22,800 km^2. Hence, it is desired to evaluate the performance of population disaggregation methods for a small and denser region. A comparison between the urban characteristics of our study area and the previous study regions is shown in Table 3.1. As it is highlighted in the last column, there is a greater density value for Enschede while it has a smaller area in comparison to other locations.

Table 3.1. Population density variation for Enschede in comparison to previous study areas

Study area	Total population	Total area (km^2)	Average density ($^{Person}/_{km^2}$)
Enschede	156,130	**143**	1091
Leicestershire	459,000	**820**	560
SEQ	2,479,000	**22,800**	109

In addition to the previous criterion for selecting Enschede as a study region, it can be mentioned that, the availability and consistency of data are an important factor for disaggregating population. If the required data are not available, or in the case of availability they do not match spatially and temporally, then the disaggregation procedure cannot be adequately run through the dataset. Although Enschede is a small area, it has an updated source of socio-economic source data which are spatially and temporally compatible with each other.

3.2. Data

Census population data: The municipality of Enschede provides population data in different administrative levels such as city, district, neighbourhood, postcode 4-digit and postcode 6-digit units in tabular format. In the Netherlands, the population data are basically based on the registration of residence in the municipality but technically are considered as census data in this research. The enumeration data from the 1st quarter of January 2009 are used for implementing in disaggregation methods as source and target zone values and also utilized for evaluating the performance of disaggregating methods.

Administrative areas: The area map of the city of Enschede is demarcated in district and neighbourhood administrative units and supplied in ESRI shape format. These municipal boundaries are ten districts and 70 neighbourhoods where for each sub-region the name, number, respective code in municipality, number of population from census record database (CBD) per sex, age, number of foreigner and households are labelled (Source: © 2009, Central Statistical Office / Land Registry, Zwolle, 2009).

TOP10NL: Is an object-oriented topographic map at the scale 1:10,000 with 2m accuracy in ESRI shape format. The production date backs to 2008 on the basis of aerial photographs, field survey and dataset. TOP10NL is a cartographic representation of buildings' foot print where each geographical object has its own unique code and is specified further by means of attributes and attribute values. A crop of this dataset is shown in Figure 3.2(a).

Cadastral parcel map: Is an object-oriented vector map of buildings in ESRI shape format. The same as TOP10, the production of this dataset is based on aerial photographs, field survey and some ancillary information but the production date is not known for sure. Each object in this dataset is numbered, coded and described based on their application and usage. A depiction of the Cadastral dataset for a part of the study area is also shown in Figure 3.2(b).

Enschede Cadastral Parcel: Is the polygon geometry with (outlines of) buildings and building blocks in ESRI shape format. The approximate scale of this dataset is 1:1,000 with the accuracy 30-60cm. This dataset is made on the basis of terrestrial surveys and large-scale aerial photographs. The parcel numbers are included in the parcel attribute table. The production date is unknown. A crop of this dataset is shown in Figure 3.2(c).

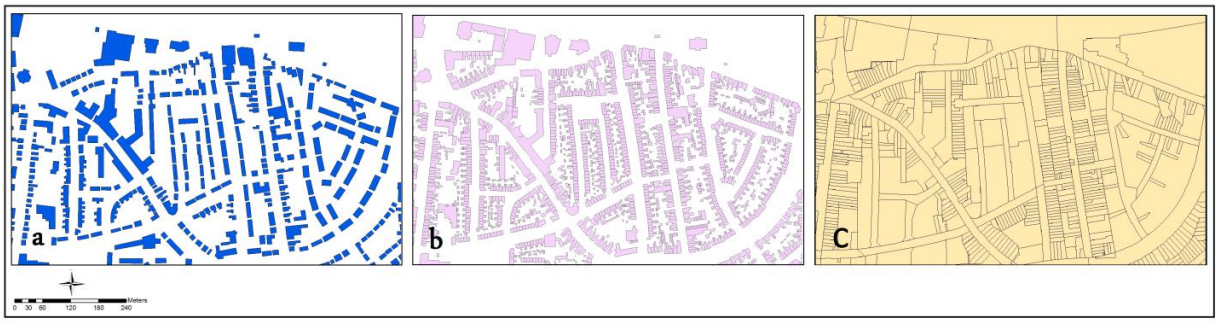


Figure 3.2. Samples of study area in (a) TOP10, (b) Cadastral datasets and (c) Enschede cadastral parcel datasets

TOP50Vector: Is a topographic dataset of roads of Enschede related to scale 1:50,000 in ESRI shape format. The data set is made on the basis of the TOP10 and other existing files in 2001; and is represented by lines.

Digital Terrain Model (DTM): Is a topographic model of the bare earth terrain relief contains the spatial elevation data of the terrain in an ASCII format with 2 meter grid cell size. The production date of this dataset is unknown.

Actual Height of the Netherlands (AHN): Actueel hoogtebestand Nederland (AHN) is translated to actual height of the Netherlands which contains detailed and precise elevation data in IMG format. During 1997 to 2003 the first version of AHN dataset were produced base on a specific and precise way for the whole of the country. The height is measured with laser altimetrie : a technique in which a plane or helicopter with a laser beam scans the surface. The measurement of the duration of the laser and reflection of the status and condition of the aircraft (via GPS) all together resulted in a very precise measurement of the height. Height of the manufactured models were based on a laser point per 4x4 or 5x5 meters and had a precision of at best 15 cm (1 point per 16 m^2). The provided raster has been derived from the original height points. This regular grid has a resolution of 5 meters and the height values are unfiltered (i.e. height of buildings and vegetations are included). AHN is using for many purposes in the Netherlands such as water and flood management, monitoring coastal erosion, hydrological models, drying out abadment, archaealogy, geomorghology, separation ground level and location and etc. (URL3). This research is utilizing AHN for extracting the height of buildings of the city of Enschede (section 3.5.3.1).

Postcode 4-digit (PC4) boundary: Postcodes in the Netherlands are alphanumeric, consisting of four digits followed by a space and two letters (NNNN AA) (URL2). The TNT post office in consultation with municipalities is responsible for adopting postcodes in the Netherlands. The smallest area-unit categorization of the city of Enschede is postcode 4-digit (the first four numbers), where based on the existed dataset the town is divided to 23 PC4 administrative units. This map is in ESRI shape format, but the production date in unknown.

Postcode 6-digit (PC6) points: The smallest categorization of postcodes in the Netherlands is postcode 6-digit points; where on average, 15 to 20 ‘addresses’ or ‘delivery points’ are allocated to each postcode 6-digit point (URL1). The city of Enschede is divided to 3948 PC6 points in the beginning of 2009. The tabular form of this dataset which is produced at the first quarter of 2009 contains the postcode 4 and 6-digit names and their relative location, number of population based on age and sex per each postcode, number of postal addresses and X and Y coordinates of each postcode 6-digit. For data consistency, it is converted to ESRI shape format. Figure 3.3 has a general overview of distribution of PC4 boundaries and PC6 points in the study area.

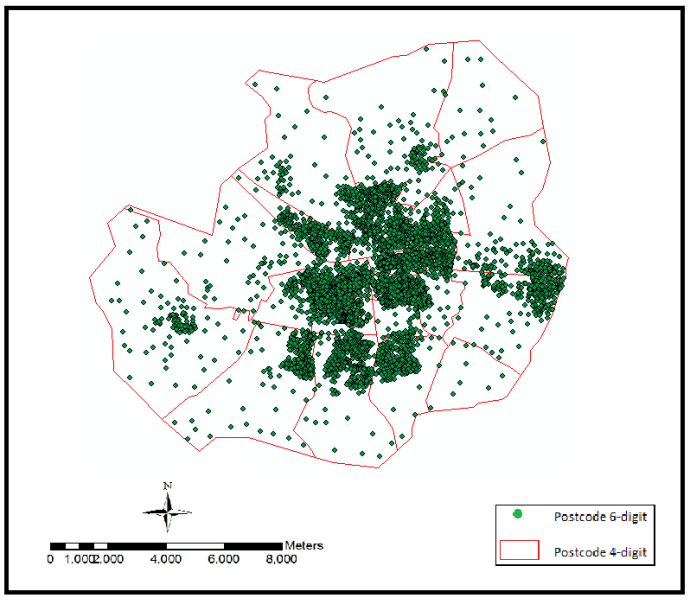


Figure 3.3. A general distribution of PC4 polygons and PC6 points in the study area

- **Information included postcode**

Postcodes in the Netherlands contain spatial information about delivery sections which are related to the different administrative units. For defining a new postcode, different criteria should take into consideration:

- It should be unalterable from a postal point of view, in other terms it should not be affected by any future changes in the postal organization.
- The postcodes should contain so much information that can be used throughout the sorting process.
- They should consist of a minimum number of positions, due to reduce the risk of error, both in the use of the postcode public and manual coding.

As it is mentioned before postcodes are defined alphanumerically, consisting of four digits followed by a space and two letters in such order NNNN AA (for instance 7514 AE). The 4-figure part identifies the place of residence, whereas the letters give more detail information regarding the delivery side. The municipalities are responsible for subdividing their territories into smaller areal units (town/district/quarter) and the TNT post office is in charge of postcode allocation to these areas. Each towns has a unique code in postal system (in the above example number 75 is related to the city of Enschede). The town first regionalize into several districts where their number can be identified by the third figure of the code number (number 1 in the above example)

as depicted in Figure 3.4. Based on the number of population within each district, they are subdivided into quarters. The quarters are identified by the forth figure code (number 4). Two letters follow the 4-figure town/district/quarter part of the postcode. The first letter indicates a group of streets roughly corresponding in size with postman's walk. The second letter indicates a subdivision of the area identified by the first letter. Together the two letters form a combination identifying a specific group of premises, which will always be unbreakable units within a postman's walk (Source: 2010, a report from TNT post office).

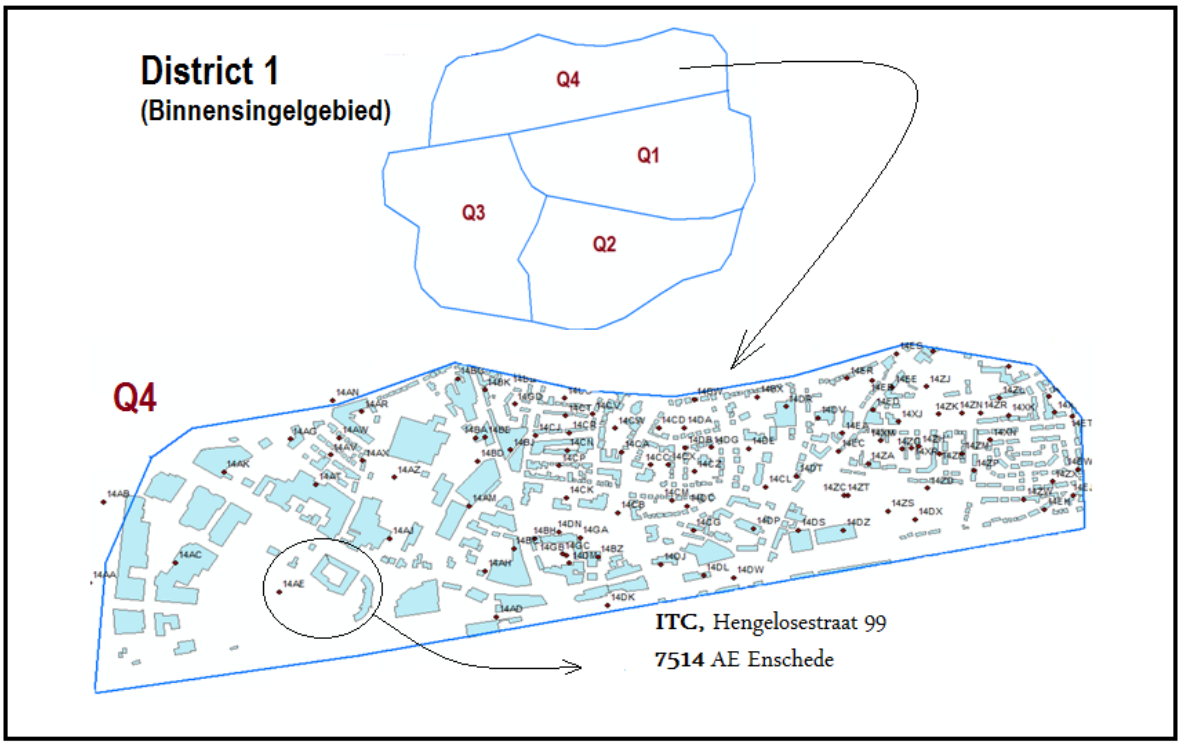


Figure3.4. An illustration of postcodes component

Since the study region is the city of Enschede in the Netherlands, most of the data are in *Rijksdriehoekstelsel_New* national coordinate system. All the data got this unique coordinate system to be spatially correlated to each other.

3.3. Finest administrative study level

The principle of spatial disaggregation methods is to transform data from a larger zone to subset zones in a disaggregating form. The original unit with known value, which is going to be disaggregated, is called the source zone and the smaller units which will get refined values are called the target zones. One of the objectives of this research is to disaggregate census population data to the finest possible level. The term "finest level" is directly related to the availability of data. As it is mentioned before, the smallest partitioning of the city of Enschede is postcode 4-digit areas; and each PC4 region contains a set of postcode 6-digit points. The aim of this section is to find out whether PC6 polygons can be produced based on PC6 points and to be considered as the finest target zone.

One approach to delineate boundaries to a set of points is using Thiessen polygon analysis (also referred to Dirichlet or Voronoi analysis in literature) (Brassel, 1979). This technique first introduced

by a climatologist A. H. Thiessen (1911) and has been commonly used in variety of disciplines. Thiessen polygon makes use of geometric distance to determine to which area the target points (PC6s) are related. In this context, this method is implemented on the PC6 points for demarcating the area of each point.
Postcode 6-digit are positioned as points and distributed heterogeneously in the study area (see Figure 3.5).

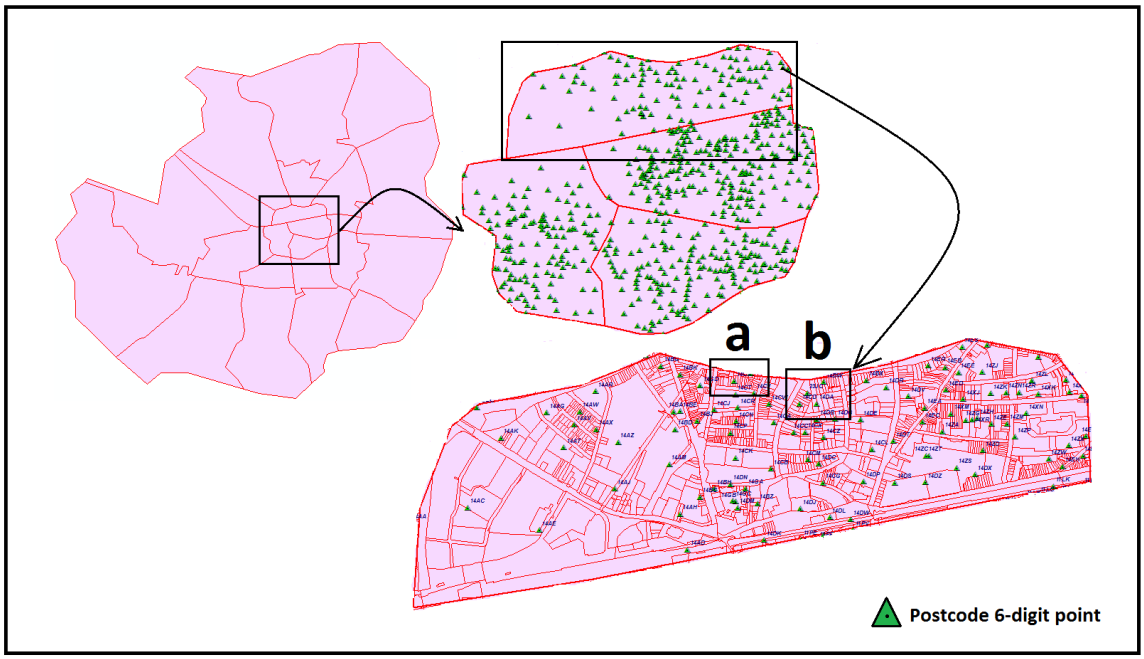


Figure 3.5. Distribution of PC6 points in a part of the study area

Two regions in Figure 3.6 are considered as samples to show how the distribution of postcode 6-digit areas in reality should be (black lines are reality). These boundaries are drawn manually based on the number of delivery addresses related to each PC6 point from a tabular dataset.

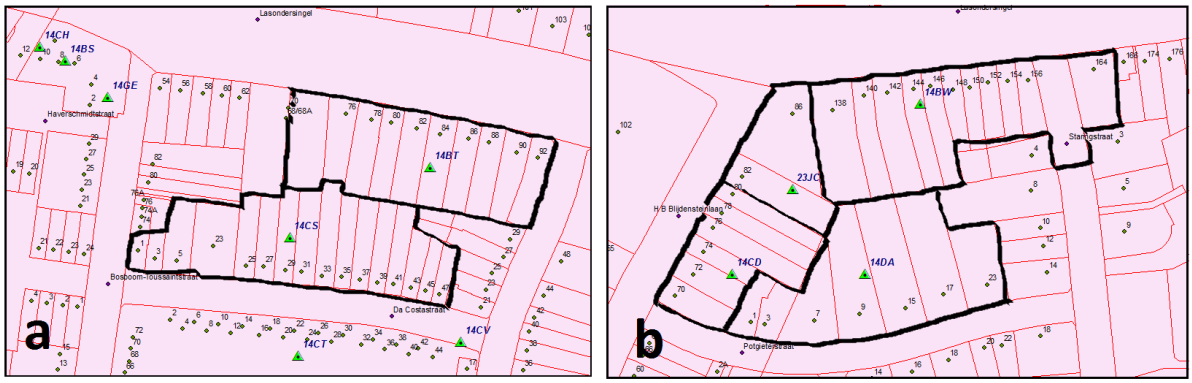


Figure 3.6. The distribution of PC6 for two locations in reality

Two types of thiessen polygons are created for PC6 points: one inside PC4 areas and the other in building block regions.

Producing Thiessen polygon for postcode 6-digit points inside postcode 4-digit zones: Since there are lots of PC6 points and also no systematic division of inhabitant areas, it is aimed at creating PC6 borders in a generic and automatic way. To this end, a model is built in the ModelBuilder environment of ArcGIS. It starts by selecting relative postcode 4-digits areas and postcode 6-digit points from the user. Then Thiessen polygons are created for the selected PC6 points. The outputs of this procedure are clipped based on selected PC4 areas. The final results are Thiessen polygons inside PC4 areas. Figure 3.7 portrays the structure of this model.

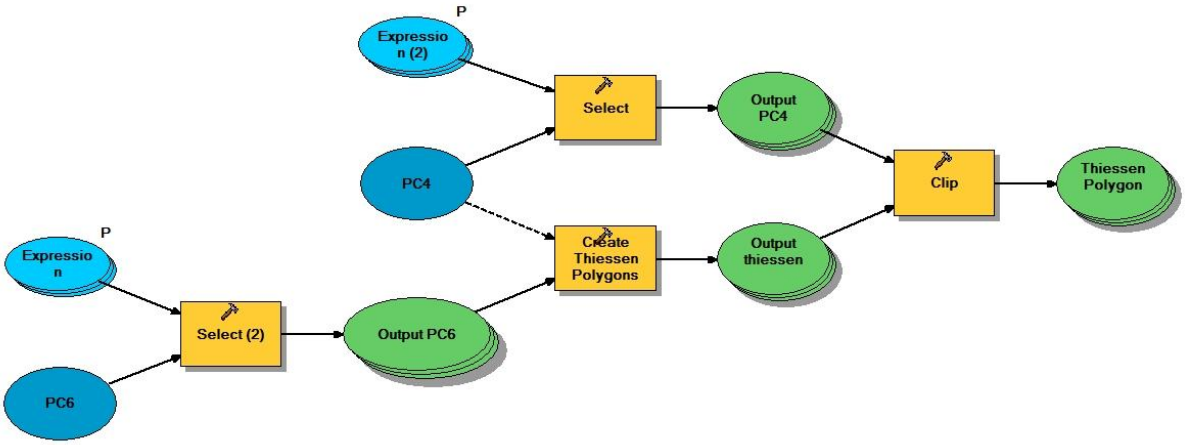


Figure 3.7. A model for producing Thiessen polygons in ModelBuilder environment (ArcGIS)

Since the inputs of the model are a list of values, the outputs which are the Thiessen polygons are a list as well. In order to union the outputs into a single layer, a script is written in Python programming language which can be found in Appendix I.

By executing Thiessen polygon technique, the study area regionalized to several portions that can be considered as postcode 6-digit boundaries. Figure 3.8 shows the result of applying this technique for the above study areas.

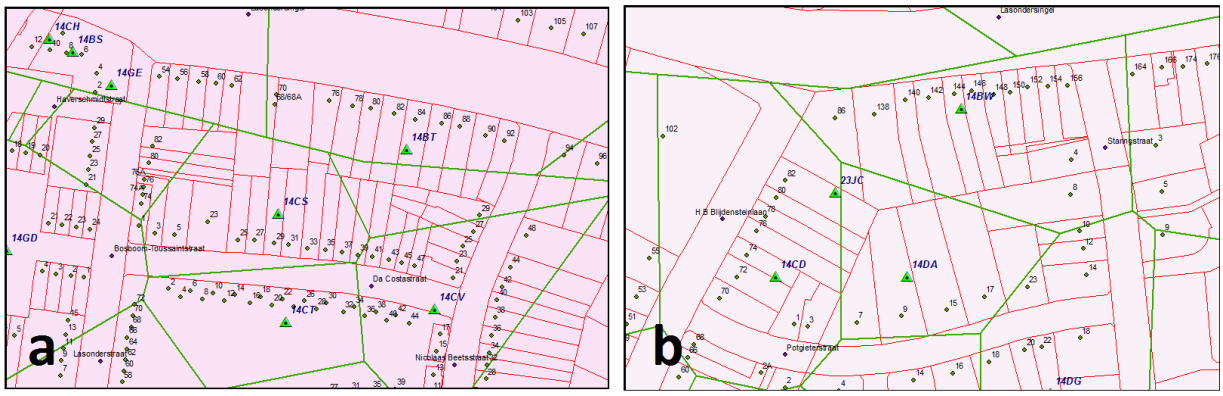


Figure 3.8. Implementation of Thiessen polygon on postcode 4-digit map

Comparing the previous outputs and what is in reality shows the accuracy of implementing Thiessen polygon technique on PC4 area (see Figure 3.9). The green lines are Thiessen polygons, the cyan lines are reality boundaries and the red lines are cadastral parcels. As it is portrayed, in some cases a large area is considered as PC6 region (e.g. 14CD) and somewhere the reality is bigger than what is

regionalized (e.g. 14CS). Both cases are far from reality and lead us to find an alternative for this situation.

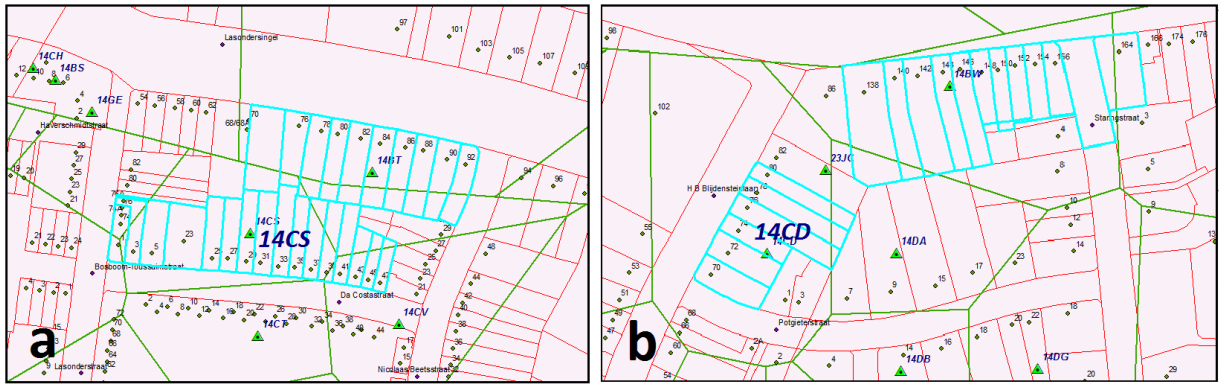


Figure 3.9. A comparison between PC6 regions in reality (cyan lines) and the result of Thiessen polygon in PC4 areas (green lines)

Producing Thiessen polygon for postcode 6-digit points inside the building block areas: Based on an obvious fact that the delivery of postal letters are done via roads, it is assumed that the houses which have same PC6 are close to each other and mostly are on the one side of the roads. In this condition, roads do not belong to PC6 areas. Hence, a 50 cm buffer is created around the roads axis sides (in TOP50 dataset) in order to remove as much as not-belonging areas to PC6 from cadastral parcels map. The map which is produced with 50 cm buffer distance has maximum removing of roads and less elimination of building areas; this gained by trial and error indeed. Remain areas are merged together and each of them considered as one building block. Each building block may contain zero to multiple PC6 points. In order to create Thiessen polygon inside the building blocks, the previous model which was created in ModelBuilder is utilized whereas it has the possibility to alter the initial values and rerun with a single click. The results (purple lines) are shown in Figure 3.10.

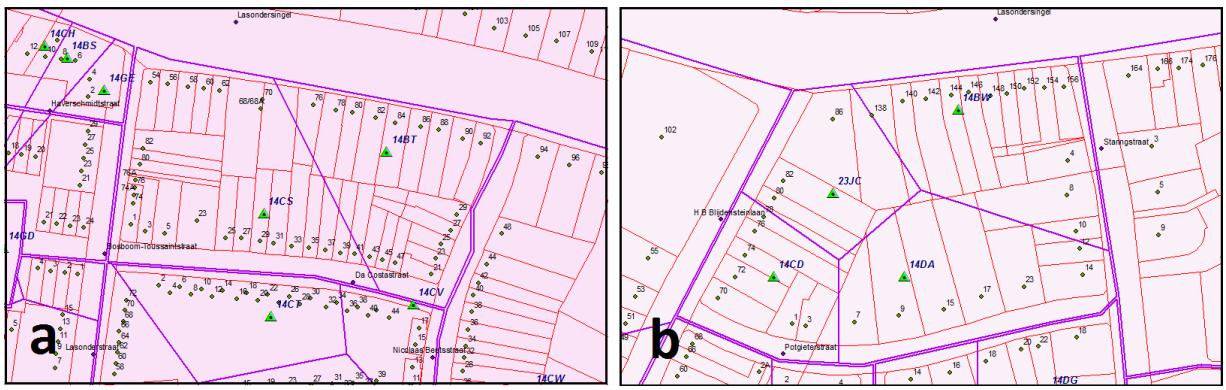


Figure 3.10. Implementation of Thiessen polygon on building blocks map

Comparing the previous outputs and what is in reality shows the accuracy of implementing Thiessen polygon technique on building block areas (see Figure 3.11). In comparison to the PC4 results, although the Thiessen polygons are created in surrounded areas (building blocks) and partitioned nicely, they also have some weaknesses. As it is portrayed, the same as what we saw in PC4 regionalization, in some cases a large area is considered as PC6 region and somewhere the reality is bigger than what is regionalized. Both cases are again far from reality.

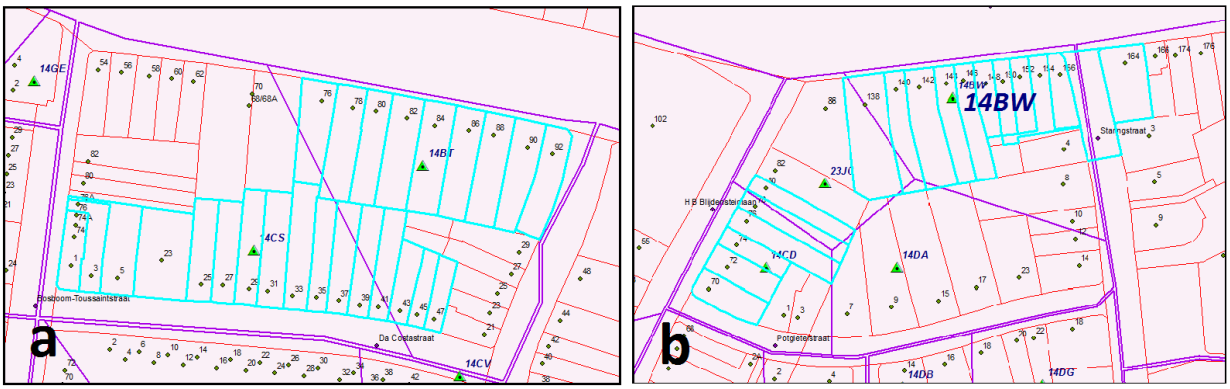


Figure 3.11. A comparison between PC6 regions in reality (cyan lines) and the result of Thiessen polygon in building block areas (purple lines)

Although it is assumed that most of the buildings which have the same PC6 number are located in one side of the roads, there are some exceptions the same as Figure 3.11 (b) in which the road crosses the one-code region (14BW). This method cannot distinguish such properties because is it a road-oriented method indeed (the first production of the dataset is based on road elimination).

Another difficulty which is related to the creation of Thiessen polygons inside the building blocks is concerning the buffer values (e.g. 50 cm) for removing the roads from cadastral map. Allocating any value may reduce the number of PC6 from the processing whereas they might be in the buffered areas which are eliminated. As an example in this case, this action reduces the number of PC6 from 3948 to 3901. This means that 47 of PC6 points are located in the roads (buffered areas), and by removing the road layers from the cadastral map these points do not contemplate in regionalization (see Figure 3.12). The chosen distance in buffer tool seems to be a proper amount, because more than this number would eliminate the building areas and less than it would allocate more road areas to postcode zones.

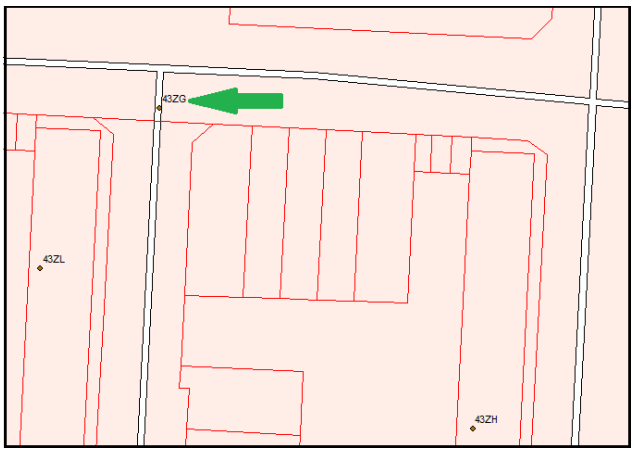


Figure 3.12. An out-of-area PC6 point

One of the demerits of creating Thiessen polygon inside the building blocks is that, there will be gaps for the areas which do not have any PC6 points inside their regions. In other words, since the allocation of PC6 points for an area is based on a minimum number of post addresses (around 15 to 20), some regions become vast due to have the minimum required number of post addresses (15 to 20) for PC6 dedication. Hence, because the building blocks are created based on road extraction, they may

not any PC6 exist inside them and they will not get any postcode 6-digit boundary in future. This produces gaps in the final result after Thiessen polygons are created (Figure 3.13a) whereas buildings are existed in empty regions (Figure 3.13b). This difficulty may be solved if a road dataset with different levels of road applications (highway, main road, alley, etc.) would be available.

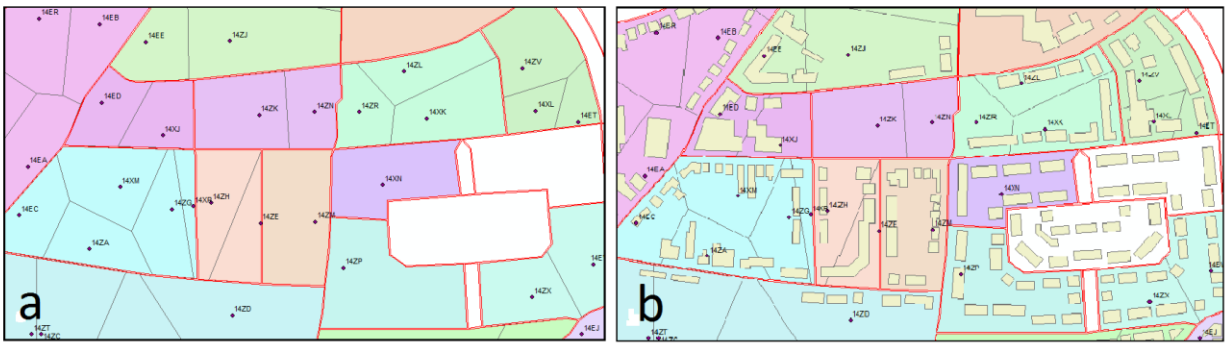


Figure 3.13. Gaps due to non-availability of PC6 points in some regions

As a visual comparison between the regionalization of postcode 6-digit boundaries in reality, in postcode 4-digit areas and in building blocks, Figure 3.14 can be considered.

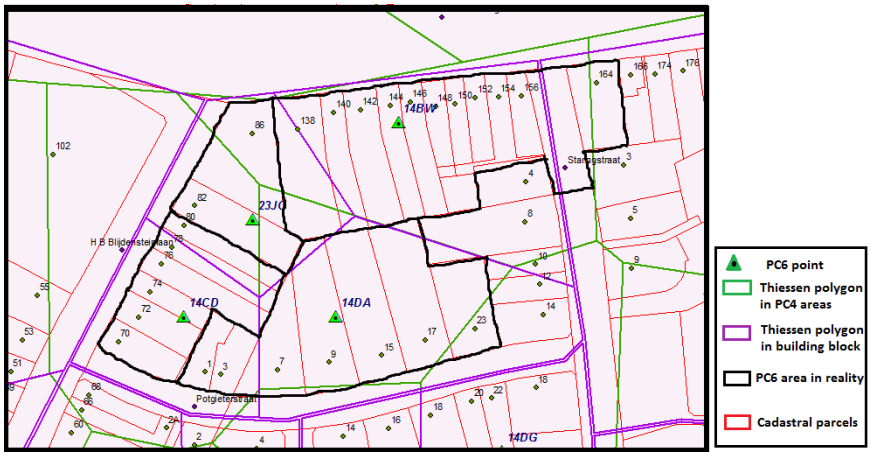


Figure 3.14. A comparison of overlaid Thiessen polygons in PC4 and building block areas vs. the reality

What can be inferred from this depiction is that, there is no accordance between created polygons and reality. The reason can be related to the fact that, postcode 6-digit allocation is largely based on the number of post addresses which should be in a finest number (e.g. 15 to 20) and is less related to the distance. Thiessen polygon is only based on distances and using it for such cases does not lead us to a proper (near to reality) and accurate result.

Based on the above results, what can be concluded is that, the PC4 and created PC6 boundaries are not suitable administrative study levels to be considering as source and target zones, respectively. Hence, two larger scale area units, the district level (as source zone) and neighbourhood level (as target zone) which have better consistency which each other are contemplated as administrative spatial units. Land cover maps, postcode 6-digit addresses and postcode 6-digit points are used as ancillary data and population per neighbourhood level is used as refined zones (target zones).

3.4. Methods

Dasymetric mapping which is translated to "density measuring" (Wright, 1936) is an area-based cartographic technique that utilizes ancillary data for transforming socio-demographic variable of interest (e.g. population) from one set of spatial unit to another. In dasymetric mapping, the administrative area (source zone) is divided into smaller spatial units in which population for each unit is averaged to get a rate, for instance population density. These smaller areas get a number of populations through utilizing land use/land cover maps (Holt et al., 2004).

In this section, two cartographic -Binary and 3-class- dasymetric methods will be discussed for disaggregation census population data and estimating the number of population per target zones. Since they employ ancillary data, they give better result than other disaggregation techniques which do not use these types of data. The TOP10 map is one type of ancillary data that is chosen due to its representation of buildings in unit/block level and also accessibility. Another type of dataset is the cadastral dataset which is a representation of the shapes and boarders of houses; and in this case, it outweighs the TOP10 dataset which is a cartographic representation of the building units. The binary dasymetric method is found simple and easy to implement in GIS. 3-class dasymetric mapping is a modified version of binary dasymetric method that takes advantage of numbers of ancillary classes, which generates higher values in comparison to the other areal interpolation techniques (Li et al., 2007). The modified 3-class dasymetric methods which utilize 3D data will be introduced in the continuous, and finally, a new type of ancillary data –postcode 6-digit- will be demonstrated for implementing in dasymetric methods. In the followings, the concept and the performance of two dasymetric methods will be explained.

3.4.1. Binary dasymetric method

One way to map population in a realistic form is to divide the source zone into two sub regions, one considered as populated and the other one as empty. Then the populated area is labelled as the source zone. This concept previously proposed by Wright (1936) and recently referred to an areal interpolation that is called Binary Dasymetric Method. It is described formulaically according to Fisher and Langford (1996) as:

$$\widehat{P}_t = \sum_{s=1}^{S} \frac{A_{tsp} P_s}{A_{sp}} = \sum_{s=1}^{S} A_{tsp} d_{sp} \qquad 3.1$$

Where $\widehat{P}_t$ is the population per target zone t; A_{tsp} is the area of overlap between target zone t and source zone s having land cover identified as populated; P_s is the number of population per source zone s; A_{sp} is the area of source zone s identified as populated; S is the number of source zones; and d_{sp} is the dasymetric density of population in source zone s which will be calculated by diving P_s by A_{sp}. Figure 3.15 shows the procedure of applying binary dasymetric mapping in ArcGIS.

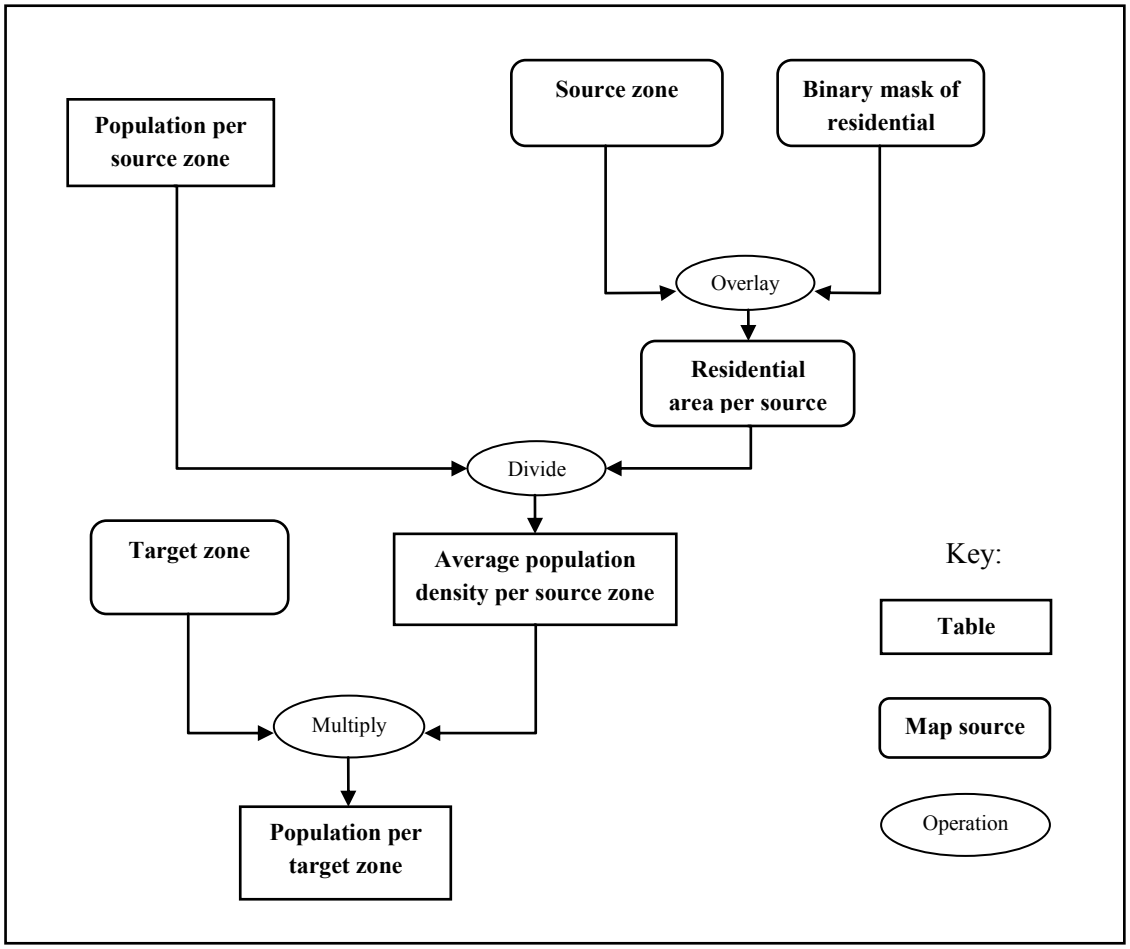


Figure 3.15. The process of implementing binary dasymetric mapping

For better perception regarding the binary dasymetric method procedure performance, an implementation of this method across the central district of the study area will be explained based on the above diagram. The central district of TOP10 dataset which has nine neighbourhoods is examined due to its complexity in the type of buildings application and also, the availability of 3D data which will be explained more in next section. Four spatial inputs are used: source zone and target zone as administrative levels which are one district and nine neighbourhoods respectively, a map of built-up areas from the TOP10 dataset and the tabular census population dataset of the central district. The first step is to mask the study region into residential and non-residential areas by overlaying the original source zone and the foot prints of inhabited building, then considering the result as the new source zone for the rest of the implementation. But, because in this case the input maps are just the foot prints of built-up areas, as depicted in Figure 3.16, there is no need for masking execution.

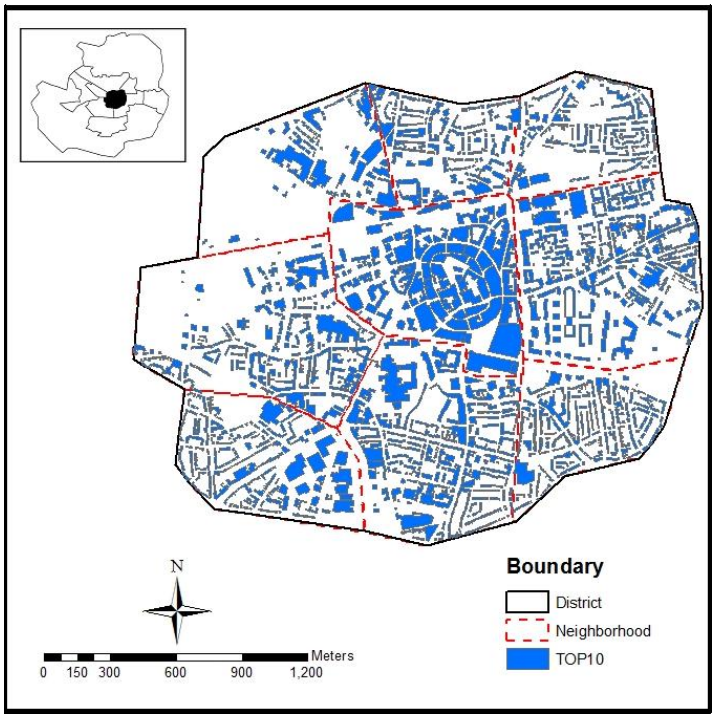


Figure 3.16. An illustration of built-up areas from central district of the TOP10 dataset

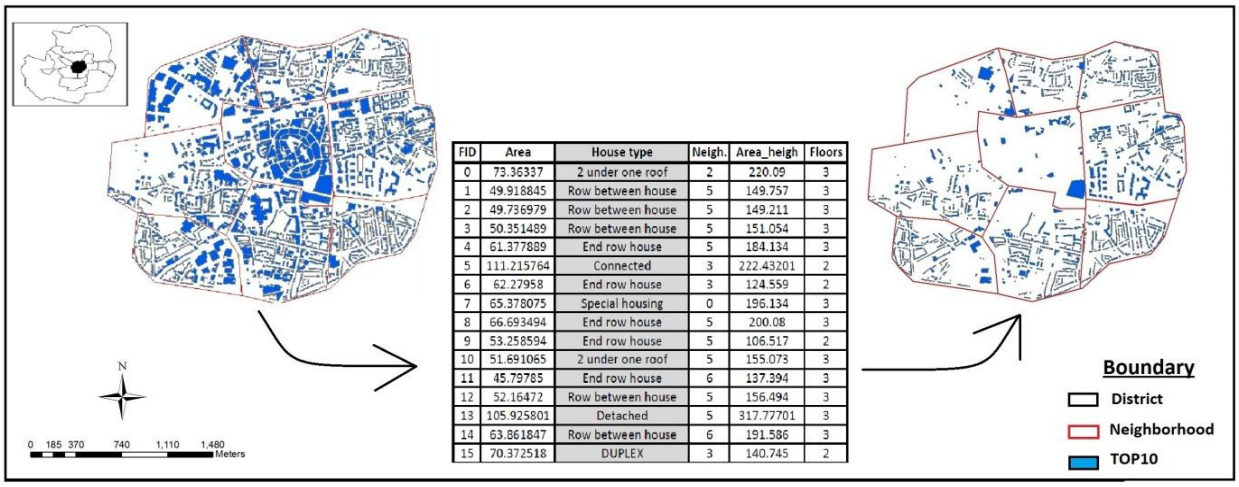
The average population density for a source zone (d_{sp}) is derived by dividing the number of population per source zone (P_s) by the area of source zone (A_{sp}) (see equation 3.2).

$$d_{sp} = \frac{P_s}{A_{sp}} = \frac{22710}{1113713} = 0.02\ (^{person}/_{m^2}) \qquad 3.2$$

Next, the population per each target zone will be achieved by multiplying the average population density by the correspond area of each target zone.

$$\widehat{P}_t = A_{tsp} d_{sp} = \begin{pmatrix} 242923.4 \\ 65827.21 \\ 51591.46 \\ 170587.5 \\ 69768.98 \\ 171388.6 \\ 97133.19 \\ 114729 \\ 129763.4 \end{pmatrix} (0.02) = \begin{pmatrix} 4955.61 \\ 1342.87 \\ 1052.46 \\ 3479.97 \\ 1423.28 \\ 3496.31 \\ 1981.51 \\ 2340.46 \\ 2647.16 \end{pmatrix} \qquad 3.3$$

As a principle, the areas of populated units are using as input in dasymetric equations. But, first effort for disaggregation population is applied to the built-up areas of the central district of the TOP10 dataset where the residential and non-residential areas are not discriminated from each other. The aim is at appraising the importance of differentiation between residential and non-residential areas in the course of population disaggregation. To this end, the original TOP10 dataset which is the foot print areas of built-up regions is modified by eliminating the non-residential units via a descriptive dataset of buildings' applications and considered as source zone (see Figure 3.17). Then, the binary dasymetric method is applied to estimate the population in the target zones with respect to the new source zone.



FID	Area	House type	Neigh.	Area_heigh	Floors
0	73.36337	2 under one roof	2	220.09	3
1	49.918845	Row between house	5	149.757	3
2	49.736979	Row between house	5	149.211	3
3	50.351489	Row between house	5	151.054	3
4	61.377889	End row house	5	184.134	3
5	111.215764	Connected	3	222.43201	2
6	62.27958	End row house	3	124.559	2
7	65.378075	Special housing	0	196.134	3
8	66.693494	End row house	5	200.08	3
9	53.258594	End row house	5	106.517	2
10	51.691065	2 under one roof	5	155.073	3
11	45.79785	End row house	6	137.394	3
12	52.16472	Row between house	5	156.494	3
13	105.925801	Detached	5	317.77701	3
14	63.861847	Row between house	6	191.586	3
15	70.372518	DUPLEX	3	140.745	2

Figure 3.17. Removing non-residential areas from TOP10 dataset by utilizing a descriptive table

In parallel to TOP10, the buildings which seem not to be residential are removed from the cadastral dataset. In addition, all the built-up units which have the area less than 41 m^2 also eliminated from this dataset. This value is related to the average living space per inhabitance in the Netherlands (URL4), so the areas downward from it are considered as non-residential.

As it is mentioned before, one of the sub-objectives of this research is to find out the right scale of ancillary dataset for implementing in dasymetric mapping. This issue is important because detailed information and high resolution data are not always reachable or easily producible; and in the cases of availability, they may not always be free of charge. To this end, binary dasymetric method is applied to the central district of residential areas of the cadastral dataset.

In order to evaluate the existing dasymetric methods, the binary method is executed for the complete study area. The result of this operation statistically will be compared by the other dasymetric method.

Binary dasymetric method is theoretically simple and easy to implement in GIS. Results show that it performs better than many alternatives (Eicher and Brewer, 2001) and it is robust for classifying errors (Fisher and Langford, 1996). Versus its advantages, binary dasymetric mapping is unable to deal with the areas with complex variety of population distribution due to the use of binary land classification (only residential and non-residential) where the residential areas might be a combination of different classes (low-density, high density, etc.). In order to improve the shortcomings of binary dasymetric method, a new disaggregation method is introduced which is called 3-class dasymetric method.

3.4.2. 3-class dasymetric method

Although the binary dasymetric method allocates population to the populated areas, population seems to be more complex than it can be modelled by the binary dasymetric method in simplistic form of populated and empty classes. In reality, residential densities may vary from place to place due to reasons such as: (i) the variety of physical shape and size of house units (e.g. flats, terraced houses, apartments, semi-detached and detached houses) and/or (ii) different socio-economic factors which may influence the occupancy rate (e.g. age, wealth and cultural groups). For instance, in semi-detached or detached houses mostly larger families tend to be inhabited, while in flat houses retired

people or young couples tend to live. This variation in inhabitancy might generate variable density indeed. Furthermore, the wealth and the age are not obviously discernable, but they can be associated with the location, size and type of house characteristics.

The 3-class dasymetric method (Mennis, 2003; Langford, 2006) improves the binary dasymetric method by considering a limited number of ancillary classes (e.g. high-density residential, low-density residential and non-urban) to display a range of residential density in the source zone. Langford (2006) implemented the relative density for each land class within each source zone as follows:

$$\widehat{P}_t = \sum_{s=1}^{S}\sum_{c=1}^{C}\frac{A_{tsc}P_s}{A_{sc}} = \sum_{s=1}^{S}\sum_{c=1}^{C}A_{tsc}d_{sc} \qquad 3.4$$

Where $\widehat{P}_t$ is the population per target zone t; A_{tsc} is the area of overlap between target zone t and source zone s having land cover identified as populated class c; P_s is the number of population per source zone s; A_{sc} is the area of source zone s identified as populated and having land cover class c; S is the number of source zones; C is the number of populated land cover classes; and d_{sc} is the dasymetric density of populated class c in source zone s.

One problem arises regarding the definition of "density" during the classification of source zones into high and low-density classes. This issue cannot be seen in the binary dasymetric method where the area is classified into populated and empty classes; and the populated class is considered as source zone. Density is a complex concept and is related to the nature of the phenomena which density is associated with. Generally, it is measured by considering different numerators and denominators. Population density is the number of people per area unit, while residential density is the number of dwelling units per given area. Also living density is the number of persons per room in a living unit. Furthermore, in each of these terms, density values can be considered relatively; for example, high-density, medium-density or low-density. The challenging task is to determine, what is called high, medium or low? Can a series of numbers be defined for these relative terms? What can be found in literature is that, there is not one solution for every area. Each field has its own characteristic and the interrelationship between variables and factors must be considered (Churchman, 1999).

In this context, Langford (2006) proposed several ways to calibrate the class density parameters:
(i) Considering a fixed proportion of the total population for each source zone class. This is the method that was utilized by Eicher and Brewer (2001) to proportionate the total population subjectively; for instance, considering 70% of the study area as urban, 20% as agriculture or forest and 10% as water resources. This approach has a weakness, because all the regions do not have same distribution of these land classes. Even a region with small proportion of urban area will get 70% of total population.
(ii) Determination ratios for high and low classes by statistically modelling the relationship between the dataset. In this case, a fixed ratio of density will be defined for each land class (Yuan et al., 1997).
(iii) By utilizing selective sampling within the dataset, relative density for high and low classes can be achieved. In this case, relative density for one land class will be computed based on the assumption that the region is entirely covered by a single land cover class, by simply dividing the total population by total area. This value can be generalized for all of the study area where different land cover classes are available.

In this research, population density is more remarkable than residential or living densities since it is aimed at disaggregating population based on areal units. In the context of determining a threshold for

density variables, through the study area, up to three-story buildings are considered as low-density and higher than this value (> three-story) as high-density. This threshold is the average number of floors (2.7) of residential buildings, but because the decimal value does not have any correlation with the floor number, it is rounded up. It should be noted that, the study region has 143 km^2 area and around 90% of the inhabited buildings have three and less than three floors. Due to the discussed weakness of method (i) and also no availability of samples to be considered as one density variable (Low/High) in the method (iii) the second method is examined to estimate the dasymetric density (d_{sc}) for each class of the study area. To this end, equation 3.5 is applied:

$$P_s = D_L * A_L + D_H * A_H \qquad 3.5$$

Where, P_s is the population per source zone, D_L and D_H are the dasymetric densities for the low-density and high-density classes respectively, A_L is the populated area of the class low-density for each source zone and A_H is the populated area of the high-density class within each source zone. This equation has two unknown parameters -D_L and D_H- and will simply be solved with two equations. In other words, if the population per two source zones and the related area of low and high-density values are known, dasymetric densities for low and high-density classes can be derived. But, since there might be several source zones, the population values and also the related areas for the source zones will be larger than the number of unknowns in this case. One solution is to utilize least square method to find optimum value of unknowns based on the observations (see equation (3.6)).

$$\hat{X} = (A^T A)^{-1} A^T P_s \qquad 3.6$$

Where $\hat{X}$, relates to the unknown parameters, and in this case is the low and high-density values; A is the coefficient matrix and in this case the elements are low and high-density areas; and P_s is the observation matrix that the inputs are the population per source zones.

$$P_s = \begin{bmatrix} D0 \\ \vdots \\ D8 \\ D9 \end{bmatrix}, \qquad A = \begin{bmatrix} A_{L0} & A_{H0} \\ \vdots & \vdots \\ A_{L8} & A_{H8} \\ A_{L9} & A_{H9} \end{bmatrix}, \qquad \hat{X} = \begin{bmatrix} D_L \\ D_H \end{bmatrix} \qquad 5.1$$

Based on the above instructions, the 3-class dasymetric method is examined for the complete city. The calculation procedure is the same as Figure 3.15 except that, for each target zone two types of population is calculated with respect to the source zone dasymetric density. These values are computed as 0.011 ($^{person}/_{m^2}$) for low-density ratio (D_L) and 0.012 ($^{person}/_{m^2}$) for high-density ratio (D_H). They multiply by their respective areas and the summed result will be considered as population per target zone.

3.4.3. Modified binary and 3-class dasymetric methods

The current dasymetric methods utilize the footprint areas of building units without considering their height or the number of floors. In this case, two buildings with same footprint areas but different in their number of floors will get equal population values during the population allocation procedure, which is far from reality. Since, one of the objectives of this research is evaluating the extra value of 3-Dimensinal data in the process of disaggregating census population data, this section makes

contribution of 2-Dimensional and 3-Dimensional data in the binary and 3-class dasymetric methods to assess how these ancillary information can improve the final result. Specifically, this information is the height of buildings and the numbers of floors. Since the TOP10 dataset is a cartographic representation of the study area and does not have the height values in its attribute table, it is attempted to calculate them based on height surface maps which are the representation of the terrain elevation.

Extracting the height of buildings

The efforts started by cropping the central district buildings (due to availability of elevation data) from the larger dataset. First aim is at, creating a digital terrain model (DTM) for representing the earth terrain relief by using current AHN dataset. This work assumed to contribute in estimating the heights by subtracting the created DTM values from AHN values. The TOP10 map is used for removing those pixels which are related to buildings in AHN dataset (see Figure 3.18).

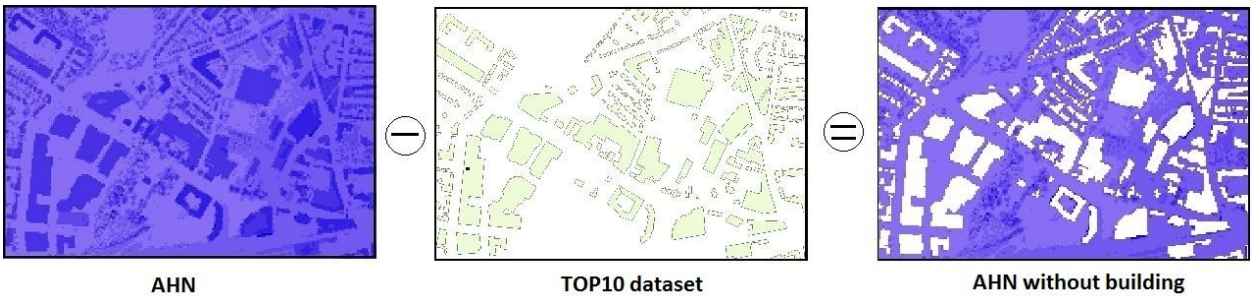


Figure 3.18. The process of removing buildings from AHN

The result is the AHN map without including building pixels. In order to fill out the gaps and create a smooth surface, the Inverse Distance Weighting (IDW) interpolation method which is a widely used and simple executing method is examined (see Figure 3.19). One of the characteristic of this method is the dependency of all the points which are going to be considered for interpolation to each other, on the basis of distance. Hence, the calculation of elevation in the region is largely depends on the elevation of sample points. The sample points which are the pixels around the unvalued areas (gaps) are weighted in the process of interpolation such that the impact of one point concerning another point decreases with distance from the unknown point (Achilleos, 2008).

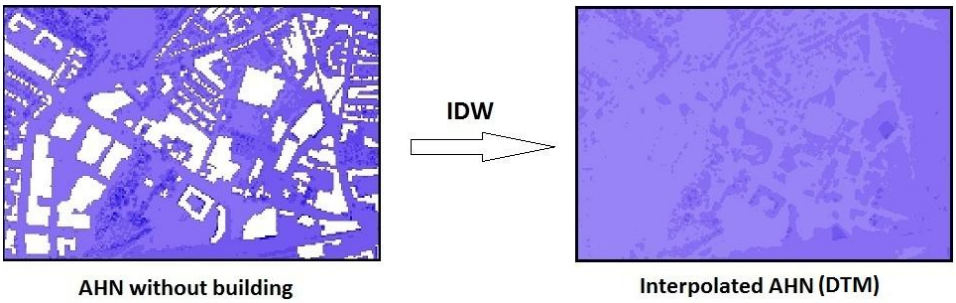


Figure 3.19. Creating interpolated AHN via IDW method

Again, the TOP10 map is implemented for clipping the buildings from the interpolated map and also AHN dataset. The cropped maps are subtracted from each other in which a raster map that contains the height of buildings is produced (see Figure 3.20).

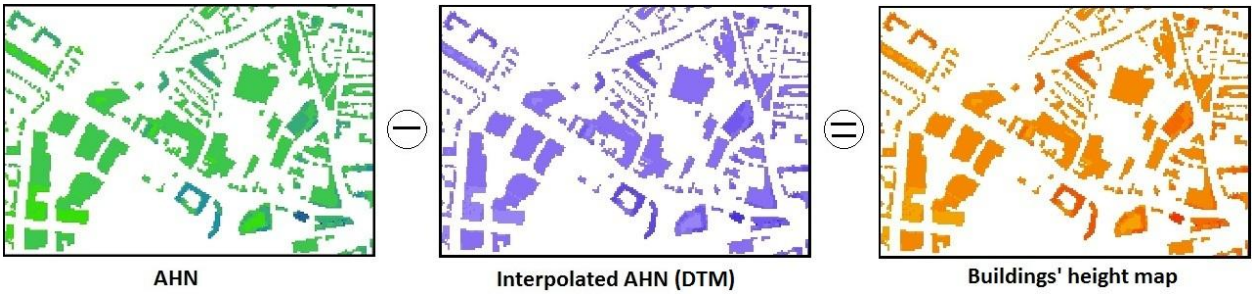


Figure 3.20. The process of creating buildings elevation

Since series of pixels are allocated to each building (based on the TOP10 polygons), the related average values of pixels could be considered as buildings' height (see Figure 3.21). By creating a new field in the attribute table of the building map named as building height, the elevation values are attached to each feature respectively.

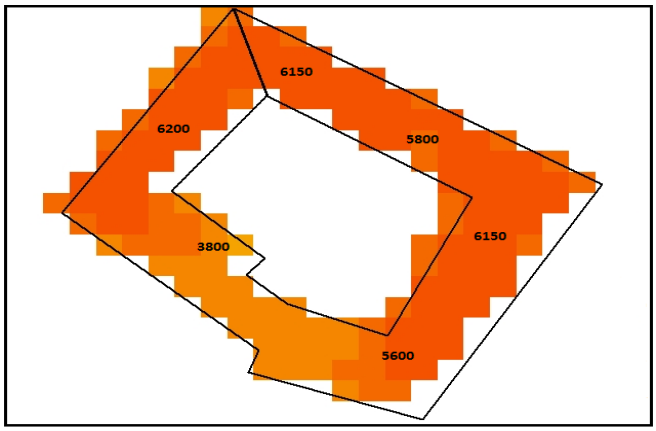


Figure 3.21. Allocation of different range of pixel values to each feature

After investigating the result, it is found out that, some buildings have negative values in their altitude which are not logical. Also, in some cases, the height values of buildings are far from what is expected in reality. Figure 3.22 compares the acheivements of a building height and the reality.

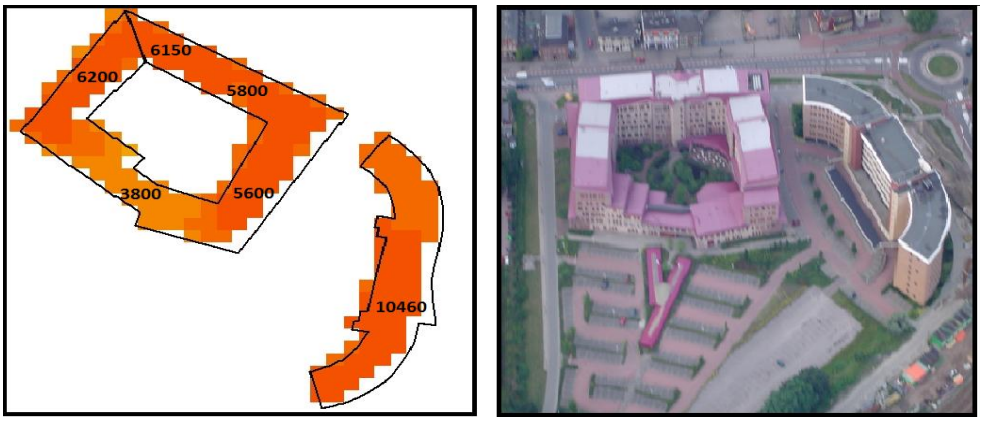


Figure 3.22. A comparison of reality and elevation results

In the above illustration, both buildings have same height in their higher level, but the average of the pixel values say something different (6.10 m vs. 10.46 m). Different reasons can be correlated to the preceding difficulties.

First, AHN dataset is realized as a probable error generator. As it mentioned before, this regular grid has a resolution of five meters and the height values are unfiltered (i.e. height of buildings and vegetations are included). During the process of pixel initializing, all the objects that have values are incorparated in the averaging process, either they are built-up objects or other features.This leads to unpurity of each pixel value in relation to one feature class (scientificually called mixed pixels); whereas the pixels which contains buildins are desired.

Second, the TOP10 dataset which is utilized during clipping AHN dataset can be a source of error generator. During cropping the building pixels, the borders of each polygon which has minimum intersection with AHN pixels, crop the related pixel. The cropped pixels may not relate to built-up regions.

Another reason for the appearance of errors in the final result can be related to IDW interpolation method. One of the disadvantages of IDW method is that, the quality of the interpolation result can decrease, if the distribution of sample data points is uneven. This also can be seen in the building-cropped AHN dataset.

Last but not least, the production dates of AHN dataset and Top10 polygons differ from each other. AHN dataset is produced among 1997 to 2003, while Cadastral map is generated in the beginning of 2008. This might leads to some mismatches during overlaying process.

Since it is desired to utilize as minimum as data types for achieving the goals (due to generality and applicability for other study areas), the DTM of the study region was produced by IDW method. While it is assumed that the error generation in the final result regarding the buildings height may be related to this interpolation method, the DTM of the Enschede is used as a replacement in the process of height derivation.

In this process, first the pixel dimension of the DTM (2x2 meter) is converted to AHN cell dimesion (5x5 meter) for data consistency. Then the calculations are pursued by subtracting the cropped building pixel values of DTM from the cropped building pixels values of AHN dataset. What remains are the corresponding building altitudes from terrain level. In order to calculate the number of floors for each building, the height value is divided by 2.9 which is the average height of each floor in the Netherlands (URL5).

In contrast to the previous work, the other vector building map, the cadastral dataset has an attribute table which contains the number of floors of each building. So, there is no need for calculating these values.

Implementing floor factor in dasymetric methods

Floor parameter (F), which is the number of floors of each building is implemented in the binary and 3-class dasymetric methods equations. The modified forms of these formulas can be seen in equations (3.7) and (3.8), respectively.

$$\widehat{P}_t = \sum_{s=1}^{S} \frac{(A_{tsp}F)P_s}{(A_{sp}F)} = \sum_{s=1}^{S} (A_{tsp}F)d_{sp} \qquad 3.7$$

$$\widehat{P}_t = \sum_{s=1}^{S}\sum_{c=1}^{C} \frac{(A_{tsc}F)P_s}{(A_{sc}F)} = \sum_{s=1}^{S}\sum_{c=1}^{C} (A_{tsc}F)d_{sc} \qquad 3.8$$

This method implements the floor area instead of the foot print areas in the process of population estimation, where allocate the population to each floor based on its area.

3.4.4. Postcode 6-digit ancillary data

One of the difficulties that dasymetric mapping deals with, is related to the type of ancillary data which corresponds well with population distribution (Bieleka, 2005). Various types of spatial data such as, topographic maps (Wright, 1936), satellite images and land use maps (Langford and Unwin, 1994; Eicher and Brewer, 2001; Mennis, 2003) and night-time city light imagery (Sutton et al., 1997) have been used as ancillary data for contributing in disaggregating population. In most of the quoted publications about dasymetric mapping, the areas of populated regions are considered as computational units. What can be inferred from the dasymetric equations is that, any spatial data that have correlation with population can be used for population estimation. It is deemed that, the 6-digit postcodes can be helpful to find out the internal distribution of populated area within the source zones; and it may be a substitution for building areas. Nowadays, postcodes are more than just a numbers where many demographic investigations are doing throughout this information. In this context, postcode 6-digit addresses and postcode 6-digit units are examined. The former is the number of postal addresses which are related to each PC6 point and the later is the number of PC6 units (points) within an areal unit. Principally, there should be around 15 to 17 post delivery addresses exist in order to get a unique PC6 number. The usability and preference of PC6 addresses and units can be highlighted by a comparison between them and the "area" of residential zones as ancillary data. The reason can be due to the production or updating the land cover (in this case built-up) maps which is doing time to time while the postcodes are defined once and is unalterable for a long time. The land cover maps are producing by different organizations in various scales and accuracies while the municipalities (and in the case of the Netherlands only by contributing TNT post office) are responsible for allocating postcodes to urban areas. In addition, there is no need for calculating the height and their corresponding number of floors as 3D data where postcodes are going to be implemented directly as ancillary data in dasymetric methods. In other words, the postcodes are indicating the corresponding addresses which are related to them.

3.5. Performance evaluation

The result of the implementation will be assessed based on an independent data source. A variety of ways for evaluating the overall accuracy of areal interpolation methods are used in literature. In this case, the mean absolute error (MAE) (as used by Langford, 2006), the root mean square error (RMSE) (as used by Eicher and Brewer, 2001) and a standardized coefficient of variation CV(RMSE) (as used by Fisher and Langford, 1995) are chosen because they can be applied to count data (e.g. population number) and also the values can be easily interpreted relatively.

Mean absolute error (MAE), is the average of differences between actual and estimated values. The formula for calculating MAE is given in equation 3.9 as below:

$$MAE = \frac{1}{n}\sum_{n}\left|Y_n - \hat{Y}_n\right| \qquad 3.9$$

Where MAE is the mean absolute error, n is the number of target zones, $\hat{Y}_n$ is the estimated values of target zone and Y_n is the known actual values.

The root mean square (RMS) error is based on the average differences between the estimated and the actual values where it is just the square root of the mean square error as shown in equation 3.10:

$$E^{RMS} = \left[\frac{1}{n}\sum_{n}\left(Y_n - \hat{Y}_n\right)^2\right]^{1/2} \qquad 3.10$$

The RMSE is more sensitive to the occasional large errors, since the squaring process gives weight to the large errors.
The calculation of RMSE has three steps in general:
1. Calculating the square of differences between true and measured values for each dasymetric zone.
2. Compute the mean of these calculated errors.
3. Take the square root of the mean of the result from step 2.
In order to evaluate the performance of areal interpolation results, the true values are required. The correct data of dasymetric units (for the 1st step) are calculated by the summation population at the neighbourhood levels.

The MAE and RMSE are both measured in the same unit as the original data and they can be used to diagnose the variation in the errors. The MAE is equal or smaller than RMSE and the larger disagreement between them means the greater differences in individual errors. If the RMSE and the MAE are equal, then all the errors are off the same magnitude. Both of them are ranged from 0 to ∞ and are negatively-oriented where the lower values are better.

In addition to the previous statistical analysis approaches, the coefficient of variation of the RMSE (in short CV(RMSE)) is obtained by dividing the RMSE values by the mean of the correct population of target zone ($\overline{N}$) (see equation (3.11)) which would be useful for the purpose of reporting if it expresses in percentage.

$$CV(RMSE) = \frac{RMSE}{\overline{N}} \qquad 3.11$$

Despite the tabular form of errors representation, visualizing the errors can be an efficient way for analyzing the results (Eicher and Brewer, 2001) and provides a comparison against the actual and estimated values.

4. Result and Discussion

4.1. Overview

In dasymetric mapping processes, ancillary data help to realize the internal distribution of populated areas within the source zones. Hence, it is desired to find the datasets which perform well with current dasymetric methods and contribute in estimating population. In this chapter, the results of executing various types of ancillary data in dasymetric methods are going to be evaluated. First, 2-Dimensional data which are the 1:10,000 scale and the cadastral maps are going to be assessed. Second, 3-Dimensional data in connection with 2-Dimensional data will be evaluated. Finally, a new type of ancillary dataset –postcode 6-digit– will be examined to appraise its robustness in the course of disaggregating population.

4.2. Evaluation of 2D ancillary data in dasymetric methods

In the context of population transformation from a large zone to the smaller sub-zones, most of the studies have utilized the area of residential units. This section will evaluate the performance of binary and 3-class dasymetric methods and their corresponding ancillary foot print maps.

Performance of binary dasymetric method

As a first attempt, the area of built-up regions from the TOP10 dataset is utilized in population estimation process to assess the role of buildings application in population propagation. In this way, the calculated dasymetric density of the central district and the estimated population per target zones are summarized in Table 4.1.

Table 4.1. Implementing binary dasymetric method by utilizing built-up areas for central district of the TOP10 dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (m^2)	Target zone code	Target zone (Neighbourhood)	Target zone area (m^2)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.02	1113713.3	**n0**	City	242923.4	4955.61	2620
					n1	Lasonder, Zeggelt	65827.21	1342.87	1240
					n2	De Laares	51591.46	1052.46	1530
					n3	De Bothoven	170587.5	3479.97	5720
					n4	Hogeland-Noord	69768.98	1423.28	2610
					n5	Getfert	171388.6	3496.31	4020
					n6	Veldkamp-Getfert-West	97133.19	1981.51	1920
					n7	Horstlanden-Stadsweide	114729	2340.46	2570
					n8	Boddenkamp	129763.4	2647.16	480

What is remarkable in the result, is the large overestimation of population in the first (n0) and the last (n8) target zones; the areas which seem to have mixed building application. In order to find out these disagreements and see the influence of buildings application in population estimation, the binary dasymetric method is tested for the populated areas of the same region (central district of the TOP10 dataset) where the non-residential areas are removed from the source data. The numeric result and actual values of this part can be seen in Tables 4.2.

Table 4.2. Implementing binary dasymetric method by utilizing residential areas for central district of the TOP10 dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (m^2)	Target zone code	Target zone (Neighbourhood)	Target zone area (m^2)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.055	414579.6	**n0**	City	36025.77	1973.4	2620
					n1	Lasonder, Zeggelt	27948.61	1531	1240
					n2	De Laares	35432.59	1940.9	1530
					n3	De Bothoven	85377.29	4676.8	5720
					n4	Hogeland-Noord	44270.48	2425.1	2610
					n5	Getfert	63842.78	3497.2	4020
					n6	Veldkamp-Getfert-West	48299.07	2645.7	1920
					n7	Horstlanden-Stadsweide	47306.21	2591.4	2570
					n8	Boddenkamp	26076.85	1428.4	480

The values of two sample target zones (n0 and n8) become much nearer to the reality by using the inhabitant building units, although still some differences exist between theses values. The first two rows of Table 4.4 relates to the statistical analysis of utilizing the residential and non-residential building areas for binary dasymetric method. What is highlighted is the large difference between error values of these two datasets, 55% vs. 25% (more than double error value) where around 30% of the generated error is related to non-residential buildings.

Regarding the sub-objective for the optimum scale for ancillary data, the binary dasymetric method is applied to the residential areas of the cadastral dataset where the results are summarized in Table 4.3. The calculated and the actual values of this performance are listed in the last two columns.

Table 4.3. Implementing binary dasymetric method by utilizing residential areas of central district of the cadastral dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (m^2)	Target zone code	Target zone (Neighbourhood)	Target zone area (m^2)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.033	675769.43	**n0**	City	101848.2	3422.73	2620
					n1	Lasonder, Zeggelt	46675.39	1568.58	1240
					n2	De Laares	47567.83	1598.57	1530
					n3	De Bothoven	116643.5	3919.94	5720
					n4	Hogeland-Noord	72674.53	2442.31	2610
					n5	Getfert	120440	4047.52	4020
					n6	Veldkamp-Getfert-West	63229.66	2124.9	1920
					n7	Horstlanden-Stadsweide	73700.28	2476.78	2570
					n8	Boddenkamp	32990.07	1108.67	480

The three previous works have done for a common study region, but by utilizing different types of dataset. In order to evaluate the performance of the binary dasymetric method and assess the efficiency of the related input data, three statistical analyses (MAE, RMSE and RMSE coefficient variation) are applied to the calculated results. The outputs for each procedure are shown in Table 4.4.

Regarding the execution of binary dasymetric method, it performs mostly the same for the TOP10 and cadastral datasets whereas the coefficient variations RMSEs for both of the datasets are 25% and 28%, respectively. Concerning the assessment of the right scale for the ancillary data, at this moment it is impossible to surely talk about the robustness of one against the other since the statistical results are mostly the same.

There is a difference between the MAE and RMSE values in the following table. As it is mentioned before, these disagreements are due to the individual errors. One type of this error can be seen in target zone code n0 from the tables 4.1, 4.2 and 4.3, where the calculated values as target zone population vary largely from the true value.

Table 4.4. Statistical analysis for implementation binary dasymetric method for central district source zone

Source zone code	Dasymetric method	Ancillary data	Population per source zone	Dasymetric density	Source zone area (m^2)	MAE	RMSE	CV(RMSE)%
D0	Binary	TOP10-Buit-up area	22710	0.02	1113713.3	**1036.1**	**1380.3**	**±55**
D0	Binary	TOP10-residential area	22710	0.055	414579.6	**532.8**	**623.1**	**±25**
D0	Binary	Cadastral dataset-residential area	22710	0.033	675769.43	**458.0**	**704.9**	**±28**

The binary dasymetric method is also examined for the whole of the city by utilizing cadastral dataset in order to test its efficiency for disaggregating the population and also to be compared by 3-class dasymetric method. The result of this section is summarized in Appendix II.

Performance of 3-class dasymetric method

The 3-class dasymetric method is tested for the complete study area to evaluate its performance in disaggregating census population data by utilizing different ancillary class variables (low-density, high-density and non-urban); and compare its robustness with binary dasymetric method. The result of disaggregated census populations through the target zone levels are shown in Appendix III. Also, a statistical comparison between the performance of binary and 3-class methods is summarized in Table 4.5.

Table 4.5. Statistical analysis for implementation binary dasymetric method for complete city

Source zone code	Dasymetric method	Ancillary data	MAE	RMSE	CV(RMSE)%
Complete city	**Binary**	**Cadastral dataset- residential area**	**1036.1**	**505.3**	**±23**
Complete city	**3-class**	**Cadastral dataset- residential area**	**532.8**	**1009**	**±45**

What is remarkable is the great error value for 3-class dasymetric method in contrast to the binary dasymetric method in all of the statistical analyses. Figure 4.1 can be considered to survey the detail RMS error propagations of these two methods within each source zone, respectively.

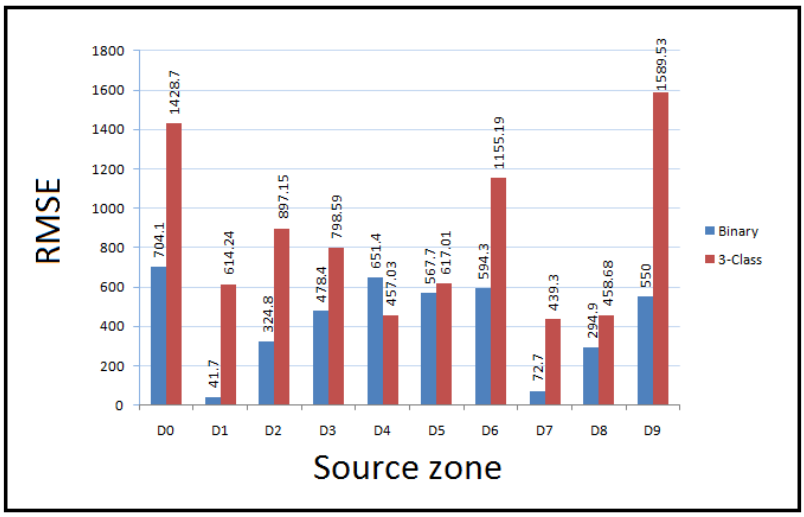


Figure 4.1. The RMSE values for binary and 3-class dasymetric methods by using 2D data

Let's investigate on the probable error sources. As a principal concept, the 3-class dasymetric method makes use of a series of land cover classes in order to distribute the population more nearer to the reality. It is important to know how these classes must be created. What is done in this study is defining a floor threshold (upper and lower than 3 floors) for classifying the study area into high and low-density classes; where by this action, around 90% of the study area is labelled as Low-density class versus 10% as high-density class. This issue can be one reason to less efficiency of 3-class versus binary dasymetric methods as it is pointed out before Churchman (1999) that, there is no necessity that high-rise buildings have higher residential density.

Another probable reason to this conceptual mismatch between what is concluded here and what is found in the literature regarding the effectiveness of 3-class dasymetric method, is the density ratio of each land cover class. This value is calculated based on the least square method by using ten equations and two unknowns. Although, when the number of observations is higher than unknown values, this method calculates the unknowns by adjustment analysis to get more accurate result, still some differences exist between what is observed in reality and what is calculated. As a numerical example, for the central district, consider the low-density ratio (D_L) as 0.011 ($^{person}/_{m^2}$), the high-density ratio (D_H) as 0.012 ($^{person}/_{m^2}$), the area of populated of low-density regions (A_L) as 1352671.8 m^2 and the area of high-density populated regions (A_H) as 1091962.7 m^2 values. Referring to equation (3.5) population per source zone is calculated as follows:

$$P_{s1} = (0.011 * 1352671.8) + (0.012 * 1091962.7)$$

Parameter P_{s1} is calculated as 27347.4 persons for central district, while it is 22710 in reality. This issue is also true for the other districts.

4.3. Evaluation of 3D ancillary data in dasymetric methods

The effectiveness of 3D data in dasymetric mapping will be assessed by using the height and the number of floors of dwelling units. The dasymetric methods are applied again to the central district and whole of the study area with the difference that, the factor *number of floors* (as integer) is added to dasymetric equations. With respect to the modified binary (eq. 3.7) and 3-class dasymetric equations (eq. 3.8), the census population is disaggregated. The related results are summarize in Appendices IV to VIII.

The statistical analyses are applied for the estimated population based on the actual values. What can be inferred from Table 4.6 is the role of residential areas versus built-up regions for estimating the population which improves the result (the same as what is concluded during the use of 2D data). This issue can be seen in the statistical values of TOP10 dataset where the built-up and residential areas in the central district are compared (64% vs. 19% CVRMSE values). Furthermore, concerning the right scale of ancillary data which was inconclusive during the use of 2D data, the execution of TOP10 dataset leads to better result in comparison to cadastral dataset. This is highlighted from the MAE and RMSE values of the residential TOP10 and the cadastral datasets. In addition to the statistical analysis interpretations, the values for the binary and 3-class dasymetric methods are considerable. 18% versus 33% are the coefficient variation of RMSE of binary and 3-class methods which represent the amount of uncertainty of each of them respectively. The same as what is concluded regarding the efficiency of

dasymetric methods in previous part (2D ancillary data section), the 3-class method generates more error in the course of population estimation than the binary method.

Table 4.6. Statistical analysis for implementation dasymetric methods by using number of floor parameter

Source zone code	Dasymetric method	Ancillary data	MAE	RMSE	CV(RMSE)%
Central district	**Binary**	**TOP10-Buit-up area**	**1243.0**	**1610.0**	**±64**
Central district	**Binary**	**TOP10- residential area**	**371.1**	**476.4**	**±19**
Central district	**Binary**	**Cadastral dataset- residential area**	**445.8**	**580.95**	**±23**
Complete city	**Binary**	**Cadastral dataset- residential area**	**310.0**	**419.4**	**±18**
Complete city	**3-class**	**Cadastral dataset- residential area**	**541.15**	**736.6**	**±33**

One of the main objectives of this research is to evaluate the added values of 3-Dimensional data in the process of population disaggregation. In this context, the error values from the binary and 3-class dasymetric are compared separately by utilizing 2D and 3D ancillary data (it is meant by 2D the footprints and 3D the floor areas).

To survey more the error propagation in the study area, the RMSE values are examined whereas the ranges of RMES and MAE vary similarly. Figure 4.2 and 4.3 are related to the error distribution of binary and Figure 4.4 and 4.5 represent the error propagation of 3-class dasymetric methods by contributing 2D and 3D ancillary data.

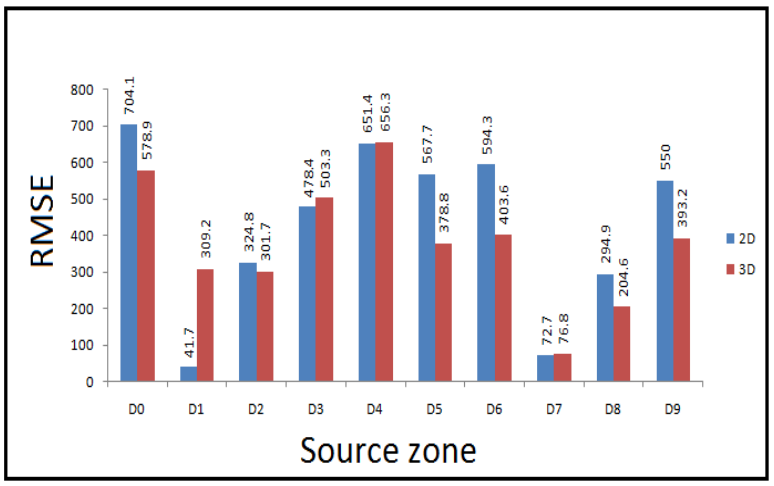


Figure 4.2.The RMSE values for each individual source zone from binary dasymetric method

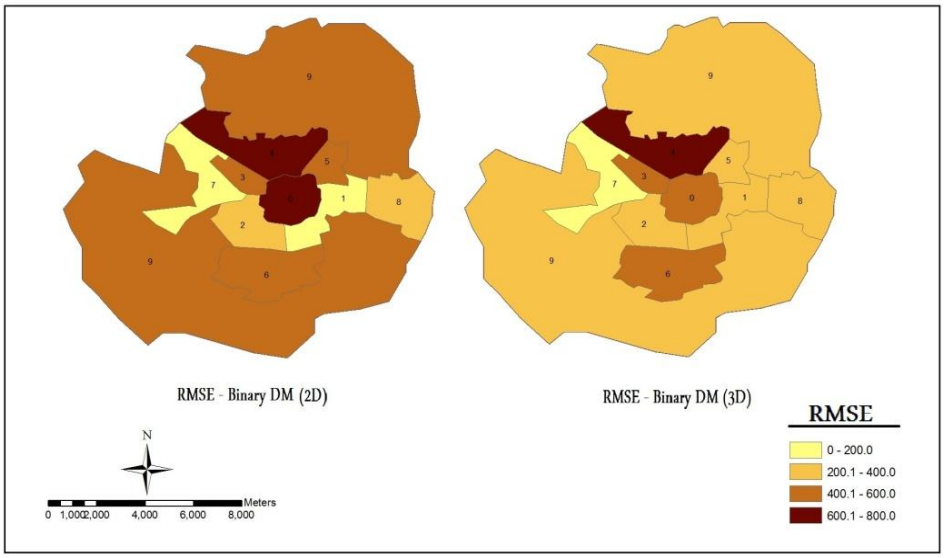


Figure 4.3. The error distribution within the study area related to the binary dasymetric method by using 2D & 3D ancillary data

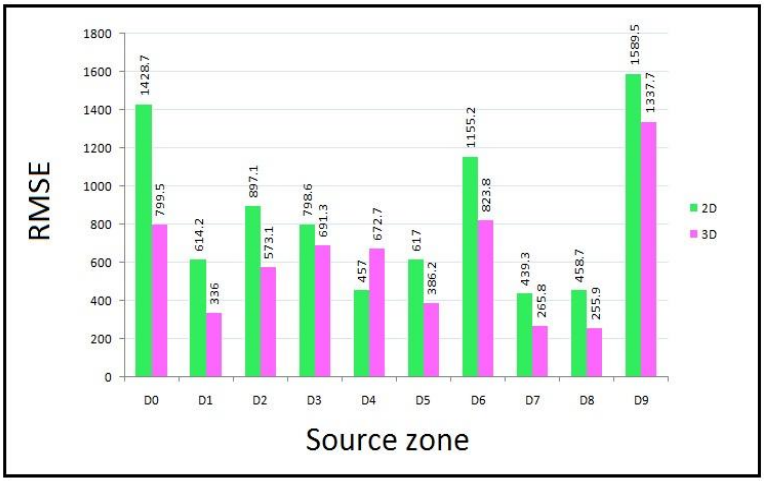


Figure 4.4. The RMSE values for each individual source zone from 3-class dasymetric method

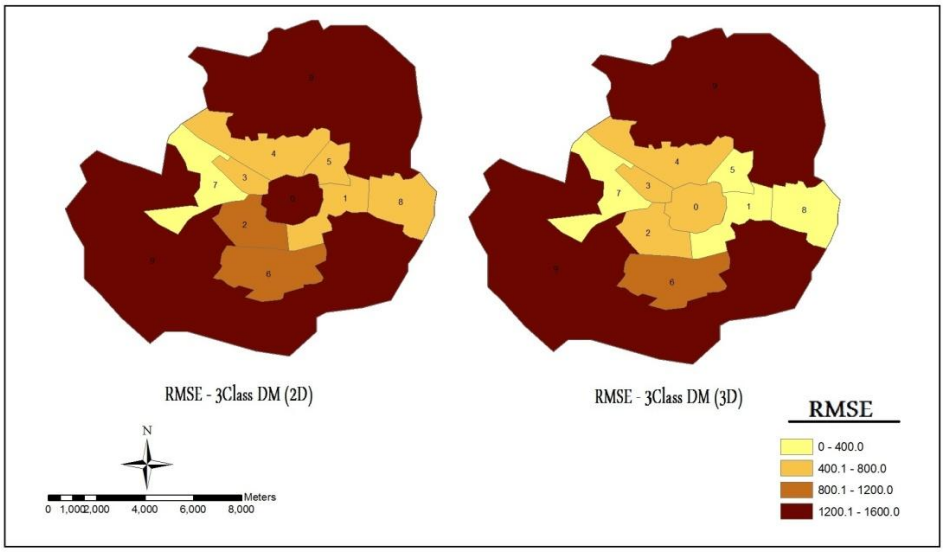


Figure 4.5. The error distribution related to 3-class dasymetric method by using 2D & 3D ancillary data

What can be inferred from the above depictions and statistics is that, more errors are propagated in the regions where still seems not to be purely inhabited. Figure 4.6 compares two districts (0 and 1) which mostly have the highest and lowest RMSE values in four previous depictions respectively. The idea is that, based on the building distribution patterns, it can be assumed that "District 1" has more inhabited buildings than "District 0". In other words, the "District 0" still suffers from the impurity of dwelling units versus district one.

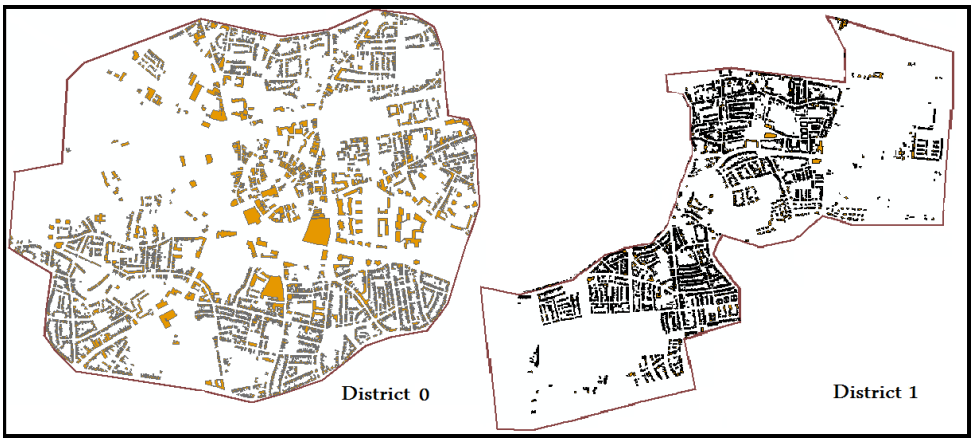


Figure 4.6. The distribution of residential buildings

Another issue that introduces errors in aforementioned processes is related to the application of each building. Especially in the central district (District 0) there are several multi-use buildings in which for example, the first floor has non-residential/commercial application, while other floors are residential (Figure 4.7).

Figure 4.7. A multi-application building

Meanwhile, if the dataset are not spatially and temporally compatible there would be mismatch between them. This makes more sense in buildings where they constructed and demolished after a while.

In addition to the previous results, in order to outline the extra value of 3D data in comparison to 2D data, a compilation of all the generated errors are shown in Figure 4.8. In most of the cases, floor areas

(3D) are effective in disaggregating population and perform much better than when only the footprint areas (2D) were put into practice.

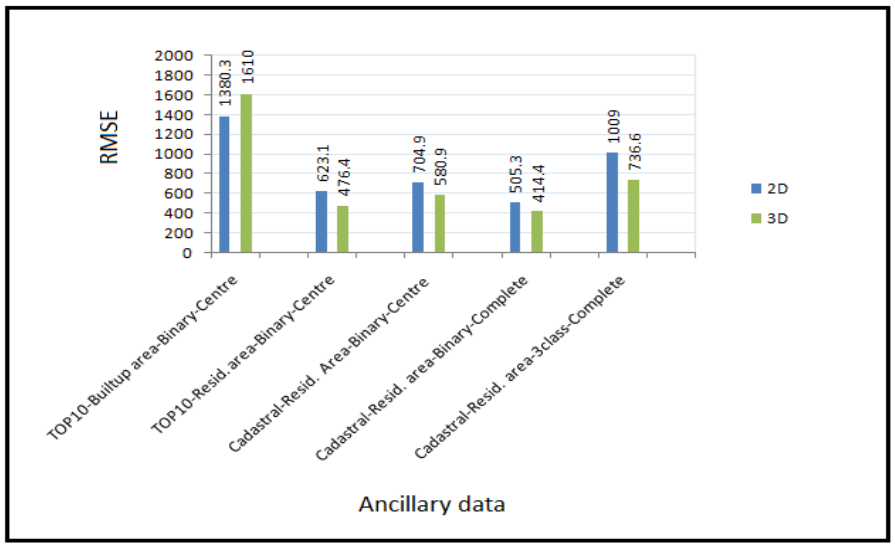


Figure 4.8. RMSE values of using 2D and 3D ancillary data

4.4. Evaluation of postcode 6-digit values

In this research, postcode 6-digit values are utilized for contributing in disaggregating population. In this way, the number of postcode addresses and also the number of postcodes within each source zone take into consideration as follows.

4.4.1. Implementing the postcode 6-digit addresses

The number of postcode 6-digit addresses within each source zone (district) and target zone (neighbourhood) are considered as ancillary data to contribute in disaggregating census population. In the binary dasymetric method (eq. 3.1), instead of administrative level areas, the related number of postcode 6-digit addresses is implemented. Table 4.7 summarizes the result for central district of the study area.

Table 4.7. Implementing binary dasymetric method by utilizing postcode 6-digit addresses

Source zone code	Source zone (District)	Population per source zone	Target zone code	Target zone (Neighborhood)	Postcode numbers	Postcode ratio	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	n0	City	2943	0.17	3842.67	2620
			n1	Lasonder, Zeggelt	938	0.05	1224.74	1240
			n2	De Laares	1131	0.07	1476.74	1530
			n3	De Bothoven	4238	0.24	5533.55	5720
			n4	Hogeland-Noord	1759	0.1	2296.72	2610
			n5	Getfert	2401	0.14	3134.98	4020
			n6	Veldkamp-Getfert-West	1236	0.07	1613.84	1920
			n7	Horstlanden-Stadsweide	2181	0.13	2847.73	2570
			n8	Boddenkamp	566	0.03	739.02	480

4.4.2. Implementing the number of postcode 6-digit units

Another type of information that is utilized in the process of population estimation is the number of postcode units (or points) within each source and target zones. The estimated population via this method can be seen in Table 4.8.

Table 4.8. Implementing binary dasymetric method by utilizing postcode 6-digit units

Source zone code	Source zone (District)	Population per source zone	Target zone code	Target zone (Neighborhood)	Postcode unit numbers	Postcode ratio	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	**n0**	City	94	0.15	3432.06	2620
			n1	Lasonder, Zeggelt	43	0.07	1569.98	1240
			n2	De Laares	42	0.07	1533.47	1530
			n3	De Bothoven	135	0.22	4929.02	5720
			n4	Hogeland-Noord	68	0.11	2482.77	2610
			n5	Getfert	95	0.15	3468.57	4020
			n6	Veldkamp-Getfert-West	55	0.09	2008.12	1920
			n7	Horstlanden-Stadsweide	71	0.11	2592.3	2570
			n8	Boddenkamp	19	0.03	693.71	480

What can be inferred from these results is the improvement of the estimated population values to the actual values in most of the target zone except the central district which also suffers from overestimation of population. Table 4.9 summarizes the statistical analyses for the aforementioned postcode variables. A slight outperform of postcodes units to the postcode addresses and an incredible improvement to the other ancillary data can be due to minimum relationship of postcode to the area.

Table 4.9. Statistical analysis for implementation binary dasymetric method for complete city

Source zone code	Dasymetric method	Ancillary data	MAE	RMSE	CV(RMSE)%
Central district	**Binary**	**PC6-addresses**	**391**	**542.84**	**±21**
Central district	**Binary**	**PC6-units**	**326.6**	**443.24**	**±17**

What can be concluded here is that, the dasymetric methods are largely sensitive to the ancillary data which are using. The area is more flexible than the number of postcodes even in a short interval.
For a general overview of the accuracy of dasymetric methods and their relative ancillary data, Figure 4.7 is created.

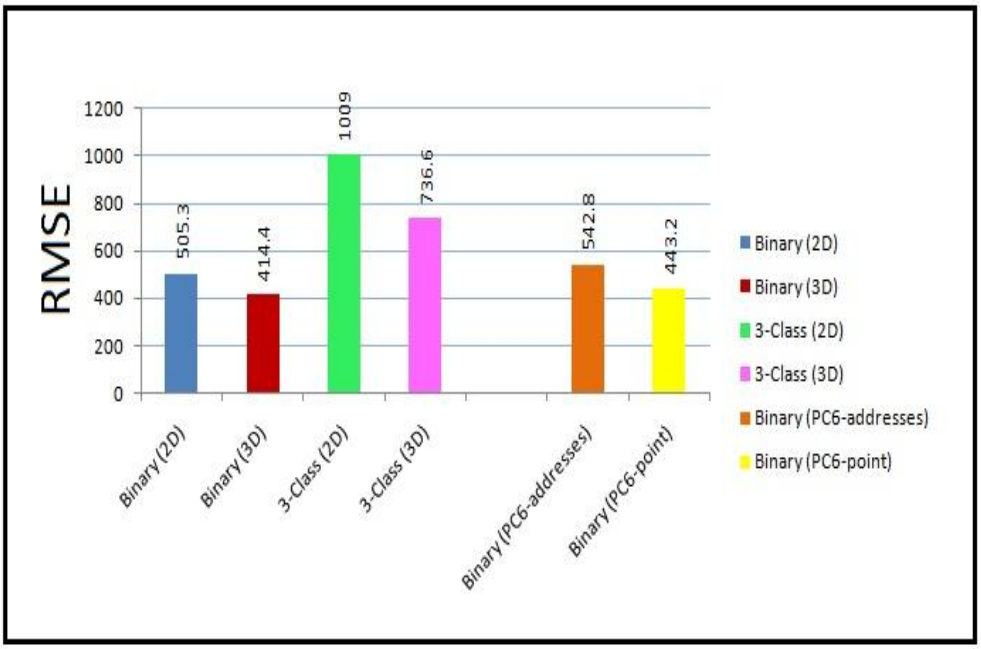


Figure 4.9. An evaluation of dasymetric methods by utilizing various types of ancillary data

All in all, the finest population disaggregation of the city of Enschede can be achieved via binary dasymetric method by contributing either 3-Dimensional TOP10 areal dataset or postcode 6-digit units with the approximate accuracy of 17% per person.

As a general comparison between the performance of dasymetric methods in this study and the previous researches, based on Figure 4.10, it is inconclusive to surely outline the robustness of a specific dasymetric method. What is observable is that, the 3-class method performs better in SEQ study region with lower population density, but is realized as accurate as the binary method in United Kingdom study area (Leicestershire), while it does not performs so well in Enschede with the higher average density value. This variation is largely based on the use of ancillary data types and the assumptions which are considering in the process of population disaggregation.

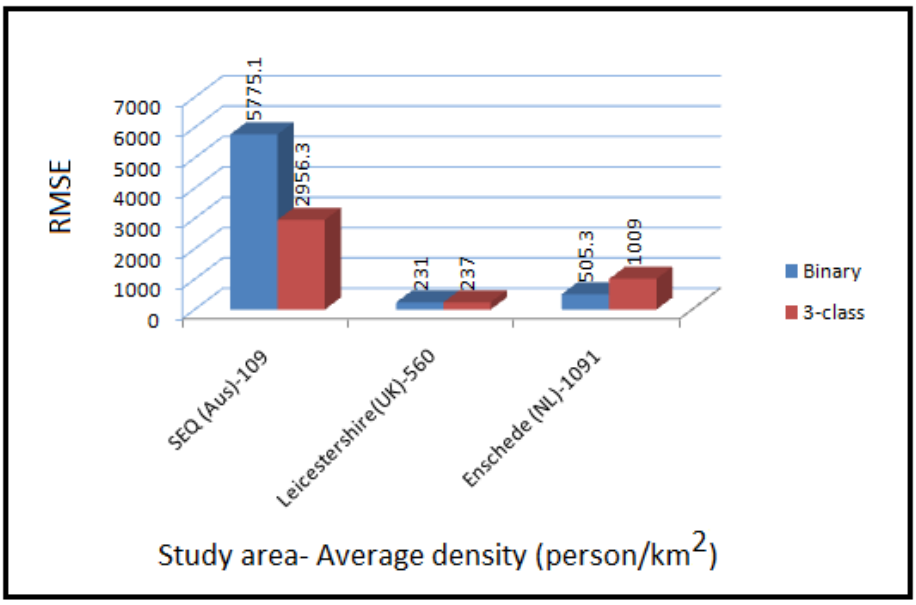


Figure 4.10. A comparison between the RMSE values of implementing dasymetric methods for three study areas

5. Conclusion and Recommendation

5.1. Conclusion

The main objective of this research was disaggregating census population data into the finest administrative level. In this context, dasymetric methods which use ancillary information for transforming data from one spatial unit to another were examined. Several types of ancillary data (2D, 3D and postcodes) were used to address their importance in the process of population disaggregation. The results of these implementations will be summarized with respect to the sub-objectives of this research as follows:

- **Evaluation of current dasymetric mapping methods**
  Dasymetric methods which are a type of areal interpolation techniques applied for disaggregating coarse census population data from the district (as the source) level to the finest possible scale which is neighbourhood (as the target) level, since they utilize ancillary data in the process of population estimation. The binary dasymetric method was found to be a simple and easy to implement in GIS method, while the 3-class dasymetric method was recognized as more complex one, whereas it makes use of a number of ancillary variables for better estimation of population. The result regarding the accuracy and efficiency of dasymetric methods contradicts previous studies which were emphasized on the better performance of 3-class method versus binary method in disaggregating census population. The reason was realized to be the sensitivity of 3-class dasymetric method to the density assumption and the quality of ancillary data. The density threshold for partitioning the populated area to low and high residential classes was defined as 3-stoory buildings. As an outline, there is no necessity that the higher buildings represent higher density values. Regarding the quality of the ancillary data, in general, spatial and temporal compatibility of the input dataset for better representing the internal structure of population distribution inside the study area were considerable.

- **Assessing the "right scale" of the ancillary datasets**
  The term "right scale" may be interpreted differently based on the intended goals and objectives. In this research, generally, it was meant as the accuracy and accessibility criteria of the input data and specifically, related to the ancillary data which had higher compatibility with source zones to better reveal the internal distribution of population within these zones. In this context, first, the attempts for defining the finest administrative units to be considered as source and target zones started by creating postcode 6-digit polygons inside postcode 4-digit boundaries which led to unsatisfactory result due to the mismatch with reality and other datasets; in addition, two larger scale areas, the district and neighbourhood levels were dedicated as source and target zone administrative zones which did not have the pervious shortages. Second, the TOP10 and the cadastral datasets which were the footprints of buildings were examined as the areal units for implementing in dasymetric method; where the employed TOP10 led to better statistical result in transferring the census data from a large zone to several sub-zones in all the cases in comparison to the cadastral dataset. Finally, a new type of ancillary data –postcode 6-digit addresses and units– were demonstrated as a

replacement to the area variable in binary dasymetric equation. Based on the previous section and what is discussed above, the finest accuracy of population disaggregation for this small study area with high population density was calculated as 17% by implementing the binary dasymetric method across either the 3-dimensional TOP10 areal dataset or the postcode 6-digit units.

- **Assessing the added value of 3D ancillary data**
  3-Dimentional ancillary data were related to the height of buildings and their corresponding number of floors. Instead of the footprint area parameter in dasymetric methods, the floor areas variable was utilized in the process of population disaggregation which mapped the population to their corresponding floor unit. This issue considerably improved the performance of dasymetric methods and led to decrease the errors values in population computing. The statistical analyses of binary dasymetric showed 23% error value for 2D data versus 18% error value of 3D ancillary data; also, regarding the 3-class method, 45% versus 33% population variation estimation with respect to the use of 2D and 3D ancillary data.

- **Identifying the best way to evaluate the results**
  In this research, two approaches were investigated for assessing the results of disaggregation methods. First, a series of statistical analysis methods are tested: the mean absolute error (MAE), the root mean square error (RMSE) and a standardized coefficient of variation CV(RMSE) are chosen because they can be applied to count data (e.g. population number) and also the values can be easily interpreted relatively. The MAE and RMSE are both measured in the same unit as the original data and they can be used to diagnose the variation in the errors. The disagreement between MAE and RMSE relates to the differences in individual errors and if no difference exists, it means all the errors are off the same magnitude. In addition, the coefficient of variation of the RMSE (in short CV(RMSE)) can address the amount of uncertainty of each method. Second approach for evaluating the result was, generating error maps to visually see and compare the error propagation through the study area.

5.2. Limitations

This research attempted to be generic where the result could be generated to other study areas. The population disaggregation process is found as a data-oriented work; where the detailed ancillary data, the detailed population estimation. Furthermore, the data should be spatially and temporally compatible with each other.

5.3. Recommendation

Various suggestions can be proposed based on the utilized dasymetric methods and their corresponding ancillary data for future studies as follows:

- Some actions can be taken based on the result of postcode 6-digit polygons. These boundaries can either be improved by modified Voronoi diagram approach or be generated by another method to delineate the territory of each postcode 6-digit point.

- This research suggests some follow-ups in the context of utilizing postcode information or other type of ancillary data except the area unit for mapping population from a large resolution area to small resolution level. So far it is found out that, postcode 6-digit addresses and units can be applied to binary dasymetric method and can be contributed in population estimation process; the next step can be testing other factors of postcodes in different dasymetric methods.

- An investigation can be carried out on how 3-Dimentional data behaves with other areal interpolation methods which use ancillary data.

- Since the 3-class dasymetric method is based on the classification of residential densities, improvement can be done by defining other criterion/criteria of density allocation in the study area (as an alternative to 3-stoory threshold in this research) based on the shape, size, census houses statistics and different socio-economic factors such as wealth, income, age, sex, etc.

- Since the 3-class dasymetric method is based on the classification of residential densities, improvement can be done by incorporating multi-density classes (e.g. very low, low, medium, high and very high-density) for solving the spatial heterogeneity in population disaggregation.

References

Achilleos, G. (2008). "Errors within the Inverse Distance Weighted (IDW) interpolation procedure." Geocarto International **23**(6): 429-449.

Bielecka, E. (2005). "A Dasymetric population density map of Poland." In proceeding of the 22nd International Cartographic Conference, July 9-15, A Coruna, Spain (CD).

Brassel, K.E. and D. Reif (1979). "Procedure to generate Thiessen polygons." In: Geographical Analysis **11**(3): 289-303.

Briggs, D. J., J. Gulliver, et al. (2007). "Dasymetric modelling of small-area population distribution using land cover and light emission data." Remote Sensing of Environment **108**(4): 451-466.

Churchman, A. (1999). "Disentangling the Concept of Density" Journal of planning literature **13**: 389-410.

Clayton, C. and J. Estes (1980). "Image Analysis as a Check on Census Enumeration Accuracy." Photogrammetric Engineering and Remote Sensing **46**: 757-764.

Donnay, J. P. and D. Unwin (2001). "Modeling Geographical Distributions in Urban Areas." In Remote Sensing and Urban Analysis, Donnay, J. P., Barnsley, M. J., and P. A. Longley (Eds.), New York, NY: Taylor and Francis, 205-224.

Eicher, C. L. and C. A. Brewer (2001). "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation." Cartography and Geographic Information Science **28**(2): 38-125.

Fisher, P. F. and M. Langford (1995). "Modeling the Errors in Areal Interpolation between Zonal Systems by Monte Carlo Simulation." Environment and Planning A **27**: 211-224.

Fisher, P. F. and M. Langford (1996). "Modelling sensitivity to accuracy in classified imagery: A case study of areal interpolation." The professional geographer **48**: 299-309.

Freire, S. M. C. (2007). "Modelling Daytime and Nighttime Population Distribution in Portugal Using Geographic information System." Master of Art thesis, University of Kansas, Department of Geography: 8-30.

Goodchild, M. F. and N S-N. Lam (1980). "Areal interpolation: a variant of the traditional spatial problem." Geoprocessing (1): 297-312.

Gregory, I. N. and S. EII. Paul (2005). "Breaking the boundaries, geographical approaches to integrating 200 years of the census." Journal of the Royal Statistic Society **168**: 419- 437.

Harvey, J. T. (2002). "Population estimation models based on individual TM pixels." Photogrammetric Engineering and Remote Sensing **68**(11): 1181-1192.

Holt, J. B., C. P. Lo, and T. W. Hodler (2004). "Dasymetric estimation of population density and areal interpolation of census data." Cartography and Geographic Information Science **31**(2): 103-121.

Kraus, S. P., L. W. Senger, and J. M. Ryerson (1974). "Estimating Population from Photographically Determined Residential Land Use Types." Remote Sensing of Environment **3**(1): 35-42.

Lam, N. S. (1983). "Spatial Interpolation Methods: A Review." The American Cartographer **10**(2): 129-149.

Langford, M. (2006). "Obtaining population estimations in non-census reporting zones, an evaluation of the 3-class dasymetric method." Computers, Environment & Urban Systems **30**:161-180.

Langford, M., D. J. Maguire, and D. J. Unwin (1991). "The areal interpolation problem: estimating population using remote sensing in a GIS framework." In Handling Geographical Information: Methodology and potential applications Eds I Masser, M Blakemore (Longman, Harlow, Essex): 55-77.

Langford, M. and D. J. Unwin (1994). "Generating and Mapping Population-Density Surfaces within a Geographical Information-System." Cartographic Journal **31**(1): 21-26.

Langford, M. and G. Higgs (2006). "Measuring Potential Access to Primary Healthcare Services: The Influence of Alternative Spatial Representations of Population." The Professional Geographer **58**(3): 294-306.

Li, T., D. Pullar, et al. (2007). "A comparison of spatial disaggregation techniques as applied to population estimation for South East Queensland (SEQ), Australia." Applied GIS **3**(9): 1-16.

Liao, Y., J. Wang, et al. (2010). "Integration of GP and GA for mapping population distribution." International Journal of Geographical Information Science **24**(1): 47-67.

Liu, X. (2003). "Estimation of the Spatial Distribution of Urban Population Using High Spatial Resolution Satellite Imagery." Ph.D. thesis, University of California, Santa Barbara, 175 p.

Liu, X., K. C. Clarke, and M. Herold (2006). "Population density and image texture: A comparison study." Photogrammetric Engineering & Remote Sensing **72**(2): 187-96.

Lo, C. P. (1986). "Applied Remote Sensing" New York, NY: Longman, 393 p.

Mennis, J. (2003). "Generating surface models of population using dasymetric mapping." The Professional Geographer **55**(1): 31-42.

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

Sutton, P., D. Roberts, et al. (1997). "A comparison of nighttime satellite imagery and population density for the continental United States." Photogrammetric Engineering and Remote Sensing **63**: 1303-1313.

Thiessen, A. H. and J. C. Alter (1911). "Climatological data for July: District No. 10, Great Basin." Mon. Wea. Rev. **39**: 1082-1089.

Weng, Q. (2009). "Remote Sensing and GIS Integration: Theories, Methods, and Applications." New York, McGraw-Hill: 295-322.

Wright, J. K. (1936). "A method of mapping densities of population with Cape Cod as an example." Geographical Review **26**: 103-110.

Wu, S., X. Qiu, and L. Wang (2005). "Population estimation methods in GIS and remote sensing: a review." GIScience & Remote Sensing **42**(1): 58-74.

Wu, S., L. Qiu, and X. Wang (2006). "Using semi-variance image texture statistics model population densities." Cartography and Geographic Information Science **33**(2): 127-140.

Wu, S., L. Wang, and X. Qiu (2008). "Incorporating GIS building data and census housing statistics for sub-block-level population estimation." Professional Geographers **60**(1): 121-135.

Yuan, Y., R. Smith, et al. (1997). "Remodeling Census Population with Spatial Information from LandSat TM imagery." Computers, Environment and Urban Systems **21**(3-4): 245-258.

URL1. http://www.geodan.com/products/geographic-data-catalogue/products/6-digit-postcode-data-set-the-netherlands/, last access 1 Oct. 2010.

URL2. http://www.enotes.com/topic/Postal_code, last access 1 Oct. 2010.

URL3. http://www.ahn.nl , last access 15 August 2010.

URL4. http://www.haus-und-grund-bayern.de/index.php?id=301&tx_ttnews[pS]=1203003392&tx_ttnews[tt_news]=555&tx_ttnews[backPid]=300&cHash=dac2fc8845, last access 24 Oct. 2010

URL5.http://www.vorm.nl/get.asp?file=Docs/bouwen_en_wonen/Bouwbesluit_integrale_tekst_jan2008.pdf, last access 13 June 2010.

Appendices

Appendix I. Union script in python

```
#   (generated by ArcGIS/ModelBuilder)

#  Import system modules

import sys, string, os, arcgisscripting

# Create the Geoprocessor object

gp = arcgisscripting.create()

# Load required toolboxes...

gp.AddToolbox("C:/Program Files (x86)/ArcGIS/ArcToolbox/Toolboxes/Analysis Tools.tbx")

# Set the Geoprocessing environment...

gp.workspace = "I:\\DATA (sample)\\Final5"

# Local variables...

XYExport_population_XY_Selec36 = "I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec36"

XYExport_population_XY_Selec35 = "I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec35"

XYExport_population_XY_Selec34 = "I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec34"

XYExport_population_XY_Selec33 = "I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec33"

print XYExport_population_XY_Selec33

# Process: Union...

gp.Union_analysis("'I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec35' #;'I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec34' #;'I:\\DATA
(sample)\\Final4\\My_Geodatabase.gdb\\test3\\XYExport_population_XY_Selec33' #",
XYExport_population_XY_Selec36, "ALL", "", "GAPS")
```

Appendix II. Binary method by utilizing residential areas of cadastral dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (sq m)	Target zone code	Target zone (Neighbourhood)	Target zone area (sq m)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	**0.034**	676156.55	**n0**	City	101848.22	3420.8	2620
					n1	Lasonder, Zeggelt	46619.03	1565.8	1240
					n2	De Laares	47958.52	1610.8	1530
					n3	De Bothoven	116628.98	3917.2	5720
					n4	Hogeland-Noord	73568.14	2470.9	2610
					n5	Getfert	119546.34	4015.2	4020
					n6	Veldkamp-Getfert-West	62920.49	2113.3	1920
					n7	Horstlanden-Stadsweide	74076.75	2488	2570
					n8	Boddenkamp	32990.07	1108	480
D1	Hogeland - Velve	12190	**0.031**	388846.22	**n9**	Velve-Lindenhof	125966.78	3949	4520
					n10	Wooldrik	55731.32	1747.1	1230
					n11	Hogeland-Zuid	72864.15	2284.2	2390
					n12	Varvik-Diekman	105604.73	3310.6	3490
					n13	Sleutelkamp	15032.49	471.3	440
					n14	't Weldink	6379.12	200	20
					n15	De Leuriks	7267.63	227.8	100
D2	Boswinkel - Stadsveld	23440	**0.039**	600993.69	**n16**	Cromhoffsbleek-Kotman	52091.25	2031.7	2280
					n17	Boswinkel-De Braker	90226.87	3519	3960
					n18	Pathmos	54251.71	2115.9	2120
					n19	Stevenfenne	115663.1	4511.1	4780
					n20	Stadsveld-Zuid	40932.16	1596.4	1770
					n21	Elferink-Heuwkamp	75664.76	2951.1	2750
					n22	Stadsveld-Noord-Bruggert	48636.03	1896.9	1850
					n23	't Zwering	77732.27	3031.7	2310
					n24	Ruwenbos	45795.54	1786.1	1620
D3	Twekkelerveld	9080	**0.035**	258562.45	**n25**	Tubantia-Toekomst	154578.26	5428.4	4950
					n26	Twekkelerveld	103984.18	3651.6	4130
D4	Enschede-Noord	18550	**0.035**	535638.35	**n27**	Walhof-Roessingh	74587.71	2583.1	2450
					n28	Bolhaar	77837.81	2695.6	1620
					n29	Roombeek-Roomveldje	118381.73	4099.7	3790
					n30	Mekkelholt	55300.99	1915.2	2250
					n31	Deppenbroek	98681.86	3417.5	4670
					n32	Voortman-Amelink	33042.79	1144.3	1200
					n33	Drienerveld-U.T.	77805.47	2694.5	2570
D5	Ribbelt - Stokhorst	8740	**0.029**	302881.08	**n34**	Schreurserve	80687.19	2328.3	2410
					n35	Ribbelt-Ribbelerbrink	54730.39	1579.3	1940
					n36	Park Stokhorst	100674.14	2905.1	3410
					n37	Stokhorst	66789.37	1927.3	980
D6	Enschede-Zuid	35700	0.041	877882.36	**n38**	Stroinkslanden Noord-Oost	110341.46	4487.2	3520
					n39	Stroinkslanden-Zuid	109714.85	4461.7	4900
					n40	Stroinkslanden Noord-West	65087.79	2646.9	2360
					n41	Wesselerbrink Noord-Oost	92447.93	3759.5	3970
					n42	Wesselerbrink Zuid-Oost	94171.88	3829.6	4520
					n43	Wesselerbrink Zuid-West	49582.8	2016.3	2530
					n44	Wesselerbrink Noord-West	116996.68	4757.8	5480
					n45	Helmerhoek-Noord	126295.37	5135.9	4200
					n46	Helmerhoek-Zuid	97909.92	3981.6	3970
					n47	het Brunink	15333.68	623.6	250
D7	Bedrijfsterreinen	330	**0.005**	60563.11	**n48**	Industrie- en havengebied	28173.95	153.5	240
					n49	Marssteden	4463.27	24.3	70
					n50	Koekoeksbeekhoek	7045.08	38.4	10
					n51	De Broeierd	20880.82	113.8	10
D8	Glanerbrug en omgeving	16980	**0.032**	523528.68	**n52**	Glanerveld	44683.36	1449.2	1130
					n53	Bentveld-Bultserve	107996.2	3502.7	3050
					n54	Schipholt-Glanermaten	85739.74	2780.9	2970
					n55	Eekmaat	51626.33	1674.4	1690
					n56	Oikos	60021.75	1946.7	2410
					n57	Eilermarke	49359.29	1600.9	1770
					n58	De Slank	14211.8	460.9	170
					n59	Dolphia	20500.7	664.9	570
					n60	Eekmaat West	89389.52	2899.2	3220
D9	Landelijk gebied en kernen	8410	**0.01**	862196.46	**n61**	Dorp Lonneker	84971.34	828.8	1740
					n62	Dorp Boekelo	106680.86	1040.6	2090
					n63	Buurtschap Lonneker-West	178589.66	1742	1200
					n64	Noord Esmarke	64751.74	631.6	360
					n65	Buurtschap Zuid-Esmarke	37303.43	363.9	230
					n66	Buurtschap Broekheurne	159936.2	1560	1380
					n67	Buurtschap Usselo	41356.1	403.4	270
					n68	Boekelerveld	106035.37	1034.3	920
					n69	Buurtschap Twekkelo	82571.75	805.4	220

Appendix III. 3-class method by utilizing residential areas of cadastral dataset

Source zone code	Population per source zone	Dasymetric low-density	Dasymetric high-density	Source zone area (sq m)-class low	Source zone area (sq m)-class high	Target zone code	Target zone area (sq m)-class low	Target zone area (sq m)-class high	Population per target zone	Population in reality
D0	22710	0.0234	0.1052	512596.06	163560.48	n0	51462.51	50385.71	**6508.29**	2620
						n1	38797.85	7821.18	**1732.99**	1240
						n2	47402.68	555.84	**1170.45**	1530
						n3	66705.77	49923.2	**6817.2**	5720
						n4	66526.79	7041.36	**2301.41**	2610
						n5	94833.74	24712.6	**4824.62**	4020
						n6	59759.82	3160.66	**1734.38**	1920
						n7	61953.08	12123.67	**2728.83**	2570
						n8	25153.81	7836.26	**1414.51**	480
D1	12190	0.0234	0.1052	378232.47	10613.76	n9	122477.87	3488.9	**3240.15**	4520
						n10	53852.88	1878.44	**1460.91**	1230
						n11	70352.58	2511.58	**1914.57**	2390
						n12	103400.69	2204.04	**2657.46**	3490
						n13	14801.8	230.69	**371.49**	440
						n14	6379.12	0	**149.64**	20
						n15	6967.52	300.11	**195.02**	100
D2	23440	0.0234	0.1052	534151.11	66842.59	n16	23003.93	29087.32	**3599.9**	2280
						n17	79447.99	10778.88	**2997.74**	3960
						n18	54251.71	0	**1272.64**	2120
						n19	111107.88	4555.22	**3085.62**	4780
						n20	36176.6	4755.56	**1348.96**	1770
						n21	71251.47	4413.29	**2135.74**	2750
						n22	42921.76	5714.28	**1608.06**	1850
						n23	72764.6	4967.67	**2229.56**	2310
						n24	43225.18	2570.36	**1013.98**	1620
D3	9080	0.0234	0.1052	239860.67	18701.78	n25	151055.86	3522.4	**3914.06**	4950
						n26	88804.81	15179.37	**3680.21**	4130
D4	18550	0.0234	0.1052	481862.97	53775.38	n27	70970.77	3616.94	**2045.37**	2450
						n28	77837.81	0	**1825.92**	1620
						n29	103082.35	15299.38	**4027.75**	3790
						n30	41371.97	13929.02	**2435.98**	2250
						n31	79587.15	19094.72	**3875.91**	4670
						n32	33042.79	0	**775.12**	1200
						n33	75970.14	1835.33	**1975.2**	2570
D5	8740	0.0234	0.1052	297010.66	5870.42	n34	79764.69	922.49	**1968.18**	2410
						n35	51438.8	3291.58	**1552.96**	1940
						n36	99017.8	1656.34	**2497.02**	3410
						n37	66789.37	0	**1566.74**	980
D6	35700	0.0234	0.1052	815561.6	62320.76	n38	108515.12	1826.34	**2737.7**	3520
						n39	104884.83	4830.02	**2968.55**	4900
						n40	65087.79	0	**1526.83**	2360
						n41	78741.4	13706.52	**3289.18**	3970
						n42	85263	8908.88	**2937.4**	4520
						n43	39119.18	10463.62	**2018.54**	2530
						n44	97190	19806.67	**4363.74**	5480
						n45	123516.66	2778.71	**3189.8**	4200
						n46	97909.92	0	**2296.77**	3970
						n47	15333.68	0	**359.7**	250
D7	330	0.0234	0.1052	57201.75	3361.36	n48	28173.95	0	**660.9**	240
						n49	4463.27	0	**104.7**	70
						n50	7045.08	0	**165.26**	10
						n51	17519.46	3361.36	**764.62**	10
D8	16980	0.0234	0.1052	495316.82	28211.87	n52	42792.97	1890.39	**1202.73**	1130
						n53	105475.5	2520.7	**2739.45**	3050
						n54	85298.58	441.16	**2047.35**	2970
						n55	45391.82	6234.52	**1720.73**	1690
						n56	58160.77	1860.98	**1560.13**	2410
						n57	47038.28	2321.01	**1347.62**	1770
						n58	14211.8	0	**333.38**	170
						n59	20500.7	0	**480.91**	570
						n60	76446.41	12943.12	**3155.03**	3220
D9	8410	0.0234	0.1052	860314.95	1881.51	n61	84971.34	0	**1993.26**	1740
						n62	105259.71	1421.15	**2618.7**	2090
						n63	178589.66	0	**4189.36**	1200
						n64	64751.74	0	**1518.95**	360
						n65	37303.43	0	**875.06**	230
						n66	159475.84	460.36	**3789.42**	1380
						n67	41356.1	0	**970.13**	270
						n68	106035.37	0	**2487.38**	920
						n69	82571.75	0	**1936.97**	220

Appendix IV. Binary method by utilizing built-up floor areas for central district of TOP10 dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (sq m)	Target zone code	Target zone (Neighbourhood)	Target zone area (sq m)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.008	2866580.98	**n0**	City	757287.12	5995.4	2620
					n1	Lasonder, Zeggelt	152629.1	1208.4	1240
					n2	De Laares	99666.64	789.1	1530
					n3	De Bothoven	478567.74	3788.8	5720
					n4	Hogeland-Noord	142316.54	1126.7	2610
					n5	Getfert	383821.72	3038.7	4020
					n6	Veldkamp-Getfert-West	187880.57	1487.5	1920
					n7	Horstlanden-Stadsweide	342245.82	2709.6	2570
					n8	Boddenkamp	322165.73	2550.6	480

Appendix V. Binary method by utilizing residential floor areas for central district of TOP10 dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (sq m)	Target zone code	Target zone (Neighborhood)	Target zone area (sq m)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.022	113167.79	**n0**	City	98671.53	2211.71	2620
					n1	Lasonder, Zeggelt	59839.56	1341.29	1240
					n2	De Laares	66367.99	1487.63	1530
					n3	De Bothoven	248624.28	5572.88	5720
					n4	Hogeland-Noord	104078.19	2332.9	2610
					n5	Getfert	143877.45	3224.99	4020
					n6	Veldkamp-Getfert-West	112937.57	2531.48	1920
					n7	Horstlanden-Stadsweide	118687.1	2660.35	2570
					n8	Boddenkamp	60084.12	1346.78	480

Appendix VI. Binary method by utilizing residential floor areas of central district of cadastral dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (sq m)	Target zone code	Target zone (Neighborhood)	Target zone area (sq m)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	0.009	2373543.89	**n0**	City	393141.31	3761.56	2620
					n1	Lasonder, Zeggelt	155503.38	1487.85	1240
					n2	De Laares	133018.15	1272.71	1530
					n3	De Bothoven	484489.91	4635.59	5720
					n4	Hogeland-Noord	251929.53	2410.45	2610
					n5	Getfert	431090.31	4124.66	4020
					n6	Veldkamp-Getfert-West	183654.81	1757.2	1920
					n7	Horstlanden-Stadsweide	237029.51	2267.89	2570
					n8	Boddenkamp	103686.98	992.07	480

Appendix VII. Binary dasymetric by utilizing residential floor areas of cadastral dataset

Source zone code	Source zone (District)	Population per source zone	Dasymetric density	Source zone area (sq m)	Target zone code	Target zone (Neighbourhood)	Target zone area (sq m)	Population per target zone	Population in reality
D0	Binnensingelgebied	22710	**0.01**	2374641.54	**n0**	City	393141.31	3759.83	2620
					n1	Lasonder, Zeggelt	155352.48	1485.72	1240
					n2	De Laares	134172.02	1283.16	1530
					n3	De Bothoven	484446.37	4633.03	5720
					n4	Hogeland-Noord	254608.89	2434.96	2610
					n5	Getfert	428410.95	4097.13	4020
					n6	Veldkamp-Getfert-West	182666.77	1746.94	1920
					n7	Horstlanden-Stadsweide	238155.76	2277.61	2570
					n8	Boddenkamp	103686.98	991.62	480
D1	Hogeland - Velve	12190	**0.011**	1073410.15	**n9**	Velve-Lindenhof	343173.62	3897.19	4520
					n10	Wooldrik	150899.07	1713.66	1230
					n11	Hogeland-Zuid	207193.39	2352.96	2390
					n12	Varvik-Diekman	298327.64	3387.91	3490
					n13	Sleutelkamp	39844.13	452.48	440
					n14	't Weldink	14232.43	161.63	20
					n15	De Leuriks	19739.86	224.17	100
D2	Boswinkel - Stadsveld	23440	**0.013**	1740128.35	**n16**	Cromhoffsbleek-Kotman	178160.29	2399.87	2280
					n17	Boswinkel-De Braker	266316.39	3587.35	3960
					n18	Pathmos	132924	1790.52	2120
					n19	Stevenfenne	333618.13	4493.93	4780
					n20	Stadsveld-Zuid	123167.15	1659.09	1770
					n21	Elferink-Heuwkamp	216987.14	2922.88	2750
					n22	Stadsveld-Noord-Bruggert	138670.39	1867.93	1850
					n23	't Zwering	219090.12	2951.2	2310
					n24	Ruwenbos	131194.74	1767.23	1620
D3	Twekkelerveld	9080	**0.013**	707083.8	**n25**	Tubantia-Toekomst	424664.74	5453.32	4950
					n26	Twekkelerveld	282419.05	3626.68	4130
D4	Enschede-Noord	18550	**0.013**	1468938.02	**n27**	Walhof-Roessingh	212697.15	2685.98	2450
					n28	Bolhaar	197105.97	2489.09	1620
					n29	Roombeek-Roomveldje	353129	4459.37	3790
					n30	Mekkelholt	189503.2	2393.08	2250
					n31	Deppenbroek	323788.61	4088.86	4670
					n32	Voortman-Amelink	81896.46	1034.2	1200
					n33	Drienerveld-U.T.	110817.62	1399.42	2570
D5	Ribbelt - Stokhorst	8740	**0.011**	773150.65	**n34**	Schreurserve	206267.49	2331.73	2410
					n35	Ribbelt-Ribbelerbrink	155961.92	1763.06	1940
					n36	Park Stokhorst	268539.8	3035.68	3410
					n37	Stokhorst	142381.45	1609.54	980
D6	Enschede-Zuid	35700	0.014	2595442.25	**n38**	Stroinkslanden Noord-Oost	300862.31	4138.33	3520
					n39	Stroinkslanden-Zuid	297787.08	4096.03	4900
					n40	Stroinkslanden Noord-West	185733.38	2554.74	2360
					n41	Wesselerbrink Noord-Oost	293707.85	4039.92	3970
					n42	Wesselerbrink Zuid-Oost	310106.17	4265.47	4520
					n43	Wesselerbrink Zuid-West	176967.23	2434.16	2530
					n44	Wesselerbrink Noord-West	393105.47	5407.12	5480
					n45	Helmerhoek-Noord	341632.77	4699.12	4200
					n46	Helmerhoek-Zuid	259031.25	3562.94	3970
					n47	het Brunink	36508.73	502.17	250
D7	Bedrijfsterreinen	330	**0.003**	105566.58	**n48**	Industrie- en havengebied	46385.27	145	240
					n49	Marssteden	10002.2	31.27	70
					n50	Koekoeksbeekhoek	10029.68	31.35	10
					n51	De Broeierd	39149.43	122.38	10
D8	Glanerbrug en omgeving	16980	**0.012**	1425851.74	**n52**	Glanerveld	118689.46	1413.43	1130
					n53	Bentveld-Bultserve	284211.77	3384.58	3050
					n54	Schipholt-Glanermaten	234813.22	2796.31	2970
					n55	Eekmaat	143101.56	1704.15	1690
					n56	Oikos	177241.24	2110.71	2410
					n57	Eilermarke	136461.87	1625.08	1770
					n58	De Slank	24632.59	293.34	170
					n59	Dolphia	50216.85	598.02	570
					n60	Eekmaat West	256483.19	3054.37	3220
D9	Landelijk gebied en kernen	8410	**0.005**	1735817.9	**n61**	Dorp Lonneker	226002.42	1094.98	1740
					n62	Dorp Boekelo	271830.67	1317.01	2090
					n63	Buurtschap Lonneker-West	330681.68	1602.15	1200
					n64	Noord Esmarke	124361.25	602.53	360
					n65	Buurtschap Zuid-Esmarke	70502.43	341.58	230
					n66	Buurtschap Broekheurne	310729.6	1505.48	1380
					n67	Buurtschap Usselo	81047.32	392.67	270
					n68	Boekelerveld	207919.97	1007.37	920
					n69	Buurtschap Twekkelo	112742.57	546.24	220

Appendix VIII. 3-class method by utilizing residential floor areas of cadastral dataset

Source zone code	Population per source zone	Dasymetric low-density	Dasymetric high-density	Source zone area (sq m)-class low	Source zone area (sq m)-class high	Target zone code	Target zone area (sq m)-class low	Target zone area (sq m)-class high	Population per target zone	Population in reality
D0	22710	0.011	0.012	1352671.84	1021969.7	n0	143507.89	249633.42	4624.11	2620
						n1	109877.87	45474.62	1763.45	1240
						n2	131948.67	2223.35	1478.56	1530
						n3	172410.68	312035.69	5703.35	5720
						n4	166877.42	87731.47	2905.98	2610
						n5	240739.93	187671.03	4937.73	4020
						n6	159860.48	22806.3	2036.7	1920
						n7	159446.49	78709.26	2714.16	2570
						n8	68002.42	35684.57	1183.38	480
D1	12190	0.011	0.012	1001343.53	72066.62	n9	314609.53	28564.09	3809.19	4520
						n10	132114.67	18784.4	1682.43	1230
						n11	194836.56	12356.83	2293.96	2390
						n12	288089.55	10238.09	3293.89	3490
						n13	38921.36	922.78	439.39	440
						n14	14232.43	0	156.56	20
						n15	18539.43	1200.43	218.58	100
D2	23440	0.011	0.012	1428198.09	311930.26	n16	41751.65	136408.65	2123.45	2280
						n17	208485.29	57831.09	2998.88	3960
						n18	132924	0	1462.16	2120
						n19	313759.22	19858.91	3693.63	4780
						n20	97033.45	26133.7	1386.2	1770
						n21	199333.99	17653.16	2408.04	2750
						n22	115813.28	22857.11	1552.8	1850
						n23	198878.26	20211.87	2434.25	2310
						n24	120218.96	10975.78	1456.31	1620
D3	9080	0.011	0.012	618539.8	88544	n25	401600.96	23063.79	4698.99	4950
						n26	216938.84	65480.21	3185.19	4130
D4	18550	0.011	0.012	1163921.21	305016.8	n27	194626.21	18070.94	2361.35	2450
						n28	197105.97	0	2168.17	1620
						n29	276123.33	77005.68	3976.83	3790
						n30	104723.71	84779.49	2186.27	2250
						n31	212573.8	111214.82	3695.13	4670
						n32	81896.46	0	900.86	1200
						n33	96871.74	13945.88	1235.73	2570
D5	8740	0.011	0.012	739222.62	33928.04	n34	201655.02	4612.47	2274.48	2410
						n35	142347.15	13614.78	1731.92	1940
						n36	252839.01	15700.79	2972.78	3410
						n37	142381.45	0	1566.2	980
D6	35700	0.011	0.012	2166284.13	429158.13	n38	292758.37	8103.94	3319.21	3520
						n39	267928.38	29858.7	3311.49	4900
						n40	185733.38	0	2043.07	2360
						n41	187274.96	106432.89	3358.51	3970
						n42	231546.37	78559.8	3505.44	4520
						n43	111310.46	65656.77	2025.43	2530
						n44	264579.58	128525.89	4478.39	5480
						n45	329612.64	12020.13	3772.38	4200
						n46	259031.25	0	2849.34	3970
						n47	36508.73	0	401.6	250
D7	330	0.011	0.012	85398.38	20168.2	n48	46385.27	0	510.24	240
						n49	10002.2	0	110.02	70
						n50	10029.68	0	110.33	10
						n51	18981.23	20168.2	454.85	10
D8	16980	0.011	0.012	1309246.29	116605.45	n52	110198.14	8491.32	1315.77	1130
						n53	274128.98	10082.79	3138.43	3050
						n54	232607.41	2205.81	2585.59	2970
						n55	118163.5	24938.06	1604.04	1690
						n56	169797.32	7443.92	1958.59	2410
						n57	127177.84	9284.03	1512.22	1770
						n58	24632.59	0	270.96	170
						n59	50216.85	0	552.39	570
						n60	202323.66	54159.53	2886.31	3220
D9	8410	0.011	0.012	1722795.97	13021.93	n61	226002.42	0	2486.03	1740
						n62	260650.16	11180.51	3003.55	2090
						n63	330681.68	0	3637.5	1200
						n64	124361.25	0	1367.97	360
						n65	70502.43	0	775.53	230
						n66	308888.18	1841.42	3420.24	1380
						n67	81047.32	0	891.52	270
						n68	207919.97	0	2287.12	920
						n69	112742.57	0	1240.17	220