MARSHALL PLAN FELLOWSHIP FINAL REPORT

Addressing the Modifiable Areal Unit Problem (MAUP) with Dasymetric Modeling in the Context of Spatial Decision Support for Real Estate Choice

By:

Garland McNew

San Diego State University, Department of Geography

San Diego, USA

M.S. Thesis Advisor:

Dr. Piotr Jankowski

San Diego State University, Department of Geography

San Diego, USA

Marshall Plan Scholarship Research Program Supervisor:

Dr. Gernot Paulus

Carinthia University of Applied Sciences, Austria

Table of Contents

Table of Figures

Addressing the Modifiable Areal Unit Problem (MAUP) with Dasymetric Modeling in the Context of Spatial Decision Support for Real Estate Choice

Abstract

Avoiding misleading or erroneous results when using aggregated census data in Spatial Models can be challenging. A host of analytical hurdles are encountered with aggregated data, not the least of which is the Modifiable Areal Unit Problem (MAUP). This research explores the integration of dasymetric modeling and Locally Weighted Linear Combination (LWLC) as a framework for mitigating negative effects associated with the MAUP.

1 Introduction

Aggregated census data is arguably one of the most ubiquitous data resources used in population based geographic analysis. The explicitly geographic nature of census data, ease of access and often times low cost, contribute to its prolific use in a myriad of spatial analysis domains. However, aggregated census data comes at a steep analytical cost. Any spatially aggregated data is subjects to a problem known as the Modifiable Areal Unit Problem (MAUP) (Fotheringham & Wong, 1991).

In order to protect individual privacy, survey data is spatially aggregated to various enumeration units (e.g. polygonal units such as tracts, blocks, and block groups or grid cells). In general, as the enumeration unit's spatial extent increases, so too does the number of attributes detailing the aggregate characteristics of the underlying population. Attribute richness makes census data with larger spatial extents very attractive for use as evaluation criteria in GIS Multi-Criteria Decision Analysis (GIS-MCDA) models. The aggregation of population attributes certainly achieves the desired result of masking respondent identity. However, analysis results derived from these zones, should be approached with great caution, due in large part to the MAUP (Malczewski, 1999). Unknown distribution of attributes within zones, as well as arbitrary zoning and boundary arrangements, all contribute to undesirable uncertainty, and even at times entirely useless results (Openshaw and Taylor 1979, 1984).

Recent work in GIS-MCDA, which incorporates parameters accounting for local variation in evaluation criteria is the main motivation behind this research. Locally Weighted Linear Combination (LWLC) is a GIS-MCDA technique that employs the range sensitivity principle to increase or decrease weights based on the variability of criterion values observed in a defined neighborhood (Fischer, 1995; Carter & Rinner, 2013). However, in applications using socio-demographic data as evaluation criterion, LWLC encounters two challenges that potentially diminishes its utility and reliability: 1) unknown MAUP-related effects on input criteria and model solution reliability, 2) restrictions imposed on neighborhood definitions due to the limited number of scales available for socio-demographic data. Though the functionality of analyzing locally variable criteria values exists in LWLC, the model must have sufficiently disaggregated data in order to be fully leveraged. Being able to quickly and accurately model granular population criteria could significantly increase model robustness and accuracy. The increase in utility would be especially beneficial in situations where decision alternatives exist in one or more of the same aggregate zones. In general, understanding of the underlying heterogeneity of evaluation criteria allows for better discrimination between decision alternatives in GIS-MCDA.

Dasymetric modelling has shown promising results as a more generalized approach for disaggregating census based population characteristics (Eicher & Brewer, 2001; Mennis, 2003; Nagle et al., 2013). Dasymetric modeling differs from traditional interpolation methods by incorporating related and limiting ancillary variables. A thorough review of the literature reveals a gap in the exploration of integrating dasymetric modeling and GIS-MCDA. Dasymetric modeling could expand the utility of LWLC by allowing for more flexible neighborhood definitions and scale arrangements. Expanded scale flexibility for evaluation criteria in LWLC would address research findings that show the necessity of evaluating different neighborhood arrangements in order to produce more reliable model results (Carter & Rinner, 2013).

The originally proposed research for this project was directed at addressing issues relating to MAUP with dasymetric modeling in the context of spatial decision support for real estate choice. However, sufficient data relevant to choice models for real-estate was unattainable. As such, an alternative case study relating to vaccination coverage and risk has been developed based on data availability. As before, the formulation of a case study is intended to serve primarily as a testbed to explore the general validity of integrating dasymetric modelling with choice models to address MAUP related challenges.

2 Literature Review

2.1 Modifiable Areal Unit Problem (MAUP)

The Modifiable Areal Unit Problem, or MAUP as it is known, has been a pervasive and long-standing challenge in nearly all spatial analysis domains. The MAUP can be categorized as a special occurrence of ecological fallacy. Ecological fallacy is a logical error that occurs when conclusions are made about individuals, or subsets of individuals, based on the results of analyzing the aggregate group (Malczewski, 1999, King, 1997). The MAUP occurs in a similar way when change in the analysis scale and/or zoning arrangements of areal units alters results and consequently the ability to produce reliable conclusions (Openshaw and Taylor 1979, 1984;). The effects of the MAUP were first articulated by Gehlke and Biehl in 1943. When investigating correlation coefficients, Gehlke and Biehl observed in bivariate analysis of census tracts that the values of the correlation coefficients were dependent on the size of the areal units and the arrangements of the sampled census tracts. Later, Openshaw and Taylor (1979) would not only coin the term Modifiable Areal Unit Problem, but clarify Gehlke and Biehl's observations as one of two more specific problems found within the MAUP. The two problems are actually very interrelated and can be thought of as the scale problem (observed by Gehlke and Biehl) and the aggregation problem (Openshaw and Taylor, 1979). The scale problem is encountered when geographic attribute values are aggregated into larger/courser areal units or disaggregated erroneously into smaller areal units. The zone problem occurs when different arrangements of aggregate zones within a shared parent boundary can produce vastly different and unpredictable summary statistics (Wong, 2004). Furthermore, as Taylor (1979) originally pointed out, the aggregation problem can be described in reference to a *zone system* or a *grouping system*, the former pertaining to contiguous arrangements of areal units and the later for non-contiguous groupings of areal units. An example of the aggregation/zoning and scale components of MAUP can be seen in Figure 1 (Bell & Schuurman, 2010); this figure illustrates the percent of population per arrangement, which hold university degrees (base population captured within a boundary divided by university degree counts within the same boundary). Here is an example where a single fixed location within a study area could have many different attribute values simply based on the scale or zoning arrangement of the data. Clearly then, these two problems individually, *and* in combination can result in inconsistent and/or unreliable decision model outcomes.

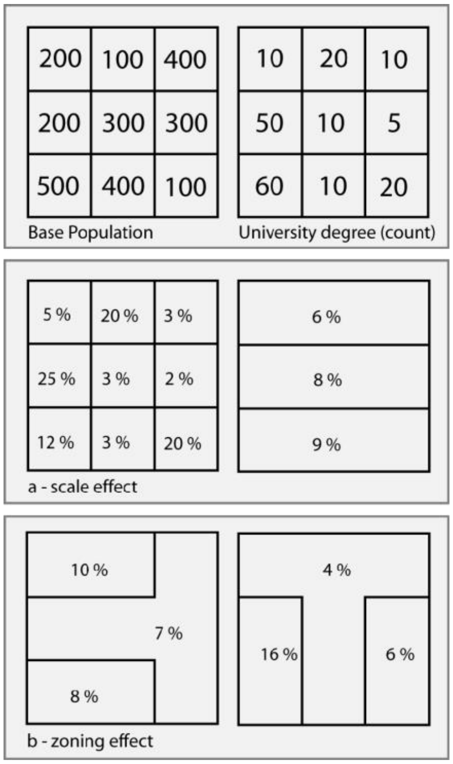

Figure 1. MAUP

Early research results using advanced multivariate statistical analysis pointed to the nearly complete unreliability of analysis performed on aggregate areal data. Unfortunately, (and with a tone of exasperation) the recommendation by the researchers to avoid MAUP related effects was simply to entirely avoid the use of aggregated enumeration units in spatial analysis (Fotheringham & Wong, 1991).

In the case of socio-demographic data, the use of aggregate data in spatial analysis is still quite commonly practiced. This is due in large part to the ease of access and relatively low cost. Enumeration units are constructed in order to protect the privacy of survey participants. Survey participants are commonly individuals, households or businesses. Access to sampled point data with individual human or business characteristics is not surprisingly highly restricted. Consequently, aggregate census data is still one of the most widely used datasets when analyzing socio-demographic population characteristics in GIS. Unfortunately, due to the complexities involved with the MAUP and lack of a generalized solution, many researchers simply omit or ignore the effects altogether (Swift et al., 2013). Fotheringham and Wong's initial pessimism regarding the ability to address the MAUP in any meaningful way seems a bit hopeless, however in the years to follow some success was seen with domain specific solutions (Dungan et al., 2002; Xu et al., 2014). Typically, these proposed solutions include a myriad of caveats and complexities, such as significant knowledge of the underlying phenomena, correct sampling unit size/shape presuppositions, and explorations of many different definitions of scale.

While many domain specific solutions for the MAUP rely solely on source areal unit attribute data and statistical interpolation methodologies, this research aims to investigate the utility of dasymetric modeling as a streamlined approach for in mitigating negative effects of the MAUP in GIS-MCDA. More specifically, this research attempts to mitigate issues arising with aggregate grid data due to the scale effect. However, since the scale and zonal effects of the MAUP are quite interrelated, potential solutions may have positive implications for addressing the aggregation component of the MUAP as well.

2.2 Dasymetric Modeling

Dasymetric mapping is a relatively old cartographic method created by the Russian Geographer Benjamin Semyonov-Tyan-Shansky in 1911 (Petrov, 2008). Semyonov-Tyan-Shansky proposed mapping population density in Russia by excluding areas where population counts would be unlikely (later termed the binary method), as well as calculating density values for inhabited regions. An example of Semyonov-Tyan-Shansky's original work from 1922 can be seen in figure 2 (Preobrazhensky, 1954). Semyonov-Tyan-Shansky's density calculations were heavily based on current knowledge of land use patterns throughout the study area (Petrov, 2012). Aside from Semyonov-Tyan-Shansky, Wright is also highly cited for his work on dasymetric mapping in an early journal contribution to Geographic Review in 1936. However, as Petrov pointed out (2012), this method sat in academic obscurity for nearly 90 years until experiencing a resurgence of interest during the early 2000's. Petrov and others have attributed this resurgence in large part to a combination of technological advancements in spatial analysis allowing for more complex tasks, and a need for higher resolution datasets detailing population distribution (Eicher & Brewer, 2001). It's important to temper the statement regarding a resurgence by pointing to the most prominent paper in the field, which was authored by Jeremy Mennis in 2003. Mennis' work has been cited 365 times (Google Scholar, 2016), The citation count highlights how this is still a method very much in the process of exploration and development.

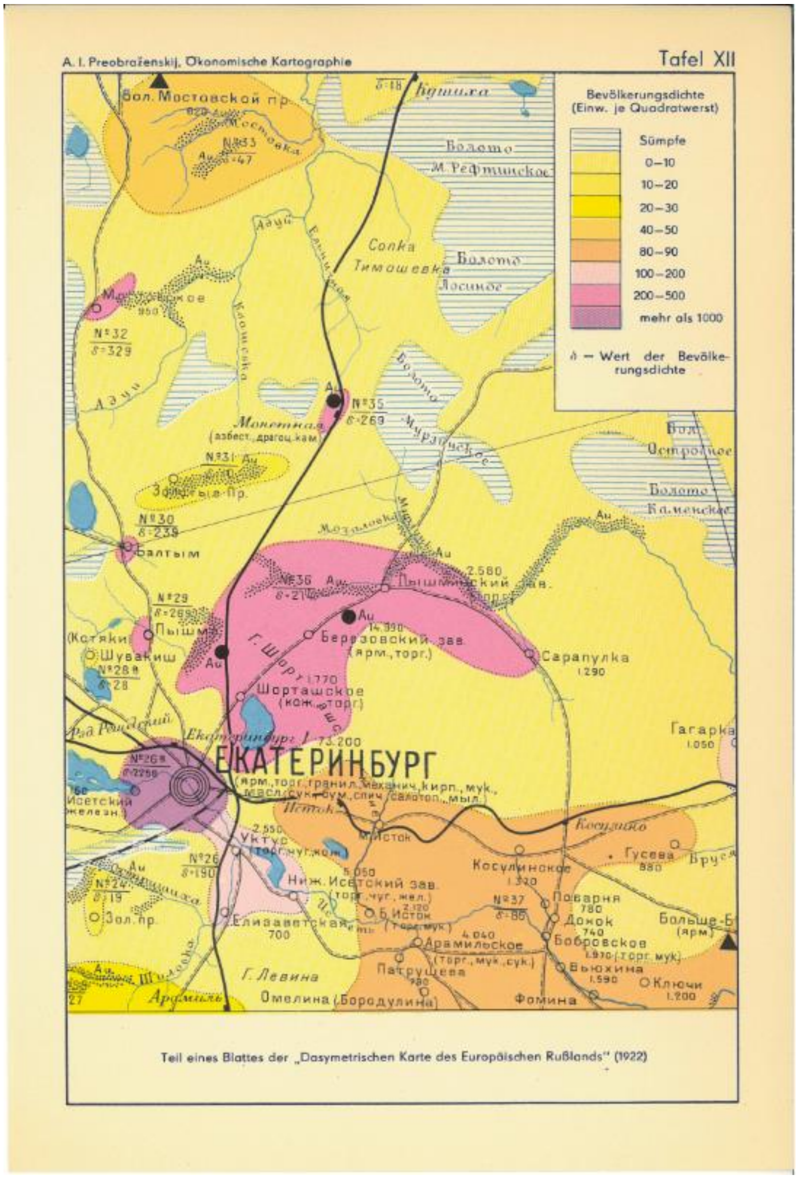
A. I. Preobraženskij, Ökonomische Kartographie
Tafel XII
Bevölkerungsdichte (Einw. je Quadratwerst)
Sümpfe
0–10
10–20
20–30
40–50
80–90
100–200
200–500
mehr als 1000
δ – Wert der Bevölkerungsdichte
ЕКАТЕРИНБУРГ
Teil eines Blattes der „Dasymetrischen Karte des Europäischen Rußlands" (1922)

Figure 2 Semyonov-Tyan-Shansky 1922 Dasymetric Map

Dasymetric mapping is generally considered a form of areal interpolation rather than strictly a cartographic technique (Qui & Cromley, 2013; Mennis, 2009). In brief, areal interpolation deals with different methods and techniques for transforming phenomena attributes from a set of source units to a set of target units (Goodchild and Lam, 1980). At the core, this is exactly what the dasymetric method does. Very simply, dasymetric mapping uses additional finer scale data (ancillary data) and empirical statistical methods to establish relationships (expressed as weights) between the ancillary data and aggregated attribute. These derived weights are then used to disaggregate source zone attributes to the higher resolution target zone scale. Ancillary data used in dasymetric mapping can be categorized as one of two variable types:

1) *Limiting variables*, which restrict or place a limit on where a source unit phenomenon can occur in the targeted zones. An example in population density modeling would be cells categorized as water in a LULC raster dataset
2) *Related variables*, which either decrease or increase the presence of a source phenomenon within a target unit. An example in population density modeling would be the cells categorized by a range of residential-use intensities (e.g. low, medium, high).

Two noteworthy examples of modern dasymetric modeling include:

1) empirical sampling of land-use/land-cover (LULC) classes in raster data to establish a relationship between target zones and source zone population attributes (Mennis & Hultgren, 2006). In traditional applications of this method, target zones (cells in this case) can only represent a single value from a range of related ancillary subtypes.
2) Ordinary Least Squares (OLS) regression weighting. In this method source zone phenomenon values are regressed against a set of target zone explanatory variables. Explanatory variable coefficients are interpreted as density values per target zone. A further scaling and fitting step is incorporated to ensure target zones estimations sum up to the value of the source zone value (Langford, Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method, 2006).

Dasymetric mapping then can be thought of a conditional probability equation in Gaussian statistics. However, this method is explicitly spatial in nature and the use of ancillary data uniquely separates this method from strictly empirical statistical methods of areal interpolation, such as Inverse Distant Weighting (Lu and Wong, 2008), Spline or Kriging (Lam, 1983).

Contemporary applications of dasymetric methods predominately revolve around mapping population density. In these applications, the source units consist of census enumeration units (e.g. grids, block groups, municipalities), while target units are typically the cells dictated by the resolution of Land Use Land Cover (LULC) data (Zandbergen & Ignazio, 2010).

In recent years researchers have started to explore new ancillary data, such as LIDAR (Aubrecht et al., 2009), cadastral data (Maantay et al., 2007), and address point location (Zandbergen, 2011) to achieve greater precision and accuracy in population density mapping. In addition to advances in spatial resolution for mapping population density, researchers have shown that dasymetric techniques provide a promising avenue for disaggregating census based socio-demographic attributes. Such research efforts include mapping diversity (Dmowska & Stepinski, 2014), crime (Poulsen & Kennedy, 2004), income and race (Leyk et al., 2013, Nagle et al., 2013) and housing prices (Eicher & Brewer, 2001).

Within the literature there are various ways researchers have tackled the discussion regarding uncertainty and accuracy of model outputs. For example, some researchers have explored uncertainty rather than error (Nagle et al., 2013). Uncertainty is calculated based on published margins of error in census data. Other research limits disaggregation to the highest resolution target zone with known population attributes in order to test model performance. Still, there are some researchers who have chosen to entirely omitted the analysis regarding the accuracy of the mapped attribute (Poulsen & Kennedy, 2004; Dmowska & Stepinski, 2014). This can be often times attributed to access limitations for actual population counts at the target unit scale being mapped. In many studies the target cell resolution is 30 meters (derived from LULC raster data), which is a significantly higher resolution than most commonly available data sources.

2.3 GIS-MCDA and LWLC

GIS can be considered an integrated decisions support and problem solving system which utilizes spatial data (Cowen 1988). GIS-MCDA is a heavily researched and active field which provides a wealth of structured methodical approaches for leveraging GIS as a decision making tool (Carver 1991, Jankowski, 1995, Malczewski, 1999). GIS-MCDA transforms, geographically represents and evaluates a decision maker's goals and preferences to aid in one, or a set of site and/or plan evaluation(s) (Carver, 1991). As Malczewski outlines (2006), there are five foundational elements of GIS-MCDA:

(i) A goal or a set of goals an individual (or a group of individuals) attempts to achieve along with associated evaluation criteria (objectives and/or attributes) on the basis of which the decision-maker evaluates alternative courses of action
(ii) The decision-maker or a group of decision-makers involved in the decision-making process along with their preferences
(iii) The set of decision alternatives (or the decision variables)
(iv) The set of uncontrollable variables or states of nature (decision environment)
(v) The set of outcomes or consequences associated with each alternative-criterion pair

GIS-MCDA models commonly use census based socio-demographic areal datasets to represent evaluation criteria. Census data can be represented using one or multiple scales, (e.g. grid, block group, municipality) and various arbitrary boundary arrangements. Due to the use of various aggregation scales and zoning schemes, GIS-MCDA is susceptible to the same analytical difficulties previously mentioned with the MAUP. Evaluation criteria are not only represented (incorrectly) as spatially homogenous in these datasets, but it has been shown that evaluating a set of alternatives using areal evaluation criteria in GIS-MCDA models can have drastic and unpredictable effects on evaluation outcomes (Malczewski & Rinner, 2015).

Malczewski and Rinner (2015) confirm these difficulties with aggregated data. Furthermore, Malczewski and Rinner point to the fact that there has been little effort directed towards creating a systematic approach to addressing the MAUP in GIS-MCDA. Additionally, Malczewski and Rinner articulate the significant value of being able to traverse criteria scales in order to find the most suitable aggregation level to address MAUP related effects. This scale selection process has also been shown to be successful in other research domains outside GIS-MCDA (Xu et al., 2014). Unfortunately, in many cases the available spatial scales for evaluation criteria are often limited and fixed to a particular zoning and boundary arrangement.

Furthermore, new GIS-MCDA methods have emerged which point to the need for higher resolution evaluation criteria. One such method is Local Weighted Linear Combination (LWLC) (Malczewski, 2011). LWLC is an extension of the popular Weighted Linear Combination (WLC) method. There is a large volume of notable research relating to the use of WLC in GIS-MCDA (Eastman et al. 1993; Malczewski 2000, 2006). One of the reasons for the extensive use of WLC is its computational simplicity and ease of implementation. However, the use of WLC for decision making can be problematic when applied to study areas with locally variable criteria. The method by which WLC is calculated can have the side effects of masking underlying variability, increasing autocorrelation, and clustering among the final decision alternative values (Carter& Rinner 2014). This masking effect is due to the global weighting scheme WLC employs (Malczewski, 2011). As Malczewski points out (2011), this global weighting scheme is unrealistic in many real world applications. LWLC in contrast employs what is known as the range sensitivity principle (Carter& Rinner 2014; Fischer, 1995). This principle is based on the notion that weights should fluctuate based on the distance from, and variability within a given neighborhood. LWLC incorporates this principle by applying criterion weights based on the range of values observed within a user defined neighborhood scheme for a given study area (Malczewski, 2011). “Neighborhood” in the case of the ArcGIS LWLC tool developed by Carter & Rinner (2014) is defined by three types of spatial contiguity; Rook, Queen and K-nearest. The ArcGIS Desktop LWLC tool uses these neighborhood definitions to measures and apply weights. Criteria with high neighborhood variability are assigned larger weight values, and conversely criterion with lower neighborhood variability are assigned proportionately lower weight values. Whereas WLC tends to suppress and mask high variability criteria, LWLC incorporates the use of contiguity, local variability, and global variability to emphasize such cases in the model output.

LWLC offers a robust exploration of local variability. However, since LWLC is noted for being sensitive to criterion scale (Malczewski, 2011; Carter, Rinner 2014). As the level of aggregation increases (e.g. going from 250m grids to 500m grids), trade-offs among alternative locations within the study area become less clear due to a higher likelihood of a multiple locations falling within the same zone. Additionally, an increase in aggregation extent only serves to further mask the underlying heterogeneity of the population by increasing the number of census respondent surveys being average. For these reasons, a high degree of criteria granularity offers a potential advantage in LWLC.

2.4 Integrating Methods

Utilizing dasymetric techniques for modeling evaluation criteria could potentially lead to a better understanding of the decision space, as well as more robust model outcomes. Dasymetrically modeled evaluation criteria could also lead to more flexibility when choosing optimal scale and/or zoning arrangements for decision criteria as well as neighborhood definitions. Additionally, there are potential implications beyond the usage in LWLC. GIS-MCDA is a broad field of research with a multitude of choice models which incorporate the use of areal census data. This research provides an exciting opportunity to add knowledge to the field of GIS-MCDA by improving criteria reliability, accuracy, precision and model robustness. After an extensive review of the literature, there is a clear gap in research addressing the analysis and testing of dasymetric modeling for the production of evaluation criteria layers in spatial choice model applications.

3 Case Study

Austria's implementation of Statutory Health Insurance (SHI) and various other supplementary insurance options for children, non-working spouses and low income households, ensures that nearly 99% of the population is covered by health insurance (Austrian Embassy). Even with the expansive implementation, coverage capacity and cost of the SHI, Austria has continually ranked one of the lowest in the world for influenza vaccination rates, with less than 10% population coverage in the 2011/2012 year (Kunze et al., 2013). According to the World Health Organization (2011), on average 400,000 people contract influenza every year in Austria, and out of those, up to 6000 die due to influenza related complications. WHO along with leading researchers point to a lack of public awareness and ineffective social marketing campaigns as a major contributor to low coverage rates (Kunze, 2008). Additionally, researcher suggests that focused, effective, research driven marketing campaigns should be adopted as part of a holistic effort to increase health literacy, and decrease flu related morbidity and mortality in the population (Kunze et al., 2013, Kunze et al. 2007; Hoffman, 2015). The multiple criteria nature of identifying the distribution and concentration of at-risk and low-coverage populations to aid decision makers in the development of more effective marketing campaigns provides a logical case study for the proposed research. The objective in this case study is to more precisely identify the geographic distribution of at-risk and low coverage population groups in order to facilitate more effective marketing campaigns.

4 Study Area

The selected study area for this research is located in the southern-most region of the federal province of Carinthia, Austria (Figure 3).

The analysis extent is defined using the Klagenfurt-Villach Nomenclature of Territorial Units for Statistics (NUTS) level 3 classification boundary. The specific NUTS 3 code for this region is AT211. This study area was selected in order test and validate a Dasymetric-LWLC framework under varying population density and land-use conditions. The Klagenfurt-Villach region contains a wide variety of land-use characteristics ranging from densely populated urban structure, to large swaths of sparsely populated or uninhabited land.

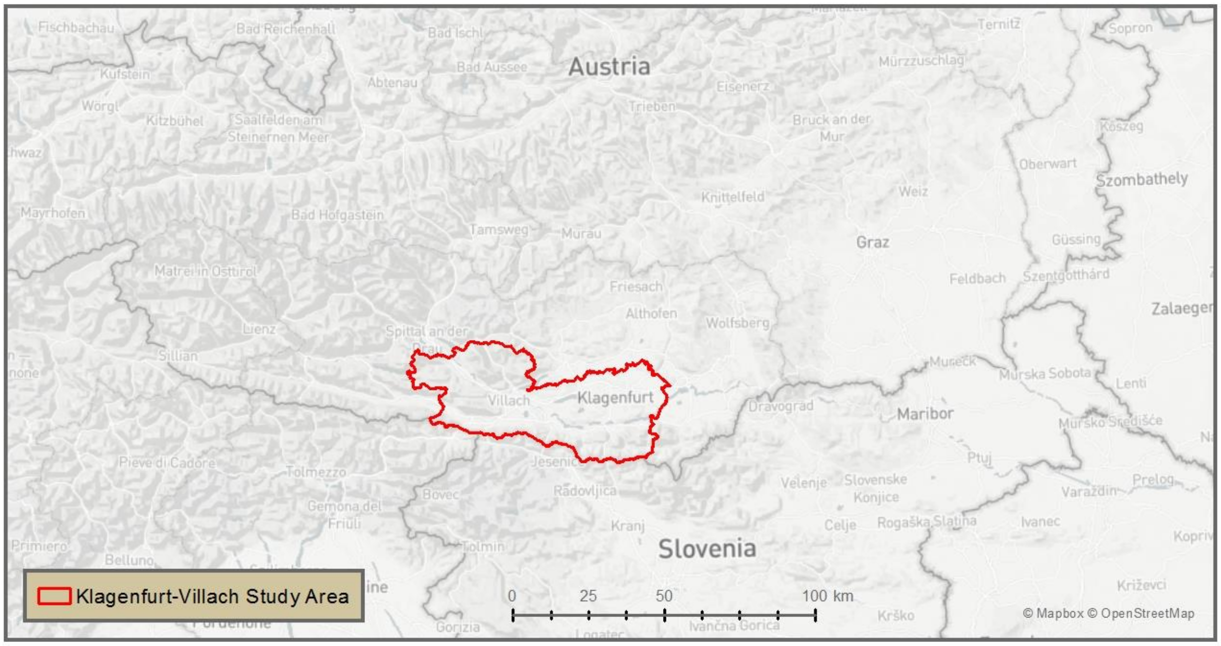

Figure 3. Klagenfurt-Villach NUTS 3 study area

5 Research Questions

In the context of the aforementioned objectives, there are two related research questions:

1) What are the performance differences between simple Linear Apportionment (LA) and OLS for producing disaggregated evaluation criteria?
2) What zonal related effect can be observed among market area alternatives when using 250m disaggregated decision alternatives (grid cells) as compared with 500m aggregated decision alternatives?

6 Data

Regional sociodemographic and building data from Statistics Austria has been provided by the Department of Geoinformation at Carinthia University of Applied Science. The following is a brief summary of the two primary datasets used in this research:

1. 250 meter regional statistical grid unit census data. The feature class includes total population counts, as well as eleven age distributions categories ranging from 0-85+. This data will provide actual population counts for model performance testing, while also serving as the source to produce aggregated data within the study area.
2. 250 meter regional statistical grid unit building data (Table 1). This feature class contains building counts and use-type counts at the 250m cell resolution. Building data is subdivided by structure type or use-types. OBJ_1 through OBJ_10) are counts of buildings according to the predominate purpose of the property (e.g. commercial or residential) . NTZ_ALL, NTZ_HWS and

NTZ_NWS are use-type counts per grid. A single property can contain multiple use types. This dataset will serve as the ancillary dataset in each dasymetric model.

Name	Description
OBJ	Building (Sum of OBJ_1 to OBJ_10)
OBJ_01	Resident building with 1 apartment
OBJ_02	Resident building with > 1 apartment
OBJ_03	Community building (e.g. retirement homes, dormitories, apprentice centers)
OBJ_04	Hotel, restaurant, pension
OBJ_05	Office building
OBJ_06	Commercial building
OBJ_07	Transportation or telecommunication building
OBJ_08	Industry or storage building
OBJ_09	Building for cultural activities, education, health
OBJ_10	Other building
OBJ_WG	Resident building (Sum of OBJ_1 and OBJ_ 2)
NTZ_ALL	All use-types (Includes NTZ_HWS and NTZ_NWS but does not equal their sum)
NTZ_HWS	Permanent residency use-type (NTZ_ALL subtype)
NTZ_NWS	Secondary residency use-type (NTZ_ALL subtype)

Table 1. Ancillary data

7 Methodology

7.1 Dasymetric Modeling

The following section outlines three specific dasymetric methodologies used in this research. All three methods were coded and implemented using Python 2.7.12 and the following libraries:

- Geospatial Data Abstraction Library (GDAL)
- Pandas
- Numpy
- Statsmodels

Maps and visualizations are produced using ArcGIS Desktop 10.4 and Mapbox.

7.1.1 Ordinary Least Squares (OLS)

The first method utilized in this research will be an adapted version of OLS regression weighted dasymetric modeling articulated by Longford (2006), as well as Reibel and Agrawal (2007).

Dasymetric modeling using OLS is a two stage process. The first step involves fitting an OLS model to derive related variable disaggregation weights. Explanatory variable coefficients from OLS results are interpreted as the global per unit area population density for each related variable. Unlike other methods for deriving weights via regression, OLS has the added advantage of directly reading coefficients as density values. For this research the dependent variable will be the 500m source zone population counts and explanatory variables will be building and/or use-type subtypes classes at the 250m scale.

In realistic applications the R^2 value obtained from OLS fitting will be below 1. In this case, less than 100% of source zone population variation is explained by the regression. To account for this a fitting procedure is utilized to scale the weighted estimates for each related variable according to its associated source zone. This fitting and scaling procedure accounts for populations not predicted in the OLS model and ensures the sum of target zone estimations equal the source zone population value, a method derived from smooth concepts in smooth pycnophylactic scaling (Tobler, 1979). The fitting procedure is describe using the following formula:

$$d_{cs} = \frac{E_s}{E_{is}} \cdot d_c$$

Where d_{cs} is the population density estimate for building type *c* in source zone *s*; E_s is the actual population count in source zone *s*, E_{is} is the estimate population count in the source zone produced by the initial OLS density estimate, and *dc* is the initial OLS density estimate for building type c.

The formula assumes a particular ancillary data structure, specifically a raster data structure. In raster applications a cell can only represent a single related variable subtype from a range of subtypes (e.g. a LULC class). However, this study utilizes vector data which allows for multiple ancillary classes to be represented within a single grid cell. As such, the formula requires an adaptation in order to leverage higher dimensional inputs. In this research a single cell can (and in most cases does) contain multiple building classes. Each of these building classes will need to be assigned an associated density estimate from OLS model . To account for this and produce valid target estimates, a new approach has been formulated which is expressed using the following formula.

$$\hat{p}_{ts} = \sum_{c=1}^{C} d_{cs} \cdot v_{ct}$$

Where $\hat{p}_{ts}$ is the sum of fitted estimated populations for target zone *t* in source zone *s*; *c* is the building type in source zone *s*; d_{cs} is the population density estimate for building type *c* in source zone *s* (obtained from the previous formula), and v_{ct} is the count value of building type *c* in target zone *t*. This method allows for multiple ancillary variables to be considered within a single vector grid cell. Lastly, two important points need to be made regarding the OLS method:

1. OLS regression for deriving disaggregation weights should be fitted without intercepts, since areas with no inhabitable building classes are expected to have no population. Reibel & Agrawal (2007).

2. The scaling and fitting of dervied weights increases the explanatory power of the OSL regression model to R^2 = 1.0. This process addresses (in part) some of the issues arising from autocorellated explanatory variables, as well as offers an alternative sollution otherwise unattainable using traditional statistical appraoches (Flowerdrew & Green 1989), (Flowerdrew & Green 1992), (Reibel & Agrawal, 2007).

7.1.2 Linear Apportionment (LA)

The second dasymetric method selected for this research is based on a very simple linear relationship model between the source population and the ancillary class variable(s) counts. The application of this model simply involves dividing each source zone population by the sum of the ancillary class variable(s) in all feasible target zones. The value obtained is then the specific disaggregation weight for each ancillary unit type within the source zone. The primary motivation behind using this approach is based on results observed in previous research which was able to produce very high accuracy estimation using building point data (Zandbergen, 2011). A criticism of this work was the use of simple mean error across all target zones as the measure of comparative model performance. This measurement of model performance can be misleading since the mean of the errors masks the effect of significant, offsetting, under/over count errors. For this reason, further performance measurements (discussed later) will be employed to test model performance and quality. Important secondary considerations for the use of this method also include the ease of implementation and low computational overhead in comparison to the OLS approach.

7.1.3 Binary Method

In order to disaggregate age-related variables, the Binary Method was implemented. This method simply removes areas of infeasible population location within a given source zone, thereby disaggregating the population density of the associated age variable within the source zone. Although the method has proven to be simple and robust (Fisher & Langford, 1996; Langford, 2006. As discussed later in the data preparation section; to expand the expression of the age variable across all source zones with population attributes, simple IDW was employed. IDW was used in order to expand the areal coverage for a comparative analysis between LWLC model outcomes. As such, it would be erroneous to develop and test the efficacy of a dasymetric model for disaggregating aggregate IDW interpolated age attributes. However, binary disaggregation is justifiable since common sense would dictate that any associated age attributes (modeled or actual) of a population can only exist where there is a presence of a population. The binary method is described as:

$$\hat{p}_{ts} = \sum_{s=1}^{S} \frac{A_{tsp} P_s}{A_{sp}} = \sum_{s=1}^{S} A_{tsp} d_{sp}$$

where $\hat{p}_{ts}$ is the estimated population of target zone *t* in source zone *s*; A_{tsp} is the area of overlap between target zone *t* and source zone s, and having *p* building class(es) identified as populated; P_s is the population of source zone *s*; A_{sp} is the area of source zone *s* having building class(es) identified as populated; *S* is the number of source zones; and $A_{tsp}d_{sp}$ is the dasymetric density of the populated class in source zone *s* (Langford, 2006).

7.1.4 Performance Testing

Two quantitative performance measurements and one qualitative method were employed in order to evaluate model performance and the efficacy of use in LWLC. These methods are described as follows:

1. Root Mean Square Error (RMSE): RMSE is the summary statistic most frequently used for error calculations and performance comparison between models in dasymetric research (Eicher, Brewer 2001; Langford, 2006). RMSE can be thought of as the standard deviation of the error distributions across a study area for a given dasymetric estimation. It is represented with the following formula:

$$\sqrt{\frac{\sum_{i=1}^{m}(\hat{p}_i - p_i)^2}{m}}$$

   where $\hat{p}_i$ is the actual population of zone *i*, p_i is the estimated population of zone i, and m is the number of zones in the study area ((Reibel & Agrawal, 2007)
2. Linear regression: A simple linear regression analysis using a 95% confidence interval is performed, where the actual population of the study is the criterion variable and the dasymetric estimates serve as the predictor variable. The R^2 will provide an added dimension of global model performance for comparison purposes, as well as an absolute measure of performance (Zandbergen, 2011).

Visual Analysis: for each of the dasymetric techniques, maps will be presented for a visual analysis of error across the study area.

7.2 Local Weighted Linear Combination (LWLC)

7.2.1 Method

As discussed previously this research will implement the LWLC method developed by Malczewski (2011) and further built upon for vector based implementation by Carter and Rinner (2014). LWLC is expressed as follows:

$$V(l_{iq}) = \sum_{k=1}^{n} W_{kq}V_{kq}(l_{iq})$$

where $V(l_{iq})$ represents the local score of location *i* within the *q*th neighborhood, w_{kq} represents the locally adjusted weight assigned to the *k*th criterion within the *q*th neighborhood, and $v_{kq}(l_{iq})$ represents the value of the *k*th criterion at location *i* as determined by standardizing the value of the *k*th criterion with respect to all values of criterion *k* within the local neighborhood, *q.*

The local weight of criterion *k* is a function of both the global weight assigned to criterion *k* and the local range of values of criterion *k*. Local weight are expressed as follows:

$$w_{kq} = \frac{\frac{w_k r_{kq}}{r_k}}{\sum_{l=1}^{n} \frac{w_l r_{lq}}{r_1}}$$

where w_{kq} represents the local weight assigned to the *k*th criterion within the *q*th neighborhood, w_k is the global weight of criterion *k*, r_{kq} is the range of values of criterion *k* within the *q*th neighborhood, and r_k is the global range of criterion *k*. The local weight is standardized by dividing by the sum of all local criterion weights (*l*) for each location such that the local weight will fall within the range $0 \leq w_{kq} \leq 1$ and the sum of all the local weights for location *i* will equal 1.

Initial testing of dasymetric evaluation criteria in LWLC will be done using an MCDA ArcGIS add-in with LWLC functionality. This tool was developed by Voss under the supervision of Rinner (2013).

7.2.2 Analysis Procedure

As with previous work in LWLC, a series of maps will be used to qualitatively evaluate and compare model outputs. Analyzing model output in this research also presents challenges not encountered in previous literature relating to LWLC. Rather than analyzing the effects associated with altering neighborhood definitions within the same spatial scale, as seen with Carter & Rinner's work (2014), this research will be evaluating disparate spatial scales using the same neighborhood definition. For this reason, measurement of performance and change between model outputs using the techniques described by Carter and Rinner are not directly applicable. To address questions relating to scale and zone effects, a quantitative evaluation of change in the areal expression, scale, and boundary relationships of criteria value pairs will be presented.

7.3 Data Preparation

7.3.1 Inverse Distance Weighting

For 250m grid cells with fewer than 30 people (discussed later in Dasymetric Implementation), a combination of Inverse Distance Weighting (IDW) and the Largest Remainder method was used to interpolate and populate missing values.

IDW is a simple method based on Tobler's first law of geography that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Unmeasured values within a study area (in this case, age category counts) are predicted based on the distance to, and value of, the surrounding locations with known values. Locations with a known value closer to the un-sampled location will have a larger influence on the predicted value than locations further away. IDW is expressed formally as (:

$$Z_j = \frac{\displaystyle\sum_{i=1}^{n} \frac{Z_i}{\left(h_{ij} + \delta\right)^{\beta}}}{\displaystyle\sum_{i=1}^{n} \frac{1}{\left(h_{ij} + \delta\right)^{\beta}}}$$

Where Z_j is the interpolated value of a grid node, Z_i are the neighboring data points, h_{ij}

are the distances between the grid node and data points, β is the weighting power, and δ is the smoothing parameter (Tomczak, 1998). In this application distance between an unknown and known point was based on Euclidian distance.

7.3.2 Largest Remainder Method

A percent of population value was derived for each age attribute from the sum of IDW interpolated age attributes. These percentages were used to distribute the actual population of a cell with 30 or fewer people into each age category. To produce integer population counts the Largest Remainder (or Hare-Niemeyer) method was employed (Niemeyer & Niemeyer, 2008). This method truncates each age estimation to an integer value, then progressively increments the each by one based on the decimal remainder value (highest to lowest) until the sum of the actual population is attained.

8 Dasymetric Implementation

8.1 Data Preparation

Since census data was provided at the 250m scale, it was necessary to manufacture an aggregate dataset to test the various dasymetric methods. A 500m aggregation level was chosen for testing. Since 250m grid cells are enumerated population counts with no margin of error, a simple summation of the population counts and associated attributes was theoretically the only needed step. However, after closer inspection of the data, it was found that any cells with population counts less than 31 did not have any associated age characteristics. Consequently, this equated to 73.3% reduction in viable aggregation cells, and a 20.1% loss of the total population count within the study area. Refer to Figure 4 for a visual representation of analysis loss in the study area. Additionally, cells with and without age attributes are commonly located within the same (500m) parent aggregate grid cell. The consequence here being that only those 500m grid cells containing 250m cells with more than 30 people could be utilized in the study. For this reason, the exclusion of the 250m cells without age attributes from the study was found to be intolerable as this would unduly constrain the study area.

In order to overcome the challenge of missing age data, a simple Inverse Distance Weighting IDW interpolation was used to create a count estimate for each missing age category across the study area. Cells with missing age attributes were appended with interpolated counts and a percentage distribution was calculated by using the sum of all interpolated age categories as the population total. The modeled age distribution percentages for each 250m cell were used as weights to allocate the actual population into respective age categories. Lastly, the Largest Remainder method was employed to produce integer population values for each age category (Niemeyer & Niemeyer, 2008)

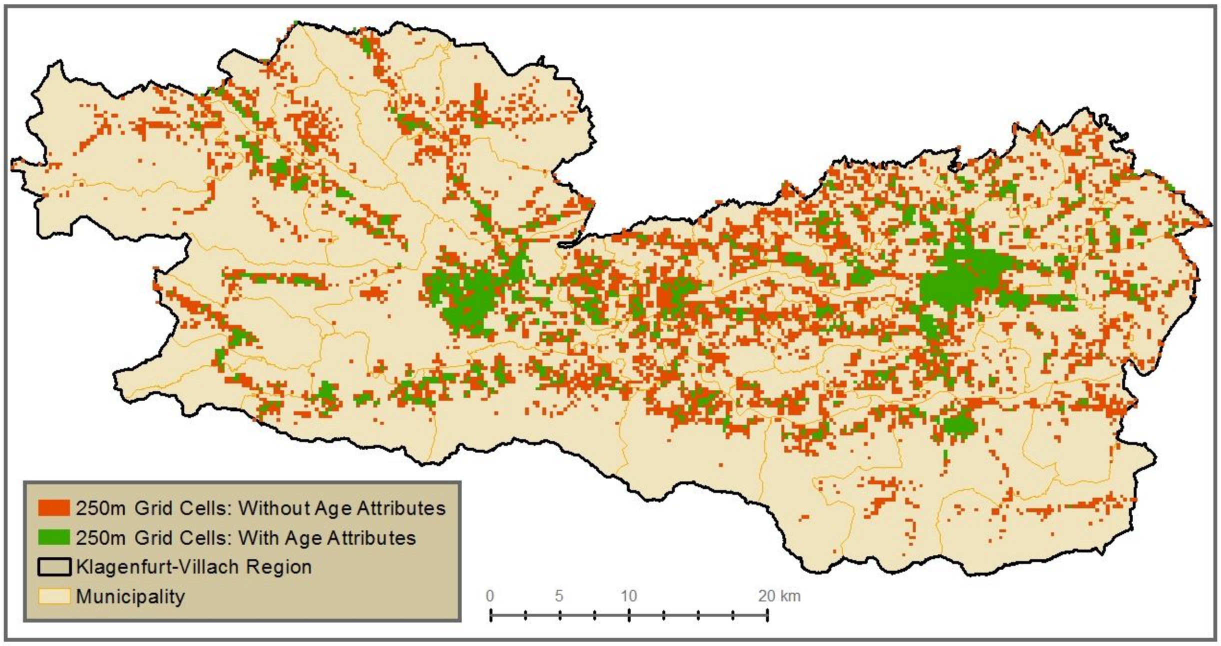
🟥 250m Grid Cells: Without Age Attributes
🟩 250m Grid Cells: With Age Attributes
□ Klagenfurt-Villach Region
□ Municipality
0
5
10
20 km

Figure 4. Data constraints

Strictly for the purpose of comparative analysis in LWLC, the need for accurate age distributions within the source population is of less concern than exact population counts. This is due to the choice of modeling age attributes using the binary method. Unlike LA or OLS, the binary method simply eliminates locations within a source zone where a population attribute (interpolated or actual) would be unlikely (e.g. grid without building attributes). For example, target zone t in source zone s contains no attributes determined to be associated with the presence of population; in which case no population counts from source zone s will be distributed to target zone t. Figure 5 provides a simple illustration of the difference between a disaggregated age attribute criteria produced via the binary method and population density criteria produced using LA or OLS.

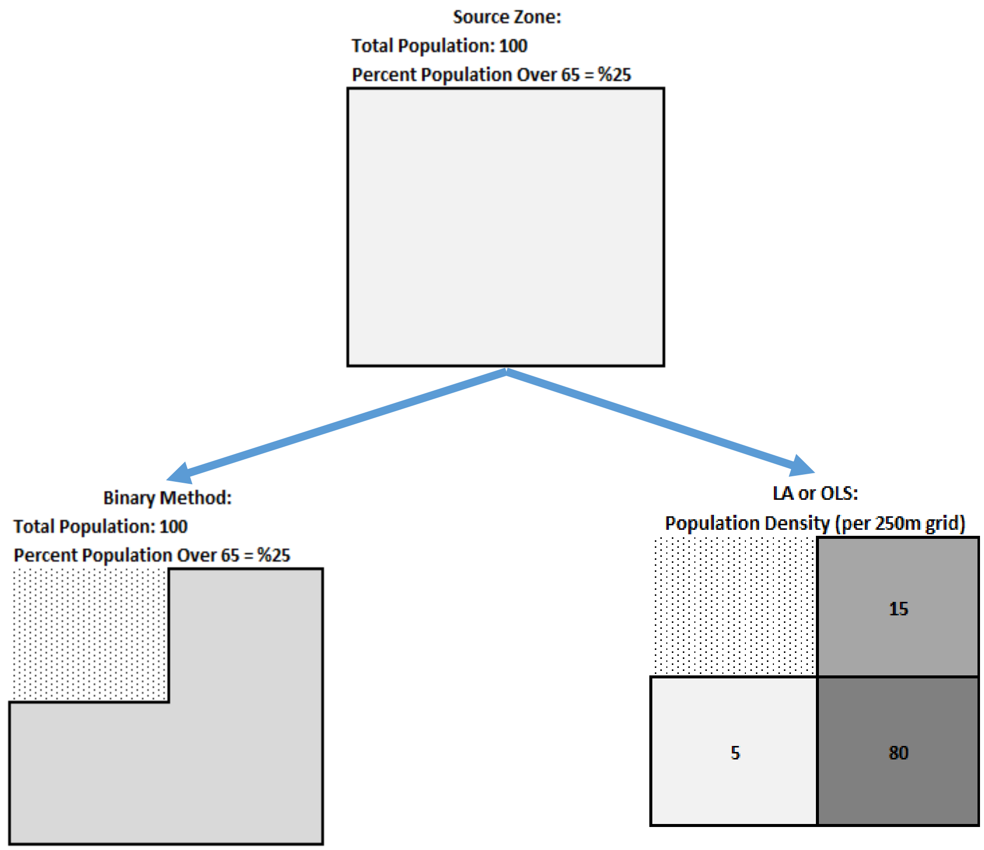
Source Zone:
Total Population: 100
Percent Population Over 65 = %25
Binary Method:
Total Population: 100
Percent Population Over 65 = %25
LA or OLS:
Population Density (per 250m grid)
15
5
80

Figure 5. Binary Method vs. LA and OLS

8.2 OLS Dasymetric Modeling

The initial OLS fitting utilized the building type subclasses outlined in Table 2 as explanatory variables.

Name	Description
OBJ_03	Community building (e.g. retirement homes, dormitories, apprentice centers)
OBJ_04	Hotel, restaurant, pension
OBJ_05	Office building
OBJ_06	Commercial building
OBJ_07	Transportation or telecommunication building
OBJ_08	Industry or storage building
OBJ_09	Building for cultural activities, education, health
OBJ_10	Other building
OBJ_WG	Resident building (Sum of OBJ_1 and OBJ_ 2)

Table 2. OLS explanatory variables

However, this resulted in OBJ_4 (Hotel, restaurant, pension) receiving a negative coefficient value. The use of OLS does have the added simplicity of interpreting coefficients as density estimates per grid cell, as well as the downside of potentially producing negative coefficients for highly locally variable phenomenon. This issue is documented and provides an example where the exact interpretation of OLS coefficients and their intuitive use in dasymetric modelling diverge (Langford, 2006). In these cases, recommended alternatives can be inequality constrained least squares estimation (Moxey & Allanson, 1994) or Poisson regression. Rather than implementing an alternative regression technique, a simplified solution was implemented by removing the *Hotel, restaurant, pension* subtype as an explanatory variable. Two justifications can be made regarding this approach: 1) Hotels, restaurants, and pensions intuitively have a low relationship with nighttime residential census population location, 2) this building subtype accounted for a relatively low percentage (2.5%) of the total number of buildings within the study area. However, the removal of this subtype does incur the cost of zero population allocation to target zones containing only this building subtype.

8.3 Linear Apportionment (LA) Dasymetric Modeling

Two attribute resolutions were used to test the LA method. The first model was implemented using the aggregated building count subtype (OBJ) in the ancillary data feature class. This subtype is the sum count of all building subtypes 1-10 for each 250m target zone. The second LA model iteration utilized the category which accounts for the sum of all reported use-types for every building in the target zone. Recall that a building subtype is a function of the classification of the main property (e.g. commercial or residential). This is not the case for the use-type attributes in the dataset. A single building can contain multiple use-types. These use-type attributes (NTZ_All, NTZ_HWS and NTZ_NWS) are the counts of the number of use-type categorizations within a given 250m grid cell. For example, if a source zone had only one building and that building contained two permanent residency units and one secondary residency unit, the *Building* attribute (OBJ) would have a count of 1 and the *All use-types* attribute (NTZ_ALL) would have a count of 3. Although the building dataset does contain primary residency use-types as a sub category, inspection of the data against the urban structure and satellite imagery of the study area revealed significant sections of residential and urbanized area classified as having zero counts for *primary residency use-type*. For this reason, the *All use-type* (NTZ_ALL) attribute was selected to ensure any habitable use-type would be captured.

9 Dasymetric Results

The following two maps (Figure 5 and 6) allow for a visual comparison between source zone population densities and modeled target zone population densities using the LA use-type method.

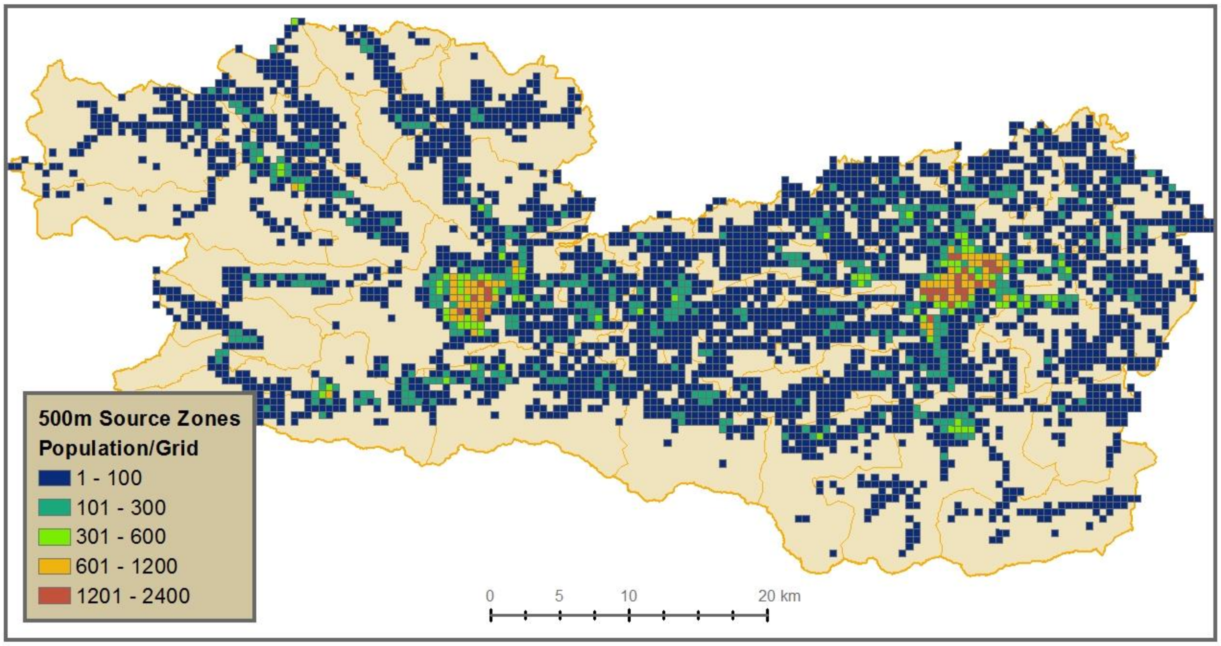

Figure 6: Dasymetric source zones

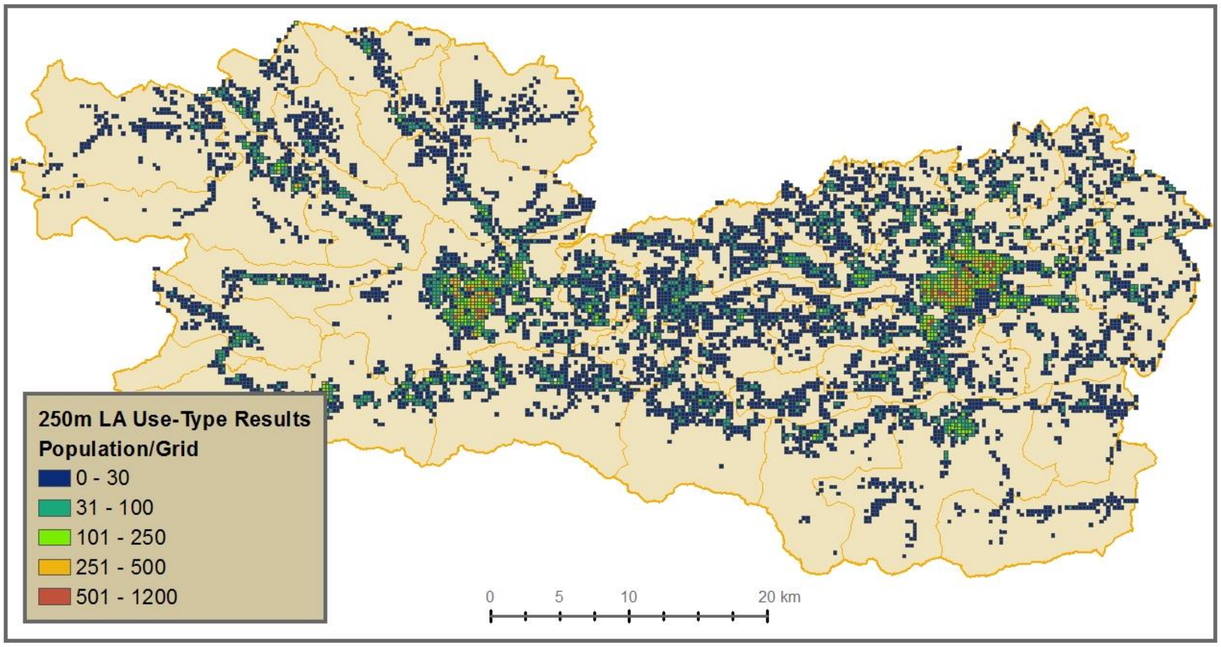

Figure 7: Dasymetric Results: LA using use-type classification counts

Table 3 provides a performance summary of each dasymetric method. In all cases, there is a significant linear relationship between the predicted cell population densities and the actual cell population densities. According to both critical statistics, RMSE and R^2, both LA methods outperform the OLS method. However, LA using all use-types outperformed alternative methods by a large margin. Additionally, it is of value to note the mean over/under count percentages among the three methods. Ranking agreement can be observed between RMSE and R^2, whereas the mean over/under percent error is negligible. This illustrates the primary issue in using this statistic as an absolute or comparative measure of performance.

Method:	RMSE	R^2	Significance	Mean over/under count percent error
LA – Use Type	12.95	.968	p = 0	+4.68%
LA – Building Count	34.26	.777	p = 0	+4.65%
OLS	37.56	.737	P = 0	+4.55%

Table 3. Dasymetric methods performance summary

Error! Reference source not found., Figure 9, Figure 10 provide the scatter plot and regression line for LA use-type, LA using building counts, and OLS respectively. Additionally, Figure 11, Figure 12, and Figure 13 are selected maps focused on Villach and the surrounding communities. The large scale perspective of these maps allow for a better understanding of the geographic distribution of errors between methods.

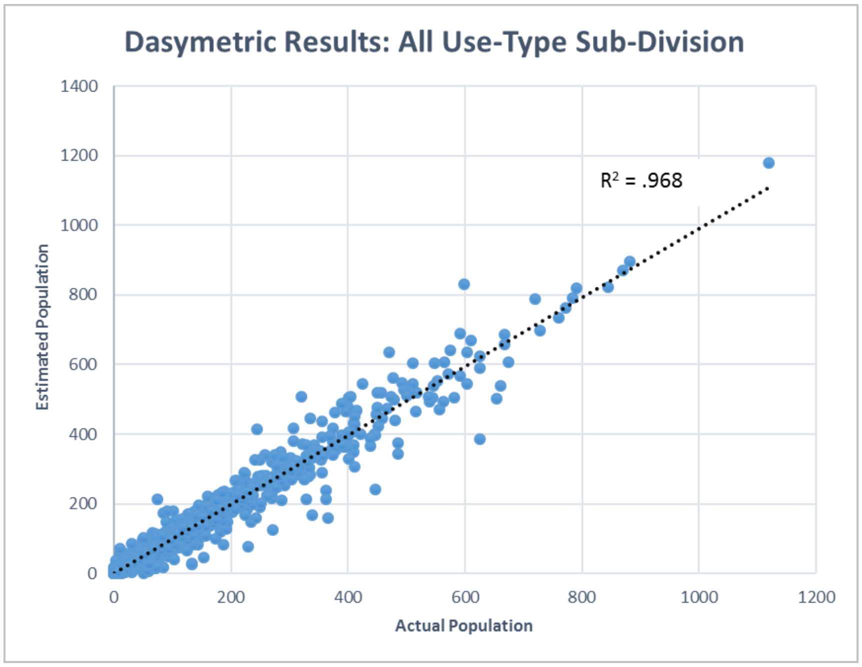

Figure 8. LA use-type scatter plot

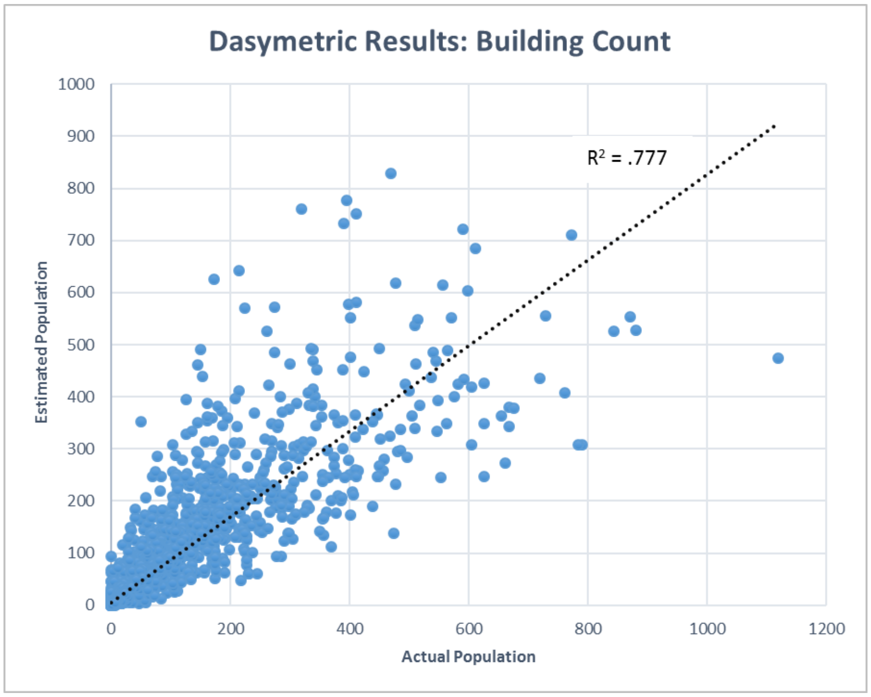

Figure 9. LA building count scatter plot

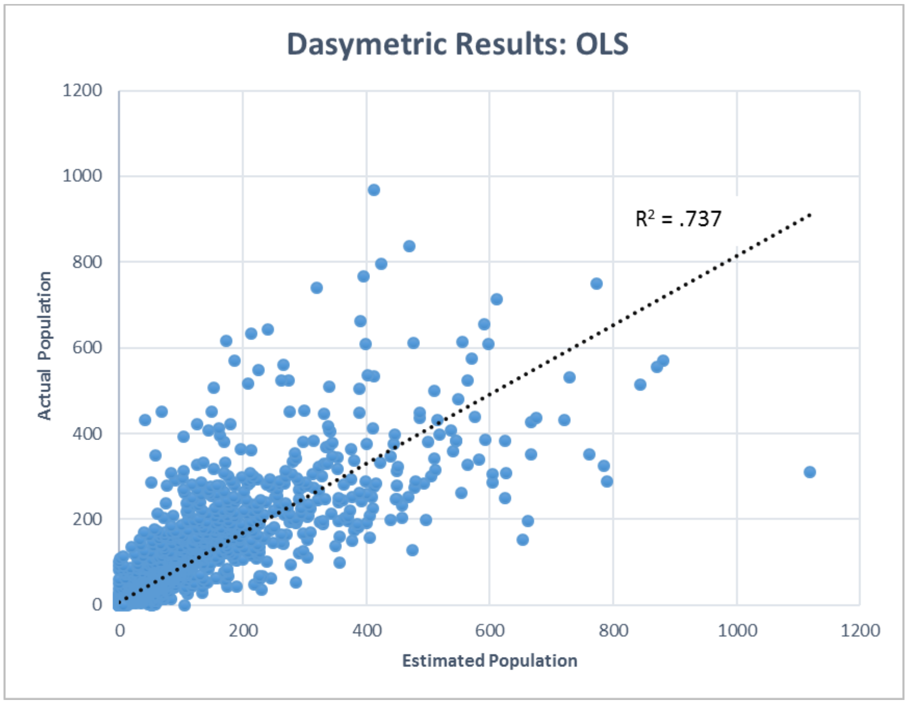

Figure 10. OLS scatter plot

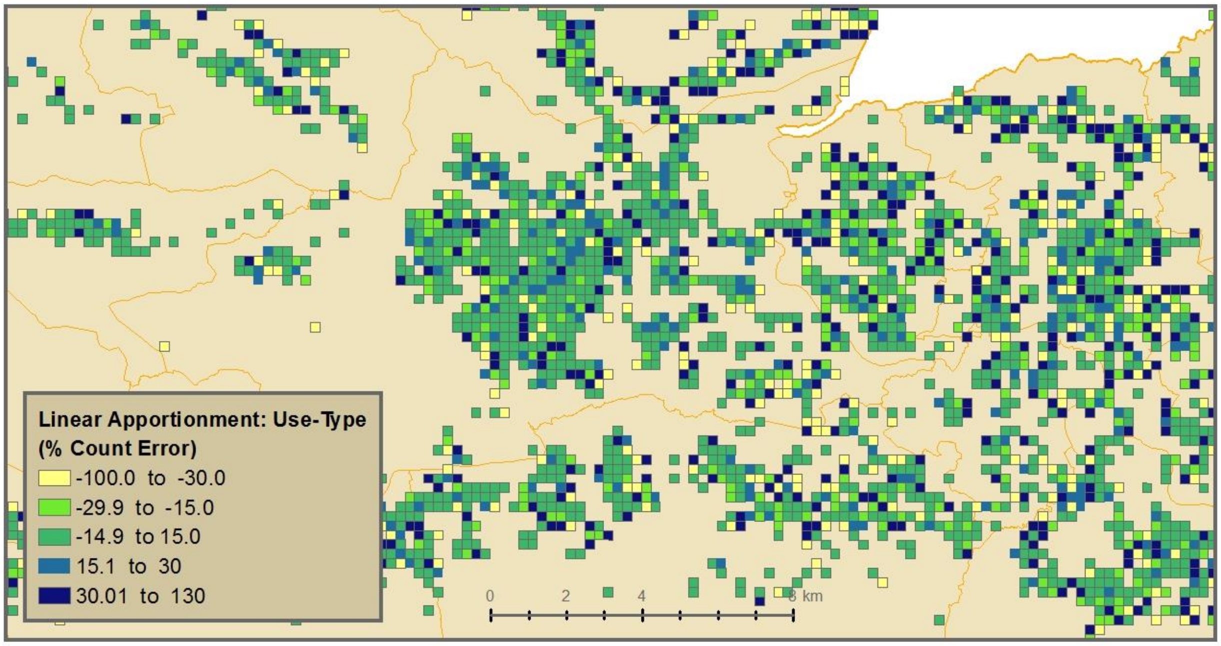

Figure 11. LA use-type map

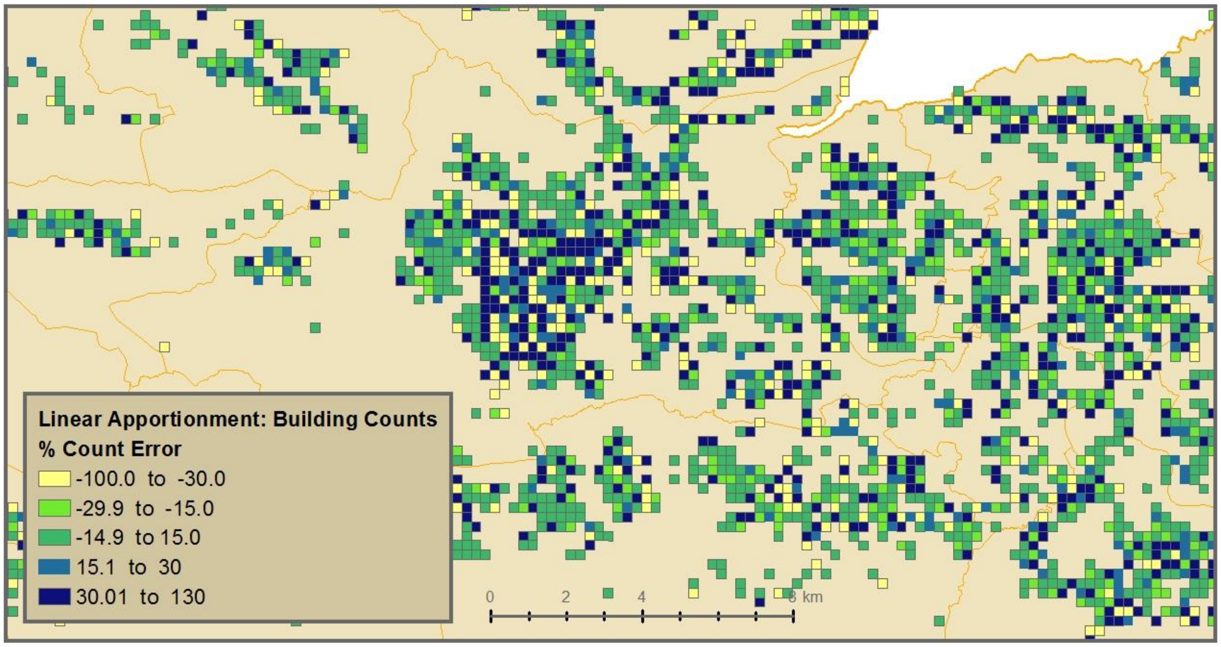

Figure 12. LA building counts map

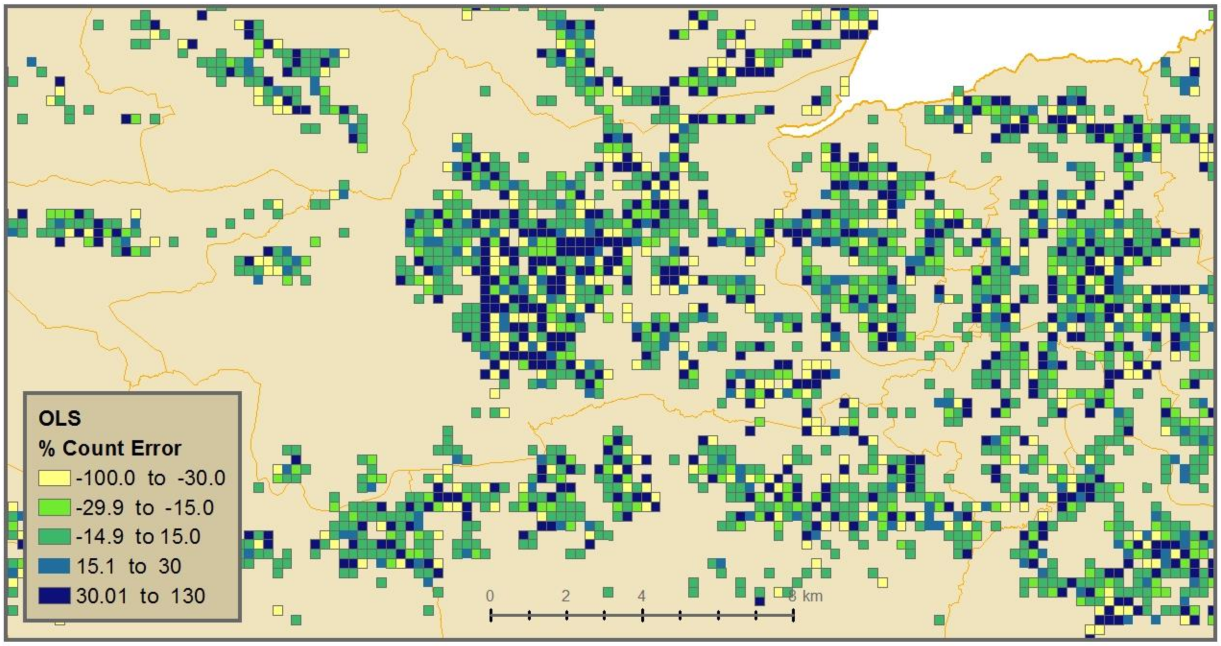

Figure 13. OLS map

An initial visual analysis reveals apparent clustering and a high degree of extreme over estimations and under estimation values in the central Villach area in both the OLS and LA building count methods. A potential contributor for the extremes could be linked to the diversity of building types and associated population densities. LA applies the same disaggregation weight for all buildings regardless of class or use. However, it would be more accurate to apply a higher a higher weight to a class such as an apartment complex versus a single family home. A high degree of diversity in building types and uses within 500m source zones is characteristic of the Villach region thus may be contributing to the pattern observed. The same observation can be made of OLS. Although OLS does distribute weights according to a statistical relationship in the building classes, recall that the predominate residential building classes were combined in order to avoid negative coefficients. Any differentiation between buildings with one unit, or more than one unit was lost in this process. Visual comparison may reveal different degrees of clustering between methods, however the Getis-Ord General G statistic indicate significant spatial clustering of high value errors for all three models.

Based on the performance of LA use-type modeling, this method's output was used as the binary restriction layer for modeling age distribution across the study region. Recall, the binary method (in this context) is simply a restriction on the probable location of an age characteristic. the LA use-type output will be used to eliminate 250m sectors within 500m source zones where a population attribute was not modeled. Additionally, the LA use-type output will serve as the population density criteria layer in the LWLC modeling step.

9.1 LWLC Implementation

Evaluation criteria for study area analysis has been derived using published research results by the World Health Organization (2011) and current research related to public health and influenza vaccination rates in Austria (Kunze, 2013; Hoffman et al., 2016). These studies have found statistically significant correlations between specific age groups and low vaccination rates in Austria. Additionally, population density has been added as an evaluation criteria based on increased transmission risk in densely populated areas (Fang, 2012, Hu, 2013). By combining age and density, not only is the model incorporating the identification of at-risk and low coverage population groups, but also environmental conditions relating to increased risk. The addition of population density provides a further level of discrimination when making decisions about the efficient and effective allocation of resources for social marketing campaigns.

In order to better conform to previously studied age ranges, the original eleven age categories were condensed into five. Global weights were applied evenly across all evaluation criteria. This was done to better illustrate the effect of local criteria weight changes between model iterations. The following is a list of selected evaluation criteria.

Criterion:	Global Weight:	Criterion Type:
Pop. Density (pop./grid)	.20	Benefit/Maximize
Age: 0 - 2 years	.20	Benefit/Maximize
Age: 3 – 19 years	.20	Benefit/Maximize
Age: 45 – 54 years	.20	Benefit/Maximize
Age: 65 years +	.20	Benefit/Maximize

Table 4. LWLC evaluation criteria

The neighborhood scheme chosen for the LWLC model was k-nearest neighbor with the k-value of 19. The choice to apply the value of k=19 was based on preliminary test results in LWLC which produced a large number of cells with no viable local score, an issue which will be discussed in greater detail in the results section.

10 LWLC Results

It was observed that applying progressively smaller k-values for each model iteration produced an increasing number of cells with no viable local score. A variation of this issue is briefly addressed by Carter and Rinner (2014). However, the issue they describe was encountered on a much less impactful scale. Carter and Rinner described the problem as the result of insufficient neighborhood diversity among one or more evaluation criteria. Carter and Rinner encountered only few instances where this occurred, one of which was an island within their study area. One of the reasons the null value outputs were not as problematic in Carter and Rinner’s study as here can be attributed to their use of highly aggregated U.S. Census tract data with large population counts and attribute values. Census tracts are constructed in such a way to ensure they contain between 1200 – 8000 people with an optimum size of 4000 people (U.S. Census

Bureau, 2016). This study utilizes significantly higher resolution data with much smaller population counts and large sections of uninhabited space. The low population counts and non-contiguous nature of the data in this study produces numerous instances where a criterion value will equal zero for all cells within a defined neighborhood; which in turn produces a division by zero error when calculating the local value function. This issue becomes especially evident in the study area when utilizing first-order rook or queen contiguity as the neighborhood definition. In the case of k-nearest neighbor, lower values of k decreased the number of cells within a neighborhood and increased the probability that all cells in that neighborhood would contain 0 values for a given criterion. Carter and Rinner propose two solutions for this problem. Below is a summary of each potential solution and a trade-off analysis for each in the context of this research:

1. **Exclude non-contiguous polygons from the study:** The high resolution, non-contiguous nature of the layers produced by dasymetric modeling created a large number of cells which would be eliminated if this method was used. The number of cells eliminated from the analysis in addition to the high degree of performance achieved in the dasymetric output made this option undesirable. This particular solution essentially penalizes successful dasymetric modeling results and as such was discarded as a viable solution.

2. **Adapt the neighborhood definition (e.g. increase k):** The problem with this solution (if increasing k is the adaptation) can be explained by the observation Carter and Rinner made later in their analysis; as k-increases so too does the spatial autocorrelation of local values. Additionally, as k increases so does the inclusion of criteria values from cells which may have little logical association. Essentially, every integer increase in k brings the LWLC model closer to the value outputs that would be expected in a global WLC solution. Although not optimal, a very large increase in k appeared to be the most feasible solution for this research. It is also worth noting that even with an increase to k=19, 115 cells (500m) and 415 cells (250m) respectively received no local value.

Due to the extent of the study area and formatting constraints, a subsection of the results from LWLC has been selected for visual comparison and analysis. The following three maps focus on the north central portion of the study area, extending from the north-eastern section of Villach to Velden am Wörthersee (Figures 13, 14, 15, and 16).

Figures 13 and 14 allow for a comparison between the value scores produced using 500m and 250m criteria. Red and orange grid cells are associated with high to medium-high local scores (respectively) and are interpreted as locations with comparatively high concentrations of at-risk and low vaccine coverage rate populations. Based upon the criteria chosen for this case study, the high scoring locations would receive prioritized consideration when allocating resources in social marketing campaign efforts. When compared to the 500m results, the 250m cells expose a more detailed picture of the variability and internal performance of specific locations within the study area. Rather than abrupt and/or extreme value changes, a smoother, more representative surface can be observed.

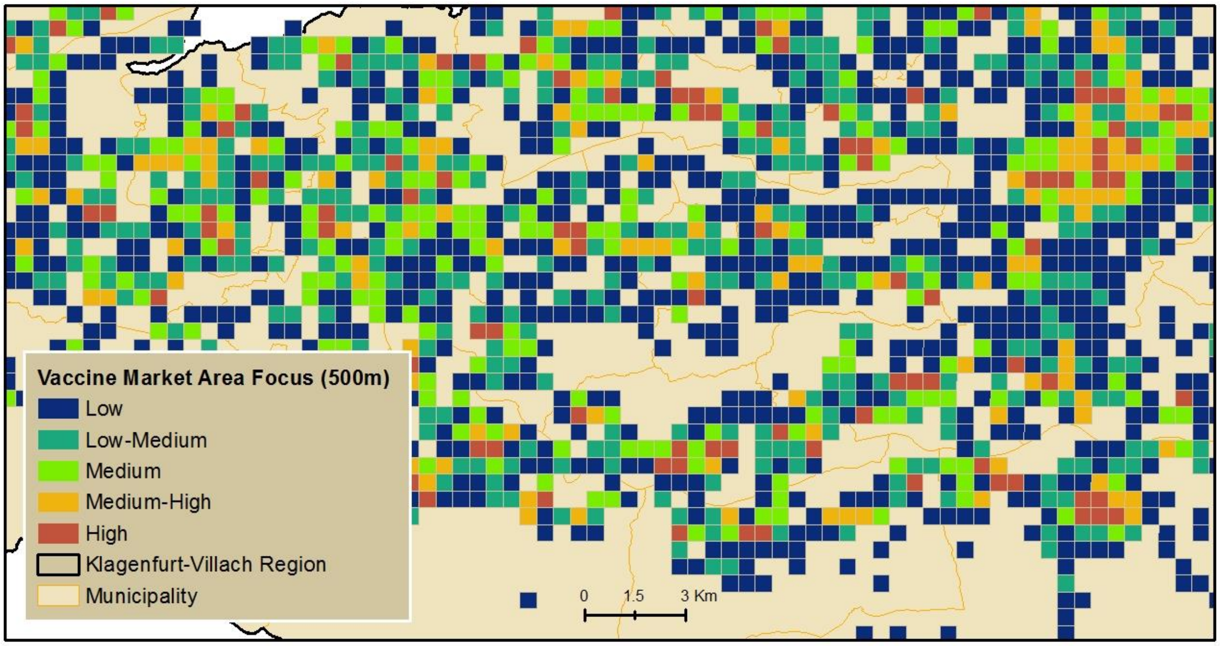

Figure 14. 500m LWLC results

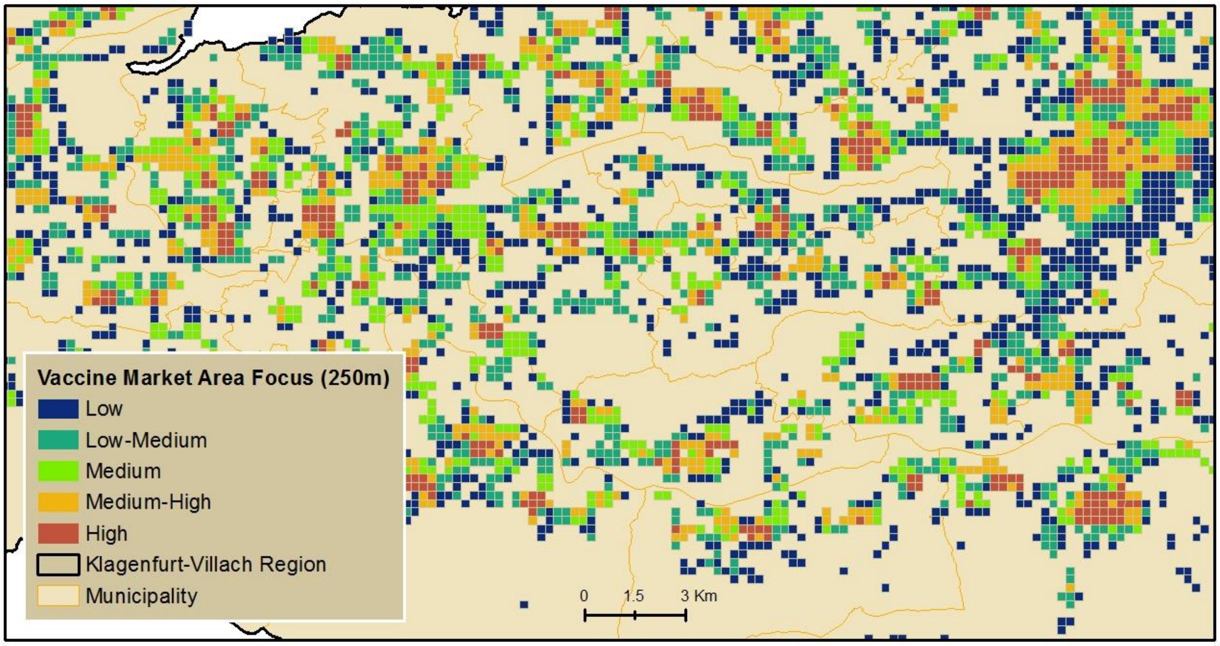

Figure 15. 250m LWLC results

The use of 250m grids also allows for the exclusion of locations which were previously characterized as high-performing in the aggregate data and conversely, the inclusion of locations which were previously attributed with being low performing locations. Figure 15 illustrates a selected location within the study area which exhibits significant underlying heterogeneity in local value ranges. This kind of general improvement in resolution can provide decision makers a far better understanding of the spatial distribution, concentration and variability of at-risk/low-coverage population groups.

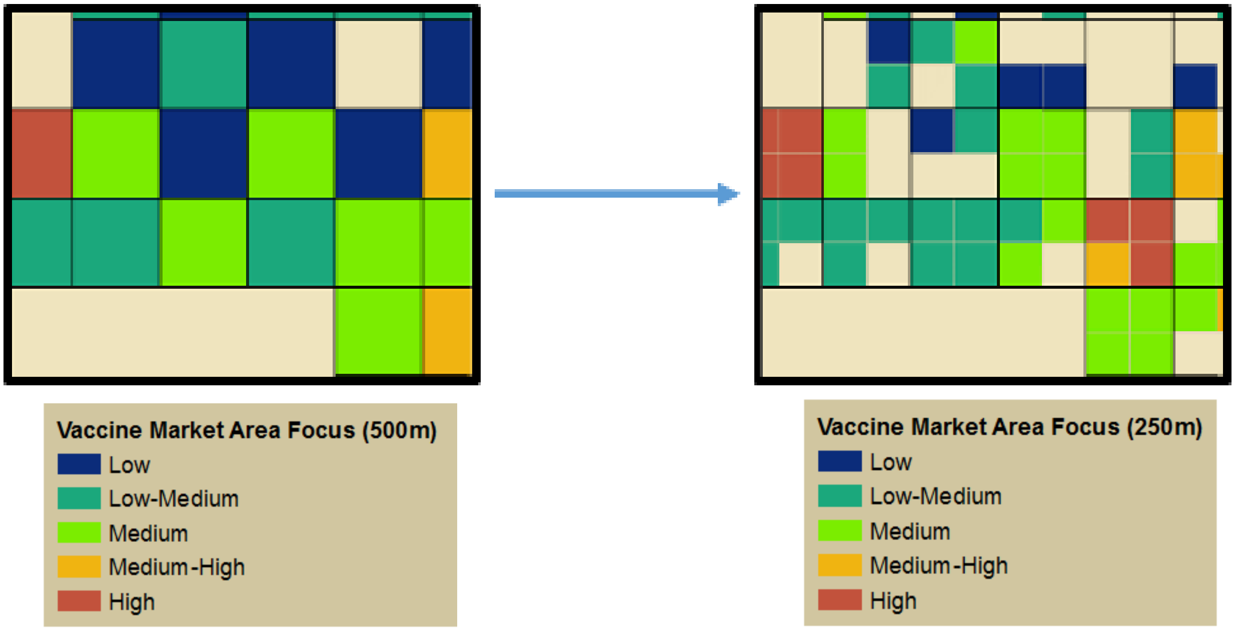

Figure 16. Observed value range change between 500m and 250m grid cells

Figure 16 provides better understanding of the difference in areal extent between the two scale outputs. Using 500m grid cells produces a 88.0 sqkm area (excluding unpopulated zones), with 3519 feasible grid cell decision alternatives. By incorporating the output from LA dasymetric modelling, the area becomes constrained to 51.3 sqkm. This represents a 42.0% reduction in what was previously categorized as feasible area in the original model. Although there is a reduction in feasible area, the higher resolution cells provide an increase in the number of potential decision alternatives. The number of feasible decision alternatives using dasymetrically modelled criteria increases from 3519 to 8204.

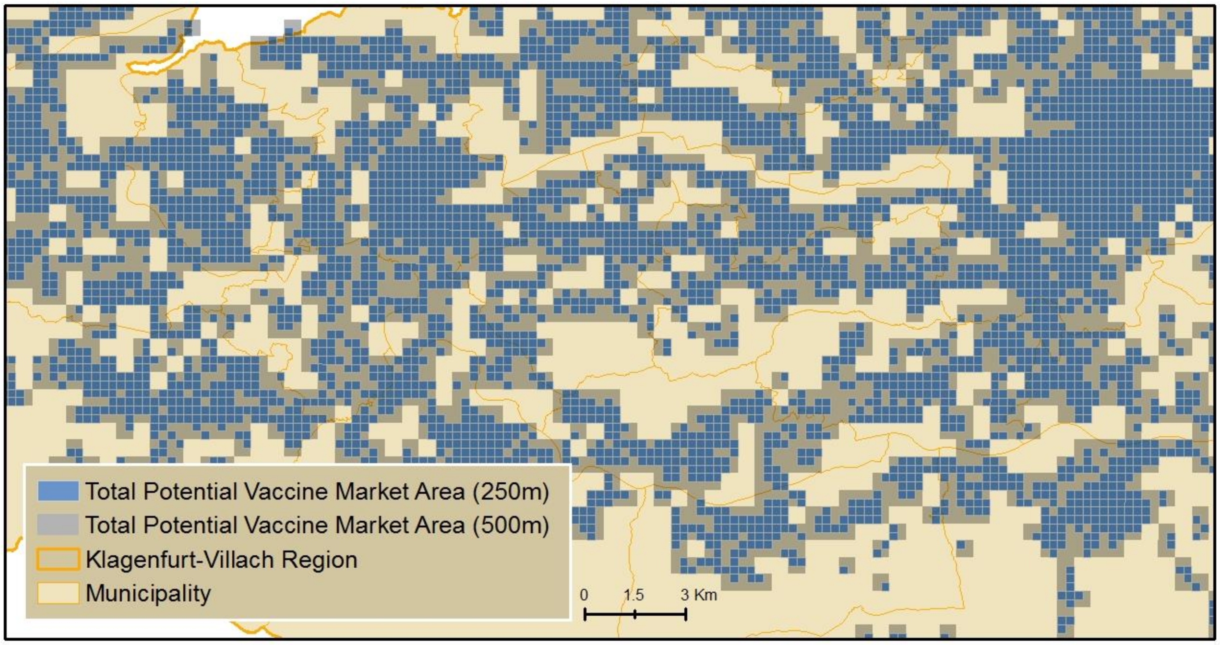
Total Potential Vaccine Market Area (250m)
Total Potential Vaccine Market Area (500m)
Klagenfurt-Villach Region
Municipality
0
1.5
3 Km

Figure 17. Coverage comparison between 500m and 250m LWLC

Furthermore, high resolution modeled data allows for boundary discrimination and directly addresses scale related effects of the MAUP. Figure 17 illustrates three particular scale related problems which can be addressed by dasymetric modeling.

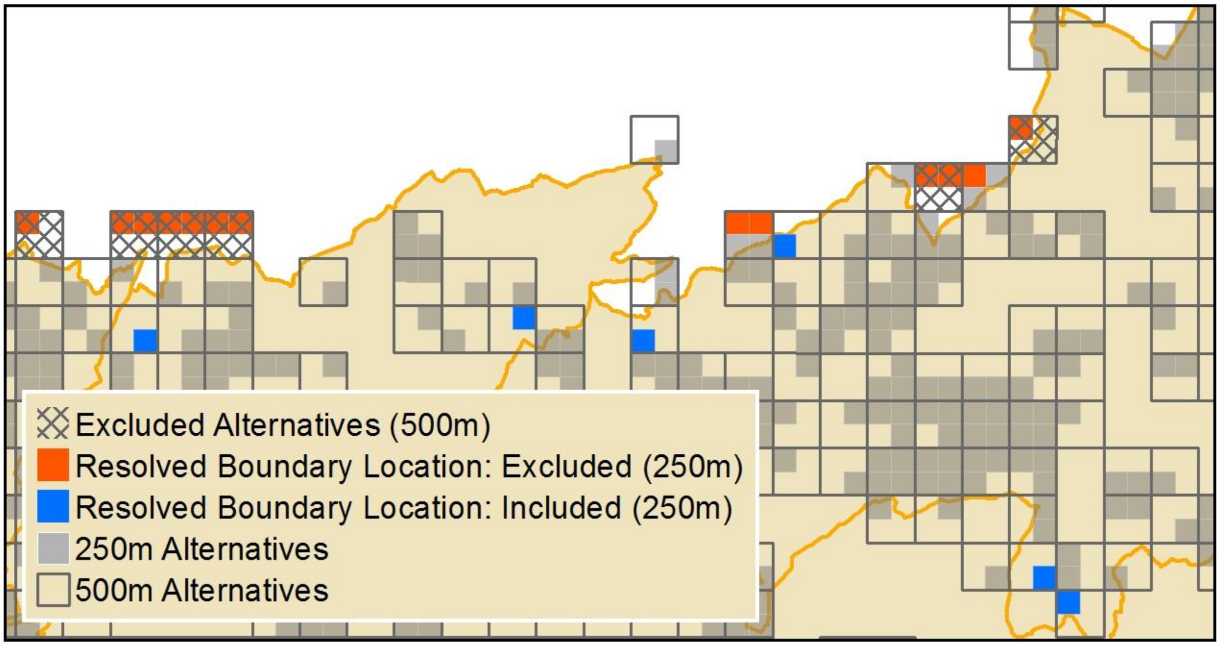
Excluded Alternatives (500m)
Resolved Boundary Location: Excluded (250m)
Resolved Boundary Location: Included (250m)
250m Alternatives
500m Alternatives

Figure 18. Scale related effects and solutions

Red cells are populated locations outside the Klagenfurt-Villach study region which were contributing to the local value calculation at the 500m grid cell level. The crosshatched 500m cells are the cells with values entirely derived from 250m grids outside the study area. Although the most conservative method of selection was used to determine study area inclusion (intersection) for each 500m grid cell, the example in the upper right corner illustrates where even alternative methods such as centroid, or percent area selection would have incorrectly classified the cell as being a potential decision alternative. In total there were 164 grid cells containing 1145 people which were erroneously contributing to local value function results.

Referring again to figure 17, the blue cells represent similar misclassifications when considering study area sub-divisions; in this case municipal boundaries. In situations where resources are allocated based on geographic patters at the sub-region level (e.g. municipality), it is of great value to understand, on which side of a boundary the attribute value belongs. It is certainly true that 250m cells can encounter the same challenge as 500m cells, however the problem is mitigated by disaggregating from 500m to 250m.

The results presented here provide encouraging evidence that evaluation criteria derived from dasymetric modeling can contribute to more informed decisions outcomes in LWLC and aid in mitigating zone related problems associated with the MAUP.

11 Conclusion, Limitations and Future Work

This research tested and evaluated the integration of dasymetric modeling methods with the LWLC choice model. Vaccine related social marketing was identified as a case study for criteria selection and model testing. Three dasymetric models were tested; OLS regression, Linear Apportionment (LA) and the Binary Method. The LA method using total counts of building use-type classifications significantly outperformed results obtained from either LA using building counts or the OLS regression method. Although a LA dasymetric model provides a simple, robust choice for population modeling, access to data such as use-type counts may be infeasible in other applications due to cost or access restrictions. Additionally, further testing should be done to study the impact on LA performance using larger source zone grids (e.g., 1000m).

This research also examined the use of LA model outputs as evaluation criteria in LWLC. Although, dasymetrically derived evaluation criteria show promise for improving decision outcomes by mitigating boundary and scale related effects of the MAUP, challenges with high-resolution, low-diversity criteria values significantly reduced the utility of the dasymetric approach in this study. In order to fully exploit the advantages of high resolution dasymetric criteria, further research is needed to explore alternative implementation of LWLC, which are able to account for low diversity or isolated data without unreasonable neighborhood expansions or data deletion.

Acknowledgements

I express my great appreciation and gratitude for the generous support of The Austrian Marshall Plan Foundation, the Department of Geoinformation at Carinthia University of Applied Science, and Dr. Gernot Paulus. I especially would like to recognize and thank Dr. Piotr Jankowski. As my thesis committee chair and mentor, Dr. Jankowksi has continually provided valuable insight, guidance and enthusiasm throughout the project.

Financial support for this work was provided by the Austrian Marshall Plan Foundation. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Marshall Plan Foundation.

References

Aubrecht, C., Steinnocher, K., Hollaus, M., & Wagner, W. (2009). Integrating Earth Observation and GIScience for High Resolution Spatial and Functional Modeling of Urban Land Use. *Computers, Environment and Urban Systems, 33*, 15-25.

Austrian Embassy. (2016, May 4). *Health Care*. Retrieved from Austria.org: http://www.austria.org/healthcare/

Bell, N., & N, S. (2010). GIS and Injury Prevention and Control: History, Challenges, and Opportunities. *International Journal of Environmental Research and Public Health*(7), 1002-1017. doi:10.3390/ijerph7031002

Bell, N., & Schuurman, N. (2010). GIS and Injury Prevention and Control: HIstroy, Challenges, and Opportunities. *International Journal of Environmental Research and Public Health*, 1002-1017.

Carter, B., & Rinner, C. (2013). Locally Weighted Linear Combination in Vector Geographic Information Systems. *Journal of Geographical Systems, 16*(3), 343-361. doi:10.1007/s101090130194-3

Carver, S. (1991). Integrating Multi-Criteria Evaluation with Geographical Information Systems. *International Journal of Geographical Information Systems, 5*(3), 321-339.

Cowen, D. (1988). GIS versus CAD versus DBMS: What are the differences. *Photogrammetric Engineering and Remote Sensing, 54*, 1551-1555.

Dungan, J., Perry, J., Dale, M., Legandre, P., Citron-Pousty, S., Fortin, M., & Rosenberg, M. (2002). A Balanced View of Scale in Spatial Statistical Analysis. *Ecography, 25*(5), 626-640.

Eastman, J., Kyem, P., Toledo, J., & Jin, W. (1991). *GIS and Decision Making.* Genva: UNITAR.

Eicher, C., & Brewer, C. (2001). Dasymetric Mapping and Areal Interpolation: Implimentation and Evaluation. *Cartography and Geographic Information Science, 28*(2), 125-138.

Erlacher, C., & McNew, G. (n.d.). This is a title.

Fang, L., Wang, L., Sake, V., Liang, S., Tong, S., Li, Y., . . . Cao, W. (2012). Distribution and Risk Factors of 2009 Pandemic Influenza A (H1N1) in Mainland China. *American Journal of Epidemiology*.

Fischer, G. (1995). Range Sensitivity of Attribute Weights in Multi Attribute Evaluation Models. *Organizational Behaviour and Human Decision Processes*, 62:252-66.

Fisher, P., & Langford, M. (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo Simulation. *Environment and Planning, 27*, 211-224.

Fisher, P., & Langford, M. (1996). Modelling sensitivity to accuracy in classified imagery: A study of areal interpolation. *The Professional Geographer, 48*, 299-309.

Flowerdrew, R., & Green, M. (1989). Statistical methods for inference between incompatible zone systems. In M. Goodchild, & S. Gopal (Eds.), *The Accuracy of Spatial Databases* (pp. 239-247). Landon, England: Taylor and Francis.

Flowerdrew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *Annals of Regional Science, 26*, 67-78.

Fotheringham, A., & Wong, D. (1991, July). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *23*(7), 1025-1044. doi:10.1068/a231025

Gehlke, C., & Biehl, K. (1934). Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association, 29*(185), 196.

Goodchild, M., & Lam, N. (1980). Areal Interpolation: Variant of the Traditional Spatial Problem. *Geo_Processing*, 297-312.

Gregory, I. (2002). The accuracy of areal interpolation techniques: standardizing 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems, 26*, 293-314.

Hoffman, K., Paget, J., Wojczewski, S., Katic, M., Maier, M., & Soldo, D. (2016). Influenza vaccination prevalence and demographic factors of patients and GPs in primary care in Austria and Croatia: A cross-sectional comparative study in the framework of the APRES project. *European Journal of Public Health*. doi:10.1093/eurpub/ckw006

HU, H., Nigmatulina, K., & Eckhoff, P. (2013). The scaling of contact rates with population density for the infectious disease models. *Mathematical Biosciences, 244*(2), 125-134.

Jankowski, P. (1995). Integrating Geographical Information Systems and Multiple Criteria Decision-Making Methods. *International Journal of Geographical Information Systems, 9*(3), 251-273.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton, NJ: Princeton University Press.

Kunze, U. (2008). Austria - Resistent Against influenza Control. *Facharzt, 4*, 26-29.

Kunze, U., Bohm, G., & Gronman, E. (2013). Influenza vaccination in Austria from 1982 to 2011: A country resistant to influenza prevention and contro. *Vaccine, 31*(44), 5099-5103.

Kunze, U., Gronman, E., Bohm, G., & Kunze, M. (2007). Kunze, U., Groman, E., Böhm, G., & Kunze, M. (2007). Influenza vaccination in Austria, 1982 2003. *Wiener Medizinische Wochenschrift Wien Med Wochensch, 157*(5-6), 98-101. doi:10.1007/s10354-007-0389-7

Lam, N. (1983). Spatial Interpolation Methods: A Review. *Cartography and Geographic Information Science, 10*(2), 129-150.

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

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

Leyk, S., Nagle, N., & Buttenfield, B. (2013). Maximum Entropy Dasymetric Modeling for Demographic Small Area Estimation. *Geographic Analysis, 45*(3), 285-306.

Lu, G., & Wong, D. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers and Geosciences, 34*(9), 1044-1055.

Maantay, J., Maroko, A., & Herrman, C. (2007). Mapping Population Distrobution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science, 34*(2), 77-102.

Malczewksi, J. (1999). *GIS and multicriteria decision analysis.* New York: J. Wiley and Sons.

Malczewski, J. (2000). On the use of weighted linear combination method in GIS: Common and best practice approaches. *Transactions in GIS, 4*, 5-22.

Malczewski, J. (2006). GIS-based multicriteria decision analysis: A survey of the literature. International Journal of Geographical Information Science. *International Jounral of Geographical Information Science, 20*(7), 703-726.

Malczewski, J. (2011). Local Weighted Linear Combination. *Transaction in GIS, 15*(4), 55-439.

Malczewski, J. (2015). *Multicriteria decision analysis in geographic information science.* New York: Springer.

Mennis, J., & Hultgren, T. (2006). Intelligent Dasymetric Mapping. *Cartography and Geographic Information Science, 33*(3), 179-194.

Moxey, A., & Allanson, P. (1994, September 8). Areal interpolation of spatially extensive variables: A comparison of alternative techniques. *Internation Journal of Geographical Information Systems*, 479-487.

Nagle, N., Buttenfield, B., Leyk, S., & Spielman, S. (2013). Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers, 104*(1), 80-95.

Niemeyer, A., & Niemeyer, H. (2008). Apportionment Methods. *56*(2), 240-253.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. In *Concepts and Techniques in Modern Geographyy Number 38.* Norwich: Geo Books.

Openshaw, S., & Taylor, P. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical Applications in the Spatial Sciences* (pp. 127-144). London: Pion.

Petrov, A. (2008). Setting the Record Straight: On the Russian Origins of Dasymetric Mapping. *Carographica: The International Journal of Geographic Information and Geovisualization, 43*(2), 133-136.

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

Poulsen, E., & Kennedy, L. (2004). sing Dasymetric Mapping for Spatially Aggregated Crime Data. *Journal of Quantitative Criminology, 20*(3), 243-262.

Preobrazhensky, A. (1954). Dorevolutsionnye i Sovietskie Karty Razmescheniya Nasieleniya [Prerevolutionary and Soviet Maps of Population Density]. *Voprosy Geografii: Kartografia, 34*, 134-49.

Qiu, F., & Cromley, R. (2013). Areal Interpolation and Dasymetric Modeling. *Geographical Analysis, 45*(3), 213-215.

Reibel, M., & Agrawal, A. (2007). Areal Interpolation of Population Counts Using Pre-classified Land Cover Data. *Population Research and Policy Review, 26*(5), 619-633.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234-240.

Tobler, W. (1979, September). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*, 519-30.

Tomczak, M. (1998). Spatial Interpolation and its Uncertainty Using Automated Anisotropic Inverse Distance Weighting (IDW) - Cross-Validation/Jackknife Approach. *Journal of Geographic Information and Decision Analysis, 2*(2), 18-30.

US Census Bureau. (2016). *Geographic Terms and Concepts - Census Tract*. Retrieved from Census.gov: www.census.gov/geo/reference/gtc/gtc_ct.html

Voss, S. (2016). *MCDA ArcMap*. Retrieved from Codeplex: https://mcda4arcmap.codeplex.com/

WHO. (2016). *Public health in Austria. An analysis of the status of public health.* Retrieved from www.euro.who.int/en/publications/abstracts/public-health-in-austrai-an-analysis-of-the-status-of-public-health

Wong, D. (2004). The modifiable areal unit problem (MAUP). In D. Janelle, K. Warf, & K. Hanson (Eds.), *Geographical Perspectives on 100 Problems* (pp. 571-575). WorldMinds.

Wright, J. (1936). A Method of Mapping Densities of Population: With Cape Cod as an Example. *Geographical Review, 26*(1), 103-110. doi:10.2307/209467

Zandbergen, P. (2011). Dasymetric Mapping Using High Resolution Address Point Datasets. *Transactions in GIS, 15*(s1), 5-27.

Zandbergen, P., & Ignizio, D. (2010). Comparison of dasymetric mapping techniques for small. *Cartography and geographic Information Science, 37*, 199-214.