university of groningen

faculty of spatial sciences

Overcoming the Modifiable Areal Unit Problem (MAUP) of socio-economic variables in real estate modelling

Abstract

This thesis develops a new methodology for associating spatial zoning characteristics to point features. Doing so, it aims to provide an applicable approach that can be used to measure location characteristics more accurately. While doing so it should reduce bias caused by the Modifiable Areal Unit Problem. Whereas traditional methods use statistical areas directly to associate spatial phenomena to point features, our suggested approach advocates for disaggregation of spatial data and utilizing optimal zoning systems created around individual properties in order to obtain functional statistical areas. By means of a case study, in which prices of 3287 commercial real estate properties across western Europe were estimated using multiple socio-economic indicators measured through varying statistical areas, it was found that our suggested approach could improve model accuracy up to 18%. Furthermore, functional statistical areas enable researchers to determine the spatial scales and shapes at which location characteristics should be measured. This should provide more control over data and allow for better comparable results.

Keywords

MAUP · Socio-economics · Location characteristics · Hedonic valuation · Optimal zoning systems

Author: P.L.Weenink
Supervisor: Prof.dr. J. van Dijk
Assessor: dr. M.N. Daams
E-mail: p.l.weenink@student.rug.nl
Date:March 2, 2022

university of groningen

faculty of spatial sciences

University of Groningen

Overcoming the Modifiable Areal Unit Problem (MAUP) of socio-economic variables in real estate modelling

Master's Thesis

To fulfill the requirements for the degree of
Master of Economic Geography
at University of Groningen under the supervision of
Prof. dr. J. van Dijk (Faculty of Spatial Sciences, University of Groningen)
and
dr. M.N. Daams (Faculty of Spatial Sciences, University of Groningen)

P.L.Weenink(s3246620)

March 2, 2022

Disclaimer

Master theses are preliminary materials to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the author and do not indicate concurrence by the supervisor or research staff.

Contents

List of Figures

List of Tables

Introduction

Statistical areas

Aggregation of spatial data to zones is common practice in numerous fields. Statistical areas are used for reporting purposes; to calculate means, variances and correlations; fit regressions; or carry out other complex tasks (Charlton 2009). Doing so, statistical areas serve a wide range of academic and social purposes. The European commission uses NUTS regions for regional policies (Eurostat 2015), statistical bureaus have spatial grids to provide detailed information while guaranteeing privacy (Leeuwen and Venema 2021) and planners use city districts for planning practice (Nouvel et al. 2015). Aggregation of data into statistical areas is necessary for "scaling-up" (Bian and Butler 1999), protecting confidentiality of personal data, reducing volume, and identifying geographical patterns (Openshaw and Alvanides 2001). Also within the field of economic geography, statistical areas are frequently used as it allows for easy incorporation of spatial phenomena in statical models (Anderson 2012).

Nevertheless, use of these statistical areas should be considered carefully. Like any other statistical dataset, statistical areas should reflect the population of interest adequately in order to obtain representative outcomes. Whereas this population of interest is clear when zones are considered as stand alone phenomena, matters become more complicated when location effects on specific features are of interest. In such a scenario, statistical areas should reflect the spatial extend at which location characteristics affect these features. However, this spatial extend is not fixed. Features can be affected by alternative location characteristics that differ in magnitudes, directions and significance at varying ranges (Chica-Olmo et al. 2019; van Duijn et al. 2016; Daams et al. 2016). Although a general consensus exist that nearby location characteristics have a stronger relation to the feature of consideration than distance characteristics (Sui 2004), there is no universal range that tells us by which location characteristics should be considered. Moreover, this distance is frequently relative and could thus vary per feature of consideration, location characteristic or geographic location (Evans 2008).

Despite the complicated nature of locations and their impact on specific features, a relation between location characteristics and features of interest is frequently sought at a single fixed scale. To do so, researchers and analysts often make use of the Standard Administrative Units (SAUs) in which spatial data is available. These cover a variety of classifications such as metropolitan areas, NUTS, municipalities, provinces, states and countries (e.g. Watts 2009; Alfieri et al. 2016; Otto et al. 2003; Xiong et al. 2020; Sundquist et al. 2011). Although such regions are widely recognized and easy to use, researchers should ask themself if these SAUs adequately represent the location characteristics of interest. Common practice is directly assigning location characteristics to features of interest based upon the SAU in which these are located. Doing so, researchers implicitly assume that these SAUs indeed reflect the spatial extend by which location characteristics affect the features of consideration. Nevertheless, justification on why these SAUs should provide an adequate measure for the feature of consideration are often failed to be mentioned (e.g. Fuerst and Haddad 2020; Grimes and Aitken 2007; Roebeling et al. 2017). This lack of justification is striking as many of these assumptions are unlikely to hold in practice. The direct association of location characteristics through SAUs fails to capture underlying differences of locations within zones. Furthermore, neighbouring areas that could have significant impact on features remain unrecognized. Surely, researchers can cover for this by acknowledging that they examine the effects of location characteristic per SAU. Nevertheless, if these SAUs have no meaningful relation to the feature of consideration one might wonder what the added value of these outcomes are.

Moreover, the direct use of these SAUs make research outcomes vulnerable to the Modifiable Unit Areal Problem (MAUP). This problem entails that aggregating data to altering zones could lead to different research outcomes (Wong 2009). This altering of zones occurs when different SAUs specification are used to conduct a similar research. However, also within a single SAU specification zoning inconsistencies occur in shape and size. Furthermore, SAU specifications are prone to changes over time. Zoning alterations can be made within SAU specifications for a number of reasons, such as political boundary changes. Examples are the NUTS classifications that differ for 2013, 2016 and 2021 (GISCO n.d.). These inconsistencies make it particularly hard to compare research outcomes across geographies, SAU specifications and over time. For such a comparison the question remains if differences are caused by actual changes in preferences over space and time, or that it is merely due to alterations in zoning measures used.

An example of how improper use of SAUs could affect research outcomes is provided in figure 1. Here A, B and C represent SAUs; person icons represent population with blue and red characteristics; and building icons are features of interest. According to the traditional method, location characteristics are assigned to features, based upon the SAU in which they are located. Therefore, buildings in 1a are assigned an absolute population of eight, zero and three, for zones A, B and C respectively. For the ratio measure of blue colour per capita, this is $\frac{3}{8}$, $\frac{0}{0}$ and $\frac{2}{3}$.

A couple of things could be noticed. First of all, remarkable differences in associated location characteristics occur along border regions. For the lower two buildings associated values are two extremes, despite their proximity. Second, features within zones are treated exactly equal despite their location within these zones. For the top two buildings, assigned location characteristics are the same, even though these features are remote. Third, the nearby population of zone C is not considered for any of the lower two features. Especially for the property in the middle this is remarkable as these location characteristics are closer by than any of the characteristics that this feature is associated with. Fourth, the building in zone B is associated with a ratio of $\frac{0}{0}$. This is an incomputable ratio, and thus this feature is at risk of being omitted. Fifth, the shape of these zones are inconsistent. If only boundaries would be slightly different, assigned values could change significantly. At last, the sizes of these zones are inconsistent. Therefore, the larger zone of A is more likely to cover a high population. Due to the law of large numbers (Kruse 1987) this would

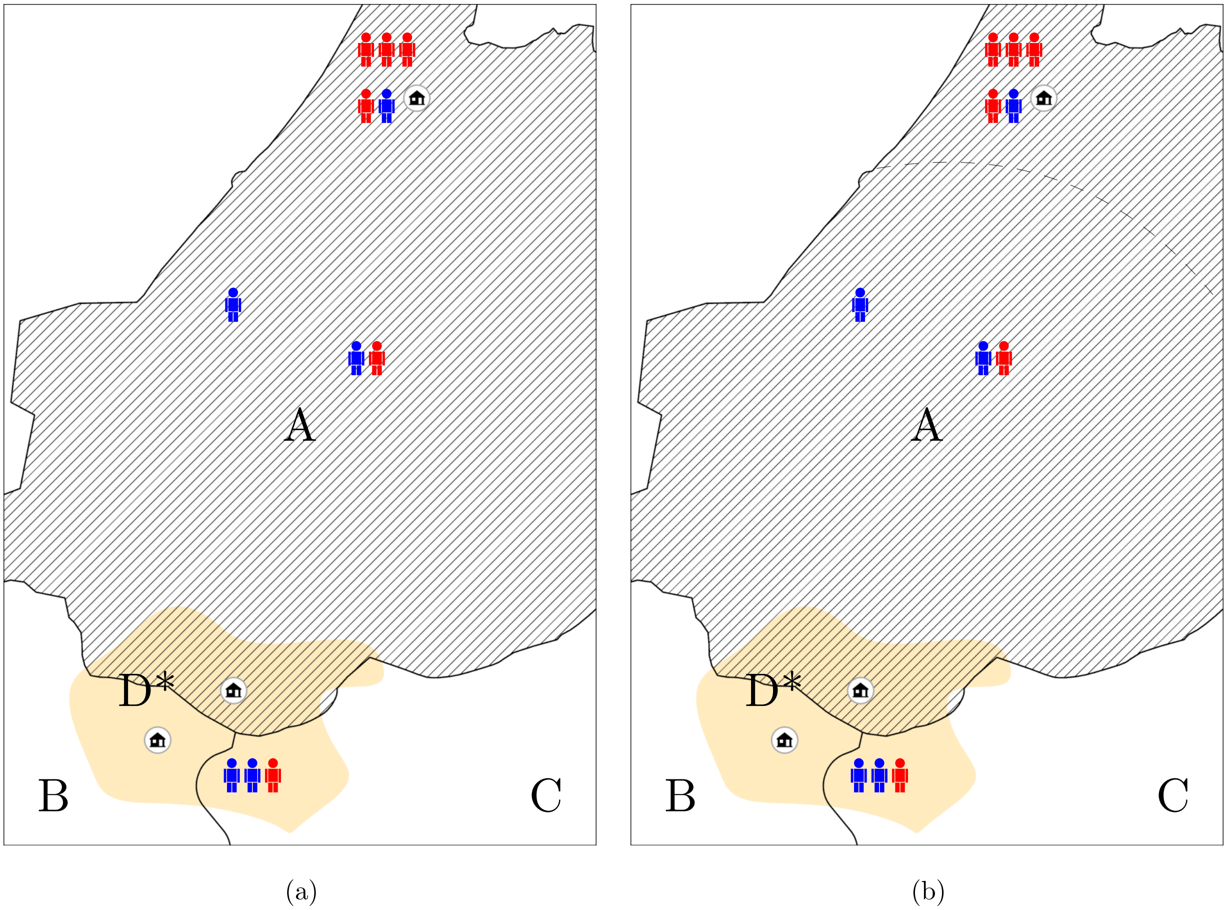


Figure 1: Measurement error can occur when spatial characteristics are wrongly associated to point features

also affect ratio values such as population characteristics, which will have a lower variance for these zones.

1b Illustrates a scenario in which areal units are indeed modified. In here, all information is equal but zone A is split into two separate areas. This single alteration has two main effects: an alteration in shape and a reduction of size for zone A. Doing so this alteration not only affects the assigned location values of features in zone A, but also changes the relative values of features in other zones.

It can be seen that the assigned location characteristics for the top two buildings change significantly. Since the zoning alteration caused these features to be located in separate zones, assigned location values no longer have to be equal. Therefore, the size and magnitude of change can vary per feature and characteristic of consideration. For the top building, assigned population and the ratio of blue population change from eight to five and $\frac{3}{8}$ to $\frac{1}{5}$. For the building in the middle, these are 8 to 3, and $\frac{3}{8}$ to $\frac{2}{3}$. It could be noticed that the magnitude of change is stronger for the building in the middle. Moreover, the direction of change differs. Whereas the ratio of blue population decreases for the top building, it is the other way around for the building in the middle.

The modification of areas also affects relative values. Due to the separation of zone A, two smaller areas are created that are more equal in size to zones B and C. Consequently, the associated populations are now more similar among zones. The opposite happens for ratios. Since populations used to calculate population characteristic rations are now smaller, variance increases. Therefore, rates are more likely to differ from the total mean.

Functional Statistical Areas

Despite the many shortcomings of assigning location characteristics directly to features by the SAUs in which these are located, it remains a frequently applied method in varying disciplines. A likely explanation for the popularity of this method is ease of use. As explained, SAUs commonly follow political boundaries that are widely recognized. The assigning of location characteristics to features based upon these zones is a relatively easy process. All that is required to perform such a join is a common denominator. Even when this denominator is missing, a spatial join can be easily performed with GIS software. Since SAUs are already defined, researchers and analysts are not bothered with considering the spatial extend of location characteristics, and can start their analysis right away. Moreover, data is frequently only limitedly available. Especially when analyses cover a large spatial extend over a wide timespan, researchers often have to rely on data that is only available as SAUs.

To provide a solution for this dilemma between ease of use, and obtaining meaningful outcomes, this thesis explores the use of “Functional Statistical Areas” (FSAs). These functional areas utilize the commonly available SAUs, and adopt these in such a way that results are more robust, easy to compare and better reflect the spatial extend by which features of interest are affected. FSAs are created for features individually, and are shaped around them. Doing so they should overcome, or at least reduce, the shortcomings that were established for the traditional method in which SAUs are directly used for assigning location characteristics.

Since FSAs are created around the features of interest, exact locations remain utilized. Therefore, actual location characteristics should be reflected more properly. Because FSAs are not restricted by existing boundaries, it can include location characteristics of adjacent SAUs while neglecting those that fall within the same SAU. To be applicable in a wide range of situations, functional statistical areas should be flexible and provide researchers more control over their data. As such, they should be easy adaptable to serve different scenarios.

An example of such a functional statistical area is zone D* in figure 1. This area is shaped around the building in the middle, and only includes location characteristics that fall within its extend. Therefore, it does include the nearby population in zone C, but no longer considers the far away population of zone A. The assigned ratio of population with a blue characteristic becomes

$\frac{2}{3}$. Since this functional area is not affected by SAU boundaries, changes in SAU specifications do not alter outcomes. When figure 1a is compared to 1b, it can be seen that assigned location characteristics remain equal.

The difficulty of creating a proper functional statistical area lies in the following points. First, when data is obtained as SAU, only total and mean values per zone are available. Therefore, actual location characteristics remain unclear. In our example this would mean that we only have the values of zone A, B and C, but do not posses the information provided by the person icons. Second, different location characteristics can have varying effects that alter per feature of consideration and distance. Therefore, the optimal shape and size of D* is unknown, and should be established per scenario. Third, since location characteristics of consideration are no longer bound by SAUs, additional data might be required of neighbouring zones. However, this should only occur if our functional statistical areas overlap SAUs that were previously not covered by features of interest.

Research setup

In order to test how such functional statistical areas could be created, and how these perform compared to traditional SAUs, a case study will be executed. In this case study, we will examine how socio-economic location characteristics affect commercial real estate (CRE) prices. CRE should provide an ideal test case. Property prices are highly affected by location characteristics; multiple types of features can be examined in one go; CRE covers a large geographical area; and a large scientific base of real estate research exists that allows us to test the performance of our modified zones.

Moreover, utilizing location characteristics to measure point data, is common practice in the discipline of real estate (Gelfand et al. 2004). Therefore, our developed methodology could be of specific benefit for this discipline. Reducing measurement error could be of value in a number of analyses, such as property price valuations, property price developments, location valuation, and valuation of specific location characteristics.

As such, this case study could serve a wide range of institutions. Individuals are interested in property prices to set up list prices, tax authorities rely on property values to estimate the basis for levying property taxes, mortgage providers and banks conduct collateral valuations to qualify the borrower for their mortgage applications and investors and portfolio managers base their investment decisions on periodic evaluations of their real estate portfolio (Liu 2012). Furthermore, incorporation of spatial phenomena through statistical areas enables researchers and analysts to optimize price development predictions in real estate markets. This serves investors, such as pension funds and real estate companies. Also, it allows policy makers to understand in which areas housing/rental prices will rise and intervene adequately. Moreover, given that real estate markets are crucial components in national economies (Pai and Wang 2020), it helps to understand overall economic development. Lastly, utilizing statistical areas in real estate models allows for valuations of non-market products. This enables municipalities to gain insight in the added value of public goods such as parks, education and nature (Daams et al. 2016).

To examine the usefulness of functional statistical areas this paper aims to answer the following research question:

> ***Could functional statistical areas provide an improved alternative for assigning spatial characteristics to point features, over the traditional method by which location characteristics are directly associated per SAU?***

Four sub-questions will be used to answer our main research question. First, since both SAUs and our proposed alternative involve aggregated spatial data, MAUP effects will be inevitable. Therefore, we will examine the MAUP more thoroughly. To do so, the following sub-question is asked:

> *What are MAUP implications and how does it affect statistical areas?*

Second, the extend by which spatial characteristics affect point features should be identified. As discussed, these could vary significantly in shape and size and are dependent upon the situation of consideration. Therefore, focus will be on steps that should be taken to create such zones, rather than specifications of these zones themselves. For this purpose we ask the sub-question:

> *Which criteria should be considered for the creation of functional statistical areas and how do these affect shapes and sizes of zones?*

Third, attention should be paid to how existing mean values of SAUs can be reaggregated to the newly created zones of our previous sub-question. As argued, frequently detailed location information is only limited available. Therefore, we will discus methods to disaggregate mean values to a more detailed level. Hence, the third sub-question will be:

> *How can mean SAU values be disaggregated to low level data, and reaggregated to newly defined statistical areas?*

At last, we should consider how functional statistical areas compare to traditional SAUs in terms of measuring location characteristics adequately. Accordingly, our last sub-question is:

> *How well do FSAs measure location characteristics, and how does this compare to SAUs?*

To answer our research questions a combination of both, literature review and empirical research will be conducted. Since our case study considers CRE as point features and socio-economics as location characteristics, focus will be on these. For the creation of multiple functional areas, we will use GIS software. The performance of FSAs are measured through comparison of model-fit of multiple hedonic models. Hedonic models should break real estate prices down to its utility bearing characteristics and is extensively explained in the methods section. The key dependent variable will be socio-economic variables, as measured by the functional areas. Model performance of different socio-economic areas will be compared to each other and to traditional single statistical areas.

Two datasets will be used to conduct our case study. The first dataset contains an extensive list of socio-economic variables that cover Europe at NUTS 2 and NUTS 3 level. This dataset will be created from multiple databases available on Eurostat. The second dataset consists of commercial real estate in Europe. It includes property information such as sales prices, yearly valuations and property characteristics. This dataset has been made available by real estate research company KR&A. Our data is further explained in the data section.

By answering these research questions, this thesis aims to contribute to our understanding of location characteristics, and its influence on point features. Although the case study examined in this thesis predominantly covers the scientific disciplines of economic geography and real estate, the examined methodology could be applicable in a number of fields. Accordingly this thesis could contribute to the development of a generic methodology that could be used for incorporation of spatial data in statistical models. Nevertheless, it is stressed that this thesis does not aim to provide an extensive guide on how functional statistical areas should be created and applied. We merely discuss options that should be considered when creating functional statistical areas, some of the major techniques available, and provide an example how such areas could be beneficial. Choices that need to be made are scenario specific, and should be carefully considered per situation. In here, a number of factors could be decisive for which technique should used. These include, but are not limited to: point features of consideration; location characteristics of interest; the desired level of sophistication; data availability; the spatial extend of the phenomena at hand; and computing power.

Researchers are urged to consider the spatial implications of using statistical areas, and encouraged to experiment creating their own functional statistical areas. The techniques discussed in this thesis could provide a starting point.

The remainder of this paper is organized as follows. Section 2 describes our conceptual model and section 3 our empirical approach. Section 4 describes the data and the exploratory analysis. Section 5 presents the results, and section 6 concludes.

Literature Review

MAUP implications

The MAUP is a spatial phenomena that occurs when data is aggregated to zones. It entails that when data is tabulated for different spatial scale levels or according to differing zonal systems for the same regions, inconsistent analysis results will be the outcome (Wong 2009). Despite its wide recognition in many social disciplines, the MAUP remains “a thorny problem that cries out for creative solutions” (Hui and Cho 2018, p.186). MAUP effects remain pervasive and their magnitudes vary per spatial variable. Therefore, Wong (2009) argues that a generic solution across different disciplines remains unlikely. Consequently, creative solutions across analysis type or discipline should be thought of to reduce bias caused by the MAUP (Hui and Cho 2018; Wong 2009). Despite the fact that the MAUP is a well recognized phenomena in economic geography and other disciplines that deal with spatially aggregated data, in practice it is rarely acknowledged (Tuson et al. 2020). Presumably less emphasizes is put on possible bias caused by the MAUP, as it is often not the key focus and socio-economic variables are merely included as control variables. Nevertheless, bias in control variables could still affect research outcomes.

Key effects caused by the MAUP are analytical results derived from geographical data being affected by variations in the spatial scale and the spatial configuration of areal units. The first is known as the ”scale effect” and implies that spatial patterns of clustering or coefficients between variables can differ when measured at different spatial hierarchies. Therefore, statistical relationships identified at a certain scale might not be valid when measured at another scale. The second effect is known as the ”zoning effect”. Even at a specific scale there is an infinite amount at which spatial data can be aggregated. Each shape could yield different statistical results (Lee et al. 2016).

The two MAUP effects caused by aggregation of spatial data are demonstrated in figure 2. In 2a and 2b, observed values are randomly distributed. However, two zonal systems are used to aggregate them. Due to the randomness of the data, two different aggregate-level data sets are likely to be produced. However, their differences are expected to be small since no significant smoothing occurs within zonal units. Accordingly, when spatial data is distributed at random, relatively little MAUP effects occur and outcomes of 2a and 2b are comparable despite differences in zones. On the other hand, 2c and 2d illustrate a scenario were observed values are distributed in clusters. While the imposed zoning systems show no dramatic differences, the zoning system in 2c matches the data clustering pattern perfectly and the zoning system in 2d is slightly off. No smoothing will occur in any of the zonal units of 2c. However, in 2d each zone includes mostly identical values with a few different observations. Therefore, data will be smoothed within each unit to create averages and illuminate the differences. Accordingly, in the presence of spatial clustering, drastic smoothing can occur when zones are not defined adequately. Therefore, outcomes of 2c and 2d can differ

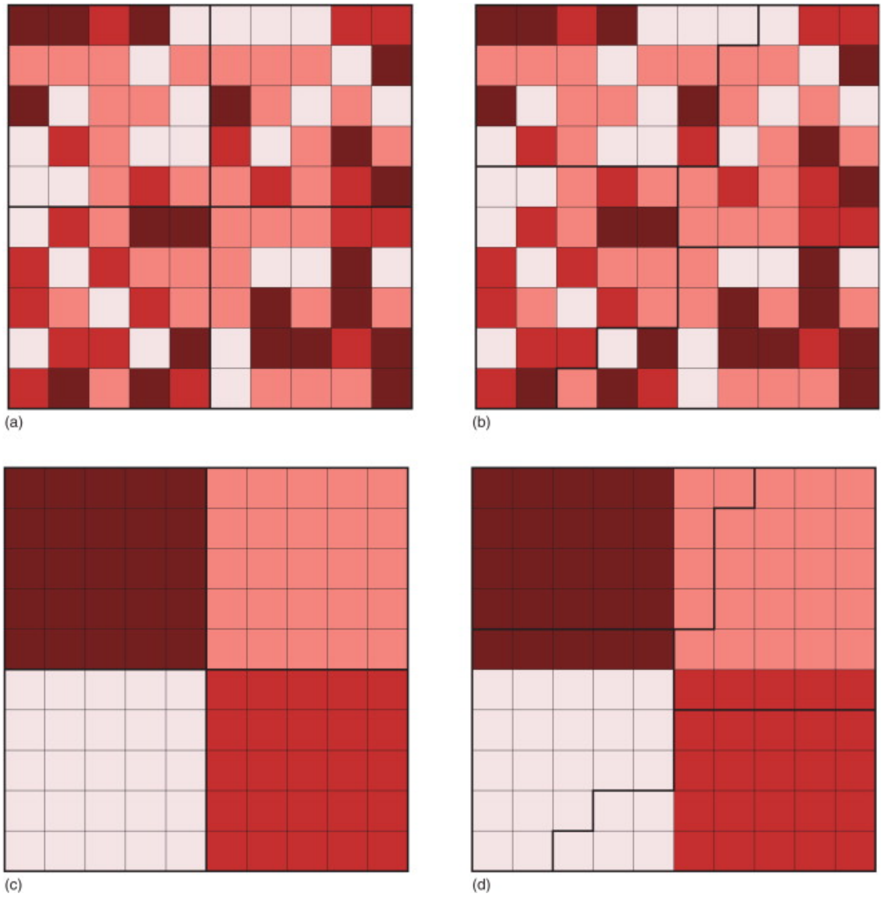


Figure 2: (a, b) Random distribution of observation values with two zoning systems; (c, d) clustering of observation values with two zoning systems (Wong 2009)

significantly because of differences in zones.

For all scenarios the magnitude of smoothing is dependent upon how much intrinsic values differ from each other. If these are significantly different from one another, smoothing effects will be strong. However, if only slight differences exist the magnitude of smoothing stays limited. Accordingly, the effects caused by the MAUP are highly related to the level of internal homogeneity of areal units. Small areal units with identical or similar values will generate less "scale effects" or smoothing when aggregated to form larger units. Zoning effects are minimized when only similar or equal units are grouped to form different zoning patterns (Wong 2009).

Solutions to the MAUP

Thus far no general solution for the MAUP has been found, and according to Wong (2009) this is not likely to happen. Since the MAUP causes inconsistency in analytical results as data is tabulated at differing scales or zonal systems, such an ideal solution should "obtain consistent results regardless of the scale or zonal patterns" (Wong 2009, p.172). Attempts and suggested solutions to deal with the MAUP have been roughly developed within the framework of two extremes in approaches. One argues that data smoothing at different levels will remain to give different outcomes and there is little researchers can do about it. Therefore, it suggests development of analytical techniques that are frame or scale independent. The other recognizes that is it almost impossible to fully overcome the MAUP and suggest that researchers should simply acknowledge its existence by showing variability of results when data of different zonal systems or scales are used. Doing so should indicate the uncertainty of results due to the spatial nature of data. Consequently, solutions to the MAUP can be roughly divided into methods that focus on statistical techniques and methods that deal with spatial data.

Particularly methods among the first category have been widely explored in the real estate literature. Models that have been used to deal with MAUP effects and the related spatial autocorrelation include: spatial lags, spatial (auto) regression, spatial error, locally weighted regressions, (mixed) geographically weighted regression, geostatistics, semi parametric analysis and the mixed spatial Durbin model (Lee et al. 2016; Yu et al. 2007; Osland 2010; Helbich et al. 2014; Kuntz and Helbich 2014; Chica-Olmo et al. 2019). Overall, these statistical techniques have proven to be satisfactory and could indeed improve model performance. Nevertheless, there are also contrasting studies that indicate no model improvement by accounting for spatial autocorrelation, such as the work of Bourassa et al. (2007).

Methods that deal with spatial data have gained less emphasizes. Nevertheless, it has been clear that SAU zones often do not reflect spatial data properly, especially for the socio-economic domain (Lee et al. 2016). According to Jones and Watkins (2009) these are sometimes too broad, so that heterogeneous features of property markets coexist, while other times they are so detailed that they fail to capture the characteristic of the underlying spatial phenomena. The solution has been sought in the identification of better zoning systems that reflect the operation of property markets. Wong (2009) argues that when a solution that deals with the data is considered, it should be remembered that the MAUP exists because data can be aggregated spatially in many ways. Therefore, commonly used approaches are characterized by the development of optimal zoning systems (OZS) according to the modelling objectives or criteria. The use of OZSs have been explored by Openshaw (1977). In his paper he compared the performance of multiple zoning systems. By examining the goodness-of-fit of these different zoning systems in spatial interaction models, it was demonstrated that zoning-system effects can have severe influence on model performance. Furthermore, it provided an optimal zoning procedure to identify approximately optimal model goodness-of-fit zoning systems for multiple spatial interaction models. Nowadays, techniques associated with the OZS have been quite mature and proven to be operational. Especially, the creation of such zoning systems for housing markets have been popular (e.g. Brown and Hincks 2008; Jones 2002; Jones et al. 2012; Jones and Watkins 2009; Royuela and Vargas 2009).

Optimal Zoning Systems for socio-economic variables

The use of OZS could also provide a solution for the proper inclusion of socio-economic variables. According to Wong (2009) zoning criteria are dependent upon the phenomena being researched. Therefore, varying spatial phenomena require the development of specific zoning systems that meet the zoning criteria. Openshaw and Alvanides (2001) argue that the exercise of creating OZS is particularly difficult when socio-economic data is considered, as the distorted patterns are nearly impossible to be revealed by mere observation of the study area. Furthermore, there is no real knowledge of what the "true" result is. Therefore, choice of zone design and proper criteria are deemed critical. Two MAUP effects that are particularly relevant for socio-economic are pointed out. First, individual spatial point data relating to non-modifiable entities are aggregated to a zoning system. Secondly, earlier spatially aggregated data are reaggregated one or more times. Doing so, data alterations can occur in the form of measurement scales, information generalisation, information loss and adding of noise. More fundamentally, it changes the entities that are available for subsequent study. In the context of human geography, data about people changes to data about spatial objects such as places and zones. It is warned that the implicit assumption could be made that such zones are now comparable. However, they might not be comparable at all, as it could still involve places that significantly differ in other aspects than its socio-economic composition. As a solution, Openshaw and Alvanides (2001) suggests the design of zonal objects as meaningful entities related to a particular purpose. Local labour markets are provided as an example of an OZS. Although reaggregation is still arbitrary, labour markets are declared to have an explicit validity of a substantive kind, relevant to one type of study at a particular spatial scale. It is argued that the problem is to decide what sort of areal objects are most useful for studying the phenomena at hand. It is therefore important that socio-economic variables are included in such a way that only the areas that indeed affect properties are included.

Meaningful entities for real estate

Although the spatial scale and possible MAUP effects have been frequently neglected when socio-economic variables are implemented in real estate models, this has not been the case for other variables. Particularly the use of isochrones has been popular as a mean to include spatial variables at multiple scales (e.g. Daams et al. 2016; Bollinger et al. 1998; Weinberger 2001; Cervero and Duncan 2002; Billings 2011; Seo et al. 2019; Can 1998; S. Yoo et al. 2012; van Duijn et al. 2016). These isochrones consist of areas around point features, that cover a certain threshold. Although thresholds are frequently based upon euclidean distance, a number of more advanced methods have been developed to properly associate spatial phenomena to certain features such as properties. Especially for variables with clear spatial boundaries at fixed locations, these advanced methods have been frequently used. In a systematic review on accessibility specification in hedonic price models, Heyman et al. (2019) analyse how locations are measured in housing valuations.

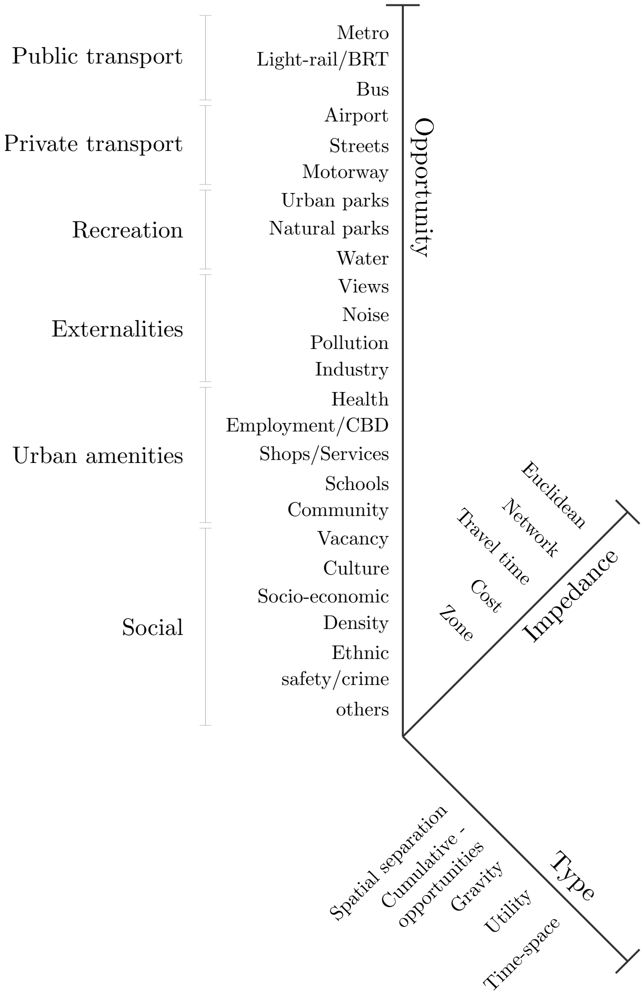


Figure 3: Accessibility measures aspects and their units (adapted from Heyman et al. 2019)

By examining 54 articles on accessibility measures in hedonic price models three main aspects of measure specifications are identified: *opportunities, type* and *impedance* (figure 3).

Opportunities consisted of the spatial phenomena being measured. 26 Opportunity categories were identified, that vary significantly in their spatial scale. Fixed spatial parameters include categories with permanent location features that can clearly be identified. Examples are public- and private transport nodes and networks, recreational places such as natural parks and water, and urban amenities like schools and shops. However, also opportunities with less clear spatial domains were identified. These parameters have less outspoken borders and are often not fixed. They include externalities such as noise and pollution, amenities like culture and community, and social factors as socio-economics, density and crime rates.

The accessibility measure type involves the main method to associate phenomena to specific properties. These include: *spatial separation, cumulative opportunities, gravity, utility and* *time-space.* Spatial separation was found to be the most common and basic accessibility measure. It only uses impedance in measuring accessibility (typically distance, cost, or travel time). A common use of spatial separation in hedonic models is the "distance to Central Business District" measure. Cumulative opportunities are also known as contour or isochrone measures, and calculate the sum of opportunities that can be reached within a particular time, distance, or cost. An example would be the amount of shops within 1000 meter walking distance. Gravitational potential, measures accessibility inversely proportional to the impedance between location and attraction, and positively proportional to the attraction size. It can be thought of as a more comprehensive accessibility measure which encompasses both cumulative opportunities and spatial separation. The utility measure interprets accessibility by monetizing the outcome of travel and destination choices. It measures the sum of all utility choices and is also known as the logsum accessibility measure. This measure is most commonly utilized by using denominators of the multinomial logit model. The time-space measure is also known as the individual-based accessibility. It measures activities in which an individual can participate within a given time. This can be achieved by putting constraints on location, duration and costs of mandatory and flexible activities. Time-space measures are the most disaggregated form of accessibility measures. However, because data at the individual level is required these measures are not commonly used in hedonic price studies.

The impedance is described as the resistance to overcome access and can be expressed as distance, cognitive resistance or time. Five types of impedance are described. These involve: *Euclidean distances, network distances, travel times, costs* and *zones.* Euclidean distance is the shortest distance in space. Network distances refer to metric distances through networks such as road center line networks. Travel time considers both distance and time, and can be measured in a number of ways. An example would be multiplying the distance by the average speed and penalizing waiting times. The costs impedance considers the costs of accessibility options, such as adding fuel prices to capture the mode choice for different socio-economic groups or at different geographical locations. Also the costs impedance can be measured in various ways. Zones are predefined areas that can take various shapes according to criteria. In practice it often involves SAUs.

Clear couplings between different opportunities, types and impedances were found. Also for socio-economic opportunities clear preferences for type and impedance were contrived. Out of the 21 cases in which socio-economic opportunities were considered, 20 accessibility measures used the cumulative opportunities type. One case used the spatial separation type. Furthermore, 19 out of the 21 cases used zones as impedance type. The other two cases used euclidean distance as impedance.

Heyman et al. (2019) warn that a strong tendency exists to less advanced measurements. These basic measures have a weaker connection to consumer perception, while this connection is fundamental to the hedonic approach. It is argued that among the impedances networks, travel time and cost can be considered to capture perceptions of consumers in a satisfactory way. However, Euclidean distance and zones in general do not.

Reaggregation of spatial data

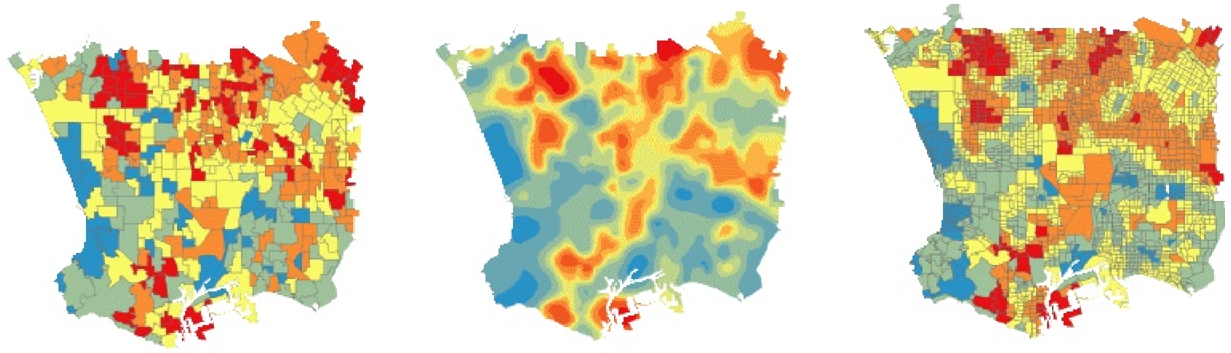
As indicated earlier, the MAUP effects are caused by aggregation of data, especially when this happens in a random way that does not recognize clusters and patterns (Wong 2009). Therefore, ideally researchers should work with low scale data and aggregate this data as meaningful entities related to the research purpose (Openshaw and Alvanides 2001). However, data is often only available at limited scales. Especially socio-economic data is often aggregated for privacy reasons and functionality. Furthermore, it is cost intensive to collect high detailed spatial data over a large time span. Therefore, such high detail socio-economic data frequently only covers a small area or a limited time span. Consequently, when considering a wider area over a larger time span, researchers often have to rely on high-scale aggregated data.

Because limitations in data availability, researchers have to rely on other approaches to obtain spatial information at the desired format. For this purpose, a number of geostatistical techniques have been developed (Arbia 1989). The transformation of spatial data has been fundamental in geographic information systems (GIS). Therefore, such geostatistical techniques are widely used in spatial analysis. Problems that involve the mismatch between observed data and target data are generally captured under the umbrella term “change of support problems” (COSP). According to E.-H. Yoo (2017), a solution to the COSP for areal objects is the use of areal interpolation.

For disaggregation of socio-economic data, which is often only available at larger areas, areal interpolation could thus provide a solution. Areal interpolation is also known as polygon-to-polygon reaggregation and is commonly used for downscaling or upscaling administrative units of data. This frequently happens in geographical or regional research in cases were available spatial data does not take the form of interest. Doing so it could provide a solution in cases were spatial data collection boundaries may change over time, or when different variables from multiple incompatible zones need to be compared (Flowerdew et al. 1991). Furthermore, it could reduce MAUP effects. Areal interpolation could result in aggregated data that behaves more predictable and understandable. Furthermore, aggregating within the range of spatial autocorrelation could reduce errors induced by averaging dissimilar units (Gotway and Young 2002).

According to Comber and Zeng (2019) a clear distinction can be made between point- and areal interpolation. The firmer is used to make predictions at locations of which values are unknown with sample points that contain empirical information. It is typically assumed that such data varies continuously over space. Point interpolation approaches include both exact methods such as IDW and Kriging, and approximate methods like trend surface analyses. The latter transfers attribute values from source zones with known values to target zones with unknown values. These target zones are usually smaller, but do not necessarily have to be. In other words, point interpolation uses measuring points to create an estimated value surface that covers the study area. On the other hand, for areal interpolation values of the complete study area are known but their resolution is inadequate.

Dependent upon the method, reaggregation of areal data is a one- or two-step process. The latter is visualised in figure 4. The initial SAUs are school zones (4a). In the first step (4b) a smooth prediction surface is made by interpolating the source areas. The second step(4c) aggregates the created surface area to a target zoning system (ArcGIS n.d.). In case of a one-step method, the first step is skipped.



(a) Initial data at SAU format (b) Continuous surface (c) Reaggregated data

Figure 4: Polgon-to-polygon data reaggregation work flow (ArcGIS n.d.)

Interpolation methods

Various areal interpolation methods have been developed. According to Comber and Zeng (2019) these can be grouped into two broad categories. Methods that use ancillary/auxiliary data to control the reallocation process from source to target zones, and methods that do not. These rely solely on the target and source zones properties. It is argued that the firmer provides more accurate prediction results. However, there is a trade-off as these methods require additional high quality data and are more difficult to model. A summary of the major areal interpolation approaches is given in table 1.

Main methods without ancillary data include: *simple areal weighting, pycnophylactic interpolation* and *area-to-point interpolation.* Areal weighting is a simple method that is frequently used when source zones and their attributes are the only available information. It allocates source zones attributes proportionately to target zones, based on the area of their intersection. Advantages are that it is easy to implement and supported in most GIS software. Furthermore, areal weighting is inherently volume preserving. Therefore, source zone values remain equal if the values of the target zones within the source zones are summed. The disadvantage is that it assumes spatial homogeneity between the source zones attribute and the target zone areas. This assumption is rarely hold true in practice (Comber and Zeng 2019).

Pycnophylactic interpolation is a slightly different approach to areal weighting. It repeatedly interpolates source zones attributes to target zones. Doing so it avoids sharp discontinuities between

Method	*Auxiliary Variables*	*Advantages*	*Disadvantages*
Simple areal weighting	×	Simplicity	Assumes spatial homogeneity
Pycnophylactic interpolation	×	Creates smooth surface	No sharp boundaries in data distribution
Area-to-point interpolation	×	Computationally inexpensive	Border issues
Dasymetric interpolation	✓	Mature, Stable performance	High data demand, Computationally expensive
Street-weighting	✓	Accurate with variables that follow road networks	Unsuitable for rural areas
(Geo)statistical	✓	Many options, Wide range of possible data	Accuracy dependent on model performance, May require advanced statistics
Point-based intelligent	✓	High accuracy, No MAUP risk in auxiliary data, Auxiliary data widely available, Computationally efficient	Scale and aggregation issues, points might not indicate measured value properly

Table 1: Areal interpolation methods

neighbouring target zones, whilst preserving overall mass or volume in counts of the course zones. Each iteration improves the smoothness further. It uses the weighted average of the target zone's nearest neighbours. The total level of smoothing is determined by the amount of iterations and the number of nearest neighbours used. Pycnophylactic interpolation offers an elegant solution to generate a smooth surface from discontinuous data. However, it assumes no sharp boundaries in the distribution of data. This might not always be the case. For example when target zones are divided by natural boundaries such as rivers (Comber and Zeng 2019).

As indicated by its name, area-to-point interpolation (also known as point-in-polygon) transforms areas to points and uses one of the many point-to-point interpolation techniques available (IDW, Natural Neighbour, Splines, Kriging, etc.). Mostly, points are based on the centroids of areas. However, other methods such as taking the vertices of cells are optional as well. Point interpolation has the advantage that it is computationally inexpensive. However, it could be inefficient if the source zones are not sufficiently small. Furthermore, border issues could arise if the estimated points do not fall within the actual zone (Do et al. 2021).

Methods that do include ancillary data are: *dasymetric, street-weighting, statistical or geostatistical approaches* and *point-based intelligent approach.* Dasymetric methods have been widely used and can be thought of as the more advanced equivalent of areal weighting. It uses areal features, including buffered lines and points, as ancillary to aid the areal interpolation process. The auxiliary data serves here as a spatial control that can identify areas to be included or excluded from the interpolation process. The advantage of doing so is that it gives more accurate results than simple areal weighting as it distributes values into different land use classes. Dasymetric interpolation has been regarded as mature and stable. However, the need for auxiliary data makes it harder to perform. Furthermore, the accuracy is dependent on the resolution of auxiliary information. High resolution auxiliary information might not only be hard to obtain, but also comes with additional computing costs (Comber and Zeng 2019).

Street-weighting methods utilize road networks to interpolate areal values. Multiple methodologies exist to do so. In the simplest form it uses networks' lengths and distributes values uniformly among the street segments within source zones. After that, it intersects the linear features with the target zones and estimates the values in those zones by summing the values along the road segments. The significant difference with areal and street-weighting is how weighting is applied. For the firmer, intersecting areas drive allocation. The latter, uses length of intersecting linear objects. Doing so, street-weighting methods perform well in urban areas. However, it is not as accurate in rural areas with less dense street networks (Comber and Zeng 2019).

Statistical and geostatistical methods involve a wide range of methods. Statistical approaches use auxiliary data in combination with mathematical techniques to derive a functional relation between spatial distributions of ancillary data and the distribution of source zones of which data is to be interpolated. Usually this is done through regression analysis. Geostatistical interpolation was originally designed for the interpolation of points. However, it was extended to areal interpolation because of its ability to accommodate spatial autocorrelation in the modelling process. Geostatistical interpolation usually takes the form of co-kriging or regression kriging (Comber and Zeng 2019). The advantage of geo(statistical) techniques is that a wide range of auxiliary data could be used. However, it requires additional modelling with advanced techniques.

Point-based intelligent approaches use point locations as auxiliary data to guide the interpolation process. This allows for more accurate interpolations. Advantages are that point data have much simpler structures than lines or polygons, and they do not require topological information to represent spatial relationships. This makes the computing process more efficient and less demanding. Furthermore, many types of point data are widely available in a number of databases. This simplifies the interpolation of varying types of variables to target zones. Moreover, since point data represent features at discrete dimensionless locations, no MAUP effects in the auxiliary data are present (Zhang and Qiu 2011; Comber and Zeng 2019).

For most approaches, a variety of choices should be made. These choices determine the sub-method which is used, and even within sub-methods many options are available. For example, with PIP one first needs to decide the point placement method. Then, the appropriate point-to-point interpolation option should be determined (e.g. kriging). And after that, there is still a choice between the many kriging approaches.

Interpolation data types

To understand how the data should be interpolated, the distinction between extensive and intensive variables is made. For an extensive variable, the value of a region is the product of all underlying values covered by that region. Every count variable, such as population counts or the number of households, is an extensive variable. An intensive variable calculates the value of a region by considering values of sub-regions, and multiplying these by a set of unequal weights. Proportions and rates, such as employment rates, are intensive variables (Do et al. 2021). According to Do et al. (2021) it is strongly recommended that when intensive variables are defined by two known extensive variables, these two extensive variables are estimated first and the intensive variables are computed from these afterwards. This should provide more accurate results than directly interpolating intensive variables.

Three dominant data structures can be recognized according to Krivoruchko et al. (2011). These are Gaussian, binomial and Poisson. The firmer stems from continuous data and the latter two from count data.

Averaged data interpolation is the simplest of the three and has traditionally been used the most. Most commonly averaged areal data interpolation is used when continuous data, measured at specific points, has been averaged into specific zones for privacy or to reduce overhead. An example would be population mean ages. Averaged data interpolation requires a Gaussian distribution and is therefore often known as Gaussian interpolation. It produces a surface with a predicted value for each point within the data domain. Gaussian interpolation requires averaged data and their specific zones.

Rates involve count data that measure one phenomena relative to another. A typical example of interpolation that considers rates is the use of samples from a known population within specific zones. More specifically, it is likely to count more highly educated people as the population size increases. However, rates could still be the same. Therefore, it should be compensated for population density. Individuals should be sampled randomly from the population within each polygon. The number of people with a particular characteristic relative to the total sample size is the number at risk. This kind of sampling is known as binomial sampling, since data is distributed binomially. Interpolation of binomial data produces a risk prediction surface for all points within the data domain. Risks at any individual point represent the probability that an individual sampled at that location has the characteristic. Interpolation of rates requires a population field, a count field and their specific zones.

Event counts also involve a count field. However, whereas rates consider population, events consider time. As sampling times increase, chances of counting more events within a zone increase as well. An example would be crime numbers in certain blocks. The longer you observe, the likelier you are to observe a crime. Consequently, event counts should compensate for the observation period. Areal interpolation for event counts produces surfaces that predict the underlying risk of observing an event at a specific location. It is not required that every single event is equal. However, the number of observed events per unit of time should be proportional to the underlying density of the phenomena being observed. In other words, the observation methodology needs to be the same for each zone. Event counts have a Poission distribution and event counts interpolation is therefore also known as Poisson areal interpolation. It requires a time field, a count field and their specific zones.

Theoretical framework

Based on the literature, two main problems of socio-economic data in relation to real estate are identified. Bias is caused by scaling- and zoning effects as consequence of the MAUP. Therefore, ideally low scale geographic data that properly reflects the radius at which properties are influenced by socio-economic data should be used. The conceptual model (figure 5) provides an overview of our proposed spatial data modifications in order reduce bias caused by the MAUP.

Two input datasets are used. The first is a dataset that contains socio-economic variables. The socio-economic variables that are considered could be of any kind. As indicated earlier there are over 400 parameters that have been associated with affecting property prices and many of these variables could be considered as socio-economic. However, the socio-economic data should contain both a spatial and temporal dimension, so that values can be associated to the right place and time. Furthermore, socio-economic variables affect which interpolation methods and accessibility measures should be used. Therefore, their spatial extend should be well considered. The second dataset contains property values and characteristics. Ideally, these property characteristics should be as encompassing as possible. However, they should at least contain the property locations and values.

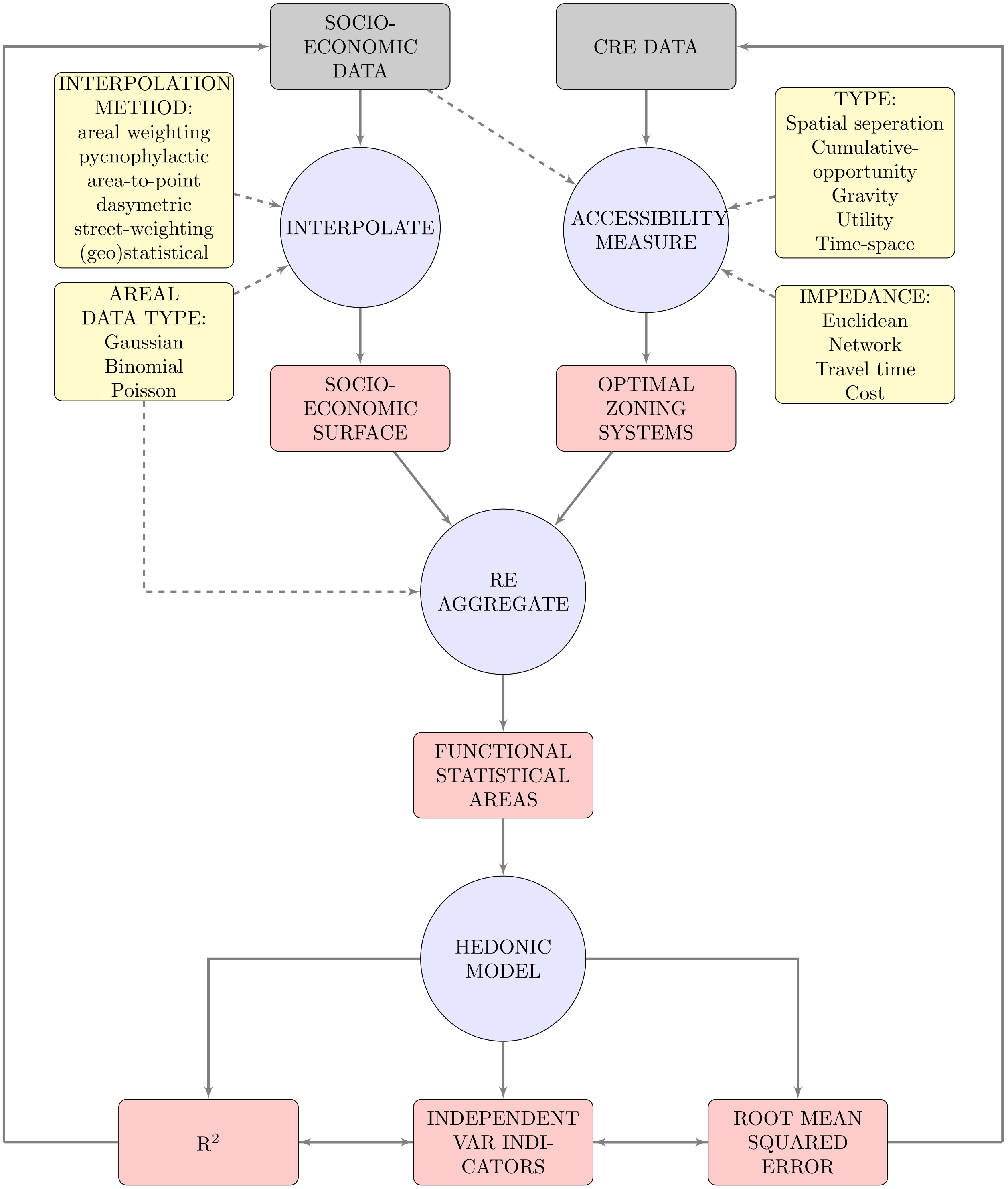


Figure 5: Conceptual model

First the socio-economic data is interpolated to create a smooth surface. The use of a smooth surface should lead to less dramatic scale effects as the drastic differences between boundaries are smoothed and zones transform to single pixels. This should lead to more predictable data behaviour and a reduction in bias as consequence of the MAUP (Wong 2009; Flowerdew et al. 1991; Arbia 1989; Gotway and Young 2002). Interpolation can be executed through a variety of methods such as areal weighting, pycnophylactic interpolation and area-to-point interpolation.

Determining which interpolation methods to use is dependent upon the situation of interest. All methods have their advantages and disadvantages. Certain interpolation methods could be favoured over others based upon variables of consideration, sizes and shapes of initial SAUs, model complexity and desired performance, and availability of auxiliary data.

Second, accessibility measures should be made around each property for each socio-economic variable. For the creation of such measures the three main aspects as described by Heyman et al. (2019) should be considered. The socio-economic data serves here as the opportunity. For types, the recommended methods of network; travel time; and cost are considered. However, also euclidean distance is included due to its simplicity. Regarding the impedance, in most cases cumulative-opportunities are the most obvious choice. However, there might be cases in which another impedance measure might be more suitable. Therefore, all impedance categories have been included. The accessibility measures should properly reflect how socio-economic phenomena at different locations affect property prices. Therefore, multiple zones per socio-economic variable and per property are very well possible. Accordingly, multiple zones could reflect different spatial scales at which properties are affected by socio-economic phenomena. Doing so allows for the qualification of relative location performance compared to the wider area.

In the next step the socio-economic surface is aggregated to the optimal zoning systems. This should provide us with a dataset containing CRE characteristics and socio-economic data that actually affects the properties at hand. Again, the socio-economic data type should be well considered here.

In our final step the modified dataset could be used for the input of a hedonic model. According to our hypothesis, the creation of FSAs should lead to higher model performance than the traditional inclusion of socio-economic variables based on SAUs. Furthermore, different FSA methods might be compared to each other. Doing so allows us to evaluate individual components of FSA measures such as the used OZS and interpolation methods. Additionally, individual subsets and multiple opportunities can be compared to serve as a robustness check.

Methodology

Disaggregating socio-economic regions

For the disaggregating of socio-economic data, the Pycnophylactic interpolation method is used. As argued, most socio-economic data is only available in SAUs format. Therefore, applying Pycnophylactic interpolation could provide us with more detailed estimations of socio-economic values. The advantage of this interpolation type is that it allows for the creation of an absolute surface. This offers functionality benefits in the process of re-aggregating the data, as OZS might overlap due to proximity of buildings. Furthermore, it preserves the volume of the values measured. Moreover, there is no additional requirement for auxiliary data. Although some methods that do require such data possibly have higher performance, they are considered to be out of the scope of this research. Of the three discussed non-auxiliary methods, Pycnophylactic interpolation has the overall highest performance according to Hawley and Moellering (2005).

Pycnophylactic interpolation has been pioneered by Tobler (1979) and has been based upon the first law of geography. According to this law “everything is related to everything else, but near things are more related than distant things” (Tobler 1970, p. 236). Pycnophylactic interpolation exploits this geographical structure over space by inferring spatial distributions of attributes. It assumes the existence of a smooth density function which is non-negative and has finite values for all locations. The Pycnophylactic property is defined in the following equation:

$$\sum_{ij} az_{ij}q_{ij}^{k} = p_k,\ \sum_{ij} aq_{ij}^{k} = A_k \text{ and } \sum_{k} q_{ij}^{k} = 1 \tag{1}$$

in which p is the original value and A is the area, of zone k; z is the density in cell ij and a is the area of each cell. q_{ij}^{k} is set equal to 1 if ij within zone k. Otherwise it is set to 0. The interpolation starts with the assignment of a mean density to each grid cell that is superimposed on the source zones. Next, the assigned values are slightly modified to bring the density closer to the value required by the governing partial differential equation, which is defined as follows:

$$\iint_{R_i} Z(x,y)dxdy = H_i \tag{2}$$

Here R_i denotes the i^{th} region and H is the total value of that region. At the end of each iteration, all density values within individual zones are incremented or decremented to enforce the volume-preserving condition (Kim and Yao 2010).

For the Pycnophylactic interpolation we make use of the *Pycno* package in R. This package provides the identically named pycno function. This function is based on the above mentioned equations and provides a spatial grid with the interpolated values (Brunsdon 2014).

As argued by Tobler (1979), the equation does not take the ellipsoidal shape of earth into account.

It therefore assumes equal area map projections. Therefore, our initial data is re-projected to the LAEA Europe system. This projection is designed for statistical purposes on the European scale, and aims to project area sizes unaltered (epsg.io 2019).

Creating Optimal Zoning Systems

For the creation of our OZS we consider the accessibility measures as defined by Heyman et al. (2019). The opportunities are the socio-economic variables of our consideration. This leaves the impedance measure and type to be decided. For the type we will use the cumulative opportunities, as is common for the measurement of socio-economic variables. The underlying reasoning for choosing this measurement is that property prices are indeed not affected by single persons or locational features, but by the whole of everything that is within its radius. To incorporate this type in our OZS we adapt the function as defined by Heyman et al. (2019) and slightly modify it so that not only a maximum impedance is considered, but also a minimum. This gives us the following function:

$$X_{ido} = \sum x_o \text{ if } r_{min} \leq d_{id} > r_{max} \tag{3}$$

in which X is the sum of opportunities for property i at distance class d and opportunity o; x is the opportunity surface; r is the threshold radius for which the opportunities are summed.

The threshold radius is determined in the impedance measure. Since, the academic literature has reached no consensus upon what scale spatial phenomena affect property prices, multiple impedance measures will be considered at varying ranges. These ranges should reflect the micro, meso and macro level at which location features affect property prices. The varying scales should not only allow for modelling the relative socio-economic performance of regions but also deals as a certainty measure that allows us to see at which ranges socio-economic values are significant. The use of multiple scales to measure location features at specific ranges has been frequently applied in the academic literature (e.g. Daams et al. 2016; Seo et al. 2019; van Duijn et al. 2016; Billings 2011).

An overview of the impedance methods and their measures at the three different scales is given in table 2. The considered impedance methods are euclidean distance, network, travel and time. Although it is argued by Heyman et al. (2019) that euclidean distance does not reflect the public perception properly, we argue that there is a trade-off between method performance and simplicity. Since euclidean distance is by far the easiest impedance measure it is still considered. On the other hand the costs method is not considered as it requires extensive additional modelling and goes beyond the scope of this thesis.

The scales of consideration involve a micro, meso and macro level. Since it is uncertain at what range location features still affect property prices, these cover a wide array. Doing so they should reflect user perception at multiple ranges. The micro scale could be considered as a measure for daily commuting; the macro scale as measure for less frequent commuting; and the meso scale falls somewhere in between. However, determining users of properties is not always obvious. In the case of offices, the company hiring office space might be the primarily user. However, employees are the ones who actually have to overcome the distance between their location of residence and where the office is located. Furthermore, the perception of distance could differ per property type and across geographies. For a logistic user, 20 kilometres might be considered as nearby while a retail user might think of it as far. The same goes for geographies. People living in a small and densely populated country such as Luxembourg might perceive distance different than people who are used to travel large distances, as is the case in more remote areas of countries like Australia. Because of this subjectivity in the perception of distance (both space and time) it is acknowledged that the assignment of scales is somewhat ambiguous. Nevertheless, it is aimed to set scales in such a way that they capture both local details, and larger phenomena.

	Euclidean Distance	*Network*	*Travel time*
Micro	10km	10km	10 min
Meso	50km	50km	30 min
Macro	100km	100km	60 min

Table 2: Impedance measures

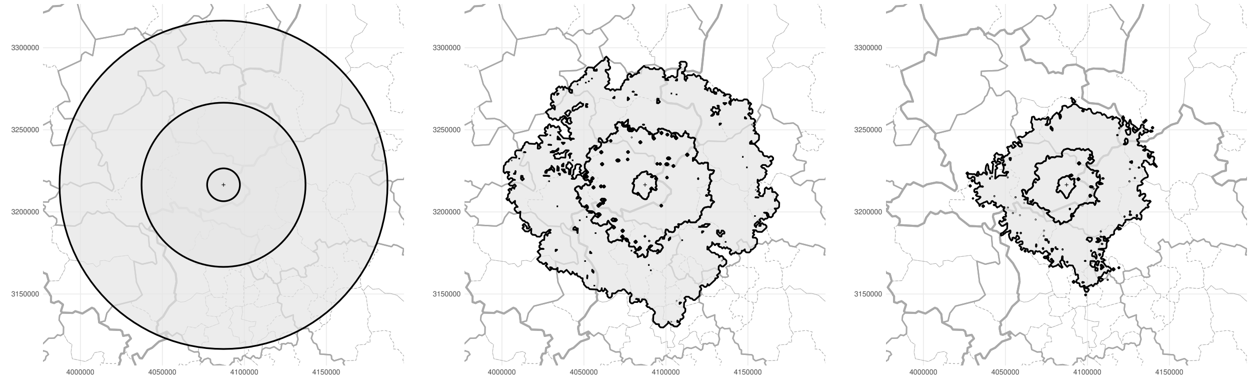
The creation of OZS is done with a number of packages that allow GIS modelling in R. The creation of euclidean distance is done with the buffer function *st_buffer*, from the *sf* package (Pebesma 2018). This function creates areas around each property, using fixed or dynamic distances. For the creation of these buffers, data will be projected in the EPSG:3035 - ETRS89-extended / LAEA Europe Coordinate Reference System (CRS). This CRS has been designed to accurately represent Europe and has an accuracy of 1.0 meter (MapTiler n.d.).

For the remaining two impedance measures we make use of the openrouteservice-r package (Oleś 2021). This package provides routing services, based on the openrouteservice API (ORS). ORS provides a number of professional routing options using geographic data from OpenStreetMap. It allows for the creation of isochrones that return service/reachability areas for multiple locations, using a variety of travel modes such as cars, heavy good vehicles, walking or cycling. Furthermore, it enables the distinction between travel time and distance. Different user types might lead to varying isochrone outcomes, as underlying assumptions such as accessibility and travel speed change. Especially travel-time isochrones are heavily affected by the user type, as they are are both dependent on travelling speed and accessibility. It is therefore important to set proper profiles for the calculation of our OZS and understand the implications of the assumptions that are made.

Since the aim of this thesis is to develop a methodology for the creation of FSAs, rather than to actually obtain one, we will limit ourself to the driving-cars profile. This profile utilizes the most complete data types from OpenStreetMap and should deliver the most accurate results. Furthermore, driving-cars are commonly used as a mean of transportation in Europe, especially for longer distances. Therefore they should reflect how accessibility is observed properly. The driving-cars profile considers all road types that are accessible for passenger cars. It estimates a travel speed for any road, based on a cascading assessment. This assessment follows a number of steps, in which maximum travel speeds for specific locations are first considered. If travel speeds at the local level are unavailable, more global indicators such as road types are considered. More details on the cascading assessment can be found in Openrouteservice (n.d.).

Limitation of ORS, is that there is no optionality for time control. Therefore, all isochrones are created with the road network available at the time of analysis. However, in practice road networks change. Therefore, the assigned OZS might not properly reflect actual accessibility at the time a property was valued. An example would be the maximum speed restriction in the Netherlands that was set in 2020. Our CRE dataset contains data from before 2020. Consequently, using the current road network to calculate historic accessibility might cause a mismatch. Also at a smaller time scale there might be a mismatch between actual accessibility and calculated isochrones. Using speed limitations might provide realistic estimations for average accessibility. However, it does not account for congestion that can occur at specific times. Furthermore, estimations might be biased for roads with varying speed limits.

Moreover, it is recognized that differentiating in travelling modes could be used to further improve OZS. An example could be the use of heavy good vehicles for retail properties. When logistic properties are considered, this user profile might reflect actual accessibility more adequately.



(a) Euclidean Distance (b) Travel Distance (c) Travel Time

Figure 6: Optimal Zoning System Methods

A graphical example of the three OZS methods is provided in figure 6. Here, a constant location is taken for the three unique zoning systems. It can be seen that the three scales differ per OZS method. The buffer has the largest range, which makes sense as it is not limited to any physical boundaries. In this case travel distance impedance has the second largest reach and the travel time impedance the smallest. However, whereas the scale for euclidean is independent to location, this is not the case for the other two measures. Therefore, it might be very well possible that at a different location, the ratio of scales is very different for the distance and time impedance. In this example there is limited highway access, but there are many small roads. Therefore, travel distance measures cover a relatively large area compared to the travel time measures. However, this might not be the case for cities with excellent highway access, or areas with little speed limitations. Furthermore, buffer zones consist of continuous circles with no dents and gaps. However, the road distance and travel time measures are oddly shaped as they indeed follow road networks. Therefore, they have dents on the outsides but also contain gaps within. Doing so they indicate that regions might not be accessible to a property at all, even though they are closer to it than other regions.

Since travel time and distance allow for this nuance it is expected that these two will indeed perform better, with travel time as most nuanced and therefore the best. However, these methods also come with additional assumptions and are therefore more prone to errors.

Validating functional statistical areas

Defining a hedonic model

To validate our proposed functional statistical areas, multiple hedonic models will be applied and compared. The hedonic model is pioneered by Rosen (1974) and is frequently used within the field of real estate (e.g. Daams et al. 2016; Sirmans et al. 2005; Herath and Maier 2010; Ekeland et al. 2002). Within the hedonic framework properties are considered as bundles of attributes that offer utilities to its users. Due to market forces, each location should be utilized by the user type who can profit the most from these location characteristics and thus is able to offer the highest bid (Evans 2008). The market equilibrium causes location features to be capitalized in surrounding property prices. Capitalisation indicates welfare benefits derived by property owners near these locations (Daams et al. 2016). Consequently, the market equilibrium determines property prices by their utility-generating attributes. Hedonic regression techniques can be used to filter out these implicit attribute prices and thus determine their values (Liu 2012). Doing so enables researchers to estimate the direction, magnitude and significance of each individual parameter included in our model. Furthermore, a number of statistics allow for quantification of model performance and comparison of different models. Through inclusion of our socio-economic parameters based on varying FSA methods in multiple hedonic models we should thus be able to compare the performance of these different FSA methods. Since no actual model changes are made and the underlying socio-economic dataset stays the same, the differences in model performance are solely based upon the FSA method. Therefore, the hedonic model should offer a reliable method to compare FSA methods with traditional methods and between each other.

Our base model is defined as follows:

$$y = \alpha + \beta x + \varepsilon \tag{4}$$

where y is the dependent variable and contains the valuation price; α is the constant; β is the estimated coefficient; x is the relevant property characteristic; and ε denotes the standard error.

The use of price estimates as dependent variable is not without debate. Within the literature there has been discussion if self evaluations could bias hedonic models, as over/under valuation could occur (Malpezzi 2002). For our research purpose it is expected that valuations could be used as dependent variable. The valuations have been made by one or more independent professionals. Valuations should thus be accurate. Furthermore, it is argued by Goodman and Ittner (1992) that although larger variances could occur because of valuations, this not necessarily leads to large bias as long as datasets are large enough. Moreover, it is not the aim of this research to make optimal prize predictions, but to find if use of FSAs could lead to more accurate location characteristics measurements. Nevertheless, possible bias from use of valuations should be considered.

For our analysis we will consider a number of socio-economic indicators that are typically recognized as important determiners of location values. Since our analysis requires spatial data that is consistently available over a large spatial area and within a certain time frame, we limit ourself to the Eurostat database. To limit zoning effects our input data should ideally be as low scale as possible. Therefore, we stick to indicators that are available at NUTS 2 or NUTS 3 level. Only extensive variables are considered. As discussed in our literature review, these give more accurate interpolation results according to Do et al. (2021). To retrieve Gaussian and Binomial parameters, it would be possible to use the interpolated results of these extensive variables as inputs for the calculation of new rates and means.

Numerous socio-economic indicators are known to influence property prices. Metzner and Kindt (2018) identified over 400 parameters that affect location values. Of these a great amount can be qualified as socio-economic. Since it will not be possible to test the performance of our supposed methodology on all of these, we stick to sixteen variables that should represent the main categories. The socio-economic variables that will be considered in our hedonic model are presented in table 3. Categories that are considered are demography, education, innovation, economy and labour. These should cover the most important location factors for each of our considered CRE type. For example, for offices it seems logical that they value a highly educated labour force that is active in a digital environment such as finance. On the other hand, retail companies might be more affected by GDP and purchase power.

In order to deal with missing values and outliers, averages for each variable are taken per SAU, for the years available. Doing so, smooths initial data and although some details mights be lost, values become more stable and predictable. Using the mean of a larger time period is justifiable, as real estate properties are long lasting assets. Therefore, prices are more likely to be affected by long term trends, rather than yearly outliers in our socio-economic data.

The inclusion of control variables for this analysis follows the conventional hedonic approach as described by Malpezzi (2002). Observed property characteristics include gross leasable areas, plot sizes and refurbish dates. These should reflect the most important building characteristics and price determiners. All three variables are expected to have a positive sign, as increases in leasable areas

Variable	SAU
Demography	
Population (nr)	NUTS 3
Education	
Low Educated (nr)	NUTS 2
Medium Educated (nr)	NUTS 2
High Educated (nr)	NUTS 2
Innovation	
Scientist/engineers (nr)	NUTS 2
Economy	
Transported Goods (1000 tonnes)	NUTS 3
Total Nights Spend in Hotels (nr)	NUTS 2
GDP (million)	NUTS 3
Purchase power (million)	NUTS 3
Labour	
Employed industrial (nr)	NUTS 3
Employed construction (nr)	NUTS 3
Employed retail (nr)	NUTS 3
Employed transport (nr)	NUTS 3
Employed food (nr)	NUTS 3
Employed communication (nr)	NUTS 3
Employed financial (nr)	NUTS 3
Employed education (nr)	NUTS 3
Employed arts (nr)	NUTS 3

Table 3: Overview of the independent variables to be used

and plot sizes generally have upward affects on property prices. The same goes for properties that are renovated more recently, as these are likely to be in better condition. Moreover, buildings are categorized by specific user categories if more than half of the rents are derived from a single use. This allows us to distinguish between different property types.

Since our base model only allows for the inclusion of a single valuation, our formula will be rewritten such that y contains the x^{th} property valuation ($x = 1, ..., n$) in user type u at time t. Furthermore, we allow for the inclusion of multiple explanatory variables. These include the control variables as described above and a single socio-economic variable. The decision to include only a single independent variable per model is made to avoid multicollinearity; keep the model simple for clear comparison; and limit over-fitting. Also, we transform our model to a semi-log functional form since the valuation data have a right-side tail. This method is described by Palmquist (2005) and allows the relation of our dependent and independent variable to vary proportionally. The semi-log form mitigates to the statistical problem of heteroskedasticity. Therefore, our dependent variable y becomes a natural log ($ln(Y)$). Despite this transformation of our dependent variable, the relation of gross leasable areas and log prices remains non linear. Therefore, a second degree polynomial is included as well. Polynomials are frequently used in linear models and allow for estimation of non-linear relationships. This allows us to measure the relation between $ln(Y)$ and GLA as a parabola instead of a linear trend. Since the added value of additional square meters to be leased has decreasing marginal returns, it is expected that this polynomial (GLA^2) has a negative sign. This gives us the following equation:

$$ln(Y_{xut}) = \alpha_u + \beta_u GLA_{xut} + \eta_u GLA^2_{xut} + \gamma_u PZ_{xut} + \delta_u RD_{xut} + \zeta_u SE_x + \varepsilon_i \tag{5}$$

in which GLA is gross leasable area; PZ is plot size; RD is the (last) renovation date and SE is the socio-economic value. $\alpha, \beta, \eta, \gamma, \delta, \zeta$ are parameters to be estimated.

When rewritten, x_a will be the a^{th} relevant property characteristic ($a = 1, ..., A$), so that we get the following model:

$$ln(Y_{xut}) = \alpha + \sum_{a=1}^{A} \beta_a C_{axut} + \eta_u GLA^2_{xut} + \zeta SE_x + \varepsilon_i \tag{6}$$

Since it aimed to test model performance of multiple impedance measures and scales for a variety of opportunities, a fit should be made for each possible combination. Therefore, we add an opportunity (o) measure that determines the socio-economic characteristic; an impedacance (i) measure that considers the shape of our zoning system; and a scale (s) measure that accounts for the size of the zone. This makes our first model as follows:

$$ln(Y_{xutois}) = (\alpha + \sum_{a=1}^{A} \beta_a C_{axut} + \eta_u GLA^2_{xut} + \zeta SE_x + \varepsilon_i)_{ois} \tag{7}$$

Additionally we consider a second model that allows for interaction of different scales. To do so, we need to consider our zoning scales as separate rings of a single zoning entity. Therefore we include all scales of s into our equation, and no longer consider estimated property prices as a function of a singular scale. This gives us our second and final model:

$$\begin{split} ln(Y_{xutoi}) = (\alpha + \sum_{a=1}^{A} \beta_a C_{axut} + \eta_u GLA^2_{xut} + \zeta_1 SE_{micro_x} + \zeta_2 SE_{meso_x} + \zeta_3 SE_{macro_x} \\ + \zeta_4 SE_{micro_x} * SE_{meso_x} + \zeta_5 SE_{micro_x} * SE_{macro_x} + \zeta_6 SE_{meso_x} * SE_{macro_x} + \varepsilon_i)_{oi} \end{split} \tag{8}$$

Model performance

To compare the performance of our OZS approaches we will use the R^2 and Root Mean Squared Error (RMSE) statistics. According to Farber and Yeates (2006) the level of “adequacy”, or fit, is often in the eye of the beholder. Nevertheless, these statistics should provide a quantitative measure of the improved performance as a result of OZS. The R^2 and RMSE or similar alternatives are frequently used measures to assess model performance and have been commonly used in hedonic modelling (e.g. Farber and Yeates 2006; Derdouri and Murayama 2020; Kuntz and Helbich 2014; Schulz et al. 2014). The R^2 compares our fitted regression line to a baseline model. The baseline model is considered to be the "worst" model and consists of a flat line that predicts every value of $\hat{y}$ by the mean value of y (Cameron and Windmeijer 1997). It can be retrieved by taking the total sum of squared errors (9). Where y is the actual property valuation; $\hat{y}$ is the fitted value of our model; $\bar{y}$ is the fitted value of the baseline mode, for property i.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{9}$$

The RMSE is a frequently applied formula to measure the error rate of a regression model. RMSE is the standard deviation of the residuals and measures how concentrated the data is around the line of best fit of our model. The RMSE is represented by equation 10 (Derdouri and Murayama 2020), where again $\hat{y}$ is the predicted value; y is the actual valuation of property; i is the property and n is the total number of validation points.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{10}$$

Data

Socio-Economics

Our first dataset contains socio-economic characteristics and has been created from multiple databases of Eurostat. Eurostat is the official statistical office of the European Union and gathers its information directly from national statistical bureaus. Therefore, it is considered as a trustworthy data source. Eurostat datasets have been used for a number of research and government applications (e.g. Tabellini 2010; Janoušková and Sobotovičová 2021; Forys and Tarczynska-Luniewska 2018). The socio-economic dataset that is created is a so called "spatio-temporal" dataset, meaning that it has both a spatial and temporal dimension. At the spatial level, the data consist of NUTS 2 and NUTS 3 regions. These are statistical areas, covering economic territory of the EU and the UK and have been designed for purpose of regional statistics and socio-economic analysis (Eurostat 2015). At the temporal level, yearly observations are considered that cover a timespan from 2015 to 2020. Such a timespan should provide adequate information on socio-economics for every acquisition and valuation year of our second dataset. Socio-economic characteristics included have been based upon whether they are typically related to real estate prices. Since the socio-economic data needs to meet certain requirements (being available on NUTS 2 or NUTS 3 level for all sample locations, cover a timespan from 2015 to 2020 and have yearly observations) the socio-economic variables included are limited to data availability. The summary statistics of our socio-economic variables are provided in table 4.

NightsSpend	EducLow	EducMed	EducHigh	Scientist	EmpAgri
Min. : 149661	Min. : 2.567	Min. : 8.783	Min. : 5.167	Min. : 1.30	Min. : 0.000
1st Qu.: 2088156	1st Qu.: 78.558	1st Qu.: 277.071	1st Qu.: 153.225	1st Qu.: 27.70	1st Qu.: 1.225
Median : 4319492	Median : 131.533	Median : 446.933	Median : 253.542	Median : 48.42	Median : 3.185
Mean : 7349114	Mean : 247.498	Mean : 557.365	Mean : 345.764	Mean : 67.03	Mean : 6.543
3rd Qu.: 7811611	3rd Qu.: 253.717	3rd Qu.: 718.688	3rd Qu.: 406.775	3rd Qu.: 84.38	3rd Qu.: 8.767
Max. :67680032	Max. :2487.600	Max. :2720.167	Max. :3407.483	Max. :533.77	Max. :69.175
NA's :1	NA's :1	NA's :1	NA's :1	NA's :2	NA's :47

EmpIndustry	EmpManufact	EmpConstrct	EmpWholesale	EmpFinancial	EmpSocial
Min. : 0.475	Min. : 0.175	Min. : 0.620	Min. : 2.28	Min. : 0.6925	Min. : 2.125
1st Qu.: 10.290	1st Qu.: 9.188	1st Qu.: 4.155	1st Qu.: 15.00	1st Qu.: 7.1262	1st Qu.: 18.128
Median : 19.600	Median : 17.435	Median : 7.303	Median : 27.65	Median : 12.7418	Median : 32.200
Mean : 28.954	Mean : 26.287	Mean : 11.012	Mean : 48.38	Mean : 29.2185	Mean : 53.516
3rd Qu.: 37.125	3rd Qu.: 34.650	3rd Qu.: 13.175	3rd Qu.: 52.51	3rd Qu.: 27.6425	3rd Qu.: 58.000
Max. :359.900	Max. :327.575	Max. :167.900	Max. :1106.70	Max. :752.8250	Max. :1064.600
NA's :47	NA's :47	NA's :47	NA's :50	NA's :50	NA's :47

TranspGoods	GDP	PurchasePower	Population
Min. : 64.5	Min. : 414.7	Min. : 634.4	Min. : 20.42
1st Qu.: 4114.0	1st Qu.: 3499.0	1st Qu.: 3582.9	1st Qu.: 135.70
Median : 7819.5	Median : 6166.7	Median : 6353.7	Median : 251.85
Mean : 10847.5	Mean : 11888.6	Mean : 11655.4	Mean : 390.63
3rd Qu.: 14123.0	3rd Qu.: 11809.4	3rd Qu.: 12148.1	3rd Qu.: 471.56
Max. :122429.0	Max. :224996.8	Max. :234813.5	Max. :6484.55
NA's :27	NA's :28	NA's :28	NA's :26

Table 4: Descriptive statistics socio-economic data

It can be noticed that regional values significantly differ from one another. The gap between minimum and maximal values is in most cases considerably large. Reason for these large differences is that it involves absolute data that does not account for the size of regions and population density. Furthermore, it can be seen that in all cases data is right tailed, as the median is considerably lower than the mean and the maximum is multiple times higher than the 3rd quadrant. Also these outliers can be explained by the absolute form of the data. It makes sense that metropolitan areas such as Paris indeed have considerably higher values than average regions, simply because the size of its population.

Commercial Real Estate

Our second dataset consists of commercial real estate in Western- and Central Europe. A geographical allocation of our CRE dataset is provided in figure 7.

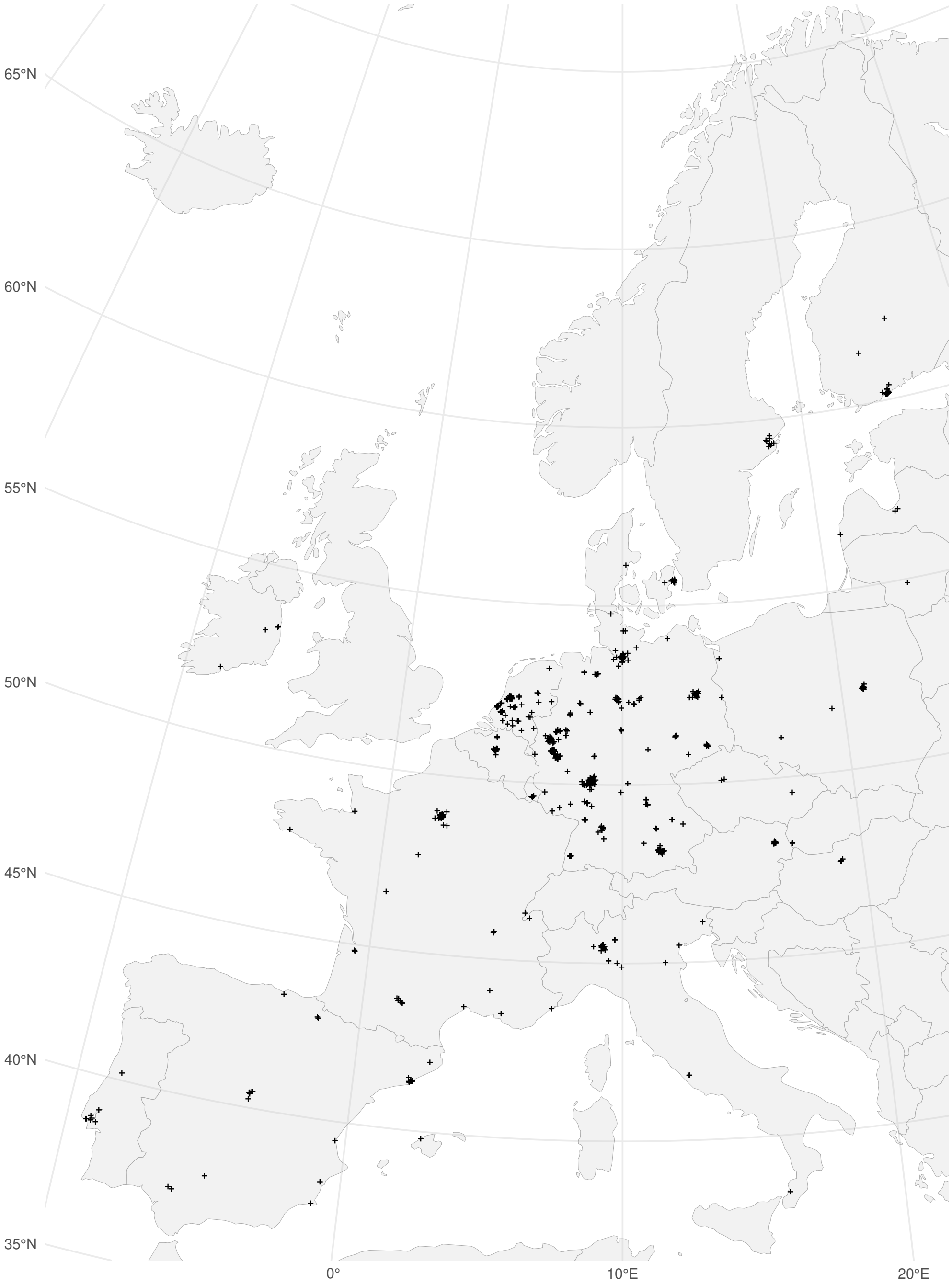


Figure 7: Commercial real estate property distribution across Western-Europe

It can be seen that properties cover a wide range of regions, which are typically associated with varying socio-economic structures. Nevertheless, density seems to be located in the Netherlands and Germany. Furthermore, properties are typically located in and around urban centres. The German Ruhr area, the Dutch Randstad, Paris, Madrid, Lisbon, Stockholm and many other economic regions can easily be identified. The over-representation of these urban areas might affect our analysis in such a way that variety in our socio-economic variables is limited. Since the socio-economic structures of these urban areas can still be very diverse, it is not expected that this concentration of CRE in urban areas will affect research outcomes. Moreover, socio-economic variables are considered at multiple scales, therefore effects of more rural areas are still included in our model.

CRE data includes yearly valuations, property characteristics and locational characteristics. The valuations have been made by at least two independent professional valuators and property characteristics span a wide range of features. The dataset has been collected by the commercial real estate research company KR&A. KR&A data is used by major institutions in the field of real estate such as Aegon, Wereldhave and ING. Furthermore, it provides data for the European Commission.

Although, KR&A data has not been used in earlier publications, it has been used for the purpose of other theses (e.g. Dröes and Minne 2015). The provided data is therefore considered trustworthy and suitable for the research aim. Nevertheless, additional data cleaning, structuring and modifications have been executed to create a workable dataset. Furthermore, geocoding was used to determine exact locations from the locational characteristics. The data modifications have been performed in R and the script is available upon request. After removing observations with missing and/or unreliable observations, a working dataset of 3287 observations remains available. The descriptive statistics of this workable dataset can be found in table 5.

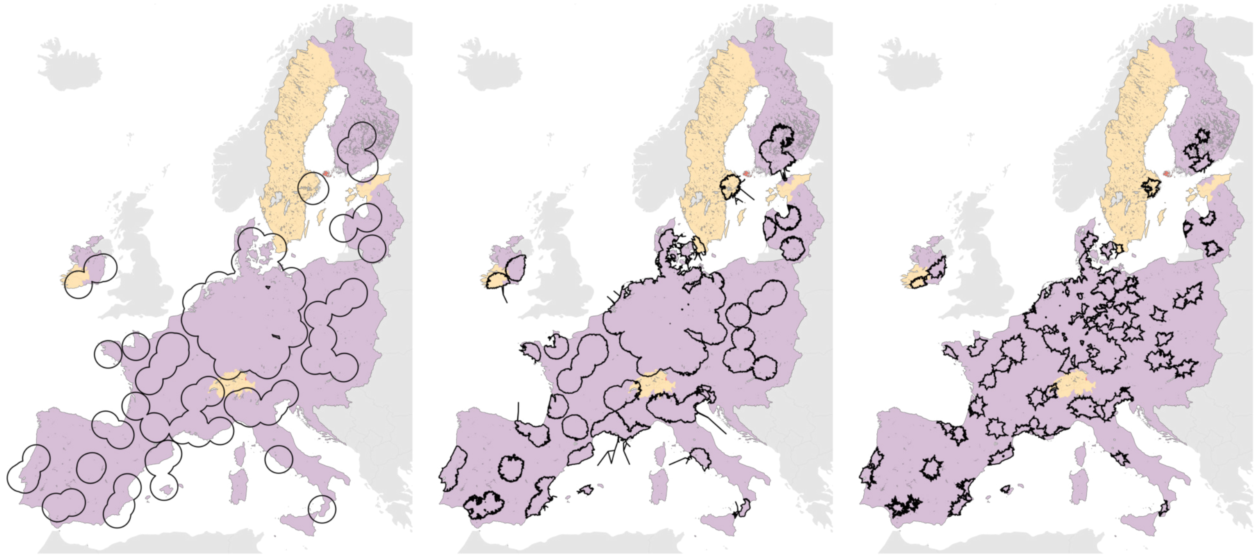
	Hotel	Logistics	Office	Residential
	(N = 177)	(N = 536)	(N = 1897)	(N = 677)
Valuation price (million €)				
min	12.38	1.12	0.25	0.78
max	253.38	476.05	781.62	82.72
mean	64.53	64.58	70.75	15.16
sd	50.06	84.27	83.4	12.44
Gross leasable area (1000 m.)				
min	2.9	0.06	0.15	0.36
max	34.77	76.33	102.9	30.6
mean	14.06	15.04	17.2	6.62
sd	7.35	17.94	12.87	5.17
Plot size (1000 m.)				
min	0.58	0.06	0.33	0.6
max	19.02	170.66	121.1	34.62
mean	5.42	16.8	7.12	6.36
sd	4.87	30.66	9.51	6.52
Renovation cohorts				
<1990	23 (13%)	58 (11%)	160 (8%)	90 (13%)
1990-2000	26 (15%)	33 (6%)	410 (22%)	165 (24%)
2000-2010	76 (43%)	310 (58%)	1,038 (55%)	155 (23%)
2010-2020	52 (29%)	135 (25%)	283 (15%)	263 (39%)
Valuation years				
2015	27 (15%)	92 (17%)	401 (21%)	84 (12%)
2016	28 (16%)	99 (18%)	389 (21%)	103 (15%)
2017	27 (15%)	90 (17%)	393 (21%)	121 (18%)
2018	30 (17%)	84 (16%)	265 (14%)	116 (17%)
2019	34 (19%)	86 (16%)	244 (13%)	131 (19%)
2020	31 (18%)	85 (16%)	205 (11%)	122 (18%)

Table 5: Summary statistics commercial real estate

Results

Optimal Zoning Systems

Results of our OZS are presented in figure 8. Because of the many overlapping OZS, at multiple scales, only united outer boundaries at macro levels are projected. Note that in practice each zone is still considered as a stand alone area. Since OZS boundaries are not mutually exclusive this overlap causes no problems and does not affect research outcomes. The figure illustrates the spatial extend at which socio-economic values are considered. Since OZS includes socio-economic values at a wider spatial level than traditional SAUs, additional data requirements should be met. This data requirement is in line with the aim of our methodology. Namely, not only the region in which a property is located affects its prices but also surrounding areas can be price influencers. Accordingly, socio-economic data should not merely cover the spatial extend of SAUs in which CRE is located, but all areas within reach of an OZS. Coloured areas in figure 8 indicate the spatial distribution of our socio-economic dataset. Blue represents complete data; beige shows regions with one or more missing variables at the NUTS 2 level; Red regions (Fasta Åland and Liechtenstein) present areas with missing values at NUTS 3 level; Grey regions have no values at all. Large lakes have been deliberately removed from the NUTS areas. Although these lakes are located within NUTS boundaries, the assumption is made that none of our considered socio-economic variables are actually located at water. Limiting the available areas for socio-economic data to land masses should better reflect actual distributions of values.



(a) Buffer (b) Drive distance (c) Drive time

Figure 8: Cumulative Optimal Zoning Systems

The majority of our OZS are located within blue areas. Aggregation for these OZS will be optimal, given our approach. However, cases do occur when OZS do overlap areas with partially missing data, or for which no data is available at all. This happens for example at OZS that overlap Switzerland, Norway and Ireland. For these OZS, particular variables will be biassed as no values would be aggregated from the zones with missing values. Accordingly, bias would be larger if the ratio of non-blue zones to blue zones increases. Furthermore, bias is dependent on the true size of the socio-economic variable at the region where data is not present. In cases of missing values, zero data from that region will be aggregated to the specific OZS that cover it. Therefore, if true values are closer to zero, less bias will occur. Since it is unlikely that true values will be close to zero, an approximation of the value is made by taking the mean surface value and multiplying it by the area of consideration, while accounting for population. This should reduce the difference between assigned value and the true values for these regions.

Disaggregation

After considering if the spatial extend of our socio-economic data adequately covers the OZS, and missing values have been estimated, the spatial extend of our data was transformed by means of Pycnophylactic interpolation. The interpolation results of our socio-economic data are presented in figure 9. In here, the socio-economic values that were initially only available at NUTS 2 and NUTS 3 level are dis-aggregated to a continuous surface. In this surface each pixel represents an estimated socio-economic value per square kilometre. For visualization, data is classed by "k-means" in which darker areas indicate low values and brighter orange areas indicate high values. Furthermore, it should be noted that the scale of each raster is independent. Accordingly, specific colours of individual socio-economic variables do not necessarily entail the same value. This classification is purely for visualisation and does not affect our actual research outcomes, in which actual interpolated cell values were applied.

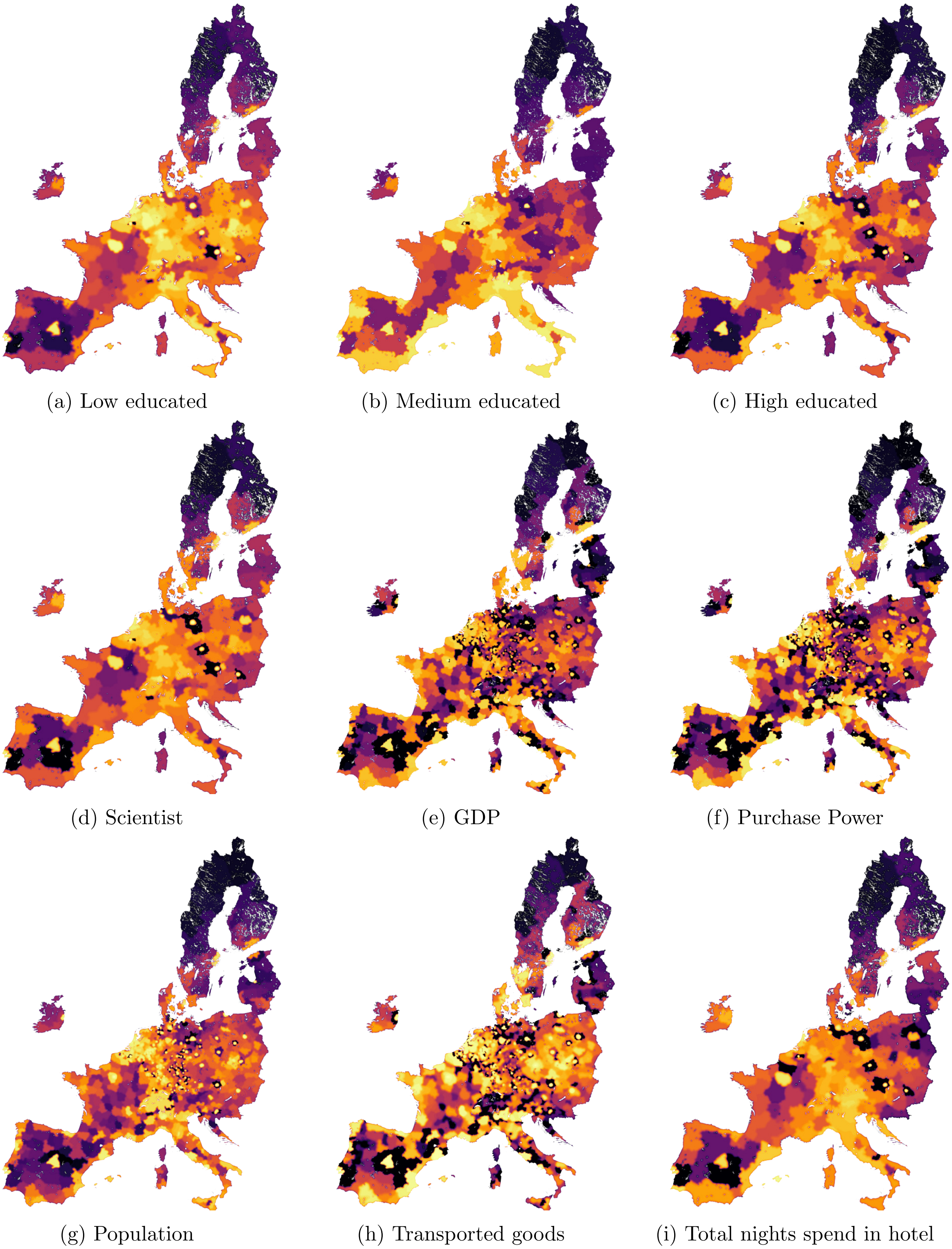
For disaggregation, pixel size was set to a squared kilometre. This level of accuracy provided a workable trade-off between detail and computing efficiency. Since our dataset covers large parts of Europe, the spatial extent of our dataset is rather large. Hence, computing time took approximately six hours per socio-economic surface. However, choosing smaller pixels and thus higher accuracy would cause computing times to grow exponentially.

Since interpolation is performed with extensive variables, which are partly dependent upon population sizes, the raster maps show clear similarities and patterns. Consequently populous areas, such as Paris, will clearly stand out for almost all variables. Nevertheless, when considering the socio-economic rasters in more detail, clear distinctions can be noticed. Figures 9a to 9d illustrate that although education levels follow a population density pattern, the center of gravity shifts westwards as the level of education increases. It can also be noticed that high educated population and scientists (figs. 9c and 9d), are more frequently located in urban areas. Especially for Berlin, Prague, Vienna and Budapest this contrast between urban regions and their surroundings is well pronounced. A similar effect can be seen for employment numbers of sectors that are typically urban. When considering figure 9l, urban regions can clearly be identified as small bright points on the map, with financial centres such as Paris, Brussels and Frankfurt as absolute outliers. On the other hand, employment numbers for sectors that are typically more often located in regional areas are more widely distributed and the image is more blurry. An example would be agriculture (9j). Although cities can still be recognized, their values become less extreme and a wider range of regions is highlighted.

As such, we can distinguish two types of spatial distributions: clustered variables and random variables. The blurrier an image gets, the more random the distribution. It might be recalled that the spatial distribution of data affects MAUP implications. In figure 2, it was illustrated that smoothing of randomly distributed data remains limited, despite the zoning system used. The opposite is true for clustered data. When measured at an improper scale, significant smoothing can occur for these variables. Therefore, clustered variables are expected to show greater heterogeneity when measured through different zones.

Despite interpolation, the distinction between variables that were initially available at NUTS 2 or NUTS 3 level can still be identified for some variables. Variables of which the initial data was available at NUTS 3 level remains more detailed and highlights more specific regions. When we for example consider the number of scientists (9d) it can be seen that the entire provinces of North Holland, South Holland and Utrecht (Netherlands) are highlighted. However, in reality it is more likely that scientist are centred in the major urban areas of the Randstad, where highly skilled jobs are available. Consequently, the presence of large differences of socio-economic values within initial statistical zones might still bias socio-economic values even after interpolation. As a consequence, the number of scientists in urban regions might be underestimated, while values in rural areas are overestimated. For socio-economic values that require less human capital and more physical attributes such as space, this might be the other way around. Especially at the micro scale, these inadequacies could bias regression results as wrong values are associated with properties. For larger scales this should be less troublesome. If the entire region is associated with a property, the mean value remains equal despite how values are distributed within regions. Furthermore, even if the macro scale does not fully encompass a region with large intrinsic difference, bias will be less severe as the errors from this region will account for a smaller amount of the total value.

Nevertheless, creation of continuous surfaces that represent distributions of socio-economic variables ever more accurately remain an objective of interest. Indeed, this could be achieved by including low scale initial data, such as NUTS 3 (or even smaller) levels. On the other hand, more sophisticated interpolation techniques, as discussed in our literature review, could provide a solution as well. An example would be to only allow values to allocate at places where this is physically possible. In our analysis this was partly done by removing large lakes. However, also the removal of other barriers such as highways and mountains could be considered. Ideally, only the spatial areas where specific socio-economics can indeed occur should be utilized. However, since this is heavily dependent on the variable of consideration, these areas should be selected carefully and come with high data requirements.



(a) Low educated

(b) Medium educated

(c) High educated

(d) Scientist

(e) GDP

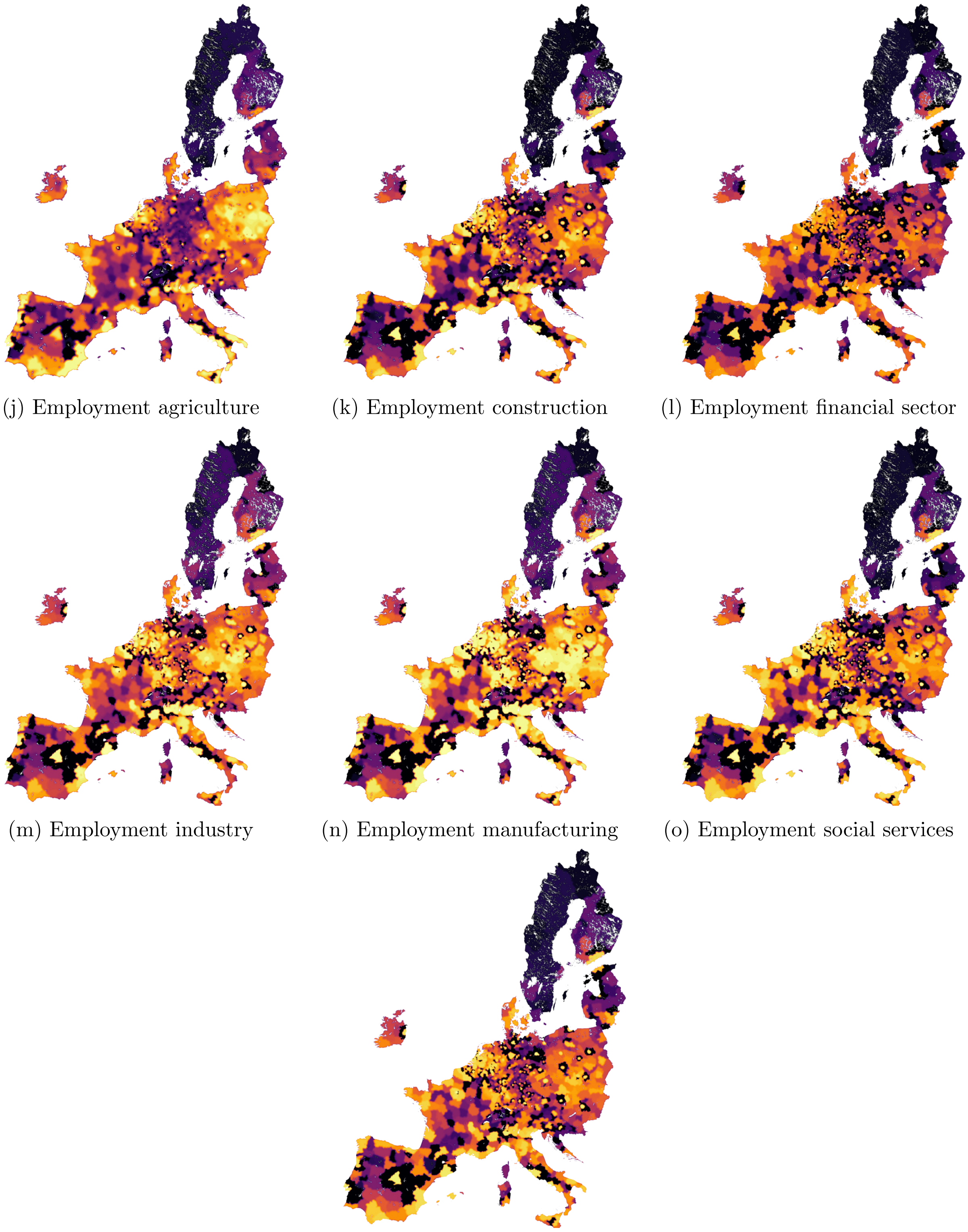
(f) Purchase Power

(g) Population

(h) Transported goods

(i) Total nights spend in hotel

Figure 9: Dis-aggregated socio-economic data to continuous surface



(j) Employment agriculture

(k) Employment construction

(l) Employment financial sector

(m) Employment industry

(n) Employment manufacturing

(o) Employment social services

(p) Employment wholesale

Figure 9: Dis-aggregated socio-economic data to continuous surface (cont.)

Reaggregation

After areal interpolation and creation of optimal zoning systems, functional statistical areas are created through reaggregation. Since only Gaussian data is considered, no additional steps are necessary to calculate rates. Reaggregation to functional statistical areas is done by simply taking all pixels that fall within an optimal zoning system and returning the sum of their values. Since FSAs are already related to specific properties, assignment of location characteristics to individual properties is easily performed by joining them. Assignment of location values for SAUs is done through the traditional method.

Summary statistics for each considered (functional) statistical area are provided in table 6. In here mean values and standard deviations are listed per opportunity, impedance and scale. Since SAUs are only considered at a single scale (which is in line with traditional research methods) values are equal across scales. Regarding the meso and macro level, minimum radius are still treated as zero. Therefore, larger scales would always include all smaller scales automatically. Consequently, values per opportunity and impedance should rank from micro ≤ meso ≤ macro. To retrieve the ringlike zones that will be used for our interaction model, one can simply deduct minor scale values from the scale of consideration.

Examining our data more closely, it can be seen that there is a huge variance within socio-economic values among property locations. Frequently, standard deviations exceed mean values by far. Nevertheless, this should not be of any concern. As argued, property locations vary significantly. Whereas some are located in remote areas, others are located in dense city centres. Therefore, outliers might seem extreme, but do grasp actual values.

Considering mean values, it can be noted that these increase as zones get larger. This happens both at the range from micro-to-macro, and at the range of time-to-buffer. Although this increase is somewhat obvious, it illustrates the discussed scale effect, that is so often forgotten. Also the standard deviation is affected by zoning sizes. Similar to mean values these increase as zones increase in size. However, ratios between SD and mean values decrease as zoning sizes get larger. This illustrates that smaller zones better reflect extreme values, which is in line with our theory.

	SAU		Buf		Dist		Time	
Variables	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Population (thsd)								
Micro	1329	1099	964	1066	720	913	177	196
Meso	1329	1099	4649	2844	3968	2702	1954	1636
Macro	1329	1099	9624	4605	8035	4100	5487	3155

Education low								
Micro	484	468	40	55	28	46	7	13
Meso	484	468	399	244	310	195	103	81
Macro	484	468	1025	611	825	513	471	296
Education medium								
Micro	1110	664	98	122	66	93	16	22
Meso	1110	664	1054	510	829	439	270	205
Macro	1110	664	2716	1274	2179	1066	1260	692
Education high								
Micro	925	857	86	119	60	95	15	24
Meso	925	857	813	539	649	456	216	182
Macro	925	857	1840	1018	1542	968	929	583
Scientist								
Micro	188	130	18	24	12	19	3	4
Meso	188	130	174	94	138	80	47	38
Macro	188	130	398	197	331	173	201	109
Nights spend (thsd)								
Micro	19527	18552	2026	3284	1449	2573	364	650
Meso	19527	18552	15324	11126	12492	9924	4516	4731
Macro	19527	18552	32908	17834	27487	17594	17068	11293
Vol. transp. goods (mil)								
Micro	47	97	18	18	13	14	3	3
Meso	47	97	90	73	77	61	37	28
Macro	47	97	173	115	147	105	100	72
GDP (mil)								
Micro	152	213	114	206	92	181	23	42
Meso	152	213	271	324	248	316	168	257
Macro	152	213	379	347	348	343	286	318
Purchase power (mil)								
Micro	145	215	105	183	84	161	21	37
Meso	145	215	251	296	230	287	155	230
Macro	145	215	351	321	323	317	265	291
Emp. agriculture (hnd)								
Micro	48	109	12	13	8	9	2	3
Meso	48	109	115	112	88	85	32	35
Macro	48	109	320	255	251	207	134	121

Emp. industry (hnd)								
Micro	1449	2933	712	848	537	694	133	162
Meso	1449	2933	2536	2515	2241	2281	1251	1277
Macro	1449	2933	4448	3401	3861	3176	2805	2443
Emp. manufacturing (hnd)								
Micro	1267	2624	602	705	453	577	112	136
Meso	1267	2624	2217	2211	1947	1987	1064	1059
Macro	1267	2624	3978	3070	3437	2850	2463	2149
Emp. construction (hnd)								
Micro	714	1429	360	489	273	399	66	87
Meso	714	1429	1337	1547	1196	1440	657	818
Macro	714	1429	2074	1787	1854	1732	1431	1508
Emp. wholesale (hnd)								
Micro	5250	9362	3409	5673	2731	5002	672	1145
Meso	5250	9362	8772	10252	7973	9597	5178	7167
Macro	5250	9362	12598	11713	11513	11429	9306	10041
Emp. financial (hnd)								
Micro	4430	6732	3248	5565	2640	4938	664	1169
Meso	4430	6732	7489	8647	6893	8346	4724	6796
Macro	4430	6732	10346	9373	9546	9210	7938	8480
Emp. social (hnd)								
Micro	5437	9445	3536	5632	2864	4978	718	1159
Meso	5437	9445	9064	10236	8201	9595	5275	7052
Macro	5437	9445	13373	11533	12145	11299	9686	9992

Table 6: Socio-economic summary statistics per statistical area specification

Consequently, zones should ideally be small while still capturing the spatial extend of consideration. A reduction of zoning size could be achieved by considering each as individual rings. E.g., the mean population value measured at the meso scale would become $\text{Buf}_{meso} - \text{Buf}_{micro} = 4649 - 964 = 3685$.

Although our research setup causes higher scales still to contain larger areas, these no longer measure overlapping values. Therefore, micro values could potentially be higher than macro level values, as is the case for purchase power measured through the buffer impedance. When treated as separate rings, the mean micro value would remain 105 million, while the macro level would be reduced to 100 million. This spatial distribution makes sense considering that CRE properties are not developed at random. Developers choose locations where demand is high and profits could be made. Purchase power could be a good estimator for this demand and is thus, not surprisingly, relatively high at the local scale. Because of this heterogeneity at different scales, measuring them individually could provide additional insights. The difficulty here is that these rings cannot be used as a stand alone. This would exclude other location characteristics that might be even more close. Accordingly, increasing zoning details comes at the cost of model sophistication.

To compare how different statistical areas behave, aggregated results are examined more closely in figure 10. In here distributions of values are visualized per impedances and opportunities in so called violin plots. These are hybrids between box plots, and kernel density plots. Unlike box plots they are not limited to summary statistics, but show densities at each value. Doing so they visualize not only standard indicators such as minimum and maximum values, but also provide an overview of how values are distributed overall. E.g. in the case of a normal distribution, the plot should contain a rotated concave parabola on either side. Since our socio-economic location characteristics contain outliers, distributions are visualised as natural logarithms. This allows us to examine all separate scales within a single frame, and keep a high level of detail.

When comparing these violin plots in shapes and lengths across impedances, scales and opportunities, a number of things can be noted. A similar pattern exists as in table 6, where values are higher at larger scales and lower at smaller scales. However, in addition to these summary statistics, violin plots indicate how different scales overlap and relate to each other. This provides an understanding of data alterations that have been made.

The extend by which different zoning measures overlap differs significantly per socio-economic variable and could provide valuable insights. While different scales are easy to distinguish for certain variables, they are almost identical for others. The amount of overlap tells us about the ratios at which specific distances affect total value outcomes. Great overlap indicates that measuring the same location characteristic at a larger scale hardly changes outcomes. Therefore, the question rises if the extra information that these large scales provide is worth the risk of additional MAUP effects. For these variables, the meso or macro scale could be considered as zoning systems that measure more than one cluster of different observations. As such, they take a format as was illustrated in the bottom right zone of figure 2 d, with high values at the local extend and low values at the wider area. Clear examples of such overlapping scales are purchase power, employment in the financial sector and GDP. Especially at the meso and macro scale, these variables are almost identical in shape and size.

The opposite is the case for scales that can easily be distinguished. Although an increase in values is expected as zones get larger, this increase could be out of proportion. In this scenario, wider areas affect total values relatively much. As such, multiple clusters could be measured with a single zone. However, instead of high values at the local extend, values in here are high at the wider area. Therefore, such areas are more similar to the bottom left zone of figure 2 d. Examples of this scenario are the education levels. Scales of these variables show little overlap. For these variables a single zoning measure might not suffice to capture the varying spatial characteristics over distance, as significant smoothing can occur. Consequently, significant differences in performance are expected for zoning measures of these variables. Because of this heterogeneity over space, including these values as separate zones, might prove to be extra valuable. Therefore, these variables are expected to profit most from our interaction models.

For the variables in which values increase proportional to the sizes of zones, we expect the opposite. This proportionality indicates that spatial characteristics are relatively equal across distance. Reason for this could be a random distribution of observations, or that all considered scales merely cover a part of the same cluster. Therefore, these variables are expected to perform relatively similar over different scales. However, these variables might benefit less from our spatial interaction model. Examples of these variables are employment in agriculture and population numbers.

The extend by which scales overlap is caused by a combination of the spatial distribution of the variable of interest (figure 9) and the locations of our CRE data (figure 7). Since properties of consideration are mostly located in urban areas, variables that are typically centralized in cities show greatest overlap between different scales. Reason for this overlap is that as long as the micro scales cover these urban areas, the most important determiners of total values are captured. Therefore, widening the spatial extend only affects total values by a limited amount. For the same reason, variables that are more intrinsic to rural areas, show least overlap. Examples of such urban variables, are employment in financial services and purchase power. An example of a more rural variable is employment in agriculture.

It could be noticed that little overlap can be found for education levels and numbers of scientists, even though these variables are not typically rural. Reason for this lack of overlap is likely to be caused by the fact that these variables used NUTS 2 areas as initial data input. Since this data is already smoothed significantly, interpolation has failed to capture specific location characteristics intrinsic to certain regions.

A similar comparison can be made for the extend by which SAU values match their FSA counterpart. Sometimes SAU values cover the whole spectrum, while they are more identical to a single scale for other variables. The extend by which the SAU measures cover micro, meso and macro scales provide an indication of the amount by which single SAUs contribute to total values. For most variables, the maximum value measured at the SAU is just a fraction lower than the maximum value measured through the macro buffer. This suggests that these outliers are predominantly obtained through values that belong to single SAUs. Nevertheless, this statement should be made carefully as outliers of different measures do not necessarily entail the same property. Moreover, this only involves the extremes.

To get an indication of the extend by which single SAUs contribute to total values it is more useful to compare the center of gravity for different zoning measures. NUTS 3 zones are typically slightly bigger than our micro scales and slightly smaller than our meso scales. Therefore, they are expected to fall somewhere in between. The same goes for NUTS 2 zones, however these are more likely to fall within the extend of meso and macro scales. For most variables this seems to be the case indeed. For these variables data alterations remain limited and although multiple SAUs can now be associated to a single point, no drastic changes are made. Nevertheless, this does not entail that no improvements have been made. Small modifications could still reflect reality more properly.

More important than the changes in the extend of values and centres of gravity are the changes in shapes and continuity. The shapes of distribution provide valuable insights in adequacy of how location characteristics are measured. Since SAU measures treat all properties within a zone equal, distributions are more grilled like. Accordingly, each zone represents a single grill in the distribution. Surely, zonal values might overlap and values get smoothed in order to obtain readable plots. Therefore, we will not be able to identify all unique values as individual grills. Nevertheless, the grilled like structure of SAUs can clearly be identified for almost all variables in our figure.

Theoretically, the travel time measure at the micro scale should reflect location characteristics of individual properties most accurately. Therefore, it is expected that this measure provides the most continuous results. Reason for this is that although features close in proximity share locational characteristics, these are slightly different. Therefore, values should change gradually over space instead of directly. Also travel distance, should reflect local location characteristics relatively well. Accordingly, little grills are expected for this measure as well. The same goes for the micro buffer measure, but then to a lesser extend.

As scales get larger, these small differences become less significant. Consequently, grill like shapes should become more visible as scales increase. To summarize, it is expected that distributions get more continuous as the accuracy of measurements improve. When we examine the different distributions, it can be seen that this hypothesis indeed turns out to be true. Overall, distributions of the travel time micro scales are most smooth. SAUs and macro scales are more grilled.

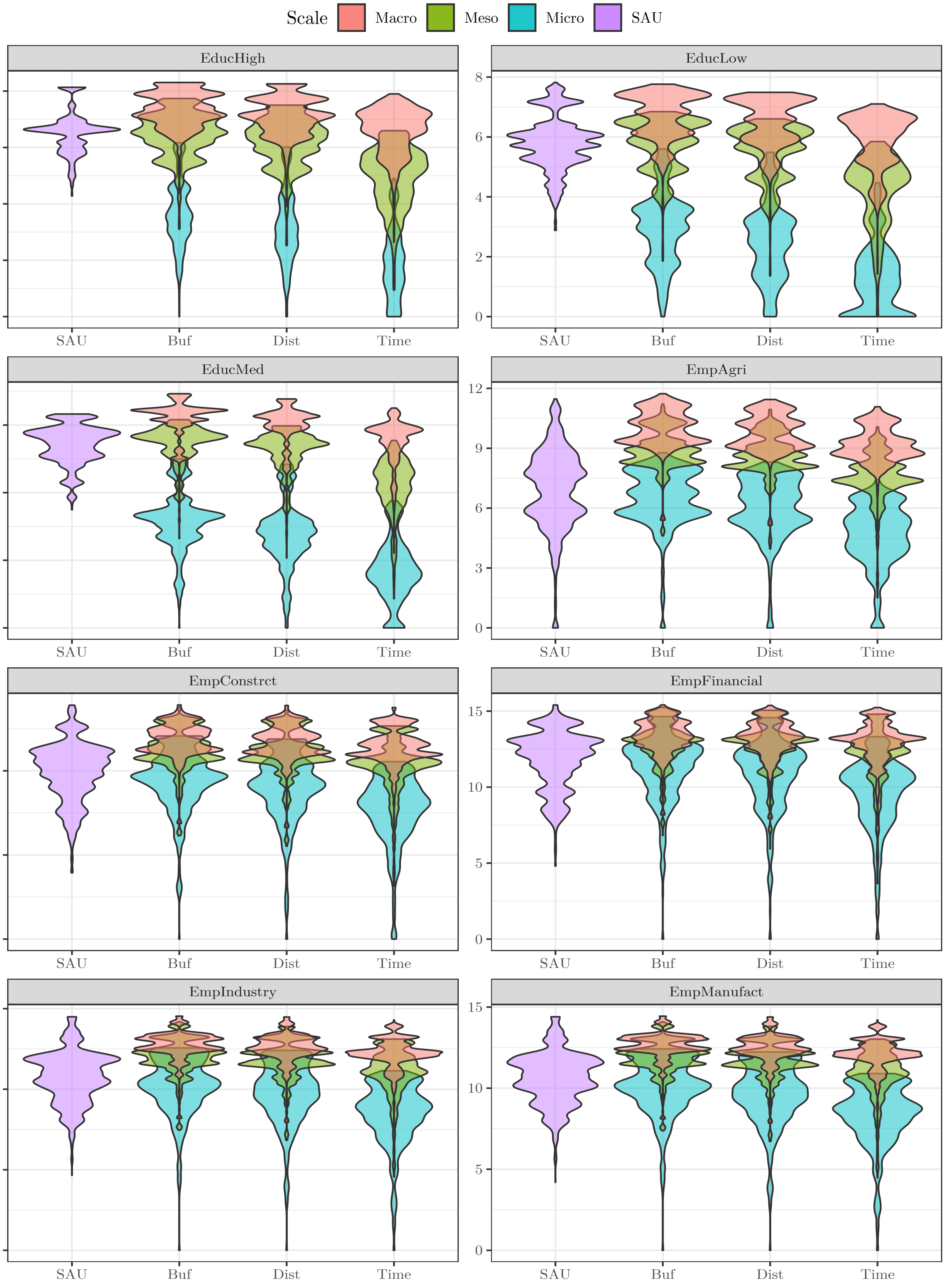


Figure 10: Distribution of socio-economic values, measured at alternative impedances

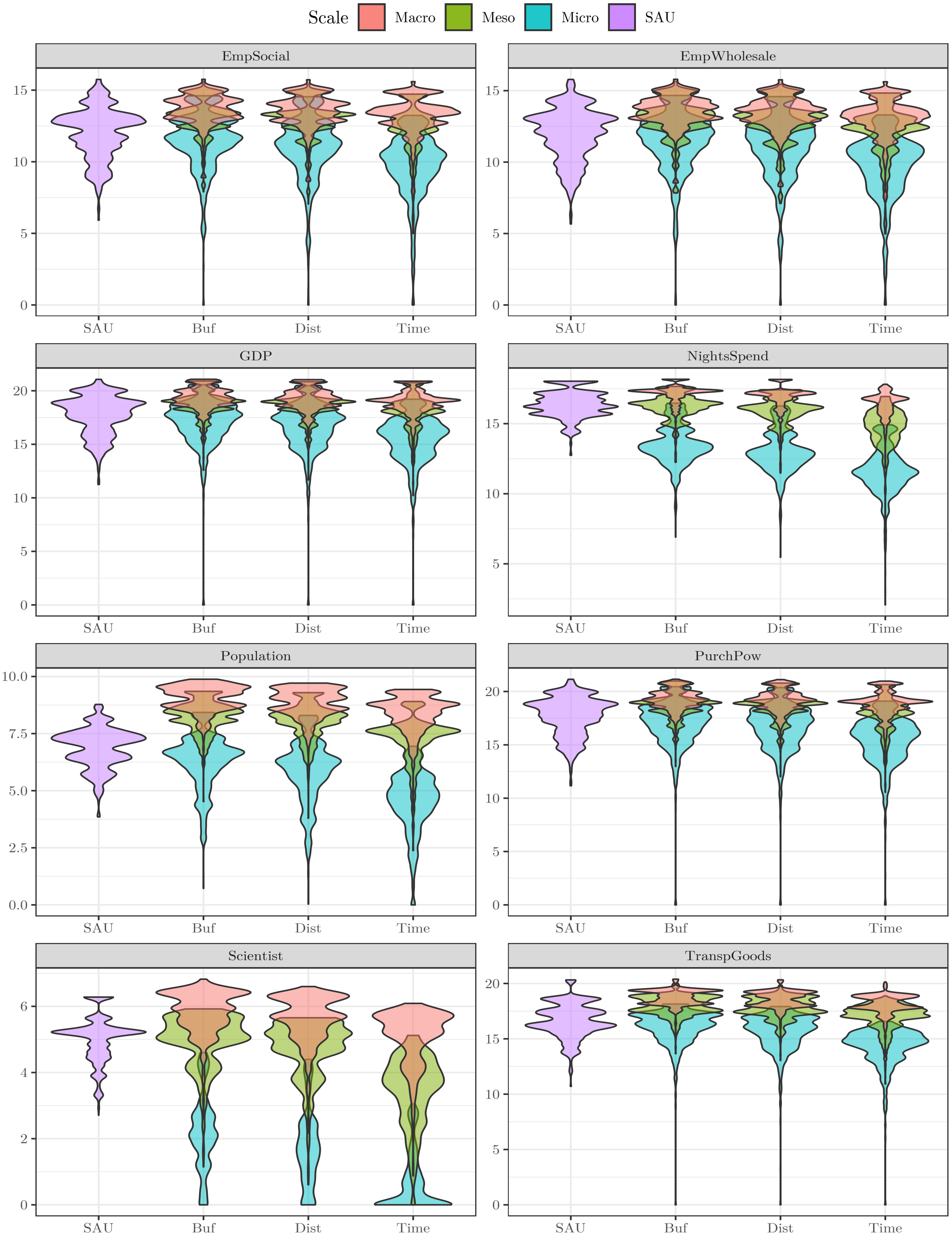


Figure 10: Distribution of socio-economic values, measured at alternative impedances (cont.)

Performance

Hedonic model

Sixteen regression results of our hedonic model are presented in figure 7. Each model considers a certain property type and population values per zoning system. Since all building characteristics that serve as control variables are held equal, the only differences per property type can be found in the zoning systems used for inclusion of the population. These involve the micro scale for our FSAs. All models as a whole are significant. Therefore, we reject the null-hypothesis that a model without independent variables fits our data equally well. Furthermore, it could be noticed that in most of our models all our coefficients are significant, with some exceptions for the refurbish date. Accordingly, we assume that there is indeed a relation between our independent and dependent variables. Moreover, no anomalies were found when testing if our models meet the ordinary least square assumptions.

	Office				Residential			
	SAU	Buffer	Distance	Time	SAU	Buffer	Distance	Time
(Intercept)	14.13^{***}	14.08^{***}	14.13^{***}	14.09^{***}	11.25^{***}	11.04^{***}	10.98^{***}	10.92^{***}
	(0.39)	(0.36)	(0.36)	(0.37)	(0.48)	(0.47)	(0.46)	(0.46)
GLA	34.68^{***}	33.43^{***}	33.51^{***}	33.49^{***}	16.72^{***}	15.95^{***}	15.80^{***}	15.73^{***}
	(0.66)	(0.60)	(0.60)	(0.63)	(0.56)	(0.55)	(0.55)	(0.54)
GLA^2	-13.19^{***}	-12.96^{***}	-13.07^{***}	-13.13^{***}	-7.39^{***}	-7.28^{***}	-7.33^{***}	-7.24^{***}
	(0.61)	(0.55)	(0.54)	(0.57)	(0.44)	(0.43)	(0.43)	(0.42)
Plot size	-0.02^{***}	-0.01^{***}	-0.01^{***}	-0.02^{***}	-0.03^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Refurbish date	0.00	0.00^{***}	0.00^{***}	0.00^{**}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Valuation date	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Population	0.17^{***}	0.27^{***}	0.31^{***}	1.20^{***}	0.06^{***}	0.32^{***}	0.46^{***}	1.69^{***}
	(0.01)	(0.01)	(0.01)	(0.06)	(0.01)	(0.04)	(0.05)	(0.18)
R^2	0.66	0.72	0.73	0.70	0.72	0.73	0.74	0.75
Adj. R^2	0.66	0.72	0.73	0.70	0.72	0.73	0.74	0.74
Num. obs.	1897	1897	1897	1897	677	677	677	677

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Hedonic Model of all impedances in micro scale with population as opportunity

	Logistics				Hotel			
	SAU	Buffer	Distance	Time	SAU	Buffer	Distance	Time
(Intercept)	15.27^{***}	15.39^{***}	15.37^{***}	15.41^{***}	14.78^{***}	14.22^{***}	14.20^{***}	14.32^{***}
	(0.94)	(0.95)	(0.95)	(0.95)	(0.81)	(0.71)	(0.70)	(0.72)
GLA	33.50^{***}	33.12^{***}	33.10^{***}	33.11^{***}	9.81^{***}	9.49^{***}	9.50^{***}	9.60^{***}
	(1.05)	(1.06)	(1.06)	(1.06)	(0.43)	(0.40)	(0.39)	(0.40)
GLA^2	-1298^{***}	-1302^{***}	-1302^{***}	-1305^{***}	-2.77^{***}	-2.42^{***}	-2.43^{***}	-2.68^{***}
	(0.79)	(0.80)	(0.79)	(0.79)	(0.41)	(0.36)	(0.36)	(0.36)
Plot size	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.05^{***}	-0.03^{***}	-0.03^{***}	-0.04^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Refurbish date	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00^{***}	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Valuation date	0.00^{*}	0.00^{*}	0.00^{*}	0.00^{*}	0.00^{***}	0.00^{***}	0.00^{***}	0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Population	0.10^{**}	-0.08^{*}	-0.10^{*}	-0.61^{**}	-0.00	0.19^{***}	0.22^{***}	0.74^{***}
	(0.03)	(0.03)	(0.04)	(0.24)	(0.02)	(0.03)	(0.04)	(0.14)
R^2	0.76	0.76	0.76	0.76	0.78	0.81	0.82	0.81
Adj. R^2	0.76	0.75	0.75	0.76	0.77	0.81	0.81	0.80
Num. obs.	536	536	536	536	177	177	177	177

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Hedonic Model of all impedances in micro scale with population as opportunity (cont.)

Coefficient sizes and directions are rather stable per property type and across zoning systems. Moreover, most coefficient signs are in line with our expectations. Exceptions are the population coefficient for logistics. It could be noted that these are negative for all three FSA measures. Furthermore, a pattern seems to exist in the magnitude of population coefficients. These get larger from SAU to Time. The firmer is likely to be explained by the fact that our FSAs at micro level cover a much smaller area than the SAU. Whereas logistics properties could profit from accessibility to a large population at bigger scale, the same could lead to congestion problems at a smaller scale. The latter could be explained by the same differences in scales. As indicated in table 6, population values at the microlevel get smaller by each zoning system. Therefore, the relative change of an increasing population differs per zoning system and magnitudes vary.

Since the defined hedonic models are relatively simplistic, the actual size of the adjusted R^{2}s by themselves are not so compelling. Indeed, some of the most defining building characteristics have been included, with the exception of location. However, there are many more unique building features which are not considered. Therefore, a predictability of 66% for offices up to 81% for hotels seems reasonable. Nevertheless, when considering how the adjusted R^2 of different models relate, we can see some interesting results. It could be noticed that the use of FSAs to include population indeed seems to improve hedonic models. In 9 out of 12 cases, the adjusted R^2 of models in which the population was included based on buffers, distance or travel time, outperformed those in which population was based upon SAUs. For the remaining three models values are only slightly below or equal to those of SAUs.

General performance comparison

To test the performance of our FSAs for the remaining socio-economic indicators, our first hedonic model was fitted for each dependent variable per property type, impedance and scale. Similar to the regressions in table 7, all control variables were held constant so that outcomes are comparable and differences can only be caused by the zoning system used. The consideration of 16 different opportunities at 3 impedance levels with each another 3 scales for 4 properties types; plus those of the SAUs which consider the same opportunities and properties but only at a single scale, provided a total of 640 regression results.

Figure 11 provides an overview of the model statistics averages per zoning system. Impedances are included as a whole (horizontal black lines) and per scale. Since SAUs are only considered at a single scale, the micro, meso and macro level are all equal. The adjusted R^2 provides a statistical measure of the relative goodness-off-fit. It is based upon the proportion of variance explained by our models, adjusted to the number of predictors, and should ideally be close to one. The RMSE is an absolute measure of the goodness-off-fit, and considers the standard deviation of the residuals. A lower RMSE is preferable. The p.value is a significance measure of the independent variable. In case of non-random measurement error, this value should be closer to zero, as values are more likely to be correlated to actual prices.

When considering the mean values per impedance, it can be seen that the use of FSAs provides overall favourable outcomes. All impedances contain a higher adjusted R^2 than their SAU equivalent, while RMSEs are reduced. As such, both measures indicate a better model fit for FSAs. Although, p.values are higher for the buffer and distance impedance, they are considerably lower at the time impedance. At the micro scale, results are even more evident. For all impedance types and model performance measures, an improvement over the traditional SAU can be seen. These results support our hypothesis. Indeed, the use of micro scale FSAs seems to provide accurate models, while using the exact same initial data input and model specification.

Considering the three impedance measures, it can be seen that travel time provides overall best results. Travel distance, performs slightly worse, and the buffers rank lasts. Also these results are line with our expectations. As might be recalled, travel distance and travel time are more advanced measures of accessibility, with the latter being most sophisticated. The improved performance of the travel methods indicate that these impedances indeed reflect consumer perception more properly (Heyman et al. 2019).

Of the three considered scales, the micro scale clearly performs best. At all impedances and considered model statistics, the micro scale outperforms the other two. Moreover, an descending

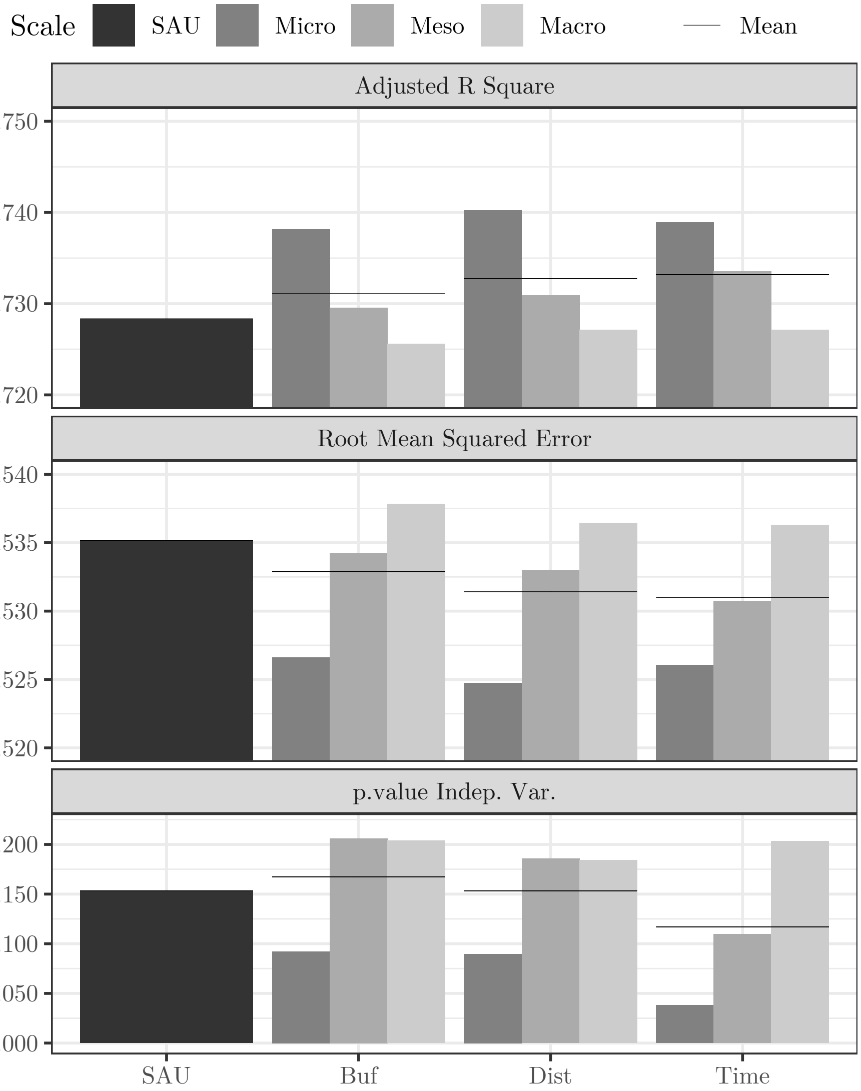


Figure 11: Average model statistics by impedance and scale (640 regression results; including all independent variables, property types, scales and impedances)

pattern in performance seems to exist when going from the micro level to the macro level. In most statistics, the meso level ranks second and the macro level ranks third. Also these results are in line with our theory. It has been indicated by Wong (2009) that zoning systems should ideally capture low level values with little variation. Decreasing values at higher scales likely imply that these are too broad and fail to capture heterogeneous features of property markets (Jones and Watkins 2009). The other extreme, in which zones are too small to capture underlying spatial phenomena, does not seem to occur.

Although the use of FSAs shows favourable effects for the whole of properties in our dataset, this does not necessarily have to be true for individual property types. Equally to the previous figure, fig. 12 shows the summary statistics of our 640 regressions. However, it separates results by property type.

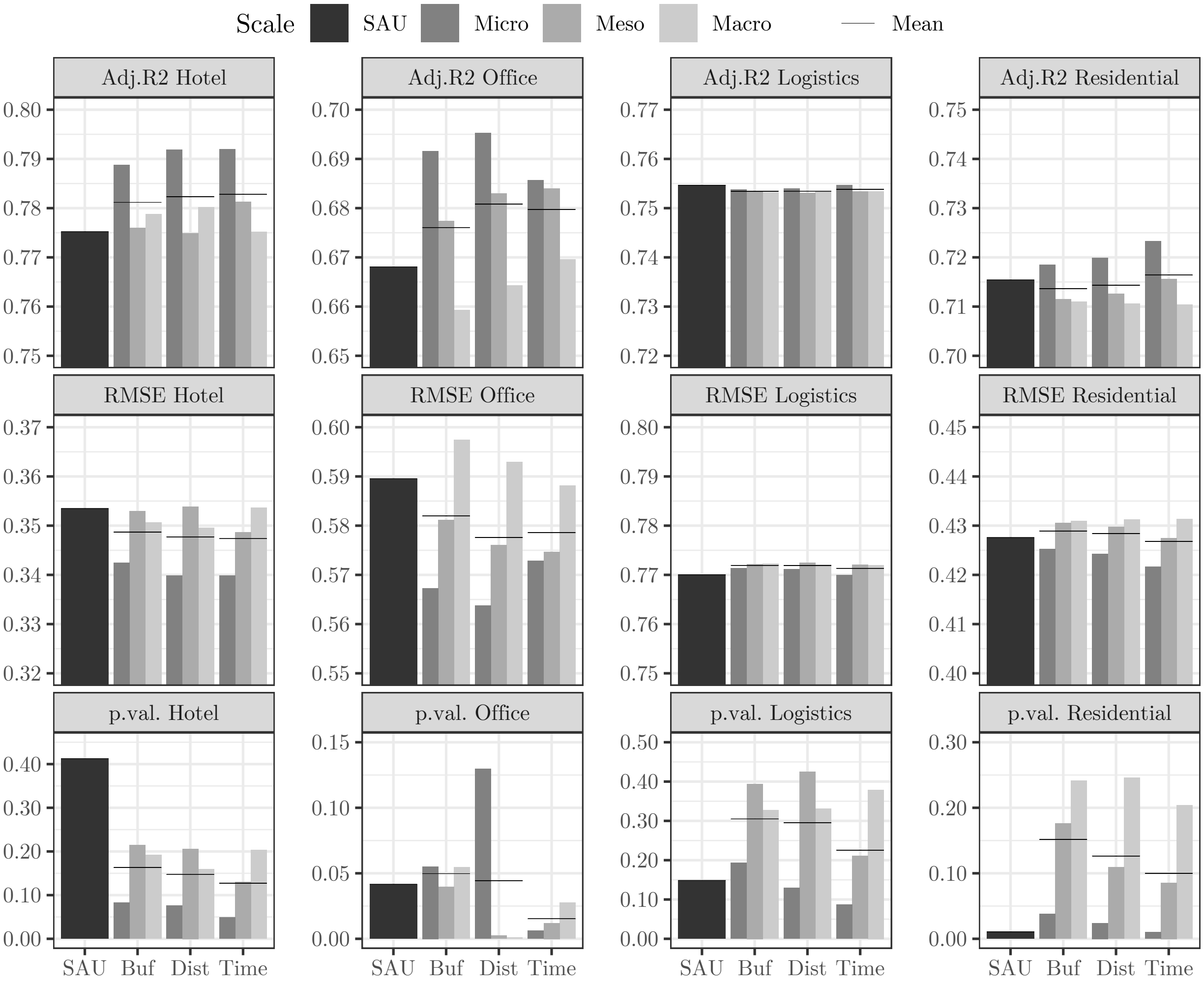


Figure 12: Hedonic model summary statistics by asset type, impedance and scale

For hotels we find a similar pattern to our previous model. All FSA means outperform the SAU in all three statistics. Again, the micro scale seems to do best. However, it could be noted that at the buffer and distance impedance the macro scale seems to do better than the meso scale. The underperformance of the meso scale could be an indication that users of hotels predominantly consider the local and national scale. This pattern indeed makes sense when considering the business model of a hotel. One could imagine that customers typically choose a destination by country and local location features, and pick a hotel that is located nearby. Therefore, semi-distant location features might be considered as less important to tenants. The outperformance of the meso scale for the travel time impedance, could be an indication that this zone is still considered to be locally accessible.

For offices, adjusted R^2 and RMSE results are in line with previous results. Again all three impedance measures outperform the SAU, and travel methods provide better results than the buffer. However, travel distance seems to be the top performer here instead of travel time. The underperformance of travel times might indicate that this impedance does not adequately reflect how office users consider accessibility. This could indeed be the case. Frequently, offices are located near excellent public transport nodes, which is not considered in our travel method. Moreover, the travel time measure does not account for structural congestions that might exist during rush hours. Although the other two measures fail to capture these factors as well, they take a more general shape. A more neutral measure could be beneficial in such circumstances. On the other hand, the p.value measures of the independent variable give a contrasting result. Socio-economic variables for offices prove to be more significant for SAUs than for buffers and travel distance. For the latter, this drop in significance is mostly caused by the micro scale, which has a remarkably high relative value when compared to the other impedances. Especially the contrast to the significance level of the travel time micro scale is striking, as this zone shows most similarities in shape and size. Interestingly enough, it is exactly at this FSA that adjusted R^2 and RMSE provide best average results. It could thus be that at this zone our hedonic model correctly recognizes that some socio-economic do not impact office prices.

For logistical properties, the use of FSAs does not seem to boost performance considerably. When considering the mean scale results, it could even be seen that model fit slightly decreases. Also the significance of the socio-economic indicators seems to drop. Only at the micro travel time impedance the use of FSAs outperforms the SAU method. However, differences are marginal. The homogenous results for logistic properties might be explained by two possibilities. First, it could be possible that none of our socio-economic variables affects logistics properties enough to cause a significant change in outcomes. In such a case, our control variables are likely to explain most of the variation in logistic property prices. Second, none of our FSAs reflects the area of interest to logistics any better or worse than SAUs. Logistic centres typically serve a large area with heavy goods vehicles. Therefore, even our macro scales might be too small to capture the areas of interest for these buildings. Moreover, use of the driving-cars profile for travel methods might not reflect actual accessibility adequately. Nevertheless, a drop in the p.value of the independent variables can be seen at the micro level for travel distance and travel time. This suggest that use of FSAs could provide an increase in model consistency for logistic properties.

For residential properties the use of FSAs only seems to improve mean model fits for the travel time impedance. Nevertheless, when considering individual scales it can be seen that model fit increases for all impedance types at the micro scale. Regarding the p.value, SAU values prove to be more significant. Only at the travel time impedance at the macro scale, values are below those of SAUs. The relatively high SAU performance might indicate that for residential properties, SAU zones could indeed provide a relevant measure for location features. This seems well possible, as NUTS boundaries do follow political boundaries, such as municipalities. Residents might be more inclined to base location decisions on subjective measures, such as belonging to a specific place. The socio-economic characteristics at the SAU level, could indeed provide as a instrumental variable, that links residential properties to specific places.

Nevertheless, a clear improvement can be seen a the micro scale, travel time impedance. The high performance of this FSA is in line with our expectations. Housing prices are indeed predominantly influenced by local location features. Moreover, the travel-time impedance should indeed properly reflect how households consider local location features, since these frequently use the travel method of consideration.

Although, differences in performance can be found for varying property types, figure 12 reaffirms that the use of FSAs could improve model accuracy. For any impedance and property type, the micro scale proves to be the best overall measure. Regarding the impedance, travel time seems to do be the best measure of accessibility. However, this impedance is prone to errors if the travel profile does not accurately reflect actual users.

Heterogeneity in zoning methods outcomes

Although the averages of the previous subsection tell us which FSAs perform best overall, they do not grand insights of their true potential as values might cancel each other out. To understand how our FSAs compare to one another at specific scales and socio-economic variables we have to zoom in. By plotting the regression results per property type, socio-economic variable and scale, model performance of individual regressions are visualized. For clarity, only variables at their best performing scale are considered. This leaves us with 256 regression outcomes.

The individual adjusted R^2s are presented in figure 13. In here shapes indicate the impedance type used, while color represent the best performing scale. Since no specific scales are applicable for SAUs, these are grey coloured (not applicable). When comparing the performance of SAUs (squares) to our FSAs (other shapes) it can be reconfirmed that the use of FSAs, indeed has the potential to improve model results. In 48 out of the 64 cases (75%), at least one of the FSA methods outperforms the SAU method. For hotels this is even 14 out of 16 (87.5%). Moreover, the differences in adjusted R^2 are significantly larger when FSAs outperform SAUs than for the other way around. Especially for hotels and offices this outperformance can be relatively large. For example, using employment in the financial sector as independent variable, predictability of office prices increases from around 67.5% with the SAU measure, up to 74.5% for the travel distance measure. This is an increase from over 10% by merely considering the same variable in a different way.

When comparing figure 12 to 13, it can be noticed that differences in adjusted R^2 are more significant for the latter. This illustrates the importance of choosing a proper zoning system per socio-economic variable and property type. Especially the scale seems to be a major determiner

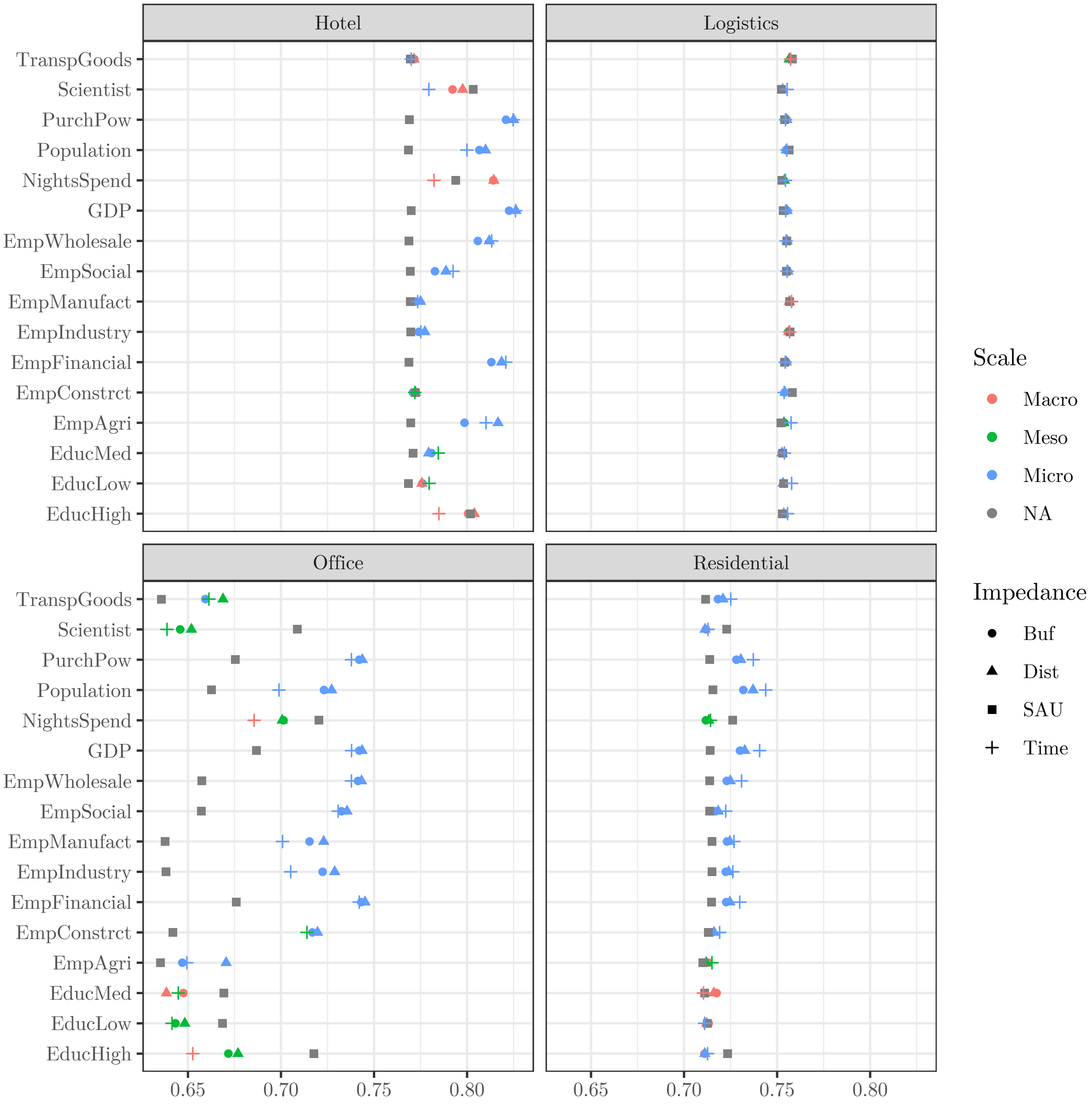


Figure 13: Adjusted R^2 comparison of simple hedonic models; by opportunity, asset type, impedance at the best performing scale

here. It should also be noted that use of FSAs does not improve adjusted R^2 unconditionally. Even if use of FSAs indeed reduces measurement error, this does not necessarily entail that models become more predictive. It could simply mean that these location values tell little about property prices. Indeed, the most significant improvements of model performance can be found for variables that are expected too have a large influence on prices. Hotels for example are not likely to be affected by the employment in manufacturing, but do retain many customers from the financial sector. As such, this location measure could explain more about these properties, and thus an improvement in the measure has a more significant effect. It could even be the case that a more accurate measure of opportunities leads to lower adjusted R^2, as these variables might no longer accidentally serve as instrumental variables for other indicators.

Despite an overall increase of adjusted R^2, figure 13 also indicates exceptions. For certain opportunities model fits of SAU values clearly surpass those of FSAs. Considering offices, it can be seen that the predictive power of the number of scientists measured by FSA stay far behind those measured by SAUs. The same goes for nights spend in hotels and the three education levels. Although the under-performance of FSAs compared to SAUs seems to be less severe for the other three property types, they remain present for certain variables. Moreover, there are cases in which some of our FSA perform better than the traditional SAUs whereas others perform worse, such as can be seen for the nights spend measure for hotel properties. In here the buffer and distance measure predict relatively well, whereas the travel time measure does not. Interestingly, any opportunity in which a SAU measure clearly outperforms one or more FSA measures, involve SAUs at the NUTS 2 level. This could indicate that the shapes and sizes of these NUTS 2 zones indeed better reflect user perception than our FSAs. Here a paradox could occur. As indicated earlier, NUTS are commonly used by both researchers and professionals. Therefore, it is very well possible that investment decisions are made upon these actual zones. If this happens on a large enough scale, these zones could potentially have a self-fulfilling prophecy, as both measure and determiner for property price. In such a case, NUTS regions are not the best measure of actual location features, but do reflect how investors perceive locations. This could potentially cause property prices to be over- or under valued, in regard to their actual location features. Another possibility is that our data disaggregation method fails to capture spatial trends at NUTS 2 level, and therefore limits data instead of improving it.

Consequently, even though FSAs have the potential to outperform traditional SAUs, they should be considered carefully. Differing situations might require varying FSAs even though the property type or socio-economic phenomena remains constant. Accordingly, even though FSAs can improve hedonic models compared to traditional SAUs, there is no guarantee they will automatically do so. For optimal results zones should be considered per property type and variables of consideration. Nevertheless, we can draw general guidelines from these results. For hotels and residential properties, travel time at the micro scale offers overall the best model fit, whereas this is travel distance at the micro scale for offices. Traditional SAUs remain the best option for logistics. Nevertheless, it is stressed that scales and travel method remain parameters which require researcher input. In our case study micro scales were set at 10 kilometres/minutes and travel methods were based on a driving-car profile. However, other cases might use walking profiles, and consider 10 kilometres to be a macro scale. Using different parameters changes zones and could therefore alter outcomes. In this regard, our supposed method does not eliminate the MAUP's zoning effect but merely provides researchers with the tools to control it. This allows researchers to link socio-economic variables at specific distances or travel times to properties. For situations in which not optimal predictability but comparability is leading, use of specific zones might therefore still be favourable over others.

Optimal results from functional statistical areas

To optimally use all information available in our FSAs, this subsection allows for interaction effects between single dependent variables at different scales. In contrast to the previous subsection, values at larger scales are no longer considered as the whole of their intrinsic scales, but as single bands that can be hollow inside. By including all three considered scales in a single model and allowing these to interact with each other, we can consider local values in relation to their wider area. Since scales are combined and no longer considered separate for every model, this leaves us with 192 new regression results. Figure 14 presents the Adjusted R^2 of our interaction model. It follows the same structure as figure 13, without the scale dimension. Previous SAU values are included for comparison.

It can be seen that the inclusion of all three scales and allowing for interaction improves prediction precision even further. Adjusted R^2s considerably increase in general and can go up to 18.18% for specific variables (employment industry in offices). Moreover, the improvements prove to be more robust. For the vast majority of socio-economic variables, interaction models outperform the SAU measure significantly. Only at offices, the number of scientists and education variables at the travel time scale stay notably behind. Especially for logistic properties, the interaction models show significant differences from earlier results. Whereas simple models showed little differences in prediction results between socio-economic indicators, this is no longer the case. It can be clearly seen that purchase power, GDP and employment in wholesale and finance improve regression results. It is remarkable that the volume of transported goods does not prove to be a determining factor. This might be caused by specific nodes that transport vast amounts by rail and are therefore not properly reflected in our zoning system. Moreover, we do not account for the supply of properties. It might be very well possible that the number of logistic facilities is highly correlated to the number of transported goods. In such a case, an increase in transported goods also increases logistic property supply and thus causes prices to remain equal.

A final overview of how FSAs compare to SAUs, is given in table 8. In here we consider the percentile improvements per impedance and property type. It can be concluded that overall the use of FSAs indeed improves our models and therefore is likely to reduce measurement errors. When scale interactions are considered, the travel distance measure performs best for Hotels, Logistics and Offices. For residential properties the travel time impedance shows best results. The difference is likely to be explained by variations in perceptions of accessibility for considered user types. The travel time measure by person cars seems to be adequate for residential users, but might not suffice for other user types. Moreover, travel time zones might not be large enough to capture macro level indicators for offices, logistics and hotels.

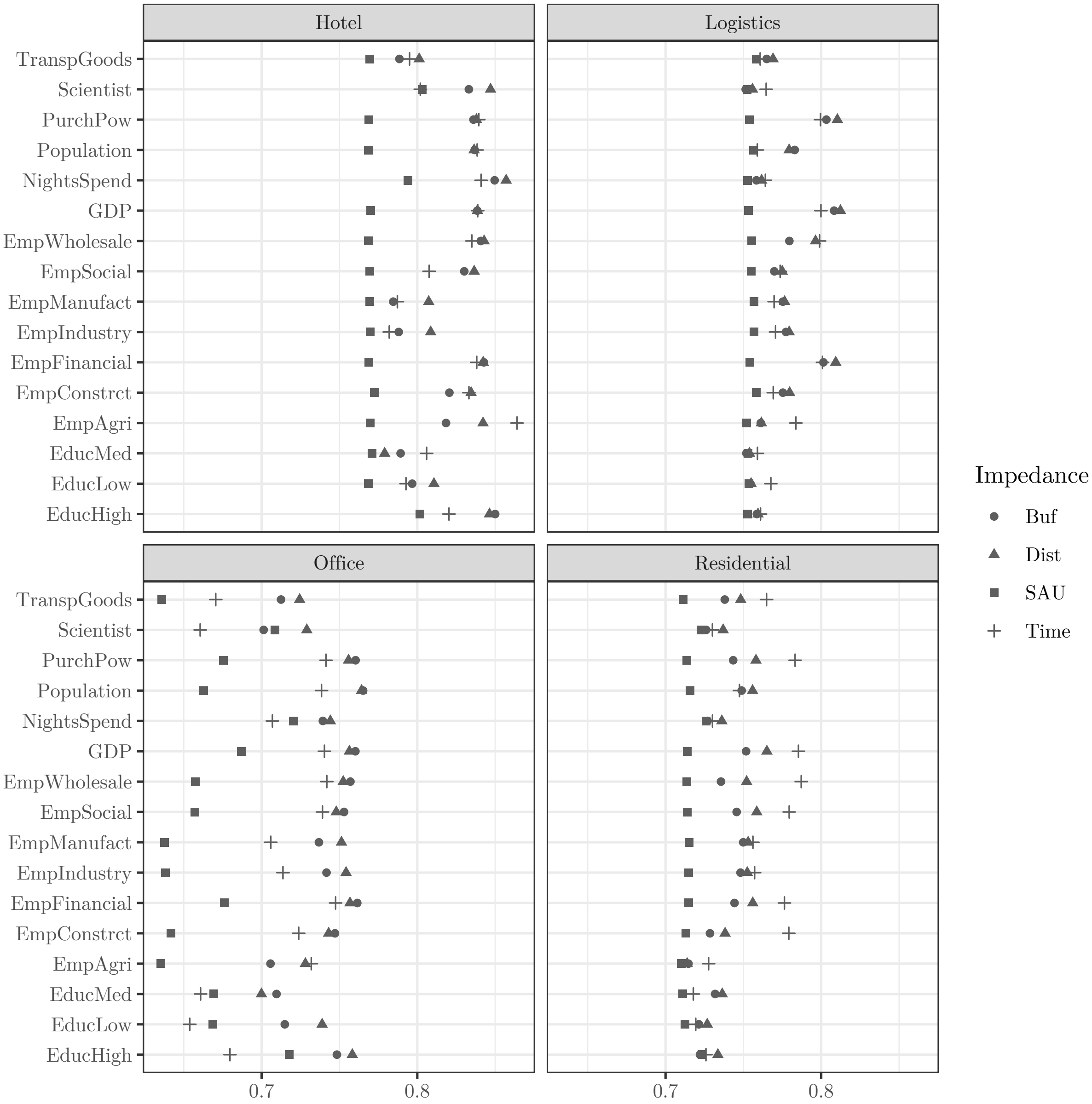


Figure 14: Adjusted R^2 comparison of hedonic models with interactive independent variables; by opportunity, asset type and impedance

	Δ adj.R^2 (%)			Δ RMSE (%)		
	Buffer	Distance	Time	Buffer	Distance	Time
Hotel	5.98	6.99	5.81	-12.36	-14.27	-11.94
Logistics	2.49	2.98	2.72	-4.43	-5.25	-4.77
Office	10.70	11.55	6.46	-11.18	-12.12	-6.45
Residential	2.90	4.16	5.44	-4.07	-5.74	-7.51

Table 8: Model performance comparison of SAU and FSA

Conclusion

With this thesis, the usefulness of functional statistical areas has been investigated. For this purpose we examined why such functional statistical areas should be used; how these should be created; when they are useful; and how they compare to the traditional method in which location characteristics are assigned per SAU.

To do so we asked one main research question and four sub-questions. Before we answer our main research question, we will go through each sub-question individually.

Sub-question 1

In our first sub-question we asked what MAUP implications are and how these affect statistical areas. Although there are various ways in which the MAUP is addressed, it basically comes down to the following: As long as statical areas are prone to change, research outcomes are at risk of being inconsistent and cannot be compared to another. Changes of zoning systems occur across research but also within single zoning specifications. Therefore, MAUP effects not only occur across research but also within. Doing so it biasses results and hampers the comparability of research outcomes. Nevertheless, the magnitude of this bias can vary significantly. MAUP effects are highly related to the level of internal homogeneity of areal units. Therefore, we should ideally use small units that are equal in shape and size and capture similar values. This should limit the amount of smoothing that occurs. When spatial data is distributed at random, smoothing will be limited automatically. However, in practice spatial phenomena are more frequently clustered in groups. Thus, caution should be paid. Measuring overlapping groups with a single areal unit could cause significant smoothing, and thus bias results.

It is acknowledged that MAUP effects still occur for FSAs. In fact, the MAUP does not affect FSAs just once, but multiple times. This happens for SAUs which are used as initial input; pixels of our disaggregated data, which are basically very tine zones; and the created optimal zoning systems. Also within our empirical results we saw that outcomes of different statistical areas varied. As such, outcomes across research might still be difficult to compare if methodologies used are not similar. Nevertheless, the use of FSAs has key advantages over SAUs in terms of the MAUP. When we consider our results it can be seen that FSAs behave in a very constant and predictable way. In figure 10 we saw that distributions of FSAs are rather consistent across impedances and opportunities. Moreover, their shapes and sizes can be well explained because it is known what these FSAs measure. In figure 11 and 12, we see the same pattern. Although results vary per scale, they remain rather constant over impedances. Reason for this consistency is explained in figure 8. Although FSAs differ in shapes and sizes, in essence they remain a circle like shape that is placed around the feature of interest. Since all FSAs within a research are created by the same methodology, they all consider a similar spatial extend. Consequently, although the MAUP might be more frequently present in the creation of FSAs, its effects are reduced.

Sub-question 2

In our second sub-question it is asked which criteria should be considered for the creation of FSAs, and how these criteria affect shapes and sizes of zones.

Unfortunately no straight forward answer can be given to this question. Criteria for when, where and how to use FSAs are highly dependent upon the situation of interest. Nevertheless, a number of guidelines can be given based upon our non-empirical research and case study.

As argued in our introduction, FSAs should solve a number of issues that typically occur when spatial characteristics are directly associated to point features by the SAU in which these are located. While doing so, FSAs should remain easy to use and be able to utilize spatial data in formats at which it is commonly available. Although any SAU can be transformed into an FSA, there is a number of decisions one should consider carefully. Ideally low level spatial data should be used as initial input. The advantage of this low level data is that it has limited smoothing effects. Although it is relatively easy to aggregate data to a lower level, disaggregating data and recovering lost variance is not. As such, preventing detail loss is better than curing it. Unfortunately, low level data is frequently unavailable. In this case the amount of data loss is dependent upon the quality of initial SAUs. If their shapes and sizes are setup in such a way that they correctly capture homogenous values, loss of variance should not be a major concern. Such SAUs correspond to 2c. Also when the spatial distribution of the phenomena of consideration is random, loss of variation remains limited. In this case the sizes and shapes of initial SAUs are of lesser influence. Such SAUs correspond to 2a and 2b.

Regarding the shapes and sizes of zones we recommend the use of optimal zoning systems. OZS should have a meaningful relation to the feature of consideration. Therefore they can vary significantly, dependent upon the phenomena and spatial characteristics of interest. A common solution to catpure this meaningful relationship is by means of accessibility measures. These measures can take a number of forms and differ in impedances and types dependent upon the opportunity of consideration.

Similar to SAUs, OZS are affected by the MAUP. Therefore they should meet similar criteria and capture homogenous areas with little variance. In practice, this often means that OZS should be small enough to capture location specific variance, while being large enough to measure spatial relationships. Since spatial distributions vary per characteristic of interest and geographical location, careful consideration of optimal scales should be made.

Also, the quality of SAU inputs and the accuracy of reaggregation outcomes could affect spatial distributions. In our case study we used NUTS regions as SAUs. These are created for reporting purposes. As such, it was expected that these correctly capture homogenous trends and correspond to the zones of figure 2c. Nevertheless, clear differences were found in the performance of NUTS 2 regions and NUTS 3 regions. In all our results it can be noted that variables measured at the NUTS 2 regions behave differently than when measured as NUTS 3 regions. Whereas FSAs correctly captured local extends for NUTS 3 regions, this was not the case for NUTS 2 regions. Results indicate that larger OZS are more optimal in scenarios where the spatial distribution of data does not reflect real life patterns.

A general guideline would be as follows. If the resolution of data becomes higher, OZS should become smaller. If resolution of data becomes lower, OZS should become larger.

Sub-question 3

For our third sub-question it was asked how SAU values can be disaggregated to lower level data, and reaggregated to our newly created zones. In our literature review we discussed some of the major methods to do so (table 1). Although all these methods can be classified as areal interpolation, large differences exist in the approaches these methods use. Techniques can be classified by two major characteristics. Is it a one-step process or a two-step process?; and does the interpolation technique use auxiliary variables or not?

Which areal interpolation technique should used most optimally is dependent upon the situation. A general rule of thumb is that more accurate results also require more sophisticated modelling. Use of geostatistical methods that require auxiliary data could significantly improve results. However, it comes at the costs of additional data requirements and might require more computational power. Deciding if possible benefits of obtaining more accurate results outweigh the costs of this additional effort, is up to the researcher.

Also the decision to go for a one-step process, or a two-step process is situation dependent. An advantage of the two-step process is that it provides a continuous surface for the entire area of consideration. This could be useful for evaluation purposes or when vast amounts of overlapping OZSs should be reaggregated. Nevertheless, it comes at the cost of a slight decrease in precision (dependent upon the pixel size). Moreover, if large parts of this continuous surface are not used for reaggregation and only limited OZS need to be estimated, it might be more efficient to only interpolate data for these zones.

It is important to note that areal interpolation has the ability to estimate distributions of values within zones. Doing so it can provide more accurate values at specific locations. However, it cannot recover lost variances. In terms of our initial example (figure 1), this means that it can estimate where location features are. However, it cannot split mean values and rates into its intrinsic components. For this reason areal interpolation should ideally use absolute data only. If rates and means are desired, these should be recalculated with outputs of absolute data.

Sub-question 4

At last we ask ourself how well FSAs measure location characteristics, and how this compares to SAUs. Given the nature of socio-economic data it is hard to quantify how well FSAs measure location characteristics exactly. Reason for this is that no comparison can be made to the "true" result (Openshaw and Alvanides 2001). Nevertheless, our results provide an approximation. Figure 10 indicated that data distributions of FSAs are in line with how we expect spatial characteristics to behave. Distributions are more continuous, and micro level values of characteristics that are typically urban show great overlap with macro level values.

When we compare model fit of FSA measures to SAU measures, we can see that FSAs have the potential to considerably improve results. Nevertheless, results can vary significantly per impedance and scale. It was illustrated by figure 13 that micro scales overall perform best, but that this boost in performance is not unconditional. Therefore, choosing the right OZS is of high importance. Figure 14 and table 8 indicate that when scales are allowed to interact, performance increases become more robust and significant. As such, allowing scales to interact provides a degree of certainty. However, this comes at the cost of more sophisticated modelling. Despite these limitations, results indicate that overall FSAs notably boost model performance over SAUs.

Main research question

We conclude that the use of functional statistical areas could offer researchers a number of benefits. It offers a flexible method that is compatible with any kind of statical zone, and could be used to study a variety of spatial phenomena. Although, this thesis considered the use of FSAs in combination with a real estate dataset, any situation in which areal zoning effects on point data are measured could be considered. Moreover, the level of sophistication is up to the user, which makes the use of FSAs widely applicable.

The following benefits of FSAs are identified. FSAs provide researchers more data control. Since zones are no longer considered upon the statistical boundaries by which data is obtained, research outcomes are less dependent upon the shapes and sizes of these statical zones. Consequently, changes in zones or improper statical areas that fail to capture the spatial extend of the phenomena of interests, will cause less bias to research outcomes. This makes results more robust, both within and between research. Moreover, since zones of interests have to be set, this method forces researchers to think about the spatial extend of their data and implications that come with it. This could be considered as negative in terms of usability. However, it incorporates a vital requirement of statics, namely that sample data should reflect the phenomena of interest properly. The specification on why certain zoning systems are chosen over others could serve as valuable input for future research, but might just as well provide additional insights to the research at hand. With the existence of the FSA method, data availability should no longer suffice as explanation for using a SAU as location measure for point data.

Moreover, FSAs take actual locations as base and create areas of interest around it. Because individual zones are created for separate locations, unique location values are obtained for each point of interest. This provides data that take a more continuous form, rather than the factor like values obtained from SAUs. Since FSAs are not limited by political boundaries, locations are not treated as stand alone zones that have no interaction with surrounding areas. Through inclusion of multiple scales, effects at specific ranges can be easily modelled. Accordingly, FSAs provide a location measure that simulate how property users actually consider locations.

Since FSAs use already available data in a more clever way it provides an accessible solution to the MAUP, when areal effects on point data are considered. Results indicate that use of FSAs could provide more accurate and precise models when compared to traditional SAUs. Nevertheless, FSAs provide no absolute solution to the MAUP. Although, FSAs create statistical zones that better reflect the spatial extend of points of consideration, these zones can still vary in shapes and sizes and should thus be considered carefully.

Moreover, no performance comparison has been made between the use of FSAs and spatial econometric models. Although both allow for inclusion of wider areas into a single statistical model, they do so in a different way. Whereas the firmer focusses on data manipulation, the second uses advanced statistical methods for this purpose. Advantages of FSAs over spatial econometric models are that by means of disaggregation, spatial phenomena cannot only be considered at a wider level but also at a smaller scale. Moreover, since FSAs actually create new datasets it allows for clear visualisation of the spatial extend being measured. Disadvantages are that the creation of FSAs can have a certain threshold. Interpolation methods need to be used, and OZS should be defined. This requires adequate choices that can differ dependent upon the context. Moreover, the creation of OZS and interpolation process could be computational expensive.

Although the use of FSAs and spatial econometric might be considered as counterparts, these are not mutually exclusive. The modified data can still be modelled in advanced ways that could provide the best of both worlds. Future research should indicate if combining these methods could indeed further advance spatial statistical models.

Discussion

Although our case study has indicated that the use of FSAs could indeed reduce measurement errors and provide more accurate and precise models, it remains difficult to identify the unique contribution of individual steps. For example, it could be possible that the interpolation of zones contributed little to total outcomes, and performance is only increased due to our OZS. If this is indeed the case, our developed methodology could be simplified while still obtaining equal results. Moreover, it is acknowledged that the amount by which the use of FSAs improve model performance is dependent upon the explanatory power of that dependent variable. Improvements of a measure that does not affect model results, cannot be measured. E.g. if variable x is measured three times as accurately, but this variable has no relation to the independent variable, no improvements in outcomes can be identified. Furthermore, adjusted R^2s and RMSEs improvements are dependent upon the initial values provided by SAUs. An increase of adj. R^2 from 0.1 to 0.2 is higher in terms of percentages than an increase of 0.8 to 0.9. However, the latter is more difficult to obtain. Accordingly, the true power of FSAs remains to be identified.

Regarding our data, there is a number of limitations of which assumptions were made to conduct this research. As discussed, the use of valuations as dependent variable has been topic of debate. According to some, valuations do not reflect actual market prices and thus biasses research outcomes. Moreover, it is recognized that the number of observations for certain properties are limited. Using complex models on limited observations could lead to over-fitting. Nevertheless, there are no indications that this is the case for this research. Results of individual property classes are in line with other outcomes. It is also acknowledged that the classification of property types could be reason for debate. These classifications have been made per building, by the user types that paid the majority of rents. As such, properties could be wrongly categorized due to changes in rents, vacancies or having mixed functions.

Concerning our socio-economic data, bias could be caused due to missing values. Nevertheless it is expected that this will not heavily affect results. SAUs with missing data only affected a small amount of properties in a limited way. Moreover, new values were estimated in a sensible way. Although omission of properties that are located in proximity to regions with missing values would have been possible as well, this would further reduce our CRE data. Moreover, the estimated values were equal for all zoning measures. Therefore, the used methodology provides an practical example how FSAs could introduce difficulties, but also provide solutions for missing values.

Regarding our hedonic model it could be possible that the omission of relevant independent variables biasses outcomes. It is recognized that a single socio-economic variable cannot explain all locational characteristics. However, inclusion of multiple socio-economic indicators within single models caused concern for multicollinearity. Although this should not affect predictability of models as a whole, it could affect coefficients of individual parameters. Moreover, inclusion of to many explanatory variables might lead over-fitting. Since no cross validation has been applied it was chosen to keep our model relatively simple and consider socio-economics individually. Nevertheless it would have been of interest to see how more complex models, with all considered socio-economic indicators, compare in terms of fit when using SAUs and FSAs. Because of the discussed implications, we leave this to future research.

Also other important property characteristics might have been omitted. Nevertheless, this should not be of major concern. As argued it was not aimed to create an optimal model. Price estimations merely served as a mean to test model-fit of our FSAs.

Future research

Although the case study of this thesis considered a number of opportunities for a varying set of property types, at three different scales and impedances, it is recognized that a lot remains to be explored. As such, the developed methodology of this thesis could guide as a starting point for future research.

Research suggestions are recommended among three pillars; Interpolation, Zoning Systems and Opportunities.

For interpolation this thesis considered pycnophylactic interpolation. However, a variety of alternatives is available. Advanced interpolation options such as spatio-temporal kriging, and methods which include auxiliary variables (e.g. satellite images, property data and districts) could guide the interpolation process and provide more optimal results. Ideally, researchers should strive for the creation of continuous surfaces that adequately reflect spatial phenomena over the time-space spectrum.

Regarding the zoning systems other impedances and scales could be explored. The costs method could be a promising option here. However, other types of measures could be considered as well. Also different kind of travel networks, such as public transport or even flight paths, could be explored. Moreover, different user types could be considered for routing options. Examples would be walking-, cycling- and heavy goods vehicle profiles.

Regarding opportunities, additional socio-economic variables could be considered. In here extra attention could be paid to relative values, and how these perform after interpolation. Furthermore, non socio-economic variables could be considered as opportunities. The only condition here is that the phenomena of interest contains a spatial dimension and can be measured as zone. Examples would be amenities such as services, shops and green spaces.

As argued earlier, the methodology could also be applied to other scenarios than commercial real estate. The condition here is that variables can be measured as points and that they are affected by locations. Possible research could be on how location features impact business performance, city development or even personal development.

At last, repetitive research that explorers the performance of FSAs at varying regions and with different data could help to establish the robustness of this methodology.

Literature

Alfieri, Lorenzo et al. (2016). “Modelling the socio-economic impact of river floods in Europe”. In: *Natural Hazards and Earth System Sciences* 16.6, pp. 1401–1411.

Anderson, William P (2012). *Economic geography*. 1st ed. London: Routledge. ISBN: 9780203114988. DOI: https://doi.org/10.4324/9780203114988.

Arbia, Guiseppe (1989). “The Accuracy Of Spatial Databases”. In: ed. by Michael F. Goodchild and Sucharita Gopal. 1st ed. Vol. 1. London: Taylor & Francis. Chap. Statistical effects of spatial data transformations a proposed general framework, pp. 249–259. ISBN: 9780429208317.

ArcGIS (n.d.). *What is areal interpolation*. Online. Accessed[21-10-2021]. URL: https://pro.arcgis.com/en/pro-app/latest/help/analysis/geostatistical-analyst/what-is-areal-interpolation.htm.

Bian, Ling and Rachel Butler (1999). “Comparing effects of aggregation methods on statistical and spatial properties of simulated spatial data”. In: *Photogrammetric engineering and remote sensing* 65, pp. 73–84.

Billings, Stephen (2011). “Estimating the value of a new transit option”. In: *Regional Science and Urban Economics* 41.6, pp. 525–536. URL: https://EconPapers.repec.org/RePEc:eee:regeco:v:41:y:2011:i:6:p:525-536.

Bollinger, Christopher R, Keith R Ihlanfeldt, and David R Bowes (1998). “Spatial variation in office rents within the Atlanta region”. In: *Urban Studies* 35.7, pp. 1097–1118.

Bourassa, Steven C., Eva Cantoni, and Martin Edward Ralph Hoesli (Aug. 2007). “Spatial Dependence, Housing Submarkets, and House Price Prediction”. English. In: *The Journal of Real Estate Finance and Economics* 35.2, pp. 143–160. ISSN: 0895-5638. DOI: 10.1007/s11146-007-9036-8.

Brown, Peter and Stephen Hincks (Sept. 2008). “A Framework for Housing Market Area Delineation Principles and Application”. In: *Urban Studies* 45, pp. 2225–2247. DOI: 10.1177/0042098008095866.

Brunsdon, Chris (2014). *Pycnophylactic Interpolation*. Online. [Accessed: 04-11-2021]. URL: https://cran.r-project.org/web/packages/pycno/pycno.pdf#page=4&zoom=100,133,508.

Cameron, A Colin and Frank AG Windmeijer (1997). “An R-squared measure of goodness of fit for some common nonlinear regression models”. In: *Journal of econometrics* 77.2, pp. 329–342.

Can, Ayşe (1998). “GIS and Spatial Analysis of Housing and Mortgage Markets”. In: *Journal of Housing Research* 9.1, pp. 61–86. ISSN: 10527001. URL: http://www.jstor.org/stable/24833659.

Cervero, Robert and Michael Duncan (2002). “Transit’s Value-Added Effects: Light and Commuter Rail Services and Commercial Land Values”. In: *Transportation Research Record* 1805.1, pp. 8–15. DOI: 10.3141/1805-02. eprint: https://doi.org/10.3141/1805-02. URL: https://doi.org/10.3141/1805-02.

Charlton, M. (2009). “Quantitative Data”. In: *International Encyclopedia of Human Geography*. Ed. by Rob Kitchin and Nigel Thrift. Oxford: Elsevier, pp. 19–26. ISBN: 978-0-08-044910-4. DOI: `https://doi.org/10.1016/B978-008044910-4.00502-2`. URL: `https://www.sciencedirect.com/science/article/pii/B9780080449104005022`.

Chica-Olmo, Jorge, Rafael Cano-Guervos, and Mario Chica-Rivas (2019). “Estimation of housing price variations using spatio-temporal data”. In: *Sustainability* 11.6, p. 1551.

Comber, Alexis and Wen Zeng (June 2019). “Spatial interpolation using areal features A review of methods and opportunities using new forms of data with coded illustrations”. In: *Geography Compass* 13. DOI: `10.1111/gec3.12465`.

Daams, Michiel N, Frans J Sijtsma, and Arno J van der Vlist (2016). “The effect of natural space on nearby property prices: accounting for perceived attractiveness”. In: *Land Economics* 92.3, pp. 389–410.

Derdouri, Ahmed and Yuji Murayama (2020). “A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan”. In: *Journal of Geographical Sciences* 30, pp. 794–822.

Do, Van Huyen, Thibault Laurent, and Anne Vanhems (2021). “Guidelines on Areal Interpolation Methods”. In: *Advances in Contemporary Statistics and Econometrics*. Springer, pp. 385–407.

Dröes, Martijn and Alex van de Minne (2015). “Time-Varying Determinants of Long-Run House Prices”. MA thesis. University of Amsterdam.

Ekeland, Ivar, James J Heckman, and Lars Nesheim (2002). “Identifying hedonic models”. In: *American Economic Review* 92.2, pp. 304–309.

epsg.io (2019). *EPSG:3035*. Online. Accessed[16-12-2021]. URL: `https://epsg.io/3035`.

Eurostat (2015). *Regions in the European Union — Nomenclature of territorial units for statistics — NUTS 2016/EU-28*. Tech. rep. DOI: `10.2785/53498`.

Evans, Alan W (2008). *Economics, real estate and the supply of land*. John Wiley & Sons.

Farber, Steven and Maurice Yeates (2006). “A comparison of localized regression models in a hedonic house price context”. In: *Canadian Journal of Regional Science* 29.3, pp. 405–420.

Flowerdew, Robin, Mick Green, and Evangelos Kehris (1991). “Using areal interpolation methods in geographic information systems”. In: *Papers in regional science* 70.3, pp. 303–315.

Forys, Iwona and Malgorzata Tarczynska-Luniewska (2018). “Multivariate analysis of the condition of the property development sector: selected local real estate markets in the European Union”. In: *International Advances in Economic Research* 24.1, pp. 1–15.

Fuerst, Franz and Michel Ferreira Cardia Haddad (2020). “Real estate data to analyse the relationship between property prices, sustainability levels and socio-economic indicators”. In: *Data in Brief* 33, p. 106359. ISSN: 2352-3409. DOI: `https://doi.org/10.1016/j.dib.2020.106359`. URL: `https://www.sciencedirect.com/science/article/pii/S235234092031252X`.

Gelfand, Alan E et al. (2004). “The dynamics of location in home price”. In: *The journal of real estate finance and economics* 29.2, pp. 149–166.

GISCO (n.d.). *Administrative units / statistical units*. Online. Accessed[31-01-2022]. URL: https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units.

Goodman, John L. and John B. Ittner (1992). "The accuracy of home owners' estimates of house value". In: *Journal of Housing Economics* 2.4, pp. 339–357. ISSN: 1051-1377. DOI: {https://doi.org/10.1016/1051-1377(92)90008-E}. URL: %7Bhttps://www.sciencedirect.com/science/article/pii/105113779290008E%7D.

Gotway, Carol A and Linda J Young (2002). "Combining Incompatible Spatial Data". In: *Journal of the American Statistical Association* 97.458, pp. 632–648. DOI: 10.1198/016214502760047140. eprint: https://doi.org/10.1198/016214502760047140. URL: https://doi.org/10.1198/016214502760047140.

Grimes, Arthur and Andrew Aitken (2007). "House Prices and Rents: socio-economic impacts and prospects". In:

Hawley, Kevin and Harold Moellering (2005). "A Comparative Analysis of Areal Interpolation Methods". In: *Cartography and Geographic Information Science* 32.4, pp. 411–423. DOI: 10.1559/152304005775194818. eprint: https://doi.org/10.1559/152304005775194818. URL: https://doi.org/10.1559/152304005775194818.

Helbich, Marco et al. (2014). "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria". In: *Urban Studies* 51.2, pp. 390–411. DOI: 10.1177/0042098013492234. eprint: https://doi.org/10.1177/0042098013492234. URL: https://doi.org/10.1177/0042098013492234.

Herath, Shanaka and Gunther Maier (2010). "The hedonic price method in real estate and housing market research a review of the literature". In:

Heyman, Axel Viktor, Stephen Law, and Meta Berghauser Pont (2019). "How is Location Measured in Housing Valuation? A Systematic Review of Accessibility Specifications in Hedonic Price Models". In: *Urban Science* 3.1. ISSN: 2413-8851. DOI: 10.3390/urbansci3010003. URL: https://www.mdpi.com/2413-8851/3/1/3.

Hui, Iris and Wendy K. Tam Cho (2018). "3.13 - Spatial Dimensions of American Politics". In: *Comprehensive Geographic Information Systems*. Ed. by Bo Huang. Oxford: Elsevier, pp. 181–188. ISBN: 978-0-12-804793-4. DOI: https://doi.org/10.1016/B978-0-12-409548-9.09665-2. URL: https://www.sciencedirect.com/science/article/pii/B9780124095489096652.

Janoušková, Jana and Šárka Sobotovičová (2021). "Approaches to Real Estate Taxation in the Czech Republic and the EU Countries". In: *International Advances in Economic Research* 27.1, pp. 61–73.

Jones, Colin (2002). "The Definition of Housing Market Areas and Strategic Planning". In: *Urban Studies* 39.3, pp. 549–564. DOI: 10.1080/00420980220112829. eprint: https://doi.org/10.1080/00420980220112829.

Jones, Colin, Mike Coombes, and Cecilia Wong (2012). “A System of National Tiered Housing-Market Areas and Spatial Planning”. In: *Environment and Planning B Planning and Design* 39.3, pp. 518–532. DOI: 10.1068/b37172. eprint: https://doi.org/10.1068/b37172.

Jones, Colin and Craig Watkins (2009). *Housing markets and planning policy.* Vol. 40. John Wiley & Sons.

Kim, Hwahwan and Xiaobai Yao (2010). “Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method”. In: *International Journal of Remote Sensing* 31.21, pp. 5657–5671. DOI: 10.1080/01431161.2010.496805. eprint: https://doi.org/10.1080/01431161.2010.496805. URL: https://doi.org/10.1080/01431161.2010.496805.

Krivoruchko, Konstantin, Alexander Gribov, and Eric Krause (2011). “Multivariate Areal Interpolation for Continuous and Count Data”. In: *Procedia Environmental Sciences* 3. 1st Conference on Spatial Statistics 2011 – Mapping Global Change, pp. 14–19. ISSN: 1878-0296. DOI: https://doi.org/10.1016/j.proenv.2011.02.004. URL: https://www.sciencedirect.com/science/article/pii/S1878029611000053.

Kruse, Rudolf (1987). “On the variance of random sets”. In: *Journal of mathematical analysis and applications* 122.2, pp. 469–473.

Kuntz, Michael and Marco Helbich (2014). “Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging”. In: *International Journal of Geographical Information Science* 28.9, pp. 1904–1921.

Lee, Gunhak, Daeheon Cho, and Kamyoung Kim (2016). “The modifiable areal unit problem in hedonic house-price models”. In: *Urban Geography* 37.2, pp. 223–245. DOI: 10.1080/02723638.2015.1057397. eprint: https://doi.org/10.1080/02723638.2015.1057397. URL: https://doi.org/10.1080/02723638.2015.1057397.

Leeuwen, Niek van and Jeroen Venema (2021). *Statistische gegevens per vierkant en postcode 2020-2019-2018.* Tech. rep. CBS.

Liu, Xiaolong (2012). “Spatial and temporal dependence in house price prediction”. In: *The Journal of Real Estate Finance and Economics* 47.2, pp. 341–369.

Malpezzi, Stephen (2002). “Hedonic Pricing Models: A Selective and Applied Review”. In: *Housing Economics and Public Policy.* John Wiley & Sons, Ltd. Chap. 5, pp. 67–89. ISBN: 9780470690680. DOI: https://doi.org/10.1002/9780470690680.ch5. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470690680.ch5. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470690680.ch5.

MapTiler (n.d.). *EPSG:3035.* Online. [accessed: 26-10-2021]. URL: https://epsg.io/3035.

Metzner, Steffen and Andreas Kindt (2018). “Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties: A literature-based approach”. In: *International Journal of Housing Markets and Analysis.*

Nouvel, Romain et al. (2015). "Combining GIS-based statistical and engineering urban heat consumption models: Towards a new framework for multi-scale policy support". In: *Energy and Buildings* 107, pp. 204–212.

Oleś, Andrzej (2021). *Query openrouteservice from R*. Online. Accessed[16-12-2021]. URL: `vignettes/openrouteservice.Rmd`.

Openrouteservice (n.d.). *Travel speeds*. Online. Accessed[17-12-2021]. URL: `https://giscience.github.io/openrouteservice/documentation/travel-speeds/Travel-Speeds.html`.

Openshaw, Stan (1977). "Optimal zoning systems for spatial interaction models". In: *Environment and Planning A* 9.2, pp. 169–184.

Openshaw, Stan and Seraphim Alvanides (2001). "Designing zoning systems for representation of socio-economic data". In: *Time and Motion of Socio-Economic Units. Taylor and Francis, London.*

Osland, Liv (July 2010). "An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling". In: *Journal of Real Estate Research* 32, pp. 289–320.

Otto, Suzie J et al. (2003). "Initiation of population-based mammography screening in Dutch municipalities and effect on breast-cancer mortality: a systematic review". In: *The Lancet* 361.9367, pp. 1411–1417.

Pai, Ping-Feng and Wen-Chang Wang (2020). "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices". In: *Applied Sciences* 10.17. ISSN: 2076-3417. DOI: `10.3390/app10175832`. URL: `https://www.mdpi.com/2076-3417/10/17/5832`.

Palmquist, Raymond B. (2005). "Chapter 16 Property Value Models". In: *Valuing Environmental Changes*. Ed. by Karl-Gran Mler and Jeffrey R. Vincent. Vol. 2. Handbook of Environmental Economics. Elsevier, pp. 763–819. DOI: `https://doi.org/10.1016/S1574-0099(05)02016-4`. URL: `https://www.sciencedirect.com/science/article/pii/S1574009905020164`.

Pebesma, Edzer (2018). "Simple Features for R: Standardized Support for Spatial Vector Data". In: *The R Journal* 10.1, pp. 439–446. DOI: `10.32614/RJ-2018-009`. URL: `https://doi.org/10.32614/RJ-2018-009`.

Roebeling, Peter et al. (2017). "Assessing the socio-economic impacts of green/blue space, urban residential and road infrastructure projects in the Confluence (Lyon) a hedonic pricing simulation approach". In: *Journal of Environmental Planning and Management* 60.3, pp. 482–499. DOI: `10.1080/09640568.2016.1162138`. eprint: `https://doi.org/10.1080/09640568.2016.1162138`. URL: `https://doi.org/10.1080/09640568.2016.1162138`.

Rosen, Sherwin (1974). "Hedonic prices and implicit markets: product differentiation in pure competition". In: *Journal of political economy* 82.1, pp. 34–55.

Royuela, Vicente and Miguel A. Vargas (2009). "Defining Housing Market Areas Using Commuting and Migration Algorithms Catalonia (Spain) as a Case Study". In: *Urban Studies* 46.11, pp. 2381–2398. DOI: `10.1177/0042098009342600`. eprint: `https://doi.org/10.1177/0042098009342600`.

Schulz, Rainer, Martin Wersing, and Axel Werwatz (2014). “Automated valuation modelling: a specification exercise”. In: *Journal of Property Research* 31.2, pp. 131–153. DOI: 10.1080/09599916.2013.846930. eprint: https://doi.org/10.1080/09599916.2013.846930. URL: https://doi.org/10.1080/09599916.2013.846930.

Seo, Kihwan et al. (2019). “Hedonic modeling of commercial property values distance decay from the links and nodes of rail and highway infrastructure”. In: *Transportation* 46.3, pp. 859–882.

Sirmans, Stacy, David Macpherson, and Emily Zietz (2005). “The composition of hedonic pricing models”. In: *Journal of real estate literature* 13.1, pp. 1–44.

Sui, Daniel Z (2004). “Tobler’s first law of geography: A big idea for a small world?” In: *Annals of the Association of American Geographers* 94.2, pp. 269–277.

Sundquist, Kristina et al. (2011). “Country of birth, socioeconomic factors, and risk factor control in patients with type 2 diabetes: a Swedish study from 25 primary health-care centres”. In: *Diabetes/metabolism research and reviews* 27.3, pp. 244–254.

Tabellini, Guido (June 2010). “Culture and Institutions: Economic Development in the Regions of Europe”. In: *Journal of the European Economic Association* 8.4, pp. 677–716. ISSN: 1542-4766. DOI: 10.1111/j.1542-4774.2010.tb00537.x. eprint: https://academic.oup.com/jeea/article-pdf/8/4/677/10313288/jeea0677.pdf. URL: https://doi.org/10.1111/j.1542-4774.2010.tb00537.x.

Tobler, Waldo R. (1970). “A Computer Movie Simulating Urban Growth in the Detroit Region”. In: *Economic Geography* 46, pp. 234–240. ISSN: 00130095, 19448287. URL: http://www.jstor.org/stable/143141.

— (1979). “Smooth Pycnophylactic Interpolation for Geographical Regions”. In: *Journal of the American Statistical Association* 74.367. PMID: 12310706, pp. 519–530. DOI: 10.1080/01621459.1979.10481647. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481647.

Tuson, Matthew et al. (2020). “Overcoming inefficiencies arising due to the impact of the modifiable areal unit problem on single-aggregation disease maps”. In: *International journal of health geographics* 19.1, pp. 1–18.

van Duijn, Mark, Jan Rouwendal, and Richard Boersema (2016). “Redevelopment of industrial heritage: Insights into external effects on house prices”. In: *Regional Science and Urban Economics* 57, pp. 91–107. ISSN: 0166-0462. DOI: https://doi.org/10.1016/j.regsciurbeco.2016.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0166046216000065.

Watts, Martin John (2009). “The impact of spatial imbalance and socioeconomic characteristics on average distance commuted in the Sydney metropolitan area”. In: *Urban Studies* 46.2, pp. 317–339.

Weinberger, Rachel R. (2001). “Light Rail Proximity: Benefit or Detriment in the Case of Santa Clara County, California?” In: *Transportation Research Record* 1747.1, pp. 104–113. DOI: 10.

3141/1747-13. eprint: https://doi.org/10.3141/1747-13. URL: https://doi.org/10.3141/1747-13.

Wong, David W.S. (2009). "Modifiable Areal Unit Problem". In: *International Encyclopedia of Human Geography*. Ed. by Rob Kitchin and Nigel Thrift. Oxford: Elsevier, pp. 169–174. ISBN: 978-0-08-044910-4. DOI: https://doi.org/10.1016/B978-008044910-4.00475-2. URL: https://www.sciencedirect.com/science/article/pii/B9780080449104004752.

Xiong, Yongzhu et al. (2020). "Spatial statistics and influencing factors of the COVID-19 epidemic at both prefecture and county levels in Hubei Province, China". In: *International journal of environmental research and public health* 17.11, p. 3903.

Yoo, Eun-Hye (Dec. 2017). "Geostatistical Approach to Spatial Data Transformation". In: ISBN: 9780124095489. DOI: {10.1016/B978-0-12-409548-9.09601-9}.

Yoo, Sanglim, Jungho Im., and John E. Wagner (2012). "Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY". In: *Landscape and Urban Planning* 107.3, pp. 293–306. ISSN: 0169-2046. DOI: https://doi.org/10.1016/j.landurbplan.2012.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0169204612001958.

Yu, Danlin, Yehua Dennis Wei, and Changshan Wu (2007). "Modeling Spatial Dimensions of Housing Prices in Milwaukee, WI". In: *Environment and Planning B Planning and Design* 34.6, pp. 1085–1102. DOI: 10.1068/b32119. eprint: https://doi.org/10.1068/b32119. URL: https://doi.org/10.1068/b32119.

Zhang, Caiyun and Fang Qiu (2011). "A Point-Based Intelligent Approach to Areal Interpolation". In: *The Professional Geographer* 63.2, pp. 262–276. DOI: 10.1080/00330124.2010.547792. eprint: https://doi.org/10.1080/00330124.2010.547792. URL: https://doi.org/10.1080/00330124.2010.547792.