The impact of spatial aggregation on urban development analyses

Chris Jacobs-Crisioni MSc[1,2] *
Prof. Dr. Piet Rietveld[2]
Dr. Eric Koomen[2]

* Corresponding author. Email: christiaan.jacobs@jrc.ec.europa.eu

[1] European Commission, Joint Research Centre,
Institute for Environment and Sustainability,
Land Management and Natural Hazards Unit,
Via E. Fermi, 2749,
21027 Ispra (VA), Italy

[2] Faculty of Economics and Business Administration,
VU University Amsterdam
Boelelaan 1105
1081HV Amsterdam

Abstract: *This paper illustrates the impacts of the scale and shape of spatially aggregated data on the analysis of urban development. It explores those impacts on the results of a variety of statistics, including multivariate explanatory analyses incorporating driving forces acting on different scale levels, and spatial econometric methods. In addition, this paper uses existing weighting methods to overcome aggregation effects that are due to uneven portions of consumable land in observed areal units. The analyses show that shape effects can be partially removed by the used weighting methods, and that even regularly latticed areal units need such weighting in practice. In the explanatory analyses, aggregating to coarser resolutions does not affect the order of magnitude of estimated coefficients when the aggregation process maintains sufficient variance within variables. We argue that small-sized areal units approximating the size of parcels are to be preferred in urban land consumption analyses, because such micro-data allows the exploration of highly local factors alongside higher scale linkages. Spatial econometric methods can ease the difficulties with replicating the spatial clustering of land uses and the worryingly low levels of explained variance that are characteristic of analyses of small-grained land-use data.*

Keywords: MAUP, spatial aggregation, spatial econometrics, urban development, land use

JEL classifications: C10, C43, 018, R14, R52

1 Introduction

Urban development or the increase in the consumption of land for urban purposes is caused by individual choices affecting individual parcels (Irwin et al., 2003). Although urban development usually concerns individual properties, urban development is analyzed in a variety of areal units that in many cases represent very different spatial delineations. If we pick some recent examples, we find that the determinants of urban land consumption have been quantified for fine-resolution rasters (Borzacchiello et al. 2010; Loonen and Koomen 2009; Verburg et al. 2004b), reciprocities between urbanization and infrastructure development have been uncovered at the municipality (Koopmans et al. 2012), borough (Levinson 2008), residential block and census tracts levels (Xie and Levinson 2010), and the factors that determine the location of urban sprawl have been analyzed at the level of individual parcels (Irwin and Bockstael 2002) and large zonal units (Irwin et al. 2006). The areal units used to analyze urban development thus vary in scale (or size, or resolution), which is a property that commonly follows from the decision to incorporate a given number of zones; and in shape, which is commonly the result of decisions to amalgamate particular predefined areal units in one observation. The usage of aggregated areal units in the study of essentially individual entities is either needed to study at a manageable level, or intrinsically part of the available data.

In recent years finer resolution data have become available for analysis, but for several reasons it is uncertain if better results may be obtained from them. First, previous studies (Irwin et al. 2006; Kok and Veldkamp 2001) have found that much of the variance in small-grained land-use data remains unexplained by statistical models that incorporate higher-scale linkages such as distance to city centres, and that the sizes of estimated parameters depend on areal unit size. Those findings demonstrate that it is unclear if fine resolution data are fit for studying the effects of high-scale linkages on land-use changes, such as urban development. Furthermore, also neighbourhood dependencies are scale dependent (Qi and Wu 1996) and particularly influential in fine resolution data (Overmars et al. 2003). This poses challenges in particular to land use modellers attempting to replicate clustered land-use patterns on fine resolutions alongside other explanatory variables; for an overview of methods to do so we refer to Verburg et al. (2004c). All in all, statistical micro data analyses pose additional challenges to the analyst because of the unclear relation with higher scale linkages, neighbourhood dependencies and low explained variances, while the outcomes of statistical analyses performed on spatially aggregated data depend on the scale and shape of the areal units that the data represent. To underpin the choice of areal units and better understand the influence that a particular spatial data configuration will have on analysis results, this study will address the effect of the scales and shapes of areal units on outcomes of urban development analyses.

The modifiable areal unit problem

The fact that statistical analysis outcomes depend on the areal units in which the analyzed data are aggregated is well known (Arbia 1989; Fotheringham and Wong 1991; Gehlke and Biehl 1934; Robinson 1950), and this phenomenon has been named the Modifiable Areal Unit Problem (MAUP) by Openshaw (1984). This dependency on the areal units of data needs to be carefully separated from the related `ecological fallacy’, which is the fact that population characteristics cannot be inferred to be the characteristics of members of that population. Gotway and Young (2002) consider the MAUP as a special case of ‘change of support problems' that come into existence when the spatial representation of variables differs from their true spatial characteristics. This is often the result of spatial data transformations such as aggregation, interpolation or multiscale modelling. Essentially, the MAUP comes into play when the outcomes of processes affecting particular areas are aggregated into different, often larger areal units. The influence of spatial aggregation can be divided into scale and shape or zonation effects, that respectively refer to “the variation in results obtained from the same statistical analysis at different levels of spatial resolution” and “different results arising from the regrouping of zones at a given scale” (Kwan and Weber 2008; p. 111). It is commonly associated with irregularly sized areal units such as census tracts or postcode areas, but it is just as persistent in regularly latticed data, in which the arbitrary and modifiable aspects of unit delineation commonly follow from technical specifications (e.g. sensor resolution, satellite trajectory) instead of zone design principles.

The influence of the MAUP on statistical outcomes has been studied repeatedly. Arbia (1989) argues that the influence of the MAUP on univariate properties such as averages, standard deviations and spatial autocorrelation can be reduced, so that distortions because of shape effects are minimized and the influence of scale effects is predictable. Arbia furthermore argues that the influence of the MAUP becomes more predictable if the analysed areal units are perfectly equivalent in terms of size, shape and neighbourhood structure. Amrhein and Reynolds (1997) add that results of the Getis statistic of spatial association change in a relatively stable manner when changing scale. Previous analyses have demonstrated that the influence of the MAUP on correlation coefficients and multivariate analyses may be worse. As Robinson (1950) and Openshaw (1984) show, when correlating the same variables in different sets of spatial units the resulting coefficients can even change in sign. Others conclude that the effects of the MAUP on multivariate analyses are “essentially unpredictable” (Fotheringham and Wong 1991; p. 1042). Amrhein (1995) and Briant et al (2010) nuance this and demonstrate that model specification has a larger influence on multivariate analysis than spatial data configuration.

Because of its possible impact on statistical outcomes, the MAUP has been the concern of many previous contributions, in which strategies to overcome its implications have been proposed. Some have suggested to use optimal zoning schemes (Openshaw 1984), but such an approach is inherently troublesome because any definition of what constitutes an optimal design is likely to be subjective. Others analyse data that describe the entities that constitute the process, for example by modelling land-use change on the parcel level (Chakir and Parent 2009), but unfortunately sufficient data is not always available for such an exercise. Furthermore, individual entities can be hard to define, for example when analysing densities (Fotheringham 1989). Another proposition is to abandon statistical methods that are affected by spatial aggregation and instead apply methods that are 'frame-independent' (Tobler 1989). For an overview of available methods we refer to Gotway and Young (2002). However, it seems like a last resort measure to discard the current methods, and all the work that has been put in their development. Therefore a number of studies have tried to find the sensitivity of commonly applied statistical tests for the MAUP (Amrhein 1995; Briant et al. 2010; Fotheringham and Wong 1991).

Aims

This paper adds to previous analyses of the sensitivity of results for the MAUP by describing the influence of spatial data configuration on the results of urban land consumption analysis. The overall aim is to provide recommendations on choosing aggregation levels of spatial data for urban land consumption analyses, and methods to cope with the influence of aggregation when areal unit choices are not available. With regard to scale effects, emphasis is put on the interplay between estimated impacts, spatial autocorrelation and explained variance. Similar to Briant et al. (2010), we attempt to reduce scale effects by including explanatory variables that address driving forces at different scale levels. With regard to shape effects, emphasis is put on weighting methods that may be used to counterbalance those differences in consumable land supply that are inherent to irregular and also regularly sized areal units. We use fine resolution data of residential land-use shares in the Netherlands to compare the univariate properties of residential land-use shares, and results of bivariate correlations and multivariate regressions. To assess scale and shape effects those dependent data are averaged into varying sizes and shapes. This choice to aggregate by averaging is not trifle. Data properties such as averages and standard deviations are known to be less sensitive to the MAUP in case of aggregation by averaging; furthermore, the MAUP is less substantial in cases where all data are aggregated similarly (Arbia 1989; Briant et al. 2010). To understand how spatial data configuration affects analysis results, it is important to understand how spatial autocorrelation in the data is affected by scale effects in particular. This will therefore be the subject of the next section.

2 Spatial autocorrelation and areal unit size

According to Legendre, spatial autocorrelation “may be loosely defined as the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observation” (Legendre 1993; p. 1659). It is symptomatic of non-random spatial distributions of observed phenomena and thus poses conceptual and technical challenges to spatial analysts. Using spatial econometric methods in explanatory analyses (Anselin 2001; Anselin 2003; LeSage and Fischer 2008), spatial autocorrelation can be accounted for alongside other factors or as a global level of correlation between residuals of neighbouring observations.

Land uses commonly exhibit considerable positive autocorrelation (Chakir and Parent 2009; Hsieh et al. 2000; Irwin and Bockstael 2002; Loonen and Koomen 2009; Overmars et al. 2003; Verburg et al. 2004a; Verburg et al. 2004b). Hsieh et al. (2000) attribute positive spatial autocorrelation in urban development to spatial spillovers of population growth, while Overmars et al. (2003) attribute spatial autocorrelation in land-use patterns to otherwise unobserved factors, such as social relations or land-use agglomeration benefits. We expect that positive spatial autocorrelation in urban land use is caused by many factors all at once; and that it indicates underlying spatial processes that interact locally with either the environment or proximate spatial processes. In the case of residential land use, the interactions that cause spatial autocorrelation may consist of unobserved neighbourhood amenities, economies of scale in building larger residential blocks, the preference of house seekers for moving to places in the vicinity, scale benefits of clustered residences for local services or the preference of planning authorities for large scale zoning.

Levels of spatial association and autocorrelation are known to decrease when aggregating to more sizeable areal units (Amrhein and Reynolds 1997; Arbia 1989; Hong Chou 1991; Qi and Wu 1996), but why? When the data at hand observes individual entities, all interdependence is captured. Under aggregation the interdependence between two neighbouring entities becomes obfuscated if both are aggregated into one areal unit. Spatial autocorrelation then becomes mainly linked to the number of individual entities that are situated at the fringe of areal units, relative to the total number of individual entities in the unit; see figure 1. Consequently, observed levels of spatial autocorrelation are `damped' by aggregating to larger spatial units (Arbia, 1989, p. 100). In the case of regularly latticed data, the share of individual entities on the fringe of areal units declines monotonically under aggregation and so do levels of spatial autocorrelation. In the case of irregular areal units such scale effects on levels of autocorrelation are less predictable, because those will more likely have irregularities in numbers of neighbouring units and in the shares of individual entities on the fringe of areal units. The potentially sizeable effect of the number of neighbouring units on the results of spatial autocorrelation tests has been demonstrated by Wall (2002).

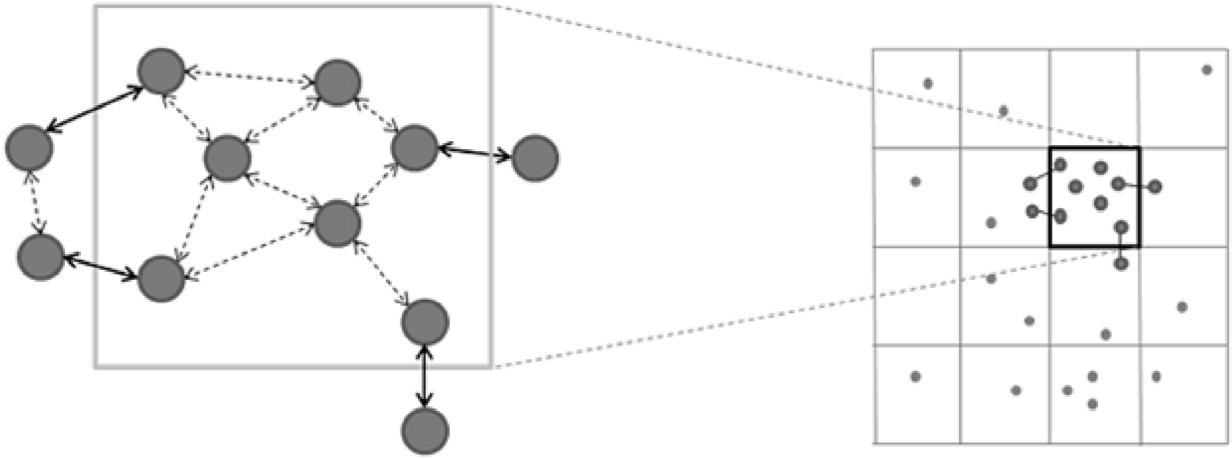


Fig. 1 Aggregation of positively spatially autocorrelated individual phenomena. Circles indicate spatial processes, arrows indicate spatial interdependencies between neighbouring processes, and gridlines indicate an aggregation scheme. The dashed arrows indicate interdependencies between individual processes that are unobservable after aggregation (based on Arbia, 1989).

3 Data and methods

In the following sections we first introduce the applied data, areal units, methods and additional variables before embarking on our demonstration of how choice of areal units affects statistical outcomes.

3.1 Urban land-use data and areal units

We analyse residential land-use shares that are derived from discretely valued land-use data in the Netherlands in 2000 in a 25 meter resolution, provided by Statistics Netherlands (2002). Residential land shares are calculated as aggregate amount of residential land per *suitable* land area in an areal unit *i* so that $Y_i = RESIDENTIAL_i/SUITABLE_i$. All the surface captured by an areal unit is considered suitable for residential land uses, except water bodies and land covered by large transport infrastructure or mining operations. The latter categories are considered not suitable. The land-use shares are averaged in various spatial data configurations; see table 2. To resemble scale effects all data are aggregated to differently sized areal units, and to resemble shape effects all data are regrouped into zone units (or zones) and three sets of regular lattices (or rasters). The selected zone units are administrative zones that are commonly applied in spatial analysis. The different sets of regular lattices only differ in their origins that (compared to the one original lattice) are moved north-westerly 25 and 50 percent of the cell width. The resolutions of these lattices resemble either resolutions (the 100 m and 1 km resolutions) that are common in land-use models (Agarwal et al. 2002; Pontius et al. 2008), or the average areas of used zone units (the coarser resolutions).

Table 1: Applied spatial data configurations. For raster units ranges of n are produced because counts of units vary with origin choice.

Raster units	N	Zone units	N
100 m.*	3,438,279		
1 km.**	36,534 – 36,585		
2 km.	9,519 - 9,548	Neighbourhoods	11,473
4 km.	2,524 - 2,550	Urban districts	2,530
10 km.	465 – 474	Municipalities	484
20 km.	137 – 138		
30 km.	65 – 70	Corop+ regions	52

All data configurations are examined in all analyses except: * not in explanatory models; ** explanatory model is based on random sample of 8,823 observations

The aggregations performed in this study cause that the studied observations are no longer linked with individual processes that act on individual residential parcels. Those parcels are 820 m^2 on average in the Netherlands[1] (Kadaster 2008). This implies that even fine resolution data such as a 1 km raster can group more than 1,000 individual processes.

3.2 Methods

Weighting methods are used to overcome shape effects that exist because of unevenly sized areal units. We have found that the applied weighting methods reduce shape effects in statistical findings. We weight both zone units and raster units for the amount of consumable land in the unit. Despite their even sizes, raster units are weighted because the edges of the study area are quite capricious, which causes differences in the average shares of relevant area covered by rasters. Larger raster units in particular can entail large portions of sea or exterior lands, which makes these units sensitive to shape effects.

Weighting is based on the comparative amount of suitable land in an areal unit, which is $S_i = SUITABLE_i / \frac{1}{n}\sum_i SUITABLE_i$. The applied weighting is similar to Robinson's method (1956). We are aware of Arbia's (1989) alternate method to compute weighting values. Arbia's weighting values supposedly perform slightly better in spatially autocorrelated data, but we expect that the performance improvement possible with those weights does not justify the greatly increased complexity involved in computing them. We are, furthermore, aware that area weighting methods do not cancel out all effects of spatial aggregation on bivariate and multivariate statistics (Arbia 1989; Thomas and Anderson 1965), but we find that, in our particular case, area weighting does reduce the overemphasis that these analyses put on those individual process-entities that are captured in the smaller units of unevenly sized aggregation schemes.

[1] This includes the parcels of apartment blocks and rental corporations. Both have multiple houses on one parcel.

Let X_i denote the value of a variable in an areal unit *i.* Then, weighted averages are computed as $\overline{XS} = \sum_i S_i X_i / \sum_i S_i$ and standard deviations as $\sigma XS = \sqrt{\sum_i S_i \, (X_i - \bar{X})^2 \, / \sum_i S_i}$. Weighted Pearson correlation coefficients (*rS*) are computed as shown in equation (1).

$$rS = \frac{S_i(X_i - \overline{XS})(Y_i - \overline{YS})}{\sqrt{\sum_i S_i \, (X_i - \overline{XS})^2} \sqrt{\sum_i S_i \, (Y - \overline{YS})^2}} \tag{1}$$

As an indication of spatial autocorrelation we apply Moran's I (Moran 1950). We weight Moran's I so that: 1) neighbours *j* with a larger suitable surface have more weight in defining the level of spatial association between an observation *i* and its neighbours *j*; and 2) observations *i* with a larger suitable surface have more weight in defining the global measure of spatial autocorrelation. The weighted Moran's I (MIS) is computed as in equation (2).

$$MIS = \frac{n \sum_i {S_i}^2}{\sum_i \sum_j S_i S_j W_{ij}} \frac{\sum_i \sum_j S_i S_j W_{ij} \, (X_i - \overline{XS})(X_j - \overline{XS})}{\sum_i {S_i}^2 (X_i - \overline{XS})^2} \tag{2}$$

Where $W_{ij} = 1$ when *i* and *j* are neighbours, $W_{ij} = 0$, otherwise.

To explore the influence of spatial aggregation on explanatory analyses we compare an Ordinary Least Squares model (OLS) with the outcome of a spatial error model (SEM). Lagrange multipliers diagnostical tests such as documented in Anselin (2005) demonstrated that a spatial error model is more suited for our particular data sets than a spatial lag model[2]. The models explaining the distribution of residential land use densities *Y* in spatial units *i* take the form of equations (3.1 – 3.2).

$$OLS\colon Y_i = \beta 0 + \beta 1 X1_i + \beta 2 X2_i + \cdots + \beta k X k_i + \varepsilon_i \tag{3.1}$$

$$SEM\colon Y_i = \beta 0 + \beta 1 X1_i + \beta 2 X2_i + \cdots + \beta k X k_i + \varepsilon_i; \tag{3.2}$$

and in the SEM model $\varepsilon_i = \lambda W_{ij} \varepsilon_j + \mu$

In the SEM model the error term ε of observation *i* consists of an independent and identically distributed (i.i.d.) disturbance term μ and the impact of spatially adjacent residuals in j. Following Anselin (2001), (3.2) can be rewritten to:

$$Y_i = \lambda W_{ij} Y_j + \beta k X k_i - \lambda W_{ij} \beta k X k_j + \mu. \tag{3.3}$$

We will use the latter form to compute predicted values Y after the model has been estimated, in order to compare accuracies of the OLS and SEM models. To obtain land-supply weighted estimates in the OLS and SEM models, we apply an exogenous constant with values ${S_i}^{1/2}$ and multiply all exogenous inputs with the same values. This weighting method is thus equal to estimating by means of weighted least squares, in which $\sum S_i (Y_i - Y_i')^2$ is minimized[3].

In the SEM model, spatial autocorrelation is located in the error term that affects the outcomes of OLS estimations (Anselin 2001; p. 11). We interpret ε as the representation of unobserved variables that are subject to spatial dependence. The spatial dependency

[2] Results of the diagnostical tests are available upon request.

[3] Weighting was not an available option in the spatial econometrics module of the used software (STATA), and we therefore resort to these prior computations.

between error terms is defined by spatial weight matrix W_{ij}, which in this exercise is based on the queen's model of contiguity (Cliff and Ord 1981; p. 247). We limit this matrix to first order neighbours because, in reverence of Tobler's first law of geography (Tobler 1970; p. 236), we expect that the most proximate observations in Y_j correlate most with Y_i.

3.3 Selected variables

In the correlation tests and the explanatory analyses a set of variables is applied that aims to capture the most important driving forces acting at different scale levels. The variables relate to local phenomena such as airport noise and transport mode proximity (distance to stations and motorway exits), and regional aspects such as spatial plans (new towns), restrictions (buffer zone and green heart) and economic opportunity (job access). All variables, although referring to different factors, are described by detailed 100 metre resolution rasters that allow aggregation to higher scale levels. Table 2 lists the most important characteristics of the variables and describes how they are aggregated. Continuous variables are aggregated by averaging, dichotomous values by predominance. Table 2 includes the standard deviations of the values of explanatory variables when aggregated into a coarser raster unit, which indicates the amount of local variance of that data. Variables with over average internal standard deviations have higher local variance, indicating that these variables represent phenomena with smaller geographic extents.

Table 2: Variable characteristics

Variable	Computed as	Aggregated by	Min	Average	Max	Internal st. dev.
Station distance	log(km)	Averaging	-2.996	1.761	3.600	0.300
Motorway exit	log(km)	Averaging	-2.996	1.561	3.729	0.298
Job access	see equation (4)	averaging	0.028	0.320	1.000	0.003
Exterior proximity	0/1	predominance	0	0.207	1	0.006
Airport noise	0/1	predominance	0	0.013	1	0.004
Buffer zone	0/1	predominance	0	0.019	1	0.007
Green heart	0/1	predominance	0	0.066	1	0.003
New town	0/1	predominance	0	0.021	1	0.005

Note: presented statistics are for the 100m raster resolution. Internal standard deviation is computed as the standard deviation of values at the 100m resolution within 1000m zones, averaged in all 1000m zones.

As indicators of ease of access we apply the natural logs of distances to nearest *railway stations* and *motorway exits.* Previous work demonstrates that the likelihood of built up land first increases and then decreases with distance to such transport system terminals (Borzacchiello et al. 2010), but such subtleties only hold on spatial resolutions finer than those applied in this study. More straightforward relations between the dependent and natural logs of distances are therefore imposed.

As an indicator of economic opportunity a potential accessibility indicator, *job access,* is applied. The used measure is similar to population potentials (Warntz 1964) or market potential (Briant et al. 2010). It can be interpreted as the number of jobs one can reach, with a *fuzzy* definition of what is *reachable* (Geurs and Ritsema van Eck 2001). In a seminal paper potential accessibility is found to positively affect the intensity of urban development

(Hansen 1959). Our measure is calculated as in equation (4) and subsequently spatially interpolated; for more details see (Jacobs 2011).

$$A_i = \sum_{j=1}^{n} \frac{P_i}{(C_{ij} + C_j)^2} \tag{4}$$

Where P_j is the number of job opportunities in zones *j* and C_{ij} is the road travel time between municipalities *j* and locations *i.* C_j is a municipality specific additional travel cost to observe the intrazonal distribution of job opportunities. Ceteris paribus, the applied distance decay function explained the most variance in residential land use shares when fitting it in the later presented regression analyses. For the sake of interpretation, the accessibility levels in this study have been rescaled in such a manner that the maximum accessibility is 1 in any spatial data configuration.

Exterior proximity indicates whether areal units are predominantly within 10 kilometres of a national border and so proxies the barrier effect that national borders have on economic interaction (Cheshire and Magrini 2008; Rietveld 2001).

A number of spatial policy indicators are applied that, except airport noise contours, are all based on administrative units. The *airport noise contours* indicate if an observation is in an area where airport noise is present and urban development is restricted. The *buffer zone* and *green heart* policy variables indicate whether areal units are predominantly in areas with severe restrictions on residential land-use development. The related policies have been considerably successful in the preservation of open space (Koomen et al. 2008). Lastly, the *new town* policy variable indicates if an areal unit is predominantly within municipalities that have been assigned residential development incentives by national planning authorities. These new town policies have positively affected Dutch urban development (Verburg et al. 2004b).

We must acknowledge that the presented models simplify the studied relations. Variables such as job access are applied exogenously, while there are more complex dependencies between, for example, the spatial distributions of residential areas and employment centres (Batey and Madden 1999; De Graaff et al. 2008), and human activities and infrastructural developments (Koopmans et al. 2012; Levinson 2008; Wegener and Fürst 1999). We assume that the lack of completeness in model specification is not problematic in our case as we focus on uncovering scale and shape effects and not so much on providing an in-depth analysis of the process of residential development.

4 Results

This section starts by demonstrating how spatial pattern properties of residential land-use shares are affected by spatial aggregation. Subsequently we discuss how spatial aggregation affects the results of bivariate correlation tests and multivariate explanatory analyses.

4.1 Spatial pattern properties

In general, the behaviour under aggregation of average values and standard deviations of land-use shares in our analysis confirms Arbia's results (Arbia 1989). Table 3 shows that the weighted average residential land use-shares ($\overline{XS}$) are unaffected by aggregation as was the case in Arbia's analyis. Note that not-weighted averages from the same data (table 7, appendix II) vary from 0.06 to 0.30, indicating the importance of weighting. The weighted standard deviations decrease monotonically under aggregation, showing that local variation is dampened by spatial aggregation. The zone units have consistently higher weighted standard deviations than their equivalent raster scales. This indicates that these zone units have a higher internal homogeneity. Apparently, the underlying design principle of the used zonal units is to achieve relatively homogenous units (e.g., cities or towns), which causes an additional shape effect next to the influence of uneven sizes. This impact of the deliberate shaping of the zonal units is also apparent in the decidedly fatter tails of residential land-use share histograms for these units; see figure 6 in appendix I.

Table 3: Properties of residential land-use shares aggregated to raster and zone data. For raster units ranges of results are produced because counts of units vary with origin choice.

Raster units (n)	Weighted average (sd)	Weighted Moran's I	Zone units (n)	Weighted average (sd)	Weighted Moran's I
100 m. (3,438,279)	0.07 (0.24)	0.82			
1 km. (~36,500)	0.07 (0.17)	0.45			
2 km. (~9,500)	0.07 (0.13)	0.38 to 0.39	Neighbourhoods (11,467)	0.07 (0.18)	0.55
4 km. (~2,500)	0.07 (0.10)	0.36	Urban districts (2,529)	0.07 (0.12)	0.42
10 km. (~470)	0.07 (0.06)	0.37 to 0.41	Municipalities (483)	0.07 (0.07)	0.35
20 km. (~140)	0.07 (0.05)	0.42 to 0.47			
30 km. (~70)	0.07 (0.04)	0.41 to 0.52	Corop+ regions (52)	0.07 (0.05)	0.46

The impact of spatial aggregation on Moran's I can be seen in table 3 and figure 2. We find that, with fewer observations, the values of Moran's I from rasters with differing origins deviate more. The weighting method does reduce these deviations, so that in all but the coarsest resolution, values of Moran's I deviate much more without weighting. We interpret the effect of changing origins on values of Moran's I as the result of higher uncertainty with smaller sample sizes. We furthermore find that in particular fine resolution zonal data have higher levels of Moran's I than comparable raster data. This is presumably caused by the higher internal homogeneity of fine resolution zonal units and the higher density of these zones in highly populated areas. These two characteristics of the applied zone units make that, if a city consists of multiple units of observation, zonal inner-city units more often border units with similar values, which results in higher levels of Moran's I.

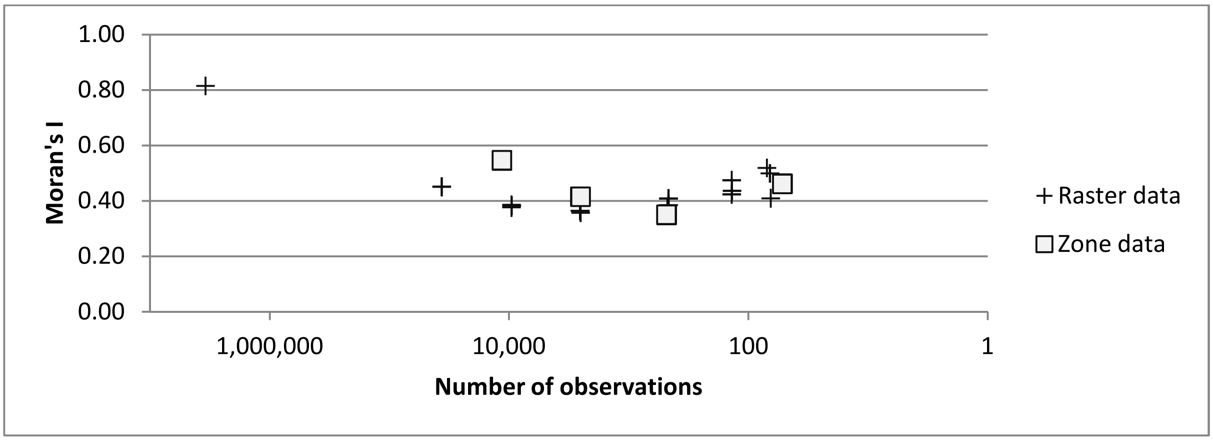


Fig. 2 Moran's I coefficient of residential land-use shares under aggregation.

We furthermore find that Moran's I decreases monotonically under aggregation up to regional levels, but increases again at coarser resolutions. Such an increase of Moran's I contradicts Arbia's expectations (1989) and previous empirical results (Hong Chou 1991; Qi and Wu 1996) and is not easily explicable. For a better understanding we visualize clustering patterns with the local indicators of spatial associations (LISA) method (Anselin 1995). The results in figure 3 demonstrate sporadic spatial associations at the municipality level, while at the coarser Corop+ level the main urban areas in the west and the peripheral northeast become identifiable as related regions (in terms of habitation). This unexpected increase in spatial autocorrelation might be due to the specific regional urbanization patterns in the Netherlands. The western part of the country is characterised by a concentration of cities at relatively short distances from each other that upon aggregation reveal the relatively densely populated urban agglomeration known as the `Randstad', while the northeast consists of fewer cities that upon aggregation become dominated by the relatively large low-density areas surrounding them.

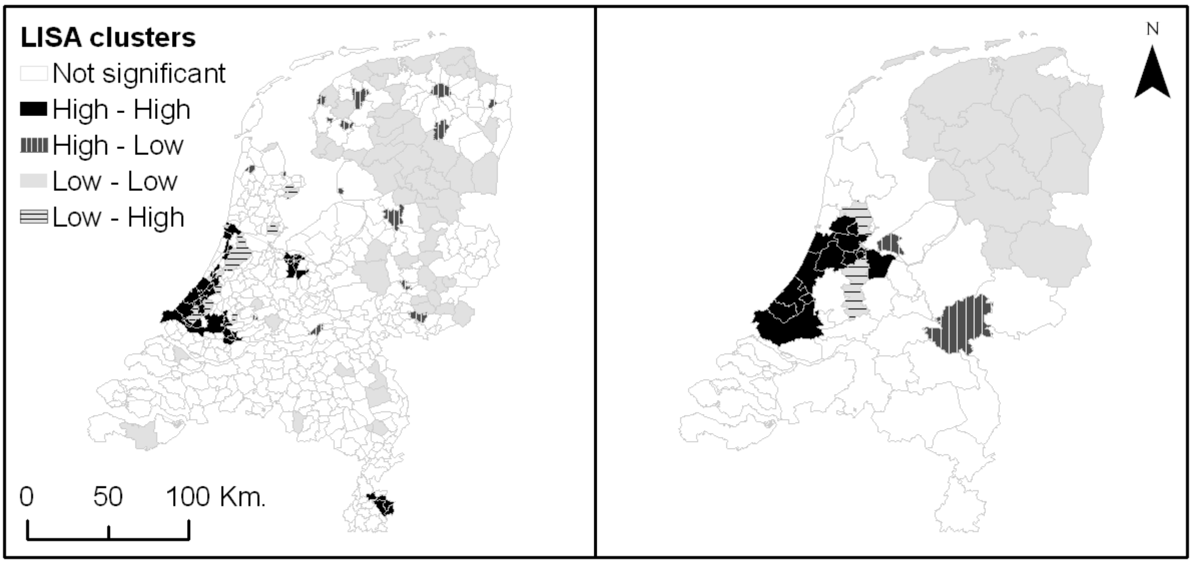


Fig. 3 Significant clustering of residential land-use shares in the Netherlands according to a LISA analysis at municipality (left) and Corop+ level (right). The Corop+ level shows substantial clusters of high and low land-use shares in respectively the west and northeast.

4.2 Bivariate correlation coefficients

We compute S_i-weighted Pearson correlation coefficients between residential land-use shares and the selected variables (see table 4). Proximity to the exterior and observed residential densities are uncorrelated, which contrasts the distinct barrier effect that national borders are expected to have on economic opportunities and urban growth. Correlations with accessibility indicators and new town policies are positive as was expected. Paradoxically, in most cases *restrictive* policies are *positively* correlated with residential land-use densities, indicating that these restrictions are imposed where housing demand is high.[4]

Table 4: Pearson correlation coefficients of residential land-use shares and individual explanatory variables under aggregation (blank in case of insufficient variation in a variable). For raster units ranges of results are produced because counts of units vary with origin choice.

Spatial units (n)	Station distance	Motorway distance	Job access	Exterior proximity	Buffer zone	Green Heart	Airport noise	New town
100 m. (3,438,279)	-0.26	-0.14	0.16	0.00	-0.04	0.03	0.01	0.07
1 km. (~36,500)	-0.35	-0.20 to -0.19	0.23	0.00	-0.02	0.04	0.02	0.09 to 0.10
2 km. (~9,500)	-0.41	-0.25	0.29	0.00	0.00 to 0.01	0.05 to 0.06	0.02 to 0.03	0.12
4 km. (~2,500)	-0.49 to -0.48	-0.33	0.38	-0.01 to 0.00	0.03 to 0.05	0.08 to 0.09	0.03 to 0.05	0.12 to 0.14
10 km. (~470)	-0.53 to -0.52	-0.44 to -0.42	0.52 to 0.54	-0.01 to 0.02	0.05 to 0.11	0.11 to 0.19	0.06 to 0.15	0.03 to 0.07
20 km. (~140)	-0.53 to -0.50	-0.47 to -0.45	0.63 to 0.65	-0.06 to 0.00		0.08 to 0.38	0.30 to 0.31	-0.07
30 km. (~70)	-0.49 to -0.44	-0.41 to -0.35	0.66 to 0.71	-0.02 to 0.05		0.05 to 0.47		
Neighb. (11,467)	-0.35	-0.21	0.21	0.00	-0.03	0.04	0.02	0.09
Districts (2,529)	-0.45	-0.30	0.32	0.00	0.01	0.06	0.06	0.14
Municip. (483)	-0.55	-0.44	0.48	0.00	0.05	0.11	0.07	0.22
Corop+ (50)	-0.66	-0.59	0.66	-0.03	0.11	0.21	0.08	0.17

Both scale and shape effects are apparent in the presented bivariate correlation coefficients. In most cases, correlation coefficients become larger[5] at coarser resolutions. The differences are substantial: most correlation coefficients become at least twice as large under aggregation. To explain this, we will look deeper into the effects of aggregation on correlation results in the subsequent section. Shape effects are apparent in the different sizes of correlation coefficients for comparable zone and raster units. However, weighting has greatly reduced those shape effects. Without weighting, the correlations with job access range from 0.48 to 0.74 in 30km rasters and Corop+ regions, and correlations with exterior proximity range from -0.14 to 0.06 in raster units and from -0.06 to -0.15 in zone units: see table 8 in appendix II. Even with weighting there are differences between raster and zone units. We link these differences to the second shape effect that exists because areal units share delineations with the studied entities and so entail data with higher internal homogeneity. This shape effect is most clearly demonstrated by the correlation coefficients of new town policies. These policies are defined at the municipality level, which makes that in both correlated variables the studied entities share delineations with the areal unit. As a result, new town policies are notably more correlated with residential land-use shares in case of municipal averages, than in case the data are aggregated to any other areal unit set. Here, higher degrees of internal homogeneity caused by the second shape effect seem to lead to overemphasised bivariate associations.

[4] As areal units get larger, the results for correlation coefficients become sensitive to the choice of origin in the system of spatial units. This result is similar to what we found for Moran's I coefficient in Figure 2.

[5] We describe *absolute* sizes of correlation and regression coefficients, so that larger has the meaning of 'farther from zero'.

The variance of correlation coefficients under aggregation is related to spatial autocorrelation in the associated processes (Arbia 1989; Openshaw 1984). Spatial autocorrelation, and bivariate associations of residential land-use shares and job access, may interplay on different resolutions, which we explore with a line profile of the associated variables on the neighbourhood and Corop+ levels (figure 4). At the Corop+ level the averaged land-use shares are smoothed and gradually increasing when job access increases, but at the neighbourhood level this smoothed pattern is replaced by spikes.

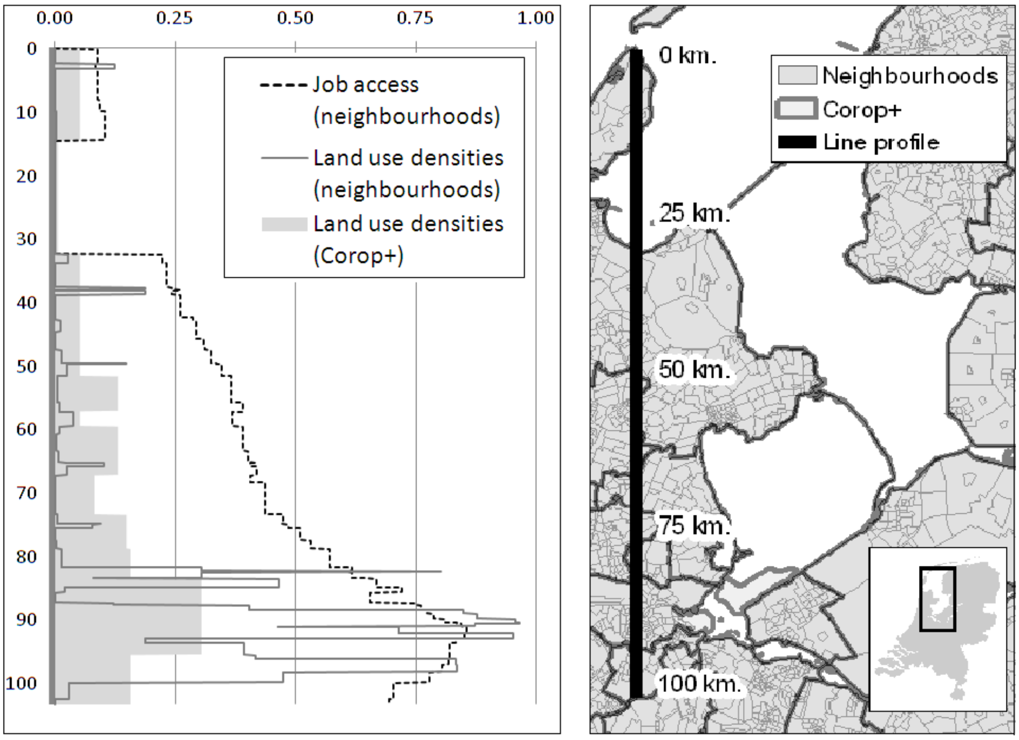


Fig. 4 Line profile of residential land-use densities and job access aggregated to neighbourhoods and Corop+ regions. In the chart (left) the Y axis indicates distance from the origin of the profile line, and the X axis indicates values of both residential land-use densities and job access. The map (right) indicates the position of the line profile in the Netherlands.

It is immediately clear that job access is far less associated with residential land-use share at the neighbourhood level, and as the previously presented correlations on the neighbourhood level show, the same goes for the other variables used in this study. In contrast, we have seen that levels of spatial autocorrelation are higher at the finer resolution of neighbourhood zones, so that we presume a trade-off between spatial autocorrelation and bivariate associations when changing areal unit sizes. To invoke the analogy of spectral densities (Curry 1966) the spiked pattern at low resolutions represents `short wavelength processes’ that are filtered out by large areal units; where many variables in the correlation tests lack sufficient local variation, spatial autocorrelation picks up the `short wavelength’. The result of aggregating to larger units is that the data are smoothed into representing only the results of higher wavelength processes, with which the correlated variables are more associated. As shown in the next section, multivariate analyses can account for both short and long wavelength processes, and such analyses may therefore be less susceptible to scale effects than correlation tests.

4.3 Multivariate analyses

We separately assess how choice of areal units affects model accuracy, spatial lag effects and other coefficients in multivariate analyses. To keep the computational tasks manageable the 100m scale is excluded and the 1km scale model is only fitted to a sample of the data.

Table 5: Explained variance and levels of spatial autocorrelation in model residuals (λ). For raster units ranges of results are produced because counts of units vary with origin choice.

Areal units (n)	R^2 OLS	Pseudo-R^2 SEM	λ
1 km. (9,766)	0.32-0.33	0.98-0.99	0.79*
2 km. (9,548)	0.37-0.38	0.94	0.63*-0.64*
4 km. (2,550)	0.52-0.53	0.91-0.93	0.59*-0.62*
10 km. (465)	0.71-0.73	0.87-0.89	0.52*-0.55*
20 km. (137)	0.82-0.84	0.87-0.97	0.53*-0.74*
30 km. (65)	0.84-0.88	0.81-0.93	0.41*-0.64*
Neighbourhoods (11,467)	0.25	0.96	0.64*
Districts (2,529)	0.43	0.92	0.56*
Municipalities (483)	0.69	0.74	0.44*
Corop+ regions (52)	0.86	0.35	0.23

Note: all spatial lag coefficients indicated with * are significant at the 0.05 level, other coefficients not.

As a measure of explained variance we present R^2 values of the fitted models[6]. We thus find that the SEM model clearly explains more variation in land-use shares at fine resolutions than the OLS model; see table 5. Note that when zones are very small, Moran's I is high and also lambda is high. In this case it is clear that by taking into account observations on neighbouring zones one can achieve a high level of precision when predicting the dependent variable in a particular zone. This explains the large difference between the (pseudo) R2 values of the OLS and SEM models when lambda is high. The results of estimating spatial autocorrelation in the error term are plotted in figure 5. Those results show that, similar to values of Moran's I, the effect of spatial autocorrelation in the error term changes unpredictably under aggregation, particularly in the case of rasters (cf. figures 2 and 5).

[6] SEM R^2 values are calculated by computing equation (3.3) with the estimated parameters and subsequently squaring the correlation between observed and estimated values of the dependent.

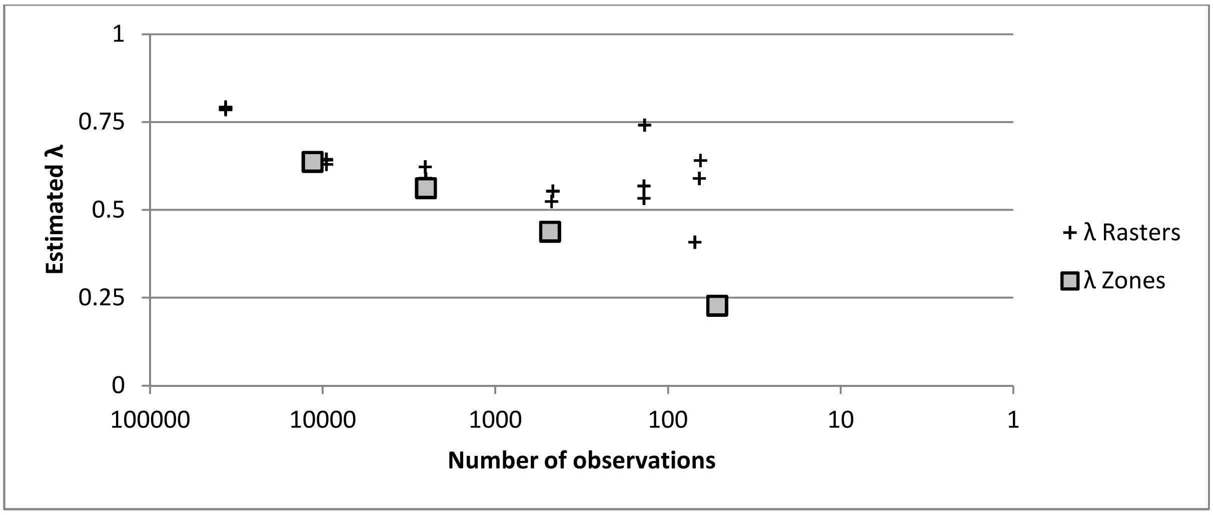


Fig. 5 Values of λ of the SEM model under aggregation. Raster results with similar counts are from rasters with different origins.

Table 6 presents estimated constant terms and coefficients of our explanatory analysis. If anything, the results demonstrate the strong `self-reinforcing' (Verburg et al. 2004b; p. 137) tendencies of human settlement. Such self-reinforcing tendencies show in the effects of spatial autocorrelation, job access and station distance that all indicate that residential land use is more likely where people have better interaction opportunities with others. Motorway exit proximity wavers between insignificant positive and negative effects. Its estimated effect apparently suffers from ambiguity: although the environs of highways are unattractive for housing, the ease of access provided by motorway exit proximity increases attraction. Exterior proximity has a significant positive effect, presumably because other variables underestimate cross-border interaction opportunities. The effects of airport noise contours on residential land-use densities are mostly insignificant. From the estimated effects of incorporated land-use policies follows that these have had substantial success. Thus, the restrictive green heart and buffer zone policies have significant negative, and new town incentives significant positive effects on residential land-use shares.

Table 6: Coefficients of the estimated multivariate regression analysis. The last column expresses the correlation of the estimated coefficients with the number of observations, while the last row indicates the average difference between OLS and SEM coefficients.

		1 km.	**2 km.**	**4 km.**	**10 km.**	**20 km.**	**30 km.**	**Neighb.**	**Distr.**	**Muncp.**	**Corop+**	**Aggr. trend**
		(*9,766*)	(9,548)	(2,550)	(465)	(137)	(65)	(11,467)	(2,529)	(483)	(52)	
Const.	OLS	0.06*	0.11*	0.11*	0.08*	0.03	-0.01	0.15*	0.13*	0.11*	0.10	0.48
	SEM	0.12*	0.18*	0.15*	0.09*	0.03	0.00	0.09*	0.13*	0.10*	0.11*	0.62
Stat. dist	OLS	-0.08*	-0.06*	-0.05*	-0.04*	-0.02*	-0.02	-0.07*	-0.06*	-0.04*	-0.04*	-0.90
	SEM	-0.13*	-0.09*	-0.08*	-0.04*	-0.02*	-0.03*	-0.06*	-0.06*	-0.05*	-0.05*	-0.88
M'way exit	OLS	0.01*	0.00	-0.01	0.00	0.01	0.02	0.00	-0.01	-0.01	0.00	-0.17
	SEM	0.03*	0.01	0.00	0.00	0.00	0.01	0.00	-0.01	-0.01	-0.01	0.39
Job acc.	OLS	0.32*	0.18*	0.16*	0.14*	0.15*	0.16*	0.17*	0.17*	0.16*	0.14*	0.67
	SEM	0.31*	0.14*	0.15*	0.16*	0.19*	0.17*	0.11*	0.12*	0.19*	0.13*	0.30
Ext. prox.	OLS	0.02*	0.02*	0.02*	0.02*	0.02*	0.03*	0.02*	0.02*	0.02*	0.03*	-0.80
	SEM	0.01	0.01	0.01	0.02*	0.02*	0.03*	0.01	0.01	0.02*	0.03*	-0.91
Buff. Zone	OLS	-0.12*	-0.06*	-0.04*	0.06			-0.10*	-0.06*	-0.03	0.04	-0.86
	SEM	-0.20*	-0.12*	-0.07*	-0.02			-0.10*	-0.08*	-0.08*	0.03	-0.95
Green Hrt.	OLS	-0.06*	-0.05*	-0.04*	-0.03*	-0.04*	-0.05*	-0.04*	-0.05*	-0.04*	-0.04	-0.46
	SEM	-0.05*	-0.04*	-0.04*	-0.03*	-0.08*	-0.06*	-0.03*	-0.05*	-0.06*	-0.04	0.44
Airp. Noise	OLS	0.00	-0.02	0.00	0.05*			-0.02	0.01	0.01	0.00	-0.29
	SEM	-0.01	-0.05*	-0.05*	0.03			-0.04*	0.00	-0.02	-0.01	-0.48
New Town	OLS	0.11*	0.07*	0.05*	0.04	-0.04		0.06*	0.06*	0.07*	0.04	0.72
	SEM	0.05*	0.06*	0.05*	0.05*	-0.04		0.03*	0.03	0.08*	0.06	0.34
λ	SEM	0.79*	0.64*	0.62*	0.55*	0.74*	0.64*	0.64*	0.56*	0.44*	0.23	0.53
Average diff.		0.03	0.03	0.02	0.02	0.01	0.01	0.02	0.01	0.02	0.01	0.94

Note 1: all coefficients are significant at the 0.05 level, unless indicated as: * not significant at 0.05 level. Blank spaces indicate coefficients that could not be estimated due to insufficient variation of the variable.
Note 2: For the 1km. raster a sample of the observations is taken. The aggregation trends are computed with the log of number of observations, with the 1km resolution set to 36,534 observations.
Note 3: because of space limitations, for each scale the results of only one set of rasters are presented. Varying origins does not affect the order of magnitude of estimated coefficients; results are available upon request.

The bottom row of table 6 shows the influence of model specification (OLS versus SEM) on estimated coefficients by listing the average absolute difference in coefficient of the different models per set of areal units. These ranges demonstrate that particularly with fine resolutions, the specification of the used models (OLS versus SEM) has a limited influence on the estimated coefficients. Parallel with the decreasing effect of lambda, the differences between estimated coefficients become smaller when aggregating to coarser resolutions. In the utmost right column of table 6, the aggregation trends computed as correlations between coefficient size and ln(number of observations) demonstrate that estimation results for some variables (e.g. station distance, buffer zone policies) are strongly associated with resolution. This demonstrates that with several of the applied variables, scale has a rather monotonous effect on coefficient size. Where coefficients vary unexpectedly, the differences in coefficient size are relatively small. Changes in coefficient sign, which possibly are the most troublesome of scale effects, are rare, and occur mostly if estimated effects vary around zero (such as motorway exit proximity) or have almost insufficient variance due to the aggregation process (such as new town and buffer zone policies at coarse resolutions). We conclude that the multivariate analysis yields very similarly sized coefficients for a wide range of resolutions if sufficient variance in variable values is maintained after aggregation. Correcting for spatial autocorrelation provides a higher explanatory power but, in this case, provides similar coefficients. Shape effects still are visible in the results; for example when comparing the coefficients estimated on 30km rasters and Corop+ zones. However, when comparing the results in table 6 with table 9 in appendix II, it is clear that weighting has decreased the impact of shape effects here.

5 Conclusion

This paper demonstrates to what extent statistically analyzing urban land consumption is impacted by the scales and shapes of aggregated areal units. To do so, residential land-use shares are averaged into regularly and irregularly shaped areal units of various sizes. In all our statistical computations, the observations are weighted to remove the biases that originate from variations in the amount of consumable land that the areal unit represents. *Scale effects* are subsequently quantified by comparing results from different resolutions and *shape effects* by comparing results obtained from zone units, and raster units with varying origins. The degree to which spatial data aggregation influences multivariate analysis results is limited when the analyzed data has positive spatial autocorrelation and all variables are aggregated similarly (Amrhein, 1995; Briant et al, 2010). We expect that the aggregation methods and the properties of the data used in this analysis are representative for many other urban land consumption analyses, and we therefore expect that the conclusions in this study are useful to better understand the impacts of spatial aggregation in similar studies.

5.1 Scale effects

Altering the sizes of areal units causes monotone increases or decreases in most results of univariate and multivariate analyses, but sometimes causes unexpected changes in the results of correlation tests and levels of spatial autocorrelation. Essentially, aggregating to coarser resolutions smooths over local variation in the spatial distribution of urban land use, so filtering out processes with `wavelengths’ smaller than the size of the aggregated areal units (Curry 1966). This smoothing dynamic makes that the estimated effects of typically local factors (such as station distance) decrease and loose significance at coarser resolutions. Vice versa, urban patterns have more variance on a fine resolution than the explanatory factors that are accounted for can likely incorporate. Higher levels of spatial autocorrelation and much higher shares of explained variance at a fine resolution demonstrate that the otherwise unobserved variance in neighbouring observations can proxy such small-grained variance. This implies that spatial econometric methods are especially needed when using such data. Next to the well documented econometric arguments for using such methods (see Anselin 2001; Lesage and Fischer 2008), incorporating neighbourhood dependencies likely increases the often disappointingly low share of explained variance of models using small-grained data (Irwin et al. 2006; Kok and Veldkamp 2001). Those methods furthermore give modellers a powerful method, for example by using (3.3), to obtain plausible clustered patterns when predicting spatial patterns with empirically estimated parameters.

Our multivariate regression results indicate that a model that incorporates explanatory variables referring to driving forces at different scales is able to account for small-grained variations, as well as overarching regional differences. The results of this study underpin Briant et al. (2010) that such a model is less susceptible to scale effects. In fact model specification turns out to be more influential than the exact choice of statistical analysis method (OLS or SEM). That results depend more crucially on model specification confirms previous findings (Amrhein 1995; Briant et al. 2010). We conclude that the areal units that approximate individual parcels the most are to be preferred for multivariate urban development analyses, because the data in such units are able to inform of processes on the widest range of ‘wavelengths’. Our results demonstrate that aggregation does not drastically affect findings when data are aggregated; thus, if the research question deals solely with regional level variables, more aggregate data can be used. However, analyzing coarse resolution data will clearly not capture the potentially large impacts of `short wavelength’ factors such as spatial dependence or proximity to transport nodes on urban land consumption.

5.2 Shape effects

We find that the regrouping of areal units affects statistical outcomes in two ways. A first shape effect exists because areal units vary in geographical size, or observed relevant area. Such varied sizes cause that equal amounts of space, and the entities that relate to that space, are not treated with equivalent weight in statistical tests. This is particularly problematic if the objective is to analyze the consumption of land, as is commonly the case in urban development analyses. Weighting methods such as applied in this paper can greatly reduce this bias, as has been demonstrated before (Arbia 1989; Robinson 1956). The second shape effect exists because the delineations of irregular areal unit schemes (such as postcode areas or administrative units) may be related to the studied individual entities, which causes that such data may have a structurally higher internal homogeneity than their regularly latticed counterparts. In regular lattices, the event that the delineations of areal units separate homogenous entities is less probable, and regular lattices are therefore to be preferred as a basis for urban development analysis. We discovered a feature of rasters that nevertheless is unattractive and that did not yet receive systematic attention in the discussion of this issue. The choice of a reference point for a raster will strongly affect the outcome of the analysis. In the case of small spatial units this effect will be negligible, but as shown in table 2 and figure 2, it may affect outcomes of analyses to some extent when raster resolution is coarse. This holds true in particular in irregularly shaped study areas, since these irregular shapes will strongly affect the weight of spatial units near borders.

Acknowledgements

The authors wish to thank Ronnie Lassche and Henk Scholten (both VU University Amsterdam) for providing data and advice, and Martin van der Beek and Maarten Hilferink (both ObjectVision) for help with using the GeoDMS software applied in this analysis. The Netherlands Environmental Assessment Agency (PBL) is thanked for providing part of the required data sets. This research was funded by the Netherlands Organisation for Scientific Research (NWO) and the Dutch National Research Programme ‘Knowledge for Climate’.

References

Agarwal C, Green G M, Grove J M, Evans T P, Schweik C M (2002) A review and assessment of land-use change models: dynamics of space, time, and human choice. U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newton Square, PA

Amrhein C G (1995) Searching for the elusive aggregation effect: evidence from statistical simulations. Environ Plann A 27(1): 105 – 119

Amrhein C G, Reynolds H (1997) Using the Getis statistic to explore aggregation effects in metropolitan Toronto census data. Can Geogr 41(2): 137 – 149

Anselin L (1995) Local Indicators of Spatial Association - LISA. Geographical Analysis 27 (2): 93 – 115

Anselin L (2001) Spatial econometrics. In: Baltagi, B H (ed) A companion to theoretical econometrics. Malden, Ma, Blackwell Publishing Ltd, pp 310 – 330

Anselin L (2003) Spatial externalities, spatial multipliers, and spatial econometrics. Int Reg Sci Rev 26(2): 153 – 166

Anselin L (2005) Exploring spatial data with GeoDa: a workbook. Spatial Analysis Laboratory and Center for Spatially Integrated Social Science.

Arbia G (1989) Spatial data configuration in statistical analysis of regional economic and related problems. Kluwer academic publishers, Dordrecht

Batey P, Madden M (1999) The employment impact of demographic change: a regional analysis. Pap Reg Sci 78: 69 – 87

Borzacchiello M T, Nijkamp P, Koomen E (2010) Accessibility and urban development: a grid-based comparative statistical analysis of Dutch cities. Environ Plann B 37(1): 148 – 169

Briant A, Combes P-P, Lafourcade M (2010) Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? J Urban Econ 67

Chakir R, Parent O (2009) Determinants of land use changes: A spatial multinomial probit approach. Pap Reg Sci 88(2): 328 – 344

Cheshire P, Magrini S (2008) Urban growth drivers and spatial inequalities: Europe - a case with sticky people. DSE University of Venice, Venice

Cliff A D, Ord J K (1981) Spatial processes. Models & applications. Pion, London

Curry L (1966) A note on spatial association. Prof Geogr 18(2): 97 – 99

De Graaff T, Van Oort F, Boschman S (2008) Woon-werkdynamiek in Nederlandse gemeenten. Ruimtelijk Planbureau, Den Haag

Fotheringham A S (1989) Scale-independent spatial analysis. In: Goodchild, M Gopal, S (eds) The accuracy of spatial databases. Taylor & Francis, London, pp 221 – 228

Fotheringham A S, Wong D W S (1991) The modifiable areal unit problem in multivariate statistical analysis. Environ Plann A 23(7): 1025 – 1044

Gehlke C E, Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. J Am Stat Assoc 29(185): 169 – 170

Geurs K, Ritsema van Eck J R (2001) Accessibility measures: Review and Applications. RIVM, Bilthoven

Gotway, C A, Young, L J (2002) Combining Incompatible Spatial Data. J Am Stat Assoc 97(458): 632 – 648

Hansen W G (1959) How accessibility shapes land use. J Am Inst Plan 25: 73 – 76

Hong Chou Y (1991) Map resolution and spatial autocorrelation. Geogr Anal 23(3): 228 – 246

Hsieh, W H, Irwin, E G, and Forster, L, 2000, "Spatial dependence among county-level land use changes", paper presented at the American Agricultural Economics Association summer meeting 2000, Tampa FL

Irwin E G, Bell K P, Geoghegan J (2003) Modeling and managing urban growth at the rural-urban fringe: a parcel-level model of residential land use change. Agric Resour Econ Rev 32(1): 83 – 102

Irwin E G, Bockstael N E (2002) Interacting agents, spatial externalities and the evolution of residential land use patterns. J Econ Geogr 2: 31 – 54

Irwin, E G, Bockstael, N E, and Cho, H J, 2006, "Measuring and modeling urban sprawl: data, scale and spatial dependencies", paper presented at the North American Regional Science Association meeting 2006, Toronto

Jacobs C G W (2011) Integration of spatially explicit potential accessibility measures in Land Use Scanner. VU University, Amsterdam

Kadaster (2008) Cadastral map of the Netherlands. Cadastral agency Netherlands, Apeldoorn

Kok K, Veldkamp A (2001) Evaluating impact of spatial scales on land use pattern analysis in Cental America. Agric Ecosyst Environ 85: 205 – 221

Koomen E, Dekkers J, Van Dijk T (2008) Open space preservation in the Netherlands: planning, practice and prospects. Land Use Policy 25(3): 361 – 377

Koopmans C, Rietveld P, Huijg A (2012) An accessibility approach to railways and municipal population growth, 1840 - 1930. J Transp Geogr 25: 98 – 104

Kwan M-P, Weber J (2008) Scale and accessibility: implications for the analysis of land use-travel interaction. Appl Geogr 28(2): 110 – 123

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecol 74(6): 1659 – 1673

LeSage J P, Fischer M M (2008) Spatial growth regressions: model specification, estimation and interpretation. Spat Econ Anal 3(3): 275 – 304

Levinson D (2008) Density and dispersion: the co-development of land use and rail in London. J Econ Geogr 8: 55 – 77

Loonen W, Koomen E (2009) Calibration and validation of the Land Use Scanner allocation algorithms. Netherlands Environmental Assessment Agency, Bilthoven

Moran P A P (1950) Notes on continuous stochastic phenomena. Biometrika 37: 17 – 23

Openshaw S (1984) The Modifiable Areal Unit Problem. Geo Books, Norwich

Overmars K P, De Koning G H J, Veldkamp A (2003) Spatial autocorrelation in multi-scale land use models. Ecol Modell 164: 257 – 270

Pieterse N, Van der Wagt M, Daalhuizen F, Piek M, Künzel F, Aykaç R (2005) Het gedeelde land van de Randstad. Ontwikkeling en toekomst van het Groene Hart. NAi Uitgevers/RPB, Rotterdam/Den Haag

Pontius R H, Boersma W, Castella J-C, Clarke K, de Nijs T, Dietzel C, Duan Z, Fotsing E, Goldstein N, Kok K, Koomen E, Lippitt C, McConnell W, Mohd Sood A, Pijanowski B, Pithadia S, Sweeney S, Trung T, Veldkamp A, Verburg P H (2008) Comparing the input, output, and validation maps for several models of land change. Ann Reg Sci 42: 11 – 37

Qi Y, Wu J (1996) Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. Landsc Ecol 11(1): 39 – 49

Rietveld P (2001) Obstacles to openness of border regions in Europe. In: Van Geenhuizen, M Ratti, R (eds) Gaining advantage from open borders. An active space approach to regional development. Ashgate, Aldershot, pp 79 – 96

Robinson A H (1956) The necessity of weighting values in correlation analysis of areal data. Ann Assoc Am Geogr 46(2): 233 – 236

Robinson W S (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15(3): 351 – 357

Statistics Netherlands (2002) Description of land use data. Statistics Netherlands, Voorburg

Thomas E N, Anderson D L (1965) Additional comments on weighting values in correlation analysis of areal data. Ann Assoc Am Geogr 55(3): 492 – 505

Tobler W R (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46: 234 – 240

Tobler W R (1989) Frame independent spatial analysis. In: Goodchild, M Gopal, S (eds) The accuracy of spatial databases. Taylor & Francis, London, pp 115 – 122

Verburg P H, De Nijs T C M, Ritsema van Eck J, Visser H, de Jong K (2004a) A method to analyse neighbourhood characteristics of land use patterns. Comput Environ Urban Syst 28(6)

Verburg P H, Ritsema van Eck J R, De Nijs T C M, Dijst M J, Schot P (2004b) Determinants of land-use change patterns in the Netherlands. Environ Plann B 31(1): 125 – 150

Verburg, P H, Schot P P, Dijst M J, Veldkamp A (2004c) Land use change modelling: current practice and research priorities. GeoJournal 61: 309 – 324

Wall M M (2002) A close look at the spatial structure implied by the CAR and SAR models. J Stat Plan Inference 121: 311 – 324

Warntz W (1964) A new map of the surface of population potentials for the United States, 1960. Geogr Rev 54(2): 170 – 184

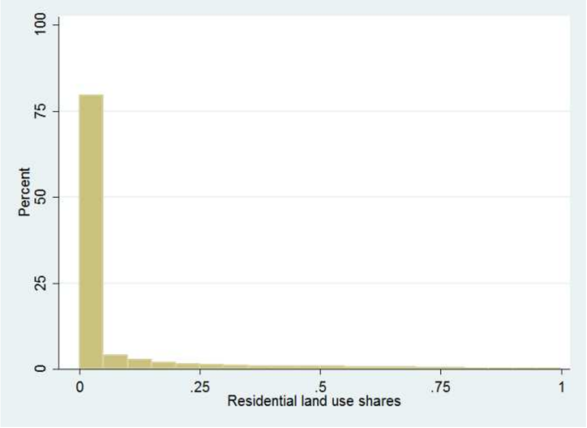
Wegener M, Fürst F (1999) Land-Use Transport Interaction: State of the Art. Fakultät Raumplanung, Universität Dortmund, Dortmund

Xie F, Levinson D (2010) How streetcars shaped suburbanization: a Granger causality analysis of land use and transit in the Twin Cities. J Econ Geogr 10: 453 – 470

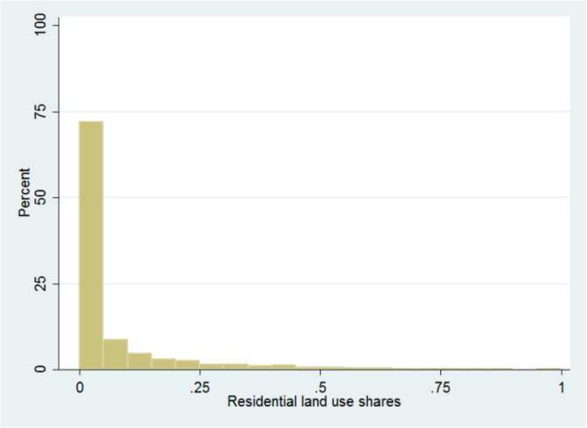
Appendix I: distribution of residential land uses in areal units

Figure 6: distribution of residential land-use share values in the spatial data configurations used in the explanatory analysis.

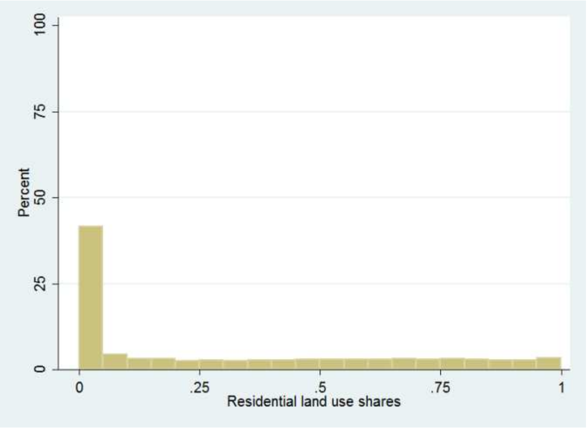
1x1km raster (9,766)



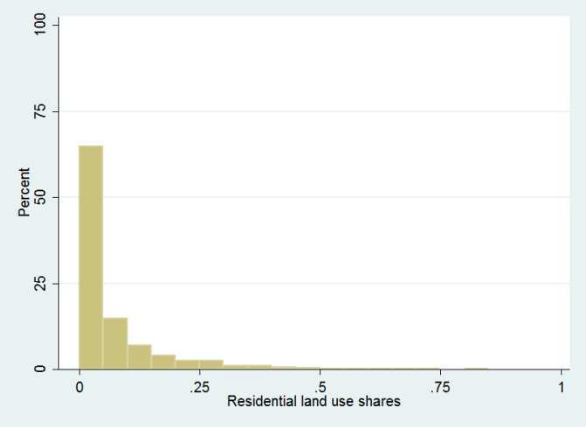
2x2km raster (9,548)



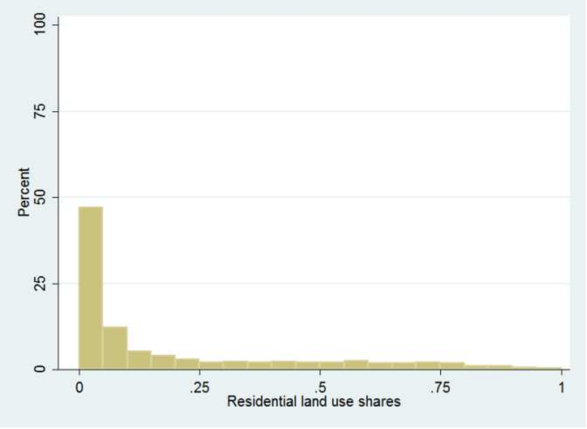
Neighbourhoods (11,467)



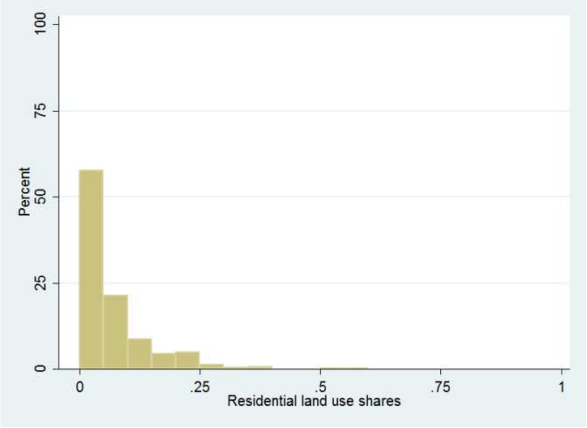
4x4km raster (2,550)



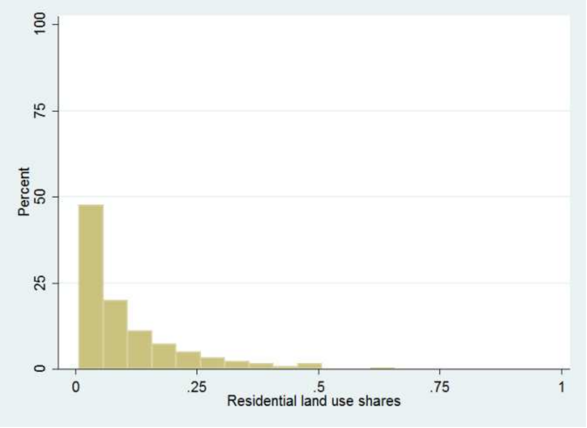
Districts (2,529)



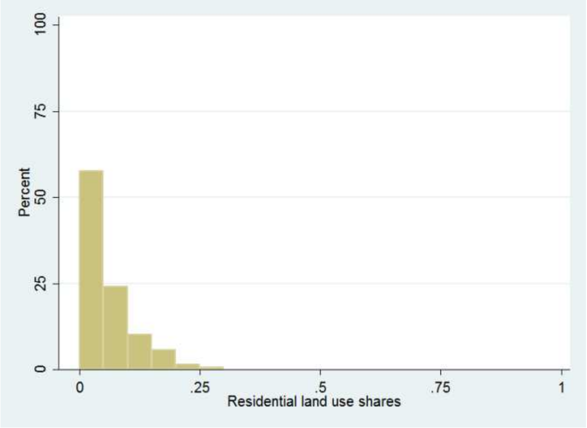
10x10 km raster (465)



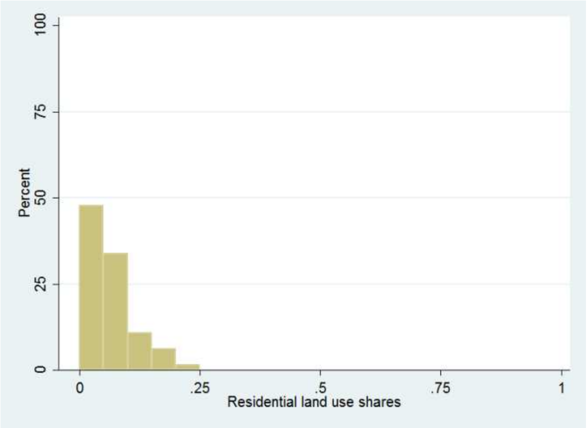
Municipalities (483)



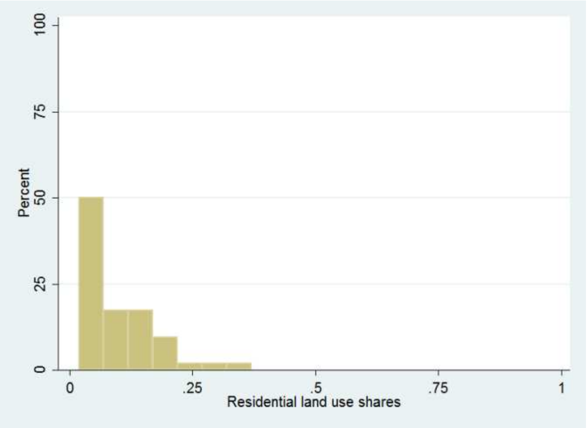
20x20 km raster (137)



30x30 km raster (65)



Corop+ regions (52)



Appendix II: results of statistics in paper when data are not weighted

Table 7: Properties of unweighted residential land-use shares aggregated to raster and zone data. For raster units ranges of results are produced because counts of units vary with origin choice.

Raster units	Mean (sd)	Moran's I	Zonal units	Mean (sd)	Moran's I
100 m. (3,438,279)	0.07 (0.24)	0.91			
1 km. (~36,500)	0.07 (0.16 to 0.17)	0.69			
2 km. (~9,500)	0.07 (0.13)	0.63 to 0.64	Neighb. (11,467)	0.30 (0.33)	0.69
4 km. (~2,500)	0.06 (0.10)	0.60 to 0.62	Urban dist. (2,529)	0.20 (0.26)	0.71
10 km. (~470)	0.06 to 0.07 (0.07)	0.62 to 0.63	Muncp. (483)	0.10 (0.10)	0.49
20 km. (~140)	0.06 to 0.07 (0.05 to 0.08)	0.42 to 0.68			
30 km. (~70)	0.06 (0.05 to 0.06)	0.40 to 0.61	Corop+ (52)	0.10 (0.07)	0.60

Table 8: Unweighted Pearson correlation coefficients of residential land-use shares and individual explanatory variables under aggregation (blank in case of insufficient variation in a variable). For raster units ranges of results are produced because counts of units vary with origin choice.

Spatial units	Stat. dist.	M'way dist.	Job acc.	Ext. Prox.	Buff. zone	Green Hrt.	Airp. noise	New town
100 m.	-0.26	-0.13	0.16	0.00	-0.04	0.03	0.01	0.07
1 km.	-0.36 to -0.35	-0.20 to -0.20	0.23 to 0.24	-0.01 to -0.01	-0.02 to -0.02	0.04 to 0.05	0.02 to 0.02	0.09 to 0.10
2 km.	-0.42 to -0.41	-0.26 to -0.25	0.29 to 0.29	-0.02 to -0.01	0.00 to 0.01	0.06 to 0.06	0.02 to 0.03	0.12 to 0.12
4 km.	-0.48 to -0.47	-0.34 to -0.32	0.37 to 0.38	-0.03 to -0.02	0.02 to 0.06	0.08 to 0.10	0.03 to 0.05	0.12 to 0.13
10 km.	-0.54 to -0.42	-0.46 to -0.33	0.44 to 0.52	-0.08 to 0.00	0.03 to 0.08	0.12 to 0.15	0.06 to 0.12	0.05 to 0.07
20 km.	-0.55 to -0.16	-0.49 to -0.09	0.32 to 0.65	-0.14 to 0.06		0.02 to 0.32	0.14 to 0.24	-0.04
30 km.	-0.49 to -0.34	-0.42 to -0.29	0.48 to 0.66	-0.03 to 0.06		0.05 to 0.31		
Neighb.	-0.40	-0.20	0.29	-0.07	-0.09	0.02	-0.01	0.12
Urban dist.	-0.53	-0.28	0.34	-0.06	-0.07	0.01	-0.01	0.19
Muncp.	-0.55	-0.38	0.43	-0.06	-0.02	0.08	0.05	0.31
Corop+	-0.80	-0.67	0.74	-0.15	0.11	0.25	0.03	0.38

Table 9: Coefficients of the estimated unweighted multivariate regression analyses. The last column expresses the correlation of the estimated coefficients with the number of observations. while the last row indicates the average difference between OLS and SEM coefficients.

		1 km.	**2 km.**	**4 km.**	**10 km.**	**20 km.**	**30 km.**	**Neighb.**	**Districts**	**Muncp.**	**Corop+**	**Corr.**
		(9,766)	(9,548)	(2,550)	(465)	(138)	(70)	(11,467)	(2,529)	(483)	(52)	
Const.	OLS	0.06**	0.11**	0.11**	0.08**	-0.06	0.03	0.28**	0.24**	0.14**	0.13*	0.42
	SEM	0.13**	0.18**	0.16**	0.08**	-0.07	0.03	0.39**	0.31**	0.14**	0.13*	0.61
Stat. Dist.	OLS	-0.08**	-0.06**	-0.05**	-0.04**	-0.02	-0.03	-0.12**	-0.13**	-0.06**	-0.06**	-0.57
	SEM	-0.13**	-0.09**	-0.08**	-0.04**	-0.01	-0.03	-0.12**	-0.15**	-0.06**	-0.06**	-0.80
M'way exit	OLS	0.01**	0.00	0.00	0.00	0.04*	0.02	0.01**	0.01	-0.01	0.00	-0.18
	SEM	0.03**	0.01*	0.00	0.00	0.04*	0.02	0.01	0.02	-0.01	-0.01	0.18
Job acc.	OLS	0.31**	0.17**	0.15**	0.15**	0.22**	0.12**	0.40**	0.31**	0.19**	0.16**	0.64
	SEM	0.26**	0.13**	0.14**	0.18**	0.24**	0.12**	0.32**	0.28**	0.21**	0.16**	0.46
Ext. Prox.	OLS	0.02**	0.02**	0.02**	0.02**	0.03*	0.03*	0.03**	0.05**	0.02	0.04*	-0.32
	SEM	0.01	0.01	0.01	0.02**	0.02	0.03*	0.01	0.03	0.01	0.04*	-0.70
Buff. Zone	OLS	-0.12**	-0.06**	-0.03*	0.00			-0.30**	-0.17**	-0.06**	0.02	-0.69
	SEM	-0.19**	-0.12**	-0.07**	-0.02			-0.30**	-0.18**	-0.09**	0.01	-0.84
Green Hrt.	OLS	-0.06**	-0.04**	-0.04**	-0.05**	-0.07	0.01	-0.09**	-0.10**	-0.05**	-0.04	-0.50
	SEM	-0.03*	-0.03**	-0.04**	-0.07**	-0.08	0.01	-0.07**	-0.09**	-0.05**	-0.04*	-0.14
Airp. Noise	OLS	0.00	-0.02	-0.01	-0.01	0.05		-0.05*	-0.01	0.01	-0.03	-0.38
	SEM	-0.01	-0.05**	-0.06**	-0.05	0.05		-0.05	-0.01	-0.02	-0.03	-0.38
New Town	OLS	0.10**	0.07**	0.05**	0.00			0.11**	0.16**	0.14**	0.05	0.54
	SEM	0.04*	0.06**	0.05**	0.01			0.05*	0.13**	0.14**	0.05*	0.24
λ	SEM	0.79**	0.64**	0.63**	0.50**	0.33**	-0.12	0.63**	0.61**	0.43**	0.22	0.88
Average diff.		0.04	0.03	0.02	0.02	0.01	0.00	0.03	0.02	0.01	0.00	0.83

Note 1: all coefficients are significant at the 0.05 level unless indicated as: * not significant at 0.05 level. Blank spaces indicate coefficients that could not be estimated due to insufficient variation of the variable.

Note 2: For the 1km. raster a sample of the observations is taken. The aggregation trends are computed with the log of number of observations with the 1km resolution set to 36.534 observations.

Note 3: because of space limitations, for each scale the results of only one set of rasters are presented. Varying origins does not affect the order of magnitude of estimated coefficients; results are available upon request.