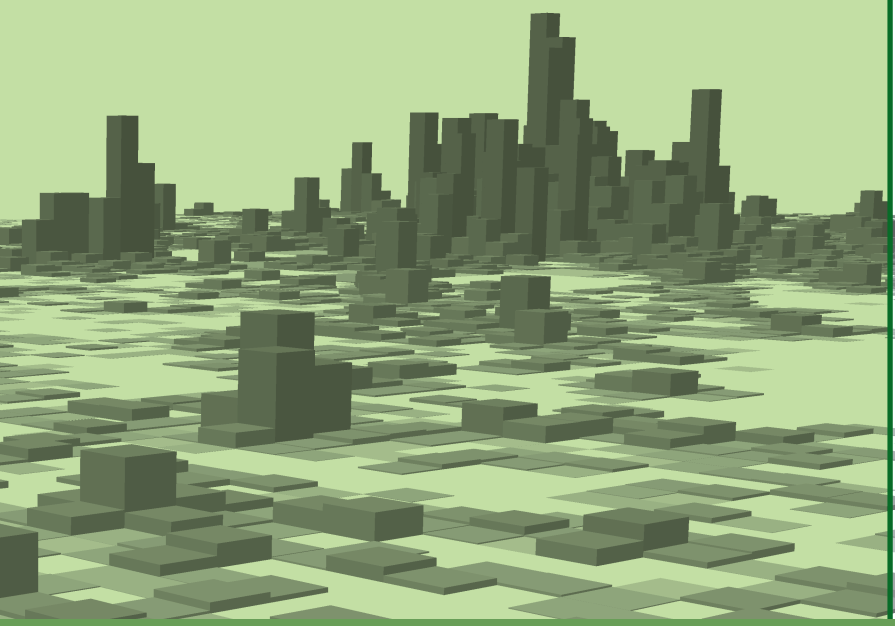
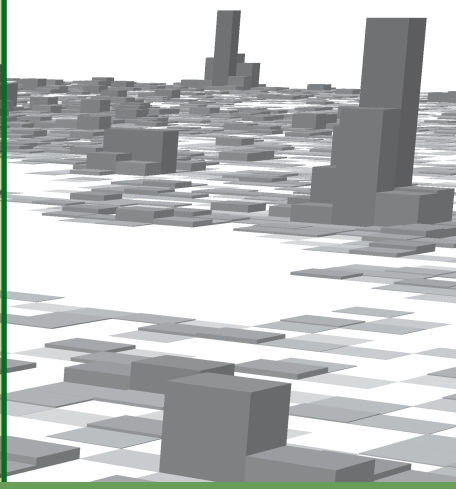
GEOGRAPHIA SLOVACA

31





SPATIAL DISAGGREGATION OF POPULATION DENSITY USING LAND COVER AND REMOTE SENSING DATA

PRIESTOROVÁ DEZAGREGÁCIA HUSTOTY ZAĽUDNENIA VYUŽITÍM MÁP KRAJINNEJ POKRÝVKY A ÚDAJOV DIAĽKOVÉHO PRIESKUMU ZEME

KONŠTANTÍN ROSINA
PAVOL HURBÁNEK

GEOGRAPHIA SLOVACA
31

GEOGRAPHIA SLOVACA

Geographia Slovaca je recenzovaný vedecký časopis, ktorý uverejňuje monografie a monotematické súbory príspevkov z rôznych oblastí geografie, environmentalistiky a regionálnej vedy. Geographia Slovaca vychádza v Geografickom ústave SAV od roku 1992.

Geographia Slovaca is a reviewed journal bringing monographs and monothematic sets of contributions from all geographical sciences including evnironmental and regional sciences. Geographia Slovaca is issued at the Institute of Geography, SAS, since 1992.

Redakcia / Editorial Office
Geografický ústav SAV, Štefánikova 49, 814 73 Bratislava, Slovenská republika (Slovak Republic), tel: +421 2 5751 0210, e-mail: geogira@savba.sk

GEOGRAPHIA SLOVACA 31

Konštantín Rosina[1], Pavol Hurbánek[2,1]

Spatial Disaggregation of Population Density Using Land Cover and Remote Sensing Data / *Priestorová dezagregácia hustoty zaľudnenia s využitím máp krajinnej pokrývky a údajov diaľkového prieskumu Zeme*

[1] Geografický ústav Slovenskej akadémie vied, Štefánikova 49, 814 73 Bratislava / Institute of Geography, Slovak Academy of Sciences, Štefánikova 49, 814 73 Bratislava

[2] Katedra geografie, Pedagogická fakulta, Katolícka univerzita v Ružomberku, Hrabovská cesta 1, 034 01 Ružomberok / Department of Geography, Faculty of Education, Catholic University in Ružomberok, Hrabovská cesta 1, 034 01 Ružomberok

Recenzenti / Reviewers:
Prof. RNDr. Jaroslav Hofierka, PhD.
Doc. RNDr. Dagmar Kusendová, PhD.

Tlač / Print: Romulus Beňo JUNIOR PRESS, Častá
Náklad / Edition: 100 výtlačkov / 100 copies
Grafická úprava / Graphic design: Konštantín Rosina
Jazyková úprava / Language editing: Resumé upravila Rút Facunová / English text was not edited

ISSN: 1210-3519
ISBN: 978-80-89548-02-6

GEOGRAPHIA SLOVACA

31 – 2016

GEOGRAFICKÝ ÚSTAV
SLOVENSKEJ AKADÉMIE VIED
BRATISLAVA

GEOGRAPHIA SLOVACA

31 – 2016

Konštantín Rosina, Pavol Hurbánek

SPATIAL DISAGGREGATION OF POPULATION DENSITY USING LAND COVER AND REMOTE SENSING DATA

Priestorová dezagregácia hustoty zaľudnenia s využitím máp krajinnej pokrývky a údajov diaľkového prieskumu Zeme

GEOGRAFICKÝ ÚSTAV
SLOVENSKEJ AKADÉMIE VIED
BRATISLAVA

CONTENTS

OBSAH

LIST OF ACRONYMS

ZOZNAM SKRATIEK

AIT – Austrian Institute of Technology

CLC – CORINE Land Cover

CORINE – Coordination of information on the environment

DEGURBA – Degree of urbanisation (classification of LAU2)

EEA – European Environment Agency

EFGS – European Forum for Geography and Statistics

EM – Expectation-Maximization

ENACT –

GIS – Geographic information system

GRUMP – Global Rural-Urban Mapping Project

GWR – Geographically weighted regression

HRIL – High resolution imperviousness layer

INSPIRE – Infrastructure for Spatial Information in the European Community

JRC – Joint Research Centre of the European Commission

LAU2 – Local administrative unit - Level 2

LULC – Land use / land cover

MAUP – Modifiable areal unit problem

MMU – Minimum mapping unit

NUTS – Nomenclature of territorial units for statistics

OECD – Organisation for Economic Co-operation and Development

OSM – OpenStreetMap project

PDIA – Population density per impervious area

RMSE – Root mean square error

RTEA – Relative total absolute error

TAE – Total absolute error

VGI – Volunteered geographic information

ACKNOWLEDGEMENT

This monograph is based on doctoral thesis of Konštantín Rosina of the same title, defended in August 2015 at the Department of regional geography, planning and environment, Faculty of Natural Sciences, Comenius University in Bratislava.

The thesis was an outcome of doctoral research carried out at the Institute of Geography, Slovak Academy of Sciences under the supervision of prof. Ján Oťaheľ. We would like to thank prof. Oťaheľ, the institute's directors prof. Ira and dr. Michniak and the whole collective of the institute's employees. Further thanks belong to the thesis opponents – assoc. prof. Marián Halás, prof. Jaroslav Hofierka and assoc. prof. Dagmar Kusendová. The latter two opponents also helped to improve this book by kindly reviewing the manuscript.

This monograph is derived, in part, from an article published in journal *Cartography and Geographic Information Science* on 16 Jun 2016, available online: http://www.tandfonline.com/doi/abs/10.1080/15230406.2016.1192487

This work was supported by the Slovak Scientific Grant Agency VEGA under grant no. 1/0275/13 *"Production, verification, and application of population and settlement spatial models based on European land monitoring services"*.

INTRODUCTION

One of the key goals of regional geographic inquiry is analysis and, ultimately, understanding of regional systems through their constituent components and interactions. Regardless of the preferred methodological apparatus used to achieve this goal (whether from the realm of qualitative or quantitative geography), the basic precondition of quality research are accurate and timely data, whose scale is appropriate for the studied phenomenon.

Human population is a basic component of regional systems that interacts intensely with other components, socio-economic as much as physical. To analyse these interactions, researchers often need to combine spatial data that are available on the basis of incompatible supports. Until recently, population data have been available almost exclusively aggregated into areal units (zones) that follow political, administrative or statistical definition; while usually unrelated to the distribution of population itself. However, researchers in geography and other spatially-aware fields of inquiry and practice (decision making, spatial planning, and emergency response among others), often need to analyze data using a set of zones that are incompatible with the administrative / statistical zones (e.g. functional urban regions, physical-geographical regions, service areas, or ad hoc analytical zones generated by a GIS such as buffers, intersections and viewsheds).

Furthermore, the spatial resolution of the finest available administrative zones (e.g. municipalities) is insufficient for many applications and these highly irregular zones, hindered by the modifiable areal unit problem, fail to capture the true nature of various settlement systems and can lead to spurious conclusions about population distribution and its interactions. Fortunately, methods have been developed to facilitate better inferences between incompatible zonal systems and enhancement of spatial resolution. With the advent of fast computers and user-friendly GIS software, the implementation of these methods has become increasingly effortless. Though, geographers and other researchers in Slovakia (and elsewhere) have been mostly hesitating to make use of these methods, which have the potential to boost value and relevance of any spatially explicit research.

The inference from a set of zones with known values (source zones) to an incongruent set of zones with missing values (target zones) is known as the areal interpolation problem (Goodchild and Lam 1980). Qiu and Cromley (2013, p. 213) identify four distinct problems that can be addressed: "Areal interpolation is needed to estimate attribute information for different geographic partitionings at the same scale (the alternative geography problem), for spatial partitionings at a finer resolution (the small area problem), for reconciling boundary changes in spatial units over time (the temporal mismatch problem), or for incomplete coverage (the missing data problem)."

In this study, we focus on the small area problem, also known as, and further referred to as spatial disaggregation. The goal of spatial disaggregation is to enhance the scale (sometimes the term downscaling is used) of the source data, typically based on some related ancillary variable whose spatial distribution is known at a finer scale. We do not intend to devise a brand new method to achieve this, but rather improve upon pre-existent method. The results should be relevant for Slovakia as well as in wider European context. Because of the large extent of the study area, the ancillary data need to be available free of charge and internationally comparable (spatially homogeneous in terms of definition, accuracy, spatial detail, etc.). Expected contributions of this study include new knowledge of factors that influence the accuracy of a chosen disaggregation method, applicable in Slovakia and other EU countries; as well as production of a raster population model based on the method.

1 BACKGROUND AND THEORY

1.1 Rationale and motivation

Limitations of aggregated population data

The arbitrary nature of administrative/statistical reporting zones

Many authors have criticised arbitrariness of administrative/statistical zones traditionally used for dissemination of census and other socio-economic data (Bracken and Martin 1989, Goodchild et al. 1993, Langford and Unwin 1994, Mennis 2003). Of course, there may be various reasons behind particular zone delineation – historical, political, physical or practical (such as efficiency of enumeration). However they often have unknown relation to the data they represent.

Unlike many examples of physical areal units (landforms, soil types, watersheds), the boundaries are usually not based on real discontinuities (natural breaks) in the spatial distribution of the phenomenon (Martin 1989). When studying residential population density, optimal zone boundaries should follow steepest slopes of underlying population density surface, thus the zones might correspond to various types of residential land use categories (non-residential land, neighbourhoods of detached housing, terraced housing, apartment buildings, and so forth). In reality, administrative/statistical zones such as counties and municipalities contain agricultural land, (semi)-natural vegetation, water, residential areas as well as various types of non-residential developed land. Thus, the resulting value of average population density in such zone can be misleading and is rather a function of the particular boundary location than of actual spatial distribution of people (Figure 1).

Further issues are caused by uneven size and irregular shape of the zones (Langford and Unwin 1994). Larger zones tend to provide lower values of population density, and on the other hand, as the area is reduced, both the peak values and the variance tend to increase. The zones act as low-pass spatial filters that generalise any high-frequency fluctuations within them. The larger the size of a zone, the stronger is the generalisation effect. If the zones have irregular shapes, the effect is also anisotropic

and difficult to quantify. For instance, the municipalities of Slovakia are very heterogeneous in size – the largest one is thousand times larger compared to the smallest one, but nevertheless they are incorrectly considered as equivalent units.

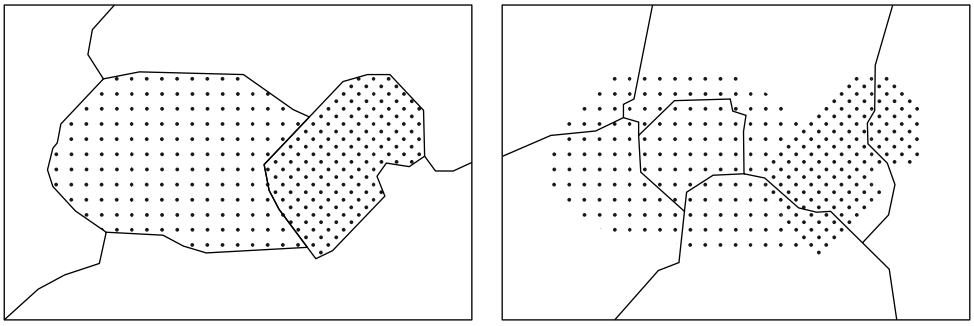


Figure 1: An example of zones that are (left) and are not (right) related to the actual spatial distribution of the phenomenon (the dots)

The modifiable areal unit problem and ecological fallacy

The use of spatially aggregated data (disregarding the level of arbitrariness of the zonal boundaries) for analysis and visual presentation of data inevitably faces two related issues – the modifiable areal unit problem (Gehlke and Biehl 1934, Openshaw and Taylor 1981, Openshaw 1983) and ecological fallacy (Openshaw 1984).

The modifiable areal unit problem (MAUP) is a situation when the result of statistical analysis or hypothesis testing can be influenced by a change in the set of boundaries used for aggregation of the analysed data. MAUP implies that statistics, relationships and patterns obtained from aggregated data are valid only for the particular mode of aggregation and cannot be generalised (Martin, 1996).

Green and Flowerdew (1996) point out two component problems of the MAUP – scale problem and zonation problem (the zonation problem is also referred to as aggregation problem by Openshaw, 1983). The scale problem is related to the variation of spatial analysis results depending on the level of spatial aggregation. For example, in correlation analysis, the strength of correlation typically increases with increasing level of aggregation (increasing size and decreasing number of zones). The zonation problem occurs when various modes of zonation at roughly the same aggregation level (the size and number of zones remain the same) produce different analysis results. Taylor (1977) suggested that the existence of the mentioned effects is given by spatial autocorrelation in the studied variables, which was demonstrated in later work of Taylor and Openshaw (1979). Green and Flowerdew (1996) further clarify that positive spatial autocorrelation is intrinsic for most socio-economic variables and therefore

the optimal level of aggregation is 'none', i.e. spatial analysis should be performed on individual data (which is, however, seldom possible).

Bezák (2008) summarised several works focused on the analytical solution of MAUP (Arbia 1989, Steel and Holt 1996), but he claimed that a final solution was not found. However, some authors (Fotheringham and Wong 1991) point at measures that can help to generalise analysis results (and to decide whether to accept or reject them). Using modern computers, numerous different zonations of study area can be created as a spatial framework for intended quantitative method. Consequently, the stability of any partial result can be evaluated by comparison of results for all automatically generated zonal systems.

A related problem to the scale problem in MAUP is the problem of cross-scale inference (Goodchild 2011) and its extreme variant – the ecological fallacy, sometimes described as aggregation bias (King 1997). It is an incorrect way of inference when properties and correlations found at the aggregate level are applied directly also to individuals within these aggregates (or to finer scale aggregate units). If cross-scale inference cannot be avoided (e.g. cases when the individual data are not available) researchers should not infer across scales directly, but rather estimate the target properties and relations by statistical modelling. King (1997) in his comprehensive monograph provides probably the most robust statistical framework for appropriate ecological inference as well as assessment of related uncertainty rates. Tapp (2010) notes that ecological fallacy is often committed by unaware users of choropleth maps, in which the portrayed zones appear to be homogeneous, i.e. that all places within a zone are similar to the zone's and that the boundaries reflect actual changes in the spatial distribution (which is usually not true for administrative/statistical units).

Other considerations

An interesting theoretical argument for favouring regular grids over conventional zones for socio-economic data is related to conceptualization of scale in geographical information science. While understanding of scale in analogue maps is intuitive, in digital representation it poses a problem. Scale in digital representation is usually defined as the ratio 'large over small', i.e. the ratio of extent (which can be easily derived from the data) and spatial resolution of a particular map. In raster (grid) representation the spatial resolution is given explicitly by its cell size, but in vector representation it is very vaguely defined and problematic to infer from the data. Therefore raster data should be preferred in rigorous research (Goodchild 2011).

Further benefits of gridded representation of data over conventional zones are of practical nature. Boundaries of administrative/statistical zones are complex polygons which may be created, distributed, and used at various levels of cartographic generalisation. Also, the zones change over time – there are unions, divisions, and

small shifts of the boundaries which may occur frequently and are prone to be overlooked. Consequently, when two researchers claim to have used the same set of zones, it is not always certain that they have used precisely the same digital representation of their boundaries. Maintenance of such vector databases – especially over larger territories – is costly, which can make spatial data that match temporally and spatially with the attribute data too expensive or even impossible to obtain in some countries. For instance, a yearly license of EU-wide EuroBoundaryMap for small organisation costs €6,000, which is restrictive for many researchers (EuroGeographics 2015).

On the other hand, regular grid cells are defined by a simple rule persistent over time (based on the rule, anyone is enabled to generate the zonal boundaries for free), independent of external factors (environmental, societal, political, administrative transformations) thus enabling for longitudinal studies without need for tedious comparison and harmonisation of the changing administrative/statistical units. Data can be stored and manipulated conveniently in raster formats and analysed using GIS tools that would not be applicable to vector formats. If the cells are relatively small, they can be used to approximate arbitrary irregular zones.

Spatial aggregation into regular grids

We recognise two approaches that fully avoid the above-described issues (including the MAUP). The first approach is the use of individual data, i.e. point-based observations not aggregated into zonal framework. That is however rarely possible – either the data from censuses are not precisely geo-referenced (only code of the respective enumeration zone is known), or such information is restricted due to confidentiality measures. Also, many variables are intrinsically defined on the basis of areal support (e.g. gross domestic product); hence it makes no sense to measure them on a point location (Martin 1996).

Another solution is modelling of variables as continuous fields (surfaces), which is possible only with spatially intensive variables, e.g. population density as opposed to population count (Goodchild et al. 2007). Continuous surfaces are independent of any zonal system and are very favourable for a visual representation of spatial structures. A disadvantage of this approach is uncertainty – the surfaces are usually imperfect models of reality (based on estimation) and therefore values derived from such surface are approximate (in contrast to actual precise counts that can be recorded by a census in a discrete zonal system). Modelling of continuous surfaces is further discussed in section 'Spatial conceptualization of population density.'

A partial solution to the outlined issues that preserves the 'ground-truth' quality of the conventional zonal data is to aggregate individual records into regular grids (so-called bottom-up approach to population grids; Figure 2). Although this approach does not avoid the MAUP-related issues (since they are inherent in all

spatially aggregate data), it is much more suitable for rigorous spatial research as the zones are regular and comparable. Furthermore, the grid cells are usually smaller than the average size of the smallest available conventional zones, allowing researchers to draw a much sharper image of the spatial patterns as well as approximate arbitrarily shaped zones required for the diversity of topics in geography. Due to the confidentiality concerns, there is a certain lower bound of grid cell size for bottom-up data, typically increasing with specificity of the variable. For instance, Statistics Austria (2012) provides general population count at a resolution as fine as 100 m, while counts by sex or age group are provided only for 250 m and coarser grids.

For some socio-economic variables, it is not possible to produce actual data on the basis of fine resolution grids, because they are collected or defined based on relatively large and self-contained spatial units. For instance gross domestic product is defined using an economy operating at the level of a country or a (functional) region, thus only top-down (estimation-based) approach is feasible. An isolated example is the 'Gross Cell Product' dataset with large 1° by 1° cells[1], produced within G-Econ project at Yale University (Nordhaus 2006).

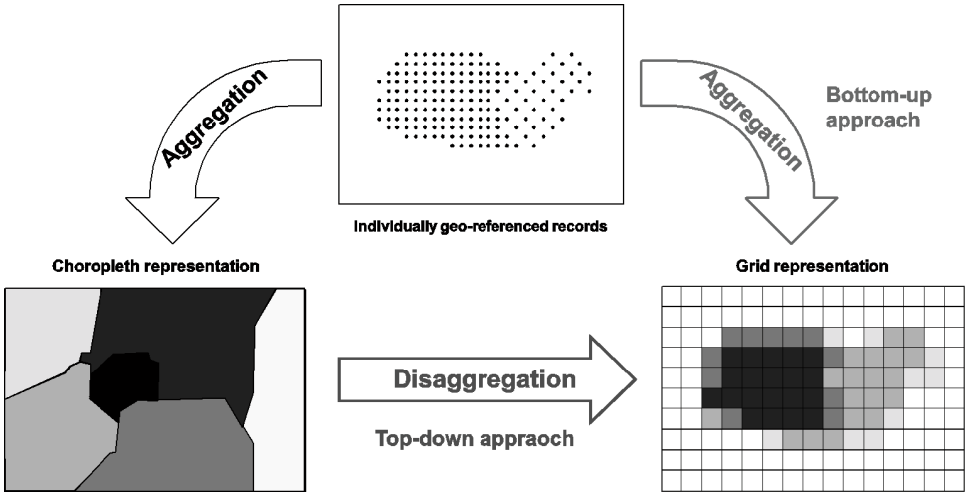


Figure 2: Illustration of disaggregation and aggregation processes

1 Gridded datasets with global coverage are often based on geographic coordinate systems (latitude-longitude), thus the cells of the grid are actually quadrilaterals with decreasing area from the equator to the poles rather than equally sized squares.

Current state of production and dissemination of gridded population data in the EU

Transition to gridded population data with finer spatial resolution is supported by Directive 2007/2/EC of the European Parliament and of the Council establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Annex I of the Directive specifies 'geographical grid systems' as one of the main themes of spatial data. Geographical grid systems are defined as a harmonised grid with multi-level resolution and common point of origin and standardised offset and size of cells. Annex III of the directive defines a spatial data theme "Population distribution – demography" as "geographical distribution of people, including population characteristics and activity levels, aggregated by grid, region, administrative unit or other analytical unit" (Official Journal of the European Union 2007).

Coordination of efforts to promote regular grids as standard spatial units for population data dissemination was led by the European Forum for Geography and Statistics[2] (EFGS). It started in 1998 as a voluntary cooperation between national statistical institutions of the Nordic countries, which pioneered the production of gridded census data. Currently, the forum has the ambition to transform from a European to a global organisation.

Eurostat initiated the GEOSTAT project aiming to produce a seamless EU-wide grid of total population counts with 1km resolution. The first phase of the project was named GEOSTAT 1A, with a reference year of 2006. The resulting grid was composed of three types of data:

a) Actual population counts aggregated from geo-referenced census records (Austria, Slovenia, Finland, Sweden and Netherlands)

b) Estimates based on detailed national data such as building, address points (Estonia, Portugal, France, Norway, Poland, England and Wales)

c) Data disaggregated from LAU2 zones (second level local administrative units, e.g. municipalities) by Austrian Institute of Technology (the AIT grid) were included for the rest of EU countries.

Slovakia belonged to the third group of countries and, according to the final report of GEOSTAT 1A, Slovakia did not pursue geo-referencing of census records in the last round of decennial censuses in 2010/2011 (EFGS, 2012, Fig. 1). Following phases of the GEOSTAT project – 1B and 1C drew upon extended number of countries that produced aggregated 1 km grid from 2010 / 2011 census data. 1B dataset is available for download and contains aggregated data from 18 countries and disaggregated

2 Formerly known as European Forum GeoStatistics, it was renamed in 2015 following frequent confusion with the scientific field of geostatistics, branch of statistics focused on prediction of spatiotemporal distributions (originally for geological and mining operations), based on the works on G. Matheron and D.G. Krige.

(modelled) data from 14 countries (including Slovakia). The GEOSTAT 1C dataset should includes also aggregated data from Slovakia (based on aggregation using geo-referenced address points) and additional variables (population breakdowns by age groups and sex).

Relevance of the disaggregation approach

Considering some of the issues raised in the section 'Limitations of aggregated population data', it seems reasonable to prefer estimated (modelled) data over real data. Without access to individual geo-referenced records, it is the only way how to cope with some of the mentioned problems. Although disaggregated data are estimates and lack the quality of being true or 'correct' (the main advantage of conventional census data), the modelled values may contain more useful geographical information for many purposes (Martin 1996).

Despite the incentives from INSPIRE directive, EFGS and Eurostat, several countries in Europe failed to produce aggregated population grids. As a result, the disaggregation remains the method of choice for some countries in the EU and even more so in non-EU countries in Europe and elsewhere. Furthermore, the current 1 km target cell size for aggregated data is too coarse for some applications, and can be seen as a starting point for further disaggregation (e.g. into 100 m or smaller grid cells) using detailed national-level ancillary data such as building footprints. It is unlikely that aggregated data will be ever produced at such fine resolutions (finer than 100 m), because of privacy concerns as well as positional inaccuracy.

Disaggregation approach also allows for the production of gridded data for past censuses when the records were not geo-referenced yet (provided that suitable ancillary data matching the respective year are available). A further advantage of disaggregation and modelling of gridded population data is the possibility of alternative temporal conceptualizations of population density. While aggregated census counts usually report the location of people in their homes (approximately night-time distribution), disaggregation methods allow for modelling of day-time or ambient population distribution (consult the section 'Temporal conceptualization of population density').

A prominent journal Geographical Analysis published a special issue on areal interpolation and dasymetric modelling in July 2013, confirming that these problems are still hot research topics, despite increasing adoption of bottom-up approaches to the production of detailed gridded data.

Application of gridded population data

Gridded population data have the potential to improve any research, planning or decision-making task that concerns population (i.e. very wide range of issues), and to facilitate studies that combine data from the realms of physical and human geography,

as well as other disciplines. The main benefits are improvement of spatial resolution at which an analysis is performed, the possibility to approximate arbitrarily shaped areal units relevant for the concerned topic, and novel raster-based data processing and analysis capabilities. Spatially more accurate estimation of population distribution is instrumental for better understanding of the interactions between population and social, economic and environmental conditions (Lu, Weng, and Li 2006). It also allows for more equitable resource allocation in infrastructure development and may improve accessibility to open space, recreation, transportation, and environmental facilities (Maantay et al. 2007). Daytime and ambient population models are particularly valuable in hazard planning and response (e.g. flood risk and vulnerability assessment), environmental impact assessment, transportation planning, and location of ambulance and fire stations. (Sutton et al. 2003).

Importantly, population density grids have been employed in policy making such as definition and characterization of metropolitan areas in OECD countries (OECD 2012, Piacentini and Rosina 2012), update of rural-urban typology for NUTS regions in the EU (Eurostat 2014), and the new DEGURBA (degree of urbanization) classification of LAU2 units (Eurostat 2012; Kusendová, Rosina, and Hurbánek 2014). These typologies and classifications are used by many policies and have a potentially far-reaching impact.

Kim and Yao (2010) refer to the application in location of business development and public health (Hay et al. 2005, Langford and Higgs 2006); as well as homeland security issues such as emergency contingency plans (Dobson et al. 2003, Garb et al. 2007). Gallego et al. (2011) point to studies measuring the exposure to noise around airports (Vinkx and Visée 2008); quantifying the impact of population density on forest fire risk (Barbosa et al. 2008); or computing indicators of accessibility to services in rural areas (Dijkstra and Poelman 2008).

1.2 Spatial disaggregation in the literature

In this section, we provide a basic theoretical body of knowledge that – we suppose – is needed to support further reading. The topics range from the theory of geographical representation, spatial and temporal conceptualizations of population density, aspects of disaggregation accuracy, as well as various approaches to areal interpolation and spatial disaggregation that were published to date.

Spatially intensive and extensive variables

In spatial disaggregation (as well as in areal interpolation generally) it is important to distinguish between spatially intensive and extensive socio-economic variable. Goodchild and Lam (1980, p. 298) characterise the two types as follows. A spatially extensive variable takes half the region's value in each half of a region (such

as population count and gross income). A spatially intensive variable is expected to have the same value in each part of a region as in the whole (such as average income or share of the male population). To every spatially extensive statistic, there corresponds a density function which is obtained by dividing by area. Thus population density is obtained by dividing population by area, and yields population when integrated over area. In other words, extensive variables express absolute quantities or counts, intensive variables express rates or ratios. Extensive and intensive variables also differ in suitable forms of their geographical representation, as clarified by general theory of geographical representation in GIS as proposed by Goodchild et al. (2007).

Theory of geographical representation in the context of spatial disaggregation

Goodchild et al. (1999) describe the geo-atom as the atomic form of geographic information since all geographic information can be reduced to it. It is composed of a point in space-time (given by its coordinates), a property, and a particular value of that property at that point. Goodchild et al. (2007) define two basic higher forms of geographic representation – geo-fields (continuous surfaces) and geo-objects (discrete entities).

Geo-field is defined as an aggregation of all geo-atoms in a spatiotemporal domain that has some property, irrespective of its particular value. It can be thought of as a function of geographic location – in each location a single value can be obtained. In GIS practise geo-field is often referred to as a surface or coverage. In principle, a geo-field consists of an infinite number of point locations (geo-atoms). Consequently, it is necessary to sample, select or approximate so that the field can represented by a file of finite size (unless the geo-field can be represented by mathematical function). Such simplified representations of a geo-field are known as discretizations, and there are six types of them. Two are based on a selected finite sample of geo-atoms from the geo-field, and four are based on discrete geo-objects.

Geo-object (discrete entity or feature) is defined as aggregation of all geo-atoms in a spatiotemporal domain that meet certain criterion, which typically is having a specific value (or range of values) of some property, i.e. a criterion of internal homogeneity. Also, geo-objects can be defined by rules related to spatial relations and flows, e.g. nodal regions (all locations with the same primary commuting destination). In case of geo-object representation, values of spatially intensive variables are assumed to be relatively uniform throughout each geo-object. The values are not varying continuously over space but step-wise, on the boundaries of geo-objects. Besides the intensive variables that are assumed to be uniform within each geo-object, new properties of the geo-objects emerge – such as length, area and volume, as well as integrals of the intensive variables (e.g. population count integrated from population density surface). These new properties are termed spatially extensive (Longley et al. 2005).

An important implication of the outlined theoretical framework is the notion that spatially extensive variables cannot be conceptualized as a continuous geo-field. This is caused by the fact that values of extensive variables cannot be obtained for a single geo-atom, and it makes sense only to define them on the basis of certain area, over which the underlying density surface is integrated. On the other hand, spatially intensive variables are measurable at a point location of an individual geo-atom, thus the geo-field representation is possible. Both intensive and extensive variables (population count and population density) can be represented by a regular grid. In case of population count, each grid cell simply represents a 2D object – a square zone, over which underlying population density surface is integrated (or number of individuals inside the zone is counted); however, in case of population density, the grid cells can be interpreted as two distinct types of geo-field discretization:

a) Sampled discretization at a set of points in a regular array – the cells substitute 0D point locations (typically cell centre) – a finite sample from the infinite number of geo-atoms constituting the geo-field, selected in a regular rectangular fashion. The cell value represents value of the function in the location of cell centre; while the actual 'support' needs not to be identical with the cell (it can be larger or smaller).

b) Piecewise constant discretization in a regular grid of cells – the cells represent 2D objects (similarly to extensive variables), in which the property is assumed to be uniform, i.e. the cell value represents average population density within the bounds of the particular cell.

Spatial conceptualization of population density

Some spatially intensive variables such as air temperature and elevation can be measured at point location (if we abstract from the dimensions of the sensor itself). With other intensive variables (such as population density), the point measurement seems impossible – hypothetically, at point location, there can be either zero or one person, but since the point has zero area, division by zero would occur. Nevertheless, certain assumptions about population density can be made at any place of the Earth's surface. To do that, one has to consider the context or local neighbourhood of the respective point to capture the phenomenon of population distribution in space.

The principle of local neighbourhood is used by kernel density estimation (Silverman 1998; Gallay, Kaňuk, and Hofierka 2014) – a moving window is placed at each location and density value is calculated from the number of persons found in the window and known area of the window. Additionally, each person can be weighted depending on their distance from the current location – based on a kernel function. The obtained result is a continuous population density surface or a two 2D function of the location. This conceptualization of population density does not allow for abrupt changes in the surface; e.g. a point found on a boundary between densely populated and uninhabited area yields intermediate density, since both types of area are found

in its vicinity. Thus a smooth transition is modelled (Figure 3). Similar approach was advocated by Nordbeck and Rystedt (1970), who argue that it is logically correct to represent population density by isopleth mapping technique (discretization of the surface by sampling along isolines). With the advent of GIS and raster data structures, point-sampled discretizations (on a regular grid) of population density surfaces prevailed in geographic studies (Tobler 1979, Bracken and Martin 1989, Langford and Unwin 1994, Thurstain-Goodwin and Unwin 2000, Kim and Yao 2010).

The smoothness of a surface resulting from kernel density estimation is largely dependent on the particular kernel function shape and bandwidth. Therefore identification of appropriate kernel parameters is necessary. Kim and Yao (2010) explain that the kernel bandwidth should be considered separately for each study area and that a uniform bandwidth is not always suitable for the whole study area. Locally adaptive kernel bandwidth was also employed by Bracken and Martin (1989), and Martin (1989). Common kernel functions include a uniform (unweighted moving window), Gaussian, Epanachnikov, quartic, cosine, and others.

The second approach to spatial conceptualization of population density is zonal or piecewise approximation. The study area is completely partitioned into zones, in which population density is assumed to be relatively homogeneous, and zone boundaries should coincide with steepest 'slopes' of the underlying density surface. Choropleth representation is an example of the zonal approach, although when administrative zones are used, the homogeneity assumption is typically far removed from reality. A more appropriate implementation of the zonal approach is so-called dasymetric mapping, coined by Russian geographer Semyonov-Tyan-Shansky and later popularised by Wright (1936). The main principle of this technique is the design of the homogeneous choropleth zones, typically derived from ancillary spatial data correlated with the distribution of the target variable (such as polygons of residential and non-residential land use in the case of population density). Consequently, values are estimated for the dasymetric zones. Frequently used ancillary data are ad-hoc classified satellite images (Fisher and Langford 1996, Mennis 2003, Kim and Yao 2010) and pre-existent maps of land cover/land use derived from remote sensing data (Reibel and Agrawal 2007, Gallego 2010). Zonal approach creates discontinuous, locally constant function (Figure 3). The value of density at each point is obtained from supports of various sizes and location of the point inside the respective areal (as opposed to kernel approach, where density at each point is usually obtained from supports of uniform size centred on each point). A disadvantage is that delineation of the dasymetric zones where density is assumed to be homogeneous can be arbitrary and as noted by Langford and Unwin (1994), the scale at which the dasymetric zones should be mapped is unclear. In practice, this is often influenced simply by (un)-availability of pre-existing ancillary data rather than careful consideration.

Since the underlying function in the zonal approach is piecewise, a dasymetric map can be represented in GIS by geo-objects (similarly to conventional choropleths).

However, if the number of dasymetric zones is high, it is often more practical and faster to use raster representation.

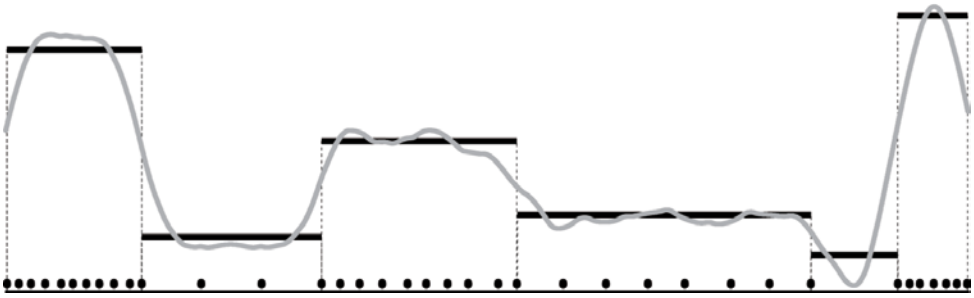


Figure 3: A piecewise density function (black) and smooth kernel-based function (grey). The black dots represent the distribution of individual objects, density of which is measured in 1D space.

Temporal conceptualization of population density

Population censuses and registers, the source of most accurate data about population distribution record location of people in their permanent or usual residences[3], thus they can capture only a fraction of overall spatial population dynamics related to the change of residence, i.e. migration. We further refer to the population present in permanent homes as residential population, and to the corresponding population density function as residential population density. While the description of residential population density is required (or at least sufficient) for some applications, others demand more fitting models of reality. Such applications include disaster management, emergency response, public health, population exposure assessment, design of telecommunication networks, and so forth.

Residential population distribution is entirely hypothetical. In reality, there is never coincidence with actual population distribution, since the population is subject to temporary movements, either cyclical (e.g. daily commuting) or turbulent (tourism). Actual population distribution fluctuates over various parts of a day, week and year. Daily commuting to work and schools is the dominant type of short-term dynamics since most people take part in it regularly. Most employees and students commute from home to workplace/school in the morning and return in the afternoon/evening (excluding the unemployed, teleworkers, retired, and night workers). This basic pattern is accompanied by less predictable movement to shopping, leisure, and other activities.

Distribution of people at night is the closest to the residential model (although far from identical). The difference between daytime and night-time distribution is most

3 Sometimes the location at the moment of census is recorded (e.g. in the GEOSTAT data).

evident in single-function parts of cities. Purely residential areas tend to be almost empty in the daytime, then density increases gradually over the course of afternoon/evening, and peaks in the night-time. An opposite pattern can be observed in areas of industrial, commercial or educational facilities. An extreme example of the contrast between daytime and night-time population density are central business districts (see Bhaduri et al. 2007, p. 109, fig. 2).

Discerning between daytime and night-time population is the first step in the transition from residential models to more realistic models (e.g. those described by McPherson et al. 2004, and Bhaduri et al. 2007). Though in theory, even more sophisticated dynamic spatiotemporal models of the population can be considered, that would enable users to obtain estimates for any temporal and spatial domain (taking into account also weekends, bank holidays, school terms and holidays). A similar approach is discussed by Ahola et al. 2007 and is pursued by a project Population 24/7 lead by David Martin at University of Southampton (Economic and Social Research Council, 2012).

Currently, an ongoing exploratory project named ENACT (Enhancing activity and population mapping) is pursued at the Joint Research Centre of the European Commission (JRC). The objective is to develop and implement a consistent and validated methodology to produce day-time and seasonal population distribution grids for Europe (Batista e Silva et al. 2015).

A different theoretical concept of population density was proposed by Dobson et al. (2000), and Sutton et al. (2003). So-called ambient population density refers to average population density at a location measured over certain time-frame (a period that is sufficiently long, so that the average occurrence of people at the particular location is stabilised). Thus, in a single figure, various types of population dynamics are indirectly captured. In an ambient population model, certain non-zero density should also be attributed to road networks, squares, public greenery and any other types of surfaces where occurrence of people is notable at least for some time periods.

Implementation of other than residential population models is exceptionally demanding in terms of input data. Demographic data of high spatial and thematic detail, geodata at the building level including footprints and attribute information on function and height, commuting flows, and other data are necessary. Furthermore, advanced algorithms and powerful computers for data processing are needed to process the data. As a result, daytime or ambient population models are still relatively rare.

Accuracy of spatial disaggregation and its assessment

All areal interpolation algorithms, either simple or intelligent, are based upon underlying assumptions regarding the likely distribution of population within the source zones. The accuracy of the interpolation is always a reflection of the validity of the assumptions made by a particular method; e.g. uniform distribution in areal

weighting, distance decay in kernel estimations or maximum smoothness in Tobler's smooth interpolation (Goodchild and Lam 1980, Langford 2013).

In intelligent areal interpolation, the assumptions about the relationship between the disaggregated and the ancillary data, as well as appropriateness, accuracy, and spatial detail of the ancillary data are the major factors influencing the accuracy of the result. According to some researchers, their importance exceeds that of particular algorithmic details of disaggregation method used (Sadahiro 1999; Martin et al. 2000).

Sadahiro (1999) identifies two aspects of areal interpolation accuracy – overall magnitude of the error and uniformity of its spatial distribution. Obviously, it is desirable to minimise both the sum of absolute errors for the whole study area as well as spatial variation of relative errors. The latter is however quite difficult to achieve if the source zones are of heterogeneous size (e.g. municipalities) since the accuracy is largely influenced by the size difference between source and target zones. The most detailed level of administrative/ statistical zones available is typically favoured in order to maximise the overall accuracy. To reduce the spatial variation of relative accuracy, we suggest that only datasets that cover the whole study area and have harmonised definition and consistent quality should be used as ancillary data (since the accuracy of ancillary data largely influences the disaggregation accuracy). On the other hand, some researchers may find more desirable to incorporate any available datasets despite covering the study area only partly, in an effort to increase overall accuracy even at the expense of spatial homogeneity (e.g. Batista e Silva et al. 2013).

Furthermore, a consistent model of population density requires that the integral of the surface over each of the source zones equals to the actual value for the respective zone. The requirement means that no people were 'lost', 'created' or 'relocated' between individual source zones. This principle is termed by Tobler (1979) as the pycnophylactic property/condition (pycnophylactic means volume-preserving). Most published disaggregation methods have this condition incorporated in their algorithm since it serves as a basic prerequisite of estimation accuracy. An interesting problem was faced by Martin (1989) when the boundaries of the source zones were unknown (only centroids with attributes were available), therefore it was not possible to verify the pycnophylactic condition.

Besides ensuring full agreement between the source data and modelled data at the level of source zones, it is also important to minimise deviations between actual and modelled data at the level of target zones (e.g. dasymetric zones or individual cells of a grid). Assessment of these deviations requires a set of reference values at the target zone level; though these are typically not available (if they were, the disaggregation would be unnecessary). A workaround for this problem is to use the second most detailed set of available census zones as the source data and retain the most detailed set for validation and calibration of the model. For instance Langford (2006) used larger wards as the source zones and preserved basic census output areas for validation; Rosina et al. (2012) disaggregated population data from the level of municipalities to a 100 meters

grid and validated the results against sub-municipal basic settlement units. Although the use of larger source zones can increase the overall estimation error, it is often the only way to quantify the disaggregation accuracy. The limitation of such approach is that obtained accuracy measures are valid only for the particular combination of source and validation zones and can be generalised.

A more robust solution of the problem was proposed by Fisher and Langford (1995), who assessed the accuracy of different disaggregation methods. Monte Carlo simulations were used to build 250 realisations of higher level zones from the set of smallest available census zones, so that the actual values could be easily summed up for each zones of the 250 zonal systems. Subsequently, pairs of the realisations were used as source and target zones in the tested areal interpolation methods. Conclusions drawn from such spatial framework have much broader validity than those obtained from analysis of a single configuration. Such approach could also be applied to disaggregation into grid cells. Pairs of random zonal systems would be used as source zones and validation zones, while keeping the same set of target zones (the grid cells), thus allowing for generally more valid comparison of various disaggregation methods and/or sets of ancillary data.

A coarser bottom-up grid can be used (if available) to validate fine resolution disaggregated grids as long as they are nested (e.g. disaggregated data into 100 m EEA reference grid could be validated by data aggregated into 1 km EEA reference grid such as those from GEOSTAT project). It is noteworthy that the measured accuracy increases with decreasing spatial resolution of validation data. This scale-dependency naturally stems from the fact that population distribution is positively spatially auto-correlated. Therefore, unless the disaggregated product has a specific scale of application, it is preferable to perform validation based on cells of increasing sizes built from the most detailed bottom-up cells (e.g. 1 km, 2 km, 4 km, 8 km, 16 km, etc.) and construct a multi-scale accuracy profile.

A separate issue is validation of population distribution models with alternative temporal conceptualization (daytime, ambient); since actual reference data are virtually non-existent. A possible avenue for future evaluation of these models (although not perfect) might be a comparison with data about the distribution of mobile phones that can be collected by mobile service providers by individual zones of the cellular network, which are getting smaller with increasing density of base transceiver stations. Thanks to the high mobile phone penetration (over 100%) in most European countries, the figures can also be reasonably representative. An obstacle to this approach is the unwillingness of service providers to release the data, as well as the huge computational demand of the data processing. Nevertheless, first studies using mobile data for population distribution modelling have already been undertaken in Italy (Ratti et al. 2006) and Estonia (Ahas et al. 2007, Ahas et al. 2008).

Overview of spatial disaggregation methods

To date, several summary studies evaluating and comparing spatial disaggregation methods were published (including related or synonymous subjects such as areal interpolation, change of support, downscaling, dasymetric mapping, population surface modelling, etc.). In Slovak provenience isolated are studies of Madajová (2010a, 2010b), although they do not focus exactly on disaggregation into fine resolution grids, but generally on areal interpolation between two incompatible zonal systems (statistical districts and functional urban regions in this case). A study by Sadahiro (1999) compared the estimation accuracy of three simple and one intelligent (dasymetric) method. Wu et al. (2005) provide an overview and classification system of population estimation methods. Two basic groups of methods are areal interpolation methods (drawing on actual source population data), and statistical modelling methods, which model population distribution based on other variables and actual population data are not directly used as a source of the estimation (though they can be used to calibrate the model). Areal interpolation methods are further classified into two main groups according to whether they use ancillary spatial data or not. Simple methods do not use any additional data and are generally less accurate than the intelligent methods, which take advantage of ancillary spatial data. We adopt this basic classification of Wu et al. (2005).

Simple methods

One of the most basic simple methods is areal weighting, which assumes a homogeneous distribution of the interpolated variable within the source zones. This assumption is seldom correct, rendering the areal weighting very inaccurate in most situations. Nonetheless, this technique is useful as a baseline for comparison with more advanced methods. Another example of a simple method is pycnophylactic (volume preserving) smoothing as proposed by Tobler (1979), which assumes that the underlying density surface is maximally smooth and the source zone population totals (the volume beneath the smoothed surface within the bounds of each source zone) are preserved. This method was later demonstrated to be practically a variant of area-to-point kriging (Yoo et al. 2010), a geostatistical approach to simple areal interpolation devised by Kyriakidis (2004) and now implemented in ESRI ArcGIS software (from the version 10.1 on). The validity of the maximal surface smoothness assumption should be considered when using this method. Nevertheless, the pycnophylactic property asserted by Tobler became an inherent part of many areal interpolation methods. Martin (1989) proposed a different solution for population density surface modelling suitable for cases when the boundaries of the source zones are not known (hence it is impossible to apply the pycnophylactic property); instead, the locations of population-weighted centroids are known for each source zone. The surface is then modelled using a distance-decay function and an adaptive kernel width, assuming that the

population density tends to decrease with increasing distance from the population weighted centroid. Estimates for any set of target zones are then simply integrated from the modelled surface.

Intelligent methods

Many methods for intelligent areal interpolation have been proposed that assume there is a relationship between the interpolated variable and some other variable, whose spatial distribution is known at finer spatial resolution (the ancillary data). The key objective of intelligent methods is to determine and quantify this relationship. Some researchers (Martin et al. 2000, Eicher and Brewer 2001, Mennis and Hultgren 2006) use the term dasymetric in a broader sense, i.e. for any method, where the reported attribute is redistributed within each source zone based on information obtained from an external control (the ancillary data). However, we prefer to restrict the use of the term to interpolation based on the principles of dasymetric mapping.

Dasymetric areal interpolation typically employs categorical land use/land cover maps (LULC, usually derived by classification of remotely sensed imagery) as ancillary data, such as CORINE Land Cover (CLC) maps used by Gallego (2010). LULC classes (control zones) are intersected with source zones to form a set of dasymetric zones, which are assumed to have a homogeneous population distribution. The population counts are redistributed into dasymetric zones and re-aggregated (e.g. by areal weighting) into the target zones to obtain the predictions. The assumption of correlation between the distribution of LULC zones and the population is intuitive, and many studies show empirically that such data are an effective interpolation guide (Schroeder and Van Riper 2013). LULC ancillary data were used, for example, by Fisher and Langford (1995), Yuan et al. (1997), Holt et al. (2004), Langford (2006), Reibel and Agrawal (2007), Gallego (2010), Lin et al. (2011), and Cromley et al. (2012).

Other disaggregation approaches rely on different types of ancillary data. Some authors tried to find a link between population and reflectance values of pixels in multispectral satellite imagery or night-time lights satellite imagery (Iisaka and Hegedus 1982, Lo 1995, Harvey 2002); length of road segments (Xie 1995, Voss et al. 1999, Reibel and Buffalino 2005), count of address points (Tapp 2010, Zandbergen 2011); impervious surface coverage data (Lu, Weng, and Li 2006; Wu and Murray 2007; Zandbergen and Ignizio 2010); image texture (Liu, Clarke, and Herold 2006); or a combination of multiple types of data (Dobson et al. 2000, Bhaduri et al. 2002, Batista e Silva et al. 2013).

In geostatistics, a cokriging intelligent method for areal interpolation was proposed by Wu and Murray (2005), where impervious surface fraction derived from Landsat Thematic Mapper imagery is used as the covariate. Another use of geostatistics in intelligent disaggregation was presented by Liu, Kyriakidis, and Goodchild (2008). Residuals from traditional regression estimation were disaggregated by area-to-point

kriging. Geostatistical methods can provide more accurate estimates and, importantly, provide uncertainty measures at target zone level; unfortunately, the intense computational demand renders them impractical for larger study areas.

In intelligent methods, complex statistical methods often need to be employed to link the ancillary data to the variable to be interpolated (e.g. the expectation-maximization algorithm, geographically weighted regression, area-to-point kriging, etc.), this topic is further discussed in the section 3.4. Langford (2007) argues that the complexity of many intelligent areal interpolation methods discourages researchers and that the simplicity of the methods proposed in the future is the key to wider adoption of such methods in the broader community of GIS users. We suggest that another option is providing users with ready-made, fine resolution population grid, which can be relatively effortlessly applied in their analysis. Using simple GIS operations values for any arbitrary zones can be obtained from the grid, and there is no need for the users to understand all the complexities of areal interpolation methods.

A product that is helpful to a wide range of users should have a large extent (continental to global) and fine resolution (roughly 1 km to 100 m, so that it can be applied at various scales of analysis).To date, only a small number of such datasets has been published. To our knowledge, there are currently two global datasets available, both with spatial resolution of 30 arc seconds. Global Rural-Urban Mapping Project (GRUMP) is an open access dataset produced by Columbia University Center for International Earth Science Information Network in collaboration with the International Food Policy Research Institute, The World Bank, and Centro Internacional de Agricultura Tropical; reporting residential population density estimates for reference years 1990, 1995, 2000 (Balk et al. 2006). LandScan™ Global is a commercial dataset developed by Oak Ridge National Laboratory. It is an ambient population density model updated on a yearly basis. A more detailed version with three arc seconds resolution is produced for the USA (LandScan™ USA), including daytime and night-time models (Bhaduri et al. 2007).

Disaggregated population datasets covering the EU were produced by JRC and the Austrian Institute of Technology (AIT). JRC has been developing a disaggregated population grid since 2000; first five versions for the reference year 2000 were based on CLC2000 data (Gallego and Peedell 2001, Gallego 2010, Gallego et al. 2011). The latest version for the year 2006 (Batista e Silva et al. 2013; JRC grid 2006 hereinafter) was produced based on the special version of CLC2006 refined by data from High-resolution imperviousness layer (HRIL), Urban Atlas, Tele Atlas, and other datasets. (Batista e Silva et al. 2012). The AIT grid (aforementioned in 1.1.3) was produced with 1 km spatial resolution for the year 2006 using on HRIL, CLC, and OpenStreetMap (OSM) as ancillary data (Steinocher et al. 2011). Both AIT grid and JRC 2006 grid are suitable as benchmarks, which we seek to exceed in overall accuracy.

2 OBJECTIVES

In this study, we aim to improve upon existing method of intelligent spatial disaggregation by modification of the method and by combining multiple sources of free ancillary data (CLC2006, HRIL2006, and OSM), characterised in the section 3.2. The method should be applicable to the territory of Slovakia, though the intended experiments with the method are possible only in the presence of suitable validation data (such as 1 km bottom-up population grid) that would temporally match the available source, validation, and ancillary datasets. Since such data for Slovakia was not available over almost the entire course of this research (it was published only in February 2015, based on 2011 census), we have preferred Austria and Slovenia as a study area.

Specific objectives of the research

- To estimate residential population density of target grid cells (corresponding to cells of 100 m EEA reference grid, based on the ETRS89-LAEA[4] projection as defined in INSPIRE directive); the source population data were provided at the LAU2 level (local administrative units, which are named Gemeinden in Austria and Občine in Slovenia). Since we aim to improve upon existing method, we also expect that the result will exceed comparable population datasets in terms of overall accuracy.
- To improve existing method by optimisation of source zone subsets used in the iterative algorithm for estimation of population density coefficients. Rather than using contiguous non-overlapping subsets of source zones (i.e. regions), we define a subset for a source zone c as the m nearest source zones to the zone c. The distance is measured in an n-dimensional feature space, while values of the n features

4 Lambert Azimuthal Equal Area (LAEA) projection from European Terrestrial Reference System 1989

(attributes) for each source zone are obtained automatically from source and ancillary data.

- To study and find optimal parameters of the selected disaggregation algorithm by means of sensitivity study, and to determine if non-spatial neighbourhoods are more appropriate in the particular setting than spatial distance-based local neighbourhoods or traditional non-overlapping contiguous regions.
- To further increase the disaggregation accuracy by reduction of non-residential impervious land captured by HRIL using an OSM-based mask. More specifically, we aim to address attribute, temporal and completeness issues of OSM data; to propose a geoprocessing chain necessary to alter values of HRIL grid cells using OSM data properly and to quantify the accuracy improvement thanks to OSM data (and its spatial variation).

3 RESEARCH METHODOLOGY AND METHODS

3.1 Study area

Austria and Slovenia were selected as a suitable study area since they present a contiguous territory with available 1 km bottom-up grid for the year 2006 that is needed to validate and compare the performance of tested methods. Austria has on average three times smaller LAU2 compared to Slovenia (35 and 105 km^2, respectively). We anticipate that this fact might be reflected in the accuracy of the results. The two countries comprise ca 2549 LAU2 zones with a total area of slightly over 100,000 km^2. As HRIL 2006 is derived from a satellite mosaic that is not completely cloud-free, it contains some unclassified cells (around 1% of the cells are unclassified and they occur in 176 LAU2 units; Figure 4). Both of the countries are spatially close and rather similar to Slovakia in terms of physical environment (hilly terrain, a large proportion of forest area, inland location), the presence of regions with scattered settlements, and shared part of history within the Habsburg Empire.

3.2 Data

Source data

Population counts as of 2006 at municipality level (Eurostat LAU2 territorial units) were used, temporally coinciding with the two primary ancillary datasets (CLC 2006 and HRIL 2006), as well as with the fine-resoulution validation data that can be obtained from the GEOSTAT 1A 2006 1 km population grid. The data were collected in register-based censuses that took place both in Austria and Slovenia in 2006.

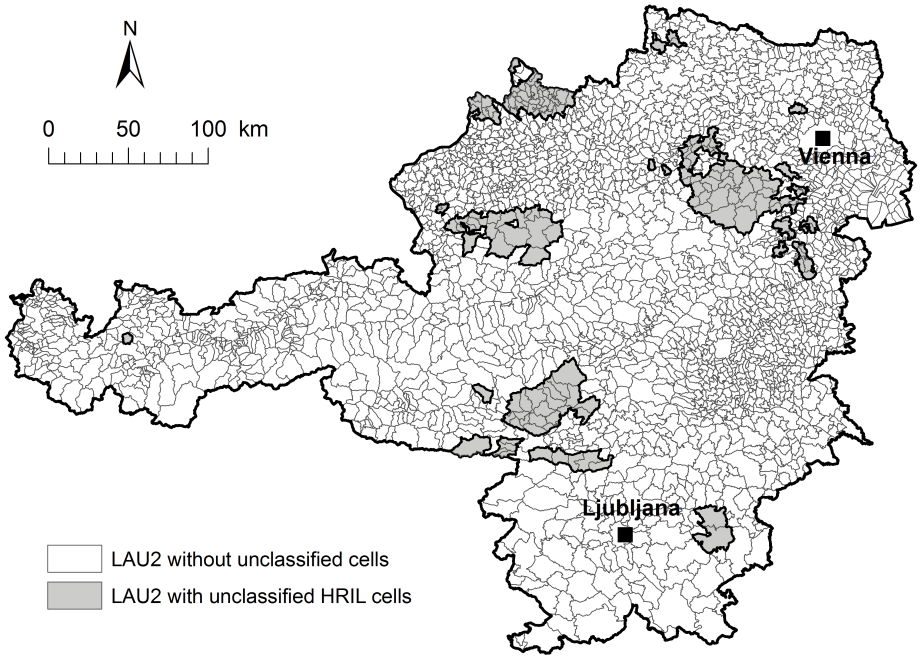


Figure 4: Study area

Ancillary data

Ancillary data for intelligent areal interpolation should be chosen carefully so that sound assumptions about the relationship between an ancillary variable and interpolated variable can be made. Many studies focusing on theoretical and methodological aspects of areal interpolation use ancillary data specifically acquired or created for the study, though covering only small study area (e.g. Langford and Unwin 1994, Mennis 2003, Wu and Murray 2007). On the other hand, more practically oriented studies that aim at developing larger extent dataset useful for a wider range of users are limited by the ancillary datasets that are already in existence because the production of special data for a large area (e.g. several countries) would normally be out of the scope of such studies. Fortunately, the EU pursue a land monitoring programme named Copernicus (formerly GMES), which provides several remote sensing derived datasets that can serve as ancillary data for population disaggregation. The Copernicus services are very convenient since they cover the entire EU, are based on harmonised definition and available free of charge. The datasets are updated periodically (HRIL every three years and CLC every six years) thus the methods developed for these datasets will have relevance also in the future.

Copernicus land monitoring services

High-Resolution Imperviousness Layer 2006

HRIL is one of high-resolution layers produced by semi-automatic classification of ca 20 m multispectral satellite mosaic of SPOT-HRIVIR and IRS-LISS imagery known as IMAGE2006. Each of these layers is focused on a single type of land cover (impervious, forest, water, wetlands, grasslands). Rather than being categorical LULC maps, the datasets provide interval values (ranging from 0 to 100%) conveying the intensity of the particular phenomenon per grid cell. HRIL has a particular potential for population downscaling since it shows 'sealed' or artificially impervious areas, which are closely related to the human presence and provides much less spatially generalised information compared to CLC. The dataset was obtained for research purposes from the EEA; a revised version was used at the aggregated 100 m resolution (the original version of HRIL is a raster derived from a 20 m satellite mosaic with the same pixel size).

In the study area, there are around 1% of unclassified pixels due to the cloud cover in the source satellite imagery. Since the missing data could bias the disaggregation result, we excluded all source zones that contained more than 1% of unclassified pixels (Figure 4). However, in the later stage of our research, HRIL 2009 data became available, and the missing data could be replaced by this data (only negligible 0.004% of pixels were unclassified in both datasets). Thus the entire territory of Austria and Slovenia was used in the analysis of OSM contribution to disaggregation accuracy (Annex I).

One of the source zones (the Austrian municipality of Gramais) did not contain any non-zero HRIL pixels, despite the fact that there are around 20 houses with 60 inhabitants (which is an example of omission error in the ancillary data). Because the applied method requires at least one non-zero HRIL pixel in each source zone, the correct values for several pixels covering the built-up area of the source zone were simulated using building footprints from OSM data.

CORINE Land Cover 2006

Previous efforts to disaggregate population data in the EU into a regular grid, performed at JRC relied mainly on the CORINE Land Cover (CLC) project data (Gallego and Peedell 2001, Gallego 2010, Gallego et al. 2011). The CLC maps are vector layers derived from satellite mosaics by visual interpretation, though we have used rasterized version of the original vector dataset with 100 m cell size. CLC uses a detailed classification scheme of 44 LULC categories in 3 hierarchical levels (Heymann et al. 1994, Bossard et al. 2000). The scheme represents a mix of land cover (coniferous forest, water areas, bare rock, and others) and land use categories (industrial and commercial facilities, ports, sport and leisure facilities, and others). Some of the classes in the latter group are very useful to diversify population densities attributed

to impervious areas according to their location within certain CLC class, especially to decrease densities in commercial and industrial areas, airports or ports, which may have high levels of imperviousness but low residential populations.

The full CLC classification scheme is overly detailed for population disaggregation. The thematic resolution can be roughly compared e.g. to 37 classes of Anderson Level II classification in GIRAS dataset used by Schoreder and Van Riper (2013, p. 219): "... for modelling population distributions; such scheme is needlessly and somewhat problematically precise. Even if significant differences exist among the population densities of the four classes of agricultural land or five classes of tundra, these differences are impossible to estimate with confidence because many of the classes occur rarely and comprise small portions of the few tracts in which they appear". The solution is reclassification to a much smaller number of classes (usually less than 10) that are potentially relevant from the population density point of view. For example Yuan et al. (1997) state that for a statistical regression (of population count on areas of land cover classes), fewer categories and more observations generally provide more reliable results. In their study, 16 classes were grouped into seven classes. CLC data used in dasymetric mapping by Gallego and Peedell (2001) were reclassified first into 17 and in the second step into nine classes. We adopted the latter scheme also in this study (Table 1; Annex II).

The major limitation of CLC in population disaggregation applications is relatively low spatial detail; the dataset is generalised by the minimum mapping unit (MMU) of 25 ha. Many LAU2s do not contain urban patch larger than 25 ha (e.g. when the settlements are rather dispersed) and therefore it is impossible to allocate the main population hot-spot within such zone based on CLC data. Gallego and Peedell (2001) observed from CLC1990 data that almost 27 thousand LAU2s with a total population of 12 million contained no urban patch in the then EU-12. The implication of the 25 ha MMU is that the LULC patches smaller than 25 ha are dissolved into surrounding dominant class. Therefore there can be, for instance, small residential land patches 'hidden' inside much larger arable land patch or vice versa. Considering this level of cartographic generalisation of CLC, we decided not to rule out any of the classes by attributing them with zero population density a priori.

OpenStreetMap Data

While HRIL suitably complements CLC in terms of spatial resolution, it is the opposite in thematic aspect. All built-up surfaces are treated equally without any distinction between various types or uses. Larger patches of built-up land used for industry, commerce, or transportation can be eliminated using CLC, but many non-residential built-up surfaces (including the most of road and rail networks captured by HRIL) are too small to be masked-out by CLC, and thus remain as target areas for population allocation. Additional spatial data can reduce this systematic error by

identifying smaller non-residential built-up areas (such as transportation network segments) and excluding them from the population downscaling process.

OpenStreetMap (OSM) is probably the most extensive and effective project of volunteered geographic information (VGI; Haklay and Weber 2008) and because we favour public domain data in our approach, using OSM for this purpose is an obvious decision. Open access data, including VGI, were advocated as a suitable data source for intelligent areal interpolation by Langford (2013). The combination of CORINE, HRIL and OSM has been used before for population downscaling by Austrian Institute of Technology (Steinocher et al. 2011) to produce 1 km population density grid (AIT Grid), which was integrated into Eurostat's GEOSTAT 1A 2006 product for several countries. According to AIT Grid's metadata, rasterized road and rail networks from OSM were used to mask the imperviousness layer. However, to our knowledge, the exact description of the methodology (how the OSM data were processed and used to alter the values of HRIL's pixels) has not been published in the literature, and the effect of this approach on the resulting population grid accuracy has not been measured. Related applications of OSM data include areal interpolation using point features (points of interest) rather than linear features as the interpolation guide (Zhang and Qiu 2014) and mapping of land use based only on OSM features without remote sensing or other sources of data (Jokar Arsanjani et al. 2013).

More types of OSM features than just roads and railways (e.g. building footprints or land use polygons) might be employed in population downscaling. However, mapping of these features relies heavily on the availability of online high-resolution imagery and volunteers determined to perform time-consuming visual interpretation. Therefore we assume that representation of such features is more prone to spatial heterogeneity and incompleteness compared to linear features such as road and rail networks, which can be mapped effectively also, for instance, using user-uploaded GPS tracks. In cases where comparability and homogeneity of the results at international scale matters, it is reasonable to restrict to major roads and railways (Annex III). For example, it was shown by Hecht et al. (2013) that the building footprints in Germany are far from completed and also that the completeness varies sign in space considerably. According to their analysis, footprint completeness in the German states of North Rhine-Westphalia and Saxony in 2011 was only 25% and 15% respectively. The completeness heterogeneity was even more significant at finer scale – while the city of Essen had 53.5% completeness, the city of Leipzig had only 28.9% completeness and rural districts scored generally worse with less than 10% completeness.

Validation data

Although the target units of disaggregation are 100 m grid cells matching with the cells of HRIL and rasterized CLC, the resulting dataset is not intended to assess population counts at the level of single target zones. The error of the estimate can

be unacceptably large at that scale due to various reasons, such as positional inaccuracy, generalisation effects as well as omission and commission errors that occur in remote sensing derived datasets like HRIL and CLC. Instead, it is rather intended to easily estimate populations of reasonably large irregular zones, which do not match municipal boundaries but can be well approximated by a subset of 100 m grid cells. Therefore, we set the target spatial resolution used in validation of the dataset at 1 km, which coincides with the bottom-up data available from GEOSTAT 1A dataset (see Annex IV).

3.3. Data preparation and processing

Combining HRIL and CLC data

CLC and HRIL are both meaningful ancillary datasets for population disaggregation and they have also been used previously for this purpose at the European level (Steinocher et al. 2011, Batista e Silva et al. 2013). In both cases the two datasets were complemented by additional proprietary or open access databases (e.g. TeleAtlas, OpenStreetMap, and Urban Atlas). The combination of CLC and HRIL is employed also in this research. The benefit of the combination is that the two datasets can partly compensate for each other's weaknesses: HRIL provides greater spatial detail and CLC enriches HRIL thematically, although the thematic enrichment of HRIL by CLC is limited to areas larger than 25 ha (and linear features wider than 100 m). Thus, the identification of non-residential built-up areas is only partial; for instance the majority of paved roads and other non-residential features captured in HRIL remain potential target for the allocation of population (to some degree, we compensated for this using OSM data in later stage of the research). Using 2006 CLC and HRIL data is convenient also because of the temporal match with 1 km validation data from GEOSTAT 1A 2006 data. Accurate source population data are available for year 2006 as well (in Austria, the majority of the study area, there was a register-based census accomplished for the first time in 2006).

Studies on combining multiple ancillary datasets are scarce compared to those dealing with a single categorical map, especially in the context of European data. In work of Batista e Silva et al. (2012), a refined version of CLC was created using multiple additional datasets (including HRIL) and subsequently it was used for population disaggregation in Batista e Silva et al. (2013). CLC data were used as a baseline layer, which was further modified by other datasets (HRIL, TeleAtlas, Urban Atlas, SRTM Water Bodies). In particular, TeleAtlas road and rail networks, as well as commercial and industrial polygons, were first used to mask out non-residential pixels in HRIL. The refined HRIL was used to identify previously undetected patches of urban land smaller than 25 ha; vacant land within the original CLC urban fabric patches; and to further classify urban fabric areas into high, medium, and low-density subclasses. A 1

km^2 disaggregated population density grid of the EU produced by Austrian Institute of Technology (Steinocher et al. 2011) used HRIL as the baseline layer for the disaggregation. CLC and data from OSM were used to mask out non-residential pixels in HRIL; however, the details of the procedure and contribution of the additional datasets to estimation accuracy were not published. Since OSM data comply with our preference of open data and its contribution to disaggregation performance is very likely, we investigate the benefit of including OSM data in this study as well; we discuss the details of data preparation and modification of HRIL in the following sections.

To combine HRIL and CLC data, we made these simplifying assumptions:

1. The data used for disaggregation reflect reality with reasonable accuracy;
2. The values of the HRIL cells (integers between 0 – 100) indicate the amount of impervious land within each cell's projection onto the Earth's surface (where a value of 100 equals 10,000 m^2; 50 equals 5000 m^2 and 1 equals 100 m^2). In theory, even a single family house that covers more than 100 m^2 might be represented in the HRIL as a cell with a value of 1 or more;
3. The population is localised exclusively on impervious surfaces. That is, only non-zero HRIL grid cells are considered (Annex VI). Effectively, this assumption is exercised by assuming a positive linear relationship between the imperviousness value and population size in a cell, with different slopes for each dasymetric zone (i.e. a unique combination of source zone and CLC class);
4. The population density per impervious area (PDIA) is the same in each dasymetric zone;
5. The ratio of PDIA for each pair of CLC classes is the same for all source zones in each subset.

Since the population is assumed to be present only in impervious areas, the variable of interest in the proposed disaggregation procedure will be PDIA instead of population density per total area. Similarly, the predictor variables used to model PDIA will be areas of impervious surface within each of the CLC classes (computed as sums of HRIL cell values) instead of total areas of each of the CLC classes.

Preparation of OSM data

Temporal issues

Unfortunately, in the case of OSM data, it is not possible to ensure a temporal match with the rest of data sources. In the feature attributes, there is only a timestamp representing the date of creation (or modification) of the database record, but no information is given on the time of creation of the respective real-world object.

Although the reason for adding a new record to the database could often be the fact that a real-world object came into existence, there is no way of 'cleaning' the database from features representing the real-world objects that did not exist in 2006.

Nevertheless, using the version of OSM data from 2014 to represent the state of the main roads and railways as of 2006 can still stand to reason. Top-level traffic infrastructure is not changing frequently, and even if there certainly were some modifications in the past eight years, the cases where the newly built infrastructure replaced residential buildings (which would result in erroneous masking of residential areas) would be rather scarce.

Completeness evaluation

Although the OSM project has been started in 2004 in London (Haklay 2010), its popularity was spreading in space relatively slowly in the first years of existence. Nies et al. (2012) evaluated the evolution of OSM in Germany evaluated the period of years 2007 to 2011. In the beginning of the period, there were almost no features in the database, but the number of new features increased in each of the years, reaching the maximum in 2011. For a comparison, in Austria only very few road segments were created before 2008, then the number of new edits was increasing every year and peaked in 2012. In Slovenia the progress was similar, but the peak was reached only in 2013 (Figure 5).

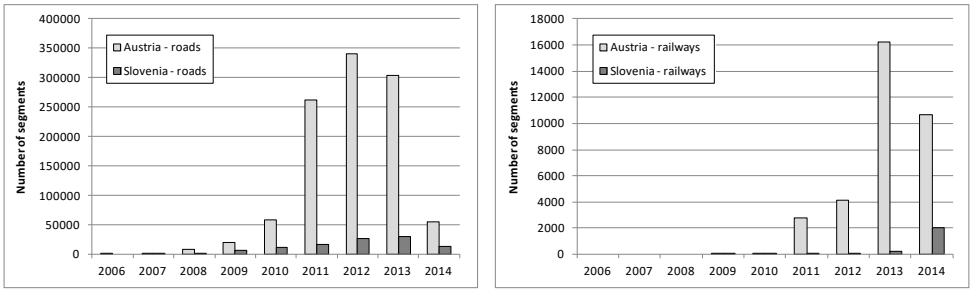


Figure 5: Yearly increase of OSM segments tagged 'highway' (left) and 'railway' (right)

When considering railway network, the evolution in the two studied countries was even more delayed, with the highest number of new features added to OSM in Austria in 2013. In Slovenia, over 90% of railway segments were added in the first five months of 2014. Thus, just a few years ago, OSM data would not be suitable for the intended purpose in the two countries. In other European countries, the completion of mapping can be still far from finished, and therefore OSM data should be used with caution.

Table 1. Reclassification scheme for CLC classes

New class	Original classes
1	Continuous urban fabric
2	Discontinuous urban fabric
3	Industrial or commercial units
4	Road and rail networks and associated land; Port areas; Airports; Green urban areas; Sport and leisure facilities
5	Non-irrigated arable land
6	Vineyards; Fruit trees and berry plantations; Complex cultivation patterns
7	Pastures; Land principally occupied by agriculture, with significant areas of natural vegetation
8	Agro-forestry areas; Broad-leaved forest; Coniferous forest; Mixed forest; Natural grasslands; Moors and heathland; Transitional woodland-shrub
0	Mineral extraction sites; Dump sites; Construction sites; Beaches, dunes, sands; Bare rocks; Sparse-ly vegetated areas; Burnt areas; Glaciers and perpetual snow; Inland marshes; Peat bogs; Salt marshes; Salines; Water courses; Water bodies; Sea and ocean

There are noticeable differences between Austria and Slovenia in density of road segments (i.e. all OSM polyline objects having certain value of the tag 'highway'; including dirt roads and tracks, pedestrian streets, hiking or cycle tracks, etc.). Austria has eight times greater length of the segments and ten times higher number of segments compared to Slovenia, although the difference in both country area and population is only fourfold. The difference may be partly due to the actual difference in road network density. However, not only the length of mapped features, but also the number of active contributors (with more than one contribution) is almost eight times higher in Austria compared to Slovenia. The factors that might contribute to these disparities include differences in broadband internet penetration and availability, computer literacy, popularity of the OSM project and willingness of people to get involved in voluntary activities. The availability and quality of base maps that can be used for mapping by the contributors (in particular aerial imagery from Yahoo! Maps and more recently Bing Maps) may also play an important role.

Apart from differences that are evident at the international level, heterogeneity can also be assumed within each country. As pointed out by Neis et al. (2012), OSM data show a much greater degree of detail in urban areas compared to rural ones. Contributors tend to use their local knowledge, focusing on areas where they live; consequently, rural areas (with lower population density, inferior broadband connection services and population with lesser computer skills) get usually fully mapped later than densely populated urban areas. This happens thanks to the 'urban' contributors who eventually shift their focus on the blank spots in the map after the urban areas are completely mapped (or even 'over-mapped' with relatively unimportant features). Similarly, Haklay (2010) observed that the centres of the big cities of England were well mapped while in suburban areas and in the boundary between the city and the rural area that surrounds it, the quality of coverage dropped very fast and many areas were not covered very well.

However, in the case of the highest hierarchy roads, which are not so numerous, but easily identifiable and mappable, the issue of spatial variation in OSM completeness – both between the countries and within them – becomes less critical. The difference in length of road segments between Austria and Slovenia is only fivefold when major road categories are considered (i.e. with the following values of highway tag – motorway, motorway link, trunk, trunk link, primary, primary link, secondary and secondary link), which is close to the actual difference between the countries both in area and population. This might indicate that for the major roads the database is almost complete in the studied area and that the marked difference in total length of roads is largely influenced by less relevant categories. Indeed, when the total length is subdivided into categories (Figure 6) it is clear that unlike Slovenia, Austria has overwhelming amount of the 'track' category. To quantify completeness in the study area more thoroughly, OSM data would need to be compared with some reliable proprietary data source, and that would require a study of its own. However, the assumption can be supported by the findings of Nies et al. (2012), who compared OSM data with TomTom navigation data in Germany. They have shown that the first category of roads to be completed (i.e. the length of OSM segments equalled or exceeded the length of TomTom segments) was motorway/highway which was by mid 2008. One year later, also secondary and tertiary roads were completed. The total length of OSM roads exceeded the total length of TomTom network in 2010, which was also thanks to many forest and field trails mapped in OSM. The exception was the category of residential roads, which was still not completed in 2011. Similar results can be expected also in our study area, possibly slightly delayed when compared to Germany.

Attribute issues

Although OSM data can in some areas have density of feature geometries exceeding that of commercial data, the attribute part of the database is rather lagging behind. The general issue with VGI according to Coleman et al. (2009) is that the amount of attribute data accompanying the contributions is usually limited to a few entries, tags, or free-form comments. The organization of such tags would typically not conform to any standard-based metadata specifications endorsed by public or private mapping organizations.

The OSM tag system uses a text string consisting of a key, equality sign and its value (k=v). Although there are some agreed conventions on standard keys and values, not all contributors are completely familiar with them (particularly the less experienced ones). Consequently, the same type of information about a feature can be expressed in several ways by different tag keys and values. Furthermore, contributors can create new tag keys or use non-standard key values. The tags are often left empty; for example, when a contributor is digitizing a street segment on top of aerial imagery without

local knowledge of its name and other properties that cannot be visually interpreted from the imagery (such as speed limit, etc.). For instance, only less than 1% of road segments in Austria have the tag 'maxspeed'.

Selection of roads for HRIL modification by road category

In order to modify the values of HRIL cells (more precisely, to decrease the imperviousness values of cells intersected by major road segments), surface area covered by relevant roads and/or railways needs to be estimated. For this purpose at least two attribute of the transport lines are necessary: category and width of road / railway. Road type (category) is needed to exclude irrelevant types such as forest trails, footways and residential roads (in the disaggregation most people should be allocated to residential areas thus values of HRIL cells intersected by residential roads should be preserved). Road types can be distinguished using values of the key 'highway'. A subset of transport lines was selected for further analysis on the basis of presence of this value; therefore there are no missing attributes (the only limitation is vagueness of some categories such as 'unclassified' or 'road'). Total length of roads in each of the categories is shown in Figure 6.

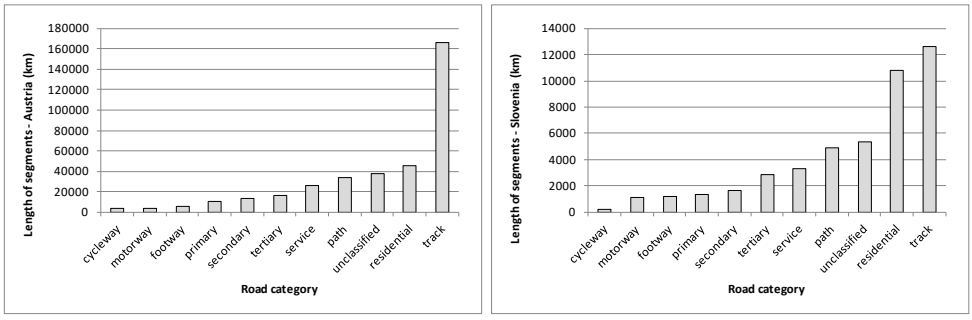


Figure 6: Distribution of OSM ways tagged 'highway' among the main road categories in Austria (left) and Slovenia (right)

It was not obvious at first glance, which road categories should be included in the analysis. For example, the inclusion of the service roads might prove beneficial when reducing high imperviousness degrees in industrial and commercial land uses. However, many of the service road segments are located close to residential areas, and their masking could decrease imperviousness levels in residential areas at the 100 m cell size. After inspection of the data (see Annex III), it seems that service roads category might be understood more ambiguously compared to the major categories. In the centre of the study area, there is a noticeable cluster with a high density of service road segments around 20 by 30 km in size, which does not correspond to an increased level

of human activity, but it is merely a local misclassification of forest roads. Although this flaw should not cause erroneous masking of residential areas, it is illustrative of issues typical for VGI.

Inclusion of secondary and primary roads could also be a matter of discussion. In theory almost every paved road should be captured by HRIL (i.e. 10 m wide road crossing a 100 m cell should result in at least 10% imperviousness). In fact, though, only a limited part of the roads is visible in HRIL outside of compact built-up areas of cities, towns and villages. The linear features clearly visible in HRIL outside of built-up areas are mostly motorways (identified by OSM tags motorway, motorway link, and sometimes trunk /trunk link) and railways.

Estimation of road and railway width

Road and railway segment width is a parameter needed for the estimation of impervious area covered by individual transport line segments. Unfortunately, the tag 'width' is very rarely recorded by the contributors. Therefore, an empirical approach was used instead to estimate this parameter. Each relevant category of roads and railways was randomly sampled and the width of the sampled segments was measured using aerial imagery. Sample size was 0.1% of segments in each category (but at least 20 segments). Results of the sampling are shown in Table 2.

Table 2: Statistics of road and railway width measured based on aerial imagery

Category	Count	Minimum	Maximum	Average	Standard Deviation
Motorway	26	11	22	15.27	2.59
Motorway link	20	5	15	9.05	2.44
Primary	39	4	19	10.92	3.15
Primary link	20	6	14	8.65	2.39
Railway	99	2	11	4.68	1.64
Secondary	30	5	18	9.70	2.98
Secondary link	20	5	13	7.35	2.01
Service	267	3	20	5.94	2.83
Trunk	20	8	14	10.80	1.88
Trunk link	20	6	10	7.95	1.32

Modification of HRIL

Half of average measured width (calculated for each of the categories separately) was used as a buffer distance to create vector areal representation of the transport network, which was then intersected with square polygons of individual nonzero HRIL cells. The proportion of cell overlapped by the vector areal representation of transport network was used as a cell value of a newly created raster representation of

the transport network. The final modified version of HRIL was derived by subtraction of this transport network raster from the original HRIL raster; any resulting negative values were changed to zero (see Annex VIII).

Visual inspection of the modified HRIL raster showed that the transport line segments captured by HRIL were not completely masked in many cases – the values of the respective cells have decreased only partly. Necessarily, in many cases, the averaged road widths were smaller than the actual ones, but overestimation of actual imperviousness levels by HRIL was also a likely cause. Supporting findings were reported in a study assessing the accuracy of HRIL (Hurbánek et al. 2010) – cells with high actual imperviousness levels tended to be overestimated by HRIL and cells with low actual imperviousness levels tended to be underestimated (while values of 0 and 100 were unrealistically numerous). To reduce the imperviousness values to a larger degree, further modifications of HRIL were performed using 1.5 and 2 times exaggerated width of the transport lines. The exaggeration of width can also reduce errors caused by positional inaccuracies in HRIL and OSM data, which can be identified as a spatial offset between corresponding features in map.

3.4. Methods

Estimation of population density coefficients in intelligent disaggregation

Many methods are described in the literature for dasymetric areal interpolation of population based on a single categorical land cover map with two or more categories (classes), such as the CLC. The key issue in dasymetric methods is to predict population density values for each LULC class used. In the binary dasymetric method (only populated and non-populated classes are used), the solution is relatively straightforward – the population is redistributed into populated zones based on the homogeneity assumption (areal weighting). If there are two or more classes with assumed non-zero population density, other assumptions have to be made. Some approaches (e.g. Mennis 2003) aim to predict the density of each class by empirical sampling from pure source zones (those that contain a single class). However, it may be impossible to find such source zones for each class (depending on the spatial detail of the source zones and ancillary data, as well as the scarcity of some classes). Sometimes, an expert guess of the respective densities can be made, provided that the study area is relatively small and familiar to the researcher.

A more frequent approach is regression analysis, where source zone population counts represent the response variable and areas of each of land cover class within source zones (obtained by GIS overlay analysis) represent the predictor variables. Intuitively, regression through the origin (non-intercept) option is chosen to find estimates

and to test the models, because there should be no population if there is no area on the Earth's surface (Yuan et al. 1997). Langford (2006) points out that model parameters βc can be loosely interpreted as population densities for each LULC class c. However, linear regression can yield negative parameters, which do not make sense as population density values. Therefore, scaling techniques are used to increase all the parameters to be non-negative. Further scaling is required so that the pycnophylactic condition is satisfied. Assuming the non-negativity of population density and a Poisson distribution of population count data, linear regression is not an appropriate model. Schoreder and van Riper (2013) suggest that alternative nonlinear models suggested by Goodchild, Anselin, and Deichmann (1993) deserve further attention, as does the quantile regression approach that Cromley, Hanink, and Bentley (2012) used to achieve properly constrained areal interpolation. Poisson regression is another reasonable alternative for interpolating population counts, even if population distributions do not strictly conform to the independence assumption of Poisson models (Flowerdew and Green 1989).

Another statistical technique used in estimation of population density coefficients is the expectation-maximization (EM) algorithm, first described by Dempster et al. (1977). It is a framework for model fitting and maximum likelihood estimation for incomplete data in which two steps are repeated iteratively. As described by Schroeder and Van Riper (2013), the expectation step 'completes' the data by computing the conditional expectation for missing data, given a set of observed data and estimated model parameters. The maximization step then fits the model, estimating model parameters by maximum likelihood given the 'complete' data from the E step. Thus, estimates from the E step are used in the M step to estimate a separate set of data, which are, in turn, used in the next E step to update the first estimate set, and so on, forming a feedback loop that repeats until convergence. The EM algorithm was first suggested for use in areal interpolation of population data by Flowerdew and Green (1989, 1992, and 1994). Schroeder and Van Riper (2013) applied an innovative geographically weighted EM algorithm (EM was applied separately in a local kernel neighbourhood of each source zone, thus, allowing for spatial variation in the population density coefficients). An algorithm used for coefficient estimation by Gallego and Peedel (2001), can be compared to a regionally applied EM algorithm; the general mechanism is analogous (running iterations with a feedback loop). However, calculations performed in the M step are unique for this approach (see description of the method in the following section). A variant of this method was applied also by Bielecka (2005) to produce a dasymetric population map of Poland.

Iterative method of population density coefficients estimation

The iterative algorithm, proposed by Gallego and Peedell (2001), was preferred in this study since it is quite easy to implement without the need to integrate specialized

statistical software and GIS environment. Various types of regression could be used as well, although the choice of particular regression type is debatable and implementation is comparably more complex. The method is a 'fixed ratio' method (Gallego et al. 2011), which assumes that ratios between population densities of land cover classes are constant either for the entire set of source zones, or for a certain subset of source zones (e.g. a region within a country).

The goal is to find the ratios (i.e. population density coefficients), and tune their values to obtain as good a performance as possible. All source zones m in a subset r share the same set of PDIA coefficients U_{cr} for CLC classes c. Initially, homogeneous PDIA was assumed for all the subsets and all the CLC classes. Thus, all Ucr values were set to 1. To further tune the U_{cr} values to be subset-specific, the following steps were taken.

PDIA values Y_{cr} for control zone cr (intersection of subset r and CLC class c) can be calculated according to Equation 1, where X_r denotes the true subset population count and S_{cr} denotes the sum of impervious areas in control zone cr. The equation ensures that the total subset population is preserved (pycnophylactic property);

$$Y_{cr} = U_{cr} \frac{X_r}{\sum_c S_{cr} U_{cr}} \qquad X_m^* = \sum_c S_{cm} Y_{cr}$$

Equation 1 *Equation 2*

Estimated population X_{m^*} for each source zone m is calculated using Y_{cr} values from the previous step and S_{cm} values of impervious area in the intersection of source zone m and CLC class c (Equation 2);

Since the true source zone counts are known, deviation of the estimate can be calculated for each subset (δ_r) and study area (δ) using Equations 3 and 4, where X_m denotes the true source zone population;

$$\delta_r = \sum_{m \in r} |X_m^* - X_m| \qquad \delta = \sum_m |X_m^* - X_m|$$

Equation 3 *Equation 4*

U_{cr} coefficients are modified (separately for each subset) to reduce the disagreement as follows:

The Pearson's correlation coefficient ϱ_{cr} is calculated for each control zone cr between (1) the relative over/underestimation of source zone population ψ_m and (2) the source zones's m share of total impervious area localized in the respective CLC class c (Equation 5 and Equation 6);

$$\psi_m = \frac{X_m^*}{X_m} \qquad \rho_{cr} = corr\left(\psi_m, \frac{S_{cm}}{S_m}\right)$$

Equation 5 *Equation 6*

The direction and magnitude of correlation indicates how to modify the U_{cr} value of the respective control zone cr. If ϱ_{cr} is positive, it can be assumed that generally population of source zones with high proportion of class c are overestimated. Thus, the coefficient has to be reduced and vice versa. The modified $U_{cr'}$ coefficients are calculated according to Equation 7. The ratio between the actual subset's TAE δ_r and the maximum possible subset's TAE (which is twice the subset population) in the equation ensures, that the magnitude of coefficient modification is decreasing as δ_r converges to a minimum;

$$U_{cr}^{'} = \left(1 - \frac{\rho_{cr} \times \delta_r}{2 \times X_r}\right) \qquad Y_{cm} = U_{cr}\frac{X_m}{\sum_c S_{cm}U_{cr}}$$

Equation 7 *Equation 8*

The modified U_{cr} values are used to disaggregate the subset population, the deviations are calculated, the coefficients are updated again and the cycle is repeated until the overall disagreement δ becomes stable.

The final values of U_{cr} coefficients are then used to calculate PDIA values for each control zone *cm* using Equation 8, which preserves the pycnophylactic property at the source zone level. If the subsets are overlapping, each source zone has its own subset and its own set of U_{cr} coefficients. In this case, the U_{cr} values obtained from each subset are applied only for the respective 'central' source zone of the subset.

The population count estimate X_t for target zone t (100 m grid cell) is calculated by multiplying the underlying control zone Y_{cm} value by the target zone's impervious area S_t (which is equal to value of the respective HRIL cell), Equation 9.

$$X_t = Y_{cm}S_t$$

Equation 9

Approaches to regional model fitting

An important question regarding the regression analysis (or other statistical method used for estimation of population density coefficients) is whether it should be applied globally to the whole set of source zones, or multiple models should be applied to several subsets (e.g. regions or local neighbourhoods) of the source zones. The limitation of the global approach is that the same parameters are applied over the whole study area. Even though the resulting population density values are scaled to preserve the known source zone population totals, and hence they can vary from source zone to source zone, the ratios between population densities of the LULC classes are fixed. It is not very realistic to assume that, for instance, the ratio of population density of residential and industrial classes would be the same everywhere. While the population density of industrial areas can be comparably low in cities of different size, the residential density can be expected to be much higher in a large city than in a smaller town.

The first approach to regional regression uses higher level administrative/census units than the source zones (e.g. county, district). Langford (2006, p. 166) questions using such an arbitrary regional scheme: "How should any given study area be subdivided into smaller regions? The counties used by Yuan et al. (1997) are, after all, just a rather arbitrary grouping of Block Groups which has little to do with population characteristics. It might be expected that the finer the level of regionalisation adopted (say County Subdivision or Place), the better the model fit would become. However, sub-regions could be developed from any combination of Block Groups not just those defined by the US Bureau of the Census, and they need not even be contiguous." A different approach was used by Gallego and Peedell (2001), who used LAU2 as the source zones and higher level administrative regions (NUTS2) for regional coefficient estimation (although they used the iterative algorithm described in 3.4.2 instead of fitting a regional regression model). Subsequently the approach was refined by stratification of LAU2 in each NUTS2 into three strata: dense (population density more than double compared with regional average), less dense (population density less than double compared with regional average), and no urban (LAU2 missing an urban patch). Again, the use of NUTS2 is unrealistic. A typical NUTS2 region contains a major city (regional centre), several other cities, smaller towns and the majority of LAU2 represent rural villages. Each of these types tends to exhibit different population density patterns, for example, as a result of different housing types, multi-storey building heights and occupancy rates. Stratification of the LAU2s according to population density and presence of an urban patch in the CLC alleviates this issue, but the pool of suitable LAU2 is still constrained by the boundaries of NUTS2 regions. There is no reason why a subset could not stretch over a NUTS2 boundary, or even country borders. Also, using LAU2 population density calculated using total area can be influenced by definition of the LAU2 border, as much as by the spatial distribution of the population (the modifiable areal unit problem).

Geographically weighted regression (GWR) is another alternative to global model fitting. GWR was used in areal interpolation problems by Lo (2008), Lin, Cromley and Zhang (2011), and by Schoreder and Van Riper (2013). In the latter study GWR was combined with the EM algorithm. GWR fits a separate model for each source zone based on its local neighbourhood, where each member of the neighbourhood can be weighted using a distance decay function. Such approach allows the population density coefficients to vary in space continuously (unlike the regional approach, where all source zones within each region share the same set of population density coefficients). Langford (2006) suggests that allowing the ratios between densities, as well as their absolute levels, to vary over space could be beneficial. GWR assumes that objects that are close to each other in space are more related than those further apart (as suggested by Tobler's first law of geography). However, this assumption is not appropriate in every setting. Each LAU2 in the EU (a source zone) typically represents a single settlement – city, town or village – positioned at a certain level in a hierarchical settlement system. We assume as well, that settlements of the same hierarchical level and with similar functions will have similar population density coefficients for land cover classes. In this setting, the larger the settlements grow in size, the greater the distance between them. Thus, while villages are usually found close together, cities are spaced much further apart. As a result, if a contiguous local neighbourhood (such as in GWR) is formed around a top level city, it will likely contain very different types of settlements than the city itself. Thus, using contiguous local neighbourhoods may be inappropriate at least for some of the source zones (if the setting was different and the source zones were smaller units – e.g. census tracts or 1 km grid cells – the contiguous neighbourhoods might be more appropriate).

Generally, the subsets of source zones used to fit regional models can be formed in these basic ways:

- contiguous, non-overlapping subsets (conventional regions);
- non-contiguous, non-overlapping subsets (e.g. stratified conventional regions);
- contiguous, overlapping subsets (such as kernel-based neighbourhoods in GWR);
- non-contiguous, overlapping subsets.

Design of non-spatial source zone neighbourhoods for coefficient estimation

The idea that parameters yielding a good fit in a regionally fitted model will perform well also at redistributing population counts within individual source zones of the respective region implies that the members of the region are somehow similar, so that the general pattern of the subset can be extrapolated to the individual source zones as well. While GWR constructs the regions so that the members are similar in spatial location, we suggest that the members should be similar rather in other aspects,

more relevant to population density variation (or that the spatial location should be used only as a secondary criterion). Because in intelligent methods of areal interpolation ancillary data about the study area is available, it is possible to characterize each of the source zones based on the ancillary data or further indices derived from the ancillary data and source population data (such as population density per impervious area). Consequently, the source zones can be clustered into subsets based on these indices. For each source zone a unique subset of the most similar source zones can be formed such that a set of spatially non-contiguous overlapping regions is obtained.

The subsets are created in following way:

1. All the source zones in the study area are in a pool of candidates for any subset;
2. Each source zone in the study area has its own subset;
3. Each source zone can be part of multiple subsets;
4. Subsets can be non-contiguous and can overlap each other;
5. A subset of a source zone *c* is defined as the m nearest source zones to the source zone *c* in an *n*-dimensional feature space, where the *n* features (attributes) are obtained automatically from source and ancillary data.

The following features of the source zones could be obtained and potentially used to construct the feature space, in which similarity of the source zones could be measured:

- From the ancillary data (CLC and HRIL):
  [A] LULC composition of each source zone (i.e. proportion of each CLC class)
  [B] Distribution of impervious areas among CLC classes (i.e. relative and absolute areas of impervious surface in each of the CLC class)
  [C] Estimate of total impervious area in the source zone
- From the source data (LAU2 boundaries and population counts):
  [D] Total population of the source zone
  [E] Total area of the source zone
  [F] Population density per total area
- From both the ancillary and source data:
  [G] Population density per impervious area
  [H] Share of impervious area in total area
- Additionally, spatial distance can be considered in combination with other features:
  [X] XY coordinates of source zone centroid

3.5. Sensitivity analysis design

Intelligently designed non-contiguous overlapping subsets based on information obtained from the ancillary data should outperform other types of subsets in terms of interpolation performance. To test this hypothesis, we conducted a sensitivity study, where the spatially non-contiguous subsets derived from the ancillary data were compared to contiguous non-overlapping subsets (randomly generated regions [Figure 10] and NUTS2 administrative regions were used) and contiguous overlapping subsets (local neighbourhood of each source zone). The same algorithm of population density coefficients estimation had to be applied to all configurations of subsets.

There are many ways to measure distance/similarity in an n-dimensional space, and many possible combinations of the above suggested dimensions that can be used to construct the space. For simplicity, we used Euclidean distance and the subsets were defined for each source zone as the m most similar source zones, where m is a predefined count of subset members. Because the variables have different units and scales of measurement, their values should be normalized in some way so that they are corresponding; z-scores normalization was used in this study. Some of the above listed attributes are not suitable for similarity assessment and can be excluded. The total area of a source zone and average population density per total area (E and F) are both largely affected by arbitrary definition of the LAU2 boundaries. Relative and absolute areas of impervious surface in each of the CLC classes (B) were also excluded due to the rarity of impervious surfaces in some of the CLC classes and the consequent small number problems. Eventually we selected the following feature spaces for the analysis: A (Figure 8), X (Figure 7), AC, AD, CDX, DGH (Figure 9), ADGH, DGHX, and ADGHX. The space X creates contiguous subsets based on spatial distance, other combinations containing X consider the spatial distance only partially. All features in all combinations were given the same weight in distance measurement.

Another factor to which the interpolation accuracy might be sensitive is the count of subset members. It is not obvious which size of subsets should be preferred, although a range of reasonable sizes can be supposed. The larger the subset (in terms of source zone count), the more dissimilar source zones will be included. On the other hand, in too small subsets it might be problematic to properly calculate the correlation coefficients needed in the iterative algorithm described in following section. Thus, for all feature spaces, as well as for randomly generated contiguous non-overlapping subsets, these subset sizes were tested: 50, 100, 150, 200, 250 and 300. Because the average size of LAU2 differs between Austria and Slovenia (and there is also marked heterogeneity present within Austria – the Alpine part of the country has much larger LAU2s than the lowland part), same-count (SC) subsets can have quite diverse areas. Therefore, also subsets based on same area principle were tested with the following sizes (in km^2): 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, and 8000. If, by chance, mostly large source zones were chosen to form a subset of e.g. 2500 km^2,

there might have been only around 10 members in the subset. To avoid this situation, we introduced a lower bound of 25 source zones for same-area (SA) subsets. As an alternative to the predefined size of subsets (based on count or area), a cluster analysis could be conducted to determine the cut-off distance (and, thus, also subset size) based on discontinuities in the feature space. However, this approach was not tested here.

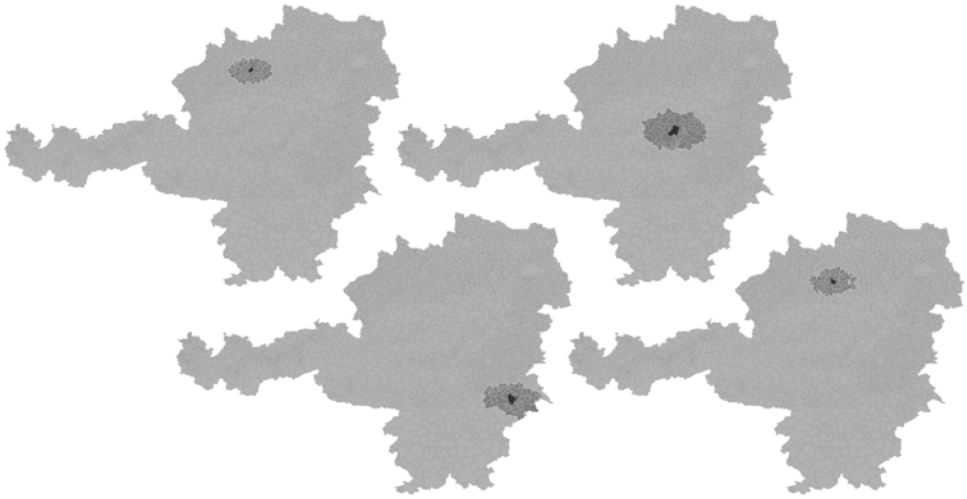


Figure 7: Examples of subsets 'X' based on spatial coordinates (75 nearest source zones)

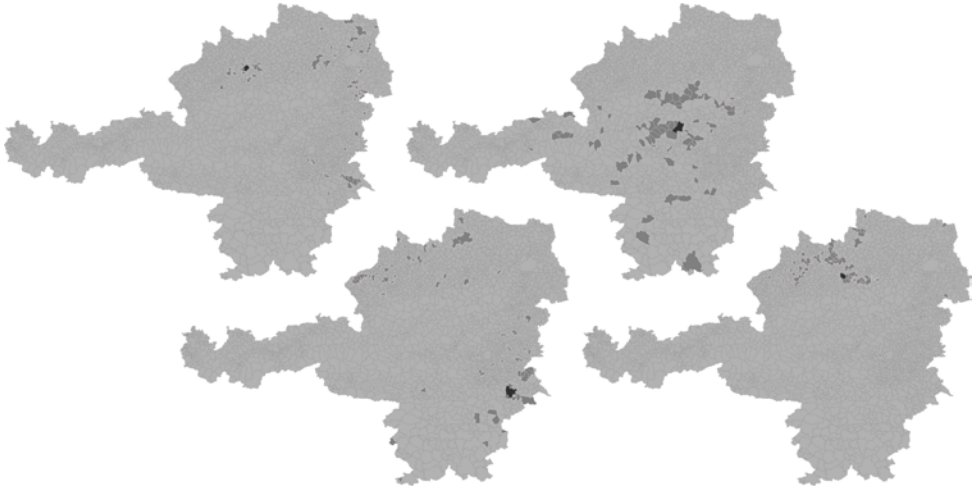


Figure 8: Examples of subsets 'A' based on proportions of grouped CLC classes (75 nearest source zones)

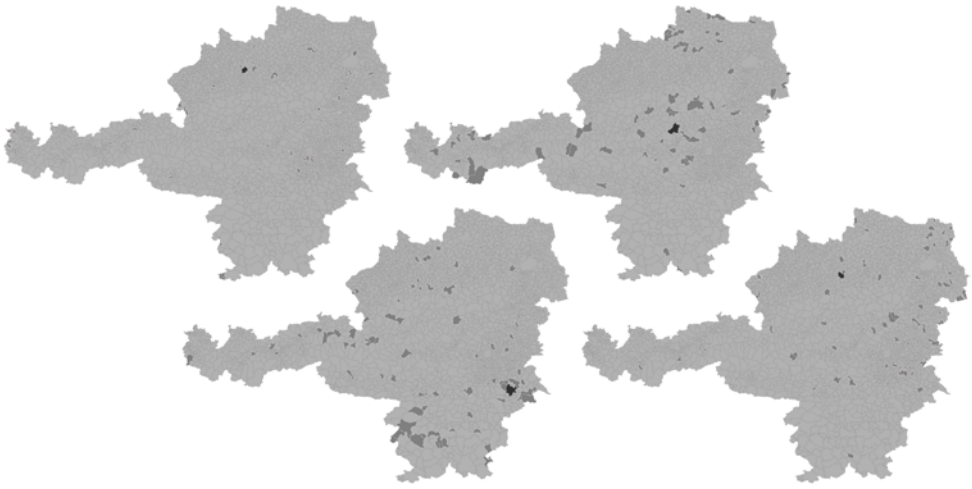


Figure 9: Examples of subsets 'DGH' based on total population, share of impervious area, and population density per impervious area (75 nearest source zones)

To find out which of the OSM road types should be masked to achieve the most accurate disaggregation, four different subsets of road networks were tested – MT, MT1, MT12, and MT12S (M-motorway, T-trunk, 1-primary, 2-secondary, S-service; the respective 'link' categories were included as well). We did not expect that masking the complete set of railway segments would deteriorate disaggregation accuracy; even in cases when the railway was not captured by HRIL because residential buildings are usually not located right next to railways. The main expected contribution of railway masking is reduction of high imperviousness degrees in major train stations and rail hubs with numerous parallel tracks. Furthermore, the four selections of road types were tested with three levels of road width exaggeration – 1, 1.5, and 2 times the measured width; resulting in 12 combinations. The comparison of these 12 combinations was performed using only the best performing scheme of source zones subsets, obtained from the preceding analysis. Otherwise, over 3000 disaggregation procedures would need to be carried out to evaluate all possible combinations of subset schemes and HRIL modifications, posing an excessive computational cost.

3.6 Validation methodology

Regardless of the principle behind the construction of source zone subsets, it is important to ask if, and to what degree the improved goodness-of-fit of regional regressions over a global regression translates into increased accuracy of areal interpolation (Langford 2006). Langford (p. 166) further notes that "the better model fit could,

for example, be attributable to smaller sample sizes; it does not necessarily follow that regional regression models perform areal interpolation any more accurately than a globally fitted model". Validation of the disaggregated products using reference population data with finer resolution than the source zones, therefore, needs to be carried out. Also, attention should be paid to scale effects in the validation process. The size difference between the source zones (LAU2 units) and target zones (100 m grid cells) is quite marked. Thus, it can be expected that validation performed using 100 m bottom-up data will result in very high total absolute errors. Nevertheless, the product is not intended to be used at this resolution; rather it is a ready-made data source for predicting population counts of arbitrarily defined areas. At 100 m resolution, the performance can be relatively sensitive to positional uncertainty and generalisation effects in the source and ancillary data and other systematic errors, but with increasing size of validation units the accuracy is expected to increase. In other words, while the error at the target resolution can be seemingly unacceptable, the product may also provide reasonably accurate population estimates for larger zones such as a buffer zone or viewshed in many applications.

To evaluate performance of areal interpolation methods various error indicators can be used. Some researchers favoured root mean square error (Eicher and Brewer 2001, Langford 2006), and mean absolute error (Goodchild et al. 2003). However, to evaluate accuracy of the disaggregation, we favoured total absolute error (TAE) as the main indicator of model performance, because it is more robust for skewed distributions as is the case of population density (Batista e Silva 2013). Gallego et al. (2011, p. 2064) also argue for the use of TAE instead of root mean square error (RMSE), since "RMSE is very sensitive to moderate errors in a small number of large values and much less sensitive to serious errors in small values."

In this case, the TAE is the sum of all absolute deviations of individual 1 km grid cells, and it is computed as shown in Equation 10; c denotes index for grid cell, n is the number of grid cells in the study area, X^* estimated population and X actual population. To compare performance between areas with different total populations e.g. Austria and Slovenia, or the full study area and the incomplete study area (source zones containing unclassified pixels excluded), relative TAE (RTAE) needs to be computed by dividing TAE by total actual population of the area (Equation 11).

$$TAE = \sum_{c}^{n} |X_c^* - X_c| \qquad \textit{Equation 10}$$

$$RTAE = \frac{TAE}{\sum_{c}^{n} X_c} \qquad \textit{Equation 11}$$

3.7 Geoprocessing and software solution

It is not feasible to implement the described method by ready-made GIS tools. Also, since many parameterizations of the method need to be tested, there is a strong need to automate the routine. To achieve this, we have employed ESRI ArcGIS 10 software package and its ArcPy module – a Python (programming language) site package for performing GIS functions. Although this package contains a multitude of useful tools, further libraries were needed to facilitate the algorithm. We have used three open source libraries – SciPy (scientific tools library), PySAL (python spatial analysis library), and matplotlib (python 2D plotting library).

Using these libraries and PyScripter (integrated development environment for Python language) we have developed a comprehensive script (over 1,000 lines of code) that takes only raw data (source zone boundaries with real population counts, the ancillary datasets and validation data) and several parameters as an input. The script automatically extracts needed information about the source zones from the ancillary data (such as share of individual CLC classes and impervious surfaces), calculates normalized values of the variables selected for construction of the n-dimensional space, computes a distance matrix, creates subsets of source zones based on proximity in that matrix, iteratively tunes population density coefficients based on the subsets, disaggregates the population counts into grid cells based on the coefficients, validates the resulting grid, and writes the output files. Each complete run of a particular parameterization of the script took between two and four hours (increasing with the dimensionality of the feature space and size of the source zone subsets), and over 250 runs were completed. An enormous amount of geographic and tabular data was produced in the process which is quite challenging to analyse and interpret, and only a small fraction of it can be included in a printed publication. Though, all the data as well as the Python source code is archived and available to readers on demand.

Disaggregation with the OSM data included took place in later phase of our research and was not fully automated. We have produced 12 modified versions of the HRIL manually; these were then used instead of the original HRIL in the above described algorithm.

4 RESULTS AND DISCUSSION

4.1 Results of disaggregation based on HRIL and CLC data

The disaggregation was preformed for nine variants of feature space and 18 subset sizes. Additionally, for each subset size, five sets of randomly generated contiguous non-overlapping subsets (Figure 10) and one set of administrative regions (NUTS2) were tested (i.e. 253 subset configurations in total). Although the iterative process of U_{cr} coefficient tuning minimizes the total disagreement δ (TAE measured at source zone level), it does not necessarily mean that the accuracy of disaggregation to target zones level will improve monotonously with the number of iterations. Thus, the disaggregation results using intermediate coefficient values were validated on every tenth iteration, so that the change in TAE could be plotted and examined. In an actual setting, the sub-source zone validation data may not be available, and the number of iterations performed must be decided beforehand. Subset configurations with validation TAE values monotonously decreasing throughout the run can be considered more robust and preferable for an actual application. 150 iterations were performed with each subset, during which the majority of the models converged to a minimum (the rest started to increase only very slightly at some point).

As a starting point (iteration #0), uniform coefficient values were applied, i.e. it was assumed that all dasymetric zones within a single source zone have the same PDIA value. Effectively, in this step the population was disaggregated within each source zone proportionally to HRIL value of the respective target zone, regardless of the underlying CLC class. The performance of this 'baseline' disaggregation can be illustrated by comparison with the performance of areal weighting (Table 3). The HRIL proportional method performed with approximately half the TAE of areal weighting. With further iteration the PDIA coefficients started to differentiate among the CLC classes, thus, further increasing the accuracy (compared to that of HRIL proportional). Variation of the accuracy gain according to size and type of subsets used is shown in Annex IX, Figure 11, Figure 12, and Figure 13; the results for contiguous non-overlapping regions ('Contig') are averaged values of five randomly generated systems of (nearly) equally-sized regions (Figure 10).

Table 3: Overview of validation results – disaggregation based on HRIL and CLC data

Validation area, spatial resolution of validation	Areal weighting TAE	HRIL proportional TAE	HRIL proportional TAE reduction	Best refined result TAE	Best refined result TAE reduction	Subsets used in best refined result
Austria + Slovenia, 1 km	10,695,440	5,093,935	52.4%	4,361,890	14.0%	DGH, 2500 km²
Austria, 1 km	8,473,131	3,819,682	54.9%	3,090,720	19.1%	DGHX, 5000 km²
Slovenia, 1 km	2,222,309	1,274,253	42.7%	1,222,100	4.1%	ADGH, 300 zones

In general, the best results were achieved using spatially non-contiguous subsets having roughly 2000 – 3000 km² in area, based on similarity of three features of individual source zones: population size, share of impervious area in total area, and population density per impervious area ('DGH' feature space). Although the feature space 'DGHX' (where also spatial distance is taken into consideration) performed comparably to 'DGH', we preferred the variant with a smaller number of variables for the assessment of accuracy improvement thanks to incorporation of OSM roads and railways (as described in the following section).

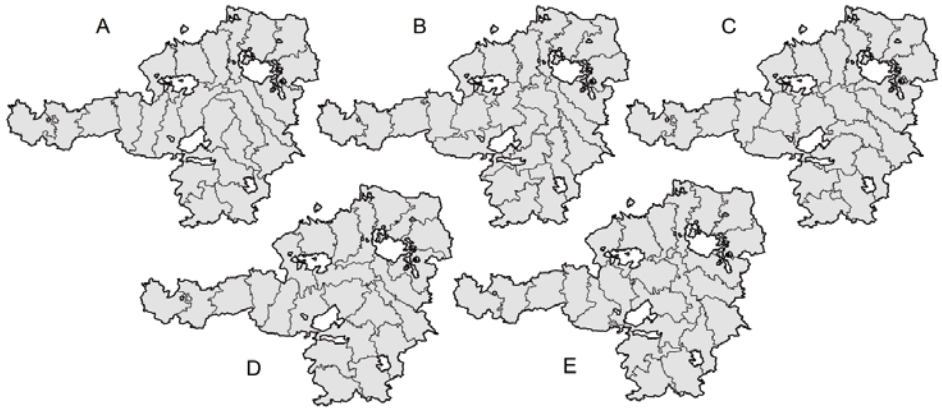


Figure 10: Five variants of randomly generated contiguous non-overlapping source zone subsets with equal area of ca 4000 km²

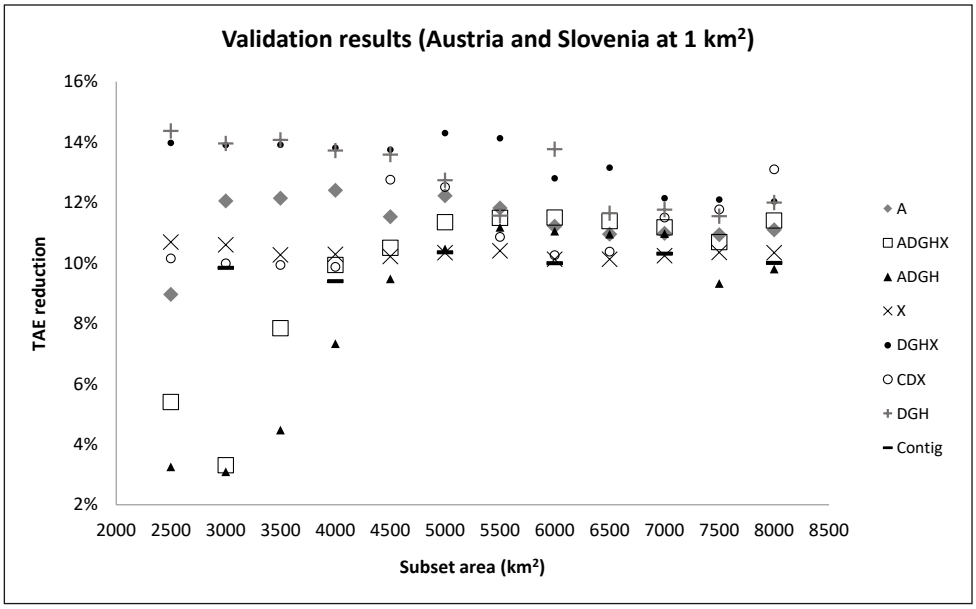


Figure 11: Validation results (Austria and Slovenia at 1 km²)

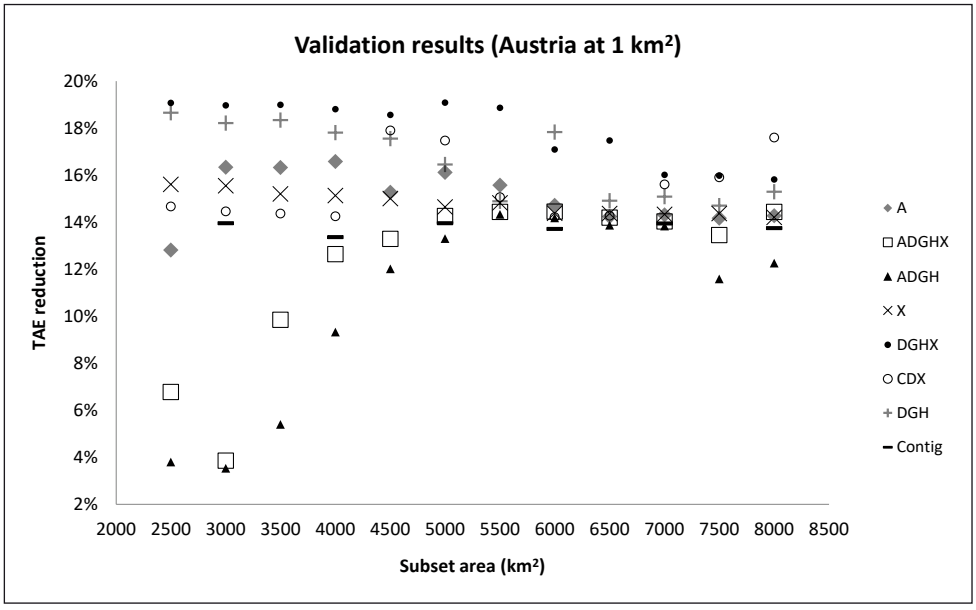


Figure 12: Validation results (Austria at 1 km²)

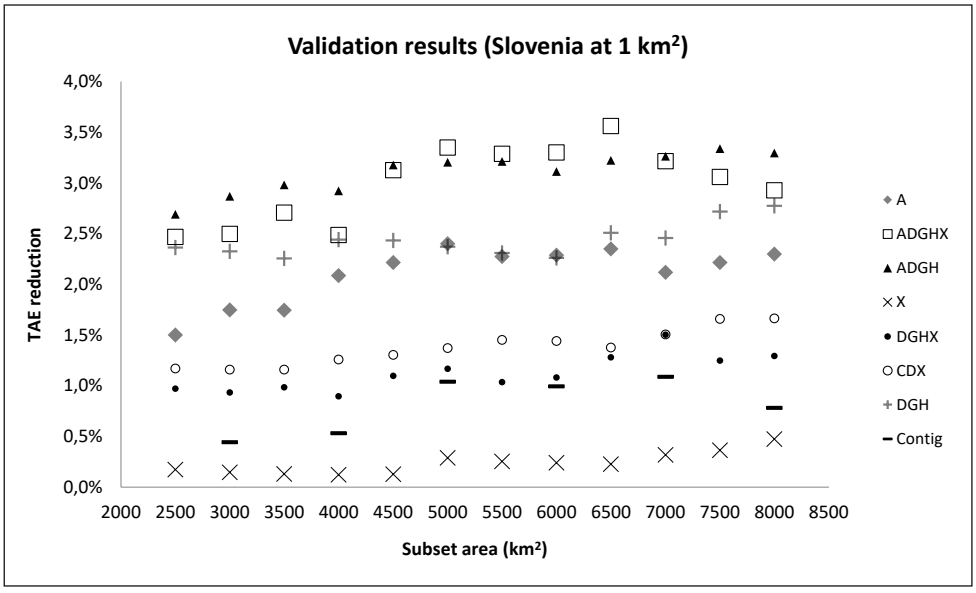


Figure 13: Validation results (Slovenia at 1 km²)

4.2. Results of disaggregation based on HRIL, OSM and CLC

Results show that the reduction of TAE thanks to OSM incorporation was significant regardless of the selection of road categories and width exaggeration when measured for the entire study area. For the 12 tested modifications of HRIL, the improvement of the baseline disaggregation ranged from 176 to 283 thousand people (i.e. 3.3 to 5.3%, Figure 14). This represents the direct accuracy gain thanks to modification of HRIL by OSM. Further reduction of TAE brought by the iterative tuning ranged from 612 to 691 thousand people (over 13%), resulting in TAE of 4,424 thousand persons. The RTAE obtained by disaggregation using the original HRIL exceeded all tested modifications of HRIL (albeit only slightly for some of them). This confirms conclusively our expectation that additional ancillary datasets improve estimation accuracy.

Similar level of TAE reduction was achieved over the iteration also with the original non-modified HRIL – ca 720 thousand (see Figure 15). Consequently, the differences in TAE after the iterations between the modified and non-modified HRIL are similar or smaller than the baseline differences – from 68 to 240 thousand people (1.5 to 5.2%; Figure 14). Interestingly, the pattern of baseline improvements is more or less copied by the pattern of improvements after the iterations, indicating that the performance of the iterative tuning was relatively insensitive to the various HRIL modifications. In other words, the initial improvement thanks to the OSM masking was mostly projected into the improvement after iteration and did not contribute significantly to better fit of the iterative tuning.

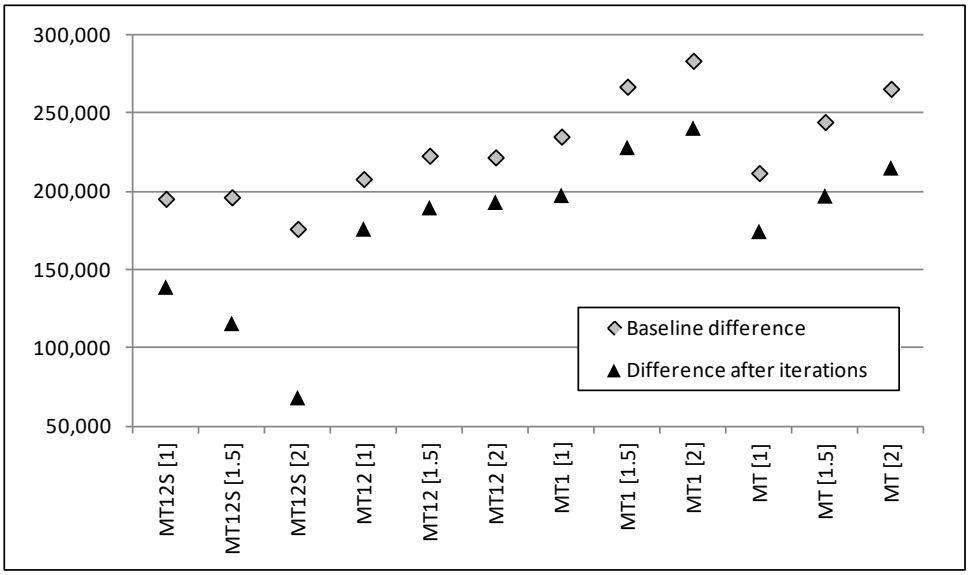


Figure 14: TAE difference between original HRIL and HRIL modifications

The best result was achieved by the HRIL modification MT1[2] – including railways, motorway, trunk and primary roads with two times exaggerated width (see Annex V and highlighted cells in Table 4). We suppose that including minor road categories was not helpful since these are seldom captured in isolation by HRIL, and elsewhere they can be found very close to residential buildings and might have a negative influence, e.g. if they intersect the same cell. Exaggeration of width is necessary, especially to cope with positional inaccuracies and perhaps spectral contamination of pixels adjacent to actual impervious pixels in the satellite imagery used to derive HRIL.

The results of disaggregation with HRIL, OSM and CLC ancillary data (reported in section 4.2) yielded RTAE in range between 38.10% and 63.11% (Table 5), depending on the subset of OSM segments used and exaggeration of their width, as well as on the concerned territory. In terms of RTAE, Slovenia scored generally much worse compared to Austria (and compared to the whole study area, since it is comprised mainly of Austria). The most likely reason for this disparity is the difference in the size of the source zones that we mentioned before (on average, the source zones in Slovenia were almost 3 times larger than in Austria), since it is one of the key factors influencing the accuracy of areal interpolation. As clarified by Sadahiro (2000), "as the proportion of source zones partially intersecting target zones increases or as the size of a target unit lying completely within a source unit diminishes, the interpolation task becomes progressively more demanding and prone to spatial prediction error."

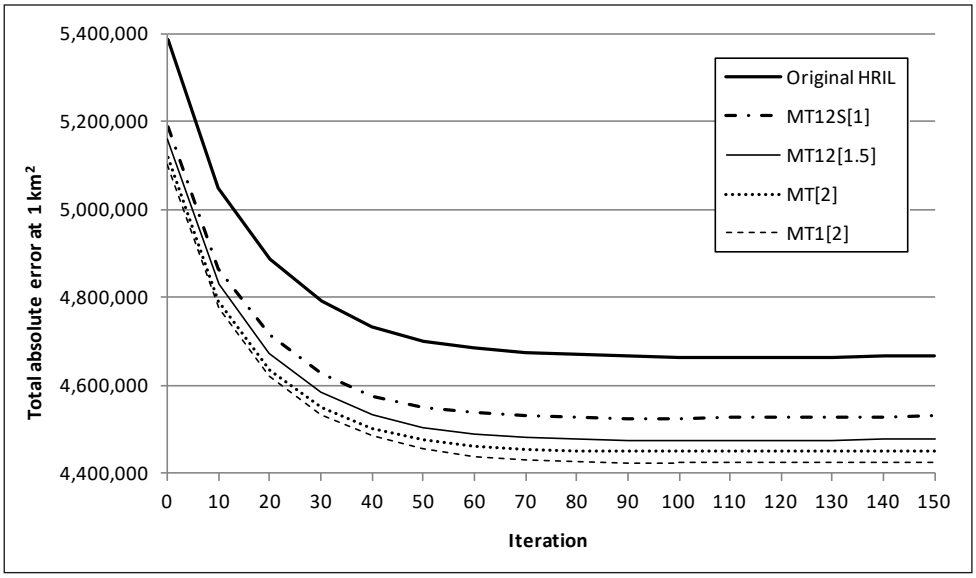


Figure 15: Reduction of TAE over the iterations for disaggregation based on the original HRIL and four examples of its modifications

More insight can be gained when the spatial variation of the estimation error is examined in detail. Annexes X and XI show maps of the estimated values (based on the MT1[2] model) and reference values respectively. For easier reading the datasets were first averaged over 8 km grid cells. A deviation grid (Annex XII) was calculated by simply subtracting the 8 km reference grid from the estimated 8 km grid. The occurrence of large absolute overestimations and underestimations is much more frequent in the lowland/urbanized parts of the study area since these have the most of the densely populated cells (see Annex XI), that are more errorprone in absolute terms. The fact that the size difference between source and target zones is usually smaller in these regions did not reverse this trend. Likewise, the extreme values of both absolute under- and overestimations are concentrated inside and around the largest cities with the highest population density. An interesting trend is common for the capitals Vienna and Ljubljana as well as for Linz, Salzburg, and Klagenfurt. Central parts of the cities tend to be underestimated while the peripheries are partly overestimated. This might be attributed to the fact that sheer amount of impervious surface can hardly approximate density gradients typical for cities. Central parts of cities differ from suburbs and peripheries in vertical development, floor area per capita ratio, and amount of non-residential impervious land (which was masked from IL imperfectly).

Another useful statistic is the relative deviation (Annex XIII), calculated as the ratio of absolute error and reference population count). A pattern almost opposite to the one described above emerges; the extreme relative underestimations and

overestimations are (with few exceptions) typical for the cells with the lowest population densities, located dominantly in the hilly regions. Cells with tiny populations are easily overestimated by more than 100%, even if the absolute number of misallocated residents is not significant. A huge size of some of the Alpine communes is a likely culprit, but the presence of ski resorts might also contribute. Number of tourists (and the amount of related built infrastructure) may greatly exceed the local permanent residents; however, a much closer inspection would be needed to confirm that. More exact decoupling of contributions of (a) heterogeneous size of the source zones and (b) inaccuracies/limitations of the ancillary data to the spatial distribution of the error indicators would be a very intriguing prospect for future research.

RTAE of the best result 42.99% (over 4.4 million people) may seem large, but compared to areal weighting RTAE of 112% (Table 6) it is a major improvement. For instance, Mrozinski and Cromley (1999) achieved 32% improvement whereas Reibel and Bufalino (2005) report improvement of only 20% when compared to areal weighting. Also, as we mentioned before, the fine resolution of the disaggregated grid should facilitate approximation of arbitrary (irregular, complex) polygons of "reasonable" size, rather than provide estimates for minuscule areas (such as individual 1 km pixels). Therefore, we examined also scale dependency of RTAE by repeating the validation at several coarser resolutions (Figure 16). Naturally, the error decreases with the spatial resolution at which it is measured. RTAE is monotonously decreasing with increasing area of the validation units, in a logarithmic fashion. At 8 km cells, it drops to 9% and gradually approaches the value of 4%. This kind of information may well support the appropriate application of the product, depending on the size of the zones and accuracy requirements.

Table 4: TAE of disaggregation based on OSM-modifed HRIL anc CLC data (in thousands of persons) and relative improvement of TAE compared to disaggregation based on non-modified HRIL

		Width exaggeration											
		1.0				1.5				2.0			
Road categories masked	TAE / Area validated	Baseline	Relative improvement	After iterations	Relative improvement	Baseline	Relative improvement	After iterations	Relative improvement	Baseline	Relative improvement	After iterations	Relative improvement
MT12S	AT+SI	5189	3.6%	4525	3.0%	5188	3.6%	4549	2.5%	5208	3.3%	4596	1.5%
	AT	3905	4.0%	3254	3.5%	3907	4.0%	3275	2.9%	3927	3.5%	3316	1.7%
	SI	1284	2.4%	1263	1.8%	1281	2.5%	1264	1.7%	1282	2.5%	1268	1.4%
MT12	AT+SI	5176	3.9%	4488	3.8%	5162	4.1%	4475	4.1%	5163	4.1%	4471	4.1%
	AT	3891	4.4%	3214	4.7%	3879	4.7%	3199	5.1%	3880	4.7%	3195	5.2%
	SI	1285	2.3%	1263	1.7%	1283	2.4%	1262	1.9%	1283	2.4%	1261	1.9%
MT1	AT+SI	5149	4.4%	4467	4.2%	5118	5.0%	4436	4.9%	5101	5.3%	4424	5.2%
	AT	3866	5.0%	3194	5.3%	3838	5.7%	3167	6.1%	3823	6.1%	3156	6.4%
	SI	1283	2.4%	1260	2.0%	1279	2.7%	1257	2.2%	1277	2.8%	1255	2.4%
MT	AT+SI	5173	3.9%	4490	3.7%	5140	4.5%	4467	4.2%	5119	4.9%	4449	4.6%
	AT	3890	4.4%	3224	4.4%	3862	5.1%	3200	5.1%	3844	5.6%	3184	5.6%
	SI	1283	2.4%	1258	2.1%	1278	2.8%	1256	2.3%	1275	3.0%	1254	2.5%

The MT1[2] modification was also the only result to slightly outperform the AIT grid, which is based on almost the same ancillary data, but different downscaling approach is used (Steinocher et al. 2011). On the other hand, all but one results outperformed the JRC 2006 grid (Batista e Silva et al. 2013), that used a version of CLC refined by HRIL, Urban Atlas, Tele Atlas and other spatial data (Batista e Silva et al. 2012). Thus this exercise can be considered successful, since the aim to exceed accuracy of comparable datasets was fulfilled, although more significant advance was expected. The most likely reason for this could be overly optimistic assumptions about quality of the ancillary data. Errors in ancillary data propagate to the resulting disaggregated map, and also hinder the fit of statistical models. The accuracy of HRIL was evaluated by producers but also by independent researches (Hurbánek et al. 2010). There is a concern that accuracy of the product may differ according to the company that processed the IMAGE2006 using semi-automated approach (each of five contractor companies processed a different group of countries, which may be another source of different results in Austria and Slovenia). CLC data, although produced on the basis of harmonised guidelines, may suffer from the fact that each country was interpreted by separate national team. Each of the teams might have interpreted the guidelines slightly differently (Buttner and Maucha 2006).

Table 5: Absolute and relative total absolute error of disaggregation based on HRIL modifications

Road categories masked	Country	Width exaggeration											
		1 ×				1.5 ×				2 ×			
		Baseline		After iterations		Baseline		After iterations		Baseline		After iterations	
		TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE
MT12S	AT+SI	5189	50.43%	4525	43.98%	5188	50.42%	4549	44.21%	5208	50.62%	4596	44.67%
	AT	3905	47.16%	3254	39.29%	3907	47.18%	3275	39.55%	3927	47.41%	3316	40.04%
	SI	1284	63.91%	1263	62.87%	1281	63.80%	1264	62.93%	1282	63.82%	1268	63.11%
MT12	AT+SI	5176	50.31%	4488	43.62%	5162	50.16%	4475	43.49%	5163	50.17%	4471	43.45%
	AT	3891	46.99%	3214	38.81%	3879	46.84%	3199	38.63%	3880	46.85%	3195	38.58%
	SI	1285	63.98%	1263	62.90%	1283	63.88%	1262	62.81%	1283	63.87%	1261	62.81%
MT1	AT+SI	5149	50.04%	4467	43.41%	5118	49.73%	4436	43.11%	5101	49.57%	4424	42.99%
	AT	3866	46.69%	3194	38.57%	3838	46.35%	3167	38.24%	3823	46.17%	3156	38.10%
	SI	1283	63.89%	1260	62.73%	1279	63.70%	1257	62.58%	1277	63.60%	1255	62.49%
MT	AT+SI	5173	50.27%	4490	43.63%	5140	49.95%	4467	43.41%	5119	49.75%	4449	43.24%
	AT	3890	46.97%	3224	38.93%	3862	46.64%	3200	38.64%	3844	46.41%	3184	38.45%
	SI	1283	63.85%	1258	62.65%	1278	63.63%	1256	62.55%	1275	63.49%	1254	62.44%

Table 6: Absolute and relative total absolute error of disaggregation based on original HRIL, AIT grid, JRC grid, and areal weighting

Country	Disaggregation with orignal HRIL				AIT grid 2006		JRC grid 2006		Areal weighting	
	Baseline		After iterations							
	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE
AT+SI	5384	52.33%	4664	45.33%	4432	43.07%	4585	44.56%	11,534	112.09%
AT	4070	49.14%	3372	40.72%	3183	38.44%	3281	39.62%	9197	111.06%
SI	1315	65.45%	1286	64.02%	1249	62.19%	1304	64.92%	2337	116.34%

4.3. Future work

Given the inconclusive improvements of disaggregation accuracy achieved above and the slight variation of accuracy among various parameter settings of the tested method, we assume that the potential of significant advance based on the data used here (open access ancillary data and municipal-level source data) is rather small. We suggest that further efforts in improving gridded population models should focus on using smaller source zones than the LAU2 system used here, or more appropriate and detailed ancillary data (e.g. residential building footprints with height information), or both. Fortunately, the production of bottom-up population data based on 1 km European reference grid is pursued strongly by Eurostat, thus the next obvious step should be changing the LAU2 for 1 km grid cells as the source zones, which would greatly improve accuracy of disaggregation into 100 m grid cells and facilitate transition to even finer target zones.

With the advent of the European Sentinel Earth observation satellites (especially the multispectral Sentinel 2), a 'flood' of ancillary data similar to those used here (but with higher spatial resolution, more frequent updates, and global coverage) can be anticipated. Another innovative dataset derived from remote sensing – the Global Human Settlement Layer (GHSL) – is being produced at the JRC. Given the fact that GHSL aims to portray only buildings (as opposed to imperious areas in HRIL), it should be even more suitable for population disaggregation.

Nevertheless, the presented results are relevant to anyone aiming at similar approach and can provide recommendations on several methodological aspects of spatial disaggregation. The principle of the iterative method using meaningful feature spaces for regional estimation of population density coefficients is not restricted to the current setting and can be used e.g. in disaggregation from 1 km grid cells into 50 m grid cells using building footprints of multiple categories.

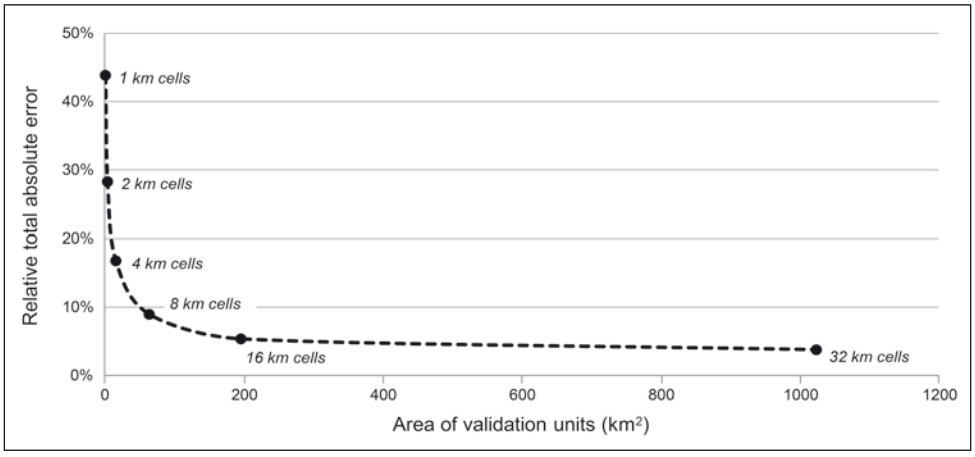


Figure 16. Effect of validation scale on magnitude of RTAE.

CONCLUSION

In this research, we aimed to improve existing method of intelligent spatial disaggregation, investigate how parameters of the method influence estimation accuracy in the particular setting, and create a fine resolution raster model of population density of reasonable accuracy. The setting here was defined by our focus on open access ancillary data and the European context, where the LAU2 units were used as the source zones, and the 100 m grid cells were used as the target zones.

The disaggregation was performed using an algorithm originally proposed by Gallego and Peedell (2001) for use with CLC as the only source of ancillary data. The method is a 'fixed ratio' method (Gallego et al. 2011), which assumes that ratios between population densities of land cover classes are constant either for the entire set of the source zones, or for a subset of the source zones (NUTS region in the original implementation). The goal was to find the ratios (i.e. population density coefficients) and tune their values to obtain as good performance as possible. We adjusted the original method so that HRIL (or its modifications) could be used along with CLC. Instead of disaggregation into all pixels covered by a patch of CLC class, we restricted only to non-zero HRIL pixels, assuming that if there is less than 1% of impervious surfaces in a 100 m square, there should not be any residences. Consequently, the variable of interest in the whole disaggregation process is not general population density, but population density per impervious area (e.g. 100 m HRIL pixel with value of 25 represents 2500 m^2 of impervious area).

The algorithm optimises coefficients as follows. First, the population is disaggregated with an initial set of coefficients (we used a uniform values for all CLC classes, thus disaggregating into pixels proportionally to the HRIL values) from the subsets of source zones into the target zones. Then the deviation from actual population counts is evaluated at the source zone level and the coefficients are modified so that the overall error of the estimate (total absolute deviation) is reduced. The modified coefficients are used to disaggregate the population from a subset of source zones again, deviations are evaluated, coefficients are modified and the whole process is repeated iteratively. When the overall error becomes stable, the fine-tuned coefficients are used to estimate

from source zone level to grid cells. As a result, population density coefficients differentiate between the CLC classes for each subset of source zones.

As a further adjustment of the method, we have proposed novel approach to optimization of source zone subsets used in the iterative algorithm for estimation of population density coefficients. Rather than using contiguous non-overlapping subsets of source zones (i.e. regions), a subset of a source zone c is defined as the m nearest source zones to the source zone c in an n-dimensional feature space, where the n source zone features (attributes) are obtained automatically from source and ancillary data. Such approach is beneficial because it enables the population density coefficients to be unique for each source zone (optimized on the basis of a unique underlying source zone subset) and assumes the same set of coefficients for source zones that are similar in relevant aspects.

An extensive sensitivity analysis of method's parameters performed in Austria and Slovenia showed that the accuracy of the estimate can be indeed improved by meaningful design of the source zone subsets and that there is no point in confining to conventional regions (spatially contiguous and disjoint subsets). The best results were achieved using spatially non-contiguous subsets having around 3000 km^2 in area, based on similarity of three features of individual source zones: population size, share of impervious area in total area, and population density per impervious area. The same configuration of the subsets was also used in this study to assess the accuracy improvement thanks to incorporation of OSM roads and railways.

The best result was achieved by the HRIL modification including OSM railways, motorway, trunk and primary roads with two times exaggerated width. This outcome, displayed in Annex V, exceeded comparable products that were in existence (JRC grid and AIT grid) in terms of overall accuracy. The fact that JRC grid used richer ancillary data compared to our approach points at better performance of the proposed method. However, the achieved increase in accuracy was only minor, suggesting that the potential to further advance the modelling accuracy based on the specific ancillary and source data is exhausted and that better sources of data should be investigated in future.

"Nonetheless, the presented findings can provide recommendations on several methodological aspects of spatial disaggregation. We think that some of the results could be useful particularly for the production of population grids in those countries that are covered by HRIL and CLC data (i.e. 38 members of EEA, including non-EU countries) and have rather limited availability of accurate, spatially detailed and homogeneous national datasets. Furthermore, as updates are planned for IL and CLC data on a 3-year and 6-year basis, respectively (the last update of both datasets was in 2012), the presented methodology is suitable for disaggregation of population data from 2011 census and possibly beyond" (Rosina, Hurbánek and Cebecauer 2016, p. 11).

RESUMÉ

Úvod

V geografickom výskume, ako aj v ďalších oblastiach výskumu a praxe (rozhodovacie procesy, územné plánovanie, krízové riadenie atď.), je často potrebné analyzovať údaje o obyvateľstve na základe územných jednotiek nekompatibilných s dostupnými jednotkami, alebo sú potrebné informácie o rozmiestnení obyvateľstva s podrobnejším priestorovým rozlíšením. Medzi takéto oblasti patrí základný výskum interakcií obyvateľstva a ďalších zložiek krajiny, priestorové plánovanie využívania krajiny, dopravnej a inej infraštruktúry, spravodlivá a efektívna alokácia služieb, krízové riadenie, hodnotenie dostupnosti, rizika a vplyvov na obyvateľstvo. V decíznej sfére sú podrobné údaje o rozmiestnení obyvateľstva potrebné na tvorbu a adresovanie politík (napr. definícia a charakterizácia funkčných mestských regiónov Organizácie pre ekonomickú spoluprácu a rozvoj, mestsko-vidiecka typológia obcí „DEGURBA" a typológia regiónov NUTS – Európska komisia).

Dáta o obyvateľstve a jeho charakteristikách sú obvykle dostupné agregované do administratívnych, resp. štatistických jednotiek (najmenšie z bežne dostupných sú obce, príp. základné sídelné jednotky – napríklad na Slovensku a v Česku). Nevýhodami takejto priestorovej reprezentácie sú nedostatočná priestorová podrobnosť, nerovnaká veľkosť jednotiek, zmeny jednotiek v čase, priebeh hraníc nesúvisiaci s rozmiestnením skúmaného javu a problém modifikovateľnej územnej jednotky (situácia, keď je výsledok štatistickej analýzy premenných závislý od použitých priestorových jednotiek, teda keď mierka a spôsob agregácie významne ovplyvňujú zistený charakter a tesnosť vzťahov a závislostí medzi týmito premennými, ktoré preto nie je možné zovšeobecniť).

Rozšírenie geografických informačných systémov a nových technológií zberu a spracovania presne georeferencovaných demografických údajov umožnilo alternatívny spôsob reprezentácie údajov formou pravidelných priestorových jednotiek – štvorcových buniek rastra s komparatívne vyšším priestorovým rozlíšením (štvorce so stranou 1 km alebo menšie). Výhodou podrobných pravidelných rastrov je ich jednoznačná definícia, jednoduchá manipulácia v prostredí GIS a univerzálnosť – je možné z nich

jednoducho odvodiť hodnoty pre ľubovoľné polygóny, takže jediná databáza je užitočná pre široký okruh používateľov.

Rastrovú reprezentáciu údajov o rozmiestnení obyvateľstva možno získať prístupom zdola nahor (bottom-up), čiže agregáciou individuálnych cenzových záznamov so známymi geografickými súradnicami, alebo ju odhadnúť pomocou areálovej interpolácie údajov. Metódy areálovej interpolácie údajov umožňujú z dostupných zdrojových jednotiek odhadnúť údaje pre ľubovoľné cieľové jednotky (napr. bunky pravidelných rastrov s vysokým priestorovým rozlíšením); ak sú cieľové jednotky menšie ako zdrojové, hovoríme o priestorovej dezagregácii alebo o prístupe „zhora nadol" (top-down).

Jednoduché metódy dezagregácie nevyužívajú iné priestorové dáta ako zdrojové jednotky (napr. areálové váženie). „Inteligentné" (niekedy nazývané dazymetrické) metódy dezagregácie využívajú pomocné priestorové údaje na presnejšiu lokalizáciu obyvateľov v rámci zdrojovej jednotky, za ktorú sú dostupné agregované údaje (obec, okres a pod.). Vďaka pomocným údajom sú obvykle presnejšie ako jednoduché. K často používaným pomocným údajom patria napríklad dáta o využití krajiny a krajinnej pokrývke (najmä o zastavaných a rezidenčných plochách indikujúcich rozmiestnenie obyvateľstva podľa trvalého, resp. obvyklého bydliska), cestná sieť, satelitné snímky (napr. nočného osvetlenia) a iné zdroje údajov.

Tvorba populačných rastrov agregáciou údajov zdola má oproti odhadu dezagregáciou výhodu pravdivosti, resp. presnosti údajov. Tento prístup však zatiaľ zostáva relatívne málo rozšírený, najmä z celosvetového pohľadu. Naopak, dezagragácia údajov síce vždy podlieha chybe (ide o modelovanie, odhad), avšak je užitočným nástrojom v širšom spektre situácií. Pri splnení určitých predpokladov umožňuje tvorbu porovnateľných (harmonizovaných) rastrových databáz na medzinárodnej úrovni, národných rastrových databáz v krajinách, kde sa zatiaľ neuplatňujú prístupy „zdola nahor" (napr. Slovensko do roku 2015 a mnohé krajiny Európy v súčasnosti), retrospektívnych databáz pre roky, keď ešte neboli uplatňované prístupy „zdola nahor", modelov s alternatívnou časovou konceptualizáciou hustoty zaľudnenia (rezidenčná vs. „ambientná", denná/nočná, 24/7 a pod.) a tiež napr. zvýšenie rozlíšenia dostupných agregovaných populačných rastrov. Výhodou je aj to, že pri dezagregácii nie je potrebné riešiť problémy ochrany dôverných údajov, ktoré sú spojené s tvorbou podrobných rastrov prístupom zdola nahor.

Ciele

V predkladanej práci si dávame za cieľ poskytnúť prehľad poznatkov teoretického a metodologického charakteru, relevantných pre riešenie problematiky priestorovej dezagregácie so zameraním na hustotu zaľudnenia a následne vytvoriť rastrový model rozmiestnenia obyvateľstva s priestorovým rozlíšením 100 m dezagregáciou počtu obyvateľov za obce, pričom na spresnenie priestorového rozmiestnenia použijeme voľne dostupné pomocné priestorové údaje. Cieľom je vylepšiť vybranú existujúcu metódu inteligentnej priestorovej dezagregácie (z hľadiska pomocných údajov, prijatých predpokladov, ako aj

algoritmu dezagregácie) a porovnaním výsledkov viacerých konfigurácií metódy zistiť, ako niektoré jej parametre ovplyvňujú presnosť odhadu. Vybraná metóda aj pomocné priestorové údaje by mali byť aplikovateľné na celoštátnej až európskej úrovni.

Naplnením uvedených cieľov chceme tiež overiť, či nasledovné hypotézy, vychádzajúce zo štúdia literatúry, budú platné aj pri skúmanej konfigurácii problému priestorovej dezagregácie:

- presnosť odhadu je možné zvýšiť zlepšením niektorých aspektov pomocných priestorových údajov (počet údajových vrstiev, vhodnosť, priestorové rozlíšenie);
- presnosť odhadu je možné zvýšiť modifikáciou (vylepšením) algoritmu metódy dezagregácie;
- zlepšenie pomocných údajov prináša významnejšie spresnenie odhadu v porovnaní so zlepšením algoritmu metódy.

Metodika

V predloženej štúdii sme pracovali s iteratívnym dezagregačným algoritmom, pôvodne navrhnutým na použitie s údajmi o krajinnej pokrývke CORINE (Gallego a Peedell 2001, Gallego 2010) a prispôsobili sme ho tak, aby mohli byť využité aj dodatočné zdroje pomocných údajov. Ďalej sme vylepšili algoritmus novým prístupom k regionálnemu odhadovaniu koeficientov hustoty zaľudnenia – predmetné regióny sme nedefinovali v konvenčnom geometrickom priestore, ale v n-rozmerných príznakových priestoroch odvodených z tých istých pomocných údajov, ktoré sme využili na priestorové spresnenie dezagregácie.

Skúmané územie a údaje

Vzhľadom na potrebu kvalitných referenčných údajov o obyvateľstve s vysokým rozlíšením pre zhodnotenie presnosti jednotlivých variant metód a tiež porovnanie s existujúcimi dezagregovanými produktmi sme uprednostnili ako skúmané územie Rakúsko a Slovinsko, kde boli takéto údaje dostupné (analýza sa vzťahovala na rok 2006, pre validáciu bolo teda potrené použiť údaje k tomu istému roku).

Ako zdrojové údaje o počte obyvateľov sme použili údaje z roku 2006 na úrovni jednotiek LAU2 (local administrative units level 2 v klasifikácii EuroStatu), resp. obcí (rakúske Gemeinde a slovinské občine) v celkovom počte jednotiek 2 549. Ako pomocné priestorové údaje pre alokáciu obyvateľov zo zdrojových do cieľových jednotiek sme použili CORINE Land Cover 2006 a európsku vrstvu nepriepustných povrchov HRIL 2006 (High Resloution Impervousness Layer), ktorá bola upravená pomocou údajov o cestnej a železničnej sieti z projektu OpenStreetMap (OSM). Presnosť výsledkov dezagregácie sme kvantifikovali porovnaním odhadnutých počtov obyvateľov s

údajmi z databázy GeoStat 2006 na úrovni 1 km buniek rastra, spolu viac ako 100 tis. validačných jednotiek.

Iteratívna metóda odhadu koeficientov hustoty zaľudnenia

Na odhad koeficientov hustoty zaľudnenia na zastavanú plochu bola použitá upravená metóda dezagregácie J. Gallega, ktorá pracovala s predpokladom fixného pomeru hustoty zaľudnenia medzi jednotlivými triedami krajinnej pokrývky v rámci NUTS2 regiónov (napr. pre všetky obce v regióne r platí, že pomer hustoty zaľudnenia medzi urbanizovanými areálmi, poľnohospodárskou pôdou, pasienkami, lesmi a vodnými plochami je 30:5:2:1:0). Výpočet finálnych koeficientov prebieha iteratívnym spôsobom:

1. stanovia sa počiatočné hodnoty koeficientov;
2. pomocou týchto koeficientov sa dezagregujú údaje, ako zdrojové jednotky sa však nepoužijú najnižšie dostupné jednotky (obce), ale ich skupiny (regióny, okresy alebo iné podmnožiny obcí);
3. z výsledného rastra sa spätne agregujú hodnoty za obce, porovnajú sa so skutočnými hodnotami a vypočítajú sa indikátory chyby;
4. koeficienty sú upravené tak, aby sa znížila chyba;
5. kroky 1 – 4 sa opakujú, pokým sa koeficienty nestabilizujú;
6. konečné hodnoty koeficientov sa použijú na dezagregáciu z úrovne najnižších zdrojových jednotiek (obcí) do buniek rastra.

Vylepšenie iteratívnej metódy

Pôvodná implementácia využívala pri odhade koeficientov hustoty zaľudnenia NUTS regióny. Tento predpoklad sme považovali za nerealistický, keďže administratívne regióny zvyčajne obsahujú heterogénne typy obcí, od veľkomiest až po najmenšie vidiecke obce, ktoré vykazujú rôzne priestorové vzorce krajinnej pokrývky a hustoty zaľudnenia. Taktiež nie je dôvod, aby boli použité zoskupenia obcí limitované administratívnymi hranicami, či už jednotiek NUTS alebo krajín.

Použitie regiónov je síce potrebné v iteratívnom procese ladenia koeficientov, avšak namiesto použitia administratívnych jednotiek možno vytvoriť unikátny región pre každú obec, takže každej obci môže byť priradený unikátny súbor koeficientov. Hodnoty vypočítaných koeficientov potom môžu varírovať v priestore kontinuálne a nebudú sa meniť skokovo na arbitrárne určených hraniciach.

Priestorovo kompaktné regióny nemusia byť pre daný účel najvhodnejšie. Ďalším navrhovaným vylepšením metódy je vyčleniť regióny (resp. podmnožiny obcí) na základe podobnosti iných charakteristík, ktoré je možné odvodiť priamo z použitých pomocných priestorových údajov (napr. štruktúra krajinnej pokrývky, počet obyvateľov, podiel

nepriepustných povrchov a pod.). Z uvedených predpokladov vyvstáva otázka, ako čo najlepšie definovať podmnožiny obcí, na základe ktorých budú prebiehať výpočty koeficientov. V rámci citlivostnej analýzy A sme testovali a porovnali 150 parametrizácií dezagregačného algoritmu s cieľom zodpovedať tieto otázky:

- Na základe ktorých atribútov obcí by mala byť identifikovaná podmnožina podobných obcí (t. j. najbližších obcí v euklidovskom n-rozmernom príznakovom priestore)?
- Aká by mala byť veľkosť podmnožín? Mala by byť určená celkovou rozlohu alebo počtom obcí (alebo príp. podľa iného pravidla)?
- Koľko iterácií je potrebné vypočítať, aby hodnoty koeficientov konvergovali k určitej hodnote?

Vylepšenie pomocných priestorových údajov

Kombinácia údajov HRIL 2006 a CLC 2006

Dátová vrstva nepriepustných povrchov HRIL 2006 bola prvou zo série monotematických rastrových vrstiev poskytovaných v rámci programu Copernicus (na rozdiel od CORINE nejde o kategorickú mapu, hodnoty predstavujú kontinuálnu premennú – percentuálne zastúpenie danej triedy v rámci bunky). Keďže nepriepustné povrchy sú obvykle asociované s ľudskými aktivitami, HRIL predstavuje vhodný pomocný zdroj priestorových údajov, ktorý sa vzájomne dopĺňa s údajmi CORINE. HRIL poskytuje vyššie priestorové rozlíšenie (1 ha oproti 25 ha v prípade CORINE), avšak nerozlišuje medzi rôznymi typmi zástavby. CORINE obsahuje 44 tried, z ktorých niektoré možno využiť na identifikáciu nerezidenčnej zástavby (najmä triedy priemyselných, obchodných a dopravných areálov). Pre kombináciu oboch vrstiev pri dezagregácii hustoty zaľudnenia sme prijali tieto základné zjednodušujúce predpoklady:

1. priestorové dáta použité na dezagregáciu odrážajú objektívnu realitu s dostatočnou presnosťou;
2. obyvateľstvo je prítomné výlučne v bunkách HRIL obsahujúcich nepriepustné (zastavané) povrchy, čiže v bunkách s hodnotou 1 – 100. Bunky s hodnotou 0 sú vylúčené z analýzy (pozri prílohu VI);
3. hodnota miery nepriepustnosti údajov HRIL reprezentuje percento nepriepustnej plochy v príslušnej bunke;
4. hustota zaľudnenia na zastavanú plochu je konštantná v každej kombinácii obce a triedy CLC;
5. pomer hustôt zaľudnenia v každej dvojici tried CLC je daný koeficientmi, ktoré sú konštantné pre všetky obce v určitej skupine obcí (t. j. priestorovo kompaktnom regióne alebo inak definovanej podmnožine obcí).

Modifikácia vrstvy HRIL pomocou údajov OSM

Nevýhodou dátovej vrstvy HRIL pri priestorovej dezagregácii rezidenčnej hustoty zaľudnenia je, že zachytáva všetky nepriepustné povrchy – nielen budovy, ale aj iné povrchy pokryté umelým materiálom (betón, asfalt a pod.) vrátane úsekov ciest, železníc, parkovísk, námestí a pod. Tento problém je možné čiastočne odstrániť použitím modelu cestnej a železničnej siete. Podobný prístup bol navrhovaný napr. v prácach Batistu et al. (2013) a Steinochera et al. (2011).

Vďaka voľne dostupným údajom o cestnej a železničnej sieti z projektu OpenStreetMap je možné modifikovať vrstvu HRIL tak, aby bola vernejšou reprezentáciou rezidenčnej zástavby – hodnotu každej bunky rastra HRIL je potrebné znížiť o podiel rozlohy bunky pokrytý cestami/železnicami.

Tento zámer sme realizovali nasledovným postupom:

- výber vhodných kategórií ciest a železníc;
- odhad priemernej šírky ciest jednotlivých kategórií (keďže údaje o šírke úsekov nie sú súčasťou OSM, táto bola určená meraním náhodnej vzorky úsekov na ortofotomape);
- konverzia vybraných polylínií na polygóny pomocou bufferu príslušnej šírky;
- výpočet plochy prekrytia medzi buffer polygónmi a jednotlivými bunkami HRIL;
- konverzia relatívnej hodnoty prekrytia do rastrovej reprezentácie;
- odčítanie rastrovej reprezentácie ciest a železníc od pôvodného HRIL rastra pomocou mapovej algebry (pozri prílohu VIII);
- pozn.: predpokladali sme, že presnosť odhadu sa môže zvýšiť, ak budú namerané šírky úsekov nadhodnotené (v snahe kompenzovať polohovú nepresnosť údajov, efekt kontaminácie spektrálnej odozvy priľahlých pixelov v zdrojových satelitných údajoch a odchýlky skutočných hodnôt šírky cesty od nameranej priemernej šírky v danej kategórii).

V rámci citlivostnej analýzy B sme testovali 4 kombinácie kategórií ciest a 3 úrovne nadhodnotenia ich šírky (t. j. 12 verzií modifikovanej vrstvy HRIL 2006), s cieľom zistiť odpovede na nasledovné otázky:

- Ktoré úseky ciest a železníc je potrebné vybrať a eliminovať z HRIL (diaľnice, cesty prvej, druhej, tretej triedy, miestne komunikácie, obslužné cesty atď.) tak, aby sa maximalizovala presnosť odhadu?
- Do akej miery má byť šírka úsekov nadhodnotená oproti nameraným hodnotám?

Výsledky a diskusia

Výsledky analýzy A

- Presnejšie a robustnejšie boli varianty dezagregácie narábajúce s nesúvislými prekrývajúcimi sa podmnožinami obcí (vytvorené na základe podobnosti niektorých vlastností odvodených zo zdrojových a pomocných údajov);
- Pre celé skúmané územie najpresnejší výsledok pri použití podmnožín zdrojových jednotiek s rozlohou cca 3 000 km^2 vytvorených na základe podobnosti počtu obyvateľov, podielu zastavanej plochy na celkovej rozlohe a hustote zaľudnenia na zastavanú plochu;
- Veľkosť zdrojových zón významne ovplyvňuje presnosť a môže mať vplyv aj na ďalšie aspekty dezagregácie (iný výsledok v Slovinsku).

Výsledky analýzy B

- Najlepší výsledok modifikácie pomocou OSM údajov bol dosiahnutý pri použití najvýznamnejších kategórií ciest (motorway, trunk, primary), všetkých úsekov železníc a dvojnásobnom nadhodnotení nameranej šírky úsekov;
- Všetky testované kombinácie zlepšili presnosť odhadu oproti pôvodnej verzii HRIL;
- Zlepšenie na území Rakúska bolo výraznejšie ako v Slovinsku (6,4 % vs. 2,5 %), pravdepodobne z dôvodu rozdielnej veľkosti zdrojových jednotiek.

Prínos práce

- Nový spôsob kombinácie CLC a HRIL s cieľom dezagregácie populačných údajov;
- Nový spôsob odvodenia mier nepriepustnosti na báze rastra z lineárnych údajov OpenStreetMap (vrátane analýzy citlivosti na viaceré parametre);
- Nový spôsob tvorby podmnožín zdrojových zón na základe informácií získaných z pomocných údajov (vrátane analýzy citlivosti na viaceré parametre);
- Vytvorenie metodiky aplikovateľnej na ľubovoľnú krajinu EÚ narábajúcu s voľne dostupnými údajmi, vrátane príslušného algoritmu a programového modulu;
- Vytvorenie rastra hustoty zaľudnenia za Rakúsko a Slovinsko použiteľného na reálne aplikácie.

Záver

Výsledky dokázali, že zlepšenie pomocných údajových vrstiev (resp. úprava surových údajov do vhodnejšej podoby) podstatne zlepšuje presnosť odhadu, napr. začlenenie OSM údajov (hoci len obmedzeného výberu objektov) znížilo celkovú chybu

o viac ako 5 %. Pre porovnanie, vylepšenie samotného mechanizmu metódy zlepšilo presnosť odhadu v menšej miere, čo je v súlade s predpokladmi. Taktiež v súlade s predpokladmi výsledky ukazujú, že pri odhade koeficientov hustoty zaľudnenia môže použitie susedstiev v zmysluplných príznakových priestoroch zlepšiť odhad v porovnaní s klasickými regiónmi.

Najlepšie výstupy mierne prekonali (pokiaľ ide o sumu absolútnych chýb) porovnateľné existujúce databázy vytvorené odlišnými metódami na základe podobných údajov, avšak nešlo o významné zlepšenie. Podobne, nevýrazná variácia presnosti medzi jednotlivými nastaveniami parametrov naznačuje, že presnosť dosiahnuteľná kombináciou použitých voľne dostupných pomocných údajov a zdrojových jednotiek na úrovni obcí je pomerne limitovaná. Domnievame sa, že možnosti ďalšieho výraznejšieho zvyšovania presnosti odhadu vylepšovaním mechanizmu metódy je obmedzené – je skôr potrebné zvýšiť priestorové rozlíšenie (a/alebo polohovú a tematickú presnosť) zdrojových a/alebo pomocných údajov, príp. počet zdrojov pomocných údajov.

Ďalšie smerovanie výskumu

Počas ostatných dvoch rokov boli publikované nové zdroje údajov, ktorých dostupnosť ovplyvňuje aj smery ďalšieho výskumu v oblasti priestorovej dezagregácie údajov o obyvateľstve. Metódu popísanú v predloženej publikácii je možné aplikovať na územie Slovenska a iných krajín Európy pre referenčný rok 2011 (zdrojové údaje o obyvateľstve z cenzu), resp. 2012 (dostupnosť pomocných priestorových údajov HRIL 2012 a CLC 2012). Avšak ešte vhodnejším zdrojom pomocných údajov (v porovnaní s údajmi HRIL) je dátová vrstva Global Human Settlement Layer (GHSL), ktorá zachytáva zástavbu budov, a tak nie je potrebné eliminovať cestné a železničné úseky, ako bolo popísané vyššie.

Z výsledkov sčítania obyvateľov, domov a bytov na Slovensku z roku 2011 bola prístupom zdola nahor vytvorená databáza počtu obyvateľov na báze buniek rastra s rozlíšením 1 km. Tento zdroj údajov predstavuje príležitosť pre priestorovú dezagregáciu do podrobnejších rastrov (napr. 100 m), a to najmä v kombinácii s podrobnými zdrojmi pomocných údajov dostupnými na národnej úrovni. Napr. databáza ZB GIS obsahuje vrstvu pôdorysov budov, vrátane údajov o ich výške a funkcii, čo umožňuje výrazné spresnenie odhadu.

Zatiaľ málo rozvinutou oblasťou výskumu je tvorba denných, nočných alebo inak časovo poňatých (ambientných, 24/7, a pod.) modelov hustoty zaľudnenia, ktorá je mimoriadne náročná z hľadiska vstupných údajov aj metód. V súčasnosti sa tejto problematike na európskej úrovni venuje exploratívny projekt ENACT realizovaný v Spoločnom výskumnom centre (JRC) Európskej Komisie (Batista e Silva et al. 2015).

REFERENCES

AHAS, R., AASA, A., SILM, S., AUNAP, R., KALLE, H., & MARK, Ü. (2007). Mobile Positioning in Space-Time Behaviour Studies: Social Positioning Method Experiments in Estonia. *Cartography and Geographic Information Science*, 34(4), 259–273.

AHAS, R., AASA, A., ROOSE, A., MARK, Ü., & SILM, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486.

AHOLA, T., VIRRANTAUS, K., KRISP, J. M., & HUNTER, G. J. (2007). A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science,* 21(8), 935–953.

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

BALK, D.L., U. DEICHMANN, G. YETMAN, F. POZZI, S. I. HAY, AND A. NELSON. 2006. Determining Global Population Distribution: Methods, Applications and Data. *Advances in Parasitology* 62, 119-156.

BARBOSA, P., CAMIA, A., KUCERA, J., LIBERTA, G., ET AL. (2008). Assessment of forest fire impacts and emissions in the European Union based on the European Forest Fire Information System. *Developments in Environmental Science*, 8, 197–208.

BATISTA E SILVA, F., LAVALLE, C., & KOOMEN, E. (2012). A procedure to obtain a refined European land use/cover map. *Journal of Land Use Science*, 8(3), 255–283.

BATISTA E SILVA, F., GALLEGO, J., & LAVALLE, C. (2013). A high-resolution population grid map for Europe. *Journal of Maps*, 9(1), 16–28.

BATISTA E SILVA, F., FREIRE, S., CRAGLIA, M., & LAVALLE, C. (2015). The ENACT project: towards spatiotemporal activity and population mapping in Europe. European Forum for Geography and Statistics Conference, Vienna, 10–12 November 2015.

BEZÁK, A. (2008). Quo vadis kvantitativna geografia? Úvahy o súčasnom smerovaní kvantitatívnej geografie. *Acta Geographica Universitatis Comenianae*, 50, 79-94.

BHADURI, B., BRIGHT, E., COLEMAN, P., & DOBSON, J. (2002). LandScan: Locating people is what matters. *Geoinformatics*, 5(2), 34–37.

BHADURI, B., BRIGHT, E., COLEMAN, P., & URBAN, M. L. (2007). LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1-2), 103–117.

BIELECKA, E. (2005). A dasymetric population density map of Poland. In *Proceedings of the International Cartographic Conferenc*e, 2005, July 9-15, A Coruña, Spain.

BOSSARD, M., FERANEC, J., & OŤAHEĽ, J. (2000). CORINE land cover technical guide – Addendum 2000. Technical report, 40. Copenhagen: European EnvironmentAgency. Available online (accessed 25 May 2015), http://www.eea.europa.eu/publications/tech40add

BRACKEN, I. & MARTIN, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A*, 21:537–43.

BÜTTNER, G., & MAUCHA, G. (2006). The thematic accuracy of Corine land cover 2000. Assessment using LUCAS (land use/cover area frame statistical survey). Technical report, 7. Copenhagen: European Environment Agency. http://reports.eea.europa.eu/technical_report_2006_7/en.

COLEMAN, D. J., GEORGIADOU, Y., & LABONTE, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4(1), 332-358.

CROMLEY, R. G., HANINK, D. M., & G. C. BENTLEY. (2012). A Quantile Regression Approach to Areal Interpolation. *Annals of the Association of American Geographers*, 102, 763–777.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B,* 39, 1–38.

DIJKSTRA, L. & POELMAN, H. (2008). Remote rural regions: how proximity to a city influences the performance of rural regions. Regional Focus n.1, 2008, EC-DG REGIO.

DOBSON, J. E., BRIGHT, E. A., COLEMAN, P. R., DURFEE, R. C., & WORLEY, B. A. (2000). LandScan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7), 849–857.

ECONOMIC AND SOCIAL RESEARCH COUNCIL (2012). Population 24/7: space-time specific population surface modeling. Available on-line (accessed 30 August 2012): http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1811/read

EFGS (2012). GEOSTAT 1A – Representing Census data in a European population grid. Final Report. 82p. Available online (accessed 30 August 2012), http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco_Geographical_information_maps/documents/ESSnet%20project%20GEOSTAT1A%20final%20report_0.pdf

EICHER, C. L. & BREWER, C. A. (2001). Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science*, 28(2), 125–138.

EuroGeographics (2015). License our products: EuroGeographics product pricing. Available online (accessed 25 May 2015), http://www.eurogeographics.org/products-and-services/license-our-products

EUROSTAT (2012). A revised degree of urbanisation classification (LAU2 level). Available on-line (accessed 16 May 2015), http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_DEGURBA

EUROSTAT (2014). Urban-rural typology update. Statistics in focus 16/2013, Regional statistics team. ISSN: 2314-9647. Available online (accessed 16 May 2015), http://ec.europa.eu/eurostat/statistics-explained/index.php/Urban-rural_typology_update

FISHER, P. F. & LANGFORD M., (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27(2), 211–224.

FISHER, P. F. & LANGFORD, M. (1996). Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping. *The Professional Geographer*, 48(3), 299–309.

FLOWERDEW, R., AND M. GREEN. (1989). Statistical Methods for Inference between Incompatible Zonal Systems. In M. F. Goodchild and S. Gopal (Eds.), *Accuracy of Spatial Databases*, 239–48, London: Taylor & Francis.

FLOWERDEW, R. & M. GREEN. (1992). Developments in Areal Interpolation Methods and GIS. *Annals of Regional Science*, 26, 67–78.

FLOWERDEW, R. & M. GREEN. (1994). Areal Interpolation and Types of Data. In S. Fotheringham and P. Rogerson (Eds.), *Spatial Analysis and GIS*, 121–45, London: Taylor and Francis.

FOTHERINGHAM, A. S. & WONG, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23, 1025- 1044.

GALLAY, M., KAŇUK, J., & HOFIERKA, J. (2014). Capacity of photovoltaic power plants in the Czech Republic. *Journal of Maps*, 11(3), 480-486.

GALLEGO, J. & PEEDELL, S. (2001). Using CORINE Land Cover to map population density. In *Towards Agrienvironmental indicators, Topic report 6/2001*. Copenhagen: European Environment Agency, pp. 92–103.

GALLEGO, F. J. (2010). A population density grid of the European Union. *Population and Environment*, 31(6), 460–473. doi:10.1007/s11111-010-0108-y

GALLEGO, F., BATISTA, F., ROCHA, C., & MUBAREKA, S. (2011). Disaggregating population density of the European Union with CORINE land cover. *International Journal of Geographical Information Science*, 25(12), 37–41.

GARB, J.L., CROMLEY, R.G., & ANDWAIT, R.B. (2007). Estimating populations at risk for disaster preparedness and response. *Journal of Homeland Security and Emergency Management*, 4, 1–17.

GEHLKE, C. & BIEHL, H. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association Supplement*, 29, 169–170.

GOODCHILD, M. (2011). Scale in GIS: An overview. *Geomorphology*, 130(1-2), 5–9.

GOODCHILD, M. F., YUAN, M., & COVA, T. J. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21(3), 239–260.

GOODCHILD, M. & LAM, N. (1980). Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, 1, 297–312.

GOODCHILD, M., ANSELIN, L., & DEICHMANN, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25, 383–397.

GOODCHILD, M.F., EGENHOFER, M.J., KEMP, K.K., MARK, D.M., & SHEPPARD, E. (1999). Introduction to the Varenius project. *International Journal of Geographical Information Science*, 13, 731–745.

GREEN, M. & FLOWERDEW, R. (1996). New evidence on the modifiable areal unit problem. In Longley, P. and Batty, M. (Eds.), *Spatial Analysis Modelling in a GIS Environment*. Cambridge: Wiley.

HAKLAY, M. & WEBER, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7(4), 12–18.

HAKLAY M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4), 682–703.

HAY, S. I., NOOR, A. M., NELSON, A., & TATEM, A. J. (2005). The accuracy of human population maps for public health application. *Tropical Medicine and International Health*, 10, 1073–1086.

HECHT R., KUNZE C., & HAHMANN, S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS International Journal of Geo-Information*, 2(4), 1066–1091.

HEYMANN, Y., STEENMANS, CH., CROISSILLE, G., & BOSSARD, M. (1994). CORINE land cover. Technical guide. Luxembourg: Office for Official Publications European Communities.

HOLT, J. B., LO, C. & HODLER, T. W. (2004). Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartography and Geographic Information Science*, 31, 103–21.

HURBÁNEK, P., ATKINSON, P. M., PAZÚR, R., ROSINA, K., & CHOCKALINGAM, J. (2010). Accuracy of built-up area mapping in Europe at varying scales and thresholds. In: N. J. Tate and P. F. Fisher (eds.): *Accuracy 2010: Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. 20-23 July 2010, University of Leicester, ISARA, 385–388.

IISAKA, J. & HEGEDUS, E. (1982). Population estimation from Landsat imagery. *Remote Sensing of Environment*, 12, 259–272.

JOKAR ARSANJANI, J., HELBICH, M., BAKILLAH, M., HAGENAUER, J., & ZIPF, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, 27(12), 2264-2278.

KIM, H. & YAO, X. (2010). Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, 31(21), 5657–5671.

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

KUSENDOVÁ, D., ROSINA, K., & HURBÁNEK, P. (2014). Mestsko-vidiecka klasifikácia základných územných jednotiek na báze analýzy rozmiestnenia obyvateľstva s použitím populačného rastra s veľkosťou bunky 1x1 km. In J. Buček et al., *Vymedzenie miest a mestských regiónov na Slovensku, vrátane zhodnotenia funkčných typov sídiel, na báze hustoty obyvateľstva, s použitím rastrovej technológie*. Ministerstvo dopravy, výstavby a regionálneho rozvoja Slovenskej republiky a Univerzita Komenského, Prírodovedecká fakulta, 64–81.

KYRIAKIDIS, P. C. (2004). A Geostatistical Framework for Area-to-Point Spatial Interpolation. *Geographical Analysis*, 36(3), 259–289.

LANGFORD, M. & UNWIN, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. *Cartographic Journal*, 31(1), 21–26.

LANGFORD, M. (2006). Obtaining Population Estimates in Non-Census Reporting Zones: An Evaluation of the 3-Class Dasymetric Model. *Computers, Environment and Urban Systems*, 30(2), 161–80.

LANGFORD, M. & HIGGS, G. (2006). Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population. *Professional Geographer*, 58, 294–306.

LANGFORD, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems*, 31(1), 19–32.

LANGFORD, M. (2013). An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data. *Geographical Analysis*, 45(3), 324–344.

LIN, J., CROMLEY, R., & ZHANG, C. (2011). Using GeographicallyWeighted Regression to Solve the Areal Interpolation Problem. *Annals of GIS*, 17(1), 1–14.

LIU, X., CLARKE, K., & HEROLD, M. (2006). Population Density and Image Texture: A Comparison Study. *Photogrammetric Engineering & Remote Sensing*, 72(2), 187–196.

LIU, X. H., KYRIAKIDIS, P. C., & GOODCHILD, M. F. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, 22(4), 431–447.

LO, C.P. (2008). Population Estimation Using Geographically Weighted Regression. *GIScience and Remote Sensing*, 45, 131–48.

LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J., & RHIND, D. W. (2005). Geographic Information Systems and Science, Second Edition. New York: Wiley.

LU, D., WENG, Q., & LI, G. (2006). Residential population estimation using a remote sensing derived impervious surface approach. *International Journal of Remote Sensing*, 27(16), 3553–3570.

MAANTAY, J. A., MAROKO, A. R., & HERRMANN, C. (2007). Mapping Population Distribution in the Urban Environment: The Cadastral-Based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science*, 34, 77–102.

MADAJOVÁ, M. (2010a). Využitie pomocných informácií pri transformácii geografických dát - vyhodnotenie vybraných postupov. *Geografický časopis: časopis Geografického ústavu Slovenskej akadémie vied*, 62(3), 259–275.

MADAJOVÁ, M. (2010b). Metódy transformácie priestorových dát: prehľad, klasifikácia a hodnotenie. *Acta Geographica Universitatis Comenianae*, 2010, 54(1), 119–135.

MARTIN, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, NS 14, 90–97.

MARTIN, D. (1996). An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems*, 10(8), 973–989.

MARTIN, D., TATE, N. J. & LANGFORD, M. (2000). Refining population surface models: experiments with Northern Ireland census data. *Transactions in GIS*, 4(4), 343–360.

MCPHERSON, T., IVEY, A., BROWN, M., & STREIT, G. (2004). Determination of the spatial and temporal distribution of population for air toxics exposure assessments. In *Proceedings of the 5th Conference on Urban Environment*.

MENNIS, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55(1), 31–42.

MENNIS, J. & HULTGREN, T. (2006). Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science*, 33(3), 179–194.

NEIS, P., ZIELSTRA, D. & ZIPF, A. (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(1), 1–21.

NORDBECK, S. & RYSTEDT, B. (1970). Isarithmic Maps and the Continuity of Reference Interval Functions. *Geografiska Annaler, Ser. B.*, 2, 92–123.

NORDHAUS, W. D. (2006). Geography and macroeconomics: New data and new findings. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10), 3510–3517.

OECD (2012). Redefining "Urban": A New Way to Measure Metropolitan Areas. OECD Publishing, Paris.

OFFICIAL JOURNAL OF THE EUROPEAN UNION (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online (accessed 26 May 2015): http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007L0002&from=en

OPENSHAW, S. (1983). The modifiable areal unit problem. Concepts and Techniques in Modem Geography 38. Norwich: GeoBooks. ISBN 0-86094-134-5.

OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16(1), 17–31.

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

OPENSHAW, S. & TAYLOR, P. J. (1981). The modifiable areal unit problem. In N. Wrigley and R. J. Bennett (Eds.), *Quantitative Geography*. Henley-on-Thames, Oxon: Routledge and Kegan Paul, 60–70.

PIACENTINI, M. & ROSINA, K. (2012). Measuring the Environmental Performance of Metropolitan Areas with Geographic Information Sources. OECD Regional Development Working Papers, No. 2012/05, OECD Publishing, Paris.

QIU, F. & CROMLEY, R. (2013). Areal Interpolation and Dasymetric Modeling. *Geographical Analysis*, 45(3), 213–215.

RATTI, C., PULSELLI, R. M., WILLIAMS, S., & FRENCHMAN, D. (2006). Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5), 727–748.

REIBEL, M. & BUFALINO, M. E. (2005). Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environment and Planning A*, 37(1), 127–39.

REIBEL, M. & AGRAWAL, A. (2007). Areal Interpolation of Population Counts Using Pre-classified Land Cover Data. *Population Research and Policy Review*, 26(5-6), 619–633.

ROSINA, K., HURBÁNEK, P., & ATKINSON, P. M. (2012). Priestorová dezagregácia populačných dát s využitím máp krajinnej pokrývky a nepriepustnosti povrchu. In *Symposium GIS Ostrava 2012. Současné výzvy geoinformatiky: proceedings*. Ostrava, VŠB – Technická univerzita Ostrava, 2012, s. 1-13. ISBN 978-80-248-2558-8.

ROSINA, K., HURBÁNEK, P., & CEBECAUER, M. (2016). Using OpenStreetMap to Improve Population Grids in Europe. *Cartography and Geographic Information Science*. In print, available online: doi:10.1080/15230406.2016.1192487.

SADAHIRO, Y. (1999). Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems*, 1(4), 323–346.

SADAHIRO, Y. (2000). Accuracy of Count Data Transferred through the Areal Weighting Interpolation Method. *International Journal of Geographical Information Science*, 14, 25–50.

SCHROEDER, J. P. & VAN RIPER, D. C. (2013). Because Muncie's Densities Are Not Manhattan's: Using Geographical Weighting in the EM Algorithm for Areal Interpolation. *Geographical Analysis*, 45(3), 216–237.

SILVERMAN, B. W. (1986). Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall, 1986.

STATISTICS AUSTRIA (2012). Regional statistical grid units. Available online (accessed 30 August 2012): http://www.statistik.at/web_en/classifications/regional_breakdown/grid/index.html

STEEL, D. G. & HOLT, D. (1996). Rules for random aggregation. *Environment and Planning A*, 28(6). 957–978.

STEINNOCHER, K., KÖSTL., M., & WEICHSELBAUM, J. (2011). Grid-based population and land take trend indicators – New approaches introduced by the geoland2 Core Information Service for Spatial Planning. NTTS conference, Brussels (CD-ROM).

SUTTON, P. C., ELVIDGE, C., & OBREMSKI, T. (2003). Building and Evaluating Models to Estimate Ambient Population Density. *Photogrammetric Engineering & Remote Sensing*, 69(5), 545–552.

TAPP, A. F. (2010). Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. *Cartography and Geographic Information Science*, 37,(3), 215–228.

TAYLOR, P. J. (1977). Quantitative Methods in Geography: An Introduction to Spatial Analysis. Boston: Houghton Mifflin Company.

THURSTAIN-GOODWIN, M. & UNWIN, D. (2000). Defining and Delineating the Central Areas of Towns for Statistical Monitoring Using Continuous Surface Representation. *Transactions in GIS*, 4(4), 305–317.

TOBLER, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. Journal of the American Statistical Association, 74, 519–530.

VINKX, K. & VISÉE, T. (2008). Usefulness of population files for estimation of noise hindrance effects. ICAO Committee on Aviation Environmental Protection. In: *CAEP/8 modelling and database task force (MODTF)*, 4th meeting, 20–22 February 2008 Sunnyvale, USA.

VOSS, P. R., LONG, D. D., & HAMMER, R. B. (1999). When census geography doesn't work: using ancillary information to improve the spatial interpolation of demographic data. CDE Working Paper No 9926. Department of Rural Sociology, University of Wisconsin.

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

WU, S., QIU, X., & WANG, L. (2005). Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience & Remote Sensing*, 42(1), 80–96.

WU, C. & MURRAY, A. T. (2005). A cokriging method for estimating population density in urban areas. Computers, *Environment and Urban Systems*, 29(5), 558–579.

WU, C. & MURRAY, A. T. (2007). Population estimation using Landsat enhanced Thematic Mapper imagery. *Geographical Analysis*, 39(1), 26–43.

XIE, Y. (1995). The Overlaid Network Algorithms for Areal Interpolation Problem. *Computers, Environment, and Urban Systems*, 19(4), 287–306.

YOO, E. H., KYRIAKIDIS, P. C., & TOBLER, W. (2010). Reconstructing Population Density Surfaces from Areal Data: A Comparison of Tobler's Pycnophylactic Interpolation Method and Area-to-Point Kriging. *Geographical Analysis*, 42(1), 78–98.

YUAN, Y., SMITH, R. M., & LIMP, W. F. (1997). Remodelling census population with spatial information from Landsat TM imagery. *Computers Environment and Urban Systems*, 21(3/4), 245–258.

ZANDBERGEN, P. & IGNIZIO D. (2010). Comparison of Dasymetric Mapping Techniques for Small-area Population Estimates. *Cartography and Geographic Information Science* 37, 199–214.

ZANDBERGEN, P. (2011). Dasymetric Mapping Using High Resolution Address Point Datasets. *Transactions in GIS*, 15(1), 5–27.

ZHANG, C. & QIU, F. (2011). A Point-Based Intelligent Approach to Areal Interpolation. *The Professional Geographer*, 63(2), 262–276.

ANNEXES

List of annexes:

Annex I: High Resolution Imperviousness Layer 2006 (ca 1% of unclassified cells were replaced by values from 2009 version)

Annex II: CORINE Land Cover 2006 (reclassified into 9 categories)

Annex III: OpenStreetMap roads and railways

Annex IV: GEOSTAT 1A – Aggregated population counts 2006

Annex V: The most accurate result of disaggregation from LAU2 level into 100 m grid cells, based on HRIL, CLC, and OSM ancillary data

Annex VI: Detail of 100 m HRIL data (a), nonzero HRIL cells (b), reclassified CLC data (c), and the combined information (d) on HRIL value and CLC class, which is then used to estimate population

Annex VII: Demonstration of positional offset between HRIL data (cells in greyscale) and OSM road segments (red lines)

Annex VIII: HRIL modification by subtraction of rasterized OSM roads and railways (the Austrian city of Linz is portrayed)

Annex IX: Absolute and relative accuracy of disaggregation using various subset configurations for the entire study area

Annex X: Estimated population density – MT1[2] model aggregated into 8 km grid

Annex XI: Reference population density – GEOSTAT 1A aggregated into 8 km grid

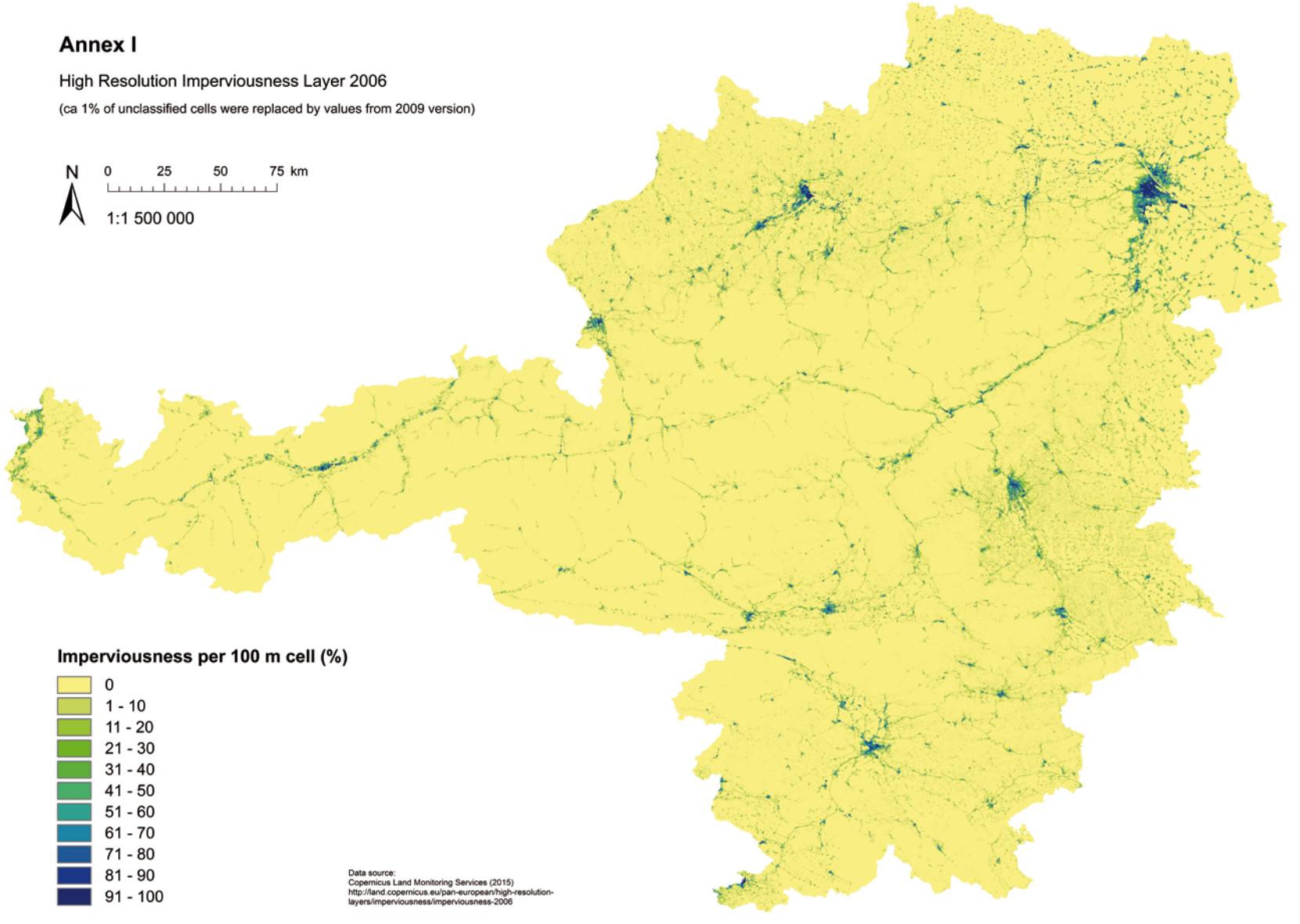
Annex XII: Absolute error of estimation – deviation of population count per 8 km grid cell

Annex XIII: Relative error of estimation – deviation of estimated population count as a share of reference population per 8 km grid cell

Annex I

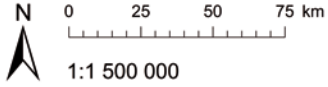
High Resolution Imperviousness Layer 2006

(ca 1% of unclassified cells were replaced by values from 2009 version)



Annex II

CORINE Land Cover 2006 (reclassified into 9 categories)

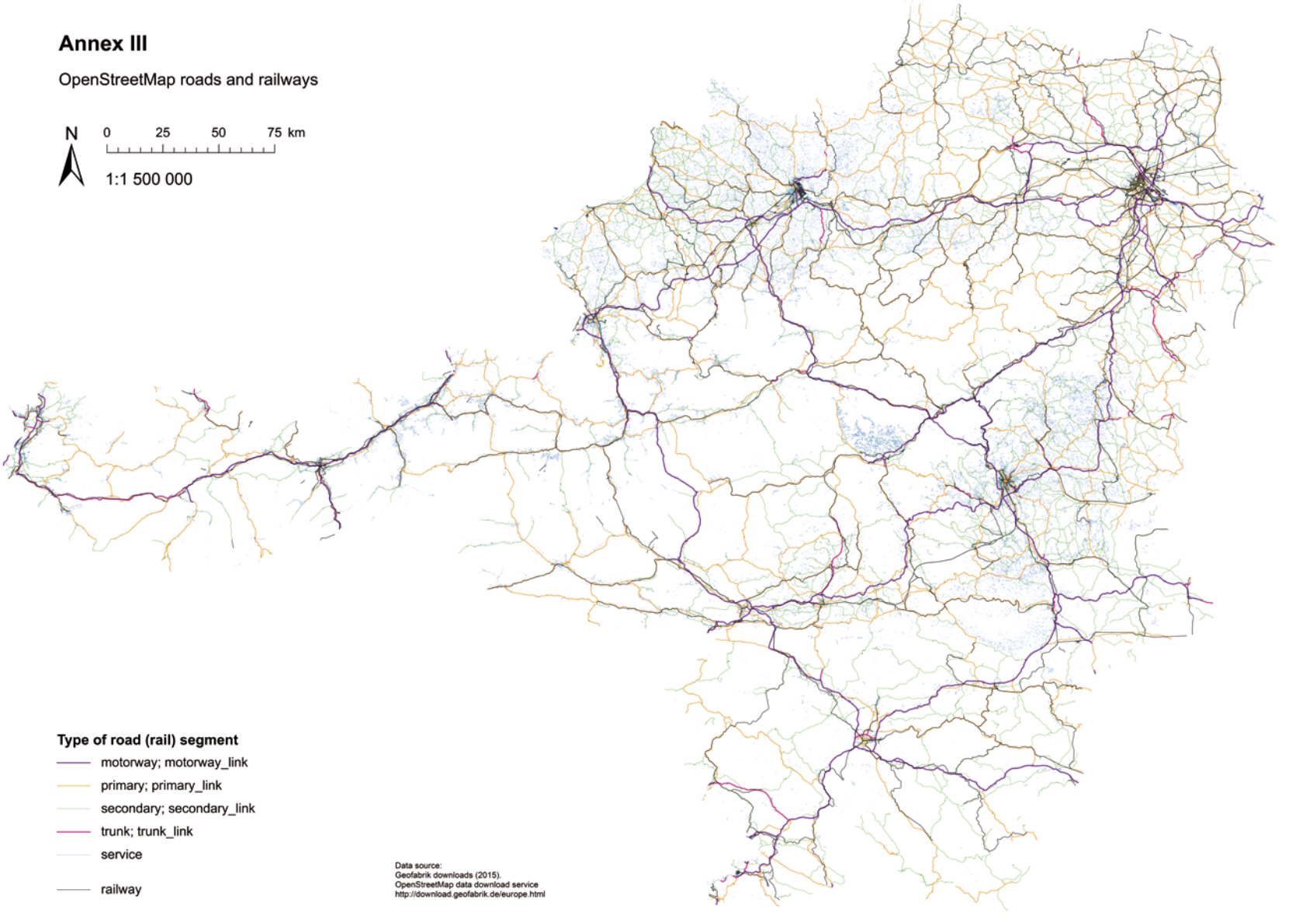


Data source:
Europe Environment Agency (2015).
Corine Land Cover 2006 raster data.
http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3

- 0 - Other
- 1 - Continuous urban fabric
- 2 - Discontinuous urban fabric
- 3 - Industrial or commercial units
- 4 - Road and rail networks and associated land; Port areas; Airports; Green urban areas; Sport and leisure facilities
- 5 - Non-irrigated arable land
- 6 - Vineyards; Fruit trees and berry plantations; Complex cultivation patterns
- 7 - Pastures; Land principally occupied by agriculture, with significant areas of natural vegetation
- 8 - Agro-forestry areas; Broad-leaved forest; Coniferous forest; Mixed forest; Natural grasslands; Moors and heathland; Transitional woodland-shrub

Annex III

OpenStreetMap roads and railways

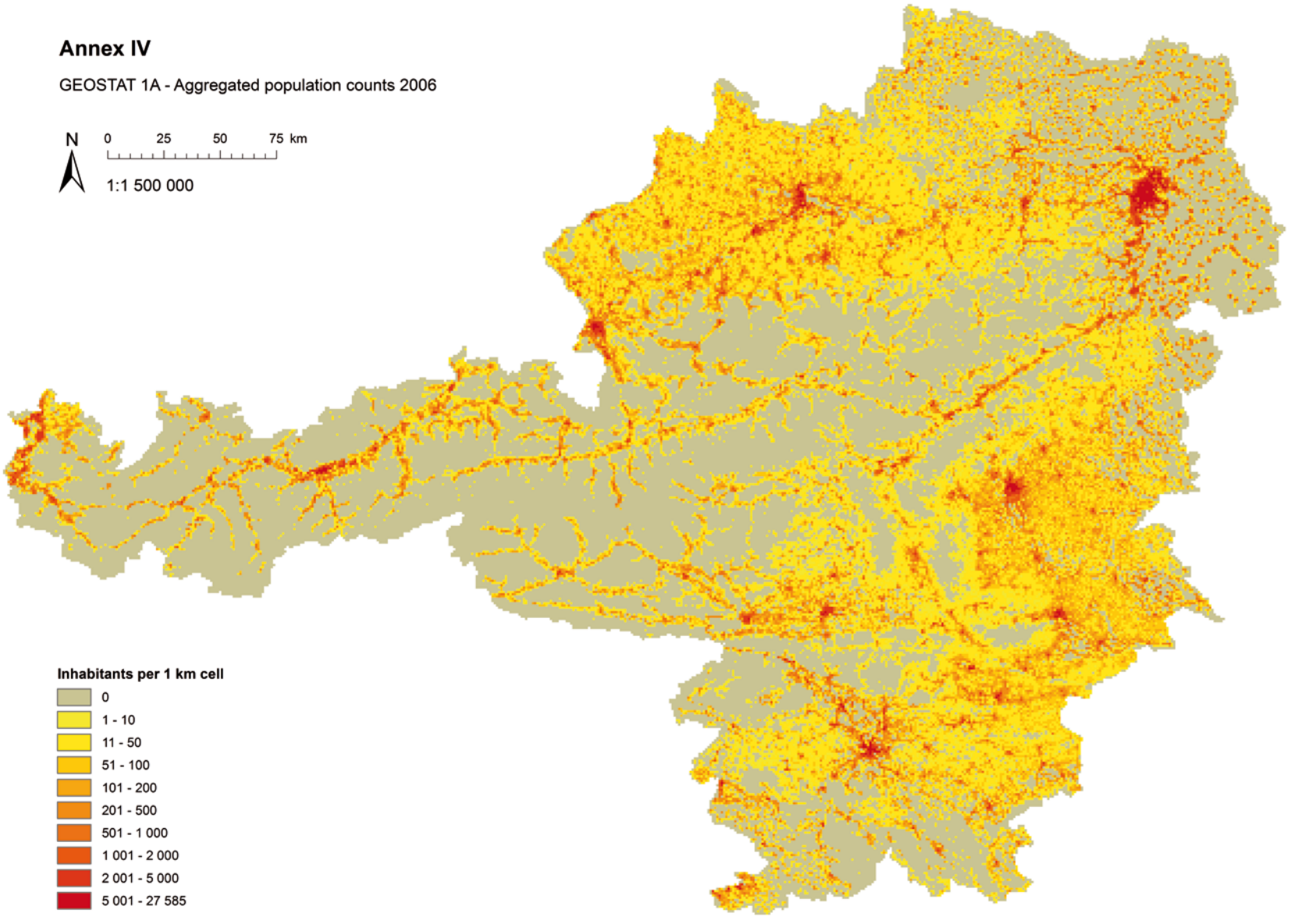


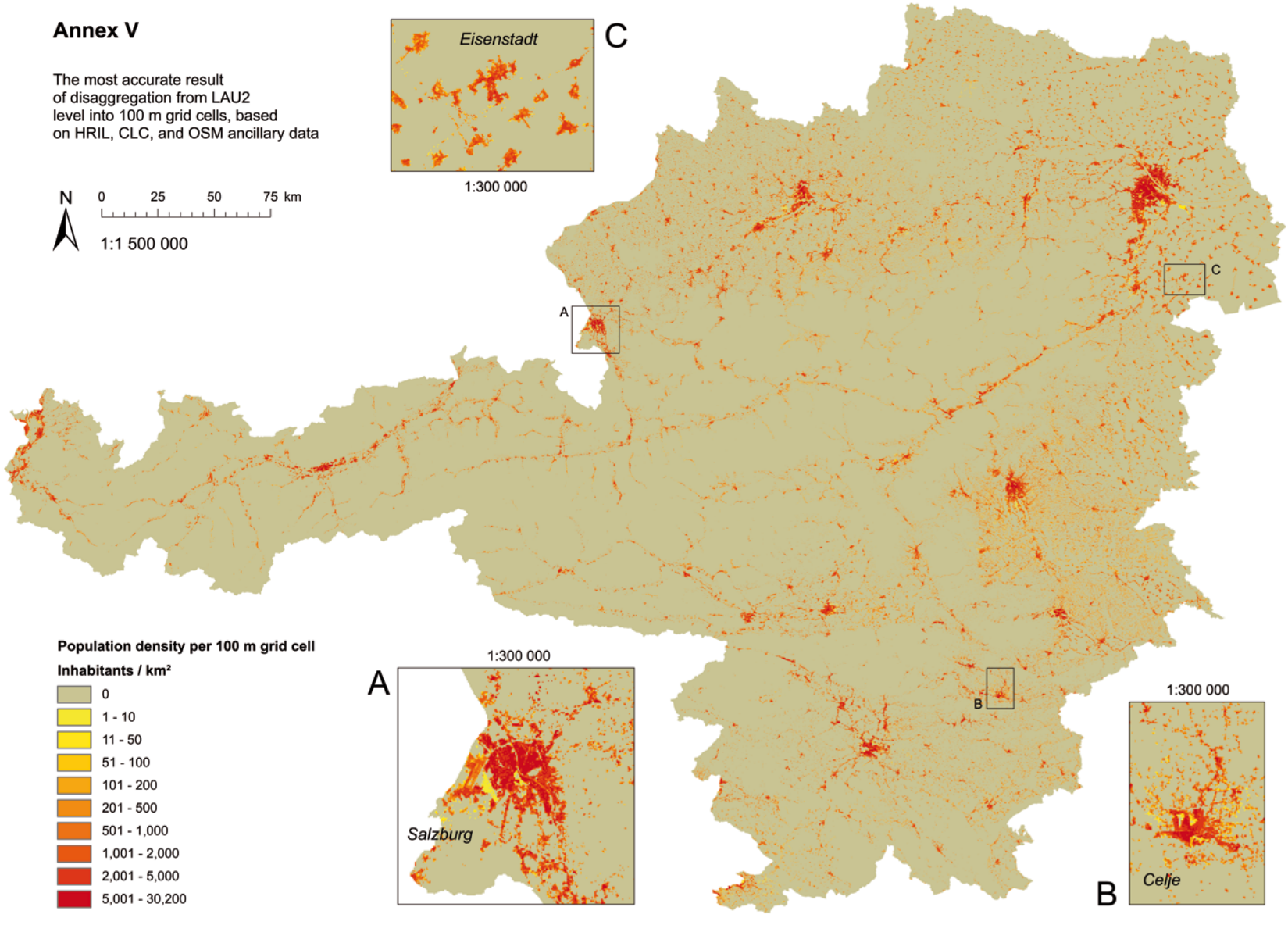
Annex IV

GEOSTAT 1A - Aggregated population counts 2006

N 0 25 50 75 km

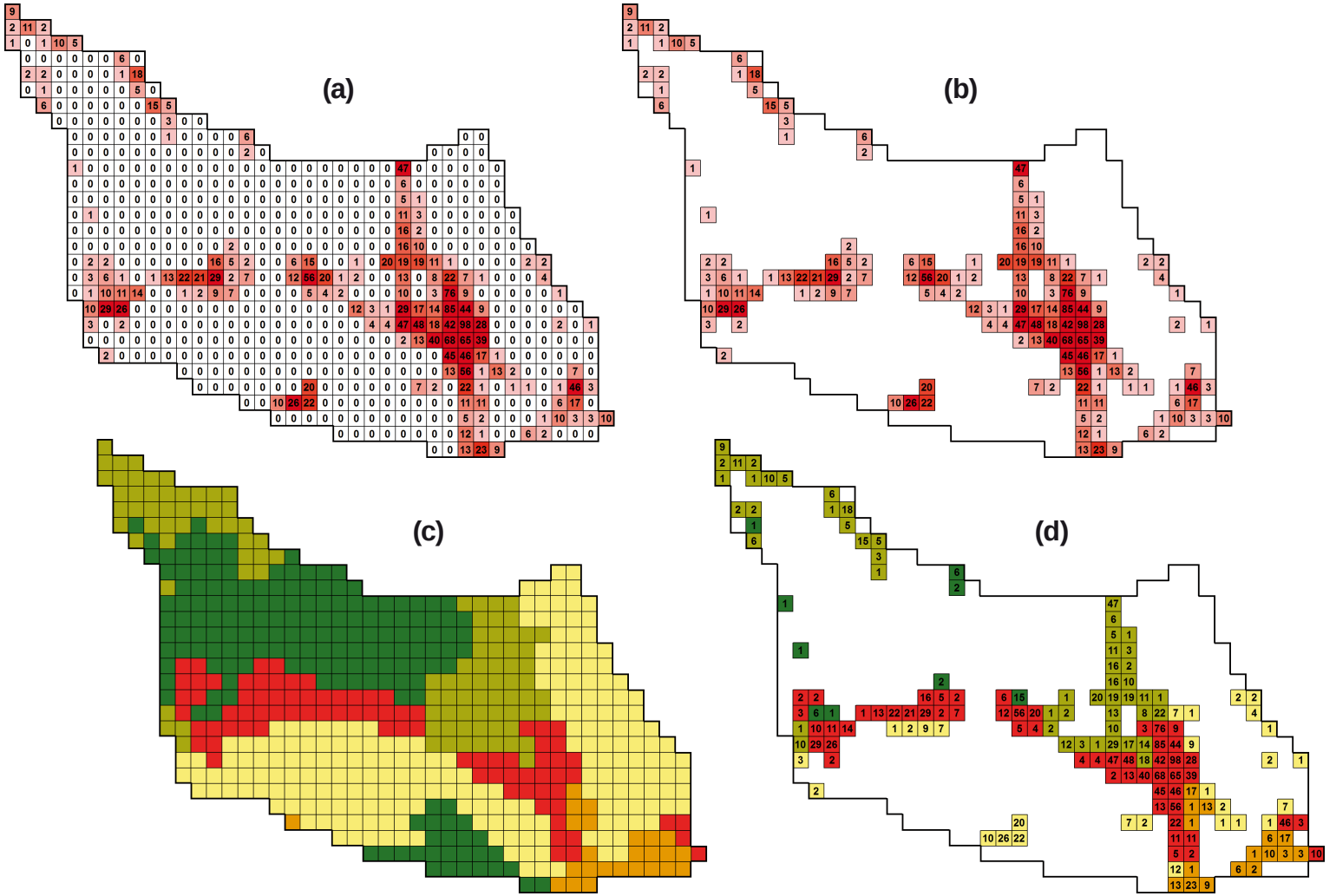
1:1 500 000




Annex V
The most accurate result of disaggregation from LAU2 level into 100 m grid cells, based on HRIL, CLC, and OSM ancillary data
N
0 25 50 75 km
1:1 500 000
Eisenstadt
C
1:300 000
A
C
B
1:300 000
A
Salzburg
1:300 000
B
Celje
Population density per 100 m grid cell
Inhabitants / km²
0
1 - 10
11 - 50
51 - 100
101 - 200
201 - 500
501 - 1,000
1,001 - 2,000
2,001 - 5,000
5,001 - 30,200

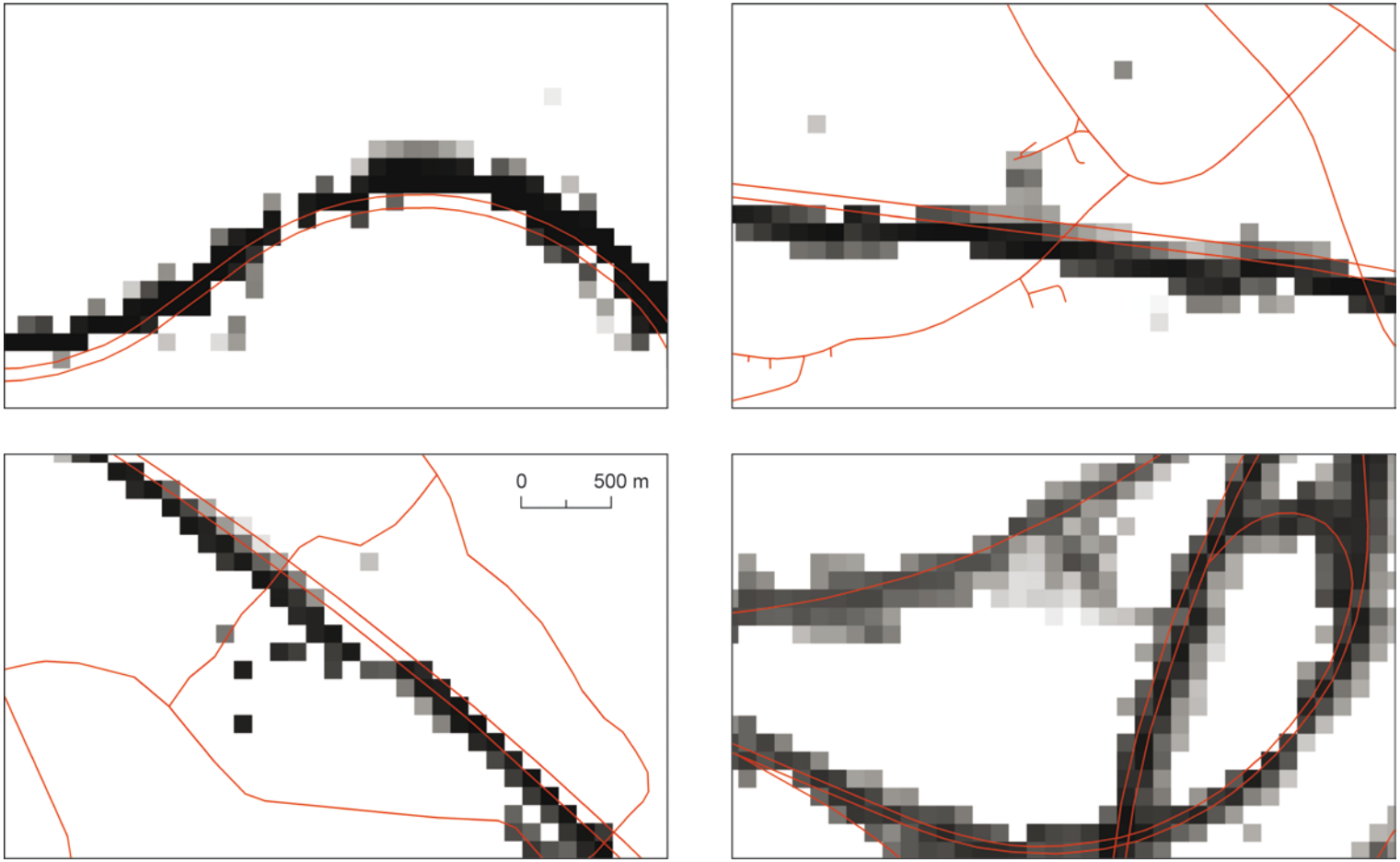
Annex VI

Detail of 100 m HRIL data (a), nonzero HRIL cells (b), reclassified CLC data (c), and the combined information (d) on HRIL value and CLC class, which is then used to estimate population



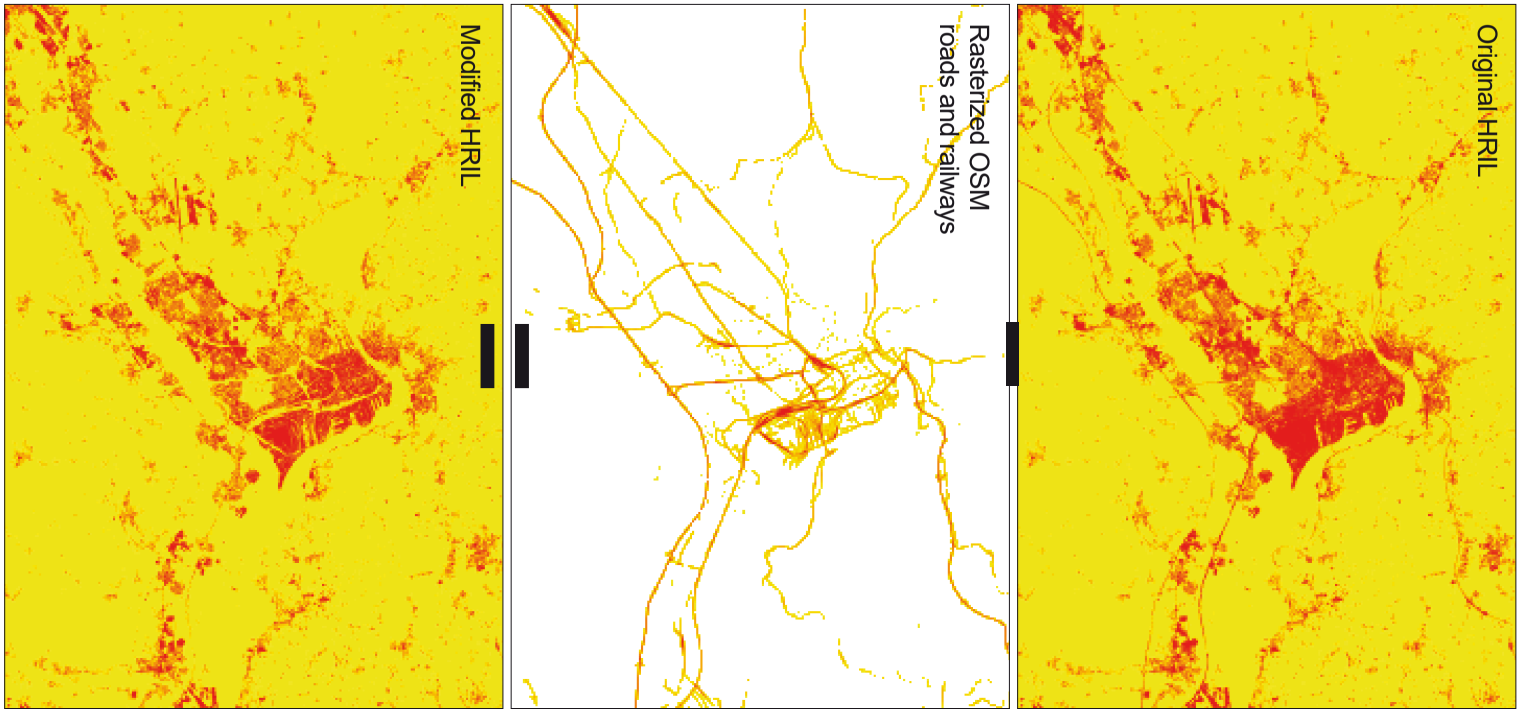
Annex VII

Demonstration of positional offset between HRIL data (cells in greyscale) and OSM road segments (red lines)



Annex VIII

HRIL modification by subtraction of rasterized OSM roads and railways (the Austrian city of Linz is portrayed)

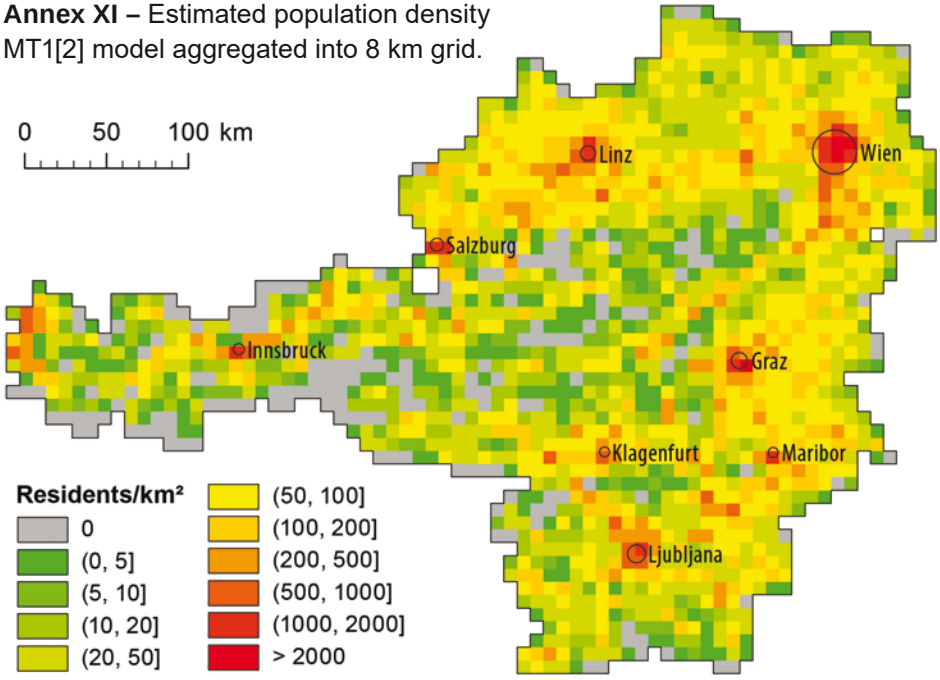


Annex IX

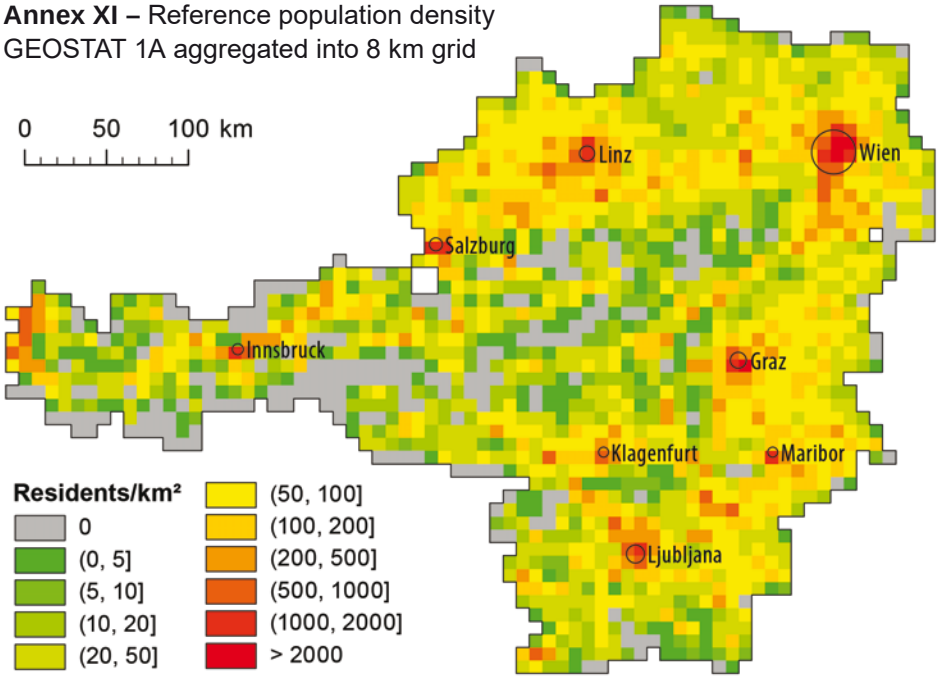
Absolute and relative accuracy of disaggregation using various subset configurations for the entire study area

	A		ADGHX		AD		ADGH		X		DGHX		CDX		DGH		Contig	
	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE	TAE	RTAE
50 LAU2	4,640,850	47.21%	4,804,450	48.87%	4,960,050	50.46%	4,856,270	49.40%	4,517,960	45.96%	4,440,140	45.17%	4,470,790	45.48%	4,480,500	45.58%	4,576,234	46.55%
100 LAU2	4,448,990	45.26%	4,446,500	45.23%	4,626,110	47.06%	4,482,980	45.60%	4,504,160	45.82%	4,381,050	44.57%	4,401,320	44.77%	4,496,230	45.74%	4,584,890	46.64%
150 LAU2	4,447,630	45.24%	4,464,910	45.42%	4,589,980	46.69%	4,567,210	46.46%	4,526,310	46.04%	4,441,100	45.18%	4,452,580	45.29%	4,508,070	45.86%	4,587,702	46.67%
200 LAU2	4,502,910	45.81%	4,480,500	45.58%	4,545,920	46.24%	4,491,050	45.69%	4,536,390	46.15%	4,446,730	45.23%	4,484,270	45.62%	4,484,040	45.61%	4,592,964	46.72%
250 LAU2	4,500,950	45.79%	4,460,500	45.37%	4,594,660	46.74%	4,487,600	45.65%	4,563,030	46.42%	4,430,280	45.07%	4,481,050	45.58%	4,492,880	45.70%	4,564,762	46.43%
300 LAU2	4,499,800	45.77%	4,453,870	45.31%	4,571,970	46.51%	4,477,950	45.55%	4,565,370	46.44%	4,411,690	44.88%	4,475,900	45.53%	4,488,660	45.66%	4,593,306	46.73%
2,500 km²	4,637,190	47.17%	4,818,610	49.02%	4,957,780	50.43%	4,928,000	50.13%	4,549,050	46.28%	4,381,870	44.57%	4,576,950	46.56%	4,361,890	44.37%	n/a	
3,000 km²	4,479,800	45.57%	4,925,420	50.10%	4,943,640	50.29%	4,936,240	50.21%	4,554,000	46.33%	4,385,680	44.61%	4,585,130	46.64%	4,383,040	44.59%	4,592,936	46.72%
3,500 km²	4,475,270	45.52%	4,694,490	47.75%	4,970,370	50.56%	4,866,210	49.50%	4,570,940	46.50%	4,385,360	44.61%	4,587,590	46.67%	4,376,860	44.52%	n/a	
4,000 km²	4,462,130	45.39%	4,587,650	46.67%	4,975,090	50.61%	4,720,320	48.02%	4,569,670	46.48%	4,390,300	44.66%	4,590,820	46.70%	4,394,900	44.71%	4,615,512	46.95%
4,500 km²	4,506,670	45.84%	4,558,990	46.38%	4,983,070	50.69%	4,611,170	46.91%	4,573,390	46.52%	4,393,410	44.69%	4,443,990	45.21%	4,401,790	44.78%	n/a	
5,000 km²	4,471,080	45.48%	4,515,820	45.94%	4,927,590	50.13%	4,560,760	46.39%	4,567,180	46.46%	4,365,710	44.41%	4,456,300	45.33%	4,444,880	45.22%	4,566,724	46.45%
5,500 km²	4,491,710	45.69%	4,508,440	45.86%	4,885,160	49.69%	4,524,300	46.02%	4,563,840	46.43%	4,374,420	44.50%	4,540,350	46.19%	4,504,750	45.82%	n/a	
6,000 km²	4,522,490	46.00%	4,508,030	45.86%	4,881,360	49.66%	4,530,730	46.09%	4,578,120	46.57%	4,441,860	45.18%	4,571,040	46.50%	4,392,710	44.68%	4,584,826	46.64%
6,500 km²	4,536,050	46.14%	4,513,830	45.92%	4,768,350	48.51%	4,535,740	46.14%	4,578,080	46.57%	4,423,800	45.00%	4,565,310	46.44%	4,500,350	45.78%	n/a	
7,000 km²	4,534,410	46.13%	4,524,600	46.03%	4,667,510	47.48%	4,534,950	46.13%	4,572,060	46.51%	4,475,120	45.52%	4,507,790	45.86%	4,494,680	45.72%	4,568,784	46.48%
7,500 km²	4,536,610	46.15%	4,548,980	46.27%	4,655,540	47.36%	4,619,140	46.99%	4,566,510	46.45%	4,477,740	45.55%	4,494,260	45.72%	4,505,720	45.83%	n/a	
8,000 km²	4,528,530	46.07%	4,512,980	45.91%	4,646,480	47.27%	4,594,630	46.74%	4,567,090	46.46%	4,480,950	45.58%	4,426,700	45.03%	4,482,480	45.60%	4,584,422	46.63%

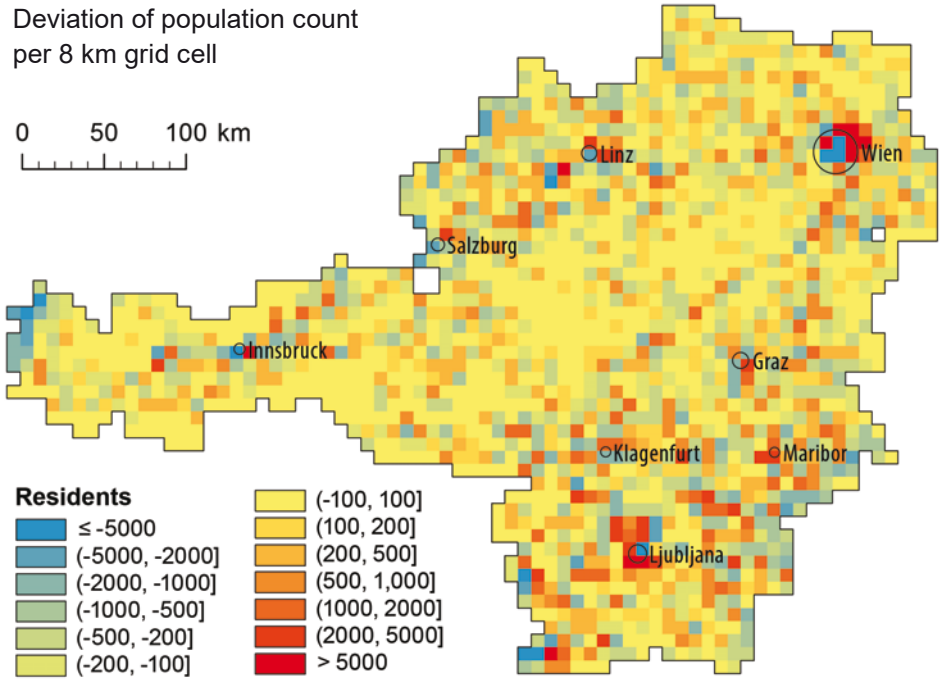
Annex XI – Estimated population density MT1[2] model aggregated into 8 km grid.



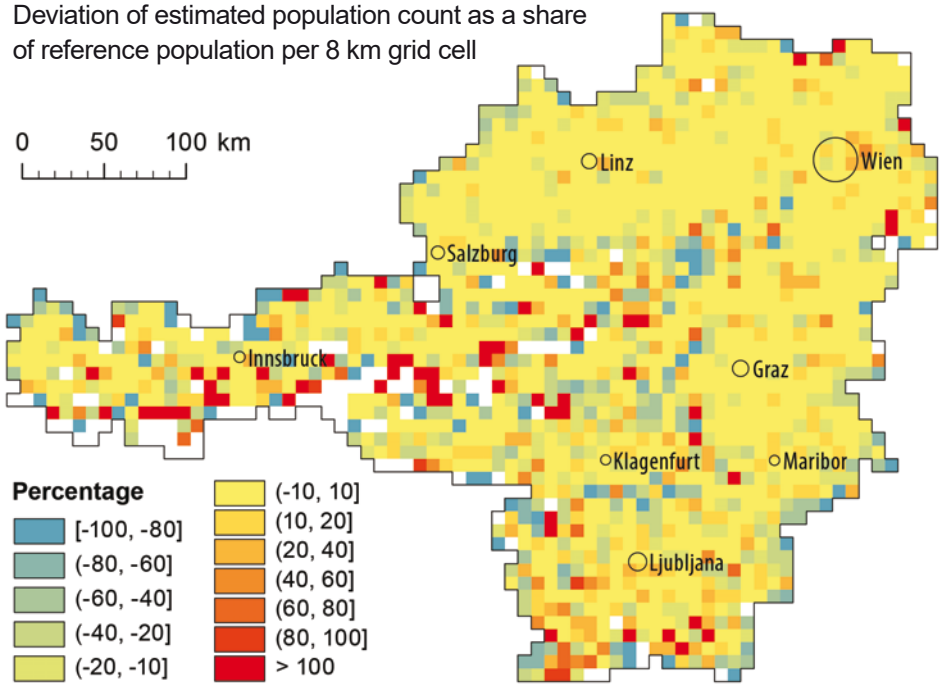
Annex XI – Reference population density GEOSTAT 1A aggregated into 8 km grid



Annex XII – Absolute error of estimation
Deviation of population count
per 8 km grid cell



Annex XIII – Relative error of estimation
Deviation of estimated population count as a share
of reference population per 8 km grid cell



GEOGRAPHIA SLOVACA 31

Posledné vydané čísla / Last issues

23. Solín, Ľ. (2006). ODHAD N-ROČNÝCH MAXIMÁLNYCH PRIETOKOV REGIONÁLNOU FKREKVENČNOU ANALÝZOU

Estimate of maximum discharges with T-return period by regional frequency analysis

24. Cebecauerová, M. (2007). ANALÝZA A HODNOTENIE ZMIEN ŠTRUKTÚRY KRAJINY (na príklade časti Borskej nížiny a Malých Karpát)

Analysis and assessment of changes of landscape structure (case study of a selected part of lowland Borská nížina and mountains Malé Karpaty)

25. Ira, V. ed. (2008). ĽUDIA, GEOGRAFICKÉ PROSTREDIE A KVALITA ŽIVOTA

People, geographical environment and quality of life

26. Ira, V., Lacika, J. eds. (2009). SLOVAK GEOGRAPHY AT THE BEGINNING OF THE 21st CENTURY

27. Ira, V., Podolák, P. eds. (2010). SÍDELNÁ ŠTRUKTÚRA SLOVENSKA(diferenciácie v čase a pristore)

Settlement structure of Slovakia (differentiations in time and space)

28. Michálek, A., Podolák, P. eds. (2014). REGIONÁLNE A PRIESTOROVÉ DISPARITY NA SLOVENSKU, ICH VÝVOJ V OSTATNOM DESAŤROČÍ, SÚČASNÝ STAV A KONZEKVENCIE

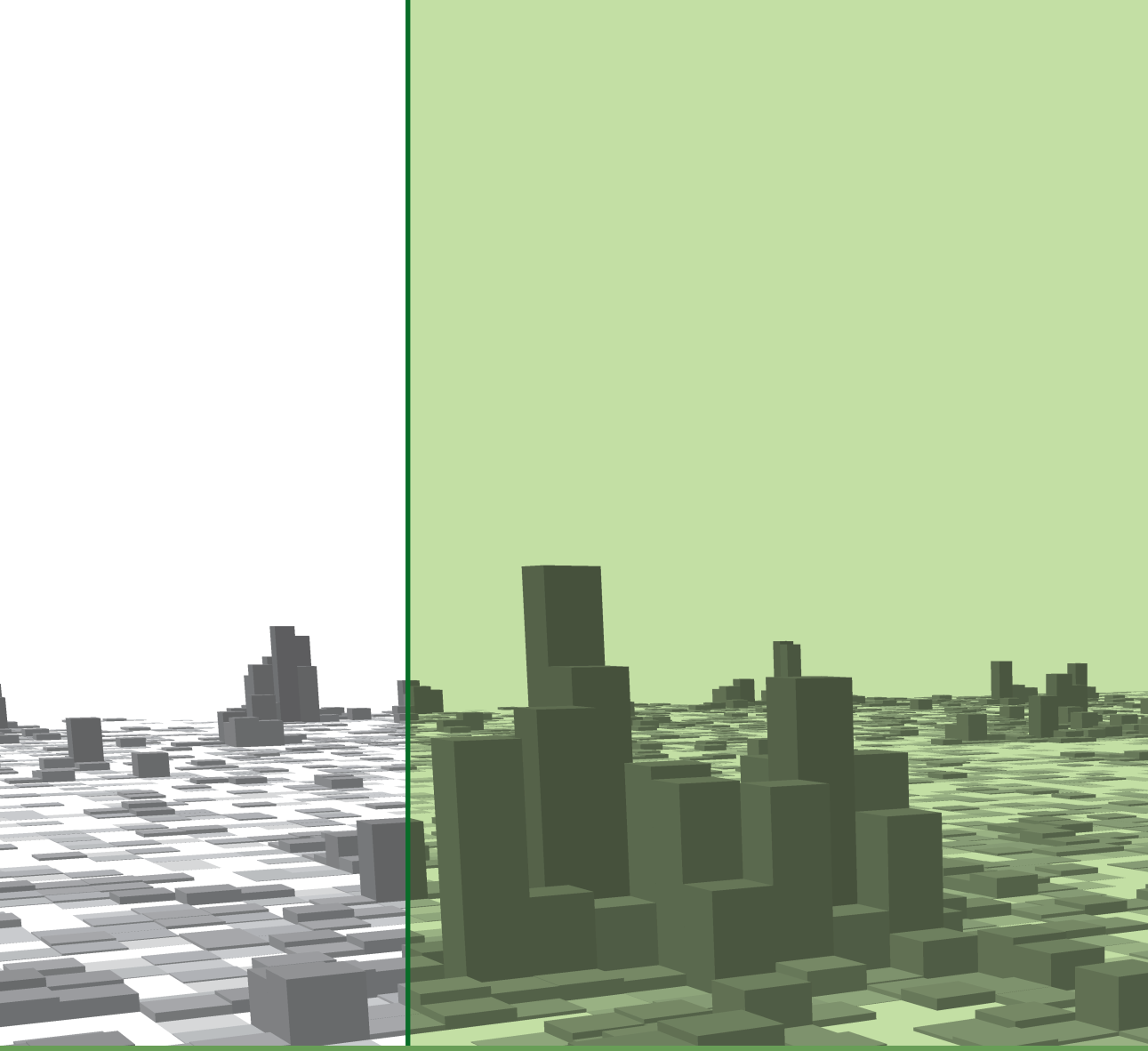
Regional and spatial disparities in Slovakia: develoment in the last decade, the present status and consequences

29. Šuška, P. (2014). AKTÍVNE OBČIANSTVO A POLITIKA PREMIEN MESTSKÉHO PROSTREDIA V POSTSOCIALISTICKEJ BRATISLAVE

Active Citizenship and the Politics of the Urban Environment Transformation in Post-socialist Bratislava

30. Kopecká, M., Rosina, K., Oťaheľ, J., Feranec, J., Pazúr, R., Nováček, J. (2015). MONITORING DYNAMIKY ZASTAVANÝCH AREÁLOV

Monitoring the Dynamis of Built-up Areas



GEOGRAFICKÝ ÚSTAV SAV
INSTITUTE OF GEOGRAPHY SAS

Bratislava 2016

ISBN 978-80-89548-02-6
ISSN 1210-3519