GSFI proposal for a national small area division model

Title: Proposal for a national small area division model
Project: The GSGF in Finland — Integration of geospatial and statistical information in Finland (GSFI)
Grant agreement number: 101112903 - 2022-FI-GEOS-GSFI

Content

1. Introduction

The purpose of this work is to draft a data needs description and a proposal for a new nationwide small-area division. A small area refers to a statistical unit that divides municipalities into smaller statistical areas. Currently, Finland does not have a unified small area system. The goal of the new division model is to provide an areal division system that is compatible with administrative boundaries and uniform across the country. The purpose of the new small area division model is not to replace municipalities' own operational divisions. The model serves as a basic dataset for more detailed area divisions, enabling the production of nationally consistent area divisions.

The most common area divisions for statistical data are administrative divisions based on municipalities. However, these are often too coarse for many user needs. Currently, statistics smaller than municipalities are produced, for example, by municipal sub-areas, postal code areas, or statistical grids. The situation regarding statistical sub-areas within municipalities is diverse. Some municipalities utilize their own sub-area divisions, fulfilling data needs within a limited area, namely the municipality itself. However, municipal sub-areas do not form a nationally consistent and comparable dataset. The national postal code area division is not compatible with municipal boundaries or other statistical divisions. Statistical grids are flexible, meaning they can be used to analyze areas tailored to each information need. However, they are not directly compatible with administrative boundaries, and they have limitations regarding data privacy when the number of statistical units within a grid cell is small.

Small areas and statistical grids are essential for area specific statistics and spatial analysis. For example, Galster (2001) has pondered the definition of neighborhoods and the challenges associated with. The variation of different variables occurs at different spatial scales, so the type of area unit needed depends on the phenomenon being examined. For instance, there can be significant differences in building stock within a few tens of meters, while education quality varies according to school districts, and air quality may be similar across large areas. Therefore, depending on the phenomenon being examined, it is important to consider what kind of areas the data should cover to make meaningful observations about the phenomenon itself (Galster 2001).

The challenge in defining small areas lies not only in the different scales but also in the constant change. From a spatial dynamics perspective, population structure and services generally change faster over time than the geographical determinants or built environment of an area. If one seeks a more permanent basis for defining an area, buildings, infrastructure, topography, or geographical barriers provide a better foundation than, for example, population structure or service network. If the formation of small areas aims to consider both scales and temporal stability, they should be based on an approach where the area to be divided is first divided as accurately as possible based on the most stable elements (Galster 2001).

This document proposes a model for small areas based on the outlined shortcomings in current statistical areal divisions, a survey of data needs, and the development of various datasets and methods to form a national dataset of small areas in Finland. The proposal promotes the goal of integrating statistics and geospatial information within the framework of the Global Statistical Geospatial Framework (GSGF). Specifically, this development work focuses on the third principle of the framework: "Common area divisions for data and statistical distribution".

During the project development process: 1. A stakeholder survey was conducted on the need and content of small areas. 2. Decision to propose creating an areal division system where small areas are based on smaller building blocks as base-areas. 3. Different approaches and datasets were tested for forming these base-areas. 4. A pilot version of the actual small areas was created based on these base-areas. The work resulted in a method where, from geospatial datasets, base-areas are formed, and from those, small areas.

The proposed small area division model is based on an approach where a dataset of very finely detailed base-areas is first created. This dataset serves as the foundation for defining the actual statistical small areas. The base-areas consider urban structure and geographical factors. With base-areas, it's possible to create small areas based on various criteria (Hugo, 2007). The definition of statistical small areas relies on population size criteria, ensuring data privacy when publishing statistical information.

The pilot version is a draft and proposal, not yet the final dataset. When forming the final dataset, it is important to note that there are some gaps in the geospatial datasets used for base-areas, requiring manual checks. Additionally, the data generated using the automated method includes some objects that are not optimally delimited. The proposal is based on a vision of a target situation where, using this dataset, different types of small areas for various needs can be easily produced with different boundary values and criteria.

2. Background

2.1. GSGF principle Common Geographies for the Dissemination of Statistics

The development of small areas is linked to Principle 3 of the GSGF framework, 'Common Geographies for the Dissemination of Statistics.' Principle 3 ensures that common geographies are available for data dissemination. These support the management and use of data from different sources as geospatial information, as well as the integration, visualisation, analysis, and interpretation of data.

In the GSFI roadmap, following goals for the principle 3 are set: 1. National territorial divisions and their production and management have been organized. 2. National territorial divisions are produced only once and made available for open use. 3. Small area divisions are based on new small area division model or statistical grids.

The first goal aims at a common hierarchical territorial division model. National territorial classifications and corresponding territorial division maps are created relying on the model. The production and maintenance of map data is agreed nationally.

The second goal is that the defined and agreed national territorial divisions (different scales, time periods, accuracy) are produced only once and freely and openly available. The users of the data can rely on the continuity of production and availability of these data. In addition to current boundaries, the material also includes historical divisions, where possible.

The third goal is that the national small-area division model serves as a basis for different types of small area divisions that divide municipalities into smaller spatial units. In addition to the national and common divisions, various stakeholders can use the model for their thematic special needs. Also, area divisions based on statistical grids can be used for various data needs.

2.2. Small area divisions as a part of national territorial divisions

In national statistics, the core of the area division model consists of hierarchical municipal-based areas and their counterparts from EU territorial classification. In addition to these, there are national and EU legislation-derived area divisions that are municipal-based but do not form a hierarchy in the same structure. Examples include the DEGURBA (Degree of Urbanization) or national administrative ELY and AVI areas.

Since Finland's municipal structure has evolved over time into a complex mixture of different-sized municipalities, they do not perform very well from the perspective of presenting spatial data or analyzing territorial data. Municipal-based area divisions are usable in the examination of national territorial differences and generally at the level of larger regional areas. They are also needed when information about municipalities or other administrative areas is specifically required.

Currently, municipalities have variously divided small areas or statistical areas according to their own needs and requirements. In this regard, there is no standardized terminology and criteria for forming areas and creating hierarchical structure. Another approach has been the use of grid cells as the basis for area divisions. Statistical grids as such form a dataset that can include diverse types of information. For example, a 1x1 km grid cell can be used directly as an area division. Statistical grids have been used to form area divisions based on thematic classifications, such as localities and the national urban-rural classification. The widely used postal code areas is a national small area division, but it is not compatible with municipality borders.

Finland does not currently have a national uniform small area division. In GSFI targets, it has been recognized as its own area division component in national system of territorial divisions. Figure 1 illustrates the vision of the territorial divisions system suggested in GSFI project. It divides into municipal-based administrative areas, base-areas for creating small areas, and statistical grids. In this areal division system, municipalities are combined to form municipal-based area divisions, and combining the base-areas forms area divisions that divide municipalities into smaller areas. In the future, it will be necessary to explore whether small areas can be aligned with statistical postal code areas or grid-based territorial divisions and classifications.

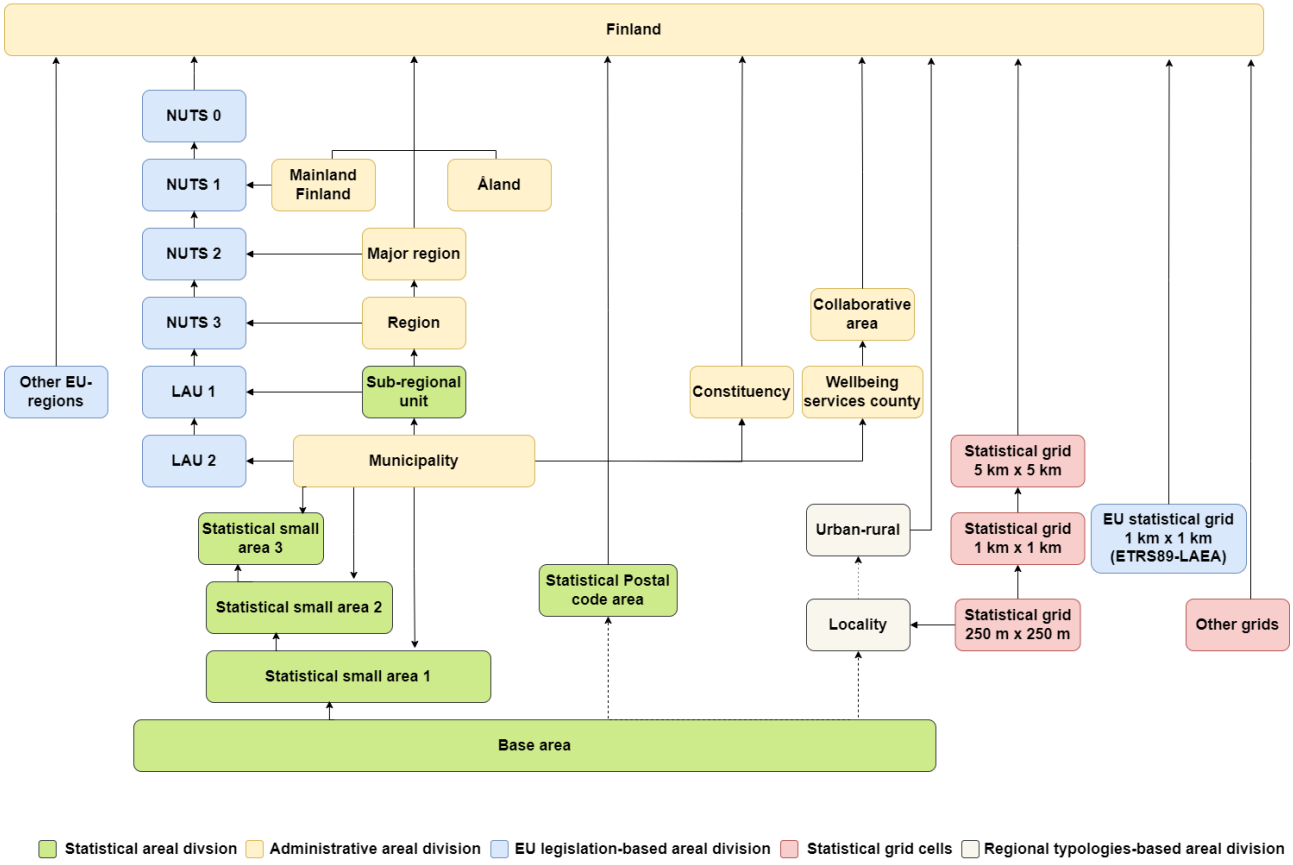

Figure 1. Vision of the future structure of territorial divisions in Finland. Structure is divided into municipal-based administrative divisions, base-areas for small area divisions and statistical grids.

2.3. Small area divisions in other countries

The definition and establishment of small areas are an integral part of statistical work in countries that use a census-style collection of population data (Martin, 2000; Walford & Hayles, 2012). Examples of such countries include England and Wales, Australia, the USA, and New Zealand. In each of these countries, the hierarchy of area divisions is updated for each census. For example, in the United States, small areas have been used to study health conditions in New York City as far back as the early 20th century (Sperling 2012). In contrast, in countries where population statistics are based on registers (i.e., continuous monitoring of demographic changes), there are different types of small area divisions that have been formed from other starting points. These often rely on existing historical area divisions (e.g., Norway) or voting districts (e.g., Sweden) (Statistics Sweden, 2018).

Methods for defining census area divisions have been systematically developed and scientific articles have been published on this topic (e.g. Cockings et al., 2013; Martin, 2000; Walford & Hayles, 2012; Hugo 2007). In general, the issue has been resolved by creating as precise a spatial granularity (parcel) as possible, which forms the basic building blocks of territorial classification. In England and Wales, these basic building blocks are postal code areas (postcode), which typically contain 17 addresses per

area. In the US, Australia, and New Zealand, the most precise spatial granule is defined using street network-based block systems (Blocks or MeshBlocks) (Australian Bureau of Statistics, 2024; Stats NZ, 2022). The differences between these approaches have been evaluated in a paper by Cockings et al., (2013). The territorial classification system is constructed on a hierarchical model, where basic building blocks are combined to form various types of statistical sub-areas. For each level of areal classification, a population base unit with the same size class is formed.

3. Data need review

3.1. Survey to gather information needs

A national small area division survey was conducted to clarify various needs related to the development of the small area division system. The survey was open from April 25th to May 12th, 2023, and received 58 responses. The respondents came from all over Finland and had expertise in spatial data. They worked in different organizations, including local government offices (18 responses), regional or local actors (10 responses), national actors (7 responses), universities (2 responses), companies (3 responses), and one private individual. The survey responses highlighted various information needs and perspectives. The respondents considered the topic to be very important and viewed the work on creating a national small area division as useful.

In the responses from municipalities, small areas were mirrored against the existing municipal small area divisions. In some municipalities, there was a need for the national data to be based on these existing divisions. However, in other municipalities, they were open to new proposals, but certain needs were mentioned in the survey. In particular, the problem of interoperability of adjacent administrative boundaries is an issue even within municipalities. At the regional level (e.g., in the counties or well-being areas), there are various types of area-specific information needs that have larger scope than a specific municipality. In these cases, small area divisions can serve as a crucial tool. In research use, the small area level was considered necessary and useful, but compared to other respondent groups, some unique perspectives emerged, especially related to flexibility and multi-dimensionality, as well as available data contents.

3.2. Using regional divisions

A large part of the respondents have utilized spatial divisions for various GIS analyses. User experiences vary, but in responses it emerged that there is a need to consider different scales and comparisons between areas. The current spatial divisions create a confusing system especially from the perspective of those who need to consider broader areas that transcend multiple administrative regions or the entire country. In municipalities, there are different operational own spatial divisions whose handling and maintenance are an integral part of the various sectors' activities in the municipality. This was highlighted in many responses, e.g., in connection with population forecasts, service network planning, or urban spatial planning. Additionally, several entities were mentioned as maintaining administrative divisions within municipalities, regions, and nationally.

In responses from actors of entities that cover multiple municipalities, the importance of statistical grid cells as an area unit is highlighted, because it enables flexible GIS analyses. Also, in urban planning, statistical area divisions are often too coarse and do not accurately define the areas being examined. Municipal own statistical areas and national postal code areas have been used in monitoring and visualizing spatial phenomena. Companies combine their data to available areas, enriching data with relevant information for the company. In addition to this, other use cases were found in responses,

such as monitoring of urban spatial structure, business analyses, research use, regional level analyses, rural development, and education.

3.3. User needs for small area divisions

Different actors have different needs for small area division. In many municipalities, there are own small area units and developed processes for their maintenance. Therefore, in responses, the distinctive needs of those municipalities emerge where they have their own self-defined small areas, while other organizations rely on using existing or externally-sourced small areas. Some organizations require an area unit from which they can flexibly form corresponding small areas to meet different needs. A typical such area unit is a statistical grid cell. Some municipal respondents noted that the current small area division in their municipality is sufficient for their own operations. However, cities also expressed a need for a common criterion, which would allow comparable data from different municipalities' areas. This would enable better comparison of the development of different urban areas. In responses from rural municipalities, the unique features of areas, such as the identification of village settlement structures and consideration of archipelago-specific characteristics, were seen as essential needs.

The small area division was seen as being relevant to state-level funding allocations, which require area classifications. Specifically, it was hoped that all stakeholders would align with the same area division. With small areas, there is potential to improve spatial precision and data access from smaller areas, leading to better differentiation between areas.

In cities, small areas are used, among other things, for analyzing segregation development, which requires sufficient population-based areas defined in terms of the phenomenon itself. However, currently combining statistical grid cells and a city's own small area divisions accurately is not possible.

Responses from regional actors and counties emphasized the need for monitoring urban structures. To this end, comparable and flexible area units are required, which include data on population, employment, construction, etc. Having data ready in existing small areas would speed up data processing in situations where information is needed from multiple municipalities' areas. From this perspective, respondents also highlighted the importance of functionality as a single criterion for boundaries, meaning that, for example, connectivity to transportation networks should be taken into account. In particular, small area division has advantages in statistical analyses where cartographic presentation is important. Such examples include population growth or construction development statistics.

In some responses, current challenges in data processing were emphasized, particularly with regards to data interoperability, consistency and integration of different area divisions, as well as the openness and permanence of datasets and small area statistics. Respondents who used area divisions for research purposes highlighted the importance that the area division is flexible and allows for multi-dimensional analyses.

3.4. Respondents' preferences and requirements for the new small area division

In many responses, it was hoped that small area division would provide clarity, consistency, and compatibility with administrative boundaries. From a data protection perspective, there was a need for an area unit that could provide as accurate area information as possible. Additionally, it was wished that attached data would be easily accessible and usable.

In the responses, good practices for defining small areas and forming the resulting dataset were described as follows:

1) Availability of boundary data (i.e., open access to boundary information)
2) Naming small areas based on place names, using logical identifier codes
3) Providing certain basic statistical information
4) Having a large number of small areas
5) Selecting boundaries that are geographically meaningful and aesthetically pleasing

Especially, it was hoped that small area division would enable comparability between areas, which would add value compared to the current area units. Thematically categorizing small areas was seen as a desirable way to facilitate comparability and usage of the data.

Alignment with administrative boundaries, specifically municipal borders, was considered an essential feature. In several municipalities, it was hoped that the areal division would be compatible with their own existing sub-area boundaries, but in some cases, respondents were open to new areal divisions or saw flaws or update needs in their current ones. Some municipal respondents also expressed a wish for collaboration with municipalities in creating small areas.

Some respondents emphasized an approach that takes into account historical municipal borders and area identity. For some respondents, identifying the current state of urban and rural development, including the characteristics of built environment and taking into account functional elements, was a crucial starting point for forming areas.

The consideration of geographical features dividing areas was almost unanimously considered an essential principle. These included roads (such as motorways), railroads, green spaces, and rivers.

In rural settings, several respondents noted the possibility of linking to historical village areas, which could be identified using property numbers, school locations, or place names.

In urban environments, it was recognized that area character is defined by its city district number, as well as geographical features such as roads and borders. Additionally, differences in building type and age were also considered.

In the responses, hierarchical structure formation was presented as a solution, which would allow for the establishment of different levels of small areas with varying sizes. From the perspective of considering different user requirements, creating a hierarchical structure was seen as a necessary solution that would allow for more flexible areas.

In a comment from a representative of a large city, it was emphasized that small areas should be internally as homogeneous as possible, which would distinguish different types of urban structures. This would enable comparisons between similar areas within and between cities.

In a dense city, the upper limit of the population criterion was considered unnecessary if size of area is small. In several responses, the mechanical population threshold was seen as a weakness both in urban (upper limit) and rural settings (lower limit). From the perspective of visual clarity and spatial differentiation, it was hoped that a certain level of territorial differentiation and spatial detail would be obtained even on less densely populated areas.

Some respondents recognized the challenge of managing changes to area units, particularly in the built-up edges of a growing city. To address this, they proposed developing a set of criteria for managing these changes.

There were also several mentions in the responses that emphasized the desire for clear and logical naming conventions, relatively similar-sized areas (in terms of land area), and as small an area size as possible.

From the perspective of a statistics producer, a stable foundation is needed to implement various small area geographies. To this end, it would be desirable to achieve as permanent and consistent basic units as possible, which could be used to manage changes and accommodate various needs. The goal would therefore be to establish a national basis for small area divisions. Small area definitions composed of basic units can be tailored to meet specific needs, and they can be implemented by different organizations according to their requirements. Pursuing a single national small area division for all user needs is considered challenging, as it would require defining it based on one specific purpose. For example, if the purpose of use is to share open data, the population-based criteria could become the primary determining factor. A small area division created from a single perspective does not work in all user needs, so the small area division system should leave opportunities for considering different purposes of use.

3.5. Evaluating goals for small area division definitions based on survey results

The survey included a question that presented different concrete goals for small areas. This resulted in clear opinions about criteria for small area zoning (Figure 2). The most important goals were considered to be: "small areas connects to municipal boundaries", "urban form is considered", and "its basic units can be easily combined for different uses". However, there are more variation in opinions regarding whether the zones should have a population size of approximately 1000 residents and whether they should follow property boundaries. The majority also favored a small areas that remain as stable and consistent as possible.

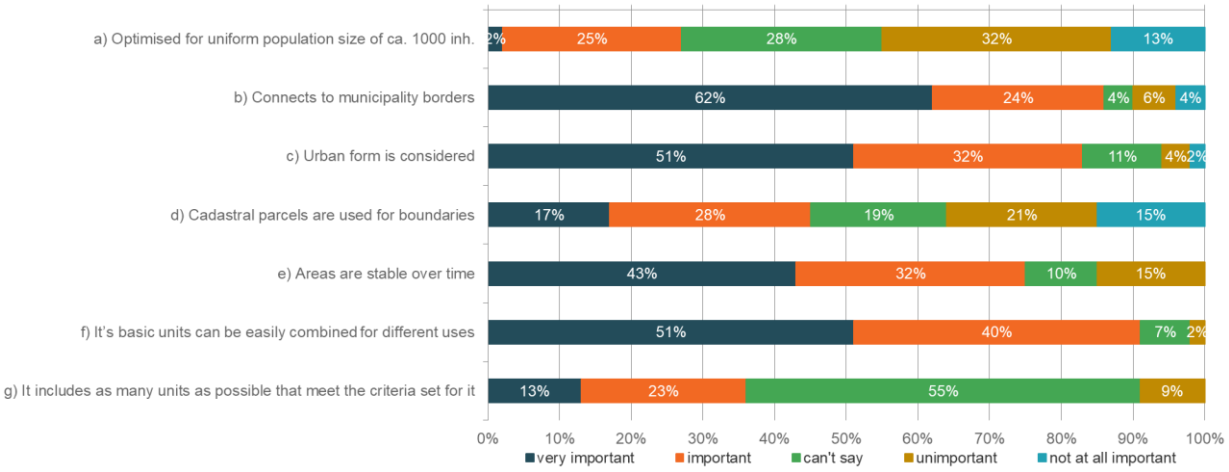

Figure 2. The results to a survey question: "We have provisionally identified the following objectives for small area division, how important they are from your perspective?"

The next section presents the respondents' views related to these provisional goals.

Population-based small area size (approximately 1000 residents)

The predetermined population-based criterion was considered a bad goal in general. However, respondents clearly pointed out that such a demographic-based approach to small area division is justified for certain purposes, but at the cost of losing other characteristics. In other words, they preferred that small areas would be distinct and unique, rather than optimized based on population structure.

The 1000-resident threshold was considered too high, especially in sparsely populated areas. In cities it might also be too low threshold. In rural municipalities, the added value of small area division would be lost if the population criterion is too high. For example, in an archipelago, such a boundary could become impractical.

In terms of uniform small areas, there was a desire for a hierarchical classification system, where small areas are combined into larger entities in cities. The use of several hierarchical levels in small area division was considered sensible.

Follows municipal boundaries

The majority of respondents considered adherence to municipal boundaries as justified and an important feature. In practice, municipalities would be divided into the smallest possible units based on chosen criteria. This also facilitates data interoperability and use, according to respondents. At the same time, areas with these boundaries are compatible with other administrative area boundaries, such as counties and well-being districts.

Consideration of urban form and spatial structure

Respondents considered urban form and spatial structure to be important aspect when defining areas. In smaller municipalities, the separation of residential areas and scattered settlements was highlighted as important, while in cities, consideration of land use and infrastructure was emphasized. In rural areas, identifying village areas was mentioned. Regarding urban form and spatial structure, it was noted that it is a prerequisite for meaningful areas, but on the other hand, changes in built environment make the goal of maintaining a consistent areal division difficult. The idea was also floated that using existing boundaries and structures would be a more straightforward way to "sell" the small area division system to potential users.

Cadastral parcels are used for boundaries

The general approach to follow property boundaries was considered as a good principle, but it was also seen problematic and difficult to implement. The small details and variability of property boundaries create problems that make maintenance and visualizing data challenging. However, in areas where following property boundaries is suitable, defining boundaries based on this approach was considered reasonable. A strictly property-boundary-based model would easily conflict with other goals. The use of classification categories from local detailed plans were seen as a viable alternative for cadastral parcels.

The property register can also serve as a starting point for boundary definition if it's possible to take advantage of city district numbers in property identifiers. In rural areas, the identification of cadastral villages through property identifiers could be theoretically feasible, but the respondent considered this approach likely to be laborious. Cadastral parcels may have very different shapes and sizes, which

according to a respondent, is an argument to abandon the strict use of boundaries especially in cases where they do not follow any natural physical or functional boundaries.

The small area division is stable over time

Temporal comparability was seen as a good feature for small areas. Small areas inevitably require updates, so the handling of changes should be systematic and consider temporal comparability in retrospect.

In change management, both the dynamics of growing areas and the changes in declining areas should be taken into account. In growing areas, new residential areas are created through planning, which requires dividing areas into smaller ones at times. On the other hand, declining areas require adjusting boundaries to avert population declines below the data protection threshold, meaning areas may need to be merged into larger ones.

Municipal merges may also lead to changes in area identification codes. One respondent noted that in potential municipal mergers, new small areas could be created from these divisions.

As for updates, there was a wish expressed that major changes should only be made relatively rarely, perhaps every 10 years. Smaller area merges and separations on the same boundaries could be done more frequently. In such cases, general criteria should be used. Some respondents emphasized the inevitability of change and wished for regular update intervals, as well as access to previous versions.

The basic units can be easily combined for different uses

Integrability was considered essential to enable small areas to be used for various purposes. The term “basic unit” caused some confusion among respondents, as they were not sure what it referred to. Several responses expressed a desire for a small area division system that would be hierarchical.

By the term "basic unit", it was meant the smallest area units that are intended to be created in large quantities. According to one respondent's view, there should be as many of these basic units as possible, so that they can be combined flexibly into different types of area divisions for various purposes. These "small blocks" would be considered the basic units. The same respondent suggested using attribute tables and combining codes, as well as area type classifications, to facilitate integration. These might include municipality codes, potential sub-area codes, urban-rural status, primary land use category, or status of building stock, etc.

Integrability is the point where the clarity of the entire small area division system is measured. It should be logical, hierarchical, and clearly integrable. In practical terms, it was emphasized that area codes should not start with zero.

The largest possible number of areas that meet the selected criteria

This point was not clearly defined in the survey question, as many respondents did not understand what was meant by it. This goal referred to having the maximum possible number of small areas that meet other specified criteria. In the responses, it was noted that it's better to have too many areas than too few. In this case, other goals are easier to achieve, such as aligning to municipality boundaries and taking into account the urban and spatial structure.

3.6. Criteria related to the delimitation of small areas

In addition to the goals, the survey also asked for opinions on criteria for delimiting small areas in general terms. Five different ways to define and delimit small areas were presented as options, and respondents were asked to select the three most important ones (see Figure 3). They were also encouraged to provide comments on these options.

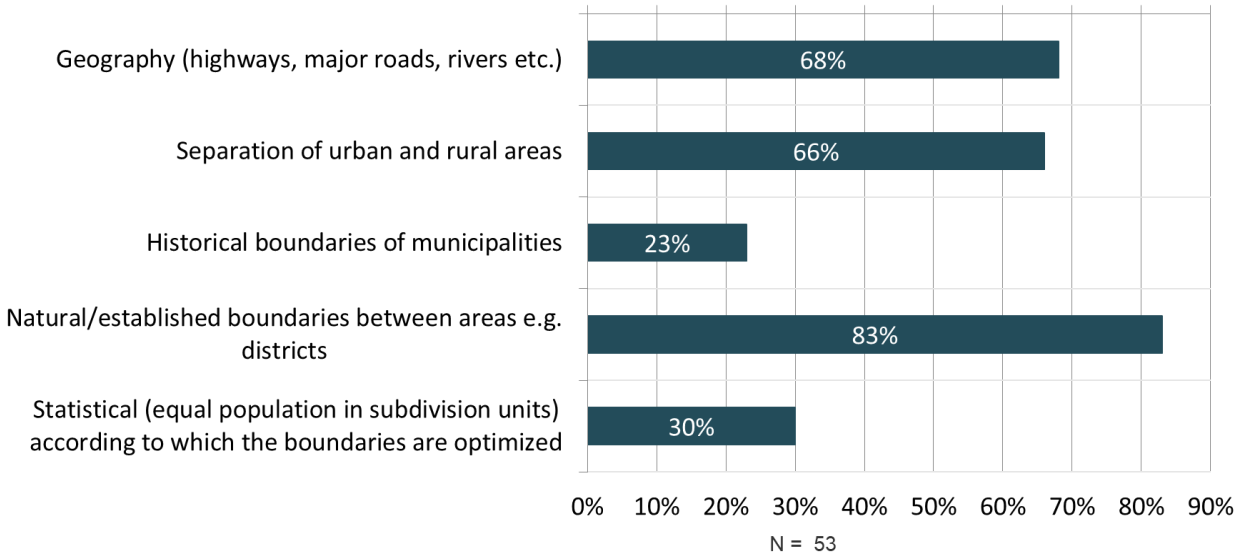

Figure 3. Response distribution to the question: "Boundary between area units can be established from various perspectives, taking into account different factors. Which of the following three are the most important?"

Most mentioned was the natural/established boundaries between areas, such as city districts, which 83% of respondents selected. Around two-thirds of respondents chose the geographical boundaries and the separation of urban and rural areas. The fewest mentions were given to historical municipal boundaries and an approach based on statistical equal population criteria. In the comments, it was noted that Finnish localities boundaries are artificial and a more detailed boundaries based on land use/land cover categories should be preferred. Similarly, sufficient accuracy for municipal user needs was asked. While the locality boundary not directly used, it can serve as an estimate of where in the urban structure a different approach for criteria is needed. This suggests that respondents value the distinction between urban and rural areas and see them as fundamentally different, requiring separate treatment.

From the respondent's perspective, historical municipal boundaries and actual/natural boundary between areas were considered essentially the same thing. One respondent saw the criteria as somewhat overlapping, while another hoped for areas that residents would also recognize. A representative from a large city noted that even current city district boundaries do not fully represent natural or actual boundaries. It is challenging to use natural borders as a basis for areal divisions because people perceive areas in different ways.

The population size criterion was seen as more of a threshold question in hierarchical small area division system, where some variation in population is allowed, but the at the same hierarchical level area's variation in population should not be too large.

One respondent hoped for service areas rather than geographical areas. Another emphasized the importance of connectivity to transportation networks as a functional feature.

Natural areas were considered suitable for definition by combining natural geography and infrastructure-related factors, as well as historical factors related to functionality. On rural areas, historical school districts could be a starting point for defining functional areas, according to one respondent. Place names were also suggested as a basis for identifying areas on rural areas. In general, it was hoped that municipalities would be divided more precisely than just into urban and rural areas.

The overall outcome of the responses was that in defining smaller areas, it is good to use multiple criteria to take into account various factors. Boundary definitions based on geographical factors can initially form rough outlines with clear borders. Statistical optimization was seen as neutral but artificial, representing areas that do not actually exist.

3.7. Summary of user needs

The survey generally showed the need for a national small area division system. Based on the responses to the survey, we identified several different categories of user needs:

1. General need for more accurate small area statistics,
2. Operational needs of municipalities and cities,
3. Regional actors' needs for consistent comparison of larger areas,
4. National actors' needs for comparable and sufficiently accurate area data,
5. Businesses' needs to integrate their own datasets with the general small area division system,
6. General need for development monitoring and forecasting across different area levels in a comparable way,
7. Concerns about open data and data security regarding access to area data,
8. Research needs for flexible areal divisions and multi-dimensional small area classification systems,
9. Clarification of the current confusing and poorly compatible small area divisions,
10. Improved interoperability in geospatial data.

4. Objectives for implementing the national small area division model

Based on international examples and survey of data needs, the project further defined the objectives for developing a national model for small area divisions.

Main objectives are:

1. Establish the basis for a small area division model that defines as stable as possible basic units for small area divisions and principles for generating small area divisions using base areas
2. The small area division system follows the principles of interoperability
3. The boundaries of the basic units of small areas are defined uniformly so that they can be flexibly used to form comparable small area divisions
4. From a data protection perspective, an area unit is created that provides as accurate as possible geographic information
5. Principles for managing the model are established

Sub-objectives and measures for implementing them have been defined under these main objectives. Not all actions will be implemented in the GSFI project, but identifying them is important for the continuation. The following table summarizes the objectives and actions identified in the project.

Table 1. Objectives for small area division model

Objective	Sub-objective	Action 1	Action 2	Action 3
1. Establish the foundations for a small area division system that defines the most stable possible base units for small area divisions and the principles governing these small area divisions.	1.1 The areal division system should incorporate a hierarchical structure, where smaller area units are combined to form larger ones.	Define and describe the hierarchical structure	Link the hierarchical structure to area identifiers	
	1.2 The model will define principles for creating various types of area divisions.	Define criteria for combining small areas from a statistical unit perspective	Propose guidelines for aggregating small areas into generalized areal divisions.	
2. The areal division system adheres to principles of interoperability	2.1 Small-areas are compatible with administrative boundary definitions	Areas are made compatible with municipal boundaries	Compatibility with property boundaries is assessed	Interoperability with municipalities' own small area divisions is assessed
	2.2 The system is based on systematic identifiers	The system uses logical identifiers, machine-readable codes, and area names	Defining logical identifiers and codes hierarchically, ensuring interoperability with administrative units	Defining the naming conventions and practices for small areas

3. The boundaries of the basic units of small areas are defined consistently so that they can be flexibly used to create comparable small area divisions	3.1 The criteria for defining the boundaries of basic units take into account different types of areas and the need to consider their specific characteristics. For example, urban areas, rural areas, urban growth zones, archipelago areas, and uninhabited areas	Areas are formed to follow naturally occurring geographical boundaries for each area and consider both the physical and functional aspects of spatial structure		
	3.2 Area boundary definitions utilize tested and suitable data sources	Testing the suitability of various geospatial datasets as basis for boundaries		
4. From a data privacy perspective, an area unit is created that provides as accurate area information as possible	4.1 Legal and ethical considerations surrounding data privacy of statistical units and ensuring interoperability guide the definition of small areas	Population thresholds are set for area units		
	4.2 Clear guidelines and rules for publishing data at the area unit level are defined	Data privacy criteria and the hierarchical structure of the areal division system are defined together		
5. Principles for model administration are drafted	5.1 Criteria and guidelines for change management are established	In the development of basic units, minimize the need for boundary updates by, for example, keeping the number of basic units large	Growing urban areas and the development of new built-up areas are taken into account when handling changes	The necessary update cycle and implementation methods for updates are evaluated
	5.2 Links and roles are identified to ensure the continuity of model maintenance and development	Maintenance needs and resources are clarified from the perspectives of different stakeholders	A proposal for model administration is drafted after the development project.	

In addition to the objectives related to the small area division model, the project has identified quality aspects for the data set of small areas. The table below summarizes the objectives related to this data set. These objectives focus on how to ensure the usability and usefulness of small areas and how to consider the user perspective in general. These can also be used as evaluation criteria for the usability of the small area divisions.

Table 2. Objectives for small area dataset

Objective	Sub-objective 1	Sub-objective 2	Sub-objective 3
Small areas are clear	Meaningfulness and aesthetics of boundaries are considered from a visualization perspective		
Suitable for handling regional variations of different phenomena	Defined areas can effectively capture and represent differences across regions for various types of data or phenomena	The areal division allows for thematic classifications describing areas	
Usability	Open availability of area boundaries and identifiers	The data linked to the small-areas is made easily accessible and usable	Clear definition and scope of data content at different spatial accuracy levels
Spatial coverage	Water areas are also classified		

5. Testing different approaches to forming base-areas

5.1. Register villages

A land registry plot number is a unique identifier for real estate in the Finnish Land Registry. A plot number consists of parts that identify each individual property. Before April 10, 2014, a plot number consisted of municipality-, location area-, group- and unit numbers. The "location area" referred to a registered village or district (source: https://www.finlex.fi/fi/laki/alkup/1996/19960970). Therefore, the plot number can be used as an identifier to describe historical territorial divisions used in property registration.

A register village refers to a concept formed during the great land reform (isojako), which defined the location area number before 2014. After 2014, the concept of register village is no longer used. Thus, plot number is no longer used to define the register village, and this information is not maintained.

The project tested the use of plot numbers as area identifiers based on the location area number. This resulted in combining larger areas within municipalities by including all properties. The data yielded promising results, and in some municipalities, the areas were divided into clear units. However, the challenge was that the resulting areas in municipalities are very diverse. Properties and register villages are defined through settlement history in different places across Finland, resulting in either very large or small areas in many cases. Especially within built-up areas, there were significant variations in the types of areas formed using plot numbers, making it impossible to establish a comparable method for defining areas based on this data. In rural areas, the main finding was that registered villages formed functional and meaningful areas in municipalities where they were based on regional clusters of shared road connections (Figure 4).

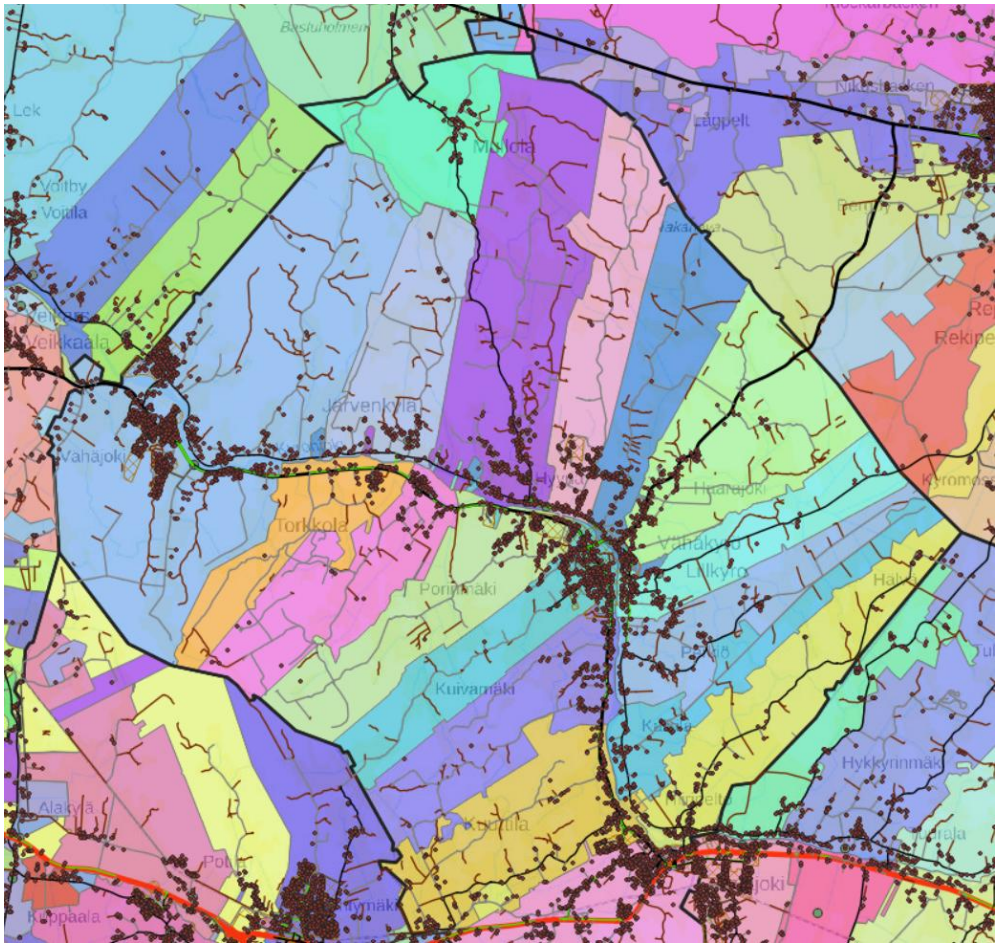

Figure 4. Example of historical register village areas based on location area numbers in registry plot number in former Vähäkyrö municipality. Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Shoreline10 and River Network: Syke, National Land Survey Finland. Division into administrative areas (vector): National Land Survey Finland. Cadastral parcels: National Land Survey Finland. Building and Dwelling Register (BDR): Population Information System, Digital and Population Data Services Agency, 1/2022. Background map: National Land Survey Finland.

5.2. Utilizing property boundaries for defining base-areas

The project experimented with using property boundaries as the basis for defining base-areas, particularly in urbanized areas. The goal was to explore how well built-up urban fabric could be divided into blocks based on property lines.

Data processing method for detecting built blocks for base-areas

- First, parcels classified as "built" were selected using geospatial methods. This involved calculating the ratio of built area to the total area for each property and selecting those where the built proportion exceeded 4%.
- Traffic areas and parks are generally their own land plots, so they surround and delimit built-up properties.

- By dissolving adjacent built plots, separate "islands" emerged that generally reflected the structure of a block system in urban context.

Results

This approach proved effective in creating precise and detailed block-level base-areas, especially within city centers (Figure 5). It also allowed for accurate delimitation of blocks not only within planned urban areas but also in suburban areas characterized by terminating roads.

Challenges

A particular challenge arose from the variability of property boundary data between municipalities. Historical differences in land use, ownership patterns, and land development practices across towns significantly affect how accurately property boundaries follow natural boundaries or group built environments into similar units. Particularly in smaller towns, street areas within planned urban zones might belong to neighboring built properties. This makes it impossible to systematically divide the area using only property data with a consistent method. Additionally, in this property-boundary-based method, undeveloped areas between buildings would need to be delimited using different data and criteria.

Figure 5. Example of delimiting built-up blocks based on property boundaries. Data sources: Cadastral parcels: National Land Survey Finland. Building and Dwelling Register (BDR): Population Information System, Digital and Population Data Services Agency, 1/2022. Background map: National Land Survey Finland.

5.3. Blocks defined by road and street networks

In planned urban areas and towns, the road network often systematically defines block structures. This is particularly true in grid-based urban layouts.

Methodology

- In the experiment, Digiroad's data on roads and streets was used, where roads are represented as lines and categorized by road type. The aim was to define areas enclosed by these roads as individual blocks or "units."
- Since a single transportation corridor might contain multiple parallel road lines in the dataset, a method was needed to estimate the centerline of the traffic area as a boundary between units. Such data does not exist pre-made.
- A trial version was developed where road lines were slightly expanded ("buffered") and then combined to define a central line for the traffic zone. This served as the approximate border between blocks.

Result

The method yielded relatively consistent results, but it sacrificed precise alignment with the actual centerline of roads. The final dataset contained some inaccuracies and randomness due to this approximation. Nevertheless, using road network data allowed for standardized criteria for defining units. Directly using only road network data, however, wouldn't allow for the formation of uniform blocks. Using solely road and street networks would result in very different sized blocks even within planned urban areas. Therefore, additional data sources are needed in some parts of urban areas to divide regions.

Challenges

Despite the method used to estimate the centerlines of traffic areas, the results included many very small units around junctions and intersections, often without buildings. Similarly, road lines rarely align with coastlines, so coastal areas remain as large, continuous units. In areas with cul-de-sacs (roads that end), built-up areas don't form clearly defined blocks, but rather relatively large areas where buildings may be located on different sides connected by various roads (Figure 6).

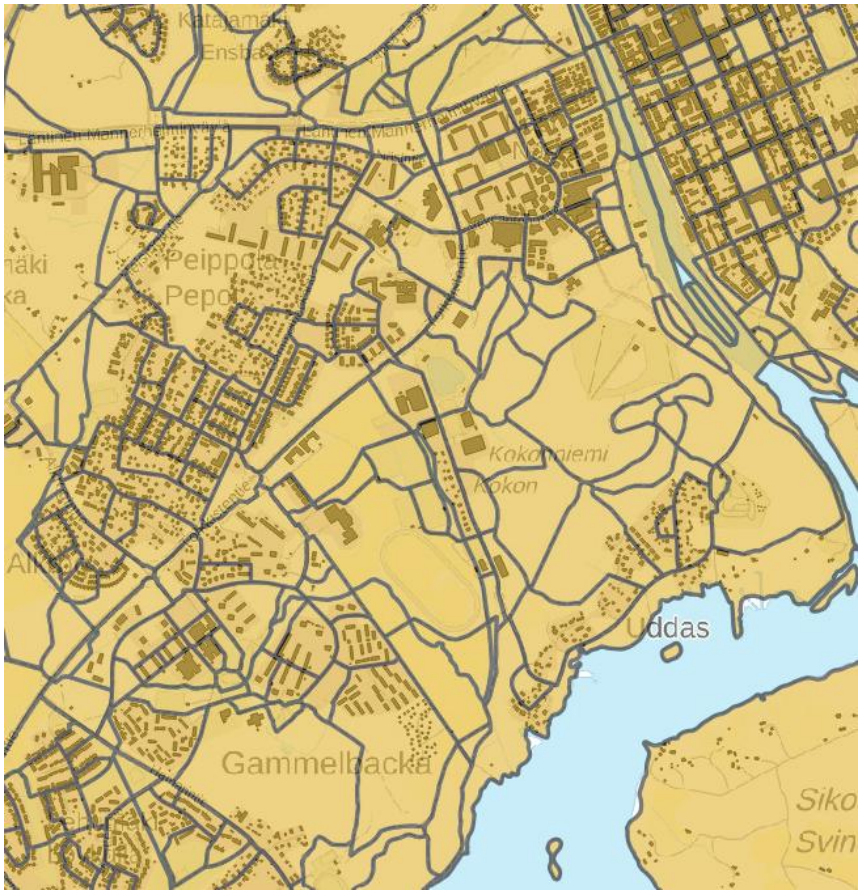

Figure 6. Example of blocks defined by road and street networks using all road categories in Digiroad dataset. Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features (vector): Topographic Database, National Land Survey Finland 04/2024. Background map: National Land Survey Finland.

5.4. Definition of small areas from base-areas

The creation of small areas or also larger area combinations from small building blocks often relies on either aggregating or dividing approaches, or a combination of these methods. However, the goals and primary use purposes set for smaller areas largely determine the basis, starting points, and how they are formed. Various methods of forming areas have been developed internationally from different starting points. Each method has its own strengths and is often best suited to its original purpose. In research literature, applications are often placed under the term "regionalisation", but depending on the starting point, they may also be placed under definitions and delimitation of neighborhood or community, or even functional region.

Regardless of the approach, the goal of applications is to identify spatially homogeneous areas under certain decision-making rules and constraints in such a way that the internal similarity or difference between areas is maximized with respect to selected variables. Approaches utilize various methods, including both spatial and non-spatial clustering techniques as well as probability- and network-based optimization and simulation methods.

Summaries of different approaches have been presented by, among others, Dugue ym. (2007), Aydin ym. (2021), Klapka & Halas (2016), Lancichinetti & Fortunato (2010).

In this project, the available pre-existing or semi-finished algorithms were compared to the identified objectives and guiding categorization principles in the project. The comparison and evaluation were not exhaustive due to the large number of alternative options. Open-source Python and R libraries (e.g., Pysal, spdeb, rgeoda) provide implementations of algorithms for area formation (regionalisation), such as MaxP (Wei ym. 2021), SKATER (Asuncao ym. 2006), and AZP (Openshaw & Rao 1995). These algorithms aim to identify either a predetermined number of areas or the maximum number of areas under certain constraints, such as a minimum population size. The algorithms attempt to form as homogeneous areas as possible with respect to the selected variable. Experiments showed that while these algorithms work well for forming homogeneous statistical areas, it is difficult to implement complex functional relationships between different base-areas (e.g., watercourse or railway barriers, consideration of islands). However, the survey highlighted the importance of considering the urban structure and various geographical obstacles in defining small areas. On the other hand, the base-area dataset itself has value and can be used for case-specific analyses using these algorithms.

6. Proposal for a small area division model

A small area delimitation model is being developed to improve the availability of statistical data at the level of more detailed spatial units than municipalities. Currently, there is no uniform national-level spatial unit of small areas in Finland. The proposed small area division model aims to create a system that forms more detailed statistical boundaries inside the municipalities. The model's goal is to provide an administrative boundary-compliant and unified small area division system. The new small area division model seeks to complement the existing municipal small areas by providing basic data sets for the flexible formation of various types of small areas, making it possible to produce uniformly defined spatial units nationwide. The system includes all aspects necessary for defining and maintaining small areas, including geospatial datasets, metadata, data architecture, maintenance and update procedures, and data dissemination systems.

In the GSFI vision developed in this project, national spatial units and their production and management are organized, using small area division model-based areas or statistical grids. National spatial units are produced only once and made available for public use.

The development of a small area delimitation model on such a large scale is still ongoing, and in this project, the related needs and objectives have been identified. The main goal of this project has been to pilot the production of small areas using available data sets, i.e., producing a geospatial data set of small area delineations.

The small area division model provides a nationally uniform and comparable data set that improves our understanding of spatial phenomena while taking into account data protection and compatibility with administrative boundaries. The model is based on an approach where first, a very detailed level base-area data set is created, which serves as the basis for defining actual statistical small areas.

The base-areas take into account urban structures and geographical factors. With the help of base-areas, it is possible to create small areas based on various criteria. In the definition of statistical small areas, reliance is made on population-based criteria that ensure data protection compliance in terms of publishing statistical data.

Below, the development work done in this project for creating base-areas nationwide and defining a pilot version for small areas is described, as well as presenting the results.

6.1. Method for defining base-areas

6.1.1. Data

The approach based on base-areas is similar to methods used in several countries that conduct population censuses (e.g., Australian Bureau of Statistics, 2024; Hugo 2007). In the implementation of base-areas, it was decided to utilize nationwide comprehensive road and transportation network data sets, building register information, shoreline and river data, as well as land registry data on parcel division. Roads and streets serve as dividing and connecting elements for areas, and they are relatively stable, making them suitable as a basis for defining the boundaries of base-areas.

The general starting point is that areas with detailed planning represent areas where building density is higher, and where it is possible to identify block-like base-areas using streets as boundaries. Since the data set describing urban-planned areas in Finland is sometimes incomplete, it has been supplemented with a data set estimating area efficiency exceeding 0.02 in a 250x250 meter grid cell. In sparsely populated areas that do not meet the criteria, road network data is used as a connecting element, i.e., on these areas buildings are grouped to seed areas based on how roads combine areas.

In the method, dense built-up areas and sparsely populated areas are first identified with sufficient human activity and physical infrastructure to create seeds (seed areas) that enable the entire study area (municipality) to be divided into a sufficient number of parts, i.e., base-areas. These base-areas form a topologically coherent surface that covers the entire area.

6.1.2. Seed areas in densely built-up areas

The seed areas of densely built-up areas are units formed by combining property boundaries and road network, and in addition, where property lines do not function well enough, building polygons are used as well. Below is a detailed description of the method, step by step.

Step 1: Calculate Ground Space Index (GSI)

Calculate the percentage of ground space covered by buildings for each property area. Select areas that are relatively compact and densely built-up, with a GSI of at least 4% and an area size of less than 3 hectares. This results in areas where properties function well as a boundary unit.

Step 2: Merge adjacent property areas

Merge the identified property areas into cohesive building blocks. Areas are merged if their polygons are topologically adjacent.

Step 3: Refine boundaries using network data

Since even small property areas may contain ambiguities, refine them further using pre-defined spatial and physical barriers (transportation networks, water bodies, and railroads). This ensures that the base-areas do not extend beyond these barriers.

Step 4: Filter out non-building areas

Remove areas that no longer meet the 4% GSI criterion. This removes non-building areas created during the refinement process. The result is a set of complete building blocks.

Step 5: Buffer building footprints of buildings located in properties with GSI lower than 4%.

Buffer the building footprint polygons that are located outside the building blocks defined using property criteria. Buffer distance is calculated based on the average distance between buildings within each building block defined using property criteria (approximately 20 meters).

Step 6: Repeat steps 1-4 for new areas

Repeat the process for newly created areas resulting from the previous step.

Step 7: Combine data into final result

Combine the data to create a final representation of building blocks.

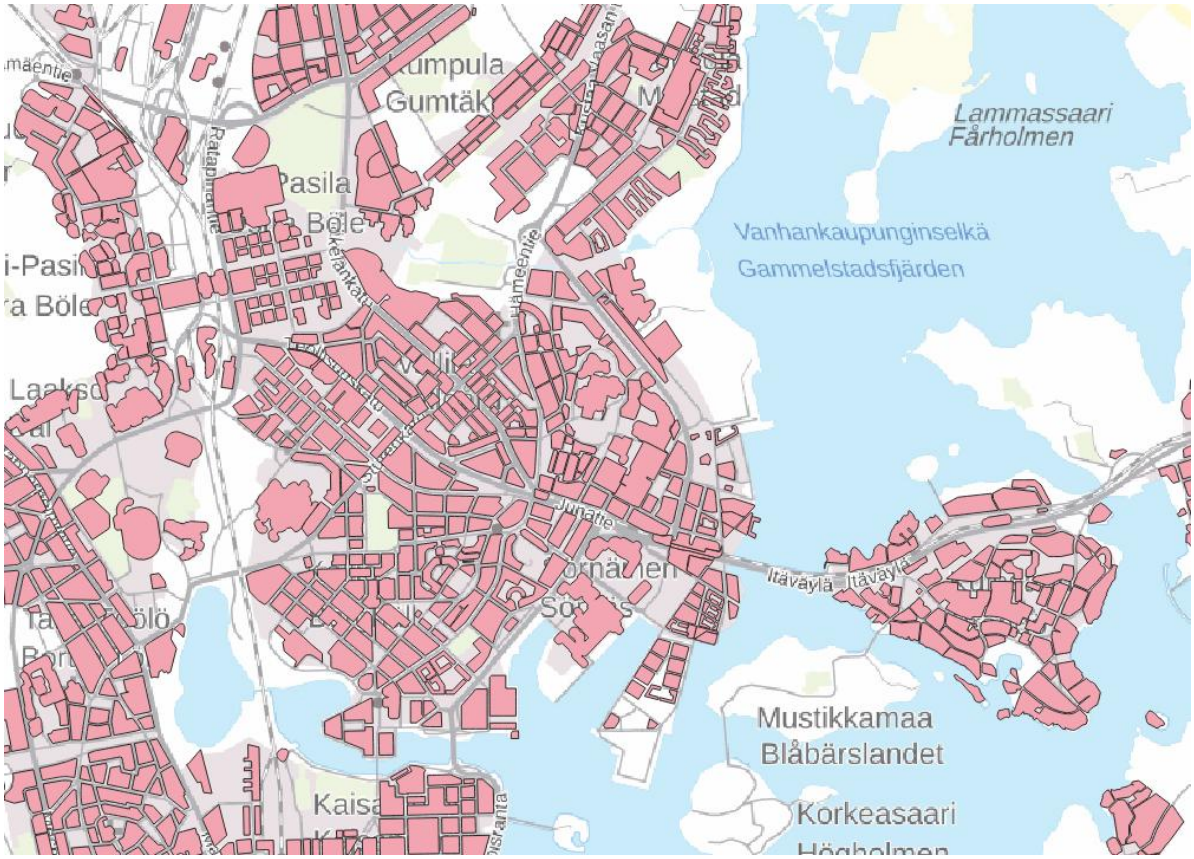

Figure 7. Example of seed areas identified using property blocks and building footprints. Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features and water bodies (vector): Topographic Database, National Land Survey Finland 04/2024. Background map: National Land Survey Finland.

6.1.3. Seed areas in sparsely built-up areas and in archipelago

The seed areas in sparse rural areas are based on the road network and its interconnected segments The method is described in detail step by step.

Step 1: Classify road network

Finnish road network is classified into functional categories based on its operational significance. Based on this data, distinguish between main roads and secondary roads.

Step 2: Create segments into main roads

Identify each 2-kilometer segment of a main road as an individual route, assigning it a unique identifier.

Step 3: Merge secondary roads into network

Secondary roads form a connecting network on sparse rural areas, linking areas to the main road network. If this network does not contain any main roads, each separate secondary road network forms its own spatially isolated sub-network.

Step 4: Split large networks into smaller sections

Since secondary road networks can be quite extensive and cover large areas, and may connect to multiple main roads, each separate secondary road network is split into smaller sections in the following two steps:

Step 5: If a secondary road network connects to only one main road segment, it remains unchanged.

Step 6: If a secondary road network connects to multiple main road segments, it is divided into sections based on which segment has the shortest distance from each point of interest. The calculation weights the lowest level of the road network ten times more heavily. This way, networks are divided more finely at areas where roads may not be drivable.

Step 7: Finally, remove secondary road networks that do not connect to other parts of the road network or consist only of a single tie element. These often represent small sections that are insignificant for connectivity and aggregation (e.g., rest stops on motorways).

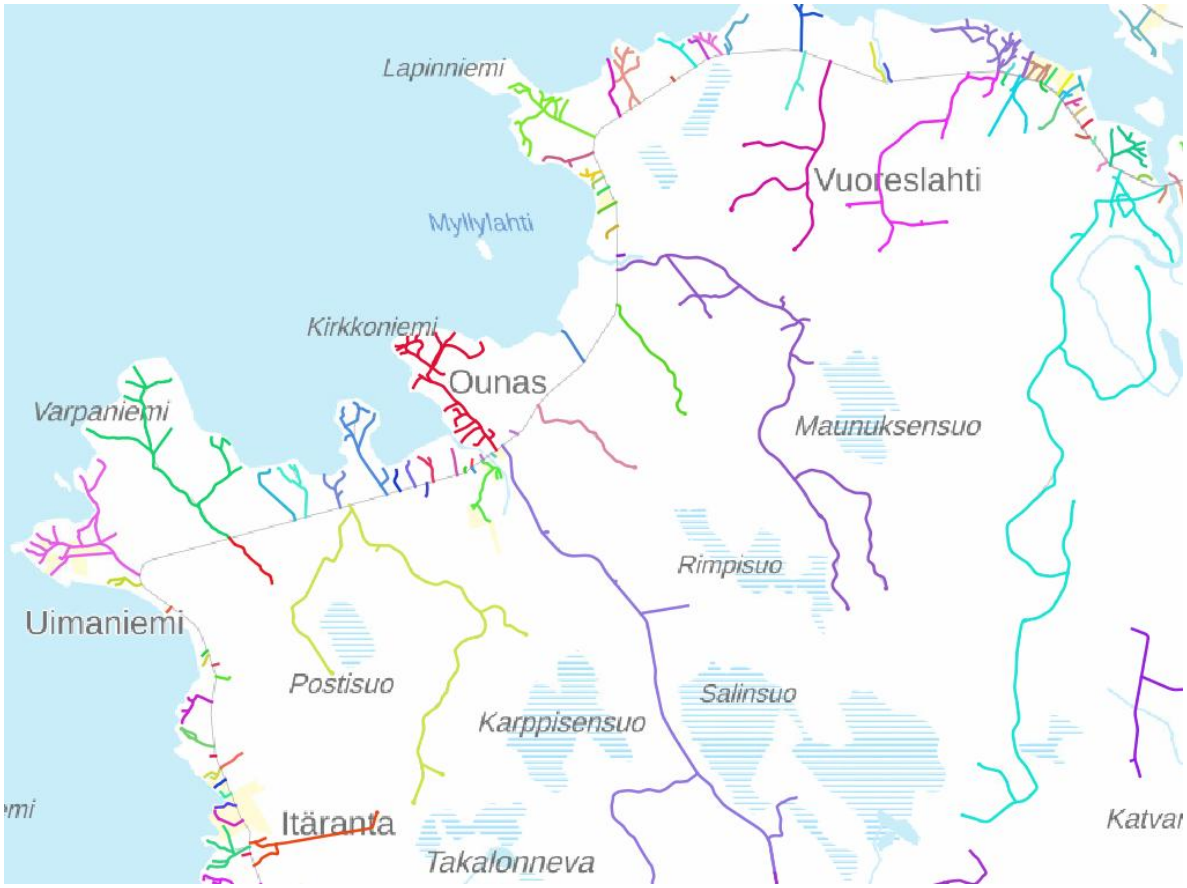

Figure 8. Example of road network based seed area division. Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features and water bodies (vector): Topographic Database, National Land Survey Finland 04/2024. Background map: National Land Survey Finland.

In archipelagos, the seed areas are primarily based on dense and sparse residential areas described earlier. However, in many areas of the archipelago, the road network or building-based approach does not adequately identify the seed areas, resulting in large base-areas being formed.

Step 8: Identify islands in Finland's archipelago data set that do not have seed areas identified using previous method of detecting dense or sparse residential areas, but contain either permanent or seasonal residents.

6.1.4. Generating base-areas using seed areas

Based on identified seed areas, base-areas are produced which form a continuous topological surface. In this method, the areas between seed areas are divided according to certain criteria and decision rules. The method uses raster-based cost-distance analysis and has been implemented with a 5-meter grid.

Here's the method described in detail:

Step 1: Formation of continuous clusters of building polygons and modified land use areas. Building clusters are formed of buildings located within 20 meters of each other. The following categories from the national data on constructed land uses are extracted, which may not be covered comprehensively

by other datasets: port areas, airport areas, areas for extraction of soil and dump sites, recreation areas, cemeteries, crop plots, construction sites.

Step 2: Creating raster: Seed areas are combined and transformed into a raster dataset, which is used for demarcating base-areas. This dataset is referred to as the "source cells" throughout the process.

Step 3: Formation of the first cost raster: The first cost raster is created, which consists of all raster cells with a value of 1 except for the defined obstacles, which have an NA value. Obstacles are water bodies, main roads, and railways.

Step 4: First allocation of costs: The first allocation of costs determines the shortest distance to any source cell. The NA values in the cost raster are not processed, and analysis does not cross over them.

Step 5: Information about the nearest source cell is transmitted to building clusters and areas with strongly modified land use data created in step 1. Now all building clusters and modified land use areas have the source cell value it is mostly integated with. Convert these to raster and and add source to cells.

Step 6: Second allocation of costs: A second allocation of costs is created without changing the cost raster but using updated source cells. This ensures that area boundaries do not pass through buildings and that adjacent areas with similar characteristics are allocated to the same base-area.

Step 7: Update of the cost raster: The cost raster is updated so that water bodies, roads, and railways are removed as obstacles but have a high cost assigned. For water bodies and railways, the maximum possible cost (10,000) is used. For roads, the cost depends on the road category, with higher categories resulting in higher costs.

Step 8: Third allocation of costs: A third allocation of costs is created using the updated cost raster and the source cells from stage 7. In this stage, previously defined obstacles are allocated to base-areas, and area boundaries are fixed to each other, creating a continuous surface.

Step 9: Final adjustment of boundaries in islands: The final step involves adjusting the boundaries of islands so that each island belongs to only one base-area, and transforming the raster dataset into polygons, which generalizes the boundaries due to the raster-based method resulting in jagged edges.

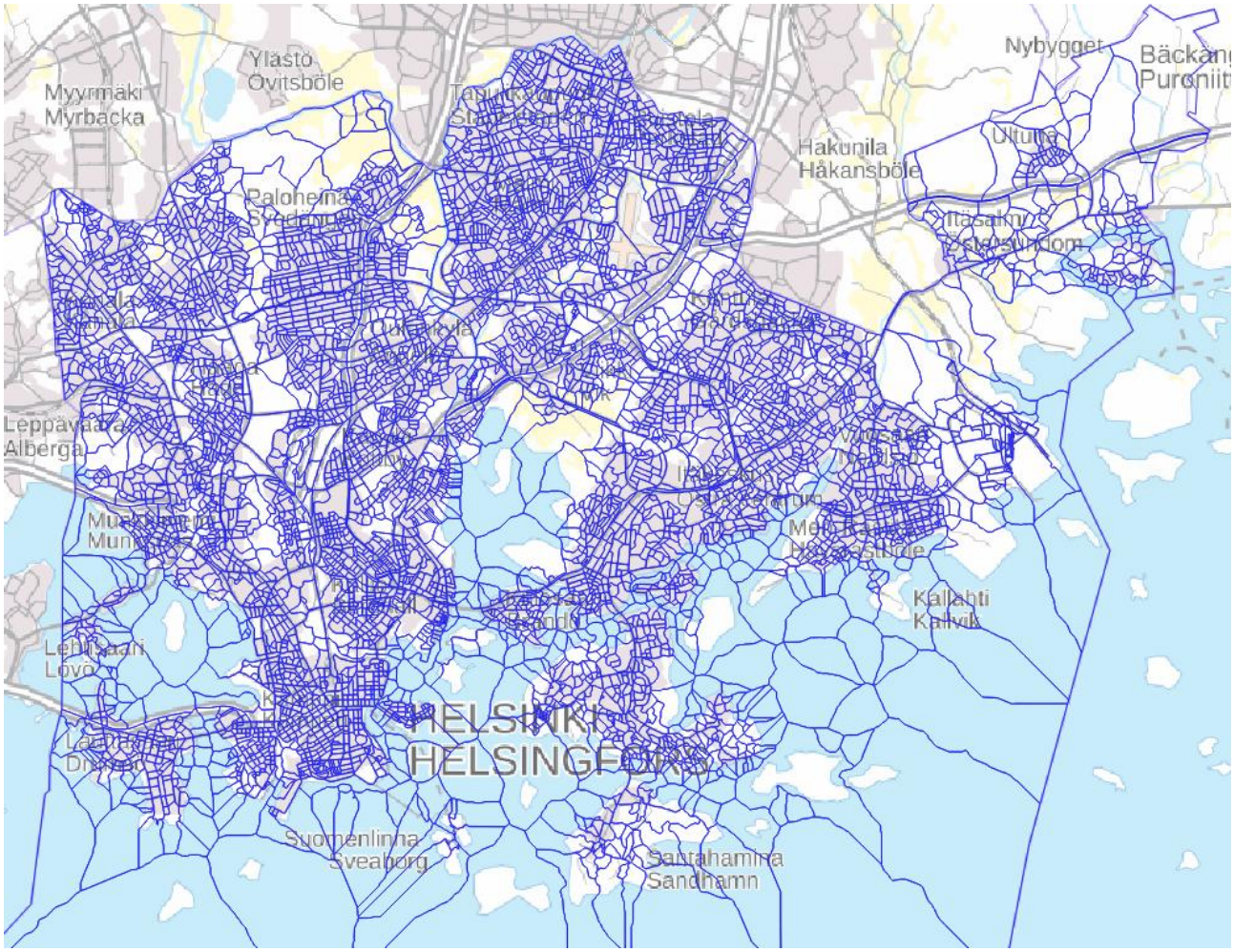

Figure 9. Pilot version of base-areas in Helsinki. Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features and water bodies (vector): Topographic Database, National Land Survey Finland 04/2024. Background map: National Land Survey Finland.

6.1.5. Requirements for improving the process of forming base-areas

Produced base-area data enables the formation of small areas, but producing a high-quality national base-area dataset requires the review of the material and improvements of certain problems related to area formation in specific situations. The problems relate mainly to the source data and specific territory types where choices have to be made in the definition phase to achieve the acceptable solution for as many areas as possible. Especially challenging are island and coastal areas, as well as edges of urban areas. These issues will be tackled in a subsequent project, which creates the test version based on pilot data and also fine-tunes the automated method. The advantage of an automated method is that the entire process can be easily documented and large amounts of data can be handled efficiently. The disadvantage of an automated method is that there are always some areas where the result does not accurately define the area as intended. In terms of base-areas, the goal is to create a national base-area dataset only once and meet its update needs by making changes only in

those sections where the built environment has changed in such a way that changes are necessary for the model's functionality.

Recognized development targets in the method and datasets:

1. The use of combination of detailed planned areas and building efficiency data to categorize densely populated areas is not an optimal solution due to shortcomings of planned area data. A better result can be achieved by producing block-based seed areas only based on building density and land-use efficiency, without being tied to these datasets. This will also make it easier to meet development target 2.
2. In the pilot version, the base-areas in densely built-up areas are mainly based on identifying property parcels and their combinations as seed units. Although the method minimizes the uncertainties of parcel-based data, they still pose a risk for the quality of the final results. A better result can be achieved by forming compact building blocks purely based on building polygons using a buffer-like method.
3. The method does not identify unbuilt compact block areas as base-areas in the best possible way. These areas are likely to become built-up blocks in the future, especially in urban areas or their surroundings. Identifying them in advance would reduce the number of cases that will need to be changed in the future.
4. In sparsely populated areas, the road network-based seed areas rely on the functional classification of transportation network data. This information is reliable and available for all road segments. However, classification criteria may differ slightly from one area to another. Alternatively, the transportation network's functional hierarchical classification can also be defined using network analytical methods.
5. By considering street addresses when forming road network-based basic units, it is possible to merge or split parts of the network into larger or smaller units.
6. The final raster-based cost analysis method is a computationally intensive process. This stage can be simplified and made smoother. However, the method is flexible because the formation of boundaries can be guided finely with different datasets emphasizing different aspects.

6.2. A method for defining the pilot version of small areas using base-areas

Base-areas are utilized to generate small areas for each municipality. The methodology comprises two stages. Initially, a neighborhood dataset is created from the base-areas, which outlines the pivotal attributes of each base-area and their inter-relationships. Subsequently, small areas are constructed based on the revised base-area dataset via an optimized algorithm. Below, the methodology is outlined in detail, stage by stage:

First stage: Constructing a Network Model

Step 1. Essential attributes of base-areas, such as population size, are calculated. Information of what seed area method (see 6.1) each base-area is based (urban, rural, island).

Step 2. Adjacency information for each base-area is added to create topologically connected small areas. This ensures that small areas will be spatially contiguous.

Step 3. The most significant structural and functional relationships between base-areas are calculated. For each base-area and its topological neighbors, information is generated on the distance between them (calculated using seed areas) as well as any barriers or connectors that exist between them.

Information about which functional road networks connect or separate areas is also added, along with details about whether there are rivers or railways in between.

Step 4. A network model is created from the data, where base-areas are represented as nodes and their links represent spatial relationships. Weighting factors are assigned to each link based on distance and adjacency characteristics, taking into account the type of base-area being considered.

Second stage: Small area formation algorithm

Based on the network model data generated in the previous stage, a pilot version of small areas is produced. The implementation uses a link-removing network-based algorithm with population size as a parameter (upper and lower bounds). The algorithm has the following main steps:

Step 1. A Minimum Spanning Tree (MST) is constructed from the network. An MST is a subset of links that minimizes the total cost while ensuring all nodes are connected and network does not contain cycles. If one link is removed from an MST, new sub-networks (clusters) emerge.

Step 2. The link is removed from the MST that minimizes the total cost of emerging sub-networks, such that all resulting clusters meet the lower population threshold.

Step 3. If no link can be removed without violating the criteria, the algorithm returns to its initial state.

Step 4. Steps 1 and 2 are repeated until all resulting clusters meet the specified population thresholds.

The algorithm does not globally optimize the outcome but rather produces locally suitable small areas by focusing on the functional and structural aspects of the solution in terms of link costs.

6.3. Pilot version of small areas

Pilot version criteria

The formation of small areas through the described method largely relies on defining limiting and connecting factors. It is essential to determine which factors represent similarity and coherence as well as diversity between areas. Geographical obstacles and elements of urban spatial structure are taken into account, but their weights need to be defined. Another crucial factor influencing the formation of small areas is the permitted minimum or maximum population size that each resulting small area must meet.

Defining link weights

In the pilot version, the definition of link weights was done in a relatively simple way, but the approach allows for considering complex connections and other factors (land use, building density, interaction between areas, etc.) to be taken into account.

The definition of link costs (weights) was done separately for each area type. On sparsely populated areas, the road network typically serves as a connecting factor between areas. In densely built-up areas with a grid-like layout, the road network often separates areas from each other. As a result, on sparsely populated areas, the cost of a link is smaller, the more functionally significant road segments connects it (excluding motorways and dual carriageways, which are obstacles in all cases). On densely built-up areas, larger roads always create a greater obstacle, so the cost is higher. The most significant limiting factors on all area types are railroads and water bodies, which receive high costs. However, on islands entirely surrounded by water bodies, the water body is not an obstacle but rather the only possible connecting factor, and the link's cost is simply its distance. Below is a table summarizing the

costs of obstacles for different land use types. The definition of costs can be adjusted to influence the results as desired, so it is essential to justify their final definition thoroughly.

Table 3. Weight values for obstacles in different area types

Area type	**Dense**	**Sparce**	**island**
Water bodies	1000 + distance	1000 + distance	distance
Railways	1000 + distance	1000 + distance	-
Road network (motorways and dual carriageways)	1000 + distance	1000 + distance	-
Road network: main roads and main streets (functional classes 1-2)	1000 + distance	distance	-
Road network: main roads and main streets (functional classes 3-4)	100 + distance	distance	-
Streets and private roads (functional classes 5-6)	distance	500 + distance	-
Small roads	distance	1000 + distance	-

Population size criteria for small areas

The population size criterion is used to determine the restrictions on the size of small areas. In the pilot version, the minimum population size was set at 1,000 residents and the maximum at 5,000 residents in the more densely populated municipalities (Helsinki, Espoo, Vantaa) and to 200 – 1000 in the more sparsely populated municipalities (Kajaani, Parainen). This means that the population size of small areas is always between these threshold values.

However, in the small area division system objectives (section 4), it is noted that distinguishing between different area types in small area definitions is essential. Therefore, for sparsely populated areas, it would be reasonable to apply lower threshold values than for densely built-up areas. This way, enough small areas can be identified from rural areas as well, and at the same time, in cities, small area division will not result in overly small units when the threshold value is higher. The population size criterion has a significant impact on the final results, so it's essential to define its final values convincingly and taking into account the area type. The implemented method allows for flexible and land-use-type-specific implementation of the population size criterion.

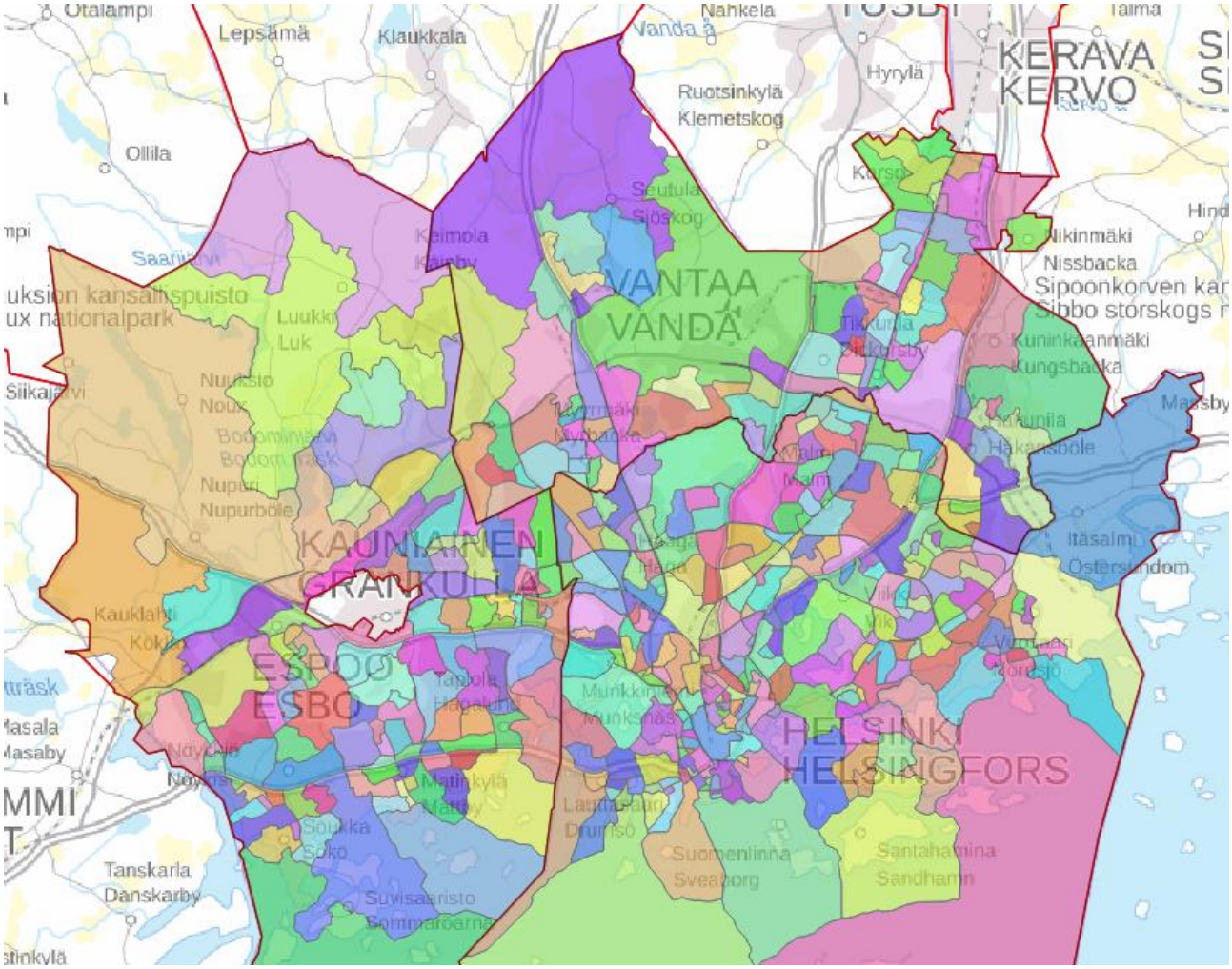

Figure 10. Pilot version of small areas in Helsinki, Espoo and Vantaa (1000-5000 inhabitants). Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features and water bodies (vector): Topographic Database, National Land Survey Finland 04/2024. Municipal boundaries: National Land Survey Finland, Background map: National Land Survey Finland.

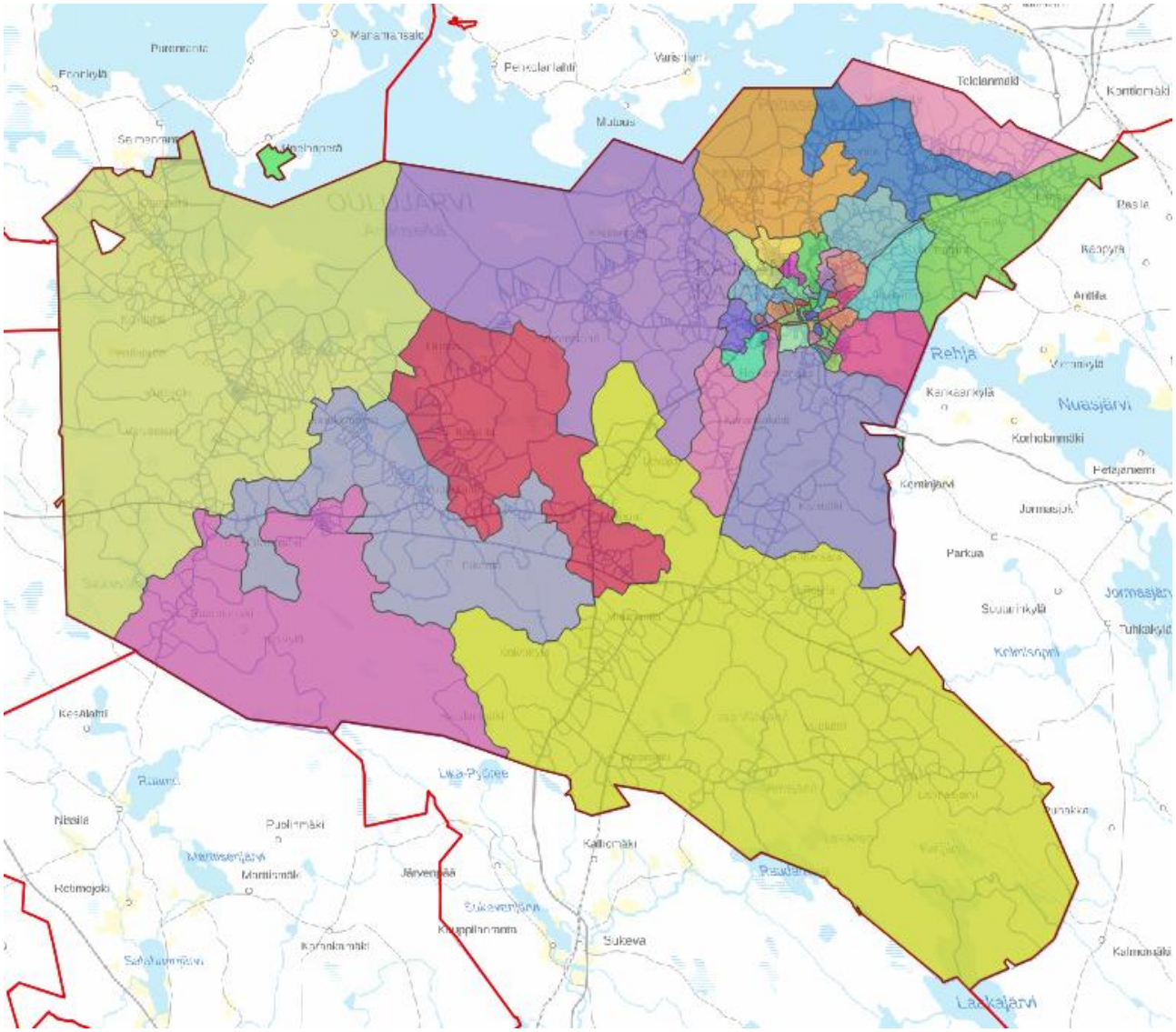

Figure 11. Pilot version of small areas in Kajaani (200-1000 inhabitants). Data sources: Digiroad: Finnish Transport Infrastructure Agency. Data is downloaded from the Download- and Viewing Service of Finnish Transport Infrastructure Agency on 04.04.2024 under the license CC 4.0 BY. Building features and water bodies (vector): Topographic Database, National Land Survey Finland 04/2024. Municipal boundaries: National Land Survey Finland, Background map: National Land Survey Finland.

Development targets for small area formation:

1. More precise definition of factors distinguishing and connecting base-areas: This involves refining the criteria used to separate and connect base-areas.

- More specific threshold values for the population size criterion by area type: This involves defining more accurate and area-type-specific threshold values for the population size criterion.
- Further development and testing of the algorithm: This involves continuing to refine and improve the algorithm's accuracy.

2. Solution based on network components, where certain links (e.g., obstacles) are removed from the outset, and the algorithm is applied to pre-formed sub-networks.
3. Optimization of small area compactness by considering the shape formed by generated areas:
4. Local swapping strategy: This involves fine-tuning the final result locally, following specific rules, by transferring base-areas between different areas in order to improve the outcome (e.g., compactness).

7. Summary

The GSFI project has initiated work on developing a small area division model for Finland. The work was started with preparatory and task-defining sub-tasks, which were carried out interactively with stakeholders to clarify the goals and practical implementation of the project. A survey with over 50 respondents provided material on the need for small area divisions, their uses, and the implementation method. Based on this information, a data need review was compiled and further developed through workshop activities with stakeholders. The progress of the work was presented at meetings of the national Network for Integration of Statistics and Geospatial Information (NISGIF), as well as other stakeholder events.

As a result of these phases, an understanding emerged of what kind of small area division model is needed for Finland. This report describes the basic principles of the model. The implementation of the model has been initiated by envisioning the future structure of national territorial units for statistics. This included considering the overall goals, specific objectives and actions required.

In the second phase of the project, different approaches to defining small area boundaries were tested using GIS methods. Based on the results of the data need review, a hierarchical structure was chosen as the implementation method, where the entire country is first divided into small base-areas, and then these are combined using selected criteria and methods to form small areas within municipalities.

This approach is similar to how small area boundaries are defined in Australia, the UK, and New Zealand. However, the developed method places more emphasis on urban structure, functionality, and geographical obstacles as defining factors for areas. In practice, the work has focused on developing methods that can produce base-areas as consistently as possible across the country using automated processes from various national data sources.

The pilot version of small areas is based on the base-area data set compiled during the project period. The pilot version demonstrates the automated generation of small areas of the same population size, by combining base-areas using a selected method and criteria.

The pilot version does not yet represent a concrete proposal for boundaries of small areas, but rather shows what kind of small areas can be produced with the developed model. The pilot version is therefore a methodological experiment that allows us to understand how the requirements of data protection affect the formation of small areas. The base-areas enable various experiments and simulations using models designed for generating small areas.

8. Literature

Australian Bureau of Statistics (2024). https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/mesh-blocks Webpage (17.12.2024).

Aydin, O., Janikas, Mark. V., Assunção, R. M., & Lee, T. H. (2021). A quantitative comparison of regionalization methods. *International Journal of Geographical Information Science*, *35*(11), 2287–2315. https://doi.org/10.1080/13658816.2021.1905819

Cockings, S., Harfoot, A., Martin, D., & Hornby, D. (2013). Getting the Foundations Right: Spatial Building Blocks for Official Population Statistics. *Environment and Planning A: Economy and Space*, *45*(6), 1403–1420. https://doi.org/10.1068/a45276

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised Regionalization Methods: A Survey. International Regional Science Review, 30(3), 195-220. https://doi.org/10.1177/0160017607301605

Galster, G. (2001). On the Nature of Neighbourhood. *Urban Studies*, *38*(12), 2111–2124. https://doi.org/10.1080/00420980120087072

Hugo, G. (2007). Space, Place, Population and Census Analysis in Australia. *Australian Geographer*, *38*(3), 335–357. https://doi.org/10.1080/00049180701639760

Klapka, P. and Halás, M. (2016) Conceptualising patterns of spatial flows: Five decades of advances in the definition and use of functional regions. Moravian Geographical Reports, Sciendo, Vol. 24 (Issue 2), (2016) pp. 2-11. https://doi.org/10.1515/mgr-2016-0006

Lancichinetti, A. & Fortunato, S. (2010). Community detection algorithms: a comparative analysis. Physical Review E 80. https://doi.org/10.48550/arXiv.0908.1062

Martin, D. (2000). Towards the Geographies of the 2001 UK Census of Population. *Transactions of the Institute of British Geographers*, *25*(3), 321–332. https://doi.org/10.1111/j.0020-2754.2000.00321.x

Openshaw, S., & Rao, L. (1995). Algorithms for Reengineering 1991 Census Geography. Environment and Planning A: Economy and Space, 27(3), 425-446. https://doi.org/10.1068/a270425

Sperling, J. (2012). The tyranny of census geography: small-area data and neighborhood statistics. Cityscape: A Journal of Policy Development and Research 14(2), 2012

Statistics Sweden (2018). Att mäta segregation på låg regional nivå [report in Swedish], https://www.scb.se/contentassets/deedfb3fbe3d4abd987cfcd67dcff2e4/slutrapport-att-matasegregation-pa-lag-regional-niva._ku2017_02404_d.pdf

Stats NZ (2022). Statistical standard for geographic areas 2023. Retrieved from www.stats.govt.nz

Walford, N. S., & Hayles, K. N. (2012). Thirty Years of Geographical (In)consistency in the British Population Census: Steps towards the Harmonisation of Small-Area Census Geography. Population, Space and Place, 18(3), 295–313. https://doi.org/10.1002/psp.658