

UNIVERSITY OF TARTU
Institute of Computer Science
Data Science Curriculum

Aleksandra Rammul

Measuring uncertainty related to ingesting data to DGGS

Master's Thesis (15 ECTS)

Supervisor(s): Alexander Kmoch,
Evelyn Uemaa

Tartu

2025

Measuring uncertainty related to ingesting data to DGGS

Abstract: Discrete Global Grid Systems (DGGS) offer a modern alternative to traditional coordinate-based spatial frameworks by dividing the Earth’s surface into equal-area cells, allowing for standardized, multiresolution geospatial analysis. However, while DGGS helps reduce geometric distortion and improves data handling, it does not eliminate the spatial uncertainty caused by the Modifiable Areal Unit Problem (MAUP). This thesis investigates how aggregation to DGGS grids influences the reliability of landscape metrics and whether uncertainty can be meaningfully quantified across different resolution levels. Using high-resolution land cover data from the Estonian Topographic Database (ETAK), landscape metrics such as patch density, percentage of like adjacencies, Shannon Diversity Index, and class proportion were calculated for hexagonal cells at multiple DGGS resolutions. The land cover class for each cell was assigned using nearest-neighbor matching with raster pixels. The analysis reveals that landscape metrics react differently to resolution changes depending on the spatial structure of the landscape. The results confirm that MAUP cannot be universally measured or avoided. Instead, spatial uncertainty must be approached through context-aware experimental design. While DGGS supports scalable and consistent analysis, its use does not remove the need for careful methodological planning to ensure valid interpretations of spatial patterns.

Key words: DGGS, MAUP, landscape metrics, spatial uncertainty, hexagonal grids, data aggregation, metric sensitivity

CERCS: P510 Physical geography, geomorphology, pedology, cartography, climatology

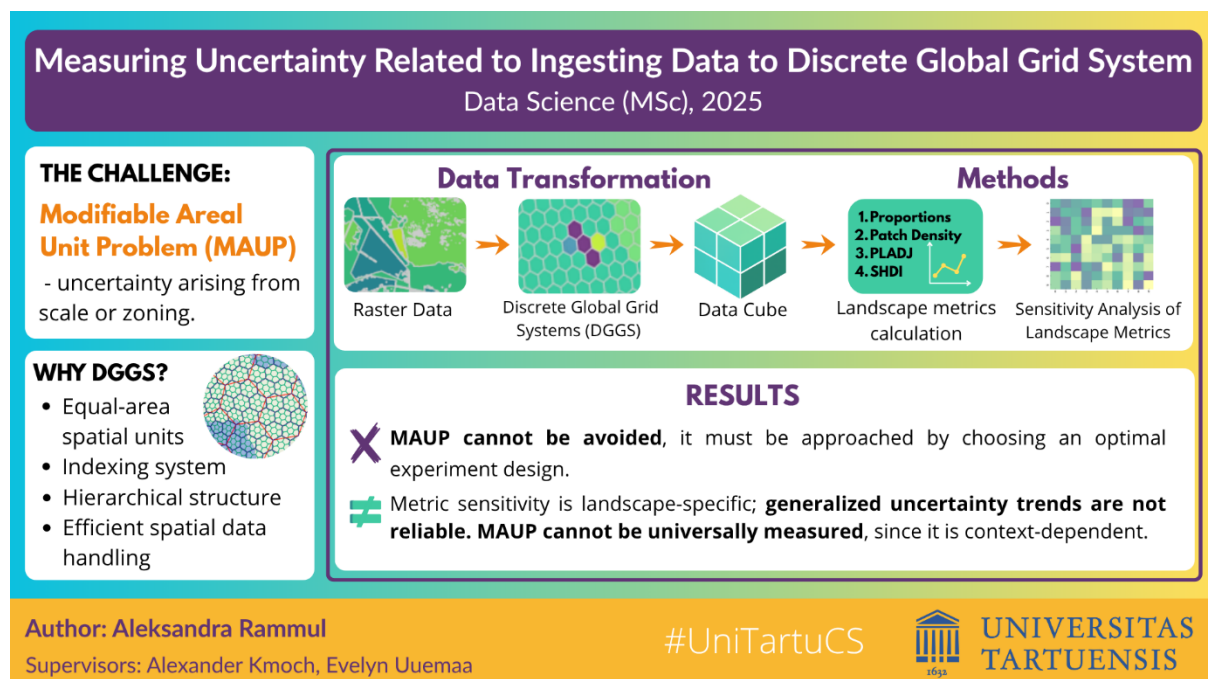


Figure 1. Visual abstract (eng).

Määramatuse mõõtmine andmete teisendamisel Globaalsesse Diskreetsesse Võrku

Lühikokkuvõte: Globaalne diskreetne võrk (DGGS) pakub kaasaegset alternatiivi traditsioonilistele koordinaatpõhistele ruumiraamistikele, jagades Maa pinna võrdse pindalaga ruumilisteks üksusteks. See võimaldab standardiseeritud ja mitmel eraldusvõimel põhinevat ruumiandmete analüüsi. Kuigi DGGS vähendab geomeetrisi moonutusi ja parandab andmetöötluse tõhusust, ei kõrvalda see ruumilist määramatust, mis tuleneb modifitseeritava ruumilise üksuse probleemist (MAUP). Käesolev magistritöö uurib, kuidas DGGS-põhine koondamine mõjutab maastikumõõdikute usaldusväärset ning kas erinevatel lahutusvõime tasemetel esinevat määramatust on võimalik sisuliselt kvantifitseerida. Uuringus kasutatakse Eesti topograafilise andmekogu (ETAK) kõrglahutusega maakasutuse andmeid ning nende põhjal arvutatud maastikumeetrikat, sh laikude tihedus, sarnaste naabrite osakaal (PLADJ), Shannoni mitmekesisuse indeks ja maakasutusklasside osakaal. Iga DGGS-i kuusnurkse lahtri maakatteliik määrati lähima rasterpiksli põhjal. Analüüs näitas, et maastikumõõdikute tundlikkus muutub sõltuvalt ruumilise struktuuri keerukusest ja lahutusvõimest. Tulemused kinnitavad, et MAUP-i ei ole võimalik universaalselt mõõta ega vältida. Ruumilise määramatuse käsitlemine peab põhinema teadlikult kavandatud katse disainil. DGGS toetab küll skaleeritavat ja ühtlustatud analüüsi, kuid ei asenda läbimõeldud meetodilist planeerimist, mis on usaldusväärsete ruumianalüüside eelduseks.

Märksõnad: DGGS, MAUP, maastikumeetrika, ruumiline määramatus, kuusnurkne võrk, andmete agregeerimine, mõõdiku tundlikkus

CERCS: P510 Füüsiline geograafia, geomorfoloogia, mullateadus, kartograafia, klimatoloogia

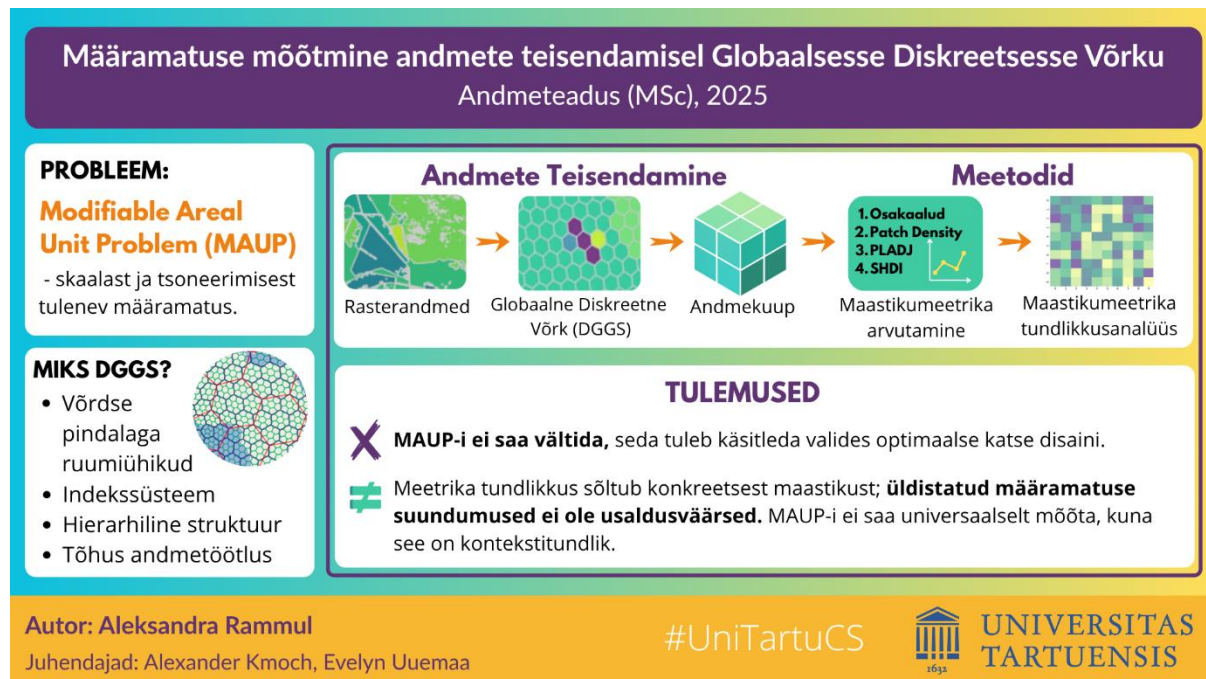


Figure 2. Visual abstract (est).

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Alexander Kmoch and Evelyn Uemaa, for their insightful guidance, continuous support, and patience throughout the writing of this thesis. Their support was essential in helping me develop and refine the focus and structure of this thesis.

Use of AI

Throughout the writing and research process of this Master's thesis, AI-based tools were used to assist in line with academic integrity standards. These tools served as supportive aids rather than substitutes for independent academic work.

ChatGPT-4 (OpenAI) was used for developing and debugging Python code for geospatial data processing and metric computation, rephrasing and clarifying text to improve academic tone and coherence, structuring explanations of complex methodological concepts. Additionally, SciSpace was used to explore relevant academic literature, helping to identify key sources, summarize theoretical frameworks, and navigate domain-specific terminology. All research design, analysis, interpretation, and writing decisions were independently conducted by the author. AI tools were used strictly to enhance productivity, clarity, and code development and are acknowledged here in the interest of full transparency.

Table of contents

1. Introduction	5
2. Abbreviations	7
3. Overview of DGGs	8
3.1. Introduction to DGGs and Its Importance in Geospatial Science	8
3.2. Advantages of DGGs	10
3.3. Indexing System in DGGs	11
4. Overview of MAUP and Attempts to Mitigate It	13
4.1. Introduction to the Modifiable Areal Unit Problem (MAUP)	13
4.2. Attempts to Measure and Mitigate MAUP	14
4.3. How can DGGs Mitigate MAUP?	16
4.4. Theoretical Conclusion: Experiment Design as the Key	17
5. Data	20
5.1. Data Description	20
5.2. DGGs creation	21
6. Methodology	22
6.1. Data Ingestion to DGGs	22
6.2. Data Cube Creation	24
6.3. Landscape Metrics Calculation	25
6.3.1. Proportions of Land Use Classes	25
6.3.2. Patch Density (PD)	25
6.3.3. Percentage of Like Adjacencies (PLADJ)	25
6.3.4. Shannon Diversity Index (SHDI)	26
6.4. Land Use Classification and Landscape Type Definitions	26
6.5. Sensitivity Analysis	27
7. Results	28
8. Discussion	39
9. Conclusion	41
Bibliography	42
License	45
Appendix 1. Code Repository	46

1. Introduction

The increasing demand for accurate and scalable geospatial data representation has led to the development of advanced spatial frameworks that enhance data integration, analysis, and visualization. Discrete Global Grid Systems (DGGS) have emerged as a transformative approach, offering a structured partitioning of the Earth's surface into equal-area, hierarchically indexed cells. Unlike traditional coordinate reference systems, which rely on latitude and longitude, DGGS provide a uniform spatial framework that eliminates projection distortions and enhances multi-resolution data analysis (Kmoch et al., 2022; Li & Stefanakis, 2020). These characteristics make DGGS particularly valuable in fields such as environmental monitoring, remote sensing, and spatial decision-making (ISO, 2021).

Despite their advantages, the use of DGGS in geospatial applications presents several challenges, primarily related to the uncertainty introduced during data aggregation and transformation. A key issue is the influence of spatial resolution and cell alignment on data representation, which can lead to information loss and variability in analytical outcomes. This phenomenon closely relates to the Modifiable Areal Unit Problem (MAUP), a well-documented issue in spatial analysis where different zoning and aggregation schemes produce varying results (Openshaw, 1984). The effects of MAUP persist even in DGGS-based models (Raposo et al., 2019), particularly in relation to resolution changes, grid shifts, and cell congruency. These uncertainties impact the accuracy of spatial classifications and influence decision-making processes that rely on DGGS-derived data.

This thesis explores how such uncertainty appears in DGGS-based spatial analysis, especially when the resolution changes. Initially, the goal was to measure how much land use data is modified when transformed across resolutions, using spatial metrics such as mode and landscape metrics (e.g. connectivity). Landscape metrics are quantitative tools used to evaluate the spatial structure and arrangement of land cover within a defined area. They offer insights into both the composition and spatial organization of different landscape elements. (McGarigal, 2015) However, during the research process, it became clear that the uncertainty introduced by DGGS aggregation might not be measurable in a universal or standardized way. Instead, the behavior of spatial metrics seems to depend heavily on the structure and complexity of the underlying data. (Traat et al., 1997; Parring et al., 1997)

As a result, the aim of this thesis is not to propose a single method for measuring uncertainty, but to develop a flexible framework for studying how uncertainty manifests across different landscapes and resolutions. The study focuses on characterizing the variability of metric behavior, assessing whether patterns in uncertainty are consistent across real-world data, and understanding how different landscapes react to DGGS-based transformations. In doing so, the research investigates how MAUP effects can be better handled, not by attempting to eliminate them entirely, but by offering practical guidance on how to design experiments that account for resolution, data properties, and metric sensitivity.

The thesis also demonstrates how the DGGS indexing system can support efficient and scalable geospatial computation. By structuring spatial data into tabular data cubes, it becomes faster and easier to calculate landscape metrics across multiple resolutions. This approach is not only practical but also necessary for analyzing large datasets and comparing metric behavior in a controlled, reproducible way.

The research is guided by the following hypotheses:

- The spatial uncertainty introduced by aggregating data to DGGS cannot be universally measured. However, its effects can be empirically characterized using metric sensitivity and resolution-dependent behavior.
- Landscape metrics exhibit resolution-dependent variation that reflects the presence of MAUP. However, these changes are not consistent across different landscapes and must be analyzed in relation to the data's structure and complexity.

The thesis is structured as follows: the introduction presents the background, motivation, and main objectives of the study. The next chapters provide an overview of DGGS, the Modifiable Areal Unit Problem, and their interaction with uncertainty in spatial analysis. The methodology chapter describes the process for evaluating metric behavior across DGGS resolutions using real-world data. The results and discussion chapters highlight key findings on how uncertainty varies across landscapes and scales. The conclusion summarizes the research outcomes and offers suggestions for future work.

By addressing the variability of metric behavior and the context-dependent nature of uncertainty, this study helps improve the way DGGS is used in geospatial research and provides tools for making spatial analyses more reliable.

2. Abbreviations

CLS	<i>Characteristic Length Scale</i> — The diameter of a circle with an area equal to that of a grid cell; used to compare resolution levels.
DGGS	<i>Discrete Global Grid System</i> — A spatial framework that partitions the Earth's surface into equal-area cells arranged hierarchically, often using hexagonal tessellation.
EPSG	<i>European Petroleum Survey Group</i> — A standard coding system for spatial coordinate reference systems.
ETAK	<i>Eesti Topograafia Andmekogu</i> — Estonia's official topographic database used for land cover and geographic information.
GPKG	<i>GeoPackage</i> — A platform-independent, compact format for storing geospatial vector and raster data.
IGEO7	A specific hierarchical aperture-7 hexagonal DGGS implementation developed with Z7 indexing.
ISEA	<i>Icosahedral Snyder Equal Area</i> — A projection method used in some DGGS implementations for uniform area distribution.
MAUP	<i>Modifiable Areal Unit Problem</i> — A source of statistical bias that arises when spatial results vary depending on the size, shape, or arrangement of aggregation units.
PD	<i>Patch Density</i> — A landscape metric indicating the number of discrete patches of a land cover class per unit area.
PLADJ	<i>Percentage of Like Adjacencies</i> — A metric measuring spatial cohesion by calculating how often neighboring cells share the same class.
SHDI	<i>Shannon Diversity Index</i> — A landscape diversity metric that accounts for both the richness and evenness of land cover classes.
Z7	A hierarchical integer DGGS cell addressing scheme encoding the hierarchical structure using base-7 digits.

3. Overview of DGGS

3.1. Introduction to DGGS and Its Importance in Geospatial Science

A Discrete Global Grid System (DGGS) is a modern framework for dividing the Earth's surface into separate, non-overlapping areas known as cells. These cells are usually shaped like hexagons, triangles, or squares, and they cover the entire surface of the planet without gaps or overlaps. The cells are organized in a hierarchical structure, allowing them to be progressively subdivided into smaller cells for higher resolution analysis. (Sahr et. al, 2003; Mahdavi-Amiri et al., 2015; Kmoch et al., 2022a; Li et. al, 2022; Thompson et al., 2022).

DGGS differs fundamentally from traditional geographic systems that rely on continuous latitude-longitude grids. Traditional systems often suffer from distortions, particularly near the poles, and produce cells of varying sizes depending on their location on the globe. This lack of uniformity complicates spatial analysis and reduces the accuracy of results. DGGS, however, uses discrete cells that are more uniform in size and shape, which simplifies data storage, retrieval, and analysis. (Mahdavi-Amiri et al., 2015; Thompson et al., 2022).

Graticules vs DGGS: A Comparison of Spatial Unit Uniformity

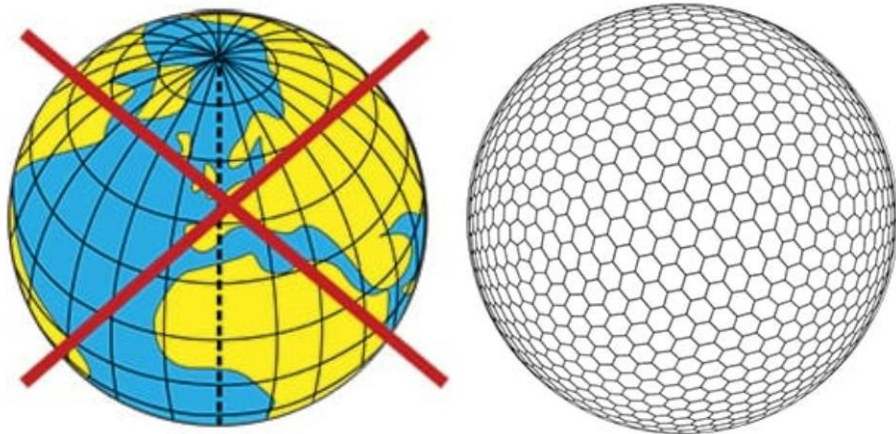


Figure 3. Traditional graticule-based coordinate systems create uneven spatial units, with severe distortion near the poles due to convergence of meridians. In contrast, DGGS partition the Earth into uniform, equal-area cells, such as hexagons, enabling consistent spatial analysis across the globe without projection-related biases. Elements of this illustration are adapted from Pelman (2016).

The idea of DGGS has been developing for many years to create a better way of mapping the Earth without distortions. Scientists were looking for new methods beyond traditional coordinate systems, which led them to use global grids for analyzing spatial data. The Geodesic Elevation Model (GEM) and the Quaternary Triangular Mesh (QTM) were two foundational models that contributed significantly to the evolution of DGGS. GEM utilized geodesic triangular quadtrees to encode elevation data hierarchically, while QTM introduced recursive triangular subdivisions of an octahedron for encoding vector geodata. These early models laid the groundwork for modern DGGS frameworks by addressing spatial hierarchy, precision, and computational efficiency. (Dutton, 2015)

The primary purpose of DGGs is to provide a consistent and scalable framework for organizing, analyzing, and integrating large amounts of geographical data. By employing a consistent structure, DGGs facilitates the integration of data from various sources, supports multi-scale analysis, and enhances the processing efficiency of large datasets. Furthermore, DGGs enables the comparison of data from different studies even when conducted at varying scales or levels of detail. This approach enables efficient handling of spatial data at different levels of detail (Sahr et al., 2003; Mahdavi-Amiri et al., 2015; Kmoch et al., 2022a; Li et al., 2022). DGGs applications extend across several key domains: landscape data analysis, remote sensing, environmental modelling, urban design, disaster prediction, health (ISO, 2021; Mahdavi-Amiri et al., 2015).

Among the different tessellation strategies, hexagonal DGGs have gained particular interest due to their ability to provide equal-area partitioning while maintaining uniform adjacency relationships between cells. Hexagonal grids allow for more accurate spatial interpolation and neighborhood analysis compared to square or triangular grids, which tend to introduce inconsistencies in their adjacency relationships (Sahr et al., 2003). In tessellation-based grid systems, aperture plays a key role in how a parent cell divides into smaller cells at a finer resolution level. For instance, shapes like squares or triangles can each be evenly divided into four smaller units, corresponding to an aperture value of 4. Hexagonal grids, however, can support several aperture values, most notably 3, 4, and 7. Among these, only aperture 7 maintains a near-uniform geometry between the parent and child hexagons, preserving the consistency and integrity of the hexagonal shape across levels. (Kmoch et al., 2022b) The use of hierarchical aperture-7 hexagonal DGGs, such as IGEO7, further enhances the efficiency of spatial indexing by preserving symmetrical tessellations across multiple resolutions, leading to better spatial analysis and data aggregation capabilities (Kmoch et al., 2025).

In the hexagonal ISEA-based DGGs implementation IGEO7, resolution levels define the spatial granularity by subdividing the Earth's surface into equal-area hexagonal cells. Each successive resolution increases the number of cells by a factor of seven and reduces the characteristic length scale (CLS), i.e., the diameter of a circle with the same area as a cell. At resolution 9, cells have an average area of approximately 1.26 km² with a CLS of ~1.27 km, whereas at resolution 12, the cells shrink to just ~0.004 km² (4,000 m²) with a CLS of ~68 m. These consistent and hierarchical subdivisions allow for scalable spatial analysis across multiple resolutions, enabling the quantification of spatial patterns and their sensitivity to scale. A detailed summary of cell characteristics across different resolution levels is provided in Table 1.

Level	Cells	Area (km ²)	CLS (km)	Hexagons	Pentagons	CLS (m)
0	12	51006562.172	8199.500	0	12	8058757.5
1	72	7286651.739	3053.223	60	12	3045924.1
2	492	1040950.248	1151.643	480	12	1151251.1
3	3432	148707.178	435.153	3420	12	435132.1
4	4012	21243.883	164.466	24000	12	164464.4

5	68072	3034.840	62.162	168060	12	62161.7
6	76492	433.549	23.495	1176480	12	23494.9
7	35432	61.936	8.880	8235420	12	8880.2
8	7648012	8.848	3.356	57648000	12	3356.4
9	03536070	1.264	1.268	403536060	12	1268.6
10	2824752492	0.181	0.479	2824752480	12	479.5
11	19773267432	0.026	0.181	19773267420	12	181.2
12	138412872012	0.004	0.068	138412872000	12	68.0
13	968890104072	0.0006	0.026	968890104060	12	25.9
14	6782230728492	0.0001	0.010	6782230728480	12	9.8
15	47475615099432	0.00002	0.004	47475615099420	12	3.7
16	332329305696012	0.00000	0.001	332329305696000	12	1.4
17	2326305139872072	0.00000	0.0005	2326305139872060	12	0.5
18	16284135979104492	0.00000	0.0002	16284135979104480	12	0.2
19	113988951853731442	0.00000	0.0001	113988951853731430	12	0.1
20	797922662976120064	0.00000	0.0000	797922662976120052	12	0.0

Table 1. This table presents various metrics related to the number and approximate size of hexagonal cells at different levels of spatial refinement. For clarity, the figures have been rounded. The Characteristic Length Scale (CLS), defined as the diameter of a circle equal in area to a cell at each resolution, is included to facilitate comparison with more familiar raster data resolutions. At the highest resolutions, levels 18 through 20, the CLS values are approximately 20 cm, 7.6 cm, and 2.9 cm, respectively.

3.2. Advantages of DGGS

DGGS represents an innovative approach to geospatial data management, while traditional Geographic Information Systems (GIS) often suffer from computational inefficiencies, projection distortions, and challenges in data interoperability. DGGS address some of these limitations and this chapter explores the key advantages of DGGS. Li and Stefanakis (2020) reviewed and categorized the advantages of DGGS, identifying key benefits such as:

1. faster data retrieval – DGGS organizes spatial data into structured grids, enabling efficient location-based queries (Goodchild, 2018);
2. higher spatial accuracy – The discrete cell-based system reduces uncertainty by representing geographic areas as defined cells rather than imprecise coordinate points (Mahdavi-Amiri et al., 2015);
3. unified data model – DGGS integrates vector and raster data within a common framework, simplifying data management and processing (Peterson, 2016);

4. multi-resolution analysis – The hierarchical structure of DGGs allows for seamless scaling between different spatial resolutions while maintaining consistency (Goodchild & Yang, 1989);
5. consideration of Earth’s curvature – DGGs ensures uniform grid sizes globally, minimizing distortions and providing a fair and accurate representation of spatial information (Goodchild, 2019).

Despite DGGs’ potential, its transformations are not without challenges. One of them is the risk of information loss when the DGGs resolution does not align with the original data’s resolution (Kmoch et al., 2022; Li et. al, 2022), thus careful resolution selection during data ingestion is very important.

Another enduring challenge lies in the persistent variance of results, shaped by the nuances of aggregation strategies and the intricate dance of spatial configurations. This problem aligns closely with the well-documented Modifiable Areal Unit Problem (MAUP), a fundamental issue in spatial analysis where different aggregation schemes lead to variable results (Openshaw, 1984). While the uniformity of DGGs grids reduces the effects of MAUP, DGGs are not entirely immune to scale and zoning biases. This uncertainty is influenced by three key factors: grid shifts, resolution changes, and grid congruency.

The lack of standardized metrics to quantify and address these uncertainties presents a significant research gap, limiting the reliability of analyses performed using DGGs (Sahr et al., 2003).

3.3. Indexing System in DGGs

In a Discrete Global Grid System, indexing serves as a fundamental mechanism for assigning unique identifiers to individual cells that partition the Earth’s surface. This process is essential for efficient data storage, retrieval, and analysis within DGGs frameworks, where the Earth is divided into potentially millions or even billions of discrete regions. The primary purpose of indexing is to provide a systematic way to locate, reference, and organize spatial data, enabling seamless integration of various datasets into a unified structure (Mahdavi-Amiri et al., 2015; Kmoch et al., 2022b, Samet, 1995).

The need for indexing arises from the complexity of managing vast amounts of geospatial data, where each cell within the DGGs represents a specific area on the Earth’s surface. By assigning a unique index to each cell, the system allows for rapid identification and retrieval of data associated with any given location, without relying on computationally intensive spatial operations (Mahdavi-Amiri et al., 2015; Sahr et al., 2003; Kmoch et al., 2022b). Furthermore, indexing facilitates the efficient construction of data cubes, where data from various resolutions and sources can be aligned and compared within a consistent framework. As noted, DGGs are increasingly being considered as a globally applicable reference system and spatial data model for hierarchical data cubes, allowing for the efficient organization of data within a structured, multiscale system (Kmoch et al., 2022b).

Indexing plays a crucial role in supporting hierarchical data structures, enabling seamless transitions between different levels of detail. This capability is particularly valuable

when conducting multiscale analysis, where information must be aggregated or disaggregated across various resolutions (Mahdavi-Amiri et al., 2015; Kmoch et al., 2022b, Samet, 1995).

The structured nature of indexing also simplifies complex queries, such as identifying neighboring cells or performing spatial joins, thereby enhancing the analytical capabilities of the DGGs. Several approaches to indexing have been developed to meet these requirements, including hierarchy-based indexing, space-filling curve-based indexing, and coordinate-based indexing. While hierarchy-based indexing is well-suited for organizing data within a multiresolution hierarchy, space-filling curves provide efficient memory storage by preserving spatial locality, and coordinate-based methods offer straightforward referencing of cells based on their geometric positions. (Mahdavi-Amiri et al., 2015, Samet, 1995). Regardless of the specific approach, indexing ultimately allows DGGs to transform geospatial data into structured, accessible, and scalable formats suitable for a wide range of analytical applications.

The aperture of a tessellation system describes how a parent cell is subdivided into smaller child cells at a finer resolution. For hexagonal cells, common apertures include 3, 4, and 7, with aperture 7 being the most effective because it divides each parent hexagon into 7 smaller hexagons, preserving an almost congruent parent-child relationship. Apertures 3 and 4 are less commonly used because they disrupt the hexagonal structure. (Kmoch et al., 2022b).

Hexagonal DGGs with aperture 7 support a unique and unambiguous hierarchical indexing structure based on Central Place Indexing (CPI). In this system, each hexagonal cell at a given resolution is subdivided into seven smaller cells at the next finer resolution, forming a compact and geometrically consistent structure. Spatial positioning within the grid is supported by a two-dimensional axial coordinate system using integer (i, j) pairs aligned along two of the three 120°-spaced axes of the hexagonal lattice. These coordinates enable efficient location referencing and neighbor finding and correspond directly to the CPI digits, which can be interpreted as directional vectors. (Sahr, 2019)

In the Z7-based IGEO7 system, each hexagonal cell is uniquely identified by a 64-bit integer index that encodes both its spatial location and hierarchical level (resolution). The first 4 bits indicate the base cell (one of 12 root cells covering the globe), while the remaining 60 bits are divided into 20 groups of 3 bits, each representing a step in the subdivision process from coarser to finer resolution. These steps are encoded as digits from 0 to 6, corresponding to the seven possible child cells in aperture 7 geometry. To determine the resolution of a given cell, the system scans these 20 sets of 3 bits from left to right and counts how many have a value between 0 and 6. Once a digit with the value 7 is encountered, it marks the end of the actual resolution path, as 7 is reserved to pad the index when fewer than 20 levels are used. (Kmoch et al., 2025)

Neighboring relationships are not encoded explicitly but can be computed algorithmically based on the structure of the index. Because the digits represent consistent directional steps in a regular hexagonal hierarchy, the relative position of a cell can be interpreted geometrically. (Kmoch et al., 2025)

4. Overview of MAUP and Attempts to Mitigate It

4.1. Introduction to the Modifiable Areal Unit Problem (MAUP)

The MAUP affects spatial analysis across disciplines like politics, economics, epidemiology, introducing statistical distortions due to scale and zoning effects. While its impact is well-documented, and researchers have developed some strategies to mitigate its influence, it still remains a challenge, since it is not fully avoidable (Tveter et al., 2022). The example with the map of Estonia (Figure 4) illustrates how changing the level of spatial aggregation can alter the results and lead to different conclusions, demonstrating the impact of MAUP on spatial data analysis. When data is aggregated at a smaller scale, the majority outcome within each area is recalculated providing a more nuanced and accurate representation of the data, but also demonstrating how the aggregation scale influences the final outcome.

The visualisation of how the scaling effect of MAUP impacts the final outcome

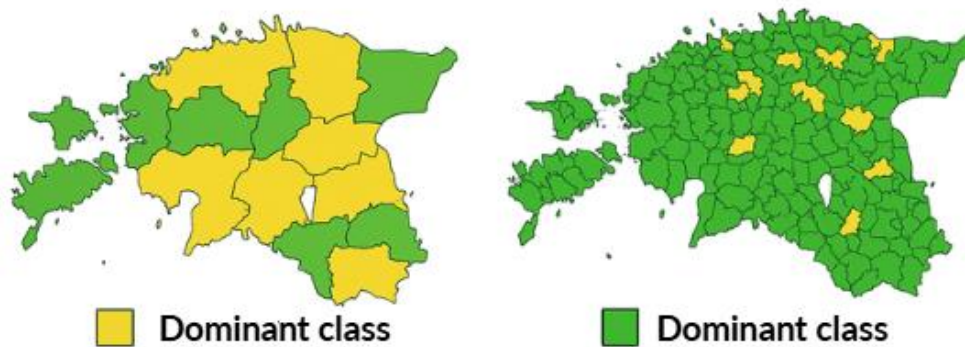


Figure 4 With the underlying data being identical in both cases, changes in aggregation and unit size lead to different dominant classes, demonstrating how spatial resolution influences the interpretation of spatial patterns.

In addition to the scale effect, the zoning effect is also clearly illustrated in the Figure 5: even when the number of spatial units remains constant, altering the shape and configuration of the zones changes how patterns appear and how dominant values are distributed. Different zoning schemes, such as grid-based versus administrative boundaries, can yield contrasting spatial patterns from the same underlying dataset. This emphasizes how subjective decisions in defining spatial units can introduce significant variation in analysis outcomes, potentially leading to misleading interpretations.

The visualisation of how the zoning effect of MAUP impacts the final outcome

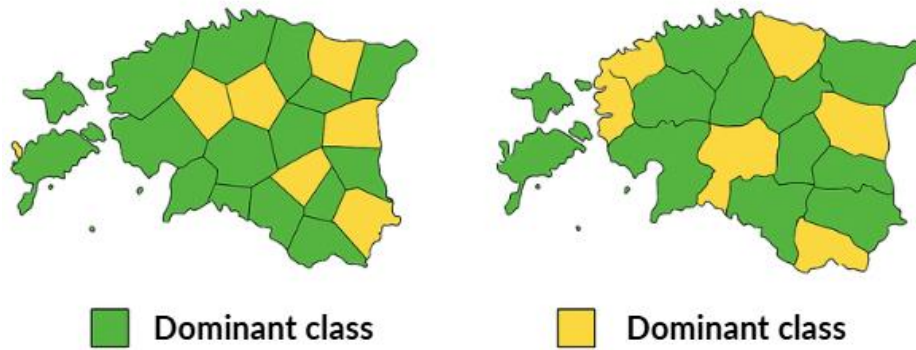


Figure 5. Despite using the same underlying data, different zoning schemes result in varying analytical outcomes, illustrating how the Modifiable Areal Unit Problem (MAUP) can influence results through changes in the shape and configuration of spatial units.

4.2. Attempts to Measure and Mitigate MAUP

Researchers of different disciplines have been trying to study the severity of the effect caused by MAUP, quantify it and try to mitigate it. This chapter synthesizes key approaches from various case studies, highlighting universal principles for reducing MAUP-related biases and identifying research gaps in the proposed methodologies.

The first and foremost principle is to **disaggregate the data and perform analysis at the finest resolution possible** (Wang & Di, 2020; Tveter et al., 2022). Wang and Di (2020) study focused on the MAUP in relation to the COVID-19 outbreak and environmental factors, specifically atmospheric NO₂ levels. They examined how the aggregation of environmental data and COVID-19 mortality data into different spatial units (cities, districts, and provinces) influences the statistical relationships between NO₂ concentrations and COVID-19 mortality. The goal was to illustrate the presence of MAUP and how it can bias findings in studies that analyze the impact of environmental factors on the severity of the COVID-19 outbreak. The authors recommend conducting epidemiological studies at the individual level wherever possible to avoid the aggregation bias of MAUP. When individual-level analysis is not feasible, researchers should attempt to conduct studies at the finest spatial scale possible and acknowledge the limitations if aggregation is necessary.

Another study supporting Wang & Di findings is Tveter et al. study aimed to understand how spatial aggregation errors affect the calculation of agglomeration benefits (economic benefits that come from businesses and workers being close together) in transport projects. Specifically, the study wanted to see how different methods of grouping data (like averaging travel costs) impact the results of calculating these benefits and to suggest the best way to avoid these errors in transport appraisals. The study concludes that MAUP-induced errors are unavoidable, despite efforts to reduce the error through better aggregation methods, the changes in the results caused by aggregation persist. The paper also recommends calculating

agglomeration benefits at the most disaggregated level possible to minimize aggregation errors. Tveter, Laird, and Aalen (2022) advise **employing weighted average method** to minimize errors caused by aggregation. However, they admit that even the weighted average method does not fully eliminate the spatial aggregation error.

Another promising approach was proposed by researchers Mirt, Reiche, Verbesselt, and Herold (2022). Their study presents a new **downsampling method** aimed at addressing the MAUP in remote sensing. The focus of the research was to create a downsampling technique that minimizes accuracy loss when reducing the resolution of images, particularly for spatial analysis. Traditional downsampling methods, such as mean, central pixel selection, and random sampling, often fail **to maintain the distribution of data** when aggregated, leading to errors in the analysis. The authors aimed to develop a method that better preserves the local and global properties of the data, thereby minimizing MAUP effects. While the novel method resulted in better accuracy, the authors note that it comes at a higher computational cost.

The study by Lee et al. (2018) also emphasizes the importance of preserving the original data distribution during aggregation. Their research explores the relationship between spatial autocorrelation (SA) and the MAUP, focusing on how varying levels of spatial autocorrelation affect the MAUP's impact on statistical measures such as means, variances, etc. Through a simulation experiment, the study examines how the initial level of SA at the finest spatial scale influences the MAUP effects when data is aggregated into larger units. The findings reveal that MAUP does not significantly affect means, except when extreme positive SA is present. In simpler terms, areas with high spatial autocorrelation have neighboring units that tend to exhibit similar values, creating patterns of strong spatial dependency. When smaller spatial units are aggregated into larger ones, the arrangement of data inevitably changes, and the distribution characteristics of the new units may differ from those of the original smaller units. This shift in data distribution during aggregation leads to altered statistical outcomes, including changes in means, variances, and spatial autocorrelation measures. Consequently, the results obtained from aggregated data may differ from those derived from the original, unaggregated data.

Briz-Redón (2022) introduces in his study a novel **Bayesian shared-effects modeling framework** to quantify the MAUP in spatial data analysis. This model helps researchers understand how changes in the way data is grouped affect the relationship between different factors (like traffic accidents or COVID-19 deaths) and the variables being studied. The model helps measure the global and local effects of these changes, making it easier to spot areas where the results might be misleading due to the aggregation of data at different scales. A Bayesian framework is a statistical approach that uses probability distributions to update prior beliefs or assumptions based on observed data. It employs Bayesian inference, a process in which prior information is refined through the incorporation of new data, leading to more accurate and updated predictions or parameter estimates. While the authors claim that the Bayesian framework improves the reliability of spatial analysis and helps identify problematic spatial units that may distort the covariate-response relationship, this conclusion appears somewhat arbitrary. The lack of detected MAUP effects in a specific dataset could stem from limited variation in the covariates across different scales or from an already stable spatial structure in the data, if it already has clear, strong, or uniform patterns that do not vary much when grouped

into different spatial units. Therefore, the absence of MAUP in one study does not necessarily imply that it is not a concern in all datasets.

One more paper that also confirms a so-called “**context-dependent aggregation strategy**” as MAUP mitigation method is a study by Comber & Harris (2022). The goal of the research was to understand how the scale of data aggregation affects the evaluation of ecosystem services. The study aimed to show how changing the scale in land use planning can influence decisions and the effectiveness of ecosystem service delivery. This paper confirms that the MAUP is unpredictable and non-linear, meaning that the effects of spatial aggregation can vary in complex ways depending on the data and the scale used. The authors argue that there is no one-size-fits-all ideal scale or spatial unit for analysis; instead, the appropriate scale should be chosen based on the specific purpose of the research. In this study, the best aggregation scale for land use decisions was determined by the spatial properties of the case study data. The highest ecosystem service scores were achieved when the ecosystem service gradient was aggregated to a 100 m scale, suggesting that this resolution was optimal for capturing the agricultural processes relevant to the study area. The paper emphasizes that selecting the right scale is crucial for making accurate and meaningful land use decisions.

In a nutshell these strategies emphasize that **no single ideal scale or spatial unit exists** for all analyses; instead, the scale should be chosen based on the research purpose and the specific data being studied. Key recommendations include the need to disaggregate data to the finest possible resolution and tailor the aggregation scale to the spatial properties of the study area. Techniques such as weighted averages, preserving data distribution when comparing the data, and Bayesian modeling have shown promise in addressing MAUP, though challenges remain, particularly with larger datasets or higher computational costs. However, the main problem is that most traditional methods use random spatial units, making them vulnerable to distortions caused by scale and zoning. This issue shows why we need new spatial frameworks like DGGS, which offer a consistent structure worldwide and allow analysis at multiple resolutions. The next chapter explores how DGGS-based spatial analysis can address these persistent challenges.

4.3. How can DGGS Mitigate MAUP?

As mentioned earlier DGGS is by their means a great mitigation factor, providing uniform spatial units. But unfortunately DGGS might not prevent all the uncertainty sources.

Uncertainty related to resolution scaling stems from the effects of upsampling and downsampling in DGGS-based spatial aggregation. Studies on spatial resolution effects indicate that coarser resolutions can lead to information loss and generalization errors, whereas upsampling finer resolutions may introduce biased results. Li et al. (2022) demonstrated in their research that higher resolution grids captured more terrain details, but computational complexity increased. Therefore, the choice of resolution should be based on problem-specific needs, balancing accuracy, computational efficiency, and the intended analytical objectives.

Many grid system designs introduce area distortions, resulting in spatial units that are not equal in size across the study region. These distortions influence the calculated values of spatial metrics by giving disproportionate influence to larger or smaller cells. In this context,

hexagonal grids offer a geometrically favorable alternative. Their compact, uniform structure reduces anisotropy, minimizes edge effects, and in properly designed DGGS systems, ensures more equal-area partitioning than rectangular or triangular alternatives. (Sahr et al., 2003)

Uncertainty related to the congruency or shape of the cells of DGGS grids arises because grid cells do not align well with the orientation, curvature, or scale of real-world features, thus the aggregation process can fragment coherent spatial patterns or merge distinct areas into a single unit, thereby distorting the resulting summary statistics.

Unfortunately, there are no existing research studies that specifically quantify how grid shifts impact data aggregation results. Grid alignment refers to the specific placement and orientation of a spatial grid over a continuous surface. In DGGS and other gridded systems, it is possible to shift or rotate the grid relative to the underlying features. The MAUP can be interpreted through the lens of sampling theory as a design-induced source of uncertainty, if we treat each spatial unit as a “sample”, where each zoning scheme or grid configuration acts like a different sampling design. (Traat et al., 1997) Just as different sample designs lead to different estimates in classical statistics, different spatial partitions produce different spatial statistics, even from the same underlying data. Therefore it is reasonable to assume that if the content within a grid cell changes due to shifts, its aggregated value may also vary. This is because alignment determines whether important features are preserved within single cells or fragmented across multiple ones. Such fragmentation affects calculations of landscape metrics, thereby altering the analytical outcome. Therefore, there is likely no universally correct or incorrect positioning of grid cells. Given the number of possible grid alignments, a practical approach could be to use mathematical models that simulate multiple grid positions to estimate the average effects of grid shifts on aggregation uncertainty.

4.4. Theoretical Conclusion: Experiment Design as the Key

From a statistical point of view, MAUP is not a problem but rather a logical outcome where the estimated parameter depends on the experimental design (Traat et al., 1997). Mathematical statistics provides methods aimed at making the most reliable conclusions based on the available data. Reliability, in this context, means making optimal use of the available data, minimizing the likelihood of making incorrect decisions, and being able to quantify the potential risk of incorrect decision (Arnab, 2017; Durrett, 2019; Parrington et al., 1997). The decision-making process is objective - anyone working with the same dataset and selecting the same acceptable error risk will reach identical conclusions (Chaudhuri & Stenger, 2005; Parrington et al., 1997).

Measuring the uncertainty caused by MAUP presents significant challenges when working with extensive geospatial datasets. Ideally, uncertainty could be measured by analyzing the entire population (entire group of items or elements of interest that we want to draw conclusions about) (Traat et al., 1997; Parrington et al., 1997; Hankin et al., 2019; Fuller, 2009), which in this case would be the whole Earth's landscape data. However, such an approach is not feasible with the currently available computational resources. A classical statistical approach is to take a sample from the population and determine the desired accuracy

and confidence interval. In mathematical statistics, samples are used because collecting data from the entire population is often either impossible or prohibitively expensive. Therefore, statistics based on samples are used to determine the parameters of the general population (Parring et al., 1997; Hankin et al., 2019).

The success of a sampling study depends on a properly designed experiment where the sample must be representative, sufficiently large, and selected using a method that ensures objectivity and reliability of the data (Traat et al., 1997; Parring et al., 1997; Fuller, 2009). Each object of the general population must have an equal probability of being included in the sample, or the effect of varying inclusion probabilities must be statistically accounted for (Parring et al., 1997). If a weighted selection is used, the sample estimates must be adjusted accordingly (Traat et al., 1997). This means that the sample to measure MAUP uncertainty should be proportional to the population to ensure the reliability of the results. However, the vast volume of geodata presents substantial difficulties in obtaining a sample that is truly representative of the entire population. It is essential to consider the number of experiments and their statistical power analysis when drawing conclusions. If the sample is too small or unrepresentative, the obtained statistics may be biased and may not reflect the actual characteristics of the general population (Chaudhuri & Stenger, 2005; Hankin et al., 2019; Parring et al., 1997; Fuller, 2009).

For example, in the study by Shan Ye, an attempt was made to quantify the uncertainty caused by MAUP by measuring sensitivity to different grid sizes. Four different equal-sized grids (50 km, 100 km, 200 km, and 400 km) were used, and fossil occurrences of marine bivalves and brachiopods were analyzed within three different areas of interest (Ye, 2024). The study shows that MAUP can significantly influence how species' spatial distribution is measured, which proved highly sensitive to grid-cell size. However, it remains unclear how representative the sample is of the broader fossil record. The study does not report whether the sample size was adequate to support broader generalizations. Therefore, while the approach offers valuable insight into MAUP-related effects, stronger conclusions would require a more detailed experimental design, including defined population parameters, sample size justification, and replication across more spatial contexts.

Another interesting research of Uemaa (2004), explored metric sensitivity through simplified, synthetic landscapes and suggested that observed trends may be extrapolated to real-world data, this thesis challenges that assumption. Synthetic landscapes, by design, lack the structural complexity, noise, and spatial heterogeneity present in real environments. As a result, any similarities in metric behavior across resolutions between idealized and real-world landscapes should be considered coincidental rather than causal. The apparent predictability in synthetic contexts does not translate into generalizable patterns, and assuming such transferability introduces a form of methodological overconfidence. This thesis therefore avoids the assumption of universal trends and instead develops a framework to analyze metric behavior empirically and contextually within real-world data domains.

Given these challenges, a legitimate question arises: “**Should and could MAUP even be measured?**” The same question is raised by Andresen (2021), who provocatively asks whether the MAUP is, in fact, a problem at all. In some cases, where for instance detailed landscape classification is the goal, using the most accurate data at the finest resolution is essential. However, in broader analyses, such as city-level studies for example, coarser levels

of aggregation may suffice, and focusing on the research purpose and experimental design might be more important than attempting to eliminate MAUP altogether. Ultimately, the focus should be on selecting the most appropriate spatial scale and aggregation strategy based on the specific research objectives, acknowledging that MAUP's impact can vary depending on the context and scale of the analysis.

5. Data

5.1. Data Description

The landscape data used in this study is sourced from the Estonian Topographic Database (ETAK), which manages vector-based topographic spatial data linked to attribute information. ETAK serves as the primary dataset for Estonia’s base map at a scale of 1:10.000 and supports various geospatial applications. (Maa-amet, 2016a)

ETAK's reality model consists of three hierarchical levels: group, feature class, and feature type . Each feature class is assigned a unique code (e.g., 303 – *arable land*), where the first digit of the code represents the broader group to which it belongs. In the data model, a real-world phenomenon is represented as an object. Objects are organized into tables based on their group classification and geometric primitives. Each object is characterized by three primary attributes that define its classification: ALAMLIIK (subtype), KOOD (code), and TYYP (type). In exceptional cases, the TYYP attribute may be missing, such as for marine areas. (Maa-amet, 2016b) In order to standardize and simplify land cover classification for spatial analysis, the original ETAK landscape classes were remapped into a custom numeric scheme (see Table 2). This remapping assigned a unique integer code to each generalized land use type, consolidating similar categories under shared labels.

Mapped code	Definition
1	Lake or pond
2	Watercourse
3	Road
4	Urban green, rocky or sandy areas and other (wasteland, small open land, clearcut under powerlines etc)
5	Built-up – urban areas, settlements, and developed land
6	Arable land
7	Grassland
8	Shrub
9	Forest
10	Wetland - includes fen, bog, marshy grassland, and peat extraction zones
11	Sea – marine areas within Estonia’s coastal zone

Table 2. Remapped land cover classification scheme used in the analysis.

To enable spatial analysis using DGGS, the vector-based ETAK data was transformed into raster format through a standard geospatial preprocessing workflow. First, the national ETAK dataset was clipped into uniformly sized landscape units that together cover the entire territory of Estonia (Figure 6). Each unit represents a fixed spatial extent and preserves complete land cover information for its area. Following the clipping process, rasterization was applied to each vector unit. During this step, the original polygon features were converted into a regular grid of cells, with each cell assigned the mapped code value of the landscape class it intersected.

Each raster tile is defined within the Estonian national coordinate reference system (EPSG:3301), also known as the Estonian Coordinate System of 1997. The spatial extent of a typical tile ranges from 650,000 to 675,000 meters east and from 6,550,000 to 6,575,000 meters

north, covering a 25 km × 25 km area on the ground. The spatial resolution is 1 meter, meaning each raster cell represents a 1 m × 1 m square of land surface. As a result, each tile consists of 25,000 columns and 25,000 rows, totaling 625 million cells. The data is stored in GeoTIFF format, where each raster cell contains an integer value representing the code of a specific landscape class, as defined earlier.

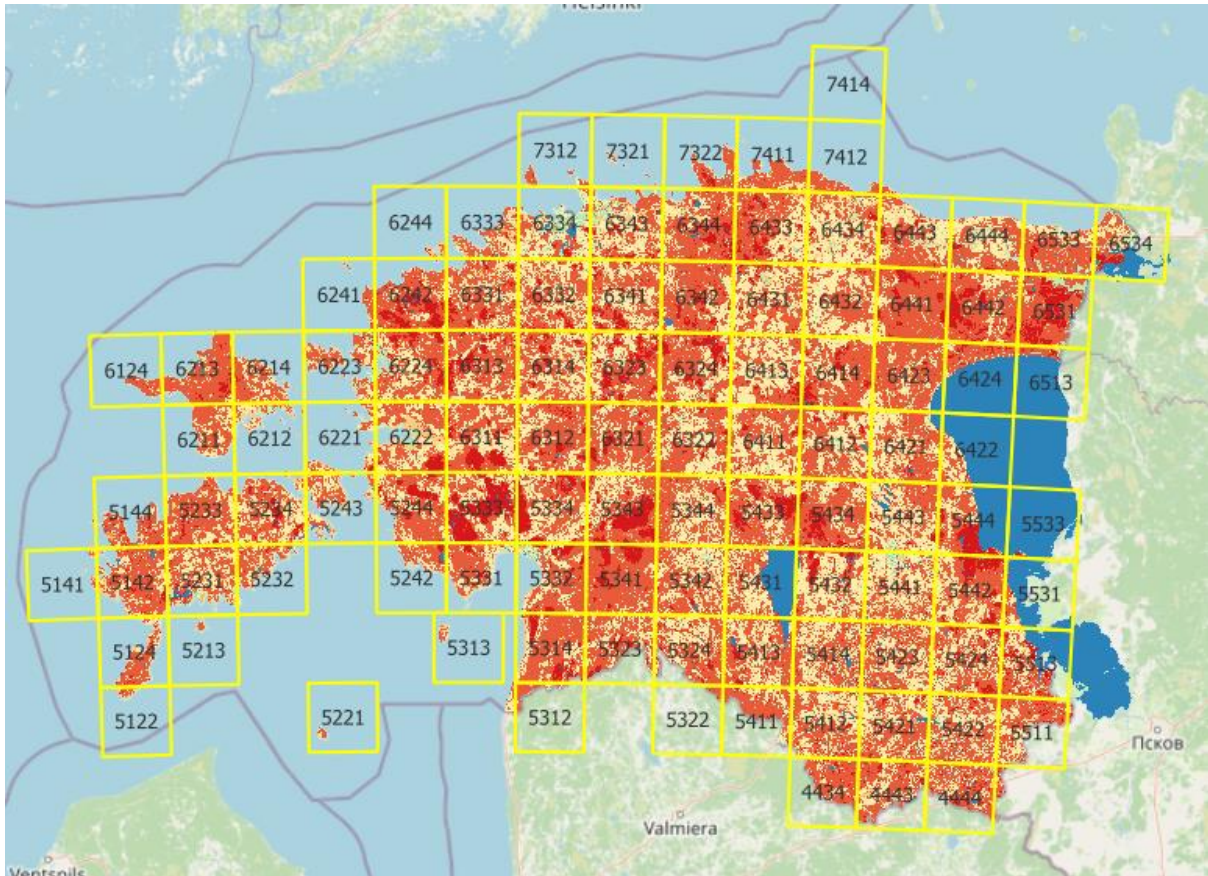


Figure 6. Rasterized land cover data grouped into tiles. Each yellow-outlined tile represents a standardized 25 km x 25 km landscape unit used for ingesting raster data into the DGGS.

5.2. DGGS creation

To facilitate spatial analysis using DGGS, a hierarchical grid was generated for each raster tile using the *dggrid4py* library - a Python interface to the DGGRID tool developed by Kevin Sahr (Kmoch, 2025). For each raster tile, a DGGS was generated by selecting an appropriate grid configuration and defining the desired resolution level to determine the size and number of grid cells. The grid was then created to encompass the extent of the raster tile, ensuring comprehensive coverage. The resulting grid cells were exported as vector geometries in the GeoPackage (.gpkg) format.

6. Methodology

6.1. Data Ingestion to DGGS

To ingest raster-based land cover data into the DGGS framework, a multi-step spatial matching process was applied. First, the raster map, containing high-resolution land cover classifications represented by land use code values, was partitioned into manageable spatial chunks. Each raster pixel was then converted into a point geometry, preserving both its spatial location and categorical land use code.

A hexagonal DGGS grid was generated over the same spatial extent at the target resolution level (i.e., level 14). Rather than performing a full zonal analysis, where the mode or average of all pixels within each hexagonal cell would be computed, a nearest-neighbor assignment method was used, not only to enhance computational efficiency, but also because the CLS of level 14 (ca 9.8m) is very close to the source raster resolution of 10m.. Specifically, for each DGGS hexagon, the land use class of the nearest raster pixel to the cell's centroid was identified and assigned as the representative value for that cell. This approach allowed each DGGS cell to be attributed with a single, categorical land cover class while minimizing the processing burden associated with more complex statistical summaries.

Raster to Grid: DGGS Data Preparation Workflow

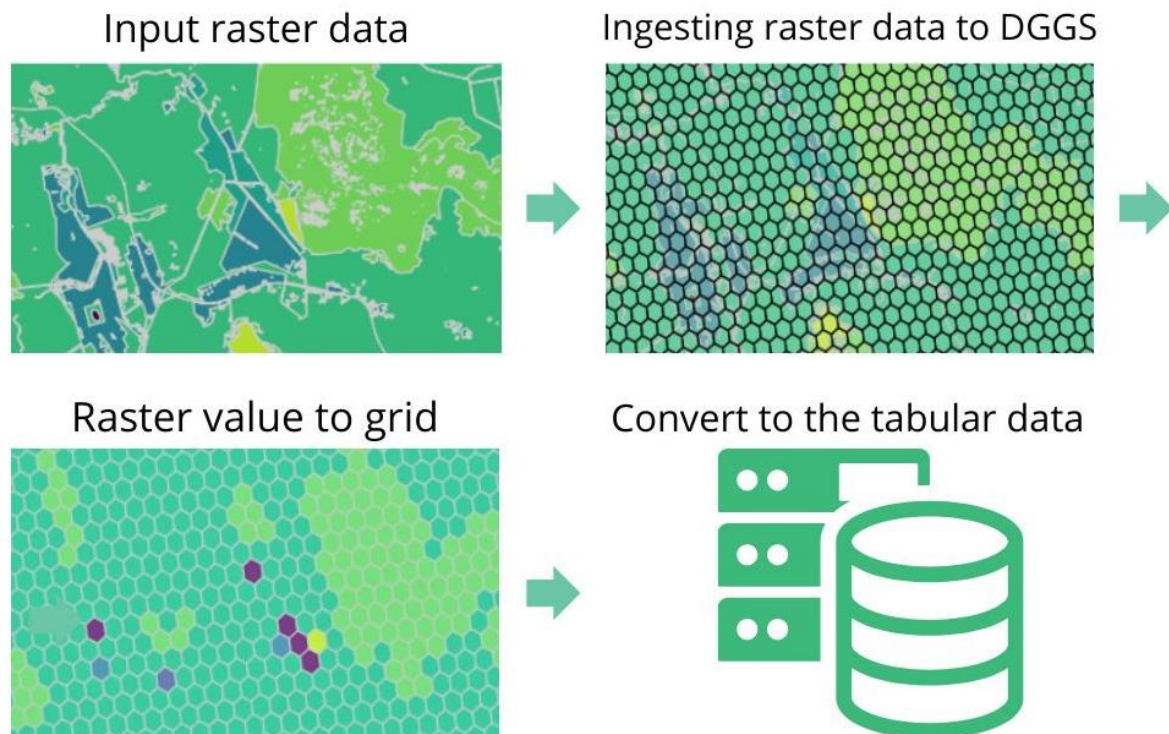


Figure 7. Overview of the data transformation process from raster land cover input to DGGS-based tabular format.

To compare how different aggregation strategies affect DGGS-based land cover classification, two methods were implemented for coarser resolution levels: flat aggregation and hierarchical aggregation. In the flat aggregation strategy, land cover values were assigned independently to

each DGGs cell at the target resolution by directly sampling from the original raster data, repeating the nearest-neighbor process for each resolution level. In contrast, the hierarchical aggregation strategy leveraged the multiresolution structure of DGGs by first ingesting raster data into a finer resolution grid and then aggregating values upward to coarser levels based on the pre-assigned child cell values. This approach preserved parent-child relationships between grid levels and allowed for a controlled assessment of how aggregation logic influences the resulting spatial patterns.

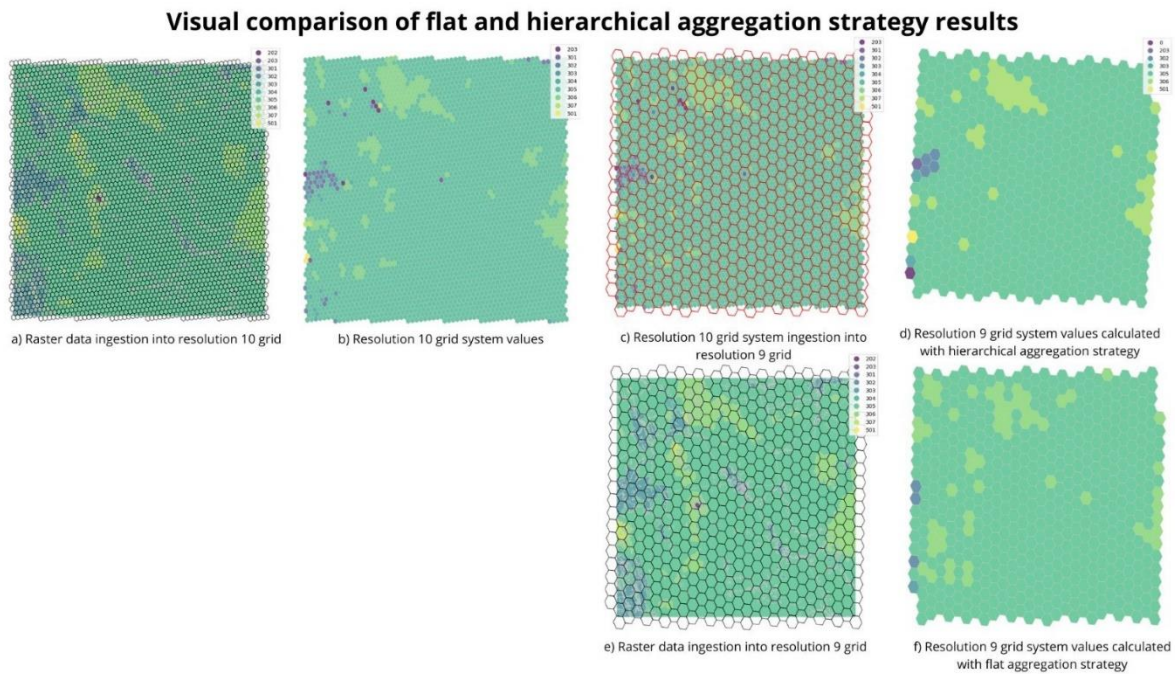


Figure 8. Visual comparison of flat and hierarchical aggregation strategies applied to DGGs grid values at resolution 9. Panels (a) and (e) show raster data ingested directly into DGGs grids at resolutions 10 and 9, respectively. Panels (b) and (f) display resulting land cover values calculated using flat aggregation, where each cell is independently assigned a value based on the underlying raster. In contrast, panels (c) and (d) illustrate the hierarchical aggregation approach, where resolution 9 cell values are derived from the already-ingested resolution 10 data (panel a). This method preserves the parent-child relationship within the DGGs hierarchy and demonstrates how using finer-resolution data as the aggregation base influences the outcome at coarser levels.

To illustrate how resolution choice affects the outcome of land cover data ingestion into the DGGs framework, Figure 9 presents a comparison between two resolutions.

Comparison of Land Cover Assignments at DGGs Resolutions 9 and 10

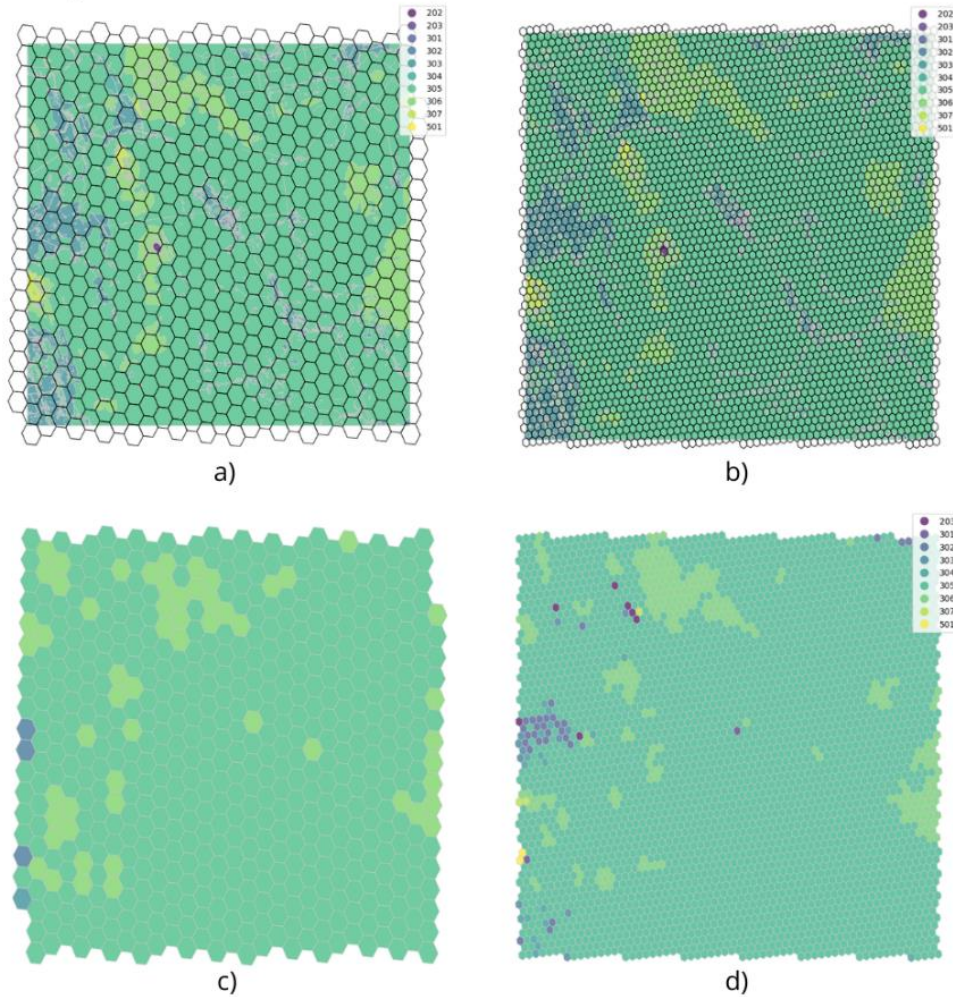


Figure 9. Comparison of DGGs grid values at resolution levels 9 and 10. Panels (a) and (b) show DGGs grids at resolutions 9 and 10, respectively, overlaid on the original raster land cover data. Panels (c) and (d) display the same DGGs grids without raster background, highlighting the differences in spatial patterns and classification outcomes across resolutions. The figure demonstrates how land cover values differ depending on the resolution of the grid system.

6.2. Data Cube Creation

To facilitate efficient computation and querying of spatial metrics across multiple resolutions, the ingested data was structured into a data cube format. A data cube, in this context, refers to a multidimensional table where each row represents a single DGGs cell and each column stores attributes such as resolution, cell index, dominant class, neighboring and parental relationships.

This tabular format was generated by assigning each DGGs cell a unique hierarchical index (based on its resolution and position), and by recording calculated or inherited properties in separate fields. Because the grid was generated separately for each landscape tile, the resulting data cubes are modular but consistent, and can be merged or queried collectively.

6.3. Landscape Metrics Calculation

To quantify spatial patterns at different spatial scales, a series of landscape metrics were calculated for each 25 km × 25 km raster tile at each resolution level of the DGGs. These metrics were computed after ingesting the raster-based land cover information into the grid structure, allowing each tile to serve as a unit of analysis for summarizing spatial composition and configuration of landscape classes. The following metrics were calculated to describe land cover proportions, fragmentation, cohesion, and diversity within each tile:

6.3.1. Proportions of Land Use Classes

The relative abundance of each landscape class was computed as the percentage of DGGs cells within the tile where a given class was the most dominant. This gives a coarse-grained yet scalable measure of land cover composition at each resolution.

$$\text{Proportion}_{c,r} = \frac{N_{c,r}}{N_r}$$

Equation 1. Formula to calculate proportions of land use classes inside a landscape tile.

where $N_{c,r}$ is the number of DGGs cells assigned class c in tile t at resolution r , and N_r is the total number of DGGs cells in that tile.

6.3.2. Patch Density (PD)

Patch Density describes the **degree of fragmentation** of a land cover class within a tile. A "patch" is defined as a contiguous group of DGGs cells that share the same majority class and are spatially connected (i.e., touching neighbors). This metric reflects how broken up or consolidated a class is spatially:

$$PD = \frac{n_{\text{patches}}}{A}$$

Equation 2. Formula to calculate patch density.

where n_{patches} is the number of such connected components, and A is the total area of that land cover class within the tile. Higher patch density implies greater fragmentation: many small, scattered patches, while lower values suggest spatial cohesion and dominance.

6.3.3. Percentage of Like Adjacencies (PLADJ)

PLADJ measures the internal spatial cohesion of a land cover class by evaluating how often adjacent cells share the same class. It is computed as the proportion of all neighbor pairs where both cells belong to the same class:

$$PLADJ = \frac{\sum_{i=1}^n \text{like_adj}_i}{\sum_{i=1}^n \text{total_adj}_i}$$

Equation 3. Formula to calculate percentage of like adjacencies.

where n is the number of DGGs cells assigned like_adj_i to the class, like_adj_i is the number of like adjacencies (neighbors of cell i that share the same class) and total_adj_i is the total number of neighbors of cell i . A high PLADJ indicates that a class tends to form spatially coherent clusters, while a low value suggests fragmentation.

6.3.4. Shannon Diversity Index (SHDI)

To measure overall landscape diversity within each tile, the Shannon Diversity Index was computed based on the proportional presence of each land cover class:

$$SHDI = - \sum_{i=1}^k p_i \cdot \ln(p_i)$$

Equation 4. Formula to calculate Shannon Diversity Index.

where p_i is the proportion of DGGs cells belonging to class i , and k is the number of unique classes in the tile. Higher SHDI values indicate that many classes are present in similar proportions (high diversity), while lower values suggest dominance by one or a few classes (low diversity).

6.4. Land Use Classification and Landscape Type Definitions

To gain insights into how different landscape configurations respond to spatial aggregation, this study identifies and categorizes representative landscape types based on the structural behavior of individual land use classes. Rather than classifying entire tiles as homogeneous entities, the classification focuses on the behavior of each land use class within each 25 km x 25 km tile at the finest resolution (level 14).

Landscape types are defined through a rule-based system, using combinations of spatial metrics: class proportion, patch density, percentage of like adjacencies, and Shannon Diversity Index. Thresholds for these metrics were determined empirically by iteratively testing combinations that match known or expected landscape patterns. For example:

- A **Dominant, Cohesive Landscape** is identified by a class proportion $\geq 75\%$, patch density ≤ 10 , and PLADJ ≥ 50 .

- A **Diverse, Fragmented, but Cohesive Landscape** is one where the class proportion is between 15% and 40%, $PLADJ \geq 50$ and the overall landscape diversity (SHDI) exceeds 1.0.
- A **Connected Minor Class** is one with class proportion $< 10\%$, patch density ≤ 5 , and $PLADJ \geq 70$.
- A **Scattered Minor Class** is defined by a class proportion $< 10\%$, patch density > 10 , and $PLADJ < 50$.

These thresholds do not represent universal definitions but rather are adapted to the observed range of metrics in Estonia's landscapes. They allow the identification of conceptually meaningful and structurally distinct land use behaviors. Once these classes are defined at resolution 14, the same class-tile combinations are traced across coarser resolutions (levels 13 to 8). If a class disappears due to aggregation (is no longer the dominant class in any DGGS cell), it is recorded with zero values for each metric. This enables a direct comparison of metric trajectories as resolution decreases, highlighting how structural simplifications or losses manifest in spatial data.

In addition to landscape-type-based analysis, a general sensitivity analysis was conducted across the entire Estonian dataset. This provides a baseline understanding of metric behavior independent of landscape structure. The purpose of this general analysis is to illustrate that MAUP is context-dependent, and the sensitivity of metrics can vary significantly depending on whether spatial structure is taken into account. This comparison reinforces the importance of structural awareness in experimental design.

6.5. Sensitivity Analysis

To evaluate the sensitivity of landscape metrics to resolution, Pearson correlation coefficients are calculated between each metric and the resolution. This approach is applied independently for flat and hierarchical aggregation strategies. Metrics with a strong positive or negative correlation are considered more sensitive to resolution changes, while near-zero correlation indicates stability across scales. Sensitivity is computed for each land use class in each tile across resolutions 14 to 8. Then, average correlation values are calculated across all tiles and across specific landscape types. This makes it possible to identify whether certain landscape configurations are more prone to metric distortion under aggregation.

7. Results

To illustrate the characteristics of different landscape types defined in this study, one representative tile was selected for visual analysis from each category. These examples serve to provide intuitive insight into how land use patterns manifest spatially in the real world. Each selected tile reflects a unique combination of landscape metrics at the finest resolution and exemplifies the structural logic behind its assigned landscape type. Further details for each landscape, including dominant land cover classes and key spatial metrics, are provided in the corresponding figure captions.

Dominant, Cohesive Landscape

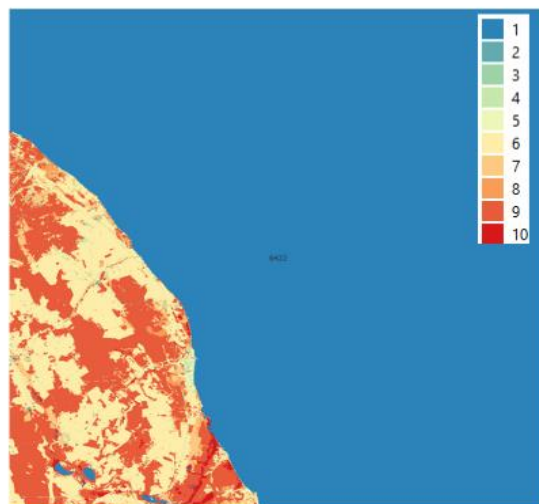


Figure 10. This tile represents a typical dominant and cohesive landscape. The land use class 1 covers nearly the entire tile, forming a single continuous patch. The visual homogeneity reflects a high PLADJ and extremely low patch density, indicating strong spatial cohesion and minimal fragmentation.

Scattered Minor Class

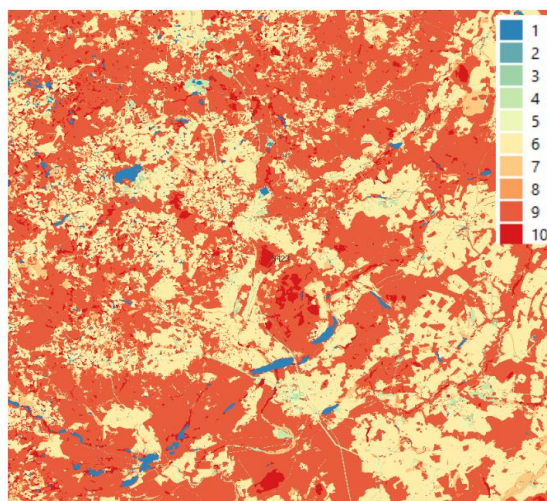


Figure 11. This tile exemplifies a scattered minor class behavior. The land use class 4 appears in small, isolated patches, making up less than 5% of the tile. These fragments are widely dispersed, yielding high patch density and low PLADJ, suggesting a high degree of fragmentation.

Connected Minor Class

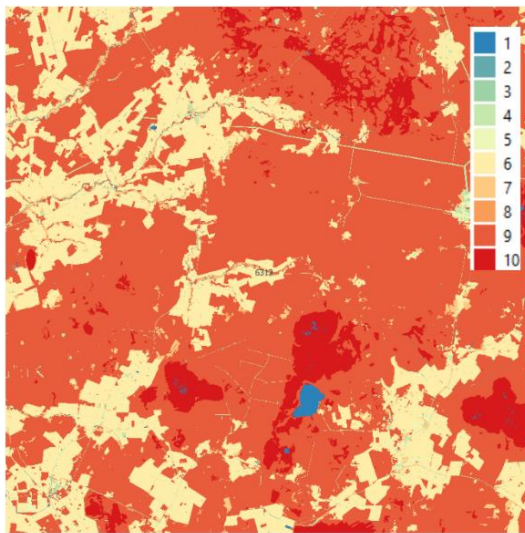


Figure 13. This tile presents the connected minor class pattern. Despite having a low proportion (<10%), class 10 appears in one or few well-connected regions, resulting in low patch density and high PLADJ. This structure allows the study of rare but spatially cohesive land use types and their metric behavior across resolution levels.

Fragmented but Cohesive Land Use Class in Diverse Landscape

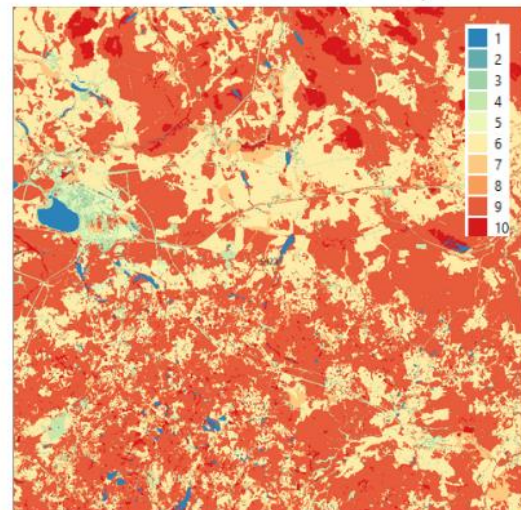


Figure 12. This tile shows a fragmented but cohesive landscape. Class 6 appears in numerous patches but tends to form loosely grouped clusters. The patch density is high, but PLADJ values are moderately elevated, indicating that fragments exhibit partial spatial connectivity.

To evaluate the influence of spatial aggregation strategies on landscape metrics and to assess the modifiable areal unit problem, we first examined the behavior of four key metrics across resolutions for all tiles covering Estonia.

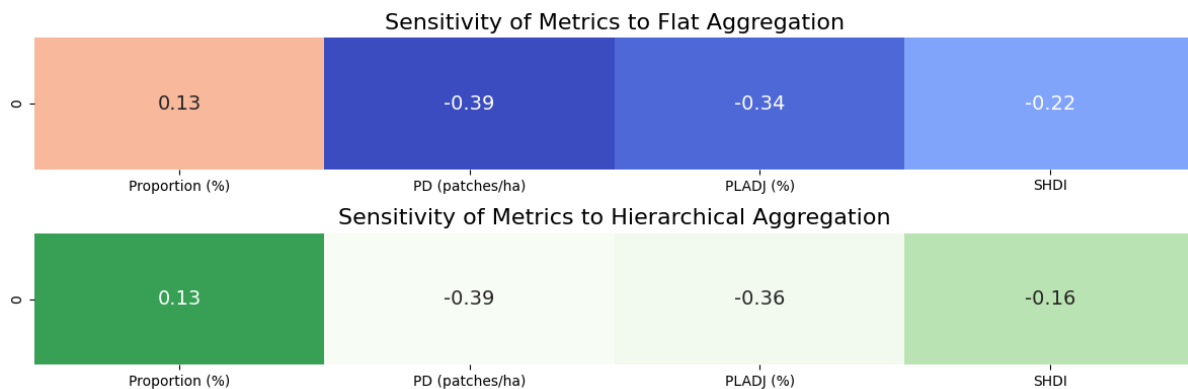


Figure 14. Sensitivity analysis results for landscape metrics across Estonia.

The analysis reveals that patch density and PLADJ are the most sensitive metrics to changes in resolution, with correlation coefficients of -0.39 and -0.34 (flat aggregation), and -0.39 and -0.36 (hierarchical aggregation), respectively (see Figure 14). These moderate negative correlations indicate a consistent decline in spatial detail and structural connectivity as resolution becomes coarser. In contrast, class proportion exhibits a weak positive correlation (0.13 in both strategies), suggesting that the dominant land cover types remain largely preserved through aggregation. SHDI, representing compositional diversity, also shows mild

sensitivity, with correlations of -0.22 (flat) and -0.16 (hierarchical), indicating a gradual loss of diversity with coarser resolution.

However, it is important to note that the range of values for each metric remains extremely broad, from 0 to 100% for proportion and PLADJ, from close to 0 to over 120 patches/ha for PD, and from 0 to nearly 2.0 for SHDI (see Figure 15).

Flat vs Hierarchical Aggregation Metric Trends (All of Estonia)

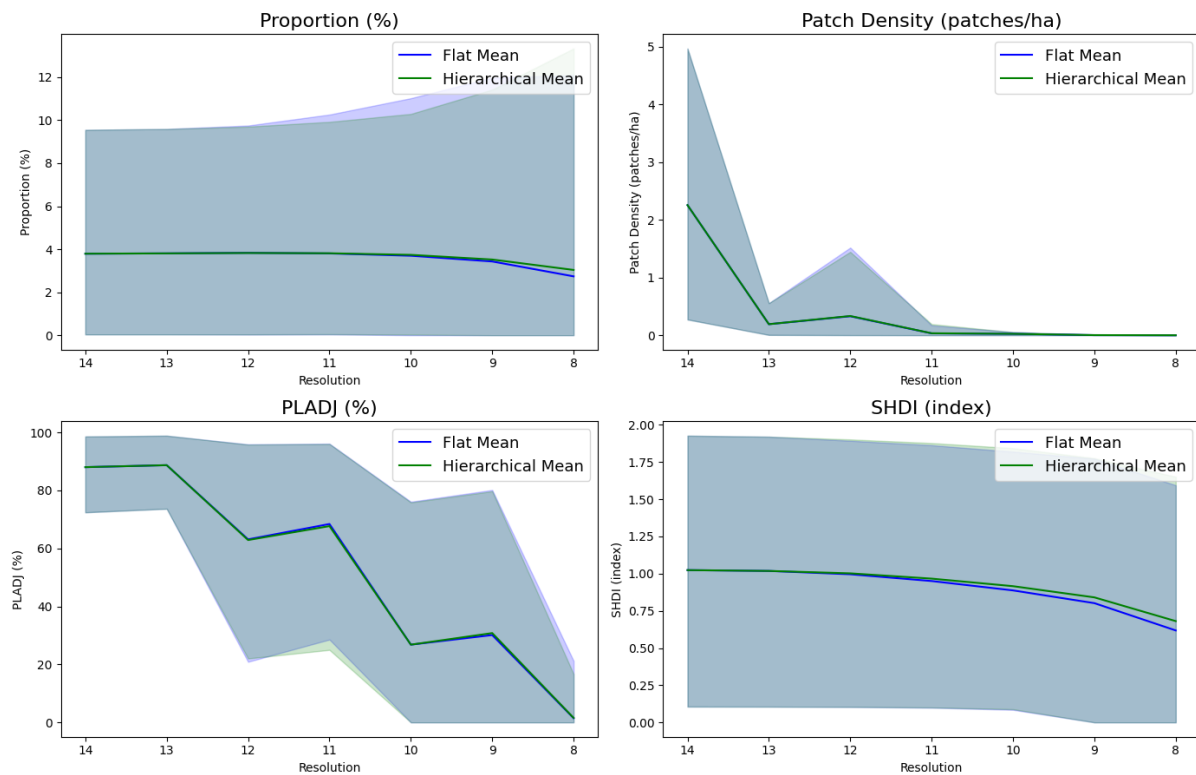


Figure 15. Mean landscape metric trends across Estonia under flat and hierarchical aggregation strategies, shown for DGGs resolutions 14 to 8. Metrics include class proportion, PD, PLADJ, and SHDI. Shaded areas represent mean value ranges across all DGGs cells.

While such variability is expected in a diverse national dataset, it significantly limits the interpretive value of aggregated trends. Results based on pooled data across all landscapes are not particularly informative when attempting to understand how metrics behave in relation to specific land cover types or spatial configurations. In essence, the averaging process masks the individual responses of different landscape types to aggregation, making it difficult to distinguish meaningful patterns from statistical noise.

To gain more insight into metric behavior, it is necessary to differentiate landscapes with distinct properties, such as dominant land cover type, spatial structure, or fragmentation level. Only by doing so can we observe how metrics respond to aggregation in a way that reflects meaningful landscape characteristics rather than general statistical noise.

Dominant Cohesive Land Use Class

This group of tiles represents landscapes where one land cover type clearly dominates the area (see), for example, large, uninterrupted forest blocks, continuous agricultural fields, or extensive wetlands. These areas are spatially uniform and exhibit little fragmentation, meaning that most of the tile is covered by a single, connected type of land use or land cover.

Flat vs Hierarchical Aggregation Metric Trends for Dominant and Cohesive Land Use Class

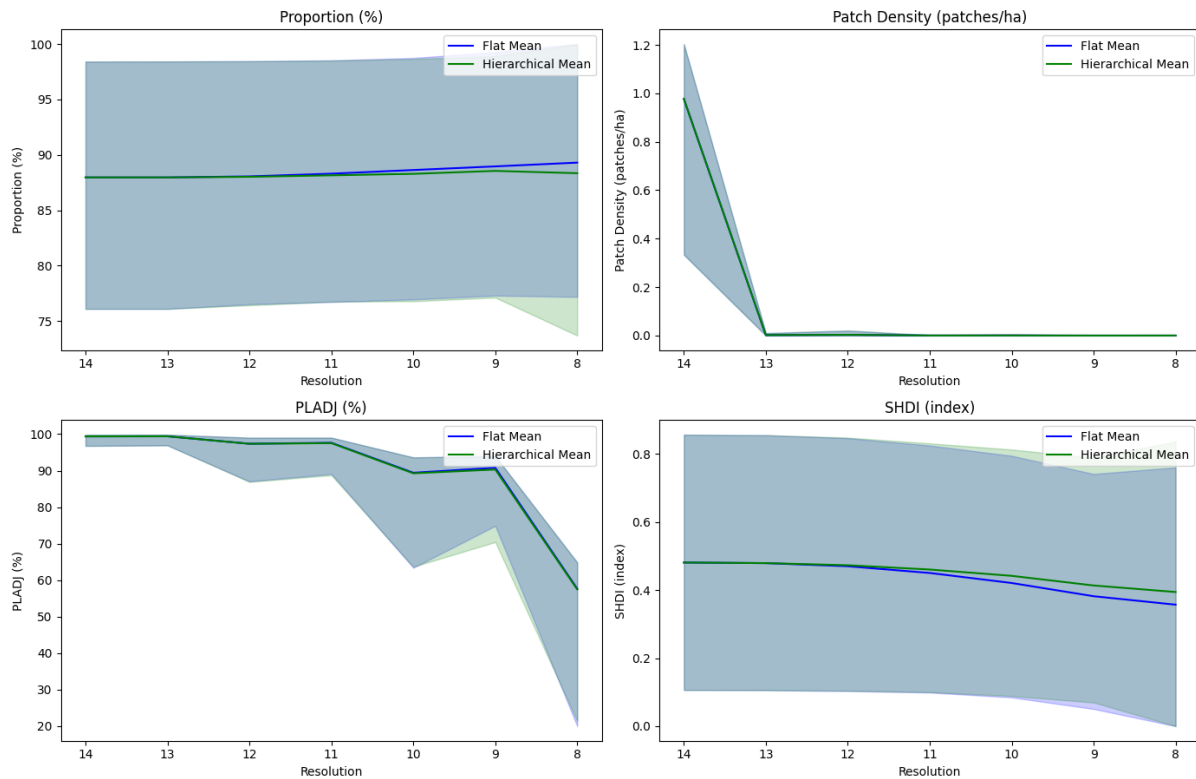


Figure 16. Mean landscape metric trends for dominant cohesive land use pattern under flat and hierarchical aggregation strategies, shown for DGGS resolutions 14 to 8. Metrics include class proportion, PD, PLADJ, and SHDI. Shaded areas represent value ranges across all DGGS cells.

The proportion of the dominant class remained consistently high across all resolution levels. The proportion metric in this case refers to the percentage of the tile covered by the dominant land cover class. Across all resolutions, from very fine to coarser scales, this percentage remained high (see Figure 16), meaning that the main land use stayed visually dominant even when the map was simplified. In real-world terms, this shows that large, uniform land blocks retain their identity regardless of how detailed or simplified the map is. Slight decreases at coarser resolutions likely result from small amounts of edge mixing with other land covers during aggregation. Both flat and hierarchical aggregation preserved high mean values, with hierarchical aggregation showing slightly broader variation between the minimum and maximum values at each resolution.

Patch density in cohesive landscapes is naturally low because the dominant cover type forms one or a few large, continuous areas. The results confirm this: PD was low at all resolutions and decreased further as the data were aggregated. This suggests that small breaks or internal boundaries within the dominant area tend to disappear when viewed at coarser

scales. In practice, this could reflect, for example, the merging of small forest clearings into a single block when zooming out on a map.

PLADJ in these landscapes started very high, close to 100%, and decreased with coarser resolutions. This drop means that although the dominant land use remains, the internal structure becomes less connected at larger scales, adjacency is lost as individual boundaries blur. This means that when zooming out or using a less detailed map, small breaks or edges between different parts of the forest are no longer visible, they get merged into one large area.

In dominant cohesive landscapes, diversity is naturally low because one class dominates. The SHDI results reflect this, with low values across all scales. A slight decrease in diversity was seen with coarser resolutions, meaning that even the small number of minor land cover types present (such as roads, clearings, or water bodies) may disappear entirely from the map when data are aggregated.

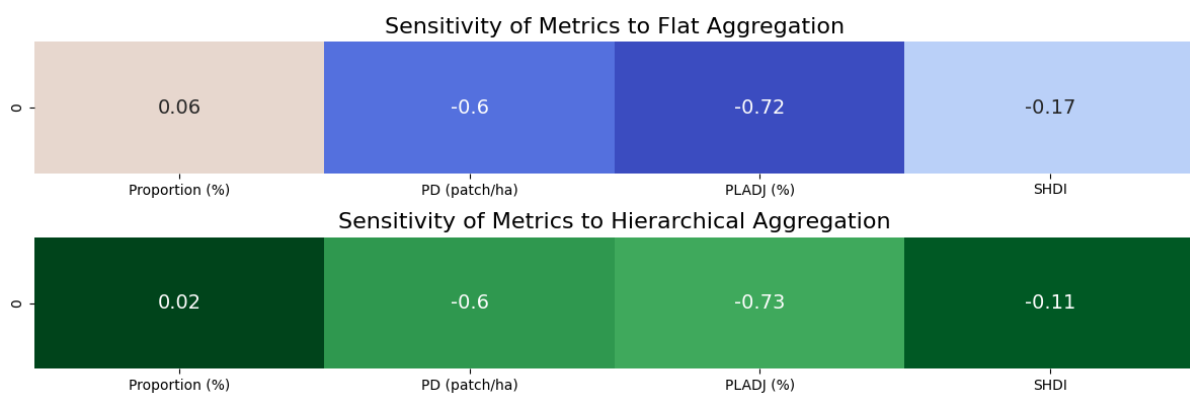


Figure 17. Sensitivity analysis results for landscape metrics for dominant and cohesive land use pattern..

Metrics like PLADJ and PD showed relatively high negative correlations ($r \approx -0.6$ and -0.7 respectively), meaning they consistently declined as the resolution became coarser. This confirms that internal structure and fragmentation details are sensitive to the resolution and get lost with aggregation, even though the overall dominance of a land cover type remains. Proportion and SHDI, in contrast, were largely unaffected, showing minimal correlation with scale (see Figure 17).

Fragmented but Cohesive Land Use Class in Diverse Landscape

This group of tiles represents landscapes where a land cover class occupies a moderate portion of the area and coexists with other land uses in a diverse, yet spatially organized way. These areas are not dominated by one class but still contain clearly recognizable patterns. Such configurations are commonly found in suburban areas, mixed-use rural landscapes, such as the edges of forests bordering open land or agricultural zones. The inclusion of a PLADJ threshold ensures that the land cover class is not only fragmented but also spatially coherent, meaning that similar land cover cells tend to be grouped together, forming visually and structurally connected clusters.

Flat vs Hierarchical Aggregation Metric Trends for Fragmented but Cohesive Land Use Class in Diverse Landscape

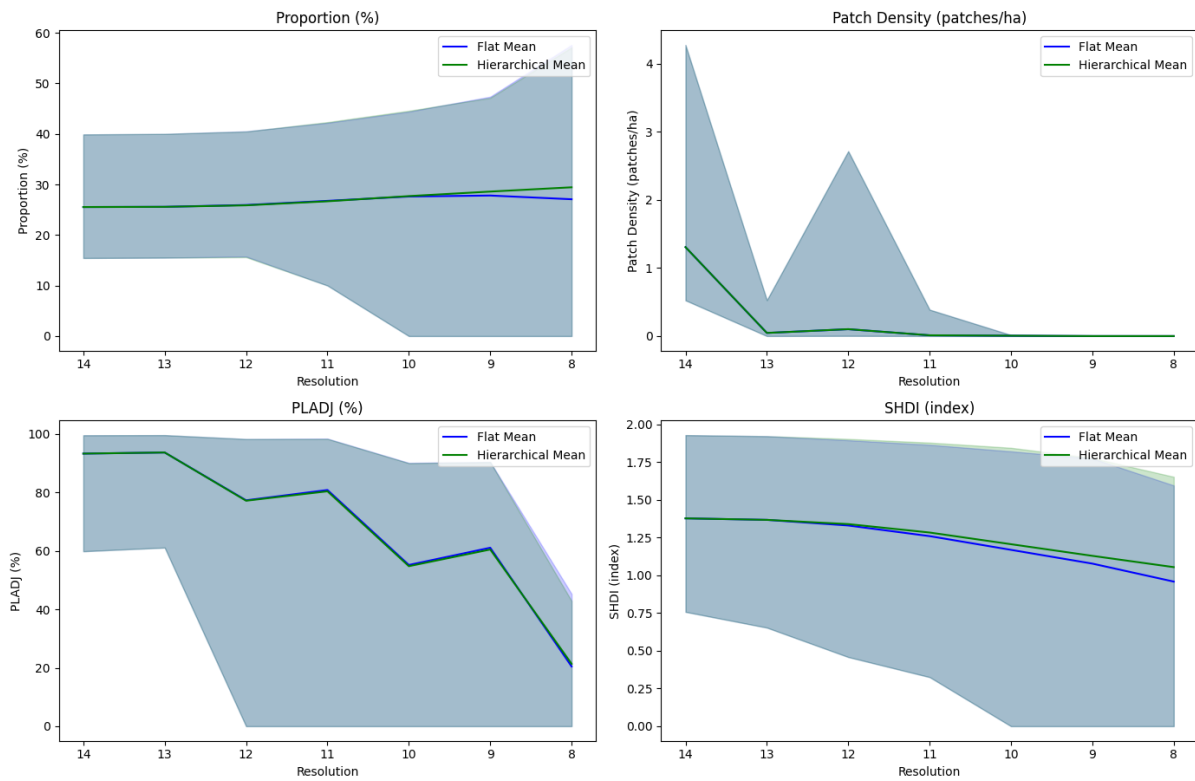


Figure 18. Mean landscape metric trends for fragmented, but cohesive land use pattern in diverse landscape under flat and hierarchical aggregation strategies, shown for DGGs resolutions 14 to 8. Metrics include class proportion, PD, PLADJ, and SHDI. Shaded areas represent value ranges across all DGGs cells.

The mean proportion of the selected land cover class remained relatively stable across resolutions, with a slight upward trend observed as resolution became coarser (see Figure 18). This trend was slightly more evident under flat aggregation, where aggregation tends to smooth and expand the footprint of the moderate class. Hierarchical aggregation yielded nearly identical trends but exhibited slightly lower variance. In real-world terms, this suggests that land covers in moderately fragmented landscapes can appear to take up more space than they actually do when the map becomes more generalized, especially with flat aggregation.

Patch density was initially moderate at fine resolutions, consistent with the fragmented nature of these landscapes. PD showed a sharp decrease between resolutions 14 and 13, after which it remained close to zero. Both aggregation strategies produced similar trends, reflecting the merging of many small patches into fewer large ones during the aggregation process. On the ground, this corresponds to the loss of detailed structure in mixed landscapes, such as smaller fields or clearings blending into larger land use blocks when the resolution is lowered.

PLADJ values were relatively high at fine resolutions, indicating that the selected land cover type, though fragmented, maintained internal cohesion through spatial clustering. As resolution decreased, PLADJ declined steadily under both aggregation strategies, with a slightly steeper drop under flat aggregation. This decline highlights how the internal grouping of similar land cover cells becomes less distinguishable at coarser resolutions. In practical

terms, this means that compact clusters of similar land uses, for instance, grouped residential areas or forest fragments, start to lose their identity and appear more diffuse or fragmented in generalized aggregated representations.

SHDI values were initially high, reflecting the balanced and heterogeneous composition of these landscapes. As resolution decreased, SHDI values also declined, indicating a simplification of the land cover mosaic. This reduction was more pronounced under flat aggregation, where minor land cover types were more likely to be absorbed into dominant neighbors during the merging process. This behavior illustrates how finer-scale features may become invisible at broader scales, especially in landscapes where visual complexity is ecologically or functionally significant.

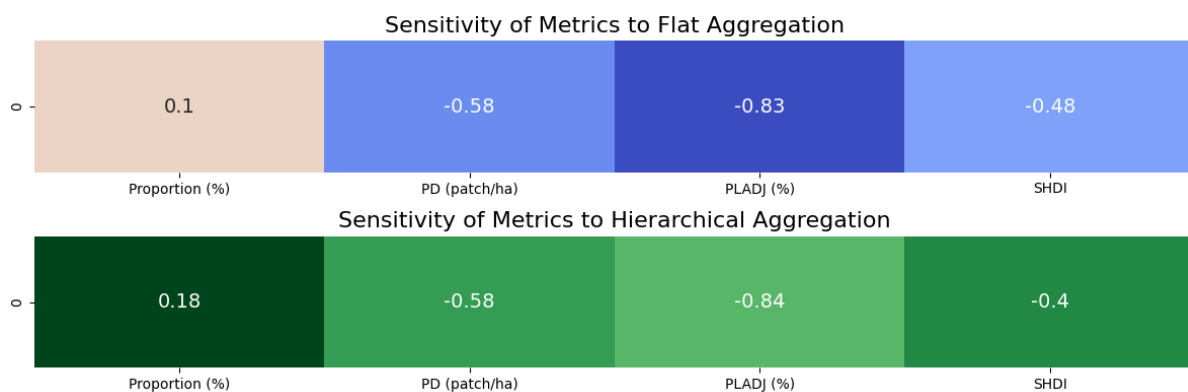


Figure 19. Sensitivity analysis results for landscape metrics for fragmented but cohesive land use pattern in diverse landscape.

The sensitivity analysis further confirms these trends (see Figure 19). The proportion metric exhibited weak correlation with resolution ($r = 0.10$ for flat, $r = 0.18$ for hierarchical aggregation), indicating minimal inflation at coarser scales. In contrast, patch density was highly sensitive to resolution changes ($r = -0.58$ in both cases), while PLADJ showed the strongest negative correlation ($r = -0.83$ for flat, $r = -0.84$ for hierarchical), reflecting the clear loss of spatial cohesion across scales. SHDI showed moderate sensitivity ($r = -0.48$ for flat, $r = -0.40$ for hierarchical), pointing to consistent simplification of landscape diversity during aggregation.

Scattered Minor Class

This group of tiles represents landscapes where a specific land cover class occupies a very small portion of the area, typically below 10%, and is spread thinly across the tile in a scattered and highly fragmented fashion. These land cover types do not form visible clusters and appear instead as isolated or sparsely distributed units. In real-world contexts, such patterns may represent small clearings, minor vegetation patches, or scattered built features embedded within dominant land uses such as forests or agriculture.

These tiles were identified based on low class proportion and a lack of spatial cohesion, with PLADJ values not exceeding 50% and patch densities often high. The resulting landscapes exhibit minimal structural organization, making them potentially especially vulnerable to data loss or distortion during spatial aggregation.

Flat vs Hierarchical Aggregation Metric Trends for Scattered Minor Classes

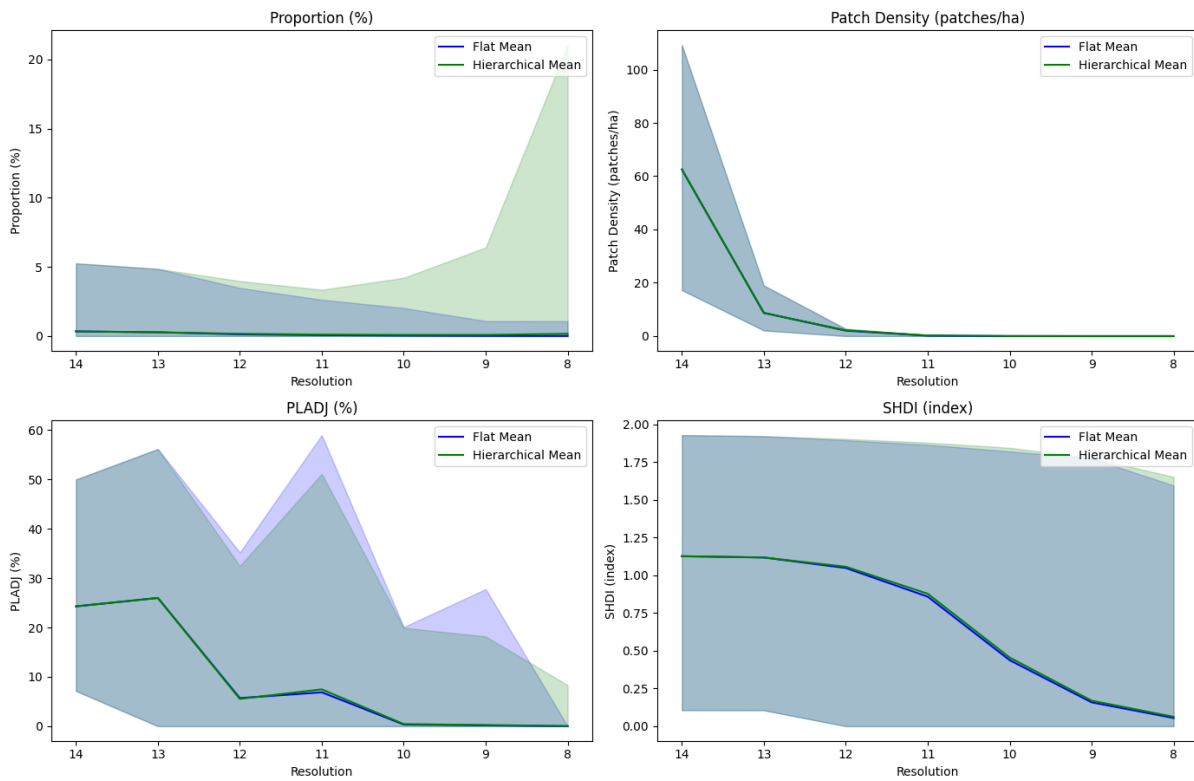


Figure 20. Mean landscape metric trends for scattered minor land use pattern under flat and hierarchical aggregation strategies, shown for DGGS resolutions 14 to 8. Metrics include class proportion, PD, PLADJ, and SHDI. Shaded areas represent value ranges across all DGGS cells.

The proportion metric remained low across all resolutions, as expected for a minor land use class (see Figure 20). Interestingly, a slight increase in proportion was observed at coarser scales, especially under hierarchical aggregation. This inflation is likely the result of how isolated small patches are merged into larger grid cells, giving the impression that the minor class occupies more area than it actually does. In real terms, this suggests that very small or narrow land use features may appear exaggerated in coarser-resolution datasets, depending on how aggregation is performed.

Patch density started at extremely high values at the finest resolution, with individual tiles showing up to 100 patches per hectare. This reflects the extreme fragmentation of the selected land cover type. As resolution decreased, patch density dropped sharply, especially between resolutions 14 and 13, and then declined to near-zero across both aggregation strategies. This dramatic loss indicates that the majority of small isolated patches are merged or eliminated during aggregation, resulting in a significant simplification of the landscape.

PLADJ values, which reflect how many adjacent cells belong to the same class, were generally low across all scales, consistent with a scattered spatial pattern. Both aggregation strategies showed noisy trends with inconsistent drops and slight rebounds, indicating instability in adjacency behavior. Flat aggregation exhibited slightly more variability than hierarchical. In practical terms, these fluctuations suggest that spatial structure is highly

unstable in scattered landscapes, and interpretations of patch connectedness can vary dramatically with resolution, offering little analytical reliability.

SHDI values were relatively high at fine resolutions, reflecting the presence of a diverse set of land cover classes. However, SHDI decreased with coarsening resolution under both aggregation strategies, as minor land covers were lost or absorbed into more dominant ones. This decline was nearly linear and similar between flat and hierarchical aggregation, suggesting a consistent reduction in visible diversity. In real-world applications, this means that heterogeneity in landscapes dominated by scattered features becomes invisible at coarser scales, potentially leading to misinterpretation in biodiversity, infrastructure, or land-use planning analyses.

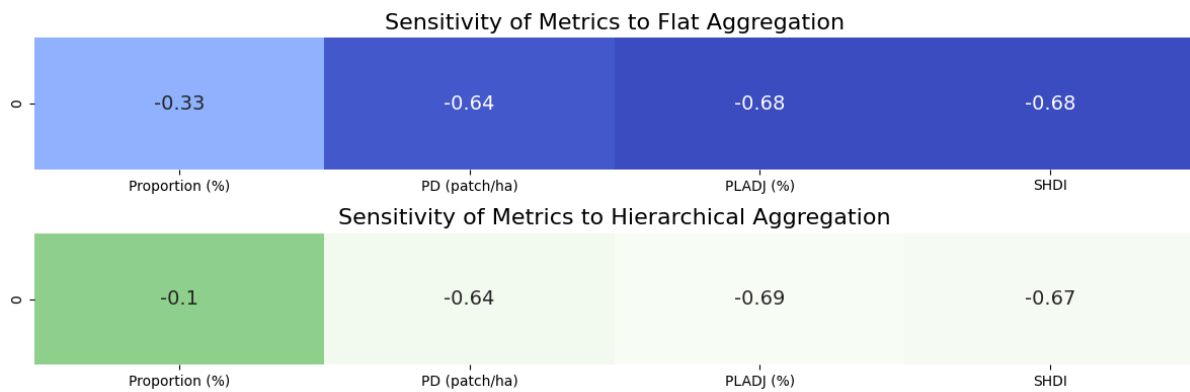


Figure 21. Sensitivity analysis results for landscape metrics for scattered minor class.

The sensitivity analysis confirms the high scale-dependence of these landscapes (see Figure 21). The proportion metric showed moderate correlation with resolution ($r = -0.33$ for flat aggregation, $r = -0.10$ for hierarchical), indicating mild inflation at coarser scales. Patch density and PLADJ exhibited strong negative correlations (PD $r = -0.64$; PLADJ $r = -0.68$ to -0.69), confirming that the metrics are highly unstable and strongly affected by aggregation. SHDI also showed substantial sensitivity ($r = -0.68$ for flat, $r = -0.67$ for hierarchical), reflecting the consistent loss of diversity as resolution decreases.

Connected Minor Class

This group of tiles represents landscapes where a minor land cover class occupies a small portion of the total area (less than 10%) but is spatially coherent, forming compact and internally connected clusters. Despite their limited extent, these features exhibit high internal adjacency (PLADJ > 70%) and low fragmentation (patch density < 5), distinguishing them from scattered or dispersed patterns. Such configurations are often seen in small, cohesive land use units, such as villages, forest groves, or isolated wetlands, which retain their structural integrity despite occupying a minor share of the landscape.

Flat vs Hierarchical Aggregation Metric Trends for Connected Minor Class

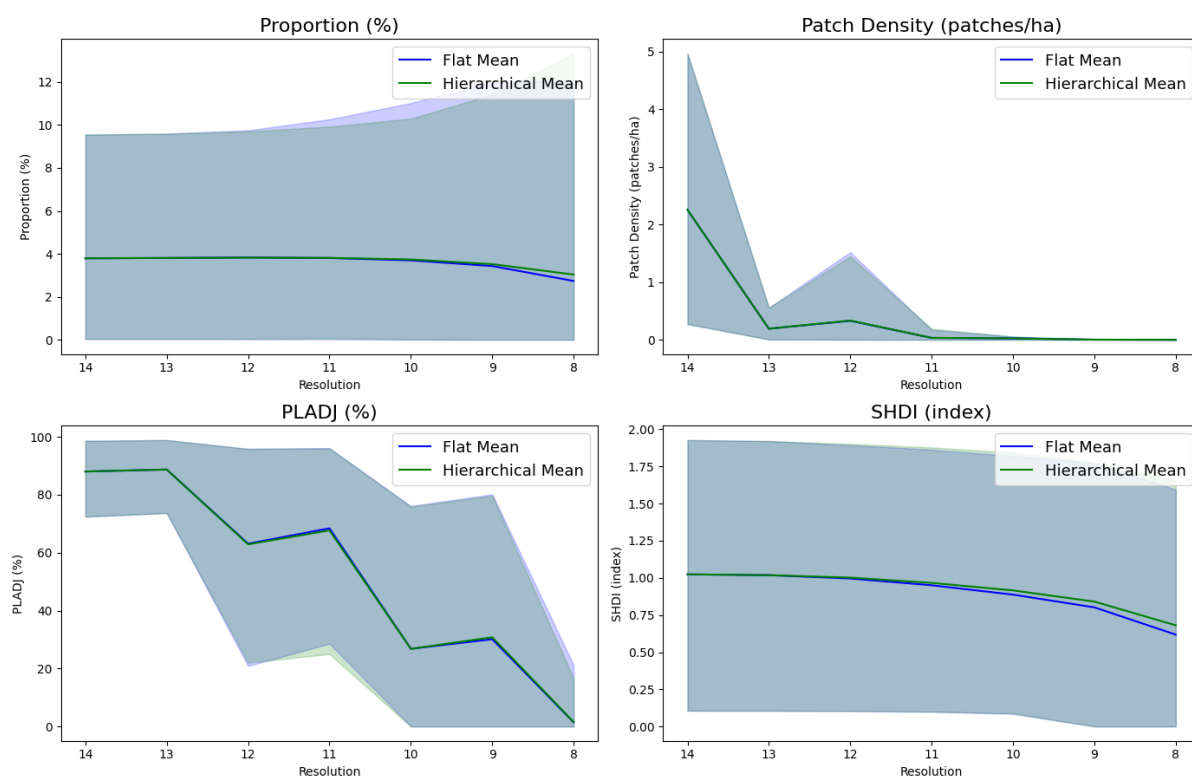


Figure 22. Mean landscape metric trends for connected minor land use pattern under flat and hierarchical aggregation strategies, shown for DGGS resolutions 14 to 8. Metrics include class proportion, PD, PLADJ, and SHDI. Shaded areas represent value ranges across all DGGS cells.

The proportion metric remained consistently low across resolutions, with both exhibiting disappearance at coarser levels to vice versa increasing in proportions. Values showed only a slight downward trend with increasing cell size, more so under flat aggregation (see Figure 22). This indicates that the spatial footprint of connected minor features is not significantly inflated or lost due to aggregation. However, proportions behavior varied across aggregation strategies - in some cases, the class proportion appeared to decrease or even disappear at coarser resolutions, while in others, it unexpectedly increased. This variability likely reflects how small, connected patches interact with grid boundaries during aggregation: they may either be absorbed into dominant surrounding classes or become overrepresented when merged with other nearby minor cells. These effects highlight the instability of minor-class area estimates when scale changes, even if the spatial structure remains internally cohesive.

Patch density values were low across all resolutions, consistent with the original filtering criteria. At finer resolutions, the metric reflects the presence of a few well-separated but internally cohesive patches. As resolution becomes coarser, patch density declined slightly under both aggregation strategies, with a sharper initial drop between resolutions 14 and 13. This reflects the merging of nearby patches or their absorption into neighboring classes during aggregation. In practice, connected features remain identifiable but may merge into larger units, particularly at lower resolutions where individual patch distinctions are lost.

PLADJ values were high at fine resolutions, but dropped considerably with decreased resolution, with similar patterns under both aggregation strategies. The decline was particularly sharp below resolution 11, indicating a loss of visible internal connectivity at coarser scales.

From a real-world perspective, this suggests that while the overall shape of connected features is retained, their internal cohesiveness becomes less apparent as resolution decreases.

SHDI values, which measure landscape diversity, started at moderate levels due to the presence of multiple land cover classes. These values showed a steady decrease with coarsening resolution, reflecting the gradual simplification of the landscape. The decline was slightly more gradual than in scattered classes, indicating that connected minor land cover types retain their distinction longer before being fully absorbed into dominant classes during aggregation.

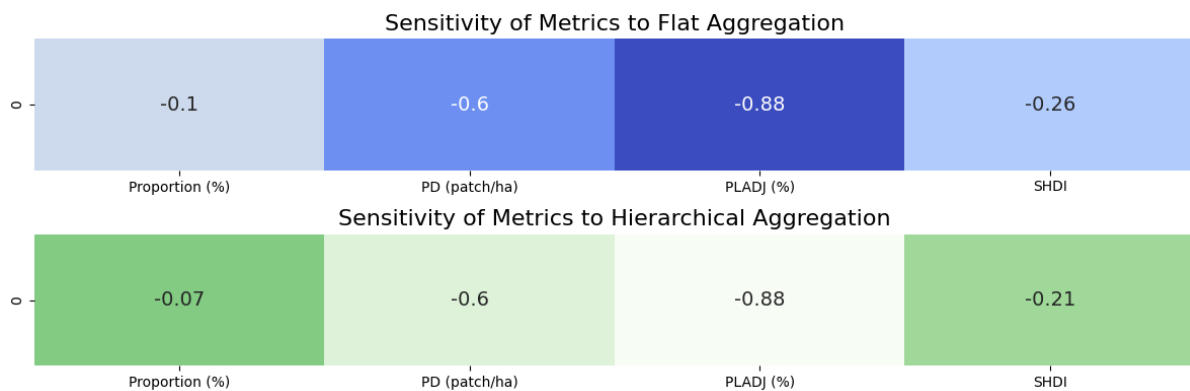


Figure 23. Sensitivity analysis results for landscape metrics for connected minor land use pattern.

The sensitivity analysis confirms the stability of the proportion metric, with low correlation coefficients ($r = -0.10$ for flat, $r = -0.07$ for hierarchical), suggesting that area coverage remains consistent across scales (see Figure 23). Patch density showed moderate sensitivity ($r = -0.60$), with a predictable drop due to patch merging. PLADJ exhibited the strongest correlation ($r = -0.88$ under both strategies), reflecting the rapid decline in visible connectivity as spatial resolution decreases. SHDI showed low to moderate sensitivity ($r = -0.26$ for flat, $r = -0.21$ for hierarchical), indicating a gradual, but not abrupt, loss of diversity.

8. Discussion

This study set out to explore how uncertainty manifests in DGGS-based spatial analyses, particularly in relation to the MAUP. The findings confirm that spatial aggregation introduces systematic, yet context-dependent, distortions to landscape metrics (see Figure 24). These effects are especially pronounced in metrics that capture structural complexity - namely patch density (PD) and PLADJ, which consistently declined as resolution decreased. This pattern was evident across all landscape types, though the extent and nature of change varied. The metrics for proportion and SHDI were more stable but still exhibited resolution sensitivity depending on landscape structure. For instance, the proportion metric in dominant cohesive landscapes remained consistent across scales, while in minor class landscapes, even small changes in resolution led to over- or under-estimation of area coverage. This supports the notion that the interpretive reliability of metrics is intrinsically tied to landscape configuration, and highlights the risk of overgeneralization when applying global averages across diverse spatial contexts.

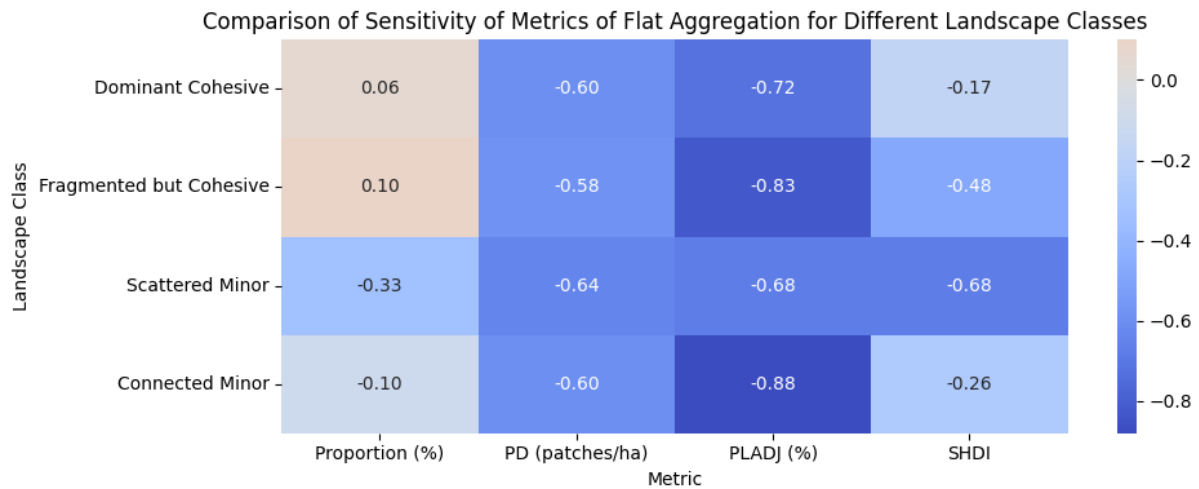


Figure 24. Sensitivity of landscape metrics to flat aggregation across different landscape classes.

One of the key hypotheses of this thesis, that the behavior of landscape metrics is dependent on underlying data characteristics was clearly supported by the stratified analysis. The grouped analysis of four landscape types (dominant cohesive, fragmented cohesive, scattered minor, connected minor) revealed that aggregation impacts are not uniform, and metric sensitivity is landscape-specific.

These results provide compelling evidence that MAUP cannot be universally mitigated or measured. Instead, it must be analyzed in the context of landscape structure, spatial heterogeneity, and resolution-specific behavior. This aligns with prior findings in the literature (Comber & Harris, 2022; Mirț et al., 2022), reinforcing the view that context-dependent strategies are essential for MAUP-sensitive studies.

Comparing flat and hierarchical aggregation methods revealed minor but notable differences. While overall trends were consistent between the two, flat aggregation typically

resulted in slightly more variance in metric values, particularly for SHDI and PLADJ. This suggests that hierarchical aggregation may offer marginal benefits in reducing metric volatility. Nevertheless, both strategies failed to fully preserve landscape detail, particularly in scattered and minor classes. This underscores that aggregation strategy alone is insufficient to address MAUP effects unless paired with thoughtful experimental design and landscape-specific interpretation.

The results of this study have significant implications for geospatial analysis, especially in contexts where data aggregation is required. Most critically, they demonstrate that landscape metrics respond differently to aggregation depending on landscape structure, and that metric behavior is not consistent across scales. This highlights the importance of carefully selecting the spatial resolution and aggregation method based on the specific goals and constraints of the study.

The choice of resolution should always be guided by the research objective. If the aim is to capture fine-scale spatial variation, then the DGGs resolution must be chosen accordingly to preserve the necessary level of detail. Using a coarser resolution in such cases could obscure important patterns and lead to misinterpretation. However, in situations where computational or financial resources are limited, a coarser resolution may be an acceptable compromise. In such cases, it is essential to recognize and justify the trade-offs involved, particularly with regard to the expected accuracy and reliability of the results. Moreover, uncertainty should not be viewed solely as a byproduct of data transformation, but rather as a function of how precise the results need to be for the intended purpose. The required level of accuracy should be explicitly stated at the outset of the analysis, and the experimental design, including the resolution and aggregation strategy, should be aligned with this expectation. When researchers define uncertainty in relation to analytical goals and limitations, they are better equipped to make context-sensitive choices that balance scientific rigor with practical feasibility.

To advance understanding of MAUP behavior and its implications for spatial metrics, future research should continue to explore the role of grid alignment and rotation, test a broader range of landscape metrics beyond those examined in this study, and evaluate strategies such as sampling design or ensemble modeling to characterize uncertainty more systematically. However, even as methodological tools evolve, one key principle stands out: uncertainty should be anticipated and minimized through careful experiment design, rather than treated as a problem to be fixed afterwards. Understanding general trends in metric sensitivity and aggregation effects will ultimately require larger and more diverse datasets, including real-world landscapes with varying structure and land cover composition. Still, no amount of data can compensate for a poorly formulated experiment. Therefore, research should always begin by clearly defining analytical goals, selecting appropriate resolution and methods accordingly, and making explicit choices about acceptable uncertainty. Only by doing so can spatial analysis move beyond reactive mitigation and toward more reliable, purpose-driven results.

9. Conclusion

This thesis explored how uncertainty emerges in DGGS-based spatial analysis due to the MAUP. By analyzing real-world land use data from Estonia across DGGS resolutions, it was found that landscape metrics exhibit resolution-dependent behavior. Patch density and PLADJ were highly sensitive to aggregation, while class proportion and SHDI showed greater stability, depending on the spatial structure of the landscape. These findings highlight that aggregation effects are not uniform but are heavily influenced by the underlying data characteristics.

The aim of this research was to determine whether spatial uncertainty introduced by DGGS aggregation could be universally measured and mitigated. The study confirmed that such uncertainty cannot be measured with a single method; more precisely, generalized assumptions about metric behavior are not meaningful, as the variability across landscapes is too wide. As a result, decisions based on aggregated or averaged results may be misleading if the underlying spatial structure is not carefully considered. Rather than eliminating MAUP effects, this thesis argues that their impact must be analyzed and anticipated within the context of the experiment design.

This research provides valuable insight into the data dependency of spatial uncertainty in DGGS systems. It underscores the limitations of assuming that equal-area grids or fine-resolution data automatically resolve MAUP issues. By showing how different landscape types respond to aggregation, this work helps geospatial analysts better understand which metrics are reliable under varying conditions and why careful methodological choices matter. The findings are especially relevant for environmental monitoring, spatial planning, and any field relying on geospatial analysis.

While the analysis revealed key patterns in metric sensitivity, it was based on a single national dataset and did not incorporate other landscape patterns not typical for Estonia. These choices were made to keep the focus on real-world applicability, but they limit the generalizability of the results. Additionally, the study concentrated on four core landscape metrics; other relevant metrics may behave differently under aggregation. Future studies should test a broader variety of landscape configurations, include additional metrics, and employ synthetic landscapes for controlled comparisons.

Ultimately, this thesis highlights that spatial uncertainty in DGGS-based analysis is not a problem that can be resolved through technical solutions or data processing improvements alone. Instead, it represents a fundamental characteristic of spatial data aggregation, one that is shaped by the interaction between spatial units, metric behavior, and the structure of the underlying landscape. As such, uncertainty must be addressed proactively through informed, context-sensitive experiment design. This involves selecting spatial resolution and aggregation strategies based not only on data availability or computational convenience, but on a clear understanding of how those choices affect analytical outcomes. By shifting the focus away from purely technical mitigation and toward conceptual clarity and interpretive awareness, geospatial research can become more reliable, transparent in its assumptions, and ultimately more aligned with its intended purpose.

Bibliography

- Andresen, Martin. (2021). *Modifiable areal unit problem*. *CrimRxiv*. DOI: <http://dx.doi.org/10.21428/cb6ab371.5c28c076>
- Arnab, R. (2017). *Survey Sampling Theory and Applications (1st ed.)*. Elsevier Science & Technology.
- Briz-Redón, Á. (2022). *A Bayesian shared-effects modeling framework to quantify the modifiable areal unit problem*. DOI: <https://doi.org/10.1016/j.spasta.2022.100689>.
- Chaudhuri, A., & Stenger, H. (2005). *Survey Sampling: Theory and Methods, Second Edition (2nd ed.)*. CRC Press. <https://doi.org/10.1201/9781420028638>
- Comber, A., & Harris, P. (2022). *The Importance of Scale and the MAUP for Robust Ecosystem Service Evaluations and Landscape Decisions*. *Land (Basel)*, 11(3), 399. <https://doi.org/10.3390/land11030399>
- Durrett, R. (2019). *Probability: Theory and Examples (Fifth edition., Vol. 49)*. Cambridge University Press. <https://doi.org/10.1017/9781108591034>
- Dutton, G. (2015). *The Making of a Global Grid—And What To Make of It*. *Digital Earth Conference*. Microsoft Word - GDutton-DigitalEarth2015-paper.docx
- Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Incorporated.
- Goodchild, M. F. (2018). *Reimagining the history of GIS*. *Annals of GIS*, 24(1), 1-8. <https://doi.org/10.1080/19475683.2018.1424737>
- Goodchild, M.F. (2019). Preface. *Cartographica: The International Journal for Geographic Information and Geovisualization* 54(1), 1-3. <https://muse.jhu.edu/article/721683>.
- Goodchild, M. F., & Yang, S. (1989). *A hierarchical spatial data structure for global geographic information systems*. [https://doi.org/10.1016/1049-9652\(92\)90032-S](https://doi.org/10.1016/1049-9652(92)90032-S).
- Hankin, D., Mohr, M. S., & Newman, K. B. (2019). *Sampling Theory: For the Ecological and Natural Resource Sciences (1st ed.)*. Oxford University Press. <https://doi.org/10.1093/oso/9780198815792.001.0001>
- International Organization for Standardization. (2021). *ISO 19170-1:2021 Geographic information — Discrete Global Grid Systems (DGGs) — Framework*. Geneva, Switzerland: ISO. *ISO 19170-1:2021(en), Geographic information — Discrete Global Grid Systems Specifications — Part 1: Core Reference System and Operations, and Equal Area Earth Reference System*
- Kmoch, Alexander & Vasilyev, Ivan & Virro, Holger & Uuemaa, Evelyn. (2022a). *Area and shape distortions in open-source discrete global grid systems*. *Big Earth Data*. 6. 1-20. <https://doi.org/10.1080/20964471.2022.2094926>
- Kmoch, A., Matsibora, O., Vasilyev, I., & Uuemaa, E. (2022b). *Applied open-source Discrete Global Grid Systems*. *AGILE: GIScience Series*, 3, 41–58. <https://doi.org/10.5194/agile-giss-3-41-2022>

Kmoch, A., Sahr, K., Chan, W. T., & Uuemaa, E. (2025). *IGEO7: A new hierarchically indexed hexagonal equal-area discrete global grid system*. *AGILE: GIScience Series*, accepted, May 2025. ISSN, 2700-8150.

Kmoch, A. (2025). *dggrid4py: A set of python modules for creating and manipulating Discrete Global Grids with DGGRID v7 (v0.4.0)*. Zenodo. <https://doi.org/10.5281/zenodo.15069400>

Lee, S. I., Lee, M., Chun, Y., & Griffith, D. A. (2018). *Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: a simulation study*. *International Journal of Geographical Information Science*, 33(6), 1135–1154. <https://doi.org/10.1080/13658816.2018.1542699>

Li, M., & Stefanakis, E. (2020). *Geospatial operations of discrete global grid systems—a comparison with traditional GIS*. *Journal of Geovisualization and Spatial Analysis*, 4(26). <https://doi.org/10.1007/s41651-020-00066-3>

Maa-amet. (2016 a). *TOPOGRAAFILISTE ANDMETE KAARDISTUSJUHEND. Eesti topograafia andmekogu | Geoportaal | Maa- ja Ruumiamet. chrome-extension://efaidnbmninnibpcjpcglclefindmkaj/https://geoportaal.maaamet.ee/docs/ETAK/ETAK_juhend2016.pdf*

Maa-amet. (2016 b). *ETAK juhend - nähtuste kataloog. Reaalsusmudel | Geoportaal | Maa- ja Ruumiamet*. https://geoportaal.maaamet.ee/index.php?lang_id=1&action=kataloog&page_id=88

Mahdavi-Amiri, A., Alderson, T., & Samavati, F. (2015) *A Survey of Digital Earth*, Computers & Graphics, Volume 53, Part B, Pages 95-117, ISSN 0097-8493, <https://doi.org/10.1016/j.cag.2015.08.005> .

McGarigal, K. (2015). *FRAGSTATS help*. University of Massachusetts: Amherst, MA, USA, 182.

Mingke Li, Heather McGrath, Emmanuel Stefanakis. (2022). *Multi-resolution topographic analysis in hexagonal Discrete Global Grid Systems*. *International Journal of Applied Earth Observation and Geoinformation*, Volume 113, 2022, 102985, ISSN 1569-8432. <https://doi.org/10.1016/j.jag.2022.102985> .

Mirț, A., Reiche, J., Verbesselt, J., & Herold, M. (2022). *A Downsampling Method Addressing the Modifiable Areal Unit Problem in Remote Sensing*. *Remote Sensing*, 14(21), 5538. <https://doi.org/10.3390/rs14215538>

Openshaw, S. (1984). *“The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography 38,”* Geobooks, Norwich. *The Modifiable Areal Unit Problem | CiNii Research*

Parring, A.-M., Käärik, E., Vähi, M., & Tartu Ülikool. (1997). *Matemaatilise statistika instituut. Statistilise andmetöötuse algõpetus*. Tartu Ülikooli Kirjastus.

Perlman, J. (2016, March 20). *Bring It All Down to Earth with DGGS*. *ApogeoSpatial*. URL: <https://apogeospatial.com/2079-2/>.

Peterson, P. R. (2016). *Discrete global grid systems. International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology, 1-10.* <https://doi.org/10.1002/9781118786352.wbieg1050>

Raposo, P., Robinson, A. C., & Brown, R. (2019). *A Virtual Globe Using a Discrete Global Grid System to Illustrate the Modifiable Areal Unit Problem. Cartographica, 54(1), 51–62.* <https://doi.org/10.3138/cart.54.1.2018-0015>

Sahr, K., White, D., & Kimerling, A. J. (2003). *Geodesic Discrete. Global Grid Systems, Cartography and Geographic Information Science, 30:2, 121-134, DOI: 10.1559/152304003100011090*

Sahr, K. (2019). *Central Place Indexing: Hierarchical Linear Indexing Systems for Mixed-Aperture Hexagonal Discrete Global Grid Systems. Cartographica: The International Journal for Geographic Information and Geovisualization 54(1), 16-29.* <https://muse.jhu.edu/article/721685>.

Samet, H. (1995). *Spatial Data Structures.* <https://www.cs.umd.edu/users/hjs/pubs/kim.pdf>

Thompson, J. A., Brodzik, M. J., Silverstein, K. A. T., Hurley, M. A., & Carlson, N. L. (2022). *EASE-DGGS: A hybrid discrete global grid system for Earth sciences. Big Earth Data, 6(3), 340-357.* <https://doi.org/10.1080/20964471.2021.2017539>

Traat, I., Inno, J., & Tartu Ülikool. (1997). *Matemaatilise statistika instituut. Tõenäosuslik valikuuring : [õpik]. Tartu Ülikooli Kirjastus.*

Tveter, E., Laird, J., & Aalen, P. (2021). *SPATIAL AGGREGATION ERROR AND AGGLOMERATION BENEFITS FROM TRANSPORT IMPROVEMENTS. DOI: 10.13140/RG.2.2.23456.25603*

Uemaa, E. (2004). *MAASTIKUINDEKSITE SÕLTUVUS LÄHTEANDMETE RUUMILISEST LAHUTUSEST NING INDIKATSIOONIVÄÄRTUS VALGLATEST TOITAINETE JA ORGAANILISTE AINETE VÄLJAKANDES. TÜ geograafia instituudi magistritöö.* <https://dspace.ut.ee/server/api/core/bitstreams/08bb98a8-e803-4547-b41d-d917e68e2ec6/content>

Yaqi Wang, Qian Di. (2020). *Modifiable areal unit problem and environmental factors of COVID-19 outbreak. Doi: https://doi.org/10.1016/j.scitotenv.2020.139984.*

Ye, S. (2024). *A Sensitivity Test on the Modifiable Areal Unit Problem in the Spatial Aggregation of Fossil Data. Geosciences, 14(9), 247.* <https://doi.org/10.3390/geosciences14090247>

License

Non-exclusive licence to reproduce thesis and make thesis public

I, Aleksandra Rammul,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
Measuring Uncertainty Related to Ingesting Data to DGGS,
supervised by **Alexander Knoch, Evelyn Uemaa;**
2. grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;
3. am aware of the fact that the author retains the rights specified in p. 1 and 2;
4. certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Aleksandra Rammul

15.05.2025

Appendix 1. Code Repository

All scripts, functions, and processing workflows used in this thesis are available in a public GitHub repository:

Rammul, A. (2025). *thesis-Rammul-2025* [Source code]. GitHub. Available at:
<https://github.com/ramasha-git/thesis-Rammul-2025>