HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES

DELINEATING ENUMERATION AREAS FOR TÜRKİYE: A TRIAL ON ANKARA

Cansu ÖZTÜRK

Department of Social Research Methodology
PhD Thesis

Ankara
January 2026

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES

DELINEATING ENUMERATION AREAS FOR TÜRKİYE: A TRIAL ON ANKARA

Cansu ÖZTÜRK

Supervisor
Prof. Dr. Ahmet Sinan TÜRKYILMAZ

Department of Social Research Methodology
PhD Thesis

Ankara
January 2026

APPROVAL PAGE

Delineating Enumeration Areas for Türkiye: A Trial On Ankara

Cansu ÖZTÜRK

This is to certify that we have read and examined this thesis, and, in our opinion, it fulfils the requirements in scope and quality of a thesis for the degree of Doctor of Philosophy in Social Research Methodology.

Jury Members:

Member (Chair): *(signature)*
Prof. Dr. Yaprak Arzu ÖZDEMİR,
Gazi University, Faculty of Sciences, Department of Statistics

Member (Supervisor): *(signature)*
Prof. Dr. Ahmet Sinan TÜRKYILMAZ
Hacettepe University, Institute of Population Studies, Department of Social Research Methodology

Member: *(signature)*
Prof. Dr. İlknur YÜKSEL-KAPTANOĞLU
Hacettepe University, Institute of Population Studies, Department of Social Research Methodology

Member: *(signature)*
Prof. Dr. Erdem KARABULUT
Hacettepe University, Faculty of Medicine, Department of Biostatistics

Member: *(signature)*
Prof. Dr. Murat YÜCEŞAHİN
Ankara University, Faculty of Language, History and Geography, Department of Geography

Member: *(signature)*
Prof. Dr. Alanur ÇAVLİN BİRCAN
Hacettepe University, Institute of Population Studies, Department of Demography

This thesis has been accepted by the above-signed members of the Jury and has been confirmed by the Administrative Board of the Institute of Population Studies, Hacettepe University.

(signature)
... /.../2026 Prof. Dr. İsmet KOÇ
Director

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES
TO THE DEPARTMENT OF SOCIAL RESEARCH METHODOLOGY

Date: 15/03/2026

Thesis Title / Topic: Delineating Enumeration Areas for Türkiye: A Trial On Ankara

According to the originality report obtained by myself/my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 15/03/2026 for the total of 235 pages including the a) Title Page, b) Introduction, c) Main Chapters, and d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 2 %.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes excluded
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Institute of Population Studies Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

Name Surname:	Cansu ÖZTÜRK
Student No:	N20143677
Department:	Social Research Methodology
Program:	Social Research Methodology
Status:	☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Prof. Dr. A. Sinan TÜRKYILMAZ

(Title, Name Surname, Signature)

SIMILARITY INDEX PAGE FROM TURNITIN PROGRAM

Cansu Öztürk

PHD Thesis

Cansu PhD
PhD
Hacettepe Üniversitesi

Document Details

Submission ID
trn:oid:::1:3507279698

Submission Date
Mar 15, 2026, 10:46 AM GMT+3

Download Date
Mar 15, 2026, 10:54 AM GMT+3

File Name
Areas_for_T_rkiye_-_A_trial_on_Ankara_PhD_Thesis_13032026_v2.pdf

File Size
6.0 MB

235 Pages

59,220 Words

371,686 Characters

turnitin Page 1 of 243 - Cover Page Submission ID trn:oid:::1:3507279698

2% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Exclusions

- 2 Excluded Sources

Match Groups

- **77 Not Cited or Quoted 1%**
  Matches with neither in-text citation nor quotation marks
- **25 Missing Quotations 0%**
  Matches that are still very similar to source material
- **0 Missing Citation 0%**
  Matches that have quotation marks, but no in-text citation
- **0 Cited and Quoted 0%**
  Matches with in-text citation present, but no quotation marks

Top Sources

1% Internet sources
1% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

ETHICAL DECLARATION

In this thesis study, I declare that all the information and documents have been obtained in the base of the academic rules and all audio-visual and written information and results have been presented according to the rules of scientific ethics. I did not do any distortion in data set. In case of using other works, related studies have been fully cited in accordance with the scientific standards. I also declare that my thesis study is original except cited references. It was produced by myself in consultation with my supervisor (Prof. Dr. Ahmet Sinan TÜRKYILMAZ) and written according to the rules of thesis writing of Hacettepe University Institute of Population Studies.

(signature)

Cansu ÖZTÜRK

DECLARATION OF PUBLISHING AND INTELLECTUAL PROPERTY RIGHTS

I declare that I give permission to Hacettepe University to archive all or some part of my master/PhD thesis, which is approved by the Institute, in printed (paper) or electronic format and to open to access with the following rules. With this permission, I hold all intellectual property rights, except using rights given to the University, and the rights of use of all or some parts of my thesis in the future studies (article, book, license, and patent).

I declare that the thesis is my original work, I did not violate rights of others and I own all rights of my thesis. I declare that I used texts with the written permit which is taken by owners and I will give copies of these to the University, if needed.

As per the "Regulation on the Online Availability, Arrangement and Open Access of Graduate Theses" of Council of Higher Education, my thesis shall be deposited to National Theses Center of the Council of Higher Education/Open Access System of H.U. libraries, except for the conditions indicated below;

- The access to my thesis has been postponed for 2 years after my graduation as per the decision of the Institute/University board.(1)
- The access to my thesis has been postponed for …. month(s) after my graduation as per the decision of the Institute/University board.(2)
- There is a confidentiality order for my thesis.(3)

15/03/2026
(signature)

Cansu ÖZTÜRK

---

[i] Regulation on the Online Availability, Arrangement and Open Access of Graduate Theses

(1) Article 6.1. In the event of patent application or ongoing patent application, the Institute or the University Board may decide to postpone the open access of the thesis for two years, upon the proposal of the advisor and the assent of the Institute Department.

(2) Article 6.2. For theses that include new techniques, material and methods, that are not yet published articles and are not protected by patent and that can lead to unfair profit of the third parties in the event of being disseminated online, the open access of the theses may be postponed for a period not longer than 6 months, as per the decision of the Institute or the University Board upon the proposal of the advisor and the assent of the Institute Department.

(3) Article 7.1. The confidentiality order regarding the theses that concern national interest or security, the police, intelligence, defense and security, health and similar shall be issued by the institution certified the thesis*. The confidentiality order for theses prepared pursuant to the cooperation protocol with institutions and organizations shall be issued by the University Board, upon the proposal of the related institutions and organizations and the assent of the Institute or the Faculty. The theses with confidentiality order shall be notified to the Council of Higher Education. Article 7.2. During the confidentiality period, the theses with confidentiality order shall be kept by the Institute or the Faculty in accordance with the confidentiality order requirements, in the event of termination of the confidentiality order the thesis shall be uploaded to Thesis Automation System.

* Shall be issued by the Institute or Faculty Board upon the proposal of the advisor and the assent of the Institute Department

ACKNOWLEDGEMENTS

Reaching this stage still feels difficult to believe. Throughout this long and demanding journey, there were many moments when I thought I might not be able to continue. Looking back today, however, I see not only the challenges but also the growth, persistence, and support that carried me here.

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dr. A. Sinan TÜRKYILMAZ. He was not only the originator of the idea behind this thesis, which we pursued wholeheartedly with the aim of contributing to the literature in Türkiye, but also a constant source of support throughout this process. His guidance extended far beyond academic supervision; at different moments, he became a psychologist, a life coach, and an elder brother to me. I will always remain grateful for his belief in me, his encouragement, and his invaluable mentorship.

I would also like to thank my friends, colleagues, and loved ones who believed in me even more than I believed in myself. Whenever I found myself filled with doubt, they were the ones who reminded me that I could succeed. Their readiness to help, their encouragement, and their unwavering support meant more than I can fully express.

My heartfelt thanks also go to my beloved family, who witnessed this long journey alongside me and supported me with patience and understanding through the years.

Finally, I would like to thank myself. This process was far from easy. I faced many hardships and made many sacrifices in terms of time, health, and personal life. Yet, as I look back now, I feel proud of myself for completing this thesis in a way that truly satisfies me and for earning my doctoral degree through years of hard work and sacrifice. The happiness and pride of successfully completing this journey, and of making those who believed in and trusted me happy as well, are feelings I will always carry with me.

ABSTRACT

This thesis develops a rule-based and GIS-supported workflow for the delineation of Enumeration Areas (EAs) in Türkiye and evaluates it through the case of Ankara. The study focuses on the lack of an EA system in Türkiye that could improve the technical processes of census and sample surveys and support field implementation planning.

The literature on EA production was reviewed, and the application was carried out in ArcGIS Pro and R. The number of residential units was used as the target variable, while physical barriers were treated as constraining elements. EA production was based on barrier-sensitive neighbourhood definition, urbanisation-sensitive targets, rule-based growing, and split-merge mechanisms. Based on the findings, a final workflow developed in R was proposed, together with a Shiny-based interface supporting its practical use.

The pilot implementation at the neighbourhood scale showed that the workflow is applicable and that some areas may remain structurally indivisible under strict constraints. At the district scale, the workflow was applied to all neighbourhoods in Çankaya under a reference scenario and was compared with alternative scenarios that varied in terms of allowing zero-target units and splitting policies. The findings indicate that the method was able to establish an EA framework at the district level and to produce reproducible spatial and tabular outputs. Full compliance with the targeted value range, however, could not be achieved in all cases because of intense urban heterogeneity, deficiencies in barrier data, and rules designed to preserve building integrity.

The main contribution of the study is the development of a transparent EA production workflow suitable for pilot implementation and institutional adaptation in Türkiye. The study shows that the use of EAs in statistical production can enable time- and cost-efficient field operations and support the production of statistics with stronger representativeness. In this respect, the adoption of EA systems in Türkiye should be regarded as a strategic investment for improving the effectiveness and quality of national statistical production.

Keywords: Enumeration Area, Türkiye, geographic information systems, barrier-sensitive delineation methodology, census and sample surveys

ÖZET

Bu tez, Türkiye için Sayım Alanlarının (Enumeration Area, EA) tanımlanmasına yönelik kural tabanlı ve CBS destekli bir iş akışı geliştirmekte ve bu yaklaşımı Ankara örneği üzerinden değerlendirmektedir. Çalışma, Türkiye'de sayım ve örnekleme araştırmalarının teknik süreçlerini iyileştirebilecek ve saha uygulaması planlamasını destekleyebilecek bir EA sisteminin eksikliğine odaklanmaktadır.

Çalışmada EA üretimine ilişkin literatür incelenmiş; ArcGIS Pro ve R yazılımlarında uygulama gerçekleştirilmiştir. Uygulamada konut sayısı hedef değişken, fiziksel bariyerler sınırlandırıcı unsur olarak ele alınmış; EA üretimi bariyer-duyarlı komşuluk, kentleşme derecesine duyarlı hedefler, kural tabanlı büyütme ve bölme-birleştirme mekanizmaları temelinde yürütülmüştür. Bulgular doğrultusunda R ortamında geliştirilen nihai bir iş akışı önerilmiş, bu yapıyı temel alan Shiny tabanlı bir arayüz geliştirilmiştir.

Mahalle ölçeğindeki pilot uygulama, iş akışının uygulanabilir olduğunu ve sıkı kısıtlar altında yapısal olarak bölünemeyen bazı alanların oluşabildiğini göstermiştir. İlçe ölçeğinde iş akışı, Çankaya'daki tüm mahallelere referans bir senaryo altında uygulanmış ve sıfır-hedef içerip içermeme ile bölme politikalarını değiştiren alternatif senaryolarla karşılaştırılmıştır. Bulgular, yöntemin ilçe düzeyinde bir EA çerçevesi oluşturabildiğini ve yeniden üretilebilir mekânsal ve tablosal çıktılar sağlayabildiğini göstermektedir. Yoğun kentsel heterojenlik, bariyer verilerindeki eksiklikler ve bina bütünlüğünü koruyan kurallar nedeniyle hedeflenen değer aralığına tam uyum her durumda sağlanamamıştır.

Çalışmanın temel katkısı, Türkiye'de pilot uygulama ve kurumsal uyarlamaya uygun, şeffaf bir EA üretim iş akışının geliştirilmiş olmasıdır. Çalışma, istatistik üretiminde EA kullanımının zaman ve maliyet etkin saha uygulamalarını mümkün kılabildiğini ve temsiliyeti güçlendirilmiş istatistiklerin üretilmesini destekleyebildiğini göstermektedir. Bu çerçevede EA sistemlerinin Türkiye'de kullanıma alınması, ulusal istatistik üretiminin etkinliğini ve kalitesini artırmak açısından stratejik bir yatırım olarak değerlendirilmelidir.

Anahtar Kelimeler: Sayım Alanı, Türkiye, coğrafi bilgi sistemleri, bariyer-duyarlı sınırlandırma, sayım ve örnekleme araştırmaları

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVATIONS

AD-EA	Automated Delineation of Enumeration Areas
ArcPy	ArcGIS Python Library
BBZ	Build Balanced Zones
CA	Census Area
CSV	Comma-Separated Values
DEGURBA	Degree of Urbanisation
EA	Enumeration Area
EU	European Union
GPKG	GeoPackage
GPS	Global Positioning System
GUI	Graphical User Interface
GIS	Geographic Information System
HTML	HyperText Markup Language
ID	Identifier
MAUP	Modifiable Areal Unit Problem
OSM	OpenStreetMap
PDF	Portable Document Format
PNG	Portable Network Graphics
QA	Quality Assurance
R	R Statistical Computing Environment
SA	Supervisor Area
SHP	Shapefile
TurkStat	Turkish Statistical Institute
UN	United Nations
UNSD	United Nations Statistics Division
UTF-8	8-bit Unicode Transformation Format
WGS84	World Geodetic System 198

CHAPTER 1. INTRODUCTION

Accurate and efficient data collection is a cornerstone of modern national statistical systems. Population censuses, sample surveys, and administrative registers constitute the main sources used for official statistics production. In many contexts, sample surveys remain essential because they can deliver regular and detailed estimates with manageable cost and within shorter production cycles than full enumeration. The performance of sample surveys, however, depends on the quality of the sampling frame and on the operational units through which data collection is organised.

Enumeration Areas (EAs) are the smallest operational spatial units used to organise enumeration and to manage fieldwork workloads. Although EAs were historically developed for censuses, they are widely used today as primary sampling units or as the lowest-level building blocks of master sampling frames for repeated household surveys. A well-designed EA system supports complete geographic coverage, transparent workload allocation, and consistent measurement. In operational terms, EAs are expected to be contiguous, to provide full coverage without overlaps or gaps, and to align with recognisable boundaries that field staff can identify on the ground.

International guidance emphasises that an EA is primarily an operational unit. The United Nations Statistics Division describes enumeration areas as the smallest geographic units created for census and survey operations, intended to be manageable by a single enumerator within a defined period (UNSD, 2010; United Nations, 2017).

In practice, statistical offices often specify indicative workload bands, usually in terms of households or dwellings, to balance field efficiency and statistical equity. Very small areas tend to increase listing, supervision, and training costs, whereas very large or internally fragmented areas can be difficult to cover and may increase non-sampling error. For household-based operations, bands around 80–120 dwellings are frequently cited as a pragmatic reference, although thresholds are adapted to local settlement structure and survey mode (Kish, 1965; UNSD, 2010). Comparable

backbone geographies include Output Areas in the United Kingdom and Dissemination Areas in Canada (ONS, 2022; Statistics Canada, 2021).

EAs link spatial organisation to survey quality because they shape both coverage and sampling properties. When EA boundaries are outdated or inconsistent, frame coverage becomes difficult to verify and fieldwork becomes harder to plan and supervise. When workloads vary widely across EAs, survey operations may require additional adjustment and can lose efficiency. In contrast, EAs that are workload-balanced, compact, and aligned with clear physical features can support more predictable fieldwork and can strengthen the comparability of repeated surveys across time and space.

Many EA systems were created under paper-map workflows and were later updated through manual revision. Rapid urbanisation, infrastructure expansion, and continuous change in the built environment make it increasingly difficult to maintain EA boundaries through manual procedures alone. At the same time, contemporary survey programmes require EAs that are sufficiently stable to support repeated selection and comparable measurement over time, while remaining responsive to demographic change.

Advances in geographic information systems (GIS) and the availability of digital geospatial layers create an opportunity to strengthen EA production through rule-based and semi-automated approaches. A central methodological tension, however, persists. On the one hand, EAs should be statistically balanced in expected workload to support efficient sample allocation and field management. On the other hand, EAs must satisfy spatial constraints such as contiguity and compliance with barriers (major roads, railways, and waterways), and they should remain interpretable for field staff and supervisors.

From a sampling perspective, EAs commonly function as primary sampling units (PSUs) within multi-stage designs, followed by the selection of dwellings or households and then individuals. The definition of the sampling unit is not a minor detail: an unsuitable PSU can inflate design effects, complicate weighting, and reduce precision even when probability sampling procedures are correctly implemented (Cochran, 1977; Kish, 1965; Lohr, 2019).

Two failure modes are particularly relevant for operational and statistical performance. First, oversized EAs can contain several distinct settlement types and therefore high within-unit heterogeneity. Second, geometrically fragmented EAs that cross major barriers or contain disconnected parts can undermine field navigation and boundary control. Both conditions increase the risk of systematic omissions, duplicate visits, and avoidable field costs.

This thesis addresses a practical research gap: the need for an operationally grounded, barrier-aware, and reproducible delineation workflow that can translate spatial address information and built-environment layers into workload-balanced EAs. The thesis further argues that delineation workflows should embed diagnostic checkpoints that detect topology errors, unrealistic adjacency relations, and workload anomalies during production rather than after deployment. Embedding diagnostics within the workflow supports transparent quality assurance and strengthens the credibility of the resulting sampling frame.

In Türkiye, household sample surveys have traditionally relied on address lists that are largely text-based. Text-only frames can constrain the ability to plan workloads using geography and can limit the use of spatial diagnostics for monitoring coverage. In response, the Turkish Statistical Institute has initiated the Spatial Address Registration System (SARS) to transform address information into a spatially referenced database aligned with national and international standards. A spatial address database creates a foundation for building geographically explicit sampling frames and for integrating geospatial diagnostics into survey operations.

EA-like operational units in Türkiye have historically been derived from address-based administrative records and street-numbering structures. This practice provides a strong register backbone, yet it may not systematically enforce spatial constraints such as barrier compliance and geometric compactness, and it can be difficult to maintain in areas experiencing rapid urban change. The present study does not frame these institutional practices as a deficiency; rather, it examines whether a spatially explicit, documented, and reproducible delineation pipeline can complement register-based systems by providing transparent geometry, diagnostics, and update mechanisms.

The transition toward spatially enabled registers also raises practical questions. Comparisons between legacy address databases and spatial datasets can reveal differences in boundary assignment, missing or misclassified building records, and heterogeneous block geometries that were not previously visible in list-based systems. Moreover, Türkiye includes diverse settlement patterns, ranging from dense urban neighbourhoods to low-density and peri-urban areas, where administrative boundaries and physical barriers interact in complex ways. These conditions motivate a delineation approach that is explicit in its rules, adaptable across contexts, and auditable through formal diagnostics.

The overall aim of this thesis is to develop and evaluate a GIS-based, rule-driven approach for delineating Enumeration Areas in Türkiye, with a focus on producing spatial units that are operationally feasible for fieldwork and suitable for use in sampling frames. The thesis proposes an Adaptive District-based Enumeration Area (AD-EA) workflow that combines building-based workload measures with barrier-aware contiguity rules and systematic quality assurance checkpoints.

Within this overall aim, the thesis addresses practical design challenges that recur in automated delineation. These include enforcing workload balance under explicit thresholds, incorporating major physical barriers as non-crossable boundaries, ensuring building indivisibility so that residential units are not split across EAs, maintaining computational performance when processing large urban datasets, and handling exceptional cases where limited barriers or homogeneous morphology allow units to exceed intended upper bounds. Together, these challenges provide an empirical basis for discussing the limits of full automation and identifying where targeted human intervention may remain necessary.

The thesis pursues four specific objectives. First, it seeks to define operational and statistical criteria for EA design that are consistent with survey sampling principles and relevant international guidance. Second, it aims to implement a reproducible GIS workflow capable of delineating contiguous, non-overlapping EAs with full coverage within administrative boundaries. Third, it examines how different barrier combinations and parameter settings influence EA geometry and workload balance. Fourth, it assesses the feasibility of scaling the proposed approach from a neighbourhood-level pilot to a district-level application. In line with these objectives,

the study is guided by a set of research questions addressing the design, implementation, and applicability of the proposed EA workflow.

Under these objectives this thesis addresses the following research questions:

RQ1. How can EAs be delineated through explicit spatial rules so that they remain contiguous, barrier-respecting, and workload-balanced at the same time?

RQ2. To what extent can building-based measures (for example, the number of residential units or housing addresses) serve as a practical proxy for enumerator workload in EA design?

RQ3. How sensitive are EA outputs to key parameters (such as target workload bands and contiguity constraints) and to alternative barrier layer combinations?

RQ4. What diagnostic checks can be integrated into the production workflow to identify topology errors, unrealistic neighbour links, and workload anomalies early in the process?

RQ5. Under what conditions can the proposed workflow be scaled from a pilot neighbourhood to a district-wide EA framework without loss of operational interpretability?

The study is conducted under several working assumptions. First, building footprints and their residential attributes provide an adequate operational proxy for expected household workloads in the study context. Second, the main physical barriers relevant for field segmentation can be represented through road, railway, and water layers, after standard cleaning and harmonisation steps. Third, EA boundaries should align with barriers and other recognisable features where possible, while still maintaining full coverage, contiguity, and practical workload targets.

The empirical evaluation is organised in two scales. The first scale is a neighbourhood-level pilot application used to develop the workflow, refine parameters, and test diagnostics under controlled conditions. In line with common survey practice, the pilot operationalises workload balance through a target cluster size defined on housing-related measures, with a tolerance band that supports practical delineation decisions. The second scale extends the same logic toward district-level

aggregation, where neighbourhood outputs are combined and assessed for spatial consistency and workload distribution.

The neighbourhood-scale pilot is implemented for Bağlıca, a rapidly developing neighbourhood in Etimesgut district (Ankara), chosen because it combines dense new housing, ongoing construction, and mixed settlement morphology. In the pilot, the workload target is operationalised as 80–120 residential units per EA, and outputs exceeding 200 units are treated as exceptional cases for diagnostic review. Delineation is based on spatial layers—administrative boundaries, residential buildings, and barrier networks—and does not rely on internal administrative coding to construct EA geometry.

The thesis concentrates on GIS-based delineation using administrative boundaries, building data with residential attributes, and barrier layers. It prioritises methodological transparency and repeatability, including the ability to reproduce outputs from the same inputs and parameters. Topics that require additional institutional decisions, such as long-term maintenance cycles, extensive field verification protocols, confidentiality rules for micro-level address data, and governance arrangements for interagency data sharing, are discussed as implementation considerations rather than as optimisation objectives within the delineation algorithm.

Following this introduction, Chapter 2 reviews the conceptual foundations of EAs and sampling frames and surveys international practices in small-area geographies and automated zoning. Chapter 3 describes the data sources and preprocessing, including building-based residential attributes and barrier layers. It presents the conceptual and operational design of the EA delineation approach, specifies the constraints and quality criteria, and details the computational workflow and implementation, including neighbour-graph construction, barrier filtering, diagnostic checkpoints, the neighbourhood-scale case study, key parameter choices, exception handling, and the district-scale application. Chapter 4 presents the findings from the applications described in Chapter 3. Chapter 5 discusses methodological contributions, limitations, and institutional implications for a spatial sampling frame in Türkiye, and concludes with priorities for future work.

CHAPTER 2. LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

This chapter provides the theoretical and empirical foundation for the Enumeration Area (EA) production approach developed in this thesis. It reviews the existing literature on the definition, purpose, and evolution of Enumeration Areas, examines international practices, and discusses the major methodological challenges associated with EA design. The chapter also introduces the key conceptual frameworks used to classify EA production approaches, including differences in spatial representation, use of Geographic Information Systems (GIS), and levels of automation.

The first part of the chapter focuses on the concept of Enumeration Areas as fundamental spatial units in population censuses and sample surveys. It discusses their definitions, types, importance, and areas of application, emphasizing their dual role as both operational units for fieldwork and analytical units for statistical dissemination. This section establishes why EA design is a critical component of national statistical systems rather than a purely technical mapping exercise.

The chapter then situates Enumeration Areas within an international context by reviewing how different countries and statistical offices implement EA systems under varying institutional, administrative, and technological conditions. These international examples highlight both common design principles and context-specific adaptations, illustrating the diversity of EA practices worldwide.

Subsequently, the chapter examines the methodological challenges inherent in EA design, including issues related to spatial aggregation, workload balance, parameter sensitivity, and the Modifiable Areal Unit Problem (MAUP). These challenges motivate the transition from traditional manual delineation methods to GIS-supported and algorithmic approaches.

Building on this discussion, the chapter introduces a conceptual classification of EA production approaches based on three orthogonal dimensions: the technical environment (GIS-based versus non-GIS-based), the spatial representation unit (cell-based versus object-based), and the level of automation (manual, semi-automated,

automated). This framework provides a structured lens through which existing methods are reviewed and compared.

Finally, the chapter reviews GIS-based, cell-based, and object-based EA production methods in the literature, including the role of linear barriers and context-sensitive constraints. The chapter concludes by identifying key research gaps and positioning the proposed method of this thesis within the existing body of work.

2.1. Enumeration Areas

This section introduces the concept of Enumeration Areas as the foundational spatial units of census and survey operations. It examines how EAs are defined in the literature, the different types of EAs used in practice, and their importance for data collection, statistical analysis, and policy-relevant applications. By clarifying the role and function of Enumeration Areas, this section establishes the conceptual basis for the methodological discussions that follow in subsequent sections. To provide context, it first outlines the historical emergence of EAs as operational field units and then summarises the core design principles that recur across national implementations.

Enumeration Areas (EAs) emerged as a practical response to a core operational need: organising census fieldwork into manageable workloads with boundaries that enumerators can recognise on the ground. Early EA systems were typically produced through manual sketch mapping and local knowledge, which provided high interpretability but limited reproducibility and difficult updating. Over time, many national statistical systems transitioned toward digital cartography and GIS-supported workflows, allowing EA boundaries to be maintained as part of a geospatial database and linked more systematically to address/building sources, administrative geographies, and census outputs. In recent practice, the evolution continues in two directions: (i) greater standardisation and auditability of boundary maintenance, and (ii) increased use of rule-based or optimisation-inspired approaches—especially where timely updating, large-scale coverage, or limited field mapping capacity necessitates scalable production methods (United Nations, 2017; United Nations, 2025).

Despite cross-country differences in terminology and institutional arrangements, the literature converges on a small set of design principles that

consistently shape EA outcomes. First, statistical balance refers to meeting target workload or population thresholds to enable efficient field operations and comparable sampling units. Second, spatial continuity and connectivity require zones to be contiguous and practically navigable; fragmentation typically increases field burden and complicates supervision and household listing. Third, operational feasibility requires boundaries to be interpretable and defensible, often by aligning them with observable linear features (e.g., roads, rivers, railways) and respecting administrative constraints when mandated. Fourth, temporal consistency and maintainability emphasise that EA systems should be updateable across census cycles, with transparent rules and QA/QC processes to avoid ad hoc boundary drift. Finally, confidentiality and disclosure control impose minimum-size and stability considerations for EA-like dissemination geographies, reinforcing the need to balance operational objectives with statistical governance requirements (Openshaw & Rao, 1995; Cockings et al., 2011; United Nations, 2017).

Against this background, Section 2.1.1 sets out the working definition of Enumeration Areas adopted in this thesis.

2.1.1. Definition of Enumeration Areas

Enumeration Areas (EAs) are geographically defined spatial units used as the primary operational framework for population censuses, household surveys, and sample-based statistical data collection. An EA is generally defined as a contiguous geographic area that can be completely enumerated by a single field enumerator within a specified period of time, under predefined operational constraints (United Nations, 2007; United Nations Statistics Division, 2009).

The fundamental purpose of defining EAs is to organize field operations in a systematic and manageable manner. By dividing larger administrative or geographic regions into smaller units, EAs ensure complete coverage of the population while minimizing the risk of omission or duplication. In addition to their operational role, EAs also function as the basic spatial building blocks for statistical aggregation, data dissemination, and sampling frame construction.

Although the general definition of EAs is widely accepted, their exact characteristics may vary across countries and institutional contexts. National statistical offices determine EA boundaries based on a combination of population thresholds, administrative considerations, geographic features, and fieldwork logistics. As a result, EAs may be referred to by different names, such as enumeration districts, census tracts, output areas, or dissemination areas, while serving functionally similar purposes (United Nations Statistics Division, 2017).

Importantly, EAs are not purely administrative constructs. Rather, they represent an interface between statistical theory, spatial organization, and practical field implementation. This multidimensional nature makes EA design a complex task that extends beyond simple cartographic subdivision.

Beyond their basic definition, Enumeration Areas can be understood as interface units that connect statistical requirements, spatial organization, and field operations. From a statistical perspective, EAs must support accurate population counts, reliable sampling frames, and consistent aggregation for data dissemination. From a spatial perspective, they must be contiguous, interpretable, and compatible with the physical structure of settlements. From an operational perspective, EAs must be feasible for field staff to enumerate within a limited time frame, taking into account accessibility, workload balance, and supervision requirements.

This multidimensional role distinguishes Enumeration Areas from purely administrative or cartographic units. While administrative boundaries are primarily designed for governance and legal purposes, and cartographic units focus on spatial representation, EAs are explicitly designed to serve statistical production processes. As a result, EA design involves trade-offs between competing objectives, such as population homogeneity versus spatial compactness, or operational simplicity versus long-term statistical consistency.

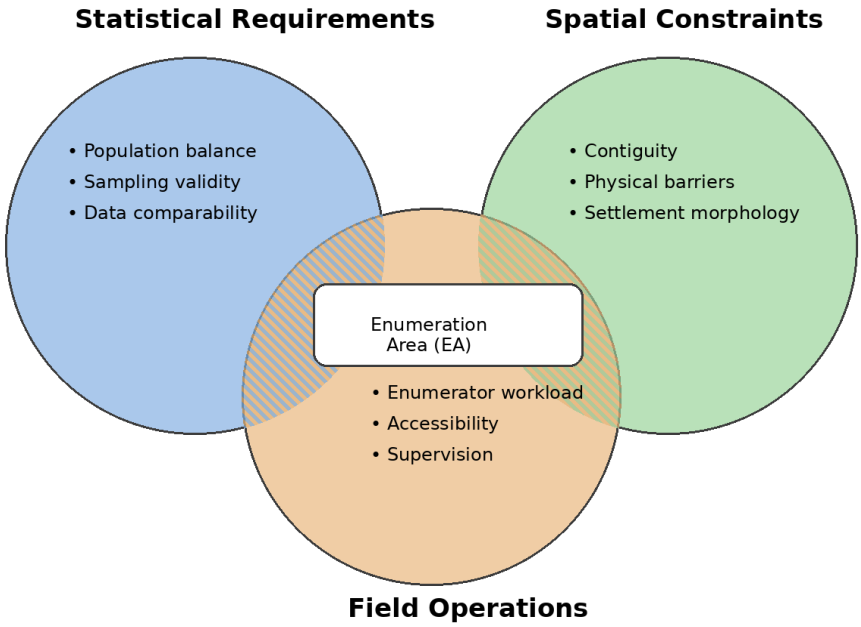
The literature consistently emphasizes that these trade-offs cannot be resolved through a single criterion or design rule. Instead, EA delineation requires a balanced consideration of multiple constraints, many of which may conflict depending on local context. For example, maintaining alignment with administrative boundaries may improve institutional integration, but it may also result in uneven population

distribution across EAs. Similarly, enforcing strict population thresholds may improve workload balance but lead to irregular or fragmented spatial units.

Recognizing Enumeration Areas as hybrid units rather than purely technical constructs has important methodological implications. It explains why EA design remains sensitive to context, why full automation is difficult to achieve in practice, and why many national statistical offices rely on semi-automated or expert-guided approaches. This conceptualization also provides a foundation for understanding later discussions on GIS-based methods, spatial representation choices, and automation levels in EA production.

The definition of Enumeration Areas extends beyond a purely operational or cartographic description. As discussed above, EAs simultaneously respond to statistical requirements, spatial constraints, and fieldwork considerations. This multidimensional role can be more clearly understood through a conceptual representation that highlights the intersection of these domains.

Figure 2.1. Conceptual role of Enumeration Areas as an interface between statistical requirements, spatial constraints, and field operations



Source: Author's conceptualization based on United Nations (2007) and United Nations Statistics Division (2009).

Figure 2.1 illustrates the conceptual role of Enumeration Areas as hybrid units formed at the intersection of statistical, spatial, and operational considerations. Statistical requirements define population balance and sampling validity, spatial constraints impose contiguity and physical boundary/conditions, and field operations determine feasibility in terms of workload and accessibility. The Enumeration Area emerges from the integration of these dimensions, rather than from any single criterion in isolation. This conceptualization explains why EA design involves trade-offs and why context-sensitive and semi-automated approaches are commonly adopted in practice.

This section has established Enumeration Areas as foundational yet inherently hybrid spatial units that integrate statistical requirements, spatial constraints, and field operations. By examining their definitions, types, importance, uses, and emerging dynamic forms, it has been shown that EA design is neither a purely technical nor a purely administrative task. This conceptual foundation provides the basis for the subsequent review of international practices and methodological approaches to EA production.

2.1.2. Types of Enumeration Areas

Enumeration Areas can be classified in several ways depending on their purpose, scale, and institutional use. In practice, multiple EA types often coexist within a single national statistical system.

2.1.2.1. Administrative Enumeration Areas

Administrative EAs are defined based on existing administrative boundaries such as provinces, districts, municipalities, wards, or villages. These areas align closely with governance structures and are commonly used to facilitate coordination between statistical agencies and administrative institutions. While administrative alignment simplifies data integration and reporting, it may not always produce statistically balanced or operationally optimal EAs, particularly in areas with uneven population distribution (United Nations Economic Commission for Europe, 2015).

2.1.2.2. Statistical Enumeration Areas

Statistical EAs are specifically designed for data collection and statistical analysis purposes. Their boundaries are defined independently of administrative divisions and are optimized to meet criteria such as population homogeneity, workload balance, and spatial compactness. Examples include census tracts in the United States, Output Areas in the United Kingdom, and Dissemination Areas in Canada. These units are widely used for census dissemination and small-area statistical analysis (Office for National Statistics, 2016; Statistics Canada, 2018).

2.1.2.3. Geographical Enumeration Areas

Geographical EAs are delineated based on physical or spatial features such as roads, rivers, railways, or natural barriers. These features provide easily recognizable boundaries for field staff and help ensure spatial continuity. Geographical considerations are particularly important in rural or topographically complex regions, where administrative boundaries may be unclear or impractical for field operations.

In addition to this functional classification, EAs may also be described by scale. Large units such as census tracts or statistical areas serve analytical and dissemination purposes, while smaller units such as blocks, meshblocks, or microdata areas support detailed population analysis and sampling. The choice of scale reflects a trade-off between statistical deail, confidentiality, and operational feasibility (Lohr, 2010).

While Enumeration Areas are commonly classified into administrative, statistical, and geographical types for analytical clarity, the literature emphasizes that these categories are rarely applied in a strictly isolated manner in operational settings. Instead, most national EA systems exhibit hybrid configurations, in which administrative boundaries, statistical criteria, and geographical features are combined according to local context, settlement structure, and institutional priorities (United Nations, 2007; United Nations Economic Commission for Europe, 2015).

Administrative Enumeration Areas provide clear advantages in terms of governance alignment and institutional coordination. Their compatibility with existing administrative units facilitates data integration across government agencies and

supports policy implementation at different territorial levels. However, several studies note that reliance on administrative boundaries alone often results in substantial variation in population size and workload, particularly in rapidly urbanizing cities or sparsely populated rural regions (United Nations Statistics Division, 2017; Martin et al., 2001). Consequently, purely administrative EAs are frequently adjusted using additional statistical or spatial constraints.

Statistical Enumeration Areas are primarily designed to optimize sampling efficiency, population homogeneity, and analytical comparability over time. Units such as census tracts, Output Areas, and Dissemination Areas exemplify this approach. While these units support robust statistical inference, the literature also reports operational challenges, including reduced spatial legibility for field staff and limited alignment with locally recognized boundaries (Office for National Statistics, 2016; Statistics Canada, 2018). This tension highlights the need to balance statistical objectives with practical field considerations.

Geographical Enumeration Areas emphasize the use of physical and infrastructural features—such as roads, rivers, railways, and natural barriers—to define EA boundaries. These features provide visible and easily interpretable limits that support field navigation and supervision. International census guidelines consistently recommend the use of recognizable physical boundaries to reduce ambiguity during enumeration (United Nations, 2007; United Nations Statistics Division, 2009). However, exclusive reliance on geographical features may lead to EAs with uneven population distribution, especially in heterogeneous urban environments.

The coexistence of administrative, statistical, and geographical EA types suggests that EA design is context-dependent. International census guidance emphasizes that small-area units should be delineated clearly enough for consistent use, while the literature on the Modifiable Areal Unit Problem shows that statistical results may vary according to the scale and configuration of areal units. Taken together, these considerations indicate that urban, peri-urban, and rural areas may require different trade-offs among administrative coherence, statistical balance, and spatial legibility (United Nations Economic Commission for Europe, 2015;

Flowerdew, 2011). As a result, effective EA systems typically adopt flexible and adaptive rule sets rather than rigid classification schemes.

This hybrid and context-sensitive understanding of EA types directly informs the methodological discussions that follow in this chapter. It explains why uniform design rules frequently fail in practice and why GIS-supported, constraint-based approaches are increasingly favored in contemporary EA production (Longley et al., 2015). The next sections build on this perspective by examining how different methodological frameworks operationalize these hybrid design principles.

2.1.3. Importance of Enumeration Areas

Enumeration Areas play a critical role in ensuring the accuracy, efficiency, and reliability of census and survey operations. Their importance extends across several dimensions.

First, EAs are essential for operational efficiency. By assigning a manageable workload to each enumerator, EAs enable systematic fieldwork planning, supervision, and quality control. Well-designed EAs reduce fieldwork duration, minimize travel time, and lower operational costs (United Nations, 2007).

Second, EAs are central to data quality and coverage. Poorly designed EAs may result in under-coverage, over-coverage, or inconsistent enumeration practices. Since EAs define the spatial scope of data collection, their configuration directly affects non-sampling error, response rates, and the reliability of population counts (Groves et al., 2009).

Third, EAs serve as the foundation for statistical sampling frames. In household and labor force surveys, EAs are commonly used as primary sampling units. Their stability and internal homogeneity are therefore critical for ensuring known selection probabilities and unbiased estimation (Kish, 1965; Lohr, 2010).

Finally, EAs support spatial analysis and policy-relevant statistics. By linking demographic attributes to geographic units, EAs enable the analysis of spatial patterns related to health, education, poverty, employment, and infrastructure. This makes them indispensable tools for evidence-based policymaking and regional planning (United Nations Statistics Division, 2009).

The importance of Enumeration Areas extends beyond their role as logistical units for census fieldwork. In the literature, EAs are consistently recognized as a central determinant of data quality, sampling validity, and the analytical usefulness of census and survey outputs. Their design directly affects both operational performance and statistical reliability across multiple stages of the data production process.

From an operational perspective, well-designed EAs enable effective planning, supervision, and quality control of fieldwork activities. By ensuring that each EA represents a manageable workload, statistical agencies can allocate enumerators efficiently, reduce travel time, and minimize fieldwork costs. International census guidelines emphasize that excessive variation in EA size or population leads to uneven workloads, increased error rates, and reduced enumeration completeness (United Nations, 2007; United Nations Economic Commission for Europe, 2015).

From a data quality perspective, EAs play a critical role in controlling coverage error and non-sampling error. Poorly delineated EAs increase the risk of household omission, duplication, or inconsistent enumeration practices. Since enumeration is conducted at the EA level, any structural deficiencies in EA boundaries propagate directly into census results. Empirical studies in survey methodology show that clear, contiguous, and interpretable EA boundaries improve response rates and reduce interviewer-related error (Groves et al., 2009).

Enumeration Areas are also fundamental to the construction of sampling frames. In most household and labor force surveys, EAs serve as primary sampling units (PSUs). The statistical properties of these units—such as population homogeneity, stability over time, and spatial coherence—directly influence selection probabilities and variance estimation. Inadequate EA design can therefore introduce bias and undermine the representativeness of survey estimates, even when sampling procedures are correctly implemented (Kish, 1965; Lohr, 2010).

Beyond data collection, EAs are essential for statistical dissemination and spatial analysis. By linking demographic and socioeconomic attributes to geographic units, EAs enable the production of small-area statistics that support evidence-based policymaking. Indicators related to health, education, poverty, employment, and infrastructure are frequently analyzed at the EA level to identify spatial disparities and target interventions (United Nations Statistics Division, 2009; World Health

Organization, 2018). The analytical value of such indicators depends heavily on the internal consistency and spatial logic of EA boundaries.

Finally, the importance of EAs is closely linked to temporal comparability. Stable EA systems allow statistical agencies to track demographic change over time and to compare census and survey results across different periods. Frequent or poorly documented changes to EA boundaries complicate longitudinal analysis and reduce the interpretability of trends. As a result, international guidelines stress the need to balance adaptability with boundary stability when updating EA systems (United Nations Statistics Division, 2017).

Taken together, these considerations demonstrate that Enumeration Areas are not neutral containers of data, but active components of statistical systems. Their design choices shape operational efficiency, data quality, analytical validity, and policy relevance. This centrality explains why EA delineation has become a key focus of methodological research and why increasing attention is being paid to systematic, transparent, and reproducible EA production approaches.

2.1.4. Uses of Enumeration Areas

Enumeration Areas are used across a wide range of statistical, administrative, and analytical applications. While their primary function is to support census and survey field operations, their utility extends well beyond data collection, forming the spatial foundation of modern statistical systems.

One of the most important uses of Enumeration Areas is in demographic and socioeconomic analysis. EA-level data allow researchers to examine population structure, household composition, migration patterns, and socioeconomic characteristics at a fine spatial scale. This level of detail enables the identification of local variations that are often masked at higher administrative levels and supports more nuanced interpretations of demographic trends (United Nations Statistics Division, 2009; Lohr, 2010).

Enumeration Areas also play a central role in survey sampling and design. In many household, labor force, and health surveys, EAs are used as primary sampling units (PSUs). Their spatial definition ensures complete coverage of the target

population and allows samples to be selected with known probabilities. Well-designed EAs facilitate efficient cluster sampling, reduce design effects, and improve the precision of survey estimates (Kish, 1965; Groves et al., 2009).

In the field of urban and regional planning, EA-based statistics support evidence-based decision-making related to land use, housing, transportation, and public service provision. Planners and local authorities rely on EA-level indicators to identify areas experiencing rapid growth, decline, or socioeconomic disadvantage. By providing spatially detailed and standardized data, EAs enable comparisons across neighborhoods and support targeted policy interventions (United Nations Economic Commission for Europe, 2015).

Public health applications constitute another major area of EA use. Health indicators such as disease prevalence, access to healthcare services, vaccination coverage, and environmental exposure are frequently analyzed at the EA level. This spatial granularity allows public health authorities to identify high-risk populations, allocate resources more effectively, and evaluate the impact of interventions. The integration of EA-based data with GIS further enhances spatial epidemiological analysis (World Health Organization, 2018).

Enumeration Areas are also increasingly used in market research and private-sector analysis. Businesses and service providers analyze EA-level demographic and income data to understand consumer behavior, define market segments, and inform location-based strategies. The standardized nature of EAs allows private-sector analyses to align with official statistics, improving comparability and reliability (Esri, 2020).

Beyond these sector-specific applications, EAs serve as a critical link between statistical data and geographic information systems. By providing stable spatial units, EAs enable the integration of census data with environmental, infrastructural, and administrative datasets. This interoperability supports advanced spatial analysis, visualization, and modeling, expanding the analytical value of official statistics (Longley et al., 2015).

Taken together, these applications demonstrate that Enumeration Areas function not only as operational tools but also as key analytical units that underpin a wide range of data-driven activities. Their versatility and centrality explain why EA

design decisions have far-reaching implications across statistical production, research, and policy domains.

2.1.5. Enumeration Areas as Dynamic Spatial Units

Traditionally, Enumeration Areas have been treated as relatively static spatial units, typically revised only in preparation for decennial censuses. This approach was largely driven by operational constraints, limited data availability, and the need for stability in fieldwork organization. However, rapid urbanization, population mobility, and the increasing use of administrative registers have challenged the adequacy of static EA systems in many national statistical contexts (United Nations, 2007; United Nations Statistics Division, 2017).

In response, the literature has increasingly discussed dynamic and adaptive EA systems, in which EA boundaries are periodically updated to reflect demographic change, urban expansion, and evolving settlement patterns. Dynamic EAs aim to reduce coverage error and workload imbalance by adjusting boundaries in areas experiencing significant population growth or decline. Such systems are particularly relevant in fast-growing metropolitan regions, where static EA boundaries may quickly become outdated (United Nations Economic Commission for Europe, 2015).

Several countries have explored adaptive EA frameworks supported by GIS and register-based data infrastructures. For example, register-integrated statistical systems enable continuous monitoring of population distribution, allowing EA boundaries to be revised more frequently and with greater precision (Statistics Estonia, 2021). These approaches reduce reliance on large-scale pre-census boundary redesigns and support more flexible fieldwork planning.

Despite their advantages, dynamic EA systems introduce important methodological and institutional challenges. Frequent boundary changes may complicate longitudinal analysis, reduce comparability across census rounds, and increase the complexity of sampling frame maintenance. The literature therefore emphasizes the need to balance adaptability with boundary stability, often recommending incremental updates focused on high-change areas rather than wholesale redesigns (United Nations Statistics Division, 2017; Lohr, 2010).

As a result, most national statistical offices currently adopt hybrid approaches, combining stable EA frameworks with targeted updates informed by demographic indicators, administrative registers, or spatial analysis. These hybrid models reflect a broader consensus that EA systems must be both robust and responsive, accommodating change without undermining statistical continuity.

The discussion of dynamic and adaptive Enumeration Areas highlights the growing importance of systematic, transparent, and reproducible EA production methods. As EA systems become more responsive to change, the role of GIS-supported and algorithmic approaches becomes increasingly central. This perspective provides a direct conceptual bridge to the methodological frameworks reviewed in the following sections of this chapter.

2.2. International Practices in Enumeration Area Design and Use

Enumeration Areas are implemented through a wide variety of institutional, spatial, and technical arrangements across countries. While their core purpose—to support census and survey field operations—remains broadly consistent, international practice demonstrates substantial variation in how these units are defined, delineated, and maintained. Differences arise from administrative structures, population distribution, legal frameworks, and the maturity of geospatial and statistical infrastructures. Consequently, Enumeration Areas should be understood not as a single standardized concept but as a family of functionally equivalent yet structurally diverse spatial units (United Nations, 2017).

In this thesis, the term “enumeration area” is used in an inclusive sense that covers both (a) operational enumeration units created to manage field workload, navigation, and supervision and (b) EA-like small-area statistical geographies that function as stable building blocks for dissemination, sampling frames, and longitudinal analysis. Although these units may differ in legal status or naming conventions (e.g., Output Areas, Dissemination Areas, IRIS, DeSO, microzones), they are comparable in the specific design dimensions that matter for census and survey implementation: workload balancing, contiguity, recognisable boundaries, confidentiality, updateability, and the ability to link microdata or registers to spatial units.

The country examples are drawn primarily from the most recent official manuals, methodological notes, and metadata available in the provided sources. Where only foundational or historical documentation is available, it is used to explain how current systems evolved from earlier approaches (e.g., transitions from legacy sketch mapping to GIS-based boundary maintenance, or from operational EAs to separate dissemination geographies).

The level of detail varies across countries because documentation depth differs by statistical system and by the purpose of the source (e.g., census operations memoranda, survey sampling designs, geocoding technical notes, or academic applications). This variation is not treated as a weakness; instead, it reflects how EA practice is institutionalised differently across contexts. The synthesis focuses on extracting transferable design principles rather than producing uniform country profiles.

To support comparability without forcing artificial uniformity, the country examples are interpreted through a common set of dimensions. Specifically, the discussion highlights (i) the primary building blocks used to define or maintain EAs (e.g., administrative units, address points, census blocks, grid cells, or settlement features), (ii) workload or size targets and any explicit threshold rules, (iii) boundary principles and recognisable features used for fieldwork (e.g., roads, rivers, neighbourhood limits, or settlement edges), (iv) the maintenance and update regime (continuous register updating versus census-cycle revision), and (v) the degree of automation and QA/QC used to ensure reproducibility. Using this lens makes the cross-country synthesis more transparent and directly relevant for the methodological positioning developed in Sections 2.3–2.5.

National statistical systems implement enumeration areas (EAs) under different institutional and technical conditions; therefore, the operational form of an EA varies across countries. In most settings, EAs serve two closely related functions: (i) they provide manageable workloads for field enumeration or listing, and (ii) they act as primary sampling units (PSUs) or small-area reporting units that support recurring household surveys and intercensal estimates. Recent practice increasingly integrates EA design with geospatial infrastructures (GIS, geocoding, satellite imagery, and administrative registers), which reduces reliance on fully manual sketch

mapping while improving boundary consistency, coverage control, and linkages between statistical and spatial datasets (UNICEF, 2013; Balistreri & Cozzi, 2015; Qader et al., 2021).

For clarity, it is useful to distinguish between (i) a GIS-based approach and (ii) an automated or semi-automated delineation approach. A GIS-based approach primarily describes the use of geospatial databases, mapping tools, and cartographic workflows to manage and update EA boundaries, support field navigation, and link census or survey attributes to small-area units. In contrast, automated or semi-automated delineation refers to the degree of algorithmic decision-making during boundary construction and workload balancing (e.g., rule-based splitting/merging, constrained optimization, or segmentation). These approaches are not mutually exclusive: automated procedures can be implemented inside GIS environments or in external scripting workflows, and they can be applied to both polygon-based and cell-based representations, depending on the available base data and the operational requirements of the census or survey programme (UNICEF, 2013; Qader et al., 2019; United Nations, 2017).

Several countries have recently strengthened the "GIS-first" management of small-area geographies by upgrading census cartography into reusable geospatial layers that support multiple statistical domains and inter-agency data integration. Nepal provides a clear example of an incremental transition. A Nepal case study reports that prior to 1991, censuses did not consistently adhere to specific boundaries of census units, while cartographic work for the 1991 census remained limited in scale and in its linkage to census attributes; the same source notes that the 2011 census used detailed EA maps for the first time in the country's census history, with EA mapping activities starting three to four years before the enumeration year (Central Bureau of Statistics Nepal, 2022). The Nepal experience highlights two practical improvements that accompany GIS-based EA mapping: stronger control of omission and duplication through location-specific referencing, and the ability to link household-level attributes to mapped units for later analysis and dissemination (Central Bureau of Statistics Nepal, 2022). In Italy, a different but complementary trajectory is seen in ISTAT's development of a "microzones" layer as an upgrade of 2011 census cartography. Mugnoli, Lipizzi, and Esposto (2018) describe the integration of geographic datasets

to produce this layer through the management and fusion of land-cover and cadastral-like sources with maps produced by regional or local authorities, emphasizing that the work is ongoing and under assessment. Together, these cases show that "updated cartography" is not only a census deliverable but also a strategic asset for broader official statistics.

In the United States, census geography is organized around Census Blocks and Block Groups, which together perform functions analogous to Enumeration Areas. Census Blocks constitute the smallest geographic units, while Block Groups aggregate blocks into operational units typically containing between 600 and 3,000 people. Block Group boundaries are defined using automated GIS procedures based on population size and contiguity, but manual adjustments are routinely applied to account for physical barriers, urban morphology, and administrative considerations. This practice reflects a balance between national standardization and local adaptation within a highly centralized statistical system (U.S. Census Bureau, 2020; U.S. Census Bureau, 2022).

In the United States, the operational geography for the 2010 census illustrates how enumeration logistics can be embedded in geographic delineation decisions. The Type of Enumeration Area (TEA) framework classifies geographic space into areas that receive different enumeration methods (e.g., mail-out/mail-back versus update/leave), reflecting the feasibility of mail delivery, housing stock characteristics, and field requirements. In this sense, the operational objective is not only to partition space, but also to standardize how enumeration is executed across different contexts, while maintaining internal consistency for planning and coverage evaluation (U.S. Census Bureau, 2011).

In the United Kingdom, census operations rely on a highly formalized statistical geography in which Output Areas serve as the smallest standard units for census data dissemination. Introduced with the 2001 Census, Output Areas were explicitly designed as statistical rather than administrative units, with the objectives of population uniformity, social homogeneity, and boundary stability. Typical Output Areas contain approximately 100–125 households and are constructed by aggregating base spatial units while respecting visible physical features such as roads and rivers. Their delineation is strongly GIS-based and combines automated zoning algorithms

with expert-led manual review to resolve local anomalies and ensure operational clarity (Office for National Statistics, 2016; Office for National Statistics, 2021).

In Canada, Dissemination Areas represent the smallest standard geographic unit for census data dissemination and function as direct equivalents of Enumeration Areas. Dissemination Areas typically contain populations between 400 and 700 persons and are explicitly designed to remain stable across census cycles in order to support longitudinal analysis. Delineation relies on GIS-based aggregation of census blocks using population thresholds, contiguity constraints, and physical boundaries. Manual validation is a core component of the process, particularly in rural and northern regions where settlement patterns are highly irregular. As a result, Dissemination Areas vary substantially in size and shape, illustrating that spatial heterogeneity is an accepted design characteristic rather than a limitation (Statistics Canada, 2018; Statistics Canada, 2021).

Canada provides a complementary perspective where the definition of small-area census units has been linked to the long-term maintenance of digital boundary systems and the re-use of census geography for analysis. Historical work on creating digital census boundary maps documents how earlier enumeration-area boundaries were reconstructed from legacy materials and converted into usable GIS boundary layers for spatial analysis and longitudinal comparison (Petrov & Ruus, 2007). In applied public health research, census small-area units have also been used as contextual socioeconomic status (SES) measures by linking census geography to postal code-based records, enabling the creation of neighborhood-level indicators for epidemiological studies (Statistics Canada, 2021; Borugian et al., 2005).

In Australia, the Australian Statistical Geography Standard (ASGS) formalizes a nested hierarchy in which Statistical Areas Level 1 (SA1s) are built from whole Mesh Blocks and aggregate to SA2s. SA1s are designed to maximize the geographic detail available for census dissemination while also meeting operational and representational criteria, such as population size, transport connectivity, and anticipated growth. Most SA1s have populations of between 200 and 800 persons, with an average of approximately 400. The standard also includes non-spatial special purpose codes for populations that are difficult to allocate geographically, as well as zero-population

SA1s for large unpopulated areas that cannot be readily combined with surrounding populated SA1s (Australian Bureau of Statistics, 2021).

In Japan, census enumeration districts are designed explicitly as operational field units and are typically delineated along clearly interpretable physical features such as streets, railways, and waterways. The emphasis on compactness, boundary readability, and workload manageability illustrates how dense urban settings can still be partitioned into practical micro-areas that support both census operations and repeated household surveys without creating ambiguous field boundaries.

In France, the IRIS (Îlots Regroupés pour l'Information Statistique) system provides the basic building block for the dissemination of infra-municipal statistics. IRIS units were originally designed around a target size of 2,000 inhabitants per elementary unit, and they are required to meet both geographic and demographic criteria, with boundaries that are identifiable without ambiguity and stable over time. For residential IRIS, boundaries generally follow major breaks in the urban fabric, such as main roads, railways, and waterways (Institut national de la statistique et des études économiques [INSEE], 2016).

Several European countries operate Enumeration Area–like units within register-based statistical systems. In Germany, census and survey operations are closely linked to population registers and address databases, enabling flexible aggregation of address-level data into operational units. Rather than maintaining permanently fixed Enumeration Areas, German practice emphasizes adaptability, with spatial units adjusted as settlement structures and population distributions change. GIS plays a central role in managing these aggregations, ensuring spatial contiguity and operational feasibility when field enumeration is required (United Nations Statistics Division, 2011; Eurostat, 2017).

In Germany, the 2022 census round strongly emphasized administrative data integration and standardization. Register-based enumeration (supported by targeted surveys) reduces the need for full-field enumeration, and reporting frameworks increasingly include grid-based statistics that are less sensitive to administrative boundary changes (Statistisches Bundesamt [Destatis], n.d.; Eurostat, n.d.).

A similar but more data-integrated approach is observed in the Netherlands, where detailed address and building registers underpin statistical geography.

Enumeration functions are fulfilled through dynamic aggregation of micro-level units rather than reliance on fixed EA boundaries. This approach allows boundaries to be adapted for specific census or survey purposes while preserving statistical consistency and confidentiality. The Dutch case demonstrates how advanced data infrastructures can reduce dependence on permanent Enumeration Areas without eliminating their functional role (Statistics Netherlands, 2019; Eurostat, 2017).

In Nordic countries such as Sweden and Finland, Enumeration Area concepts are embedded within register-based census systems that rely heavily on geocoded address points and grid-based spatial frameworks. Although traditional field enumeration is limited, GIS remains essential for linking population registers to spatial units, supporting sampling design, and managing confidentiality. These systems illustrate an alternative pathway in which Enumeration Area functions persist even as their geometric representations evolve (Statistics Sweden, 2018; Statistics Finland, 2019).

In Sweden, Demografiska statistikområden (DeSO) were introduced as a national small-area division for demographic and socioeconomic analysis. DeSO divides Sweden into 6,160 areas, most designed to fall within a population range of 700 to 3,000 inhabitants, and boundaries are designed to follow physical features such as streets, watercourses, and railways where feasible (Statistics Sweden [SCB], n.d.).

In Estonia, a register-based census model has been developed around secure data exchange and register quality improvement, with the goal of reducing reliance on full field enumeration. Recent metadata describe approaches that combine multiple administrative registers and an index-based logic for defining the usual resident population, supported by national data exchange infrastructure (Eurostat, n.d.).

In Spain, census sections (secciones censales) remain a core small-area unit in official statistics and survey sampling. They are used by the National Statistics Institute whenever an infra-municipal division is required, and in the Labour Force Survey they serve as first-stage sampling units. Their continued use makes stable maintenance and consistent linkage to address-based and population-register information essential for updating and analysis (Instituto Nacional de Estadística [INE], n.d.).

In South Africa, the Small Area Layer (SAL) supports small-area statistics and operational planning, with boundaries intended to be interpretable, stable, and consistent with settlement structure. Recent information products link SAL maintenance to dwelling-frame updates and broader statistical reporting needs (Statistics South Africa, n.d.).

Taken together, the mature systems reviewed above tend to exhibit three operational regularities. First, they rely on stable identifiers and nested hierarchies that support both dissemination and operational planning. Second, even where register-based approaches reduce reliance on full-field enumeration, countries still emphasise boundary legibility and settlement-aware design because these characteristics support survey operations, quality assurance, and interpretability in small-area analysis. Third, modernisation efforts increasingly formalise update rules and metadata management (e.g., versioning, change logs, and validation reports), which strengthens temporal consistency across cycles.

Italy demonstrates a shift from EA use as a purely field-operational construct toward a broader geostatistical infrastructure. ISTAT has developed a "microzones" layer by integrating geographic datasets to support land cover statistics and other spatial analyses at a fine scale, thereby providing a structured small-area geography for statistics beyond the census (Mugnoli et al., 2018). In parallel, ISTAT's geocoding work for the 9th Industry and Services Census describes how address archives and GIS packages were integrated to georeference enterprise records and to associate recognized addresses with census enumeration areas, supporting both territorial analysis and the updating of census EA characteristics (Balistreri & Cozzi, 2015).

Beyond census cartography, some systems use EA-like units as an explicit integration layer between registers and census/survey operations. In Slovenia, register and census sources were merged at the EA level for national coverage, and EA-level datasets were enriched with geographic and operational proxies (e.g., building centroids, altitude) to support geodemographic stratification and nonresponse modelling at the small-area scale (Zaletel & Vehovar, 2000).

In Brazil, the Instituto Brasileiro de Geografia e Estatística (IBGE) uses census tracts (setores censitários) as foundational enumeration and reporting units, and these units also support specialized modules such as the survey of the physical surroundings

of households. The 2022 census cycle continued this tract-based approach supported by updated digital cartography and boundary products (Instituto Brasileiro de Geografia e Estatística, 2024).

In Brazil, census enumeration areas remain central to probability sampling for large-scale household surveys. An explicit example is the use of the 2010 census enumeration area frame to design complex multi-stage household samples, with attention to design effects and intra-class correlation derived from census-based frames. This approach highlights the role of EAs as stable PSUs that can be re-used across survey programs while controlling for clustering and fieldwork cost (Arantes & Silva, 2013).

Similarly, multi-stage survey programs in other middle-income contexts adopt census EAs as basic units in national sampling frames. For instance, Mongolia's Labour Force Survey documentation describes the use of the Population Census 2000 frame, where enumeration areas are treated as basic units but may require cleaning (e.g., removing EAs associated with institutional living quarters or EAs without households). This illustrates that even when EA boundaries exist, operational quality assurance and frame maintenance are necessary to ensure that EAs function as valid sampling units (National Statistical Office of Mongolia, n.d.).

International guidance documents also standardize key terminology and practical steps that connect these country cases. The MICS Manual defines an EA as the smallest geographical statistical unit for which census information is available and clarifies that a survey "cluster" may correspond to an entire EA or to a segment of a large EA with well-defined boundaries (UNICEF, 2013). The same manual introduces the base map as the reference map describing the location and boundaries of an EA and highlights segmentation as a pragmatic solution when complete household listing in large EAs is not cost-effective (UNICEF, 2013). These definitions are important because they make explicit that "EA" is not a fixed geometric form; rather, it is a field-operational construct that can be subdivided, merged, or segmented depending on workload targets, boundary legibility, and available resources.

International survey guidance further clarifies how EA-type clusters are operationalized under standardized field protocols. The MICS Manual for Mapping and Household Listing emphasizes that a listing operation typically updates cluster

maps (including location and sketch mapping) and records every structure and household on listing forms, because listing quality directly affects representativeness and survey cost (UNICEF, 2013). In post-enumeration survey (PES) settings, field manuals similarly rely on EA maps and systematic listing/verification to assess coverage and to validate the census, demonstrating the importance of consistent EA boundaries and enumerator training for quality control (Ghana Statistical Service, 2021).

Low- and lower-middle-income contexts often emphasize the operational role of EAs through detailed map use, listing, and field navigation procedures. Ghana's 2021 Population and Housing Census Post-Enumeration Survey (PES) manual positions the EA map as the most important aid for achieving complete enumeration within selected EAs and stresses that omissions of households and dwellings directly affect PES results (Ghana Statistical Service, 2021). The manual also provides a practitioner-oriented classification of EA maps and notes that, regardless of type, maps typically display locality identifiers, higher-level administrative names and codes, and supervision area and EA identifiers (Ghana Statistical Service, 2021). Such procedural detail demonstrates how EA cartography is operationalized in the field: it is used to identify boundaries, list structures, plan systematic coverage, and organize logistics and supervision (Ghana Statistical Service, 2021). Nigeria-oriented applied work similarly frames EAs as operational entities that benefit from digital geodatabases; a Lagos case study describes the development of a census enumeration information system intended to manage census data in a consistent digital database, and it demonstrates how GIS and network analysis can support the efficient routing and organization of enumeration activities (Amusa et al., 2017). While these contexts may differ in institutional maturity, they share a common emphasis on EA maps as tools for field control, coverage assurance, and cost containment.

In India, enumeration blocks (EBs) are the main operational units for household listing and enumeration. Official census mapping materials show that EB-based field organisation is closely tied to administrative atlases and mapped coverage units designed to support complete and non-overlapping field operations. This illustrates a strongly operational approach in which mapped control units are integral

to census implementation (Office of the Registrar General & Census Commissioner, India, 2012).

In developing and transitional contexts, Enumeration Area practices are often shaped by constraints related to data availability, cartographic quality, and field logistics. In countries such as India and South Africa, Enumeration Areas are primarily delineated to support decennial census operations, with strong emphasis on balancing enumerator workload and accessibility. GIS adoption has increased substantially, but manual and semi-automated methods remain prevalent due to heterogeneous settlement patterns and informal housing structures. As a result, Enumeration Areas are frequently redesigned between census rounds, prioritizing operational feasibility over long-term boundary stability (United Nations, 2017; Statistics South Africa, 2014).

Low-income and fragile contexts often face incomplete, outdated, or missing EA demarcations; this drives innovation in GIS-assisted and semi-automated delineation. In Somalia, a semi-automatic workflow has been proposed to generate pre-census EAs and population sampling frames by combining satellite imagery with road networks, settlement information, and available administrative boundaries. The study explicitly frames EAs as both operational units for census data collection and as building blocks for national sampling frames, arguing that semi-automation can reduce cost and time while improving coverage in contexts where manual delineation would be slow and inconsistent (Qader et al., 2021).

Fragile and data-scarce settings highlight another strand of recent practice: using openly available geospatial data and population models to (re)construct small-area units when conventional census cartography is missing, outdated, or infeasible to update through large-scale field mapping. In Somalia, Qader et al. (2019) describe an automated approach that combines georeferenced features (including crowdsourced road and natural feature data from OpenStreetMap) with high-resolution population models (e.g., WorldPop) to design a complete set of pre-defined census EAs or to update existing EAs. The authors term their automated process “split and merge,” inspired by image segmentation in mathematical morphology, where the country is first split into the smallest possible regions that follow visible boundaries (such as roads) and is then progressively re-merged under explicit constraints (e.g., area and

population size) while ensuring that EA boundaries do not cross obstacles and administrative boundaries (Qader et al., 2019). From a conceptual perspective, this case illustrates the growing feasibility of automated EA reconstruction in contexts where official base data are limited, and it directly motivates the thesis emphasis on transparent constraints, barrier-respecting boundaries, and reproducible workflows.

Nigeria-based case studies also emphasize the role of remote sensing and GIS for census planning and EA management, particularly in settings where traditional mapping and boundary verification are resource-intensive. One documented application mapped EAs using satellite remote sensing data and GIS in Enugu, with the objective of improving census mapping, dataset management, and contingency planning (Eze, 2009). A related "Census Enumeration Information System" case study in Lagos State presents an EA information system aimed at reducing under-enumeration and improving management and retrieval of census data, showing how geospatial information systems can support both field operations and post-collection data use (Amusa et al., 2017).

In Nigeria, the literature documents a sustained shift from analogue mapping towards remote sensing and GIS-based census mapping. Case studies report the use of high-resolution imagery and digitization to build geodemographic databases for managing enumeration areas and supervisory areas, and they argue that GIS-based EA databases improve updateability and transparency for national census management (Eze, 2009; Amusa et al., 2017).

In Kenya, recent census rounds have relied on digital cartography and extensive GIS workflows for EA delineation and field operations. The mapping process typically combines administrative layers, satellite imagery, and GNSS-supported ground verification to produce navigation-ready EA maps and to manage coverage and workload more systematically (Kenya National Bureau of Statistics [KNBS], 2019).

In lower-capacity and data-scarce contexts, the same principles are pursued under different constraints. Documentation frequently emphasises the sequencing of mapping, listing, and enumeration; the use of GNSS-enabled field verification; and pragmatic reliance on remotely sensed settlement information and open geospatial layers. As a result, the "EA system" is often defined as much by the production

workflow and data integration strategy as by the final polygon boundaries, which reinforces the importance of auditable rules and standardised QA/QC outputs.

The international evidence also indicates that EA geographies are increasingly combined with gridded population approaches for small-area estimation and planning, especially when census data need to be integrated with land cover or remotely sensed settlement layers. In East Africa, high-resolution population mapping work demonstrates how census data can be combined with land cover information to produce more detailed population surfaces for low-income nations, illustrating a pathway where EA-based census information supports model-based spatial allocation and validation (Tatem et al., 2007).

Some national systems explicitly document the modernization of EA cartography as part of official statistical system development. For example, Nepal's Central Bureau of Statistics highlights geospatial data and enumeration-area mapping as a distinct topic in its official statistical practices, indicating an institutional recognition that EA mapping is not only a field task but also a core component of statistical infrastructure and data integration (Central Bureau of Statistics Nepal, 2022).

In Nepal, official statistical practice documents describe a progressive transition from paper or sketch-based maps toward digital census mapping, including converting sketch maps into GIS, collecting GPS points for key features, and conducting full mapping programs where resources permit (Central Bureau of Statistics Nepal, 2022).

Across these cases, three themes recur: (1) the separation between operational enumeration units and dissemination units (even when they share boundaries), (2) the increasing use of GIS/GNSS and digital boundaries to support consistent maintenance between rounds, and (3) the emergence of automated or semi-automated zoning approaches where conventional field mapping is costly or infeasible. These themes provide the comparative basis for positioning the proposed Türkiye-focused framework in later sections.

Overall, cross-country practice suggests that "EA implementation" should be understood as a spectrum rather than a single template. Some countries emphasize operational stratification (e.g., the TEA concept), others prioritize maintaining fine-

scale statistical geographies for analysis (e.g., microzones), and many programs rely on EAs primarily as PSUs for household surveys. In contexts where delineations are weak or missing, semi-automated workflows based on imagery and available transport/barrier layers can provide a pragmatic route to establishing an EA framework, but still require strong validation and field feedback to ensure usability for enumeration and sampling (U.S. Census Bureau, 2011; Mugnoli et al., 2018; Qader et al., 2021).

Overall, the reviewed practices show a clear technological trajectory from paper-based EA sketch mapping toward georeferenced, database-driven, and increasingly reusable small-area geographies. At the same time, the cases illustrate that the most appropriate EA design is context-dependent. Stable hierarchical geographies and register linkage support continuity and comparability (e.g., the U.S. and Slovenia), upgraded spatial layers improve multi-purpose integration (e.g., Italy), census-based frames enable efficient two-stage household sampling (e.g., Brazil and Mongolia), and detailed EA map procedures improve field control and coverage in census and PES contexts (e.g., Ghana). Finally, Somalia demonstrates that automation can be used not as a substitute for GIS-based practice but as a complementary strategy when open data and population models are leveraged under explicit constraints. This distinction supports the methodological positioning adopted in the next chapter: GIS-based implementation refers to the broader use of geospatial databases, tools, and cartographic workflows, while automated or semi-automated delineation refers to the degree of algorithmic decision-making in boundary construction and unit balancing, which can be applied both within GIS environments and in external scripting workflows.

These international practices motivate the conceptual distinction made in the next section between (i) the technical environment (GIS-enabled versus non-GIS), (ii) the chosen building blocks (administrative units, cells, or objects), and (iii) the automation strategy (manual, semi-automated, automated). Across countries, successful implementations typically combine elements from all three dimensions, which suggests that method evaluation should consider both statistical compliance and operational legibility rather than treating EA delineation as a purely geometric optimisation task.

In Türkiye, there is currently no publicly documented and standardised enumeration area (EA) layer used as a stable spatial framework for routine survey and census operations. Instead, household-based survey practice relies on operational clusters formed within existing settlement units, typically by grouping occupied addresses recorded in the address-based registration system. These clusters function as primary sampling units, but they are not defined as permanent, spatially explicit EA polygons.

In current practice, occupied addresses are grouped primarily according to address components such as street or avenue name, building number, and apartment number, and are then partitioned into clusters intended to contain approximately similar numbers of units. Because this procedure is not based on geographic coordinates, contiguity rules, or explicit barrier constraints, the resulting units do not constitute a fixed small-area geography and may vary over time as the sampling frame is updated and clusters are reconstructed. This limits transparency and makes it difficult to enforce design criteria such as workload balance, spatial coherence, and barrier compliance in a systematic and reproducible way. Figure 2.2 provides a schematic illustration of the existing non-spatial blocking output.

Figure 2.2. Schematic illustration of the existing non-spatial blocking method used in survey practice in Türkiye

Source: Author's illustration based on the current operational logic of address-based clustering.

An early academic effort nevertheless examined the feasibility of establishing a dedicated census geography for Türkiye using GIS-based delineation principles. In a case study conducted for Çankaya (Ankara), Kırlangıçoğlu (2005) explored a candidate hierarchy of standardised small statistical areas intended to support census dissemination and small-area statistics beyond existing administrative boundaries. However, this proposal remained methodological and was not adopted as an official operational system in routine statistical production.

Taken together, these observations indicate that Türkiye has a strong address-based statistical infrastructure but still lacks a stable, standardised, and spatially explicit EA framework. The automated approach developed in this thesis addresses this gap by operationalising widely used EA design principles through spatial layers, barrier-aware neighbourhood structure, and building-based residential counts.

To facilitate structured cross-country comparison, Table 2.1 provides an integrated synthesis of the country examples discussed above. Rather than repeating narrative detail, the table distils each example into (i) the unit terminology used in practice, (ii) the functional role of the unit, and (iii) the dominant data inputs and maintenance direction. This synthesis clarifies why EA implementation is best understood as a spectrum of institutional arrangements and technical workflows, even when the underlying objective—manageable fieldwork and reusable small-area statistical infrastructure—remains shared.

Table 2.1. Cross-country examples of EA implementation (integrated synthesis)

Setting / programme	Unit used (term)	Main function and use	Design basis / key inputs	Recent direction / note
United States (Census Bureau)	Census blocks; block groups; tracts; TEA	Statistical small-area hierarchy, operational enumeration framework, and survey/sampling-frame support.	Hierarchical feature-based geography; TIGER/Line; PSAP; address- and delivery-based operational classification.	Stable census geography combined with TEA-based operational strategy.
United Kingdom (ONS)	Output Areas (OAs); (S)OAs	Operational workload unit and small-area statistical/sampling building block.	OAs aggregated to higher units; digitally maintained boundaries and lookups.	Comparability across rounds with selective updates.
Canada (Statistics Canada)	Dissemination Areas (DAs); historical enumeration areas	Operational workload management, small-area dissemination, sampling-frame construction, and applied GIS reuse.	Dissemination blocks; digital boundary reconstruction; postal-code linkages.	Boundary updates and reuse of census geography in longitudinal GIS applications.
Australia (ABS)	SA1	Operational workload unit and small-area statistical/sampling building block.	Mesh Blocks; criteria include transport, growth, and special/zero SA1s.	Population-sized small areas designed for stable operational use.
France (INSEE)	IRIS	Operational workload unit and small-area statistical/sampling building block.	Micro-neighbourhood groupings designed for interpretable dissemination geography.	Strong emphasis on interpretability and stability.
Sweden (SCB)	DeSO	Operational workload unit and small-area statistical/sampling building block.	Regional divisions combined with physical features.	Periodic revisions while preserving feature-following boundaries.
Germany (Destatis)	Small-area outputs; grids	Small-area statistical production, integration, and confidentiality-oriented reporting.	Administrative and register sources; grid-based reporting in some outputs.	Register integration and standardized reporting structures.
Brazil (IBGE)	Census tracts / setores censitários / EA-based frame	Operational workload unit and small-area statistical/sampling building block for census and household survey sampling.	Tracts as operational/reporting base; digital cartography; census-based frame; design-effect planning.	Systematic use of tract/EA frames across survey programmes.
Estonia (Statistics Estonia)	Register-based outputs	Small-area statistical production with minimal field burden.	Administrative registers, spatial links, and secure data exchange.	Strong register-based integration rather than field enumeration.
Kenya (KNBS)	Enumeration Areas (EAs)	Operational workload unit and small-area statistical/sampling building block for census operations.	Administrative layers, imagery, and GNSS.	Digital cartography and navigation-ready EA maps.
South Africa (Stats SA)	SAL / EAs	Operational workload unit and small-area statistical/sampling building block.	Small-area layer combined with dwelling frames.	Stable small areas linked to updated dwelling-frame infrastructure.

Table 2.1. Cross-country examples of EA implementation (integrated synthesis) (continued)

Setting / programme	Unit used (term)	Main function and use	Design basis / key inputs	Recent direction / note
India (ORGI / Census)	Enumeration Blocks	Operational workload unit and small-area statistical/sampling building block.	Household listing units; GIS-supported mapping; urban growth management.	Explicit urban/rural household-based workload sizing.
Nigeria (NPC)	Enumeration Areas; Supervisory Areas	Operational workload control, census mapping, and small-area statistical/sampling support.	Digitized imagery, EA boundaries, GIS databases, and digital planning tools.	GIS-supported updating and management of EA frames.
Nepal (CBS)	EAs; census mapping units; detailed EAs	Census mapping, EA-level spatial linkage, and broader statistical infrastructure development.	Sketch maps evolving into GIS; GPS features; geospatial data; multi-year pre-census mapping.	Progressive digitization and formalization of EA mapping.
Somalia (pre-census / WB-supported applications)	Predefined EAs (automatic or semi-automatic)	Rapid EA creation, census mapping, and sampling-frame recovery in constrained contexts.	Gridded population, OpenStreetMap, imagery, roads/barriers, and settlements.	Automated delineation used where official demarcations are missing or outdated.
Slovenia	Enumeration Areas	Register-census integration, geodemographic stratification, and nonresponse modelling.	EA-level register linkage, including cost-related contextual variables.	EA linkage used for integrated statistical analysis.
Italy (ISTAT)	Microzones; census EAs linked via geocoding	Multi-purpose spatial layer for land-cover/statistical integration and fine-scale geostatistics.	Geodata fusion, cartography upgrading, integrated geodatasets, address archives, and GIS tools.	Expansion of geocoding infrastructure and microzone layers.
Mongolia (LFS 2002-2003)	Enumeration Areas as PSUs	Labour force survey sampling with domain estimation.	Two-stage probability sample based on the census frame.	Census-based EAs used directly as PSUs in labour-force sampling.
Ghana (PHC/PES 2021)	EAs with standardized map types	Operational control for complete enumeration and post-enumeration listing.	EA maps used as navigation and coverage tools; standardized identifiers.	Standardized EA map products for census and PES operations.
MICS (global guidance)	EAs; segments of large EAs	Standardized mapping, listing, and segmentation rules for surveys.	Base maps, segmentation of large EAs, and cluster definition rules.	Survey guidance centred on EA-based mapping and listing.

2.3. Conceptual Framework for EA Production and Positioning of the Proposed Method

Sections 2.1–2.2 established that Enumeration Areas (EAs) are both operational units for fieldwork management and, in many national systems, stable building blocks for statistical dissemination and sampling frames. This section builds a conceptual framework to classify EA production approaches in a way that is directly usable for methodological positioning in later chapters. The framework is designed to avoid common terminological conflations—particularly the tendency to treat "GIS-based" as synonymous with "automated"—and to make explicit the choices that drive EA boundary outcomes (United Nations, 2025; Cockings et al., 2011).

In Türkiye, the statistical system has benefited from strengthened address-based registers and continuous administrative updating of address information. However, the presence of address-based registers does not automatically imply a stable and workload-balanced operational micro-geography for field surveys. Where an explicit EA layer is not formally defined and maintained, survey operations may default to heterogeneous proxy units (administrative subdivisions or ad hoc partitions), limiting standardisation across programmes and over time. This thesis therefore treats EA delineation as a missing infrastructural layer that can be derived from available spatial inputs (buildings and barriers) while remaining compatible with register-based sampling frames.

Geographic Information Systems (GIS) refer to a set of concepts and tools for capturing, storing, managing, analysing, and visualising data that are referenced to locations on the Earth. In census and survey contexts, GIS typically supports (i) the maintenance of geospatial databases for administrative boundaries, buildings and addresses; (ii) cartographic production of enumeration maps; (iii) field logistics through recognisable boundaries and navigation aids; and (iv) integration of statistical attributes with small-area geometries (United Nations, 2025; UNICEF, 2013). Importantly, GIS is an enabling technical environment: it can support purely manual delineation (e.g., digitising sketch maps), rule-based semi-automation (e.g., scripted

splitting and merging), or fully automated zone design (e.g., optimisation under constraints).

Conceptually, EA production can be described along three orthogonal dimensions. First, the technical environment distinguishes GIS-supported workflows from non-GIS workflows. Non-GIS workflows may still use digital mapping artefacts, but they do not rely on a geodatabase and spatial analysis as the backbone of production. Second, the spatial representation distinguishes polygon-based approaches (using administrative units, parcels, blocks, or building footprints as building blocks) from cell-based approaches (using regular grids such as 100 m or 1 km cells, or fishnet segmentation). Third, the degree of automation distinguishes manual delineation, semi-automated delineation (where algorithms propose or adjust boundaries subject to human supervision), and automated delineation (where optimisation or algorithmic rules determine boundaries with limited discretionary editing) (Openshaw & Rao, 1995; Martin, 2001; Cockings et al., 2011).

For operational clarity, “manual”, “semi-automated”, and “automated” delineation are defined here by where boundary decisions are made. Manual delineation means that an analyst or mapping team directly draws or edits EA boundaries and resolves exceptions case-by-case, even if GIS is used for digitising and validating topology. Semi-automated delineation means that explicit rules or algorithms generate candidate splits/merges or propose boundary adjustments to meet constraints (e.g., target workload, contiguity, barrier adherence), while a human supervises parameters and intervenes for edge cases or quality assurance. Automated delineation means that an algorithm (typically heuristic or optimisation-based) determines final boundaries under a formally specified constraint set, with minimal discretionary editing beyond QA/QC. This decision-authority definition supports transparent comparison of methods across countries and software environments.

These three dimensions should not be nested as if one implies another. In particular, GIS-based workflows are not necessarily automated; they often formalise manual practice by improving data integration, map production, and reproducibility. Conversely, automation does not require a GIS interface: automated zone design can be implemented in external software and later imported into GIS for validation and map output. Similarly, cell-based representations are frequently implemented within

GIS because raster and grid operations are a standard part of modern GIS analysis; however, cell-based EA generation is conceptually distinct from polygon-based EA delineation because it changes the underlying unit of aggregation and, therefore, the constraint set and error modes (Eurostat, 2024; Openshaw & Rao, 1995).

This separation also clarifies how core EA design principles translate into concrete method specifications. Statistical balance motivates explicit target constraints and reporting of deviations; spatial continuity/connectivity motivates topological rules and exception handling for edge cases; operational feasibility motivates the use of recognisable boundaries and barrier-aware logic; and temporal consistency motivates versioned workflows, stable identifiers, and documented update procedures. These principles recur across international guidance and national practice, even where the underlying data infrastructure differs (United Nations, 2017; United Nations, 2025).

A practical implication of this framework is that debates about "best practice" should be reframed as debates about which dimension is being optimised. For example, operational enumeration emphasises manageable workload and navigable boundaries, while dissemination geographies emphasise stability, confidentiality, and comparability over time. Automated zone design literature explicitly recognises this multi-objective character by formulating zone design as constrained optimisation: zones must be contiguous, respect hard boundaries where required, achieve target population or workload sizes, and sometimes maximise internal homogeneity of social or built-environment attributes (Openshaw & Rao, 1995; Cockings et al., 2011; Haynes et al., 2007).

A further implication concerns statistical validity and confidentiality. Zone design choices interact with the Modifiable Areal Unit Problem (MAUP): aggregation to different boundary configurations can change estimated relationships and spatial patterns, even when the underlying microdata are identical. For this reason, many statistical systems emphasise stability and comparability over time for dissemination units, while also applying minimum-size and disclosure-control considerations to protect respondent confidentiality. In practice, EA and EA-like geographies therefore balance operational manageability with statistical robustness, confidentiality thresholds, and the governance requirement that units be maintainable and auditable

across census cycles (Openshaw, 1984; Fotheringham & Wong, 1991; Cockings et al., 2011).

Boundary constraints are especially important for making zones interpretable to enumerators and acceptable to statistical authorities. In operational EA design, linear features such as rivers, railways, and roads are commonly used because they are observable on the ground and tend to correspond to real movement barriers. The relative "cutting strength" of these features is context dependent: major roads can be strong barriers in urban areas, while lower-class roads may still provide meaningful separation in rural settlement patterns. One way to make this context dependence explicit is to tie barrier rules to a settlement typology such as the Degree of Urbanisation (DEGURBA), which classifies territory along an urban–rural continuum using population grid cells and density thresholds (Eurostat, 2024). In this thesis, such typology-driven barrier logic is treated as a transferable design principle for reconciling navigability with spatial coherence.

Positioning the proposed method within this framework, the thesis adopts a GIS-centred but algorithmically controlled workflow that is best characterised as rule-based semi-automation operating on polygon building blocks. GIS is used primarily as an analytical and data-management infrastructure (geodatabase, topology checks, map outputs), while the delineation logic is implemented through explicit, reproducible rules and optimisation-inspired constraints (e.g., workload targets, contiguity, barrier respect, and controlled handling of zero-target cells). Accordingly, the method is GIS-based in its technical environment, but it is not defined by GIS alone; and it is semi-automated in its boundary construction logic, but it remains designed for human interpretability and operational defensibility (United Nations, 2025; Cockings et al., 2011).

This positioning matters for two reasons. First, it aligns with the cross-country evidence that successful EA systems combine stable cartographic infrastructure with explicit decision rules for boundary maintenance rather than relying on ad hoc edits. Second, it enables a transparent methodological comparison: later chapters can evaluate how different constraints (e.g., road hierarchies by settlement type, administrative boundary protection, or target thresholds) shape EA outcomes, without

ambiguity about whether observed differences stem from data environment, spatial representation, or automation level (Eurostat, 2024; United Nations, 2025).

Across national statistical systems, a quality assurance and quality control (QA/QC) typically includes: (i) topology checks (no gaps/overlaps, valid geometries), (ii) contiguity verification, (iii) constraint compliance checks (workload/population targets, mandatory boundary protection), (iv) barrier coherence checks (whether major linear features are used consistently as boundaries), and (v) field usability checks (legibility of maps, stability of identifiers, and supervisor review). Where semi-automated or automated procedures are used, QA/QC also requires a documented audit trail of parameters and rule versions so that boundary outcomes remain reproducible across iterations (United Nations, 2017; United Nations, 2025; UNICEF, 2013).

A recurring theme in the literature is that many EA failures are not conceptual but operational: they arise from incomplete input data, uncontrolled parameter sensitivity, or inconsistent exception handling. For example, when road networks or building layers are incomplete, algorithmic splitting can generate implausible boundaries or disconnected fragments. Similarly, if a method treats all barriers with equal strength, it may over-fragment urban areas or under-separate rural settlements. The framework therefore motivates two practical design requirements for the proposed method: (i) explicit, context-aware barrier rules and (ii) explicit handling of edge cases (e.g., low-target or zero-target areas) under auditable decision rules, rather than ad hoc manual edits.

Finally, comparability across time is strengthened when production workflows separate ‘stable’ components (e.g., administrative boundaries or dissemination geographies) from ‘adaptive’ components (e.g., workload-driven operational EAs). This separation allows a census programme to update operational units as settlement patterns change while maintaining consistent dissemination reporting where required. It also clarifies why different countries may adopt different unit portfolios without contradicting shared design principles: they are optimising different combinations of operational feasibility, statistical robustness, confidentiality, and maintainability (Office for National Statistics, 2021; Cockings et al., 2011).

2.4. Application Approaches and Method Families in EA Production

Building on the cross-country evidence in Section 2.2 and the conceptual dimensions in Section 2.3, this section reviews the main method families used to delineate and maintain EAs (and EA-like small-area geographies). The objective is not to catalogue software tools, but to summarise transferable workflow patterns, typical data requirements, and known strengths and failure modes. This synthesis supports the later methodology chapter by clarifying which choices are genuinely methodological (rule sets, objective functions, constraints) versus those that are primarily data- or infrastructure-driven (United Nations, 2017; United Nations, 2025). Although presented as method families for clarity, these approaches are frequently combined in practice; therefore, they should be read as recurring workflow patterns aligned with the conceptual dimensions in Section 2.3 rather than as mutually exclusive categories.

2.4.1. GIS-supported manual delineation and database-centred maintenance

The most widely documented approach remains GIS-supported manual delineation. Here, GIS is used to maintain a geodatabase of base layers (administrative boundaries, roads, water features, settlements, buildings/addresses where available) and to produce standardised enumeration maps and field materials. Delineation decisions are largely manual, but they are implemented within a controlled environment that supports topology validation, consistent identifiers, versioning, and supervisor review (United Nations, 2025; UNICEF, 2013).

This approach is operationally robust and defensible because boundaries can be aligned with recognisable features and exceptions can be resolved with local knowledge. Its limitations are scalability and consistency: manual editing becomes expensive at national scale, and different mapping teams may apply implicit rules differently unless a formal specification and QA/QC checklist is enforced (United Nations, 2017).

In practice, many statistical offices treat this workflow as the baseline infrastructure even when advanced automation is introduced. Automated or semi-

automated outputs are often imported into GIS for validation, cartographic refinement, and final publication of EA maps and identifiers (Office for National Statistics, 2021). This persistence reflects the fact that GIS-based boundary and identifier governance (geodatabases, topology rules, version control, and cartographic production) remains the institutional backbone of EA systems even when delineation becomes more algorithmic.

The cross-country evidence in Section 2.2 shows that even in systems with advanced delineation practices, the operational production pipeline typically culminates in a GIS environment for validation, map outputs, and the management of stable identifiers and metadata.

2.4.2. Rule-based semi-automation (scripted split/merge under constraints)

Rule-based semi-automation occupies an intermediate position between manual delineation and fully automated zone design. The method family is defined by explicit decision rules that propose boundary changes—typically splitting large units or merging small adjacent units—until target thresholds are met (e.g., households, population, dwellings, or workload proxies). The rules frequently include contiguity requirements and hard constraints that protect administrative boundaries or other mandated limits (United Nations, 2025).

Semi-automation is especially attractive when the input building blocks are reliable (e.g., blocks, parcels, or buildings) but the number of units makes manual balancing inefficient. Because human supervision remains integral, the workflow can accommodate context-specific exceptions and can prioritise interpretability of boundaries (e.g., ensuring alignment with major roads rather than arbitrary cuts).

A key risk is parameter sensitivity: small changes in thresholds, neighbour definitions, or barrier ‘strength’ can change the resulting EA set. As a result, semi-automated workflows should be paired with QA/QC that reports constraint violations, summary statistics (min–max–mean sizes), and diagnostic maps that reveal edge effects or disconnected fragments. The emphasis is therefore on reproducible rule versioning rather than on producing a single ‘optimal’ solution (United Nations, 2017;

Cockings et al., 2011). In this family, two recurring implementation challenges are particularly salient: (i) the treatment of low-target or zero-target areas that are spatially necessary for contiguity but neutral with respect to workload goals, and (ii) the definition of barrier hierarchies (e.g., differentiating major roads from minor streets) so that boundary formation remains both defensible and operationally legible. These issues are therefore addressed explicitly in the proposed method through auditable exception rules and context-aware barrier logic.

Illustrations of this logic appear in the international practices reviewed in Section 2.2, where semi-automated or rule-guided split/merge workflows are used to maintain workload targets while respecting mandated administrative limits and recognisable features such as road and water networks.

2.4.3. Automated zone design (optimisation and heuristic methods)

Automated zone design treats EA delineation as a constrained optimisation problem: building blocks are allocated to zones so that each zone is contiguous, meets target sizes, respects required boundaries, and may also optimise secondary criteria such as internal homogeneity of socio-demographic characteristics. In the literature, this family is closely linked to discussions of the Modifiable Areal Unit Problem (MAUP), because alternative admissible partitions can yield different statistical patterns (Openshaw, 1984; Fotheringham & Wong, 1991).

Optimisation-based methods can deliver consistent outcomes at scale and provide transparent objective functions. However, they can also generate zones that are difficult to interpret operationally if ‘ground-recognisable’ boundary constraints are weak. For census operations, this has led many systems to integrate optimisation with practical constraints such as using major linear features as preferred cut-lines and enforcing administrative boundary protection where required (Openshaw & Rao, 1995; Cockings et al., 2011).

In applied settings, automated zone design is rarely an opaque ‘black-box’ procedure. Instead, it is typically embedded within a broader production system that includes stakeholder constraints, iterative parameter testing, and GIS-based QA/QC before final adoption.

The examples summarised in Section 2.2 indicate that, where optimisation concepts are adopted, they are typically constrained by operational requirements (e.g., preferred cut-lines along major linear features) and implemented within a governance process that includes iterative testing and formal approval, rather than being treated as a purely technical exercise.

2.4.4. Cell-based (grid/fishnet) and gridded-population approaches

Cell-based approaches use regular grid cells (e.g., fishnets) as the base unit for workload balancing and spatial aggregation. In contexts where building footprints, addresses, or block-level units are incomplete or inconsistent, grid-based representations can provide a uniform, nationally consistent foundation. Grids also align naturally with modern urban–rural typologies and population-surface products, supporting both sampling-frame construction and analytical stratification (Eurostat, 2024).

The main methodological distinction is that cell-based approaches change the unit of aggregation and therefore shift the constraint set. While they can support automated balancing, they can also create operational challenges when grid boundaries are not recognisable on the ground. Practical implementations often require converting grid-based outputs into field-usable polygons by snapping to roads or administrative lines, which can introduce conversion artefacts and additional QA/QC needs. In operational settings, grid-based outputs are therefore commonly used as an intermediate analytic layer and subsequently translated into field-usable polygonal units by snapping boundaries to roads or administrative lines, producing a hybrid grid–polygon workflow.

The cross-country practices in Section 2.2 also suggest that gridded products are most effective when they complement, rather than replace, administratively grounded geographies—serving as inputs for stratification, coverage checks, or preliminary balancing where detailed address or building data are limited.

2.4.5. Object-based (building- and address-centric) approaches

Object-based approaches treat individual buildings, addresses, or housing units as primary inputs and aggregate them into EAs. Where high-quality address registers or building footprints exist, this approach can improve workload precision and reduce reliance on coarse population proxies. It is also compatible with frequent updates, because new buildings can be integrated continuously and then rebalanced under defined rules.

However, object-based methods are data inten,sive and are vulnerable to systematic omissions (informal settlements, rapidly changing peri-urban areas) unless complemented by field verification or auxiliary sources. For this reason, many national systems combine building-centric layers with administrative and road-based constraints to maintain interpretability and governance compatibility (United Nations, 2025). This approach is particularly feasible in statistical systems with mature address registers and routine geocoding pipelines, where building-centric updates can be integrated continuously and then rebalanced under explicit rules, while still maintaining compatibility with administrative governance requirements.

This institutional pattern is reflected in several country examples discussed in Section 2.2, where address or building layers support either EA maintenance directly or the linkage between operational units and dissemination geographies.

To support comparison across method families, Table 2.2 summarises typical inputs, expected automation levels, recurring strengths and failure modes, and the QA/QC emphases most frequently recommended in the literature. The table also anticipates the evaluation metrics reported in later chapters.

Table 2.2. Summary of EA method families, typical inputs, and quality-control emphases

Method family	Typical building blocks / inputs	Automation level (typical)	Main strengths	Common limitations / risks	QA/QC emphasis (illustrative)	Illustrative sources
GIS-supported manual delineation	Admin boundaries, roads/water, settlements; optional buildings/addresses	Manual	High interpretability; strong institutional defensibility	Labour-intensive; inconsistent decisions without formal rules	Topology, contiguity, identifier management, supervisor review	United Nations (2017, 2025); UNICEF (2013)
Rule-based semi-automation	Polygons (blocks/parcels/ admin units) and/or buildings; barrier layers	Semi-automated	Scalable balancing with human oversight; reproducible rule versions	Parameter sensitivity; edge-case instability	Constraint compliance reports; diagnostic maps; parameter logging	United Nations (2017, 2025); Cockings et al. (2011)
Automated zone design	Fine-scale building blocks with attribute vectors; constraints	Automated (with QA/QC)	Consistency at scale; explicit objective functions	Operational 'meaning' may degrade without strong constraints; MAUP sensitivity	Objective/constraint auditing; sensitivity checks; disclosure considerations	Openshaw (1984); Fotheringham & Wong (1991); Openshaw & Rao (1995)
Cell-based / gridded approaches	Regular grid cells; population surfaces; land cover; DEGURBA-type typologies	Semi-automated to automated	Uniform coverage when admin/object data are limited	Poor ground-recognisability; polygon conversion artefacts	Grid-to-polygon validation; barrier snapping checks; size distribution	Eurostat (2024); United Nations (2025)
Object-based aggregation	Buildings/addresses; optional registers; roads/admin constraints	Semi-automated to automated	Precise workload accounting; update-friendly	Data completeness bias (informal/peri-urban); integration complexity	Coverage audits; register–map reconciliation; exception documentation	United Nations (2025)

Taken together, the literature suggests that robust EA systems rarely depend on a single technique. Instead, they combine an auditable GIS infrastructure with explicit rules and QA/QC, while selecting polygon-, grid-, or object-based building blocks

according to data availability and operational requirements. This observation motivates the methodological choices of the proposed approach in later chapters, where the focus is placed on rule transparency, barrier-aware spatial constraints, and defensible exception handling rather than on an unconstrained notion of optimality (United Nations, 2017; United Nations, 2025).

2.5. Problems and Research Gaps Identified in the Literature

The literature reviewed in Sections 2.1–2.4 shows that EA production is rarely constrained by a single objective. Instead, it is shaped by a set of competing requirements that involve field logistics, statistical governance, spatial data infrastructure, and temporal maintenance. Accordingly, the most frequently reported problems do not merely concern how to 'draw boundaries', but how to maintain a defensible, auditable, and updateable small-area system under imperfect data and heterogeneous settlement patterns. This section synthesises the dominant challenges reported in international guidance and academic studies, and it clarifies the main research gaps that motivate the proposed approach in later chapters (United Nations, 2017; United Nations, 2025; Cockings et al., 2011).

2.5.1. Tension between statistical balance and spatial continuity

A recurring challenge is the trade-off between statistical balance (e.g., workload or population targets) and spatial continuity/connectivity. From an operational perspective, EAs are expected to be contiguous and navigable units that enumerators can traverse efficiently, often using recognisable boundaries such as major roads or water features. From a statistical perspective, EAs are expected to remain within target ranges to reduce field burden variability and support comparable sampling units. These objectives can conflict in heterogeneous urban contexts, where dense housing clusters may force repeated splitting to meet workload targets, potentially producing irregular shapes or boundaries that are less legible on the ground (United Nations, 2025).

The literature on zone design highlights that balancing constraints often yields multiple feasible partitions, and that 'optimality' depends on how objectives are weighted. For example, prioritising size targets can reduce workload variance but may increase the number of boundary segments that do not correspond to ground features. Conversely, prioritising barrier alignment and recognisability can preserve operational legibility but may create EAs that deviate from target thresholds, especially in areas with sparse or unevenly distributed settlement. The implication is that EA methods must explicitly state which constraints are 'hard' (non-negotiable) versus 'soft' (optimised where possible), and must disclose how trade-offs are resolved in practice (Openshaw & Rao, 1995; Cockings et al., 2011).

A further tension concerns confidentiality and disclosure control in published statistics. While operational EAs are designed primarily for fieldwork efficiency, dissemination or output geographies often require minimum-size and stability criteria to reduce disclosure risk. In practice, these requirements can pull in different directions: enforcing minimum sizes may encourage merging that worsens operational legibility, whereas strict workload balancing may produce very small units that are unsuitable for release. The literature therefore distinguishes between operational EAs and dissemination geographies, or it introduces additional constraints and post-processing steps to align the two where a single geography is expected to serve both functions (Openshaw, 1984; United Nations, 2017; United Nations, 2025).

A related research gap concerns the translation of 'field usability' into measurable criteria. While international manuals emphasise readable maps and ground-recognisable boundaries, fewer studies formalise these as quantitative constraints comparable to population/workload targets. This gap is relevant for semi-automated workflows, where operational feasibility must be encoded into rules (e.g., barrier hierarchies) rather than left to discretionary manual edits.

These trade-offs are also visible in the international practices summarised in Section 2.2, where countries combine workload targets with ground-recognisable boundaries and administrative constraints through iterative review and standardised mapping procedures.

2.5.2. The impact of zero- and low-target areas, sparsity, and edge effects

Zero-population or low-target areas (hereafter also referred to as zero-target or 0-RES units when the target variable is residential dwellings) appear frequently in practice, particularly when EAs are constructed from regular grids (fishnets) or when building blocks include non-residential land uses, industrial zones, parks, water bodies, or buffer corridors. Such units can be spatially necessary to preserve contiguity and to avoid gaps, yet they are 'neutral' with respect to the workload objective. If not handled explicitly, zero/low-target units can distort split/merge logic by acting as artificial barriers or by forcing inefficient merges that create long, thin, or operationally implausible EAs (United Nations, 2017; Eurostat, 2024).

The literature identifies at least three common failure modes. First, an algorithm may isolate zero-target units as standalone polygons that are operationally meaningless. Second, the algorithm may attach them to a neighbouring EA using arbitrary adjacency rules, producing outcomes that do not respect barrier logic or administrative constraints. Third, when zero-target units occur at administrative boundaries or near complex road intersections, boundary effects can trigger fragmentation, generating small 'sliver' EAs or disconnected components. These patterns are often underreported because they may be resolved informally during cartographic clean-up; however, such ad hoc fixes reduce reproducibility and make it difficult to compare results across iterations (Cockings et al., 2011; United Nations, 2025).

A clear gap is therefore the limited standardisation of exception-handling rules for zero/low-target units. International guidance encourages logical attachment to neighbouring units and the use of recognisable boundaries, but the literature offers fewer explicit, auditable decision rules that can be reused across contexts. This motivates methods that treat zero/low-target units as first-class edge cases, with deterministic attachment rules and documented QA/QC checks, rather than as incidental artefacts to be repaired manually.

Section 2.2 provides practical illustrations of these edge effects, particularly in contexts where sparse settlement patterns and mixed land-use produce 'empty' units

that must be attached under defensible rules to preserve contiguity and operational readability.

2.5.3. Parameter sensitivity, instability, and reproducibility

Another widely reported issue is parameter sensitivity. Many EA workflows—manual, semi-automated, or automated—depend on design choices such as target thresholds, neighbour definitions, barrier ‘strength’, and the sequencing of split/merge operations. Small changes in these parameters can produce materially different EA configurations, even when the same input layers are used. This sensitivity is closely related to the fact that multiple feasible solutions exist under the same constraint set, and that algorithms may converge to different local optima depending on initial conditions and tie-breaking rules (Cockings et al., 2011; Openshaw, 1984).

For statistical governance, instability undermines temporal consistency and complicates longitudinal comparisons. If EAs change substantially between updates without clear documentation, it becomes difficult to interpret changes in indicators as ‘real’ social change versus boundary change. For field operations, instability complicates training and supervision because enumerators and supervisors rely on stable identifiers and map familiarity. Consequently, guidance documents increasingly emphasise version control, parameter logging, and formal QA/QC reporting as part of the EA production process (United Nations, 2017; United Nations, 2025).

A research gap concerns the standard reporting of sensitivity and uncertainty. While optimisation-based zone design literature discusses MAUP and alternative admissible partitions, census practice often treats the produced EA set as a single deterministic output. In semi-automated settings, it is particularly valuable to report diagnostic summaries (e.g., min–max–mean unit sizes, number of units outside target range, count of disconnected polygons, and frequency of constraint violations) to support transparent comparison of parameter settings across iterations.

The need for reproducibility is reflected in Section 2.2, where several national systems emphasise versioned workflows, stable identifiers, and documented QA/QC reporting to support maintenance across census cycles.

2.5.4. Tool dependency, transparency, and institutional feasibility

Finally, the literature highlights institutional constraints that can dominate technical choices. Many countries rely on a combination of commercial GIS platforms, bespoke scripts, and manual cartographic workflows. While these toolchains can be effective, they may also create dependency on specific software, limited reproducibility across teams, and difficulty in transferring methods to new contexts. Tool dependency is particularly salient when methods are described at a high level (e.g., "use GIS to delineate EAs") without specifying the underlying rules, data checks, and exception-handling procedures that determine the final boundaries (United Nations, 2017; UNICEF, 2013).

Transparency is not only a scientific concern but also an operational requirement. EAs constitute official geographies that influence resource allocation, sampling frames, and the interpretation of published statistics. Therefore, the literature increasingly calls for auditable methods with clear documentation of inputs, constraints, rule versions, and QA/QC results. In practice, this implies separating (i) the conceptual method specification (rules and constraints), (ii) the production environment (GIS infrastructure and data sources), and (iii) the automation strategy (manual, semi-automated, automated). This separation supports replication and cross-country comparison and aligns with the conceptual positioning developed in Section 2.3.

These institutional constraints are also evident in Section 2.2, where capacity, software ecosystems, and data availability shape how far automation can be adopted without compromising transparency and governance.

Operationally, EA systems are implemented through combinations of GIS platforms and scripting environments. Commercial GIS software can provide robust topology tools and interactive editing, but it can constrain full reproducibility outside licensed settings. Open-source stacks (e.g., QGIS, R with sf/lwgeom, and PostGIS) offer strong transparency and automation potential, yet they require explicit management of geometry validity, coordinate reference systems, and computational performance when scaling to large urban datasets.

From a governance perspective, cross-country experience also suggests that sustainable EA systems require institutional components beyond a one-off delineation algorithm. A Türkiye-oriented implementation would likely need (i) a national EA geometry layer designed around workload thresholds and field constraints; (ii) a documented update protocol that absorbs new buildings into existing units and defines controlled split/merge rules; (iii) confidentiality-aware design choices for small-area outputs; and (iv) systematic versioning and governance to support auditability and repeated survey use.

To consolidate the synthesis, Table 2.3 maps each major challenge reported in the literature to its typical causes, operational/statistical consequences, and common mitigation strategies. The final column indicates how each challenge informs method design choices in this thesis, thereby linking the literature review to the methodological specification and evaluation metrics presented in later chapters.

Table 2.3. Key challenges in EA production, typical causes, and implications for method design

Challenge	Typical causes	Operational/statistical consequences	Common mitigations reported in the literature	Implications for this thesis
Balance vs recognisability	Heterogeneous settlement; strict targets; weak barrier constraints	Irregular or hard-to-navigate EAs; inconsistent workload	Define hard vs soft constraints; prefer major barriers; supervisor review (United Nations, 2025)	Use barrier-aware rules and explicit trade-off logic; report compliance summaries
Zero/low-target units and sparsity	Non-residential land; fishnet cells with 0 dwellings; missing building data	Standalone 'empty' polygons; arbitrary attachments; sliver units	Deterministic attachment rules; post-processing clean-up with documentation (United Nations, 2017)	Treat 0-target/0-RES as explicit edge cases with auditable merge rules
Parameter sensitivity	Threshold choices; neighbour definitions; split/merge order; barrier weighting	Instability across runs; weak temporal consistency	Version control; parameter logging; diagnostic statistics (Cockings et al., 2011)	Define reproducible rule versions; add sensitivity-aware evaluation metrics
Data incompleteness	Incomplete roads/buildings; outdated admin boundaries; misaligned layers	Disconnected fragments; implausible boundaries; field confusion	Coverage audits; topology checks; integration QA/QC (UNICEF, 2013)	Layer validation pipeline and explicit exception handling for missing inputs
Tool dependency / low transparency	Proprietary toolchains; undocumented manual edits; ad hoc fixes	Low replicability; difficult transfer and review	Document rules, constraints, QA/QC; standard outputs (United Nations, 2017)	Separate method specification from platform; produce traceable outputs

In summary, the literature indicates that the key gaps are not limited to proposing new delineation algorithms. Rather, they concern the operationalisation of design principles into transparent rule sets, the standardisation of exception handling (notably for zero/low-target areas), and the reporting of reproducible QA/QC metrics. These gaps are central for contexts where data are heterogeneous and where the institutional environment requires methods that are auditable, transferable, and maintainable across cycles (United Nations, 2017; United Nations, 2025).

This chapter reviewed the concept, functions, and design logic of Enumeration Areas (EAs) as the foundational spatial units of census and survey operations. Section 2.1 clarified the role of EAs as operational workload units and as intermediaries between fieldwork and statistical outputs, including their importance for sampling frames, data integration, confidentiality-aware reporting, and longitudinal comparability. Section 2.2 demonstrated that EA implementation differs across countries in terms of unit types, maintenance strategies, and the balance between operational and dissemination geographies, but that recurring principles can still be identified (United Nations, 2017; United Nations, 2025).

Section 2.3 positioned EA design within a three-part conceptual space: the technical environment (GIS-enabled versus non-GIS environments), the spatial representation used as building blocks (polygon-, cell-, or object-based), and the automation strategy (manual, semi-automated, automated). This positioning highlighted why GIS-based does not imply automation, why automation can also be implemented in non-GIS computational settings, and why the chosen building blocks strongly influence both the feasible constraint set and the practical failure modes. Section 2.4 then consolidated the major method families used internationally—GIS-supported manual delineation, rule-based semi-automation, automated zone design, cell-based approaches, and object-based aggregation—showing that robust EA systems are typically hybrid and depend on auditable QA/QC procedures as much as on delineation algorithms (UNICEF, 2013; Office for National Statistics, 2021; Cockings et al., 2011).

Building on this foundation, Section 2.5 synthesised the dominant problems and research gaps emphasised in the literature: the persistent trade-off between workload balance and ground-recognisable boundaries; the operational disruption

caused by zero/low-target units and edge effects; parameter sensitivity and instability across runs and cycles; data incompleteness and layer misalignment; and institutional constraints associated with tool dependency and limited transparency. These gaps motivate a methodological focus on explicit rule specification, deterministic exception handling (particularly for zero/low-target or 0-RES units), barrier-aware logic, and reproducible QA/QC reporting rather than reliance on ad hoc manual corrections or opaque optimisation outputs (United Nations, 2017; United Nations, 2025; Openshaw, 1984).

The literature review provides three direct implications that guide the methodological design of the proposed approach. First, the method should treat EA delineation as an auditable production workflow rather than a one-time partitioning task, and therefore must include traceable rule versions, parameter logging, and standard QA/QC outputs. Second, it should integrate barrier information through an explicit hierarchy (e.g., distinguishing major roads from minor streets) to improve operational legibility and reduce arbitrary boundary formation. Third, it should define deterministic exception rules for zero/low-target units so that spatial continuity is preserved without creating stand-alone 'empty' units or arbitrary attachments.

The next chapter operationalises these implications by specifying: (i) the required spatial input layers and their validation checks; (ii) the building-block representation used in the proposed workflow; (iii) the split/merge and attachment rules, including barrier-aware decisions and edge-case handling; and (iv) the QA/QC and evaluation metrics used to compare outputs across parameter settings and study areas. Consistent with the challenges identified in this chapter, the evaluation emphasises both statistical targets (e.g., compliance with workload thresholds) and spatial-operational criteria (e.g., contiguity, barrier coherence, and reduction of artefacts such as sliver polygons), with transparent reporting designed to support replication and maintenance across cycles (United Nations, 2017; United Nations, 2025; Cockings et al., 2011).

CHAPTER 3. METHOD

The primary objective of the methodology developed in this study is to design an automated, reproducible, and barrier-aware Enumeration Area (EA) delineation framework suitable for large metropolitan regions, with a specific application to Ankara, Türkiye. Enumeration Areas constitute the fundamental spatial units for population censuses and large-scale household surveys, and their design directly affects data quality, field operability, and statistical comparability (United Nations, 2017).

Traditional EA delineation practices in many countries rely on manual cartographic processes, local expert judgment, or semi-automated tools that still require extensive human intervention. While such approaches may yield acceptable results for limited geographic extents, they are increasingly inadequate for large and rapidly evolving urban environments. Urban expansion, mixed land-use patterns, and complex transportation infrastructures introduce spatial heterogeneity that cannot be efficiently addressed through static or purely administrative partitioning schemes.

In response to these conditions, the methodology is designed to address three central challenges observed in contemporary EA design. First, target size heterogeneity requires threshold settings that remain operationally meaningful across urban, semi-urban, and rural contexts. Second, delineation must respect physical and functional barriers—such as major roads, railways, and waterways—that constrain feasible field movement and should not be crossed by enumerators during field operations. Third, the workflow must be operationally robust, enabling restartable execution, auditable decision trails, and scalability to metropolitan-level datasets.

Rather than treating EA delineation as a purely geometric partitioning problem, the proposed approach frames it as a constrained spatial aggregation task. Atomic spatial units are first generated based on physical barriers, after which these units are aggregated under explicit population-based constraints. This two-stage structure reflects international best practices recommended by the United Nations Statistics Division (UNSD) and Eurostat, while allowing adaptation to local data availability and administrative classifications such as DEGURBA (Eurostat, 2021).

The scope of this methodology is limited to the pre-fieldwork phase of census operations. It does not attempt to optimize enumerator routing or workload balancing at the individual enumerator level. Instead, it focuses on producing spatially coherent, non-overlapping, and operationally feasible EA polygons that satisfy predefined size constraints and respect real-world barriers.

The preference for ArcGIS Pro is based on two main reasons. First, the ArcGIS ecosystem is still dominant in institutional GIS practice in Türkiye, including municipalities and central public institutions. For applications associated with official statistical processes such as EA production, developing a method in a software environment that is widely used institutionally increases the applicability and transferability of the outputs. The second reason is that ArcGIS Pro provides reproducible, script-based automation through ArcPy. This feature enables the systematic testing of different parameters and the verification of results in multi-stage and iterative EA production processes.

Within the tool ecosystem provided by ArcGIS Pro, there are several geoprocessing tools that may potentially be suitable for EA production. In particular, the ability to organize spatial data in a structured manner through Feature Dataset structures; the ability to analyze barrier effects and neighborhood relationships using the Pairwise Buffer and Polygon Neighbors tools; and the ability to convert building and housing information into spatial attributes using the Summarize Within tool led to the initial assessment of ArcGIS Pro as a strong implementation environment.

Within this toolset, the Build Balanced Zones (BBZ) tool initially emerged as a particularly notable solution for EA production. BBZ is an automated zoning tool that aims to generate spatially contiguous regions by balancing a target quantitative variable (e.g., population or number of dwellings). Because it conceptually overlaps with EA requirements defined in the literature and can produce results quickly, BBZ was initially considered a reasonable reference tool in this study (Esri, n.d.).

However, applications conducted with BBZ showed that the tool could not fully deliver the EA behavior targeted in this thesis. In particular, the level of user control remained limited in areas such as applying specific thresholds, handling structures with zero dwellings in the zoning process, and defining special merging rules. In addition, the tool's cost and license dependency and its black-box structure

created important constraints in terms of methodological transparency and reproducibility.

At this point, ArcGIS Pro was re-evaluated not only as an environment for using ready-made tools, but also as a platform where open, rule-based, and customizable workflows can be developed through ArcPy. Since BBZ was not adopted as the final solution, the study transitioned to an ArcPy-based workflow with full building-level control, explicitly defined constraints, and auditable steps for EA production. This transition constitutes the starting point of the methodological evolution that is discussed in detail in the subsequent sections.

At the initial stage, it was assumed that EAs that both preserve spatial continuity and comply with dwelling-unit constraints could be produced within neighborhood boundaries by combining ArcGIS Pro's advanced geoprocessing tools. However, during implementation it was observed that this assumption could not be satisfied in every neighborhood due to license-level constraints of tools for planar partitioning and full-coverage production, as well as output stability problems (Esri, n.d.). These observations revealed the limitations of approaches based on ready-made tools and strengthened the need for more flexible methodological solutions.

3.1. Practical Evaluation and Limitations of the Build Balanced Zones (BBZ) Tool

This section explains how the Build Balanced Zones (BBZ) tool in the ArcGIS Pro environment was evaluated in practice in the context of Enumeration Area (EA) production and why it was not adopted as the final solution in this thesis. The discussion focuses on the differences between the opportunities BBZ theoretically offers and the behaviors encountered in practice; it also addresses the methodological implications of BBZ not being sufficient on its own for "EA area production."

BBZ is an optimization tool that groups "input features" into a smaller number of "zones" according to a selected zoning method, while attempting to keep these zones spatially contiguous. According to Esri documentation, BBZ can accept point or polygon data as input and produces an output layer showing which inputs are grouped under which ZONE_ID (Esri, n.d.). Accordingly, BBZ primarily answers the question

of "which units should be grouped together" for EA production; however, the problem of "which areas will be represented by boundaries" (full-coverage polygon production) may remain outside the scope of BBZ.

This creates a critical gap for the operational use of EAs. In practice, EA outputs are generally expected to be used as polygon areas (for map production, field planning, and census organization). However, when point inputs are provided to BBZ, the output layer still retains point geometry and carries the ZONE_ID information; the tool does not automatically generate area boundaries from points. BBZ documentation explicitly states that neighborhood constraints vary by input geometry: if the input is polygon, neighborhood is handled through contiguity; if the input is point, a point-based neighborhood definition such as trimmed Delaunay triangulation is used (Esri, n.d.). This technical distinction indicates that the input geometry and preprocessing steps are decisive for representing EAs as areas.

In this thesis, because building and dwelling-unit information is mostly carried through point- or building-based data structures, using BBZ alone did not guarantee producing EAs as polygon areas. Therefore, even in scenarios where BBZ could be used, it became necessary to develop additional methods for area representation either before or after zoning. In practice, this need was evaluated through two main strategies: (i) producing a preliminary area layer to be used as the census unit (e.g., fishnet/grid), summarizing dwelling information into these areas, and providing polygon input to BBZ; (ii) after grouping point/building-based units with BBZ, using planar partitioning (full coverage) approaches to create an areal representation for each group. Both strategies demonstrated that BBZ can produce a meaningful output for EA production only when combined with complementary geoprocessing steps.

In previous trials conducted with BBZ (in a separate implementation process outside the thesis text), inconsistencies were observed, particularly in keeping EA sizes within the targeted 80–120 dwelling range. In some areas, EAs remained clearly below the lower threshold, while in other cases the upper thresholds were exceeded substantially. Such deviations stem from the user's inability to closely control BBZ's internal decision mechanisms and optimization preferences during the zone-growing process. Esri states that BBZ uses a genetic algorithm-based search process and evaluates different constraints using "fitness" scores (Esri, n.d.). While this structure

provides strong optimization capability, it can make it difficult to apply the specific threshold and exception rules targeted in this study in a deterministic manner.

Another major limitation of BBZ is that it does not provide detailed building-level control. In this thesis, EAs are intended to be created based on dwelling units defined at the building level. However, BBZ does not make visible in detail how building-based attributes are incorporated into the zoning process; it offers limited capability to define "special merging" rules for structures with zero dwellings, mixed-use buildings, or outliers. For this reason, BBZ remains a more general-purpose zoning tool rather than a framework where EA behavior is finely tuned using building-based constraints.

Tool dependency and explainability also contributed to BBZ not being selected as the final solution. As a black-box tool, BBZ does not expose intermediate decisions such as the zone-growth order, prioritization of neighborhood relationships, and tie-breaking rules. This makes it difficult to explain why the produced EAs take a particular form and limits methodological auditability. In the context of official statistics and population censuses, methodological transparency and reproducibility are not merely technical preferences; they are also institutional quality assurance requirements (United Nations, 2017).

Due to these practical requirements, the workflow was attempted to be automated by moving to Python/ArcPy scenarios that included BBZ. The aim was to manage steps such as area production before BBZ (e.g., fishnet) and area representation after BBZ (full coverage via planar partitioning) within a single script. However, during implementation, serious difficulties were encountered, particularly under boundary conditions (fragmented geometries of neighborhood boundaries, intersections of road/barrier lines, gap formation, disjoint parts) and due to license/stability problems of planar partitioning tools. These difficulties showed that BBZ must be run not alone but together with auxiliary processes that require "area production," and thus the solution increases tool dependency.

At this point, the role of ArcPy should be briefly explained. ArcPy is a site package that provides access to ArcGIS Pro's geoprocessing tools through Python; it enables script-based execution of tasks such as geographic data analysis, data

management, transformation, and map automation (Esri, n.d.). The main advantages of ArcPy are end-to-end automation of workflows, systematic testing of parameters, preservation of processing logs, and the ability to define rule-based post-processing steps. In contrast, ArcPy depends on the ArcGIS licensing model; access to some tools may vary by license level, which can affect long-term reproducibility. In addition, complex geoprocessing steps can become fragile with respect to geometric validity and topology and may require additional error handling.

In summary, while BBZ is valuable as a conceptual reference and starting point for EA production, it was not adopted as the final solution in this thesis because (i) it requires complementary methods for the areal representation of EAs, (ii) it remains limited in applying building-level special constraints deterministically, and (iii) it increases tool/license dependency. For these reasons, the study moved from BBZ-including scenarios to a BBZ-independent, open, rule-based, and building-centered approach. The next section discusses in detail how this transition was transformed into a customized ArcPy-based workflow.

3.2. Transition to a Rule-Based ArcPy Workflow Independent of BBZ and the Design of the EA Production Process

This section addresses the methodological evolution followed for Enumeration Area (EA) production in the ArcGIS Pro environment within an integrated framework. The practical and methodological limitations of trials initially carried out based on the Build Balanced Zones (BBZ) tool made it necessary, in later stages of the study, to re-formulate EA production without dependence on ready-made tool behavior. In line with this need, the EA production process was re-designed using ArcPy as a rule-based, traceable, and reproducible workflow.

The main limitation of BBZ-based approaches is that they reduce EA production to two separate and disconnected problems: (i) which spatial units should be grouped together and (ii) how these groups should be represented as areas. BBZ can address the first problem to a certain extent; however, for the second problem it directs the user to additional tools and complex pre-/post-processing steps such as fishnet, Thiessen/Voronoi polygons, or planar partitioning. This has made it difficult

to produce gap-free, topologically valid, and operational EA polygons within neighborhood boundaries. The “area” representation that is mandatory for field applications of EA production emerged as a structural gap that BBZ does not solve directly (Esri, n.d.).

At this point, EA production was reconsidered not merely as a zoning problem, but as a multi-dimensional process requiring the simultaneous management of spatial, statistical, and topological constraints. This reframing required moving beyond the closed optimization logics offered by ready-made tools and defining the EA production process through rules specific to the research question. In this context, ArcPy was adopted as the primary implementation environment that made the methodological reconstruction of EA production possible.

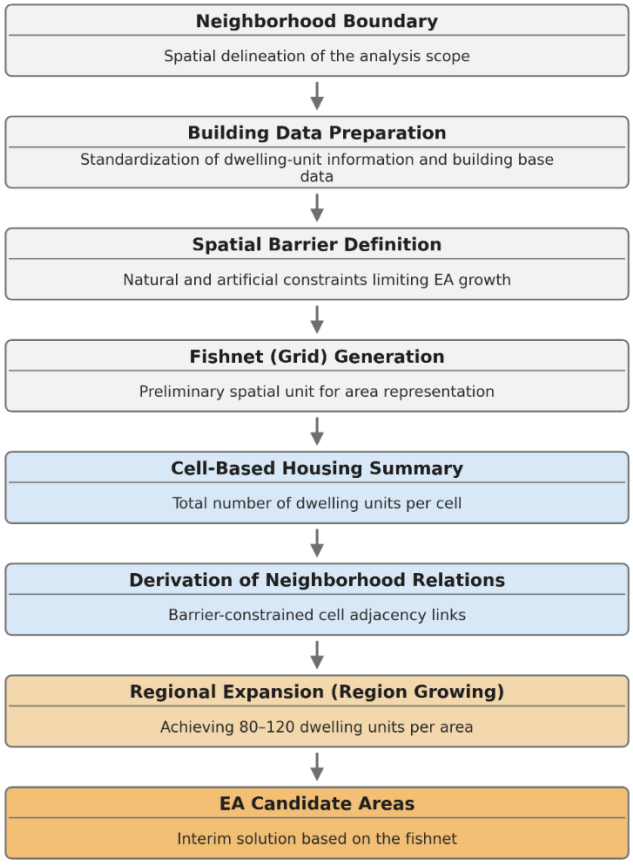
ArcPy is the Python-based application programming interface (API) of ArcGIS Pro and enables script-level automation of geospatial data processing workflows (Esri, n.d.). Spatial operations performed through the graphical user interface can be transformed into parametric, reproducible, and auditable scripts via ArcPy. In this thesis, the main reasons for preferring ArcPy are the ability to control the EA production process step by step, systematically test different parameter sets, and increase the comparability of the produced results (Longley et al., 2015). In this way, it became clearly traceable which spatial operation was applied in which order, which rules were activated at which stage, and how EA boundaries were formed.

This flexibility offered by ArcPy also brings certain structural limitations. A large portion of ArcGIS tools depend on license level, and some critical operations (for example, planar partitioning or advanced area production) have become inaccessible in certain periods. In addition, the internal optimization logic of closed-source tools not being visible to the user limits methodological transparency in complex spatial problems (Openshaw, 1984; Goodchild, 2007). Nevertheless, in this thesis ArcPy is evaluated not as a final solution, but as a transition and analysis environment that enables systematic and auditable execution of methodological trials within the ArcGIS environment.

The ArcPy-based EA production workflow was initially built on a fishnet-based design that aimed to solve the area representation problem from the outset. The primary motivation for this approach is that BBZ does not directly produce areas and

that there is a mandatory need for polygon representation in the operational use of EAs. Fishnet was evaluated as a suitable precursor structure because it provides a geometrically regular starting unit that is gap-free and ensures full coverage within neighborhood boundaries. The general structure of this workflow is presented as a methodological process diagram in Figure 3.1.

Figure 3.1. Methodological representation of a fishnet-based EA production workflow developed in ArcGIS Pro and ArcPy environments.



Source: Generated by the author.

In the figure, gray tones represent data definition and preliminary preparation stages, blue tones represent spatial analysis and calculation steps, and orange tones represent merging and optimization stages based on housing-unit constraints.

In the first stage of the workflow, the administrative boundary of the neighborhood included in the analysis was defined and the building dataset within this boundary was prepared. Building polygons were cleaned by removing records that do

not contain housing-unit information or that are geometrically invalid. They were then dissolved using unique identifiers to obtain building-level analytical units. This step is critical because it allows EA size constraints (e.g., 80–120 housing units) to be monitored directly through building attributes.

Following the preparation of the building data, spatial barriers intended to guide EA growth were defined. Natural and artificial obstacles—such as primary roads, railways, and water bodies—were combined to create a barrier mask. This mask was incorporated into the workflow to be used when defining neighborhood relationships. The purpose of this approach is to ensure that EAs remain within logical and operationally accessible boundaries for field operations.

Fishnet generation is one of the most sensitive stages of the workflow. Equal-sized square cells were generated within the neighborhood boundary, and cell size was defined as a parametric variable. Selecting an appropriate cell size was treated as a trade-off between the flexibility of EA boundaries and computational stability. While smaller cell sizes enable more detailed EA boundaries, they also increase the complexity of the neighborhood graph and reduce the stability of the results.

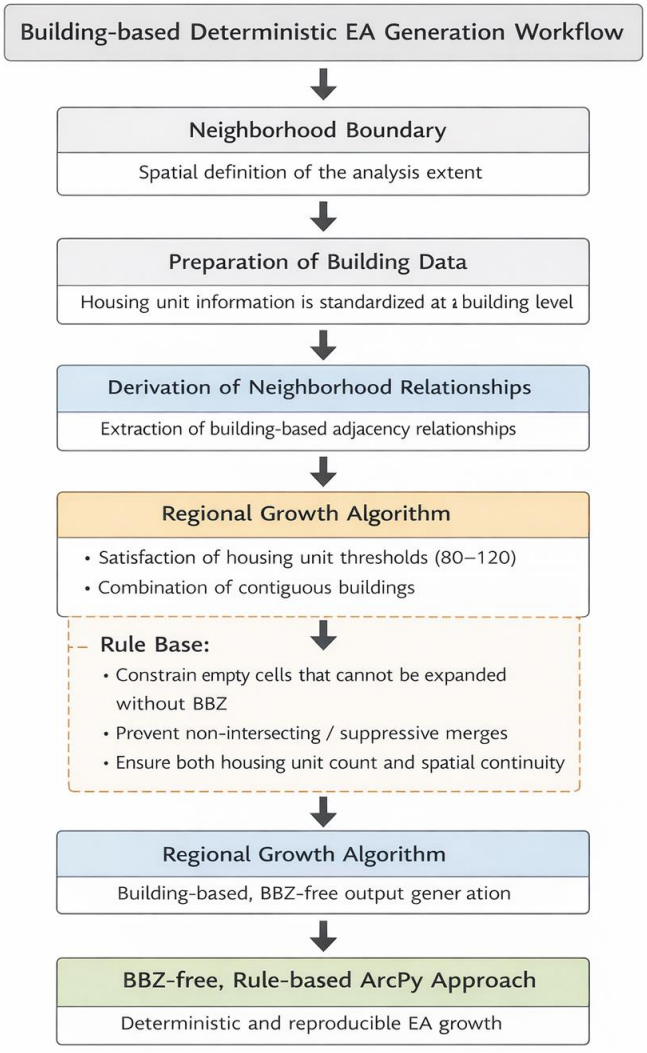
After the grid cells were created, building-based housing-unit information was summarized at the cell level. Using the Summarize Within tool, the total number of housing units was calculated for each cell, and the fishnet cells became the core analytical units for EA production, carrying both spatial and quantitative information. At this stage, the presence of cells with zero housing units became evident and later emerged as a key source of methodological problems.

Neighborhood relationships were extracted using the Polygon Neighbors tool, and selected connections were disabled with the support of the barrier mask. A cell-based region-growing logic was then applied to this neighborhood graph. Starting from initial cells, adjacent cells were gradually merged in an attempt to reach the targeted housing-unit band. However, during implementation it was observed that factors such as cell size, the neighborhood definition, and the influence of zero-housing areas made EA outputs highly sensitive and unstable.

Overall, the fishnet-based workflow made the spatial representation and quantitative balancing requirements of EA production explicitly visible; however, it was not adopted as the final solution due to practical limitations. These experiences

demonstrated that EA production should be treated as a spatial integration problem defined directly by rules, without dependence on antecedent units such as fishnets. The transition from the fishnet-based approach to the rule-based approach is summarized in Figure 3.2.

Figure 3.2. The final methodological workflow demonstrating the transition from a fishnet-based approach to a rule-based and building-oriented EA production process.



Source: Generated by the author.

The colors used in this manner are consistent with the color coding defined in Figure 3.1; green tones are used to emphasize the final EA production approach proposed in this thesis.

In conclusion, the ArcGIS Pro– and ArcPy-based EA production workflow presented in this section represents a methodological transition phase within the scope of this thesis. The fishnet-based approach clearly revealed the points at which EA production becomes unstable; ArcPy, by making the causes of these problems visible, established the conceptual basis required for the data preparation and preprocessing processes discussed in the next section.

3.3. Data Preparation and Preprocessing Processes

Enumeration Area (EA) production, although it may appear on the surface to be a spatial zoning problem, is fundamentally a process that is highly sensitive to prerequisites such as data quality, representational consistency, and attribute accuracy. EAs are not merely areas drawn on a map; they are the basic building blocks of operational processes such as population censuses, household statistics, and field organization. For this reason, the building data used for EA production must be carefully prepared, both geometrically and in terms of attributes.

In this thesis, for applications conducted in ArcGIS Pro and ArcPy, the data preparation process was initially treated as a secondary technical step. However, as the implementation progressed, it became evident that many problems encountered in EA production were directly related to decisions made at this stage. In particular, the fragmented structure of building geometries, the spatial distribution of housing-unit information, and the inaccurate definition of dwelling characteristics emerged as key factors determining the statistical and spatial stability of EA results.

Therefore, data preparation is considered in this thesis not only as "preprocessing," but as a methodological stage that shapes the behavior of EA production.

3.3.1. Merging (Dissolving) Building Polygons

Building datasets produced at the urban scale often do not reflect physical reality on a one-to-one basis. Different sections, floors, or functional areas of the same structure may be represented as separate polygons. This is common especially in

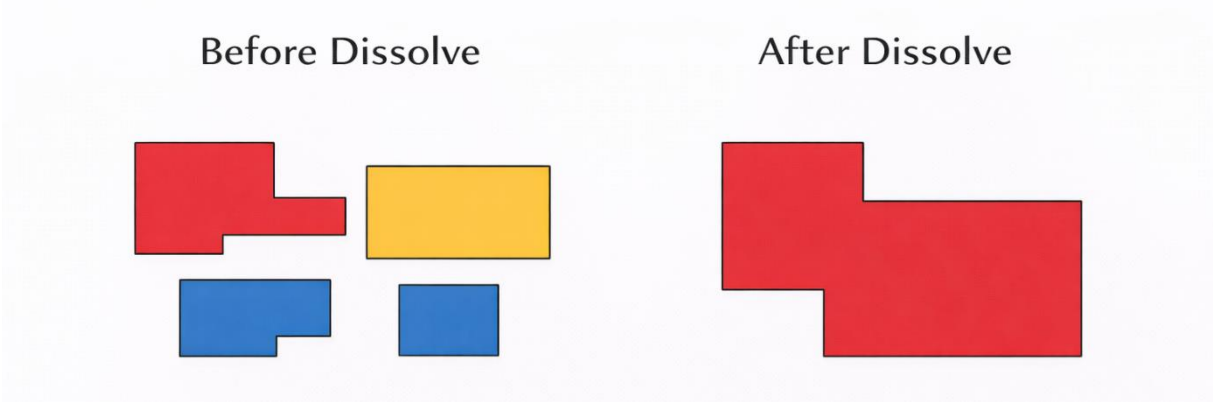
apartment-type buildings and mixed-use parcels. While such fragmented representations may be acceptable for certain spatial analyses, they create major methodological problems in a threshold-based task such as EA production.

A core constraint in EA production is keeping the number of housing units within a specified range (for example, 80–120) within each enumeration area. When housing units belonging to the same physical building are distributed across multiple polygons, this count is artificially divided. As a result, EA algorithms attempt to optimize a spatial fragmentation that does not exist in reality. Consequently, EA boundaries take forms that respond to errors in the data structure rather than to statistical requirements.

To eliminate this problem, building polygons were merged (dissolved) based on the yapi_kimli field. During this operation, the bb_sayisi field, or an equivalent housing-unit field, was aggregated into a single value using the SUM statistic. As a result, each physical building is represented by a single geometric feature and a single housing-unit value.

This step ensured that the smallest analytical unit used in EA production was not arbitrary geometric fragments, but physical buildings consistent with on-the-ground reality. The spatial effect of the Dissolve operation is presented visually in Figure 3.3, which shows fragmented polygons belonging to the same building being consolidated under single building geometries.

Figure 3.3. Representation of building polygons before and after the dissolve process.



Source: Generated by the author.

The literature also emphasizes that enumeration areas should, as far as possible, be based on spatial units that are meaningful, operationally definable, and consistent with on-the-ground reality (United Nations, 2017).

3.3.2. Distinguishing Residential and Non-Residential Structures

Building datasets typically include not only structures that accommodate population, but also buildings used for commercial, public, and industrial purposes. However, the primary objective of Enumeration Area (EA) production is to create spatial units that support the reliable production of population- and household-based statistics. For this reason, the assumption that every building geometry contributes equally to EA size is not methodologically valid and can directly distort statistical outputs.

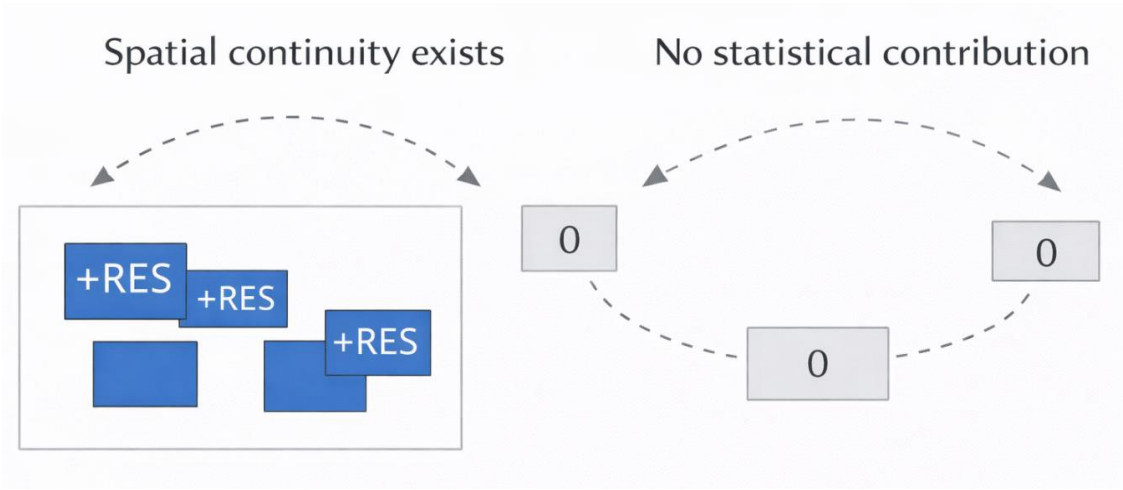
In this thesis, structures were classified as residential or non-residential using the "type" field (or an equivalent land-use/use-category attribute) in the building dataset. Only buildings with a residential function were included in the housing-unit (RES) calculation. Buildings with commercial, public, or industrial functions were retained in the workflow but assigned a RES value of 0. This approach ensured that the statistical content of EAs was defined directly through population-carrying units and enabled EA size constraints (e.g., 80–120 housing units) to be monitored in a more meaningful way.

Nevertheless, non-residential structures were not fully excluded from the analysis. Although they do not carry resident population, non-residential buildings are important components of spatial continuity within a neighborhood. Removing them entirely can create gaps within EA boundaries, produce irregular geometries, and result in areas that are problematic from the perspective of field operations. Therefore, non-residential structures were geometrically preserved while being treated as statistically neutral.

The effect of this distinction on EA production is illustrated schematically in Figure 3.4. On the left side of the figure, residential buildings that carry housing units contribute directly to EA size, and spatial continuity aligns with statistical content. On the right side, non-residential buildings are shown to have no housing-unit contribution

while still forming part of spatial continuity and thus playing an indirect role in shaping EA boundaries. This visualization clearly captures the distinction between the "absence of statistical contribution" and "spatial presence" for 0-RES structures.

Figure 3.4. Schematic representation of the different roles of residential and non-residential buildings in the EA production process.



Source: Generated by the author.

This methodological choice played a decisive role in the later stages of the fishnet and region-growing approaches. In particular, the fact that 0-RES areas can cause spatial expansion during EA growth without providing any statistical contribution emerged as one of the main sources of methodological problems. Accordingly, distinguishing residential and non-residential structures in this manner is not merely a data-cleaning step; it is a critical design decision that directly influences the behavior of EA production algorithms.

The literature similarly emphasizes that enumeration areas should be defined on the basis of population-carrying units, while the preservation of spatial continuity is indispensable for field implementation (United Nations, 2017; Eurostat, 2019). The approach adopted in this study aims to establish a balance between these two requirements.

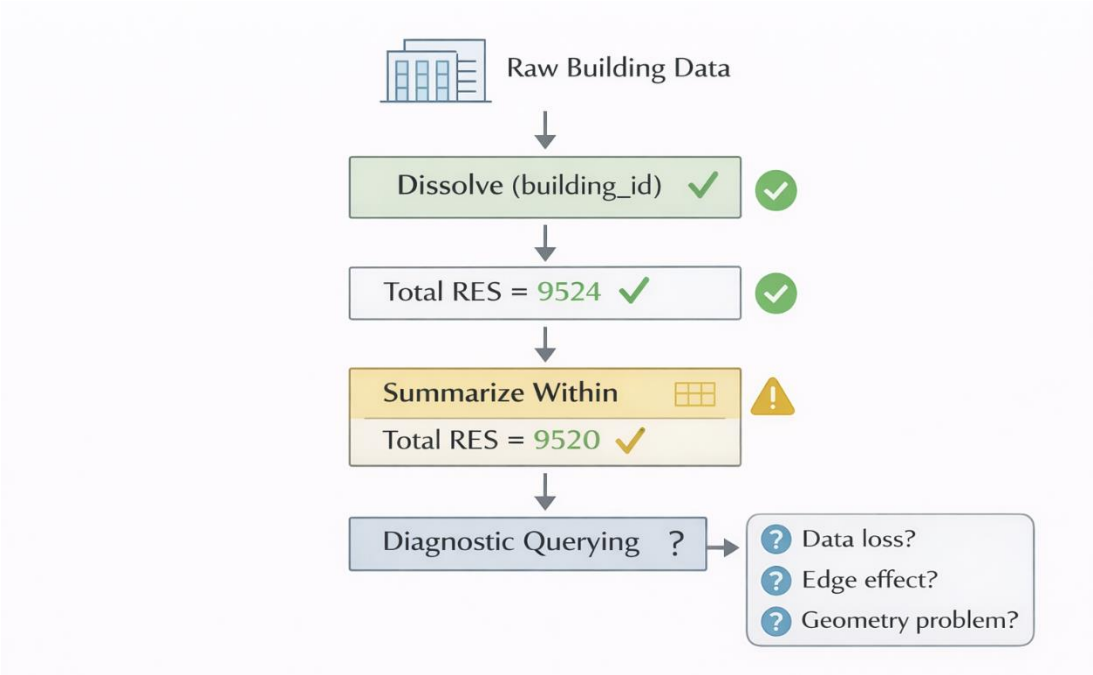
3.3.3. Total RES Validation and Diagnostic Monitoring (Methodological Framework)

At each stage of the data preparation process, the systematic calculation and recording of the total number of housing units (RES_UNITS) was designed as a methodological control mechanism. The purpose of this approach is to enable early detection of whether the spatial transformations applied in the EA production workflow produce unexpected effects on housing-unit information. In this context, the total RES value is treated not as an outcome indicator, but as a diagnostic variable used to audit data integrity and the internal consistency of the processing chain.

Within the scope of this thesis, the total RES value was recalculated after each major processing step, including the dissolution of building polygons, the separation of residential and non-residential structures, spatial summarization, and key operations related to areal representation. These calculations were planned to monitor whether housing-unit information is carried consistently through the workflow at every stage of data preparation. This control mechanism was considered critical, particularly because steps involving areal representation and spatial summarization can indirectly affect quantitative attributes.

The logic of total RES monitoring is presented in Figure 3.5 as a methodological flow diagram. The diagram shows how the total RES value is used not only as a final output, but also as a diagnostic reference tracked across successive spatial operations in the EA production process. At each stage of the workflow, the total RES value is compared with the values from preceding steps to verify whether quantitative consistency is maintained.

Figure 3.5. Methodological flow for diagnostic monitoring of total residential unit (RES) value in the data preparation process.



Source: Generated by the author.

The total RES value was recalculated after successive spatial operations and used as a diagnostic reference variable to monitor data integrity and the consistency of the processing chain.

The green markers shown in the figure represent cases in which the total RES value is preserved consistently with the previous stages at the corresponding processing step. The warning symbols indicate methodological control points showing that a discrepancy in the total RES value has emerged after specific spatial operations. The query stage at the end of the diagram represents a diagnostic process aimed at methodologically assessing possible sources of these discrepancies (such as boundary effects, geometry problems, or the internal behaviors of spatial summarization tools).

Within this approach, changes in the total RES value were not interpreted as direct results or findings; they were treated only as technical indicators intended to monitor the interaction between data and method. The diagnostic monitoring logic aims to identify which processing steps are more sensitive in terms of quantitative information and to guide, in an informed manner, the methodological decisions to be taken in subsequent stages of the EA production process.

Such diagnostic control mechanisms are consistent with quality assurance approaches recommended for ensuring data integrity in complex, multi-stage spatial workflows. In particular, monitoring consistency across intermediate outputs is widely regarded as one of the core methodological principles that supports the reliability of final results (Longley et al., 2015).

3.3.4. Potential Sources of Deviation and Methodological Risks in Data Preparation

The data preparation process is not merely a technical preliminary stage of Enumeration Area (EA) production; it is a critical methodological component that directly determines the statistical validity and spatial consistency of the produced areas. In this context, deviations that may emerge during data preparation should be evaluated not only as implementation errors, but also as methodological risks inherent to the nature of the EA production process. This subsection addresses, at a conceptual level, the main deviation sources that are likely to arise during data preparation operations conducted in ArcGIS Pro and ArcPy, and discusses their potential effects on EA production.

The first major source of deviation encountered in data preparation is boundary effects. Because neighborhood boundaries are often irregular and defined for administrative reasons, they may not coincide exactly with building geometries. Buildings that partially intersect the neighborhood boundary may be either fully excluded or only partially accounted for during spatial summarization or areal representation operations. This can cause housing-unit information to change unexpectedly after spatial transformations. Boundary effects should be considered a methodological risk in all cases where the spatial units used in EA production necessarily intersect administrative boundaries.

The second major source of deviation is geometric validity problems. Building datasets produced at the urban scale may include various geometric inconsistencies, such as self-intersections, overlapping polygons, sliver geometries, or unclosed boundaries. Such geometric problems can produce unexpected outputs, particularly when tools such as Dissolve, Summarize Within, or neighborhood analysis are applied.

Geometric validity issues are often not detectable through visual inspection alone; however, they can lead to methodological inconsistencies through their indirect effects on quantitative outputs.

A third source of deviation concerns the behavior of spatial summarization tools in edge cases. Tools such as Summarize Within rely on certain assumptions regarding when a geometry is considered "inside" or "outside." In edge cases—such as point touches, partial intersections, or very small area overlaps—these assumptions can lead to decisions that are not explicitly controllable by the user. This, in turn, can prevent quantitative information from being fully preserved throughout the spatial analysis chain.

A common characteristic of these deviation sources is that they often do not appear as a single, isolated error, but rather as small inconsistencies that accumulate across successive operations. For this reason, deviations that may occur during data preparation must be monitored not only through final outputs, but also through methodological controls applied at intermediate stages. In this thesis, using the total housing-unit (RES) value as a diagnostic reference variable was developed as a methodological strategy aimed at making such risks visible at an early stage.

In conclusion, deviations encountered during data preparation were treated in this thesis not as "errors," but as a natural consequence of the complex and multi-stage nature of EA production. Systematic monitoring of these deviations and their evaluation with methodological awareness are critical for making the EA production process more predictable, explainable, and reproducible. This approach is also consistent with quality assurance and data integrity principles recommended in geographic information systems (Longley et al., 2015).

The data preparation steps and methodological risks discussed in this section demonstrate that the EA production process depends not only on statistical thresholds, but also strongly on forms of spatial representation. Cleaning, classifying, and diagnostically monitoring housing-unit information at the building level are necessary for EA production, but they are not sufficient on their own. For enumeration areas to be usable in field applications, this quantitative information must be transformed into gap-free, topologically valid, and administratively consistent areal representations within neighborhood boundaries.

This requirement moves EA production beyond the question of "which units should be grouped together" and also makes methodologically necessary the question of "how these groups should be represented spatially." Although the previous steps conducted in ArcGIS Pro and ArcPy enabled building-based data to be controlled in quantitative terms, the problem of areal production had not yet been resolved at this stage. Therefore, in the next phase, the use of spatial precursor structures that can provide gap-free coverage within neighborhood boundaries and enable polygon representation of EAs was evaluated.

In this context, the fishnet (grid)-based areal production approach was considered as the first systematic solution attempt for the areal representation of EAs. In the next section (3.4), the reasons why the fishnet approach appeared attractive for EA production, how it was implemented, and which methodological problems it introduced are examined in detail.

3.3.5. Data Sources, OSM-Derived Inputs, and Cross-Layer Consistency

Data preparation constitutes a decisive phase in the Enumeration Area (EA) delineation process, often exerting a greater influence on final outcomes than the aggregation algorithm itself. In complex urban environments, deficiencies in data structure, spatial consistency, or attribute completeness propagate downstream, resulting in unstable geometries, infeasible EA configurations, and excessive manual intervention.

In this study, data preprocessing is not treated as a purely technical prerequisite but as an integral methodological component. The preprocessing strategy is explicitly designed to support the barrier-aware, target-driven EA delineation framework introduced. Each preprocessing step is therefore motivated by a specific operational or methodological requirement, rather than by generic GIS conventions.

Three guiding principles shape the preprocessing workflow. First, data integrity must be preserved to ensure that every building and household-equivalent unit is accounted for exactly once. Second, topological robustness must be ensured to prevent geometry-related failures during atomic unit generation. Third, process

reproducibility must be maintained at metropolitan scale, enabling the workflow to be restarted, audited, and extended without loss of consistency.

These principles are particularly relevant in the context of Ankara, where heterogeneous building typologies, large administrative extents, and dense transportation networks combine to create a challenging data environment.

OpenStreetMap (OSM)-derived inputs bring both advantages and structural risks. OpenStreetMap (OSM) is an attractive alternative because it provides a highly up-to-date and freely accessible spatial data source worldwide. In particular, building footprints, road networks, and administrative boundaries provide critical inputs for EA production. However, OSM is not a system designed for the production of official census data. This brings several structural risks in using the data.

Studies in the literature on the accuracy and completeness of OSM show that the data can provide high accuracy in urban areas, but it may also contain anomalies that are not expected for the intended use (Haklay, 2010). The problems encountered in this thesis make this general observation concrete at the neighbourhood scale.

A related issue concerns scale mismatch and cross-layer consistency. In the case of Bağlıca Neighbourhood, the building and road data downloaded from OSM are the product of a schema designed at the national scale. When these data are processed at the neighbourhood scale, excessive detail, a high number of geometries, and topological complexity emerge. In particular, a large number of small building polygons and complex road nodes dramatically increased the computational cost of geometric operations.

This corresponds to the problem referred to as "scale mismatch" in the literature (Goodchild, 2011). Scale mismatch arises from the difference between the context in which the data were produced and the context in which they are used, and it typically results in performance bottlenecks.

From an operational perspective, the EA delineation workflow relies on three primary data categories: administrative boundary data, building-level attribute data, and physical barrier data. Each category plays a distinct role in the methodological framework.

First, administrative boundary data define the spatial scope within which EAs are generated. In this study, neighbourhood boundaries constitute the primary

administrative units for EA delineation. These boundaries serve two purposes. First, they ensure alignment with existing administrative and statistical reporting units. Second, they provide an upper-level spatial constraint that prevents EAs from crossing neighbourhood borders.

Neighbourhood boundaries are treated as hard constraints rather than soft guidelines. All subsequent spatial operations—including barrier filtering, atomic unit generation, and EA aggregation—are performed within the confines of these boundaries. This decision reflects standard census practice, where EAs are nested within established administrative hierarchies (United Nations, 2017).

Each neighbourhood is uniquely identified by a persistent identifier, which enables consistent linkage across datasets and supports restartable processing at scale.

Second, building-level data form the quantitative foundation of the EA delineation process. Rather than relying on population estimates aggregated to coarse spatial units, this study uses detailed building attributes to derive household-equivalent counts.

Each building record contains multiple attributes representing the number of independent units by usage type. Attributes beginning with the prefix *kt_* denote distinct functional categories. Two categories—residential units (*kt_1110*) and residential-purpose units (*kt_11*)—are identified as primary contributors to enumeration workload. All other *kt_* categories are explicitly retained in the dataset but contribute zero to the household-equivalent target count.

This design choice ensures that non-residential buildings are spatially represented within EAs without artificially inflating target sizes. It also guarantees that no building is excluded from the EA framework, thereby preserving full spatial coverage.

Building-level data are spatially linked to neighbourhoods through a common identifier. This linkage is validated during preprocessing to detect orphan records and ensure consistency between administrative and building datasets.

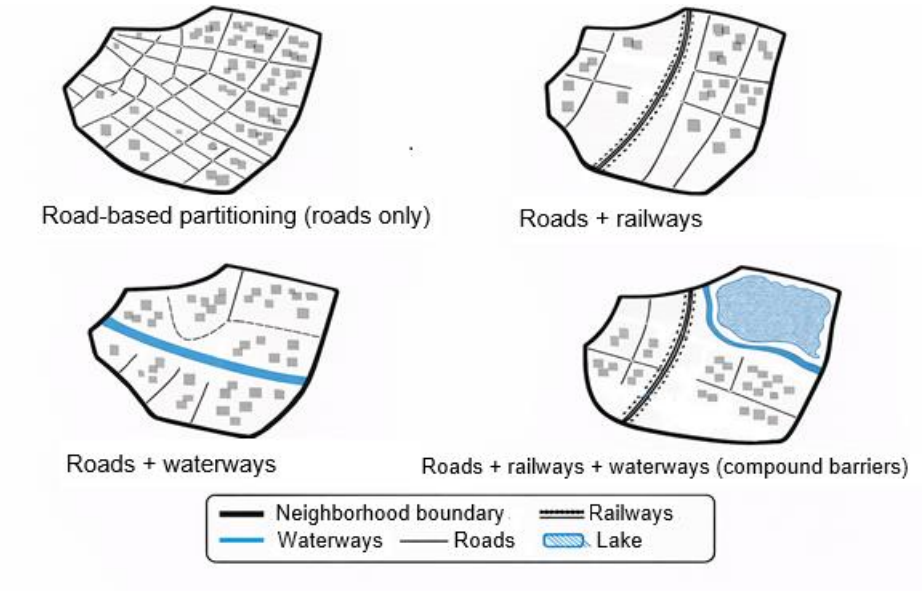
Third, physical barrier data are derived from OpenStreetMap (OSM), an openly available and continuously updated spatial dataset. OSM provides detailed representations of transportation and hydrological features that are critical for barrier-aware EA design.

Three barrier categories are utilised:

- Major road classes (motorways, trunk roads, primary and secondary roads)
- Railways
- Waterways

These categories are selected based on their operational relevance for census fieldwork. Minor streets and local access roads are intentionally excluded to avoid over-fragmentation, consistent with international recommendations for census geography design (United Nations, 2017). In some neighborhoods, the road network is the main determinant, while in others railways, waterways, or water surfaces directly shape EA production. Taking this diversity into account, the EA algorithm handles barriers as combinations.

Figure 3.6. Physical barrier combinations



Source: Generated by the author.

The four basic combinations shown in Figure 3.6 visually summarize this context-sensitive approach. Road-based partitioning represents the baseline case in neighborhoods with a regular road network. In this setting, EAs tend to form around the natural cells created by the road network.

In neighborhoods where roads and railways coexist, the railway is treated as a hard and impermeable barrier. Built development is expected to be limited around

railways, and these areas are typically evaluated separately in the EA growth process. This approach is consistent with considerations of field accessibility and safety. Barrier data are used exclusively as geometric constraints and do not contribute attribute information to EA statistics.

3.3.9. DEGURBA Integration and Target Definition Preprocessing

A distinguishing feature of the preprocessing workflow is the integration of the DEGURBA classification prior to EA generation. DEGURBA (Degree of Urbanisation) is a settlement typology that classifies areas along an urban–rural continuum using population grid cells and density thresholds (Eurostat, 2024; see also Section 2.3). In this study, DEGURBA functions as the contextual variable that determines the target size thresholds applied to each neighbourhood. For this reason, DEGURBA integration is not treated as a secondary descriptive step, but as a core preprocessing component that directly shapes the subsequent aggregation process.

The first stage of this integration involves the attribution of DEGURBA values to neighbourhoods using an external reference table that links administrative units to their degree of urbanisation. This attribution is performed through a non-spatial join based on standardised district and neighbourhood names. Because administrative naming conventions are not always fully consistent across datasets, extensive validation is carried out to detect mismatches arising from spelling variations, abbreviations, formatting differences, or typographical inconsistencies. This validation step is important for ensuring that the urbanisation context assigned to each neighbourhood is both accurate and operationally usable within the EA production workflow.

In addition to standard matching procedures, special handling is applied to organised industrial zones (OSB). Although such areas may not always align neatly with conventional urban-rural classifications, they are systematically classified as rural for enumeration purposes because of their low residential density and limited relevance as household-based enumeration environments. This rule-based override ensures that DEGURBA is interpreted in a way that is consistent with the operational

requirements of census geography rather than being applied only in a purely nominal or administrative sense.

Neighbourhoods that remain without DEGURBA attribution after automated matching are flagged for further review. This makes it possible to identify unresolved naming inconsistencies or limitations in the reference table before the target-setting stage begins. Through this combination of automated matching, validation, and manual review where necessary, the final dataset is produced without missing DEGURBA values. As a result, target thresholds can be assigned to all neighbourhoods in an unambiguous and internally consistent manner.

Once DEGURBA values have been assigned and verified, neighbourhood-specific target ranges are derived. These ranges define the minimum and maximum household-equivalent counts used for EA aggregation. Importantly, the target ranges are assigned at the neighbourhood level rather than globally across the entire study area. This distinction is methodologically significant because it allows the aggregation framework to respond to local settlement structure and degree of urbanisation instead of imposing a single uniform rule on areas with very different spatial and residential characteristics.

By incorporating DEGURBA into preprocessing in this way, the EA aggregation algorithm operates from the outset with explicit and context-sensitive constraints. This improves the internal coherence of the workflow and reduces the need for post hoc corrections after EA generation. In methodological terms, DEGURBA integration therefore serves as the bridge between the administrative context of the study area and the operational logic of workload-based EA design.

3.3.10. Coordinate Reference Systems and Topological Consistency

Spatial preprocessing places particular emphasis on coordinate reference system (CRS) management and topological consistency. All geometric operations involved in EA delineation, especially line union, snapping, and polygonization, are highly sensitive to CRS choice. For this reason, this section addresses three closely related issues that directly affect the reliability of the preprocessing workflow: the use of a projected CRS for geometry-intensive operations, the treatment of invalid and

near-degenerate geometries, and the implications of adopting planar assumptions when working with data that are originally stored in geographic coordinates.

A first consideration concerns the CRS framework within which geometric processing is conducted. Geographic coordinate systems based on longitude and latitude are unsuitable for planar topological operations because they are affected by angular distortion and can introduce numerical instability in operations that depend on distance, adjacency, and boundary alignment. Since EA delineation requires repeated execution of such operations, including the construction of line networks, boundary integration, snapping, and polygonization, all spatial datasets are transformed into a projected CRS appropriate for the study area before any geometry-intensive processing is performed. For Ankara, a transverse Mercator projection (EPSG:5254) is used. This projection minimises distortion across the metropolitan extent and supports metre-based distance and area calculations, thereby providing a more stable basis for topological processing. Accordingly, all barrier processing, atomic unit generation, and EA aggregation are carried out exclusively in this projected CRS environment.

A second consideration relates to the internal geometric quality of the input data. Urban spatial datasets frequently contain minor geometric inconsistencies, including self-intersections, nearly coincident vertices, duplicated segments, and sliver polygons. Although such issues may appear negligible at first inspection, they can create substantial problems during automated preprocessing, particularly when the workflow relies on the exact closure of linework and the valid construction of polygonal units. If left unaddressed, these inconsistencies may cause polygonization to fail, generate incomplete faces, or produce invalid atomic units that cannot be used reliably in subsequent EA aggregation steps. To mitigate these risks, preprocessing includes explicit geometry validation and regularisation procedures. Line geometries are snapped to a fine grid prior to polygonization in order to remove micro-gaps and improve line continuity, while invalid geometries are corrected using standard topological repair functions. These interventions are not merely technical refinements; they are essential for ensuring that atomic unit generation remains stable, reproducible, and comparable across all neighbourhoods included in the study.

A third issue concerns the relationship between planar computation and geodesic reality. Many of the topological operations implemented in the R sf package operate, by default, under an assumption of planar geometry. In contrast, OSM-derived data are generally stored in a geographic coordinate reference system based on latitude and longitude. This mismatch led to warnings such as "st_union assumes planar" during processing. As Pebesma (2018) emphasises, the distinction between planar operations and s2-based geodesic operations within the sf ecosystem is particularly important when spatial data are large, complex, and topologically demanding. In the present study, however, geodesic accuracy was treated as a secondary concern in comparison with computational stability, reproducibility, and the practical feasibility of large-scale EA production. Planar assumptions were therefore accepted in a controlled and methodologically explicit manner. This choice was justified by the scale and operational purpose of the analysis, since the principal objective was not geodetic precision in a strict cartographic sense, but the stable production of contiguous and topologically consistent operational units. Nevertheless, as discussed in later sections, the controlled acceptance of planar assumptions also contributed to some unexpected outcomes in the processing results.

Taken together, these considerations show that CRS selection and topological consistency are not peripheral technical matters, but central conditions for the successful implementation of the preprocessing workflow. The stability of atomic unit generation, the validity of polygonized outputs, and the reproducibility of EA delineation all depend on maintaining an appropriate projected CRS, repairing problematic geometries, and adopting a coherent computational framework for spatial operations. In this sense, CRS management and topology control form a foundational layer of the overall methodology rather than a routine preparatory step.

3.3.11. Large-Scale Processing, Restartability, and Preprocessing Outputs

Processing an entire metropolitan area such as Ankara in a single, uninterrupted workflow poses substantial computational and operational risks. The volume of spatial data, combined with geometry-intensive operations, increases the likelihood of software interruptions, memory exhaustion, and unexpected topological failures.

These risks become even more critical when preprocessing is designed not merely as a one-time technical exercise, but as a reproducible workflow intended to support repeated testing, parameter adjustment, and large-scale implementation. For this reason, the preprocessing strategy adopted in this study is explicitly designed to support incremental execution and restartability.

A central feature of this strategy is the decomposition of the metropolitan area into smaller and operationally manageable units. Rather than treating Ankara as a single spatial entity, the preprocessing workflow decomposes the metropolitan area into neighbourhood-level units. Each neighbourhood is processed independently, producing self-contained outputs that can be reused in downstream stages. This design choice is important not only from a computational perspective, but also from a methodological one, because it allows the workflow to maintain continuity even when local processing problems arise. In other words, the structure of the workflow is intentionally organised so that the failure of one unit does not invalidate the processing logic or outputs associated with the others.

This decomposition yields several advantages. First, it localises errors: failures encountered in one neighbourhood do not compromise the entire processing run. This is particularly valuable in large-scale spatial workflows, where the probability of local geometric anomalies or unexpected data-specific issues cannot be assumed to be negligible. Second, it enables parallel inspection and validation of intermediate outputs. Since each neighbourhood produces a separate and interpretable result, quality control can be conducted progressively rather than postponed until the end of a full metropolitan-scale run. Third, it allows the overall workflow to be paused and resumed without reprocessing completed neighbourhoods. This greatly improves the practical usability of the method, especially in environments where long-running processes may be interrupted by software instability, system limitations, or the need for interim methodological revisions. The neighbourhood-based approach also aligns with established practices in large-scale census geography production, where hierarchical processing is preferred to monolithic execution (United Nations, 2017).

The operational logic of restartability is further supported by explicit process monitoring. In addition to spatial outputs, the preprocessing workflow maintains a progress log recording the status, processing time, and completion timestamp for each

neighbourhood. This log enables real-time monitoring of long-running processes and also supports post hoc performance analysis by making it possible to identify which units required longer runtimes or produced interruptions. In this sense, the progress log serves both as a technical monitoring instrument and as a diagnostic record of workflow behaviour across the study area.

Building on this logging structure, checkpointing is implemented by comparing completed neighbourhood identifiers against the progress log at runtime. Neighbourhoods already marked as completed are automatically skipped, allowing the workflow to resume precisely from the point of interruption. This mechanism ensures that restartability is not dependent on manual tracking or ad hoc intervention, but is embedded directly into the logic of the preprocessing routine. As a result, the workflow can recover efficiently from interruptions while preserving the integrity of outputs that have already been generated.

Taken together, these design choices show that large-scale processing in this study is approached not as a single continuous run, but as a controlled and modular workflow architecture. Incremental execution, neighbourhood-based decomposition, progress logging, and checkpoint-based restartability jointly provide the operational resilience required for metropolitan-scale preprocessing. This makes the workflow more robust, more transparent, and more suitable for reproducible EA production under real-world computational conditions.

A further dimension of this large-scale processing strategy concerns performance considerations and computational trade-offs. Beyond spatial and statistical diagnostics, the pre-analysis stage also includes a systematic examination of computational performance. Given the scale of the Ankara dataset and the geometric complexity introduced by barrier-aware operations, runtime behaviour is a critical design consideration.

Initial exploratory runs reveal that spatial clipping, intersection, and polygonisation operations constitute the dominant share of computational cost. In particular, repeated clipping of large OpenStreetMap (OSM) datasets to neighbourhood boundaries is identified as a major bottleneck. These operations scale nonlinearly with both neighbourhood size and barrier density.

Pre-analysis timing experiments further demonstrate substantial variability in processing time across neighbourhoods. Dense urban neighbourhoods with complex barrier networks require significantly longer processing times than smaller or more homogeneous areas. This variability highlights the impracticality of monolithic, city-wide processing pipelines.

As a result, performance considerations emerge as a first-order design constraint rather than a secondary optimization concern. The methodology must be structured in a way that accommodates long runtimes, supports interruption and resumption, and provides continuous feedback on progress.

The findings of the pre-analysis stage directly inform several key design decisions formalized in the methodology.

First, the decision to preprocess and store neighbourhood-level datasets independently is motivated by both spatial logic and computational efficiency. By performing expensive geometric operations once and reusing the results, the methodology significantly reduces redundant computation.

Second, the introduction of incremental output writing and progress logging addresses the observed runtime variability. These mechanisms ensure that partial results are preserved even if processing is interrupted, enabling robust long-running execution.

Third, the decision to avoid aggressive optimization during the primary EA delineation stage reflects insights gained from pre-analysis. Attempting to enforce strict target compliance in structurally constrained neighbourhoods would increase algorithmic complexity without guaranteeing meaningful improvement in outcomes.

Finally, the pre-analysis supports the choice to decouple area-based refinement from the core methodology. By deferring secondary subdivision to a later stage, the primary algorithm remains interpretable and grounded in empirically observed constraints.

A related output of this strategy is the production of persistent preprocessing datasets for each neighbourhood. For each neighbourhood, preprocessing produces an intermediate spatial dataset containing the neighbourhood boundary, DEGURBA classification, and summary statistics such as the number of buildings. These

intermediate outputs are stored as persistent files and serve as stable inputs for subsequent EA delineation stages.

Crucially, these outputs are not transient objects held only in memory. By writing them to disk, the workflow ensures that progress is preserved even in the event of system failure or forced termination. This design decision proved essential in practice, as full preprocessing of Ankara required extended computation time.

The pre-analysis stage establishes a detailed empirical foundation for the EA delineation methodology developed in this thesis. Through systematic exploration of the Ankara dataset, it identifies the spatial structures, statistical distributions, and computational constraints that shape feasible EA design.

Key findings include the presence of extreme building-level values, strong barrier-induced fragmentation, substantial variation across DEGURBA classes, and significant runtime heterogeneity. These findings collectively demonstrate that EA delineation cannot be approached as a purely numerical problem.

By explicitly linking pre-analysis outcomes to methodological choices, this chapter ensures that the design decisions presented in Section 4 are both empirically justified and transparent. The transition from diagnostics to formal methodology thus represents a progression from observation to structured action rather than an abstract leap.

3.4. ArcGIS Pro–Based EA Production Trials: Fishnet Discretisation, Region Growing, and Failure Analysis

The ArcGIS Pro trial was conducted as an initial, single-neighbourhood experiment to test whether off-the-shelf zoning functionality could produce workload-balanced, barrier-compliant EAs with minimal custom development. The limitations observed in this environment motivated the transition to a fully reproducible, script-based workflow in R, first stabilised at the neighbourhood scale and subsequently generalised for district-scale processing.

In this section, following the completion of the data preparation and preprocessing stages, the fishnet (grid)-based approach implemented in the ArcGIS Pro environment for Enumeration Area (EA) production is examined in detail. The

fishnet approach was initially evaluated as a strong and attractive method for EA production because it provides full spatial coverage within neighborhood boundaries and because its initial units are geometrically regular.

Within this ArcGIS trial, “fishnet-based production” refers to the discretisation of the study area into uniform cells and the allocation of RES to those cells, whereas “region growing” refers to the subsequent graph-based aggregation of cells into candidate EAs. They are therefore treated as sequential components of a single ArcGIS-stage approach, rather than competing methods.

The data preparation steps described in the previous section (Section 3.3) enabled housing-unit information to be defined consistently at the building level and allowed quantitative integrity to be diagnostically monitored throughout the workflow. This preparation ensured that the spatial units to be used for EA production were no longer only geometric objects, but also analytical entities carrying housing-unit information. The fishnet approach was the first method selected to place this quantitative information within an areal framework.

The main assumption of the fishnet-based approach is that, if the study area is divided into equal-sized cells, these cells can later be merged in line with specific statistical constraints (e.g., 80–120 housing units). This assumption aligns with regular grid-based methods widely used in the GIS literature, especially in spatial sampling, density analysis, and zoning studies (Longley et al., 2015; Openshaw, 1984). In the context of EA production, this approach was considered methodologically advantageous because it enables complete and gap-free areal representation within neighborhood boundaries.

Nevertheless, the fishnet approach was evaluated not only through theoretical assumptions, but also through the behaviors observed during implementation. The objective of this section is not to label the fishnet method as “successful” or “unsuccessful,” but rather to systematically identify under which conditions it operates, where it becomes unstable, and which structural issues it makes visible in terms of EA production. This evaluation represents a critical intermediate stage for understanding why region-growing algorithms and building-based approaches became necessary in later sections.

Accordingly, Section 3.4 addresses the fishnet approach step by step by examining (i) its theoretical rationale, (ii) how it was implemented in the ArcGIS Pro environment, (iii) the logic of cell-based housing-unit calculation, and (iv) the methodological limitations that emerged during implementation. The intermediate assessment presented at the end of the section clarifies the role of the fishnet-based approach in the EA production process and provides a methodological basis for the transition to the next stage.

3.4.1. Theoretical Rationale of the Fishnet Approach

Fishnet (grid)-based approaches have long been used in GIS to standardize spatial analysis and reduce it to controllable units. The core assumption of this approach is that, if the study area is partitioned into regular and equal-sized cells, spatial operations conducted over these cells become more predictable both algorithmically and statistically. In the context of EA production, fishnet was considered a theoretically attractive starting point, particularly because it directly addresses the problem of areal representation.

One of the key requirements in enumeration area design is full coverage of the study area and the absence of gaps or overlaps between EAs. At the neighborhood scale, where administrative boundaries are irregular and complex, meeting this requirement can necessitate additional correction steps when working directly with building or parcel geometries. The fishnet approach, by guaranteeing that every point within the neighborhood boundary belongs to a cell, provides gap-free spatial coverage. This property was seen as an important advantage in terms of operational requirements related to the areal representation of EAs.

A second major rationale of the fishnet approach is the geometric homogeneity of the initial units. Because all cells have the same size and shape, the initial conditions of region-growing or merging algorithms used for EA production become standardized. This aims to ensure that the behavior of the algorithm is driven by defined rules and parameters rather than by the geometry of the input data. The literature also emphasizes that regular grids can simplify complex spatial patterns and facilitate analytical processes (Openshaw, 1984; Longley et al., 2015).

Another theoretical advantage of the fishnet approach is that it enables parametric control. Cell size functions as an indirect but powerful control variable in EA production. Reducing cell size increases the potential to produce more detailed and flexible EA boundaries, whereas increasing cell size results in fewer and coarser initial units. This property makes it possible to systematically test different scenarios rather than assuming a single “correct” solution. In this respect, fishnet was considered a methodological experimentation space.

A further theoretical strength of regular grid structures is that neighborhood relationships between cells are relatively simple and predictable. In a fishnet composed of square cells, the number of possible neighbors per cell is limited and these neighborhood relationships can be defined topologically. This was considered a factor that can simplify computation for region-growing algorithms to be applied in later stages. In particular, in neighborhood-based algorithms, having a regular initial graph is typically regarded as a methodological advantage.

Based on these rationales, the fishnet approach was adopted as the first main method for EA production. However, this choice was not based on the assumption that fishnet constitutes the final solution for EA production. On the contrary, the aim was to systematically evaluate the extent to which the theoretical advantages of the approach are realized under practical implementation conditions. For this reason, within the scope of this thesis the fishnet approach is treated both as a method and as a comparative reference framework for alternative approaches developed in later stages.

In the next subsection (3.4.2), the implementation of the fishnet approach in ArcGIS Pro, the parameters used for cell generation, and how these parameter choices are reflected in the EA production process are explained in detail.

3.4.2. Fishnet Generation Process and Parameter Selection

The implementation of the fishnet-based EA production approach is based on generating a regular grid structure within the neighborhood boundary included in the analysis in the ArcGIS Pro environment. The main purpose of this stage is to define the initial spatial units to be used in the EA production process in an explicit, reproducible, and parametrically controllable manner. For this reason, fishnet generation was treated not merely as a technical preprocessing step, but as a methodological decision that directly influences the behavior of all subsequent steps. Working with a regular grid was intended to provide "controllability" and "scenario comparability," and in particular to enable repeated runs in the same neighborhood with different parameter settings.

Fishnet generation was performed using the Create Fishnet tool in ArcGIS Pro. This tool creates a grid composed of equal-sized rectangular or square cells within a user-defined extent. In this thesis, the extent was defined as the administrative boundary polygon of the neighborhood to be analyzed. In this way, it was intended that the fishnet structure would not generate cells outside neighborhood boundaries and would overlap with the analysis area as closely as possible. Nevertheless, due to the irregular structure of administrative boundary geometries, partial cutting of cells along the boundary line can be unavoidable. This causes the theoretical regularity of the fishnet approach to be disrupted in practice and produces a structural mechanism that should be treated as a "boundary effect" (discussed in detail below).

The main parameters used for fishnet generation were configured as follows:

• Extent: Neighborhood boundary polygon

• Cell shape: Square

• Coordinate system: Metric projection consistent with the neighborhood dataset

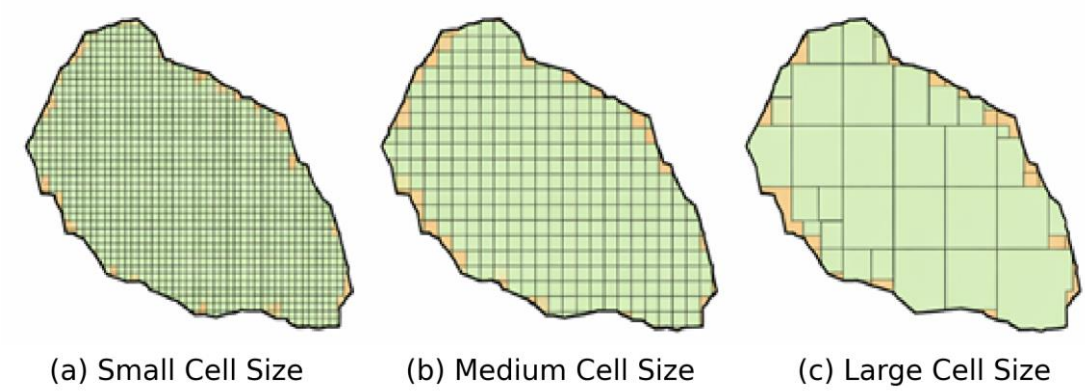
• Cell size: A fixed value determined experimentally

Among these parameters, cell size stands out as the most critical and sensitive component of the fishnet approach. Cell size functions like a meta-parameter that determines (i) the number of initial units, (ii) the density of the neighborhood graph,

and therefore (iii) the “order and direction” in which the region-growing algorithm will expand. In other words, cell size shapes not only geometric resolution, but also the operating regime of the merging/neighborhood logic used for EA production. For this reason, cell size selection was treated not as a one-time setting, but as a methodological design variable, and it was tested systematically across different values.

Within this scope, the effect of cell size on spatial representation was evaluated not only through numerical outputs, but also by comparing geometric representation behavior. For the same neighborhood boundary, fishnet structures generated with different cell sizes were compared in terms of cell count, fragmentation along the boundary, and representational resolution. This comparison indicates that there is no single “ideal” cell size in the fishnet approach; rather, cell size should be treated as a methodological sensitivity parameter in EA production (Figure 3.7).

Figure 3.7. Effect of different cell sizes on fishnet structure and spatial representation.



Source: Generated by the author.

During implementation, cell size was tested through trial and error across different values. Relatively small cell sizes (e.g., 20–30 m) allowed the settlement pattern within the neighborhood to be represented in greater detail, but they substantially increased the number of initial units. This not only increased computational burden, but also produced a denser topological structure in neighborhood relationships. In contrast, larger cell sizes (e.g., 50 m and above) reduced the number of initial units; however, the cells then produced a coarser representation that masked spatial heterogeneity in the distribution of housing units.

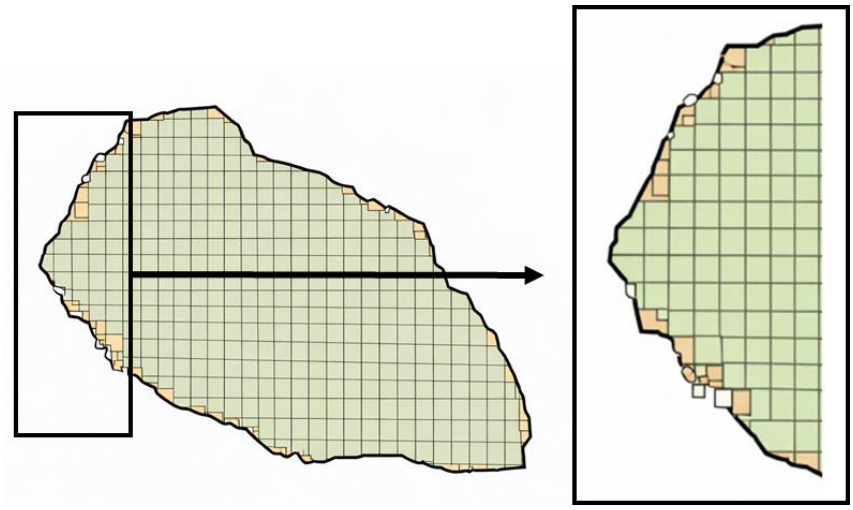
Accordingly, each change in cell size altered not only the “number of cells,” but also the operating conditions of the neighborhood-based algorithms to be used for EA production.

These observations are consistent with the literature emphasizing that scale selection in regular grid-based analyses can fundamentally change outcomes. In the context of the “scale problem,” it is noted that different resolutions applied to the same dataset may produce different spatial results (Openshaw, 1984; Longley et al., 2015). In the context of this thesis, this scale problem demonstrates that fishnet cell size is not merely a technical parameter, but a decision that shapes the methodological orientation of EA production.

Another methodological issue considered in fishnet generation is the geometric irregularity of the neighborhood boundary. The indented and protruding structure of administrative boundaries can cause fishnet cells to be partially cut along the boundary or split into very small fragments. These boundary cells can become “weak” initial units both geometrically (very small or thin polygons) and statistically (often carrying low or zero housing units). Therefore, the regularity that the fishnet approach theoretically provides can be partially undermined in practice due to boundary effects.

The main reason boundary effects are critical for EA production is that these weak cells can unexpectedly influence EA growth decisions in later steps (for example, neighborhood graph construction and region-growing algorithms). How boundary effects emerge in practice in the fishnet approach is illustrated in Figure 3.8 through the interaction between the neighborhood boundary geometry and the regular fishnet structure. The figure clearly shows that some cells along the boundary are partially cut and transformed into very small polygons with weak functional integrity. Although these cells are part of a geometrically regular grid structure, they often carry low or zero housing units (0-RES) in terms of statistical content. This situation methodologically weakens the fishnet approach’s assumption of “equivalent initial units.”

Figure 3.8. Boundary effects of neighborhood boundary geometry on fishnet cells.



Source: Generated by the author.

On the left, the fishnet structure generated over the neighborhood boundary is presented; on the right, a zoomed-in area is provided in which the cells that were cut along the boundary and whose functional integrity has weakened are more clearly visible.

At this stage, the fishnet cells were not yet treated as EAs. The generated grid was considered only as a precursor spatial structure, and it was acknowledged that, for the cells to become meaningful for EA production, building-based housing-unit information had to be transferred to these cells. For this reason, fishnet generation was planned as a geometric step independent of quantitative information, and the statistical dimension of the EA production process was deferred to the next stage.

To document methodologically the effect of cell size on the EA production process, a series of trials was conducted using different cell sizes during fishnet generation. The purpose of these trials was not to identify a single “best” cell size, but rather to observe how cell size influences the number of initial units, boundary fragmentation, and the way in which the distribution of housing units is reflected in the cells. The summary indicators derived in this context were evaluated not to produce quantitative results, but to document, at the methodological level, the parameter sensitivity of the fishnet approach.

Table 3.1. Methodological summary indicators for fishnet generation across different cell sizes

Cell size (m)	Total number of cells	Mean RES per cell	Share of 0-RES cells (%)	Methodological observation
25	↑ high	↓ low	↑ high	High resolution, excessive fragmentation, and dense neighborhood.
35	medium	medium	Medium	Balanced representation, but boundary effects remain evident.
50	↓ low	↑ high	↓ low	Coarse representation; spatial heterogeneity is masked.

Note: The qualitative terms "high / medium / low" used in the table are not intended for numerical comparison; they are used only to describe methodological tendencies associated with cell size.

This table shows that cell size in the EA production process is not only a geometric preference; it is also a design variable that directly affects the neighborhood structure, the way quantitative information is distributed across cells, and the proportion of weak units created along boundaries. In particular, the change in the share of 0-RES cells indicates that boundary effects and scale selection must be evaluated jointly in the fishnet approach. For this reason, cell size selection was treated as a decisive prerequisite for the subsequent steps of cell-based housing-unit calculation and region growing.

In summary, the fishnet generation process was defined as a methodological stage that aims to meet the basic requirements of areal representation in EA production, but is highly sensitive to parameter choice. Factors such as cell size and boundary effects were observed to determine the operating conditions of the neighborhood and growth operations applied in later stages; therefore, fishnet generation was evaluated as an "initial design" step that shapes the behavior of the entire workflow.

In the next subsection (3.4.3), the logic of cell-based housing-unit (RES) calculation on the generated fishnet cells is discussed in detail, and the role of the spatial summarization step in the EA production process is explained.

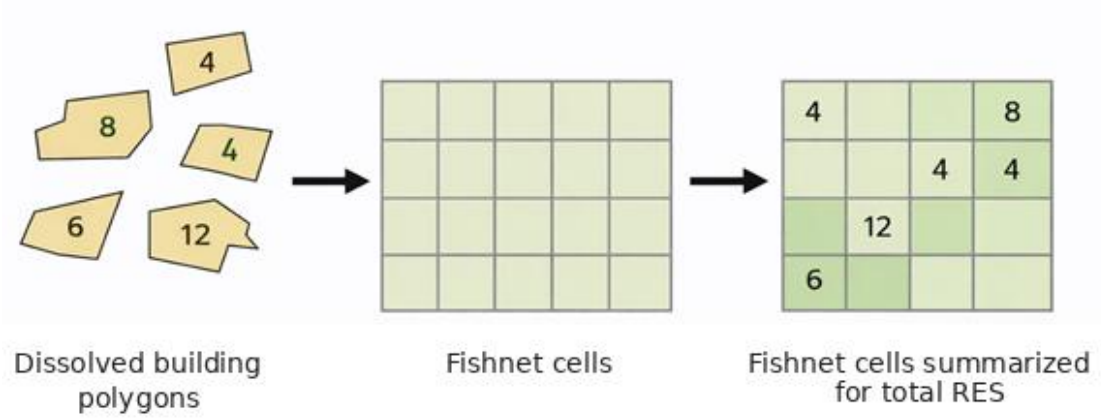
3.4.3. Logic of Cell-Based Housing-Unit (RES) Calculation

After the fishnet generation stage, building-based housing-unit (RES) information must be transferred to the generated cells so that the cells can be converted into analytical units usable for EA production. This step is the core methodological stage that enables the fishnet approach to move beyond being a purely geometric precursor structure and to become an analysis layer carrying quantitative content.

In this thesis, cell-based RES calculation was performed using dissolved building polygons. As described in previous sections, building geometries were merged based on building identity, and the total number of housing units for each building was defined as a single attribute. This preparation aims to prevent housing-unit information from being artificially fragmented during spatial summarization and to ensure that each building contributes to EA production with an appropriate weight.

Cell-based RES calculation was implemented in the ArcGIS Pro environment using the Summarize Within tool. In this operation, the RES values of building polygons falling within each fishnet cell were summed, and the result was assigned to the cell as an attribute. As a result, each fishnet cell gained a quantitative variable representing the total number of housing units within its covered area.

Figure 3.9. Transfer of building-based residential unit (RES) information onto fishnet cells via spatial summarization.

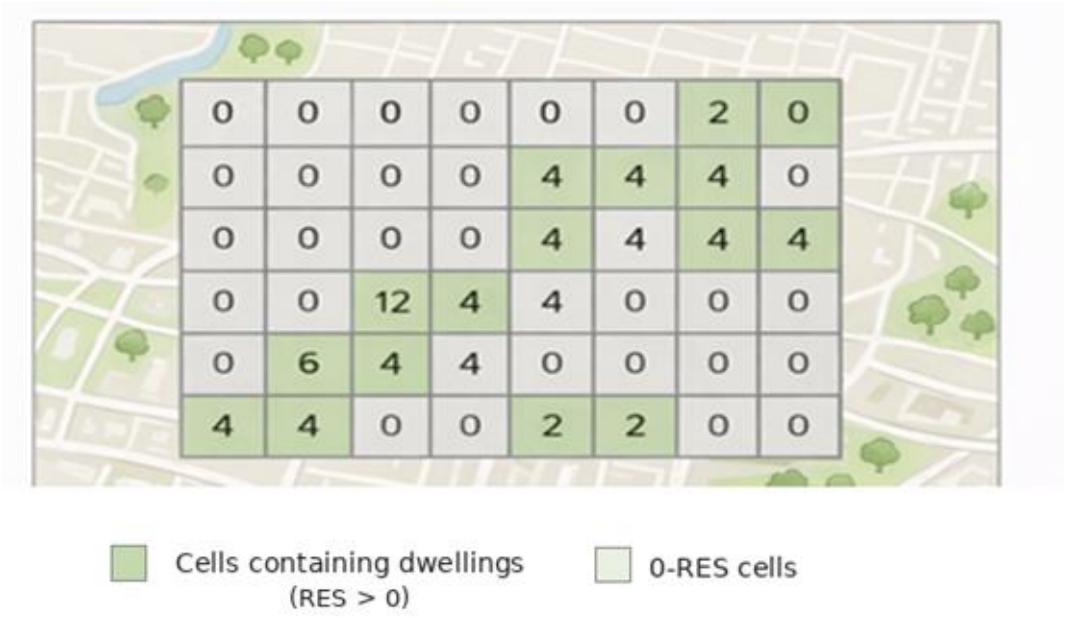


Source: Generated by the author.

The methodological operation of this process can be explained by considering Figures 3.9 and 3.10 together. Figure 3.9 schematically illustrates how building-based housing-unit (RES) information is transferred to regular fishnet cells through spatial summarization (Summarize Within). At this stage, each cell acquires a quantitative attribute by taking the sum of the RES values of the building polygons that fall within its boundaries.

Complementing this basic transfer logic, Figure 3.10 conceptually shows how the same spatial summarization operation produces a heterogeneous RES distribution at the cell level. Although fishnet cells have a regular geometric structure, once quantitative information is transferred they cease to be statistically equivalent units: some cells carry high housing-unit values, whereas cells corresponding to roads, open spaces, or areas without buildings receive a zero housing-unit (0-RES) value.

Figure 3.10. Schematic representation of the spatial distribution of 0-RES cells and residential cells after cell-based RES calculation.



Source: Generated by the author.

In this context, Figure 3.10 makes visible a critical methodological tension inherent in the fishnet approach: even if cells are geometrically equal in size, the RES distribution exhibits high variance across cells, and this directly affects the statistical

quality of the initial units used for EA production. In particular, the presence of 0-RES cells becomes one of the key factors shaping the behavior of the neighborhood-based growth algorithms applied in subsequent stages.

For this reason, the cell-based RES calculation step was evaluated not as a "balancing" mechanism in the EA production process, but rather as an analytical threshold that reveals the spatial and statistical limitations of the fishnet approach. The 0-RES cells that emerge at this stage were not treated as errors or anomalies; they were interpreted as a natural output of the method and as a structural characteristic that must be addressed in later stages.

Accordingly, Section 3.4.3 represents the stage at which fishnet cells become not only geometric units, but also quantitative initial units to be used in EA production. However, this transformation also formed the basis for the methodological instabilities that are discussed in detail in the next subsection.

In the next subsection (3.4.4), it is discussed in detail how the cell-based RES distribution—particularly through 0-RES cells—affects EA production and why this creates structural instability within the fishnet approach.

3.4.4. Structural Effects of 0-RES Cells

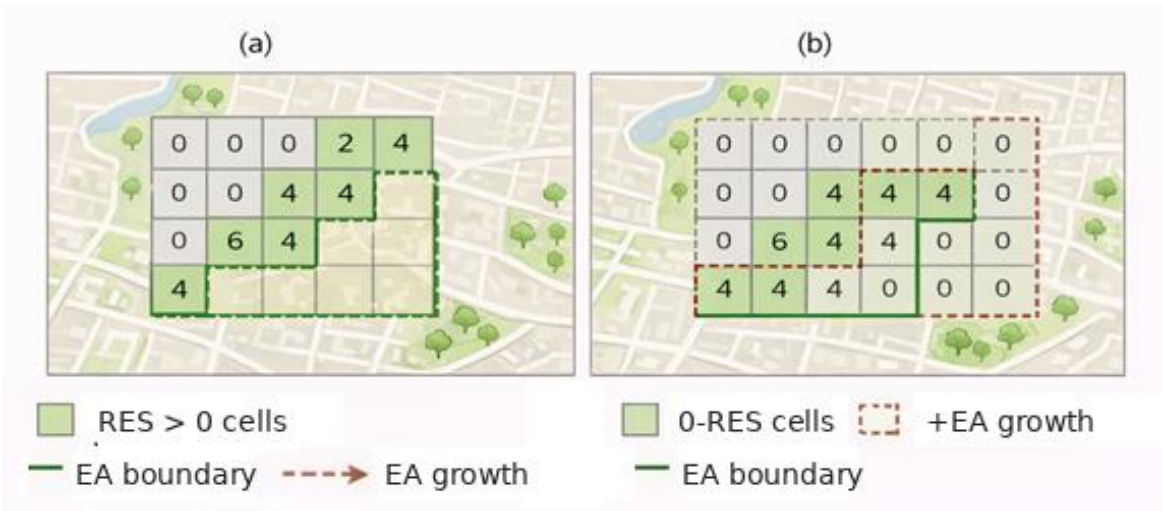
After cell-based housing-unit (RES) values were calculated for fishnet cells, one of the most critical methodological issues determining the behaviour of the EA production process was the presence of 0-RES cells. These cells are an integral geometric component of the fishnet structure, yet statistically they carry no housing units. In other words, 0-RES cells are necessary for spatial continuity, but they are neutral with respect to quantitative targets.

One theoretical assumption of the fishnet approach is that initial units can be treated as equivalent. However, after cell-based RES calculation, this assumption lost its methodological validity. Cells with the same size and shape were observed to differ radically in terms of their capacity to carry housing units. This demonstrates that EA production is not merely a geometric problem, but a multi-layered problem in which statistical, topological, and algorithmic constraints must be managed simultaneously. In this context, the structural effect of 0-RES cells on EA production emerges at three

closely related levels. These concern areal expansion without statistical contribution, the topological bridge effect within the neighbourhood graph, and the distortion of EA geometry together with its operational implications.

The first of these effects is that 0-RES cells can cause areal expansion during EA growth without providing any statistical contribution. Because region-growing algorithms must keep EAs spatially contiguous, selecting only cells with RES is often not feasible. For this reason, the algorithm may be forced to include 0-RES cells in the growth process in order to preserve the spatial integrity of the EA. This leads to an expansion of the area covered by an EA while the total RES value remains unchanged. As a result, EA growth can become effectively decoupled from quantitative targets such as the targeted RES band. While the geometric boundary of the EA expands, its statistical content does not change, producing a methodological tension that disconnects the area-statistics relationship. This effect is illustrated schematically in Figure 3.11. The figure compares two configurations for the same EA: a narrower EA boundary that includes only RES>0 cells, and an expanded EA boundary that results from adding 0-RES cells. The visualisation clearly demonstrates that the total RES value does not change even though the EA area increases, thereby making areal expansion without statistical contribution methodologically visible.

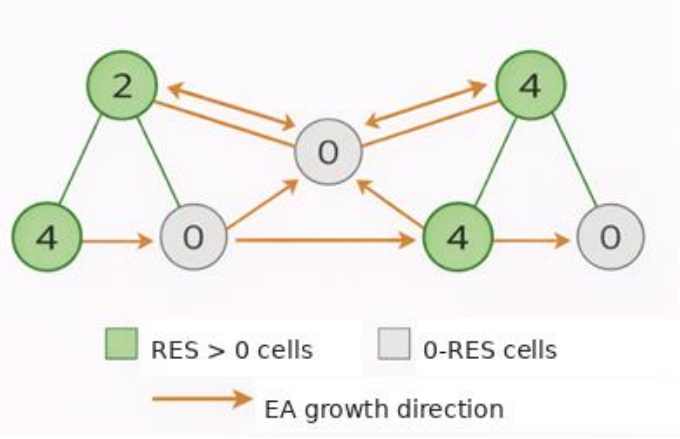
Figure 3.11. 0-RES cells causing areal expansion during EA growth without providing any statistical contribution.



Source: Generated by the author.

A second effect concerns the role of 0-RES cells in steering neighbourhood-based growth decisions. Within the cell-to-cell neighbourhood graph, 0-RES cells often function as topological bridges between multiple RES-carrying cells. This can cause the decision mechanism of the EA growth algorithm to be driven more by topological accessibility than by RES density. In other words, the algorithm may determine its growth direction not on the basis of the cell that provides the highest quantitative contribution, but on the basis of the adjacent cell that is easiest to connect. This shifts the growth process away from a quantitative optimisation logic and makes it dependent on the structural properties of the neighbourhood graph. This methodological issue is illustrated in Figure 3.12 using a schematic neighbourhood graph. The figure shows 0-RES cells providing connectivity between two RES>0 cells and thereby indirectly determining the direction of EA growth. The growth directions indicated by arrows clearly demonstrate that the algorithm prioritises topological continuity over quantitative contribution. This visualisation constitutes critical methodological evidence that the problem arises not from the data itself, but from the topological properties of the fishnet structure.

Figure 3.12. 0-RES cells acting as bridges in the neighborhood graph and thereby indirectly determining the direction of EA growth.

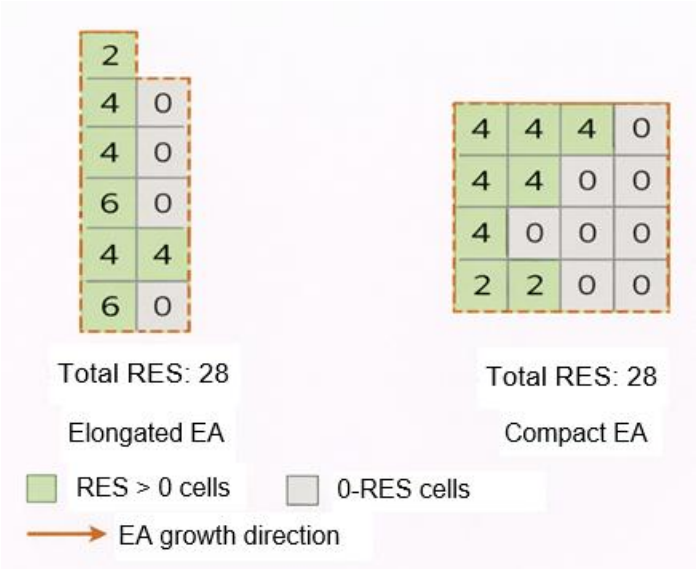


Source: Generated by the author.

A third effect relates to the direct influence of 0-RES cells on the spatial form of EA geometries. In particular, 0-RES cells that concentrate along neighbourhood boundaries, around road networks, or in sparsely built-up areas can cause EAs to take elongated, narrow, or otherwise irregular shapes. This can result in EAs that have

similar RES totals being represented with very different spatial forms. Figure 3.13 presents two EA examples with the same total RES value. The EA on the left has an elongated and narrow geometry due to the influence of 0-RES cells, whereas the EA on the right has a more compact spatial structure. Although the two EAs are statistically equivalent, they differ substantially in geometric and operational terms. This difference matters for field implementation. Elongated and fragmented EA geometries can hinder access for enumeration teams, create ambiguity in area definitions, and reduce operational efficiency. Therefore, the influence of 0-RES cells produces not only methodological consequences, but also practical, implementation-oriented outcomes.

Figure 3.13. EAs with the same total RES value producing different spatial geometries due to the influence of 0-RES cells.



Source: Generated by the author.

Taken together, these observations show that 0-RES cells should not be treated as an error or a data problem, but rather as a natural and unavoidable output of the fishnet-based approach. Spatially representing areas without housing is necessary to produce EAs that are gap-free and topologically valid. However, if these cells are not managed in a way that is consistent with quantitative targets, they can significantly, and often unpredictably, distort the behaviour of the growth algorithm. This indicates

that the problem stems not primarily from data quality, but from the structural properties of the chosen areal representation approach.

Within the scope of this thesis, different strategies regarding 0-RES cells were evaluated. Fully excluding such cells was not considered methodologically appropriate because it risks creating gaps and spatial discontinuities along EA boundaries. Conversely, including 0-RES cells in EA growth without conditions led to systematic deviations of EAs from the targeted housing-unit ranges. This dual outcome clearly demonstrates a structural tension between areal representation and statistical balance in the fishnet-based approach. As a result, the structural effect of 0-RES cells was assessed as one of the key methodological factors explaining why the fishnet approach struggles to produce stable and predictable outcomes in EA production. In the next subsection (Section 3.4.5), the design of the region-growing algorithm applied over fishnet cells, the rules used, and the reasons why this algorithm produced unstable results in practice are discussed in detail.

3.4.5. Fishnet-Based Region Growing Trial and Diagnostic Reporting

Following the completion of data preparation and pre-processing steps, this section describes the fishnet (grid)-based approach implemented in ArcGIS Pro to produce Enumeration Areas (EAs). The fishnet approach was initially considered methodologically attractive because it provides full spatial coverage within neighbourhood boundaries and offers a geometrically regular set of starting units. In principle, a regular grid can support repeatable scenario testing by enabling the same neighbourhood to be processed multiple times under different parameter settings, such as cell size, thereby improving comparability between runs.

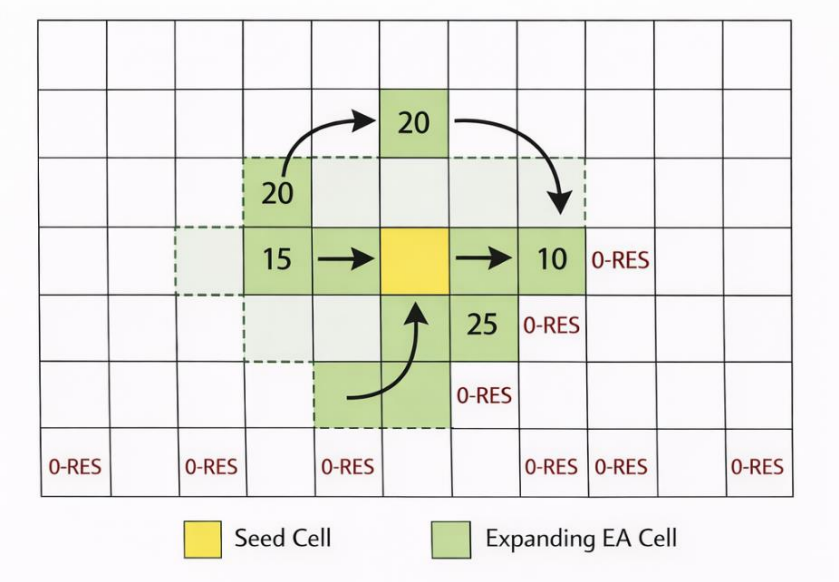
In this workflow, fishnet cells are not treated as EAs by default. Instead, they are defined as precursor spatial units that later gain statistical meaning after residential unit information is transferred from building-based data. Therefore, the fishnet layer is best understood as a controllable spatial frame that supports subsequent aggregation and growth operations under quantitative constraints. Within this overall logic, the discussion below brings together the main components of the fishnet-based trial in an integrated manner: the rationale for adopting a fishnet as an initial spatial frame, the

implementation and observed behaviour of the region-growing algorithm, the role of seed selection and adjacency construction, the decision rules and residual-cell handling procedures, and the reporting protocol used to document and evaluate the resulting outputs. The holistic methodological evaluation and the transition to the subsequent approach are provided in Section 3.4.6.

A central stage in the fishnet-based trial is the implementation of a region-growing algorithm designed to produce contiguous EAs under a target residential band, for example a desired interval such as 80–120 residential units. Region growing was selected because it can iteratively aggregate neighbouring cells to form contiguous zones while tracking a running quantitative sum. In the implemented logic, an EA begins from a seed cell and expands by adding neighbouring cells until the cumulative RES reaches, or approaches, the target band. At each iteration, the algorithm evaluates a candidate set of neighbours based on pre-defined rules, including contiguity, adjacency, and optional preferences. In theory, such a process can balance quantitative targets and spatial coherence; however, in practice, several interacting factors caused the growth behaviour to become unstable across runs.

Figure 3.14 illustrates the general logic of the fishnet-based region-growing workflow used for EA production.

Figure 3.14. Fishnet-based region growing workflow for EA production.

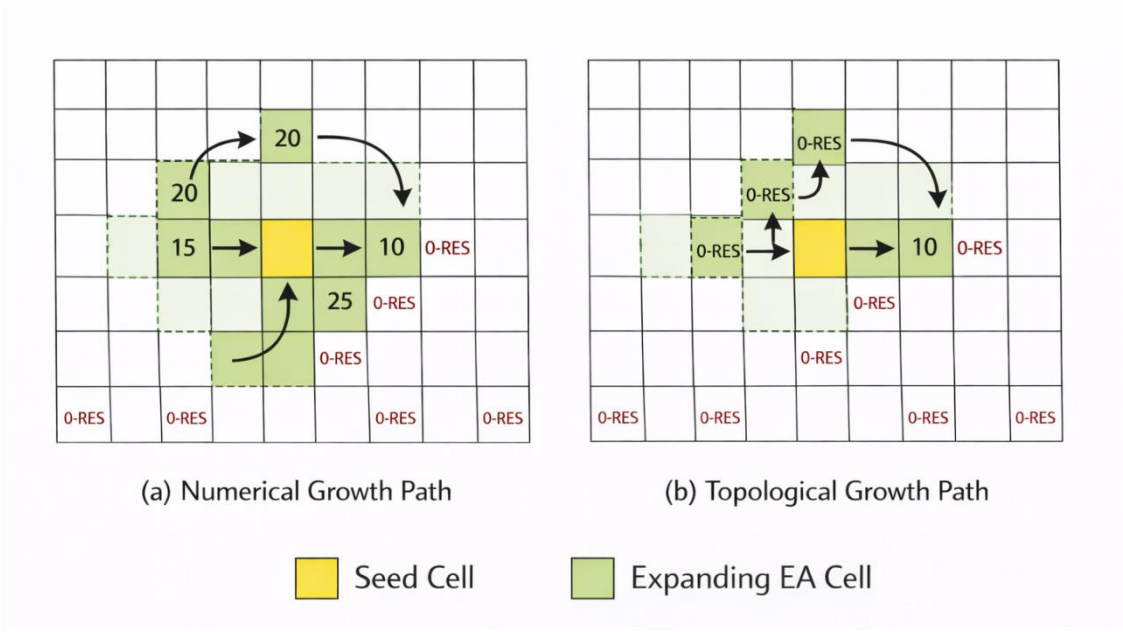


Source: Generated by the author.

The schematic illustrates fishnet cells carrying RES attributes, the selection of a seed cell, and iterative neighbour addition until a target RES band is reached while maintaining contiguity. A core methodological difficulty is that the growth process is highly sensitive to the structure of the adjacency graph and to the distribution of 0-RES and boundary cells. When 0-RES cells are prevalent, the algorithm can expand spatially without improving the RES total, which increases the likelihood of irregular geometries and of failing to reach the target band efficiently.

This behavioural tension is further clarified in Figure 3.15, which contrasts quantitative-driven growth with topology-driven growth.

Figure 3.15. Contrast between quantitative-driven growth and topology-driven growth.



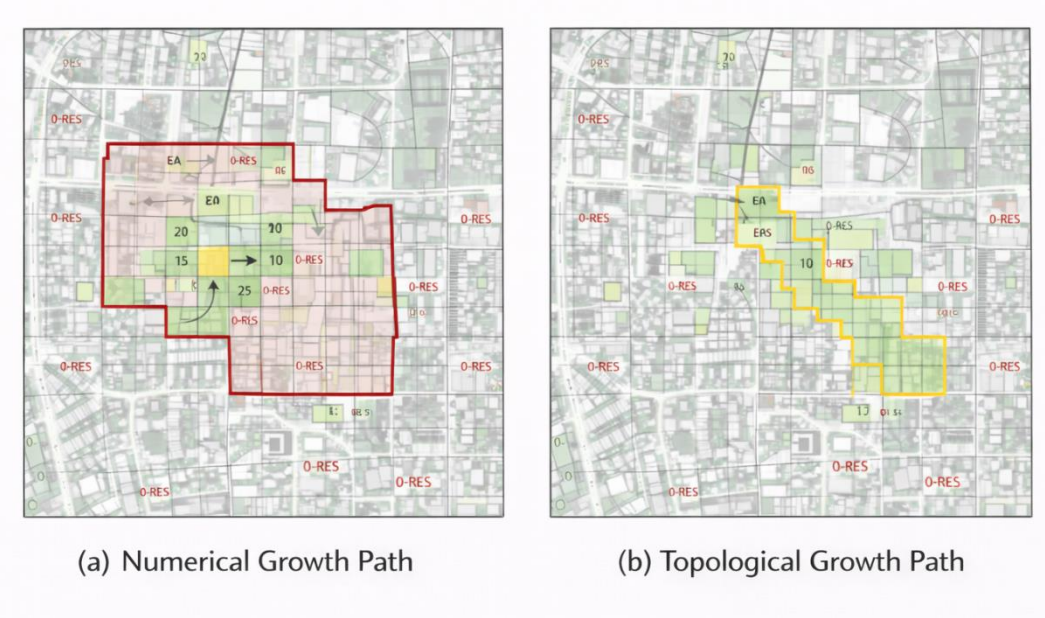
Source: Generated by the author.

The left panel conceptualises a growth path dominated by maximising RES contribution, whereas the right panel shows a case in which 0-RES bridge cells steer growth through connectivity. As shown in Figure 3.15, the diagram highlights the key behavioural tension: even when the algorithm is intended to prioritise quantitative contribution, connectivity constraints can force the inclusion of 0-RES cells, which in turn may redirect the growth path. This effect is amplified when boundary cells are

clipped into small polygons that create narrow corridors or fragmented neighbour relations.

The consequences of this divergence in growth logic are also visible at the level of produced EA geometries. Figure 3.16 presents illustrative EA outputs under different growth-dominance regimes.

Figure 3.16. Illustrative EA outputs under different growth dominance regimes.



Source: Generated by the author.

As illustrated in Figure 3.16, the schematic map comparison shows that EA growth can follow different spatial paths under numerical versus topology-driven logic, even when the final target RES is similar. Quantitative-driven growth tends to produce more compact shapes, while topology-driven growth, often mediated by 0-RES and boundary cells, can yield elongated and operationally weaker geometries. These figures therefore function not merely as visual aids, but as methodological evidence of the structural instability encountered in the fishnet-based trial.

Although the fishnet-based workflow is expressed as iterative growth from a seed cell, the observed behaviour is strongly conditioned by how seeds are placed and in what order they are initialised. In operational regionalisation, seeding affects the growth frontier, exposure to barrier-induced constraints, and the likelihood of consuming low-information areas early in the process (Openshaw & Rao, 1995;

Duque, Ramos, & Suriñach, 2007). For this reason, seed selection is treated as an explicit methodological choice rather than an implementation detail. Seeding is applied to the admissible fishnet cells remaining after barrier filtering and any pre-defined exclusions. To preserve contiguity, seeds are initiated within connected components of the allowed-neighbour graph, and growth is not permitted to jump between components.

Within this framework, four pragmatic heuristics are supported. Random seeding is applied over admissible cells. High-RES seeding prioritises the highest RES_UNITS cells (Assunção, Neves, Câmara, & da Costa Freitas, 2006; Guo, 2008). Centrality-based seeding prioritises structurally interior cells in order to support compactness. Barrier-distance seeding preferentially initiates near barrier edges so as to stress-test the feasibility of compact growth under constrained adjacency. The purpose of using multiple seeding heuristics is not to tune outcomes post hoc, but to reveal whether the growth and stopping rules are robust to plausible initialisations. Centrality-based and barrier-distance strategies in particular approximate two operational extremes: interior starts often promote compact zones, whereas barrier-proximal starts test whether contiguity can be maintained without producing corridor-like geometries when the frontier is compressed by prohibited links. The compactness rationale follows the procedural interpretation of compactness metrics such as the Polsby–Popper family (Polsby & Popper, 1991), while acknowledging that compactness is not the sole operational criterion for EAs.

The behaviour of the growth algorithm is also shaped by how neighbourhood relationships are defined and operationalised. Neighbour relationships between candidate cells are represented through first-order contiguity and are stored explicitly as an adjacency graph used in the growth loop. In ArcGIS Pro, the Polygon Neighbors tool produces a neighbour table under either strict edge contiguity (rook adjacency) or edge-and-vertex contiguity (queen adjacency) (Esri, n.d.). Given the interpretability requirements of EA boundaries, strict edge contiguity is adopted as the default, while queen adjacency is retained for diagnostics and sensitivity checks. The raw neighbour table is transformed into an adjacency list and then into a graph representation, $G = (V, E)$, enabling constant-time neighbourhood queries during iterative growth and

supporting QA diagnostics such as component size, degree distribution, and bridge-like links.

Barrier compliance is enforced before growth by filtering E using the barrier layers, including roads, railways, and water bodies, so that prohibited crossings are removed from the admissible neighbour set. This design makes barrier compliance a hard constraint rather than a soft preference, and aligns the spatial feasibility checks with the same adjacency structure used for scoring. Because first-order contiguity alone can leave small residual pockets near barriers or administrative boundary cuts, secondary graph diagnostics are also computed to support residual handling, for example component membership and distance-to-nearest EA measured in graph steps. These diagnostics are archived together with parameter settings so that any post-allocation steps reported in Chapter 4 can be traced to explicit intermediate graph states.

Within the growth loop itself, candidate selection follows a two-stage decision structure consisting of admissibility screening followed by scoring and tie-breaking. Admissibility is evaluated first in order to exclude candidates that would violate barrier constraints, contiguity, or explicitly prohibited links. Only admissible candidates are scored, ensuring that optimisation is performed within the feasible set and does not rely on ex post repairs to restore feasibility. The primary score operationalises the statistical objective by reducing the distance of the EA's cumulative RES_UNITS to the target band, for example 80–120 units. However, the fishnet setting introduces a characteristic failure mode. When many 0-RES cells are present, adding a 0-RES cell may not change the distance-to-band metric and can therefore appear harmless. If added consecutively, these cells can expand the EA footprint without improving statistical balance and can propagate the instability mechanism described in Section 3.4.6. Accordingly, the scoring logic is complemented by explicit exclusion rules and secondary penalties that discourage repeated 0-RES additions unless such additions are required to preserve contiguity.

Stopping conditions were initially defined in two ways: growth stops when the EA enters the target band, or growth stops when no admissible neighbours remain. In practice, these conditions are insufficient on their own, because maintaining contiguity can require additional cells after the target is reached, and certain seeds can exhaust

their local neighbourhood before reaching the band. The workflow therefore logs the reason for termination, distinguishing between band reached and frontier exhausted, together with the final distance-to-band and compactness or fragmentation diagnostics. This makes it possible to interpret later why a particular EA stopped growing and whether the termination was statistically satisfactory or topologically forced.

Residual cells frequently remain after growth terminates for all seeds. These residuals are typically low-RES or 0-RES units located near barriers, boundary cuts, or disconnected components. Naïvely assigning residuals to the nearest EA tends to deteriorate geometry and can reintroduce instability by forcing narrow corridors or detached appendages. Residual handling is therefore treated as a distinct post-allocation step with its own rules. Residuals are attached only if contiguity and barrier compliance are preserved and if the attachment does not trigger extreme geometric degradation; otherwise, they are flagged for targeted review in the diagnostic outputs reported in Chapter 4.

The combined influence of these design elements is summarised in Table 3.2, which documents the main factors shaping fishnet-based region-growing behaviour.

Table 3.2. Factors influencing fishnet-based region growing behaviour.

Factor	Definition	Effect on the algorithm
Cell size	Scale of starting units	As cell size decreases, adjacency density increases and growth decisions become more complex; computational load typically rises.
Share of 0-RES cells	Proportion of cells with zero residential units	Creates area–RES decoupling: the EA can expand spatially without increasing RES, reducing convergence efficiency and increasing shape irregularity risk.
Boundary cells	Clipped/weak cells along the neighbourhood boundary	Distorts growth direction and neighbour availability; increases fragmentation and irregular geometries due to edge effects.
Seed cell	Initial EA core cell	Strongly influences the final EA geometry and the convergence path; different seeds can lead to materially different outputs under the same rules.

As summarised in Table 3.2, region-growing behaviour is jointly shaped by grid resolution, represented through cell size, the share and spatial configuration of 0-RES cells, weakened boundary cells produced by clipping, and the choice of the initial seed cell. The table is therefore used as a methodological documentation tool rather than as a results statement. Its purpose is to make the main sensitivity drivers explicit and to explain why repeatability becomes difficult under certain spatial conditions.

Overall, the region-growing algorithm did not fail because the data were incorrect, but because the method design places strong and sometimes competing demands on contiguity, scale, boundary handling, and quantitative balancing, especially under high 0-RES presence. This observation provides a clear methodological explanation for why the fishnet-based approach struggled to produce stable EA outputs across repeated runs.

For this reason, diagnostic reporting becomes an essential part of the fishnet-based trial rather than a supplementary documentation step. The reporting protocol is designed to ensure that every methodological trial can be audited and compared on a consistent basis: which parameters were used, what intermediate graph and growth states were produced, what statistical and spatial indicators were computed, and how outputs were archived for later interpretation.

First, a neighbourhood-level RES baseline is recorded before EA delineation. The baseline is computed as the sum of RES_UNITS across all candidate cells after applying the same preprocessing rules used by the growth algorithm, including exclusions, masks, and barrier-informed filtering. Recording the baseline provides a reference point for evaluating whether any losses or duplications are introduced by the workflow and supports comparability across neighbourhood contexts.

Second, EA-level summary indicators are computed for every delineated unit. At minimum, these include total RES_UNITS, number of cells, number of boundary segments, and basic compactness proxies based on area and perimeter derived from the cell union. Additional indicators can be computed where needed in order to compare multiple trials or parameter sets, including within-EA variance, distributional summaries of EA size, and counts of out-of-band EAs. The use of comparable indicators aligns with established practice in zonal system evaluation and regionalisation diagnostics (Openshaw, 1984; Duque, Anselin, & Rey, 2012).

Third, spatial diagnostics are generated to support operational review of the produced EA geometries. These diagnostics focus on compactness and fragmentation, for example the presence of corridors and appendages, adjacency consistency, and the identification of residual pockets that remain unassigned after the growth loop terminates. Compactness is treated as a practical safeguard because excessive perimeter-to-area ratios and corridor structures are closely associated with fieldwork inefficiencies and unstable EA boundaries (Polsby & Popper, 1991; Duque, Ramos, & Suriñach, 2007).

Finally, all outputs are archived together with their parameter settings and diagnostic summaries, enabling reproducibility and structured comparison across trials. This includes saving the final EA polygons, the intermediate growth logs, including seed order and termination reasons, and any residual-allocation actions taken after the main growth phase. The descriptive and comparative statistics produced under this protocol are reported in Chapter 4, Findings, while the present section defines how those outputs are generated and recorded.

3.4.6. Holistic Methodological Evaluation of the Fishnet-Based EA Production Workflow

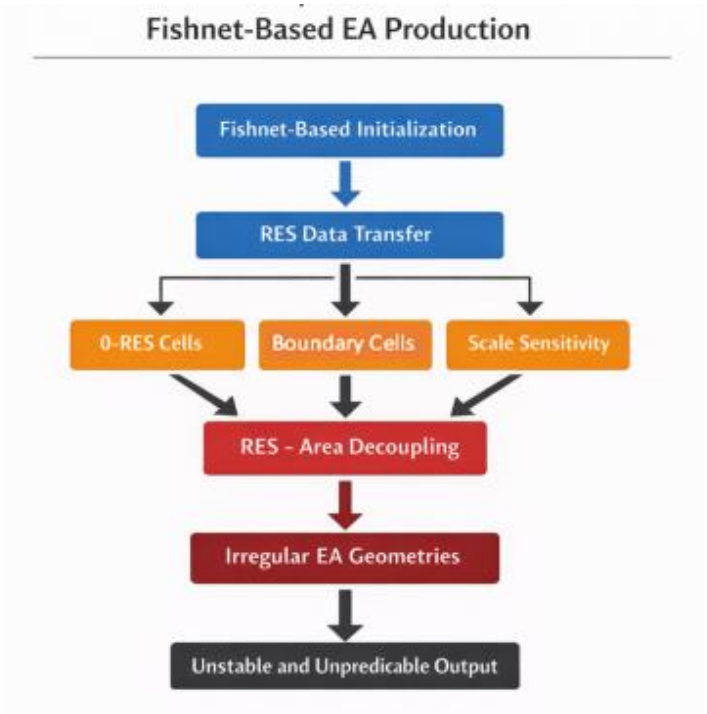
This section provides an integrated methodological evaluation of the fishnet-based EA delineation workflow described in Sections 3.4.1–3.4.5. The aim is not to assess single examples in isolation, but to explain—at the level of process design—why the overall workflow struggled to deliver stable, predictable, and reproducible EA outputs that simultaneously satisfy quantitative balance (RES targets), geometric consistency (compact and coherent boundaries), and operational usability for field enumeration. Accordingly, the evaluation highlights that the dominant drivers of instability are structural (grid resolution, boundary effects, topology constraints, and the behaviour of 0-RES cells) rather than solely data-quality limitations.

Although the fishnet approach offers clear advantages in terms of full spatial coverage and formalised starting conditions, the workflow exhibits an inherent instability mechanism driven by the interaction of scale sensitivity, boundary effects, and 0-RES cell structure. At the design level, cell size determines the resolution of

representation and the complexity of the adjacency graph. At the boundary, clipped cells introduce geometrically weak units that frequently carry low or zero RES values. After RES transfer, the grid becomes statistically heterogeneous, and the initial assumption of equivalent starting units is no longer valid.

As the region growing algorithm operates on this heterogeneous grid, growth decisions become increasingly constrained by topological connectivity rather than by quantitative contribution. This shifts the control of EA formation away from the intended RES target and toward graph-based accessibility. Consequently, EAs may expand substantially in area without proportional RES gains, and shapes can become elongated or irregular, undermining field interpretability.

Figure 3.17. Instability mechanism in fishnet-based EA production.



Source: Generated by the author.

The diagram summarises the cause–effect chain linking cell size sensitivity, boundary clipping, 0-RES prevalence, adjacency-graph dominance, and the resulting loss of stability and repeatability in EA outputs.

Figure 3.17 makes explicit that the core difficulty is structural: the fishnet approach simultaneously enforces gapless coverage and contiguity while seeking quantitative balance under heterogeneous RES distribution. In this context, 0-RES

cells are unavoidable for full representation of non-built space; however, they must be governed by additional rules if the method is expected to converge reliably to a quantitative target band without compromising geometric quality.

Therefore, the key methodological takeaway is that the fishnet-based workflow is well suited for guaranteeing spatial coverage and enabling controlled scenario testing, but it becomes unstable when it is used as the primary mechanism for producing quantitatively balanced EAs in heterogeneous urban fabric. This evaluation provides the methodological justification for moving toward alternative EA production strategies that reduce dependency on grid topology and that align more directly with the underlying residential structure (e.g., building-based or hybrid approaches).

As a transition to the next section, the limitations identified here motivate the methodological shift introduced in Section 3.5, where alternative approaches are presented to improve quantitative stability, geometric coherence, and operational feasibility under the same EA design constraints.

3.5. Final Building-Based EA Delineation Method

Sections 3.4.1–3.4.6 documented the ArcGIS Pro prototyping phase, in which fishnet discretisation and region growing were used to test the feasibility of automated EA production under realistic operational constraints. The principal outcome of that phase was not a stable final EA product, but a clarified methodological specification. EA generation had to be anchored directly to building-level residential counts, had to rely on an explicit barrier-filtered neighbour structure, and had to produce outputs that were both operationally interpretable and reproducible. The final delineation method developed in response to these findings was subsequently implemented in R, and a local Shiny-based interface was developed to support execution, parameter control, and output review. Accordingly, the present section defines the methodological structure of the final approach, while Section 3.6 presents its implementation in R.

The decisive methodological change concerns the analytical unit. Instead of treating fishnet cells as the starting units of aggregation, the final method treats buildings carrying residential workload as the atomic demand units of the system. This reduces dependence on arbitrary discretisation, prevents zero-information cells from

dominating growth behaviour, and aligns the balancing logic more directly with census workload. At the same time, the method preserves the requirement that final outputs must constitute gap-free and topologically valid areal partitions within neighbourhood boundaries.

In conceptual terms, the final approach separates two tasks that were partially entangled during the ArcGIS trial. The first task is workload-sensitive EA assembly, carried out on building-level units under explicit adjacency and barrier rules. The second task is areal completion, in which those assignments are converted into full-coverage EA polygons through residual-area attachment and topology-oriented cleanup. This separation improves interpretability, strengthens diagnostics, and allows statistical balancing decisions to be evaluated independently from coverage repair operations.

3.5.1. Rationale for the Building-Based Formulation

The move to a building-based formulation was motivated by both methodological and operational considerations. In the fishnet-based trial, the regularity of grid cells offered an attractive analytical frame, but the resulting system remained highly sensitive to cell size, boundary clipping, and the behaviour of 0-RES cells. By contrast, buildings correspond much more closely to the actual spatial distribution of census workload. Treating buildings as the atomic demand units therefore improves the substantive meaning of the balancing process and makes the resulting EAs easier to interpret in relation to field listing and enumeration practice.

This choice also reflects the conceptual meaning of constraints in EA design. Enumeration Areas are not defined by geometry alone, but by workload, accessibility, and administrative feasibility. International census practice emphasises approximate equality of workload, respect for obvious physical boundaries, and the maintenance of clear operational units rather than geometrically regular partitions imposed for their own sake (United Nations, 2017; Eurostat, 2017). In the present study, these principles are encoded directly in the building-based assembly logic instead of being approximated through an antecedent grid.

A further advantage of the building-based formulation is that it makes the smallest indivisible operational unit explicit. Each building, identified through a persistent building identifier, is treated as a non-splittable unit. This avoids the artificial division of a single physical structure across multiple EAs, which would complicate enumerator assignment, increase the risk of omission or duplication, and weaken the intuitive spatial logic of the output. The method therefore accepts that exact numerical balance is sometimes impossible and treats such cases as structural consequences of real urban form rather than as algorithmic defects.

3.5.2. Inputs, Pre-processing, and Barrier-Constrained Neighbour Structure

The final method uses four core input classes: neighbourhood administrative boundaries defining the scope of delineation, building footprints with residential-unit attributes, barrier layers representing operational separators such as major roads, railways, and waterways, and optional auxiliary layers used for control or interpretation. Before delineation begins, all layers are harmonised to a common projected CRS and checked for geometry validity and key-attribute completeness. This ensures that metric operations, topology checks, and neighbourhood relations remain internally consistent throughout the workflow.

Pre-processing also ensures that the building layer represents physical structures rather than arbitrary geometric fragments. Where the same building is represented by multiple polygons, attribute-based consolidation is applied so that each physical building is carried forward with a single geometry and a single residential-unit value. Residential and non-residential structures are distinguished at this stage. Non-residential buildings are retained geometrically in order to preserve spatial continuity, but they contribute zero to the household-equivalent target count. This maintains full spatial representation without inflating statistical workload.

Neighbour relations are then constructed at the building level using contiguity and, where necessary, controlled near-touch proximity. These candidate relations are filtered through the barrier model so that the final allowed-neighbour graph contains only those connections that remain operationally admissible. In this way, barrier

compliance is treated as a hard constraint rather than a soft preference. Graph diagnostics such as connected components, node degrees, and edge density are used to identify over-fragmentation or under-constrained connectivity before aggregation begins.

3.5.3. Rule-Driven EA Assembly, Target Ranges, and Exception Logic

EA assembly is implemented as deterministic region growing on the barrier-filtered building graph. Each EA begins from a seed building or seed cluster and expands by iteratively attaching admissible neighbouring buildings until the cumulative residential-unit total reaches, or approaches, the target band. Candidate additions are evaluated hierarchically. Hard constraints—administrative boundary adherence, barrier integrity, and contiguity—are enforced before balance objectives are optimised. Within the feasible set, candidates are ranked according to their contribution to the target range and, secondarily, according to geometric preferences that discourage unnecessarily elongated or corridor-like growth.

The primary workload constraint is the number of residential units (RES), which functions as the most direct proxy for enumeration effort. Alternative secondary constraints, particularly a broader total-unit variable that included non-residential uses, were evaluated but not retained in the final method because they complicated the optimisation problem without improving field relevance. Within the final framework, a target around 100 household-equivalent units is treated as a planning reference rather than as a universal hard quota, and neighbourhood-level target ranges are differentiated through the DEGURBA classification. Dense urban, intermediate, and rural contexts therefore operate under related but non-identical admissible bands.

This target structure is deliberately implemented as soft balancing rather than strict equality. The empirical structure of the data makes exact compliance neither realistic nor desirable in all cases. Buildings are indivisible, barrier-defined neighbourhood fragments can be spatially isolated, and compactness cannot always be preserved if the algorithm is forced to chase a numerical threshold at any cost. Consequently, the method minimises deviation from the target band while accepting controlled departures when they arise from higher-priority operational constraints.

Two exception mechanisms are particularly important. First, buildings with RES_UNITS = 0 are treated as neutral elements: they are not preferred for quantitative progress, but they may be used to preserve contiguity or to support complete areal coverage. Second, exceptionally large buildings are handled as explicit structural cases. When a single building contains a residential-unit count that exceeds the upper EA target, it is designated as a standalone EA rather than being forcibly split or combined in a way that would violate building integrity. Oversized EAs caused by barrier sparsity or locked large buildings are therefore retained and flagged transparently, not hidden through forced rebalancing.

3.5.4. Coverage Completion, Residual Handling, and Topological Validation

After building assignments have been finalised, the method converts these assignments into full EA polygons. This stage is kept analytically separate from balancing. Buildings are first dissolved by EA identifier to create the primary EA geometries, and only then are residual non-building areas inside the neighbourhood boundary attached through explicit coverage-completion rules. This sequencing is methodologically important because it prevents topological repair operations from silently altering the building-based workload logic that governed the primary delineation stage.

Residual areas may arise from barrier segmentation, boundary-adjacent slivers, non-residential spaces, or local geometric effects introduced during polygon construction. These areas are attached only when contiguity and barrier compliance are preserved and when the attachment does not cause disproportionate geometric degradation. In this way, the method pursues complete areal coverage while preserving the priority structure established during EA assembly.

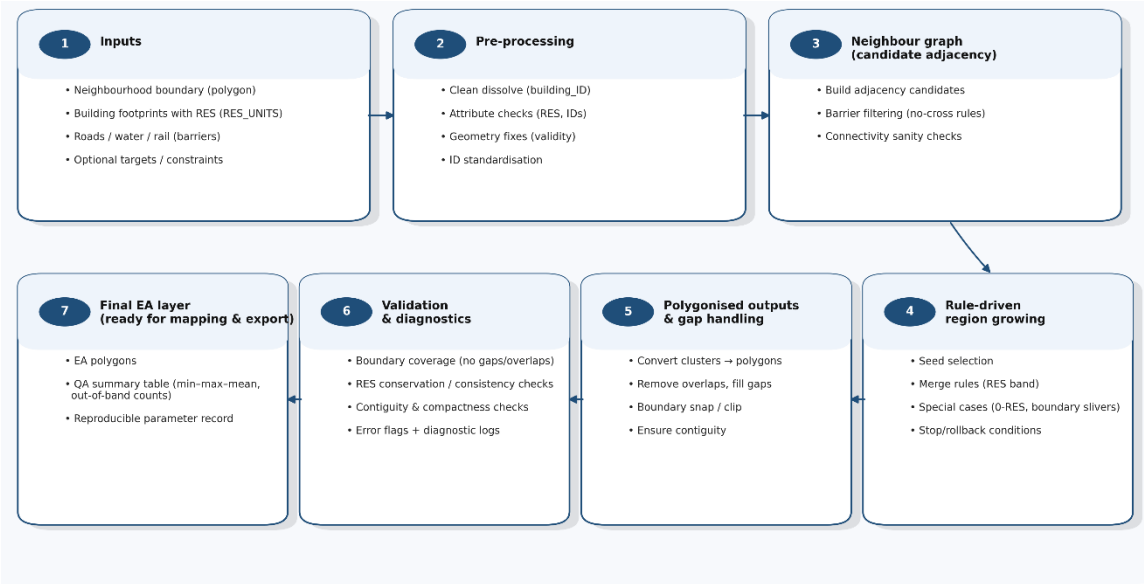
The resulting EA polygons are then subjected to topology-oriented validation. Checks address invalid geometries, gaps, overlaps, multipart artefacts, isolated fragments, and unexplained changes in total RES across stages. Where issues are detected, repair operations are applied as deterministic rule-governed interventions rather than as manual edits. This preserves reproducibility, makes corrective actions

auditable, and ensures that the final EA layer remains interpretable both statistically and operationally.

3.5.5. Documentation Outputs and Pilot-to-Scale Rationale

To support thesis-level transparency, the final method produces both substantive outputs and methodological records. Final EA polygons and their attribute tables are complemented by intermediate layers, graph diagnostics, merge records, and structured run logs. Together, these outputs make it possible to reconstruct how a given delineation was produced, which rules were activated, where exceptions occurred, and how parameter settings influenced the result. Figure 3.5.1 and Table 3.3 summarise the workflow and its documentation structure in compact form.

Figure 3.18 Schematic of the building-based, rule-driven EA delineation workflow implemented after the fishnet trials.



Source: Generated by the author.

Table 3.3. Documentation structure for the building-based workflow (stages, decisions, intended effects, risks, and diagnostics)

Stage	Main decision / parameter	Brief definition	Intended effect	Potential methodological risk	Logged diagnostics (examples)
Input assembly	Input layers & IDs	Collect boundary, buildings (RES), barriers, optional targets; enforce unique IDs.	Ensure consistent spatial extent and traceable units.	Mismatched CRS/IDs; missing fields leading to silent drop-outs.	CRS check; feature counts; missing-field report.
Geometry normalisation	Repair & dissolve rules	Repair invalid geometries; dissolve building fragments by building ID.	Prevent artificial fragmentation of RES and adjacency.	Over-dissolve merges distinct buildings; geometry repair shifts edges.	Pre/post counts; RES totals by stage; invalid-geometry log.
RES definition	Residential vs non-residential	Assign RES to residential buildings; keep non-residential as RES=0 while retaining geometry.	Maintain coverage while keeping the RES signal meaningful.	RES=0 areas dominate topology; misclassification biases RES.	RES summary by type; share of RES=0 features.
Barrier model	Barrier inclusion rules	Create a barrier mask from roads/rail/water; define what blocks adjacency.	Reduce unrealistic crossings; improve operational plausibility.	Over-blocking fragments the area; under-blocking enables crossings.	Blocked-edge counts; barrier intersections reported.
Neighbour candidates	Adjacency definition	Generate candidate neighbours (touching/within tolerance) for units used in growth.	Build a reproducible graph for subsequent growth.	Sensitivity to tolerance creates unstable graphs.	Degree distribution; tolerance notes.
Filtered graph	Barrier filtering	Remove candidate edges that cross a barrier; keep only allowed connections.	Align graph connectivity with operational constraints.	Disconnected components lead to underfilled EAs.	Component count; isolated nodes list.
Seeding	Seed selection strategy	Select initial cores (e.g., high-RES nodes, spatial spread, deterministic order).	Stabilise growth path and improve reproducibility.	Seed choice drives final geometry; may introduce bias.	Seed list; random seed if applicable.
Growth rule	Target band (e.g., 80–120)	Iteratively add neighbours to reach the target RES band while maintaining contiguity.	Support workload balance and comparability.	Greedy traps; instability near thresholds.	Step-wise RES trace; stop reasons.

Table 3.3. Documentation structure for the building-based workflow (stages, decisions, intended effects, risks, and diagnostics) (continued)

Stage	Main decision / parameter	Brief definition	Intended effect	Potential methodological risk	Logged diagnostics (examples)
RES=0 handling	Inclusion logic	Allow RES=0 units only when required for connectivity or boundary coverage.	Avoid area–RES decoupling while preserving continuity.	Area inflates without RES gain; 'bridge' effects steer growth.	RES=0 additions flagged; bridge cases counted.
Underfilled reconciliation	Merge / attach rules	Resolve zones below minimum RES by merging/attaching to the best neighbour under rules.	Avoid many small, non-operational EAs.	Merges violate barriers or compactness; cascade effects.	Merge decisions logged; before/after RES.
Polygonisation & gaps	Coverage enforcement	Convert assignments to EA polygons; fill gaps within boundary; remove overlaps.	Produce a valid, gapless EA layer for field use.	Slivers/disjoints; boundary artefacts.	Gap stats; disjoint/overlap report.
Validation & export	QA thresholds	Run topology checks; export EA and support layers; write diagnostics and metadata.	Deliver repeatable outputs and auditability.	Passing QA but poor usability; missing metadata.	QA summary; min/mean/max RES; export manifest.

Source: Generated by the author.

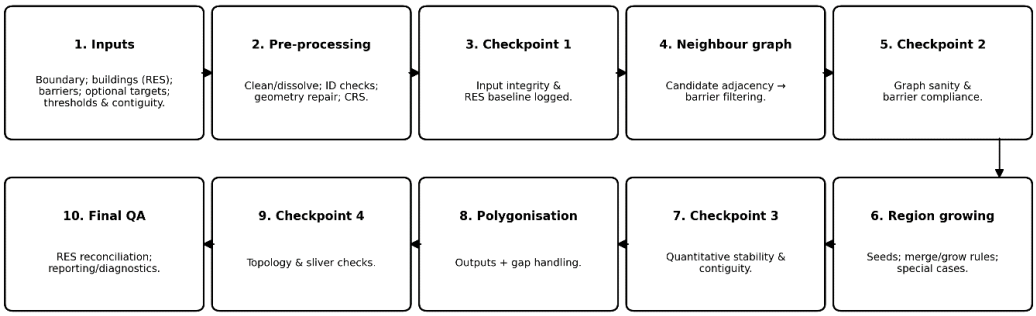
The Bağlıca pilot played a critical role in demonstrating how this documentation structure supports scaling. The pilot did not function as a one-off showcase, but as the stage in which data heterogeneity, barrier-neighbour conflicts, large-building exceptions, and topology-driven repair needs became visible in a controlled setting. The methodological value of the pilot lies in showing that district-scale execution does not require identical datasets everywhere, but does require stable constraint classes, explicit diagnostics, and a workflow whose decisions remain interpretable across neighbourhoods.

3.6. R Implementation, GUI, and Quality Assurance of the Final Method

Whereas Section 3.5 defined the final EA method conceptually, this section documents how that same method was operationalised in the open-source R environment, deployed through a local Shiny interface, and controlled through

structured validation and diagnostic logging. The implementation should therefore be read as the execution environment of one final method rather than as a separate methodological alternative.

Figure 3.19. QA and diagnostic checkpoints integrated into the EA delineation workflow.



Source: Generated by the author.

3.6.1. Open-Source Implementation Rationale and Software Environment

R was selected as the primary production environment because it supports script-based reproducibility, explicit parameter control, and a mature ecosystem for spatial data processing and graph-based operations. Packages aligned with the Simple Features standard provide consistent vector geometry handling, while graph libraries support deterministic neighbour-based growth and diagnostic analysis. Compared with GUI-driven execution chains, this environment makes rule sets, parameter values, and intermediate outputs explicit and therefore more suitable for thesis-level auditability and future replication.

Long-term portability was also an important consideration. A scripted workflow can be archived together with the thesis artefacts and rerun on different machines under the same parameter settings, whereas point-and-click procedures are more difficult to reconstruct precisely. For this reason, the R implementation was designed not as a visualisation layer, but as the operational core in which delineation, logging, and topology-oriented post-processing are executed.

3.6.2. Workflow Architecture, Minimal Working Dataset, and Execution Logic

To stabilise execution across neighbourhoods, the implementation uses a minimal working dataset that standardises the essential inputs required by the delineation routine. Buildings carrying RES attributes provide the demand signal, neighbourhood boundaries define the spatial scope, and barrier layers constrain candidate adjacency. Optional exclusion or control layers can be added where operationally necessary, but the main algorithm is insulated from source-schema idiosyncrasies through this minimal and consistent data model.

The workflow itself is organised as a modular pipeline. Input assembly and validation, graph construction, rule-driven EA growth, coverage completion, and export are executed as separable stages, each of which produces intermediate artefacts that can be checked independently. Deterministic ordering rules are imposed wherever candidate sets might otherwise be ambiguous, so that identical inputs and parameters yield identical outputs. In addition to the main delineation sequence, a controlled post-processing stage is invoked when diagnostics reveal contact errors, residual boundary slivers, or disjoint assignments that cannot be resolved by parameter tuning alone.

This architecture also clarifies the relation between method and implementation. The conceptual rules defined in Section 3.5 remain unchanged; what the R workflow provides is a stable execution framework in which those rules can be run repeatedly, logged systematically, and extended to district-scale production without dependence on proprietary geoprocessing chains.

3.6.3. Shiny-Based Operational Deployment and Parameter Structure

In addition to direct script execution, the final workflow is exposed through a local Shiny interface. The interface does not introduce a distinct delineation logic. Instead, it collects runtime parameters, builds a run-specific configuration, validates critical inputs, and executes the same core R workflow used in batch mode. In this sense, the Shiny layer should be interpreted as an operational deployment interface

that improves usability, transparency, and controlled experimentation rather than as a separate method.

The interface groups parameters in a way that is methodologically useful. A first group contains scenario-defining controls that materially affect delineation behaviour, such as threshold bands, split and merge settings, and the activation of optional hierarchy-construction modules. A second group contains guardrail controls that protect the workflow against specific geometric artefacts, such as boundary-adjacent slivers, multipolygon formation, or problematic zero-target residuals. This distinction is important for interpretation because scenario-defining controls are suitable for comparative analysis in the Findings chapter, whereas guardrail options mainly improve robustness and revision traceability.

Because the underlying script remains the same in both GUI-based and script-only execution, the interface contributes operational convenience without blurring the methodological distinction between the final delineation logic and its implementation environment. Detailed parameter groups and interface organisation are therefore relegated to Appendix A rather than treated as independent conceptual sections in the main methodological narrative.

3.6.4. Validation Framework, Diagnostic Logging, and Controlled Post-Processing

Quality assurance is embedded directly into the workflow rather than applied only after outputs have been produced. The validation framework is organised around three complementary dimensions: attribute integrity, spatial integrity, and operational suitability. Attribute integrity concerns the preservation and correct aggregation of residential-unit totals across stages. Spatial integrity concerns the production of a gap-free, non-overlapping, and contiguous areal partition within each neighbourhood boundary. Operational suitability concerns whether the resulting EA geometries remain interpretable and usable in field conditions, even when they satisfy the first two dimensions.

These dimensions are implemented through repeatable checks at defined control points. Total RES is recalculated after major processing stages to detect loss,

duplication, or distortions introduced by joins, dissolves, or geometry failures. Topology checks verify coverage, overlap, contiguity, and multipart structure. Additional geometric warnings identify narrow corridors, slivers, and extreme elongation that may not invalidate the output formally but can still indicate weak operational quality. Table 3.4 summarises the minimum validation criteria used in this thesis.

Table 3.4. Validation dimensions and minimum acceptance criteria used in this thesis

Stage	Checkpoint objective	Indicator / metric	Pass–fail rule (example)	Action if flagged
Pre-processing	Input data integrity	Geometry validity; CRS match; missing/duplicate IDs	Any invalid geometry or CRS mismatch = fail	Repair geometries; reproject; enforce unique IDs
Pre-processing	RES baseline control	Total RES (global sum) recorded	Baseline must be logged before delineation	Log baseline; stop run if baseline missing
Neighbour graph	Barrier compliance	Count of links crossing barriers	Any barrier-crossing adjacency = fail	Fix barrier layer; re-run barrier filtering
Neighbour graph	Graph sanity	# connected components; degree extremes	Excess components or abnormal degrees = warning	Inspect disconnected zones; adjust adjacency rules
Region growing	Quantitative stability	Running RES per EA vs target band	Persistent drift / repeated overshoot = warning	Adjust seed rules; merge rules; 0-RES handling strategy
Region growing	Contiguity enforcement	Disconnected EA parts during growth	Any non-contiguous EA = fail	Force contiguity constraint; rollback step
Polygonisation	Topology quality	Gaps/overlaps; slivers	Any overlap = fail; gaps/slivers above tolerance = warning	Gap handling; dissolve/clean; reassign residuals
Final QA	Reconciliation & reporting	Total RES after output; min–max–mean; disjoint polygons	Total RES must match baseline within tolerance; no disjoint polygons	Investigate loss/gain sources; fix assignment; regenerate outputs

Source: Generated by the author.

Diagnostic logging complements these checks by recording the conditions under which each run was produced. At minimum, the logs retain the parameter set, EA counts, RES distributions, topology flags, and exception events. This makes revision traceability possible: when a rule is changed or a guardrail is introduced, its effect can be evaluated against concrete and comparable diagnostics. The controlled

post-processing routine used for contact errors and residual assignments is integrated into this same QA logic. It is not treated as a second delineation method, but as a deterministic repair layer invoked only when diagnostics justify it.

3.6.5. Metropolitan Deployment, Checkpointing, and Computational Controls

The final workflow is designed to operate at metropolitan scale through neighbourhood-level isolation, incremental execution, and structured checkpointing. Each neighbourhood is processed as an independent unit, and outputs are written to disk immediately upon completion. This design reduces redundant computation, limits error propagation across administrative boundaries, and allows interrupted runs to resume without repeating completed work. It also aligns with the operational logic of census geography production, in which neighbourhoods function as administratively meaningful containers for EA generation.

District-scale deployment reuses the preprocessed neighbourhood datasets generated in earlier stages, thereby avoiding repeated clipping, validation, and barrier preparation for every run. Progress logs record identifiers, runtime, DEGURBA class, and status flags, while failed neighbourhoods are isolated and reported without halting the full processing chain. This safe-failure discipline preserves the integrity of long-running executions and makes exceptional cases available for targeted review rather than burying them inside a monolithic run.

At scale, quality control also requires summary classifications. EAs are therefore labelled according to their relation to the target band, for example as within-range, under-sized, over-sized, or single-large-building cases. These labels do not function as automatic correction triggers, but as diagnostic summaries that support interpretation and prioritised review. From a computational perspective, the method prioritises predictability over raw speed. Runtime may remain substantial for dense metropolitan datasets, but in the context of census preparation such cost is acceptable because the gains in reproducibility, transparency, and reduced manual editing outweigh the expense of long but controlled batch execution.

3.7. Formal Properties, Methodological Scope, and Limitations of the Final Workflow

This section does not introduce a separate delineation approach. Instead, it consolidates the formal properties of the final workflow already established in Sections 3.5 and 3.6. The aim is to restate the governing principles, clarify the scope within which the method should be interpreted, and document the main trade-offs that remain even after the transition from the ArcGIS fishnet prototype to the final building-based implementation.

3.7.1. Governing Principles and Priority Structure

The final workflow is governed by a clear priority structure. First, EAs must remain mutually exclusive and spatially contiguous. Second, they must respect neighbourhood boundaries and the barrier model used to encode meaningful movement constraints. Third, they should approximate the relevant DEGURBA-specific workload bands as closely as possible. Compactness is treated as a secondary preference rather than as a dominant objective; compact shapes are preferred when they emerge under the higher-priority constraints, but compactness is never pursued by crossing a major barrier, fragmenting a building, or forcing an otherwise incoherent merge.

This priority structure is methodologically important because it makes trade-offs explicit. The workflow does not pretend that all desirable EA properties can be satisfied simultaneously in every urban context. Instead, it defines which rules are non-negotiable and which objectives may be relaxed when conflicts arise. In this respect, the method is better understood as a constrained decision framework than as a single mathematical optimisation problem with one globally optimal solution.

3.7.2. Target Ranges, DEGURBA Differentiation, and Exceptional Large-Building Cases

A central formal property of the method is the differentiation of target size ranges according to DEGURBA. This reflects the empirical observation that dense urban, intermediate, and rural settlement structures cannot be managed through a single uniform workload threshold without creating either overloaded urban EAs or excessively large and inefficient rural ones. DEGURBA therefore functions as a categorical control variable that translates settlement context into target-setting logic.

Within this framework, target size is interpreted as an admissible band rather than as an exact quota. The method seeks to minimise deviation from that band while preserving building integrity and barrier compliance. This is why exceptionally large buildings are formalised as explicit structural cases rather than treated as failures. When a single building exceeds the normal upper band, it is retained as a standalone EA. Likewise, EAs that remain oversized because a barrier configuration prevents any operationally reasonable split are kept and flagged transparently. These cases are methodologically consistent with the logic of the workflow because they reflect real spatial constraints rather than arbitrary exceptions invented after the fact.

3.7.3. Failure Modes, Trade-offs, and Limits of Greedy Aggregation

Despite the structured and rule-explicit nature of the final workflow, certain failure modes remain inherent to the problem. The most common is the barrier-target conflict: in some neighbourhoods, barrier-defined local units or large buildings make it impossible to reach the target band without violating a higher-priority rule. Under-sized EAs can also remain in highly fragmented settings where no admissible merge preserves contiguity and barrier integrity. In both cases, the method chooses transparent retention and flagging over concealed forced correction.

A second limitation concerns sensitivity to the specification of barrier classes. Including additional minor linear features may over-fragment the graph and inflate the number of under-sized EAs, whereas excluding meaningful barriers may create geometrically convenient but operationally implausible outputs. Barrier selection is

therefore context-sensitive and must be interpreted as a design choice with substantive consequences rather than as a neutral preprocessing step.

A third limitation is tied to the local, greedy character of the aggregation logic. The workflow makes deterministic and interpretable local decisions, but it does not search for a globally optimal arrangement across the entire neighbourhood. This simplicity is deliberate: more aggressive optimisation techniques could in principle improve numerical balance in some cases, but would also introduce greater opacity, heavier parameter dependence, and higher computational cost. The final method therefore accepts that certain extreme cases—such as very large EAs arising in barrier-poor environments—must be documented as structural limits of the chosen design philosophy rather than eliminated at any cost.

3.7.4. Reproducibility, Scalability, and Transferability

The final workflow is designed to be reproducible and auditable. Because each stage is governed by explicit rules, recorded parameters, and deterministic ordering where ambiguity might arise, the same inputs can be rerun under the same conditions to produce the same outputs. This is a central requirement in official statistical contexts, where methodological transparency is not only a scientific preference but also an institutional quality-assurance expectation.

The workflow is also scalable. Neighbourhood-level isolation, incremental output writing, progress logging, and safe-failure handling allow the same method to be applied across large metropolitan datasets without requiring uninterrupted city-wide execution. This makes the workflow suitable for district-scale and metropolitan-scale production, where repeatability and controlled diagnostics matter more than single-run elegance.

Finally, the framework is transferable. Certain operational parameters—particularly target ranges and barrier classifications—are necessarily context-specific, but the overall architecture remains adaptable to other settings. By substituting locally appropriate settlement typologies, workload targets, and barrier definitions, the same two-stage logic of constraint-aware assembly followed by controlled areal completion can be applied beyond Ankara. In this sense, the proposed method should be

understood as a generalisable workflow template whose concrete calibration remains open to national and institutional adaptation.

Taken together, Sections 3.5–3.7 establish one final EA delineation workflow composed of a building-based methodological core, a reproducible R implementation environment, and a clearly stated set of formal properties and limits. This organisation resolves the ambiguity between method and implementation that emerged during earlier drafting and provides a more coherent basis for the findings presented in the following chapter.

CHAPTER 4. FINDINGS

This chapter reports the empirical outputs obtained by applying the proposed EA delineation approach under two complementary implementation tracks: (i) ArcGIS Pro–based trials using existing zoning tools and grid/Thiessen pre-structures, and (ii) an R-based rule-driven pipeline that operationalises barrier-aware adjacency, iterative growth, and district-scale batch production. The purpose of the chapter is to present the produced EA geometries, diagnostic artefacts, and summary statistics in a coherent analytical flow, and to document practical implications observed during the applications.

4.1. Findings of ArcGIS Pro Applications

This section presents the outputs obtained from the ArcGIS Pro applications introduced in the Methodology chapter. All ArcGIS Pro-based trials reported here were conducted at the single-neighbourhood scale in Baglica. The findings are organised around two main implementation paths: BBZ-based applications and ArcPy-based applications without BBZ. This structure helps clarify the practical differences between the built-in zoning tools of ArcGIS Pro and the custom delineation workflows developed during the study.

4.1.1. Results of BBZ-Based Applications

The BBZ trials served as an exploratory stage for assessing how alternative input geometries affected the form and interpretability of the resulting zones. In particular, the applications examined whether fishnet- and Thiessen-based representations could yield outputs with a clearer areal character than the original point- and building-based inputs.

4.1.1.1. Results of BBZ Trials with Original and Fishnet-Based Inputs

The first BBZ trials were carried out using the original Bağlıca data and its fishnet-based representation. Figure 4.1 presents the population-point view of the dataset, whereas Figure 4.2 shows the fishnet representation derived from the same area.

Figure 4.1. View of the population points of the Bağlıca dataset

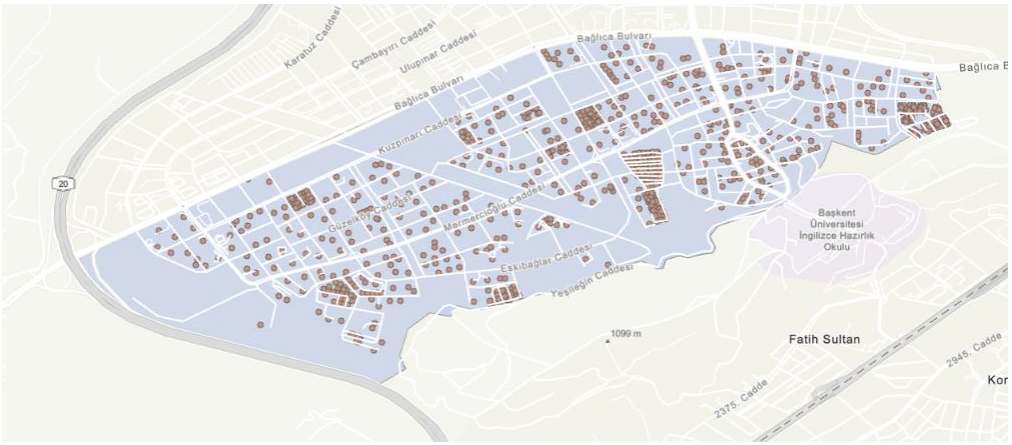
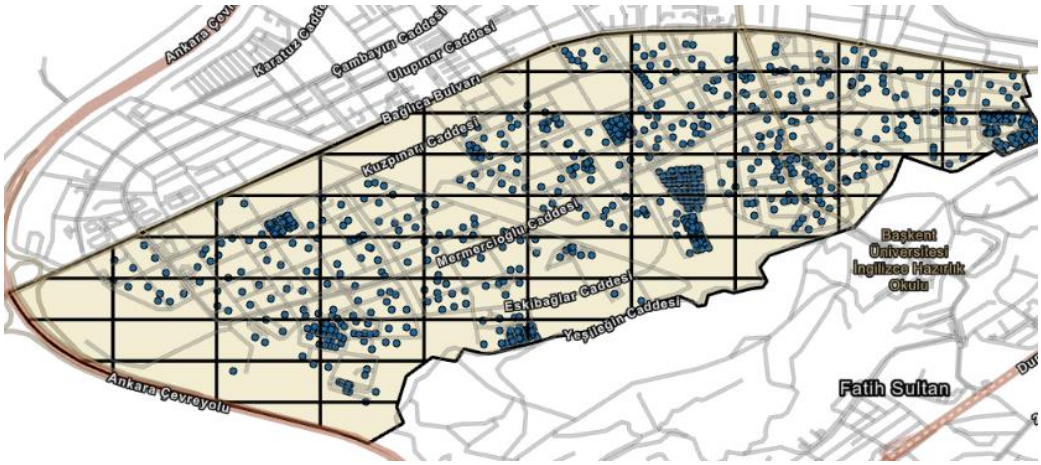


Figure 4.2. Fishnet representation of the Bağlıca dataset



These views are important because the BBZ tool largely preserves the geometry of the data supplied to it. When the input is point-based, the balanced outputs remain tied to point geometry; when polygon-based inputs are used, the outputs are likewise generated in polygon form. This pattern became clearer when the updated

Bağlıca data, which include building footprints as polygons, were used in the BBZ trial.

Figure 4.3. Updated Bağlıca dataset used in the BBZ trial

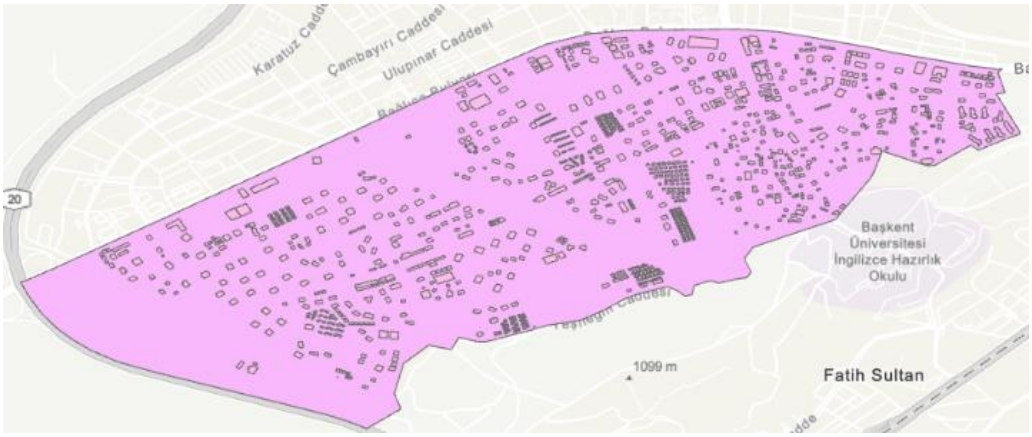
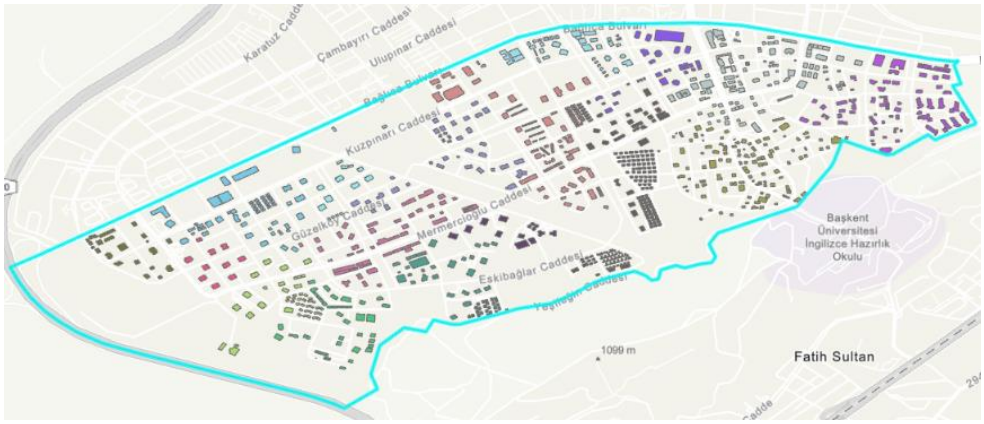


Figure 4.4. BBZ output for the updated Bağlıca dataset



Note: Polygon colours are cartographic-only; they do not encode magnitude, ranking, or category.

Figure 4.3 shows the updated Bağlıca dataset used in the BBZ application, in which building footprints were represented as polygons rather than points. When this polygon-based input was processed through BBZ, the resulting zones were likewise generated in polygon form, as shown in Figure 4.4. This comparison confirms that the geometry of the BBZ output is strongly shaped by the geometry of the input layer. Although the polygon-based result appears more areal than the point-based representation, it still reflects the structure of the source data rather than an independently constructed EA surface.

The statistical summaries generated after the BBZ procedure were also useful for evaluating the composition of the resulting zones. Figure 4.5 presents the sum of hh_nufus by zone, whereas Figure 4.6 shows the number of buildings associated with each ZONE_ID. These outputs indicate that BBZ can produce balanced summaries in a statistical sense; however, they do not by themselves demonstrate that the resulting units satisfy the operational requirements of EA delineation.

Figure 4.5. Sum of hh_nufus by zone for the BBZ trial

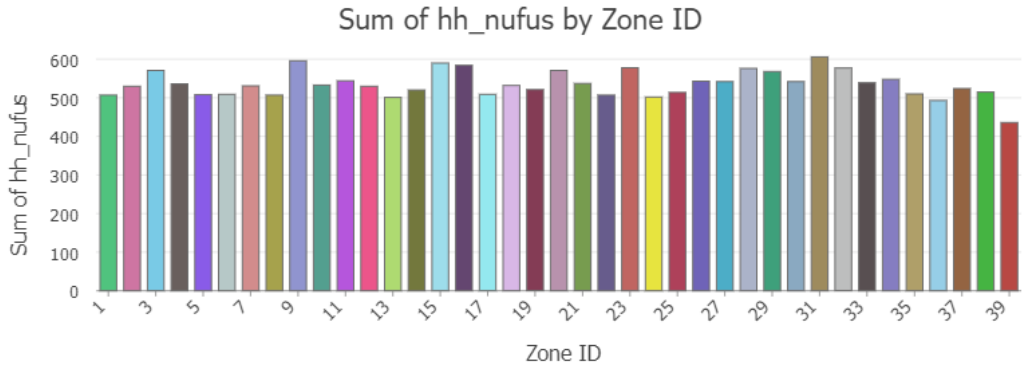
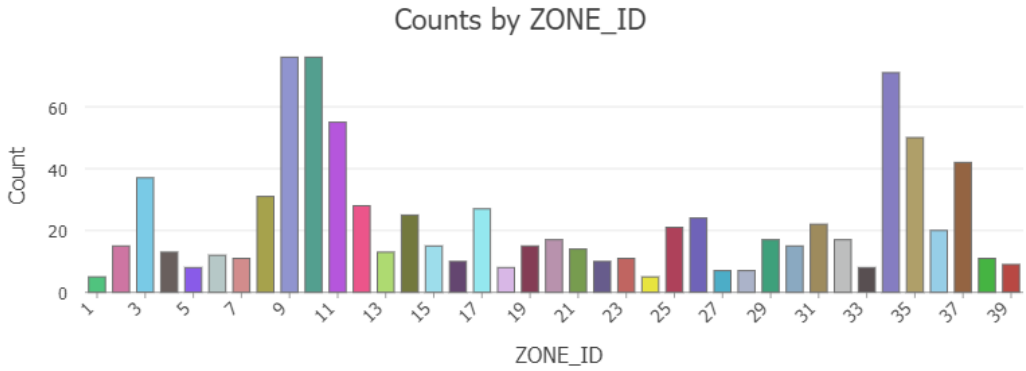


Figure 4.6. Building counts by ZONE_ID for the BBZ trial



4.1.1.2. Results of BBZ Trials with Thiessen Polygon Inputs

A separate BBZ trial was conducted using Thiessen polygons in order to obtain a more continuous areal representation prior to zoning. Figure 4.7 shows the initial Thiessen-polygon representation of the Bağlıca dataset. After preprocessing, the Thiessen layer was colourised and clipped to the neighbourhood boundary, as illustrated in Figures 4.8 and 4.9. These steps provided a clearer areal representation

of the point-based input within the limits of the study area. Figure 4.10 shows the same Thiessen surface with the population points overlaid, making it possible to assess the spatial correspondence between the derived polygons and the original population-point distribution.

Figure 4.7. Thiessen-polygon representation of the Bağlıca dataset

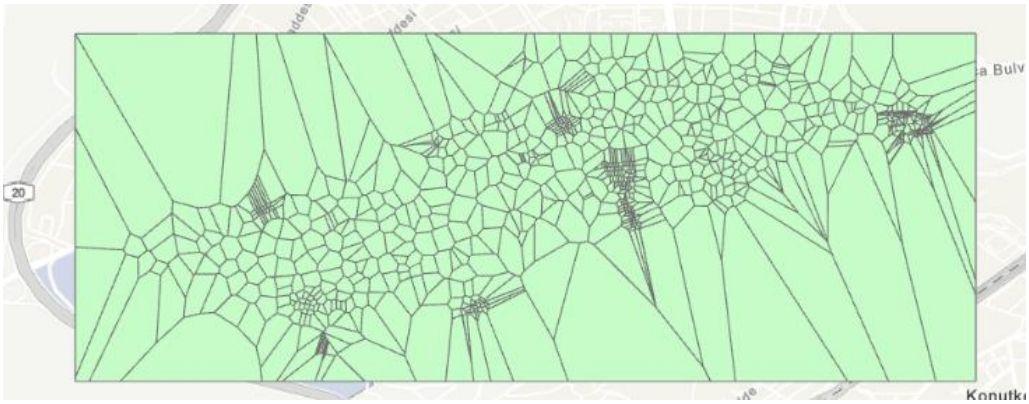
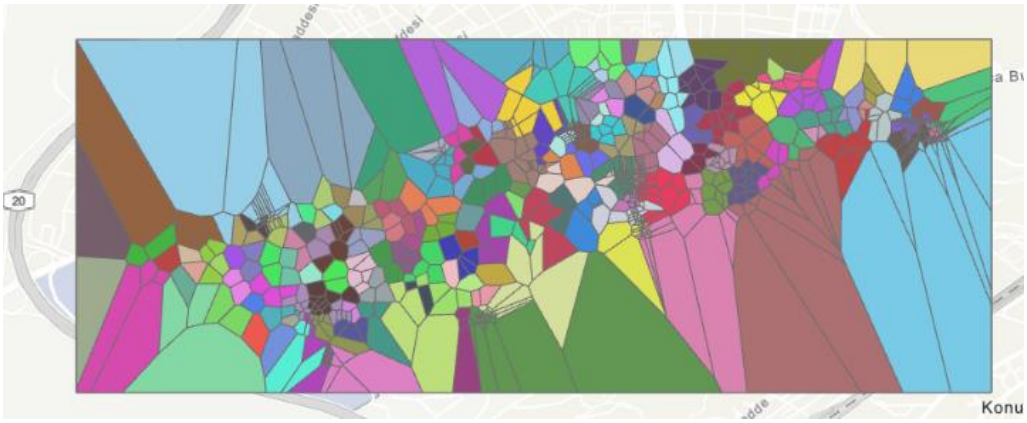
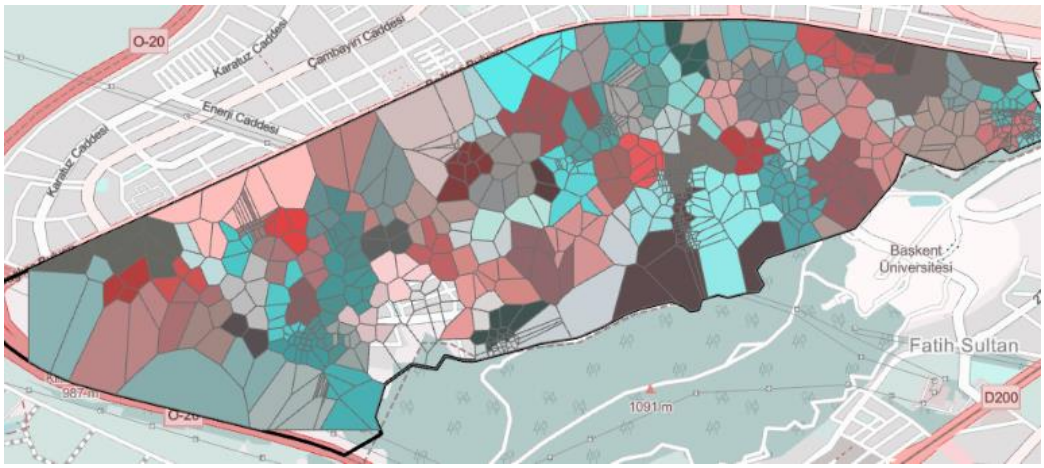


Figure 4.8. Colourised Thiessen polygons derived from the Bağlıca dataset



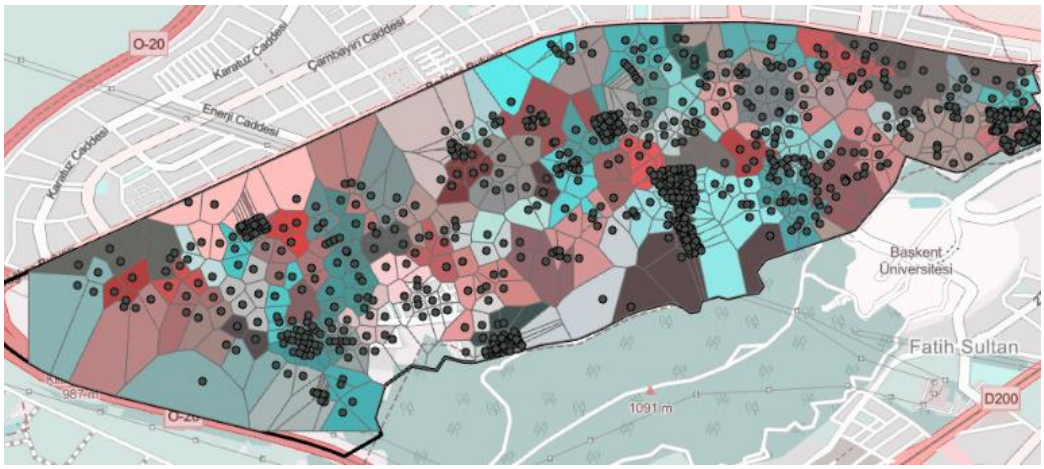
Note: Polygon colours are cartographic-only; they do not encode magnitude, ranking, or category.

Figure 4.9. Thiessen Polygons Image of Bağlıca data colored and clipped to the neighborhood border



Note: Polygon colours are cartographic-only; they do not encode magnitude, ranking, or category.

Figure 4.10. Thiessen polygons clipped to the neighbourhood boundary with population points added



Compared with the original point-based view, the Thiessen approach produced a more continuous and visually interpretable surface. However, this remained a representational transformation rather than a direct solution to the methodological requirements of EA delineation. The trial therefore showed that, although Thiessen polygons improved visual continuity, the BBZ output was still fundamentally shaped by the geometry of the input data.

4.1.2. Results of ArcPy-Based Applications Without BBZ

In addition to the BBZ trials, several ArcPy-based workflows were tested to assess whether EA delineation could be performed more explicitly and under tighter procedural control. These applications bypassed the BBZ framework and instead relied on custom logic based on building aggregation, distance-based assignment, and grid-based region growing.

4.1.2.1. Results of Building-Polygon-Based Delineation

This trial was carried out directly on building polygons, without using fishnet cells, Thiessen polygons, or the FeatureToPolygon tool. Buildings were first dissolved by building identifier, residential-unit counts were computed on the resulting polygons, and polygon-based adjacency relations were then constructed. Seed EAs were assigned to buildings falling within the target residential-unit range, and the remaining buildings were grouped through adjacency-based aggregation.

The resulting outputs indicate that this approach preserved the original building geometry but did not produce full spatial coverage. As shown in Figure 4.11, the generated EAs remained tied to individual building polygons. This limitation becomes more evident in Figure 4.12, where the overview and zoomed view show that gaps between buildings were left unfilled. In addition, Figure 4.13 presents an example of the attribute structure associated with a building-based EA and demonstrates that the output remained building-based rather than forming continuous areal units.

Figure 4.11. Building-based EA polygons

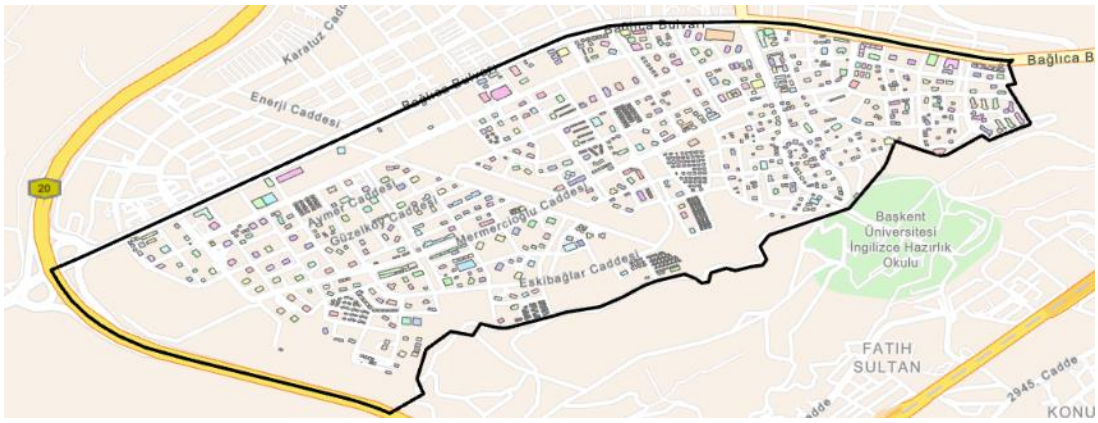
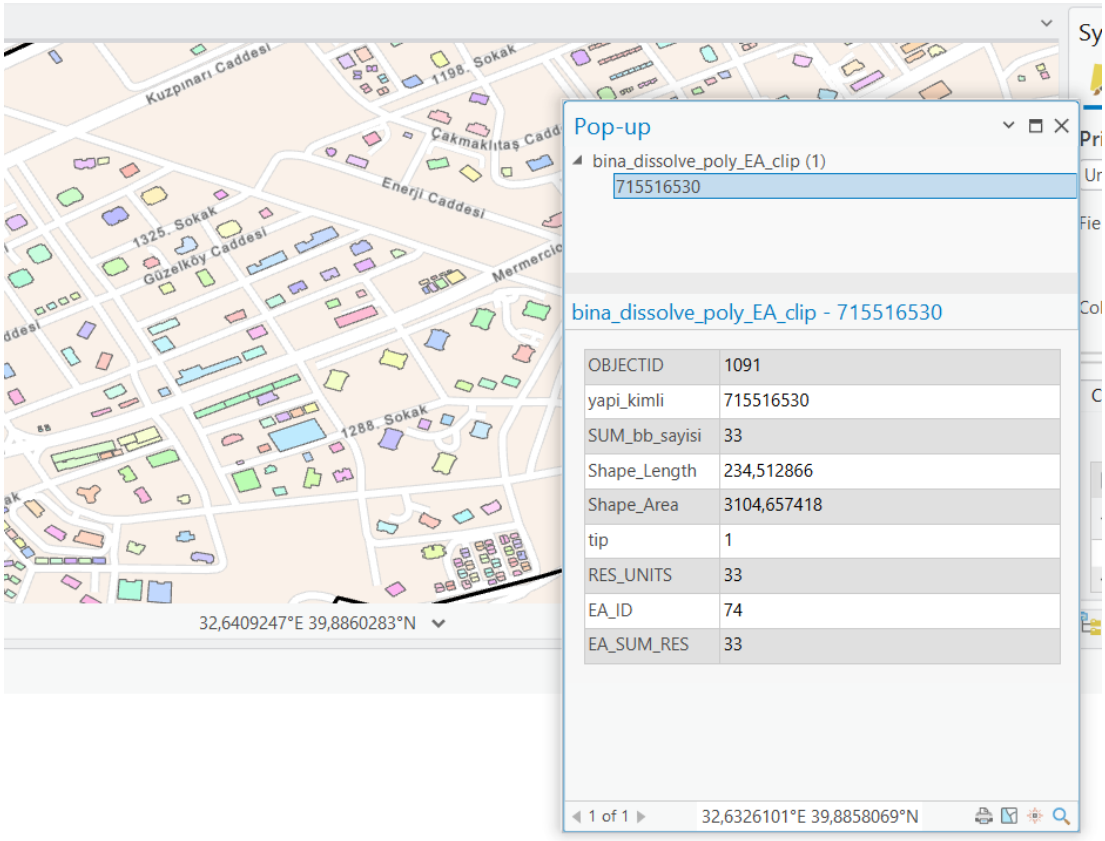


Figure 4.12. Building-based EA polygons: overview and zoomed view, showing gaps between buildings

Figure 4.13. Example pop-up and attribute information for a building-based EA



The diagnostic summary of the trial is presented in Table 4.1. The results show that only two seed EAs could be created initially and that the final output consisted of 940 EA regions. The residential-unit distribution was also highly unbalanced: only two EAs fell within the preferred range of 80-120 residential units, whereas 938 EAs remained below this interval.

Table 4.1. Diagnostic summary of the building-polygon-based EA trial

Indicator	Value
Total residential units in dissolved buildings	9,524
Seed EAs created	2
Remaining buildings to group	1,129
Total EA regions created	940
Minimum EA residential-unit count	0
Average EA residential-unit count	10.1
Maximum EA residential-unit count	109
EAs within target range [80,120]	2
EAs below 80	938
EAs above 120	0
EAs above 180	0

The diagnostic results confirm that the building-polygon-based approach produced a highly fragmented EA structure. Although the method did not exceed the hard upper thresholds, it failed to generate balanced operational units and was therefore not suitable for full-coverage EA delineation in its current form.

4.1.2.2. Results of Building-Point-Based Delineation with Thiessen Polygons

The second ArcPy trial was based on building points and distance-based adjacency, followed by an attempt to convert the resulting point assignments into Thiessen polygons. Neighbour relations were established through GenerateNearTable using an 80.0 m search radius. Seed EAs were assigned to buildings with residential-unit counts in the interval [80,120], and the remaining buildings were grouped iteratively into EA candidates. This procedure produced 186 EA regions at point level before polygon creation.

Figure 4.14 shows the EA assignments on building points prior to polygon generation. Figure 4.15 presents the same output together with the associated attribute table, making the point-based grouping structure more explicit.

Figure 4.14. EA assignment to building points prior to polygon creation

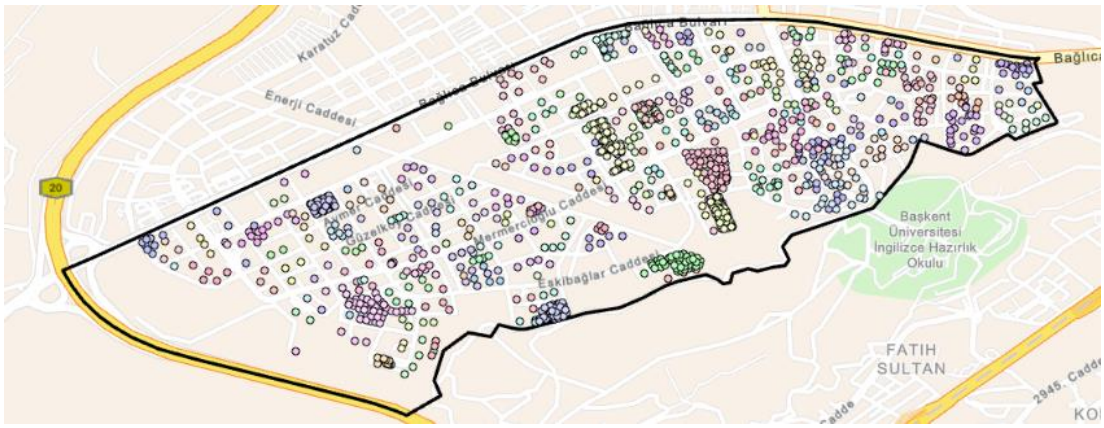
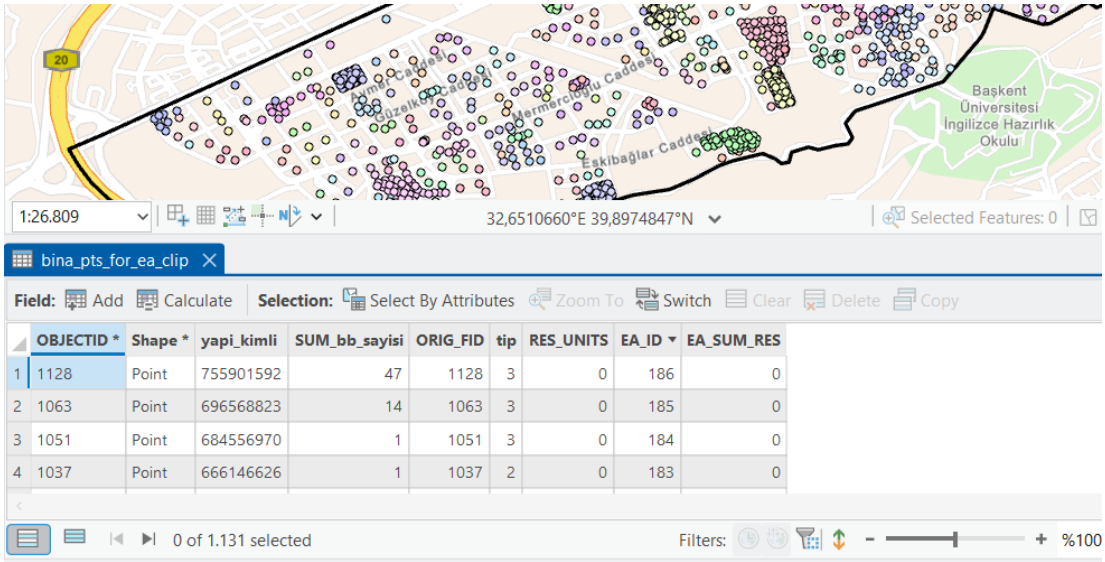


Figure 4.15. EA assignment to building points prior to polygon creation: overview and attribute table



The diagnostic summary of the trial is presented in Table 4.2. The results show that only two seed EAs could be created initially, while 1,129 buildings remained to be grouped. The final point-based output consisted of 186 EA regions before the Thiessen-polygon stage.

Table 4.2. Diagnostic summary of the building-point-based Thiessen trial

Indicator	Value
Search radius	80.0 m
Seed EAs created	2
Remaining buildings to group	1,129
Total EA regions created on building points	186
Polygon generation status	Failed at Thiessen stage
Error type	ERROR 000824

Despite the successful point-level assignment, the workflow could not be completed at the polygon-conversion stage because the CreateThiessenPolygons function was not available under the active ArcGIS Pro licence. As a result, the method could be evaluated only up to the point-assignment stage. Nevertheless, the trial demonstrated that the grouping logic itself was capable of producing candidate EA structures before the areal-conversion step failed.

4.1.2.3. Results of Grid-Based Region Growing

A third and more advanced ArcPy application employed a barrier-aware grid and region-growing logic. This workflow consisted of four main steps: generation of a barrier-aware grid, population-based region growing, contiguous assignment of zero-population cells, and mild merging of very small EAs into neighbouring units. Building adjacency was established through polygon-based neighbour relations, after which a greedy region-growing procedure was applied to form candidate EAs without using BBZ.

Figure 4.16 presents the grid generated for the workflow. Figure 4.17 shows the output table produced by the Summarize Within operation, while Figure 4.18 presents the summarised grid results after clipping to barriers. Taken together, these intermediate outputs show how the grid was prepared for region growing and how residential-unit values were transferred to the grid structure before final EA assignment.

Figure 4.16. Grid generation for the region-growing workflow

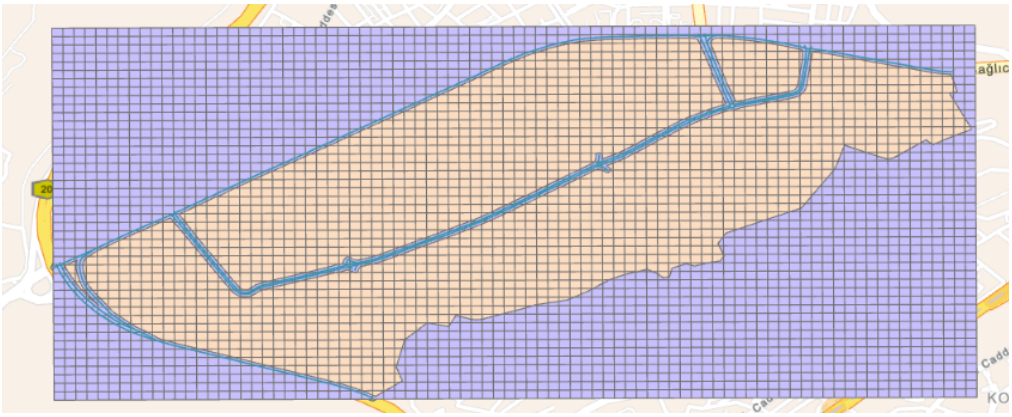
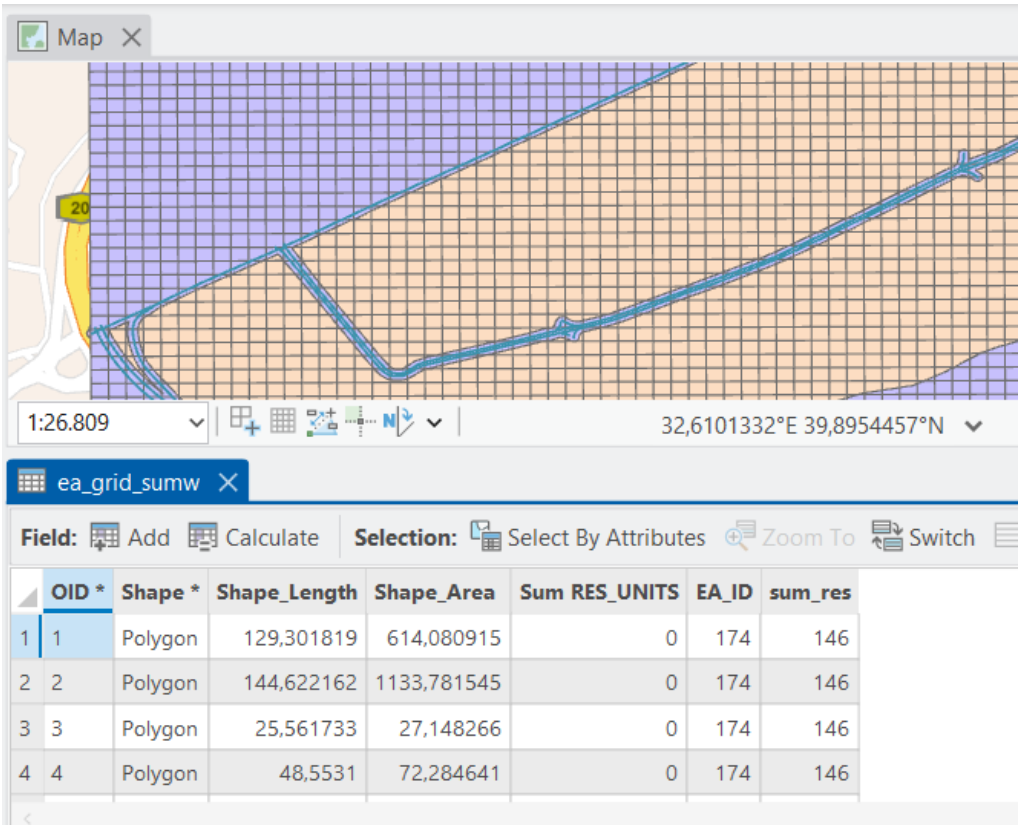
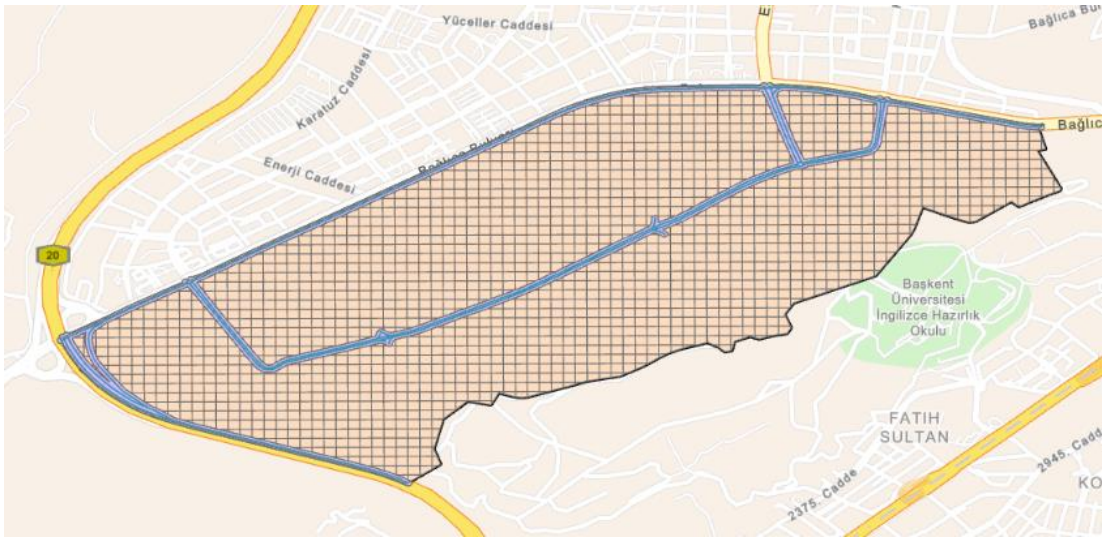


Figure 4.17. Output table from the Summarize Within operation



	OID *	Shape *	Shape_Length	Shape_Area	Sum RES_UNITS	EA_ID	sum_res
1	1	Polygon	129,301819	614,080915	0	174	146
2	2	Polygon	144,622162	1133,781545	0	174	146
3	3	Polygon	25,561733	27,148266	0	174	146
4	4	Polygon	48,5531	72,284641	0	174	146

Figure 4.18. Summarize Within result for grids clipped to barriers



The resulting EA delineation is shown in Figures 4.19 and 4.20, while the relationship between the generated EAs and the building layer is illustrated in Figure 4.21. These figures indicate that the workflow was able to produce continuous areal outputs and to assign the grid structure to final EA units.

Figure 4.19. EA delineation using region growing

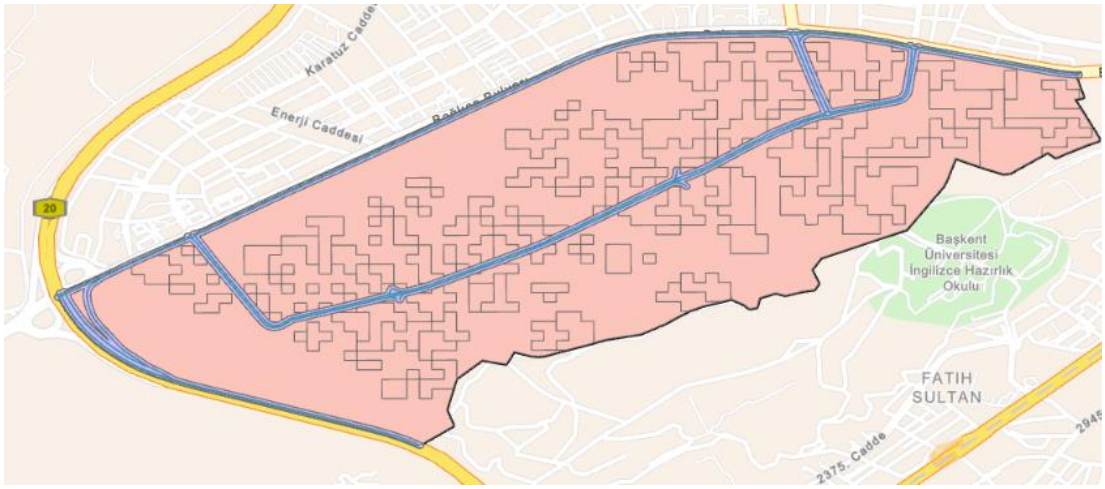
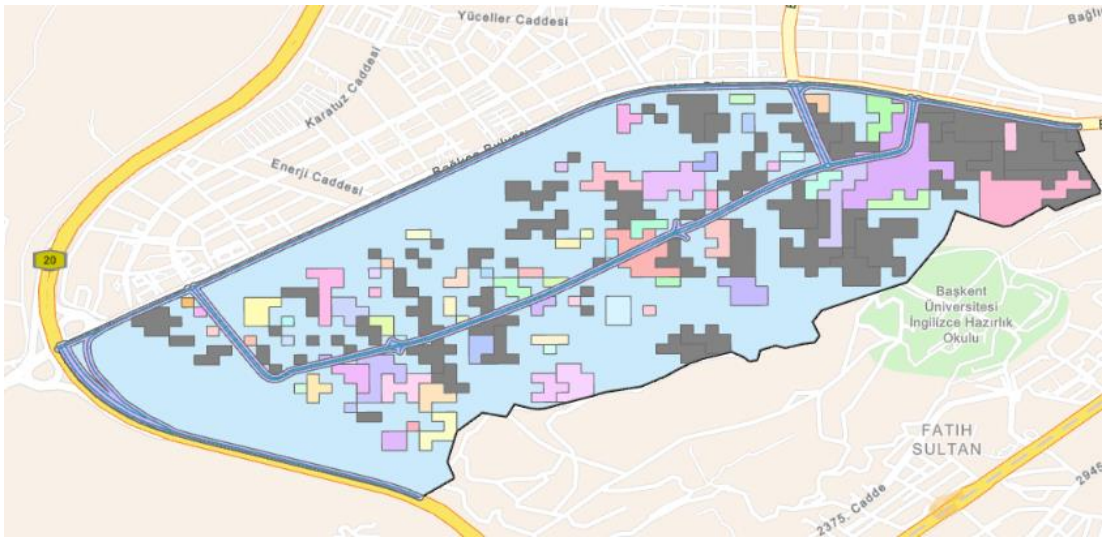


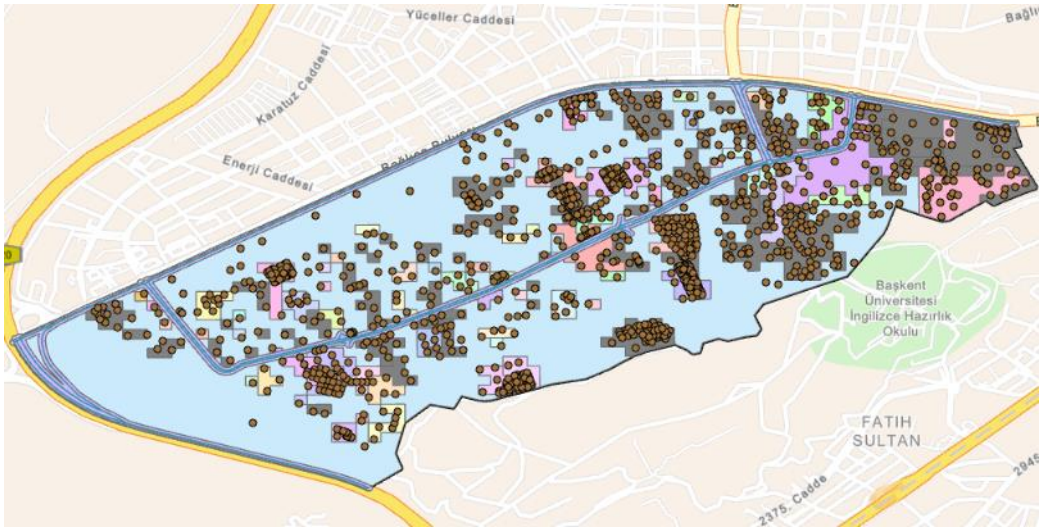
Figure 4.20. EA delineation using region growing in colour-coded form



Note. Polygon colours are cartographic only; they do not encode magnitude, ranking, or category.

Because the number of EAs exceeded 100, ArcGIS Pro displayed some polygons with identical colours; however, these similarly coloured polygons do not belong to the same EA.

Figure 4.21. Display of EAs and buildings



The diagnostic summary of the trial is presented in Table 4.3. The procedure initially produced 203 EA regions before zero-population cells were assigned. After the contiguous assignment of 1,271 zero-residential-unit cells, the number of EA

regions increased to 204. Following the mild merge of very small EAs into neighbouring units, the final number of EAs was reduced to 129.

Table 4.3. Diagnostic summary of the grid-based region-growing trial

Indicator	**Value**
EA regions before zero-cell assignment	203
Zero-residential-unit cells assigned	1,271
EA regions after zero-cell assignment	204
Final EA count after small-EA merge	129
Minimum EA residential-unit count	18
Average EA residential-unit count	73.8
Maximum EA residential-unit count	149
EAs with sum_res < 50	41
EAs with sum_res < 70	64
EAs with sum_res > 100	38
EAs with sum_res > 150	0
EAs with sum_res > 200	0

Despite improving areal continuity, the region-growing procedure also revealed a major limitation. As shown in Figure 4.22, apart from a limited number of small areas, most of the remaining neighbourhood was ultimately represented as a single dominant zone. This outcome is inconsistent with the intended purpose of EA delineation, which requires subdivision into manageable operational units rather than retention of one extensive residual area. Figure 4.23 complements this visual result by presenting the corresponding EA attribute table, including target-unit counts, shape length, and area values. Taken together, these outputs show that the method improved spatial continuity but did not yet produce a satisfactory EA structure.

Figure 4.22. Single large-area output, showing the dominant EA.

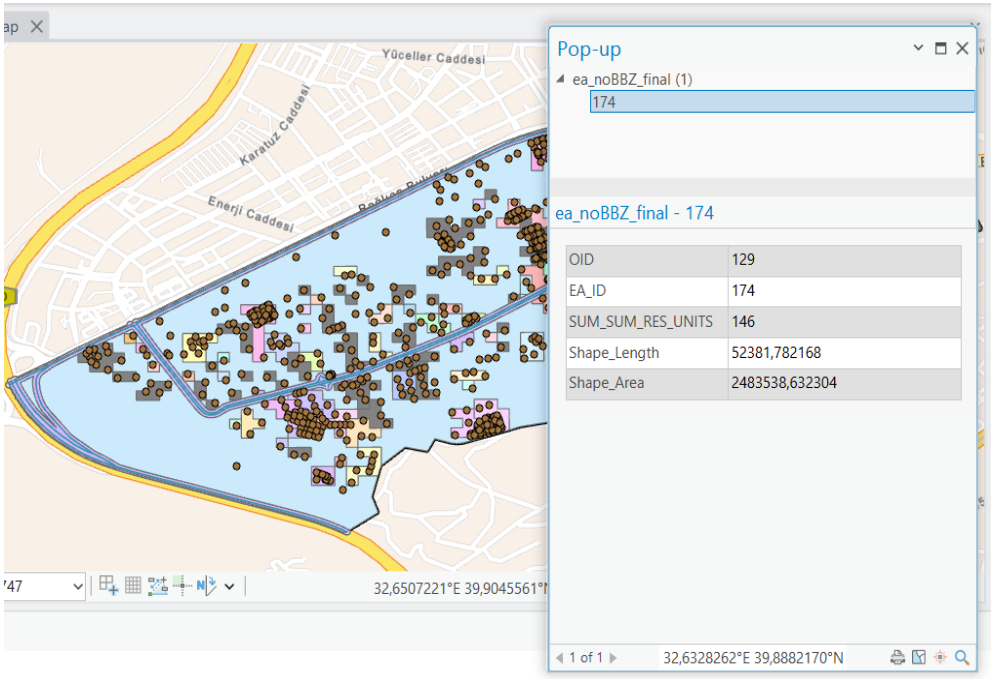
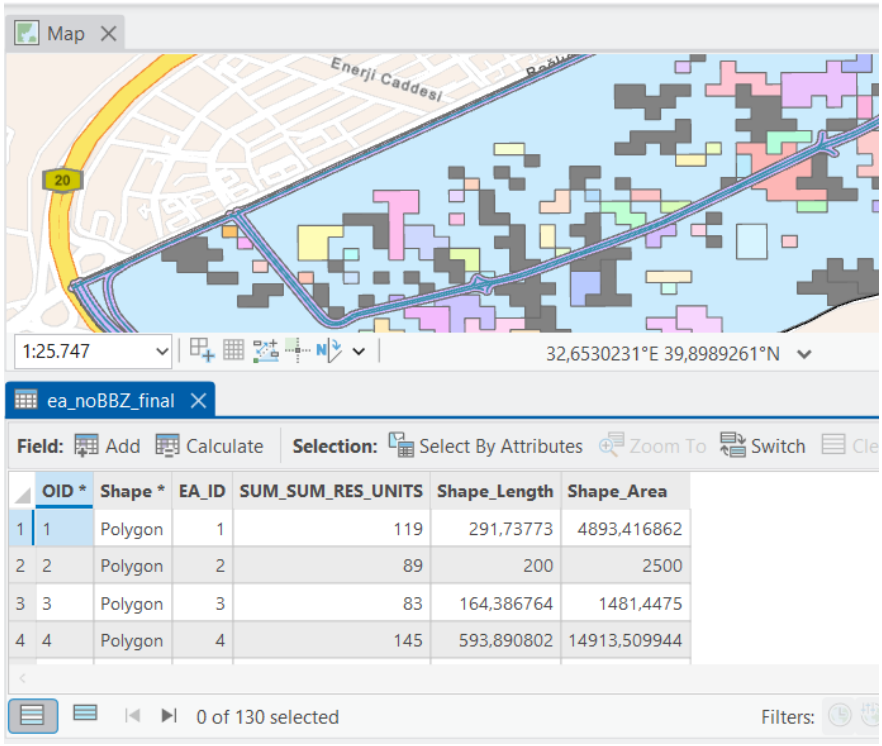


Figure 4.23. EA attribute table including target-unit counts, shape length, and area.



4.1.2.4. Results of License Constraints and Failed Functions

During the thesis period, several scripts that had previously been used to test alternative criteria became non-functional because some ArcGIS Pro functions were no longer available under the active license level, as discussed in the Methodology chapter. These failures affected the feasibility, reproducibility, and comparability of the ArcGIS Pro-based workflows. The principal examples were as follows.

- ea_all_in_one_BBZ_major_rect_v3: ERROR 000357 - End type option invalid with a Basic or Standard license. Failed to execute (Buffer).
- ea_building_thiessen_v3: ERROR 000824 - The tool is not licensed. Failed to execute (CreateThiessenPolygons).
- ea_building_fullcover_v1: ERROR 000824 - The tool is not licensed. Failed to execute (FeatureToPolygon).
- SayimAlani_Olustur: ERROR 000824 - The tool is not licensed. Failed to execute (PolygonToLine).
- ea_pipeline_bina_v6n_bbz_compact_v4_attr_fix: ERROR 000824 - The tool is not licensed. Failed to execute (Erase).

Taken together, the ArcGIS Pro findings show that, although the software environment offered useful exploratory capabilities, both built-in and custom workflows encountered substantial limitations in practice. BBZ-based applications were strongly shaped by input geometry, whereas ArcPy-based alternatives either failed to produce fully satisfactory EA structures or were interrupted by license restrictions. These findings help explain the later transition to a more controlled and reproducible R-based workflow.

4.2. Findings of R Applications

This section presents the findings obtained from the R-based delineation workflow, which constitutes the main production environment of the study. In contrast to the ArcGIS Pro trials discussed in the previous section, the R implementation was designed to generate final EA outputs in a parameterised, reproducible, and batch-oriented manner. The findings are presented at two scales. First, a single-neighbourhood example is used to show the character of the generated outputs and the type of local issues that arise during delineation. Second, the workflow is examined at district scale through a batch application for all neighbourhoods of Çankaya under a reference scenario. Before moving to the scenario comparison, the reference scenario is also interpreted through DEGURBA-based neighbourhood illustrations so that the local meaning of the output is established more clearly. Finally, the district-wide results are compared across alternative scenario settings in order to assess the sensitivity of the workflow to key zero-target and split policies and to identify the strongest-performing configurations under a target-first interpretation.

4.2.1. Single-neighbourhood application

The R workflow generates EAs by combining road- and barrier-aware partitioning with building-based target-unit counts, in accordance with the algorithm described in the Methodology chapter. Figure 4.24 presents a typical neighbourhood-level output, whereas Figure 4.25 illustrates a local case in which an EA remained above the desired target threshold. These figures are included not only to show the cartographic appearance of the output, but also to demonstrate the type of workload and geometry information produced by the workflow at neighbourhood level.

Figure 4.24. Overview of neighbourhood EAs.

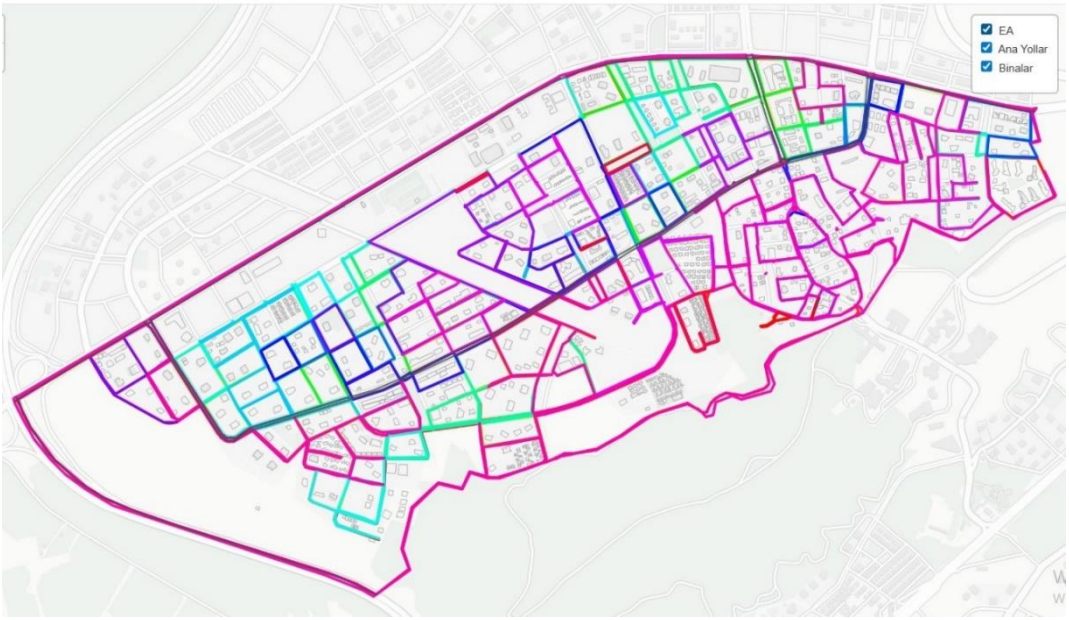
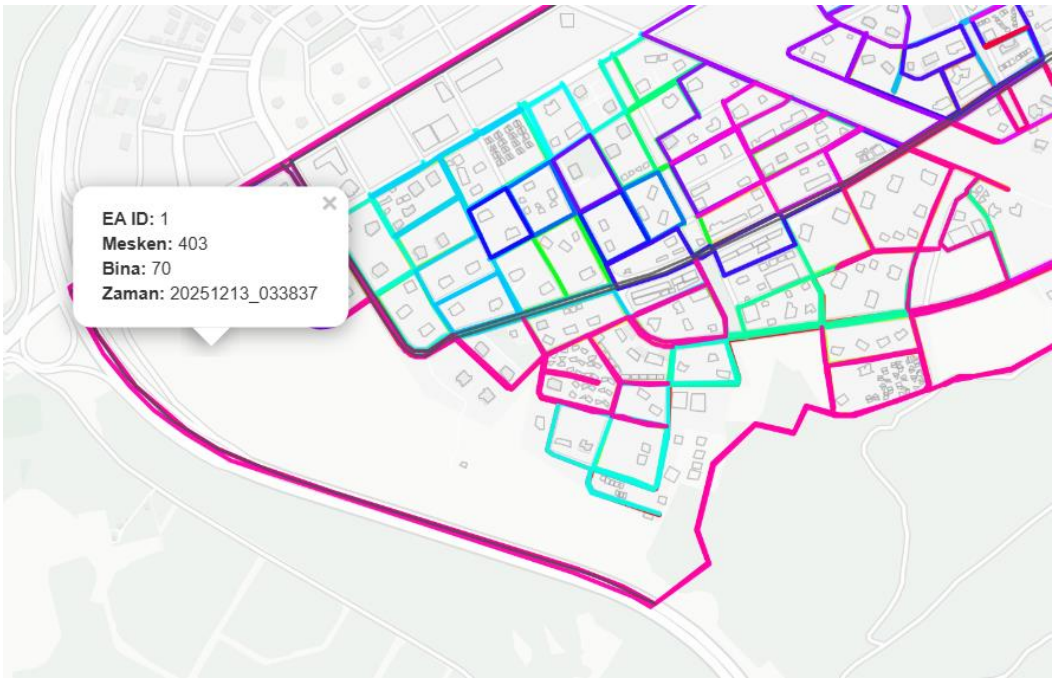


Figure 4.25. Overview of an EA with 403 target units.



In addition to the map output itself, the script produces popup-based attribute summaries and tabular output files for each neighbourhood. These include workload statistics for the generated EAs and supporting files that can be used for quality control and subsequent comparison. The neighbourhood-scale outputs therefore form the most detailed inspection level within the workflow, making it possible to evaluate both the

local strengths of the delineation and the locations where the target balance could not be fully achieved.

As seen in Figure 4.25, some localised areas could not be subdivided as intended under the current barrier and face structure. In such cases, the absence of a suitable internal partition, together with the spatial distribution of buildings, prevented the generation of additional operationally meaningful EAs. The single-neighbourhood example is therefore useful not only as an illustration of successful output production, but also as an introduction to the structural limitations that later become visible at district scale.

Table 4.4. R single-neighbourhood Excel output table 1

Class	**N**
0	38
<80	6
80-120	66
121-150	12
151-200	4
>200	2

Table 4.5. R single-neighbourhood Excel output table 2

EA_sayisi	Min	Median	Mean	Max
128	0	89.5	74.41	403

4.2.2. District batch application under the reference scenario

After the neighbourhood-level example, the workflow was further refined and then implemented in batch mode across all 124 neighbourhoods of Çankaya to assess its operational performance at district scale. The aim at this stage was not only to generate a complete district-wide EA layer, but also to evaluate whether the refined workflow could produce consistent outputs, summary statistics, and map products in a structured and repeatable manner.

The reference scenario used in this district batch run retained zero-target cleanup, enabled hard split, disabled both soft-split options, and retained low-target merging. In the scenario notation used in this chapter, this configuration is denoted as S4 and functions as the main production run against which the alternative settings are interpreted. S4 is retained as the reference scenario not because it is necessarily the strongest-performing configuration under every criterion, but because it provides a comparatively transparent baseline before the more aggressive refinement behaviour associated with the soft-split scenarios is introduced.

4.2.2.1. Output structure and file organisation

The district batch run produced outputs at both neighbourhood and district scales. At neighbourhood level, the workflow generated EA layers, optional SA layers and SA summary files, neighbourhood-level summary tables, and map products in interactive and static formats. At district level, it produced an aggregated EA statistics file and district overview maps. The output structure is methodologically important because it allows the analyst to move directly from district-scale summaries to the corresponding neighbourhood-level files and inspect the local geometry and workload distribution of any case that appears problematic.

The same core workflow could be executed either directly from script or through a local Shiny-based graphical user interface developed for parameter management and operational convenience. Because the interface does not alter the delineation logic itself, its technical details are presented in the appendix rather than in the present findings chapter. In the findings chapter, emphasis is placed on the output products created by the workflow rather than on the interface used to launch it.

Table 4.6 summarises the main output products generated by the district batch application under the reference scenario. These outputs include both spatial layers and supporting summary files, as well as interactive and static map products prepared for inspection and reporting.

Table 4.6. Main output products of the district batch application under the reference scenario

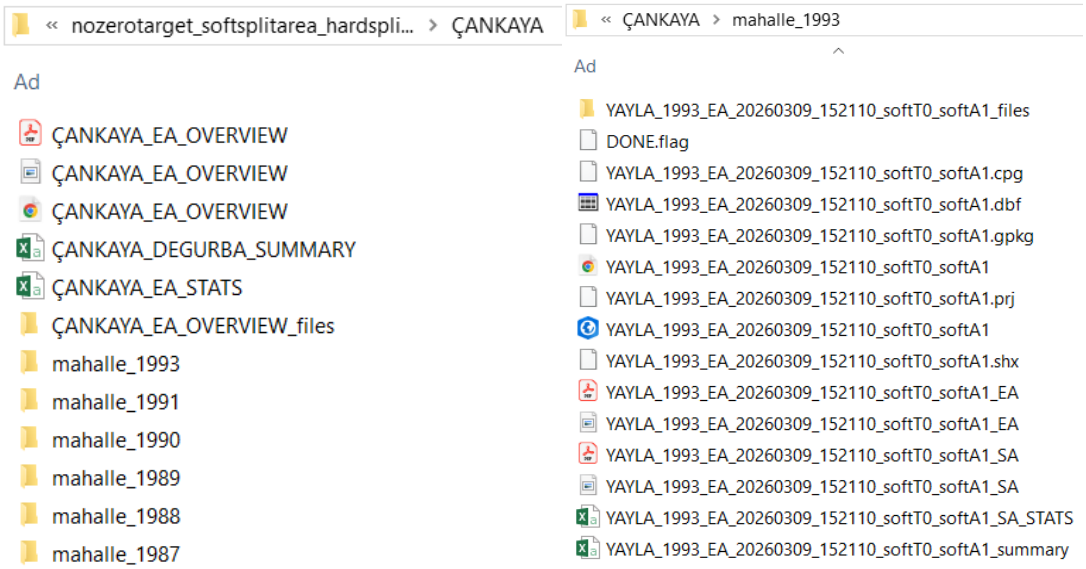
Spatial level	Output product	Purpose	Example file
Neighbourhood	EA GeoPackage (.gpkg)	Stores the final EA layer; the same file can also contain the SA layer when SA construction is enabled.	<Neighbourhood>_<ID>_EA_<timestamp>.gpkg
Neighbourhood	EA summary CSV	Provides neighbourhood-level summary statistics for the generated EAs.	<Neighbourhood>_<ID>_EA_<timestamp>_summary.csv
Neighbourhood	SA statistics CSV	Provides per-SA summary statistics when SA construction is enabled.	<Neighbourhood>_<ID>_EA_<timestamp>_SA_STATS.csv
Neighbourhood	Leaflet HTML / PNG / PDF	Provides interactive and static map products for local inspection and reporting.	<Neighbourhood>_<ID>_EA_<timestamp>.html / _EA.png / _EA.pdf

Table 4.6. Main output products of the district batch application under the reference scenario (continued)

Spatial level	Output product	Purpose	Example file
Neighbourhood	DONE.flag	Marks neighbourhood runs that have already completed successfully in batch mode.	DONE.flag
District	District EA statistics CSV	Combines EA records for the district and supports district-level summary analysis.	ÇANKAYA_EA_STATS.csv
District	District overview HTML / PNG / PDF	Provides district-scale map outputs for rapid inspection and reporting.	ÇANKAYA_EA_OVERVIEW.html / .png / .pdf

The corresponding file organisation is shown in Figure 4.26, where the district-level outputs are presented on the left and the neighbourhood-level outputs on the right.

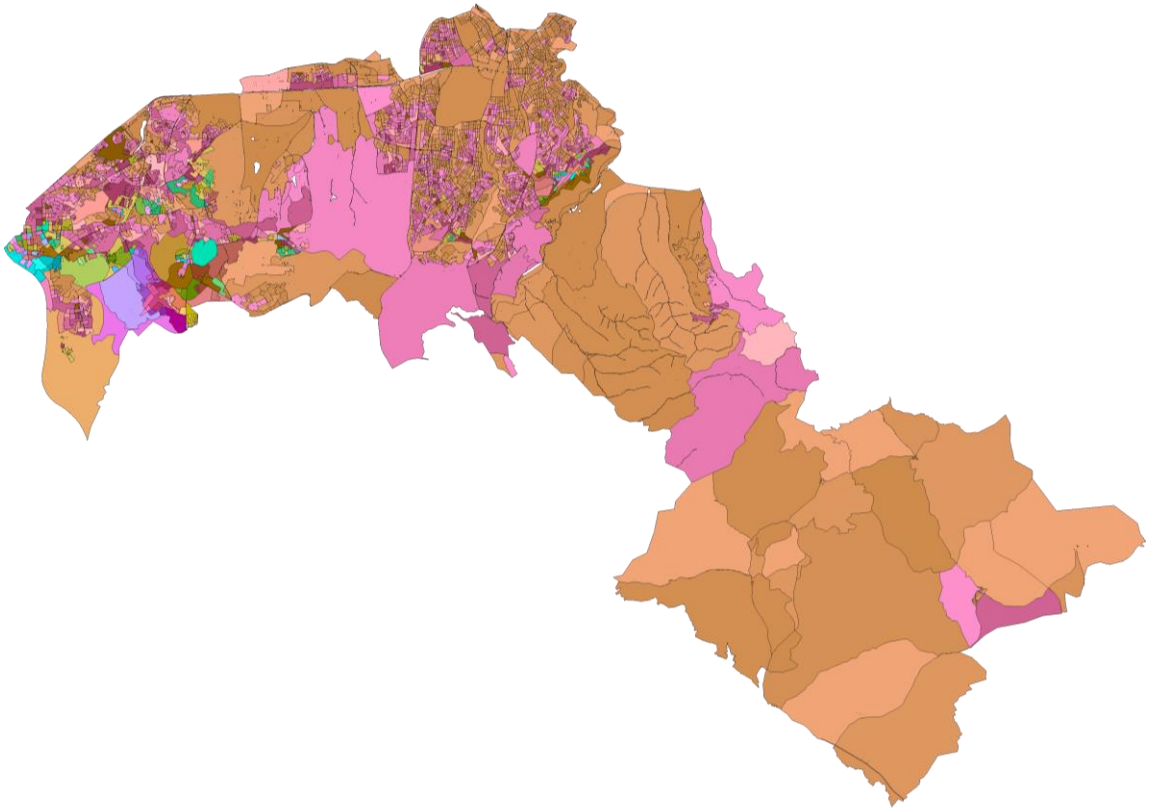
Figure 4.26. Example file structure of the district batch outputs under the reference scenario.



4.2.2.2. District overview output, summary tables and basic statistics

At district level, the workflow produces an overview map that makes it possible to inspect coverage, spatial continuity, and the general distribution of the generated EAs at a glance. This output is analytically useful because it verifies that the workflow has progressed beyond a single-neighbourhood prototype and can operate as a district-scale production system. It also provides a rapid visual entry point for identifying neighbourhoods that may require closer inspection through the corresponding summary tables and neighbourhood-level files.

Figure 4.27. District-level overview of Çankaya EA outputs under the reference scenario (S4)



Here, scenario labels are used only as output identifiers; their detailed definitions and comparative interpretation are provided in Section 4.2.3.

The reference district batch run covers all 124 neighbourhoods of Çankaya. Under the DEGURBA classification used in the workflow, 111 neighbourhoods are classified as dense urban, 5 as medium-density urban, and 8 as rural. Table 4.7 summarises this basic composition together with the corresponding EA counts produced under the reference scenario. As expected, dense-urban neighbourhoods dominate the district both in terms of neighbourhood count and in terms of final EA count.

Table 4.7. Basic district composition of Çankaya by DEGURBA class and reference-scenario EA count

DEGURBA	Neighbourhood count	EA count (S4)	Neighbourhood share (%)	EA share (%)
Dense urban	111	2912	89.52	93.51
Medium-density urban	5	142	4.03	4.56
Rural	8	60	6.45	1.93

Under the reference scenario, the batch application produced 3114 EAs in total. Only 2 zero-target EAs remained district-wide, while 1354 EAs remained above the DEGURBA-specific TARGET_MAX thresholds and 870 positive-workload EAs remained below the corresponding TARGET_MIN thresholds. The district-wide maximum target-unit value remained 9696, which indicates that a small number of extreme local cases persisted despite the split and cleanup rules. At the same time, the district-wide mean target-unit count was 157.40, and the mean calculated only over positive-workload EAs was 157.50.

Table 4.8 presents the reference-scenario EA output summary by DEGURBA class. The table is especially useful because it shows that the district-wide results should not be interpreted through a single undifferentiated average. Dense-urban EAs dominate the output, medium-density urban EAs remain relatively close to the intended workload regime, and rural EAs reflect a much lower workload pattern in line with the chosen thresholds. This table therefore functions as a bridge between the district-level totals and the later scenario-based comparisons.

Table 4.8. Reference-scenario EA output summary by DEGURBA class

DEGURBA	EA count	Zero-target EAs	Min tgt_units	Max tgt_units	Mean tgt_units
Dense urban	2912	2	0	9696	162
Medium-density urban	142	0	1	1777	115
Rural	60	0	1	180	24

4.2.2.3. DEGURBA-based illustrative cases under the reference scenario

To make the reference scenario more interpretable before the chapter moves to explicit scenario comparison, selected neighbourhood cases were reviewed within each DEGURBA class. In this reading, the primary criterion is whether EA workloads fall within the DEGURBA-specific target interval. The area threshold is treated as a secondary control rather than as the principal success criterion, because the main operational aim of the delineation is to achieve a reasonable target-unit balance and area is mainly used as an auxiliary rule indicating whether further subdivision may still be warranted.

The purpose of these examples is therefore not to claim that any neighbourhood achieves perfect compliance under the reference setting. Rather, the aim is to distinguish relatively better cases from clearly problematic ones within each DEGURBA class. This provides a clearer bridge between the district-level statistics and the later scenario comparison.

Table 4.9. DEGURBA-based illustrative neighbourhoods selected from the reference scenario (S4) under a target-first interpretation

DEGURBA class	Relatively better example	Clearly problematic example	Why selected
Dense urban	ATA (1520)	MEBUSEVLERİ (1577)	ATA shows the strongest dense-urban fit to the target interval under S4, whereas MEBUSEVLERİ provides a clear overloaded counterexample.
Medium-density urban	AHLATLIBEL (1619)	ÜNİVERSİTELER (1592)	AHLATLIBEL is the strongest medium-density case under S4, while ÜNİVERSİTELER remains clearly problematic because of persistent overload.
Rural	ÇAVUŞLU (1987)	YAYLA (1993)	ÇAVUŞLU is the least problematic rural case in target terms, while YAYLA is a clear low-workload counterexample.

The dense-urban contrast is illustrated first by the relatively better case of ATA (1520) in Figure 4.28 and then by the clearly problematic case of MEBUSEVLERİ (1577) in Figure 4.29.

Figure 4.28. Reference-scenario dense-urban example: ATA as the relatively better case.

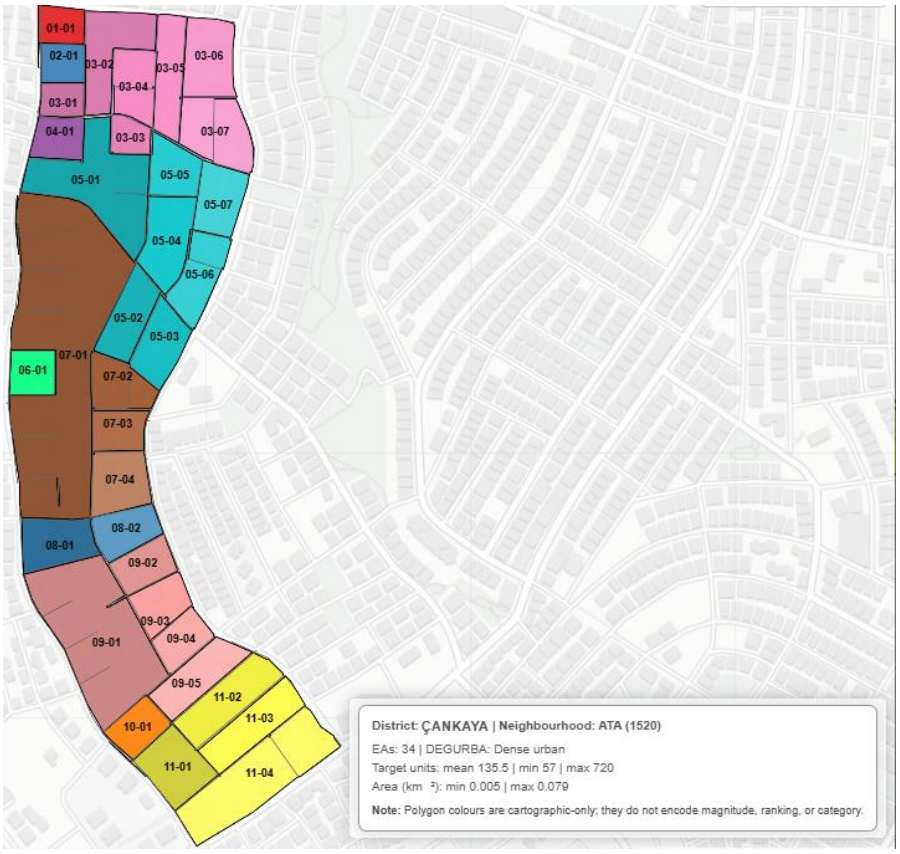
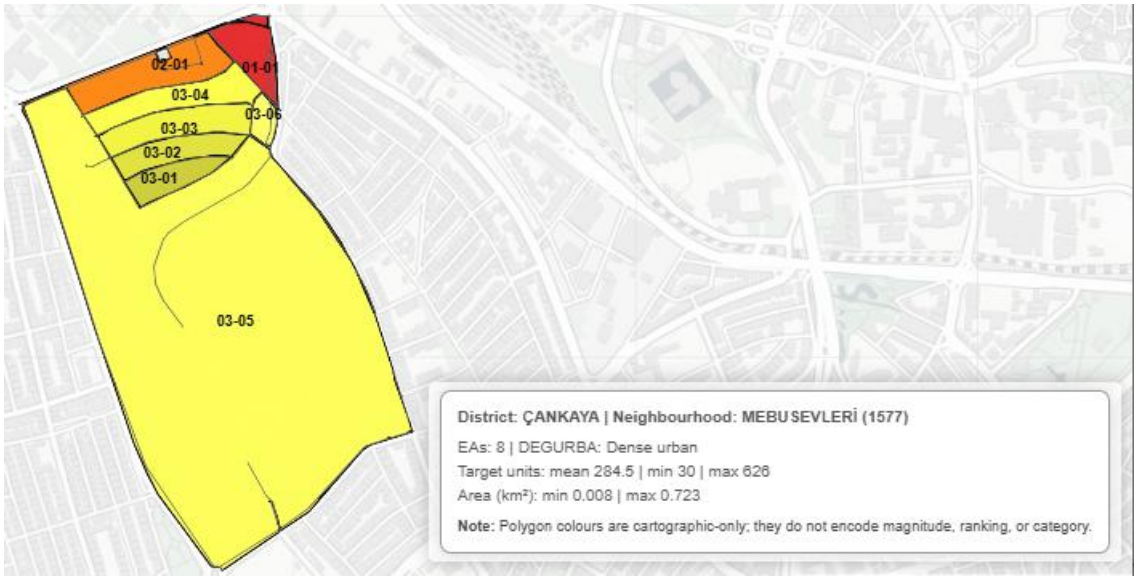


Figure 4.29. Reference-scenario dense-urban example: MEBUSEVLERİ as the clearly problematic case.



The medium-density urban contrast is then shown through AHLATLIBEL (1619) in Figure 4.30 and ÜNİVERSİTELER (1592) in Figure 4.31.

Figure 4.30. Reference-scenario medium-density urban example: AHLATLIBEL as the relatively better case.

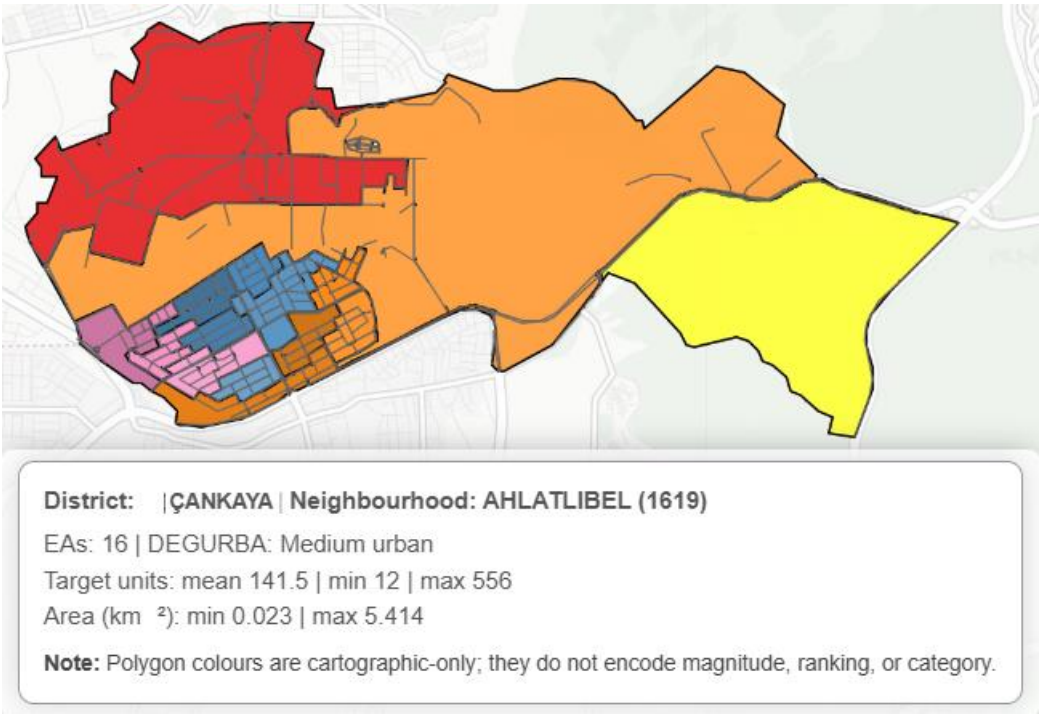
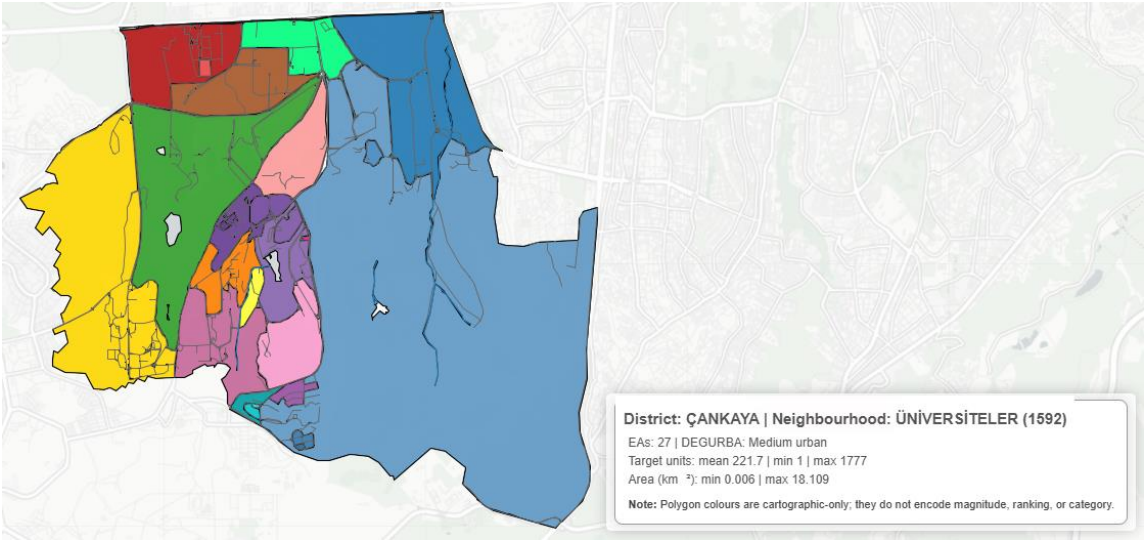


Figure 4.31. Reference-scenario medium-density urban example: ÜNİVERSİTELER as the clearly problematic case.



Finally, the rural contrast is shown through ÇAVUŞLU (1987) in Figure 4.32 and YAYLA (1993) in Figure 4.33.

Figure 4.32. Reference-scenario rural example: ÇAVUŞLU as the least problematic case.

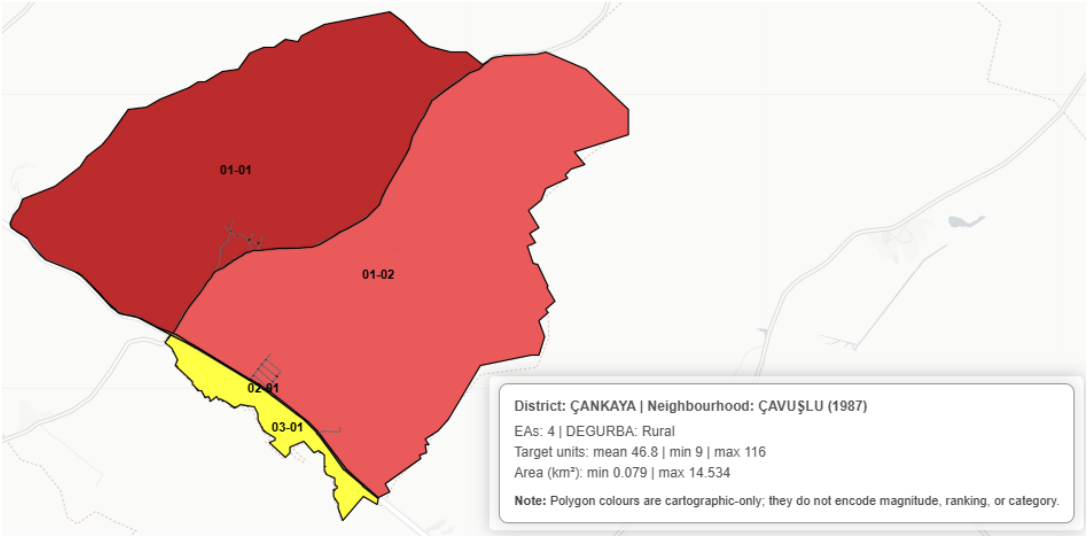
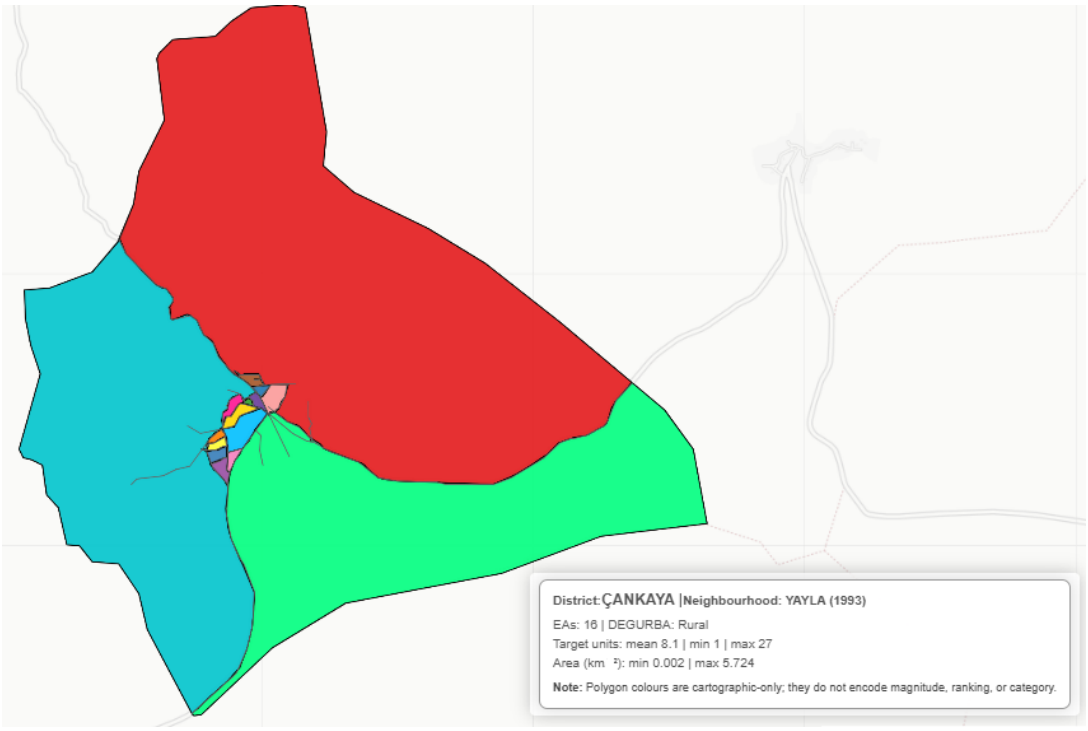


Figure 4.33. Reference-scenario rural example: YAYLA as the clearly problematic case.



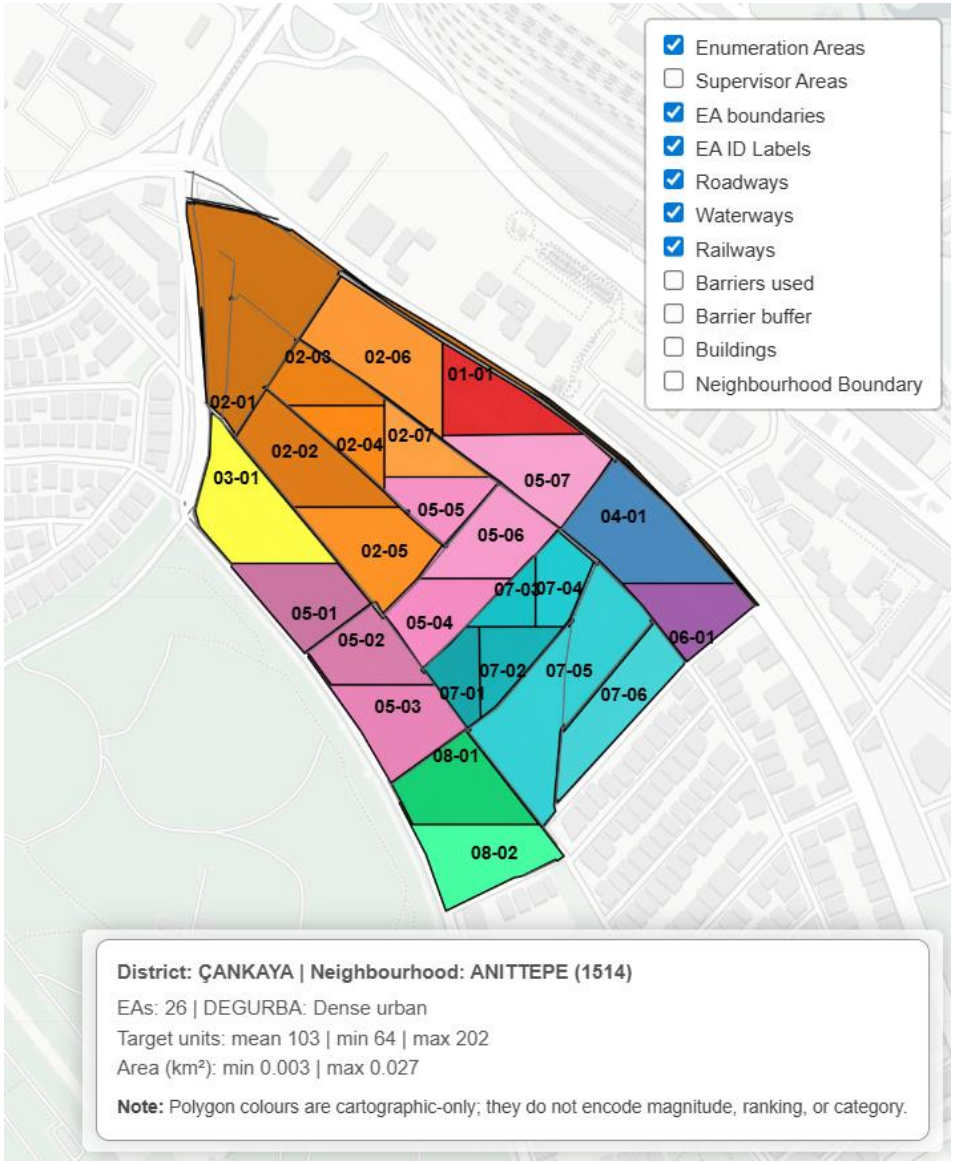
These examples clarify the logic of the reference scenario before the chapter moves to explicit scenario comparison. In dense urban and medium-density urban settings, the contrast between relatively better and clearly problematic neighbourhoods is already visible under S4. In rural settings, by contrast, even the least problematic examples remain weak, which suggests that rural difficulty is not simply a scenario issue but reflects a broader mismatch between observed workload patterns and the imposed design thresholds.

4.2.2.4. Neighbourhood-level EA, SA output and SA summary table

After the DEGURBA-based reference-scenario illustrations, it is also useful to retain one neighbourhood-level output example to show the structure of the products created by the workflow. In this subsection, the neighbourhood figures are used primarily to illustrate output format and hierarchical structure rather than to function as the main substantive evidence for the evaluation of S4. For that reason, the ANITTEPE example is retained here as an output-structure illustration.

The example in Figure 4.34 presents an illustrative neighbourhood-level EA output for ANITTEPE (1514) under Scenario S8. In this case, the workflow produced 26 EAs in a dense-urban setting. The neighbourhood summary panel indicates a mean target load of 103 units, with a minimum of 64 and a maximum of 202. The same panel also shows that EA areas remain relatively small, ranging from 0.003 to 0.027 km². This output is useful because it allows the analyst to evaluate the final EA fabric as an operationally legible local product rather than only as an abstract district-level count.

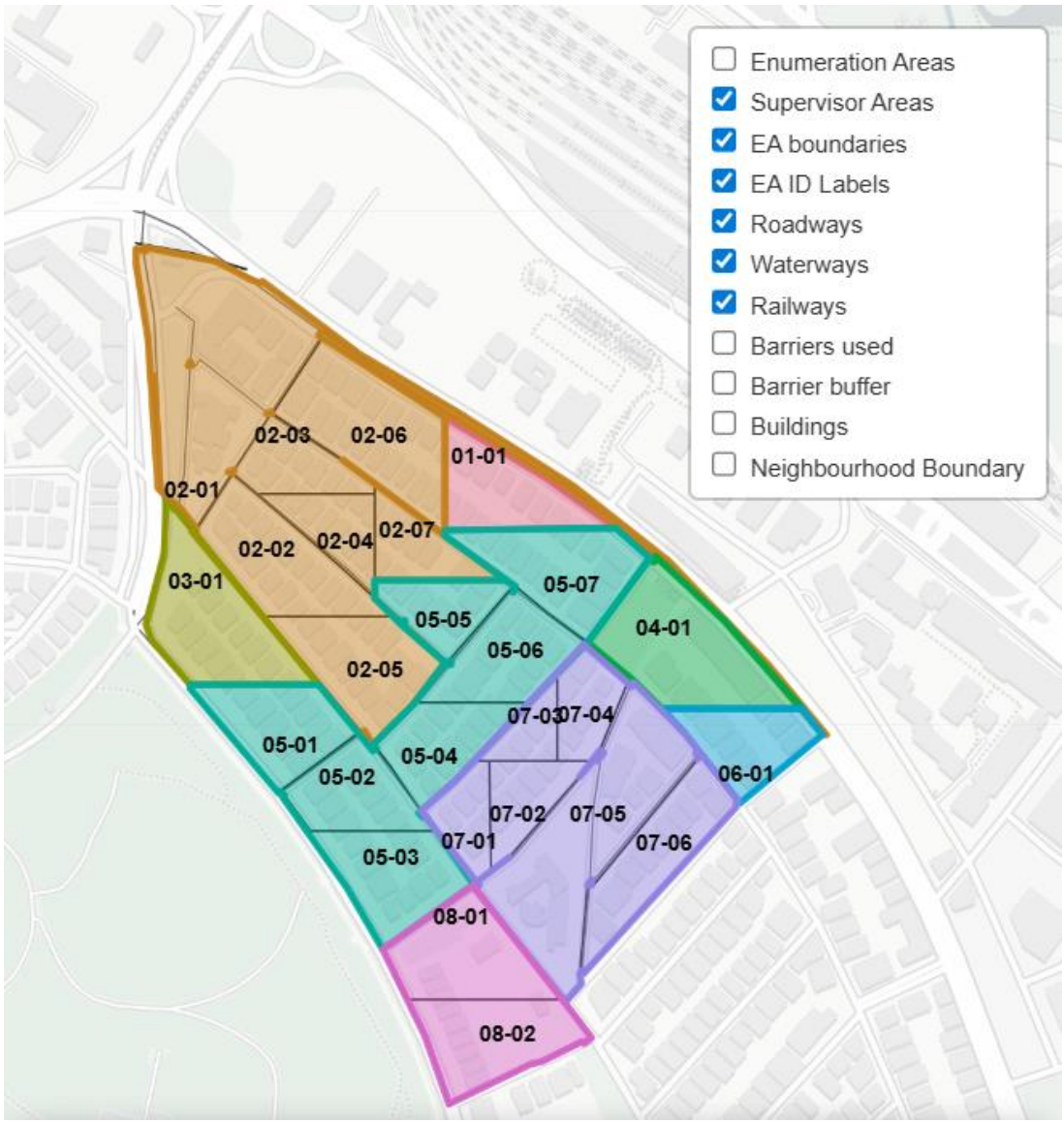
Figure 4.34. Neighbourhood-level EA output and summary panel for ANITTEPE under Scenario S8.



In addition to the EA layer, the workflow also produces a neighbourhood-level SA view that groups the generated EAs into a higher-level supervisory structure. This output is methodologically useful because it demonstrates that the workflow does not stop at EA generation but can also support hierarchical aggregation for operational supervision and review. The SA map provides a clearer view of how lower-level EAs are organised into larger field-management units while preserving neighbourhood-wide coverage.

Figure 4.35 shows the corresponding SA output for ANITTEPE (1514). In this example, the 26 EAs are grouped into 8 SAs. Relative to the EA map, the SA view simplifies the internal structure and makes the broader supervisory pattern easier to inspect. This is especially helpful when evaluating whether EA production also yields a usable second-level operational geography rather than only a set of fine-grained polygons.

Figure 4.35. Neighbourhood-level SA output example for ANITTEPE (1514) under Scenario S8.



To complement the map-based outputs, the workflow generates a neighbourhood-level SA summary table. This table is analytically useful because it translates the spatial grouping visible on the map into a compact tabular form that can be reviewed quickly. In particular, it shows how many EAs were assigned to each SA and how the resulting target-unit and building totals are distributed across the supervisory structure.

Table 4.10 presents the SA summary for ANITTEPE (1514) under Scenario S8. The table shows that the neighbourhood was partitioned into 8 SAs containing between 1 and 7 EAs each. SA target loads range from 76 to 781, while the number of buildings ranges from 8 to 89. The largest supervisory workloads are concentrated in SA-02 and SA-05, whereas several SAs consist of only a single EA. This kind of table is useful because it links the visual SA structure to an explicit operational summary and supports rapid identification of supervisory units that may require closer review.

Table 4.10. Example SA summary table for ANITTEPE under Scenario S8.

SA_ID	SA_LONG_ID	Area (km²)	Number of EAs	Target units	Number of buildings
1	1231-1514-01	0.0083	1	113	10
2	1231-1514-02	0.0783	7	781	89
3	1231-1514-03	0.0125	1	107	12
4	1231-1514-04	0.0140	1	89	11
5	1231-1514-05	0.0570	7	697	78
6	1231-1514-06	0.0069	1	76	8
7	1231-1514-07	0.0478	6	620	53
8	1231-1514-08	0.0196	2	195	17

Taken together, the neighbourhood-level EA map, SA map, and SA summary table show that the workflow produces not only district-scale totals but also locally interpretable hierarchical outputs. This is important because it allows evaluation to move back and forth between district-wide summaries and neighbourhood-specific internal structure.

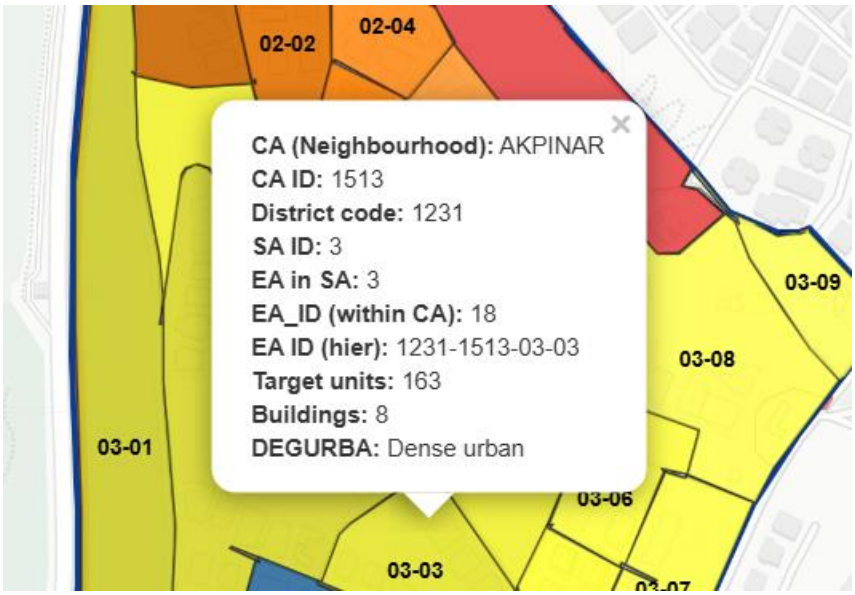
4.2.2.5. EA–SA–CA numbering and ordering scheme

The district batch output does not only generate polygons; it also produces a hierarchical identification and ordering structure that supports field use, quality control, and traceability between files. Within the workflow, the CA corresponds to

the neighbourhood boundary, the SA represents the higher-level grouping of EAs within each CA, and the EA is the final operational enumeration unit. This hierarchy is reflected in both the attribute structure and the output filenames.

At neighbourhood level, CA identifiers correspond to the neighbourhood code, SA identifiers are assigned within each CA, and EA identifiers are assigned within each SA. The long identifier field links these levels explicitly in a district-CA-SA-EA format, such as 1231-1513-03-03. This structure makes it possible to move directly from a district summary table to a specific neighbourhood folder, and from there to a specific CA, SA, or EA record in the output layer. The ordering logic is therefore not only cartographic, but also operational and documentary. This numbering and ordering logic is illustrated in Figure 4.36.

Figure 4.36. Example of the EA–SA–CA numbering and ordering scheme under the reference scenario.



4.2.2.6. EA output summaries under the reference scenario

The reference-scenario summaries show that the district batch workflow is capable of producing a complete output system rather than a single map. The combined use of neighbourhood folders, district-level summaries, and hierarchical identifiers allows the analyst to move between overview and detail in a reproducible manner. This

is one of the main practical strengths of the R implementation when compared with the earlier exploratory GIS trials.

At the same time, the reference scenario also reveals that a complete output system does not automatically imply complete workload resolution. Even under S4, a limited number of neighbourhoods remain problematic, especially where the internal structure available for splitting is weak or where the building distribution remains highly concentrated. These reference-scenario patterns provide the baseline against which the alternative scenario runs are interpreted in the next subsection.

4.2.3. Scenario-based district batch comparison

Once the district-wide reference run had been established, a scenario-based comparison was conducted to evaluate how the workflow responded to alternative zero-target and split policies. The purpose of this comparison was not to treat each scenario as a different method, but to assess the sensitivity of the same delineation workflow under different operational settings. For that reason, the reference scenario introduced in the previous subsection serves as the baseline against which the remaining configurations are interpreted.

4.2.3.1. Scenario design and evaluation criteria

Eight district-wide scenario runs were retained for interpretation. These scenarios vary the zero-target policy, the hard-split setting, and the use of target-based and area-based soft split. In all of them, low-target merge remained enabled so that the comparison could isolate the main effects of the split and zero-target policy settings. Table 4.11 summarises the scenario matrix.

The scenario outputs were evaluated by comparing both district-level totals and neighbourhood-level EA summaries. The main indicators used in the comparison were total EA count, number of zero-target EAs, number of EAs above TARGET_MAX, number of EAs below TARGET_MIN, maximum target-unit value, and mean target-unit value. Because these indicators operate on different numerical scales, a visual

synthesis was also prepared in normalized form so that similarities and contrasts across scenarios could be interpreted more directly.

Table 4.11. Scenario matrix used in the Çankaya comparison

Scenario	Zero-target policy	Hard split	Soft split	Short description
S1	Allow zero-target EAs	Off	Off	Allow-zero / soft off / hard split off / low-target merge on
S2	Allow zero-target EAs	On	Off	Allow-zero / soft off / hard split on / low-target merge on
S3	Merge zero-target EAs where possible	Off	Off	No-zero / soft off / hard split off / low-target merge on
S4	**Merge zero-target EAs where possible**	**On**	**Off**	**No-zero / soft off / hard split on / low-target merge on (reference)**
S5	Merge zero-target EAs where possible	Off	Target only	No-zero / soft split by target / hard split off / low-target merge on
S6	Merge zero-target EAs where possible	Off	Area only	No-zero / soft split by area / hard split off / low-target merge on
S7	Merge zero-target EAs where possible	Off	Target + area	No-zero / soft split by target+area / hard split off / low-target merge on
S8	Merge zero-target EAs where possible	On	Target + area	No-zero / soft split by target+area / hard split on / low-target merge on

At district level, the overall comparison is easier to read visually than through raw values alone. Figure 4.37 therefore presents a normalized comparison of the district-level summary metrics across S1-S8, while retaining the actual values inside the cells for reference.

Figure 4.37. Normalized comparison of district-level scenario summary metrics across S1-S8.

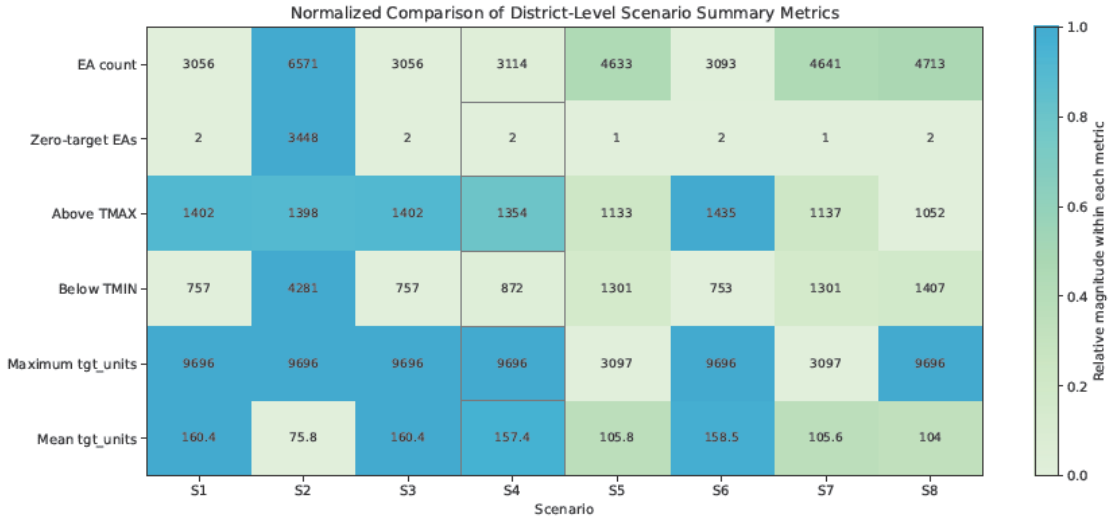


Figure 4.37 visualises the district-level summary metrics in normalized form so that similarities and contrasts can be compared across indicators with different scales. The actual metric values are retained inside the cells for reference, while the colour intensity shows the relative magnitude of each scenario within a given metric. The boxed S4 column marks the reference scenario used in the remainder of the comparison. The figure also confirms the district-level contrast numerically. S1 and S3 are identical across all reported summary metrics, each producing 3,056 EAs, 2 zero-target EAs, 1,402 EAs above TMAX, 757 below TMIN, a maximum target load of 9,696, and a mean target load of 160.4. By contrast, S2 is clearly the most problematic scenario: it produces 6,571 EAs and 3,448 zero-target EAs while leaving the district maximum unchanged at 9,696. S4 was therefore retained as the reference configuration not because it is uniformly the strongest-performing scenario, but because it provides a transparent baseline for neighbourhood-scale interpretation before more aggressive refinement is introduced. Among the refinement scenarios, S6 remains comparatively weak. By comparison, S5 and S7 reduce the district maximum to 3,097 and bring the mean target load down to 105.8 and 105.6, respectively, whereas S8 yields the lowest number of EAs above TMAX at 1,052 but retains the district maximum at 9,696. These contrasts indicate that no single refinement scenario is uniformly strongest across all district-level indicators and that the more specific

differences are better interpreted later through a DEGURBA-wise target-first comparison.

The neighbourhood cases selected for more detailed scenario-focused interpretation are listed in Table 4.12.

Table 4.12. Recommended neighbourhood cases for scenario-focused EA outputs

Scenario family	Recommended neighbourhood	Why it is useful	Main message for discussion
Zero-target policy under hard split on	ÜNİVERSİTELER (1592)	Most striking medium-density case; S2 multiplies EA count and zero-target EAs while leaving the main overload unchanged.	Allow-zero becomes consequential mainly after hard split and may inflate the partition without solving overload.
Zero-target policy under hard split on	ALACAATLI (1899)	Strong dense-urban companion case; S2 nearly doubles the partition and produces many zero-target units.	Allowing zero-target fragments may create many weak or empty units without resolving the main workload peak.
Hard split effect	KIRKKONAKLAR (1568)	Shows a modest positive effect: overloaded EAs decrease slightly, but underloaded EAs increase.	Hard split can improve overload only modestly and often through a trade-off with smaller units.
Hard split effect	İLKBAHAR (1598)	Shows a null effect under the soft-off comparison.	In some neighbourhoods, hard split has no effective action, likely because no internally valid split structure is available.
Soft split by target	YUKARI DİKMEN (1610)	Most dramatic partial-improvement case.	Target-based soft split is the strongest corrective mechanism, but it does not fully solve structurally difficult overload.

Table 4.12. Recommended neighbourhood cases for scenario-focused EA outputs (continued)

Scenario family	Recommended neighbourhood	Why it is useful	Main message for discussion
Soft split by target	EMEK (1543)	Balanced trade-off case: overload is reduced, but underloaded EAs increase.	Soft target split improves balance at the cost of added fragmentation.
Soft split by area only	BEYTEPE (1616)	Shows that area-only soft split is weaker than target-based soft split in a high-pressure dense-urban case.	Area-based soft split alone is comparatively weak where workload pressure is driven by concentrated building stock.
Target-only versus both	KIRKKONAKLAR (1568)	One of the few neighbourhoods where target-only and both are not identical.	Adding area-based soft split on top of target-based split yields only marginal extra benefit in most cases.
Interaction between hard split and later refinement	İLKBAHAR (1598)	Strong interaction case: S7 improves the case, whereas S8 partly reverses that improvement.	Hard split should not be narrated as universally beneficial; its interaction with later refinement can be non-monotonic.
Persistent unresolved hotspots	İLKBAHAR (1598)	Retains a very high maximum target load in several scenarios.	Some overloads are structurally persistent and cannot be removed by toggling split/merge options alone.
Persistent unresolved hotspots	BEYTEPE (1616)	Changes under many scenarios, but never reaches a fully satisfactory balance.	Partial mitigation is not the same as resolution; this points to structural or data-driven limits.
Persistent unresolved hotspots	ALACAATLI (1899)	Useful both for the zero-target policy and for unresolved dense-urban overload.	High-pressure dense urban areas remain difficult even under more aggressive refinement rules.

4.2.3.2. Zero-target policy under hard split on

The first comparison isolates the effect of zero-target policy while keeping hard split active. The comparison between S2 and S4 shows that allowing zero-target EAs can increase the number of locally fragmented outputs, whereas merging zero-target EAs where possible produces a cleaner operational structure. This contrast is illustrated by ÜNİVERSİTELER in Figure 4.38 and ALACAATLI in Figure 4.39.

Figure 4.38. EA outputs for ÜNİVERSİTELER under Scenarios S2 and S4.

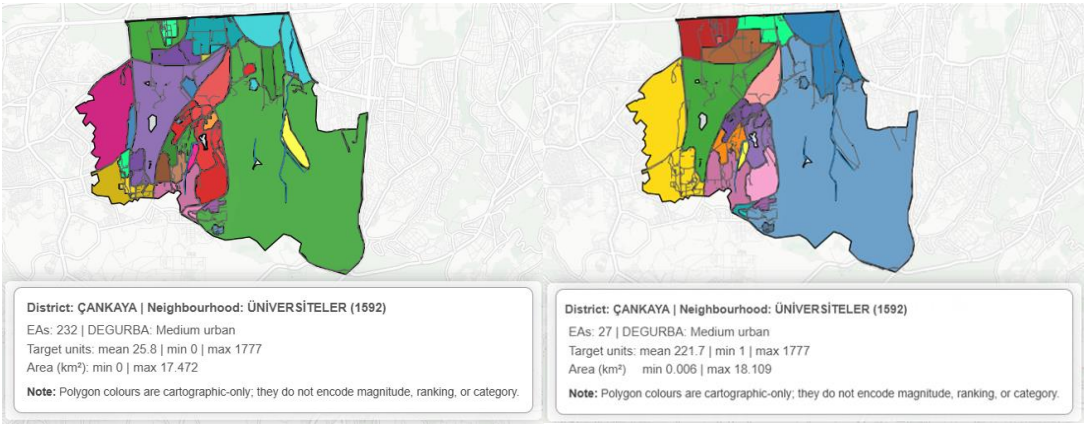
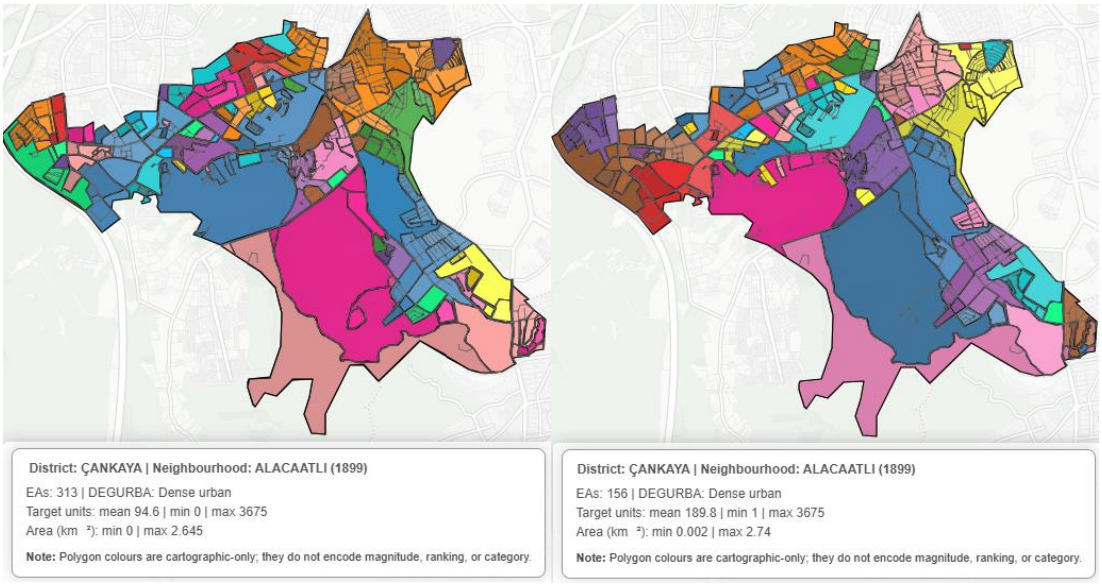


Figure 4.39. EA outputs for ALACAATLI under Scenarios S2 and S4.



Under hard split on, the contrast between the allow-zero and no-zero policies becomes sharp and visually meaningful. In ÜNİVERSİTELER (1592), S4 produces 27 EAs,

whereas S2 produces 232 EAs, of which 205 are zero-target, while the maximum target load remains 1,777 in both cases. In ALACAATLI (1899), S4 produces 156 EAs, whereas S2 produces 313 EAs, including 161 zero-target EAs, again without reducing the main workload peak, which remains 3,675 in both cases. These cases show that the zero-target policy should be discussed together with the split structure that makes such fragments visible; allowing zero-target EAs under hard split can substantially inflate the final partition without improving the most critical overloads.

4.2.3.3. Hard split

The second comparison isolates the effect of hard split by comparing S3 and S4. The resulting examples are analytically important because they show that the hard-split switch does not necessarily produce a visible change in every neighbourhood. In some places, the internal geometry and barrier structure simply do not generate a valid additional split opportunity, so the outputs remain effectively unchanged even when hard split is enabled. This null-effect pattern is illustrated by KIRKKONAKLAR in Figure 4.40 and İLKBAHAR in Figure 4.41.

Figure 4.40. EA outputs for KIRKKONAKLAR under Scenarios S3 and S4, a null-effect case.

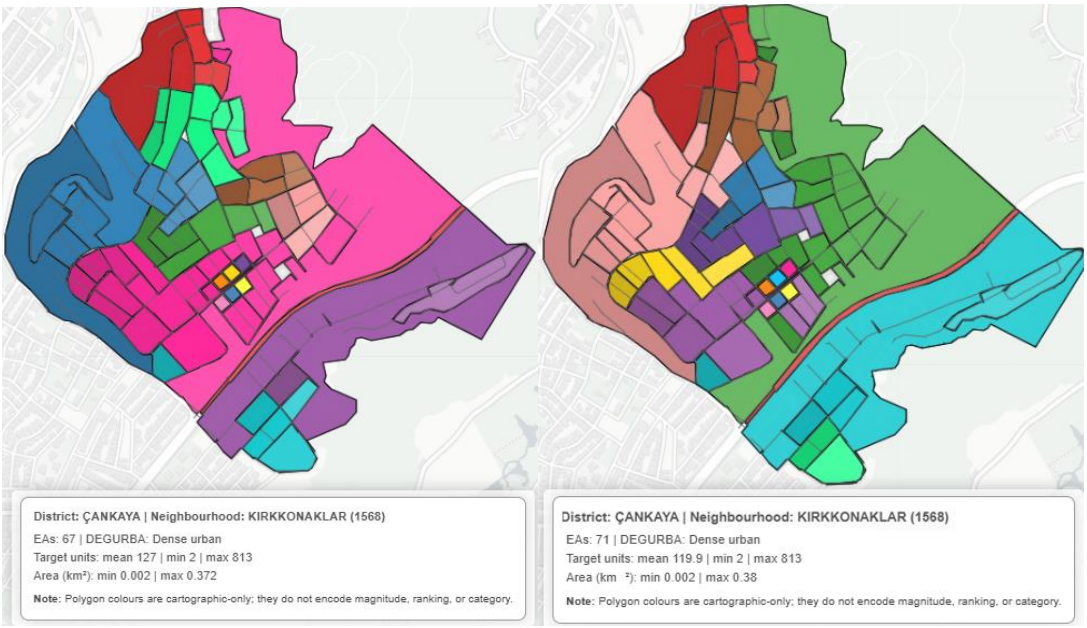
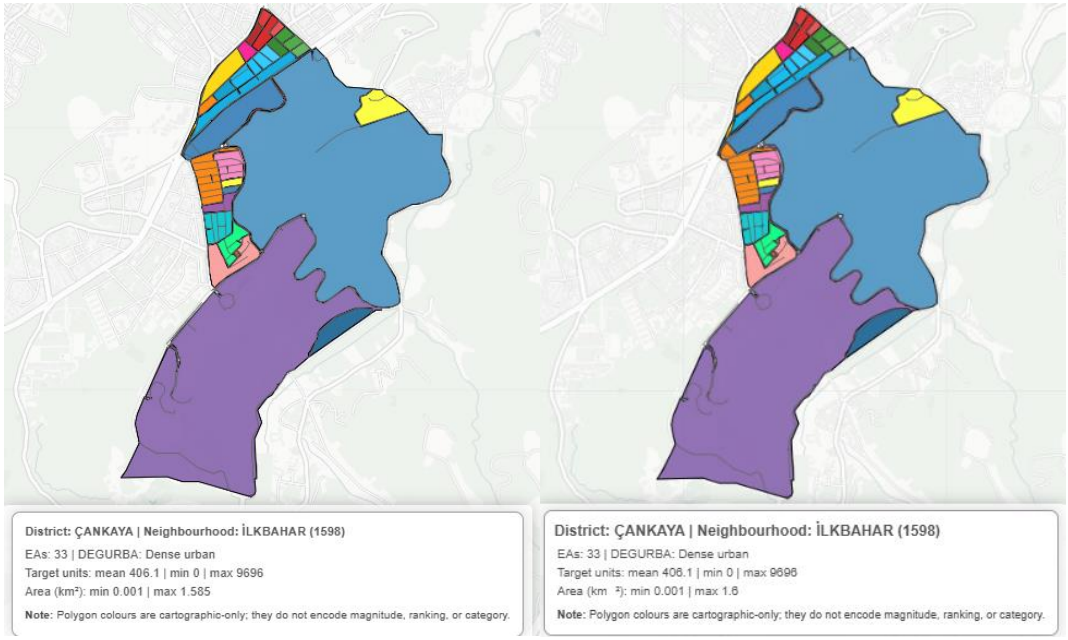


Figure 4.41. EA outputs for İLKBAHAR under Scenarios S3 and S4, a null-effect case.



The hard-split comparison shows that its effect is uneven and sometimes null. In KIRKKONAKLAR (1568), the shift from S3 to S4 increases the EA count from 67 to 71, reduces the number of EAs above TMAX from 23 to 19, but increases the number below TMIN from 20 to 30, while leaving the maximum target load unchanged at 813. By contrast, İLKBAHAR (1598) is effectively unchanged under the same comparison: both S3 and S4 produce 33 EAs, 1 zero-target EA, 6 EAs above TMAX, 21 below TMIN, and the same extreme maximum of 9,696. These results show that hard split can sometimes produce a modest local improvement, but in other neighbourhoods it has no effective action because no internally valid split structure is available.

4.2.3.4. Soft split by target

The comparison between S3 and S5 isolates the effect of target-based soft refinement. In this setting, the workflow attempts additional subdivision specifically where workload concentration remains high and where a valid local improvement can still be achieved. This makes S5 analytically important because it represents the clearest workload-driven refinement logic among the retained scenarios. This

refinement effect is illustrated by YUKARI DİKMEN in Figure 4.42 and EMEK in Figure 4.43.

Figure 4.42. EA outputs for YUKARI DİKMEN under Scenarios S3 and S5.

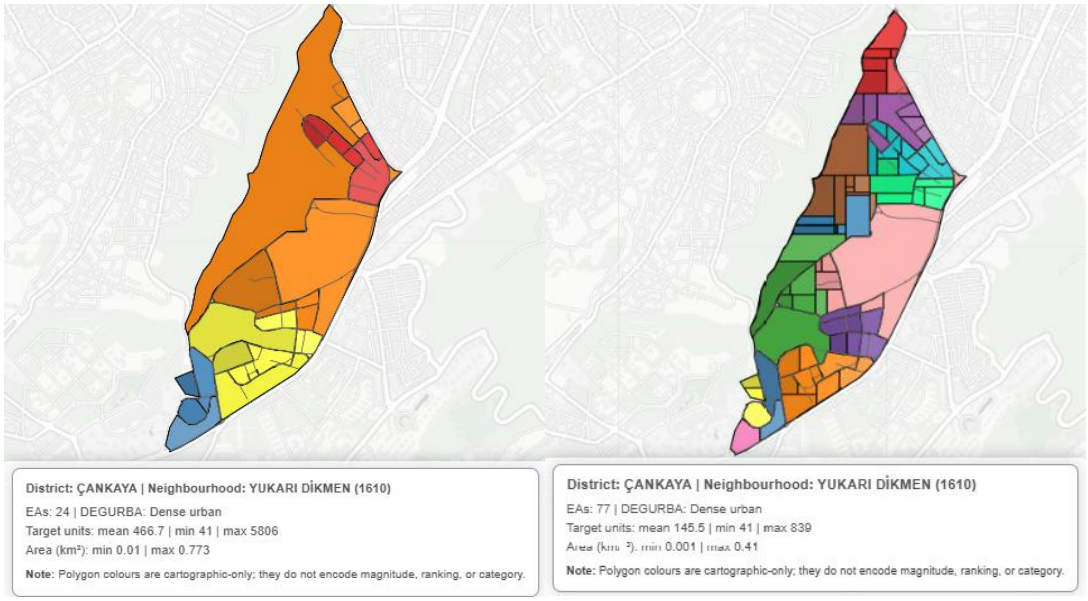
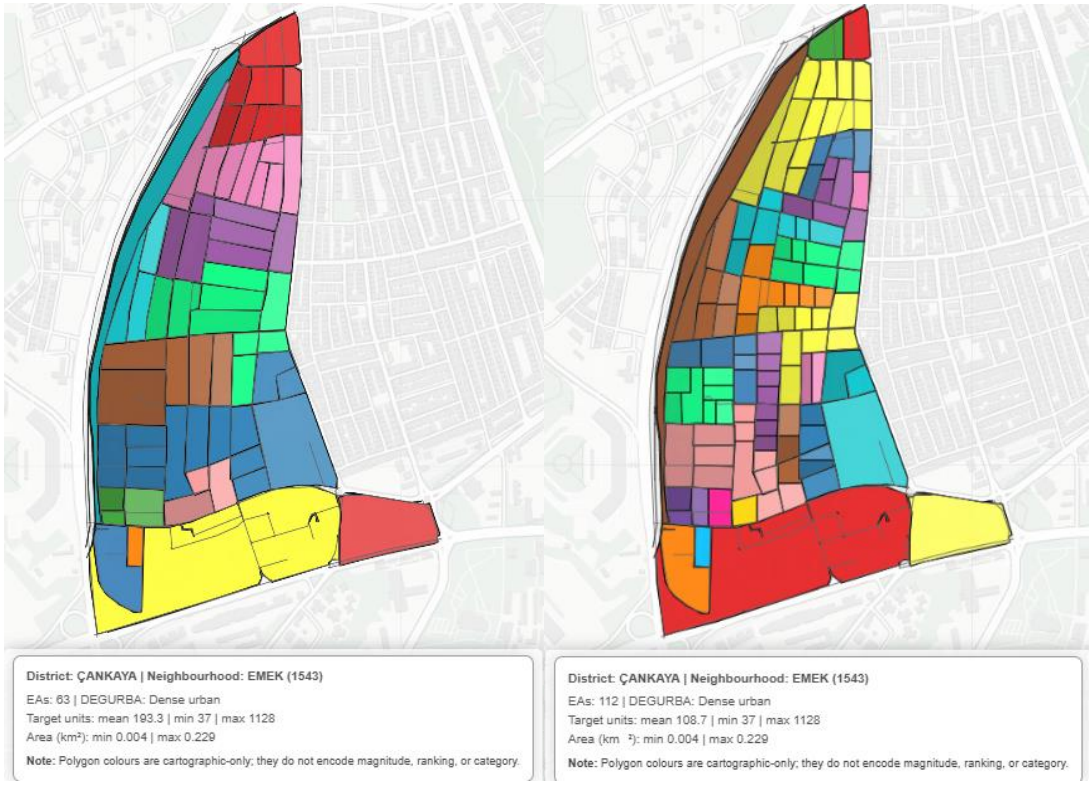


Figure 4.43. EA outputs for EMEK under Scenarios S3 and S5.

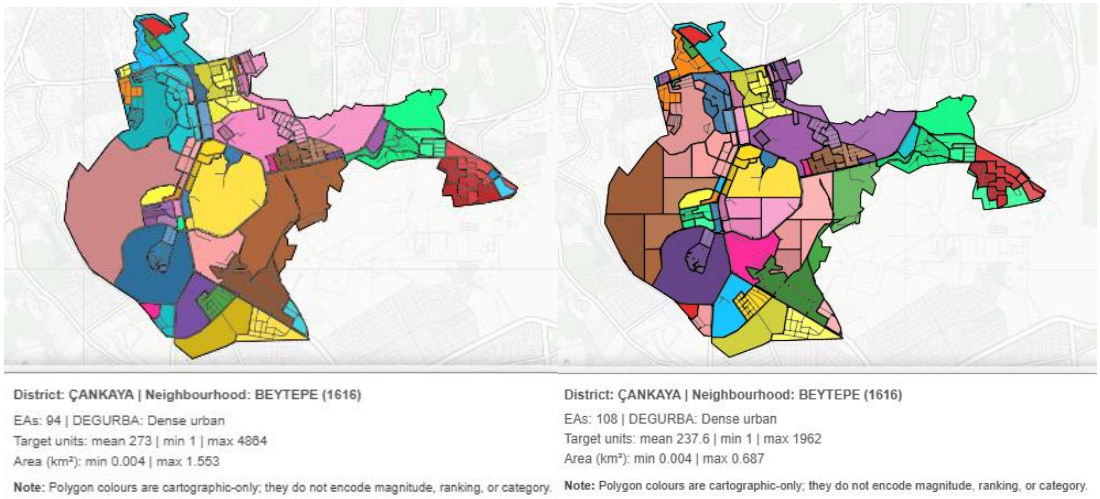


Target-based soft refinement produces the clearest workload-driven correction, but the effect is not uniform across indicators. In YUKARI DİKMEN (1610), the shift from S3 to S5 increases the EA count from 24 to 77 and reduces the maximum target load very sharply from 5,806 to 839, while also lowering the mean target load from 466.7 to 145.5. However, the number of EAs above TMAX still rises from 17 to 40 because the dense-urban target interval remains difficult to satisfy even after the large overload is broken down. In EMEK (1543), S5 increases the EA count from 63 to 112, reduces the number of EAs above TMAX from 44 to 26, and lowers the mean target load from 193.3 to 108.7, but increases the number below TMIN from 8 to 32 while leaving the maximum unchanged at 1,128. Taken together, these cases show that target-based soft split is a strong corrective mechanism, but that improvement in overload is often achieved at the cost of added fragmentation and additional underloaded units.

4.2.3.5. Soft split by area only

The comparison between S3 and S6 isolates the effect of area-based soft refinement. Unlike target-based refinement, area-only soft split is driven by geometric size rather than by direct workload concentration. This makes it useful for understanding whether some large units should be subdivided even where target imbalance alone does not force the split. The effect is illustrated by the BEYTEPE example in Figure 4.44.

Figure 4.44. EA outputs for BEYTEPE under Scenarios S3 and S6.



The area-only comparison confirms that geometric refinement alone is weaker than target-based correction. In BEYTEPE (1616), the shift from S3 to S6 increases the EA count from 94 to 108 and reduces the maximum target load from 4,864 to 1,962, while the mean target load declines from 273.0 to 237.6. At the same time, however, the number of EAs above TMAX increases from 33 to 47, and the number below TMIN remains essentially unchanged at 32 in both cases. This indicates that area-only soft split can alter geometry and reduce the size of the most extreme units, but it does not provide the clearest improvement in target balance when workload pressure is driven mainly by concentrated building stock.

4.2.3.6. Target-only versus target+area

The next comparison evaluates whether adding area to target-based refinement materially changes the result. The contrast between S5 and S7 is important because it shows whether the area rule provides an additional corrective effect after target-based refinement has already been applied. This contrast is illustrated by KIRKKONAKLAR in Figure 4.45 and ALACAATLI in Figure 4.46.

Figure 4.45. EA outputs for KIRKKONAKLAR under Scenarios S5 and S7.

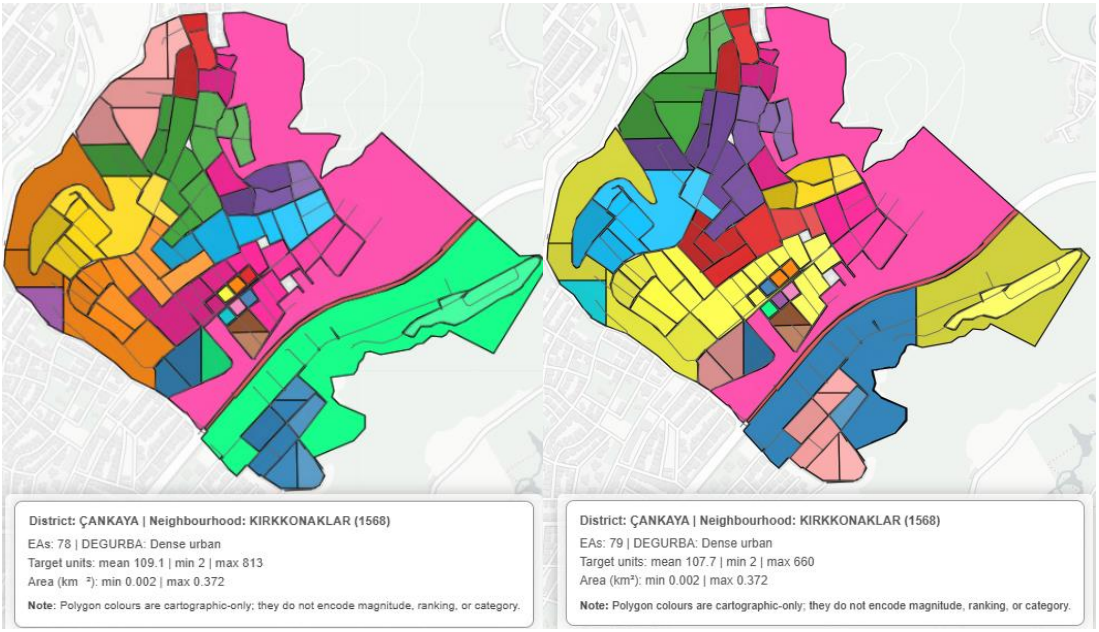
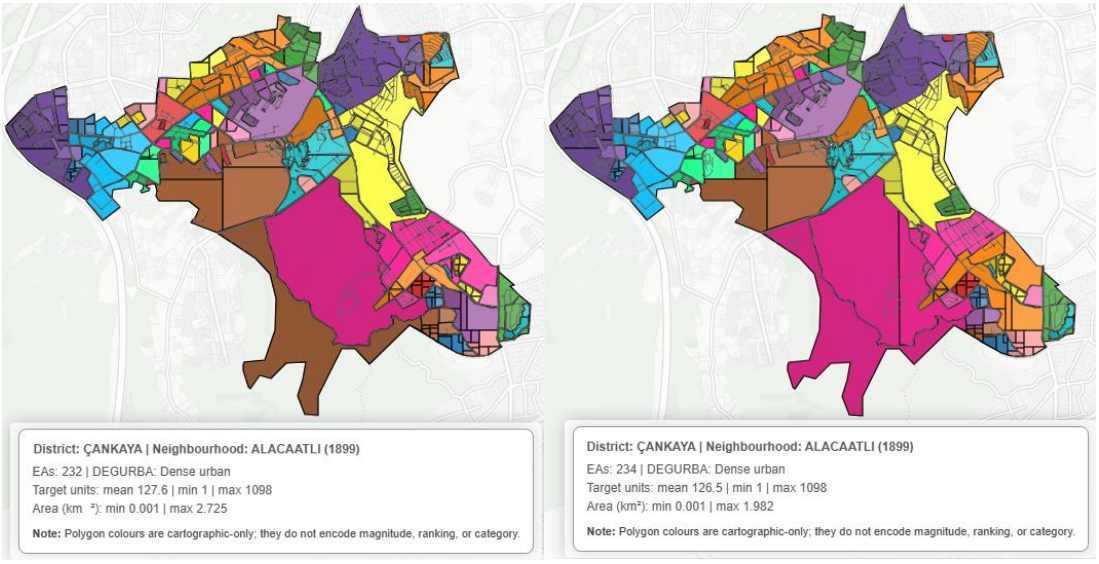


Figure 4.46. EA outputs for ALACAATLI under Scenarios S5 and S7.



The comparison between S5 and S7 shows that adding area to target-based refinement often changes little. In KIRKKONAKLAR (1568), the EA count changes only from 78 to 79, the number above TMAX changes from 20 to 21, and the mean target load shifts only slightly from 109.1 to 107.7, although the maximum decreases from 813 to 660. In ALACAATLI (1899), the difference is even smaller: S5 produces 232 EAs and S7 produces 234; both have zero zero-target EAs, the maximum remains

1,098 in both cases, and the mean changes only from 127.6 to 126.5. These examples support the interpretation that target+area should be treated as a refinement family rather than as a guaranteed improvement over target-only splitting in every neighbourhood.

4.2.3.7. Interaction between hard split and later refinement

The comparison between S7 and S8 evaluates the interaction between hard split and later soft refinement. This is analytically useful because it shows that early structural decisions can condition the form of the subsequent refinement process. Hard split may create or remove later opportunities for local adjustment, depending on how the geometry is reconfigured. This interaction is illustrated by İLKBAHAR in Figure 4.47 and YUKARI DİKMEN in Figure 4.48.

Figure 4.47. EA outputs for İLKBAHAR under Scenarios S7 and S8.

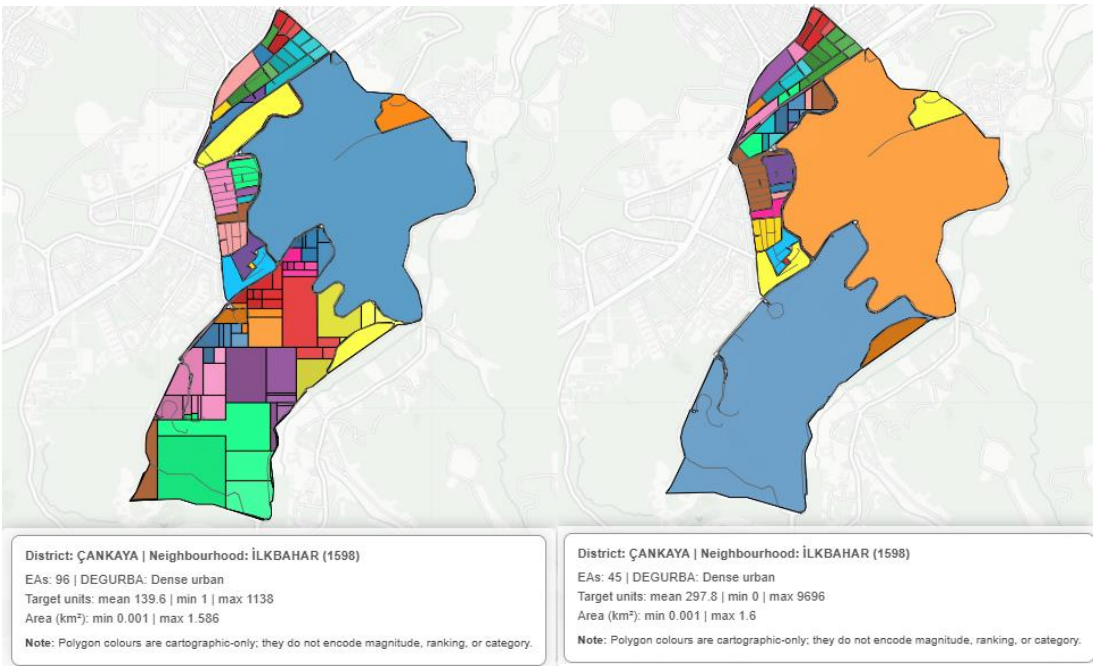
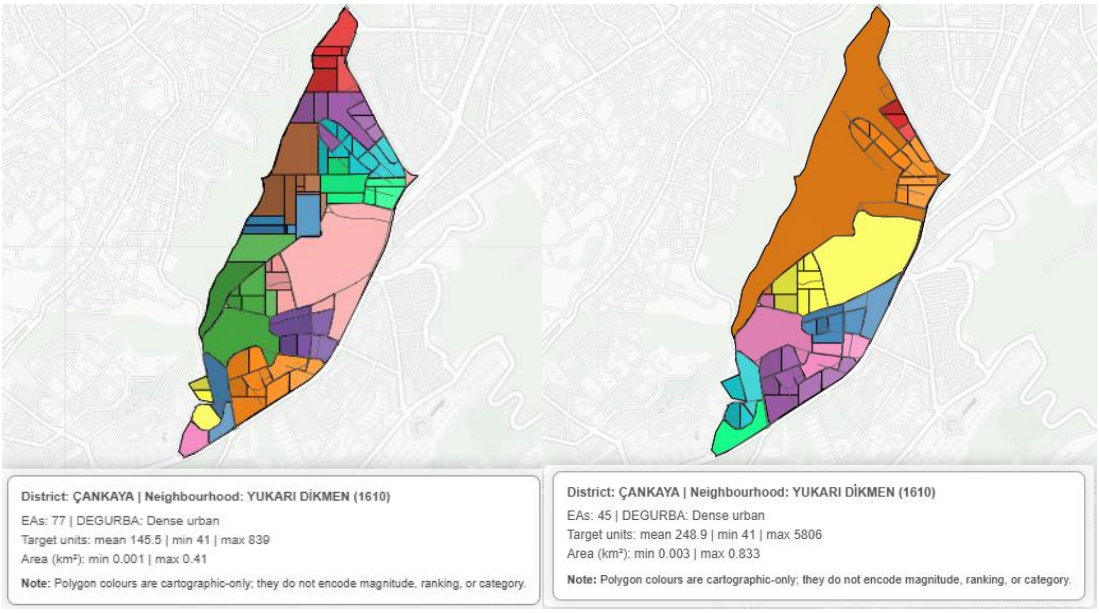


Figure 4.48. EA outputs for YUKARI DİKMEN under Scenarios S7 and S8.



The interaction between hard split and later refinement is clearly visible in both example neighbourhoods. In İLKBAHAR (1598), S7 produces 96 EAs with no zero-target EA, a maximum target load of 1,138, and a mean of 139.6, whereas S8 produces 45 EAs, reintroduces 1 zero-target EA, and raises the maximum back to 9,696 while also increasing the mean to 297.8. In YUKARI DİKMEN (1610), S7 produces 77 EAs with a maximum of 839 and a mean of 145.5, while S8 reduces the EA count to 45 but raises the maximum back to 5,806 and the mean to 248.9. These results show that hard split should not be narrated as universally beneficial; its interaction with later refinement can be strongly non-monotonic and can partly reverse earlier improvements.

4.2.3.8. Persistent unresolved hotspots

Several neighbourhoods remain problematic across multiple scenarios and should therefore be presented as persistent unresolved hotspots rather than as isolated failures of a single setting. İLKBAHAR (1598) retains a very high maximum target load in several configurations. BEYTEPE (1616) changes under many scenarios, but never reaches a fully satisfactory balance. ALACAATLI (1899) combines dense-urban pressure with limited resolution even under more aggressive refinement. These hotspot cases are analytically important because they shift the interpretation away from

toggle-level success or failure and toward the structural limits of the current data and partition logic.

Figure 4.49. Maximum EA target load across scenarios for the persistent hotspot neighbourhoods: İLKBAHAR, BEYTEPE, and ALACAATLI.

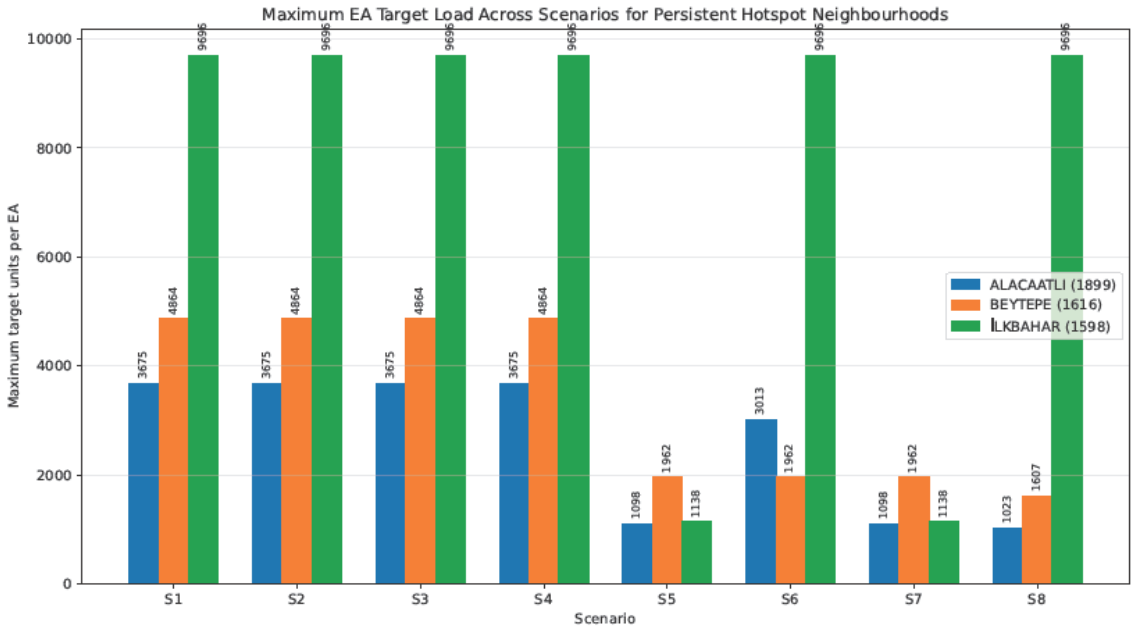


Figure 4.49 shows that the maximum EA target load remains structurally high in İLKBAHAR across most scenarios. This pattern is not only a scenario effect, but also a data-structure effect: one building carries an exceptionally large target count, and the current EA system does not permit subdivision of individual buildings. As a result, the maximum value has limited potential to decline even when other partition settings are changed. The numerical contrast is also clear. In İLKBAHAR (1598), the maximum target load remains 9,696 in S1, S2, S3, S4, S6, and S8, and only falls to 1,138 under S5 and S7. In BEYTEPE (1616), the maximum remains 4,864 under S1-S4, falls to 1,962 under S5-S7, and reaches its lowest value of 1,607 under S8. In ALACAATLI (1899), the maximum remains 3,675 under S1-S4, falls to 3,013 under S6, to 1,098 under S5 and S7, and to 1,023 under S8. These values show that extreme overload can be reduced substantially in some hotspot cases, but that the scale of improvement varies and is partly constrained by local structure.

Figure 4.50. Mean EA target load across scenarios for the persistent hotspot neighbourhoods: İLKBAHAR, BEYTEPE, and ALACAATLI.

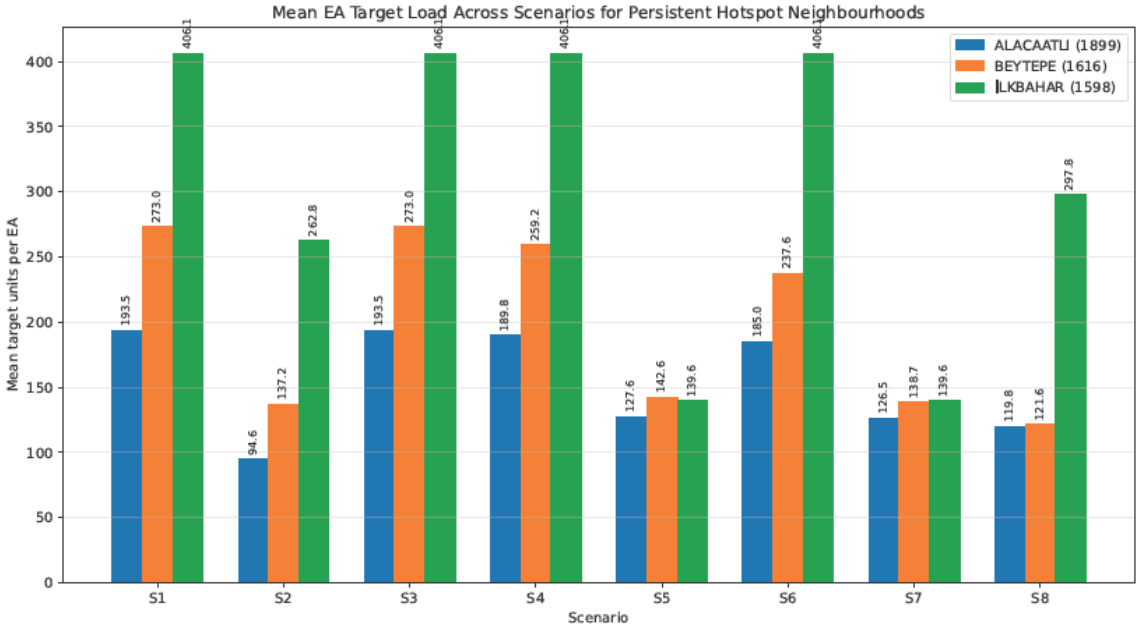


Figure 4.50 provides a more interpretable comparison of overall workload balance by showing the mean EA target load across scenarios for the three persistent hotspot neighbourhoods. Unlike the maximum metric, which in some cases is dominated by indivisible high-target buildings, the mean values reflect how far scenario changes improve the broader internal distribution of workload. In İLKBAHAR (1598), the mean remains very high at 406.1 in S1, S3, S4, and S6, falls sharply to 139.6 in S5 and S7, and rises again to 297.8 in S8. In BEYTEPE (1616), the mean declines from 273.0 in S1 and S3 to 259.2 in S4, then falls further to 142.6 in S5, 138.7 in S7, and reaches its lowest value of 121.6 in S8. In ALACAATLI (1899), the mean decreases from 193.5 in S1 and S3 to 189.8 in S4, 127.6 in S5, 126.5 in S7, and 119.8 in S8. The figure therefore shows that S5, S7, and, in several respects, S8 produce clearer improvements in average EA balance, even though none of the three hotspot neighbourhoods can be regarded as fully resolved across the scenario family.

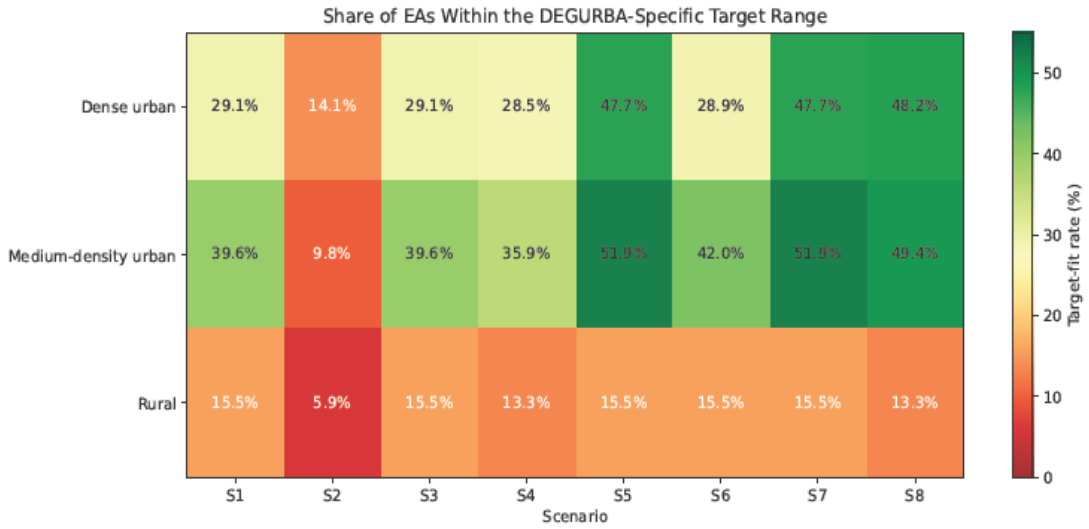
Taken together, these results suggest that the district-scale comparison is more appropriately interpreted through scenario families than through isolated toggle settings. Figures 4.49 and 4.50 likewise show that certain neighbourhoods are better understood as persistent unresolved hotspots than as isolated failures of specific parameter combinations. The İLKBAHAR case is especially instructive in this respect,

as its extreme maximum values appear to be driven partly by atomic building-level constraints that cannot be resolved within an EA system that does not subdivide individual buildings, whereas the mean values still show that broader workload balance can improve under selected refinement settings. The scenario comparison thus helps distinguish parameter-sensitive imbalance from structurally persistent workload concentration.

4.2.3.9. DEGURBA-wise target-fit comparison across all scenarios

The scenario-specific figures above show how individual settings affect selected neighbourhoods. To reintroduce the district-wide comparative perspective, all eight scenarios were also re-evaluated by DEGURBA class under a target-first interpretation. In this synthesis, the principal criterion is the share of EAs falling within the DEGURBA-specific target interval, while the area rule is treated as a secondary control rather than as the main indicator of success. The resulting class-based comparison is presented in Figure 4.51.

Figure 4.51. Share of EAs falling within the DEGURBA-specific target range across scenarios by DEGURBA class.



The heatmap shows that scenario performance differs by DEGURBA class when the target interval is treated as the primary evaluation criterion. Dense urban areas perform best under S8, medium-density urban areas under S5 and S7, while rural

areas remain weak across all settings. This supports retaining S4 as the analytical reference scenario while presenting S8 and S5/S7 as the strongest refinement outcomes under a target-first interpretation.

The DEGURBA-wise pattern is also numerically clear. In dense urban areas, the share of EAs within the target range is 29.1% in S1 and S3, 28.5% in S4, and rises to 47.7% in S5 and S7, reaching its highest value of 48.2% in S8. In medium-density urban areas, the corresponding rate is 39.6% in S1 and S3, 35.9% in S4, 42.0% in S6, 49.4% in S8, and reaches its highest value of 51.9% in both S5 and S7. In rural areas, by contrast, the target-fit rate remains weak throughout, varying only between 5.9% in S2 and 15.5% in S1, S3, S5, S6, and S7. These values confirm that no single scenario is uniformly strongest across all settlement types: S8 performs best in dense urban areas, S5 and S7 in medium-density urban areas, and no clearly satisfactory winner emerges in rural areas.

The DEGURBA-wise best-performing scenarios under the target-first interpretation are summarised in Table 4.13.

Table 4.13. Best-performing scenario by DEGURBA class under a target-first interpretation

DEGURBA class	Reference scenario (S4) target-fit rate	Best-performing scenario	Best target-fit rate	Interpretation
Dense urban	28.5%	S8	48.2%	Hard split combined with target+area refinement produces the strongest dense-urban improvement.
Medium-density urban	35.9%	S5 / S7	51.9%	Target-based refinement produces the strongest fit; adding area does not materially change the result.
Rural	13.3%	No clear winner	15.5%	All scenarios remain weak, indicating a structural limitation rather than a scenario-specific failure.

The target-fit synthesis therefore does not point to a single universally superior configuration. Instead, it shows that the stronger-performing settings vary by settlement type. Dense urban areas benefit most clearly from S8, medium-density urban areas from S5 and S7, and rural areas remain difficult under all tested configurations.

4.2.3.10. Preferred scenario under a target-first interpretation

Under a target-first interpretation, the findings do not support replacing the current reference scenario with a single new universal reference. S4 remains useful as the baseline configuration because it provides a transparent and methodologically clean point of comparison before aggressive refinement is introduced. At the same time, the comparative results show that refinement scenarios can improve target fit more clearly in specific settlement contexts. In dense urban areas, S8 is the strongest-performing configuration, whereas in medium-density urban areas the strongest results are obtained under S5 and S7. Rural areas remain weak under all settings and therefore should not be interpreted as evidence for a simple scenario substitution.

On this basis, the most defensible presentation is to combine district-level synthesis with a limited number of neighbourhood cases illustrating zero-target policy, hard split, target-based refinement, interaction effects, and persistent unresolved hotspots, followed by a DEGURBA-wise target-fit synthesis. The implications of these findings are discussed in the following chapter.

CHAPTER 5. DISCUSSION AND CONCLUSIONS

This chapter interprets the findings reported in Chapter 4 in relation to the conceptual and operational objectives of the thesis. The purpose is not merely to restate the outputs of the ArcGIS Pro and R applications, but to explain what those outputs mean for EA delineation under heterogeneous urban conditions, what kinds of methodological trade-offs they reveal, and how the developed workflow can be evaluated from the perspective of reproducibility, institutional usability, and future implementation in Türkiye.

The discussion is structured to remain closely tied to the empirical evidence. The ArcGIS Pro findings reported in Sections 4.1.1 and 4.1.2 are used primarily as a diagnostic baseline showing why several existing or intermediate approaches were insufficient for full institutional use. The R-based findings reported in Sections 4.2.1, 4.2.2, and 4.2.3 provide the main basis for evaluating the final workflow, because they demonstrate not only neighbourhood-scale feasibility and district-scale production under a reference scenario, but also DEGURBA-based illustrative interpretation and a target-first synthesis of the scenario comparison. Accordingly, the present chapter interprets the thesis contribution not as a single map output, but as a transparent and scalable delineation framework whose strengths and limits can be assessed through structured diagnostics.

This interpretive strategy is especially important because the thesis was not designed as a purely geometric zoning exercise. Rather, the study addressed EA production as a constrained planning problem situated at the intersection of statistical balance, spatial continuity, field practicality, and data realism. For that reason, the key questions in this chapter concern not only whether EAs were produced, but also under what conditions they were produced well, why certain deviations persisted, and which governance and data conditions would be required for wider operational use.

Taken together, the findings show that EA delineation in a large metropolitan setting cannot be reduced to a single optimisation problem with a universally valid answer. The pilot and district-scale applications demonstrate that the feasibility of an EA configuration depends on a combination of local morphology, barrier structure,

settlement density, and the granularity of the candidate units used during delineation. This conclusion emerges clearly from the contrast between the ArcGIS Pro trials in Sections 4.1.1–4.1.2 and the final R-based workflow in Sections 4.2.1–4.2.3.

The ArcGIS Pro strand was valuable because it clarified which classes of solution were not sufficient on their own. The BBZ-based trials reported in Sections 4.1.1.1 and 4.1.1.2 showed that outputs remained strongly dependent on input geometry. Point-based, fishnet-based, and Thiessen-based inputs changed the visual and areal character of the output, but did not by themselves solve the substantive requirements of EA production. Likewise, the ArcPy-based experiments in Sections 4.1.2.1–4.1.2.4 showed that building-polygon aggregation, point-based grouping followed by Thiessen conversion, and grid-based region growing could each illuminate part of the problem, yet each also exposed a different structural weakness. In the building-polygon trial, the output remained tied to building footprints and did not achieve full areal coverage. In the point-based Thiessen trial, the logic of grouping could be observed, but the workflow could not be completed because polygon creation was interrupted by licensing constraints. In the region-growing trial, full areal coverage was improved, but the resulting structure still risked being dominated by very large residual units.

These early findings matter analytically because they show that the problem is not simply one of drawing boundaries around a target number. If the atomic units are poorly aligned with building structure, if the barrier network does not provide meaningful internal separation, or if the algorithm grows mainly to preserve contiguity rather than to control workload, then the resulting geography may become visually complete while remaining operationally weak. In this sense, the ArcGIS Pro applications functioned as an explicit calibration stage. They demonstrated why the final workflow needed to privilege building-based atomic units, barrier-aware subdivision, and diagnostic transparency over a purely generic balancing logic.

The R-based findings provide the complementary evidence required to support the central thesis claim. At single-neighbourhood scale, Section 4.2.1 showed that the workflow could produce coherent EA outputs while keeping the logic of growth, exception handling, and review visible to the analyst. At district scale, Sections 4.2.2.1–4.2.2.6 showed that the same core logic could be executed as a structured batch

production system capable of generating neighbourhood-level files, district-level statistics, hierarchical EA–SA–CA identifiers, DEGURBA-based reference-scenario illustrations, and review-oriented map products. The scenario comparisons in Sections 4.2.3.1–4.2.3.10 then demonstrated that the remaining deviations were not random. Rather, they reflected recurrent and interpretable tensions between target compliance, spatial coherence, and the operational meaning of barriers and building integrity, while also showing that scenario performance varies by settlement type when interpreted under a target-first lens.

What the findings therefore reveal is not that metropolitan EA delineation can be made perfectly uniform, but that it can be automated to a high degree when automation is understood as supervised, constrained, and diagnostically explicit. This is a different claim from a universal optimisation claim. The final workflow does not remove complexity; it organises complexity into auditable outputs. That distinction is crucial for institutional adoption, because it means that the value of the workflow lies not only in its successful cases, but also in its ability to identify and explain the difficult cases in a systematic way.

5.1. Interpreting deviations and scenario differences as methodological trade-offs

A central result of the study is that deviations from the preferred target band are systematic and interpretable rather than arbitrary. This is a methodological strength, not a weakness. In institutional practice, the most problematic zoning systems are often not those that contain deviations, but those in which deviations cannot be explained, traced, or reviewed under explicit rules. The scenario-based district comparison reported in Section 4.2.3 shows that different parameter sets do not simply generate outputs that can be ranked from best to worst along a single scale. Instead, they expose recurring trade-offs between numerical balance, areal integrity, continuity, and field plausibility.

This interpretation is also consistent with the design of the application itself. The workflow was deliberately structured so that DEGURBA-specific target ranges and area limits could be adjusted, and so that several correction options could be

enabled or disabled by the user, including soft split, hard split, zero-target handling, and small-EA merging. These options were not included merely as technical conveniences. They were included because the study proceeded from the assumption that no single static rule-set would remain equally appropriate in all settlement contexts. The findings confirm this assumption.

The scenario differences therefore need to be read as methodological trade-offs that help define the operational envelope of the workflow. The most important of these trade-offs concern the zero-target policy, hard split, soft split, the interaction between split mechanisms and persistent hotspots, and the way these effects change across DEGURBA classes when the target interval is treated as the primary evaluation criterion.

One of the clearest results is that the practical meaning of the zero-target policy depends strongly on whether hard split is active. As shown in Section 4.2.3.2, the allow-zero and no-zero alternatives become nearly indistinguishable in the hard-split-off comparison, which indicates that allowing zero-target EAs has limited practical effect when the workflow is not already generating additional structural fragmentation. Once hard split is enabled, however, the difference becomes clearer. In neighbourhoods such as ÜNİVERSİTELER (1592) and ALACAATLI (1899), allowing zero-target units can increase the number of resulting EAs and generate empty or operationally weak fragments without necessarily resolving the main overload problem.

At the same time, the discussion should not be reduced to a simple rule that zero-target areas are always undesirable. The earlier development process also showed why the option was made user-controllable in the first place. In dispersed or morphologically irregular areas, especially where barriers, steep terrain, rail corridors, or sparsely settled margins interrupt the built fabric, forcing all zero-target or very low-target areas to merge can create very large and operationally implausible units. In such contexts, a strict no-zero rule may preserve numerical neatness while weakening the spatial logic of the resulting geography. The discussion therefore supports the current interpretation of the allow-zero option as a diagnostic or conditional instrument rather than as the main production default.

This is a useful example of why EA delineation should be evaluated through both numerical and geographic reasoning. A policy that looks desirable from the perspective of workload counts may produce unreasonable territorial outcomes when applied without regard to the actual structure of settlement and barriers. The main production configuration is therefore better aligned with a no-zero logic, but the availability of the alternative remains methodologically justified because it helps reveal where the geometry itself resists strict balancing.

The corrected district-scale implementation confirms that hard split is important, but not universally effective. As discussed in Section 4.2.3.3, enabling hard split can reduce overload in some neighbourhoods, yet the improvement is often partial and may also create additional underloaded units. KIRKKONAKLAR (1568) is an instructive example, because overload is alleviated to some degree but not removed entirely. By contrast, neighbourhoods such as İLKBAHAR (1598) show that enabling hard split does not automatically lead to meaningful change.

These patterns suggest that the effectiveness of hard split is fundamentally conditioned by the internal face structure produced during delineation. The workflow can only split a unit when a valid internal partition exists that is compatible with barrier logic, contiguity, and building integrity. If the available barrier-derived subdivision does not provide an operationally meaningful internal cut, then hard split has little space in which to act. In other words, hard split is not a general optimisation switch. It is a conditional rule whose effectiveness depends on whether the underlying geometry offers an acceptable partition opportunity.

The discussion also needs to emphasise that the limits of hard split are not purely algorithmic. Some EAs remain above the preferred threshold because a single building or a compact building cluster already contains a very high residential-unit count. Since the workflow was intentionally designed to preserve building integrity, these cases cannot be treated as ordinary split failures. In areas dominated by high-rise or high-density residential structures, a threshold exceedance may reflect the real morphology of the built environment rather than an avoidable delineation error. This is one reason why the discussion supports retaining hard split in the workflow while interpreting its contribution as selective and context-dependent rather than universal.

The scenario comparison further shows that soft split by target and soft split by area should not be treated as interchangeable corrective mechanisms. As reported in Sections 4.2.3.4–4.2.3.6, target-based soft split is the stronger refinement option in cases where overload is driven mainly by concentrated residential-unit density. Neighbourhoods such as YUKARI DİKMEN (1610) and EMEK (1543) illustrate that target-based refinement can reduce extreme overload more effectively than area-only correction. However, this improvement often comes at the price of additional fragmentation or a larger number of units falling below the preferred range. The later target-first synthesis in Section 4.2.3.9 reinforces this point by showing that the strongest-performing refinement settings differ by DEGURBA class rather than collapsing into a single universally superior scenario.

Area-based soft split, by contrast, is more limited when the main pressure is created by concentrated building stock rather than by large areal extent. The findings show that it may still be useful in specific contexts, particularly where unusually large polygons persist despite moderate target values, but it is not an efficient primary response to concentrated workload peaks. This is why the combined target+area scenarios often behave much more like target-based soft split than like a balanced hybrid in which both components contribute equally, and why the dense-urban and medium-density results in Section 4.2.3.9 favour different refinement families under a target-first reading.

This interpretation is also consistent with the broader development history of the workflow. The interface was designed so that users could enable or disable soft split and choose whether refinement should respond to target limits, area limits, or both. The discussion confirms that this flexibility is justified, because the relative value of these options changes with urban form. Yet the findings also indicate that target-based soft split deserves the main analytical attention, whereas area-based correction is better understood as an auxiliary mechanism whose contribution is situation-specific rather than dominant. In this respect, the final results support retaining S4 as a transparent analytical baseline while interpreting S5, S7, and S8 as context-dependent refinement outcomes rather than as simple replacements for the reference case.

One of the most important findings of the scenario comparison is that the interaction between hard split and later refinement is not monotonic. As shown in

Section 4.2.3.7, enabling more corrective options does not automatically yield a better result. In some neighbourhoods, a soft-split-only scenario may produce a more acceptable workload pattern than a formally more aggressive configuration that also includes hard split. The workflow should therefore be understood less as a one-directional optimisation routine and more as a sequence of interacting guardrails whose effects depend on local partition structure, the order of operations, and the specific morphology of the neighbourhood.

The persistent hotspot neighbourhoods identified in Section 4.2.3.8 are especially important in this respect. İLKBAHAR (1598), BEYTEPE (1616), and ALACAATLI (1899) remain difficult under multiple scenario settings even though the exact number of units and the workload distribution vary across runs. These cases are analytically valuable because they show where the present workflow reaches its boundary conditions. The paired maximum and mean comparisons in Figures 4.49 and 4.50 are especially instructive here, because they show that some extreme values remain structurally persistent even when average workload balance improves under selected refinement settings. Not every overload can be eliminated by changing split switches or target thresholds alone.

Several substantive factors appear to contribute to this persistence. First, the available barrier network may not provide enough internal structure to support meaningful subdivision in very concentrated areas. Second, the use of all residential units rather than occupied residential units may intensify local peaks in ways that are operationally important. Third, the current face structure may fail to provide a divider that satisfies both workload logic and geometric coherence. Finally, some cases are driven by single buildings whose internal residential-unit count already exceeds the preferred threshold. Persistent hotspots should therefore be interpreted as indicators of structural difficulty rather than as simple evidence of algorithmic failure. The later DEGURBA-wise target-fit synthesis in Section 4.2.3.9 supports the same conclusion from a different angle by showing that no single scenario resolves all settlement types equally well.

5.2. Reproducibility, transparency, and the role of the implementation environment

The findings emphasise that reproducibility is not only a conceptual principle but also a concrete implementation property. One of the practical lessons of the thesis is that a delineation framework intended for institutional use must be re-runnable, inspectable, and auditable under stable rules. This point became clear during the ArcGIS Pro applications, where several otherwise informative workflows were interrupted by license-dependent tools such as CreateThiessenPolygons, FeatureToPolygon, PolygonToLine, Buffer, or Erase (Section 4.1.2.4). These interruptions were not merely technical inconveniences. They represented a direct threat to auditability, updateability, and inter-regional replication.

The R implementation addresses this problem more directly. As shown in Sections 4.2.2.1–4.2.2.6, the workflow produces not only spatial outputs but also a structured file organisation, neighbourhood-level and district-level summaries, hierarchical EA–SA–CA identifiers, DEGURBA-based illustrative cases, and map products in both interactive and static formats. This output environment matters analytically because it allows the reviewer to move from district-level diagnostics to individual neighbourhood files and back again without losing the logic of the production process. Quality assessment is therefore not dependent on visual inspection alone; it is supported by standardised summaries and diagnostic tables.

The role of the Shiny-based interface should be interpreted in the same light. The interface does not define a separate delineation logic; it provides a local graphical front-end for parameter management and operational convenience while calling the same core workflow. This means that options such as DEGURBA-based target ranges, area limits, hard split, soft split, zero-target handling, and small-EA merge thresholds can be managed in a more accessible way without sacrificing script-level reproducibility. For the discussion, the important point is not the interface as software design in itself, but the fact that the implementation environment supports supervised automation: a repeatable baseline geography is produced under explicit rules, and expert judgement is reserved for a transparent set of flagged cases. The distinction

between a stable baseline scenario and context-dependent refinement scenarios is therefore an implementation strength rather than a methodological inconsistency.

This combination of scriptability, diagnostic visibility, and interface-supported parameter control is one of the main reasons the final workflow is institutionally more promising than the exploratory ArcGIS Pro strand. In a national statistical context, repeatability and traceability are as important as local geometric quality. A system that can explain why an EA is difficult, record which option set generated it, and produce consistent review outputs is substantially more valuable than a system that offers visually plausible units without a clear audit trail.

5.3. Evaluation of aims and working hypotheses

The findings support the thesis aims in a qualified but meaningful way. First, the study aimed to define operational and statistical criteria for EA design consistent with survey and census practice. This objective was met by translating the design problem into explicit constraints related to barriers, contiguity, building integrity, target ranges, and DEGURBA-sensitive interpretation. The results show that these criteria can be implemented computationally and evaluated diagnostically rather than left at the level of purely descriptive guidance.

Second, the study aimed to implement a reproducible GIS-based workflow capable of generating contiguous and non-overlapping EAs within administrative limits. The evidence in Sections 4.2.1 and 4.2.2 shows that this objective was largely achieved. The final workflow can produce neighbourhood-level and district-level outputs repeatedly under the same rule-set, while preserving neighbourhood boundaries and providing a complete production structure rather than a one-off visual output.

Third, the study aimed to evaluate how barrier combinations and parameter settings affect EA geometry and workload balance. This aim was strongly supported by the scenario-based findings in Section 4.2.3. The comparison between hard split, soft split, allow-zero, and no-zero settings demonstrated that the behaviour of the workflow changes in interpretable ways across neighbourhood contexts. Importantly, the findings show that trade-offs between spatial coherence and target compliance are

not accidental side effects; they are central properties of the delineation problem itself. They also show that no single scenario is uniformly strongest across all DEGURBA classes, which further supports the thesis emphasis on supervised, context-sensitive automation rather than a single universal rule-set.

Fourth, the study aimed to assess whether the approach could be extended from a pilot neighbourhood to district-scale implementation. The district batch application under the reference scenario, together with the structured outputs reported in Sections 4.2.2.1–4.2.2.6, provides strong evidence in favour of this claim. The workflow does not merely run once outside the pilot setting; it scales to a production logic in which district overviews, DEGURBA-based interpretive cases, local review files, summary statistics, and hierarchical coding are generated in a consistent manner.

The working hypotheses are therefore supported in a qualified form. Barrier-aware delineation improves operational plausibility because it avoids cross-barrier aggregation that would be difficult to implement in the field. Settlement-class differentiation is necessary because feasible EA size and workload structure differ substantially across DEGURBA contexts. Finally, reproducible automation is feasible at scale, but not complete in the sense of eliminating all exceptional cases. The findings support a supervised automation model in which explicit rules generate most of the geography, while a minority of difficult outputs are handled through review-oriented diagnostics and controlled intervention. In this sense, the thesis supports not a single universally optimal scenario, but an auditable framework in which a stable baseline can be combined with context-sensitive refinement.

5.4. Limitations and boundary conditions for interpretation

The limitations observed across the pilot and district-scale applications help define the conditions under which the methodology can be interpreted and improved. Making these limits explicit is important because the thesis does not claim that every deviation can be eliminated under the current data environment. Rather, it claims that the workflow makes those deviations visible and manageable under transparent rules.

A first limitation concerns the target variable itself. The workflow operates on residential units rather than occupied residential units because occupied-unit information was not available for direct use in the study. From a statistical and fieldwork perspective, however, actual workload is more closely related to occupied residential units than to total recorded residential-unit stock. This means that some overload or underload patterns may reflect the difference between structural housing stock and effective field workload. The use of all residential units was a defensible practical choice under data constraints, but it is also one of the clearest limits of the present implementation.

A second limitation concerns the quality and continuity of the publicly available barrier data, especially the OpenStreetMap-derived road layers used during delineation. Because roads function as one of the most important barrier inputs, omissions, discontinuities, or locally inconsistent representation can directly affect the quality of the resulting atomic units and the ability of the workflow to identify meaningful internal splits. When roads are under-mapped, large EAs may remain non-splittable even though the target count exceeds the preferred range. Conversely, fragmented or over-detailed linework may generate excessive subdivision. The findings therefore need to be interpreted in light of the barrier data that were actually available to the study rather than as if an authoritative transport layer had been used.

A third limitation concerns the treatment of highly concentrated structures and the preservation of building integrity. The workflow deliberately avoids splitting buildings, since building-level integrity is important both geometrically and operationally. However, this rule means that some EAs remain above the preferred threshold simply because a single building, especially a high-rise or very dense residential structure, already exceeds the target range. In such cases, the output may look suboptimal when judged against the target band alone, even though it is consistent with the logic of the built environment and with the methodological decision not to divide buildings artificially.

A fourth limitation concerns the alternative subdivision strategies that were explored in order to address such difficult cases. One tested approach involved subdividing problematic areas into 25 m^2 grid cells, matching buildings to those cells, and then re-growing units from the grid structure. While this strategy created

additional internal partition opportunities, it weakened areal integrity and could produce spatially unsatisfactory configurations after re-aggregation. An additional attempt using 10 m² cells offered greater granularity, but the computational burden became unreasonable relative to the expected gain. Another tested approach relied on forced splitting of overloaded areas by dividing them into two or more parts based on target or area criteria alone. This approach also proved unsatisfactory, because it could generate straight-line divisions that ignored local morphology, passed through buildings, or cut across building groups that should have remained together. These results are important because they show that not every technically available subdivision method is methodologically acceptable for EA production.

A fifth limitation concerns the absence of parcel data from the delineation logic. Parcel boundaries could, in principle, provide a more structured auxiliary layer in cases where barrier information alone is insufficient for meaningful subdivision or where grid-based re-aggregation weakens areal integrity. However, parcel data were not incorporated into the current workflow. The discussion therefore cannot assume that all subdivision problems would disappear with parcel-based refinement; it can only note that this remains an important avenue for future improvement, subject to data quality, completeness, and currency.

Finally, there are operational boundary conditions that extend beyond geometry and data. EA delineation is intended for use within field organisation, staffing practice, confidentiality requirements, and repeat-survey continuity. For that reason, target thresholds cannot be interpreted as absolute constraints in every local context. Institutional use would require governance rules defining when exceptional cases are accepted as structurally legitimate, when additional review is necessary, and how updates are documented over time. In this sense, a methodological boundary is also an organisational boundary: automation can produce a transparent baseline, but accountable institutional use still requires explicit review protocols.

5.5. Implications for scaling to Ankara and institutional deployment

The movement from a single-neighbourhood application to district batch processing has implications that go beyond the mere increase in geographic coverage. The findings show that scaling should be understood as a methodological and institutional transition. In this study, the workflow was organised so that it could be executed not only for the 124 neighbourhoods of Çankaya, but, if desired, for other selected districts, selected neighbourhoods, or, in principle, the whole of Ankara under the same production architecture. This scalability matters because it suggests that the workflow is not limited to a bespoke pilot exercise; it can function as a modular production system.

At the same time, the district findings make it clear that metropolitan deployment cannot rely on a single rigid rule-set. As spatial heterogeneity increases, the workflow must rely more on context-sensitive parameterisation and explicit exception handling. The district summaries reported in Sections 4.2.2.2–4.2.2.6 and the scenario evidence in Sections 4.2.3.2–4.2.3.10 support this conclusion by showing that dense urban, medium-density, and more dispersed contexts differ not only in the number of resulting EAs, but also in the kinds of deviations that remain acceptable or unavoidable.

This implies that scaling to Ankara should not be interpreted as the uniform mechanical repetition of one target band. Rather, it implies a controlled production logic in which thresholds remain nationally interpretable but are applied through explicit DEGURBA-sensitive reasoning. In dense urban cores, somewhat narrower EAs may remain operationally manageable because walking distances are short and the built environment is compact. In sparse or peripheral contexts, somewhat larger EAs may remain reasonable if the settlement structure and movement costs make strict target compliance unrealistic. Such variation does not weaken standardisation; it clarifies the circumstances under which standardisation remains meaningful.

A second implication concerns the handling of exceptional units. Large or persistent hotspot EAs should not automatically be treated as erroneous outputs at district or metropolitan scale. The findings indicate that some of these units are

structurally legitimate under the existing building-integrity and barrier constraints. A staged review model is therefore preferable: first, run the barrier-aware framework under transparent rules; second, identify and flag exceptional units through diagnostics; third, apply targeted review only to the flagged cases, using additional evidence or local refinement when operationally justified.

A third implication concerns the institutional environment in which such outputs would be used. The structured outputs reported in Sections 4.2.2.1–4.2.2.6 already resemble the kind of production and review materials needed for institutional deployment: district overviews, neighbourhood files, summary tables, hierarchical identifiers, DEGURBA-based interpretive examples, and visual review outputs. These products support the idea that EA production should not be treated as an isolated mapping task, but as part of a wider workflow linking spatial production, quality review, and field planning.

From the perspective of official statistics, this is especially relevant for the Turkish Statistical Institute (TurkStat). The value of the workflow lies not only in the creation of polygons, but also in its potential to provide a maintainable and operationally meaningful backbone for census operations and sample surveys. Used in this way, EAs would not replace administrative coding systems; they would complement them by adding a spatially coherent operational layer that can support listing, workload planning, supervised field assignment, and the consistent organisation of repeated statistical activities.

5.6. Synthesis, conclusions, and recommendations

Overall, the findings confirm that an EA framework can be automated to a high degree when the delineation process is explicitly constrained by barriers, contiguity, building integrity, and neighbourhood-boundary preservation, and when target ranges are treated as context-sensitive objectives rather than as rigid requirements. The study also shows that the most informative outputs are not only the final polygons themselves, but also the diagnostics that reveal where spatial structure, data limitations, or operational constraints prevent strict balance. The final target-first

synthesis further indicates that stronger-performing settings vary by settlement type rather than converging into a single universally preferable scenario.

The principal contribution of the thesis therefore lies not only in producing a set of EA outputs for a pilot area, but in developing and testing a transparent, scalable, and diagnostically explicit EA production workflow suitable for pilot implementation and institutional adaptation in Türkiye. The workflow is scriptable, reproducible, and capable of generating structured district and neighbourhood outputs in a form that supports both technical review and operational interpretation. In this respect, the study demonstrates that EA delineation in official statistics should not be understood merely as a technical GIS exercise. It should be understood as a planning and design framework capable of supporting time- and cost-efficient field operations and of contributing to more spatially coherent, operationally consistent, and statistically representative production. Equally importantly, the study shows that such a framework can incorporate both a stable reference configuration and context-dependent refinement scenarios without losing transparency or auditability.

On this basis, the principal recommendation of the thesis is that an EA system should be established and gradually integrated into official statistical practice in Türkiye. More specifically, the study recommends that such a system should be adopted and used by TurkStat in census operations and sample surveys. The findings indicate that the use of EAs in these activities would strengthen field organisation, improve operational efficiency, and support more time- and cost-effective implementation in the field. At the same time, because EAs provide more consistent and better-structured operational units, their use would also strengthen representativeness and contribute to the production of higher-quality statistics. In this sense, EA systems should be regarded as a strategic institutional investment with the potential to improve the effectiveness, transparency, and quality of national statistical operations. Operationally, this does not imply that a single scenario should be frozen as a universal production rule. Rather, it suggests a governance model in which a transparent baseline configuration is maintained and context-sensitive refinement is applied where diagnostics show that it is justified.

Beyond this principal recommendation, the thesis also suggests several improvements that could increase the quality and operational performance of future

implementations. First, where feasible, EA design should rely more directly on occupied residential units rather than on all residential units, since actual field workload is more closely related to occupied residential units than to total recorded residential-unit stock. This is one of the most important practical improvements indicated by the study, because it would align balancing logic more closely with field reality.

Second, road data should, where possible, be obtained from service providers or other institutional infrastructure holders capable of supplying more complete and operationally relevant transport layers than the publicly available data used in the present implementation. Stronger road data would improve barrier-sensitive delineation, reduce the risk of false continuity, and increase the likelihood of identifying meaningful internal subdivisions in difficult neighbourhoods.

Third, parcel data may be considered as an auxiliary spatial layer in future implementations. If parcel data obtained through the General Directorate of Land Registry and Cadastre (Tapu ve Kadastro Genel Müdürlüğü, TKGM) are sufficiently complete, current, and suitable for this purpose, they may help address cases in which barrier information alone is insufficient to subdivide space appropriately, or in which grid-based subdivision followed by re-aggregation weakens areal integrity. In such cases, parcel boundaries could support cleaner subdivision and more coherent final units.

Fourth, future implementations should include a more explicit analytical treatment of single-building exceedance cases. Where a single building already exceeds the preferred residential-unit threshold, forced subdivision should not be treated as the default response. Such cases should first be identified and reviewed separately in order to distinguish genuine structural exceptions from problems arising elsewhere in the delineation logic. On that basis, an explicit exception-handling rule or review protocol could be developed for high-rise or very dense structures, allowing the system to preserve building integrity while still supporting informed operational decision-making.

Finally, the effective and sustainable implementation of EA systems would require stronger inter-institutional cooperation. This need applies not only to parcel data, but more broadly to all institutions holding relevant administrative records and

spatial infrastructure data, including address data, building records, road and barrier layers, update procedures, and field implementation rules. Such cooperation is necessary not only for successful EA implementation, but also for the broader development and advancement of the national statistical system in Türkiye. For this reason, institutional coordination should be understood as a core enabling condition rather than as a secondary administrative issue.

Taken together, the discussion supports a staged path forward. The first stage should focus on establishing the core EA framework under transparent and reproducible rules, including a stable reference configuration that can function as an auditable production baseline. The second stage should involve structured review of flagged exceptional cases and gradual improvement of the supporting data environment, especially with respect to occupied residential units, stronger road layers, and, where appropriate, parcel data. The third stage should focus on broader institutional integration so that EAs can function not as a one-time pilot product, but as a stable and operationally meaningful component of census and sample-survey production in Türkiye. Under such a model, the central value of the framework lies not in promising a perfectly balanced geography everywhere, but in providing an auditable, scalable, and institutionally usable basis for managing spatial statistical operations under real-world constraints, while allowing context-sensitive refinement where the evidence supports it.

REFERENCES

Amusa, I. A., Adelakun, G. I., Akinbami, D. S., Adebowale, S. B. N., Ogundele, O. M., & Oyelakin, L. O. (2017). Census Enumeration Information System – A case study of part of Ikeja GRA residential layout, Ikeja of Lagos State. *World Scientific News, 81*(2), 132–149.

Arantes, S. B., & Silva, P. L. N. (2013). Designing household samples in Brazil using the 2010 census enumeration area frame. In *Proceedings of the 59th ISI World Statistics Congress* (Session CPS016), Hong Kong.

Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science, 20*(7), 797–811. doi:10.1080/13658810600665111

Australian Bureau of Statistics. (2021, July 20). Statistical Area Level 1. In Australian Statistical Geography Standard (ASGS) Edition 3: Main Structure and Greater Capital City Statistical Areas, July 2021-June 2026. Retrieved from https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/statistical-area-level-1

Balistreri, A., & Cozzi, S. (2015, September). Geocoding process of the 9th Industry and Services Census data: Sources and methods used. Paper presented at the *Meeting of the Group of Experts on Business Registers*, Brussels.

Borugian, M. J., Spinelli, J. J., Mezei, G., Wilkins, R., Abanto, Z., & McBride, M. L. (2005). Childhood leukemia and socioeconomic status in Canada. *Epidemiology, 16*(4), 526–531.

Central Bureau of Statistics Nepal. (2022). *Official Statistics of Nepal: Issues and Practices*. Government of Nepal, National Planning Commission.

Cochran, W. G. (1977). *Sampling techniques (3rd ed.)*. New York, NY: John Wiley & Sons.

Cockings, S., Martin, D., & Harfoot, A. (2011). Maintaining existing zoning systems using automated zone-design techniques: Methods for creating output geographies. *Environment and Planning A, 43*(10), 2399–2418. doi:10.1068/a43601

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Duque, J. C., Anselin, L., & Rey, S. J. (2012). The max-p-regions problem. *Journal of Regional Science, 52*(3), 397–419. doi:10.1111/j.1467-9787.2011.00743.x

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review, 30*(3), 195–220. doi:10.1177/0160017607301605

Esri. (2020). The role of GIS in official statistics. Esri Press.

Esri. (2024). *ArcGIS Pro documentation* (ArcPy; Create Fishnet; Summarize Within; Build Balanced Zones). Retrieved from https://pro.arcgis.com/

Eurostat. (n.d.). *Population and housing censuses*. Retrieved from https://ec.europa.eu/eurostat/web/population-demography/population-housing-censuses

Eurostat. (2017). *Statistical areas and population grids in Europe*. Publications Office of the European Union.

Eurostat. (2019). Methodological guidelines and description of EU-SILC target variables: 2018 operation (Version July 2019). Eurostat.

Eurostat. (2021). Applying the degree of urbanisation – A methodological manual to define cities, towns and rural areas for international comparisons (2021 ed.). Luxembourg: Publications Office of the European Union.

Eurostat. (2024). *Territorial typologies manual – Degree of urbanisation*. Eurostat Statistics Explained. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/

Eze, B. E. (2009). The role of remote sensing and GIS in census mapping. *Nigeria Journal of Surveying and Geoinformatics, 3*(1), 15–28.

Flowerdew, R. (2011). How serious is the Modifiable Areal Unit Problem for analysis of English census data? *Population Trends, 145*, 106-118. doi:10.1057/pt.2011.20

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025–1044. doi:10.1068/a231025

Ghana Statistical Service. (2021). Ghana 2021 Population and Housing Census: Post-Enumeration Survey (PES) field officer's manual. Accra, Ghana: GSS.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211-221. doi:10.1007/s10708-007-9111-y

Goodchild, M. F. (2011). Scale in GIS: An overview. *Geomorphology, 130*(1-2), 5-9. doi:10.1016/j.geomorph.2010.10.004

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology (2nd ed.)*. Hoboken, NJ: Wiley.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science, 22*(7), 801–823. doi:10.1080/13658810701674970

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design, 37*(4), 682-703. doi:10.1068/b35097

Haynes, R., Daras, K., Reading, R., & Jones, A. (2007). Modifiable neighbourhood units, zone design and residents' perceptions. *Health & Place, 13*(4), 812–825.

Hijmans, R. J. (2023). *terra: Spatial data analysis* [Computer software]. Retrieved from https://CRAN.R-project.org/package=terra

Institut national de la statistique et des études économiques. (2016, October 13). IRIS. Retrieved from https://www.insee.fr/fr/metadonnees/definition/c1523

Instituto Brasileiro de Geografia e Estatística. (2024). *Censo Demográfico 2022: Malha de setores censitários*. Rio de Janeiro, Brazil: IBGE.

Instituto Nacional de Estadística. (n.d.). *Labour Force Survey: Sample Design and Evaluation of Data Quality*. Retrieved from https://www.ine.es/en/inebaseDYN/epa30308/docs/epa21_disenc_en.pdf

Kenya National Bureau of Statistics. (2019). 2019 Kenya Population and Housing Census Volume I: Population by county and sub-county. Nairobi, Kenya: KNBS.

Kırlangıçoğlu, C. (2005). *A new census geography for Turkey using geographic information systems: A case study on Çankaya District, Ankara* (Master's thesis). Middle East Technical University, Ankara, Türkiye.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley & Sons.

Lohr, S. L. (2010). *Sampling: Design and analysis (2nd ed.)*. Boston, MA: Brooks/Cole.

Lohr, S. L. (2019). *Sampling: Design and analysis (2nd ed., reprint)*. Boca Raton, FL: Chapman and Hall/CRC.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems (4th ed.)*. Hoboken, NJ: John Wiley & Sons.

Martin, D. (2001). Developing the automated zoning procedure to reconcile incompatible zoning systems. In *Proceedings of GeoComputation 2001*. Retrieved from https://www.geog.leeds.ac.uk/groups/geocomp/2001/papers/martin.pdf

Martin, D. (2002). Geography for the 2001 Census in England and Wales. *Population Trends, 108,* 7–15.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science, 17*(2), 181-196. doi:10.1080/713811750

Martin, D., Nolan, A., & Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A, 33*(11), 1949-1962. doi:10.1068/a3497

Mugnoli, S., Lipizzi, F., & Esposto, A. (2018). New ISTAT 'microzones' layer: A new way to read land cover statistics. *Journal of Research and Didactics in Geography (J-READING), 7*(2), 95–104. doi:10.4458/1563-08

National Statistical Office of Mongolia. (n.d.). *Mongolia - Labour Force Survey 2019: Sampling*. Retrieved from https://web.nso.mn/nada/index.php/catalog/120/sampling

Office for National Statistics. (2016). 2011 Census: Methods and quality report for England and Wales. London, UK: ONS.

Office for National Statistics. (2021). *Census 2021 geographies*. Retrieved from https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies

Office for National Statistics. (2022). *Annual small area population estimates*. London, UK: ONS.

Office of the Registrar General & Census Commissioner, India. (2012). *Census of India 2011 - Administrative Atlas*. Retrieved from https://censusindia.gov.in/census.website/data/atlas

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, UK: Geo Books.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and Planning A, 27*(3), 425–446. doi:10.1068/a270425

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal, 10*(1), 439–446. doi:10.32614/RJ-2018-009

Petrov, A., & Ruus, L. (2007). Geocoding and mapping historical census data: The geographical component of the Canadian Century Research Infrastructure. *Historical Methods, 40*(2), 76–91.

Polsby, D. D., & Popper, R. D. (1991). The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & Policy Review, 9*(2), 301–353.

PostGIS Documentation. (2020). *PostGIS 3.0 manual*. Retrieved from https://postgis.net/documentation/

Qader, S. H., Lefebvre, V., Ninneman, A., Himelein, K., Pape, U., Bengtsson, L., ... Bird, T. (2019). A novel approach to the automatic designation of predefined census enumeration areas and population sampling frames: A case study in Somalia (World Bank Policy Research Working Paper No. 8972). Washington, DC: World Bank.

Qader, S. H., Lefebvre, V., Tatem, A. J., Pape, U., Himelein, K., Ninneman, A., ... Bird, T. (2021). Semi-automatic mapping of pre-census enumeration areas and population sampling frames. *Humanities & Social Sciences Communications, 8*(1), 3. doi:10.1057/s41599-020-00670-0

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Statistics Canada. (2018). *Dictionary, Census of Population, 2016*. Ottawa, Canada: Minister of Industry.

Statistics Canada. (2021). *Dissemination area (DA): Census dictionary*. Retrieved from https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

Statistics Estonia. (2021). *Register-based census model and data exchange infrastructure*. Tallinn, Estonia: Statistics Estonia.

Statistics Finland. (2019). *Population and housing census based on administrative registers*. Helsinki, Finland: Statistics Finland.

Statistics Netherlands. (2019). *The Dutch virtual census: A register-based approach*. The Hague, Netherlands: CBS.

Statistics South Africa. (n.d.). *Small Area Layer (SAL) for Census 2011: Metadata*. Pretoria, South Africa: Stats SA.

Statistics South Africa. (2014). *Census 2011: Methodological report*. Pretoria, South Africa: Stats SA.

Statistics Sweden. (n.d.). *Open geodata for DeSO - Demographic statistical areas*. Retrieved from https://www.scb.se/en/services/open-data-api/open-geodata/

Statistics Sweden. (2018). *Register-based census methodology*. Stockholm, Sweden: Statistics Sweden.

Statistisches Bundesamt [Destatis]. (n.d.). *Census 2022*. Retrieved from https://www.destatis.de/EN/Themes/Society-Environment/Population/Census2022/_node.html

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLOS ONE, 2*(12), e1298. doi:10.1371/journal.pone.0001298

U.S. Census Bureau. (2011, December 22). *2010 Census operational assessment for Type of Enumeration Area delineation* (2010 Census Planning Memoranda Series No. 164). Retrieved from https://www2.census.gov/programs-surveys/decennial/2010/program-management/planning-memo-series/2010-memo-164.pdf

U.S. Census Bureau. (2020). *Geography and the American Community Survey: What data users need to know.* Retrieved from https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_geography_handbook_2020.pdf

U.S. Census Bureau. (2022). *Census geography overview*. U.S. Census Bureau.

UNICEF. (2013). Multiple Indicator Cluster Surveys (MICS): Manual for mapping and household listing. New York, NY: UNICEF. Retrieved from https://mics.unicef.org/

United Nations Economic Commission for Europe. (2015). *Conference of European Statisticians recommendations for the 2020 censuses of population and housing*. Geneva, Switzerland: United Nations.

United Nations Statistics Division. (2009). *Designing household survey samples: Practical guidelines*. New York, NY: United Nations.

United Nations Statistics Division. (2010). *Handbook on geospatial infrastructure in support of census activities*. New York, NY: United Nations.

United Nations Statistics Division. (2011). Report on the results of a survey on census methods used by countries in the 2010 census round. New York, NY: United Nations.

United Nations Statistics Division. (2017). *Handbook on geographical classification*. New York, NY: United Nations.

United Nations. (2007). Principles and recommendations for population and housing censuses (Rev. 2). United Nations.

United Nations. (2017). Principles and recommendations for population and housing censuses (Rev. 3). United Nations.

United Nations. (2025). Principles and recommendations for population and housing censuses (Rev. 4). United Nations.

Wickham, H. (2019). *Advanced R (2nd ed.)*. Boca Raton, FL: Chapman and Hall/CRC.

World Health Organization. (2018). *2018 global reference list of 100 core health indicators (plus health-related SDGs)*. Retrieved from https://apps.who.int/iris/handle/10665/259951

WorldPop. (n.d.). *WorldPop methods*. Retrieved from https://www.worldpop.org/methods/

Zaletel, M., & Vehovar, V. (2000). Nonresponse and socio-demographic characteristics of enumeration areas. In A. Ferligoj & A. Mrvar (Eds.), *Developments in Survey Methodology* (Metodoloski zvezki 15, pp. 201-215). FDV.

Zhang, J., Yuan, X., Tan, X., & Zhang, X. (2021). Delineation of the urban-rural boundary through data fusion: Applications to improve urban and rural environments and promote intensive and healthy urban development. *International Journal of Environmental Research and Public Health, 18*(13), 7180. doi:10.3390/ijerph18137180

APPENDICES

APPENDIX-A. ALGORITHMIC WORKFLOW OF THE EA DELINEATION PROCEDURE

This appendix presents the operational workflow implemented in the R script used for the delineation of Enumeration Areas (EAs). The statement is intentionally written as a thesis-style algorithmic description rather than a code listing, so that the computational logic can be read independently of the software syntax.

The neighbourhood is treated as the basic processing unit. Within the hierarchical design adopted by the script, the Census Area (CA) is not an optimised output but the administratively existing neighbourhood boundary itself. Enumeration Areas (EAs) are generated first within this fixed CA. Supervisor Areas (SAs) are then constructed by grouping the final EAs inside the same neighbourhood. The target number of EAs per SA is determined by the local DEGURBA class, with 4 EAs per SA in rural areas, 5 in medium-density urban areas, and 7 in dense urban areas. Accordingly, the desired number of SAs is obtained from the total EA count in the neighbourhood. In the implemented configuration, SA construction follows an axis-order grouping procedure: EAs are first separated into barrier-aware connected components, then ordered along the principal spatial axis of each component, and finally grouped into contiguous SA segments so that spatial coherence is preserved and workload remains balanced in terms of target units. The full procedure therefore combines polygonization, barrier-aware adjacency, workload-constrained growth, controlled split and merge operations, and optional topology correction.

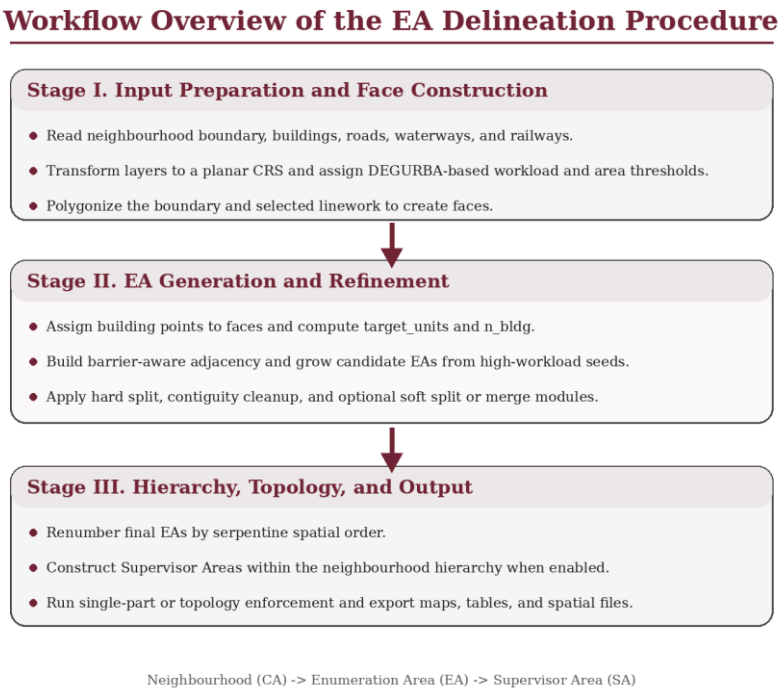


Figure A.1. Main Stages of the EA Delineation Procedure.

Implementation note. In the current script version, the core sequence is defined as faces -> barrier-aware adjacency -> grow -> hard split -> cleanup. The hard split stage is a conditional core pass controlled by HARD_SPLIT_ON. When enabled, it acts on EAs whose workload exceeds the DEGURBA-specific TARGET_MAX or whose area exceeds AREA_MAX. SA construction is enabled, zero-target and low-target merges are enabled, and the final topology-enforcement pass is defined but disabled by default.

Table A.1. DEGURBA-Based Operational Thresholds Used by the Script.

DEGURBA class	TARGET_MIN	TARGET_MAX	AREA_MAX (m²)
Dense urban	80	120	200,000
Medium-density urban	70	140	350,000
Rural	50	200	800,000

The threshold set above is directly linked to the DEGURBA class assigned to the neighbourhood. These values determine the lower and upper workload limits used during EA construction and the maximum polygon area tolerated before a split or corrective intervention is considered.

Algorithm A.1. Stepwise Statement of the EA Delineation and SA Construction Procedure.

Step	Operation
1	**Define the processing unit.** Select one neighbourhood as the operational unit and read the neighbourhood boundary, building layer, road network, waterway layer, and railway layer where available.
2	**Transform and standardize the inputs.** Convert spatial layers to a common planar coordinate reference system, repair invalid geometries, and recover the local DEGURBA class from the building dataset.
3	**Assign threshold values.** Load the DEGURBA-based values of TARGET_MIN, TARGET_MAX, and AREA_MAX that will constrain workload accumulation and maximum EA size.
4	**Prepare the barrier system.** Separate the transport network into linework used for face construction and stricter linework used as hard barriers during adjacency and merge decisions; add waterways and railways when available and buffer the barrier geometry.
5	**Construct faces.** Polygonize the neighbourhood boundary together with the selected line network to create small spatial units called faces, then remove invalid or very small polygons and clip the result to the neighbourhood boundary.
6	**Attribute buildings to faces.** Convert each building to an interior point, assign the point to a face, and compute face-level statistics for target_units and n_bldg.
7	**Build barrier-aware adjacency.** Create a face adjacency graph by retaining only true shared borders that exceed the minimum border-length criterion and rejecting links that cross the buffered barrier system except for explicitly allowed boundary cases.
8	**Initialize EA seeds.** Select high-workload faces as initial seeds for EA construction and assign temporary EA identifiers.
9	**Grow candidate EAs.** Expand each seed by absorbing adjacent unassigned faces until the cumulative workload reaches an acceptable level relative to TARGET_MIN, while preserving contiguity and respecting the barrier-aware adjacency graph.
10	**Fill remaining unassigned faces.** Assign residual faces, including zero-target faces, to neighbouring EAs by using an area-balanced scoring rule that penalizes excessive growth beyond TARGET_MAX or AREA_MAX.

Step	Operation
11	Dissolve and hard-split oversized EAs. Dissolve faces by EA identifier to form polygons and, when HARD_SPLIT_ON is enabled, split any EA whose workload exceeds the DEGURBA-specific TARGET_MAX or whose area exceeds AREA_MAX; if the split yields zero-target fragments, immediately reassign them to suitable neighbours.
12	**Enforce strict contiguity.** Identify disconnected components inside each EA with respect to the face graph and reassign detached components to the best-touching neighbouring EA, preferably the one sharing the longest common border.
13	**Apply optional refinement passes.** When enabled, perform soft split passes for target or area control and execute zero-target or low-target merge modules under conservative acceptance conditions.
14	**Renumber final EAs.** Reorder the EA identifiers by a serpentine spatial sequence running generally from northwest to southeast, alternating left-to-right and right-to-left order across rows.
15	**Construct higher-level hierarchy and outputs.** Generate SA membership and hierarchical identifiers within the CA, optionally enforce topology and single-part geometry, and write the final spatial, tabular, and cartographic outputs.

Table A.2. Post-processing Modules and Acceptance Conditions.

Module	Trigger	Operational rule
Hard split	Conditional core pass; HARD_SPLIT_ON = TRUE and EA workload > TARGET_MAX or EA area > AREA_MAX	Split the EA geometrically under the DEGURBA-specific workload and area limits; if the split yields zero-target fragments, reassign them before continuing.
Soft split (target)	Optional pass; tgt_units exceeds TARGET_MAX	Accept the split only if both parts remain meaningful, both parts have positive workload, and the maximum overload is reduced sufficiently.
Soft split (area)	Optional pass; area exceeds AREA_MAX	Accept the split only if both parts remain valid polygons and satisfy the same minimum-size safeguards used in the target pass.
Zero-target merge	Optional pass; tgt_units = 0	Merge the EA into a barrier-safe neighbour, preferably by nearest or best-touching assignment; fallback can be disabled.
Low-target merge	Optional pass; 0 < tgt_units < TARGET_MIN x threshold fraction	Merge only when the receiving EA does not grow excessively in area and does not exceed the allowed post-merge workload fraction.
Single-part enforcement	Multipart or island-like EA geometry	Keep the largest part in the original EA and reassign smaller detached fragments to the best neighbouring EA.
Topology enforcement	Optional final QA module	Clip the final set to the CA, remove overlaps, fill gaps, and test the result against a gap and overlap tolerance.
Serpentine renumbering	Final EA stage	Assign readable EA_ID values by row-wise spatial ordering with alternating direction.
SA construction	Optional hierarchy module	Group EAs into Supervisor Areas under DEGURBA-based target EA counts and balancing rules; the uploaded script uses AXIS_ORDER as the current SA method.

The script also generates a formal output package for review, quality control, and operational use. Because these products show how the delineation is implemented at neighbourhood and district scales, they should be mentioned briefly in the appendix even if the thesis discusses their analytical use elsewhere.

Table A.3. Output Products Generated by the Script at Neighbourhood and District Levels.

Level	Main output products	Purpose in the workflow
Neighbourhood	EA layer; optional SA layer; CA identifier fields; local statistics tables; review maps.	To store the final units, preserve identifiers, and support local inspection of contiguity, barriers, numbering, and workload balance.
District	Merged district layers; consolidated summary tables; DEGURBA overviews; district maps; reporting sheets.	To support district-wide quality control, compare neighbourhood outputs, and produce implementation and documentation summaries.
Tabular exports	CSV and spreadsheet-compatible summaries of counts, target-unit totals, area measures, and diagnostics.	To provide an audit trail for threshold compliance, workload summaries, and links between spatial outputs and thesis tables.

Taken together, the procedure is a constrained spatial partitioning algorithm that seeks to produce field-usable, barrier-respecting, contiguous, and topologically stable EAs. In the corrected script configuration, the hard split stage is both toggle-controlled and tied to DEGURBA-specific workload ceilings rather than to a single global hard threshold. The appendix format used here is intended to support thesis readability by translating the operational logic of the script into a formal methodological statement.

APPENDIX-B. SUPPLEMENTARY NOTE ON SOFTWARE IMPLEMENTATION AND THE GRAPHICAL USER INTERFACE

A local graphical user interface (GUI) was developed in R using the Shiny framework to facilitate the operational use of the EA delineation workflow. The interface was designed as a front-end layer over the same core delineation script rather than as a separate computational method. Accordingly, the algorithmic logic, threshold structure, split-merge rules, and output regime remain identical to those of the script-only execution mode.

In practical terms, the interface collects the input and output paths, district and neighbourhood filters, split and merge options, topology guardrails, and DEGURBA-specific workload thresholds. These user selections are assembled into a configuration object, written into a run-specific script instance by patching the template script, subjected to a pre-parse validation step, and then executed in-process. During runtime, the application records the generated script path, updates execution status, and streams the most recent log lines to the user.

Because the same script is executed, the GUI produces the same analytical outputs as the script-only workflow, including EA layers, optional shapefiles, summary tables, and neighbourhood-level or district-level overview products. The interface should therefore be interpreted as an operational deployment layer that simplifies the use of the method without altering the underlying delineation procedure.

Figure B.1. Operational Architecture of the Shiny-based Interface Developed for the EA Delineation Workflow.

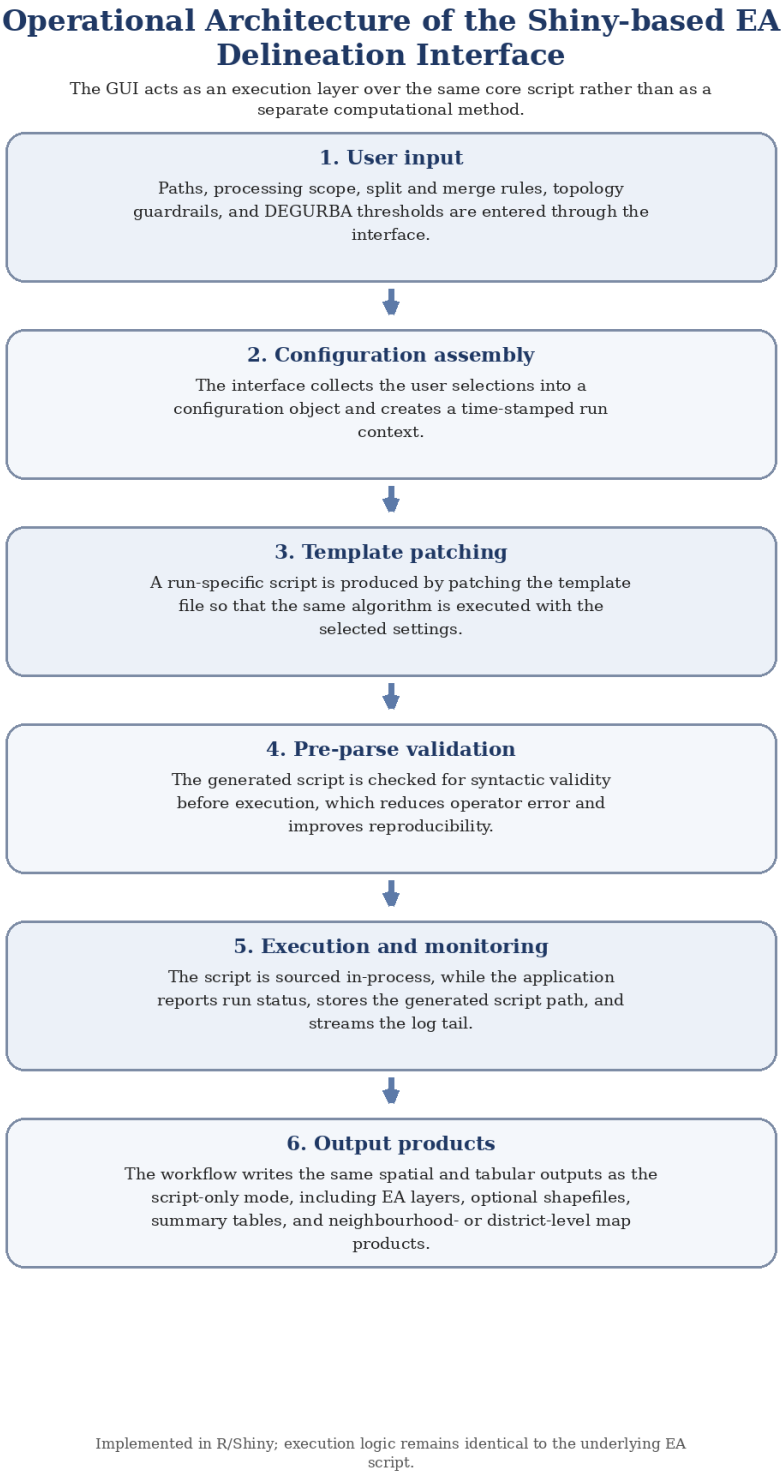
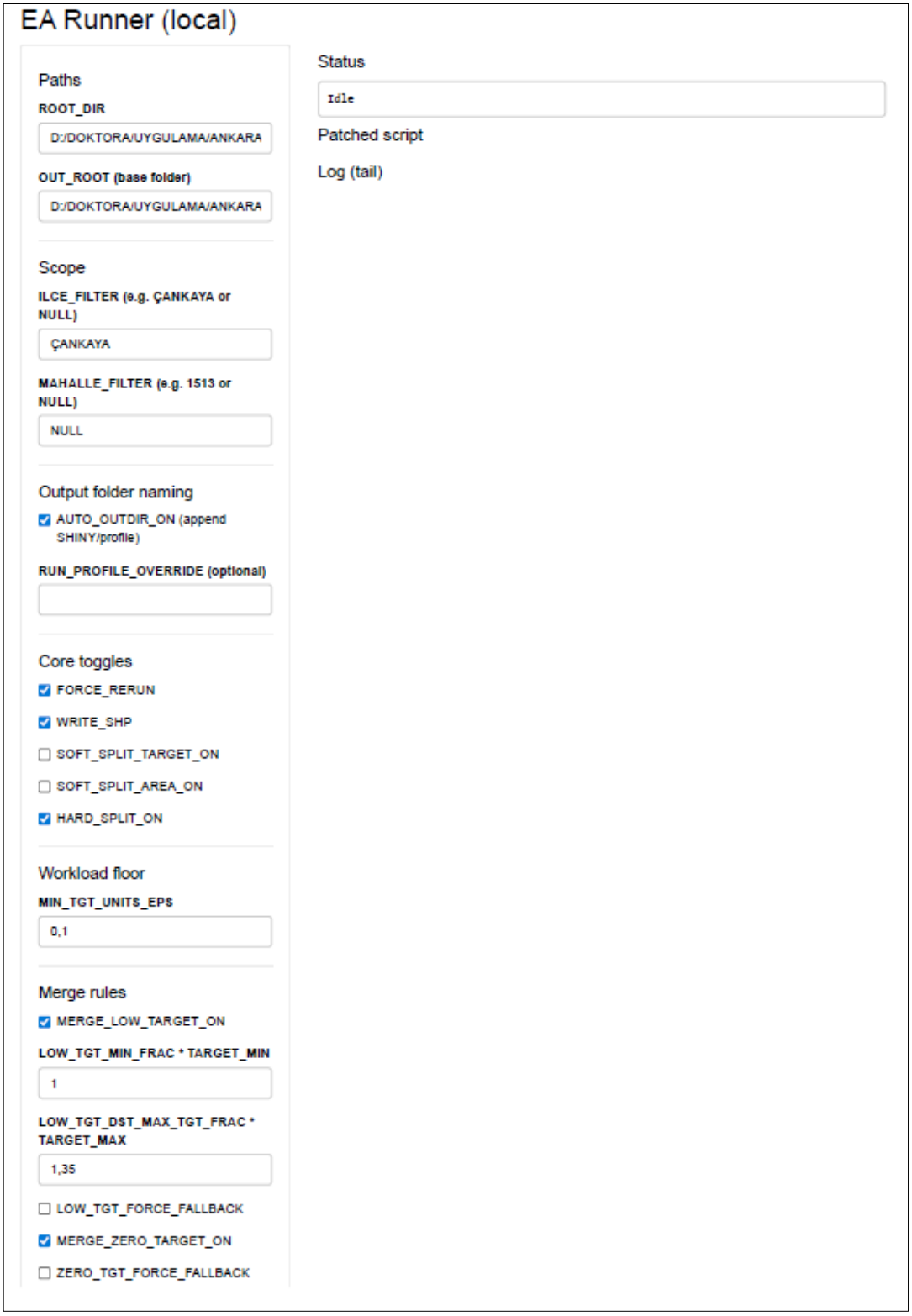


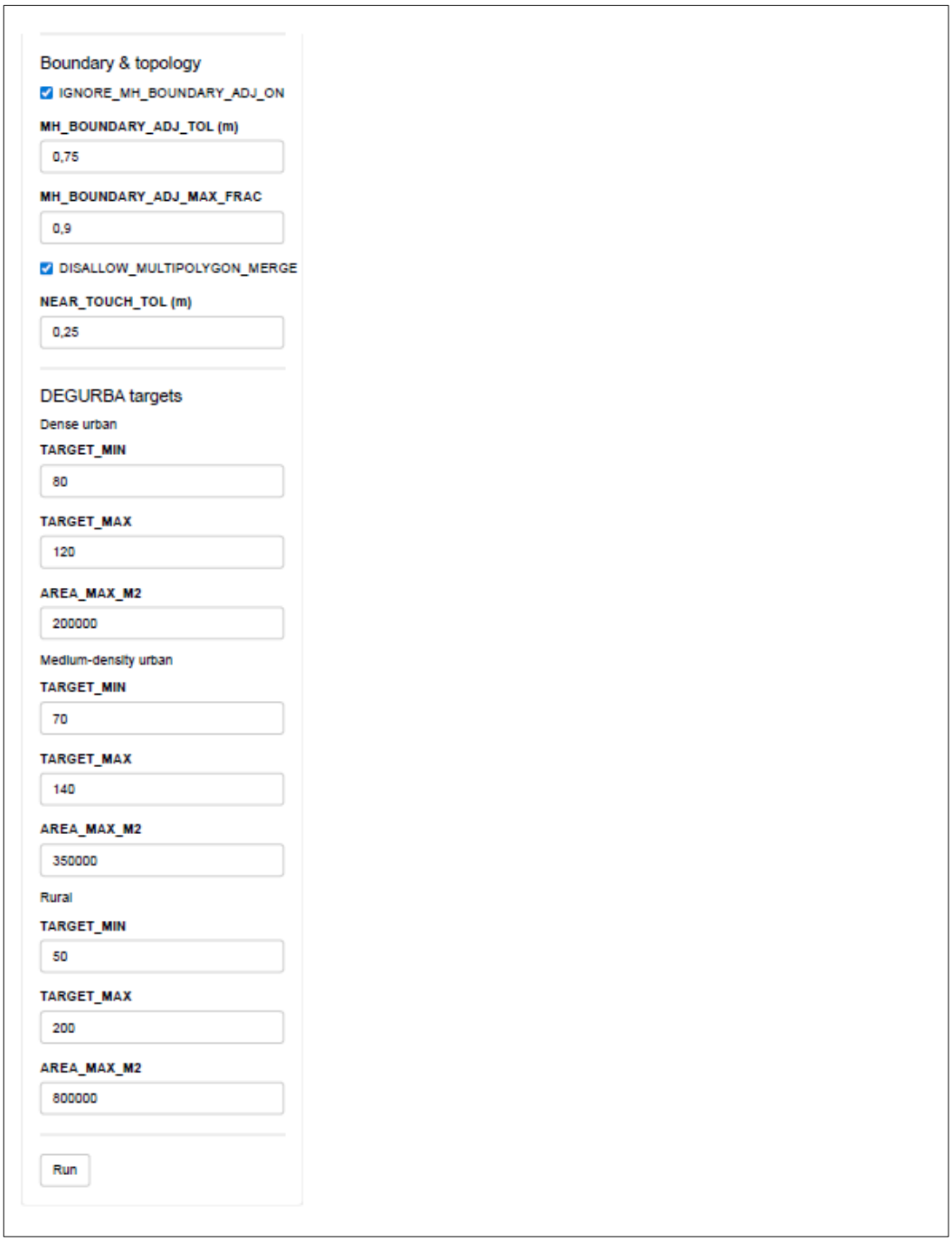
Table B.1. Main GUI Parameter Groups and Their Operational Roles in the Workflow.

Interface group	Operational role	Representative controls
Paths, scope, and output location	Defines the data root, processing extent, and output folder for a run.	ROOT_DIR, OUT_ROOT_BASE, ILCE_FILTER, MAHALLE_FILTER
Core toggles	Switches major processing and export options on or off.	FORCE_RERUN, WRITE_SHP, SOFT_SPLIT_TARGET_ON, SOFT_SPLIT_AREA_ON, HARD_SPLIT_ON
Workload and merge controls	Sets the workload floor and the handling of low-target or zero-target units.	MIN_TGT_UNITS_EPS, MERGE_LOW_TARGET_ON, LOW_TGT_MIN_FRAC, LOW_TGT_DST_MAX_TGT_FRAC, MERGE_ZERO_TARGET_ON
Boundary and topology guardrails	Controls boundary-adjacency tolerance and topology-safe merge restrictions.	IGNORE_MH_BOUNDARY_ADJ_ON, MH_BOUNDARY_ADJ_TOL, MH_BOUNDARY_ADJ_MAX_FRAC, DISALLOW_MULTIPOLYGON_MERGE, NEAR_TOUCH_TOL
DEGURBA thresholds	Defines class-specific workload and area limits used by the core algorithm.	TMIN_DENSE/TMAX_DENSE/AMAX_DENSE; TMIN_MED/TMAX_MED/AMAX_MED; TMIN_RUR/TMAX_RUR/AMAX_RUR
Run monitoring outputs	Displays execution state and recent run information to the operator.	Status, patched script path, log tail

In thesis presentation, the architecture figure is best placed in the methodology chapter under a subsection such as "Software Implementation and User Interface", whereas the parameter table may be retained in the appendix as implementation documentation.

APPENDIX-C. SCREENSHOTS OF THE SHINY-BASED INTERFACE DEVELOPED FOR THE EA DELINEATION WORKFLOW




Boundary & topology
IGNORE_MH_BOUNDARY_ADJ_ON
MH_BOUNDARY_ADJ_TOL (m)
0,75
MH_BOUNDARY_ADJ_MAX_FRAC
0,9
DISALLOW_MULTIPOLYGON_MERGE
NEAR_TOUCH_TOL (m)
0,25
DEGURBA targets
Dense urban
TARGET_MIN
80
TARGET_MAX
120
AREA_MAX_M2
200000
Medium-density urban
TARGET_MIN
70
TARGET_MAX
140
AREA_MAX_M2
350000
Rural
TARGET_MIN
50
TARGET_MAX
200
AREA_MAX_M2
800000
Run

APPENDIX-D. ORIGINAL ARTICLE

(starts with the next page)

To cite:, Öztürk, C., Türkyılmaz, Ahmet Sinan (2026.) , GIS-Based Enumeration Area Delineation: A Trial for Türkiye with Selected Examples from the United States, *Coğrafi Bilimler Dergisi/ Turkish Journal of Geographical Sciences, 26 (1)*, doi: 10.33688/aucbd.1805466

Coğrafi Bilimler Dergisi

Turkish Journal of Geographical Sciences

e-ISSN:1308-9765

GIS-Based Enumeration Area Delineation: A Trial for Türkiye with Selected Examples from the United States

CBS Tabanlı Sayım Alanlarının Oluşturulması: Türkiye için Bir Deneme ve Amerika Birleşik Devletleri'nden Seçilmiş Örnekler

Cansu Öztürk*[a], Ahmet Sinan Türkyılmaz [b]

Makale Bilgisi

Araştırma Makalesi

DOI:

Makale Geçmişi:
Geliş:
Kabul:

Anahtar Kelimeler:
Nüfus,
Sayım alanı,
Örnekleme çerçevesi,
Demografi,
Coğrafi Bilgi Sistemleri

Öz

Bu çalışma, Türkiye'de Sayım Alanlarının (SA) Coğrafi Bilgi Sistemleri (CBS) tabanlı tanımlanmasına yönelik bir yaklaşım önermektedir. Çalışmada, Amerika Birleşik Devletleri ve Türkiye'den olmak üzere dört veri seti üzerinde uygulama yapılmış; girdi değişkenlerinin dağılım özellikleri ile mekânsal bitişiklik kısıtlarının alan oluşturma performansını nasıl etkilediği incelenmiştir. Uluslararası uygulamalar, Türkiye'nin konumunu bağlamlandırmak ve iyi uygulamaların benimsenmesine yönelik çıkarımlar üretmek amacıyla gözden geçirilmiştir. Uygulamada; farklı eşik değerler, bitişiklik kuralları ve gelişmiş parametre ayarları altında, coğrafi olarak bütüncül, sahada uygulanabilir ve istatistiksel açıdan dengeli sayım alanları oluşturulması hedeflenmiştir. Bulgular, CBS tabanlı alan tanımlamanın sınırları net, maliyet etkin ve daha yönetilebilir alanlar oluşturmayı destekleyebileceğini göstermektedir. Çalışma, SA sistemlerinin güçlendirilmesinin yalnızca teknik bir süreç olmadığını; ulusal istatistik operasyonlarının verimliliği ve kalitesi açısından stratejik bir yatırım olduğunu ortaya koymaktadır.

Article Info

Research Article

DOI:

Article History:
Received:
Accepted:

Keywords:
Population,
Enumeration area,
Sampling frame,
Demography,
Geographic Information Systems

Abstract

This study proposes a Geographic Information System (GIS)-based approach to delineate Enumeration Areas (EAs) in Türkiye. Analyses were conducted on four datasets from the United States and Türkiye to examine how the distributional properties of input variables and spatial contiguity constraints influence the performance of zone creation. International practices are reviewed to contextualize Türkiye's position and to derive lessons for best-practice adoption. The workflow tests alternative threshold values, contiguity rules, and advanced parameter settings to produce EAs that are geographically coherent, operationally workable, and statistically balanced. The findings indicate that GIS-based delineation can support the creation of clearly bounded, cost-effective, and more manageable areas. The study concludes that strengthening EA systems is not merely a technical exercise but a strategic investment in the efficiency and quality of national statistical operations.

***Corresponding Author:** ozturkcansu07@gmail.com

[a] Hacettepe University Institute of Population Studies, Department of Social Research Methodology, Ankara/Türkiye ORCID Number (http://orcid.org/0009-0007-3096-1565)

[b] Hacettepe University Institute of Population Studies, Department of Social Research Methodology, Ankara/Türkiye, ORCID Number (http://orcid.org/0000-0002-2783-932X)

1. Introduction

An enumeration area (EA) is a small, clearly bounded geographic unit used as the basic building block for censuses and household surveys. It is designed so that one enumerator (or a small team) can cover all dwellings within its boundary during a defined fieldwork period. EAs are typically constructed to ensure complete coverage without overlap or gaps, to contain a manageable and broadly similar workload (often defined by a target number of households or dwellings), and to follow recognizable features such as roads, rivers, or administrative boundaries so that they can be located and worked on consistently in the field.

Census geography refers to the standardized set of geographic units and their hierarchical relationships that a national statistical office uses to plan field operations and to produce, tabulate, and disseminate census and survey statistics. It provides a consistent geographic framework by organizing space into nested units, enabling data to be aggregated and compared across places and over time. A well-defined census geography supports full territorial coverage, improves operational planning, and strengthens the comparability of published statistics. Within this framework, EAs function as the core operational building blocks: they translate the census geography structure into actionable field units for workload allocation and coverage control, while also providing small-area units that can be aggregated into higher-level reporting geographies(United Nations, 2**1**009; United Nations, 2021).

Accurate and efficient data collection is the cornerstone of modern national statistical systems. In both censuses and sample surveys, EAs serve as the smallest operational units, providing the basic spatial framework for organizing fieldwork and ensuring systematic coverage of the target population. A well-designed EA system contributes to complete population coverage, reduces data collection costs, and improves the overall quality, comparability, and consistency of statistical outputs. As the United Nations Statistics Division (United Nations, 2009) emphasizes, EA delineation should be guided by principles such as spatial compactness, manageable workload for enumerators, and alignment with administrative or natural boundaries. These principles are essential for ensuring that enumeration can be performed efficiently and that data collection activities remain standardized across diverse regions.

Beyond the technical dimension of delineation, the importance of EAs lies in their connection to the broader logic of sample surveys. In modern statistics, surveys are indispensable for generating timely, detailed, and representative information about populations (Groves et al., 2009). Full population enumeration, while theoretically desirable, is often infeasible due to the immense financial and logistical resources required. As a result, statistical offices rely on sample surveys, which collect data from a carefully selected subset of the population and generalize the findings to the entire population with measurable accuracy (Kish, 1965; Lohr, 2010). For this reason, the design of EAs directly affects the ability to construct reliable sampling frames—comprehensive lists or databases that serve as the basis for drawing representative samples.

The rationale for conducting surveys, particularly sample-based surveys, extends beyond resource efficiency. They allow statistical agencies to generate continuous and specialized insights into economic, demographic, and social processes without waiting for the decennial census cycle. For example, household budget surveys, labor force surveys, and health surveys rely on accurate EA

delineation to ensure representative sampling at national and subnational levels (Dillman, Smyth, & Christian, 2014). In this sense, EAs serve as the operational link between the statistical concepts of representativeness and the practical execution of fieldwork.

A critical methodological element of surveys is the sampling frame. The sampling frame is the operational listing of all units in the population from which the sample is drawn, and its quality directly determines the representativeness of survey results (Bethlehem, Cobben, & Schouten, 2011). A good sampling frame must be complete, accurate, current, and free of duplication. When frames are incomplete or outdated, coverage errors arise—leading to the systematic exclusion of certain population groups and thereby biasing estimates. In the context of Türkiye, the Spatial Address Registration System(SARS)provides a promising foundation for building modern sampling frames, as it combines geographic precision with administrative reliability. By shifting from traditional, text-based enumeration lists to spatially defined digital frameworks, SARS has the potential to reduce coverage errors and enhance operational efficiency.

Türkiye, like many other countries transitioning to register-based statistical systems, faces the challenge of modernizing its EA methodology in line with changing settlement patterns, rapid urbanization, infrastructure development, and increasing availability of digital geospatial data. Traditional EA systems, based on paper maps and descriptive boundary definitions, often fail to capture the complexity of urban sprawl or the fluidity of administrative boundaries. The introduction of spatially defined EAs using Geographic Information Systems(GIS) addresses this limitation by allowing automated delineation, dynamic updating, and integration of multiple data sources. International experience demonstrate that GIS-based EA systems not only improve efficiency but also support innovations in data collection, such as handheld devices, GPS-based field monitoring, and real-time quality assurance mechanisms (Kalton, 2020).

The role of EAs becomes even more significant when viewed from the perspective of data quality dimensions. According to Groves et al. (2009), survey quality depends on minimizing total survey error, which includes coverage error, sampling error, nonresponse error, and measurement error. The design and maintenance of EAs directly affect the first two components: coverage and sampling. Spatially consistent and up-to-date EAs ensure that every household has a known and nonzero probability of selection, thereby enhancing the representativeness and accuracy of survey results. At the same time, compact and manageable EAs reduce the workload for enumerators, lowering the risk of data collection errors.

The aim of this study is to propose a GIS-based approach for delineating Enumeration Areas (EAs) for Türkiye through a trial implementation. Selected case applications from the United States illustrate how properties of input variables, dataset scale, and spatial contiguity constraints influence automated zone creation. The Türkiye application illustrates how the workflow can be configured to produce EAs that are geographically coherent, operationally workable, and statistically balanced, and it has potential relevance for future census and household survey operations in Türkiye. In addition, the analysis considers institutional prerequisites for sustainable implementation, including data governance and interagency coordination, and reviews international practices to contextualize Türkiye’s position

and derive practical lessons for adoption. In this context, strengthening EA systems is regarded as an area for operational and governance-oriented improvement of national statistical systems.

2. Literature Review

The global literature demonstrates a significant evolution in the methods used for delineating Enumeration Areas(EAs) moving from traditional manual cartography to advanced, GIS-supported, and even automated processes. EAs are no longer viewed only as operational units for census fieldwork but as fundamental components of national statistical systems that interact with geospatial infrastructures, administrative registers, and digital technologies. This transformation reflects the growing demand for high-quality, timely, and spatially precise data for evidence-based policymaking, economic planning, and social research (United Nations, 2009; Esri, 2020).

In the early stages of census-taking, EAs were defined primarily by hand-drawn boundaries on paper maps, often following visible physical landmarks or administrative units. While functional, these approaches were highly dependent on local knowledge and difficult to replicate consistently. Errors were frequent, especially in rural or remote areas, where natural boundaries were vague and settlement patterns were scattered. By the mid-20th century, national statistical offices(NSOs)increasingly realized the need for more standardized methods, particularly as censuses grew in scope and complexity. The growing demand for household and labor force surveys further highlighted the importance of precise EA delineation, as survey quality is directly linked to the accuracy of the underlying sampling frame (Groves et al., 2009; Kish, 1965).

In developed countries, the transition from manual to digital approaches was accelerated by the availability of robust administrative registers and technological infrastructure. The United States, for instance, institutionalized the use of Census Blocks and Census Tracts through the TIGER/Line system, which, since the 1980s, has enabled detailed mapping, consistent boundary updates, and the integration of statistical data with digital geographies (U.S. Census Bureau, 2021). Canada's Dissemination Areas (DAs), established in 2001, provide standardized units for detailed socio-economic analysis across provinces and territories, facilitating cross-regional comparisons and supporting federal funding allocations (Statistics Canada, 2022).

The United Kingdom introduced Output Areas (OAs) in 2001, which have since become the backbone of small-area statistics, supporting not only census dissemination but also a wide variety of social and economic indicators (ONS, 2021). France's IRIS units, developed in 1999, similarly illustrate how statistical geography can be adapted to urban planning and sub-municipal policymaking (INSEE, 2021). Germany's Erhebungsbezirke, established in 1987, show how alignment with administrative and infrastructural realities supports high-quality labor force and household surveys.

Australia offers another important example with its Statistical Area Level 1 (SA1) units, designed to provide consistent small-area data for policy planning across a vast and diverse territory. These units are particularly effective in supporting local planning because they can be aggregated into larger statistical areas, ensuring flexibility in both design and application. Estonia, meanwhile, represents a pioneering case of digital innovation, in which EAs are dynamically generated through the

integration of population registers and GIS. This adaptive approach allows for continuous updating and significantly reduces coverage error (Statistics Estonia, 2021).

In contrast, developing countries have faced greater challenges due to weaker administrative infrastructures, limited technical capacity, and financial constraints. Somalia's pilot use of high-resolution satellite imagery to delineate EAs (Qader et al., 2019) demonstrates how geospatial innovation can partially compensate for institutional fragility. Botswana (Statistics Botswana, 2016) has also used geospatial solutions to update enumeration boundaries, underscoring the potential of digital tools in resource-limited environments.

Kenya's 2019 census illustrates how GPS-supported enumeration geography can strengthen coverage control in rapidly changing urban environments. In the Kenyan case, georeferenced EA maps were loaded onto tablets, and household location data (GPS coordinates) were used to support quality checks and completeness verification during field operations (Wanyoike, 202**2**). India, a country with a population exceeding one billion, relies on Enumeration Blocks (typically designed to cover 120–150 households) and has increasingly integrated GIS-based databases into its census mapping and fieldwork organization (Chakravorty, 2007). Indonesia is not a "billion-plus" country; yet it similarly institutionalizes GIS-enabled small-area operational units through Wilayah Kerja Statistik (Wilkerstat), supported by satellite imagery, GPS devices, and GIS for census and survey mapping activities (BPS-Statistics Indonesia, 2018; Worldometer, 2025). In South Africa, Small Area Layers (SALs) were developed by aggregating small enumeration areas to meet confidentiality and usability thresholds for dissemination, demonstrating how GIS-enabled statistical geographies can extend beyond enumeration to support broader small-area statistical products (Mokhele, Mutanga, & Ahmed, 2016; Statistics South Africa, n.d.).

Brazil faces a distinct set of challenges because rapid urban growth and the dynamism of informal settlements can outpace intercensal boundary maintenance, creating divergences between census basemaps and municipal settlement maps. For example, evidence from São Paulo shows that differences in identification and delimitation of favelas between federal census and municipal basemaps can be substantial, which may contribute to underestimation and difficulties in maintaining stable small-area boundaries over time (Pedro & Queiroz, 2019). In response to these territorial dynamics, Brazil's national statistical office (IBGE) reports a major expansion and a decentralised update of its census cartography (Territorial Base) for the 2022 Census, supported by high-resolution imagery and locally sourced operational information, which are intended to reduce the risk that newly occupied areas are missed (IBGE, 2022). Nevertheless, the official post-enumeration coverage analysis for the 2022 Census documents material coverage gaps and explicitly identifies difficulty in covering large urban centres — including "slums and urban communities" — as one of the operational challenges associated with higher error rates (IBGE, 202**3**). Collectively, these observations underscore a broader point: the sustainability of EA systems depends not only on GIS technology but also on stable institutional arrangements that ensure continuous maintenance, adequate funding, and coordinated statistical–geospatial integration across agencies. Without these, GIS tools alone cannot guarantee data quality (United Nations, 2008; PARIS21 & Statistics Sweden, 2021).

A common thread in the literature is that EAs are not only essential for census-taking but also central to sample surveys, household surveys, and labor force studies. They ensure that representative samples can be drawn with known selection probabilities, thereby minimizing bias and improving the reliability of estimates (Lohr, 2010). Moreover, EAs facilitate spatial analysis by linking demographic and geographic information, which has become increasingly important in fields such as health, education, environmental monitoring, and infrastructure planning (United Nations, 2009; Esri, 2020).

Advantages of well-designed EAs include enhanced operational efficiency, reduced fieldwork costs, minimized risks of double-counting or omission, and improved integration of demographic and geospatial data (UNECE, 201**4**). By providing standardized geographic units, EAs allow policymakers to compare conditions across regions, identify vulnerable populations, and allocate resources more effectively. The visualization capabilities enabled by GIS further expand the utility of EAs by supporting decision-making through maps and spatial analytics.

Nevertheless, significant limitations and challenges remain. One persistent difficulty is the resource-intensive nature of maintaining EA boundaries. Updating requires not only technical expertise but also coordination among government agencies, local authorities, and field staff. In rapidly urbanizing countries, informal settlements and migration flows often outpace official boundary revisions, generating coverage errors. Privacy and confidentiality concerns also intensify as geospatial precision increases. Balancing the demand for small-area statistics with the need to protect individual anonymity remains a key methodological and ethical challenge (UNECE, 2015).

Emerging trends suggest several promising directions. First, adaptive and dynamic EA systems are gaining attention. Estonia exemplifies how continuous integration of administrative registers and GIS technologies can create real-time updating processes, reducing coverage errors and improving efficiency (Statistics Estonia, 2021). Second, the increasing role of artificial intelligence and remote sensing in automated EA delineation offers opportunities for timelier updates, particularly in countries with limited staff capacity. Third, interoperability between statistical and geospatial systems is becoming a priority, as it allows data from multiple sectors to be linked and analyzed at the EA level.

In Türkiye, there are currently no officially defined enumeration areas. Instead, in sample surveys—particularly household-based ones—artificial clusters are created within the same settlement unit (quarter or village) when household addresses are grouped by the Turkish Statistical Institute (TurkStat). These clusters serve as the primary sampling units. During the clustering process, the occupied addresses within each settlement are sorted by their address components (such as street or avenue name, building number, and apartment number) and then divided into clusters designed to contain approximately equal numbers of units (~100 households each). The sampling frame is updated and clusters are reconstructed twice a year. Since this process is not based on geographic coordinates, the clusters and the number of units they contain are not fixed and may vary over time.

The output of the current blocking procedure, which does not account for spatial characteristics, is illustrated in Figure 1. The author prepared the figure to align with the method's output.

Figure 5. Schematic illustration of the existing non-spatial blocking method

An early academic effort discussed the feasibility of designing a dedicated census geography for Türkiye using GIS-based delineation principles. In this study, a candidate small-area hierarchy was explored through a case study in Çankaya(Ankara),with the aim of producing standardised statistical units for census dissemination and for small-area statistics beyond the constraints of existing administrative boundaries (Kırlangıçoğlu, 2005). Importantly, this design was proposed as a methodological alternative but was not adopted as an official operational system in Türkiye's routine official statistical production. Nevertheless, it remains a useful reference point because it frames EA-like units as the backbone of national statistical geography and, by contrast, clarifies the institutional trade-off observed in practice. While boundary-independent statistical areas can be analytically attractive, official statistics and policy reporting often benefit from EA designs that can be consistently related to administrative hierarchies (e.g., through nesting or robust correspondence tables) and can also satisfy operational constraints such as workload balancing.

In conclusion, the literature clearly demonstrates that EA systems are evolving from static, manually defined units into dynamic, geospatially integrated infrastructures that underpin modern statistical systems. Developed countries leverage robust registers and advanced GIS tools to maintain up-to-date, highly detailed EAs, while developing countries employ innovative, resource-constrained approaches, such as satellite imagery or grid-based models (Table 1). Across contexts, the fundamental role of EAs in ensuring representativeness, accuracy, and efficiency remains constant, even as the methods and technologies evolve. Their continued development will be critical for the future of evidence-based policymaking, social research, and national statistical capacities worldwide.

Table 1. Comparative overview of EA systems

Country	EA Unit Type	Data Source/Frame	Technology Used	Update Frequency	Advantages	Challenges
United States	Census Blocks/Tracts	TIGER/Line database	GIS, digital mapping	Every 10 years (census)	Detailed small-area statistics, integration with mapping	Costly updates, confidentiality issues
United Kingdom	Output Areas (OAs)	National census registers	GIS-enabled, statistical registers	Every 10 years	Supports local policy and deprivation indices	Static boundaries, slow updates
Canada	Dissemination Areas (DAs)	Household and address registers	GIS, geocoding	Every 5 years (census)	Granular socio-economic analysis	Coverage issues in remote areas
France	IRIS Units	Census and administrative registers	GIS-based statistical system	Every census and major update	Sub-municipal level analysis for urban planning	Resource-intensive maintenance
Germany	Erhebungsbezirke	Administrative boundaries	GIS, integrated mapping	Before major surveys	Efficient for labor force surveys	Dependent on local admin data quality
Kenya	Enumeration Areas	Census lists	GPS-enabled, GIS-based	2019 census	High geospatial accuracy	Sustainability of GPS systems
Somalia	Grid-based EAs	Satellite-derived maps	High-resolution imagery	Pilot since 2014	Innovation under constraints	Fragile institutions, limited coverage
Brazil	Custom EAs	Field-based mapping	Mixed GIS/manual	Irregular updates	Flexibility in diverse contexts	Urban sprawl, informal settlements
Estonia	Adaptive EAs	Population registers	Dynamic GIS integration	Continuously updated	Real-time adaptation, low coverage error	Requires strong digital infrastructure
South Africa	Small Area Layers	Census and admin data	GIS-based, digital mapping	Every census	Supports census and ongoing surveys	High resource needs for updates
India	Enumeration Blocks	Census operations	Manual + GIS integration	Every 10 years	Covers large rural population	Challenging in remote areas
Australia	Statistical Area Level 1 (SA1)	National address file	GIS, ABS systems	Regular census cycles	Policy-relevant small-area data	Requires large administrative effort

Source: Compiled from national statistical office reports and international guidelines.

3. Methodology

The ArcGIS Build Balanced Zones (BBZ) tool was used to delineate enumeration areas (EAs) in four case-study applications—Richmond County, Seattle, Story County, and Ankara—to examine how properties of input variables, dataset scale, and spatial contiguity constraints affect automated zoning performance. To assess the tool's operational behavior and its sensitivity to extreme or atypical

conditions, the methodology first applies controlled trials to international open datasets with stable schemas and well-documented attributes. These preliminary experiments are used to validate the core workflow, identify common failure modes, and clarify how boundary definitions, barrier configurations, and settlement patterns shape the resulting zones. Building on these insights, we then apply the same workflow to the Ankara case study, which is based on a single quarter.

In many countries, GIS platforms—and frequently ArcGIS—have become core components of the production and maintenance of census geography. The U.S. Census Bureau, for example, integrates GIS workflows with the TIGER/Line database to support nationwide geographic consistency for census tracts and blocks (U.S. Census Bureau, 2019). In Canada, ArcGIS Pro is used in the management of Dissemination Areas, which provide a standardized framework for small-area census dissemination and socio-economic analysis (Statistics Canada, 2022). Similar applications are reported in Kenya, where the 2019 census employed GPS-enabled field mapping to digitize and verify enumeration areas, aiming to improve coverage control and reduce operational errors (Kenya National Bureau of Statistics, 2020). In India, census geography and mapping activities have also been supported through GIS-based approaches, including the development and operational use of census geographic databases for enumeration planning and field implementation (Chakravorty, 2007). In Australia, the Australian Bureau of Statistics maintains Statistical Area Level 1 (SA1) units as the smallest geographic units for many demographic and socio-economic outputs, which are supported by ArcGIS Pro-based spatial management workflows (ABS, 2021). In France, IRIS units represent sub-municipal geographic divisions that support detailed census dissemination and urban analysis (INSEE, 2021). Taken together, these cases indicate that ArcGIS Pro is often used not only for cartographic production, but also as part of the operational infrastructure that supports the design, maintenance, and use of statistical geographies.

Within ArcGIS Pro, one of the most relevant tools for EA delineation is the Build Balanced Zones(BBZ)function. BBZ automates the creation of compact and balanced spatial units by optimizing user-defined variables. The tool aggregates small base units, such as census blocks or polygons, into larger meaningful zones. Unlike traditional manual delineation, which is time-consuming and prone to inconsistencies, BBZ relies on optimization algorithms that minimize shape irregularities while ensuring that each unit adheres to the balancing criteria. Users can define population thresholds, numbers of households, or land area as balancing variables, and specify a target value for each EA. For instance, zones can be designed to contain roughly equal resident populations while maintaining contiguity and compactness. BBZ also allows the integration of multiple balancing variables simultaneously, making it possible to create zones that consider not only the population but also the number of households, housing units, or service facilities.In the Build Balanced Zones (BBZ) method, “spatial constraints” define how neighboring features are identified and merged as zones expand (Esri, 2023). Zones are allowed to grow only into features adjacent to at least one feature already included in the same zone, ensuring geographical contiguity and consistency. When the input features are polygons, the default constraint is Contiguity Edges Corners, while for point-based inputs, the default is Trimmed Delaunay Triangulation. The BBZ tool offers several spatial-constraint options, including Contiguity Edges Only, Contiguity Edges Corners, Trimmed Delaunay Triangulation, and Get Spatial Weights from File. By defining these spatial relationships, the BBZ algorithm ensures that zones grow logically and maintain contiguity according to the selected constraint type.

The BBZ tool also provides a set of "Zone Characteristics Criteria" that shape the internal properties of the resulting zones. The main parameters are Equal Area, Compactness, and Equal Number of Features, which respectively aim to balance surface area, geometric regularity, and the number of input features. Adjusting these criteria allows users to fine-tune the trade-offs between spatial compactness, uniformity, and statistical balance (Esri, 2023).

Complementing these criteria, the BBZ tool includes "Advanced Parameters" that control the optimization process. The most influential parameters are Population Size, Number of Generations, and Mutation Factor, which define the scope of iterations, search diversity, and variability in achieving balanced zones. Together, these parameters enable users to balance computational efficiency with spatial and statistical precision.

This tool has found applications across a wide range of sectors. In the United States, education authorities have used BBZ to delineate school districts that maintain equitable student populations. In Europe, statistical offices have integrated BBZ into automated EA creation, particularly in densely populated urban areas where demographic variations are pronounced. The healthcare sector has also adopted the tool to delineate hospital service zones in ways that balance population coverage with geographic accessibility. Transportation planners use BBZ to define traffic analysis zones that support the equitable modeling of road usage and infrastructure demands (Goodchild & Li, 2021). These cross-sectoral applications underscore the flexibility of BBZ and its potential to ensure fairness, efficiency, and reproducibility in zone design.

The advantages of BBZ are particularly evident in large-scale statistical operations such as censuses. By creating compact and balanced zones, the tool ensures a more equitable workload distribution among field staff, reduces the risk of omission or duplication in data collection, and guarantees that the delineation process remains transparent and replicable. This is especially critical for developing countries, where limited resources make efficient EA design a priority. Automating zone creation using BBZ reduces the human and financial costs associated with traditional manual methods while enhancing data quality. Furthermore, the capability to incorporate spatial and demographic variables simultaneously ensures that the resulting EAs are not only statistically sound but also geographically meaningful.

This study applies the BBZ methodology to the Turkish context. Türkiye is modernizing its statistical infrastructure and has already established key register-based components through the Spatial Address Registration System (SARS). In recent years, SARS has been routinely used by municipalities in official administrative processes and by TurkStat in statistical production, indicating a level of operational maturity rather than an early-stage transition. The availability of SARS therefore provides a practical foundation for moving from traditional text-based enumeration frameworks toward spatially defined and more systematically managed EAs. In this study, SARS is used as a primary administrative source to support EA delineation. However, to fully capitalize on this opportunity, Türkiye must adopt methodological tools that ensure efficiency, accuracy, and scalability. The integration of ArcGIS Pro and the BBZ function directly addresses this need by allowing for automated delineation of compact and balanced EAs that align with demographic realities.

In practice, the methodology involves integrating demographic data, administrative boundaries, and other relevant spatial datasets within the ArcGIS environment. Base units, such as address points or census blocks, are aggregated into larger enumeration areas using BBZ. Criteria such as population size, area thresholds, and spatial contiguity are applied to ensure that the resulting units are both operationally feasible and analytically meaningful. For example, a target of 2,500 residents per EA might be established, with adjustments made to accommodate urban density patterns or rural sparsity. In addition, by setting parameters for compactness and contiguity, the tool avoids irregular or fragmented zones that could complicate field operations.

International best practices further illustrate the potential of this approach for Türkiye. In Estonia, adaptive EAs have been developed by integrating national registers with GIS, allowing real-time updates that reflect demographic changes (Statistics Estonia, 2021). In Somalia, high-resolution satellite imagery has been combined with automated delineation techniques to create grid-based EAs under severe resource constraints (Qader et al., 2019). These cases demonstrate that GIS-enabled automation is not only a luxury of developed countries but also an achievable innovation in developing and transitioning economies. By adopting BBZ within ArcGIS Pro, Türkiye positions itself alongside countries that are leveraging advanced spatial methodologies for statistical modernization.

The benefits of this methodological approach extend beyond census operations. Spatially balanced and well-defined EAs provide a foundation for representative household surveys, socio-economic studies, and administrative monitoring. They also facilitate integration with broader policy areas such as healthcare, education, infrastructure planning, and disaster risk management. In Türkiye, where rapid urbanization and demographic shifts create ongoing challenges for statistical agencies, an adaptive and GIS-based EA system ensures that statistical operations remain both reliable and relevant.

In conclusion, this methodology builds upon global experience with ArcGIS Pro and the BBZ function to address the challenges of EA delineation in Türkiye. By automating the creation of compact, contiguous, and demographically balanced zones, the approach enhances efficiency, reduces costs, and improves the accuracy of statistical outputs. Furthermore, it aligns Türkiye's statistical practices with international standards, ensuring interoperability with global data systems. Through the integration of demographic data, administrative registers, and spatial analysis tools, this study demonstrates how GIS technologies can serve as a catalyst for the modernization of national statistical infrastructures.

4. Application and Findings

This section presents the empirical applications of the proposed workflow and summarizes the outputs obtained across the case datasets. It first reports results from datasets from the United States, which illustrate how BBZ behaves across different input-variable distributions, dataset sizes, and levels of spatial contiguity. It then presents the main application conducted in a single quarter of Ankara(Türkiye),where multiple parameter configurations were tested using integrated administrative and spatial layers. The section documents the implemented thresholds, contiguity rules, and advanced parameter settings and reports the resulting EA configurations in terms of geographic coherence and statistical balance. Figures and maps are used throughout to demonstrate the input layers, intermediate outputs, and final zones produced in each case.

4.1. Application on the United States Datasets

4.1.1. Richmond County Data

The first dataset used for testing was obtained from the North Carolina Integrated Cadastral Data Exchange project. It includes parcel boundaries together with attributes such as ownership, parcel size, and assessed land value. Figure 2 illustrates the parcel–point structure.

.

Figure 6. View of the parcel and point data of the Richmond County

For this application, the variable 'Land Value' was selected as the balancing criterion using the BBZ tool. In this step, the tool was used to aggregate parcels into zones based on the distribution of land values. Table 2 presents the descriptive statistics of this variable.

Table 2. Summary statistics of the Land Value variable in the Richmond County data

Fieldname	Count	Min	Max	Mean	Median	Mode	Outliers	Iqr	Q1	Q3
LANDVAL	32518	0	30718148	33871	12800	15000	3353	19855	6037	25893

This dataset was primarily used to practice using the BBZ tool and to evaluate its performance on parcel-level cadastral data.

4.1.2. City of Seattle Data

The second dataset was obtained from the City of Seattle open data portal. It contains parcel boundaries with attributes including address, land use, and the number of housing units per parcel. Figure 3 illustrates the parcels and their housing unit data.

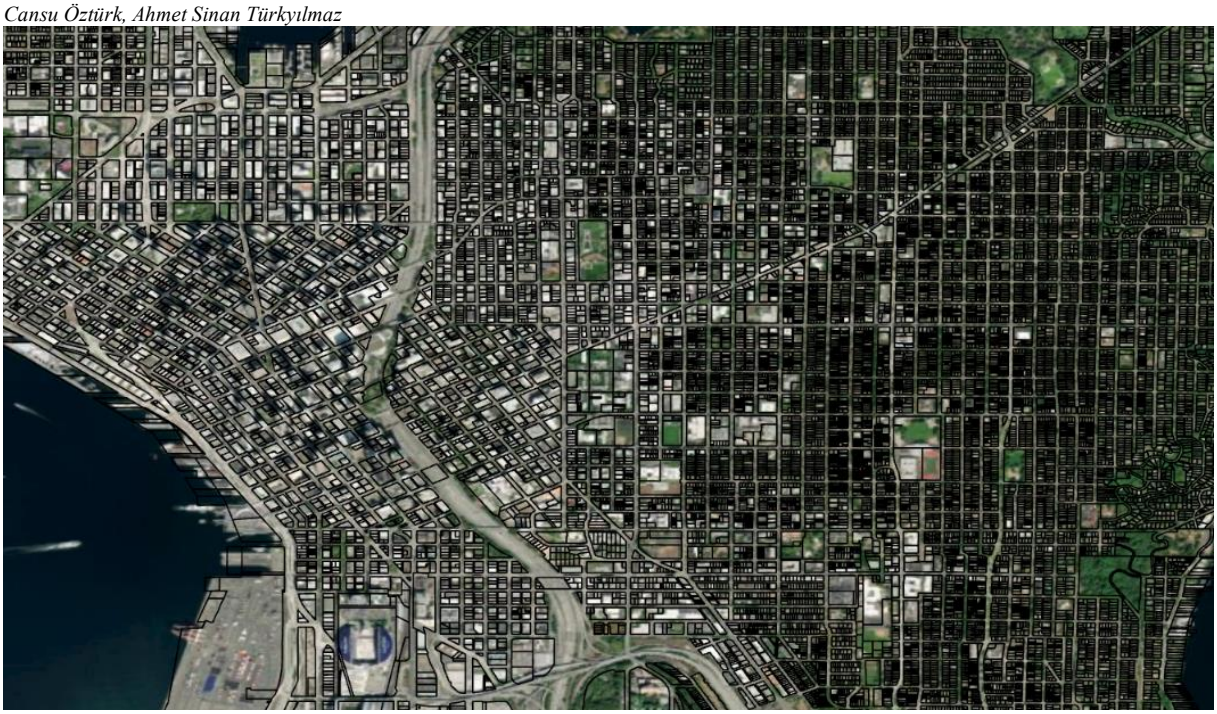


Figure 7. View of the parcel data of Seattle

In this application, the variable Number of Housing Units in the parcel (Exist_Unit) was used as the balancing criterion in the BBZ tool. The tool was applied to group parcels into zones so that each zone would have a comparable number of housing units. Table 3 provides descriptive statistics for the variable.

Table 3. Summary statistics of the Exist_Unit variable in the Seattle data

Fieldname	Count	Min	Max	Mean	Median	Outliers	Sum	Range	Iqr	Q1	Q3
EXIST_UNIT	178.897	0	707	1.8	1	35.485	330.527	707	0	1	1

This dataset was used to evaluate how the BBZ tool performs when applied to housing-related parcel attributes.

4.1.3. Story County, Iowa Data

The third dataset was obtained from publicly available cadastral records for Story County, Iowa. It comprises more than 43,000 parcels with attributes such as ownership, parcel identification numbers, assessed values, and land-use types. The spatial structure of the dataset is shown in Figures 4 and 5.

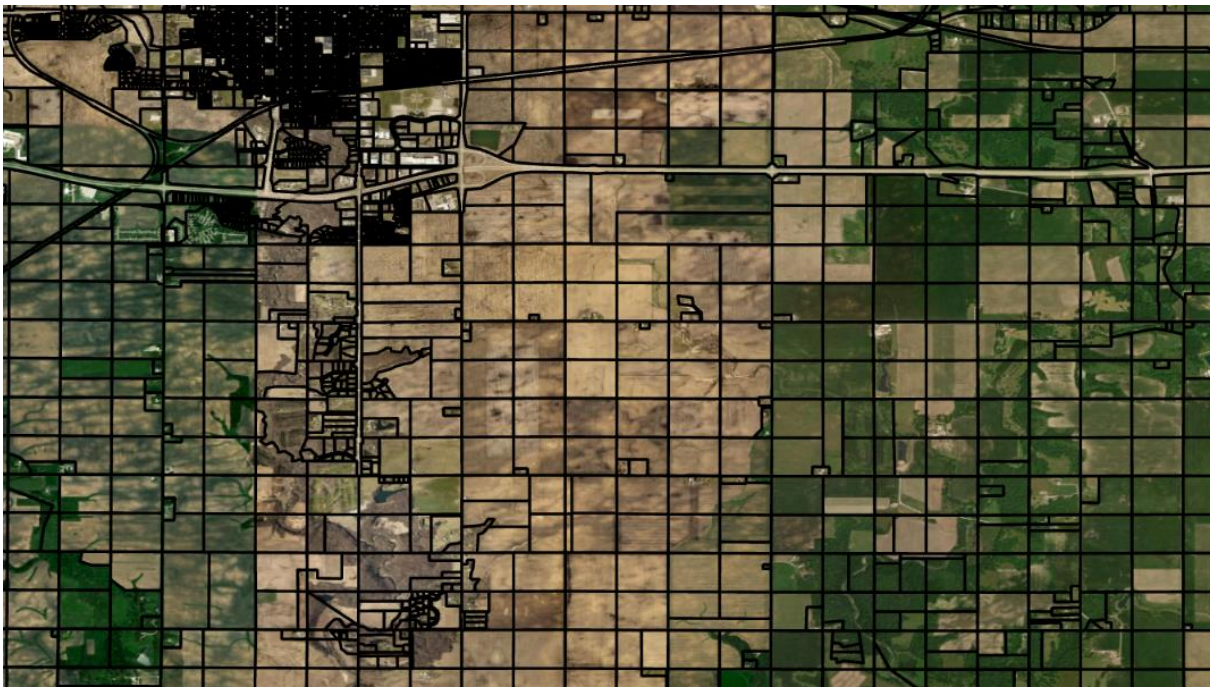


Figure 8. View-1 from Story County parcel data

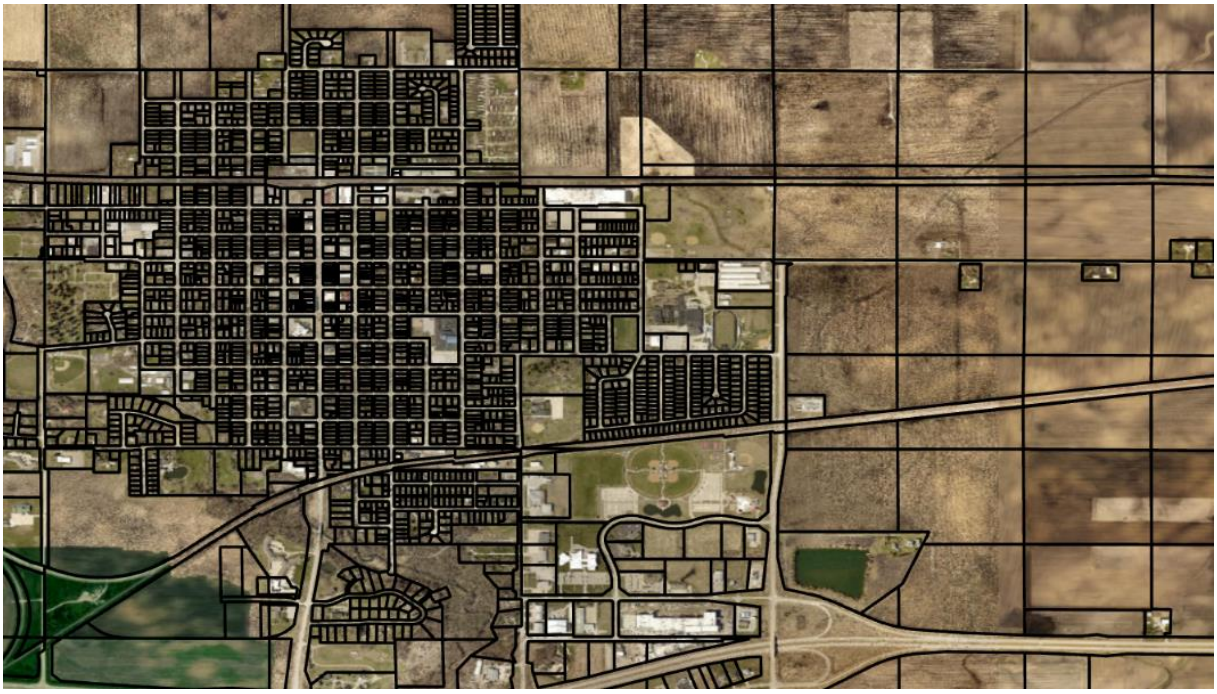


Figure 9. View-2 from Story County parcel data

For the application, the BBZ tool was run using parcel-level attributes as balancing variables. The "Homestead" and "FullDwelli" variables were used separately as target variables when creating enumeration areas from Iowa data. Thresholds were set for grouping

parcels into zones, and different parameters were tested to observe how the tool handles a dataset of this size. Table 4 summarizes the descriptive statistics of the selected variables

Table 4. Summary statistics of the Homestead and FullDwelli variables in the Story County data

Fieldname	Count	Min	Max	Mean	Median	Outliers	Sum	Iqr	Q1	Q3
FullDwelli	43.112	0	10.132.400	87.469	72.100	1.077	3.770.962.644	124.925	0	124.925
Homestead	43.112	0	366.686	1.824	0	42	78.644.608	4850	0	4.850

This dataset was primarily used to explore the use of BBZ with large-scale parcel data comprising many individual records.

4.2. Application on a Dataset from a Single Quarter in Ankara

The final dataset used in this study comes from Ankara (Türkiye). After sufficient experience had been gained from the United States cases, the analysis focused on the Ankara dataset as the primary application for Türkiye. The case study was conducted in a single quarter, consistent with TurkStat's practice of applying the blocking procedure within quarter and village boundaries. The selected quarter is a rapidly developing urban area that has expanded in recent years and is expected to continue to change in the near future. The name of the quarter is not disclosed due to confidentiality considerations, given the relatively small size of the study area.

Several spatial datasets were used in this application, including quarter boundaries, road centerlines, numbering data linking parcels to structures, building data containing information on structures and independent sections, and parcel boundaries. Figure 6 illustrates the study area and the integrated representation of these datasets. Because the datasets were obtained from different sources, minor boundary inconsistencies were observed.

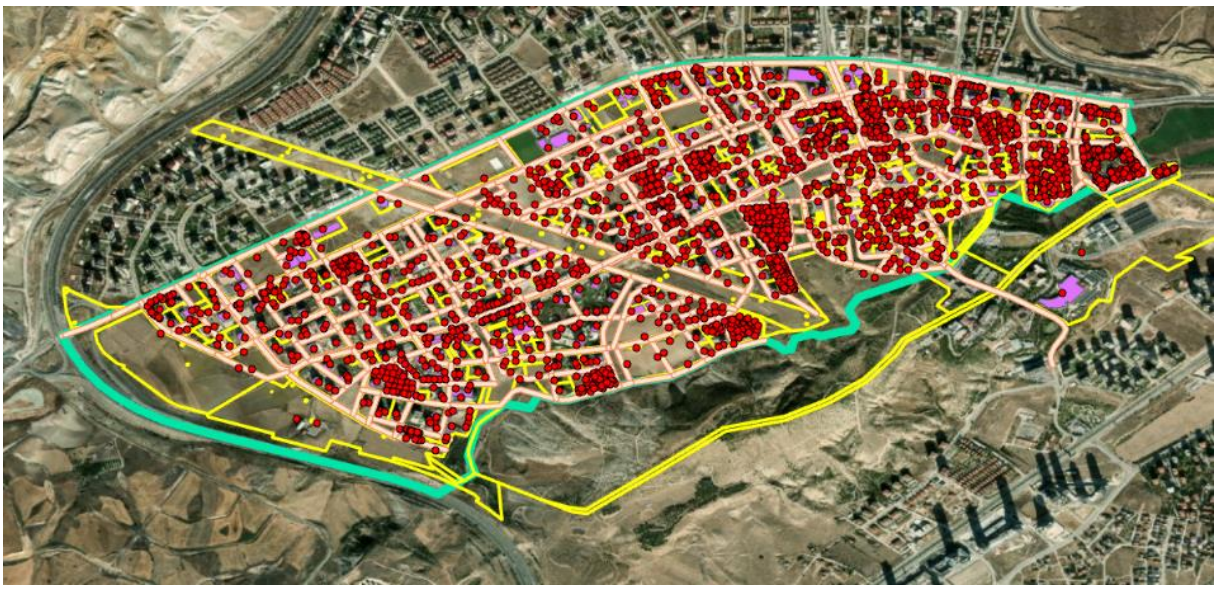


Figure 10. Spatial view of the quarter dataset

For the Ankara application, parcels were grouped into enumeration areas(EAs)using the ArcGIS Pro Build Balanced Zones(BBZ)tool. The primary balancing criterion was the Number of Independent Sections per parcel (total_bb), where independent sections consist of residential units(households), commercial units(workplaces),and public buildings. The total count (total_bb) was used as the main

target variable, while the counts of residential, commercial, and public units were also available at parcel level as separate attributes. To support replication and clarify the trial structure, the BBZ configurations tested in this study are summarized as follows. The baseline specification used total_bb as the target variable with a threshold of 120 units per zone. This threshold was informed by TurkStat's household survey practice, in which clusters are typically formed within quarter boundaries, with approximately 100 occupied household addresses and roughly equal cluster sizes. In addition to the baseline, an alternative specification was tested by including household counts as a target variable with a threshold of 100 per zone. Across trials, multiple combinations of thresholds, spatial contiguity settings, zone-shape preferences, and advanced optimization parameters were explored, together with optional distance-based constraints derived from quarter borders and the road network.

Tested BBZ parameter set:

(1) Input features: parcel polygons.

(2) Primary balancing variable (target attribute): total_bb (Number of Independent Sections per parcel).

(3) Target threshold(s): total_bb = 120 (baseline); household count = 100 when used as a secondary target.

(4) Spatial contiguity constraints (adjacency definition): Contiguity (Edges Only); Contiguity (Edges and Corners); and proximity-based adjacency where applicable (e.g., triangulation-based option), with additional constraints implemented using quarter borders and road network layers.

(5) Zone Characteristics (shape/structure preferences): Compactness; Equal Area; Equal Number of Features (tested as alternative settings).

(6) Advanced Parameters (optimization controls): Population Size, Number of Generations, and Mutation Factor (tested under multiple combinations to assess sensitivity and improve convergence).

(7) Distance option (when activated): Distance to Consider, using quarter borders and road centerlines as reference layers.

When enabled, the Distance to Consider option accounts for proximity to selected reference features. For each input parcel, it calculates the distance to the nearest feature in the specified reference layer and uses this distance as an additional constraint when selecting the final zone configuration. In this application, quarter boundaries and road centerlines served as reference layers for the distance parameter.

5. Results and Discussion

The results obtained from both the United States datasets and the Ankara application provide meaningful insights into BBZ's performance and the prospects for geospatial zoning within Türkiye’s statistical system. Each dataset highlighted unique aspects of the methodology, offering lessons on variable suitability, dataset size, parameter sensitivity, and operational feasibility.

The Richmond County dataset included parcel boundaries and land value attributes. Applying the BBZ tool to this dataset revealed the disruptive effects of extreme variable distributions (Figure 7). Some parcels had exceptionally high land values while others had almost none, preventing the algorithm from generating statistically and geographically balanced zones(Figure 8).This outcome highlighted the methodological challenge posed by highly skewed variables and emphasized the importance of selecting indicators that are both representative and statistically stable. In practice, census planners are advised to use variables that directly and consistently reflect population or household characteristics rather than those influenced by irregular market dynamics (Flowerdew & Feng, 2005).

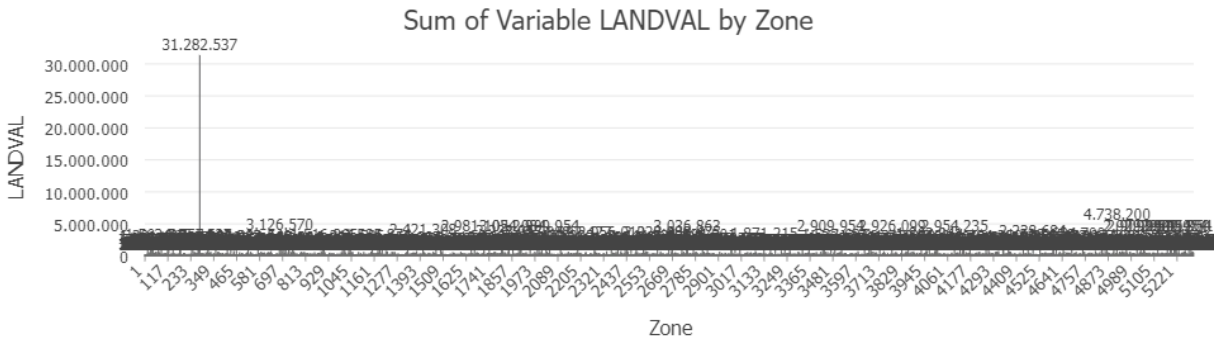


Figure 11. Sum of Land Value by Zone

Figure 12. The result of the BBZ for Richmond County data

The second dataset, from Seattle, used the number of housing units per parcel as the balancing variable. Since most parcels contained only one housing unit, variation was extremely low, limiting the BBZ tool's ability to redistribute features and achieve balance. The Seattle dataset revealed limitations caused by insufficient variability, resulting in zones of limited operational and analytical significance (Figure 10). Various alternatives were tested, but no meaningful results were obtained, as the values were clustered around 1. This finding underscored the importance of selecting input variables that

provide sufficient diversity for the algorithm to function effectively (Flowerdew & Feng, 2005). The issue is particularly relevant for Türkiye, where certain quarters may exhibit highly uniform housing patterns. Such conditions necessitate careful preprocessing or the use of alternative variables to ensure meaningful and interpretable outputs (United States Census Bureau, 2020).

The Story County case highlighted how dataset size affects the stability of algorithmic zoning. With over 43,000 parcels, the tool created thousands of fragmented and impractical zones (Figures 9 and 10). The operation of the BBZ tool was time-consuming and produced inconsistent results. Adjusting thresholds did not fully resolve the instability, reinforcing the idea that zoning algorithms require tailored parameterization for large-scale applications.

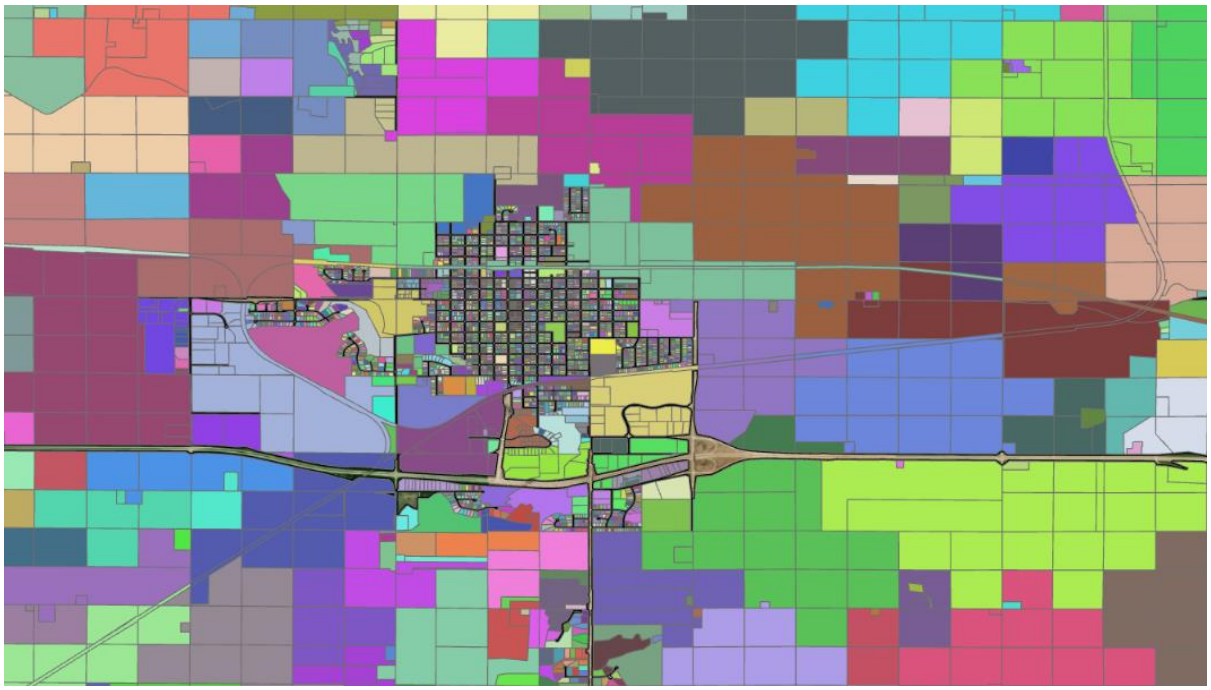


Figure 13. View-1 from BBZ result of Homestead variable

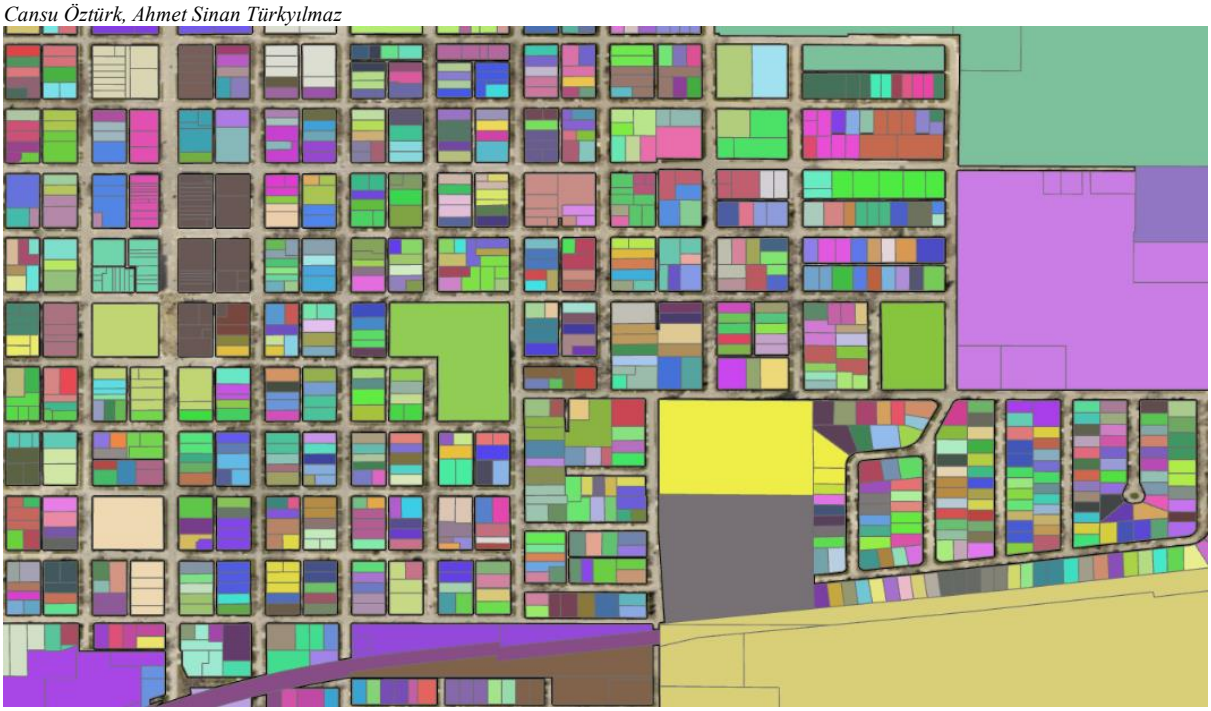


Figure 14. View-2 from BBZ result of Homestead

The influence of threshold selection was also evident in this case. When the threshold for FullDwelli was set to 100.000, approximately 20.000 zones were generated for 43.000 parcels, corresponding to nearly half of the total dataset. Increasing the threshold to 1,000,000 reduced the number of zones to about 3,000, which still represented a relatively high proportion (Figure 11).

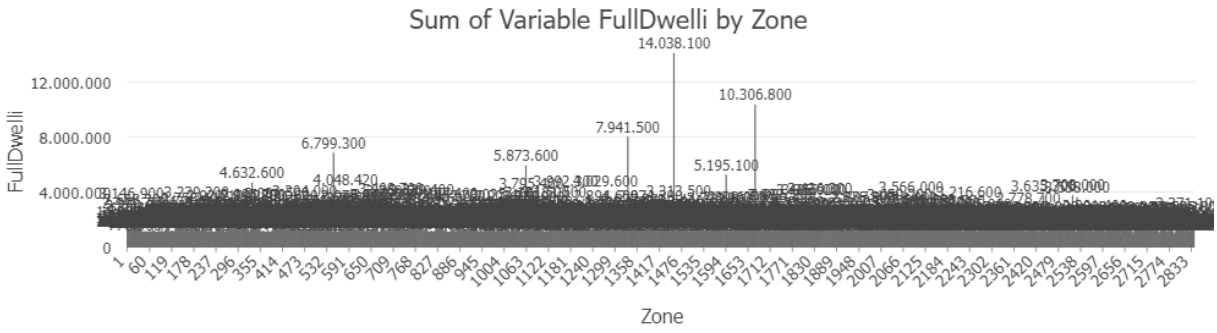


Figure 15. BBZ Result for FullDwelli when threshold is 1 million

These findings demonstrate that the configuration of threshold values critically affects computational performance and spatial aggregation, particularly when working with extensive parcel datasets. This observation aligns with broader international discussions on the Modifiable Areal Unit Problem (Openshaw, 1984).

The Quarter dataset was used to evaluate the applicability of the Build Balanced Zones tool in Türkiye's spatial and statistical context. During data preparation, minor inconsistencies were detected between different boundary and parcel layers, which created small alignment mismatches. These technical discrepancies highlighted common challenges in spatial data management and emphasized the importance of thorough data cleaning and integration before applying automated zoning tools.

Several spatial constraint configurations were tested to determine their influence on the zoning process. Using only contiguity edges produced compact, geometrically regular zones but offered limited stability in statistical balance. Incorporating both edges and corners increased connectivity between parcels, although this occasionally led to irregular shapes. The Trimmed Delaunay Triangulation option produced proximity-based zones, yet these were not always consistent with defined statistical targets.

The analysis of Zone Characteristics Criteria revealed the trade-offs between geometric fairness and statistical adequacy. The Compactness criterion generated visually regular, contiguous zones, but made balancing the independent sections more difficult. The Equal Area and Equal Number of Features settings provided slight improvements, but did not fully eliminate imbalances, reinforcing the need to balance simultaneously in EA design (Esri, 2023).

Among the Advanced Parameters, increasing the Mutation Factor resulted in the most visible improvements, producing outcomes closer to the desired threshold. Introducing household counts as a secondary balancing variable, with a threshold of 100, enhanced demographic alignment. Incorporating distance constraints based on quarter borders and road networks further improved spatial coherence, producing zones that corresponded more closely to recognizable urban features. This configuration improved the operational feasibility of EAs, enabling field staff to more easily identify and navigate them during data collection.

Overall, the Quarter case demonstrated that the BBZ tool can be effectively adapted to Turkish datasets, given adequate data quality and parameter optimization. The tool produced zones that were both geographically coherent and statistically balanced, confirming its potential for census and survey applications (Figure 12). However, even small inconsistencies in boundaries and spatial layers significantly affected outputs, emphasizing the necessity of standardized data preparation and harmonized spatial frameworks for national-scale implementation.

Figure 16. The final BBZ Result for Ankara data

The discussion of zone characteristics also confirmed the multi-objective nature of EA delineation. Compactness enhances geometric regularity but often conflicts with demographic balance,

whereas equal-area options ensure visual fairness but lack statistical adequacy. These findings highlight that EA delineation cannot rely on a single optimization criterion but must balance geography, demography, and operational manageability (Esri, 2023; United Nations, 2017).

From a methodological perspective, the Ankara study demonstrated that BBZ is a flexible and scalable tool for EA generation but remains sensitive to parameter settings such as Mutation Factor and Number of Generations. Hybrid approaches that combine BBZ with manual verification or supplementary geospatial tools could offer the most practical balance between automation and expert control.

From a policy standpoint, adopting geospatial methods could strengthen Türkiye's statistical infrastructure by enabling standardized, reproducible, and auditable delineation of primary sampling units and by reducing reliance on ad hoc manual boundary production. Automated delineation, as shown in the Ankara case, offers reproducibility, transparency, and scalability—qualities that align with modernization goals set by the United Nations Statistics Division (United Nations, 2017). Nevertheless, automated tools should complement, not replace, local expertise, as contextual knowledge remains essential for designing meaningful and functional enumeration areas.

In summary, the results highlight both the opportunities and the limitations of using geospatial methods for EA delineation. The United States datasets illustrated methodological risks linked to poor variable selection or dataset imbalance, while the Ankara application demonstrated that, with proper data and parameterization, the method can yield zones that are both statistically meaningful and operationally feasible. For Türkiye, the main challenges lie in harmonizing spatial datasets across institutions and establishing clear guidelines for parameter optimization. With these measures in place, geospatial zoning can significantly enhance the efficiency and quality of survey and census operations.

6. Conclusions

This study proposed a GIS-based approach for delineating Enumeration Areas (EAs) for statistical field operations in Türkiye through a trial implementation designed to test its feasibility under different data conditions. The empirical structure combined three case applications from the United States with one main application, conducted in a single quarter of Ankara (Türkiye). Rather than claiming nationwide applicability, the Ankara-quarter application demonstrated the workflow in a realistic local setting, while the United States cases illustrated how automated zone creation behaves under contrasting input-variable distributions and dataset scales. Taken together, the cases clarify the conditions under which GIS-based delineation can support geographically coherent and operationally workable zones, and they identify limitations that must be addressed before wider institutional implementation.

The case applications in Richmond County, Seattle, and Story County provided a diagnostic foundation for understanding methodological strengths and weaknesses. Across these cases, three practical "screening conditions" emerged that should be checked before applying automated delineation in routine production. First, extreme variation and skewness in balancing variables can dominate the optimization problem and contribute to fragmented or unstable outputs, indicating that selected variables should be both representative and distributionally stable for EA purposes. Second, insufficient variability can mechanically constrain the balancing process because the workflow has limited degrees

of freedom to redistribute features, thereby reducing the likelihood of operationally meaningful outputs. Third, very large datasets introduce computational and stability challenges: threshold choices can yield either an excessive number of zones or overly coarse aggregation, and long run times and fragmentation become more likely. These findings reinforce the importance of selecting suitable balancing variables, assessing distributional properties in advance, and ensuring that the intended scale of delineation matches the structure of the input data. They also are consistent with the Modifiable Areal Unit Problem, which emphasizes that scale choices and distributional properties can lead to substantially different spatial outputs (Openshaw, 1984; Flowerdew & Feng, 2005).

Building on these lessons, the Ankara quarter application illustrated how a GIS-based delineation workflow can be configured to reflect field-operational realities. By using independent sections per parcel and household counts as balancing variables and incorporating road networks and quarter borders as spatial constraints, the workflow supported the production of zones that were closer to intended workload targets and easier for field operations to interpret than zones produced by ad hoc manual delineation. This application, therefore, serves as an applied demonstration of how a Türkiye-relevant EA workflow can be parameterized around workload-oriented targets and recognizable geographic features.

At the same time, the Ankara application highlighted a set of critical contradictions that must be discussed with respect to both methodology and application. Inconsistencies between the Spatial Address Registration System and the General Directorate of Land Registry and Cadastre —particularly mismatches in quarter boundaries and parcel alignments—directly affected zone membership and adjacency. Methodologically, these cross-source inconsistencies challenge core assumptions of automated delineation, because the workflow depends on a coherent spatial topology in which parcels, boundaries, and barriers define stable neighbor relations and a consistent spatial support for the balancing variables. When boundary lines do not align across sources or when parcel or building geometries are shifted, the induced contiguity relationships may change, and the same parameter settings can yield different configurations depending on which layer is treated as authoritative. In practical terms, boundary mismatches can reassign parcels at the margins and alter zone totals, while parcel–building misalignment can weaken attribute linkage and introduce local distortions in balancing variables. Accordingly, evaluation in the Turkish context requires reporting not only zoning parameters but also the pre-delineation harmonization decisions and spatial validation steps that define the effective analytical support for the workflow.

The study further confirms that EA delineation is inherently multi-objective. Geometric preferences such as compactness or equal area can improve certain shape properties, but they must be balanced against demographic relevance and operational constraints. More workable configurations were obtained when demographic criteria were combined with geographic and field-operational constraints, consistent with international guidance emphasizing multi-criteria design for census geography(United Nations, 2017).In addition, the findings suggest that reproducibility and transparency are achievable only when workflows are standardized and data integration is treated as a prerequisite rather than a post-processing step(Cockings & Martin, 2005; United States Census Bureau, 2020).For this reason, a core practical output of the study is not only the zoning results identification of the need

for clear parameter reporting—at minimum, target variables, threshold values, contiguity rules, optional constraints, and optimization settings—to support replication and diagnosis across quarters and update cycles.

From a policy and implementation perspective, the results indicate that the potential benefits of GIS-based EA delineation—standardization, reproducibility, and operational manageability—depend on institutional coordination and data governance. Establishing shared standards for dataset integration, implementing routine cross-agency validation procedures, and formalizing cooperation mechanisms are central prerequisites for sustained adoption. Such steps are consistent with broader census modernization recommendations that emphasize integrated geospatial infrastructure and documented procedures for boundary management and fieldwork geographies (United Nations Statistics Division, 2017).

The main limitation of this study is its scope. The application in Ankara was conducted over a single quarter, and the United States cases were chosen to illustrate contrasting data patterns rather than to provide national benchmarks. Future work should, therefore, extend the workflow to multiple quarters with different urban forms, test stability under data updates, and evaluate scalability using larger official datasets. A hybrid production model—automated delineation followed by structured expert review—also appears essential for balancing efficiency with operational realism. Additional research may explore dynamic updating mechanisms and complementary data sources to improve adaptability in rapidly changing urban settings (Krebs & MacQueen, 2016).

Overall, the study supports the conclusion that GIS-based EA delineation is a feasible methodological direction for Türkiye when (i) balancing variables have sufficient variability and distributional stability, (ii) contiguity and constraint rules reflect field-operational requirements, and (iii) input layers are harmonized with adequate geometric consistency. Under these conditions, the approach offers a practical foundation for enhancing the production of clearly bounded, operationally workable small-area units for future census and household survey operations, explicitly acknowledging that methodological rigor and data governance are inseparable components of reliable automated delineation.

Note: This article is derived from the PhD thesis prepared by Cansu ÖZTÜRK at Hacettepe University Institute of Population Studies, Department of Social Research Methodology

References

ABS. (2016). *Census of Population and Housing: Mesh Block Counts, Australia, 2016*. Australian Bureau of Statistics. Retrieved from https://www.abs.gov.au

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Wiley.

BPS-Statistics Indonesia. (2018, July 12). BPS builds Wilkerstat with geospatial technology. Retrieved from https://www.bps.go.id/en/news/2018/07/12/201/bps-builds-wilkerstat-with-geospatial-technology.html

Chakravorty, S. (2007). Geography in India's census: A GIS-based approach (ESA/STAT/AC.115/17). United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems. Retrieved from https://unstats.un.org/unsd/demographic-social/meetings/2007/egm-census-mapping/docs/esa_stat_ac115_17.pdfCockings, S., & Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine, 60*(12), 2729–2742.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.). Wiley.

Esri. (n.d.). *How Build Balanced Zones works*. Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/learnmore-buildbalancedzones.htm

Esri. (2020). *ArcGIS Pro: Spatial Statistics Tools*. Redlands, CA: Environmental Systems Research Institute.

Esri. (2021). *ArcGIS Pro: Geostatistical Analyst*. Redlands, CA: Environmental Systems Research Institute.

Esri. (2023). *Build Balanced Zones (Spatial Statistics)*. ArcGIS Pro 3.2 Documentation. Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/build-balanced-zones.htm

Flowerdew, R., & Feng, Z. (2005). Enumeration districts, output areas and geographical information systems. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 168*(1), 49–67.

Goodchild, M. F., & Li, W. (2021). Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences, 118*(35), e2105274118.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Wiley.

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N. (2004). Small area estimation and the geography of poverty. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 167*(3), 319–341.

INSEE. (2019). *IRIS: Ilots Regroupés pour l'Information Statistique*. Institut National de la Statistique et des Études Économiques. Retrieved from https://www.insee.fr

Instituto Brasileiro de Geografia e Estatística (IBGE). (2022). Country report of Brazil: Twelfth session of the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM). United Nations.

Instituto Brasileiro de Geografia e Estatística (IBGE). (2024, August 22). IBGE releases coverage analysis of the Post-Enumeration Survey of the 2022 Population Census. IBGE News Agency.

Kalton, G. (2020). Developments in survey sampling. *Journal of Official Statistics, 36*(4), 855–879.

Kırlangıçoğlu, C. (2005). *A new census geography for Turkey using geographic information systems: A case study on Çankaya District, Ankara* (Master's thesis). Middle East Technical University, Ankara, Türkiye.

Kish, L. (1965). *Survey Sampling*. Wiley.

KNBS. (2019). *Kenya Population and Housing Census 2019: Enumeration Areas and Sampling Frame*. Kenya National Bureau of Statistics.

Krebs, C. J., & MacQueen, J. (2016). Spatial sampling. In *Ecological Methodology* (pp. 381–403). Harper & Row.

Laaribi, A., & Peters, A. (2019). *GIS and the sustainable development goals*. United Nations Economic and Social Council Report. United Nations.

Lohr, S. L. (2010). *Sampling: Design and Analysis* (2nd ed.). Brooks/Cole.

Martin, D. (2011). Geography for the 2011 Census in England and Wales. *Population Trends, 145*(1), 7–15.

Mokhele, T., Mutanga, O., & Ahmed, F. (2016). Development of census output areas with AZTool in South Africa. South African Journal of Science, 112(7/8). https://doi.org/10.17159/sajs.2016/20150010Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Geo Books.

ONS. (2020). *Output Area Boundaries (2020) for England and Wales*. Office for National Statistics. Retrieved from https://www.ons.gov.uk

Qader, S. H., Dierkes, C., & Schneider, C. (2019a). Integrating remote sensing and GIS for delineating urban areas. *Journal of Geospatial Information Science, 22*(2), 45–60.

Qader, S. H., Dierkes, C., & Schneider, C. (2019b). Remote sensing-based assessment of land cover change in Baghdad. *International Journal of Remote Sensing, 40*(3), 1120–1145.

PARIS21, & Statistics Sweden. (2021). Guide on geospatial data integration in official statistics. Partnership in Statistics for Development in the 21st Century (PARIS21).

Pedro, A. A., & Queiroz, A. P. (2019). Slum: Comparing municipal and census basemaps. Habitat International, 83, 30–40. https://doi.org/10.1016/j.habitatint.2018.11.001

Statistics Canada. (2016). *Census Dictionary: Census Year 2016*. Statistics Canada. Retrieved from https://www.statcan.gc.ca

Statistics South Africa. (n.d.). Small area statistics. Retrieved from https://www.statssa.gov.za/?page_id=4086UNECE. (2015). *Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing*. United Nations Economic Commission for Europe.

United Nations, Department of Economic and Social Affairs, Statistics Division. (2008). Handbook on geographic databases and census mapping (Draft, 7 March 2008). United Nations.

United Nations. (2009). *Handbook on Geospatial Infrastructure in Support of Census Activities*. New York: United Nations Statistics Division.

United Nations. (2017). *Principles and Recommendations for Population and Housing Censuses* (Rev. 3). New York: United Nations Statistics Division.

United Nations. (2021). Handbook on the management of population and housing censuses, Revision 2 (Studies in Methods, Series F No. 83/Rev.2). New York, NY: United Nations.

United States Census Bureau. (2019). *Geographic Areas Reference Manual*. Washington, DC: U.S. Government Printing Office.

United States Census Bureau. (2020). *Census Enumeration Areas and Geographic Concepts*. Washington, DC: U.S. Government Printing Office.

United States Census Bureau. (2021). *Geographic Areas Reference Manual Update*. Washington, DC: U.S. Government Printing Office.

Wanyoike, H. (2023). Use of geospatial information in census-taking: 2019 Census Kenya experience. United Nations Statistics Division Expert Group Meeting on the Use of Geospatial Information in Census-taking. Retrieved from https://unstats.un.org/unsd/demographic-social/meetings/2023/egm-20230523/docs/s03-03-KEN.pdf

Worldometer. (2025). Indonesia population (2025). Retrieved from https://www.worldometers.info/world-population/indonesia-population/