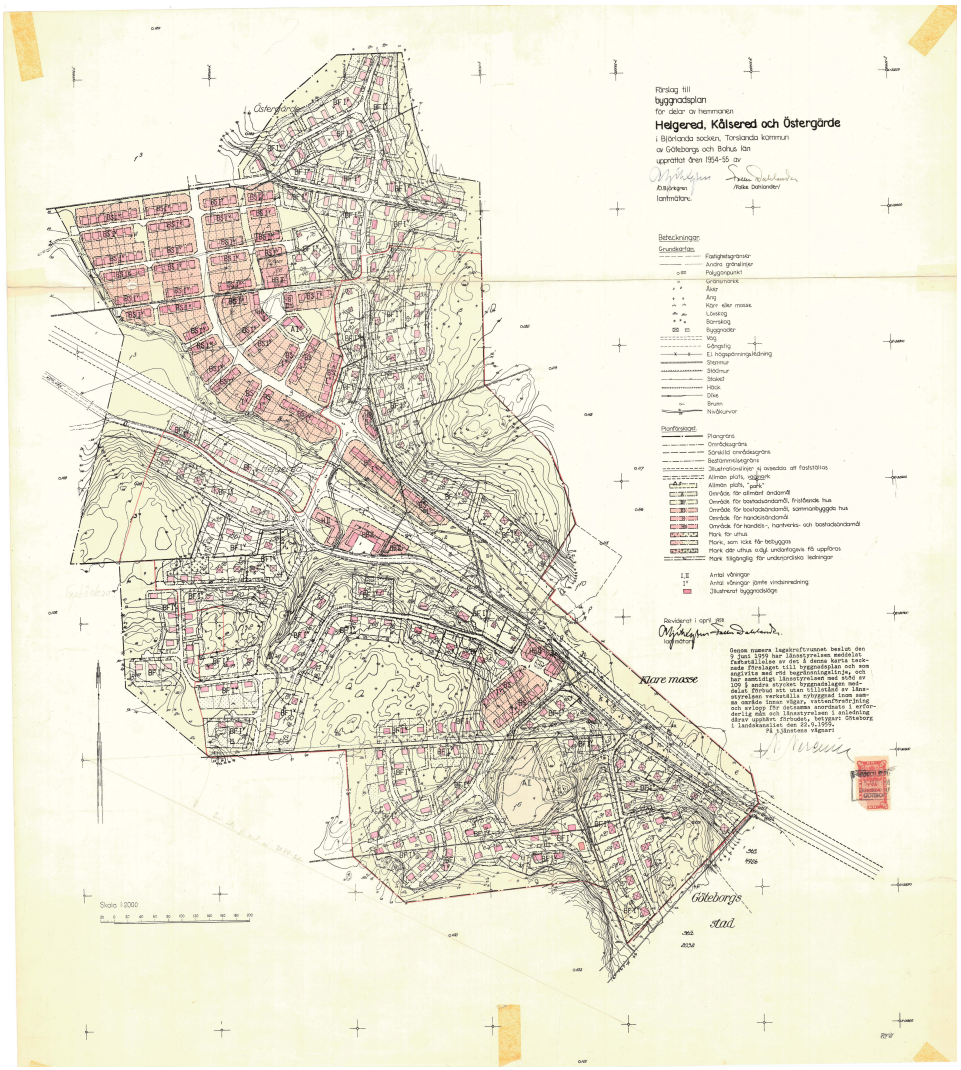
CHALMERS
UNIVERSITY OF TECHNOLOGY



Assessment of Zoning Plan Metadata using AI

Automating the Assessment of Land Parcels by Leveraging Large Language Models

Master's thesis in the Master's Programme Data Science and AI

ISAC MJÖRNELL

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING
DIVISION OF BUILDING TECHNOLOGY

CHALMERS UNIVERSITY OF TECHNOLOGY
MASTER'S THESIS ACEX30
GOTHENBURG, SWEDEN 2025

MASTER'S THESIS ACEX30

Assessment of Zoning Plan Metadata using AI

Automating the Assessment of Land Parcels by Leveraging Large Language Models

Master's Thesis in the Master's Programme Data Science and AI

ISAC MJÖRNELL

Department of Architecture and Civil Engineering
Division of Building Technology
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025

Assessment of Zoning Plan Metadata using AI
Automating the Assessment of Land Parcels by Leveraging Large Language Models
Master's Thesis in the Master's Programme Data Science and AI

ISAC MJÖRNELL

Examensarbete ACEX30
Institutionen för Arkitektur och Samhällsbyggnadsteknik
Chalmers Tekniska Högskola, 2025

Department of Architecture and Civil Engineering
Division of Building Technology
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 31 772 1000

Department of Architecture and Civil Engineering
Göteborg, Sweden, 2025

Assessment of Zoning Plan Metadata using AI
Automating the Assessment of Land Parcels by Leveraging Large Language Models

Master's thesis in the Master's Programme Data Science and AI

ISAC MJÖRNELL

Department of Architecture and Civil Engineering
Division of Building Technology
Chalmers University of Technology

ABSTRACT

Zoning plans contain critical information about land use, building rights, and development potential, but analyzing these documents manually is time-consuming and often impractical at scale. This thesis investigates how automation and large language models (LLMs) can be used to extract meaningful data from Swedish zoning plan (*detaljplan*) metadata to support early-stage site selection and investment decisions. By combining spatial data analysis and economic indicators, the developed methods enables identification of underutilized parcels with high development potential.

The study compares a regular expression keyword matching algorithm with Google's LLM *Gemini* or interpreting regulations and extracting key parameters such as maximum building height and area utilization. Results show that the LLM achieves better overall performance and is consistently interpreting irregularly phrased regulations that traditional methods overlook.

A tool has been created that integrates market data from recent real estate transactions to estimate local attractiveness, offering an economic dimension to the evaluation. Case studies from Järfälla, Halmstad, and Sundsvall demonstrate how the tool can improve the efficiency of zoning plan analysis, reduce manual workload, and provide a foundation for evaluating development sites.

Key words:
Urban Planning
Large Language Models
Natural Language Processing
AI
Zoning Plan
Nationella Geodataplattformen
Digital Twin City Centre
UI

Utvärdering av detaljplaners metadata med hjälp av AI
Automatisering av utvärderandet av outnyttjad mark genom utnyttjande av stora språkmodeller

Examensarbete inom masterprogrammet Data Science and AI

ISAC MJÖRNELL

Institutionen för Arkitektur och Samhällsbyggnadsteknik
Avdelningen för Byggnadsteknik
Chalmers Tekniska Högskola

SAMMANFATTNING

Detaljplaner innehåller stora mängder information om tillåten markanvändning och byggrätter, men analysen av dessa dokument är ofta tidskrävande och kräver manuellt arbete. Detta arbete undersöker hur automatiserad tolkning av metadata från detaljplaner kan stödja tidiga skeden i stadsutveckling, särskilt när det gäller att identifiera exploaterbara markområden och ge ett beslutsunderlag för potentiella investeringar.

Arbetet bygger på en verktygsprototyp som integrerar flera datakällor, inklusive digitala detaljplaner via Lantmäteriets Nationella geodataplattform, information om befintlig bebyggelse samt marknadsdata om bostadspriser. En central komponent i analysen är användningen av stora språkmodeller (Large Language Models), som används för att tolka planbestämmelser i detaljplaner och extrahera nyckelvärden såsom bruttoarea (BTA). Detta jämförs med mer traditionella metoder som semantisk sökning och regelbaserade nyckelordsmatchningar.

Fallstudier i Järfälla, Halmstad och Sundsvall visar hur verktyget kan effektivisera urvalsprocessen av planer och minska den manuella arbetsinsatsen. Resultaten tyder också på att stora språkmodeller erbjuder en högre träffsäkerhet vid tolkning av fritextregler, vilket ökar precisionen i analysen.

Sammantaget visar studien att automatisering med hjälp av tillgänglig planeringsdata och moderna AI-modeller kan ge ett värdefullt beslutsunderlag för aktörer inom stadsutveckling, särskilt i tidiga skeden då många potentiella projekt behöver jämföras snabbt och på ett strukturerat sätt.

Nyckelord:
Stadsplanering
Large Language Models
Natural Language Processing
AI
Detaljplan
Nationella Geodataplattformen
Digital Twin City Centre
UI

Contents

ABSTRACT I

SAMMANFATTNING II

CONTENTS IV

PREFACE VI

NOTATIONS VIII

1 INTRODUCTION 1
- 1.1 Background 1
- 1.2 Previous research 2
- 1.3 Research questions 3
- 1.4 Aim 3
- 1.5 Target user 4
  - 1.5.1 Limitations 5
  - 1.5.2 Demarcations 5

2 THEORY 6
- 2.1 Zoning plans 6
- 2.2 Geographic data 9
- 2.3 Large language models 10

3 METHODS 12
- 3.1 Methods for meeting the tool's requirements 12
- 3.2 Fetching of data 16
  - 3.2.1 Zoning plan data 16
  - 3.2.2 Buildings in Sweden 16
  - 3.2.3 Pointcloud of areas in Sweden 17
  - 3.2.4 Market price of real estates 17
- 3.3 Analysis of metadata attached to the zoning plan 17
- 3.4 Data description 18
- 3.5 Leveraging LLMs on the metadata attached to zoning plans 20
- 3.6 Extracting currently built GFA 21
- 3.7 Presentation and visualization of the results 21
- 3.8 Market value estimation 23
- 3.9 Case studies 25
  - 3.9.1 Järfälla 25
  - 3.9.2 Halmstad 26
  - 3.9.3 Sundsvall 26

4 RESULTS 27
- 4.1 The tool 27
- 4.2 Case studies 28
  - 4.2.1 Järfälla 29

4.2.2 Halmstad 30
4.2.3 Sundsvall 31
4.3 Examples of regulations 31
4.4 Accuracy metrics 34
4.5 Discussion 34
4.5.1 Data quality 34
4.5.2 Tool evaluation 36
4.5.3 Tool efficiency 37

5 CONCLUSION 38
5.1 Further research 38
5.1.1 Terrain analysis 38
5.1.2 Infrastructure 39
5.1.3 Proximity analysis 39
5.1.4 Alignment to overview plan 39
5.1.5 Demography 39

6 REFERENCES 40

Preface

This master's thesis has been carried out during the spring of 2025 as part of the Master's Programme Data Science and AI, in the Department of Architecture and Civil Engineering at Chalmers University of Technology. The work was conducted in collaboration with a private architecture firm Arkitekterna Krook & Tjäder, whose interest in digital methods for early-stage site evaluation inspired much of the focus of this thesis.

The project explores how zoning plan metadata can be analyzed with the help of automation and large language models, to support early decision-making in urban development processes. The work combines planning law, geospatial data, and natural language processing, and aims to address practical challenges faced by professionals working with large volumes of planning documents.

I would like to express my deepest gratitude to my supervisors at Chalmers, Sanjay Somanath and Yinan Yu, as well as Omar Zalloum from Krook & Tjäder, for their continuous support, valuable feedback, and critical questions throughout the process. I am also thankful to Krook & Tjäder for their interest, insights, and encouragement. Special thanks go to my colleagues, friends and family who provided useful discussions and moral support along the way.

Finally, I want to acknowledge the various public agencies and organizations whose open data and documentation have made this work possible. In particular, I would like to thank Lantmäteriet for providing access to essential geospatial and cadastral data, and Booli for offering structured information on real estate transactions, which was instrumental in evaluating market conditions for development sites.

Gothenburg, June 2025

Isac Mjörnell

Notations

LLM	Large Language Model
NLP	Natural Language Processing
UI	User-Interface
GFA	Gross Floor Area
BYA	Building Footprint Area
NGP	National Geodata Platform
DTCC	Digital Twin City Centre
CAD	Computer-Aided Design
PDF	Portable Document Format
API	Application Programming Interface
JSON	JavaScript Object Notation
HTML	HyperText Markup Language
KPI	Key Performance Indicator

1 Introduction

1.1 Background

In Sweden there are tens of thousands of zoning plans, the majority of which have already been exploited. In this thesis, the term zoning plan refers to the Swedish term *detaljplan*, a legally binding planning document that regulates land use and building rights within a defined area. Zoning plans are commonly developed by municipalities, sometimes in collaboration with private entities, especially when new development initiatives are being considered. In other instances, municipalities independently generate zoning plans for parcels they deem to hold development potential. Notably, not all initiated plans reach the execution phase, some remain unbuilt or partially realized. Therefore, even solely identifying unexploited zoning plans is a valuable step in finding land for new projects. Additional analyses such as current exploitation rates and available floor area are important but can typically be conducted manually, which is feasible as long as the pool of potential projects is not overwhelmingly large. Although methods for analyzing zoning plan data exist, they are typically only feasible for a limited number of plans due to the manual effort involved.

Zoning plans have long played a central role in urban planning by regulating the permissible uses of land parcels. These legally binding documents provide strong preconditions for obtaining a building permit, as long as the proposed development complies with the regulations outlined in the plan. The legal foundation for this system is defined in the "Plan- och bygglag (2010:900)" (2010), which governs how land and water areas in Sweden are to be used and developed. In recent years, substantial efforts have been directed towards digitizing these zoning plans, specifically by converting them into geometric vector formats. This transformation enhances their accessibility for automated analyses and enables a range of applications that were previously impractical or impossible.

Traditionally, zoning plans have been created either by hand or with the aid of software tools such as CAD. Many of these plans are now available in digital format, often as scanned or digitally produced PDFs. In Sweden, a considerable number of such documents are publicly accessible through *Lantmäteriet*, which is the Swedish national mapping, cadastral, and land registration authority responsible for managing geographic and property-related information across the country. However, the majority of these plans have yet to be transformed into fully digitized, vectorized formats enriched with metadata. The digitization process in this context, that are currently ongoing, refers specifically to zoning plans that are vector-based and linked to structured metadata, accessible via the API of the *National geodata platform*, (NGP).

Zoning plans contain vast amounts of information, but sifting through these documents is cumbersome and time-consuming. Analyzing a single zoning plan can often be done in a relatively short amount of time, but in order to understand the broader development within a city, one typically needs to review hundreds or thousands of plans. This cumulative workload quickly becomes a bottleneck when spatial metadata needs to be interpreted manually. Today there is a demand from developers, urban planners, and architects for efficient methods to filter large volumes of zoning plans to discover underutilized land parcels, that are suitable for development. Additionally,

providing key financial metrics, such as potential gross floor area and revenue estimates, can offer substantial value to stakeholders by facilitating informed investment decisions. Such metrics help guide early feasibility assessments, support prioritization between sites and provide a common language for discussing opportunities between financiers and municipalities, which ultimately reduce uncertainty and accelerate the planning process.

Over several decades, a substantial amount of data has been amassed in Sweden regarding zoning plans and the development of the built environment. One of the critical challenges lies in effectively analyzing the metadata associated with these plans. Of particular importance are the gross floor area and the current built environment on a given site, both of which are essential in evaluating a site's development potential. Approximately half of the zoning plans have already put the value and units of the plan regulations in the metadata, which makes automatic analysis of the metadata possible at a large scale. The data quality differ for different municipalities in Sweden, so three cities are analyzed for reference. These cities are Järfälla, Halmstad and Sundsvall. These cities were selected because they collectively offer a diverse representation of data quality, area size and geographical spread. Extracting the unit and value from zoning plan regulations is the most important and difficult challenge of this project. With these, key performance indicators (KPIs) can be automatically calculated for each zoning plan. The KPIs of most interest are the allowed gross floor area, current utilization within the zoning plan, and market estimation for the economic attractiveness of the neighborhood.

One of the most crucial aspects is to assess the economic dynamics of neighborhoods. Understanding how much individuals are willing to pay to reside in a specific area is vital before committing to a development project. This type of market analysis is essential not only for the real estate sector but for any domain involving high-stakes investments. There are numerous soft qualities to consider when evaluating a residual site such as proximity to parks and green spaces, sea view, schools, availability of public transport and other public services. These qualitative aspects are difficult to automatically estimate due to their variability across contexts. One promising approach to estimate this is to evaluate market interest by examining the volume and selling price of recently sold residential properties in a given neighborhood. By calculating the price per square meter, it becomes possible to estimate not only local demand but also the area's market valuation. Housing prices within a single city can vary significantly, even between adjacent neighborhoods. Several variables influence where people choose to live, ranging from proximity to societal functions to social features like street vibe and closeness to city center. Cultural and historical attributes also contribute to an area's attractiveness. These multifaceted considerations further underscore the importance of integrating both automated and manual analysis in site selection.

1.2 Previous research

In the United States large language models (LLMs) have been successfully used to interpret and extract values from textual zoning regulations. Since the structure and wording of zoning plan metadata can vary significantly over time and between municipalities, LLMs help extract key parameters such as permitted building height or footprint area, even when these are expressed ambiguously in free-form text. This

extraction enables a more consistent and scalable analysis across a diverse set of plans.

For example Axelrod et al. (2023) explored the use of natural language processing (NLP) and machine learning to automate the collection of zoning information from municipal documents in the United States. The authors propose a semi-automated pipeline for extracting, classifying, and standardizing zoning data across jurisdictions, emphasizing the complexity introduced by inconsistent terminology and document structure. Their conclusion highlights the potential of hybrid models which combine automated methods with human oversight. This problem is not present in Sweden at the moment since zoning plans are digitally available to a much broader extent.

Another study by Salazar-Miranda and Talen (2025) investigated the adoption of form-based codes (FBCs) in American municipalities and assesses their prevalence using NLP techniques. Rather than relying solely on explicit labels, the authors use semantic text analysis to detect regulatory patterns indicative of form-based zoning. The relevance to this thesis lies in the methodological parallel: using language models to infer zoning characteristics that are not consistently structured. Their work supports the scalability and effectiveness of NLP-based methods in analyzing zoning regulations across diverse planning contexts, but the implementation of a tool is not explicitly mentioned.

Zheng et al. (2025) introduces a benchmark framework for evaluating large language models in urban planning tasks, combining semantic text data with geospatial features. It emphasizes the importance of domain-specific datasets and evaluation metrics for assessing LLM performance in planning-related applications. This paper is particularly relevant as it illustrates the growing need for structured benchmarking when applying AI to urban planning, and supports the use of LLMs for interpreting complex, unstructured planning regulations.

However, there is no present research on how to scale these types of analyses. Many previous attempts are working with data interpretation, rather than making use of the data for taking more informed decisions. An advantage with the data from Lantmäteriet is that the majority is already categorized which makes it possible to take this to the next step to make key performance indicators for exploitable zoning plans.

1.3 Research questions

The overarching research question is if automated analysis of zoning plan metadata can support early-stage site selection and investment decisions in urban development. This is broken divided into two sub-questions.

- RQ1: Can automation reduce the time and labor typically involved in zoning plan evaluation?
- RQ2: How effective is LLM-based analyses compared to traditional semantic keyword matching at extracting key performance indicators from regulations in zoning plans?

1.4 Aim

The aim is to develop a tool designed to automate the analysis of zoning plan metadata. The tool is based on analyses on parameters such as existing land utilization, the potential gross floor area (GFA), and the overall market attractiveness of the site.

These factors are crucial for assessing the viability of proposed developments.

The primary goal of the project is to develop methods that allows users to define a geographic area within Sweden and automatically analyze all zoning plans located within that area. While the analysis does not need to be perfectly accurate, it must be sufficiently reliable to support effective sorting and filtering. The filtering should involve key performance indicators that will give an estimation of the potential for exploitation on the land parcel. This will enable users to identify a handful of zoning plans that are most likely to meet their project criteria. The secondary aim is to integrate economic indicators, to provide users with a solid foundation for making informed, strategic investment decisions.

1.5 Target user

The primary user group of the tool consists of property developers. For these professionals, timely identification of suitable properties is essential. The current practice involves significant manual effort to locate and interpret zoning plan metadata, often resulting in missed opportunities due to the sheer volume of data that needs to be assessed. Of particular interest are zoning plans that remain largely unexploited, offer generous allowances for development, and are situated in high-demand areas.

The secondary user group includes architects who are searching for parcels of land suitable for new construction. Their needs center around the ability to quickly and accurately filter zoning plans to locate sites that have not yet been fully exploited. Ideal candidates for development may be sites that are only partially built or areas where the current built environment utilizes only a small fraction of the permitted capacity as specified in the zoning plan. Architects must consider numerous additional factors when evaluating a potential building site, including topography, economic conditions, infrastructure, and the surrounding urban context. Each municipality also maintains an overarching plan that outlines strategic goals for future development, which can signal whether a specific site is implicitly encouraged for exploitation. Ideally, this comprehensive evaluation is done manually. However, applying such scrutiny across thousands of zoning plans is infeasible. This is where the tool's filtering capabilities become invaluable.

By automating the filtering process, the tool can scan large portions of Sweden to identify under-exploited land parcels that fall within existing zoning plans. The accompanying metadata allows users to distill these findings down to a handful of promising candidates, which can then undergo deeper manual analysis. This second-level review includes qualitative considerations such as terrain, infrastructure, local context, and alignment with municipal planning strategies. If a suitable zoning plan is found, a direct link to the legally binding document is provided, enabling users to verify the information and proceed to contact the property owner or a potential investor.

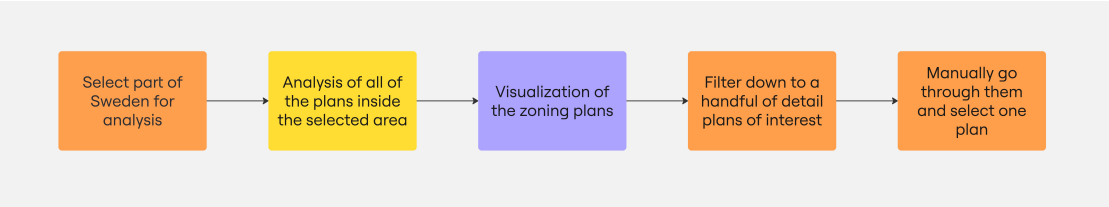


Figure 1.1: Overview of the analysis workflow implemented in the tool.

The thesis begins with an introduction, followed by a presentation of the theoretical background and the methodological framework. These methods are then applied to three case studies on the cities Järfälla, Halmstad, and Sundsvall. The case studies are done to illustrate how the approach functions in practice. The outcomes from these case studies are subsequently analyzed and discussed in relation to the research questions.

1.5.1 Limitations

Tool efficiency is a significant concern. If the area analysis process takes too long, the tool's utility diminishes. One of the primary computational bottlenecks is the retrieval and processing of data related to existing buildings, particularly their heights. Currently, the most efficient method involves accessing a point cloud for the area surrounding each building and identifying the highest point within the building's footprint. This method, however, requires downloading large volumes of data from a database, especially when boundary of the building footprints are spread across extensive geographic areas.

Another limitation lies in the historical variability of zoning plan formats. In Sweden, some legally binding plans date back over a century. Over the years, standards and visual conventions for drafting these documents have evolved significantly, leading to inconsistencies in their structure and appearance. As a result, it is difficult to develop a single method capable of reliably analyzing all types of plans. To ensure accuracy and feasibility, this project limits its scope to vectorized zoning plans that are available in a digitized format, containing metadata.

1.5.2 Demarcations

Municipalities exhibit considerable variation in how they draw and publish zoning plans, where the older plans are scanned and uploaded in PDF format. Differences in PDF rotation, orientation, scale, and quality further complicate automated processing. Analyzing such documents would require image preprocessing, which is beyond the scope of this project.

Finally, the ability to scale the analysis is not merely a technical preference, it is essential for the tool's practical value. It should be possible to analyze an entire municipality or city within a couple of hours. Ensuring that the sequence of methods can handle such up-scaling efficiently is therefore a central requirement of the project.

2 Theory

2.1 Zoning plans

A zoning plan is a legally binding document that governs the use and development of land and water areas within a specified area of a municipality. It regulates not only the permitted land uses but also the design, size, and placement of buildings. Zoning plans form the foundational legal framework for land development in Sweden and are essential tools in municipal urban planning.

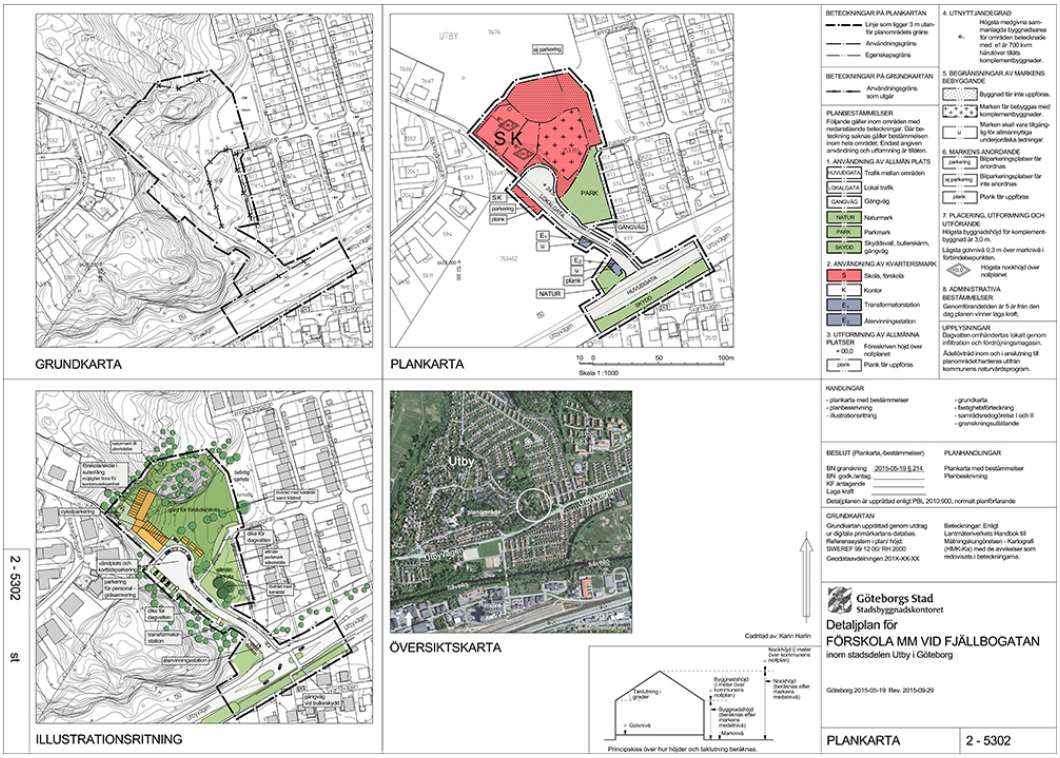


Figure 2.1: Example of a Swedish zoning plan and its components. Source: Göteborgs Stad, *Hur du läser och förstår en detaljplan*

Zoning plans have been employed for a long time to regulate land use across Sweden. It is the responsibility of each municipality to produce and maintain these plans, sometimes in collaboration with private entities. Zoning plans are occasionally revised to align with a municipality's updated development goals and vision for future land use. However, many plans have remained unchanged for decades, with some even dating back over a hundred years. As a result, there is significant variation between historical and contemporary zoning plans. Older plans were drawn by hand and their visual style, symbols and formatting differ considerably from modern standards.

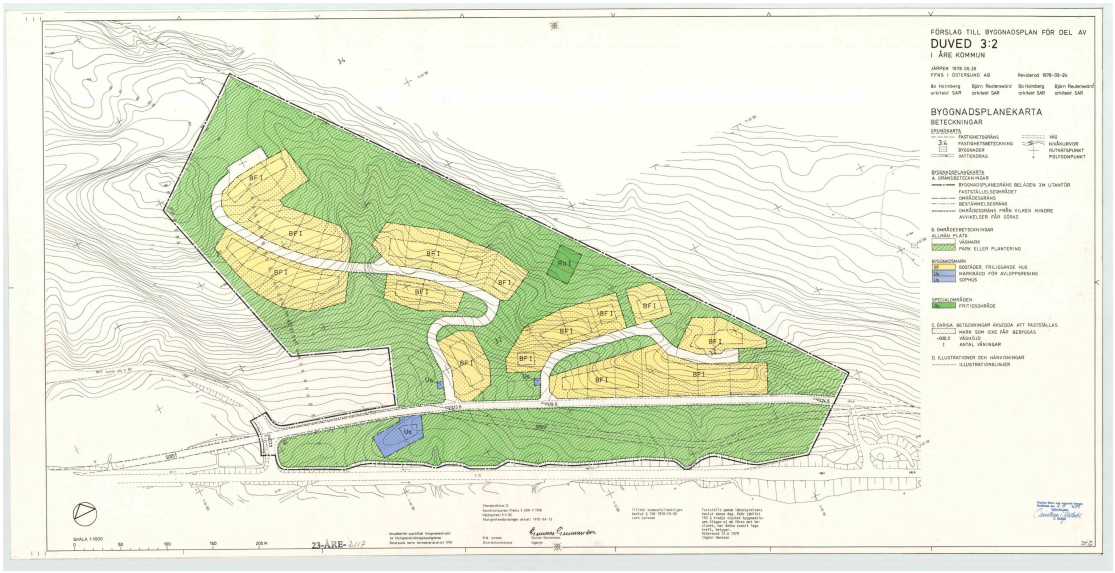


Figure 2.2: Example of a zoning plan from the 1970s. The plan is from *Duved*, in the municipality of Åre.

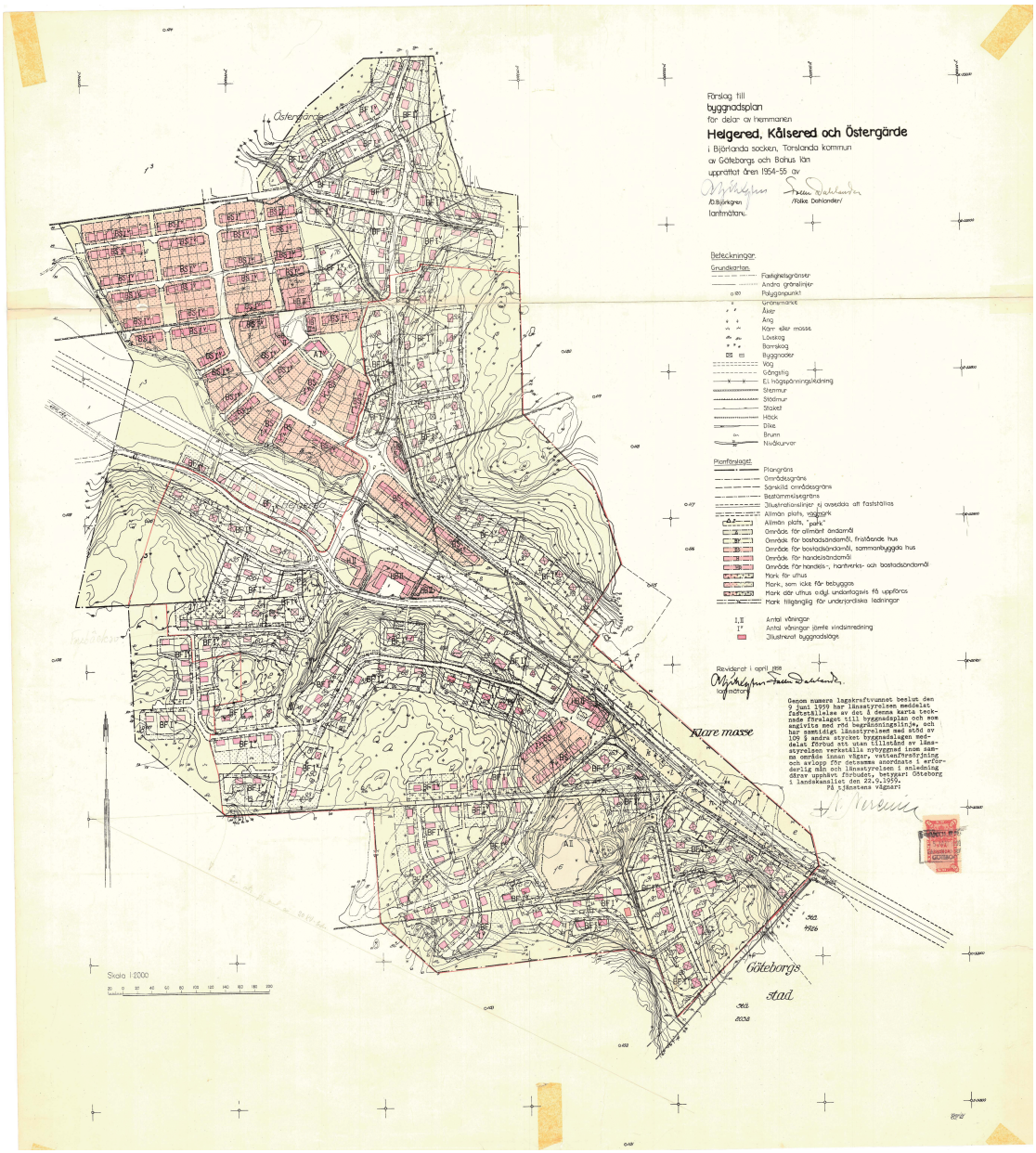


Figure 2.3: Example of a zoning plan from the 1950s for the areas Helgered, Kålsered and Östergärde in the municipality of Gothenburg.

Zoning plans typically consists of several components: the plan map, development regulations (sometimes called planning description), terrain map, overview map, and in some cases, a proposed illustration. Among these, the plan map and development regulations are the most critical, as they define the permissible uses and development intensity for each section of land.

According to Boverket, the Swedish National Board of Housing, Building and Planning, a zoning plan should include a planning description, plan map, and property register, which together provide the legal and practical foundation for implementing the plan (Boverket, n.d.).

The plan map, which is called *plankarta* in Swedish, outlines how land within the zoning area may be used by assigning different land use categories, such as residential,

school, office, commercial, park, or industrial areas. Modern zoning plans clearly and consistently employ standardized colors and alphabetical codes to differentiate between land uses. The terrain map provides a two-dimensional representation of the landscape's elevation using contour lines, offering insight into topographical constraints that may affect development. The overview map situates the zoning area within a broader geographic context, helping users understand its location relative to surrounding neighborhoods and infrastructure. Some zoning plans also include illustrative proposals, giving a conceptual view of how the land might be developed.

The development regulations specify both the extent and the manner in which construction is allowed. These can include limits on building footprint, either expressed as a percentage of the land parcel or as an absolute value in square meters. Regulations also address building height, which can be stated in terms of number of floors or maximum height in meters, depending on the era and drafting practices of the plan. Not all land within a zoning plan is intended for construction. Some areas are deliberately excluded to preserve space for roads, public green spaces, or setbacks. Additionally, zoning plans frequently distinguish between primary and complementary buildings, the latter often being smaller and subject to separate regulatory constraints.

Beyond individual zoning plans, each municipality in Sweden is required to maintain an overview plan, which is called *översiktsplan* in Swedish. This is a strategic document outlining long-term goals and visions for the municipality's overall development. While not legally binding in the same way as zoning plans, the overview plan plays a critical guiding role and helps contextualize zoning decisions within broader municipal planning efforts.

The (Nationella geodataplattformen, n.d.) (NGP) is a digital infrastructure managed by Lantmäteriet that facilitates structured access to spatial planning data across Sweden. It is designed to support municipalities, governmental agencies, and other stakeholders in the planning and development process by enabling standardized data sharing. Through NGP, municipalities can upload, manage, and distribute vectorized zoning plans along with associated metadata. This centralized approach improves data interoperability, promotes transparency, and provides a critical foundation for automating urban planning analyses and decision-making processes. In addition to zoning plans, NGP also offers access to other spatial databases, including detailed buildings and topographic features, further enriching its utility as a comprehensive planning resource.

2.2 Geographic data

The coordinate reference system used for geospatial positioning in Sweden is called *SWEREF 99* (Swedish Reference Frame 1999). It is the official geodetic reference system for all mapping and surveying in the country and is based on the *ETRS89* (European Terrestrial Reference System 1989), which ensures compatibility with global navigation satellite systems such as GPS. The system is maintained by Lantmäteriet and relies on a nationwide network of permanent GNSS reference stations known as *SWEPOS*, which provide high positional accuracy across Sweden.

In this system, locations on the Earth are described using latitude and longitude, which are angular measurements. Latitude indicates how far north or south a point is from the equator, while longitude indicates how far east or west a point is from the Prime

Meridian in Greenwich, England. These two values uniquely identify any point on the globe.

For practical use in mapping and spatial analysis, *SWEREF 99* is typically applied in a projected format called *SWEREF 99 TM* (Transverse Mercator). This projection translates the spherical coordinates into a flat, two-dimensional coordinate system expressed in meters. Each position is defined by an easting and a northing, representing distances in the east-west and north-south directions, respectively. The mathematical foundation and implementation details of this projection are comprehensively documented by Lantmäteriet in their publication on the Gauss Conformal Projection (Lantmäteriet, 2009).

Consistency in the use of coordinate reference systems is essential in spatial data processing. If multiple systems are used within the same project without proper transformation, a single geographic point can appear at different positions depending on the reference system. This can result in significant spatial errors and confusion. Therefore, to ensure accuracy and interoperability, it is crucial to adopt and maintain a single coordinate reference system throughout the entire analysis and development process.

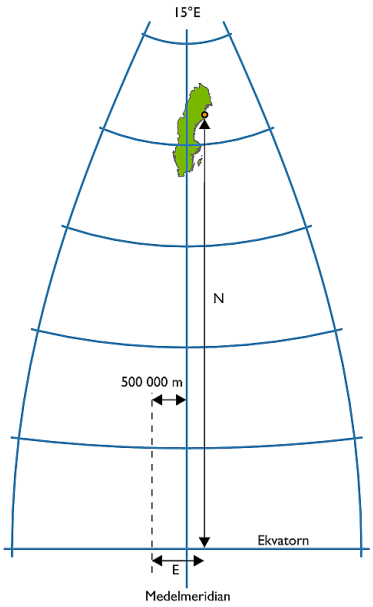


Figure 2.4: Illustration of the SWEREF 99 TM projection system used in Sweden. Source: SWEREF 99

2.3 Large language models

Large language models (LLM) are a class of machine learning models designed to understand and generate human language. Built using transformer-based architectures, these models are trained on vast corpora of text data to learn the statistical relationships between words, phrases, and broader semantic structures. Recent advancements in both model size and training methodologies have enabled LLMs to perform a wide variety of language-related tasks, including translation, summarization, information retrieval, and question answering with a high degree of fluency and contextual awareness.

In the context of this thesis, LLMs are applied to analyze the metadata associated with zoning plans. The goal is to support the automatic interpretation and categorization of large volumes of planning data, enabling more efficient filtering and identification of development opportunities. For example, zoning plan metadata might contain regulatory text, land use designations, and development constraints which can vary significantly in structure and terminology. Here, LLMs can assist in extracting structured meaning from such semi-structured or unstructured data sources.

Despite their capabilities, LLMs also present a number of limitations. One major challenge is the phenomenon of hallucination, where the model, in the absence of clear context or certainty, generates plausible but incorrect or fabricated information. This behavior is especially problematic when working with legal or technical documents such as zoning plans, where factual precision is critical.

Another key limitation lies in the nature of the model's training data. LLMs acquire their knowledge from the text data they are trained on. If the training dataset is outdated or lacks domain-specific accuracy, the responses produced by the model may also be outdated or misleading. Moreover, the authority of the sources used during training is not always guaranteed. In some cases, the model might generate answers based on content from non-authoritative or even incorrect sources, which can be particularly problematic in urban planning contexts where decisions have regulatory and financial consequences.

Terminology ambiguity is an additional issue, particularly in specialized fields like zoning and urban development. Different municipalities or planning documents might use the same term in slightly different ways, or use different terms to refer to similar concepts. This variability can confuse even advanced LLMs, leading to misinterpretations or oversimplified outputs.

To enhance understanding, LLMs often rely on vector embeddings, a method for representing words, phrases, or entire documents as points in a high-dimensional space. This allows the model to capture semantic relationships between words based on their contextual usage. For instance, words related to building regulations or land use categories can be mapped in such a way that their similarities and distinctions become quantifiable, which is useful when interpreting zoning plan metadata with nuanced or overlapping categories.

While LLMs are not without their challenges, their adaptability and language understanding capabilities make them powerful tools for handling large volumes of textual metadata. In this thesis, the use of LLMs is not meant to replace expert interpretation but to serve as a scalable aid for helping users filter and interpret zoning plan data, identify key regulatory parameters, and support data-driven site selection across extensive geographic areas.

3 Methods

Figure 3.1: Overview of methods for identifying and evaluating zoning plans to meet the tool's requirements.

3.1 Methods for meeting the tool's requirements

For the tool to be efficient for the user, it is important that it is user-friendly and with minimal amount of inputs. The tool is based on only one input from the user, which is a bounding box for the queries. The tool begins by allowing the user to define an area of interest within Sweden. This is achieved either by drawing a region on an interactive map interface or by specifying two geographic coordinates that together can be converted into a rectangular bounding box. Each coordinate is composed of a latitude and longitude, formatted according to the SWEREF 99 reference system. This bounding box is used in subsequent POST queries to retrieve data from Lantmäteriet, Sweden's national mapping agency.

The first step is to fetch data from various databases. One query fetches all zoning plans that have obtained legal force, located within the bounding box. A separate query retrieves relevant development regulations from these plans, focusing specifically on land use categories deemed significant for assessing exploitation potential. For this purpose, land designated for residential, industrial, commercial, and school use is prioritized.

To retrieve regulations, two specific queries are constructed. The first filters for plan regulations categorized under "Omfattning" with subcategories such as "Höjd på byggnader," "Utnyttjandegrad," and "Begränsning av markens utnyttjande." The second query focuses directly on categories like "Höjd på byggnadsverk," "Utnyttjandegrad," and "Begränsning av markens utnyttjande."

One of the queries is fetching the zoning plan regulations that are categorized as:

```
"planbestämmelse.kategori": {
  "in": ["Omfattning"]
},
"planbestämmelse.underkategori": {
  "in": [
    "Höjd på byggnader",
    "Utnyttjandegrad",
    "Begränsning av markens utnyttjande"
  ]
},
```

The second query are fetching the zoning plan regulations that are categorized as:

```
"planbestämmelse.kategori": {
  "in": [
    "Höjd på byggnadsverk",
    "Utnyttjandegrad",
    "Begränsning av markens utnyttjande"
  ]
},
```

Since regulations on building height can be expressed in different formats such as building height in meters, number of floors or height above sea level, it is crucial to correctly interpret each format. In cases where elevation above sea level is relevant, the tool fetches topographic data from the *EU-DEM* dataset via the Open Topo Data API. This dataset provides elevation values across Europe with a spatial resolution of 25 meters. In addition to zoning plans, data on the existing built environment is retrieved. Building footprints are collected in the form of *SWEREF 99* coordinates and transformed into two-dimensional polygonal shapes. These geometries are then spatially overlaid with zoning plans to calculate the extent of land within each plan that is currently unexploited.

To ensure accuracy, areas where no construction is allowed such as setbacks, green spaces or protected zones are excluded from the analysis. These are filtered out using set difference operations, which remove non-developable zones from the zoning plan geometry before the intersection with building footprints is computed. The result is a cleaned and quantified representation of how much of the permitted building area is currently in use, laying the groundwork for further estimations such as available gross floor area (GFA) and development potential. The resulting geodataframe, which includes all zoning plans and their associated development regulations, is sorted according to user-defined metrics and visualized for further analysis.

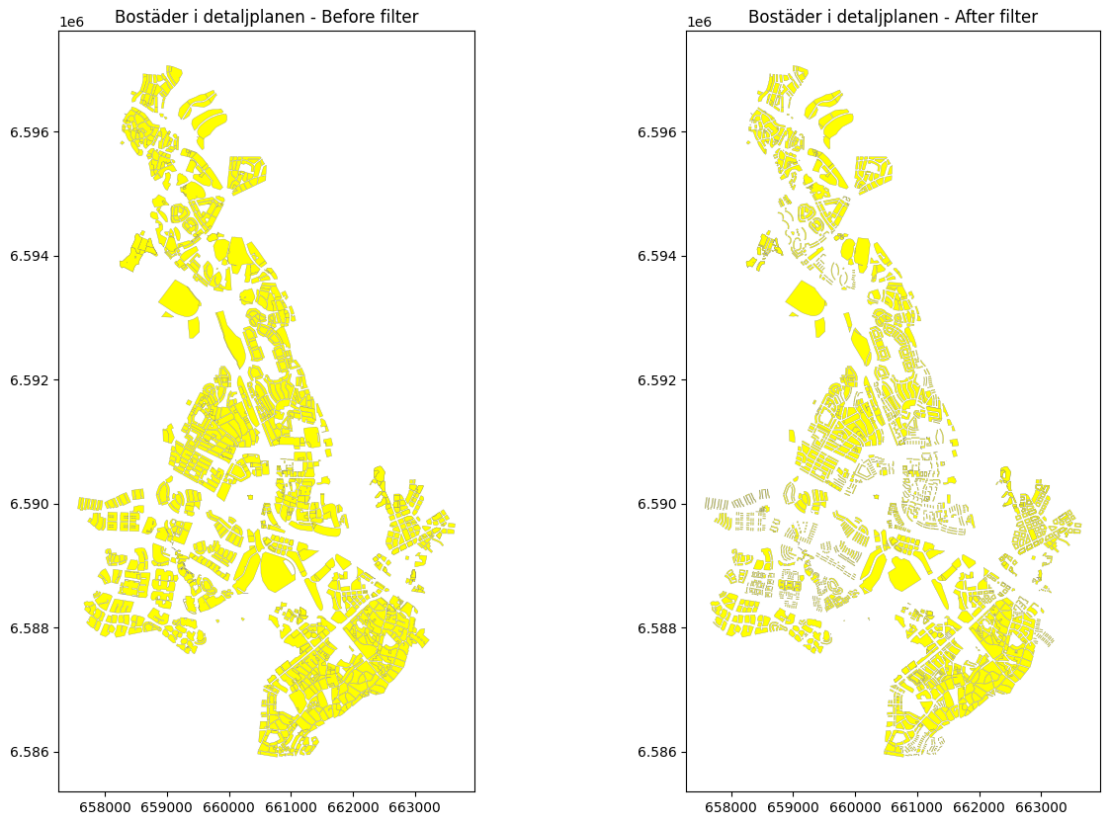


Figure 3.2: Visualization of residential zoning plans.

To assess the extent to which each zoning plan has been utilized, the existing built environment is compared to the areas designated for development within the plan. This process involves spatial analysis using geometric data structures. Both zoning plans and building footprints are loaded into *GeoDataFrames* using the *GeoPandas* library. These *GeoDataFrames* allow for advanced geometric operations. The building geometries are then overlaid onto the zoning plan geometries to calculate their intersection. This operation identifies the portions of the zoning plans that are currently occupied by buildings. The intersection area is then divided by the total area of the zoning plans exploitable land, resulting in a ratio that indicates the level of exploitation. This percentage provides a straightforward metric for identifying underutilized zoning plans, which are of particular interest for potential projects.

Figure 3.3: Overlay of zoning plan polygons and building footprints.

3.2 Fetching of data

Table 3.1: Overview of datasets used in the study

Dataset	Source
Zoning plans	National geodata platform
Building footprints and point cloud	Lantmäteriet/DTCC
Residential transactions	Booli

3.2.1 Zoning plan data

Zoning plan data is accessed through the API of the *National Geodata Platform (NGP)*, administered by Lantmäteriet (see Table 3.1). This platform provides access to a comprehensive dataset of zoning plans across Sweden. The data can be converted into JSON format and contains both geometric representations and associated regulatory metadata. The geometric data consists of coordinate pairs defined in the *SWEREF 99* reference system. These coordinates outline the boundaries of each zoning plan and are used to construct polygon objects. The transformation from coordinate data to polygonal shapes is performed using the Python library *Shapely*, which is designed for manipulation and analysis of planar geometric objects.

The API allows for filtering of zoning plans based on spatial parameters, such as a bounding box defined by the user. Additional filters ensure that only plans which have obtained legal force are retrieved, which is critical for the legal credibility of any subsequent analysis. To manage and process the retrieved data efficiently, the *GeoPandas* library is used to structure the geospatial data in the form of a *GeoDataFrame*. This format allows for structured integration of both geometric and regulatory information and supports advanced spatial operations such as intersection analyses and set differences between multiple geometries.

3.2.2 Buildings in Sweden

The (Digital Twin Cities Centre, n.d.) (DTCC) is a research center in Sweden focused on developing digital twin technologies for urban planning, combining geospatial data, simulation, and visualization to support smarter city development. Data on currently existing buildings is retrieved through an API for Lantmäteriet's building database, accessed through the DTCC. This database contains detailed information about the footprints of buildings across Sweden, formatted using the same *SWEREF 99* coordinate reference system as the zoning plans. Each building record includes geometric data in the form of coordinates, which are converted into polygonal shapes using the *Shapely* library. This allows for spatial set overlay intersection with zoning plan geometries. In addition to the building footprints, each building is associated with a unique identification number, which facilitates cross-referencing and further data processing. The building data plays a crucial role in assessing the current level of land utilization. When overlaid with zoning plans, the building geometries can be used to determine which parts of the plan are already developed and which remain available for future construction. This is essential for identifying underutilized zoning plans and for estimating the built area in relation to the regulatory potential of the land.

3.2.3 Pointcloud of areas in Sweden

Retrieving point cloud data across extensive geographic areas is significantly more resource-intensive than fetching zoning plans or building footprints. Point clouds consist of a dense collection of elevation points that represent the three-dimensional shape of both terrain and buildings. This data is essential for estimating building heights when no direct height attribute is available from the building footprints. The point cloud files are organized in a grid structure, with each tile covering an area of 2,500 by 2,500 meters. The point cloud has a resolution of approximately 1–2 points per square meter, offering sufficient detail for analyzing the vertical dimensions of buildings and terrain. These tiles are stored in a database managed by Lantmäteriet and are accessed via API requests (see Table 3.1). The request queries include bounding boxes, but even when a small subset of a tile is needed, the entire file must be downloaded and loaded into memory. For each request, the file is first cached locally to improve performance during repeated access.

3.2.4 Market price of real estates

Estimating the market value of a potential development area is a critical step in determining its financial feasibility. To support this analysis, data on recently sold residential properties in the vicinity of each zoning plan is collected. This data provides insight into local housing demand and price levels, which are key indicators for assessing the attractiveness of an area. The data is retrieved from a property transaction database maintained by (Booli, n.d.), which includes detailed records of sales prices and floor areas for apartments and houses across Sweden (see Table 3.1). By focusing on properties sold within a defined radius of 500 meters from the centroid of each zoning plan, a localized price per square meter can be calculated.

Only transactions from the past 15 years are considered to ensure the data reflects current market conditions. For each zoning plan, the square meter prices of the selected transactions are averaged to produce an estimated market value metric. This metric enables comparisons between different areas and helps identify neighborhoods where people are willing to pay a premium to live, which in turn informs development potential. The market data serves as a proxy for many qualitative factors that are otherwise challenging to quantify such as access to amenities, views of green areas or water and the overall desirability of the neighborhood.

3.3 Analysis of metadata attached to the zoning plan

A central component of this project is the extraction and interpretation of metadata linked to each zoning plan. Among the various metadata types, the most critical for assessing development potential are the development regulations. Especially the development regulations that are related to land use, building height, site coverage, and usage restrictions. The initial step involves removing areas within the zoning plans where development is prohibited. This is done using overlay difference operations in *GeoPandas*, which exclude these areas from further analysis, allowing for a more accurate estimate of exploitable land. Once cleaned, the remaining areas are analyzed for their potential gross floor area (GFA), which is a crucial metric for evaluating the scale of development that could be realized on the site.

GFA is typically estimated by multiplying the allowed building footprint area (BYA)

by the number of floors permitted. However, the regulatory data provided in the zoning plan metadata exhibits substantial variability. In some cases zoning plan regulations are explicitly stated in numeric form, such as a maximum of 3 floors on 40 square meters. In other cases, the information is only implied, for example through restrictions on building height above sea level or through vague descriptions such as "building height shall vary". Some zoning plan regulations already include key-value pairs of unit and value within their JSON metadata, while others lack this information. For those that do not, additional processing is required to infer these details. In this thesis, results of two different methods for extracting this information are presented. The first is a traditional regular expression keyword matching algorithm where certain keywords such as "meter", "%" and "m^2" are searched for. The second is making use of a large language model (LLM). Specifically, the LLM that is used is *gemini-2.5-flash-preview-05-20*, which is developed by (DeepMind, n.d.).

3.4 Data description

The metadata is organized in JSON format, with entries for plan categories (*kategori*), and occasionally also subcategories (*underkategori*), and the corresponding regulation text (*bestämmelseformulering*). Height regulations may appear under the category “Höjd på byggnadsverk” without a subcategory. The same regulation text could also have been in the category "Omfattning" with the subcategory of "Höjd på byggnader". Similarly, usage utilization is listed under “Utnyttjandegrad.”, which is both an own category and also a subcategory in the category "Omfattning". Several examples of well-structured metadata entries are shown in the JSON data, where variables such as the allowed number of floors or footprint size per building are clearly defined. However, there are also cases where the regulation is ambiguously formulated, lacks a numeric value, or provides only a general directive.

Because of the variability, it is essential to design API queries that are robust and inclusive so that they capture as many relevant regulation types as possible. The land use regulations are then parsed and the meaning is distinguished based on key terms in their formulation. Before the categorization methods, approximately half of the regulations texts are manually extracted, depending on the municipality that the plan in located in. What is occasionally missing is the key-value pair of "bestämmelsevarde", where the value is at least two other key-value pairs, where the keys are called "variabelvarde" and "enhet". In english, there words correspond to value and unit. After applying the following sequence of methods, most of the regulation texts are fully complete JSON strings.

The data from the national geodata platform is converted to JSON format:

Listing 3.1: Example of zoning regulation metadata for number of floors

```
{
  "planbestammelsekatalogreferens": "5eb359f5-4bb8-4ea1-b2cb-36
     ddf190a6f0",
  "bestammelseformulering": "Högsta antal våningar är angivet
     som 1. Bestämmelsen har digitaliserats för att möjliggöra
     analyser, men används inte idag.",
  "bestammelsevarde": {
    "variabelvarde": 1,
    "vardetyp": "max",
```

```
    "enhet": "antal"
  },
  "anvandningsform": "Kvartersmark",
  "kategori": "Höjd på byggnadsverk",
  "kvalitet": {
    "korrigeradeGranser": false,
    "kontrolleratPlaneringsunderlag": false
  }
}
```

Listing 3.2: Example of zoning regulation metadata

```
{
  "planbestammelsekatalogreferens": "dc4c4b93-5135-4dc0-9ca9
     -108b686db3a9",
  "bestammelseformulering": "Största byggnadsarea är 40 m2 per
     komplementbyggnad",
  "bestammelsevarde": {
    "variabelvarde": 40,
    "vardetyp": "max",
    "enhet": "kvadratmeter"
  },
  "anvandningsform": "Kvartersmark",
  "kategori": "Utnyttjandegrad",
  "underkategori": "Area per byggnad",
  "kvalitet": {
    "korrigeradeGranser": false,
    "kontrolleratPlaneringsunderlag": false
  }
}
```

Listing 3.3: Example of zoning regulation metadata with missing value and unit

```
{
  "planbestammelsekatalogreferens": "e1423d9f-e4e6-4f2c-9df7-0
     e9970a119ad",
"bestammelseformulering": "Byggnadshöjden ska variera",
"bestammelsevarde": None
  "anvandningsform": "Kvartersmark",
  "kategori": "Höjd på byggnadsverk",
  "underkategori": "Högsta höjd på byggnadsverk",
  "kvalitet": {
    "korrigeradeGranser": false,
"kontrolleratPlaneringsunderlag": false
  }
}
```

A regular expression keyword matching algorithm are applied to classify regulations. For instance, any regulation in the category of "Höjd på byggnadsverk" that is also containing the word “våning”, (floor) will have the unit value of "antal", (quantity) and the first number occurring in the formulation will be assigned as the value. To clarify, the purpose of this is to complete the "bestämmelsevarde" value of the JSON that

stems from the data that are fetched from the national geodata platform. However, regular expression keyword matching alone is insufficient in cases with ambiguous or non-standard language. This is where large language models (LLMs) are introduced to enhance accuracy. LLMs are capable of interpreting the semantic meaning of a sentence by utilizing vector embeddings and understanding contextual relationships, rather than relying solely on predefined keyword patterns. This significantly improves the understanding of regulations that are difficult to interpret with rule-based algorithms alone.

3.5 Leveraging LLMs on the metadata attached to zoning plans

While semantic keyword matching offers a baseline approach for classifying zoning plan metadata, it falls short when confronted with ambiguous phrasing, inconsistent terminology, or implicitly stated regulations. To overcome these limitations, large language models (LLMs) are introduced to improve both the accuracy and reliability of metadata interpretation. The task of the LLM in this process is to identify the correct regulatory value and its corresponding unit.

The LLM *gemini-2.5-flash-preview-05-20* is provided with free-form text regulations from zoning plans. It is prompted to extract a structured JSON representation of the free-form text. The prompt for the building utilization (BYA) is as follows:

> You are an expert in Swedish building regulations. You consistently understand regulation rules on building utilization.
>
> The input consists of a list of rules. You must return one JSON for each of the rules in the input.
>
> If the regulation is not interpretable or does not contain both unit and value, return null.
>
> Output only the list of JSON objects.
>
> Input Rules:
>
> *Utilization regulations*
>
> Each JSON object in the list must match the following schema:

Listing 3.4: Schema for Building Utilization (BYA)

```
{
  "type": "object",
  "properties": {
    "value": {
      "description": "This is where the actual value
         is inserted.",
      "type": "number"
    },
    "unit": {
      "description": "This is where the unit is
         inserted.",
      "type": "string",
```

```
      "enum": ["kvadratmeter", "procent"]
    }
  },
  "required": ["value", "unit"]
}
```

3.6 Extracting currently built GFA

To evaluate the current utilization of each zoning plan, the tool estimates the built gross floor area (GFA) by integrating spatial and height data. This estimation is performed by identifying all intersections between building footprints and each respective zoning plan boundary. Each building footprint intersection is multiplied by the number of floors of the building, yielding an approximation of the total developed floor space inside the zoning plan. This method provides a scalable approach to compute how much of the zoning plan's development potential has already been realized. While building footprints are readily available and can be converted into polygons for spatial analysis, the number of floors is not currently included in any publicly accessible datasets. Although the National Geodata Platform (NGP) provides a designated database for detailed building data, it currently does not contain data.

A practical method for estimating building heights utilizes point cloud data for the area of interest. For each building footprint, the tool identifies all elevation points (Z-values) within the polygon and extracts the highest point, assumed to correspond to the rooftop. To determine the ground level, the tool isolates ground-classified points within the same footprint. The building height is then estimated as the difference between the rooftop and ground elevations. To approximate the number of floors, a standard assumption of 3 meters per floor is applied. While effective for small areas, this approach becomes computationally intensive when applied at scale across thousands of buildings. To overcome this challenge, efforts are underway to precompute and store building height data in a centralized dataset, enabling more efficient and scalable analyses in future applications.

3.7 Presentation and visualization of the results

After processing and analyzing the zoning plan data, the results are compiled into a structured *GeoDataFrame* where each row represents a distinct zoning plan. The *GeoDataFrame* includes both geometric data in the form of a polygon and metadata such as name, municipality, the year when the plan entered legal force, as well as derived metrics such as built percentage, estimated GFA, and market value indicators. To facilitate spatial interpretation and user interaction, the *GeoDataFrame* is exported to *GeoJSON* format and visualized using *Kepler.gl*, an open-source geospatial analysis platform. An interactive HTML file is generated in which zoning plans are represented as colored polygons on a map, enriched with associated metadata. With this map, users are able to intuitively explore, filter, and assess zoning plans based on spatial, regulatory and economic parameters.

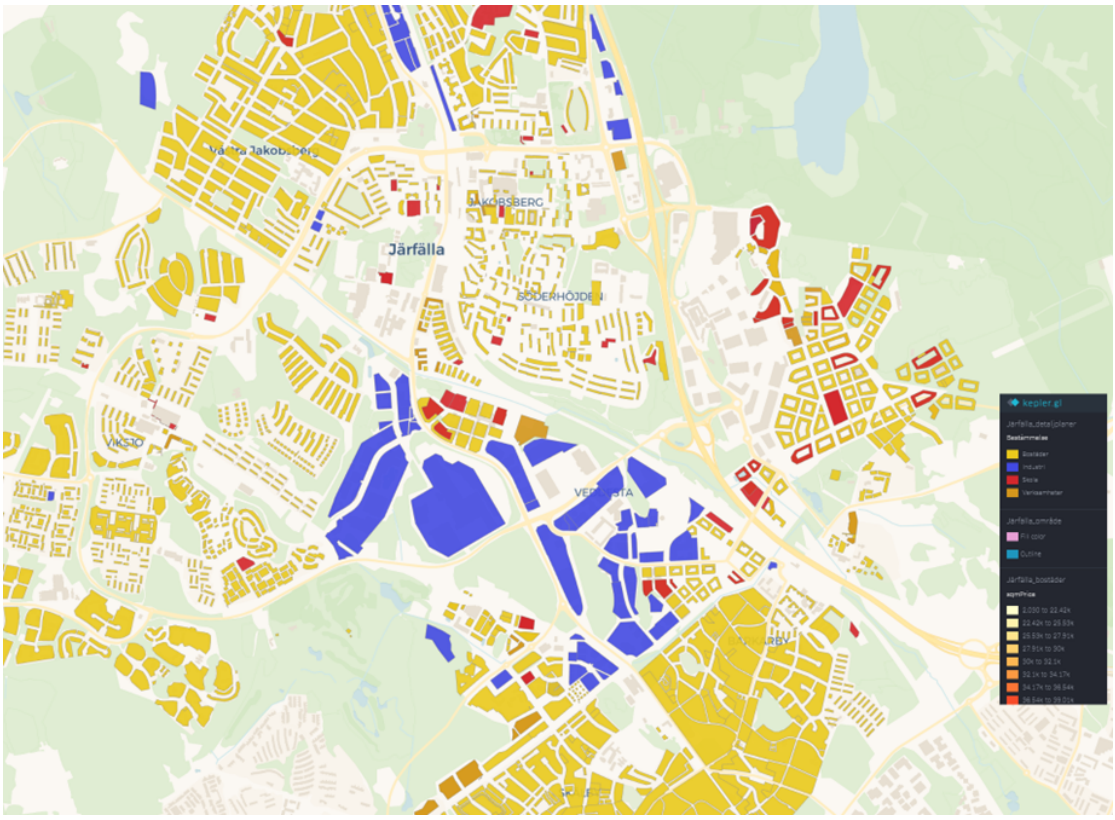


Figure 3.4: Visualization of digitized zoning plans in Järfälla.

A configuration file allows customization of the visualization by specifying color schemes, tooltips, filtering parameters, and map styles. This includes defining how zoning plans are color-coded based on their use regulations and setting up pre-configured filters for attributes such as estimated GFA or market value indicators. These filters are particularly useful for quickly identifying zoning plans with high development potential. The resulting HTML file is portable and can be opened locally or hosted online for collaborative access, with all predefined settings embedded. This interactive map transforms complex spatial and regulatory data into an intuitive visual format, supporting informed decision-making for developers, planners, and architects.

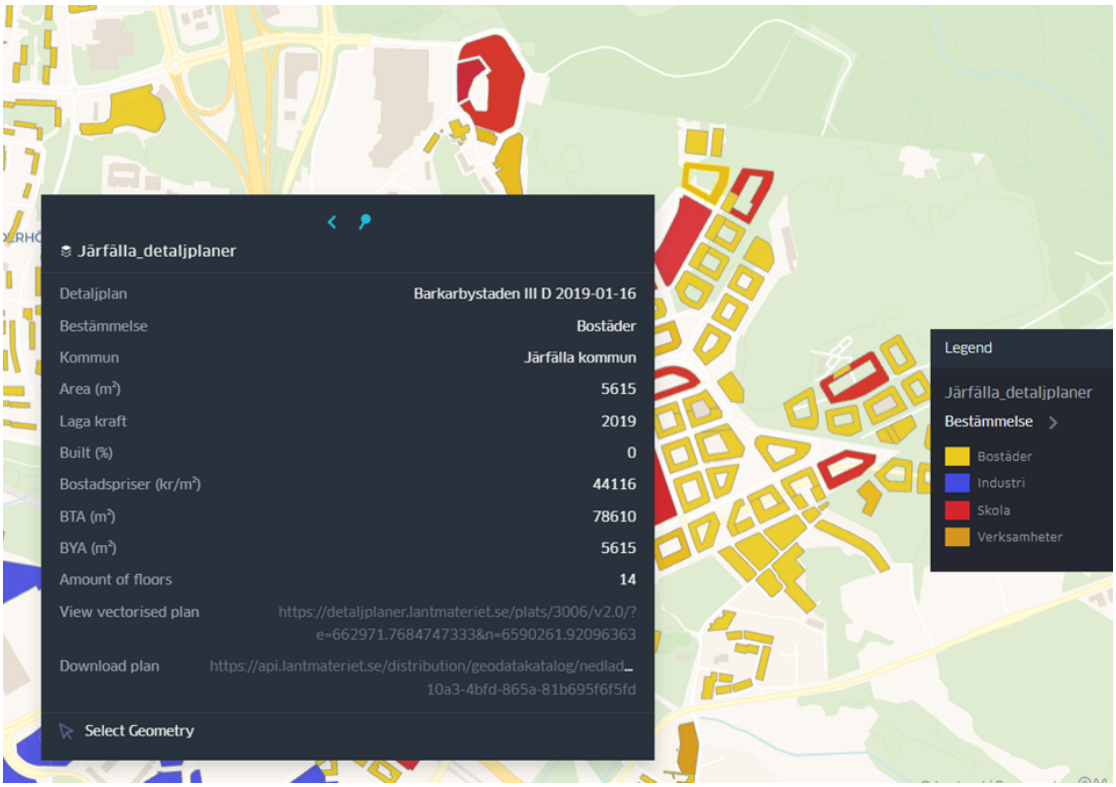


Figure 3.5: Tooltip showing detailed zoning plan metadata in the interactive Kepler map.

3.8 Market value estimation

Data on residential transactions is fetched from a database maintained by *Booli*, which is a Swedish real estate data provider offering detailed information on property sales, pricing trends, and housing market dynamics across the country. The fetched data includes a point with metadata of sale price, construction year, date of sale, and floor area. To ensure relevance, only properties sold within the last 15 years are considered. For each zoning plan, the tool identifies transactions that occurred within a 500-meter radius of its centroid. The average square meter price of these sales is then calculated, providing a localized estimate of market value.

To visualize market conditions spatially, the map incorporates property transactions represented as points, each color-coded according to its square meter price. A pre-configured legend defines price intervals and assigns them a gradient of colors, enabling intuitive visual interpretation of housing market trends. This color-coding makes it easy to detect variations in demand, even within small urban segments. By overlaying this market data onto the zoning plan layer, the tool supports comparative assessment of different areas based on localized market value. Ultimately, integrating market indicators into the analysis helps bridge the gap between regulatory potential and actual demand, providing a more informed foundation for investment decisions.

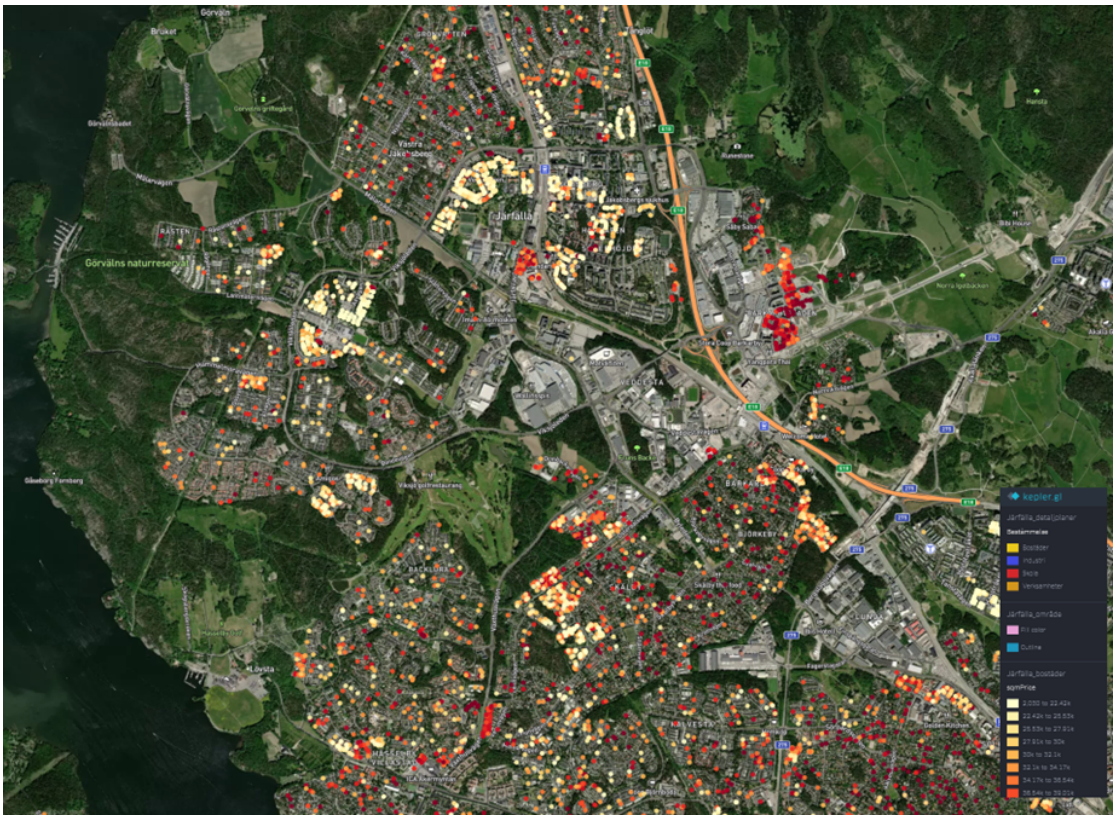


Figure 3.6: Visualization of recent residential transactions from Booli's database in Järfälla.

The tooltip function in the Kepler map provides an interactive way for users to explore detailed information about each data point. When hovering over a property transaction, a tooltip appears showing metadata such as address, sale price, living area, and square meter price. This functionality enhances user interaction by providing immediate access to detailed information without leaving the visual environment.

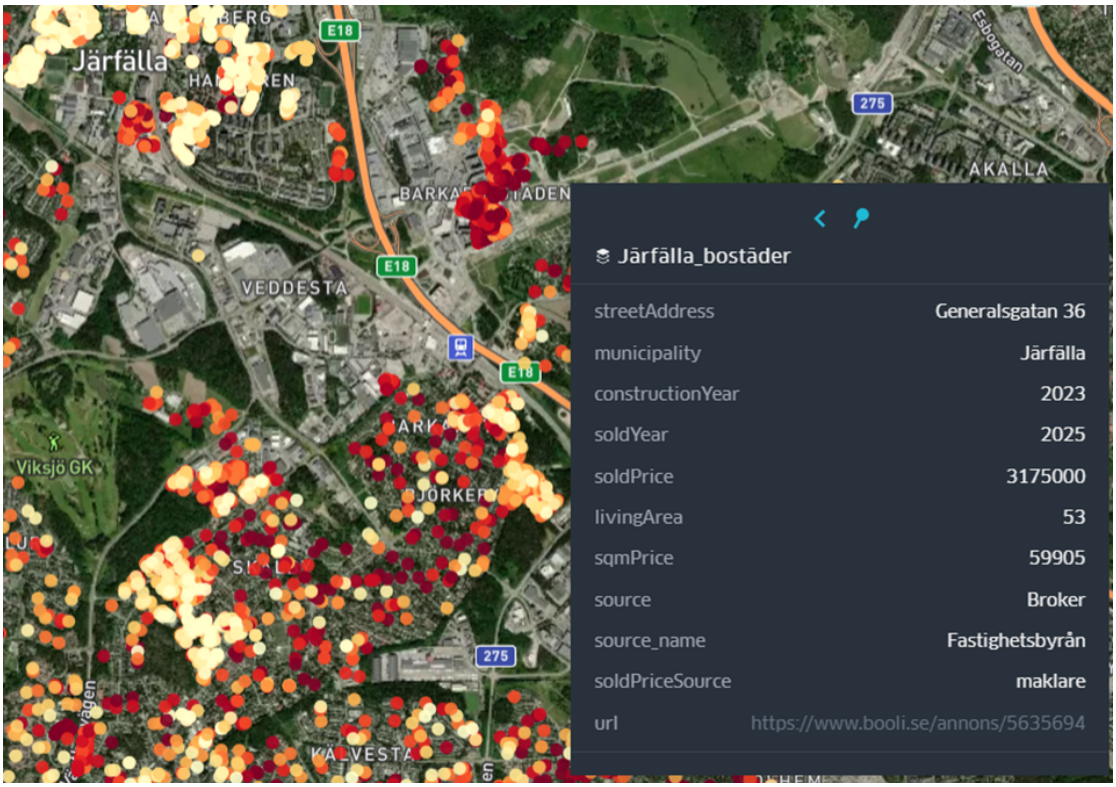


Figure 3.7: Tooltip showing detailed transaction information of a residence Järfälla.

3.9 Case studies

To evaluate the performance and practical applicability of the tool, a series of case studies were conducted across selected municipalities in Sweden. These case studies serve as proof-of-concept demonstrations, highlighting how the tool can be used to identify underutilized zoning plans, estimate development potential, and assess market feasibility based on location-specific data. Each case study involves a predefined geographic area where the complete workflow of the tool is applied, from data retrieval and spatial analysis to GFA estimation and market value evaluation. The selection of municipalities reflects a range of urban contexts, including both growing cities and smaller towns, to demonstrate the tool's flexibility and scalability.

The zoning plan regulations in the three case studies differ significantly, reflecting the common variability between municipalities. These cities were selected based on the extent of their zoning plan coverage and their geographic distribution across Sweden. Despite their relatively modest size, all three locations contains a sufficient amount of zoning plans for a comparative analysis. Still, processing and generating a *Kepler* map for each area is not particularly time-consuming. Building height and land utilization values are extracted using both a traditional regular expression keyword matching algorithm and a large language model (LLM). These two approaches allows for a comparative evaluation between natural language processing (NLP) techniques and conventional methods for interpreting zoning regulations.

3.9.1 Järfälla

Located just northwest of Stockholm, Järfälla is a rapidly developing municipality that features a blend of newly planned urban expansion areas and established residential

districts. The area was chosen for this case study due to the diversity of its zoning plans, which range from recently produced documents to plans that are several decades old. Its appeal is further strengthened by its close commuting distance to Stockholm, where residential property prices are significantly higher, making Järfälla an attractive alternative for both residents and developers.

3.9.2 Halmstad

Halmstad is a coastal city located on Sweden's west coast, known for its maritime economy, and a balanced mix of residential and commercial development. The municipality features a wide variety of zoning plans, ranging from central urban areas to more suburban and industrial zones. Halmstad's attractiveness is supported by its seaside location and its role as a regional hub in Halland County. In this case study, the tool was applied to identify zoning plans with low levels of exploitation and high development potential, particularly in proximity to the coastline and areas with good transport access. The market analysis revealed significant variation in property prices, especially between waterfront zones and inland neighborhoods, highlighting the importance of location-specific valuation in development assessments.

3.9.3 Sundsvall

Sundsvall is a northern Swedish city, known for its strong industrial heritage, expansive forest sector, and growing service economy. The city features a distinct urban core flanked by surrounding residential areas and green spaces. Zoning plans in Sundsvall exhibit a range of characteristics, from compact central districts to large, sparsely developed plots on the urban fringe. Its location, within commuting distance of regional industries and educational institutions, makes certain neighborhoods particularly attractive for future housing development. In this case study, the tool was used to evaluate how existing plans align with current market conditions. Several underutilized zoning plans were identified, particularly in areas with moderate residential prices and access to public infrastructure, suggesting potential for targeted infill development.

City	Height regulations extracted (%)	Utilization regulations extracted (%)
Järfälla	33.3	35.1
Halmstad	47.2	12.7
Sundsvall	48.5	93.8

Table 3.2: Percentage of regulations containing a manually extracted unit and value.

4 Results

4.1 The tool

A tool has been developed to automate the analysis of zoning plans across Sweden. It is designed for ease of use, featuring a minimal interface with a single button which initiates the entire workflow. The only thing that is expected from the user is to define the area of interest. Once the analysis is executed, the tool automatically generates a fully configured *Kepler* map and proceeds to open it in a browser window. This map includes interactive layers, filtering options, and visual styling that allow users to explore and interpret the results. In most urban areas of Sweden, the tool performs sufficiently well in identifying unexploited zoning plans, particularly those with much remaining development potential.

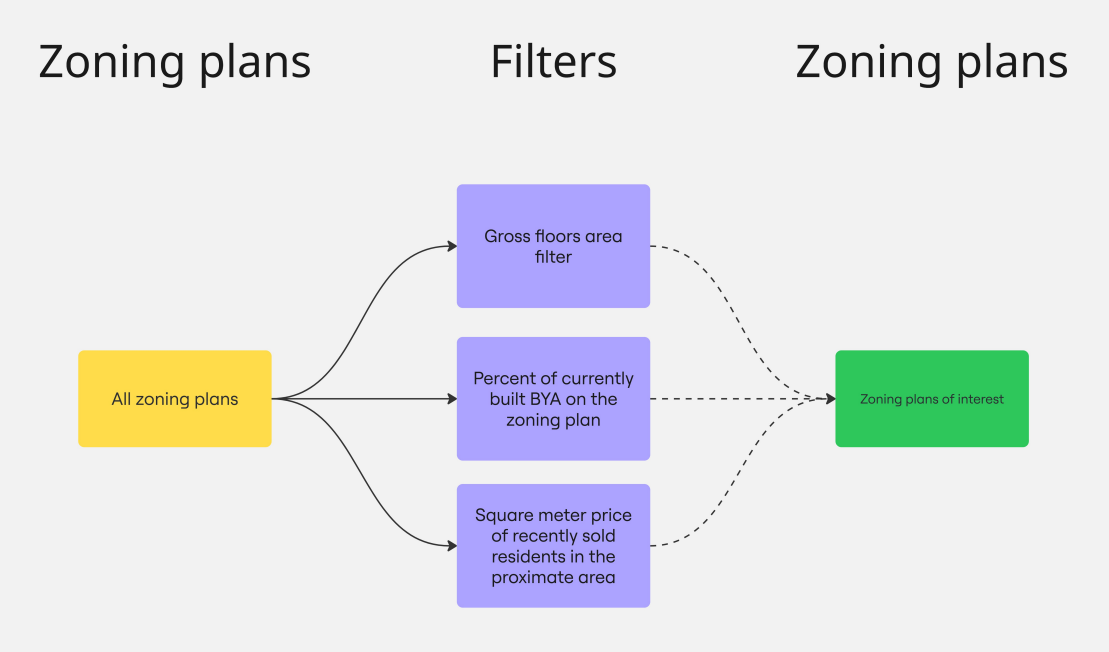


Figure 4.1: Multi-parameter filter for identifying zoning plans of interest.

The filter function of the tool enables users to efficiently identify zoning plans of interest by applying constraints across multiple parameters. By combining filters for regulation categories such as building height or estimated gross floor area, along with contextual indicators like the percentage of current site coverage and recent selling prices per square meter in the surrounding neighborhood, the user can filter out irrelevant plans. This multi-parameter filtering allows for targeted searches, making it possible to identify zoning plans that meet specific development criteria.

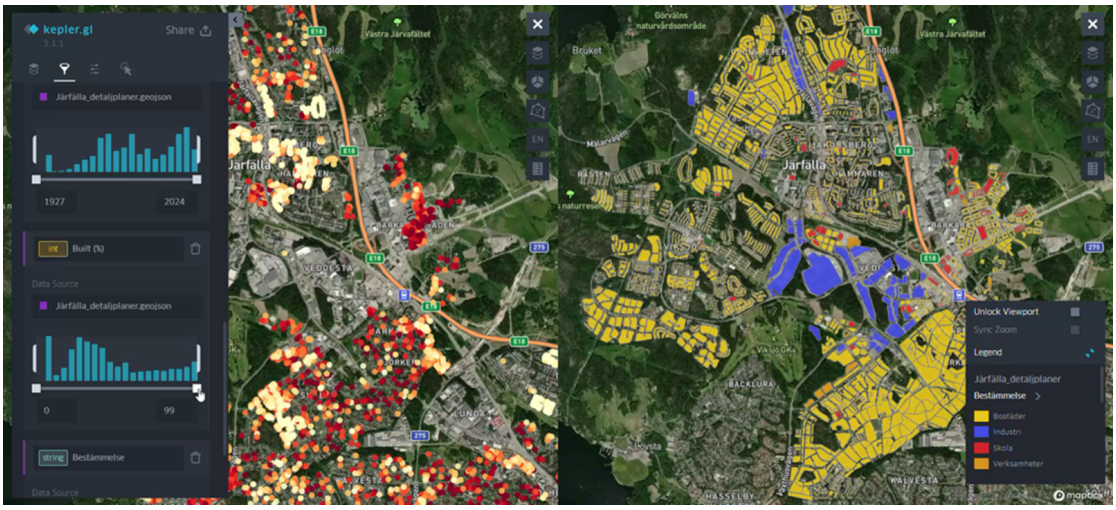


Figure 4.2: Dual map displaying both market prices and zoning plans in Järfälla. Filter based on building footprint utilization is lenient.

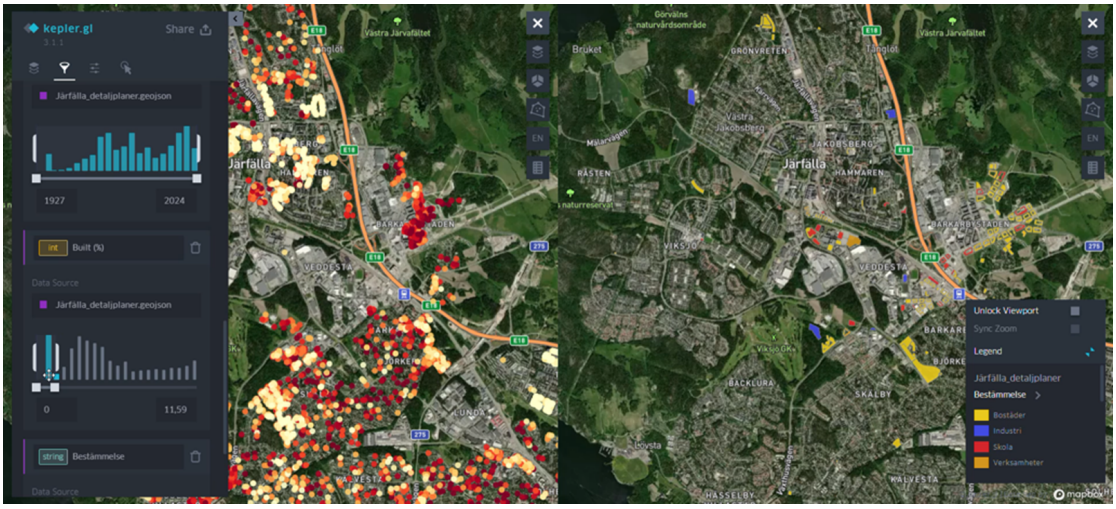


Figure 4.3: Narrowing down zoning plans in Järfälla using a strict filter on built percentage to identify unexploited plans.

4.2 Case studies

Confusion matrices have been produced to evaluate the performance of the two classification methods on zoning regulations where values and units have not been manually provided. The classification logic is based on a binary distinction: *positive* and *negative* category. Regulations in the *negative category* are those that the model or algorithm should not attempt to extract, because they lack sufficient information that contribute to the zoning plan key performance indicator (KPI). The *positive category* includes regulations that clearly and explicitly state both a unit and a value. These are regulations where extraction is possible, straight-forward and provide information on the KPI of the zoning plan. A *false negative* occurs when the model fails to extract the value and unit from a regulation that meets the criteria for the positive category. A *false positive* categorized regulation are one where a value and unit have been extracted, without the sufficient specificity for either unit or value in the regulation.

For regulations where it is uncertain wether the user would expect the information to be extracted or not, they are put in the *negative category*.

4.2.1 Järfälla

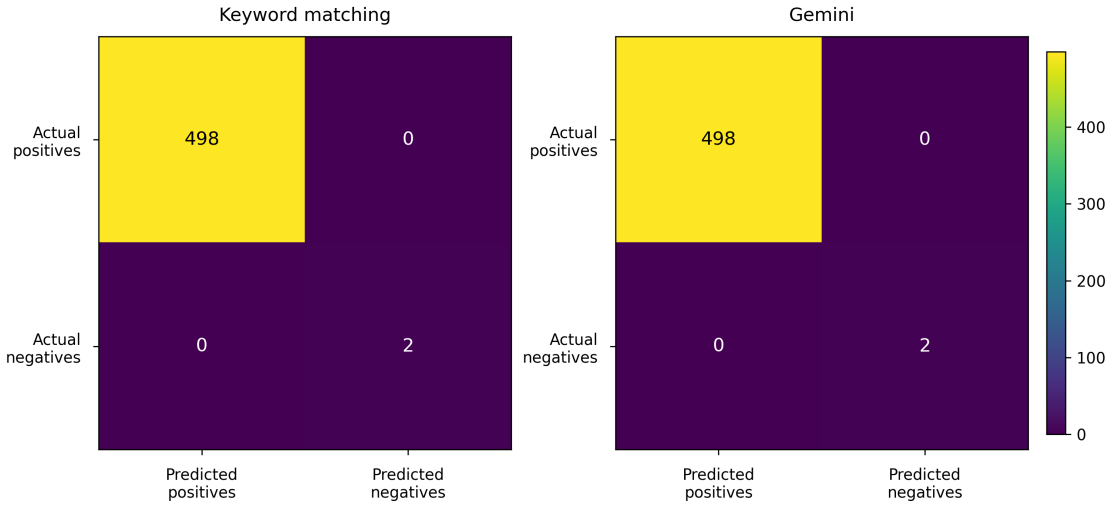


Figure 4.4: Building height regulations extracted from Järfälla.

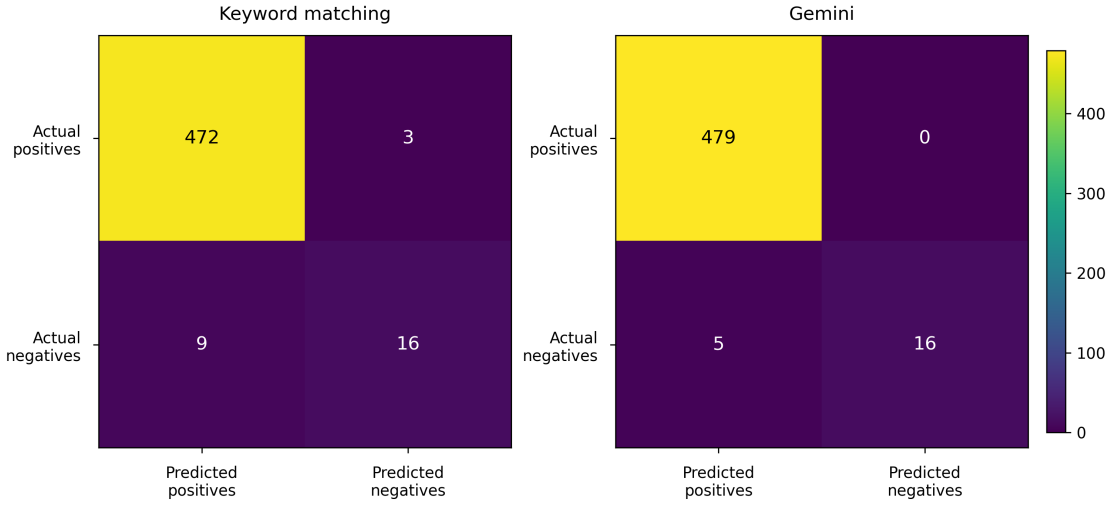


Figure 4.5: Building utilization regulations extracted from Järfälla.

4.2.2 Halmstad

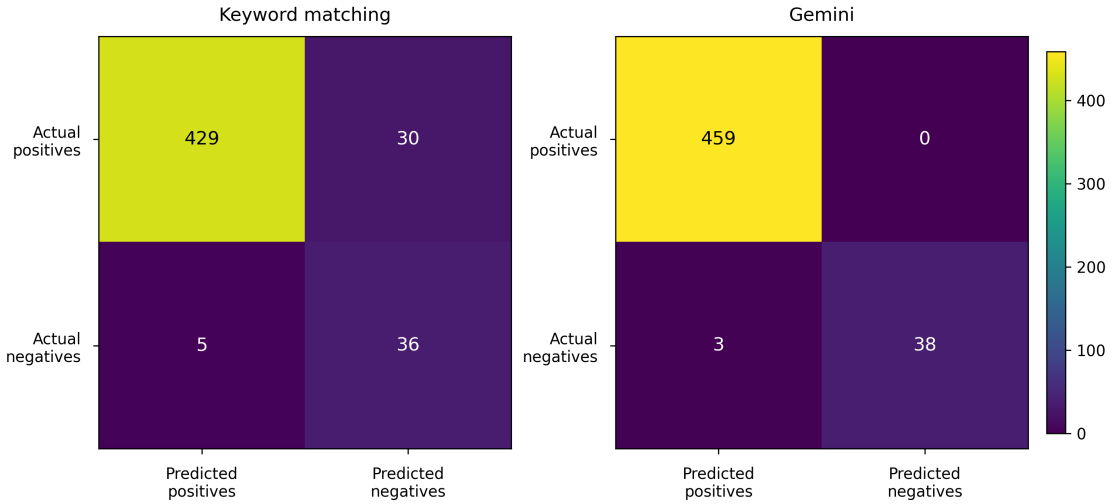


Figure 4.6: Building height regulations extracted from Halmstad.

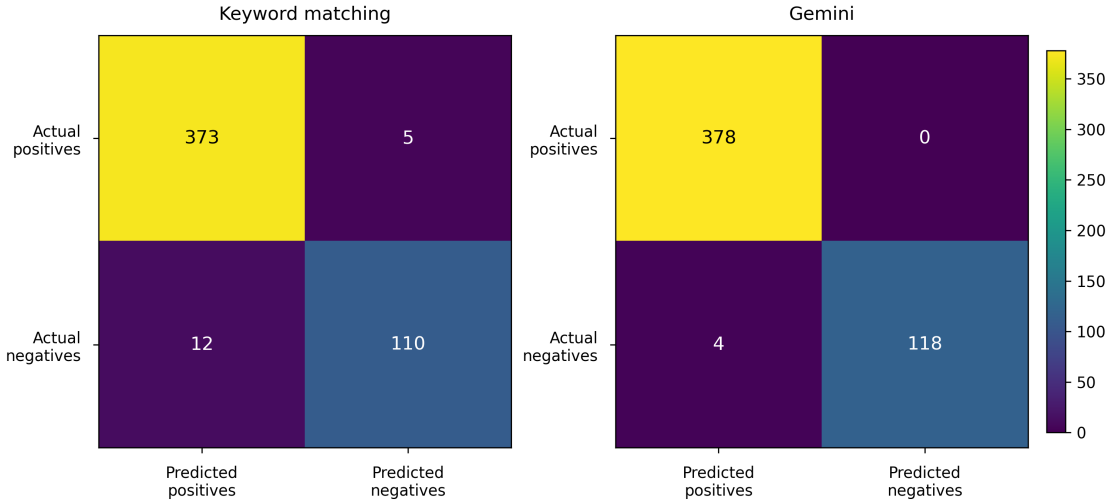


Figure 4.7: Building utilization regulations extracted from Halmstad.

4.2.3 Sundsvall

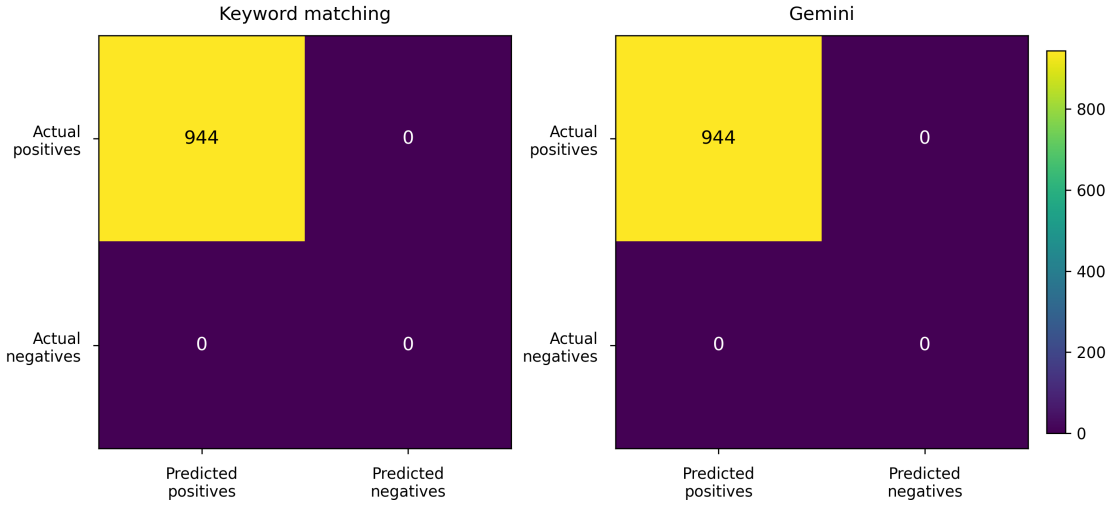


Figure 4.8: Building height regulations extracted from Sundsvall.

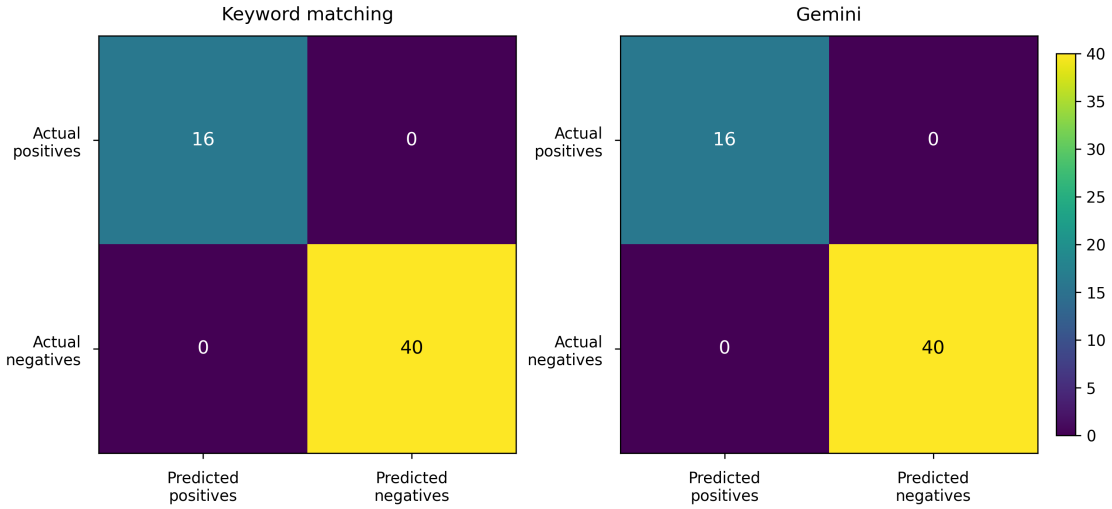


Figure 4.9: Building utilization regulations extracted from Sundsvall.

4.3 Examples of regulations

Tables 4.1–4.4 compare regulations against the outputs of the keyword-matching and LLM methods.

Table 4.1: *True positive* regulations extractions by keyword matching and LLM

1. *Högsta totalhöjd på fast föremål av typ skorsten, antenn, träd eller dylikt är +65 meter över angivet nollplan med hänsyn till flygsäkerheten, oaktad annan höjdangivelse*
Keyword matching: {variabelvarde: 65, enhet: meter}
LLM: {variabelvarde: 65, enhet: meter}

2. *Högst en åttondel av tomtplats får bebyggas*
Keyword matching: null (*false negative*)
LLM: {variabelvarde: 12.5, enhet: procent}

3. *Högsta exploateringsgrad i bruttoarea per fastighetsarea är 0,8*
Keyword matching: null (*false negative*)
LLM: {variabelvarde: 80, enhet: procent}

4. *Tillägg 1380K-Ä46 Bestämmelse tillkommer: Högsta tillåtna byggnadshöjd för huvudbyggnad är 12,5 meter. Bestämmelse upphävs: "Högsta tillåtna byggnadshöjd för huvudbyggnad är 12,0 meter."*
Keyword matching: {variabelvarde: 1380, enhet: meter} (*false positive*)
LLM: {variabelvarde: 12.5, enhet: meter}

5. *Tillägg 1380K-Ä46 Bestämmelse tillkommer: Största byggnadsarea är 45 % av fastighetsarean*
Keyword matching: {variabelvarde: 1380, enhet: procent} (*false positive*)
LLM: {variabelvarde: 45, enhet: procent}

6. *Största tillåtna sammanlagda (ytor markerade med "e3") bruttoarea är 3600 m² ovan mark*
Keyword matching: {variabelvarde: 3, enhet: kvadratmeter} (*false positive*)
LLM: {variabelvarde: 3600, enhet: kvadratmeter}

7. *Största sammanlagda bruttoarea för fastigheten Karl XI 7 är 7800 m² ovan mark*
Keyword matching: {variabelvarde: 7, enhet: kvadratmeter} (*false positive*)
LLM: {variabelvarde: 7800, enhet: kvadratmeter}

8. *För fastigheter mellan 750 och 1000 kvm är största byggnadsarea 1/7 av fastighetsarean. För fastigheter större än 1000 kvm är största byggnadsarea 1/8 av fastighetsarean.*
Keyword matching: {variabelvarde: 750, enhet: kvadratmeter} (*false positive*)
LLM: {variabelvarde: 14.29, enhet: procent}

Table 4.2: *True negative* regulations extractions by keyword matching and LLM

1. *Största byggnadshöjd är 4,0*
Keyword matching: null
LLM: {variabelvarde: 4.0, enhet: meter} (*false positive for LLM*)

2. *Se planhandling. Därest byggnadszonen omfattas av flera tomter än en, skall byggnadsarean fördelas på tomterna i proportion till de ytor som den i stadsplanekartan inritad illustration upptager å respektive tomter.*
Keyword matching: null
LLM: null

3. *500*
Keyword matching: null
LLM: null

4. *Befintlig byggnad får ha det våningsantal som gällande lov medgivit före planens fastställelse*
Keyword matching: null
LLM: null

5. *Byggnadshöjden ska variera*
Keyword matching: null
LLM: null

6. *Högsta höjd på byggnadsverk är angiven som "Byggnad får inte uppföras till större höjd (byggnadshöjd) än 4,0 merter". Bestämmelsen har inte tolkats eftersom begreppet ”byggnadshöjd” inte längre omfattas i aktuell planbestämmelsekatalog.*
Keyword matching: null
LLM: {variabelvarde: 4.0, enhet: meter} (*false positive for LLM*)

7. *Högsta höjd på byggnadsverk är angiven som 12,7. Bestämmelsen har inte tolkats - äldre plan.*
Keyword matching: null
LLM: {variabelvarde: 12.7, enhet: meter} (*false positive for LLM*)

8. *Största sammanlagda bruttoarea är 18 000 kvmexklusive garage*
Keyword matching: null
LLM: {variabelvarde: 18000, enhet: m^2} (*false positive for LLM*)

Table 4.3: *False negative* regulations extractions by keyword matching and LLM

1. *Högsta höjd på byggnadsverk är angiven som "Byggnad får uppföras till högst 7,5 meters höjd". Bestämmelsen har inte tolkats eftersom begreppet "byggnadshöjd" inte längre omfattas i aktuell planbestämmelsekatalog.*
Keyword matching: null
LLM: {variabelvarde: 7.5, enhet: meter} (*true positive for LLM*)

Table 4.4: *False positive* regulations extract ed by keyword matching and LLM

1. *Sammanbyggda småhus uppdelade i tre grupper. Placering i huvudsak enligt illustrationskarta. Avstånd mellan husgrupperna minst 7 meter. Största byggnadsarea per bostad är 125 kvm* **Keyword matching:** {variabelvarde: 7, enhet: kvadratmeter} **LLM:** {variabelvarde: 125, enhet: kvadratmeter} (somewhat correct)
2. *Största antal lägenheter är 250, där bruttoarean i medeltal är högst 120 kvm per lägenhet exklusive balkong. För radhus/parhus gäller bruttoarea högst 180 kvm. Inglasning av balkong är tillåtet utöver angiven bruttoarea* **Keyword matching:** {variabelvarde: 250, enhet: kvadratmeter} **LLM:** {variabelvarde: 120, enhet: kvadratmeter} (somewhat correct)

4.4 Accuracy metrics

Accuracy is the percentage of all regulations for which the method's extraction matches the ground truth. Precision is the ratio of true positive extractions to the total number of extractions attempted. Recall measures how many of the true positives the method was able to find. Both techniques work reliably, but the LLM yields significantly better performance.

Table 4.5: Accuracy metrics using the two different methods

Model	Accuracy (%)	Precision (%)	Recall (%)
Keyword matching	97.9	99.1	98.6
LLM	99.6	99.6	100

4.5 Discussion

4.5.1 Data quality

The quality of the data is essential. This varies a lot for different municipalities and cities, which can be seen from the case studies. The major difficulty is information extraction from the plan regulations, when the regulations are not already manually extracted. There is a clear difference of how the regulations are expressed for different cities, since the regulations are written free-form. This is due to each municipality having their own way of expressing these regulations. This is not a problem in itself but it is important to consider when developing the methods for extracting the information from the regulation. If there is a simple action that would make the algorithm catch a lot of errors on a set of regulations in a city, this would not necessarily increase the accuracy for another set of regulations in another city. At the moment there are still a significant amount of cities that have not yet digitized their zoning plans so it is not possible to know the performance of the methods on future data.

The most common reason that a regulation end up in the *negative* category is that that they are missing a unit or a value, but it can also be due to ambiguous wording, misspelling or concatenation of words. There is a fine line wether the regulations should be categorized as positive or negative if words are misspelled or concatenated,

but also when a unit is not explicitly provided. The following regulations could have been extracted but are slightly ambiguous, in this study they have been placed in the negative category: The regulation "Största byggnadshöjd är 4,0" is obviously referring to the height in the unit meter and not amount of floors, given the decimal in the value. A regulation where words have been concatenated is "Största sammanlagda bruttoarea är 18 000 kvmexklusive garage", where the words *kvm* and *exklusive*, which is translanted to square meter and exclusive. A regulation that are misspelled is "Byggnad får inte uppföras till större höjd (byggnadshöjd) än 4,0 merter", where the word "merter" should have been written as "meter".

Recall is occasionally important for the model. An example is the regulation "Största sammanlagda bruttoarea är 18 000 kvmexklusive garage" which is a very crucial regulation, and it was overlooked by the semantic keyword matching model, but not by the LLM. The reason it was overlooked by the semantic keyword matching algorithm was because the unit square meter was not found, thus it could not confidently interpret the whole regulation. The unit is present in the regulation but it is not found because of a concatenation error in the regulation. Another regulation that the regular expression keyword matching algorithm did not manage to extract was the regulation "Byggnad får uppföras till högst 7,5 meters höjd". The reason is that the algorithm was searching for the word "meter" rather than "meters", this information could easily have been extracted by developing the algorithm to be more robust. If the user needs a high reliability of the key performance factors (KPI), it is important to capture these slightly ambiguous regulations as well. The regulation above is very important from a usefulness point of view, since it enables the exploitation of 18000 GFA, which might be the difference that makes the zoning plan very interesting.

Precision is arguably even more important than recall for the model, since to much noise could make the filter less ineffective. If many of the zoning plans that pass the filters the handful of plans that the user will manually go through might be noise. For the example where the regulation is: Tillägg 1380K-Ä46 Bestämmelse tillkommer: Högsta tillåtna byggnadshöjd för huvudbyggnad är 12,5 meter Bestämmelse upphävs: "Högsta tillåtna byggnadshöjd för huvudbyggnad är 12,0 meter.", the keyword matching algorithm are extracting the value of 1380 and unit meter, which infer that the building is allowed to be 1380 meter tall. This is likely to erroneously not be filtered out by the building height, or GFA parameter. A more sophisticated keyword matching algorithm could be developed where the number keyword that is closest in the sentence to the unit keyword is selected. To many of underestimated regulations could make the pool of promising zoning plans to large and the manual work to laborious.

This problem could also be seen as a regression problem where the error of the algorithm/model is the absolute difference between the estimated value and the ground truth value. For the regulation "Största tillåtna sammanlagda (ytor markerade med "e3") bruttoarea är 3600 m² ovan mark" it would make sense to put it in the *postive* category. It should be possible to extract the correct value 3600, but for the keyword matching algorithm that does not understand meaning and context, it makes as much sense to extract the value 3. Hence this regulation would fit better in a regression metric where the aim is to come as close as possible to the ground truth value. The reason why this was not investigated further was that the amount of regulations where this metric is relevant is very few. For regulations such as "Största antal lägenheter är 250, där bruttoarean i medeltal är högst 120 kvm per lägenhet exklusive balkong. För

radhus/parhus gäller bruttoarea högst 180 kvm" is it a lot to ask an algorithm to be able to first interpret the text and then compute the result successfully, even for a LLM. With respect to the efficiency of the tool and the scope of this study, regulations such as this are considered as inside the negative class, which means that they are preferably skipped.

4.5.2 Tool evaluation

Understanding the economic context of a zoning plan is crucial for assessing its development potential. One of the most direct indicators of market attractiveness is the price of nearby residential properties. By analyzing the prices of recently sold homes in the vicinity of each zoning plan, the tool estimates what people are willing to pay to live in the area. This price serves as a proxy for numerous soft qualities that are otherwise difficult to quantify. These include proximity to public transport, schools, parks, healthcare facilities, and cultural landmarks. Some areas may also command higher prices due to aesthetic factors such as views of the sea or architectural character. Cultural and historical character also play a role, as some neighborhoods have a distinct identity or atmosphere that adds to their appeal. These factors are deeply social in nature and subjective across individuals, making them hard to model directly. Although preferences vary greatly among individuals, the market price reflects the aggregate value that buyers assign to these factors.

This thesis advances the analysis of zoning plans beyond what has been achieved in previous research, largely due to the relatively high quality of available data in Sweden. The structured format of zoning metadata provided through national platforms such as the NGP, along with consistent planning practices and extensive digital records, creates a favorable environment for developing and evaluating automated analysis methods. This data enables a practically useful interpretation of parameters due to the relatively high data quality.

Despite its strengths, the tool's performance is partly constrained by the quality and coverage of available data, particularly within the national geodata platform. Continued digitization of zoning plans and improved access to comprehensive geospatial datasets would significantly enhance the tool's utility and scalability. One of the major limitations of the tool lies in the lack of consistent available data through the national geodata platform. While some urban areas are comprehensively covered with digitized and vectorized zoning plans, others have sparse or incomplete datasets. This uneven coverage directly affects the tool's ability to provide reliable results across all regions, and highlights the importance of continued national efforts in data digitization and standardization.

A central challenge in zoning plan analysis lies in estimating exploitation parameters, such as the maximum number of floors or the allowable building footprint (BYA). These values are occasionally not provided explicitly in the metadata. Instead, they are embedded in formulations, described through indirect references to building height limits or land use constraints. As a result, semantic keyword matching approaches sometimes struggle to identify these values, particularly when the phrasing or formatting deviates from expected patterns. This highlights the need for more sophisticated methods capable of interpreting the underlying intent of regulatory language. LLMs offer a significant advantage in this context. Instead of relying solely on exact keyword matches, they interpret the semantic structure of a sentence by

employing vector embeddings that capture the contextual meaning of words and phrases. This enables the model to recognize equivalent meanings expressed in different ways. For example, it can deduce that "byggnadshöjd får ej överstiga 12 meter" and "Byggnad får uppföras till högst 7,5 meters höjd" both convey constraints on vertical development, even though the phrasing and expression of the unit differ. Integrating LLMs into the analysis pipeline enables a more robust interpretation of zoning regulations. By focusing on extracting essential parameters such as building height limits or maximum footprint areas, even when these are not explicitly or consistently stated, LLMs enhance the accuracy of GFA estimations and overall exploitation assessments. In summary, this methodology supports more reliable filtering of zoning plans and enables the scalable identification of development opportunities across municipalities. Ultimately, it bridges the gap between rigid rule-based parsing and the nuanced, often ambiguous language found in planning documents.

4.5.3 Tool efficiency

Since the intended use of the tool is for early-stage site screening and exploration, there is no requirement for perfect accuracy. Users are generally looking to narrow down a large number of zoning plans to a manageable set of promising candidates, rather than making final investment decisions based solely on the automated output. Therefore, even though large language models (LLMs) offer a higher degree of interpretative precision, a simpler keyword matching algorithm may be sufficient for most users. The trade-off between simplicity, speed, and acceptable accuracy makes keyword matching a viable and pragmatic choice in many real-world applications, particularly when a slight margin of error is tolerable in exchange for rapid, large-scale filtering.

The current methodology for estimating building heights involves extracting Z-values from point cloud tiles based on building footprints. While effective for small-scale use, this approach becomes inefficient when scaled across hundreds or thousands of buildings, due to the time required to cache and process full 2,500 x 2,500 meter point cloud tiles per query, which is required even if only a small portion of the tile is queried. To address these performance challenges and improve scalability, efforts are currently underway to construct a centralized dataset of precomputed building attributes. This dataset would store key metrics such as building IDs, building footprints, building heights and ground level elevation. For tools like the one presented in this thesis, such a resource would eliminate the need for repeated data fetching and computations, thus enabling significantly faster analyses. Additionally, it holds potential value for broader research efforts related to urban planning, demographic analyses, simulations and digital twin applications.

5 Conclusion

This thesis has presented the development and application of a tool designed to automate the analysis of zoning plans in Sweden, with the goal of identifying underutilized land parcels suitable for development. By integrating geospatial data, development regulations and residential market prices, the tool enables efficient filtering and evaluation of zoning plans at a scale that would be unfeasible through manual processes alone. The methodology combines spatial analysis using geographic information systems with semantic interpretation of regulatory metadata. The use of large language models (LLMs) has proven particularly valuable in extracting meaningful parameters, such as building height limits or floor area from often unstructured or inconsistently formulated plan texts. The optimal presentation strategy depends on the intended user. Seasoned planners and analysts may favor a high-recall output that they can review and prune. Less technical stakeholders might prefer a concise shortlist with high precision, to avoid information overload. Implementing a hybrid approach is likely sufficient for both users. These insights are further enriched by local market data, allowing for a more informed assessment of economic viability in each location.

In informal pilot tests with a couple of architects, both reported that the tool was simple to use while requiring minimal input. The users appreciated the automatically generated interactive Kepler map for visualizing results. Case studies conducted in municipalities like Järfälla, Halmstad, and Sundsvall demonstrate the tool's effectiveness in revealing promising development opportunities, while also highlighting the variability in zoning plan availability across regions. In conclusion, the work presented here provides a foundation for data-driven site selection and urban development planning. It demonstrates how automation and intelligent data interpretation can support more strategic, evidence-based decisions in a field that still heavily relies on manual analysis. Future improvements could focus on expanding building height data, refining regulatory interpretation through improved LLM training, and integrating more nuanced urban quality indicators into the analysis.

5.1 Further research

While this thesis is focused on presenting a functional prototype for analyzing development potential based on zoning plans, several extensions can enrich the scope and utility of the tool. These include terrain evaluation, infrastructure availability, proximity to amenities, alignment with municipal plans and demographic considerations.

5.1.1 Terrain analysis

Topography can significantly impact construction feasibility and cost. Terrain analysis can be conducted by transforming point cloud data into a mesh. The slope of the terrain is then estimated by calculating gradients at selected points across the mesh. The average of these gradients provides a representative measure of the overall terrain slope for each zoning plan.

5.1.2 Infrastructure

Evaluating infrastructure involves assessing the availability and accessibility of water, sewage, electricity, and roads. While a comprehensive infrastructure inspection is complex and beyond the current scope, a proxy approach is to analyze the proximity of the zoning plan to existing residential developments. A high concentration of neighboring residential areas suggests that basic infrastructure may already be in place, lowering the cost and complexity of new developments.

5.1.3 Proximity analysis

Proximity to amenities such as schools, healthcare, green spaces and public transport affects both the desirability and sustainability of a development. This analysis involves computing distances from each zoning plan to these facilities. A cumulative score can be derived by applying weighted sums to these distances, based on their relative importance.

5.1.4 Alignment to overview plan

Each of Sweden's 290 municipalities maintains a comprehensive overview plan (*översiktsplan*) that outlines long-term land use strategies. Lantmäteriet are in the process of collecting and digitizing these documents into a structured database. By employing a Retrieval-Augmented Generation (RAG) model, an LLM could be used to extract relevant excerpts and assess how well a development aligns with municipal goals. This would provide not only regulatory insights but also strategic guidance for developers.

5.1.5 Demography

Understanding local demographics is crucial for tailoring development projects to the needs of the population. Key indicators might include age distribution, income levels, employment rates, and population growth. Integrating demographic datasets with zoning plan analysis could help prioritize projects that meet emerging demands, such as housing for young families, elderly residents, or student populations. This layer of insight would support more inclusive and adaptive urban planning.

6 References

Axelrod, J., Lo, L., & Bronin, S. C. (2023, February). *Automating zoning data collection: Results from a pilot effort to automate national zoning atlas methodologies* (tech. rep.). Urban Institute. https://www.urban.org/research/publication/automating-zoning-data-collection

Booli. (n.d.). *Booli – sveriges största samlade utbud av bostäder till salu* [Accessed May 27, 2025]. https://www.booli.se/

Boverket. (n.d.). *Planhandlingar i detaljplan* [Accessed May 27, 2025]. https://www.boverket.se/sv/PBL-kunskapsbanken/detaljplan/handlingar/

DeepMind. (n.d.). *Introducing gemini: Our largest and most capable ai model* [Accessed May 27, 2025]. https://deepmind.google/models/gemini/

Digital Twin Cities Centre. (n.d.). *Digital twin cities centre – a vinnova competence centre* [Accessed May 27, 2025]. https://dtcc.chalmers.se/

Lantmäteriet. (2009). *Gauss conformal projection – grid coordinates in the swedish reference frame sweref 99* (tech. rep.) (Retrieved May 27, 2025). Lantmäteriet. https://www.lantmateriet.se/globalassets/kartor-och-geografisk-information/geodesi-och-swepos/sweref99-projektion_eng.pdf

Nationella geodataplattformen. (n.d.). *Nationella geodataplattformen* [Accessed May 27, 2025]. https://www.lantmateriet.se/sv/nationella-geodataplattformen/

Plan- och bygglag (2010:900) [Accessed May 27, 2025]. (2010). https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/plan-och-bygglag-2010900_sfs-2010-900/

Salazar-Miranda, A., & Talen, E. (2025). Zoning in american cities: Are reforms making a difference? an ai-based analysis. *Nature Cities*, *2*, 304–315. https://doi.org/10.1038/s44284-025-00214-0

Zheng, Y., Liu, L., Lin, Y., Feng, J., Zhang, G., Jin, D., & Li, Y. (2025). Urbanplanbench: A comprehensive urban planning benchmark for evaluating large language models [arXiv preprint arXiv:2504.21027]. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. https://arxiv.org/abs/2504.21027

Appendix A

Code snippets and Kepler maps are available on GitHub:
`https://github.com/Misac98/Assessment-of-Zoning-Plans`

DEPARTMENT OF ARCHITECTURE AND
CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY