ESSnet Trusted Smart Statistics – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 3

New Use-cases

Deliverable 3.6:

Reports on methods and feasibility to track construction activities based on real estate web portals

Version, 2025-03-06

Prepared by:

UC coordinators:
Tobias Gramlich (HSL, Germany)
Alexandra Ils (HSL, Germany)

Contributors:
Kerstin Erfurth (SSI-BBB, Germany)
Pär Hammarström (SCB, Sweden)
Nicole Jurisch (SSI-BBB, Germany)
Dr. Gitta Lasslop (HSL, Germany)
Dr. Holger Leerhoff (SSI-BBB, Germany)
Andreas May-Wachowius (SSI-BBB, Germany)
Pieter Vlag (SCB, Sweden)

This document was funded by the European Union.

Table of Contents

1. Background

This document is part of the Work package 3 (WP3) *New use-cases* from the ESSnet Trusted Smart Statistics – Web Intelligence Network project (ESSnet TSS-WIN). The overall objective of WP3 is to explore the potential of new types of web data sources for official statistics, with each use-case focused on a specific use-case (UC). The set of use-cases being explored are:

- **UC1** Characteristics of the real estate market
- **UC2** Construction activities
- **UC3** Online prices of household appliances and audio-visual, photographic and information processing equipment (and generalising the data-collection to other activities)
- **UC4** Experimental indices in tourism statistics (hotel prices)
- **UC5** Business register quality enhancement
- **UC6** Faster Economic Indicators using new data sources

This deliverable focusses on UCS specifically and describes methods, data sources and the feasibility to gather and analyse data on construction activities in the context of official statistics on construction activities.

This report starts with an introduction to the Use-case, which is thematically split into three parts. The first part highlights the role of the official statistic on construction, the overall aim of this use-case, as well as raising a few remarks on the connection between use-case 1 and use-case 2. The second part of the introduction is concerned with a basic summary of the use-case, where the definition of "newly constructed" properties, the real estate web portals, from which the data is collected, as well as a minimal set of indicators to extract and the various IT Choices of the partners are presented. Lastly, the introduction presents overarching questions regarding quality aspects of the data, as well as drawing a conclusion. Following the introductory section, the partners HSL, SSI-BBB and SCB present their country specific data sources, the data preparation as well as results, and discuss quality aspects, such as missings, duplicates and coverage. The report closes with a conclusion. [1]

2. Introduction

This use-case investigates methods, data sources and the feasibility to gather and analyse data from real estate web portals on construction activities in the context of official statistics on construction activities. Overall, three partners worked on this use-case: Hesse Statistical Office (HSL, Germany), Statistical Office Berlin-Brandenburg (SSI-BBB, Germany), Statistics Sweden (SCB, Sweden). Partners have been linked by the common goal of this use-case, however, work is organized independently since resources and conditions are different for each partner. Organizational structures, staff and IT resources, vary, as well as the situation on the real estate market for each partner regarding types of as well as relevance of different real estate web portals.

This use-case was closely related to use-case 1: two partners (HSL, SSI-BBB) also contributed results to use-case 1. Both use-cases focused on scraping the same or similar websites, the same data sources. This means that for these partners not only have been data sources the same, but also many of the decisions and steps necessary to gather this data as well as challenges collecting, processing and analysing this data. The main difference between use-case 1 and use-case 2 has been its focus on new buildings and aims to derive an early indicator for 'construction activities' on the one side, and

[1] The authors are grateful for appreciative and valuable comments made by a reviewer to an earlier draft of this report. We gladly accepted and incorporated comments whenever in scope of this use-case. We would like to express our appreciation for the efforts to improve this report.

the focus on characteristics of the real estate market in general (e.g. trends in prices) on the other side.

2.1. Official Statistics

This use-case is situated in the context of construction statistics. Typically, there are several official statistics in this area: (1) statistics on construction permits for residential and non-residential buildings. Results of this statistic are published monthly. (2) Statistics on new buildings, and (3) statistics on conversions and demolitions. The yearly statistic on newly constructed (residential) buildings (2) forms this use-case's official statistics background.

Statistics on construction activities are an important aspect to understand and picture the economic as well the social situation of a region or a whole country. Construction activities not only reflect the current strength of a country's economy, but also serve as an indicator of growth. Together with other data, these statistics are an important basis for political as well as economical decisions.

It is expected that data from real estate web portals can provide information on construction activities with sufficient coverage of the population of interest, with higher spatial accuracy and possibly at earlier stages than official statistics can: often, construction projects are advertised even before a formal permit is finally granted. Therefore, there is a delay in the data collection of the official statistics on newly constructed buildings. Often, construction of buildings has been completed during the year but the report of this completion is filed at the end of the year. As a result, in Germany, the yearly statistic on newly constructed buildings is published in the second quarter of the following year (typically in May). By using data from real estate web portals, the results from this data source could even be published just a few days after the end of each reference period (e.g. at the beginning of January). While official statistics is currently published yearly, using data from real estate web portals could potentially even allow for more frequent publication of results, such as monthly, to offer more frequent up-to-date data.

Many national official statistics are based on European legislation, for example the monthly statistic on building permits, which is part of the Regulation (EC) 1165/98. For the statistic on newly constructed buildings, there is no common specific EU legislation but only specific national legislation (e.g. in Germany the "Hochbaustatistikgesetz"). Consequently, the legal basis for two of the three partners (SSI-BBB and HSL) is identical. While for SCB similar national legislation applies, the situation for Germany and Sweden is comparable, and as a result, products of official statistics look very similar. [2]

[2] For Sweden, see https://www.scb.se/en/finding-statistics/statistics-by-subject-area/housing-construction-and-building/housing-construction-and-conversion/statistics-on-building-permits-new-construction-and-conversion/ for more information about the official statistics context. The quality report is available here https://www.scb.se/contentassets/34907586e4474520bd074ffdeb0de61b/bo0102_kd_2018_201217_engelsk.pdf. Tables can be found in the Statistical Database here https://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__BO__BO0101__BO0101A/LagenhetNyAr/. A For Germany in general, see https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Bauen/_inhalt.html. For Germany, the quality report is documented here https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bauen/baufertigstellungen.pdf?__blob=publicationFile. Results at the level of the German federal states (NUTS 1) are available here https://www-genesis.destatis.de/genesis/online?sequenz=statistikTabellen&selectionname=31121#abreadcrum. Results on regional level (NUTS 3, LAU) are available here https://www.regionalstatistik.de/genesis/online?operation=statistic&levelindex=0&levelid=1727694523186&code=31121#abreadcrumb

2.2. Overall Aim

The goal of use-case 2 'Construction Activities' was to derive early estimates or forecasts of construction activities based on data from real estate web portals. More specifically, the purpose of use-case 2 was:

- to provide a method for collecting data for early estimates of construction activities,
- to provide a method for early estimates of construction activities,
- to carry out real-time observation of trends on construction activities in general, on prices for objects for sale or rent, on sizes (area and rooms),
- to provide estimates of construction activities with spatial accuracy reaching lower levels of territorial aggregation (NUTS Level 3) beyond administrative borders (e.g. postal code, city district).

2.3. Limitations

Use-case 2 is strongly related to use-case 1 regarding data sources (two partners of use-case 2, HSL and SSI-BBB, contribute results based on collected data this use-case to use-case 1). However, while data sources for use-case 1 and use-case 2 are identical their target population differs: data collection and processing for use-case 2 focuses on the subset of newly constructed objects and not on all residential buildings as in use-case 1. More specifically, the focus of use-case 2 lies in newly constructed residential buildings. This excludes existing buildings, renovated buildings, as well as business buildings without any living space. Buildings with mixed use (business premises with at least one apartment) are part of the study population.

One can assume that data from internet portals does not cover the total population of all / all newly constructed residential buildings by design. Some of the newly constructed buildings will appear on real estate web portals only many years after construction, since they are constructed and used by their owners for many years. Having in mind that official statistics on newly constructed residential buildings cover all residential buildings – and not only buildings available on the market but also self-owned – then some amount of undercoverage can inherently be expected by comparing results from this data source to official statistics. However, if there is a stable relationship between data from the web (with undercoverage) and official statistics (no undercoverage), data from real estate web portals can serve as an indicator for official statistics.

Advertisements can describe one or more objects (e.g. within a larger construction project). For deduplication, the full address is the most important variable. For many – but not all – advertisements, full address information is given. However, due to legal restrictions, it is not possible to use and link official micro data but only published aggregates for deduplication or comparison.

2.4. Basic Summary

2.4.1. Defining 'Newly Constructed Properties'

Use-case 2 is about construction activities. This means that among all advertisements for houses and apartments appearing in real estate web portals, a decision needed to be made whether an advertised object refers to a "newly constructed building" or not. In principle, there are several criteria that must be fulfilled by an advertised object must in order to be considered as newly constructed:

- A building year is available and corresponds to the reference year (2022, 2023 or 2024). By itself, this would be a rather weak criterion and would include the risk of overcounting the same object when being advertised two or more times within the year of reference.

- Additionally, objects in recently constructed buildings often are labelled as “New Building” (e.g. “Neubau” in Germany). Again, relying only on this criterion would include the risk of overcounting objects when advertised several times within the year of reference.
- To avoid overcoverage and overcount, advertisements should refer to objects that are available for the first time. In many portals, this corresponds to the “condition” of “First Occupancy” (“Erstbezug” in Germany) of the advertisement.

In general, it has been decided to use a narrow definition of “newly constructed building”. This means that the year of construction may be missing or must refer to the year of reference but an advertisement must meet the condition of “First Occupancy”. This strict definition may lead to false negatives, i.e. advertisements from the portal being rejected that in fact belong to the target population. This happens for example if the year of construction refers to the reference year but the more specific reference to the condition of a “First occupancy” is missing. This seems to be favourable to the risk of including numerous false positives and duplicates – also due to the fact that it has been shown that identifying duplicates as well as deduplication is hardly possible in many cases.

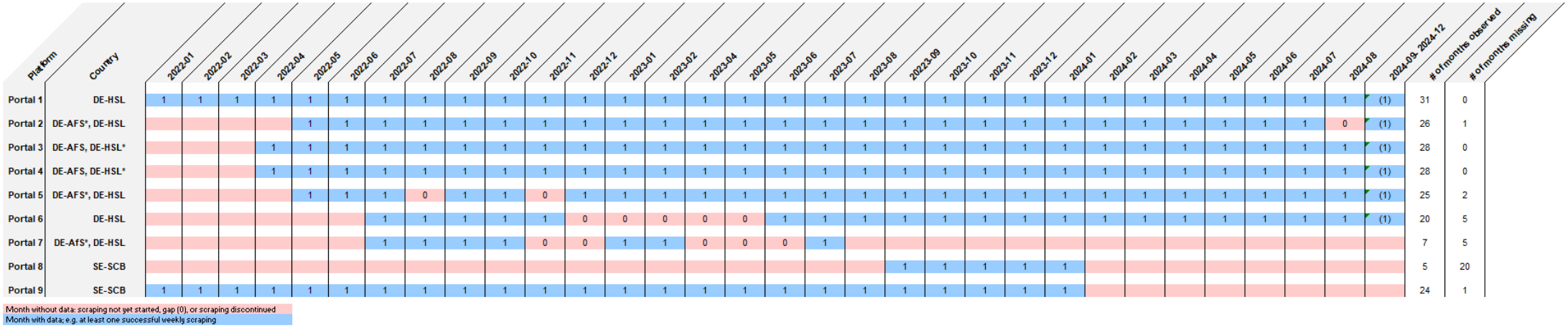
Official statistics typically use a specific concept of “newly constructed houses or apartments”. Typically this means that some specific construction steps have been finished (“Rohbau” or “unter Dach” in German) – and typically this is some time before people can move into an apartment or house to actually live in it. On the other side, advertisers are free when to advertise an object – developers typically advertise objects (long) before constructions even being. At the latest, objects have to be advertised in the year of construction to be counted as “newly constructed” for this use-case. It is assumed that the date / year of construction of an advertised newly constructed object refers to a state of the object when someone actually could move in and live in this apartment or house (so this is not a “Rohbau” anymore).

2.4.2. Portals and Data Collection

Overall, for the two German offices, HSL and SSI-BBB, there are seven real estate web portals covering the three states / NUTS1 areas Berlin, Brandenburg and Hesse. For Sweden, there are two data sources used during the project. In general, sources cover the most popular portal in each country. Additional sources have been chosen in order to investigate possible overlap and undercoverage by using only one or two sources, or to highlight special characteristics of one of the additional sources, say a specific regional focus or a focus on larger construction projects.

Figure 1 shows the collected dataset for use-case 2 over time.

Figure 1: Overview over data sources of UC2 by month



Platform	Country	2022-01	2022-02	2022-03	2022-04	2022-05	2022-06	2022-07	2022-08	2022-09	2022-10	2022-11	2022-12	2023-01	2023-02	2023-04	2023-05	2023-06	2023-07	2023-08	20223-09	2023.10	2023.11	2023.12	2024.01	2024.02	2024.03	2024.04	2024.05	2024.06	2024.07	2024.08	2024-09- 2024-12	# of months observed	# of months missing
Portal 1	DE-HSL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	(1)	31	0
Portal 2	DE-AFS*, DE-HSL					1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	(1)	26	1
Portal 3	DE-AFS, DE-HSL*				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	(1)	28	0
Portal 4	DE-AFS, DE-HSL*				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	(1)	28	0
Portal 5	DE-AFS*, DE-HSL					1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	(1)	25	2
Portal 6	DE-HSL							1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	(1)	20	5
Portal 7	DE-AfS*, DE-HSL							1	1	1	1	0	0	1	1	0	0	0	1															7	5
Portal 8	SE-SCB																				1	1	1	1	1									5	20
Portal 9	SE-SCB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1									24	1

Month without data: scraping not yet started, gap (0), or scraping discontinued
Month with data; e.g. at least one successful weekly scraping

Overall, for Hesse, one data source completely covers the years 2022 and 2023 (Portal 1). Two other sources (Portals 3 and 4) cover the year 2023 for Hesse, Berlin and Brandenburg. During the project, because of a large overlap with another portal (Portal 2) it has been decided to stop reporting results from Portal 7 but to continue scraping advertisements from the portal (with lower frequency) in order to keep track of changes in this portal in case its counterpart stops working. For Sweden, portal 9 has the most continuous data collection; however, it ends in early 2024. Gaps in the data are mostly due to changes in the portals' site structures, for which scrapers needed to be adapted.

Portals have been chosen according to their size / relevance for the online real estate market (market leaders) as well as their specialization (developers, regional focus, no agency). For this use-case, also accessibility was important in order to see how frequent changes in the site may occur – with or without affecting scrapers. This also means that there may be some selection bias in the choice of portals.

For the data collection, different approaches to scraping were used:

- **Screen scraping**: automated tools for webscraping visit each portal's start page and gather data from overview pages (e.g. for specific regions) to extract links to the individual advertisements. This requires a thorough inspection of the website, its source code as well as network traffic in order specify XPATH expressions that extract the relevant information from each overview page or individual advertisement page. Additionally, extracted data has to be processed and to be put in tabular form. This automated process is very similar to a person manually visiting each and every relevant page using a browser.
- **API**: sometimes, portal providers offer an access to their data that is specifically designed for programmable software. By using an API, screen scraping is not necessary. Instead, it is possible to request specific information directly from the server, without the need to record and extract information from source code for display in a browser window. Typically, this way of data collection is often much faster and more efficient than screen scraping. Typically, using an API, data can be extracted more easily and needs less processing. Sometimes, when starting to inspect a website for screen scraping, it turns out that there is actually an API available (but often not described) that provides the data displayed on the screen that would otherwise be extracted with screen scraping.
- **Agreement**: portal providers may provide access to data themselves, commercially or non-commercially. HSL has entered an agreement with one of the largest portal providers to receive a monthly dataset containing all new advertisements for objects in newly constructed buildings (for Hesse).

2.4.3. Minimal Set of Indicators

Typically, advertisements for residential buildings or apartments independently of country or real estate web portal share a set of common attributes they present to potential customers. This leads to the idea of a set of "mandatory variables" to collect for analysis.

In general, such a set of variables consists of object ID or advertisement ID, type of advertisement (object for rent or sale), type of building (apartment, detached house, semi-detached house), number of rooms, size or surface in m^2. For identifying duplicates as well as deduplication, addresses are needed. For regional analyses as well as comparison with official statistics, at least some indicator of region is necessary. Additionally, many other characteristics would be of interest to produce experimental statistics or for widening the scope of official statistics that already exist in the area of construction activities. For instance, the type of heating as well as energy consumption could be of

great interest, as well as specific amenities or characteristics of an offered object, e.g. barrier-free access to a building or apartment, barrier-free bathroom and so on.

However, as soon as scraping started and first analyses of data began, it became clear that for all of these “mandatory variables”, missing data occurs (since it is also not mandatory to advertisers to provide such information). Excluding advertisements with any amount of “mandatory” information missing, would discard a large portion of data that would otherwise potentially widen the scope of official statistics. However, it turned out that for characteristics other than a minimal set of indicators the amount of missing data can become prohibitively high. More concretely, some portals or advertisements may allow specifying the information about an accessible entrance to the object or an accessible bathroom in a standardized way – but the information given may not be very reliable, meaning just because the information is not provided does not mean that the amenities do in fact not exist.

Advertisements on real estate web portals typically present a large list of information to potential customers. Many of them are presented in a standardized way, which in terms of scraping means that for every advertisement on a specific portal, certain information appears at the same position on the page, i.e. they all have the same XPATH. When there is no standardized way of presenting information as to whether an apartment has specific amenities (e.g. a balcony by including a specific icon at a specific location in an advertisement), additional information often can be specified as free text in larger text boxes. These text boxes again typically have a standardized position on the page while their content is not standardized. Checking for both, standardized XPATH extracts or extracts from free text boxes, advertisements typically contain information about dozens or hundreds of amenities. As a rule, however, only few of these characteristics are common to all or most of the advertisements on real estate web portals.

Overall, the minimum set of indicators that is collected and delivered to use-case 1 contains:

- Unique object / advertisement ID: necessary to exclude obvious duplicates within one data source.
- Access time: necessary in order to decide which duplicate advertisement is the first / last one collected. Additionally, an analysis such as a “duration analysis” would need this information in order to record the time span an advertisement is available online.
- City name / address / geo coordinates: for deduplication within as well as between portals, address information is necessary. Additionally, results (number of advertisements, number of newly constructed buildings and apartments) have been aggregated to NUTS 3 level for comparison with official statistics.
- New building: to identify newly constructed buildings, information, such as the year of construction of an object or some information on the condition of the object (e.g. “new building”, “first occupancy”, but not “first occupancy after renovation”), need to be given in the advertisements.

In use-case 2, a specific minimal set of variables is required to derive the main indicator of construction activity. Additionally, optional variables have been used for additional analysis or have been helpful to investigate or improve the quality of the main indicator. In general, this set of minimal variables is sufficient. However, it turns out that missing information within these variables poses challenges. Especially complete address information or location of an object within a house may be missing.

2.4.4. IT Choices

Each of the partners in use-case 2 was free to decide which tools to use to accomplish the task of data collection as well as data processing and aggregation.

SSI-BBB and HSL shared some of the data sources. It was decided that offices should share the workload by being responsible for development, application, and maintenance of software for specific portals. By sharing workload resources could be saved effectively.

The following table 1 summarises the IT-choices and solutions per partner, which are explained in more detail below.

Table 1: Comparison of IT choices and solution of UC2 partners

WP3 Partner	**New data sources exploration**	**Programming, production of software**	**Data acquisition and recording**	**Data processing**	**Modelling and interpretation**	**Dissemination of the experimental statistics and results**
IT choices and solutions						
HSL	*R*	*R/Python*	*R/Python* *Scrapy* *CSV*	*R tidyverse*	*R tidyverse*	*R tidyverse*
SSI-BBB	*R, Python*	*R, Python*	*R, Python, rvest, polite, httr, xml2, tidyverse, xpath, Scrapy, CSV, json,* paho-mqtt, requests, jmespath	*R, Python, Pandas, numpy, os, re tidyverse*	*R, Python, Jupyter Notebook, Pandas, scikit-learn, tidyverse,* matplotlib,	*R, Python, Pandas, tidyverse*
SCB	R,SQL, Python	R, Python	R: httr, rvest Python: requests, json	R: Tidyverse Python: Pandas	R, Python	*R tidyverse*

HSL

Within HSL, the following environment has been used to set up and run the scrapers:

- Windows Server 2012
- Python 3.8.3 – Modules: re 2.2.1, csv 1.0, unicodedata, bs 4 4.9.1, dataclasses, Scrapy 2.5.0

Necessarily for scraping, this environment is connected to the internet. After scraping, the gathered data is transferred to the internal area for further processing and analysis.

HSL used the *Scrapy framework* (https://scrapy.org/) to develop the scrapers for the real estate web portals. The Scrapy framework is available as a Python package. The Scrapy framework also offers important parameters or options in a specific settings file, which the HSL applied to lower the risk of attracting negative attention by the portal providers or hinder other users of the portal. More specifically, parameters have been used to define a unique

- User Agent (USER_AGENT): In the HSL implementation, scrapers identify themselves to the servers by means of a user-agent string in the HTTP header of the page request, leading to a webpage (https://statistik.hessen.de/ua) containing information about the origin of the scraper, purpose of scraping as well as contact information.

- Robots.txt (ROBOTSTXT_OBEY=TRUE): The HSL adhered to the guidelines set by a website in its robots.txt.
- Download delay (DOWNLOAD_DELAY=3): The download delay was set to 3 seconds, meaning that between the individual page requests to the same server the scraper pauses for 3 seconds. This delay is introduced to make sure that the scraping is not interpreted as an attack to the server, when there are many requests to the same server. While this leads to significantly longer scraping times, waiting times are an effective measure in order to avoid being mistaken as an attack to the server(s).

Additionally, scraping takes place at early morning hours when it is assumed that only few other users access servers. The HSL contacted and informed portal owners at the beginning of the project about the scraping of their portals. As of now, no partner has reported complaints of portal providers or problems with blocking of the scraper.

The scraping process has been divided into two parts. The first part of the scraper extracts the URLs of the individual exposes from overview pages. The second part uses these URLs as input and finally retrieves the information on the specific real estate objects (project, house, apartment).

The first part uses the spider type CrawlSpider, which is efficient in retrieving URLs. The spider starts from a specified starting URL, then craws through the portal and retrieves URLs according to rules that are specifically defined for each individual portal. The rules for all implemented portals have been specifically defined in a user-defined class „ImmoPagesDict".

The second part uses the most simple spider class scrapy.spider. The spider reads the URLs that have been retrieved in the first step. The HTML code of the URLs then is retrieved one after the other and the desired information on location, construction year etc. is extracted. The class „Item" serves functions as a container for the scraped data. A class „ItemLoader" provides a mechanism to populate the item object with the scraped data. A portal-specific dictionary defines how the information for a specific variable can be retrieved from the HTML source code, providing the portal specific XPATH. While there is some basic pre-processing by the scraper, more data processing and cleaning follows in an additional step after scraping and exporting data to text files (CSV).

In summary, HSL developed a multi-portal web scraper that could be extended easily to other portals by adding rules for the URL extraction and a dictionary with information on how to retrieve the information for the individual variables from the expose page. The HSL scraper has collected data weekly for three portals with only minor adaptions to respond to changes on one portal's website.

SSI-BBB

SSI-BBB used the following environment:

- Windows 7 – Workstation machines
- Python 3.8.12 – Libraries datetime, os, pickle, glob, requests, urllib, numpy, pandas, scipi, scrapy, beautifulsoup4 4.10.0, glob2 0.7, requests 2.26.0, urllib3 1.26.7, numpy 1.21.2, pandas 1.3.4, scipi 1.7.1, scrapy 2.4.1, matplotlib, paho-mqtt, requests, jmespath
- R 4.0.2, xml2, rvest, polite, httr, tidyverse

Given the experimental nature of the work package, SSI-BBB adopted a comprehensive approach to the webscraping task. This strategy allows for the comparison of various tools regarding user-friendliness and facilitates the development of broader expertise in this domain. Three distinct approaches utilizing different web scraping tools and technologies were evaluated for scraping the portals relevant to both use-case 1 (UC1) and use-case 2 (UC2), due to the significant overlap in their offerings:

1. **R 4.0.2 with rvest and xml2**: For extracting information from the well-structured portal Portal 5, R was employed in conjunction with the rvest package and the XPath expressions provided by the xml2 package. This combination enables precise navigation to specific HTML elements and the evaluation of results. However, it should be noted that modifications to the R code may become time-consuming in the event of changes to the HTML structure.
2. **Python 3.8.12 with Scrapy**: The Scrapy framework, which was also employed by HSL, was applied to Portal 2. Scrapy offers extensive configuration options and is regarded as a convenient solution for web scraping tasks. For further details regarding this approach and the configuration of the Scrapy framework, refer to HSL's section on Scrapy.
3. **Python with API Requests**: This approach was utilized to access data via the API from Portal 7. Given that the data from this portal largely overlaps with that of Portal 2, it was used primarily as a backup source to address any data gaps that may arise from inaccuracies in scraping Portal 2.

Since all three approaches generate plain CSV files (or alternative formats such as JSON) for subsequent data analysis, utilizing different technologies does not pose any significant challenges. In summary, the decision between approaches (1) and (2) largely hinges on personal preference regarding the choice of tools (R vs. Python) and is deemed appropriate for smaller portals and rapid prototyping. Conversely, approach (2) emerges as the preferred solution, particularly for scraping larger portals to gain more information. However, two notable drawbacks of this approach include its difficulty in integration within Jupyter Notebook environments and its steep learning curve.

Additionally, it is important to acknowledge the general advantages and disadvantages between screen scraping (1) and (2) and API scraping (3), which will be discussed below. Screen scraping often allows for greater flexibility in data extraction from web pages that do not provide an API, but it can also lead to challenges related to data consistency and maintenance due to changes in HTML structure. In contrast, API scraping tends to offer more reliable and structured data access, though it may be limited by the availability and scope of the API endpoints provided by the data source.

Advantages of APIs over Screenscraping

APIs offer several significant advantages over screen scraping, particularly in terms of efficiency, reliability, and ease of integration. One of the primary benefits is the reduced maintenance and higher stability associated with APIs. Unlike web scraping, which requires constant monitoring and adjustment due to changes in webpage structure or content, APIs are less prone to breakage because they are designed specifically for machine communication. While website designs and formats may change frequently, affecting the structure of scraped data, APIs tend to be more stable, as changes in website design usually do not disrupt the underlying API.

Another major advantage is that APIs provide structured and machine-readable data. APIs typically deliver data in formats such as JSON or XML, which are well-suited for automated processing and can be easily integrated into other systems. This stands in contrast to screen scraping, where the data is often unstructured and embedded in the HTML of a webpage, making it harder to parse and requiring additional steps to extract the relevant information.

APIs also enable faster and more efficient data retrieval. For example, APIs can return large volumes of data in a single request, reducing the number of necessary queries. In contrast, screen scraping often requires multiple individual requests to extract each element separately, leading to higher traffic and slower data acquisition. This efficiency is especially important when handling large

datasets, such as retrieving hundreds of property listings, which can be done in one API call instead of several separate scrapes.

In terms of scalability, APIs are inherently designed to support high-volume data retrieval, making them better suited for applications requiring large datasets or high-frequency updates. Screen scraping, by comparison, can struggle with performance issues as the scale increases, often leading to bottlenecks or timeouts.

Furthermore, less code is required when working with APIs, especially compared to custom-built screen scraping solutions. While Python-based tools like Scrapy can streamline the scraping process, developing a robust and adaptable screen scraping tool often requires more extensive programming effort than integrating an API, which typically involves simpler calls and data handling.

Limitations of API Usage

Despite these advantages, there are also several limitations associated with API usage. One key drawback is the requirement for an available API interface. If an API does not exist for a given data source, screen scraping may be the only option to access the desired data. While APIs offer structured access, their availability is not guaranteed, particularly for smaller or less technologically advanced websites.

Another potential issue is the need for comprehensive API documentation. Well-documented APIs provide detailed information on endpoints, methods, parameters, response formats, error codes, and authentication processes. In the absence of such documentation, utilizing and integrating the API can become difficult and time-consuming.

Authentication and access represent another challenge. Most APIs require some form of authentication, typically through API keys or tokens. In some cases, APIs may be gated behind paywalls or require specific access permission. However, in the context of our use-case, a public key provides access, and there are no associated costs for using the API. This may not always be the case with other APIs, where usage fees can apply.

Lastly, APIs may offer limited information compared to screen scraping. In some scenarios, APIs do not provide the full range of data available on the website, which can be a disadvantage if more detailed or niche information is required. In our case, the API returns fewer data points than what might be accessible via screen scraping, but this limitation is context-dependent and does not apply universally to all APIs.

Overall, APIs are typically the preferred method for extracting data due to their stability, efficiency, and ease of use, particularly for large-scale or automated applications. However, the limitations of API availability, documentation, and potential data restrictions should be considered carefully, especially in cases where screen scraping may provide more flexibility.

SCB

Within SCB, two different web scraping solutions were implemented to retrieve and process data from the real estate web portal website. Each solution employs a different approach, with one utilizing HTML scraping through R and the other leveraging the portal's GraphQL API with Python.

HTML scraper (R)

The first solution is built using R 4.2.1 and designed to scrape, parse, process, and integrate real estate data into a SQL Server database.

- Key R packages used:

- `dplyr` for data manipulation
- `readr` for reading data
- `crul` for HTTP requests
- `tibble` for data frames
- `rvest` for web scraping
- `stringr` for string manipulation
- `tidyr` for tidying data
- `odbc` and `DBI` for database connections
- `lubridate` for date-time manipulation

This web scraper reads advertisement listings directly from the HTML content of web pages. It constructs URLs and retrieves data from specified pages, setting necessary HTTP headers and parsing the HTML to extract advertisement links and summaries. The newly scraped data is then compared with existing database records to identify new and existing advertisements. Detailed data is extracted from individual advertisement pages, classified, and cleaned using regular expressions to ensure consistency and accuracy. The cleaned data is saved as parquet files for backup, and the SQL Server database is updated with new and existing advertisements.

GraphQL API Scraper (Python)

The second solution is built using Python 3.9.7 and designed to gather real estate data using the real estate's GraphQL API. This scraper is particularly efficient due to its use of structured GraphQL queries, which allow for more precise data retrieval. By querying the API, we fetch only structured and necessary data.

- Key Python libraries used:
  - `requests` for handling HTTP requests
  - `json` for parsing JSON responses
  - `pandas` for data manipulation and analysis
  - `time` for time-related functions

The web scraping process is encapsulated in a function which retrieves data based on the type of object (either "sold", "for_sale", or "projects"). The function accepts several parameters, including the page range, area, and a creation date range to limit the query to only read adverts that have not already been read.

Key parameters and options are set within the function:

- URL: The real estate web portal GraphQL endpoint.
- Payload: The GraphQL query and variables are dynamically constructed based on the type of object (sold or for sale) and other parameters.
- Headers: Necessary headers for the HTTP request.

The function iterates through the specified page range, sending POST requests to the API and collecting the results. The JSON responses from the API are parsed, and relevant data is extracted and stored in a list. Finally, the retrieved data is normalized into a pandas DataFrame for further processing and analysis.

Comparison

The two solutions differ significantly in their approach. The Python-based scraper uses a GraphQL API, which provides a structured and efficient way to retrieve data associated with real estate objects

that are for sale and that have been sold. This method is faster and less prone to errors related to changes in the website's layout. On the other hand, the R-based solution scrapes HTML content directly from web pages, potentially offering greater flexibility in terms of data access.

Summary

Developing and maintaining webscrapers for this use case requires and required skilled staff familiar with a programming language, like R or Python here. Many statistical offices built and build up knowledge in webscraping in other areas e.g. for online based enterprise characteristics (OBEC), enhancing (business) register quality, supporting NACE classification of businesses. This use-case has been conducted not in the light of production, so there has been put less emphasize on infrastructure that would be necessary for production. Accordingly, no suggestions should and can be made regarding infrastructure.

3. General Quality Aspects

There are some general data quality aspects to consider when working with webscraped data as data source.

- **Specification error**

There may be a mismatch between the targeted concept as used by official statistics (newly constructed properties) and the concept underlying behind the mechanism of advertising apartments and houses on online real estate platforms. For official statistics there is a defined state, when a building is "newly constructed" and when it should be counted for the specific official statistic (even though reports may be late). On the other side, advertisements for newly constructed houses of apartments may appear at any time –long before the building is built as well as up to the date when an object is ready to actually move in.

This specification error may lead to (validity) bias – especially if there is a systematic difference and the difference is large (in relation to the number of objects).

- **Undercoverage**

First of all, there is a mismatch of survey and target population. For example, objects that are constructed and used by their owners are not covered by this kind of data source. A bias would only arise, if these objects differed systematically in their characteristics from the observed i.e. advertised objects. Considering that official statistics on newly constructed residential buildings cover all residential buildings – and not just those offered on the (online) market – a certain degree of undercoverage is naturally to be expected by comparing results from this data source to official statistics. This leads to undercoverage, which could in principle be addressed by weighting, for example.

Second, there is additional undercoverage due to the fact that data is not collected using webscraping methods continuously but only at specified time intervals (e.g. weekly, monthly, every two days). By choosing a longer interval, some amount of "short-term advertised" objects are missing from the dataset. Again, this leads to bias if these "short term advertised objects" differ systematically from objects that are covered by a longer scraping interval (e.g. in price).

- **Duplicates**

Next, there are duplicates within the data. Duplicates can be identified by their ID (within a data source) and/or by comparing other characteristics, such as address, size, price, and other amenities. Since information can be missing as well as concepts can differ between sources (or change within

one source), it can be hard to identify and remove duplicates by comparing characteristics that are missing or use different definitions within or between the sources.

- **Missings**

Even with objects covered, there is missing data (item non-response), such as missing address information (street name or house number) or price. For this use-case, especially address data is relevant in order to identify and remove duplicates as well as to aggregate results in order to compare aggregates to official statistics.

Missings (in the sense of unit non-response) can also occur due to setting scraping intervals too long and missing "short term advertisements". This type of missing could be avoided by setting shorter scraping intervals.

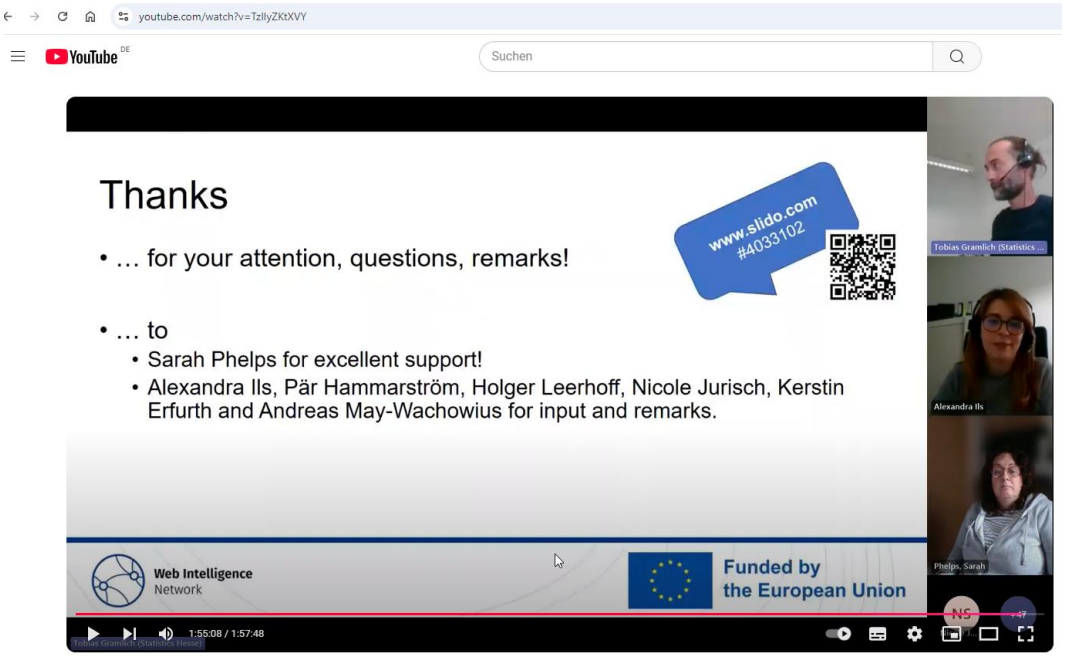
4. Additional Activities

Use-case 2 contributed three articles to the ESSnet WIN Blog series[3] and was responsible for one event in the ESSnet WIN webinar series[4], where about 60 participants followed a two-hour presentation on how to gather data from real estate web portals.

Screenshot from the ESSnet WIN blog

[3] The ESSnet WIN blog articles are available here https://cros.ec.europa.eu/book-page/web-intelligence-network-blog.
[4] The recordings of the ESSnet WIN webinar series also are available online (https://www.youtube.com/@ESSnetWIN).

Screenshot from the webinar about "Measuring Construction Activities using Data from the Web"



ESSnet WIN Webinar Measuring Construction Activities using Data from the Web New Date 20231016 1

5. Country Specific Part: HSL

5.1. Data Sources

Several portals have been investigated for this project, from which seven portals have been chosen in order to reflect the largest and most popular portals, but also to select rather small but specialized portals that have a regional focus or concentrate on larger construction projects.

Table 2 shows the portals used during this project, as well as the coverage, size, type of advertisement the web portals specialize in, the main advertisers and the method with which the data has been collected.

Table 2: Data sources used by HSL for UC2 and UC1

Portal	Coverage	Size	Type of Advertisements	Advertiser	Method	Remarks
Portal 1	Germany	large, popular	houses, apartments, complete projects; sale, rent	private owners, real estate agents	agreement	
Portal 2	Germany	large, popular	houses, apartments, complete projects; sale, rent	private owners, real estate agents	API	
Portal 3	Gemany	small	houses, apartments; sale, rent	private owners, no agents	screen scraping	Portal 3 and 4 share the same IDs for some advertisements
Portal 4	Germany	small	houses, apartments; sale, rent	mainly private owners	screen scraping	Portal 3 and 4 share the same IDs for some advertisements
Portal 5	Germany, larger cities only	medium	houses, apartments, complete projects; sale, rent	mainly developers, agents	screen scraping	
Portal 6	Rhine-Main region	small	houses, apartments, complete projects; sale, rent	developers	screen scraping	
Portal 7	Germany	large, popular	houses, apartments, complete projects; sale, rent	private owners, real estate agents	screen scraping	discontinued during project phase

For Portal 1, HSL entered an agreement for regular – i.e. monthly – data delivery for all advertisements describing newly constructed objects in Hesse appearing on the portal during the previous month. This data is also considered as “data from the web”, since this data could be collected from this real estate web portal using webscraping methods. This data source has proven to be more complete then scraped data as no data is missing due to set scraping intervals missing “short term advertisements” or failure of scrapers. Within the data, information tends to be more complete as well, for example there is complete address data for all advertisements of this source. Especially address information is hard to come by (e.g. because the advertiser didn’t make it available to the online ad or is made available only to registered users to the portal), but crucial for further steps such as deduplication.

Portals 2 and 7[5] are very large and popular portals covering all federal states. However, gathering data from Portal 7 has been discontinued during the project phase because of a large overlap of advertisements between both portals.

[5] For more information on portals 2, 5, and 7 see section 6.1. “Data Sources”.

Portals 3 and 4 also cover the whole of Germany but are comparatively small. They have been chosen in order to investigate coverage and overlap with larger real estate web portals. Additionally, Portal 3 is specialized in transactions, where no real estate agent is engaged. Interestingly, Portal 3 and 4 share the same IDs for some advertisements.[6]

Portal 5 is a small portal that mainly contains complete and larger construction projects advertised by developers. It covers all federal states but especially objects from larger cities.

Portal 6 is a small portal where mainly large construction projects as a whole are advertised. This portal has a focus on the Rhine-Main region. This real estate web portal was chosen as it represents advertisements from larger project developer companies, which might not be available on other real estate web portals. One of the questions regarding this portal was to investigate the overlap of advertisements from this portal with larger portals.

All portals cover houses and apartments for sale as well as for rent. Portals 1, 2, 5, 6 and 7 additionally include units of complete projects in their advertisements for sale as well as to rent.

Table 3 shows the number of advertisements for newly constructed objects for the years 2022 and 2023. The table reflects the different sizes of the portals, with portals 1 and 2 exceeding the amount of advertisements from portals 3, 4 and 5 by far. Between 2022 and 2023, advertisements on all real estate web portals increased. Note that only Portal 1 completely covers the two reference years.

Table 3: Number of advertisements: newly constructed objects in Hesse (2022, 2023)

	Number of Advertisements	
Portal	2022	2023
1	4808	5135
2	3467	6474
3	130	152
4	131	156
5	--	372
Hesse	8536	12289

5.2. Data Preparation

Since portals are scraped weekly, there should be some true duplicates by design for the combined monthly dataset. A weekly scraping frequency has been chosen in order not to miss objects that are advertised only for a short time – but obviously misses advertisements being online only for a few days or even for some hours only. These true duplicates can be identified easily through a specific ID that is used by the portal and is often part of an URL or otherwise identifiable in the source code of the advertisement. It has been decided that only the last (i.e. newest) advertisement should be kept. Here, it is assumed that the last advertisement is also the most accurate, e.g. after an update in the advertisement.

In order to obtain meaningful results and to be able to compare the scraped data with official statistics, the month and year of completion of the new building must be assigned to the scraped advertisements. Therefore, it needs to be decided at which month an advertisement has to be counted as part of the newly available monthly supply of apartments and houses: typically, the month an advertisement first appears at is not the month when the object actually becomes available. Advertisements of newly constructed houses and apartments are usually advertised weeks or months in advance. However, the information about the time at which an object is ready to be

[6] Both portals are owned by the same company. Despite objects sharing IDs, overlap between the portals is small.

used is often neither standardized (free text) nor filled with an absolute date ("from now on"/"ab sofort", "by arrangement"/"nach Vereinbarung", "3rd quarter of 2022"/"3. Quartal 2022"). A crucial part of cleaning the data is to attach a specific date to the advertisements.

Since the real state web portals' pages are highly structured for different advertisements within one portal, after scraping and extracting information from specific HTML tags using specific XPATHs from each page there only is minor data cleaning and editing necessary. Most of the data cleaning can be done applying a set of rules in form of simple regular expressions. More specifically, the following variables can be cleaned easily:

- Standardising price to numeric value, e.g. deleting € symbols or other character strings like "EUR" or "EURO; deleting character strings, e.g. "upon request"/"auf Anfrage".
- Standardising surface to numeric value, e.g. deleting "m²".
- Standardising floor numbers to numeric value, e.g. substituting character strings with the according numeric values, e.g. "Ground Level"/"Erdgeschoss" becomes 0, "1. Floor"/"1. Obergeschoss" becomes 1.
- Standardising number of rooms to numeric value, e.g. deleting character strings, "2 Rooms"/"2 Zimmer" becomes 2.
- Standardising "half rooms" to numeric value, e.g. rounding up half rooms to the next full number "2.5 rooms"/"2.5 Zimmer" becomes 3.[7]
- Standardising building types as "houses", "detached houses", "apartments", "other" according to the definition agreed upon in use-case 1. For some advertisements, this information is not available from the advertisement itself but is encoded in the URL of the advertisement.
- Standardising advertisement types to either rent or sale. For some advertisements, this information is not available from the advertisement itself but can be derived from the URL or, if this is not possible, from the price of an object. Typically, there are no objects to buy with a price lower than 10.000 Euro – on the other side, typically objects to rent have lower prices than 10.000 Euros. For some advertisements, the information on prices is not given and either remain empty or contain a note stating "upon request"/"Auf Anfrage".

While the above mentioned characteristics of advertisements can be cleaned with comparatively little effort and fixed rules, standardising addresses is a time consuming manual task, as building a set of rules involves reviewing and solving numerous individual cases. Cleaning and standardising address information includes standardising city names, standardising abbreviations and standardising house numbers. More specifically, this involves:

- Standardising house numbers: house number ranges (e.g. "12-14)" have been dissolved, using the first part of the range (e.g. "12-14" becomes "12"). Suffixes to house numbers have been removed (e.g. house number "12c" becomes "12"). Standardizing house numbers in such a way increases the number of potential duplicates within and between portals, as well as the number of potential false positive matches. Since deduplication

[7] In Germany, it is still common in advertisements to state "half rooms" (e.g. "1.5 rooms","2.5. rooms"). This tradition stems from an outdated norm that stated rooms of a size between 6 and 10 m² should only be counted as "a half room". Some are still used to indicate that one or more rooms are quite small. Even the search page of real estate web portals often provide the possibility to search for objects with half rooms. Statistically as well as legally speaking, there are no "half" rooms. Any room has to count as one room or isn't considered a room at all. For the official statistic (in Germany "Statistik der Baufertigstellungen"), all rooms greater than 6 m² have to be counted as a room.

has to take place anyway, and assuming that such inconsistencies exist, deduplication and matching procedures have to take this into account.

- Standardising city names: all variants of a city's name as well as possible spelling errors must be standardized to one city name to correctly aggregate the data at different NUTS levels to compare webscraped data to official data. For some real estate web portals quarters, districts, neighbourhoods and suburbs are used in combination or even instead of the city name. In some cases, city names given in the portals are not unique (even on NUTS1 level), e.g. only the NUTS3 name is given. In some of these cases, a city name can be assigned manually using the given postal code from the advertisement. To illustrate this problem, table 4 presents spelling variations, the use of quarters/districts/neighbourhoods/suburbs, as well as spelling errors for just three out of 426 Hessian cities.

Table 4: Spelling variations of city names

Frankfurt am Main	**Wächtersbach**	**Hofheim am Taunus**
Frankfurt	Wächtersbach	Hofheim
Frankfurt (Bockenheim)	Wächtersbach-Hesseldorf	Hofheim (Hofheim am Taunus)
Frankfurt (Gallus)	Wächtersbach, Main-Kinzig-Kreis	Hofheim am Taunus
Frankfurt (Griesheim)	Wächtersbach/ Afenau	Hofheim am Taunus / Wallau
Frankfurt a.M.		Hofheim Langenhain
Frankfurt a/M		
Frankfurt am Main		
Frankfurt am Main (Bockenheim)		
Frankfurt am Main (Griesheim)		
Frankfurt am Main (Innenstadt)		
Frankfurt am Main (Nordend)		
Frankfurt am Main / Bahnhofsviertel		
Frankfurt am Main / Gallusviertel		
Frankfurt am Main / Griesheim		
Frankfurt am Main / Harheim		
Frankfurt am Main / Kalbach		
Frankfurt am Main / Nordend-Ost		
Frankfurt am Main / Oberrad		
Frankfurt am Main / Preungesheim		
Frankfurt am Main / Unterliederbach		
Frankfurt am Main Eckenheim		
Frankfurt am Main, Gallus		
Frankfurt am Main.		
Frankfurt am Mian		
Frankfurt Bergen-Enkheim		
Frankfurt Eckenheim		
Frankfurt Main		
Frankfurt-Zeilsheim		
Frankfurt/ M.		
Frankfurt/M		
Frankfurt/Main		
FrankfurtamMain		
Frankfzrt am Main		
Frankfzrt am Main (Griesheim)		

- Standardising addresses by using geocoordinates: often, when a complete address of an object is not available through HTML tags, but only a street name or postcode, accurate geocoordinates are still available from map applications imbedded in the website showing the location of an object. These geocoordinates can then be used to retrieve standardized

address information using tools which are regularly used for geocoding ("reverse geocoding"). In general, accurate geocoordinates could be used to retrieve standardized addresses – even when address information is available from the portal. This reverse geocoding can help to standardise addresses or complete address information from real estate web portals. However, when dealing with newly constructed buildings or projects where entire streets are being build, the street names may not yet be part of the official databases typically used for geocoding.

5.3. Results

5.3.1. Number of all Advertisements

The two plots in Figure 2 show the raw number of advertisements for 2022 and 2023 by month for all scraped real estate web portals. In total, 8536 unique advertisements have been collected in 2022, in 2023 the number of collected advertisements was 12289. The figure clearly demonstrates how the number of advertisements varies throughout the year: in summer and autumn there seem to be more advertisements for newly constructed objects than during winter months, whereas the frequency of scraping stayed constant.

Note that not all portals have been observed over the complete period 2022-2023. The pattern of lower total numbers of advertisements at the beginning of 2022 is due to the fact that one of the larger portals (Portal 2) was only included from May onwards in the data sources.

Figure 2: Number of collected advertisements in 2022 and 2023

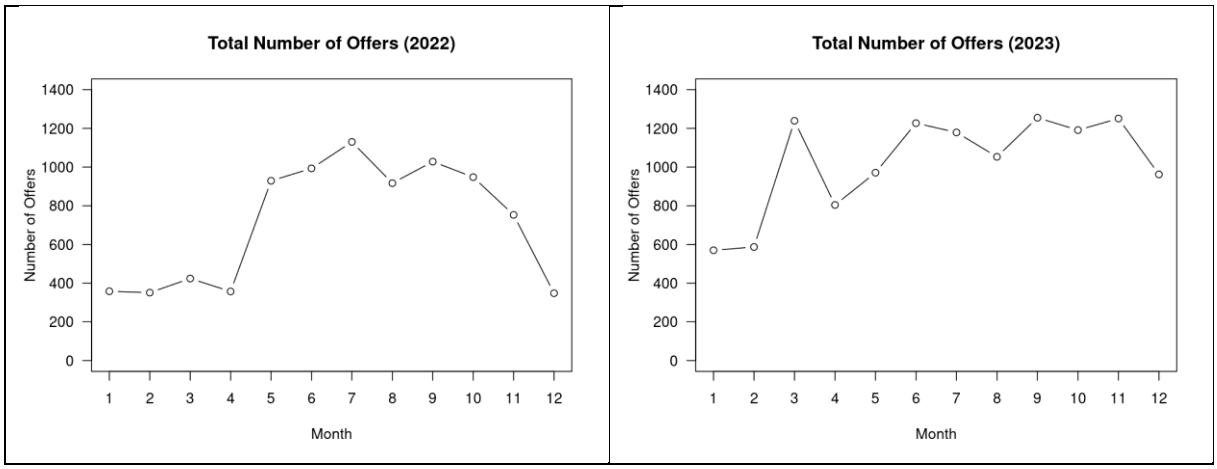
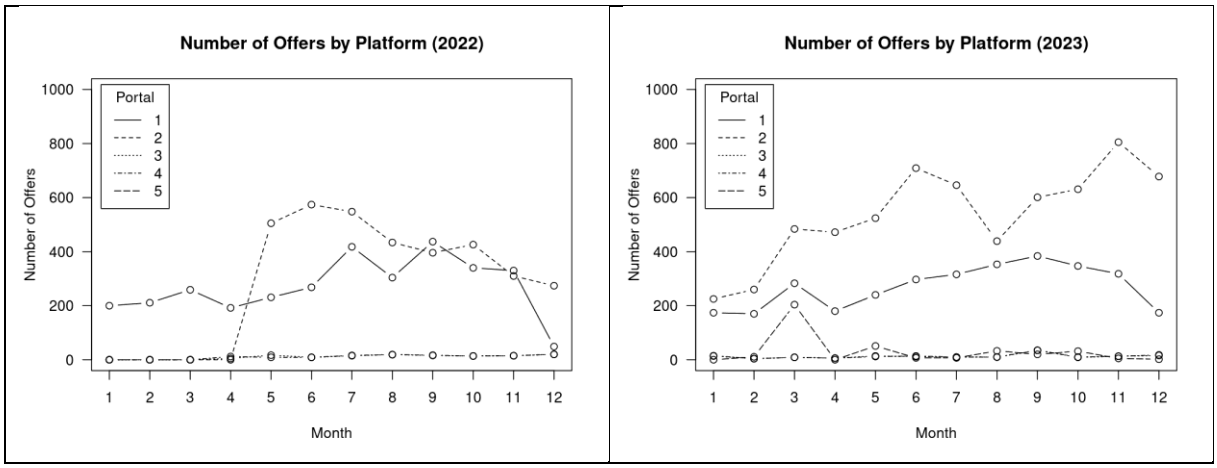


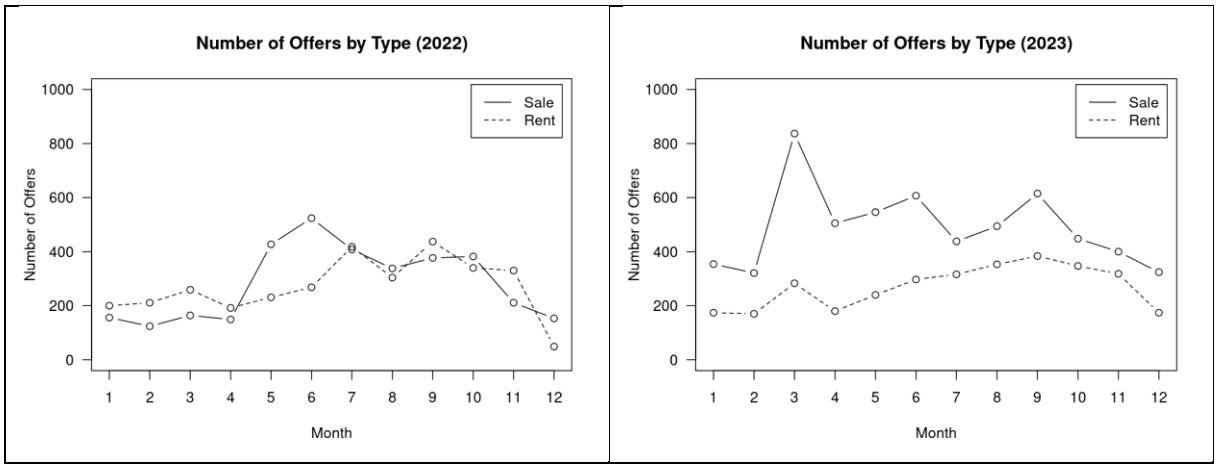
Figure 3 shows the number of collected advertisements for 2022 and 2023 by portal. Here, again, it is clearly visible that Portal 2, which was only collected from May 2022 onwards, offers the highest number of advertisements, with Portal 1 following. Portals 3, 4 and 5 are – as expected – lower in number, as they are smaller portals with specific foci.

Figure 3: number of collected advertisements in 2022 and 2023 by portal



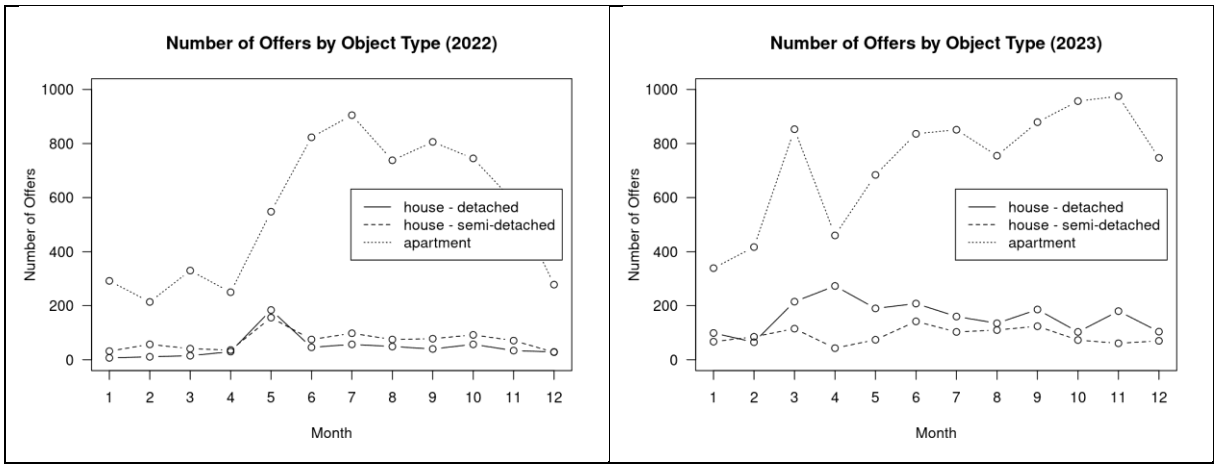
Inspecting the data further (figure 4), the number of advertisements offered for sale overall rose from 2022 to 2023. While advertisements to buy and to rent are roughly at the same level in 2022, advertisements to buy increase in 2023. At the same time, rental advertisements remain at a similar level.

Figure 4: Number of collected advertisements in 2022 and 2023 by advertisement type



The scraped data can also be divided by building type (figure 5). Patterns here are quite similar for both 20222 and 2023. Most often, apartments are offered in the advertisements, while detached and semi-detached houses are advertised less frequent.

Figure 5: Number of collected advertisements in 2022 and 2023 by building type



In summary, these initial analyses show that the data set mainly consists of advertisements from Portal 1 and 2, with the overall frequency of advertisements increasing in 2023. Most often, apartments are offered for rent.

5.3.2. Summary of all Collected Advertisements

Table 5 shows a summary of all collected advertisements for 2022 and 2023 by type of advertisement and building. Categories for type of advertisement (object to rent or for sale) as well as building (houses: categories 01 and 02. Apartments: 10. Others: 03) follow the definitions of use-case 1.

Table 5: Summary of advertisements in 2022 and 2023, by type of advertisement and type of building

		# of ads		median size (m²)		median room number		median price (€)	
advertisement type	**building type**	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**	**2023**	**2022**
Sale	01	1760	436	143	169	5	5	412070	629000
Sale	02	809	636	140	145	5	5	629000	739000
Sale	03	202	197	85	107	3	3	428000	657000
Sale	10	3118	2144	91	92	3	3	520855	499000
Rent	01	79	42	137	148	5	5	1577	1820
Rent	02	188	141	153	142	5	5	2100	1950
Rent	03	291	378	75	61	3	2	1200	1074,5
Rent	10	2678	2678	81	81	3	3	1225	1140
unknown	01	79	81	154	138	5	5		
unknown	02	70	62	144	140	5	4		
unknown	03	58	47	61	44	2	2		
unknown	10	2957	1694	61	74	2	3		

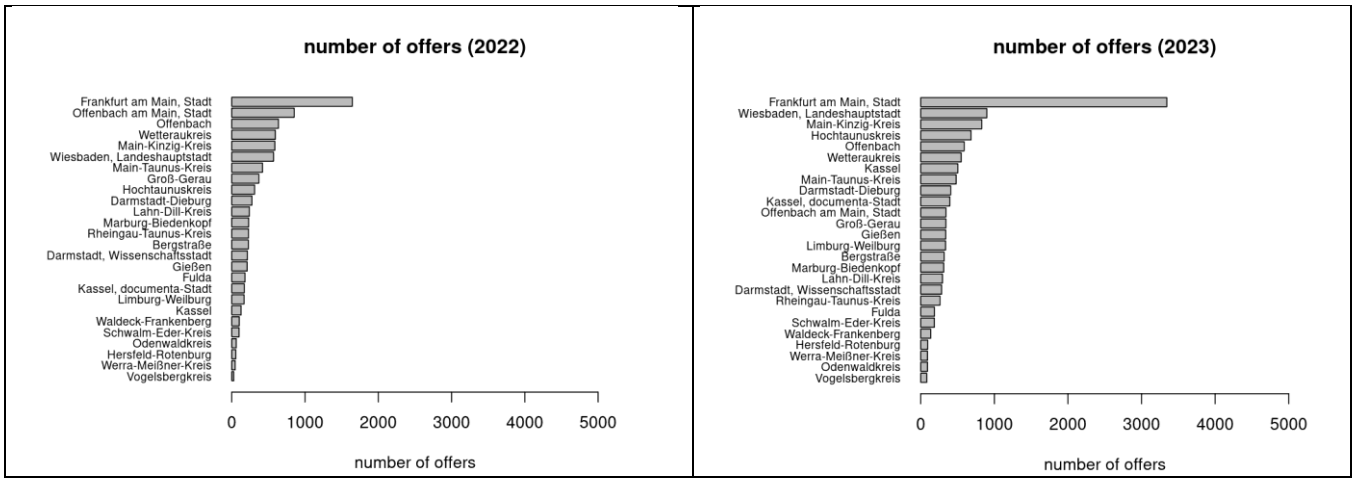
For some advertisement with unknown type of offer from the advertisement itself, the type can be derived from the URL of the advertisement or from the price (here, the assumption is that no objects for sale with a price less than 10.000 Euro; see more on this in section 5.2 “Data Preparation”). But still, for some advertisements from certain portals, the URL has not been stored and the price is missing, making it impossible to derive this information.

5.3.3. Aggregated Results (NUTS)

Figure 6 shows the number of advertisements for newly constructed objects (houses and apartments) in 2022 and 2023 by NUTS3 region (“rural districts and urban districts”/ “Landkreise und kreisfeie Städte”). For both years, most advertisements describe objects in the Rhine-Main region of Hesse, with

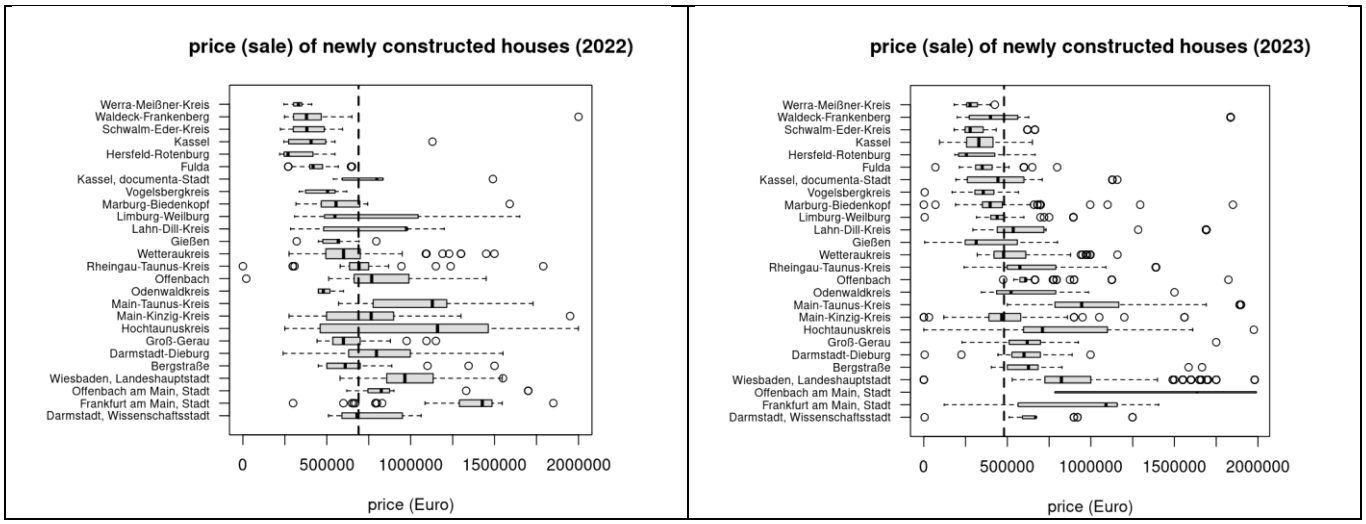
Frankfurt am Main on top of the list followed by neighbouring regions (Offenbach, city and county; Wiesbaden; counties Main-Kinzig, Main-Taunus, Wetterau, Darmstadt-Dieburg and Groß-Gerau). Only the northern city and county of Kassel (city and county) appear in the top ten counties in 2023. At the end of the sorted list with only very few are mainly the rural northern and eastern counties. This picture seems plausible to reflect the fact that there are more construction activities in urban regions in general, and additionally the share of objects built by their owners – for their own use and therefore not advertised in real estate web portals – may be higher in rural areas.

Figure 6: Number of advertisements for newly constructed objects in 2022 and 2023 by NUTS3 region



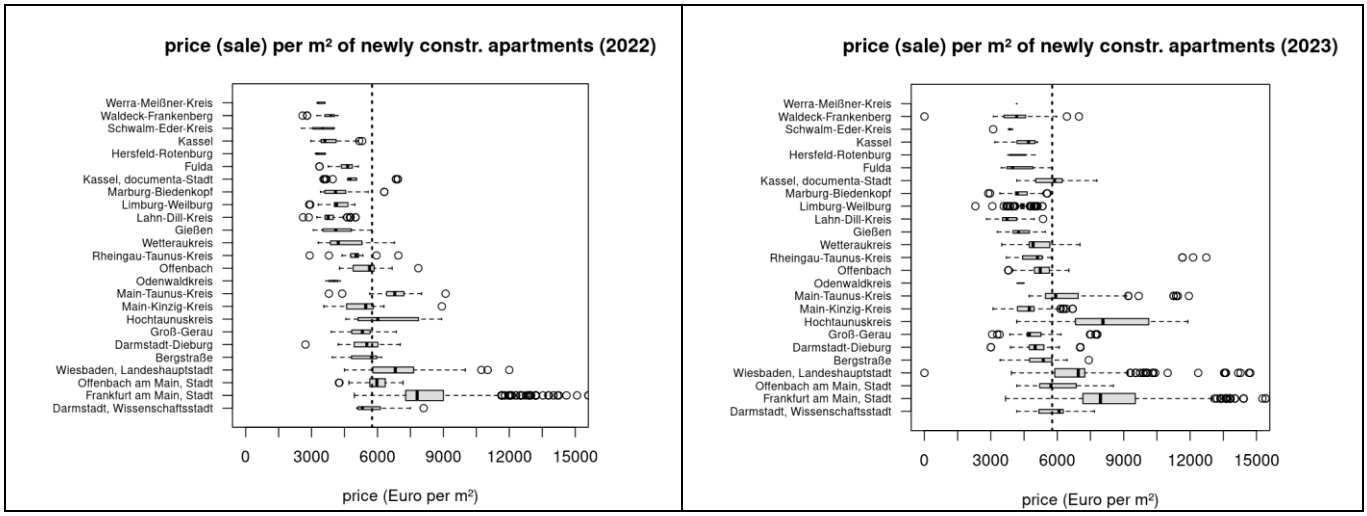
A rural-urban divide also shows when looking at prices, e.g. for houses to sale (figure 7; note that the axis showing the NUTS3 region is ordered by regional key: north-eastern counties are shown at the top. Also, each figure box width is proportional to the number of advertisements). Typically, advertisements for newly constructed houses in urban regions of Hesse show higher prices. In particular, prices for houses are higher in the cities of Frankfurt am Main and Wiesbaden, as well as in the counties Main-Taunus, Hochtaunus. Since the number of advertisements for houses is quite small in total in both years – and information on the price may be missing for some advertisements – it does not seem to be very reliable to interpret differences between NUTS3 regions in detail.

Figure 7: Boxplots of prices of newly constructed houses in 2022 and 2023 by NUTS3 regions



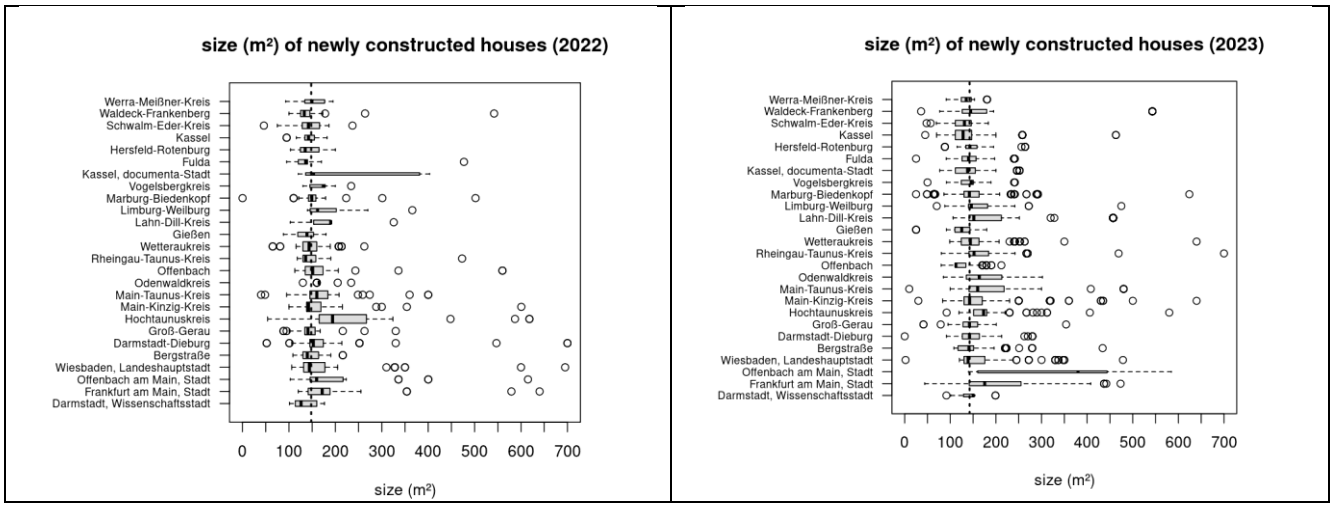
Similar patterns as in figure 7 are repeated in figure 8. At first, when there is a difference in the overall median price for houses for sale between 2022 and 2023, looking at prices per square metre, there is no such a difference between the two years (around 5900€/m²). But there are large and plausible differences between NUTS3 regions with prices per m² in urban regions being higher than in counties that are more rural. Again, the number of advertisements for houses is quite small, even too small to investigate and discus differences between NUTS3 regions in detail.

Figure 8: Boxplot for the price per m² of newly constructed houses for sale in 2022 and 2023 by NUTS3 regions



Interestingly, there are less differences when looking at the sizes of the advertised houses (to rent or to sale). Figure 9 shows variation between NUTS3 regions but especially within regions. Overall, median size is quite stable at around 150m² for both years.

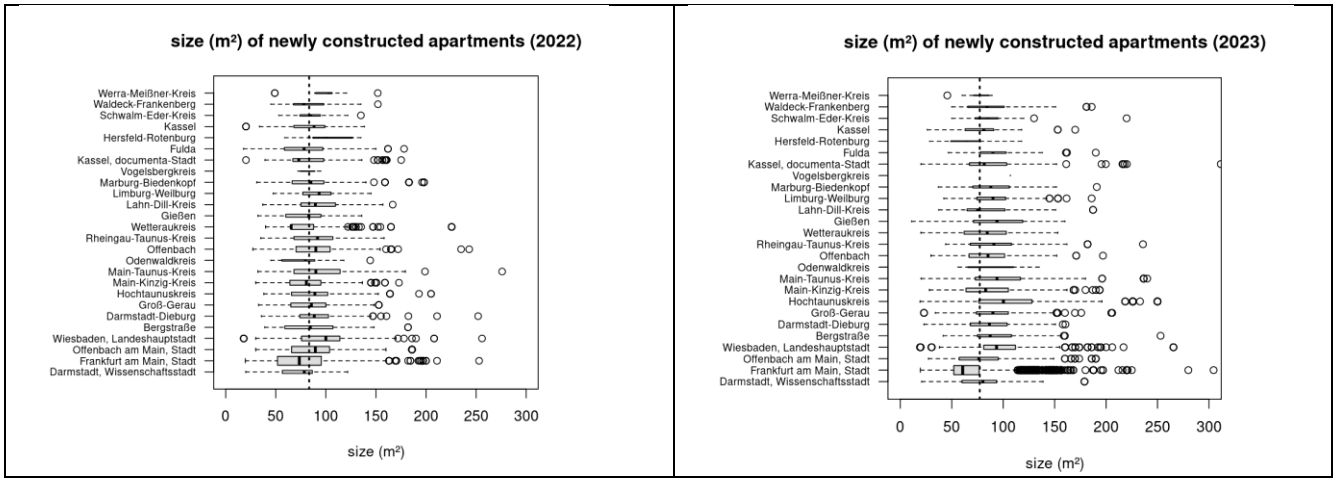
Figure 9: Boxplot for the size of newly constructed houses for sale in 2022 and 2023 by NUTS3 regions



For all the analyses for houses, it should be noted that the figures may present misleading artefacts due to the small number of observations within NUTS3 regions. For apartments, the situation is different, since most of the collected advertisements refer to apartments.
Figure 10 show the sizes of newly constructed apartments by NUTS3 region for the years 2022 and 2023. Again, there is considerable variation within NUTS3 regions and also between regions – but with

less emphasis. The overall median size is similar for both years (about 82m²). Most striking is the lower size of apartments for Frankfurt am Main.

Figure 10: Price per m² for newly constructed apartments in 2022 and 2023 by NUTS3 region

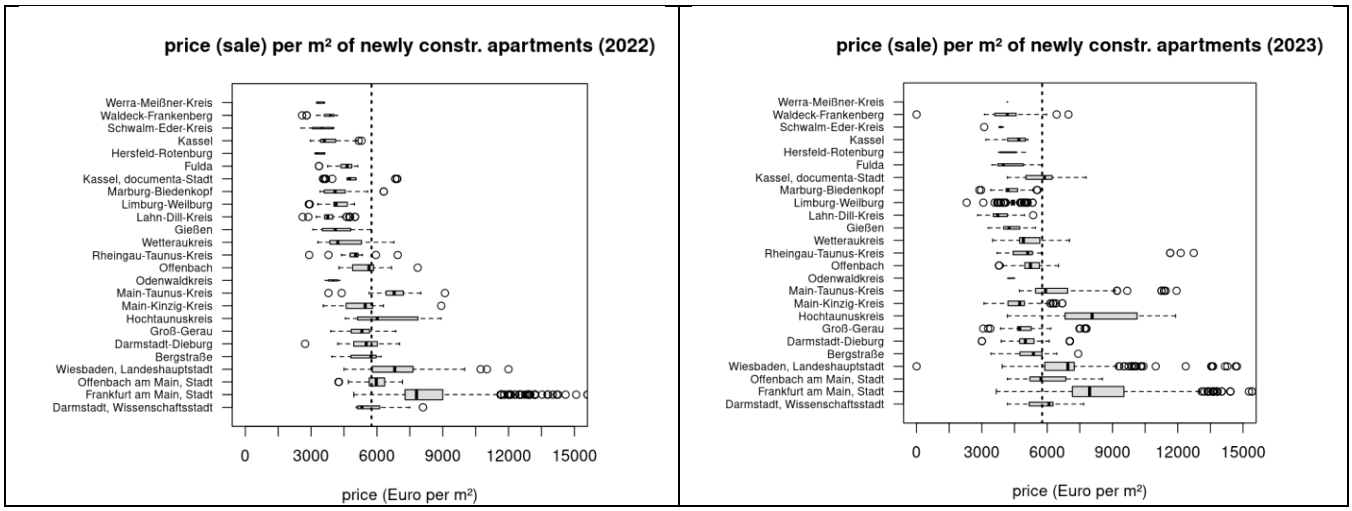


Very similar pictures for the years 2022 and 2023 are shown in figure 11: prices per square metre for apartments to rent. For both years, the median price is about 15€/m². As expected, there are difference between NUTS3 regions. Prices are higher for apartments in urban areas, especially in Frankfurt am Main and the Rhine-Main region in general. On the other side, rental prices are lower in the rural areas of Hesse in the north and east as well as south-west.

Figure 11: Rental price per m² for newly constructed apartments by 2022 and 2023 by NUTS3 region

A similar pattern to the prices per square metre for apartments to rent is shown in figure 12: sale prices per square metre by NUTS3 region. Overall, the median price for 2022 and 2023 is nearly the same (about 5.900€/m²) and the typical rural-urban divide is evident with the Rhine-Main region in general as well as the cities Frankfurt am Main and Wiesbaden in particular exhibiting the highest prices per square metre.

Figure 12: Sale price per m² for newly constructed apartments 2022 and 2023 by NUTS3 region



5.3.4. Aggregated Results by NUTS3 Level

Table 6 shows aggregated results by NUTS3 level for the year 2022. Columns G and H show a first measure of coverage: the share of apartments or buildings from official statistics for 2022 covered by the number of advertisements from real estate web portals or unique addresses.

It becomes obvious that on this aggregated level, coverage in general is low: assuming that the number of newly constructed buildings (from official statistics, column A) should correspond with the number of unique addresses of newly constructed objects from the advertisements (column E), results (column H) show that coverage reaches only 38 % for Hesse in total. Similarly, when the number of apartments in newly constructed buildings (from official statistics, column B) should correspond in some way to the number of advertised objects (column C), coverage of apartments is about 46 % (column G). However, at this point no deduplication between portals has been conducted yet. Therefore, the number of unique advertisements can be assumed to be lower than stated in column C. This would lower the overall coverage (column H) as well.

However, the presented first result for coverage heavily varies between different NUTS3 regions, ranging from 9% to 110% for apartments covered by the number of ads (column G), respectively from 12% to more than 200% for buildings covered by unique addresses (column H). Coverage seems to be much higher in urban NUTS3 regions than in rural NUTS3 regions.

Table 6: Aggregated results for 2022 by NUTS3 areas

Kreis / NUTS3	AGS	A) Number of Buildings	B) Number of Apartments	C) Number of Ads	D) Number of Ads by 1000 Inh.	E) Number of Addresses	F) Number of Addresses by 1000 Inh.	G) Ads / Apartments *100	H) Addresses / Buildings *100
Darmstadt, Wissenschaftsstadt	06411	39	195	213	1,3	85	0,5	109,2	217,9
Frankfurt am Main, Stadt	06412	261	2498	1645	2,1	266	0,3	65,9	101,9
Offenbach am Main, Stadt	06413	81	812	852	6,4	134	1	104,9	165,4
Wiesbaden, Landeshauptstadt	06414	377	1839	570	2	195	0,7	31	51,7
Bergstraße	06431	380	814	222	0,8	108	0,4	27,3	28,4
Darmstadt-Dieburg	06432	351	670	272	0,9	103	0,3	40,6	29,3
Groß-Gerau	06433	295	1049	368	1,3	124	0,4	35,1	42
Hochtaunuskreis	06434	267	604	312	1,3	113	0,5	51,7	42,3
Main-Kinzig-Kreis	06435	604	1565	581	1,3	178	0,4	37,1	29,5
Main-Taunus-Kreis	06436	214	678	420	1,7	133	0,5	61,9	62,1
Odenwaldkreis	06437	112	227	57	0,6	28	0,3	25,1	25
Offenbach	06438	151	846	637	1,8	193	0,5	75,3	127,8
Rheingau-Taunus-Kreis	06439	185	538	226	1,2	67	0,4	42	36,2
Wetteraukreis	06440	464	1242	588	1,9	131	0,4	47,3	28,2
Gießen	06531	284	758	208	0,7	87	0,3	27,4	30,6
Lahn-Dill-Kreis	06532	261	526	234	0,9	75	0,3	44,5	28,7
Limburg-Weilburg	06533	319	568	158	0,9	52	0,3	27,8	16,3
Marburg-Biedenkopf	06534	287	477	224	0,9	93	0,4	47	32,4
Vogelsbergkreis	06535	144	158	14	0,1	22	0,2	8,9	15,3
Kassel, documenta-Stadt	06611	56	286	171	0,8	28	0,1	59,8	50
Fulda	06631	331	522	170	0,7	60	0,3	32,6	18,1
Hersfeld-Rotenburg	06632	90	117	50	0,4	33	0,3	42,7	36,7
Kassel	06633	312	570	121	0,5	48	0,2	21,2	15,4
Schwalm-Eder-Kreis	06634	250	396	91	0,5	53	0,3	23	21,2
Waldeck-Frankenberg	06635	315	455	94	0,6	38	0,2	20,7	12,1
Werra-Meißner-Kreis	06636	109	121	38	0,4	23	0,2	31,4	21,1
HESSEN	06	6539	18531	8536	32	2470	9,7	46,1	37,8

5.3.5. Early Indicator for 'Construction Activities'

Measuring and predicting construction activities using data from the web is the heart of this Use Case. Table 7 compares numbers from official statistics, more specifically the number of newly constructed houses and number of newly constructed apartments, to the aggregated number of advertisements and addresses of the advertisements from 2022. More specifically, only data from Portal 1 was used for this analysis, as it is assumed to be most complete and without many internal duplicates. Additionally, Portal 1 has no missing advertisements because of a scraping interval: this source covers all advertisements of newly constructed objects within the year of reference that appeared during the month of reference.

Based in the comparison between official data and scraped data, a first and naive weighting factor was calculated. This weighting factor – based on data from 2022 – was then applied to scraped data from 2023.Finally, differences between this first "prediction" and the observed data from official statistics for 2023 are calculated and discussed.

Column (2) and (3) of table 7 show official statistics on newly constructed buildings for 2022. Columns (4) and to (5) show the number of all collected advertisements or the number of advertisements from one portal (Portal 1). Columns (6) and (7) show the number of unique addresses of these advertisements. Unique addresses have been used as a method of deduplication and micro-aggregation of apartments to buildings.

Columns (8a) to (9b) show "coverage rates" for apartments or buildings by dividing the number of advertisements from Portal 1 by the number of buildings or apartments from the official statistics. Since not all advertisements refer to apartments or buildings alone, this only gives a rough indication of a degree of "coverage". For columns (10a) and (10b) micro-aggregated advertisements from Portal 1 have been divided by the number of buildings from official statistics.

Comparing the number of advertisements with the official number of new buildings, there is an overall "overcoverage" of 131% with a huge range between different NUTS3 regions: there are 10 times more advertisements for objects in the city of Offenbach am Main (within the urban Rhine-Main region) than there have been new buildings according to official statistics. While this may be due to buildings consisting of several apartments, the number still seems quite high. On the other side, for the rural county "Vogelsbergkreis" the coverage is less than 10 percent: according to official statistics, there are ten times more new buildings than indicated by advertisements from real estate web portals in 2022. While the indicated coverage is poor (either too high or too low), the pattern behind it is plausible: in urban regions coverage is better, however too high, than in rural regions, where it can be assumed that newly constructed objects are primarily built and used by their owners and therefore do not appear on real estate web portals.

Coverage in this context does not mean that identical objects are compared but that there is a stable relationship between these objects and these numbers. However, a comparison of advertisements with buildings from the official statistics is inflated by the fact that advertised buildings on real estate web portals can contain more than one apartment, which can also mean more than one advertisement on a real estate web portal. Looking at a coverage rate not based on the number of buildings but on the number of apartments (columns 9a and 9b), rates drop accordingly. Overall, there is a low coverage rate of 46%. In general, one would expect a coverage rate below 100% since, one, it can be assumed that some newly constructed apartments are never advertised. Two, another set of apartments may be advertised on other portals not covered by the use-case. Third, the scraping interval could be too long and thereby not capture the advertisement while it is online.

Having that in mind, an overall coverage rate of 46% does not seem too bad: about every second newly constructed apartment is covered by data from real estate web portals.[8]

However, there is a large variation between NUTS3 regions. Again, counties in or near the Rhine-Main region show a higher coverage rate (i.e. the number and share of advertised apartments is higher – which one would expect). Cities of Darmstadt and Offenbach even slightly exceeding a rate of 100%. This rate also can be considered too high having in mind that not all advertisements refer to apartments but also houses. However, the share of advertisements for apartments at all collected advertisements is much higher than that for houses. Additionally, duplicates between portals have been ignored / neglected in the comparison (column 9a). Column 9b takes into consideration only advertisements from the largest portal, Portal 1. Accordingly, overall coverage decreases (about 26% of all apartments are covered by advertisements from Portal 1). There are cities and counties (mainly from the urban Rhine-Main area) that are covered at a higher rate, like the cities of Darmstadt (71%) and Offenbach (58%), counties Rheingau-Taunus and Main-Taunus. The city of Frankfurt am Main with the most advertisements reaches a coverage rate of 38%.

But still, this comparison is flawed: not all advertisements refer to either houses (columns 8a and 8b) or apartments (columns 9a and 9b). For a more realistic comparison, advertisements need to be deduplicated by address (columns 10a and 10b): regardless of type of advertisement, only unique addresses have been taken into consideration for a comparison with the number of newly constructed buildings from official statistics, which are assumed to only have unique addresses as well. Coverage rates for the cities of Darmstadt, Frankfurt am Main and the city and county of Offenbach exceed 100%: there are more unique addresses for advertisements in these regions than there are new buildings according to official statistics. The overall picture remains the same: on the one side there is a high(er) coverage for urban rural cities and counties especially in the Rhine-Main area, but on the other side there is a low coverage for rural counties outside the Rhine-Main region.

Since address data is complete for all advertisements from Portal 1, the number of deduplicated addresses from Portal 1 can be compared to the number of newly constructed buildings from official statistics in column 10b: once more, the overall pattern clearly indicates overcoverage in urban areas, while there is a clear undercoverage for rural areas. However, the low coverage of buildings in rural counties becomes even more apparent: the coverage drops from less than one in five buildings outside the Rhine-Main area to less than one in ten buildings or even fewer (Vogelsbergkreis, Kassel, Waldeck-Frankenberg, Schwalm-Eder). Again, although the coverage of data from real estate web portals is low, it is still possible to derive a relationship between advertisements on real estate web portals and official statistics on construction activity.

In a next step, a naïve correction factor was calculated in order to weight the number of advertisements from Portal 1 with the number of buildings from official statistics (column 11); the number of advertisements from Portal 1 with the number of apartments from official statistics (column 12); and, finally, the number of unique addresses from all advertisements from Portal 1 with the number of buildings (column 13).

These naïve correction or weighting factors reflect the inverse of the columns 8b, 9b and 10b, respectively the number of objects (apartments or houses) divided by the number of advertisements (unique addresses) from Portal 1. Accordingly, this factor reflects variation in the coverage rate between NUTS3 regions.

[8] Again, it needs to be stressed that there is no one-to-one linkage or comparison. The presented rate is only the relationship or ratio of two numbers.

Table 7: Comparison of official statistics with advertisements for newly constructed objects 2022 and derived weighting factors by NUTS3 regions

	Numbers from Official Statistics		Dataset from Portal 1		Number of unique addresses		Coverage rate in % for newly constructed buildings		Coverage rate in % for newly constructed apartments		Coverage rate in % for adresses of newly constructed buildings		Weighting factors for buildings, apartments and addresses		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8a)	(8b)	(9a)	(9b)	(10a)	(10a)	(11)	(12)	(13)
NUTS 3	**# of buildings**	**# of apartments**	**# of advertisements**	**# of advertisements (Portal 1)**	**# of addresses**	**# of adresses (Portal 1)**	**advertisements / buildings in %**	**advertisements (p1) / buildings in %**	**advertisements / apartments in %**	**advertisements (p1) / apartments in %**	**addresses / buildings in %**	**addresses (p1) / buildings in %**	**factor: p1 advertisements to buildings**	**factor: p1 advertisements to apartments**	**factor: p1 addresses to buildings**
Darmstadt, Wissenschaftsstadt	39	195	213	139	85	78	546,2	356,4	109,2	71,3	217,9	200,0	0,28	1,40	0,50
Frankfurt am Main, Stadt	261	2498	1645	937	266	207	630,3	359,0	65,9	37,5	101,9	79,3	0,28	2,67	1,26
Offenbach am Main, Stadt	81	812	852	473	134	86	1051,9	584,0	104,9	58,3	165,4	106,2	0,17	1,72	0,94
Wiesbaden, Landeshauptstadt	377	1839	570	418	195	166	151,2	110,9	31,0	22,7	51,7	44,0	0,90	4,40	2,27
Bergstraße	380	814	222	125	108	71	58,4	32,9	27,3	15,4	28,4	18,7	3,04	6,51	5,35
Darmstadt-Dieburg	351	670	272	172	103	70	77,5	49,0	40,6	25,7	29,3	19,9	2,04	3,90	5,01
Groß-Gerau	295	1049	368	268	124	95	124,7	90,8	35,1	25,5	42,0	32,2	1,10	3,91	3,11
Hochtaunuskreis	267	604	312	188	113	82	116,9	70,4	51,7	31,1	42,3	30,7	1,42	3,21	3,26
Main-Kinzig-Kreis	604	1565	581	310	178	111	96,2	51,3	37,1	19,8	29,5	18,4	1,95	5,05	5,44
Main-Taunus-Kreis	214	678	420	310	133	105	196,3	144,9	61,9	45,7	62,1	49,1	0,69	2,19	2,04
Odenwaldkreis	112	227	57	26	28	15	50,9	23,2	25,1	11,5	25,0	13,4	4,31	8,73	7,47
Offenbach	151	846	637	416	193	137	421,9	275,5	75,3	49,2	127,8	90,7	0,36	2,03	1,10
Rheingau-Taunus-Kreis	185	538	226	189	67	49	122,2	102,2	42,0	35,1	36,2	26,5	0,98	2,85	3,78
Wetteraukreis	464	1242	588	222	131	82	126,7	47,8	47,3	17,9	28,2	17,7	2,09	5,59	5,66
Gießen	284	758	208	123	87	50	73,2	43,3	27,4	16,2	30,6	17,6	2,31	6,16	5,68
Lahn-Dill-Kreis	261	526	234	131	75	47	89,7	50,2	44,5	24,9	28,7	18,0	1,99	4,02	5,55
Limburg-Weilburg	319	568	158	79	52	27	49,5	24,8	27,8	13,9	16,3	8,5	4,04	7,19	11,81
Marburg-Biedenkopf	287	477	224	98	93	49	78,0	34,1	47,0	20,5	32,4	17,1	2,93	4,87	5,86
Vogelsbergkreis	144	158	14	7	22	4	9,7	4,9	8,9	4,4	15,3	2,8	20,57	22,57	36,00
Kassel, documenta-Stadt	56	286	171	33	28	11	305,4	58,9	59,8	11,5	50,0	19,6	1,70	8,67	5,09
Fulda	331	522	170	91	60	34	51,4	27,5	32,6	17,4	18,1	10,3	3,64	5,74	9,74
Hersfeld-Rotenburg	90	117	50	15	33	11	55,6	16,7	42,7	12,8	36,7	12,2	6,00	7,80	8,18
Kassel	312	570	121	5	48	5	38,8	1,6	21,2	0,9	15,4	1,6	62,40	114,00	62,40
Schwalm-Eder-Kreis	250	396	91	17	53	11	36,4	6,8	23,0	4,3	21,2	4,4	14,71	23,29	22,73
Waldeck-Frankenberg	315	455	94	15	38	10	29,8	4,8	20,7	3,3	12,1	3,2	21,00	30,33	31,50
Werra-Meißner-Kreis	109	121	38	1	23	1	34,9	0,9	31,4	0,8	21,1	0,9	109,00	121,00	109,00
HESSEN	**6539**	**18531**	**8536**	**4808**	**2470**	**1614**	**130,5**	**73,5**	**46,1**	**25,9**	**37,8**	**24,7**	**1,36**	**3,85**	**4,05**

Table 8: Weighting factors from 2022 applied to 2023 web data and comparison to official data

	weighting factors from web data to observed data 2022			web data 2023		prediction for 2023			observed 2023		difference					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
NUTS 3	factor1: P1 ads to buildings	factor2: P1 ads to apartments	factor3: P1 adressses to buildings	# of ads (P 1)	# of adresses (Pl 1)	buildings1) based on factor 1	buildings2) based on factor 3	Apartments based on factor 2	buildings	apartments	buildings1)	buildings1), rel.	buildings2)	buildings2), rel.	apartments	apartments, rel.
Darmstadt, Wissenschaftsstadt	0,28	1,40	0,50	171	50	48	25	240	69	464	-21	-30,4	-44	-63,8	-224	-48,3
Frankfurt am Main, Stadt	0,28	2,67	1,26	1084	218	302	275	2890	231	2842	71	30,7	44	19,0	48	1,7
Offenbach am Main, Stadt	0,17	1,72	0,94	185	47	32	44	318	44	305	-12	-27,3	0	0,0	13	4,3
Wiesbaden, Landeshauptstadt	0,90	4,40	2,27	534	185	482	420	2349	250	1328	232	92,8	170	68,0	1021	76,9
Bergstraße	3,04	6,51	5,35	172	69	523	369	1120	320	611	203	63,4	49	15,3	509	83,3
Darmstadt-Dieburg	2,04	3,90	5,01	208	74	424	371	810	196	405	228	116,3	175	89,3	405	100,0
Groß-Gerau	1,10	3,91	3,11	214	76	236	236	838	221	976	15	6,8	15	6,8	-138	-14,1
Hochtaunuskreis	1,42	3,21	3,26	368	133	523	433	1182	206	589	317	153,9	227	110,2	593	100,7
Main-Kinzig-Kreis	1,95	5,05	5,44	395	129	770	702	1994	491	1337	279	56,8	211	43,0	657	49,1
Main-Taunus-Kreis	0,69	2,19	2,04	302	111	208	226	661	202	543	6	3,0	24	11,9	118	21,7
Odenwaldkreis	4,31	8,73	7,47	36	21	155	157	314	149	270	6	4,0	8	5,4	44	16,3
Offenbach	0,36	2,03	1,10	295	127	107	140	600	122	894	-15	-12,3	18	14,8	-294	-32,9
Rheingau-Taunus-Kreis	0,98	2,85	3,78	131	51	128	193	373	175	483	-47	-26,9	18	10,3	-110	-22,8
Wetteraukreis	2,09	5,59	5,66	217	87	454	492	1214	406	758	48	11,8	86	21,2	456	60,2
Gießen	2,31	6,16	5,68	130	59	300	335	801	241	646	59	24,5	94	39,0	155	24,0
Lahn-Dill-Kreis	1,99	4,02	5,55	123	44	245	244	494	328	775	-83	-25,3	-84	-25,6	-281	-36,3
Limburg-Weilburg	4,04	7,19	11,81	127	51	513	603	913	258	744	255	98,8	345	133,7	169	22,7
Marburg-Biedenkopf	2,93	4,87	5,86	132	52	387	305	642	359	731	28	7,8	-54	-15,0	-89	-12,2
Vogelsbergkreis	20,57	22,57	36,00	8	7	165	252	181	156	193	9	5,8	96	61,5	-12	-6,2
Kassel, documenta-Stadt	1,70	8,67	5,09	92	26	156	132	797	42	155	114	271,4	90	214,3	642	414,2
Fulda	3,64	5,74	9,74	77	44	280	428	442	366	777	-86	-23,5	62	16,9	-335	-43,1
Hersfeld-Rotenburg	6,00	7,80	8,18	46	16	276	131	359	118	175	158	133,9	13	11,0	184	105,1
Kassel	62,40	114,00	62,40	30	16	1872	998	3420	282	698	1590	563,8	716	253,9	2722	390,0
Schwalm-Eder-Kreis	14,71	23,29	22,73	18	14	265	318	419	329	590	-64	-19,5	-11	-3,3	-171	-29,0
Waldeck-Frankenberg	21,00	30,33	31,50	26	19	546	599	789	278	410	268	96,4	321	115,5	379	92,4
Werra-Meißner-Kreis	109,00	121,00	109,00	14	4	1526	436	1694	61	96	1465	2401,6	375	614,8	1598	1664,6
HESSEN	**1,36**	**3,85**	**4,05**	**5135**	**1730**	**6984**	**7009**	**19791**	**5900**	**17795**	**1084**	**18,4**	**1109**	**18,8**	**1996**	**11,2**

Table 8 shows the correction factors based on the comparison of 2022 official and web data (columns 2, 3 and 4). There is one factor to weight advertisements to buildings, another factor to weight advertisements to apartments, and a third factor to weight addresses to buildings. In columns 5 and 6 there are the number of advertisements with unique addresses from Portal 1 for 2023. Columns 7 to 9 show the number of buildings or apartments to expect for 2023, after weighting factors from 2022 have been applied to data from real estate web portals from 2023.

In total in 2023 Portal 1 contains 5.135 advertisements for newly constructed objects at 1.730 unique addresses. Applying the weighting factors for Hesse to advertisements gives 6.984 (based on advertisements) or 7.009 newly constructed buildings (based on unique addresses). Compared with the official statistics of 2023, which show 5.900 newly constructed houses respectively 17.795 apartments, the indicator overestimates the number of buildings by 19% and the number of new apartments by 11%. Given that the weighting scheme is very naïve, this difference is large but still surprising.

On one hand, the usefulness of such a simple approach in predicting construction activities would be evident not only at a highly aggregated level, but also at lower regional levels. On the other hand, number of advertisements and coverage of advertisements have shown to vary heavily between NUTS3 regions. This can be problematic since small numbers of advertisements and poor coverage lead to large weighting factors.

Looking at the results at NUTS3 level, there is no clear pattern visible because of large variation between regions, e.g. that the prediction for urban counties would produce smaller differences. To be more specific, for the biggest city in Hesse, Frankfurt am Main, Portal 1 contains 1084 advertisements for newly constructed objects in 2023. These advertisements have 218 unique addresses, most likely because there are some buildings with more than on apartment in it. Applying the weighting factors based on results from 2022 returns 2890 apartments in 302 or 275 buildings (depending on whether number of advertisements or number of addresses are weighted). Official statistics for 2023 have reported 231 new buildings with 2.842 apartments for Frankfurt am Main. Based on addresses, this would overestimate the official number of newly constructed buildings by 44 buildings (+19%). Looking at apartments, the difference is even smaller: an overestimation of 48 apartments corresponds to a relative difference of only 1.7%. A similar picture arises for the city of Offenbach am Main. Looking at the number of predicted buildings, there is no difference, while for apartments the difference is only 13 (+4.3%).

However, this pattern cannot be generalised. There are larger cities from the urban Rhine-Main region that show considerable overcoverage – e.g. Wiesbaden with an overcoverage of 170 buildings (+68%) and 1021 apartments (+77%) – but also examples for overcoverage in rural counties, e.g. the Odenwaldkreis with an overcoverage of 8 buildings (+5.4%) and 44 apartments (+16.3%).

Overall, the results show that the indicator should be interpreted and used very cautiously when predicting construction activity. The calculated factors are not assumed to be stable over time, since the weighting and correction is based on the previous year's construction activities and web data. Table 9 shows the correction or weighting factors based on 2022 respectively 2023 data from Portal 1 and the relative difference between factors from 2023 to 2022, exemplifying the volatility of the data even further.

Table 9: Comparison of weighting factors based on data from 2022 and 2023

	weighting factors from web data to observed data 2022			weighting factors from web data to observed data 2023					
(1)	**(2a)**	**(3a)**	**(4a)**	**(2b)**		**(3b)**		**(4b)**	
NUTS 3	**factor: p1 advertisements to buildings**	**factor: p1 advertisements to apartments**	**factor: p1 addresses to buildings**	**factor: p1 advertisements to buildings**	**Relative diff. to '22**	**factor: p1 advertisements to apartments**	**Relative diff. to '22**	**factor: p1 addresses to buildings**	**Relative diff. to '22**
Darmstadt, Wissenschaftsstadt	0,28	1,40	0,50	0,40	43,8	2,71	93,4	1,38	176,0
Frankfurt am Main, Stadt	0,28	2,67	1,26	0,21	-23,5	2,62	-1,7	1,06	-16,0
Offenbach am Main, Stadt	0,17	1,72	0,94	0,24	38,9	1,65	-4,0	0,94	-0,6
Wiesbaden, Landeshauptstadt	0,90	4,40	2,27	0,47	-48,1	2,49	-43,5	1,35	-40,5
Bergstraße	3,04	6,51	5,35	1,86	-38,8	3,55	-45,4	4,64	-13,3
Darmstadt-Dieburg	2,04	3,90	5,01	0,94	-53,8	1,95	-50,0	2,65	-47,2
Groß-Gerau	1,10	3,91	3,11	1,03	-6,2	4,56	16,5	2,91	-6,4
Hochtaunuskreis	1,42	3,21	3,26	0,56	-60,6	1,60	-50,2	1,55	-52,4
Main-Kinzig-Kreis	1,95	5,05	5,44	1,24	-36,2	3,38	-33,0	3,81	-30,1
Main-Taunus-Kreis	0,69	2,19	2,04	0,67	-3,1	1,80	-17,8	1,82	-10,7
Odenwaldkreis	4,31	8,73	7,47	4,14	-3,9	7,50	-14,1	7,10	-5,0
Offenbach	0,36	2,03	1,10	0,41	13,9	3,03	49,0	0,96	-12,8
Rheingau-Taunus-Kreis	0,98	2,85	3,78	1,34	36,5	3,69	29,5	3,43	-9,1
Wetteraukreis	2,09	5,59	5,66	1,87	-10,5	3,49	-37,6	4,67	-17,5
Gießen	2,31	6,16	5,68	1,85	-19,7	4,97	-19,4	4,08	-28,1
Lahn-Dill-Kreis	1,99	4,02	5,55	2,67	33,8	6,30	56,9	7,45	34,2
Limburg-Weilburg	4,04	7,19	11,81	2,03	-49,7	5,86	-18,5	5,06	-57,2
Marburg-Biedenkopf	2,93	4,87	5,86	2,72	-7,1	5,54	13,8	6,90	17,9
Vogelsbergkreis	20,57	22,57	36,00	19,50	-5,2	24,13	6,9	22,29	-38,1
Kassel, documenta-Stadt	1,70	8,67	5,09	0,46	-73,1	1,68	-80,6	1,62	-68,3
Fulda	3,64	5,74	9,74	4,75	30,7	10,09	75,9	8,32	-14,6
Hersfeld-Rotenburg	6,00	7,80	8,18	2,57	-57,2	3,80	-51,2	7,38	-9,9
Kassel	62,40	114,00	62,40	9,40	-84,9	23,27	-79,6	17,63	-71,8
Schwalm-Eder-Kreis	14,71	23,29	22,73	18,28	24,3	32,78	40,7	23,50	3,4
Waldeck-Frankenberg	21,00	30,33	31,50	10,69	-49,1	15,77	-48,0	14,63	-53,6
Werra-Meißner-Kreis	109,00	121,00	109,00	4,36	-96,0	6,86	-94,3	15,25	-86,0
HESSEN	**1,36**	**3,85**	**4,05**	**1,15**	-15,5	**3,47**	-10,1	**3,41**	-15,8

5.4. Quality Aspects

5.4.1. Missing Data

One major indicator of data quality is the number of ads with missing information. Table 10 gives an overview over the different data sources and some relevant characteristics from the ads. For some portals, not all mandatory variables are available or only available with a high share of missing data. For one of the portals, address information (at street or house number level) and prices are often missing, which poses a challenge as this information is especially useful for deduplication.

Dealing with this missing data is challenging, as there needs to be decided whether a complete case analysis is necessary or whether imputations are feasible. One setback of using a complete case analysis would be losing the majority of data. Imputation is a challenge since homogenous imputation cells of sufficient size are needed. Imputation cells could be formed (at least) e.g. by region, building type, floor number of rooms, surface / size. The number of imputation cells quickly becomes quite large and the number of units within each cell can get quite small or result in empty cells. This decision is not only relevant for numbers of remaining cases, but has an impact on identifying duplicates as well.

Table 10: Number of collected advertisements for 2022 with missing information (item nonresponse)

		Number of ads with missing information on …									
	Total number of ads	**… city name**	**… postal code**	**… street name**	**… house number**	**… floor number**	**… number of rooms**	**… size (m2)**	**… price (Euro)**	**… building type**	**… object type**
Portal1	4808	0	0	0	0	1408	64	10	38	602	38
Portal2	3467	0	0	2083	2083	1170	20	5	1825	2	1820
Portal3	130	0	0	0	0	26	4	3	6	9	6
Portal4	131	0	0	0	0	26	4	3	20	9	20

5.4.2. Duplicates

Deduplication of advertisements is a major challenge. Advertisements have to be deduplicated for two reasons: one, the weekly scraping interval leads to scraping one advertisement multiple times and therefore deduplication is needed within a real estate web portal. Two, deduplication between portals is needed when combining data from several real estate web portals in order to not include the same objects multiple times.

5.4.2.1. Deduplication within Portals

Deduplication within a portal uses the portal's own ID for an advertisement or object: only the last / newest advertisement is kept. This makes sure that corrections of an advertisement are included (say correction of a price or other characteristics) as long as the ID does not change.

However, the very same advertisements (and objects) may appear under different IDs within the same portal or within different portals. Therefore, in a next step, deduplication is based on address information: the assumption is that duplicates share the same address and main or even all other characteristics like size, price, and location within a building.

Unfortunately, missing data makes deduplication challenging and even with complete address and additional information, two separate advertisements may appear as potential duplicates.

Example 1:

Example 1

adresse2	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price_rent.x	price_sell.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price_rent.y	price_sell.y
Bad Emstal - 34308 - wolfhager - - 2 - 155		282LV56	155.0	2	4	NA	449000		28GXY56	155.0	2	4	NA	299900
Bad Emstal - 34308 - wolfhager - - 2 - 155		28GXY56	155.0	2	4	NA	299900		282LV56	155.0	2	4	NA	449000

In example (1), two advertisements from one portal with distinct IDs share the same address information (the linking key is formed by city name, postal code, (standardised / cleaned) street name, (standardised / cleaned) house number, floor number, number of rooms and size. In this example, the floor number is missing for both advertisements. Only the selling price differs by a larger amount (449.000€ vs. 299.000€). Given that the complete address matches as well as the number of rooms and size, these two advertisements may show a pair of duplicates.

Example 2:

In a similar example (2), floor numbers are not missing and prices do not differ by a larger amount (340.000€ vs. 332.000€). This makes it even more reasonable to treat these two advertisements as a pair of duplicates.

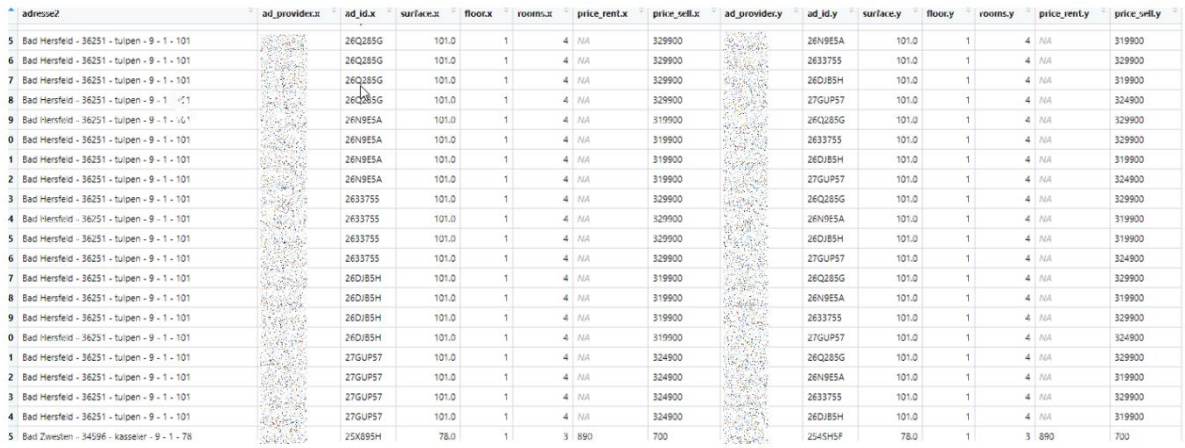
Example 2

	adresse2	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price_rent.x	price_sell.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price_rent.y	price_sell.y
3	Bad Endbach - 35080 - land - 2 - 2 - 156		2755L59	156.0	2	4	NA	340000		27Q4F59	156.0	2	4	NA	332000
4	Bad Endbach - 35080 - land - 2 - 2 - 156		27Q4F59	156.0	2	4	NA	332000		2755L59	156.0	2	4	NA	340000

Example 3:

Another illustrative example (3) presents one advertisement with four potential duplicates. All these advertisements are located at the very same (complete) address, have the same size (101m², 4 rooms) and location within the building (first floor). They only differ in the price (329.000€, 329.900€, 324.900€, 319.900€). When matching these five advertisements by address, location and size, the result gives the following 5x4 combinations:

Example 3



	adresse2	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price_rent.x	price_sell.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price_rent.y	price_sell.y
5	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26Q285G	101.0	1	4	NA	329900		26N9E5A	101.0	1	4	NA	319900
6	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26Q285G	101.0	1	4	NA	329900		2633755	101.0	1	4	NA	329900
7	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26Q285G	101.0	1	4	NA	329900		26DJB5H	101.0	1	4	NA	319900
8	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26Q285G	101.0	1	4	NA	329900		27GUP57	101.0	1	4	NA	324900
9	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26N9E5A	101.0	1	4	NA	319900		26Q285G	101.0	1	4	NA	329900
0	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26N9E5A	101.0	1	4	NA	319900		2633755	101.0	1	4	NA	329900
1	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26N9E5A	101.0	1	4	NA	319900		26DJB5H	101.0	1	4	NA	319900
2	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26N9E5A	101.0	1	4	NA	319900		27GUP57	101.0	1	4	NA	324900
3	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		2633755	101.0	1	4	NA	329900		26Q285G	101.0	1	4	NA	329900
4	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		2633755	101.0	1	4	NA	329900		26N9E5A	101.0	1	4	NA	319900
5	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		2633755	101.0	1	4	NA	329900		26DJB5H	101.0	1	4	NA	319900
6	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		2633755	101.0	1	4	NA	329900		27GUP57	101.0	1	4	NA	324900
7	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26DJB5H	101.0	1	4	NA	319900		26Q285G	101.0	1	4	NA	329900
8	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26DJB5H	101.0	1	4	NA	319900		26N9E5A	101.0	1	4	NA	319900
9	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26DJB5H	101.0	1	4	NA	319900		2633755	101.0	1	4	NA	329900
0	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		26DJB5H	101.0	1	4	NA	319900		27GUP57	101.0	1	4	NA	324900
1	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		27GUP57	101.0	1	4	NA	324900		26Q285G	101.0	1	4	NA	329900
2	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		27GUP57	101.0	1	4	NA	324900		26N9E5A	101.0	1	4	NA	319900
3	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		27GUP57	101.0	1	4	NA	324900		2633755	101.0	1	4	NA	329900
4	Bad Hersfeld - 36251 - tulpen - 9 - 1 - 101		27GUP57	101.0	1	4	NA	324900		26DJB5H	101.0	1	4	NA	319900
5	Bad Zwesten - 34596 - kasseler - 9 - 1 - 78		25X895H	78.0	1	3	890	700		254SH5F	78.0	1	3	890	700

One way to assist in the decision-making process is to use a map service, which – in the case of example (3) – actually shows a larger construction site (figure 13):

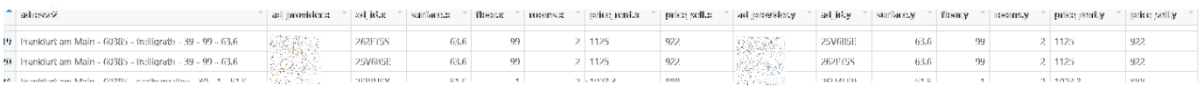
Figure 13: Map service pictures (Example 3)

Given the pictures from the building, the five advertisements could at least refer to two distinct objects (two separate four-room apartments at the first floor).

Example 4:

Another example (4) displays that even with no or with some missing data, advertisements with distinct IDs within a portal can be identified as a potential pair of duplicates. Two advertisements with different IDs are located at the same address, share the same location, room number, size and the same rental prices.

Example 4



Example 5:

The next example (5) illustrates the difficulty of deciding whether a difference or deviation between advertisements is significant for finding duplicates or not. Two advertisements only differ by room number (4 vs. 5) but share price and size – which could mean that one of the advertisements is erroneous. Two other advertisements share size, price and room number. Moreover, two other objects again share price, size and room number with the difference that size and price being slightly increased. For all of the advertisements, the location within the object (floor) is missing. At first, three distinct apartments could be plausible.

Example 5

	adresse2	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price_rent.x	price_sell.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price_rent.y	price_sell.y
35	Biblis - 68647 - heifrichsgaertel - 3 - NA - 112		2/CJN5S	112.0	NA	5	NA	529900		2/J8U5T	112.0	NA	4	NA	529900
36	Biblis - 68647 - heifrichsgaertel - 3 - NA - 112		2/J8U5T	112.0	NA	4	NA	529900		2/CJN5S	112.0	NA	5	NA	529900
37	Biblis - 68647 - heifrichsgaertel - 3 - NA - 132		2/YHN5S	132.0	NA	6	NA	599900		2/39U5T	132.0	NA	6	NA	599900
38	Biblis - 68647 - heifrichsgaertel - 3 - NA - 132		2/39U5T	132.0	NA	6	NA	599900		2/YHN5S	132.0	NA	6	NA	599900
39	Biblis - 68647 - heifrichsgaertel - 3 - NA - 139		2/NHN5S	139.0	NA	6	NA	609900		2/98U5T	139.0	NA	6	NA	609900
40	Biblis - 68647 - heifrichsgaertel - 3 - NA - 139		2/98U5T	139.0	NA	6	NA	609900		2/NHN5S	139.0	NA	6	NA	609900

However, using a map service again shows a detached house (figure 14). The object appears to have two floors. Therefore, 6 x 3 = 18 rooms seem not plausible, as well as the first impression of three 3-room apartments. It is rather plausible that these 6 ads refer to two apartments: e.g. a 3-room apartment (139 m²) at the first floor (four ads with two different prices), and a 3-room apartment at the second floor (112 m², two ads).

Figure 14: Map service pictures (Example 5)

These above shown examples illustrate the challenge to decide on duplicates given the information available from advertisements from real estate web portals, especially, if some information is missing (e.g. location in a building).

Table 11 shows the results of deduplication for Portal 1 for the year 2022, where there is no missing information on the address level (see table 10).

There are 4808 unique ads referring to newly constructed buildings in 2022. There are 529 combination of address, location, and size with more than one advertisement (in total, there are 2409 advertisements with one of these combinations). For half of these combinations, there is only one potential duplicate. When inspecting these potential duplicates, it turns out that most of them have to be considered true duplicates, since in most cases, prices are also identical (or at least similar enough to assume that a correction for the price information is the reasons for the second – new – advertisement for this object. Overall, there are 3298 advertisements with unique combinations of

address, location and size. The total number of potential duplicates is 2409, while 529 advertisements have at least one duplicate (in terms of address, location within building, size, and number of rooms).

Table 11: Deduplication by address and other characteristics within Portal 1

Portal	**Portal 1**
Total number of advertisements	4808
Total number of potential duplicates (number of advertisements with same address, location, price, size)	2409
Number of advertisements with at least one duplicate	529
Median number of duplicates	1
Mean number of duplicates	4.6
Number of advertisements without potential duplicate	3298
Deduplicated number of advertisements	3827
Rate of duplicates	20.4

Example 6:

Some of the potential duplicates can be classified as clear duplicates. To illustrate this, example 6 lists key combinations with more than 50 occurrences. Figure 15 shows pictures of the address from a map service. Clearly, the building is newly constructed but does not hold 91 three-room apartments at a first floor.

Example 6: Key combinations with 50 or more occurrences

	adresse2	dubletten
1	Frankfurt am Main - 60598 - moerfelderland - 128 - 3 - 3 - 142.99	136
2	Offenbach am Main - 63069 - buchrain - 139 - 1 - 3 - 107	91
3	Offenbach am Main - 63069 - buchrain - 139 - 4 - 3 - 145	91
4	Offenbach am Main - 63069 - buchrain - 139 - 1 - 3 - 108	78
5	Offenbach am Main - 63069 - buchrain - 139 - 1 - 3 - 106	66
6	Offenbach am Main - 63069 - buchrain - 139 - 4 - 4 - 186	66
7	Offenbach am Main - 63069 - buchrain - 139 - NA - 3 - 106	66
8	Frankfurt am Main - 60327 - europaallee - 2 - 39 - 3 - 89	55
9	Frankfurt am Main - 60596 - garten - 134 - 3 - 2 - 87.91	55
10	Frankfurt am Main - 60596 - garten - 134 - 3 - 3 - 142.99	55
11	Offenbach am Main - 63069 - buchrain - 139 - NA - 3 - 99	55

Figure 15: Map service pictures (Example 6)

After deduplication, 3827 unique combinations from 4808 unique ads remain. This equates a share of about 20% of duplicates for Portal 1.

5.4.2.2. Deduplication between Portals

Between portals, distinct advertisements may actually refer to the same object. In order to avoid double counting of advertisements, the data needs to be deduplicated not only within but also between portals. Again, the linking key for deduplication is formed by address, floor number, room number and size of the object. Since price information may differ in concept between portals – or is missing at a higher frequency in one of the portals –, it is not used for matching but may be useful to tip the decision whether two otherwise identical advertisements refer to the same object.

Example 7:

In the following example (7), there are two advertisements in two portals at the same address. The floor number is missing in both sources. The number of rooms (four rooms) is the same in both sources as well as the price (539.000€). Both advertisements, however, differ in size (150m² vs. 142m²).

Example 7

	adresse.x	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price_rent.x	price_sell.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price_rent.y	price_sell.y	quelle
35	Bad Soden-Salmünster - 63628 - andenmittelaeckern - 17 - NA - 142	NA	NA	NA	NA	NA	NA	NA	[illegible]	1787	142.0	NA	4	NA	539000	nur [illegible]
36	Bad Soden-Salmünster - 63628 - andenmittelaeckern - 17 - NA - 150	[illegible]	27AGV5D	150.0	NA	4	NA	539000	NA	NA	NA	NA	NA	NA	NA	nur [illegible]

Without additional information, it seems reasonable to treat these two advertisements as a pair of duplicates – since the price is the same even when size differs by a few square metres. Using a map service providing satellite images, it seems that these two advertisements refer to a newly built detached or semi-detached house. Still, while they may refer to the very same detached house or part of a semi-detached house, it is also possible that they refer to two different parts of the same building.

Images from three different providers (figure 16) show the challenge of deciding on duplicates, especially with newly constructed buildings. Sometimes, the source of these tools shows different states of construction or do not cover recent changes at all. For example, one of the map services does not indicate the street name, another map service does not indicate house numbers.

Figure 16: Map service images (Example 7)

Example 8:

For the two biggest portals used for Hesse, there are 4808 objects in newly constructed buildings for Portal 1, and 3467 for Portal 2. Combining both by a matching key results in 526 advertisements matching (including multiple matches A1 -> B1 and A1 -> B2). Example 8 shows 28 different advertisements at the same address – up to street name.

Example 8

	adresse2	ad_provider.x	ad_id.x	surface.x	floor.x	rooms.x	price.x	ad_provider.y	ad_id.y	surface.y	floor.y	rooms.y	price.y	quelle	strasse
251	Bad Vilbel - 61118 - friedberger - 77 - NA - 5 - 171.74	NA	NA	NA	NA	NA	NA	[illegible]	1926	171.74	NA	5	1450000	nur [illegible]	TRUE
252	Bad Vilbel - 61118 - johannesgutenberg - 07 - 1 - 2 - 64	NA	NA	NA	NA	NA	NA	[illegible]	5313	64.00	1	2	1090	nur [illegible]	TRUE
253	Bad Vilbel - 61118 - johannesgutenberg - 1 - 2 - 2 - 63	[illegible]	27Y6R56	63.00	2	2	NA	NA	NA	NA	NA	NA	NA	nur [illegible]	TRUE
254	Bad Vilbel - 61118 - johannesgutenberg - 1 - 3 - 3 - 93	NA	NA	NA	NA	NA	NA	[illegible]	823	93.00	3	3	1400	nur [illegible]	TRUE
255	Bad Vilbel - 61118 - johannesgutenberg - 11 - 3 - 3 - 93	NA	NA	NA	NA	NA	NA	[illegible]	8258	93.00	3	3	1500	nur [illegible]	TRUE
256	Bad Vilbel - 61118 - johannesgutenberg - 11 - 3 - 4 - 112.9	NA	NA	NA	NA	NA	NA	[illegible]	2142	112.90	3	4	1860	nur [illegible]	TRUE
257	Bad Vilbel - 61118 - johannesgutenberg - 3 - 1 - 2 - 61	NA	NA	NA	NA	NA	NA	[illegible]	3081	61.00	1	2	1020	nur [illegible]	TRUE
258	Bad Vilbel - 61118 - johannesgutenberg - 3 - 1 - 2 - 61.2	NA	NA	NA	NA	NA	NA	[illegible]	3681	61.20	1	2	980	nur [illegible]	TRUE
259	Bad Vilbel - 61118 - johannesgutenberg - 3 - 1 - 3 - 84.7	NA	NA	NA	NA	NA	NA	[illegible]	3253	84.70	1	3	1420	nur [illegible]	TRUE
260	Bad Vilbel - 61118 - johannesgutenberg - 3 - 2 - 2 - 62	NA	NA	NA	NA	NA	NA	[illegible]	896	62.00	2	2	975	nur [illegible]	TRUE
261	Bad Vilbel - 61118 - johannesgutenberg - 3 - 2 - 3 - 83.73	NA	NA	NA	NA	NA	NA	[illegible]	4763	83.73	2	3	1260	nur [illegible]	TRUE
262	Bad Vilbel - 61118 - johannesgutenberg - 3 - 2 - 3 - 84.7	NA	NA	NA	NA	NA	NA	[illegible]	5365	84.70	2	3	1355	nur [illegible]	TRUE
263	Bad Vilbel - 61118 - johannesgutenberg - 3 - 4 - 2 - 56.8	NA	NA	NA	NA	NA	NA	[illegible]	2067	56.80	4	2	960	nur [illegible]	TRUE
264	Bad Vilbel - 61118 - johannesgutenberg - 3 - 4 - 2 - 56.8	NA	NA	NA	NA	NA	NA	[illegible]	1389	56.80	4	2	960	nur [illegible]	TRUE
265	Bad Vilbel - 61118 - johannesgutenberg - 5 - 3 - 2 - 62.1	NA	NA	NA	NA	NA	NA	[illegible]	2798	62.10	3	2	950	nur [illegible]	TRUE
266	Bad Vilbel - 61118 - johannesgutenberg - 7 - 0 - 2 - 64	NA	NA	NA	NA	NA	NA	[illegible]	4508	64.00	0	2	1090	nur [illegible]	TRUE
267	Bad Vilbel - 61118 - johannesgutenberg - 7 - 2 - 2 - 63.4	NA	NA	NA	NA	NA	NA	[illegible]	1304	63.40	2	2	1005	nur [illegible]	TRUE
268	Bad Vilbel - 61118 - johannesgutenberg - 7 - 4 - 2 - 58	NA	NA	NA	NA	NA	NA	[illegible]	6001	58.00	4	2	970	nur [illegible]	TRUE
269	Bad Vilbel - 61118 - johannesgutenberg - 7 - 4 - 3 - 99	NA	NA	NA	NA	NA	NA	[illegible]	5517	99.00	4	3	1680	nur [illegible]	TRUE
270	Bad Vilbel - 61118 - johannesgutenberg - 7 - 4 - 3 - 99	NA	NA	NA	NA	NA	NA	[illegible]	8277	99.00	4	3	1920	nur [illegible]	TRUE
271	Bad Vilbel - 61118 - johannesgutenberg - 7 - 4 - 3 - 99	NA	NA	NA	NA	NA	NA	[illegible]	7831	99.00	4	3	1680	nur [illegible]	TRUE
272	Bad Vilbel - 61118 - johannesgutenberg - 9 - 0 - 2 - 61.8	[illegible]	27MK75N	61.80	0	2	NA	NA	NA	NA	NA	NA	NA	nur [illegible]	TRUE
273	Bad Vilbel - 61118 - johannesgutenberg - 9 - 0 - 2 - 62	NA	NA	NA	NA	NA	NA	[illegible]	6992	62.00	0	2	975	nur [illegible]	TRUE
274	Bad Vilbel - 61118 - johannesgutenberg - 9 - 1 - 2 - 61.1	NA	NA	NA	NA	NA	NA	[illegible]	352	61.10	1	2	945	nur [illegible]	TRUE
275	Bad Vilbel - 61118 - johannesgutenberg - 9 - 2 - 2 - 61	[illegible]	27LVT5T	61.00	2	2	NA	[illegible]	7992	61.00	2	2	1000	beide Portale	TRUE
276	Bad Vilbel - 61118 - johannesgutenberg - 9 - 2 - 3 - 83.8	NA	NA	NA	NA	NA	NA	[illegible]	7611	83.80	2	3	1400	nur [illegible]	TRUE
277	Bad Vilbel - 61118 - johannesgutenberg - 9 - 2 - 3 - 86	NA	NA	NA	NA	NA	NA	[illegible]	731	86.00	2	3	1500	nur [illegible]	TRUE
278	Bad Vilbel - 61118 - johannesgutenberg - 9 - NA - 2 - 61.8	NA	NA	NA	NA	NA	NA	[illegible]	78	61.80	NA	2	990	nur [illegible]	TRUE
279	Bad Vilbel - 61118 - kurtmoosdorf - 52 - NA - 6 - 171	NA	NA	NA	NA	NA	NA	[illegible]	3449	171.00	NA	6	1094250	nur [illegible]	TRUE

Obviously, this location is an area with several newly constructed buildings (figure 17), many of them presumably very similar in size and price. Given the matching key (consisting of city, postal code, street name, house number, floor number, number of rooms, and size in m²), there is only one match (marked green), and only two advertisements, that could refer to the same object since the square metres differ only by a small amount (marked red). In general, advertisements come mainly from one portal, only few advertisements referring to this address stem from the other portals (3 advertisements out of 28 advertisements). In general, this constitutes a typical example for the combination of advertisements from these two portals. Overlap between both seems to be small.

Figure 17: Map service images (Example 8)

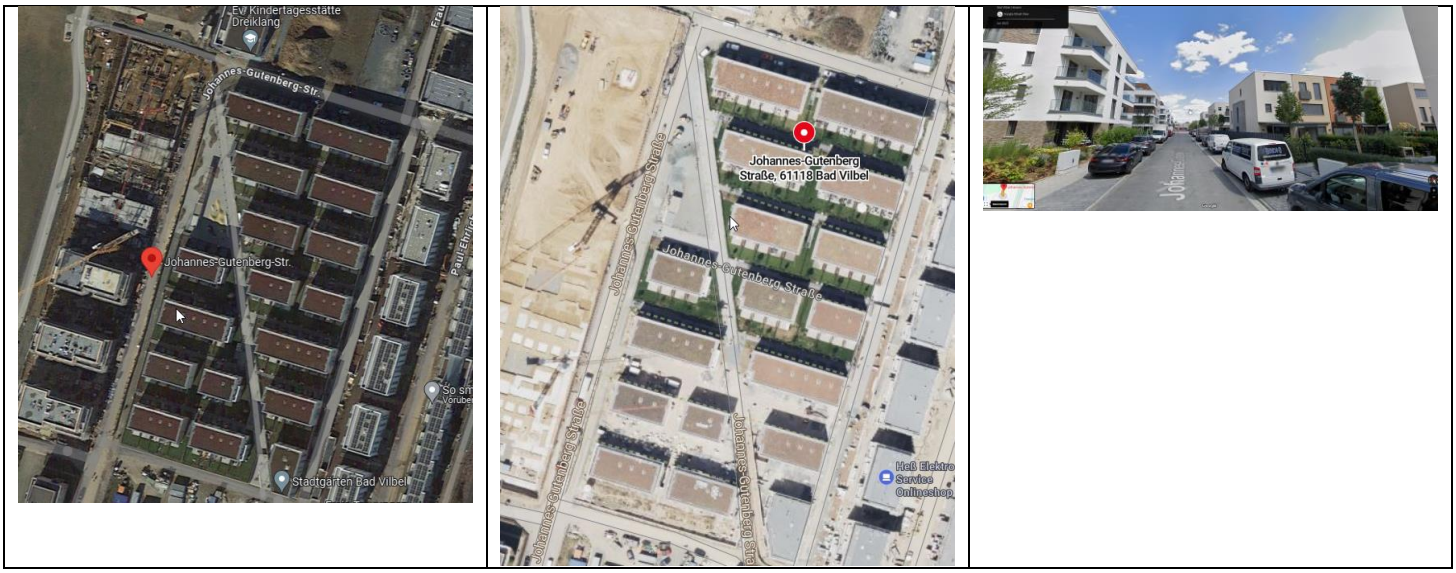


Table 12 shows the number of advertisements for 2022 where different strategies are used to get a rough estimate of the potential number of deduplicated advertisements between Portals 1 to 4 (see table 3).

Table 12: Deduplication of advertisements from 2022 by address aggregation

Raw Number of Advertisements …	**8536**	
… deduplicated by number of portals with an advertisement at the same address		7460
… deduplicated: a third of all ads with the same address are duplicates		6450
… deduplicated: the half of all ads with the same address are duplicates		4938
… deduplicated: two thirds of all ads with the same address are duplicates		3821
… deduplicated: number of unique addresses		2072

Overall, 8.536 advertisements in 4 portals refer to 2.072 unique addresses (unique up to street name). This number seems to be too low, or put differently, the approach seems too strict to treat all advertisements at a given address as duplicates. However, it gives the lower bound of such a

number. Heuristically or empirically, it could turn out that about a third of all advertisements at the same address are duplicates of one object.

Another heuristic approach to deduplication between portals could correlate the number of duplicates at a given address with the number of portals that report an advertisement at this given address: for example, if two portals report 6 advertisements at one address. One could correct the number of advertisements by the factor of two (as there are two portals) and only count 3 objects at this address. Similarly, if 4 portals in sum report 12 advertisements at a given address, one would end up with 12 advertisements at 4 portals, equating 3 distinct objects. Treating the 8.536 advertisements at 2.072 different addresses like this, one would end up with an estimate of 7.460 different objects and 1.075 duplicates.

5.4.3. Overcoverage, Undercoverage

Especially in areas with a high demand of apartments advertisements on real estate web portals only appear for short period of time – meaning days or even hours. When scraping advertisements from real estate web portals, there is a risk that these short-term advertisements are missed if the scraping interval is longer than the period in which a short-term advertisement is online. This leads to undercoverage. One partner, SSI-BBB investigated undercoverage by adjusting the scraping interval (see section 6.4.4 “Completeness of Scraped Data”).

A scraping interval of one week was selected for scraping the real estate web portals used in this use-case. Advertisements that are online shorter than one week therefore are missed. However, short-term advertisements may be covered if they happen to be online at the day and time of scraping.

With one data source – Portal 1 –, there is no undercoverage, since all advertisements for newly constructed objects that have appeared in a specific month on this portal are part of the data set delivered by the portal owner. For the other portals, the amount of undercoverage is unknown. The problem of overcoverage is assumed to be more prominent in urban areas and for apartments.

5.4.4. Other Quality Issues

5.4.4.1. Maintaining Scrapers

Maintaining scrapers is a challenge and involves checking regularly for minor or major changes of the portals’ sites’ structure. Smaller changes (e.g. changes of single HTML tags or XPATH expressions) may stay undetected for a while since even file sizes after scraping and data processing will not change dramatically. Major changes of a portal due to a complete re-design of the page can often be detected easily and immedeately but adapting the scraper to this new structure takes some time. Adapting the scraper sometimes only means that adjusting one or several XPATH expressions have to be readjusted in order to extract information from the new page structure. During the project phase, this was the case with one of the portals for Hesse and the process of fixing the scraper took some time. In other cases, not only one or more XPATHs must be adapted, but the entire process of accessing the portal must be adapted.

6. Country Specific Part: Statistical Office Berlin-Brandenburg (SSI-BBB)

6.1. Data Sources

Evaluation of Relevant Websites

In collaboration with a second German partner, Hesse, participating in both use-case 1 (UC1) and use-case 2 (UC2) within Work Package 3 (WP3), a comprehensive list of relevant websites was compiled. From the outset, sites with a regional focus outside of Hesse and the Berlin-Brandenburg area were excluded from consideration.

Given that both partners are involved in UC1 and UC2, and to leverage potential synergies, the checklist was designed to prioritize portals that meet the requirements of both use-cases, assigning them a higher score. The objective was to ensure that the web scrapers developed could be utilized across both use-cases. Consequently, despite the high performance ratings of certain portals, they were ultimately excluded from further consideration in the project.

The evaluation of the list was conducted in close collaboration with Hesse, utilizing a checklist based on predetermined criteria. To facilitate this assessment, a specially developed R script was employed alongside manual verification processes. Following this initial evaluation, only a limited number of portals remained for further consideration. The final selection of portals was made collaboratively with Hesse to ensure alignment with the project objectives.

Overview of Key Real Estate Web Portals for Use-Case 1

Three major portals, referred to as Portal 2, Portal 7, and Portal 5, which finally were considered for use-case 1, offer distinct features that differentiate them from one another. The following outlines the key characteristics of these portals, providing a deeper understanding of their specific roles within the real estate landscape.

Portal 2 is a comprehensive real estate web portal, covering a broad range of real estate, with a particular emphasis on rental and purchase properties. Its extensive database is regularly updated, ensuring users have access to the latest advertisements. With nationwide coverage, Portal 2 lists properties in numerous cities and regions across Germany. It maintains strong partnerships with various stakeholders in the real estate market, such as real estate agents and developers, thereby expanding the breadth of its listings. Portal 2 has a user-friendly search interface, offering multiple filters (e.g., price, size, number of rooms) to tailor searches to individual needs. This portal serves as the primary data source for this use-case by Berlin-Brandenburg.

The second portal, Portal 7, while similar in scope, covers a wide range of listings, including residential properties, commercial spaces, and land. Portal 7 also boasts nationwide coverage, with a strong presence in urban areas. This extensive geographic reach ensures that users can find properties in both rural and metropolitan regions. Since there is a large overlap, this portal was not used as a primary data source.

In contrast to the broader focus of the previous portals, Portal 5 specializes in new construction projects. It specifically targets users interested in newly built properties, including ongoing and planned residential developments. Although Portal 5 offers nationwide coverage, its focus is primarily on larger cities and regions with active construction projects. This targeted approach makes it particularly useful for users seeking new-built properties in urban areas. This portal was not considered in use-case 1; instead, it was utilized only for use-case 2, as it primarily focuses on new constructions.

Ensuring Stability and Coverage

As previously mentioned, Portal 2 was used as the primary portal for webscraping, while the second major portal (Portal 7) served as a backup. This approach ensures that potential data gaps, caused by disruptions, are minimized as they can be compensated by data from the second portal. Since websites tend to change rather frequently, data gaps are always a risk when such changes occur, as the scraping code needs to be adjusted accordingly. This process may take more than 24 hours, depending on the complexity of the changes. The data was scraped daily, which represents a relatively high frequency.

The two real estate web portals belong to a shared parent company and offer nearly identical content. This parent company is the second largest provider on the German market, following the market leader. Additionally, the company is active in Austria and Switzerland, which means that the scrapers developed within the project can be adapted for these countries with minimal effort. In addition to gaining knowledge in webscraping, the size and nationwide coverage of the portals, allow to produce statistically reliable results for Germany.

After a location- or postal-code-based search, which can be further refined by criteria such as transaction type (sale, rental) and property type (apartment, house, room, commercial property, garage, land, new construction project), the search results are displayed in sets of 20 listings per page. Each listing is identified by a unique identification number (ID) and includes images of the property, location details, and basic information. This is followed by a text description, the agency's advertisement number, and the listing's publication date.

The listings share a consistent HTML structure and similar class names. The textual description is available as plain HTML text, while the location information includes postal codes that can be mapped to NUTS-3 regions.

Regional Focus and Collaborative Approaches

The German statistical offices focused their data collection efforts on specific regions, namely Hesse and the Berlin/Brandenburg metropolitan area. HSL and SSI-BBB established a systematic and consistent data exchange protocol for the data they collected, ensuring that each office was responsible for gathering data from designated real estate web portals. This approach not only reduced the workload for the statistical offices but also promoted closer collaboration between the two German offices.

SSI-BBB continued to scrape data from two real estate web portals. However, due to a significant overlap in listings between these portals (for more details, see use-case 2), data from Portal 7 was no longer included in the analysis. Nonetheless, data scraping from this portal remains active for backup purposes. For the remaining two portals, Portal 2 and Portal 5, monthly data has been consistently available since May 2022.

Although the colleagues from Hesse also collected data for the Berlin-Brandenburg region, these were not included in the present analysis. There are two main reasons for this decision: First, the amount of data collected was relatively small, meaning it would have had little impact on the statistical validity of the analysis. Second, cross-portal deduplication posed a significant challenge, as the extent of overlap between the portals was difficult to estimate. Given the unclear benefit of including this additional data and the disproportionate effort required, it was decided not to incorporate it into the analysis.

6.2. Data Preparation

Data cleaning and editing

Only variables that can be directly accessed in the scraped HTML code were analysed. Information that is only contained in larger free-text fields (e.g. "… the house has a large garage") is saved, but not evaluated.

Classification of the mandatory variables is done in accordance with the pre-agreed between the partners method.

The data obtained from scraping requires a comprehensive data cleaning process to be usable for analysis. This process applies not only to data obtained through screen scraping but also to data acquired via APIs. For example, string elements need to be removed from value fields, such as converting a price into an integer since it is treated as a string due to the euro symbol. Additionally, it is necessary to convert the entry "no information" into missing values.

The portals differ heavily when it comes to the characteristics that are present in the advertisements. Whilst in some portals there is a huge number of very special characteristics present, some ads even lack information on pre-agreed mandatory variables.

Localization

In general, every advertisement in the portals scraped by SSI-BBB contains fine grained data on the object's address – at the very least the postal code of the object is available. Since the postal code areas in Germany are much more fine-grained than the NUTS 3-level, it is a very simple task to aggregate the data to the NUTS levels for the project's purposes.

The fine-grained local information is one of the key elements for deduplication and can be the basis for very detailed maps. Unfortunately, deduplication is still not so simple, as much more accurate address data would be necessary for that purpose. Further details are mentioned below (see "Deduplication").

Deduplication

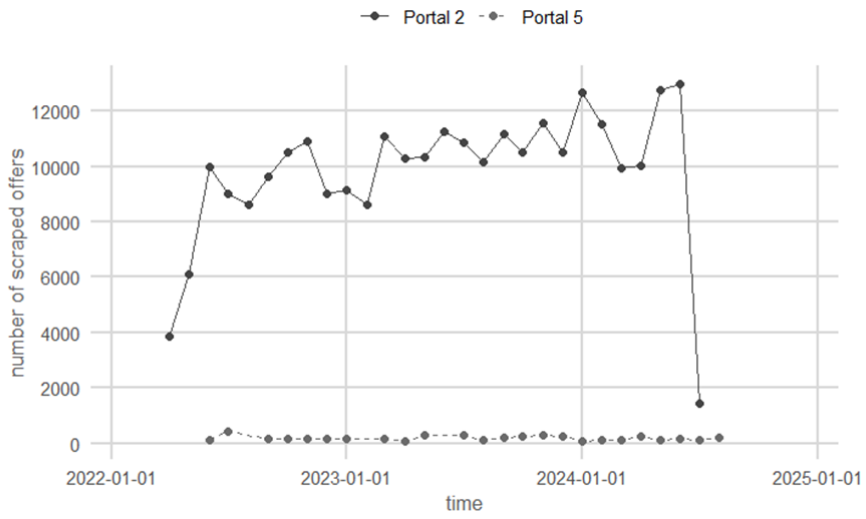
One of the major challenges for SSI-BBB is deduplication: Within the portals a simple object id-based approach was used to detect duplicates. The detection of duplicates across different portals is still a problem with no reliable solution in sight. This issue is described in detail in section 6.4.3 "Duplicates".

6.3. Results

6.3.1. General Results

During the entire observation period, advertisements from Portal 2 and Portal 5 were continuously scraped from April 2022 to July 2024. The monthly obtained data from both portals are shown in figure 18. Due to complications with the development of the scraper, data from Portal 2 appears to be incomplete in the first two months, as initial experiments were conducted to ensure the code functioned properly. Furthermore, in the final month, significant changes to the website led to the discontinuation of the data collection method, resulting in further gaps in the time series. In order to produce reliable results data were trimmed for presenting average values, but entirely used for cumulative comparatives of recorded new construction (construction year 2023) and data of the official statistics. The effect can be observed in the following figure, which presents the complete time series. In contrast, with the first and last months omitted, the time series is also displayed for comparison.

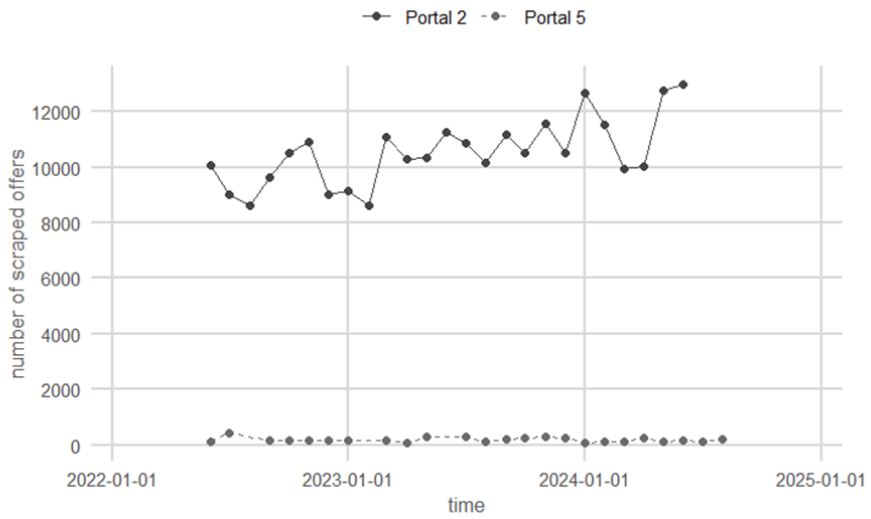
Figure 18: Scraped date (April 2022 to July 2024) for Berlin and Brandenburg (Portal 2, Portal 5)



The following figure 19 illustrates the reduced time series of data scraped for a specific date (by month). It is evident that the volume of data from the two portals significantly differs. However, it is essential to note that Portal 5 exclusively focuses on new construction projects, whereas Portal 2 encompasses the entire real estate market, with only a portion of the collected data representing new builds.

Although the fluctuations appear to be substantial, a slight upward trend can be observed. However, it remains unclear whether this trend is substantiated or merely coincidental, as the time series is not excessively long. The monthly variations may also be influenced by the differing number of days in each month, theoretically providing more opportunities for data to be posted online in months with more days. However, this remains a hypothesis and requires further investigation to confirm its validity.

Figure 19: Scraped date (June 2022 to June 2024) for Berlin and Brandenburg (Portal 2, Portal 5)



As previously mentioned, the data from Portal 2 must first be filtered to identify the relevant cases for new constructions. This filtering is necessary because the raw data contains a large number of listings, only a portion of which pertains to new builds. The goal of this step is to isolate those listings that (presumably) represent new constructions. This filtering process is crucial, as it determines the quality and reliability of the final dataset.

In contrast, this filter is not required for Portal 5. The reason for this is that the listings in Portal 5 exclusively pertain to new constructions, making additional filtering in this case unnecessary.

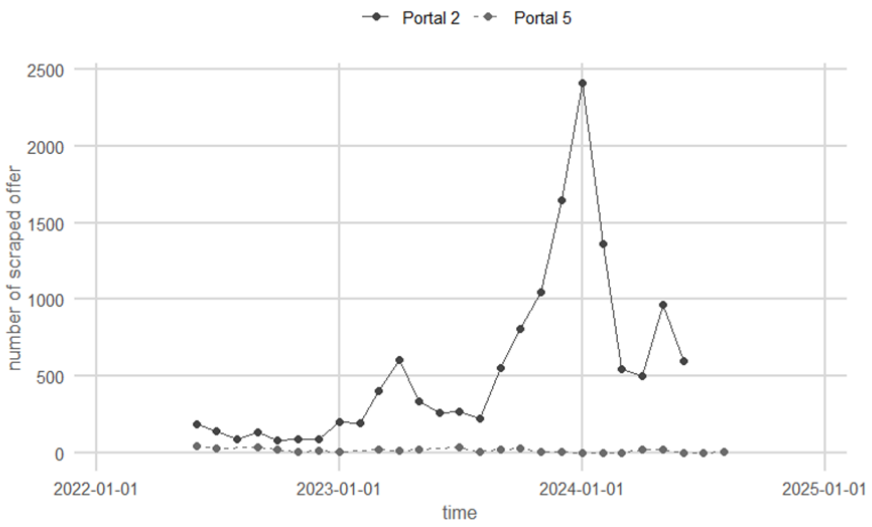
The following figure 20 provides an illustration of the filtering effect for Berlin, showing how many cases remain after extracting the presumed new constructions. The term "presumed" is used because it cannot be determined with absolute certainty during the filtering process whether the cases in question truly represent new constructions. Despite this uncertainty, the filtering process is essential for the use-case, as it forms the data basis.

Further information on this topic and a detailed discussion of the filtering methods can be found in Section 6.4.2 “Selection criteria” and in Section 2.4.1 “Defining ‘Newly Constructed’ Properties”.

A significant increase in advertisements around the turn of the year to 2024 can be observed in Berlin in particular, although it is not entirely clear why this is occurring. Furthermore, even before 2023, there are new constructions listed with a 2023 construction year. It is likely that properties are often listed before their actual completion, with a future construction year indicated. This means that properties can appear on the market well in advance of their actual construction year (in this case, 2023). Similarly, new constructions completed in 2023 may not be listed and sold until 2024 or later. Such delays in marketing generally seem plausible.

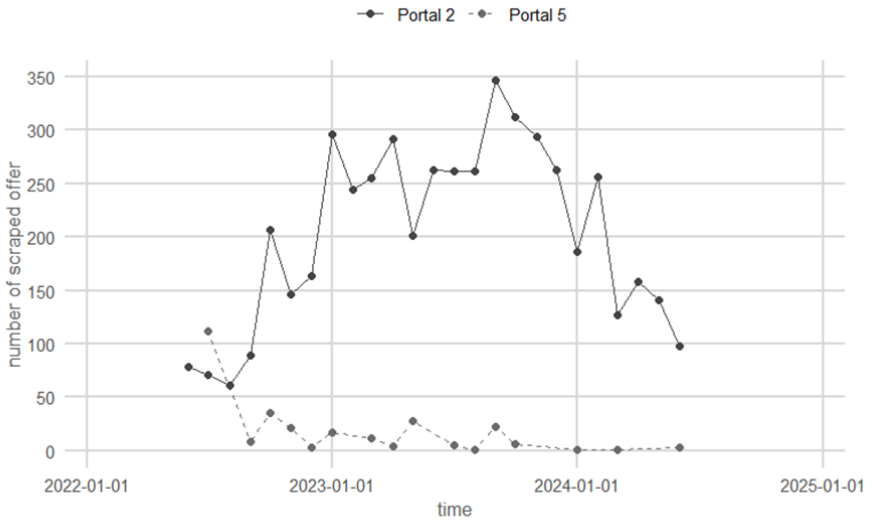
However, it remains unclear to what extent listings are posted prematurely or delayed, as the data for the previous year, 2022, and the following year, 2024, are not yet fully available. This temporal shift presents an additional challenge, as the new constructions of a given year are often spread across several years on the online real estate market.

Figure 20: Scraped data of new construction advertisements (construction year 2023) for Berlin (Portal 2, Portal 5)



For Brandenburg, a slightly different picture emerges, which is understandable given the significantly lower number of cases compared to Berlin. Nevertheless, the general phenomenon is also evident here: new constructions from 2023 appear in listings both in the previous year and in the following year. This effect is evident in the following figure 21.

Figure 21: Scraped data of new construction advertisements (construction year 2023) for Brandenburg (Portal 2, Portal 5)

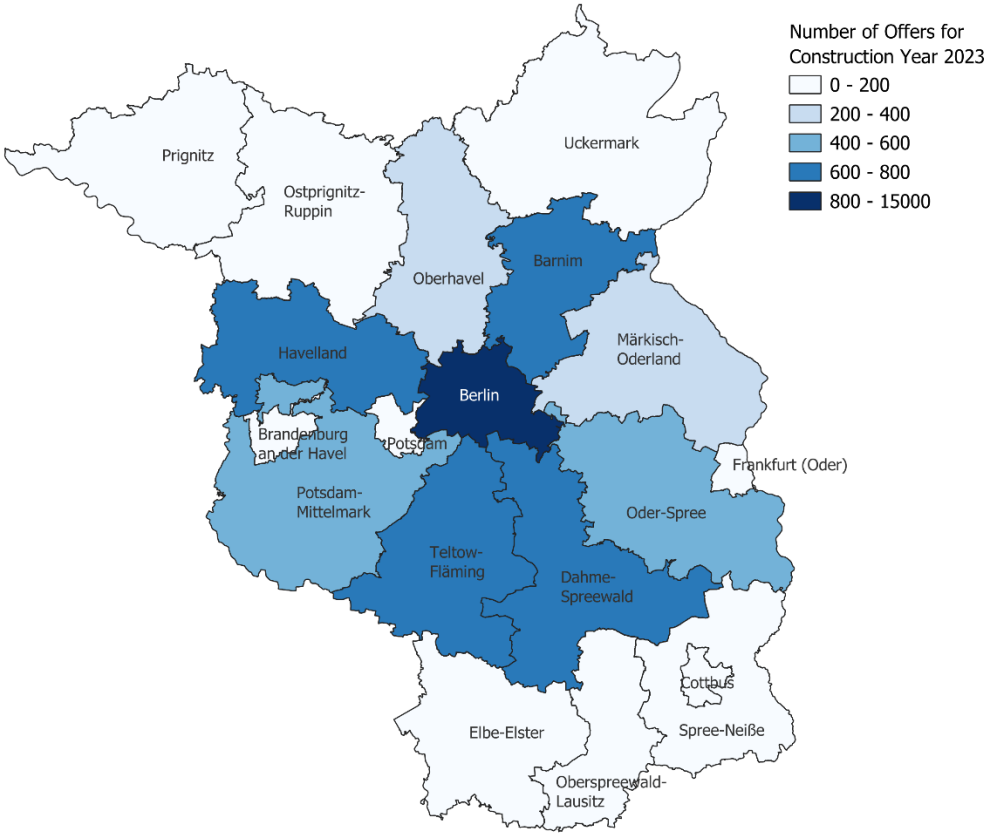


In summary, it can be stated that the proportion of new constructions relative to the total dataset in Portal 2 is relatively low and can be roughly estimated at a maximum of 25 %. However, this value varies depending on how "new constructions" are defined and filtered from the dataset. There is considerable room for interpretation here, which cannot be clearly categorized as "right" or "wrong." Nevertheless, the remaining dataset provides a significant foundation for further analysis.

6.3.2. Comparative Analysis

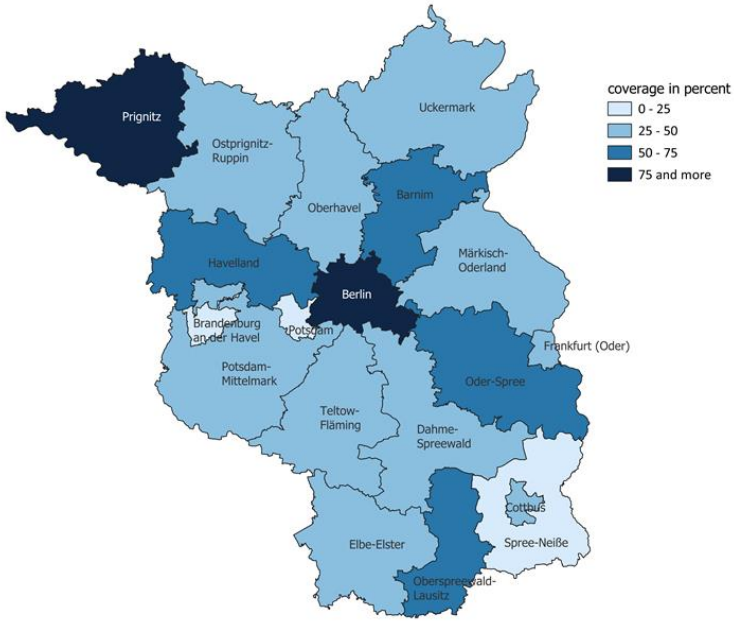
The map below (Figure 22) illustrates the average monthly volume of new construction listings at the NUTS-3 level across Berlin and Brandenburg for the year 2023. Notably, there is a significant variation in the volume of new constructions among different NUTS-3 regions within Brandenburg. This disparity highlights the diverse dynamics of the real estate market across the region, which may reflect varying demand, local development policies or economic conditions. Understanding these differences is crucial for accurately interpreting the data and for making informed decisions regarding future construction projects and investment strategies in the area.

Figure 22: Monthly average of newly listed properties (built in 2022) for sale or rent on Portal 2, April 2022 to July 2024, by region



The second map (Figure 23) offers a comprehensive comparison between the number of newly constructed listings with a construction year of 2023, scraped from April 2022 to July 2024, and the officially released figures for newly constructed properties. As previously noted, discrepancies may arise primarily due to properties not being advertised and, to a lesser extent, data inaccuracies. However, for the purposes of nowcasting, it is more crucial to maintain a consistent ratio between the number of scraped listings and the official statistics over time than to achieve precise coverage. This approach allows for a more reliable interpretation of current market trends and assists in making informed statements regarding future developments.

Figure 23: Proportion of newly constructed properties (built in 2023) listed for sale or rent on Portal 2 compared to totally new constructions in 2023 according to official Statistics, by region, April 2022 to July 2024.



6.4. Quality Aspects

6.4.1. Missing Data

Missing data presents a significant challenge, particularly concerning the analysis of relevant information. It is crucial to determine whether there are complete cases or if only certain attributes are missing from the listings. The absence of complete cases essentially leads to undercoverage, which is mentioned in the section 6.4.5 "Overcoverage and Undercoverage." This, in turn, leads to the dataset being less valuable for forecasting.

However, the consequences of missing individual attributes can vary. Some attributes are less relevant, such as whether a new construction includes a fitted kitchen or not. On the other hand, certain attributes can be highly significant for the substantive evaluation, particularly when information like price is missing. The situation becomes especially critical when details about the exact location of the new construction are missing, as this makes deduplication across multiple portals impossible.

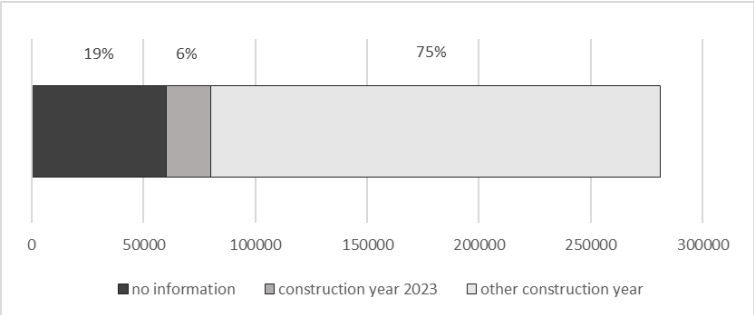
A missing construction year also poses a significant problem in this context. Since the dataset covering the entire real estate market must necessarily be filtered to focus on new constructions, the lack of this information could lead to biases in the analysis results. More on this can be found in the following section, "Selection Criteria for 'New Construction'." In such cases, it is essential to consider appropriate methods for data enrichment or cleaning to ensure the robustness of the analyses.

6.4.2. Selection Criteria ‘New Construction’

The web scraping operation for this project encompasses the entirety of available real estate listings as these were needed for UC1. For UC2, one task is to accurately identify new constructions within this extensive collection. The primary challenge lies in establishing a definition that encapsulates the

largest possible proportion of newly constructed properties. In the absence of a reliable attribute in the scraped data that directly labels listings as new constructions, the year of construction emerges as the most viable identifier. However, almost 20 % of our listings do not contain this information, a common issue in web scraping where data often is incomplete or incorrect due to reliance on non-standardized text descriptions (see figure 24). In this case, missing values represent a significant challenge, as they are crucial for identifying the data relevant to the analysis.

Figure 24: Proportion of individual properties newly listed for sale or rent on Portal 2 and Portal 5 from April 2022 to July 2024, categorised by availability of the feature Construction Year 2023



Another notable difficulty arises in urban areas, where properties are often sold well before their completion. Consequently, the listings of those properties in online portals often lack precise information on the year of construction, as this detail is often undecided at the time of market listing. Listings without a construction year were simply excluded. In cases where multiple years are provided (e.g., "2022 or 2023", "2022-2024"), we default to using the latest year mentioned (see the description of HSL above for another, more strict definition, that includes the condition of “first occupancy” additionally to the year of construction).

6.4.3. Duplicates

Quality assessment: address completeness and duplicate detection

In assessing data quality, two critical areas are the completeness of address information and the identification of duplicate listings. For individual portals, duplicates are efficiently detected using portal-specific identification numbers. However, cross-portal duplicate recognition primarily relies on address data, which does not guarantee absolute certainty, especially when multiple units within the same address are marketed separately. This scenario is particularly common in Berlin, where numerous residential units in a building may be individually listed for sale or rent upon completion.

Additionally, properties can be listed twice in scenarios where individual apartments or even entire buildings are initially sold and later, individual apartments within these buildings are offered for rent. Rapid turnovers in such cases could lead to double counting. The challenge extends beyond identifying duplicates at a single point in time to tracking them over a period.

The crux of duplicate identification lies in having precise address details. Regrettably, in our dataset, complete addresses are available for only about 22 % of the listings. This limitation stems from various factors. For instance, addresses are often provided by real estate agents only upon request, and in the case of new constructions, addresses may not yet exist, especially in newly developed residential areas where house numbers and even street names are often being established during or even after construction.

The availability of address data is presented below, with distinctions between overall advertisements (figure 25) and new constructions (figure 26).

Figure 25: Proportion of scraped data for sale or rent on Portal 2 from April 2022 to July 2024, categorised by completeness of address information

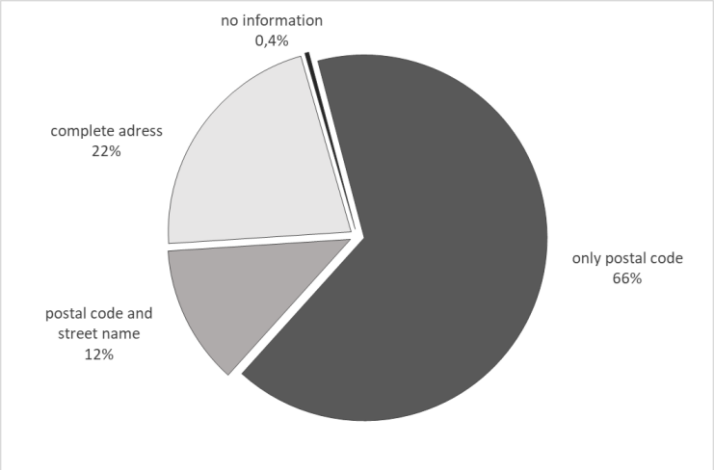
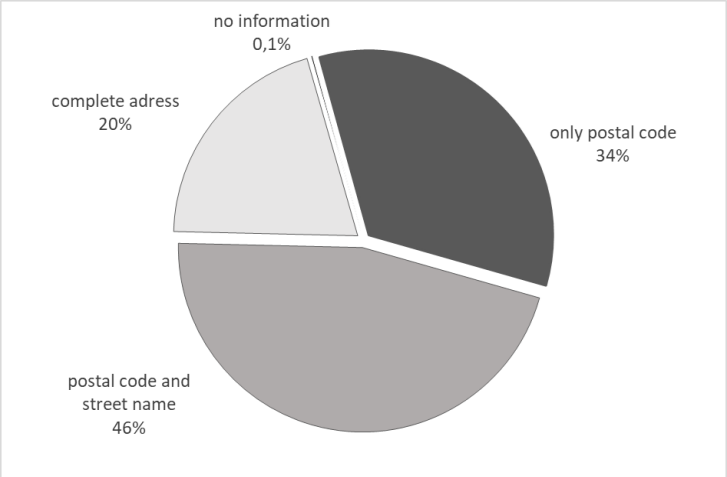


Figure 26: Proportion of newly listed properties (built in 2023) for sale or rent on portal 2 from April 2022 to July 2024, categorised by completeness of address information



Furthermore, even in cases where a complete address is available in both portals, deduplication is only possible to a limited extent or rather not reliable, especially concerning new buildings or complete refurbishments of entire buildings.

Example 1:

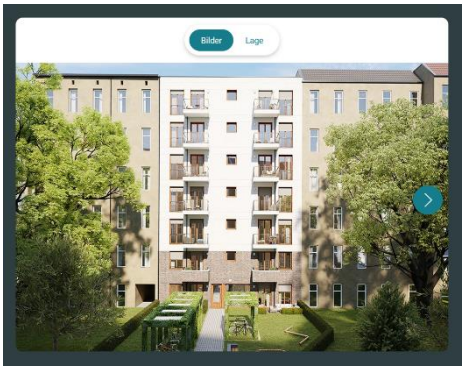
One example is a construction project in Berlin, which is advertised on both Portal 5 and Portal 2.

	A	B	C	E	F	G	H	I	J	L	M	N	O	Q	S
1	ad_provider	date_scraped	ad_id	location	postal_code	street	number	federal_state	building_type	offer_transact	offer_floor	offer_room	offer_price_sale	price_sell_meter	offer_surface
2	Portal 2	12.04.2024 08:30		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	1. Geschoss	1	259900	918.374.558.303.887	28,30
3	Portal 5	16.07.2024 00:00	268515	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	1.OG	1	259900	9272.21	28,03
4	Portal 5	25.10.2023 00:00	241260	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	259900	9272.21	28,03
5	Portal 2	23.06.2024 08:38		Birkenstraße 12A, 10559, Berlin	10559	Birkenstraße	12A	Berlin	wohnung	sale	1. Geschoss	1	261400	913.986.013.986.014	28,60
6	Portal 2	12.04.2024 08:30		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	1. Geschoss	1	262000	916.083.916.083.916	28,60
7	Portal 5	25.10.2023 00:00	241257	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	262000	9337.13	28,06
8	Portal 2	12.04.2024 08:31		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	2. Geschoss	1	265000	939.716.312.056.738	28,20
9	Portal 5	25.10.2023 00:00	241264	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	265000	Mai 89	449,85
10	Portal 2	22.04.2023 07:50		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	zinshaus_rendliteobjekt	sale	2 Geschosse	1	267000	933.566.433.566.433	28,60
11	Portal 2	19.09.2023 08:51		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	2. Geschoss	1	267000	933.566.433.566.433	28,60
12	Portal 5	15.08.2023 00:00	237229	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Apartment	sale	2.OG	1	267000	9335.66	28,60
13	Portal 5	25.10.2023 00:00	241261	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	267000	9515.32	28,06
14	Portal 5	16.07.2024 00:00	268523	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	3.OG	1	271000	9692.42	27,96
15	Portal 5	25.10.2023 00:00	241266	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	271000	9692.42	27,96
16	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	3. Geschoss	1	271900	972.460.658.082.976	27,96
17	Portal 2	12.04.2024 08:31		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	3. Geschoss	1	273900	100.735.564.545.789	27,19
18	Portal 5	16.07.2024 00:00	268528	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	3.OG	1	273900	10073.56	27,19
19	Portal 5	25.10.2023 00:00	241265	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	273900	10073.56	27,19
20	Portal 5	16.07.2024 00:00	268527	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	4.OG	1	279000	9978.54	27,96
21	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	4. Geschoss	1	279900	100.107.296.137.339	27,96
22	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	4. Geschoss	1	282900	104.045.605.001.839	27,19
23	Portal 5	16.07.2024 00:00	268525	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	4.OG	1	282900	10404.56	27,19
24	Portal 5	25.10.2023 00:00	241267	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	282900	10404.56	27,19
25	Portal 2	23.06.2024 08:39		Birkenstraße 12A, 10559, Berlin	10559	Birkenstraße	12A	Berlin	wohnung	sale	6. Geschoss (D	1	286500	100.174.825.174.825	28,60
26	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	5. Geschoss	1	289900	103.683.834.048.641	27,96
27	Portal 5	16.07.2024 00:00	268524	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	5.OG	1	289900	10368.38	27,96
28	Portal 5	25.10.2023 00:00	241271	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	289900	10368.38	27,96
29	Portal 2	22.04.2023 07:50		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	5. Geschoss	1	292900	107.723.427.730.783	27,19
30	Portal 5	15.08.2023 00:00	237231	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Apartment	sale	5.OG	1	292900	10772.34	27,19
31	Portal 5	16.07.2024 00:00	268520	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	5.OG	1	292900	10772.34	27,19
32	Portal 5	25.10.2023 00:00	241268	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	292900	10772.34	27,19
33	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	6. Geschoss	1	297000	106.223.175.965.665	27,96
34	Portal 5	16.07.2024 00:00	268522	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	6.OG	1	297000	10622.32	27,96
35	Portal 5	25.10.2023 00:00	241275	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	297000	10622.32	27,96
36	Portal 5	16.07.2024 00:00	268521	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	5.OG	1	299000	10996.69	27,19
37	Portal 2	12.04.2024 08:29		Birkenstraße 12a, 10559, Berlin	10559	Birkenstraße	12a	Berlin	wohnung	sale	6. Geschoss	1	299000	109.966.899.595.439	27,19
38	Portal 5	16.07.2024 00:00	268526	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	6.OG	1	299000	10996.69	27,19
39	Portal 5	25.10.2023 00:00	241272	Birkenstraße 12a, 10559 Berlin	10559	Birkenstraße	12a	Berlin	Etagenwohnung	sale	NA	1	299000	10996.69	27,19

Many of the residential properties can be found in both portals with almost identical information on living space and purchase price. However, key features such as the floor are missing in order to clearly and reliably identify the property as a duplicate. Added to this is the different date of registration by us and, in this case, a different provider of these properties on the respective portals.

In any case, it is uncertain whether the scraped objects are duplicates.

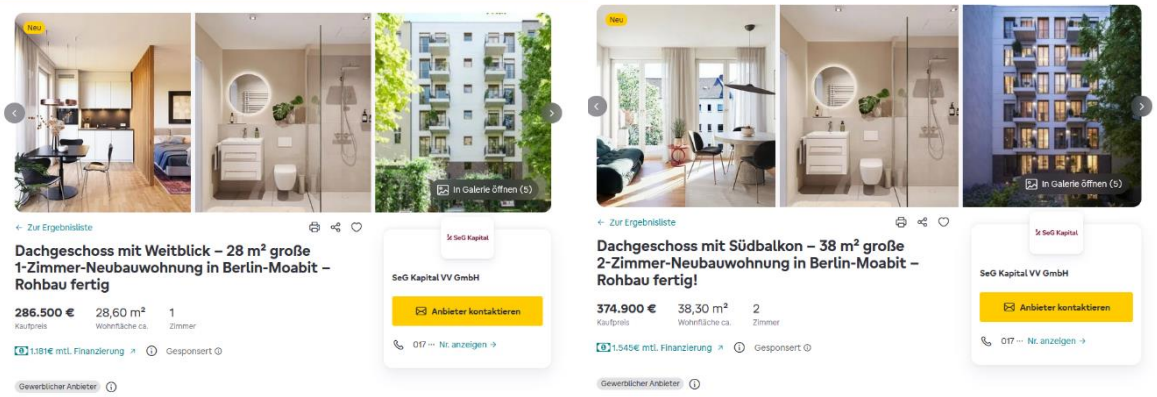
Portal 5



Projektdetails Baubeginn erfolgt

Adresse	Birkenstraße 12a, 10559 Berlin / Moabit
Wohntyp	Eigentumswohnung, Kapitalanlage, Mikroapartment
Preis	259.900 € - 649.000 €
Zimmeranzahl	1 - 2,5 Zimmer
Wohnfläche	27,19 m² - 75,62 m²
Bezugsfertig	4.Quartal 2024
Einheiten	26
Kategorie	Gehoben
Entfernungen	anzeigen

Portal 2

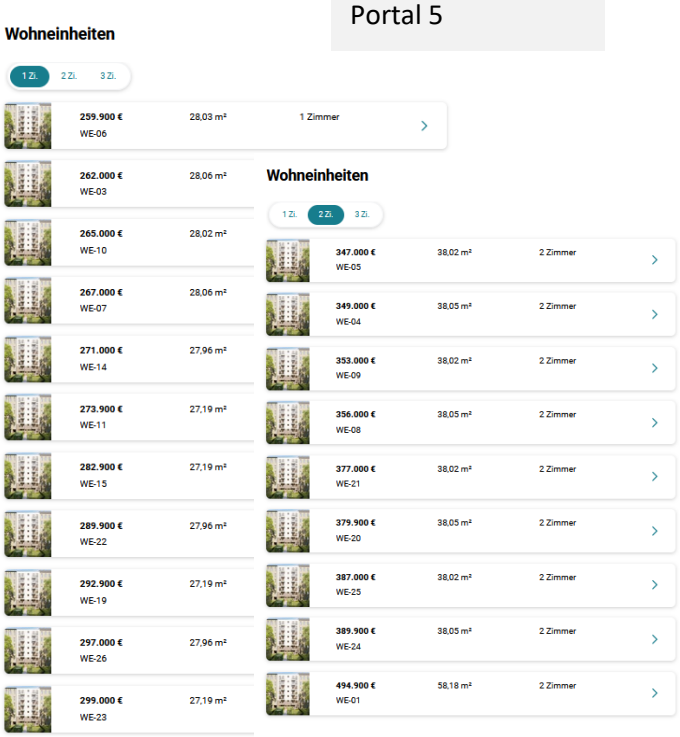


Example 2:

Detecting potential duplicates is possible if the most important features of the residential property, such as number of rooms, living space, price and floor, have been correctly specified. However, the decision whether these objects are actually duplicates is uncertain and can lead to incorrect exclusions of residential properties.

A major residential construction project in Berlin comprises several apartment blocks, each with 21 apartments. The scraped data show two completely identical residential properties with different IDs on each floor. With such large apartment blocks, it is not unusual to find apartments with almost or exactly identical layouts, in some cases only mirror-inverted. As this is a construction project that is only advertised in one portal, deduplication can be carried out using the existing property ID. But if the property is advertised in another portal at the same time, deduplication based on the characteristics is hardly possible.

Portal 5



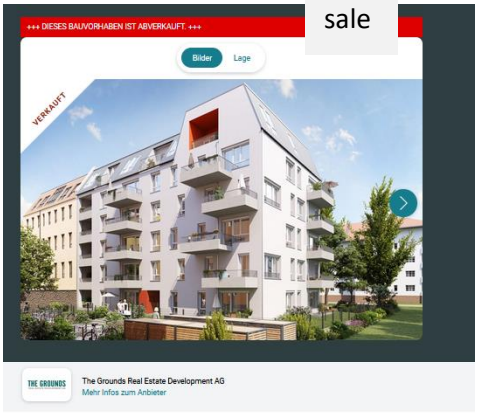
Example 3:

Another new construction project was advertised on both portals in January 2023, with a completion date of 2024. Also in this case there were many apartments with almost identical details. However, identification of duplicates was made more difficult by a different house number in the address.

Another problem that arose in connection with this construction project is the rental of previously purchased apartments. One apartment in this building can be found as a rental property on Portal 2 in June 2024.

	A	B	C	E	I	L	M	N	O	P	Q	S
1	ad_provider	date_scraped	ad_id	location	federal_state	offer_transaction_cat	offer_floor	offer_rooms	offer_price_sale	offer_price_qm	price_sell_meter	offer_surface
2	Portal 5	06.10.2022 00:00	204000	Margaretenstraße 24-25, 10317 Berlin	Berlin	sale	5.OG(DG)	2	412800	7050.38	7050.38	5855
3	Portal 5	06.10.2022 00:00	203987	Margaretenstraße 24-25, 10317 Berlin	Berlin	sale	2.OG	2	468600	6550.18	6550.18	7154
4	Portal 5	06.10.2022 00:00	203995	Margaretenstraße 24-25, 10317 Berlin	Berlin	sale	4.OG	2	468600	6550.18	6550.18	7154
5	Portal 5	25.01.2023 00:00	203991	Margaretenstraße 24-25, 10317 Berlin	Berlin	sale	3.OG	2	468600	6550.18	6550.18	7154
6	Portal 5	25.01.2023 00:00	203983	Margaretenstraße 24-25, 10317 Berlin	Berlin	sale	1.OG	2	475700	6649.43	6649.43	7154
7	Portal 2	14.01.2023 08:22		Margaretenstraße 24 A, 10317, Berlin	Berlin	sale	1. Geschoss	2	475700	0	664.942.689.404.529	7154
8	Portal 2	14.01.2023 08:22		Margaretenstraße 24 A, 10317, Berlin	Berlin	sale	3. Geschoss	2	468600	0	655.018.171.652.222	7154

sale



Projektdetails

Adresse	Margaretenstraße 24-25, 10317 Berlin / Lichtenberg
Wohntyp	Eigentumswohnung
Preis	Auf Anfrage
Zimmeranzahl	2 - 2,5 Zimmer
Wohnfläche	51,2 m² - 78,88 m²
Bezugsfertig	2024
Einheiten	27
Kategorie	Gehoben
Entfernungen	anzeigen

rent

← Zur Ergebnisliste

Hochwertige Neubauwohnung: moderne EBK mit Miele Geräten, Parkett und Fußbodenheizung zum Erstbezug

1.346 € Kaltmiete zzgl. NK | 56,07 m² Wohnfläche ca. | 2 Zimmer

Gewerblicher Anbieter

Margaretenstraße 24 A
10317 Berlin (Lichtenberg)
Auf Karte ansehen →

THE GROUNDS
THE GROUNDS Real Estate Development AG
Vivian Buchholz
Anbieter kontaktieren
+49 ··· Nr. anzeigen →

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ad_provider	date_scraped	ad_id	location	street	number	federal_state	building_type_ca	offer_transaction_ca	offer_floor	offer_rooms	offer_price_sale	offer_price_qm	offer_surface
2	Portal 5	07.11.2022 00:00	218785	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	1	277865	8500	3269
3	Portal 5	07.11.2022 00:00	218788	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	1	277865	8500	3269
4	Portal 5	07.11.2022 00:00	218786	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	2	379164	7600	4989
5	Portal 5	07.11.2022 00:00	218787	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	2	379164	7600	4989
6	Portal 5	07.11.2022 00:00	218784	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	3	614291	7608.26	8074
7	Portal 5	07.11.2022 00:00	218783	Carl-Spindler	Carl-Spindler	14	Berlin	apartments	sale	1.OG	3	614291	7608.26	8074

6.4.4. Completeness of Scraped Data

Duration of Online Listings in a Major Real Estate Web Portal

This study presents a temporary analysis conducted over nearly two months, aimed at investigating the visibility duration of online listings on one of the largest real estate web portals. The primary focus of this analysis is to address the question of how many listings may be "missed" if data is collected solely on a daily or even weekly basis. Given the high volatility in the housing market, capturing the dynamic changes in listing availability is of particular interest.

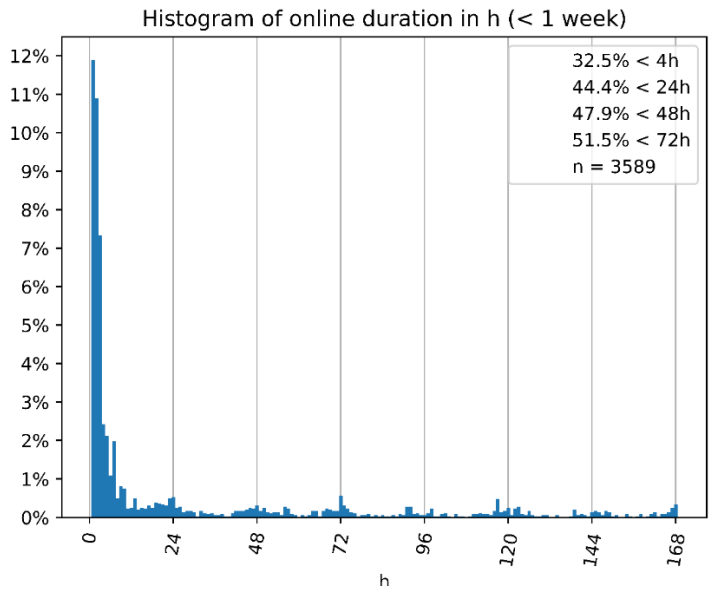
To achieve this, hourly data collection of online listings was implemented to determine which listings were active and which were no longer available. Based on this data, two critical pieces of information were incorporated into the dataset: the date and time of the first viewing of a listing ("first seen on") and the date and time of the last viewing ("last seen on"). This information facilitates the determination of the online duration of each listing. Listings that remain online for less than 24 hours may be considered unrecorded within the framework of a daily scraping process, suggesting that the market representation is incomplete.

It is essential to note that only listings for which the online duration could be determined were included in this study. Listings that were active prior to the observation period or were not taken offline during this period were excluded. This exclusion has direct implications for the results and their interpretation. It is possible that listings with a duration of up to four months may not have been captured, given that the observation period was shorter. Consequently, a trend may be observed suggesting that the determined percentages could be subject to downward correction with an extended observation duration.

The results of the analysis are presented in the following visualisations, which further differentiate between apartments and houses, as well as between sales and rentals. Notably, the most interesting findings are observed in Berlin (as a city-state), particularly regarding rental apartments, and in Brandenburg (as a rural state), where houses for sale show significant results. One of the graphs illustrates a histogram depicting the age of deleted IDs (listings) in hours, focusing on the period of one week. This histogram includes percentage representations of the distribution of all deleted IDs grouped by the following time intervals:

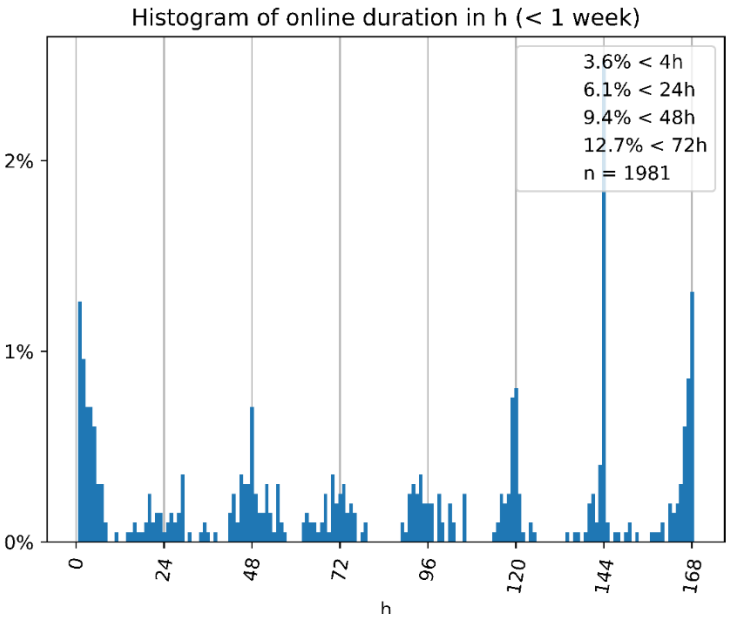
- < 4 hours: Listings that were online for less than 4 hours,
- < 24 hours: Listings that were online for less than 1 day,
- < 48 hours: Listings that were online for less than 2 days,
- < 72 hours: Listings that were online for less than 3 days

Figure 27: Histogram of recorded IDs (listings) deleted during the study period, categorised by hours of their duration online



This histogram focuses on Berlin's rental apartments. The data indicate that a significant portion of listings is removed quickly from the portal, likely due to rapid rentals. The total number of observations (n) is 3,071. This representation covers only the first week (168 hours) of the observation period.

Figure 28: Histogram of recorded IDs (listings) deleted during the study period, categorised by hours of their duration online



This analysis focuses on houses for sale in Brandenburg. Notably, approximately 7% of listings are missed if data is scraped only on a daily basis. The total number of observations (n) is 1589. This representation covers only the first week (168 hours) of the observation period.

In summary, the findings indicate that the results vary significantly depending on the evaluation level (houses/apartments, rental/sale, Berlin/Brandenburg). Specifically, for the Berlin rental market, a relatively high number of listings are deleted early. In contrast, houses for sale in Brandenburg tend to

remain on the real estate web portal for a longer duration before being removed. This discrepancy highlights the importance of context-specific analysis in understanding the dynamics of the housing market across different regions and property types.

6.4.5. Overcoverage and Undercoverage

The objective of UC2 "Construction Activity" is for real estate web portals to provide information about construction activities with greater spatial accuracy and, in some cases, at earlier stages than is possible through official construction planning statistics. Based on this information, early estimates or forecasts regarding construction activity are to be derived.

A crucial component of the analysis involves assessing whether the construction activity statistics can be adequately covered. Previous analyses have already identified various aspects that may lead to over- or undercoverage of these statistics. These aspects will be briefly summarized in this section.

In summary, the following reasons for overcoverage of webscraped data compared to official statistics are listed:

- Duplicate Entries: Webscraping can result in the capture of duplicate data, for instance, when the same listing appears on multiple portals or in different locations. These duplicate entries lead to an overestimation of the actual values.
- Broad Definition of Categories: In webscraping, the definitions of categories such as "new constructions" on real estate web portals can be very broad. In some cases, properties are listed as new constructions even though they are not true new builds in the formal sense (e.g., renovated old buildings or properties that have only been partially renovated). This can lead to overcounting, as objects are classified into categories to which they technically do not belong.
- Unverified Data Sources: Webscraped data often comes from unregulated portals that lack strict quality control measures. This can lead to a higher number of entries that are neither realistic nor accurate (e.g., fake listings, outdated entries, etc.).
- Lack of Timeliness in Official Data: In official statistics, there often is a time lag between data collection and publication. Webscraping can provide more up-to-date information, resulting in apparent overcoverage, as new entries are already captured but have not yet been reflected in official statistics.
- In contrast, there is undercoverage. The following are reasons for the underrepresentation of webscraped data compared to official statistics:
- Incomplete Data Source Capture: Webscraping is limited to specific websites or portals. If important data sources are not available on the web or cannot be scraped, those data will be missing. This can lead to an underrepresentation of certain categories (such as new constructions).
- Missing Content within a Portal
    - Relevance of Content: Missing current information, such as recently completed new constructions or currently available properties, can diminish the relevance of the collected data, leading to insufficient coverage.
    - Content Gaps: If certain categories of data or relevant information are missing within a portal, analyses can become skewed or inaccurate. For example, a real estate web portal may not list all new constructions (or vice versa), resulting in an underrepresentation of construction activity.
    - Insufficient Data Diversity: A portal that captures only a portion of the market or specific types of properties may not provide the full range of relevant data. This limits the ability to make informed decisions or forecasts.

- Restrictions Due to Privacy and Access Rights: Some data sources are not scrapeable for legal or technical reasons, such as being protected by login requirements or CAPTCHAs. As a result, relevant information may be missing.
- Incorrect or Missing Metadata: Scraped data often lacks metadata necessary for accurate classification. If important information, such as location or classifications, is not captured, it can lead to distortion and undercoverage.
- Dynamic or Rapidly Obsolete Data: Websites frequently change, and information is quickly updated or removed. If scraping occurs at a specific point in time, relevant data may be missing at another time, resulting in incomplete data collection.
- Narrow Definitions: If the definitions of categories, such as "new constructions," are too strict or narrowly defined, many properties that should be included may be excluded. For instance, if only properties that are newly built from the ground up are considered, renovated buildings may be overlooked, leading to an incomplete representation of the market.

6.5. Conclusion, Discussion

The use of webscraped data compared to official statistics presents both opportunities and challenges. Overcoverage and undercoverage can arise from methodological differences, technical limitations, and variations in data sources. While official statistics typically exhibit higher methodological control, web scraping offers the flexibility to quickly and comprehensively gather data. However, this flexibility often comes with an increased susceptibility to biases.

A central issue in utilizing scraped data is data quality. Since these data often originate from unregulated sources, they can be inconsistent or inaccurate. Additionally, ambiguities in definitions and categorizations can lead to further distortions. For example, the definition of "new constructions" can be interpreted differently across various sources, thereby impairing comparability.

Moreover, there are technical and legal aspects that can restrict access to certain websites. Technical barriers such as CAPTCHAs or login requirements, as well as legal constraints, can complicate the collection of relevant data.

Despite these challenges, scraped data provide valuable insights into the real estate market. Whether they can help generate more accurate forecasts regarding construction activity and ensure improved spatial coverage remains to be seen. It is essential to consider the quality and limitations of scraped data and to critically incorporate them into the analysis. A well-founded combination of scraped data and official statistics can thus contribute to a more comprehensive and precise picture of construction activity.

Although it has been shown that real estate web portals are only partially suitable for forecasting within the framework of official construction activity statistics, the applied methodology remains of high value. Webscraping could be particularly valuable as a supplementary tool for validating price statistics. The experiences gained, as well as the increased awareness of the associated challenges, are also of great importance for future projects of a similar nature – especially regarding the differences between screen scraping and API scraping. The insights gathered will allow for better assessment of whether new project ideas are feasible and realistic in the future.

7. Country Specific Part: Statistics Sweden (SCB)

7.1. Data Sources

Early in the project, we encountered a limitation regarding the availability of suitable portals for scraping rental listings. We were unable to identify a set of portals for constructing a complete picture of the rental market and consequently shifted our focus towards properties for sale.

Two Swedish portals, Hemnet and Booli, offer nearly complete national coverage of new construction real estate advertisements. These portals provide detailed and up-to-date information on properties available for sale, including key variables such as listing dates, property types, number of rooms, and prices.

After contacting the owners of both portals, an informal agreement was established to allow the scraping of these websites. As a result, no formal assessment of portal availability was necessary.

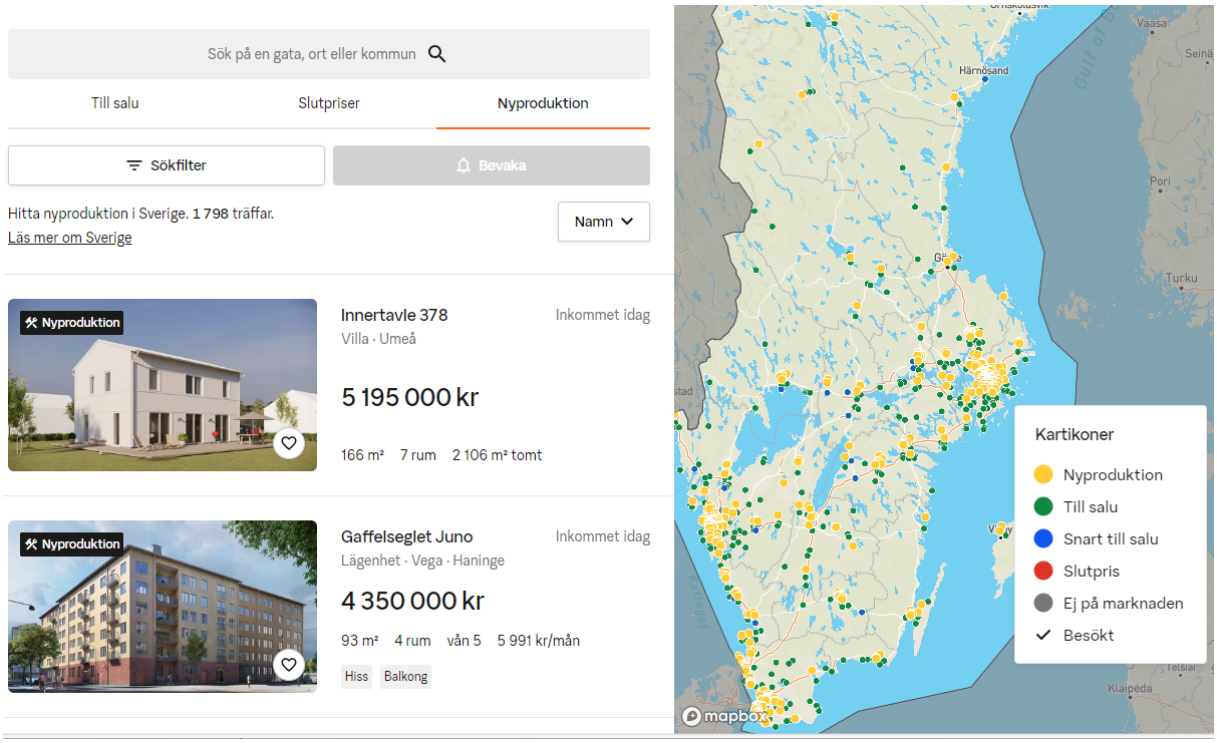
7.1.1. Choosing Portals to Scrape

Hemnet and Booli are two prominent Swedish portals that provide nearly complete national coverage of new construction real estate advertisements. Hemnet primarily focuses on real estate listings from various agencies and has comprehensive coverage of properties for sale across Sweden. Booli, on the other hand, describes itself as "Sweden's largest combined selection of homes, regardless of whether they are for sale, soon to be sold or completely newly produced" and covers a greater number of properties. Booli gathers its information from real estate agencies' websites and systems, as well as from real estate developers.

While the exact coverage has not been formally verified, samples from both sites indicates that Booli has greater coverage and is believed to cover almost the entire Swedish real estate market. Consequently, a decision was made to focus the efforts made within the project on the Booli portal.

On the Booli "Nyproduktion" (New Constructions) page, users can search for properties by city or zip code, with the option to further refine their search using various filters such as type of dwelling (e.g. apartment, house), number of rooms, living area, floor and other characteristics. Each search result page displays up to 40 listings, each including a thumbnail image and some key details about the listing.

Figure 29: Screenshot of the Booli new constructions page



Each property listing page contains further key details such as location, price development and information on the date the property was listed, upcoming viewing dates, and any special notes (e.g., “Inflyttningsklart” meaning “Ready to move in”).

While the overall HTML structure of each advertisement is consistent across listings, CSS class names are not informative and variables are often not easy to locate using CSS selectors. The description of each property is presented as plain text in HTML, and contain an irregular set of information.

Data is collected via web scraping from three different datasets on a portal: objects for sale, new production projects, and sold objects. The initial web scraping began in 2022 but was halted due to technical issues. It briefly resumed with a new approach in January 2023 and then again later in September 2023. The data is gathered on an irregular basis, but at least once a month. The scrapers are written in R and Python.

There are 43 different characteristics available for the “sold” dataset and 32 characteristics for the “for sale” dataset. Most advertisements contain all the characteristics and the data covers objects in Sweden.

Although there is no formal agreement, the scraping activities have been discussed with the portal provider and they have been informed about the operations. Newly constructed buildings are defined by their construction year, the condition of “first listing,” or the address being classified as a new production project.

The collected data consists of three different datasets representing different states of real estate advertising:

- Active new construction projects
  - This dataset contains aggregated information about active new construction projects and their available objects collected from real estate developers. Limited information on individual linked objects is also available and could be scraped, but as of the time of writing this data is not collected.
    These advertisements can be collected very early in the construction process (upwards of five years ahead of availability) but contain only limited aggregated information about available objects and suffers from rather poor accuracy in some key variables (e.g. date of residence availability).
- New construction objects
  This dataset contains all currently listed individual new construction ads and is comparable with the dataset collected using the old method.
  These can also be collected early and generally have better accuracy and completeness.
- Sold objects
  The data is scraped monthly to collect newly registered sold objects during this time interval. Older data could be collected, but our initial observations indicate that the coverage and completeness of the data diminishes beyond this timepoint.
  Sold objects contain the most complete data but can naturally only be collected once the purchase of an object has been reported. It also lacks some crucial characteristics which makes it challenging to use for certain purposes (e.g. there is no information regarding how long before the date of purchase it had been available).

There are multiple possible advertisement flows for a new construction object. Most are initialized either as part of an aggregated project listing created by the developer or an individual object listing created by a realtor. Some pass through both states before finally being collected as sold objects once the ownership change has been reported. Not enough data has been collected to properly analyse the coverage overlap between the datasets, but we assume that their intersections are non-empty sets when collected over time. At any given timepoint however, the datasets are assumed to be mutually exclusive.

Due to constraints in the in-house technical portal, the periodic scraping of the “currently available” datasets (“for sale” and “active projects”) could not be automated and was consequently discontinued.

The “sold data” dataset was initially collected covering a period between 2002 and 2023, with subsequent incremental scraping to handle backfills and new purchases. The “for sale” and “active projects” datasets cover a five month long period from September 2023 to January 2024. All three datasets were useful as parts of the analysis, but the main focus during the project was on the data describing the “sold objects” dataset and its observed transfer of ownership.

Table 13 gives an overview of the data collected from the “sold” dataset.

Table 13: Overview of sold objects classified as new construction by year

Year sold	Object type	Unique ads	Unique ads classified as new construction
2019	Apartment	86024	5444
2019	Other object type	110787	1706
2020	Apartment	97264	7198
2020	Other object type	118516	2189
2021	Apartment	107602	9737
2021	Other object type	128852	2735
2022	Apartment	90800	8781
2022	Other object type	107423	2248

7.2. Data Preparation

The advertisements in the "sold" dataset each have an identifier and a supplementary identifier for the object in question. Since only one portal is used and, and the fact that the advertisements are not collected periodically to reflect the current market availability but rather the observed transfer of ownership of newly constructed real estate, no intricate de-duplication process is necessary.

In order to identify observations that correspond to our definition of "new construction", the data is processed according to the following heuristics:

1. A construction year should indicate that the advertisement indeed relates to a newly constructed building. About 83% of all advertisements have valid construction years, but the quality of the variable varies. While the construction year variable cannot sufficiently be used to represent a precise definition of "first occupancy availability", it can be used to roughly indicate that an observer transfer of ownership happened relatively close to the time of finished construction.
   Condition 1: an object should be sold within +- 2 years of its construction year in order to be considered a new construction advertisement.
2. Objects that otherwise indicate that they belong to recently constructed buildings should be considered. There are no explicit indicator variables or labels for this characteristic, but by linking the advertisements to the set of collected active new construction project by geographic location, a number of additional candidates were identified.

3. To avoid overcoverage and overcount, advertisements should refer to objects that are available for the first time. The portal tracks the advertisement history for each object, making it simple to identify observations that represent the first recorded sale of each object.

To summarise: an advertisement is classified as a new construction if it is the first recorded instance of an object in the portal and the construction year is close to the date of sale. This is a rough set of conditions, but due to the high degree of coverage, the first recoded observation of an object can be considered a significant enough indicator to warrant such a pragmatic approach at this stage.

The data returned from the GraphQL API request results is structured, clean and easy to parse. It allows us to specify the data needed, reducing the need for cleaning and editing outside of standardizing characteristics according to the definitions used in use-case 1.

Some important steps of the editing and cleaning process include:

- Renaming and standardizing column names, e.g. removing unnecessary pre- and postfixes.
- Converting missing values to `None`.
- Converting the data type of some variables from strings to integers and floats.
- Extracting area variables from a jagged array into separate columns.

Table 14 lists all the location variables in the "sold" dataset along with descriptions and their respective data types and complete rates.

Table 14: Area variables in the "sold" dataset

Data type	Variable	Description	Complete rate
numeric	addressId	A unique identifier assigned to each address	1.000
character	county	Swedish NUTS3 region name	1.000
character	descriptiveAreaName	Neighborhood	1.000
character	indexArea	Custom Swedish NUTS2 cluster name	1.000
numeric	latitude	Latitude coordinate	1.000
numeric	longitude	Longitude coordinate	1.000
character	municipality	Swedish municipality name	1.000
numeric	postcode	Five digit postcode	1.000
character	primaryArea	Main area or region	1.000
character	country	Country	0.999
character	electricityBiddingZone	Electricity market region	0.999
character	streetAddress	Street name and number	0.983
character	streetName	Street name	0.946
character	userDefined	User defined location description	0.854
character	populatedArea	Associated population centre (city)	0.629
character	locality	Distinct district within a municipality	0.306
character	suburb	Residential area	0.245
character	deso	Statistical area used for demographic analysis	0.031
character	desoCluster	Group of DeSo areas	0.031

The most complete variables are clean and distinct enough to offer great possibilities for geolocation and regional aggregation. Some of the variables with lower complete rates, such as user-defined, locality and suburb tend to be more arbitrary and less consistently recorded and often contain redundant information.

7.3. Results

7.3.1. Monthly Number of Ads

Figure 30: Number of collected sold advertisements in 2021 and 2022 by object type (apartments and other)

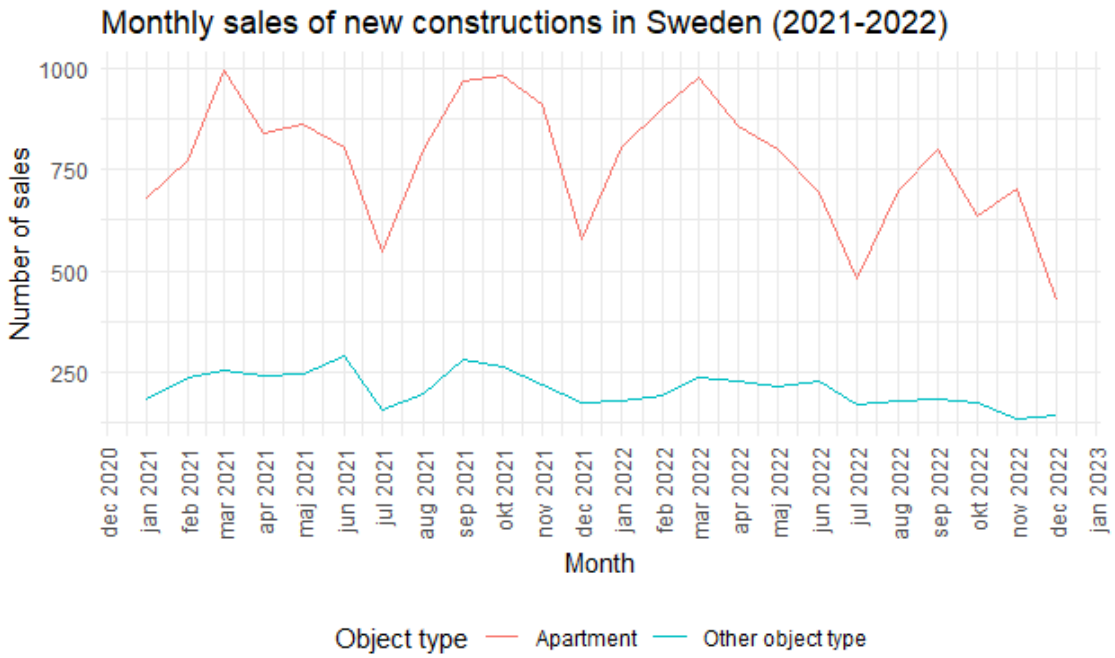


Figure 30 shows the number of new construction real estate sales for 2021 and 2022 by month and object type. The number of sales fluctuates from month to month, but here appears to be a seasonal pattern with sales dipping in the summer months and around the end of the year. Table 15 shows a further differentiation of advertisements for 2021 and 2022 by object type, number of ads, sizes (square metres and rooms), and price.

Table 15: Summary of advertisements in 2021 and 2022, by type of building

Year	Object type	Number of ads	Size (m2, median)	Size (rooms, median)	Price (SEK, median)
2021	Apartment	9737	60.0	2	2750000
2021	House	873	135.0	5	4610000
2021	Other	1172	114.0	5	3600000
2022	Apartment	8781	59.5	2	2750000
2022	House	846	135.0	5	5100000
2022	Other	1080	116.0	5	3797500

Figures 31 and 32 show the sold prices for apartments and houses by NUTS level (NUTS3, “Län”) in Sweden.

Figure 31: Boxplot of prices (sale) for apartments by NUTS3 regions in Sweden

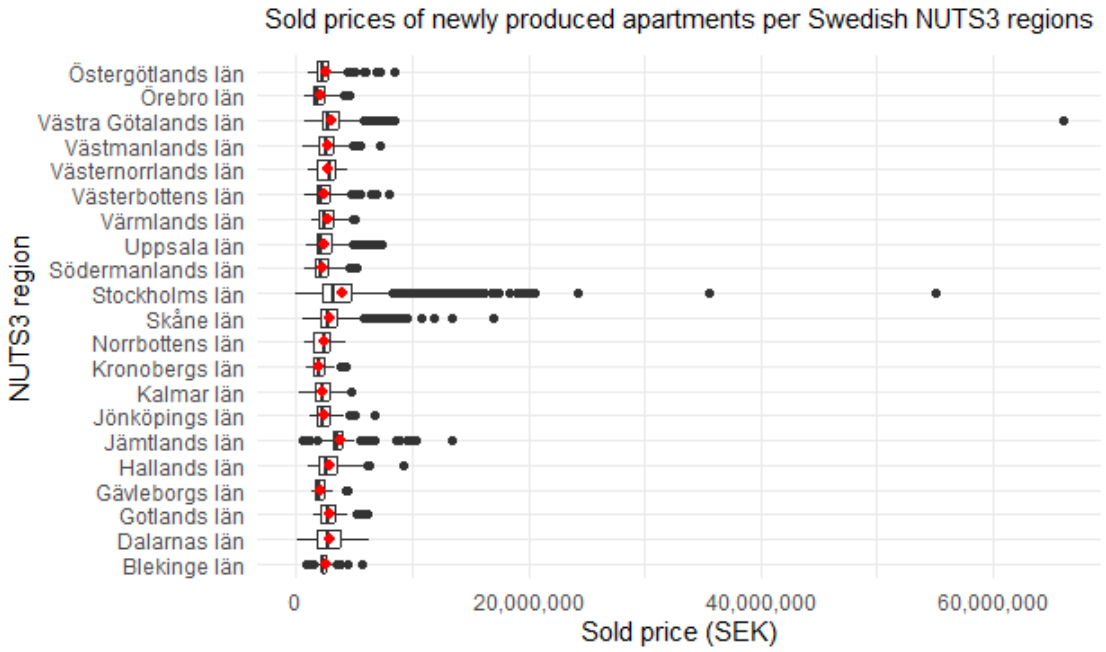
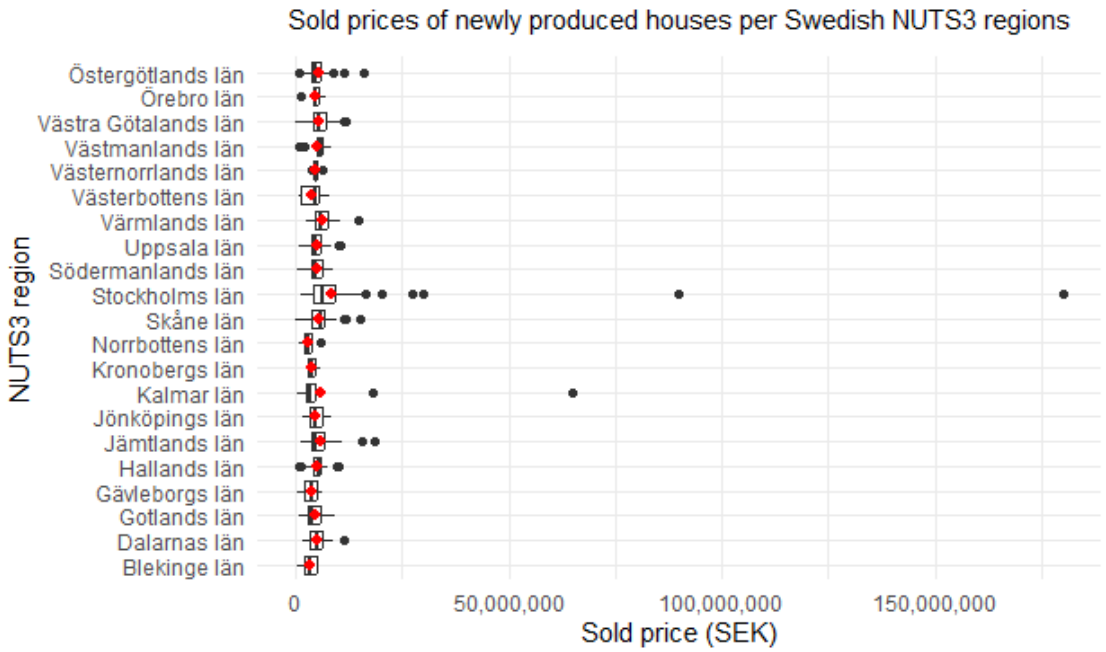


Figure 32: Boxplot of prices (sale) for houses by NUTS3 regions in Sweden



Figures 33 and 34 show the living areas for houses and apartments by NUTS level (NUTS3, “Län”) in Sweden.

Figure 33: Boxplot of living areas (m2) for apartments by NUTS3 regions in Sweden

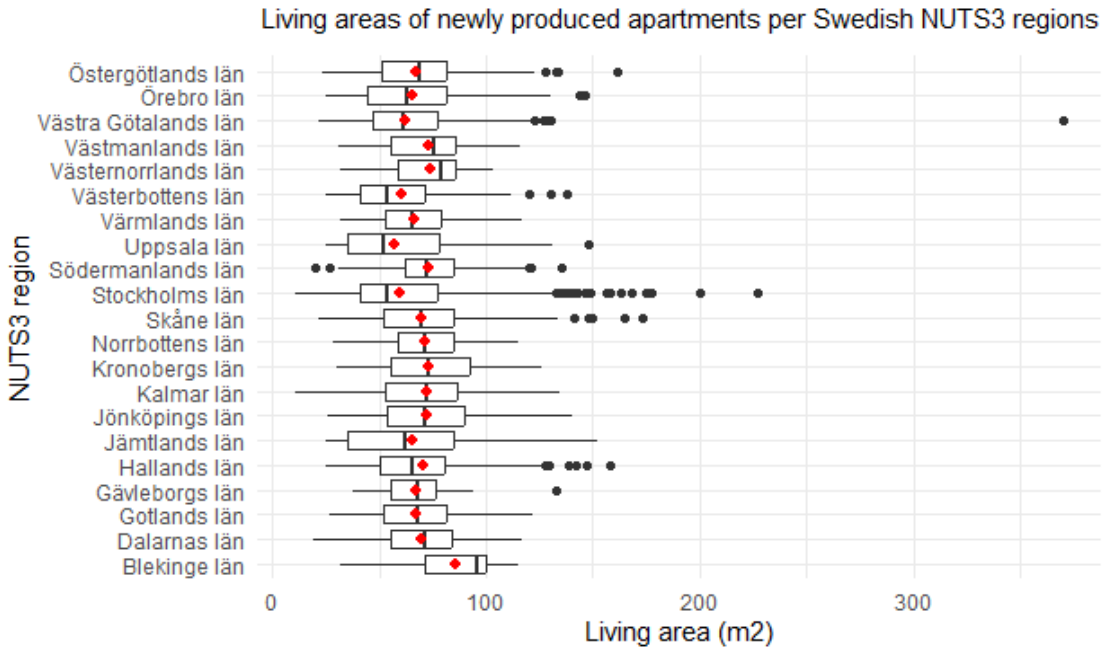


Figure 34: Boxplot of living areas (m2) for houses by NUTS3 regions in Sweden

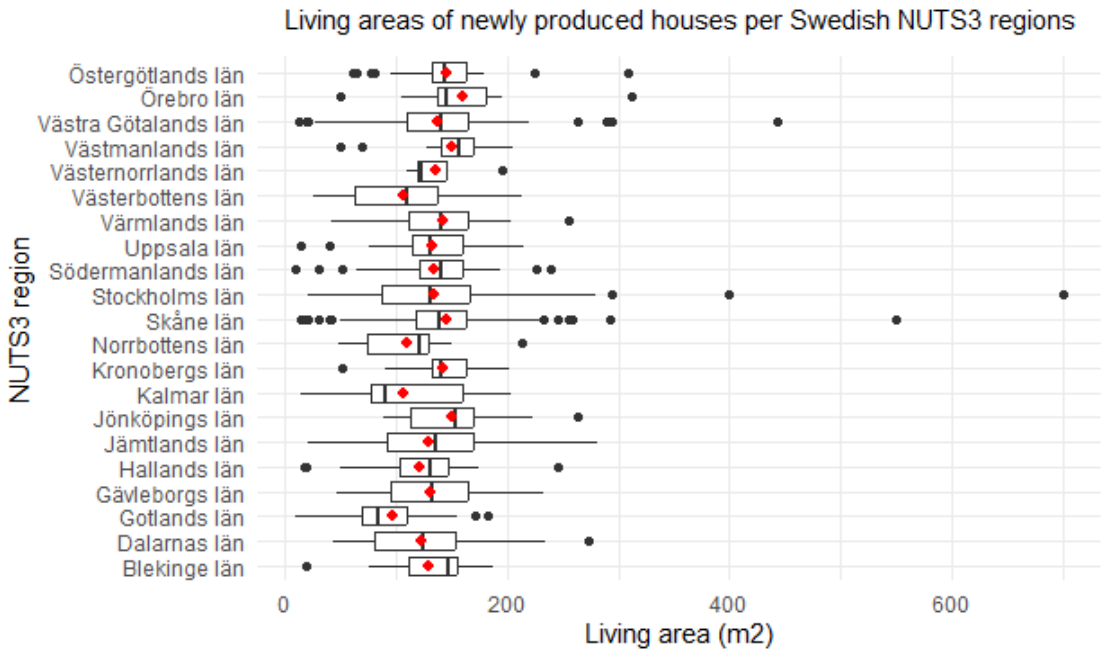


Figure 35 presents the mean number of sold objects by month and NUTS3 area for the years 2021 and 2022, with most sold objects being located in the Stockholms Län.

Figure 35: Mean number of sold objects per month for the years 2021 and 2022

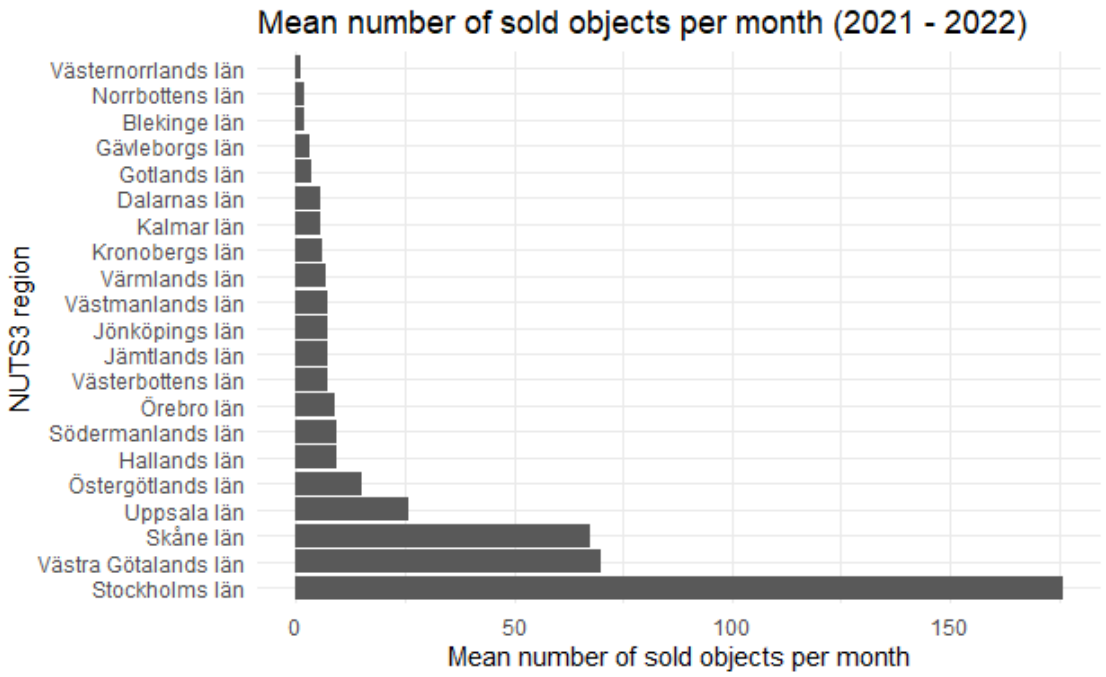


Figure 36 and figure 37 compare the number of sold objects by year with official statistics (Completed dwellings in newly constructed buildings by type of building, tenure and size of dwelling. Year 1991 – 2023)[9]. It can be seen that the scraped data consistently shows lower numbers compared to the official figures, confirming the incomplete coverage, especially concerning larger houses.

The discrepancy in coverage between larger houses (5+ rooms) could be attributed to some yet to be identified market dynamics, but the main reason is likely a high incidence of self-built homes. Newly produced self-built houses do not enter the market, as they are constructed by individuals or families for their own use. Consequently, they bypass the typical sales process and do not generate listings on real estate web portals.

In contrast, apartments are usually developed by commercial builders and are more likely to be sold through real estate agents and listed online, leading to better coverage in scraped data.

[9] Completed dwellings in newly constructed buildings by type of building, tenure and size of dwelling. Year 1991 - 2023. PxWeb (scb.se)

Figure 36: Number of newly constructed apartments by number of rooms and data source

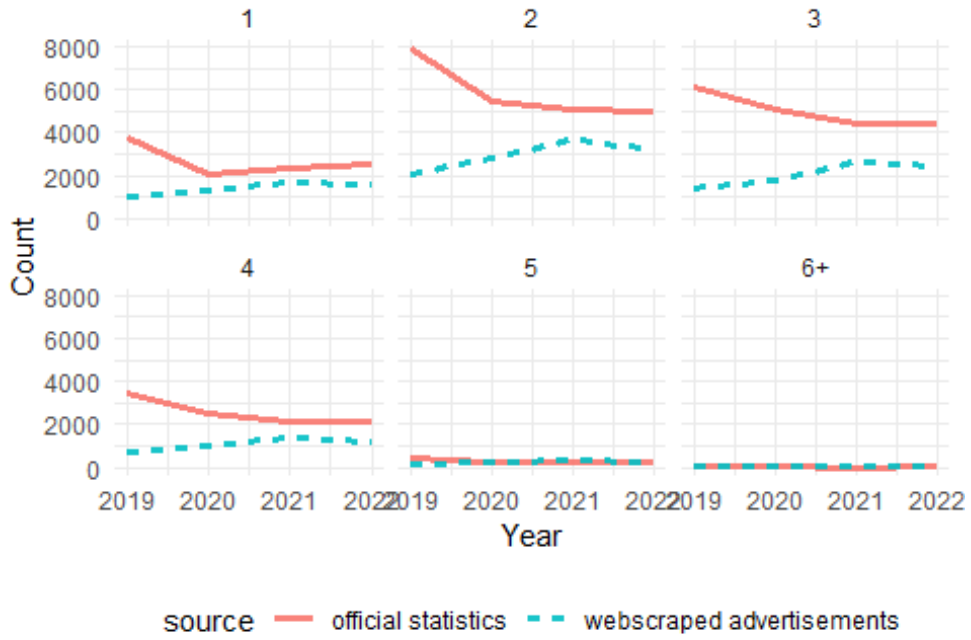


Figure 37: Number of newly constructed houses by number of rooms and data source

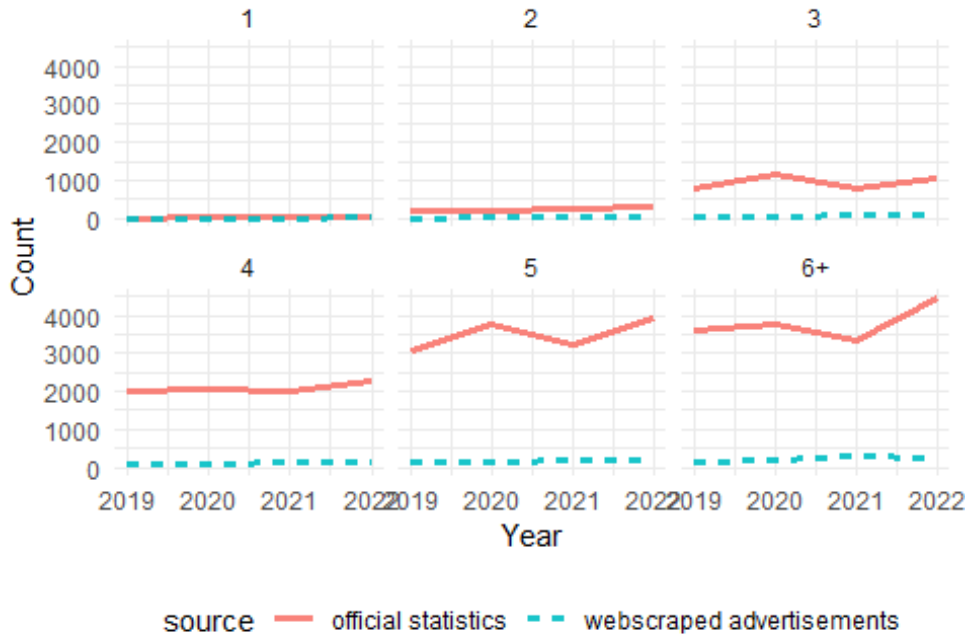


Figure 38 and figure 39 compare the proportion of sold objects by number of rooms and year with official statistics (Completed dwellings in newly constructed buildings by type of building, tenure and size of dwelling. Year 1991 – 2023)[10]. One can see that despite the incomplete coverage, the market representation of the scraped data appears to be quite comparable with the official statistics.

[10] Completed dwellings in newly constructed buildings by type of building, tenure and size of dwelling. Year 1991 - 2023. PxWeb (scb.se)

Figure 38: Proportion of newly produced apartments by number of rooms and data source

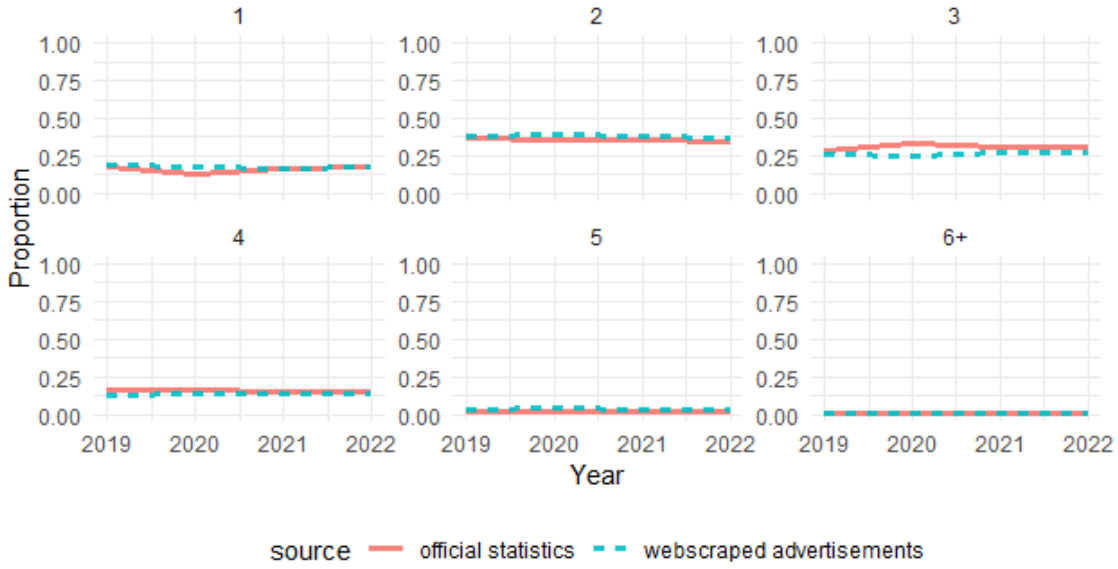
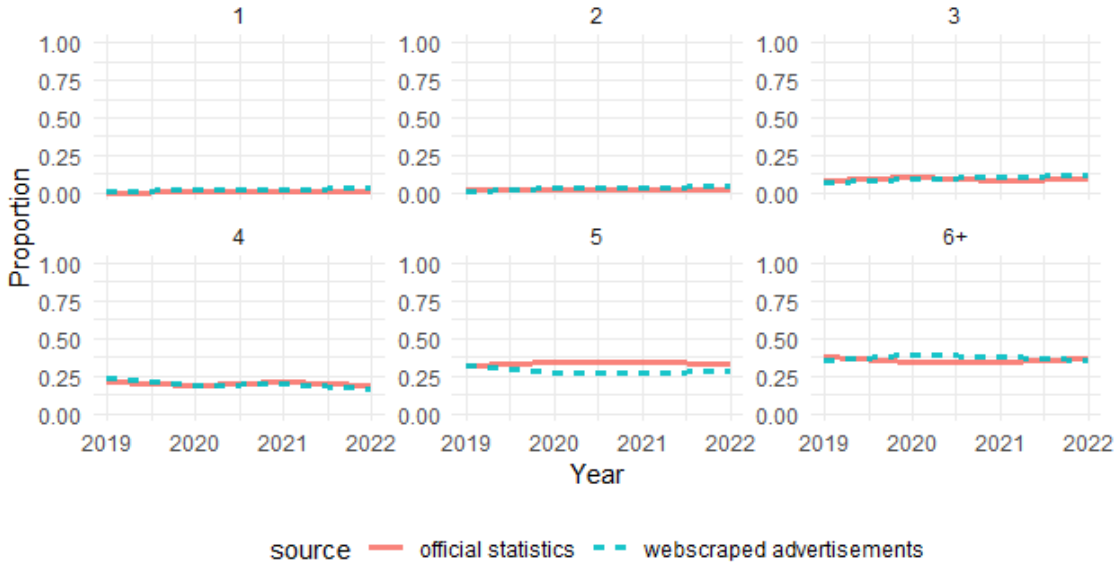


Figure 39: Proportion of newly produced houses by number of rooms and data source



7.3.2. Early Indicator for 'Construction Activities'

The dataset on new construction projects provides forward-looking information about the supply side of the housing market. This dataset includes macro data and summaries of projects that will be available for occupancy within up to five years and could be used for identifying trends in the initiation and expected completion of construction projects. Metrics such as the number of new projects started, the scale of these projects (e.g. number of units), and their expected completion dates could serve as indicators of future housing supply.

7.4. Conclusions and Discussion

The data sourced from real estate web portals has proven to be potentially relevant and useful for our office. Although we have not yet implemented any findings from the project into our statistics production, the insights gained have highlighted the potential of this data source. The economic statistics department has shown continued interest in exploring future applications of this data, recognizing its potential value in providing early estimates and as an ancillary data source for validation.

The methods of data collection employed in the project have demonstrated their utility, not only for construction activities but potentially for other areas as well. Solutions for web scraping as a data collection method is being developed and integrated into the Official Statistics production process.

Experiences from the project have led us to realize the need for a more mature portal and more sophisticated ways of working to explore innovative data sources, technologies, and methods. These experiences have and will continue to help us to identify requirements for being able to successfully leverage web data for early estimates and real-time observations.

8. Conclusion

In summary, webscraping advertisements from real estate web portals has been an effective method for data collection. However, it does come with pitfalls that need to be addressed and considered when working with this kind of data.

One, issues regarding data collection can arise, as scrapers require maintenance and quick action when website structures change in order to prevent or minimize gaps in the dataset.

Two, some important quality issues need to be considered.

- **Technical Undercoverage due to Scraping:** Real estate is currently in high demand, particularly in urban regions, where advertisements on well-known real estate portals can attract a large number of applicants within a short period of time. As a result, the market is highly volatile, and many listings are available for only a brief duration, often just a few hours. This has a direct impact on the intervals at which scraping needs to be conducted to avoid data gaps and undercoverage caused by technical limitations. A careful balance must be struck between achieving comprehensive coverage of the listings and minimizing disruptive traffic for both the agencies conducting the scraping and the portal operators. The severity of this issue likely depends on the specific goals of the scraping project, but it should be thoroughly examined before transitioning the scraping process into a production phase. An alternative to implementing very short scraping intervals could involve establishing contact with portal operators to develop a mutually acceptable solution—possibly within the framework of a formal agreement. For instance, one option could be to retain listings that are no longer available on the portal for 24 hours, marking them with a “No longer available” notice or making them accessible via an API.
- **Definition of Variables:** As the aim of this use-case was to identify newly constructed buildings, the attribute “new construction” needed to be defined as precisely as possible to avoid including non-new constructions in the analysis or overlooking actual new constructions in the available data. Most portals are not exclusively focused on new constructions but also list existing properties. Even when portals offer clear attributes for identifying new constructions, these are usually filled in by the property advertisers and may not always be accurate or aligned with the intended definition. For instance, renovated apartments in first-time occupancy after renovation may be incorrectly marked as “new constructions.” While human users can often easily identify new constructions based on (for example) photos or free-text descriptions, this level of certainty is not achievable for simple automated programs. This issue is likely to vary depending on the specific use-case of web scraping. As outlined earlier, there are some pragmatic approaches to handling this challenge. However, for the specific case of identifying new constructions, a more complex but potentially effective technical alternative could involve employing machine learning methods.
- **Deduplication Issues:** In many cases, data is scraped from multiple portals, resulting in some properties appearing in the datasets of several portals. This creates the challenge of identifying such duplicates and excluding them to avoid distortion in the results (see, e.g., 6.4.3 above). This task is inherently complex and is further complicated by the fact that not all portals and listings provide complete information about the property’s location (such as geo-coordinates or exact addresses). Moreover, even when such data is available, it does not always suffice to reliably identify duplicates. A pragmatic approach, similar to the one described under “Definition of Variables,” is even less likely to yield robust results in this context. In a side project, the use of machine learning methods based on manually classified

test data was explored to address this issue (Meyberg, C., Rendtel, U., & Leerhoff, H. (2024). *Flat rent price prediction in Berlin with web scraping.* AStA Wirtschafts- und Sozialstatistisches Archiv. https://doi.org/10.1007/s11943-024-00340-6). The findings indicate that machine learning can be a viable and effective solution if sufficiently high-quality data is available. However, it raises the legitimate question of whether the benefits justify the effort required by official statistics to implement such methods. A similar issue arises when webscraped data at the object level needs to be merged with data from official statistics.

Three, issues of genuine over- and undercoverage need to be addressed. When comparing the scraped data to official statistics, a clear urban-rural divide emerges, with urban areas significantly more represented, suggesting that rural regions may be underreported. For other regions, a clear overcoverage is visible.

Four, the transferability of tools is not clear. While Data from real estate web portals and tools for data collection could be used in other countries as well, neither sources themselves nor tools used during this project are assumed to be easily adaptable. In principle, it can be expected that there are many country specific details of legal, organizational, and technical matters to take into consideration.

Offices interested in investigating this data source should consider these questions:

- What is the target population of the study? Is the survey population (advertised objects) different from the target population (newly constructed apartments and houses)?
- What data / which variables are actually needed from the advertisements? Which information would be necessary in order to identify duplicates within one data source as well as between different data sources? Is this information even available?
- What kind of data portals do exist? What is their size / coverage? What are the largest portals? Are there smaller but specialized portals to consider?
- What kind of access to portal data is possible? (Screen scraping vs API vs. agreement)
- Are there differences of concepts and definitions between official statistics and web data portals? Are concepts and definitions consistent between portals?
- Are there legal barriers in gathering this data through webscraping methods? Is it preferable to have an agreement with portal providers? Is it even possible to have an agreement with portal providers?
- What kind of technical barriers are there regarding collecting data from these portals using webscraping methods? Can these barriers be overcome? Is it desired to overcome these barriers?
- Is it possible (and legally allowed) to combine microdata on houses and advertisements from advertisements and official data (i.e. using full address information) in order to identify duplicates, identify undercoverage, to build and test a model and to build training data, e.g. for imputation purposes?