

Work Package 3

New use-cases

Deliverable 3.4: Report on the results of the new data sources exploration and the conditions for using the data

Characteristics of the real estate market

Version, 2025-02-24

Prepared by:

UC1 coordinator: Dominik Dąbrowski (GUS, Poland)

Contributors:

Galya Stateva (BNSI, Bulgaria)
Kostadin Georgiev (BNSI, Bulgaria)
Klaudia Peszat (GUS, Poland)
Bartosz Grancow (GUS, Poland)
Jacek Kotowski (GUS, Poland)
Marta Kruczek-Szepel (GUS, Poland)
Krystyna Piątkowska (GUS, Poland)
Gitta Lasslop (HSL, Germany)
Tobias Gramlich (HSL, Germany)
Alexandra Ils (HSL, Germany)
Nicole Jurisch (SSI-BBB, Germany)
Holger Leerhoff (SSI-BBB, Germany)
Andreas May-Wachowius (SSI-BBB, Germany)
Kerstin Erfurth (SSI-BBB, Germany)
Elina Peltoniemi (SF, Finland)
Katja Löytynoja (SF, Finland)
Elina Vuorio (SF, Finland)
Sini Liukkonen (SF, Finland)
Ville Auno (SF, Finland)
Pierre Lamarche (INSEE, France)
Aurelie Goin (INSEE, France)



This document was funded by the European Union.

The content of this deliverable represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.



Web Intelligence
Network



**Funded by
the European Union**

CONTENTS

| | |
|---|-----------|
| Project partner organizations..... | 5 |
| 1. Background | 6 |
| 2. New data sources exploration | 7 |
| 2.1. Bulgaria..... | 11 |
| 2.2. Poland..... | 12 |
| 2.3. Germany – HSL | 13 |
| 2.4. Germany – SSI-BBB..... | 16 |
| 2.5. Finland | 18 |
| 2.6. France | 19 |
| 3. Programming, production of software | 20 |
| 3.1. Bulgaria..... | 21 |
| 3.2. Poland..... | 25 |
| 3.3. Germany – HSL | 27 |
| 3.4. Germany – SSI-BBB..... | 27 |
| 4. Data acquisition and recording | 29 |
| 4.1. Bulgaria..... | 29 |
| 4.2. Poland..... | 30 |
| 4.3. Germany – HSL | 31 |
| 4.4. Germany – SSI-BBB..... | 32 |
| 4.5. Finland | 35 |
| 4.6. France | 35 |
| 5. Data processing | 36 |
| 5.1. Bulgaria..... | 37 |
| 5.2. Poland..... | 39 |
| 5.3. Germany – HSL | 43 |
| 5.4. Germany – SSI-BBB..... | 45 |
| 5.5. Finland | 51 |
| 5.6. France | 53 |
| 6. Modelling and interpretation..... | 55 |
| 6.1. Bulgaria..... | 55 |
| 6.2. Poland..... | 56 |
| 6.3. Germany – HSL | 60 |
| 6.4. Germany – SSI-BBB..... | 63 |
| 6.5. Finland | 64 |
| 6.6. France | 70 |



| | | |
|------------|---|-----------|
| 7. | Dissemination of the experimental statistics and results | 71 |
| 7.1. | Bulgaria..... | 73 |
| 7.2. | Poland..... | 74 |
| 7.3. | Germany – HSL | 75 |
| 7.4. | Germany – SSI-BBB..... | 76 |
| 7.5. | Finland | 77 |
| 7.6. | France | 78 |
| 8. | Conclusions | 79 |
| 9. | Annexes | 83 |
| 10. | List of tables | 83 |
| 11. | List of figures | 84 |

Project partner organizations

| No. | Name | Short name | Country |
|-----|---|------------|---------------|
| 1 | Główny Urząd Statystyczny | GUS | Poland (PL) |
| 2 | National Statistical Institute | BNSI | Bulgaria (BG) |
| 3 | Tilastokeskus | SF | Finland (FI) |
| 4 | Institut National De La Statistique Et Des Etudes Economiques | INSEE | France (FR) |
| 5 | Amt Fur Statistik Berlin-Brandenburg | SSI-BBB | Germany (DE) |
| 6 | Hessisches Statistisches Landesamt | HSL | Germany (DE) |



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 (TSS-WIN). The overall objective of WP3 is to explore the potential of new types of web data sources for official statistics. The work is organised in a number of use cases (UCs), each focused on a specific application. The 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 focuses on UC1, the use of web data sources in real estate statistics. The aim was to examine the possibility to produce experimental statistics on sales and rental market on the lowest possible territorial level with monthly frequency, using the same methodology in multiple countries. The project focused on elements, such as the surface area, number of rooms and the price of the property, supplemented with additional qualitative information, such as e.g. parking space, security and other existing amenities, if present in the available web data. The resulting data can be used in the calculation of the Harmonised Index of Consumer Prices (HICP) or to complete real estate market studies with new indicators.

Project participants used different approaches to collect publicly available data, including various techniques for data acquisition. The intention of the project was to cover the whole process from conducting research, building up the necessary infrastructure to processing the data into meaningful indicators. In addition, issues and lessons learned were written down for other entities that would like to use the results of this project to produce experimental statistics on the real estate market.

The use-case involved 6 national and regional statistical organizations from 5 countries: Bulgaria, Finland, France, Germany (SSI-BBB, HSL) and Poland. The organisations obtained data by web scraping or directly from the web portal owners based on the bilateral agreements. Despite different data sources, all partners followed the same steps of the BREAL model:

1. New data sources exploration
2. Programming, production of software
3. Data acquisition and recording
4. Data processing
5. Modelling and interpretation
6. Dissemination of the experimental statistics and results.

The process steps listed above are the phases, expressed in terms of GSBPM 5.1 phases recognized as part of the big data lifecycle. These steps are part of the business processes and functions layer derived from the BREAL model¹. Although the presentation of the steps follows the logical sequence, each step may have occurred in a different order and in different circumstances, depending on a country's own purposes.

¹ Deliverable F1, ESSnet on BD II project, WPF.

2. New data sources exploration

The new data sources exploration step includes:

- definition and integration of criteria for assessment of new data sources;
- individual assessment of web data sources by partner;
- creation of the inventory list of new data sources assessed and selected to be the target of the web-scraping for the data acquisition and recording phase;
- definition of a minimal set of possible indicators.

At the beginning of the project, it was considered to obtain information on the market share of a given portal. However, the results of such work would have limited value on large portals, which present not only real estate but also vehicles, electronics, fashion, etc. In case of smaller portals, focused on real estate only, such information can be a useful source of knowledge, but the lack of data about the share of these largest companies means that calculating the market share is impossible. An alternative could be to use meta-information regarding the popularity of a given website from Google Trends. This may also be helpful in assessing the importance of chosen portal over years.

However, examining the representativity of the source should not be done without monitoring its general stability. That would require longer observation, to be started before the data acquisition process begins. This should be done at least by analyzing the responses to queries sent to the website in terms of whether there are changes in the structure of the website or lack of responses over the course of the day, month and seasons.

The trustworthiness of a source may also be derived from previous cooperation or previous acquisition of information based on an agreement. Portals with which previous cooperation has lasted a long time are more trustworthy. For instance, France and Finland used data from large portals operating in their countries, with which they have been cooperating for some years.

Therefore, during the project, the WP3 partners developed the *Checklist for assessing web data sources*. The *Checklist* is divided into three parts: **common criteria, mandatory and optional variables**. Some criteria groups were mandatory to pass, while some were increasing the final score of assessed website. The exhaustive list includes 30 criteria, covering technical aspects of the website, the way information is presented and the number of advertisements that can be collected (Table 1). The *Checklist* makes it possible to assess any website regarding its responsiveness and content and can be used to evaluate usefulness.

Table 1. Common criteria list for assessing utility of web data sources

| Criteria | Description |
|--------------------|---|
| Captcha | Whether a web source uses captcha or not |
| Robots blocking | Whether a web source blocks robots or not |
| JavaScript | Whether a web source uses JavaScript or not |
| List of pages | The web source has a list of pages with pagination |
| Filter criteria | If a web source offers a content filtering functionality relevant to the use case |
| GET HTTP method | If a web source uses GET method for HTTP requests |
| Up to date content | Whether the content of web source has new user content published last month |

| | |
|---|--|
| Number of ads > X | Whether the number of ads on the web source is bigger than X e.g. 100, 1000, 10000, etc. |
| Structured description | Whether the web source has structured presentation of the content or just a plain text |
| HTML Microdata | Whether the web source uses HTML Microdata https://www.w3.org/TR/microdata/ |
| Description schema | Whether the web source uses description schema https://schema.org/IndividualProduct |
| HTML code changed every X | Frequency of HTML code change e.g. 1 year, 2 years, 3 years etc. |
| Specific time period filter | Whether a web source allows scraping the content published during specific time period selected via the content filter |
| Scraping of yesterday publications | Whether a web source allows scraping the content published only yesterday |
| Multilanguage | Whether a web source have option to change language and currency |
| Ratings | Whether a web source have option to rate an offer and leave a comment |
| Cookies and tracking | Whether a web source force to accept cookies and tracking information |
| Aggregator | Whether a web source display information gathered from many portals |
| Dynamic class tags | Whether a web source code is generated automatically |
| Terms of use | Whether a web source terms of use allow web scraping |
| robots.txt | Whether a web source lists relevant pages as disallowed in robots.txt |
| Offers API | Whether the website offers an API |
| CDN | Blocking by content delivery network services (like Cloudflare) |
| File extension | File extensions that do not contain renderable websites (e.g. .xlsx, .docx) |
| HTTP error | A URL returning a temporary (HTTP) error |
| Sale URL | A URL that is 'for sale' |
| Scope of the data | Whether the data are representative for the entire territory |
| Frequency of the data delivery/transmission | Whether the data are provided at least every month |
| Representativity of the data | Whether the data stands for a significant part of the entire rental offers (could be 20 % or 30%) |
| Data description – metadata | Whether metadata/a minimal set of data description is delivered along with the data |
| Data completion | Whether the rate of non-response/non completion does not exceed a given threshold for given variables that are considered as critical – should be declined for rents/surface/type of dwellings |

Due to high differentiation of websites and their structure or level of security, organisations could define the set of criteria to assess the web data sources individually. The minimal criteria chosen by all participants were: „GET HTTP method” and „number of advertisements“. Additionally, some participants also choose „structured description” (each offer presented in structured form), as well as „the list of pages with pagination” (the site does not present the data by uploading more content while navigating through the site).

Two of the participating countries (France and Finland) entirely obtained the data directly from the data providers based on bilateral agreements. Thus, the lists of criteria for assessment of web data sources differ in those cases.

The set of mandatory variables was agreed by all partners. It included the most important data on real estate that were collected, processed and analyzed by all partners. Optional variables from the common pool of variables were chosen individually.

Mandatory variables for this use case, the same for every country participant, were:

- ad_provider - the name of the website or data provider
- ad_id - unique ID code of advertisement/offer on the scraped website
- building_type - a type of real estate (apartment building, house, row house, access from outside, duplex, detached house, summer house, wooden house share, other)
- location - location obtained (may include multiple information on different territorial levels)
- NUTS3 – territorial division on NUTS3 level
- LAU - local administrative unit (in some cases equals city)
- offer_transaction - whether the offer is for rent or sale
- offer_price - price from the offer
- offer_surface - the surface area of the offered object
- offer_floor - floor number on which the object is situated
- offer_rooms - number of rooms in the offered object.

The summary results of the individual assessment of web data sources is presented in the country-related sub-chapters. More detailed assessment results are available in annex 1 and 2.

For the project's purposes it was also necessary to prepare a set of definitions common for all partners. They were developed on the basis of the official *ESS Quality Glossary*² and *Methodological guidelines of EU-SILC*³. The definitions covered all the needed information about the dataset and the main terms related to real estate offers.

Table 2. General concept and definitions for real estate market data

| Concept | Definition |
|--------------------------|---|
| Unit non-response | <p>Means that a sampled unit that is contacted may fail to respond (Total non-response).</p> <p><i>Note: In case of this project this refers to a situation when it is known that an offer exists on the website, but the scraping program is not able to achieve a success response for any reasons (code=4**).</i></p> <p><i>For data that is acquired on the base of an agreement there are no unit non-responses.</i></p> |

² <https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/ess-quality-glossary>

³ [Methodological guidelines 2021 operation v4 09.12.2020.pdf \(europa.eu\)](#)

| Concept | Definition |
|--------------------------------------|---|
| Item non-response | Means that the unit may respond to the questionnaire incompletely (Partial non-response). <i>Note: In case of this project this refers to a situation when the offer is obtained from the website, but in the set of mandatory variables not every variable is filled (a scraping program may find a tag on the page with no value assigned). It refers only to the mandatory variables as the optional may not be filled.</i> |
| Missing value | A variable has no value or when the variable is filled incorrectly (e.g. by putting a value that is out of the data type template). |
| Over-coverage error | Frame over-coverage is the case if it entails duplicated, non-existent or out-of-scope elements. When units are included, which from a conceptual viewpoint should be not included. |
| Offer | An existing offer on the data provider's website presenting a property for sale or rent. |
| Collected offer | Number of offers with positive code response (successfully scraped) or offers acquired from data provider. |
| Outlier | An outlier is a data value that lies in the tail of the statistical distribution of a set of data values. In the distribution of raw data, outliers are often regarded as more likely to be incorrect. <i>Note: In the case of this project an outlier is an observation that deviates significantly in the data distribution. In fact, it may be a correct or incorrect value.</i> |
| Room | A room is defined as a space in a housing unit, or in living quarters other than housing units, enclosed by walls reaching from the floor to the ceiling or roof covering, or at least to a height of 2 meters above the ground, of a size large enough to hold a bed for an adult (4 square meters at least) and at least 2 meters height over the major area of the ceiling. Normal bedrooms, dining rooms, living rooms, habitable cellars and attics, servants' rooms, kitchens and other separate spaces used or intended for habitation all count as rooms. Passageways, verandas, lobbies, bathrooms, and toilet rooms should not be counted as rooms, even if they meet the criteria. <i>Note: In the case of this project it is assumed that all the rooms counted in the offer meets the criteria.</i> |
| Surface area (Net floor area) | Constitutes the floor area which can actually be used by the occupants of the building. It is referred to as the internal area and excludes all construction features. |
| House | Means that no internal space or maintenance and other services are normally shared with other dwellings. Sharing of a garden or other exterior areas is not precluded. |



| Concept | Definition |
|--|---|
| Detached house | Means the dwelling has no common walls with another dwelling. If it is a separate building, without any common walls or ceiling with other dwelling, it is counted as a detached house. |
| Semi-detached or terraced house | Refers to two dwellings sharing at least one wall, and 'terraced' refers to a row of (more than two) joined-up dwellings, considering houses in which are more than one dwelling, sharing at least one wall (or ceiling) but have separate entrances. |
| Apartments or flats | Apartments or flats in a building normally share some internal space or maintenance and other services with other units in the building. Commonly there is also a shared entrance to the building as such. |

The set of common basic indicators for partners to produce was created from the variables possible to obtain. The final set of basic indicators is as follows (category classes are presented in Table 11.):

- number of offers
- average price per square meter (sale)
- share of offers by price per square meter divided into classes (sale)
- average surface area in square meter
- share of offers by surface area classes
- average number of rooms
- share of offers by number of rooms
- average price (rent)
- share of offers by price classes (rent).

2.1. Bulgaria

For two years before the project launched the BNSI has been extracting data on real estate rents from national websites as the request of statistical experts from the Price Statistics Department. The list of these initially chosen websites was taken as a starting point for UC1 purposes. Obtained data is used as an additional source of information for CPI, but it is not directly involved in its calculation. The list of Bulgarian websites was evaluated using the Checklist to assess web data sources. All websites from the list offer real estate sales and rentals.

Table 3. Assessed real estate portals with highest score (maximum = 100)

| Web portal | Score (June 2021) | Score (July 2024) |
|--------------|-------------------|-------------------|
| imot.bg | 100 | 0 |
| imoti.com | 96 | 89 |
| address.bg | 93 | 0 |
| mirela.bg | 89 | 96 |
| holmes.bg | 86 | 71 |
| nedvijim.com | 86 | 0 |

Two of the assessed portals scored 0 points in the evaluation process. The reason for the negative scoring was that some of the mandatory variables *offer_surface*, *offer_floor*, *offer_rooms*, were assessed negatively, which automatically led to rejection of the whole source. From the remaining six web portals, two have been deemed not viable - imot.bg and address.bg. The first website later occurred to not allow access to filtered content via the scripts and necessitated manual filtering. The second was impossible to scrape due to the abundant presence of JavaScript, which is not allowed by the software used for acquiring data. That is why finally only four data sources were chosen for scraping process to produce experimental statistics for real estate. The mirela.bg was scraped regularly, but the data was not processed nor aggregated. It would require a different approach and software for data processing and editing. Whereas nedvijim.com showed to be an unstable source which later the number of offers decreased below the assumed acceptable level. The lower or higher scores for 2024 are due to non-structured information about the optional variables found in ads. The mandatory variables scores were the same.

All the chosen websites presented real estate for sale and rent in lists, paginated with possibility to use filters to narrow the scope. Information from robots.txt on all the websites were very simple, giving possibility to use all of the site areas. What is more BNSI has no legal constraints for data scraping from Internet sources.

2.2. Poland

Due to the project's large scope, Statistics Poland decided to choose national web portals and exclude portals limited to regional offers. Initially, ten portals of different sizes were reviewed, however only five were chosen due to the time constraint needed to prepare a dedicated program for scraping and processing collected data. All of the chosen portals presented data with advertisements referring to objects for rent and sale. The selected portals were not the most popular but presented different opportunities due to their diversity. One of the portal provided access by API. A summary of the assessment of these portals is presented in the Table 4.

Table 4. Assessed real estate portals with highest score (maximum = 100)

| Web portal | Score (June 2021) | Score (July 2024) |
|--------------|-------------------|-------------------|
| gratka.pl | 80 | 80 |
| domy.pl | 89 | 89 |
| szybko.pl | 82 | 82 |
| domoferty.pl | 75 | 74 |

The website of gratka.pl presents advertisements for products from various fields (real estate, automotive, job advertisements, animals, house and garden and many more). The offers cover new products and second-hand ones. In the real estate category are advertisements of apartments, houses, garages, rooms, fields, commercial premises and development investments. Gratka.pl presents more than 90000 offers of apartments for sale and over 15000 for rent (as for September 2024). The advertisements are also automatically shown on several regional newspaper websites. The entire list of suggestions is accessible through the link (<https://gratka.pl/nieruchomosci/mieszkania>). There is also the possibility to access links by referring to the localization of the offered object. The list is divided into pages with 32 advertisements on each of them, with maximum 312 pages, which leads to a problem with acquiring offers from localization with more than 32*312 offers. At the top of the page is a group of filter items to narrow the amount of presented results. The filter works within the presented page dynamically. Each advertisement has its own unique number and is presented with pictures of the real estate, localization and basic information followed by text description. Each ad has the same HTML structure and class names. The advertisement description

is in plain HTML; the localization information is also a plain text giving names of NUTS levels. This information is not the same for every offer. All other necessary variables are in a text format on the page.

Domy.pl is a portal presenting only advertisements of real estate. The offers are categorized into groups regarding the type of the object (apartment, houses, fields, garages, rooms, commercial objects) or regarding the owner (new constructions or secondary market). The website covers over 90000 apartment offers from Poland for sale and over 11000 for rent (as for September 2024). All offers for the entire country are available by <https://domy.pl/mieszkania>, and offers for sale are accessible through the <https://domy.pl/mieszkania-sprzedaz--pl>, and the rent offers starts with <https://domy.pl/mieszkania-wynajem--pl>. Further, the offers may be accessed by a dedicated location link or by filtering. The portal also presents advertisements from outside of Poland. Every advertisement has its own unique number and is presented with pictures of the real estate, localization and basic information followed by text description and further: agency's offer number, publication date and actualization date. Each advertisement has the same HTML structure and class names. The advertisement description is in plain HTML; the localization information is also a plain text giving names of NUTS levels. This information is not the same for every offer. All other necessary variables are in a text format on the page. Additional information is presented in two columns, with each row displaying different variable pair.

Szybko.pl presents only real estate advertisements (apartments, houses, garages, rooms, fields) with more than 27000 offers of apartments for sale and over 4000 for rent (as for September 2024). It is a group of several thematic websites that allow posting various types of real estate offers on the Internet. Each advertisement is published on several portals belonging to szybko.pl, including 16 regional portals (for every region-voivodship in Poland), and is automatically exported to cooperating portals. The entire list of offers is accessible through one, visible link (<https://szybko.pl/l/na-sprzedaz/lokal-mieszkalny/Polska>). The list is divided into pages with 12 advertisements on each of them. On top of the page is a group of filter items to narrow the amount of presented results. Each advertisement has its own unique id number and is presented with pictures of the real estate, localization and basic information followed by text description and further agency's offer number and publication date. The advertisements have the same HTML structure and class names. The descriptions are in plain HTML and the localization information is also a plain text giving names of NUTS levels. This information is not the same for every offer. All other necessary variables are in a text format on the page.

The portal called Domoferty.pl in the beginning provided an access by API, which was later cancelled. This portal presents only real estate advertisements with more than 14000 offers for sale and very small number of offers for rent – less than 1000. It achieved the lowest score in the assessment due to less variables available in the offers and dynamic load of offers. The offers were possible to scrape only using tools like Selenium (for dynamic pages), but thanks to the availability of a publicly accessible API it was possible to start data collection.

During the stage of assessing websites a popularity check has been done. Portals like szybko.pl and domy.pl were located in the middle of the top 10 in 2020 and were close to the first ten portals with the highest number of views in 2021 (according to similarweb.com stats).

2.3. Germany – HSL

Note: For HSL, Use Case 1 and Use Case 2 are closely related, since both use the same data sources. Description of these data sources, data collection and challenges during data collection, as well as

preparation and analysis apply to both use cases. However, selection of data sources and data collection primarily focuses on Use Case 2 – new constructions.

Sources include the largest and most popular general platforms as well as more specialized platforms with a focus on the Rhine-Main region and covering mainly complete larger construction projects. Several portals have been assessed, but finally seven have been chosen for regular data collection. Besides relevance and coverage, the choice of platforms was also based on accessibility in terms of (missing) access restrictions e.g. in form of captchas: Gathering data from the site must not bypass any obvious barriers for automatic software for data collection.

Table 5. Portals general information

| 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 | Germany | 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 | | discontinued during project phase |

Table 6. Assessed real estate portals with highest score (maximum = 100)

| Web portal | Score (June 2021) | Score (September 2024) |
|------------|-------------------|------------------------|
| Portal 1 | 96 | 100 |
| Portal 3 | 83 | 93 |
| Portal 4 | 96 | 93 |
| Portal 6 | 87 | 74 |

Regarding developing and maintaining scrapers, HSL and SSI-BBB have shared the workload: HSL was responsible for developing and maintaining as well as running scrapers for three data sources, SSI-BBB for three (during project phase, gathering data from one data source was discontinued).

Although most of the used portals cover Germany, only data from advertisements for objects in Hesse (Berlin and Brandenburg respectively) have been collected.

For one data source from the HSL (portal 1), there was no scraping as there was an agreement for regular data delivery (only for Hesse) with a platform provider. However, since this data in principle is the same that would have been collected from the website, this also is considered “data from the web”. This data source has the advantage that all advertisements within a month are part of the data delivery and no

advertisement is missing because of the scraping interval. This is one reason why some of the analyses for Use Case 1 and Use Case 2 are based on this source alone.

Typically, each of the portals can be searched for (and return search results) by federal state.

The result pages contain an overview list of all objects according some search criteria (e.g. region like federal state, “sale” or “rent”, “house” or “apartment”, year of construction). This overview list already contains some basic information on each object, e.g. ID number, type of object (sale, rent), building type (e.g. house or apartment), size (number of rooms or surface size in m²), prices. Each of the listed objects in this result list links to the specific advertisement page. These individual pages are the main data source for this use case.

Usually, these advertisements from one source contain a standardized set of information (i.e. fixed positions on the page with fixed locators, e.g. XPath selectors, size in square meters, number of rooms, a price, year of construction, a condition of the object, energy type and consumption, address information). However, beside these pre-defined, a different possibilities of providing these information are permitted. Advertisers are even free to fill in a text field or not give detailed information at all.

Figure 1 gives an example for an overview page after a search for objects in “Hesse”, and one example for a specific advertisement for one of the portals used.

Left Screenshot: Search Results Overview

| Suchergebnisse | Exposétitel | Preis |
|----------------|--|--|
| | Exklusives Penthouse mit Taunus-Panorama und Badelandschaft Objekt-Nr.: CM-315305 Adresse: 61184 Karben Zimmer: 4.50 Wohnfläche: 137.80 m² | 788.000 € Kaufpreis Privatangebot |
| | Top RMH mit Dachterasse + Keller und bis zu 4 PKW-Stellplätzen in Riedstadt Objekt-Nr.: CM-315269 Adresse: Parkstraße 5a, 64560 Riedstadt Zimmer: 5.00 Wohnfläche: 120.00 m² Grundstücksfäche: 185.00 m² | 649.900 € Kaufpreis Privatangebot |
| | Schmuckstück: Maisonette mit Seeblick zur Vermeidung oder Selbstnutzung Objekt-Nr.: CM-315293 Adresse: 63303 Dreieich Zimmer: 2.00 Wohnfläche: 82.00 m² | 229.000 € Kaufpreis Privatangebot |
| | Schöne Praxis auch als Wohnung nutzbar zu verkaufen Objekt-Nr.: CM-293923 Adresse: Hauptstr. 218a, 65517 Gonsheimertal Zimmer: 7.00 Wohnfläche: 125.00 m² | 236.000 € Kaufpreis |
| | 3-Zimmer-Wohnung als Kapitalanlage in Darmstadt-Eberstadt zu verkaufen Objekt-Nr.: CM-315157 Adresse: 64297 Darmstadt Zimmer: 1.00 Wohnfläche: 50.50 m² | 136.000 € Kaufpreis Privatangebot |
| | Großzügige 3-Zimmer-Erdgeschosswohnung mit Süd-Ost-Garten in Darmstadt-Eberstadt Objekt-Nr.: CM-270369 Adresse: Ludwigshofstr. 113, 64285 Darmstadt Zimmer: 3.00 Wohnfläche: 102.00 m² | 688.000 € Kaufpreis Privatangebot |
| | Hochwertig saniert 5/2024 - Wohnung in Eschersheim für Kapitalanleger - ohne Maklergebühr! Objekt-Nr.: CM-315210 Adresse: 60433 Frankfurt Zimmer: 3.00 Wohnfläche: 59.00 m² | 360.000 € Kaufpreis Privatangebot |
| | Hochmoderne Bauhausvilla in traumhafter Blicklage nahe Königstein im Taunus Objekt-Nr.: IE-230487 Adresse: 61462 Nähe Königstein Zimmer: 6.00 Wohnfläche: 235.00 m² Grundstücksfäche: 850.00 m² | 1.980.000 € Kaufpreis |
| | Bestlage Georgenborn: Penthouse-Maisonette mit Traumblick Objekt-Nr.: IE-280452 Adresse: 65398 Schlangenbad Zimmer: 5.00 Wohnfläche: 173.00 m² | 950.000 € Kaufpreis |
| | Wertiges Einfamilienhaus sucht dich! Objekt-Nr.: CM-315183 Adresse: 63667 Nidda Zimmer: 4.00 Wohnfläche: 156.00 m² Grundstücksfäche: 675.00 m² | 275.000 € Kaufpreis |

Right Screenshot: Detailed Advertisement

Kapitalanlage oder in das eigene Zuhause? Neubau 3-Zimmerwohnung, individuell, zentrumsnah, modern.

Objekt-Nr.: IE-272670
Objekt-Nr. des Maklers: 10300
Objektart: Wohnung
Objekttyp: Elagenwohnung

Kaufpreis: 361.200 €

Übernahme ab: nach Absprache

Objektschreibung:
Hier entsteht eine 3-Zimmer-Erdgeschosswohnung mit Gartenanteil in einem Zweiparthenhaus mit Energieeffizienz der Systemklasse KfW-förderfähige. Energieeffizienz mit Wohnklima zum Gesehen, sparsam in Energieverbrauch und Kosten.
Nur 10 Minuten vom Zentrum entfernt. Einkaufs-, Fußgängerzone, Wochenmarkt und S-Bahnhof mit guter Verbindung nach Frankfurt. Der Kinderspiplatz und ein Kindergarten sind fast vor der Haustür. Dennoch ruhig gelegen erwartet Sie Ihre Traumwohnung in grüner Umgebung. Hoher Lebens- und Freizeitwert für Sie alleine oder zu zweit.

Angaben zur Ausstattung:
KfW Effizienzhaus 40 NH, KfW-förderfähig
Wärmepumpe mit Kälteelektrik für warme Winter und kühle Sommer
Photovoltaikanlage
Zellulose-Woll-Dachdämmung
3-fach Verglasung
Elektrische Rollläden
Multimedia in allen Zimmern
Möbelausstattung Elektro/Sanitär
Dosenanlage nach Wunsch
Maisonetteboden

Courtage:
Vom Käufer 3,57 % des Kaufpreises inkl. 19 % MwSt.

Zustand: Einberaub
Wohnfläche: 56.00 m²
Zimmer (Anzahl): 3.00
Wesentlicher Energieträger: Holz
Baujahr: 2024

Lagebeschreibung: Ehemals unter dem Altort "Häselstadt" bekannt und heute ein Begriff als "Kiesstadt im Grünen", das ist Hofheim im Taunus. Die 42 000 Einwohnerstadt hat eine ideal zentrale Lage im Städtedreieck Mainz, Wiesbaden und Frankfurt am Main, mit jeweiligen direkten S-Bahnanschlüssen. Außerdem gibt es einen direkten Autobahnanschluss. Sowohl zum kulturellen und sportlichen Ergehen lädt Hofheim ein. Anziehungspunkt für historisch oder städtebaulich interessierte Gäste ist die Altstadt. Vom Kindergarten bis zum Altort sind alle Schulformen in Hofheim vorhanden.

Objektadresse
65719 Hofheim
Hessen, Deutschland

Anbieter
Claus Blumenauer Immobilien GmbH
Claus Blumenauer Immobilien
Impressum des Anbieters
alle Rechte des Anbieters
09174 58100

E-Mail an Anbieter
Ihre Name*
Arbeits- / Vorname
Nachname
Wohnort*
Strasse / Haus Nr. / PLZ
Ort
Kontaktform*
E-Mailadresse
Telefon
Ihre Nachricht*

☒ Eine Kopie der E-Mail an mich schicken
Sicherheitsabfrage*
10 + 1 =
Die verlinkte Rechenaufgabe ist zu lösen und das Ergebnis anzugeben. (28. März 2024, 10:11 Uhr, 10 Sekunden)

E-Mail abschicken

Figure 1. Search result overview and specific advertisement

In this example, the search for objects in “Hesse” resulted in 36 overview pages with up to 10 individual objects listed on each of these pages. The advertisement shown even fulfills the requirements of Use Case 2 as it describes a newly constructed house (Year of construction “2024”, condition “First occupancy”). Note that object type (“sale”), building type (“apartment”) as well as price (“361,200 €”), size in square meters (“56.00m²”) or the number of rooms (“3.00”) is given in a standardized way (i.e. they have a fixed XPATH on every page). Other information is given in free text fields. Also note that there is no complete address given but only postal code and city name. When combining different sources, this makes it very hard to decide if two advertisements refer to the same object.

2.4. Germany – SSI-BBB

Evaluation of Relevant Websites

In collaboration with a second German partner, HSL, 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 were 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 HSL, 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 HSL to ensure alignment with the project objectives.

Overview of Key Real Estate Platforms for Use Case 1

Three major platforms, 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 platforms, providing a deeper understanding of their specific roles within the real estate landscape.

Portal 2

Portal 2 is a comprehensive real estate platform, 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 offerings. 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, surface area, number of rooms) to tailor searches to individual needs. This platform serves as the primary data source for this use case by Berlin-Brandenburg.

Portal 7

The second platform, 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 platform was not used as a primary data source.

Portal 5

In contrast to the broader focus of the previous platforms, 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.

Table 7. Assessed real estate portals with highest score (maximum = 100)

| Web portal | Score (June 2021) | Score (July 2024) | Description |
|------------|-------------------|-------------------|--|
| Portal 2 | 97 | 97 | Comprehensive real estate platform; Focus on rental and purchase properties; Extensive, regularly updated database; Nationwide coverage across Germany; Primary data source for this use case |
| Portal 7 | 0 | 97 | Wide range of listings: residential, commercial, land; Nationwide coverage, particularly in urban areas; Significant overlap with Portal 2; Not used as a primary data source due to redundancy |
| Portal 5 | 97 | 97 | Specializes in new construction projects; Focus on newly built and planned residential developments; Nationwide coverage, with emphasis on larger cities and active construction regions; Used specifically in use case 2 |

Ensuring Stability and Coverage

As previously mentioned, Portal 2 was used as the primary platform 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 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 offer 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 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 portals. However, due to a significant overlap in listings between these portals (for more details, see Deliverable 3.6 Report on methods and feasibility to track construction activities based on real estate web portals), 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 HSL 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.

2.5. Finland

Possibility to use online data on real estate market to produce additional statistics begun couple of years before the UC1 project started. In the Finnish real estate market, there are two main providers online. Negotiations were held with each of them, in which mutual data collaboration was discussed. Thus, during the project the data was gathered based on contract instead of web scraping.

It was found that the proportion of rental advertisement data from private individuals was insufficient on one of the providers. Moreover, restrictions imposed on the dataset would have poorly suited the requirements for statistical production. Instead, dataset from the other provider (oikotie.fi) covered private individuals' rental advertisement information more comprehensively and, what was also crucial, the dataset was available without publication restrictions. Thus, Finland started gathering data from Oikotie in 2020 based on an agreement.

Oikotie.fi contains a vast coverage of advertisements from all of Finland. Many kinds of agents use this service, for example private persons, companies, and municipalities. The data has approximately 40000 to 60000 houses and apartments for sale and 20000 to 40000 for rent every month. The service also includes summer cottages and other vacation residences, garages, plots, and forest land. For Statistics Finland however, only data on houses and apartments are gathered, vacation homes excluded.

This data source was assessed with full 100 points, both in 2021 and 2024 using the prepared Checklist. The data structure has stayed the same throughout the project's years. Data delivery has been on-time, up to date, with good representativity and stable, high coverage. On rare occasions problems occurred, but were always solved quickly with a participation of a provider representative.

2.6. France

It was decided beforehand not to scrape websites, as for France web scraping was considered as posing legal and technical issues, such as copyright, instability of the website architecture, etc. Hence the way to go was to foster agreements between Insee and the data providers. Several providers were contacted (SeLogger, pap.fr and leboncoin), following the partnerships that were signed for a project conducted by the French Ministry of Housing (<https://www.ecologie.gouv.fr/carte-des-loyers>). However, only SeLogger have responded to solicitations by Insee. After a one-year negotiation round, Insee and SeLogger managed to agree on a partnership giving way to data transmission for expertise and evaluation by Insee, aiming at disseminating experimental statistics on rents⁴.

The platform SeLogger has a wide audience across the French territory, and aggregates ads for the entire country. Its use and coverage may vary for some specific areas, where regional platforms may reach a broader audience (e.g. Ouest France Immo for Brittany region). However, the platform is generalist enough and widely well-known.

The website is fed with ads posted by realtors, and recently individuals who do not want to use intermediary for their real estate transactions or renting out. It offers the possibility for realtors to bulk upload the ads they want to publish and is specifically focused on real estate activities (transactions, rentals). It is possible to browse the website for specific parts of the territory (down to the municipality level). The ads encompass information on the dwelling that is to be sold or rent: size, number of rooms, number of bedrooms, price, equipment, etc. A number of information may also be retrieved through the short descriptive text that goes with the ads. Every advertisement is uniquely identified in the database and may be retrieved on the website provided that it is still available at the time of the data processing.

⁴ A few months after signing the contract, SeLogger have merged with MeilleursAgents. In national public communication INSEE uses the name SeLogger/MeilleursAgents for this data provider.

3. Programming, production of software

Building a system for data collection and processing may be based on different tools or needs of the users. Establishing a project for web data acquisition need to fit into both the need of the statistical organization, but also has to be adequately adjusted to the type of data that are going to be collected. At different levels there is a need to provide control system to monitor the programming errors or acquired data. Which refers both to the information that is already obtained, as well as information on the quality of publicly available web data. It is necessary to periodically check the terms and conditions relevant to the scraped websites, follow the changes in terms of copying and use of the gathered data, and check for the availability of APIs. Those processes may be launched before the actual process of data collection and keep executing later in temporal checks before the actual scraping – if it occurs rarely (e.g. once every 6 months) or adjust it if scraping is performed often (e.g. it is performed weekly).

In the efforts to enhance data extraction techniques, an exploration of both screenscraping and API scraping has been executed. Organisations used tools and programming languages in which they were most experienced. By experimenting on both solutions it was possible to identify their respective advantages and disadvantages, ultimately gaining a clearer understanding of how they can complement each other in data retrieval tasks. Screenscraping 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.

APIs offer several significant advantages over screenscraping, 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 screenscraping, 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, screenscraping 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. Screenscraping, 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 screenscraping solutions. While Python-based tools like Scrapy can streamline the scraping process,

developing a robust and adaptable screenscraping tool often requires more extensive programming effort than integrating an API, which typically involves simpler calls and data handling.

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, screenscraping 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 this use case, a public key provides access, and there are no associated costs for using the APIs. This may not always be the case with other APIs, where usage fees can apply.

Lastly, APIs may offer limited information compared to screenscraping. 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. Sometimes, the API may return fewer data points than what might be accessible via screenscraping, 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 carefully considered, especially in cases where screenscraping may provide more flexibility.

3.1. Bulgaria

Implemented IT infrastructure and software

Based on the previous experience, BNSI chose the following environment for data scraping and processing:

- Windows 10/11 – Development Workstation machines
- Windows 2019 Server – Production Virtual machine dedicated for the project
- Initial set of inventory: Python 3.7.0 – Libraries json, re, datetime, os, sys, logging, urllib, glob, smtplib, mimetypes, email, Scrapy~=2.5.1, w3lib~=1.22.0, pytz~=2021.3, itemadapter~=0.4.0, pandas~=1.3.4, numpy~=1.21.3.
- Actual set of inventory: Python 3.11.0 – Libraries json, re, datetime, os, sys, logging, urllib, glob, smtplib, mimetypes, email, Scrapy~=2.9.0, w3lib~=2.1.1, pytz~=2023.3, itemadapter~=0.8.0, pandas~=2.0.3, numpy~=1.25.0.

BNSI first started developing the software for data scraping based on Python in 2019 for an internal project analyzing price statistics. The software has been working steadily for the next two years, and that was why the same approach was undertaken in the project conducting within UC1. Starting in autumn 2021, a process of updating the software had begun. There was a need to adjust the software to be less country-

specific, more user friendly, and to increase its flexibility. The scraping software works with JSON configuration files, where data scraping selectors for indicators or text parts from the web pages are defined, and options for scraping policies and processes are established. If the process fails, the software logs the scraping process in log files for debugging purposes.

Additionally the webscraper.io Chrome browser extension was used for the scraping of JavaScript sites, because that kind of websites cannot be scraped with Bulgarian Python software.

Scripts/codes

The scraping policies and processes options that the software uses are:

- URLs to be scraped
- Bot name (scraper name)
- Delay between requests
- Depth of scraping
- Robots.txt obey
- Stakeholders' e-mails
- Date of history data in days

The software for data processing uses four predefined functions:

- Reading the scraped data
- Validate the scraped data
- Save the processed data
- E-mail the results to the stakeholders

The processing script for each data source is developed using the predefined functions and specific processing instructions for the mandatory and optional variables present at the source. The scraped data are stored in separate CSV files by date of acquisition, with fields generally different from the scraped variables agreed upon beforehand between partners. Some of the fields could be the agreed variables, but most of them are text content from different parts of the scraped pages. Then the last scraped data is processed with the dedicated script, and the agreed variables are derived from it and saved in separate file(s). Then scraped data undergoes some general validation (check the last file to be predefined scraped date format for the moment), and the result is stored in a file with the name *_noerrors_xsl.csv. File is then processed, and the variables are derived and stored in a CSV file, with the date of acquisition in the name. The process is shown in the Figure 2.

At the end the software sends an email with summary logs from the data scraped and the processed data in the attached file(s). The software and basic documentation (data scraper, processing functions and processing examples) is available on the restricted project page.

The scraping and processing of chosen websites is done automatically on regular basis with the help of Windows Task Scheduler. A batch script is executed with the website of interest passed as the parameter, which calls the scraping software and immediately after completing it calls the processing software.

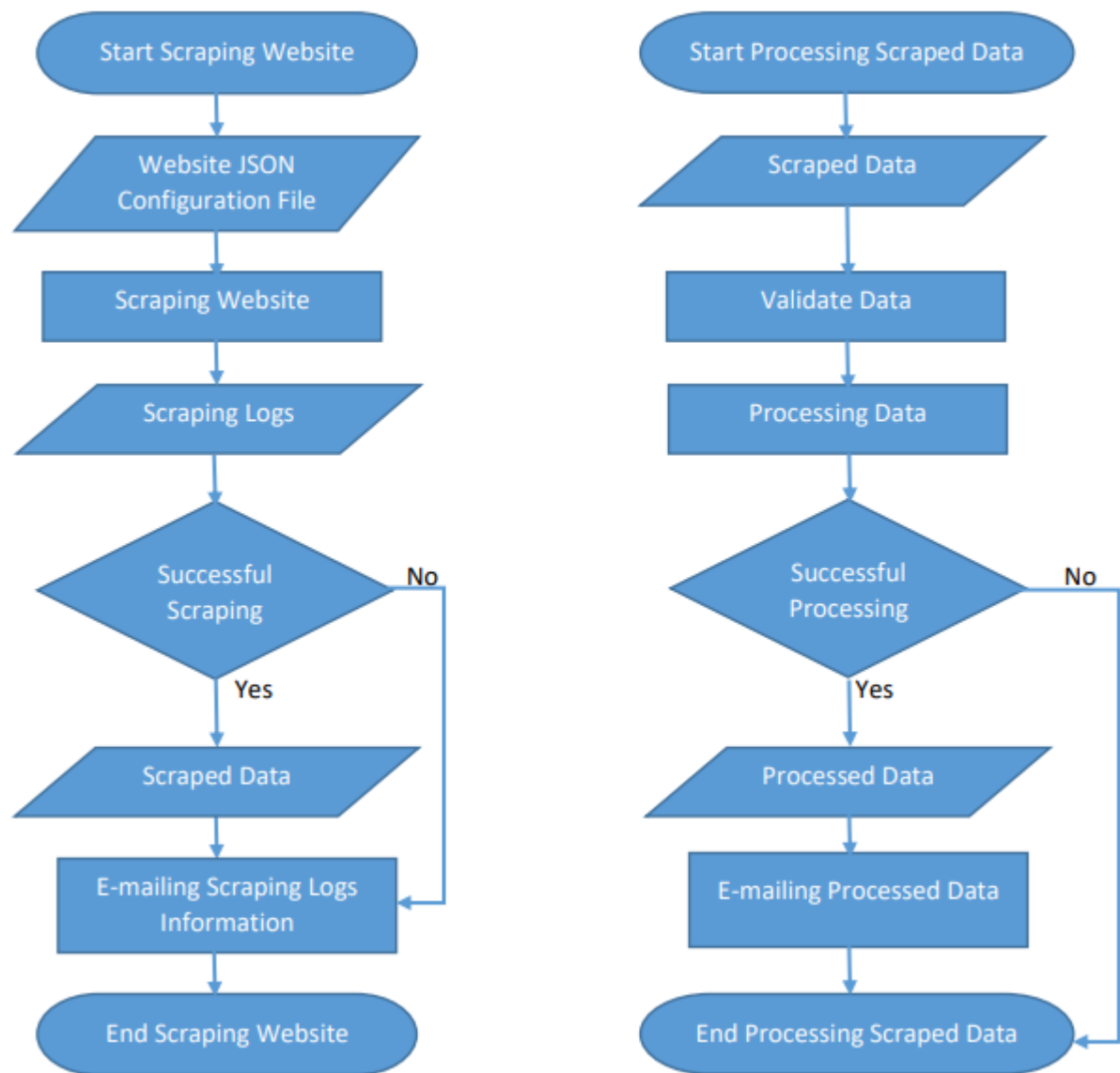


Figure 2. Data collection process steps

| Name | Status | Triggers | Next Run Time | Last Run Time | Last Run Result | Author |
|--|--------|--|----------------------|----------------------|-----------------|---------|
| NSI Scraper --- online.prices.techmart.bg | Ready | At 4:12 every Friday of every week, sta... | 20.9.2024 r. 4:12:46 | 13.9.2024 r. 4:12:46 | (0x0) | NSI.BG\ |
| NSI Scraper --- online.prices.technopolis.bg | Ready | At 3:42 every Friday of every week, sta... | 20.9.2024 r. 3:42:46 | 13.9.2024 r. 3:42:46 | (0x0) | NSI.BG\ |
| NSI Scraper --- online.prices.zora.bg | Ready | At 3:32 every Friday of every week, sta... | 20.9.2024 r. 3:32:46 | 13.9.2024 r. 3:32:46 | (0x0) | NSI.BG\ |
| NSI Scraper --- real.estate.holmes.bg | Ready | At 3:14 on day 4 of January, February,... | 4.10.2024 r. 3:14:55 | 4.9.2024 r. 3:14:55 | (0x0) | NSI.BG\ |
| NSI Scraper --- real.estate.imot.bg | Ready | At 3:23 on day 4 of January, February,... | 4.10.2024 r. 3:23:13 | 4.9.2024 r. 3:23:13 | (0x1) | NSI.BG\ |
| NSI Scraper --- real.estate.imoti.com | Ready | At 3:33 on day 4 of January, February,... | 4.10.2024 r. 3:33:46 | 4.9.2024 r. 3:33:46 | (0x0) | NSI.BG\ |
| NSI Scraper --- real.estate.mirela.bg.loop | Ready | At 5:22 on day 4 of January, February,... | 4.10.2024 r. 5:22:22 | 4.9.2024 r. 5:22:22 | (0x0) | NSI.BG\ |
| NSI Scraper --- real.estate.nedvijim.com | Ready | At 3:13 on day 4 of January, February,... | 4.10.2024 r. 3:13:13 | 4.9.2024 r. 3:13:13 | (0x0) | NSI.BG\ |

Figure 3. Windows Task Scheduler view

Also, during the project, the software was implemented with a module for processing and visualization of logged information about the scraping process. Twenty logged variables are showed and filtered in table with a graphical visualization for six of them (see Figure 4).



Figure 4. View of Bulgarian NSI's scraper log for monitoring collection of data from multiple sources

Problems encountered during executing the scraping software

- For two of the sources the scraping software worked without problems with only few non-responses and timeouts;
- One of the sources had a lot of HTTP 500 errors till the April 2023 and fewer after that, but these were server errors and most likely were not related to the scraping software;
- One of the sources was scraped differently, thus has no logs analysis;
- One sources (imot.bg) was initially configured to be scraped, but since the website had been using JavaScript for filtering the content, no data was ever scraped.

3.2. Poland

Pre-scraping tasks

The initial task involved checking the robots.txt files, sitemaps, and terms of reference to uncover hints from website owners about how to use the information presented on the sites. Also the complexity of potential websites for scraping was assessed by identifying the most suitable areas to acquire apartment offers.

Subsequently, a deeper assessment of the websites has been conducted by examining the HTML code for rules relevant to the needed information and identifying scraper-friendly areas on the site. Four scraping programs were prepared: three built from scratch and one received from BNSI (requiring adjustments for the new website). One of them was based on a publicly available API connection provided by the website owners.

Initially, prototypes of the scrapers were created to test data acquisition on a smaller sample, ensuring error-free processes. Later, just before April 2022, it was decided to perform a full bulk download of offers.

To establish a common working environment for the scrapers a new infrastructure was built. This decision was primarily driven by the IT department's desire to avoid using the same external IP address for scraping as for regular institutional work.

Implemented IT infrastructure and software

The web scraping environment consists of an application server (which runs software or scripts that scan websites), an application server (to process the downloaded data) and a database server:

- Application server 1 – a server on which scraping software is executed – proprietary Python scripts, Scrappy software and others indicated by users. The server is located in the DMZ zone, has a direct connection to the Internet, bypassing the proxy server and a separate IP address assigned. Scraping applications have direct access to the SQL Server database and disk resources, where data downloaded from the Internet in the form of text files is stored. Installed software and libraries - Python 3.7.11 – re, datetime, os, sys, logging, urllib, beautifulsoup4==4.10.0, numpy==1.21.5, pandas==1.4.1, requests==2.27.1
- Application server 2 – a server in the internal network, on which software will be installed for viewing and processing data downloaded from the Internet. The software will be made available to employees of public statistics responsible for data processing. The server will have access to one disk resource – with processed data and to the SQL Server database.
- Network Attached Storage 3TB – a resource intended for storing data downloaded from the Internet, in txt or csv files.
- SQL Server 2019 (Dev. Edition) database server – a database that stores and processes data downloaded by scraping applications. The server has 2 disks, each 1 TB, with the possibility of expanding them.

Scripts/codes

All the prepared scrapers are manually executed every month due to the experimental nature of the project. During the initial phase of data acquisition, the programs were frequently adjusted and upgraded in response to errors or new decisions made during the methodological development of the project. Currently, the scrapers are maintained by a single employee, which could potentially become a challenge in the future.

The scrapers utilize BeautifulSoup 4.12.0 libraries. The scrapers' logic is divided into several parts: response checking, links gathering, data retrieval, basic data processing (such as removing special alphanumeric characters), exporting to a CSV file on the scraping server, and creating a copy on the NAS. In the first step the scraper downloads links with non-special offers iterating by regions and from every pagination available on the provider's site. In the next step scraper downloads data and retrieves selected variables from every offer link until it reaches the last earlier gathered link. Each execution of the scraper generates a log saved on the server.

The scraper received from BNSI is based on Scrapy solutions and its actions are described in chapter 3.1.

Problems encountered during executing the scraping software

- In one of the portals, an issue was encountered with duplicate naming of poviats (at the NUTS3 level). Specifically, repeated names in different voivodeships (NUTS2) had URL addresses with additional numerical designations, which did not correspond to the classification names, e.g.:
Powiat-bielski & Powiat-bielski-308
Powiat-brzeski & Powiat-brzeski-216
It occurred that an additional number was added to the poviat that appears in the second alphabetically sorted voivodeship.
- It was also considered to first scrape the sitemap with links to all the locations. However, querying such a large number (20000) of pages increased the program's runtime. In most of the cases there was no offers in a given location, but it generated a request and extended the execution time. It was also contrary to basic assumptions to burden the website as little as possible.
- In the initial phase of the project, a few problems were encountered while adapting the internal infrastructure, which initially did not anticipate problems resulting from: monthly password changes, changing users during the project, updating software while scripts were executing, sharing files between users, moving scripts from an external location to an area inside the dedicated infrastructure.
- In December 2022 there were some technical issues with getting responses from the source, which eventually disabled collection of data from the source 2
- In August 2024, for the gratka.pl, there were two changes to the html structure on the page. First, a change, which required significant modifications to the scraper. Then a return to the old version of the page. Finally, the original scraper was used. The page stayed under observation as it has been already seen that it still uses its new version in other product categories. In September 2024 once again the page layout was changed.

3.3. Germany – HSL⁵

HSL used one scraper for all three portals, programmed in Python using the Scrapy-Framework.

In the beginning of the project, email notification was investigated as another method of collecting data from real estate websites. Many real estate platforms offer this kind of service: whenever there is a new offer according to some search criteria, an email with the offer (or a link to the offer) is sent to a registered recipient. These emails could also be processed in order to extract information on newly constructed objects. This way, no offer could be missed due to a too long scraping interval.

Since now, a scraper capable of processing static HTML was sufficient in order to collect the data needed during this Use Case. The scraper has been run automatically each week and output data has been stored in text files. The scraper already included some basic pre-processing, but extensive data cleaning and further pre-processing followed after data have been transferred to an the internal area.

The scraping interval of one week includes the risk of missing “short-term advertisements”: offers that are online only for some days or even only hours. Typically, this could be the case in areas with large demand especially of apartments to rent (and presumably less risky for houses to buy). Portal 1 uses data directly from one platform provider without scraping, so there is no risk of missing offers or undercoverage of short-term advertisements.

3.4. Germany – SSI-BBB

Implemented IT infrastructure and software

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 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.
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

⁵ More information about used IT infrastructure and software is presented in Deliverable 3.6 Report on methods and feasibility to track construction activities based on real estate web portals.

(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.

Portal 2 has a huge advantage when scraping, since it grants access to the offer as a JSON file through the source code of the page. Identifying the selector corresponding to the attribute reduced the coding complexity for this portal via Spyder. A Scrapy Spider is a programmable web crawler that efficiently extracts and stores data from websites, leveraging Scrapy's robust tools and features. This was not the case for Portal 7 and Portal 5, where each attribute had to be identified separately. This also required some previous cleaning of symbols inherent to the HTML code that were present with the attributes to extract.

Adapting scraper to website structure

Due to changing website structures SSI-BBB had to adapt the scrapers frequently. Checks of the scraped data regularly to recognize such changes and adjustments were also frequent.

One portal had some offers on pages with HTML structures differing from those of the standard offers. SSI-BBB always had to decide if the number of those non-standard ads is high enough to adapt the scraper. However, the scrapers were adapted several times during the entire observation period.

4. Data acquisition and recording

Data acquisition began in April 2022. Due to the fact that web scraped data is highly unstructured a need to implement various techniques for their improvement arose. During scraping data from various portals different issues were encountered, so each country defined and implemented their own individual cleaning rules and techniques.

Collected data were pre-processed and cleaned to calculate nine indicators for rent and sale real estate market. Mandatory variables were recoded into categorized variables to calculate the aggregated tables with indicators.

4.1. Bulgaria

The regular collection has started in April 2022 and was scheduled for the 4th of every month. Each of the chosen sources occurred to be on different steps problematic. In some cases it was impossible to keep the scheduled data acquisition plan and in case of one source a decision has been made to finally reject it due to complications with too many manual implementation needed in the scraping software.

For holmes.bg around 45000 pages have been scrapped monthly. Since changes in the structure of the data source in July 2022, the advertisement related to rent of properties, have been unavailable until November 2022 and that resulted in necessary adjustment of the script's properties. During this period, around 15000 advertisements have been collected monthly, but only those related to sales of properties. In November, due to another data source structure change, advertisements related to rent of properties became available again and the number of scrapped pages increased to 20000 monthly. That is why this source is aggregated starting from November 2022.

Imoti.com has proven a stable source of data and no challenges have been encountered during the data collection phase. The only downside of the source is that it provides a very limited number of data due to its small number of users. Around 6200 advertisements monthly, related to both rent and sales of properties have been scrapped for the period April 2022 - August 2024. Specific for this source is an over 40% of over-coverage ratio.

Nedvijim.com has proven to be an extremely unreliable data source. During the collection period April 2022 - January 2023, the scheduled scraping processes fluctuated between 0 and 10000 scrapped offers monthly. Several manual reboots of the process were required after each unsuccessful attempt for scheduled web scraping, in order to collect data. Analysis of the log files on the collection showed multiple internal server errors 500. Nevertheless, the scheduled scraping remained and since April 2023 the source went relatively stable with no major scraping issues except low number of server errors. However this source was ultimately rejected due to low number of offers.

Mirela.bg is the dedicated website of a real estate agency and possesses some peculiarities. Data has been collected on a regular basis during the period April 2022 - August 2024, but due to the individual specifics of the data source, an individual approach have been implemented in both collection and processing. The collection phase issues were overcome, but the processing needs further additional efforts and extra steps to produce the same results as the other data sources. The processing of the data is still not done for this source.

4.2. Poland

Data is collected from four portals, once a month, starting from April 2022. All data includes information about apartments only. The Table 8 presents the information that is covered by different web portals. All the variables to produce mandatory indicators are covered by all the data sources.

Table 8. Information coverage by data source (*X, if the information exist in the source*)

| Information acquired | gratka.pl | domy.pl | szybko.pl | domoferty.pl |
|---|-----------|---------|-----------|--------------|
| Date of the beginning of acquiring the set | X | X | X | X |
| Name of the website or data provider | X | X | X | |
| Unique ID code of advertisement | X | X | X | X |
| URL of the offer | X | X | X | |
| Title | X | X | X | X |
| Price from the offer | X | X | X | X |
| Currency of price | X | X | | X |
| Surface area of the offered object | X | X | X | X |
| Number of rooms in the offered object | X | X | X | X |
| Floor number on which the object is situated | X | X | X | X |
| Number of floors in the building | X | X | | X |
| Date of the offer actualization | X | X | | X |
| Date of the publication of the ad | X | X | | |
| Type of construction (new development, not new) | X | X | X | X |
| Type of ownership (private property, cooperative) | X | X | | X |
| Year of construction of the building | X | X | | X |
| Type of real estate | X | X | | X |
| Material used for construction | X | X | | X |
| Condition of the apartment (to renovate, high standard) | X | X | | X |
| Extended description of the offer | X | X | | X |
| Type of the kitchen in the offer | X | X | | X |
| If the object has a parking space | X | X | | |
| Number of parking spaces | X | X | | |
| Date from which you can rent the real estate | X | X | | X |
| Location obtained | X | | X | X |

Some problems were encountered during data acquisition process regarding the scraping procedure. Only once (in December 2022) there were some technical issues which disabled collection of data from the domy.pl. Since December 2023 domoferty.pl stopped responding. After investigating the case it occurred that previously used API connection is off. After a short try to establish the connection again it was decided to abandon this data source, because of its low usability (it was not the most abundant source and not the most variable-covering). In August and September 2024 major changes were observed in the gratka.pl website. This source changed its layout which forced to implement adjustments to the scraping software.

In overall data acquired from April 2022 is stable in terms of the number of offers. Approximately 200000 sales offers and 25000 rental offers per month were obtained from all sources. Since December 2023

(when one of the sources dropped out), the number of offers for sales is approximately 175000 offers and for rent 30000.

4.3. Germany – HSL

Data is collected weekly from three portals using webscraping techniques. For one portal, data is made available by the portal owner shortly after the beginning of the next month. Data includes information on all types of objects, houses as well as apartments or “investments”. “Investments” typically consist of several houses in one area – with the same address. Often these houses again consist of several apartments – often with very similar or even identical characteristics.

In general, the scraper for the three portals worked well and only minor changes have been necessary over a period of two years, and only a lack of resources – but not the severity of changes – caused a gap for portal 6 for some time.

The amount of data collected weekly using webscraping techniques is quite small. There have been no complaints by portal providers. The scraper used an identifying User-Agent string that linked to a HSL webpage in order to provide additional information about the origin of the scraper and reason of scraping.

In general, dozens of characteristics are available from real estate advertisements, but typically, most advertisements present only a few in a standardized way. Since data stems from another background – Use Case 2 – there is no long list of mandatory variables. In principle: all characteristics of an object can be missing since there is no duty for advertisers to provide it on the platform. Only some kind of paradata from the process of data collection is present for each advertisement (e.g., date of scraping, URL, name of source).

Even type of offer (object is for sale or rent) or building type (house or apartment) can be missing from some advertisements. In some case missing information can be derived from other characteristics. E.g. typically, one cannot buy (but rent) an 3-room apartment for a price less than 10000 €. Usually there is at least some regional information like postal code or city name, but often the complete address is not available from the advertisement at first.⁶

Table 9 gives a list of the most basic variables that are available after pre-processing datasets with even more characteristics (but typically with very high amounts of missing data).

Table 9. Information coverage in scraped datasets

| Information acquired | collected | can be missing |
|---|-----------|----------------|
| Date of data collection (time or month) | yes | |
| Name of the website or data provider | yes | |
| Unique ID code of advertisement | yes | |
| URL of the offer | n.a./ yes | |
| Type of offer (sale, rent) | yes | yes |
| Type of real estate (house, apartment, other) | yes | yes |
| Price from the offer | yes | yes |
| Surface of the offered object | yes | yes |

⁶ Only for Portal 1, there is a complete address information for all advertisements (because of data acquired on an agreement). Collecting the same data using webscraping would result in missing address data, since the address information would not be available for all advertisements as it could be hidden and/or only be available for registered users.

| | | |
|---|-----|-----|
| Number of rooms in the offered object | yes | yes |
| Floor number on which the object is situated | yes | yes |
| Type of construction (new development, not new) | yes | yes |
| Condition of the apartment (to renovate, high standard) | yes | yes |
| Year of construction of the building | yes | yes |
| Date from which you can rent the real estate | yes | yes |
| Location / Address | yes | yes |

Variables from Table 8. correspond to the “mandatory variables” and have been used to produce the monthly set of the Use Case’s “mandatory indicators” for each of the portals.

4.4. Germany – SSI-BBB

Two (Portal 2 und Portal 7) of the largest real estate portals in Germany were selected by SSI-BBB which clearly dominate the online market. It was not possible to scrape the third largest platform due to the blockade of the content of the page against web scrapers. HSL negotiated an agreement with this portal nonetheless, so that the data for Hesse was available and could be evaluated for the purposes of this use case. Initially to get a good coverage of the German real estate market in general, it was decided to scrape both large platforms. It turned out that both portals overlap significantly, and therefore the second portal was used only as a backup. An API was utilized here to experimentally explore this data collection method. It is important to mention that these large portals belong to the same owner, but structure and administration are individual on local level. This is due to the fact that they were direct competitors at the beginning and after the unification of the companies it was decided to keep both portals running independently to guarantee experience and preferences to the customers of both portals. These two larger portals, Portal 2 and Portal 7, are extracted with Scrapy and via API (section “Programming, production of software”). A third portal with a major focus on urban regions and a limitation to new constructions is Portal 5, for which a dedicated Scraper in R was developed.

Overall, in Portal 7 and Portal 2, the given information regarding different attributes of an offer for apartment is quite similar. However, there are some huge differences in the format in which the variables are presented. Portal 7 comes up with a standardized format, which is way more straightforward to process. It is not still free of typing and plausibility errors, but the structure of the data offers easier access. On the other hand, Portal 2 gives the users more freedom when inserting a new offer by allowing text descriptions in numerical variables. From the point of view of data analysis, this is unstructured and requires a lot of more effort to process the data; an example of this problem is the required deposit of an offer for an apartment. Portal 5 has a focus on new buildings and residential property. With a standardized structured website it makes it easier to identify essential HTML tags. Nevertheless, continuous modifications in the web data source require frequent adjustments of the web scraper.

The web scrapers have been running weekly since April 2022 with only few failures, so there is data from mid-2022 available.

The two larger portals do not have a regional focus: SSI-BBB extracts the data for the states Berlin, Brandenburg, and Hesse. The data for offers in Hesse is passed on to HSL in an unprocessed form. Due to performance reasons recurring offers were not collected, so original offers always are part of the analysis.

Duration of Online Listings in a Major Real Estate 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 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 visualizations (Figure 5, Figure 6) 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. The graphs illustrate histograms 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

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 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.

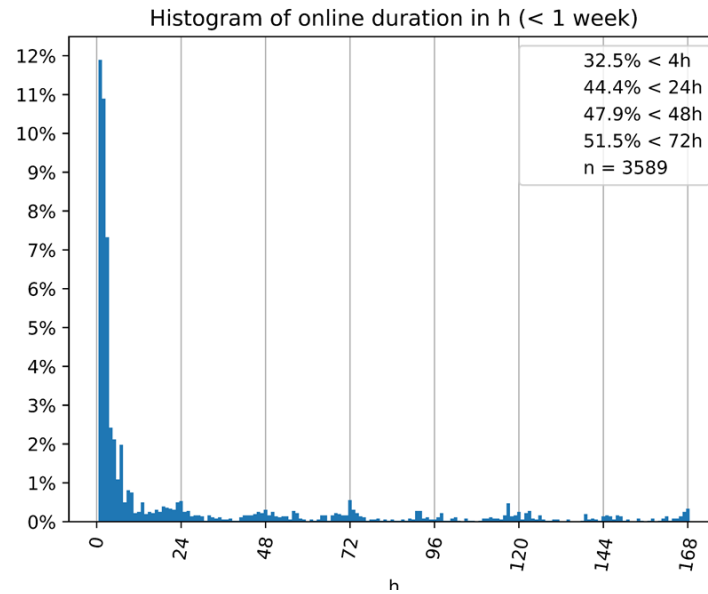


Figure 5. Histogram of online duration of Berlin's rental apartments in h (<1 week)⁷

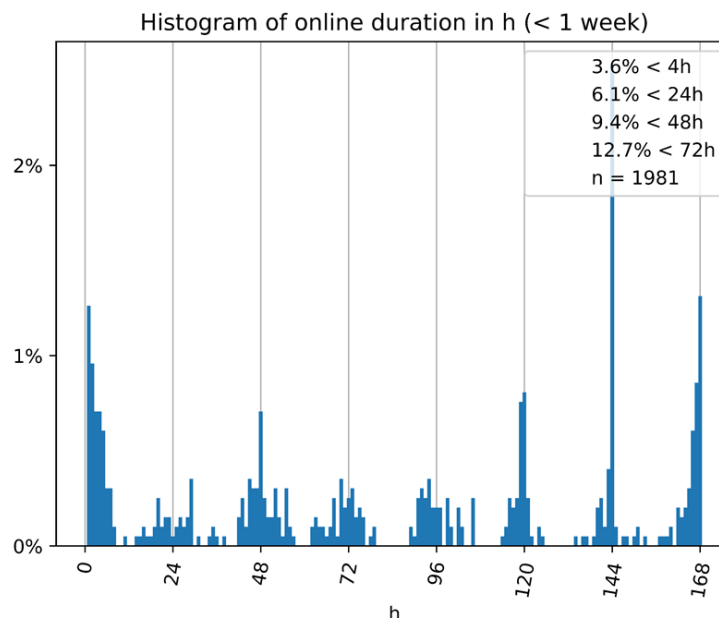


Figure 6. Histogram of online duration of Brandenburg's houses for sale in h (<1 week)⁸

⁷ Histogram of recorded IDs (listings) deleted during the study period, categorised by hours of their duration online. This diagram 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 3589. This representation covers only the first week (168 hours) of the observation period.

⁸ 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 1981. This representation covers only the first week (168 hours) of the observation period.

4.5. Finland

Oikotie data is collected monthly and started in 2019. So far, the agreement has been discussed and renewed every year. Statistics Finland each year buys the data from the provider with a negotiated price for a coming year.

Data are gathered in csv files via an automatic interface search. In case of data problems, Statistics Finland can contact the provider's specific help email for fast problem solving. Fortunately, issues have not occurred often. For example, once there was exceptionally small number of observations in a data set. Within one to two days, new fixed data sets were received.

The data files contain all real estate offers available each month on the provider's website. They contain offers that have been listed or delisted on the month in question, and offers that were already listed before and were still online. The numbers contain both active and inactive offers.

There are about 70 variables in each sale and rent datasets. Both of them contain the same variables, except for a few specific ones regarding rental house market only, like whether a dwelling is rented furnished or unfurnished, and if it's rented for long-term or short-term period. The datasets also contain identifiers for each offer, as well as timestamps of the advertisement announcement and date of its removal from the website. All mandatory variables within this project's scope are well covered.

Most variables describe properties of the apartments and buildings and are similar to those collected from different portals in other countries. Locational variables are very wide, providing information on country, street, postal code, coordinates, city and district.

4.6. France

The partnership between Insee and Seloger/MeilleursAgents established regular data provision at the point where the use of this data was considered mature enough to be integrated in a production-like process. By this time, the partner has provided Insee with several datasets, some of which time-wise overlapping, for the sake of the evaluation of the possible use of these data. The last provision included data ranging from 2018 up to 2023. As agreed in the partnership, the scope of the data is restricted to web ads for housing rental. The table below shows the volume of ads per year. The drop in number for 2023 should be regarded rather as the result of delayed information resulting in a non-complete picture for this specific year, rather than an actual drop in offers.

Table 10. Number of advertisements per year

| Time period | Offers for rent |
|-------------|-----------------|
| 2018 | 969306 |
| 2019 | 823979 |
| 2020 | 815518 |
| 2021 | 979452 |
| 2022 | 1077215 |
| 2023 | 476606 |

The partner has provided Insee with already cleaned and partially treated data, in particular with regards to duplicate handling. As some of the data transmissions included raw data, it is also possible for Insee to apply alternative data treatments and processing, so as to assess the sensitivity of the results to the partner's choices.

5. Data processing

The data processing step consisted of four sub-processes:

- Data cleaning and editing
- Integration with the territorial division register
- Duplicates detection
- Classification of mandatory variables.

To ensure comparability between indicators calculated by partners, all mandatory variables were classified appropriately. The classification categories were discussed between partners, preceded by a review of the content of variables and their distributions.

Table 11. Classification of mandatory variables

| Categorized variable | Categories | | Description |
|----------------------|---|---|---|
| | Code | Label | |
| building_type | 01 02 03 10 | Detached house Semi-detached or terraced house Other Apartment | Type of real estate |
| NUTS3_id | European nomenclature of territorial units for statistics (NUTS): https://ec.europa.eu/eurostat/web/nuts/background | | |
| NUTS3 | Name of the NUTS III region | | |
| LAU | LAUs are the building blocks of the NUTS, and comprise the municipalities and communes of the European Union. In the datasets, it refer to the lowest possible level of territorial division: https://ec.europa.eu/eurostat/web/nuts/local-administrative-units | | Local administrative unit (in some cases equals city) |
| offer_transaction | 01 02 | Rent Sale | Offer for rent or for sale |
| offer_rooms | 01 02 03 04 05 | 1 room 2 rooms 3 rooms 4 rooms 5 and more | Number of rooms in the offered object |
| offer_price | Only offers „for rent“ 01 02 03 04 | Up to 350 EUR 350.1 - 600 EUR 600.1 - 950 EUR 950.1 and more EUR | Offer price |
| price_sell_meter | Only offers „for sale“ 01 02 03 04 05 06 07 | up to 800 EUR/m2 800.1 – 1400 EUR/m2 1400.1 - 2000 EUR/m2 2000.1 – 2600 EUR/m2 2600.1 – 3200 EUR/m2 3200.1 – 3800 EUR/m2 3800.1 and more EUR/m2 | Offer price per square meter |
| offer_surface | Apartments: 01 02 | up to 40 m2 40.1-60 | Surface area of the offered object |



| | | | |
|--|---------|-------------------|--|
| | 03 | 60.1-80 | |
| | 04 | 80.1-100 | |
| | 05 | 100.1 and more | |
| | Houses: | | |
| | 01 | up to 100 m2 | |
| | 02 | 100.1 – 150 m2 | |
| | 03 | 150.1 – 200 m2 | |
| | 04 | 200.1 and more m2 | |

Every organization has encountered a problem with data redundancy in the acquired data sets. Possible ways that were tested to monitor the duplicates were at first the most basic - to use an object id to detect duplicates within a given portal. Then, the offers grouped by various type of information (market type, location, etc.) were assessed. However, detection sometimes resulted only in finding potential duplicates across portals. The experts were not able to decide whether they were actual duplicates or not. Portals belonging to the same capital group may share even more than one third of the offers posted on both platforms. And even identifying a set of possible duplicates may in fact indicate not duplicate offers, but a group of apartments in one construction investment that have the same set of values in all variables, or even the same descriptions. Moreover, in the case of buildings that are not newly constructed, the offer may be published by several realtors, with different id and descriptions. The examples of above mentioned cases are presented in country-specific subchapters.

5.1. Bulgaria

The data processing was done in two steps. First step, immediately after scraping, was dedicated to Data cleaning and editing, Integration with the territorial division register and Classification of mandatory variables. In some cases, when errors were found (within variables or changes in data source structures), an adjustment has been done to the processing code, so this step was repeated for some of the sources. The second step was dedicated to Duplicates detection. This step was done together with data aggregation and calculating of final indicators.

Data cleaning and editing

The variables from the structured text data of the source offers were extracted in two ways.

Extraction of predefined key words found in the text, such as house, apartment, sale, rent, BGN, EUR and others:

```
dfn['offer_transaction'] =
    dfne['Title'].str.strip().str.extract(r'(продажба|под наем)')
dfn['offer_transaction'] =
    dfn['offer_transaction'].replace('под наем', 'rent')
dfn['offer_transaction'] =
    dfn['offer_transaction'].replace('продажба', 'sale')
```

Figure 7. Extraction of predefined key words

Extraction of information by splitting the text by strings (such as *price per m2*, *floor*, *air-conditioner* or others) and then splitting again or perform a regular expression to process the remaining string in order to achieve expected result:

```
dfn['offer_surface'] =
    dfne['txt-1'].str.split('Квадратура: ').str[1].astype(str).str.split(
        '\s|м2|Цена|Етаж|Допълнителна информация|ТЕЦ|ГАЗ|Публикувана|Особености|
        Допълнителна информация').str[0].replace(' ', np.nan).astype('float')
```

Figure 8. Extraction of information by splitting the text

Although there were filters for offers to choose only apartments and houses at scraping stage, additional control with cleaning process was added. The data was cleaned from offers that were out of the scope, thus presents offers of land, stores, garages or others.

Integration with the register of the territorial division

After Data cleaning and editing process the addresses and NUTS3 names were completed by matching with their NUTS and LAU codes.

Duplicates detection

Deduplication was carried on during the calculation of quality indicators and right before calculation of statistical indicators. It was done on the each scraped dataset. The data was deduplicated by seven variables describing – *NUTS code, transaction type, building type, surface area of the offered object, number of rooms in the offered object, location obtained and month when the offer was valid*. The first occurrence was kept, the others were removed.

```
qil = dfv.loc[(dfv['month'] == period)].copy().shape[0]
dfv = dfv.drop_duplicates([
    'nuts_code',
    'offer_transaction',
    'building_type',
    'offer_surface',
    'offer_rooms',
    'location',
    'month'
]).reset_index()
dfvm = dfv.loc[
    (dfv['month'] == period) & (
        dfv['building_type'].isin(['apartments', 'houses'])
    )
].copy()
```

Figure 9. Script example for basic deduplication process

Classification of mandatory variables

Classification of mandatory variables was done simultaneously during the Data cleaning and editing process in accordance with Table 11 to ensure comparability with other partners.

5.2. Poland

Data cleaning and editing

The aim of this step was to prepare the dataset for analysis, which included: removing unnecessary (e.g. white spaces) or wrong inserted characters, converting separators, changing data types and replacing descriptive values (such as “ground floor” or “ask for price”) into numerical categories, dealing with missing values and outliers. All replacement rules, applicable for the four data sources, selected by Statistics Poland, are indicated in the Table 12.

Table 12. Cleaning and editing rules

| Before editing | After editing |
|-----------------------|---------------|
| m2 | |
| ask for price | NULL |
| low ground floor | 0 |
| ground floor | 0 |
| residential basement | 0 |
| attic | 1000 |
| more than 8 [rooms] | 8 |
| more than 30 [floors] | 30 |
| comma | dot |

Missing values have been replaced with other values, most often 0, but in the case of the floor variable, where 0 is reserved for the ground floor, to 9999. Due to the large number of missing values for new apartments, further analyses focused on the secondary and rental markets.

During this task outliers were also removed. After consultations with the subject-matter statisticians, it was decided that the thresholds would be set at the 1st and 99th percentiles of the price per m² and per 1000 m² on the variable *surface area*. For instance, in the case of variable *price*, after completion of the cleaning process, the maximum value of price for sale decreased from around 4.5 million EUR to 2 million EUR.

Exchange rates from Polish currency (PLN) into Euro were acquired via API from the website of the National Bank of Poland⁹ in JSON formats as the date of data acquisition.

Integration with the register of the territorial division

Next step was to prepare the dataset with the standardized information of localization of properties. In most cases the web-scraped datasets contained information of localization divided into three elements: address1, address2, address3. Basically, it should be three levels of administrative division: “gmina” (municipality, LAU2), “powiat” (county and city county, LAU1) and “voivodeship” (region, NUTS2), but in practice the order of these elements and their accuracy can be different (for example one can refer to the locality, city district, street or contain error). Therefore, the first task was to clean and edit each element of localization variable. Then, the unique key of these names, referring to the administrative units had to be created. The key had to be created in such a way as to enable a linkage the dataset with the National Official Register of the Territorial Division of the Country (TERYT). TERYT register contains, *inter alia*, systems of: identifiers and names of units of territorial division (TERC) and identifiers and names of localities (SIMC). In both SIMC and TERC databases the unique identifiers had to be prepared, based on the names of units

⁹ <http://api.nbp.pl/en.html>

of territorial division and localities. The process of linking two datasets was deterministic and iterative (Figure 10).

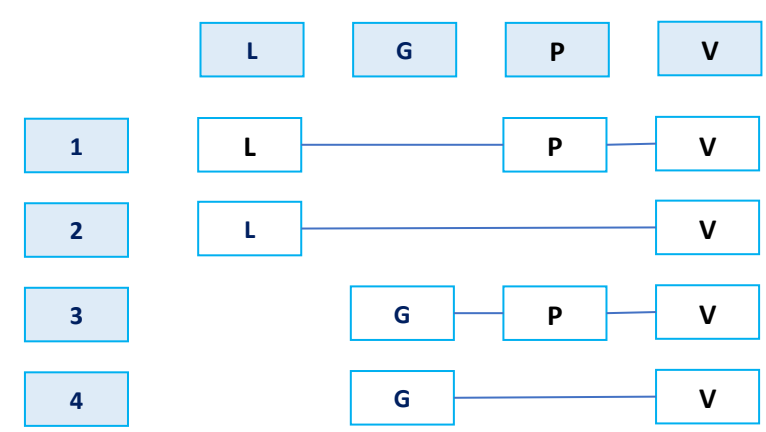


Figure 10. The iterative process of linkage dataset with the register of the territorial division (keys)¹⁰

At the first and second linking stage, all duplicates were dropped, due to the fact that in Poland there are the cases of the same name of locality in the same powiat, but in a different gmina. Therefore, it is impossible to predict in which gmina such real estate is located. The next linking stage was different, as duplicates were dropped, except for the first occurrence. In Poland, it is common that the urban gmina and surrounding rural gmina have the same name. They have got a different TERYT identifier, but the key which is based only on the names in such cases is the same. Therefore, it was decided to remove from the TERC database the raw of the rural gmina, as the probability that the real estate is located right there is low. On the basis of the exploratory phase of the dataset, it can be assumed that when the apartment is located in the village the real estate’s owner indicate it in the address offer and rather insert the exact name of locality. The last linkage stage was most flexible, but helped to matched rows which had a correct name of gmina and voivodeship, but wrong name of powiat (for example the same name as gmina).

As a result, 99.88% of the records were paired correctly. At the end, the dataset was merged with the nomenclature of territorial units for statistics (NUTS) to aggregate data into NUTS3 level. The dataset contains also geometry of the territorial units, which enables visualization the results on a map.

Duplicates detection

The problem of duplicates was approached after standardizing certain information based on the data contained in the original set from gratka.pl scraped in January 2024. The analysis was divided into three steps.

Step 1: First, the analysis of available information and fields selection was performed based on which duplicates were identified. The fields used were: *Surface area of the offered object*, *Number of rooms in the offered object*, *Floor number on which the object is situated*, and *Location obtained* which was divided into three separate variables defining: voivodeship, powiat, gmina (respectively NUTS2, LAU1, LAU2 according to Nomenclature of Territorial Units for Statistics).

¹⁰ L = Locality; G = Gmina (LAU2); P = Powiat (LAU1); V = Voivodeship (NUTS2)

Unfortunately in the scraped information from portals is no precise localization data. Even when the street address is provided it often refers to the address of the brokers office. That is why this data cannot be used, and it forces to refer only to higher levels of territorial classification in the process of aggregation. Nevertheless obtained addresses are used in the deduplication process to combine offer into groups on the later stage.

Step 2: It was decided that all offers that have in common all the above mentioned variables were going to be marked as a duplicate. The floating point data were compared to integer precision.

Additionally due to the relatively large location areas of the offer, a field with description of the offers was used ('Description of the offer'). This field allows to identify the same offers in the subsets determined by the assumptions of the first step.

Then, to perform deduplication, libraries based on Python software were analyzed to identify duplicate records. After initial analysis, three libraries were selected:

1. pandas.DataFrame.duplicated – description: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.duplicated.html>
2. Dedupe – description: <https://github.com/dedupeio/dedupe>
3. Recordlinkage – description: <https://recordlinkage.readthedocs.io/en/latest/about.html>

Step 3: After testing, the dedupe library was initially employed. However, during this process, a challenge in identifying duplicate records emerged. Numerous announcements pertained to substantial investments, resulting in individual real estate agencies producing nearly identical descriptions. This generated a lot of information noise hindered the development of an effective model for detecting duplicate units. While the program successfully identified duplicate records within a single real estate agency, it struggled to do so across different offices.

Hence, it was opted to perform deduplication among offers from the new-construction market through two sub-processes. First, individual records and duplicates were identified within the investment by analyzing their descriptions and the acquired approximate location data. Next, a 'unique' selection (in the context of fuzzy logic) was applied to descriptions and conducted the deduplication process between these distinct descriptions. The first part was based on the Recordlinkage library, the second one was based on the Dedupe library. On the base of the analysed duplicated records a set of examples was prepared.

Table 13. Offers with same values, except floor number

| id | price sell | address | descr | surface | rooms | floor | building year |
|----|------------|----------------------------|---|---------|-------|-------|---------------|
| 35 | 449000 | Jelenia Góra, dolnośląskie | Bazując na naszym wieloletnim doświadczeniu stworzyliśmy wyjątkowe miejsce, łącząc udogodnienia (...) | 36.92 | 2 | 4 | |
| 86 | 499000 | Jelenia Góra, dolnośląskie | Bazując na naszym wieloletnim doświadczeniu stworzyliśmy wyjątkowe miejsce, łącząc udogodnienia (...) | 36.92 | 2 | | |

Table 14. Offers with same values in all variables

| id | price sell | address | descr | surface | rooms | floor | building year |
|-----|------------|--|---|---------|-------|-------|---------------|
| 106 | 347340 | Jelenia Góra, Cieplice Śląskie-Zdrój, dolnośląskie | OFERTA ZAREZERWOWANA W TRAKCIE REALIZACJI Inwestycja Panorama (...) | 57.89 | 2 | | 2020 |
| 109 | 335762 | Jelenia Góra, Cieplice Śląskie-Zdrój, dolnośląskie | OFERTA ZABLOKOWANA W TRAKCIE REALIZACJI Inwestycja Panorama (...) | 57.89 | 2 | | 2020 |
| 111 | 335762 | Jelenia Góra, Cieplice Śląskie-Zdrój, dolnośląskie | OFERTA ZABLOKOWANA W TRAKCIE REALIZACJI Inwestycja Panorama (...) | 57.89 | 2 | | 2020 |

Only by the description of the offer it could be identified that the offers are not duplicates.

The offers' translated description (id 109 and 111):

"OFFER RESERVED DURING CONSTRUCTION Investment Panorama Śnieżka Building 5, (...) The investment is scheduled to be completed by the end of June 2021."

"OFFER RESERVED DURING CONSTRUCTION Investment Panorama Śnieżka Building 6 (...) The investment is scheduled to be completed by the end of 2020."

Table 15. offers with exactly the same values in all variables

| id | price sell | address | descr | surface | rooms | floor | building year |
|-----|------------|--------------------------------------|---|---------|-------|-------|---------------|
| 203 | 395000 | Łaziska, bolesławiecki, dolnośląskie | Polnoc Nieruchomosci O/BOLESŁAWIEC oferuje do sprzedaży nowo budowane (...) | 58.81 | 3 | 1 | 2023 |
| 204 | 395000 | Łaziska, bolesławiecki, dolnośląskie | Polnoc Nieruchomosci O/BOLESŁAWIEC oferuje do sprzedaży nowo budowane (...) | 58.81 | 3 | 1 | 2023 |

The offers' translated description:

"Północ Nieruchomości O/BOLESŁAWIEC offers for sale newly built, rent-free, 3-room apartments in developer standard near Bolesławiec. DETAILS OF THE OFFER: - The investment consists of five two-unit buildings with gardens located at the back of the buildings and parking spaces (two spaces for each apartment) in front of the buildings. - Apartments located on the ground floor with a surface area of 55.11 m2 have a garden. - Apartments located on the first floor with a surface area of 58.81 m2"

The use of the plural in the description suggests that these may be separate offers.

Table 16. group of offers that may contain duplicate information

| id | price sell | address | descr | surface | rooms | floor | building year |
|-----|------------|-------------------------------|--|---------|-------|-------|---------------|
| 296 | NULL | Sienna, kłodzki, dolnośląskie | Lokal zarezerwowany. Sprawdź dostępność innych w ofercie dewelopera. (...) | 24.7 | 1 | 1 | |
| 347 | NULL | Sienna, kłodzki, dolnośląskie | Lokal zarezerwowany. Sprawdź dostępność innych w ofercie dewelopera. (...) | 24.7 | 1 | | |
| 349 | NULL | Sienna, kłodzki, dolnośląskie | Lokal zarezerwowany. Sprawdź dostępność innych w ofercie dewelopera. (...) | 24.7 | 1 | 2 | |
| 374 | NULL | Sienna, kłodzki, dolnośląskie | Lokal zarezerwowany. Sprawdź dostępność innych w ofercie dewelopera. (...) | 24.7 | 1 | 3 | |

It is unable to confirm it based on the data contained in the database. The unit with index 347 have missing information in the *floor* field. The remaining information is exactly the same as for other units.

Table 17. Offers with same values except Price

| id | price sell | address | descr | surface | rooms | floor | building year |
|-----|------------|---|--|---------|-------|-------|---------------|
| 397 | 435393 | Świeradów-Zdrój, lubański, dolnośląskie | APARTAMENTY BRAMA DO NATURY to wyjątkowa inwestycja składająca się z dwóch (...) | 42.86 | 2 | | 2023 |
| 407 | 453967 | Świeradów-Zdrój, lubański, dolnośląskie | APARTAMENTY BRAMA DO NATURY to wyjątkowa inwestycja składająca się z dwóch (...) | 42.86 | 2 | | 2023 |

Table 18. Offers with very similar values

| id | price sell | address | descr | surface | rooms | floor | building year |
|------|------------|----------------------------------|---|---------|-------|-------|---------------|
| 1327 | 741846 | Wrocław, Swojczyce, dolnośląskie | kupujący nie płaci prowizji, ani PCC* stan deweloperski (...) | 71.40 | 3 | 2 | 2022 |
| 1323 | 742469 | Wrocław, Swojczyce, dolnośląskie | kupujący nie płaci prowizji, ani PCC* stan deweloperski (...) | 71.46 | 3 | 2 | 2022 |

This is an example of offers for which decision-making processes based on fuzzy logic may not provide satisfactory results. These offers may not be duplicates. However, the difference in price of 0.08 percent and in the surface of 0.06 meters may not be sufficient for correct interpretation.

Classification of mandatory variables

To ensure comparability with other partners the classification of mandatory variables was done in accordance with Table 11.

5.3. Germany – HSL

Data cleaning

Data from all sources has been processed in order to comply with the requirements agreed with other partners.

A set of rules was employed to clean data, e.g. to turn verbal descriptions into numbers (“ground floor”, “First floor”). In Germany, it is still common in advertisements to indicate and count small rooms as a “half room”. According to an outdated norm, “half rooms” indicated a room size between, 6 and 10 m². German official statistics concepts do not distinguish rooms by their size: all rooms with 6 m² and more count as “a room”. In order to make concepts comparable, the room number has been rounded to the next larger number whenever an advertisement used the concept of a half room.

To merge official NUTS information and to aggregate and compare aggregated offers to official data, address information (especially city names) needed comprehensive data cleaning. Portal data typically contained many variants of city names, e.g. including/solely suburb names, and besides spelling variants contained many spelling errors – if the platform does not check for valid input.

For some characteristics (i.e. *offer_price*, *price_sell_meter*, *offer_surface*), the classification scheme described in Table 11 leads to the case that many or all of the offers for Hesse lie within one of the highest categories.

Besides data cleaning, deduplication within as well as between portals is a major challenge and yet there is no single or simple solution to the problem of undetected duplicates.

Duplicates detection¹¹

Duplicates within a one portal and one monthly data set can occur due to the weekly scraping, when the same offer is scraped each week. Portal's unique ID has been used to deduplicate the data within a portal. Only one offer per ID was kept and since there may be updates/corrections to an advertisement, it has been decided to keep the last (newest) offer per ID.

Additionally, there may be duplicated offers with different IDs, when there is another new advertisement for the very same object that uses a new unique ID. These duplicates could only be identified by comparing other key characteristics like complete address (postal code, street name, house number), location within building (floor number), size in m², number of rooms, price).

Depending on building type, identification of duplicates with different IDs but many identical other attributes is not trivial. For example, typically there is not more than one newly constructed detached house at one specific address (street name and house number), but there may be more than one advertised apartment within a larger newly constructed building at a specific address. Typically, all of these apartments may have identical or very similar characteristics (price or surface area information is not identical but very similar, other amenities may be very similar or identical).

Deciding on duplicates gets even more challenging when there is missing data (e.g. missing information on the floor number of one or more advertised apartments). Deduplication – in fact even identifying potential duplicates - becomes very hard, if address information is missing.

When it comes to deduplication between portals the process has to rely on characteristics of the offer or object itself as different platforms use different IDs to identify advertisements uniquely. First of all, this is address information (postal code, street name, house number) as well as location within the building (floor number). Additionally, size (number of rooms, surface in m²) and price is used to identify duplicates.

A situation may occur when two portals (e.g. "OM" and "IE") share the same set of IDs – they are owned by the same company – at least for some advertisements. One part of the portal's ID points to the portal itself, say "OM-#####" or "IE-#####", and advertisements with an ID with the part "OM" also appear within the portal "IE" as well as the other way round. But this is not the case for many advertisements of both of these platforms. However, the two portals do not share the same ID part after the portal identifier, this means "OM-12345" from portal "OM" is not the same advertisement as "IE-12345" – given such a match of "12345" would exist in the dataset.

Table 19 shows an example of a potential duplicate within one data source where the linking key (formed by city, postal code, street name, house number, floor number, surface size) matches for two distinct IDs, but selling price differs by a larger amount. Since other characteristics match, this looks like a duplicate where one of the prices given is erroneous – which might have been the reason for another additional advertisement of the same object.

Table 19. Example of a duplicate offer within one data source

| address | ad_id | surface area | floor | rooms | price |
|---|---------|--------------|-------|-------|--------|
| Bad Emstal – 34308 – wolfhager -- 2 - 155 | 282LV56 | 155.0 | 2 | 4 | 449000 |
| Bad Emstal – 34308 – wolfhager -- 2 - 155 | 28GXY56 | 155.0 | 2 | 4 | 299900 |

¹¹ More information on deduplication process and examples are presented in Deliverable 3.6 Report on methods and feasibility to track construction activities based on real estate web portals.

Again, identifying duplicate houses is easier – given their address information is complete and standardized between sources in an extensive data preparation step - than identifying duplicate apartments at the same address. Especially in larger construction projects, different apartments share very similar characteristics (especially size, price). Having in mind that underlying concepts (say for specifying a price or a surface size) may be different between portals, characteristics do not have to match exactly. Finally, besides real differences in some characteristics, say price, information needed for deduplication may be present in one portal but is missing in the other.

Table 20 shows the amount of missing information for all gathered advertisements for newly constructed objects within 5 portals for 2023. Especially for one of the largest portals, information on street name and house number is missing in many of the advertisements. This makes it very hard to identify duplicates with the other sources.

Table 20. Amount of missing information by source and variable for newly constructed offers in 2023

| Variable name | Portal 1 | Portal 2 | Portal 3 | Portal 4 | Portal 5 |
|----------------------------------|----------|----------|----------|----------|----------|
| Number of missing information | | | | | |
| city | 0 | 0 | 0 | 0 | 0 |
| postal code | 0 | 0 | 0 | 0 | 0 |
| street name | 0 | 3424 | 0 | 0 | 0 |
| house number | 0 | 3424 | 0 | 0 | 0 |
| floor number | 1763 | 1993 | 27 | 30 | 150 |
| number of rooms | 74 | 42 | 5 | 5 | 0 |
| surface m ² | 11 | 7 | 5 | 5 | 0 |
| price (rent / sale) | 57 | 3106 | 0 | 1 | 0 |
| construction year | 573 | 99 | 69 | 72 | 213 |
| building type | 502 | 0 | 6 | 6 | 37 |
| type of offer | 57 | 3106 | 0 | 1 | 0 |
| Number of overall offers scraped | | | | | |
| scraped offers | 5135 | 6474 | 152 | 156 | 372 |

5.4. Germany – SSI-BBB

Data cleaning and editing

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

The data obtained from scraping required a comprehensive data cleaning process to be usable for analysis. This process applied not only to data obtained through screenscraping but also to data acquired via APIs. For example, string elements needed 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 was necessary to convert the entry "no information" into missing values.

The portals differed heavily when it came to the characteristics that were present in the advertisements. Whilst in some portals there was a huge number of very special characteristics present, some ads even lacked information on pre-agreed mandatory variables.

Table 21. Number of observations available per indicator (Portal 2 and Portal 5) collected from April 2022 to January 2023

| ID | Indicator | Berlin | | Brandenburg | |
|----|------------------------------------|--------|------|-------------|------|
| | | Sale | Rent | Sale | Rent |
| 1 | Number of offers | 2875 | 2125 | 362 | 782 |
| 2 | Average price | 2856 | 2123 | 357 | 780 |
| 2 | Average surface area in m2 | 2868 | 2122 | 360 | 779 |
| 5 | Share of offers by number of rooms | 2847 | 2121 | 354 | 780 |

Outliers pose a particular challenge in that their removal or the criteria used to determine them have a massive influence on the results, especially the arithmetic mean. Instead of using a threshold for the outliers, it would make more sense to assess the content of the properties. In fact, it is difficult to research more detailed information about individual offers afterwards.

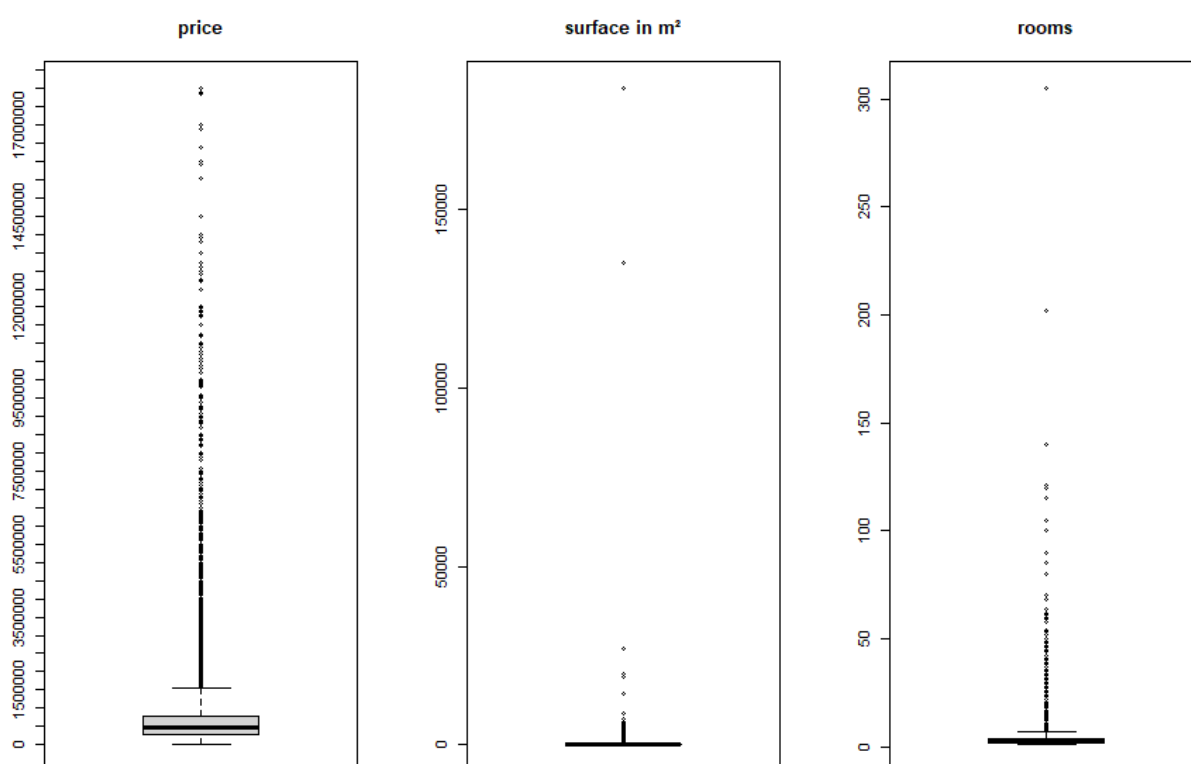


Figure 11. Outliers in objects for sale in Berlin

Localization

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

The fine-grained local information was one of the key elements for deduplication and can be the basis for very detailed maps. Unfortunately, deduplication was still an issue, and much more accurate address data would be necessary for that purpose.

Duplicates detection

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 offers (Figure 12) and new constructions (Figure 13).

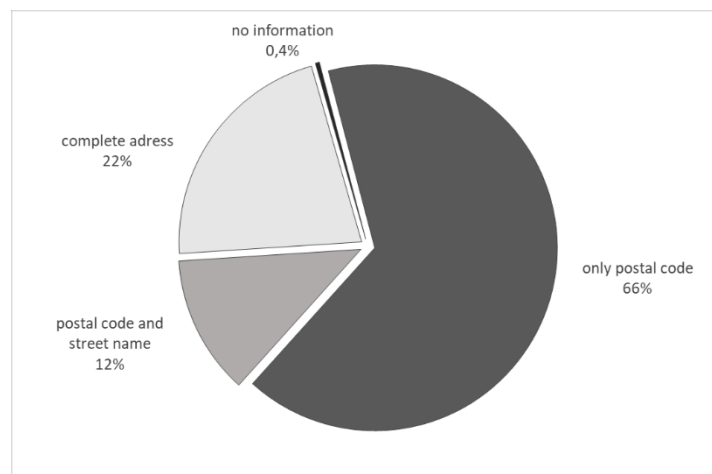


Figure 12. Proportion of scraped data for Sale or Rent on Portal 2 from April 2022 to July 2024, Categorised by Completeness of Address Information

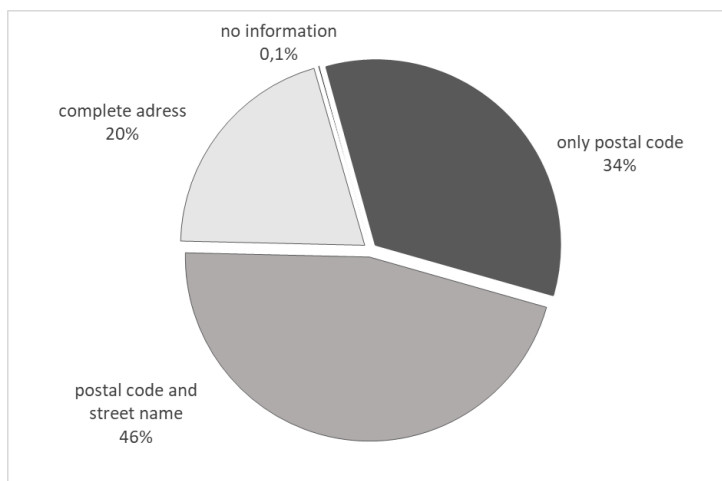


Figure 13. Proportion of Newly Listed Properties (Built in 2023) for Sale or Rent on Portal 2 from April 2022 to July 2024, Categorised by Completeness of Address Information

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:


| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|-------------|------------------|--------|---------------------------------|-------------|--------------|--------|---------------|------------------------|----------------|-----------------|------------|------------------|---------------------|-------------------|--------------------|-------------------|---------------|---|
| ad_provider | data_scraped | ad_id | location | postal_code | street | number | federal_state | building_type | offer_transact | offer_floor | offer_room | offer_price_sale | offer_price_rent | offer_price_buyer | offer_price_seller | offer_price_meter | offer_surface | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| Portal 5 | 25.10.2023 00:00 | 241264 | Birkenstraße 12a, 10559 Berlin | 10559 | Birkenstraße | 12a | Berlin | Etagenwohnung | sale | NA | 1 | 265000 | | | | | 449.85 | |
| Portal 2 | 22.04.2023 07:50 | | Birkenstraße 12a, 10559, Berlin | 10559 | Birkenstraße | 12a | Berlin | zinshaus_renditeobjekt | sale | 2. Geschoss | 1 | 267000 | 933.566.433.566.433 | | | | 28.60 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| Portal 5 | 16.07.2024 00:00 | 268529 | Birkenstraße 12a, 10559 Berlin | 10559 | Birkenstraße | 12a | Berlin | Etagenwohnung | sale | 4.OG | 1 | 282900 | 10404.56 | | | | 27.19 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |
| 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 | |

Figure 14. Example of a construction project in Berlin, advertised on both Portal 5 and Portal 2

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 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.





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 5

Wohneinheiten

1 Zi. 2 Zi. 3 Zi.




| Bild | Preis | Fläche | Zimmer |
|------|-----------|----------|----------|
| | 259.900 € | 28,03 m² | 1 Zimmer |
| | 262.000 € | 28,06 m² | |
| | 265.000 € | 28,02 m² | |
| | 267.000 € | 28,06 m² | |
| | 271.000 € | 27,96 m² | |
| | 273.900 € | 27,19 m² | |
| | 282.900 € | 27,19 m² | |
| | 289.900 € | 27,96 m² | |
| | 292.900 € | 27,19 m² | |
| | 297.000 € | 27,96 m² | |
| | 299.000 € | 27,19 m² | |

Wohneinheiten

1 Zi. 2 Zi. 3 Zi.

| Bild | Preis | Fläche | Zimmer |
|------|-----------|----------|----------|
| | 347.000 € | 38,02 m² | 2 Zimmer |
| | 349.000 € | 38,05 m² | 2 Zimmer |
| | 353.000 € | 38,02 m² | 2 Zimmer |
| | 356.000 € | 38,05 m² | 2 Zimmer |
| | 377.000 € | 38,02 m² | 2 Zimmer |
| | 379.900 € | 38,05 m² | 2 Zimmer |
| | 387.000 € | 38,02 m² | 2 Zimmer |
| | 389.900 € | 38,05 m² | 2 Zimmer |
| | 494.900 € | 58,18 m² | 2 Zimmer |

Portal 5

Dachgeschoss mit Weitblick – 28 m² große 1-Zimmer-Neubauwohnung in Berlin-Moabit – Rohbau fertig

286.500 € 28,60 m² 1 Zimmer

Kaufpreis Wohnfläche ca. Zimmer




1,54% mtl. Finanzierung Gesponsert

Gewerblicher Anbieter

SeG Kapital VV GmbH

Anbieter kontaktieren

017 -- Nr. anzeigen

Dachgeschoss mit Südbalkon – 38 m² große 2-Zimmer-Neubauwohnung in Berlin-Moabit – Rohbau fertig!

374.900 € 38,30 m² 2 Zimmer

Kaufpreis Wohnfläche ca. Zimmer

1,54% mtl. Finanzierung Gesponsert

Gewerblicher Anbieter

SeG Kapital VV GmbH

Anbieter kontaktieren

017 -- Nr. anzeigen

Portal 2

Figure 15. Example of duplicates in different portals

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.

| | 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_cat | offer_transaction_cat | offer_floor | offer_rooms | offer_price_sale | offer_price_rent | 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 |



© 2024 Google

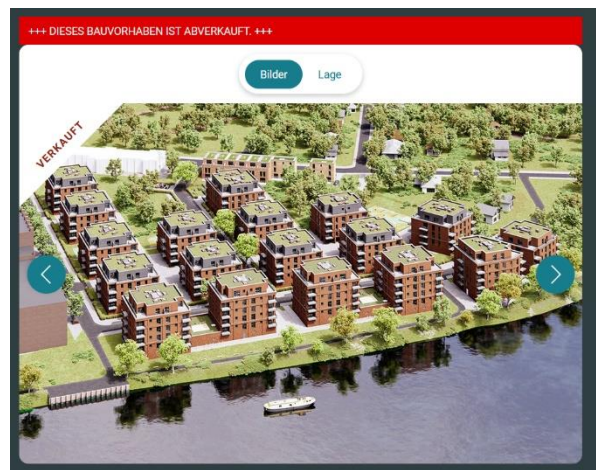


Figure 16. Example of offers in a residential construction project

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
- Wohnstyp: 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

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

1.346 € 56,07 m² 2 Zimmer

Katzenste zzgl. NK Wohnfläche ca. Zimmer

Gewerblicher Anbieter

Margaretenstraße 24 A 10317 Berlin (Lichtenberg)

Auf Karte ansehen →

THE GROUNDS Real Estate Development AG Vivian Buchholz

Anbieter kontaktieren

+49 ... Nr. anzeigen →

Figure 17. Example of previously purchased rental offer

Classification of mandatory variables

To ensure comparability with other partners the classification of mandatory variables was done in accordance with Table 11.

5.5. Finland

Data cleaning and editing

Since the data sets are acquired directly from the provider, the data are already quite structured and clean with key variables categorized. The analysis covered only offers of dwellings located in Finland, thus the ones located abroad were excluded. The number of these offers is small, about 0.2% of overall.

The acquired data correspond to the offer advertisements that have been active online. Even the highest and lowest price demands seem valid, based on dwelling descriptions and other variables, and they reflect the price differentiation that has been trending in the past decade in the Finnish house market.

The indicators were calculated excluding missing values, negative values, and obvious outliers. Those outliers that skewed the NUTS level aggregations were excluded from calculations, after which the exceptional means returned to normal. For rents data, observations with total asking rent of over one million euros were excluded as outliers.

Regarding sales data, very high price ranges occur, for example, in new apartments in the centers of growing cities. On the contrary, the price range can be very low in the outskirts towns in northern and eastern Finland, where population is decreasing, and houses may be in poor condition. Year of construction and condition

also affect asking prices, as well as the type of housing. For example, right-of-residence apartments (intermediate form of owner-occupied housing and rental housing) have a low asking price because they only cost about 15% of usual market price. Applying simple price limits could rule out many valid observations like these, which is undesirable. When tested deleting below the 1st and above the 99th percentile from the whole data, average prices would not change a lot, a few percent at maximum on NUTS3 level.

Oikotie data has some offers that have a large surface area and number of rooms. These observations either contain many buildings on the same lot, many apartments, or a whole block of apartments. These observations where many apartments or buildings are listed in one advertisement, are considered as one offer, since all the objects will be sold or rented at once. Number of such observations seems to be quite small, about a few dozen per month.

Localization

Regional classification was added to the data according to postal codes. Incorrect or missing postal codes were fixed manually as much as possible, using the house address and city variables.

First, incorrect postal codes were found by joining regional classification to the data. Then, observations that did not get classified were searched by their addresses and postal codes in an online service available by Finnish post office. After that, the regional classification was joined again, with almost all observations correctly classified. Only a few dozen observations have an incorrect or missing postal code each month. These not-merged were not used in the analysis for this project.

The NUTS classification from year 2020 was merged in the data by municipality codes. Municipality classification was first merged by the postal codes. After that, NUTS classes were merged by these municipalities.

Duplicates detection

Deduplication is part of the monthly production process of experimental statistics on rent and sale offers. By default, the monthly data sets do not have any duplicates to be edited out as it is controlled by the data provider. However, a small number of observations may be included as duplicates in both data sets in consecutive months, so these are deleted with the help of time period parameters. Each offer occurs only once per month.

Classification of mandatory variables

The mandatory variables were classified as agreed between UC1 partners. The most challenging category for Finland was the variable *offer_floor*, since there are only few buildings with more than 10 floors in the whole country. So, the number of observations in the highest floor category remained in less than 1% of all observations.

There are only a few dozen offers per month which have building type classified as 'Others', and based on apartment descriptions, these mostly seem like houses or apartments. These other buildings were included in the building type category with houses.

5.6. France

Data cleaning and editing

The data, for most of it, have been treated and cleansed by the data provider. Some winsorization techniques have been applied on a case-by-case basis, as it may happen that some outliers still remain in the data. It is foreseen to work (also in coordination with the partner) on techniques using NLP processing for the textual pieces of information extracted from the ads, to offer a possibility to correct the data. In particular, it turned out that close look to the data and discussion with the partner had shed light on some biases in the database (e.g. regarding localization). Indeed, advertising strategies for realtors may include fuzzy localization, aiming at displaying some of the ads as belonging to a more attractive area than it is actually. A clever use of textual content may help in addressing this issue.

Duplicates have been treated using the different features of the ads, such as descriptive variables, but also textual elements. Applying a set of rules regarding distance between two ads helps in identifying duplicates. A rule-of-thumb is used for determining thresholds below which two ads will be considered as targeting the same dwelling at the same time. Different scenarios and thresholds have been tested so as to estimate the sensitivity of the indicators to the duplication treatment.

For the rest, the data suffer from a significant rate of missing values. For the time being, those missing values have been partially treated with imputation, but there is no systematic procedure to produce a fully imputed dataset. The volume of ads available makes it possible to proceed with data analysis without a strong imputation effort.

Localization

Most of the time, postal codes were provided within the ads, to specify the municipality (LAU 2) to which the dwellings belong. Moreover, the data provider offers a first localization below the municipality level using his own analytical procedure of data processing. However, Insee used the postal codes to compute a proper municipality code corresponding to the official spatial location (for instance, postal codes may cover several municipalities, and do not refer to a solid and shared partitioning of the territory). The procedure currently implemented relies on processes that are usually used by Insee for geolocalization of external data, such as administrative data.

Duplicates detection

Data provided by the company are deemed to be treated for duplication, which was challenged through alternative methods. Mainly it rests on textual description constituting the advertisement; these descriptions were compared two by two what restricted the field of comparison between ads in terms of localization, surface area, types of dwellings. A first rough approach was to compare the characters of the ads and consider as duplicates those with less than 10% different characters; other approaches relying on Levenshtein distance were also tested.

Validation and quality assessment

Insee has conducted quality assessment through a comparison between the volume of ads and the yearly share of residential moves among the private rental sector. In particular, thanks to the Fidéli database¹², which compiles administrative information regarding the localization of every individual known by the French fiscal administration, it is possible to estimate at a very granular level the share of dwellings that

¹² <https://www.insee.fr/en/metadonnees/source/serie/s1019>

welcome new inhabitants every year. This database played as a benchmark. Then it was possible to compare that share estimated at the municipality level, with the volume of ads observed in the database, to estimate a coverage rate of the data. Below a comparison for NUTS 3 areas restricted to NUTS-2 Île-de-France region between data from tax data and data from SeLoger is displayed. Tax data make it possible to identify dwellings being rent in the private sector (as opposed to social dwellings, that do not belong to the field for the data from SeLoger) and measure for these dwellings year-on-year relocations. We then expect data from SeLoger to compare with this natural benchmark so to have an idea of the “market share” of SeLoger in the total of relocations for the rental private sector. As shown below, the data compare well with the benchmark, even after deduplication.

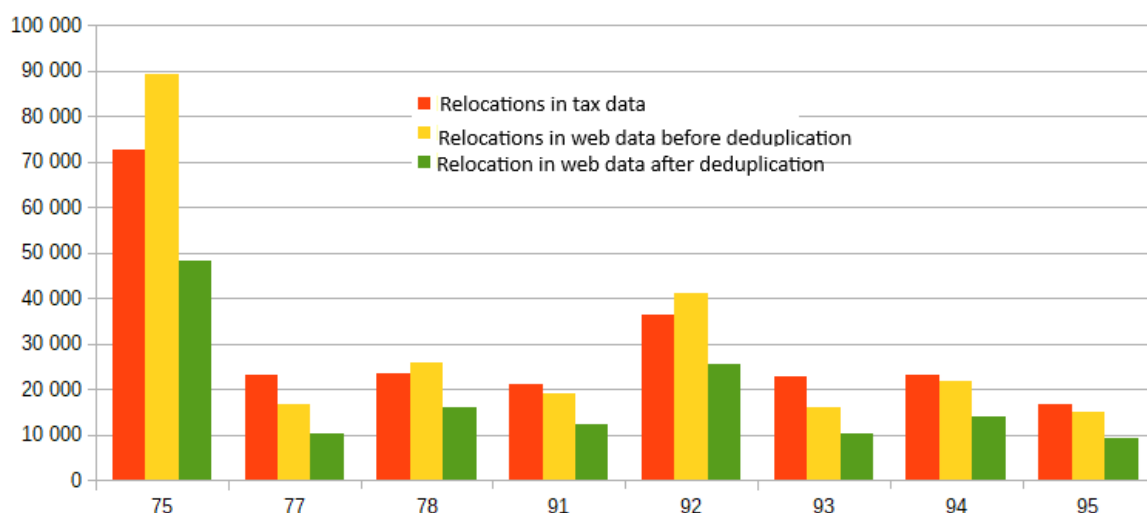


Figure 18. Number of relocations of dwellings by the department number in Île-de-France region

Also, Insee conducts on a quarterly basis a household survey among the tenants in order to measure the evolution of rents over time, especially for the computation of the CPI. The evolution observed on the data has been compared with the one measured with the survey, so as to assess the plausibility and coherence of the information, bearing in mind that there is a conceptual shift between those two sources: while web ads provide likely information on market prices for rent, indices derived from the survey also take into account in-place tenants who will face limited and regulated changes in their rents. Below results on the computation of such an index at the level of NUTS-2 Île-de-France region has been performed, showing a yet not perfect correlation between the index computed on web data and the data from the traditional survey.

Table 22. Comparison of price index in Île-de-France region

| Quarter | Webdata index | Survey index |
|---------|---------------|--------------|
| 2018Q2 | 0.5 | 0.1 |
| 2018Q3 | 0.8 | 0.4 |
| 2018Q4 | 0.0 | 0.2 |
| 2019Q1 | 0.9 | 0.5 |
| 2019Q2 | -0.6 | 0.2 |
| 2019Q3 | -0.2 | 0.3 |

See <https://www.insee.fr/en/information/6050999?sommaire=6049874> for more information on the sources that may act as benchmarks.

6. Modelling and interpretation

The process includes development and application of models, as well as data analysis and quality assessment. Partners had different approaches reflecting their own needs.

6.1. Bulgaria

Stability over time

After establishing working procedure for scraping and processing the data, the three sources of interest seem to be rather stable for certain periods:

- Source 1 – sales are increasing steadily at total of around 6% during the stable period from April 2023 till August 2024, while the rents are fluctuating within 12% for the same period. This source is offering twice as much sales offers then rent offers. There were problems with source structure and stability in the beginning of data scraping and processing till the April 2023. The source could be considered middle-to-big for Bulgarian market;
- Source 2 – the most well established source with less problems, but with only around 3000 offers per month. The sales and rents offers are fluctuating within 15-20% since April 2022 till August 2024. During the period the sale offers are generally decreasing, while the rent offers are generally increasing;
- Source 3 – small source with less than 3000 offers and most of them for rents at around 85% from all. The offers are steadily decreasing with around 10% total since April 2023 till August 2024. Before this period, the source was quite unstable with a lot of server errors and not responding during the scraping of data.

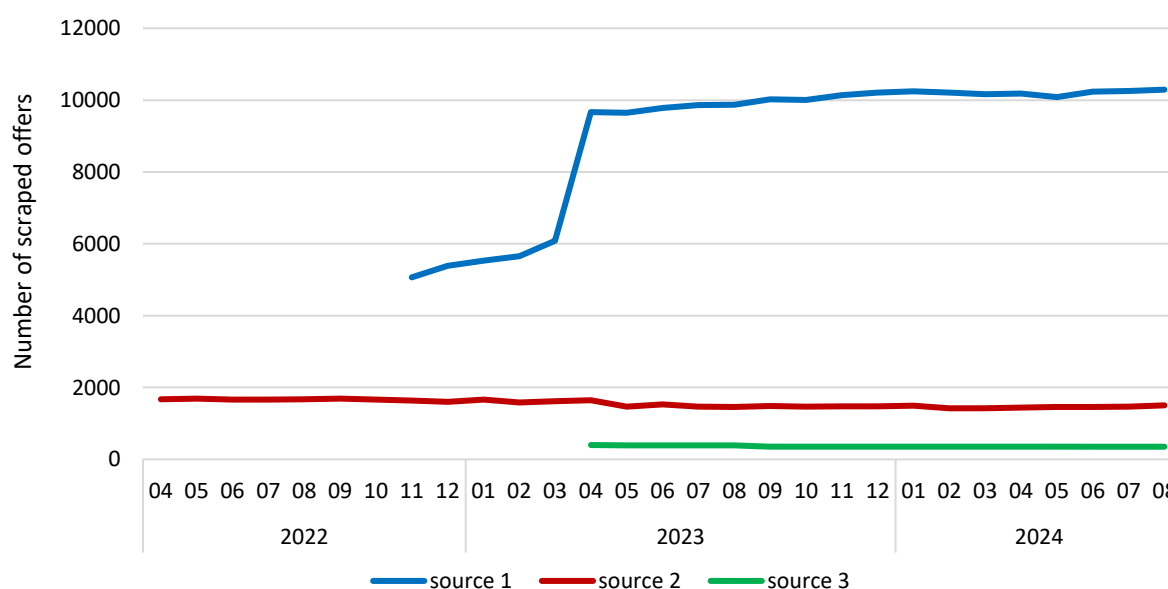


Figure 19. Number of offers for sale in Bulgaria

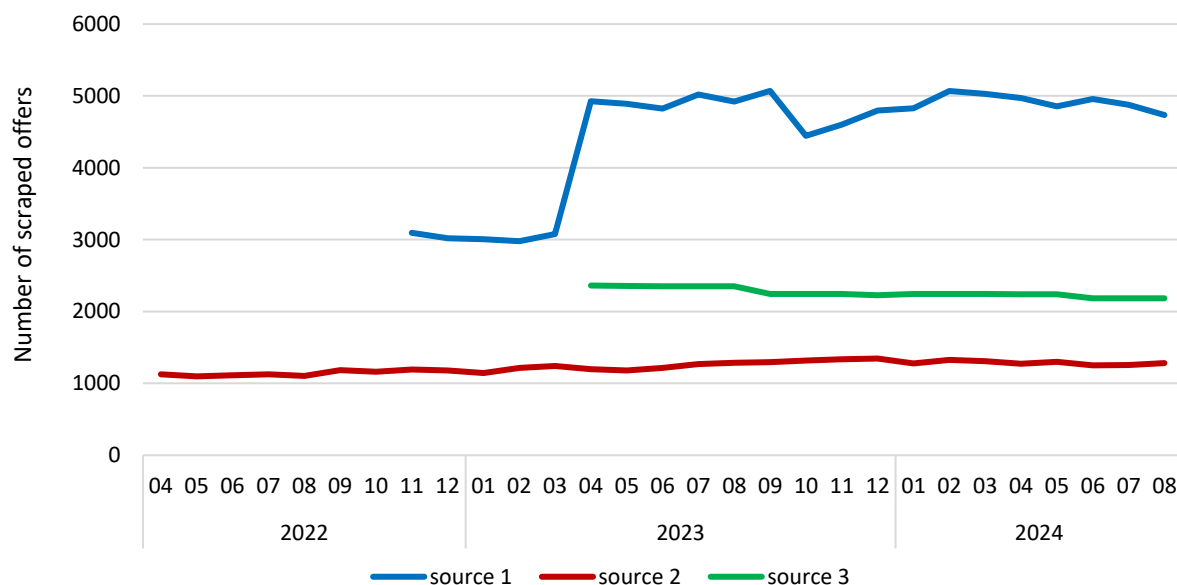


Figure 20. Number of offers for rent in Bulgaria

6.2. Poland

Stability over time

The web data sources used for the project are stable. Only one of the sources is no longer available due to the blocking of the possibility of data acquisition in its current form (via API). The average monthly number of web scraped offers from other data sources is rather stable and resembles the market shifts.

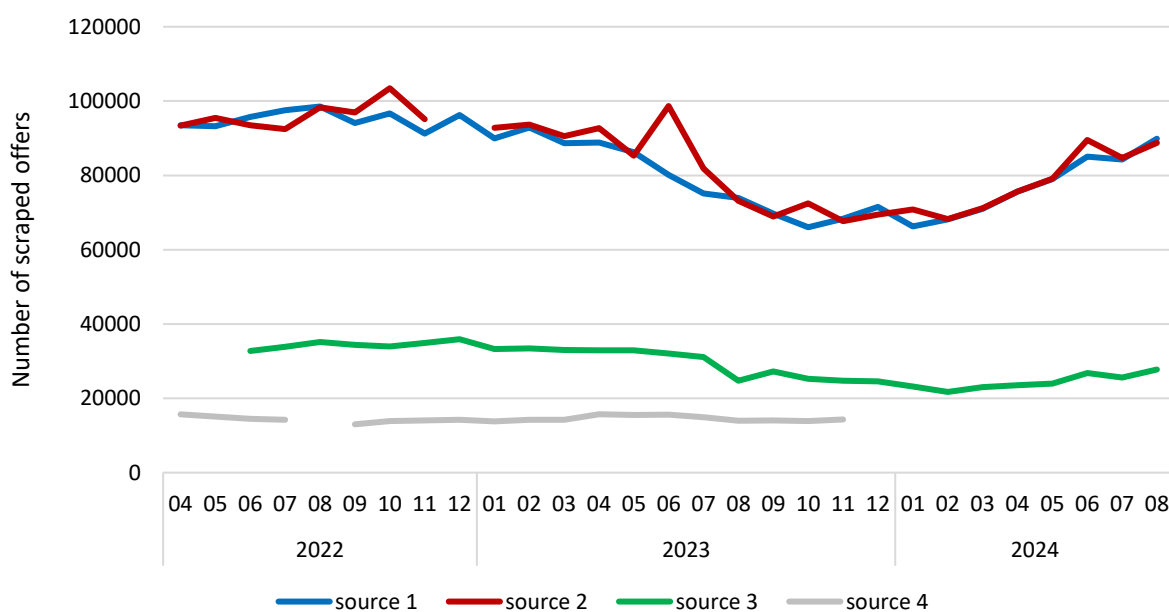


Figure 21. Number of offers for sale in Poland

This is particularly visible in the case of the rental market. The number of offers for rent in the first period of collecting data was small, but started increasing in 2023 and stabilizing in 2024. Due to the lack of the high-quality reference source on the rental market in Poland it is not possible to compare any trends on this phenomena¹³. However, the small number of offers occurred at the beginning of 2022 may indicate high demand for apartments at this time, when the war in Ukraine started and the massive inflow into Poland of Ukrainian refugees looking for short-term accommodation were observed. As the war continued, some immigrants moved further West or returned to Ukraine, and the real estate market began to stabilize.

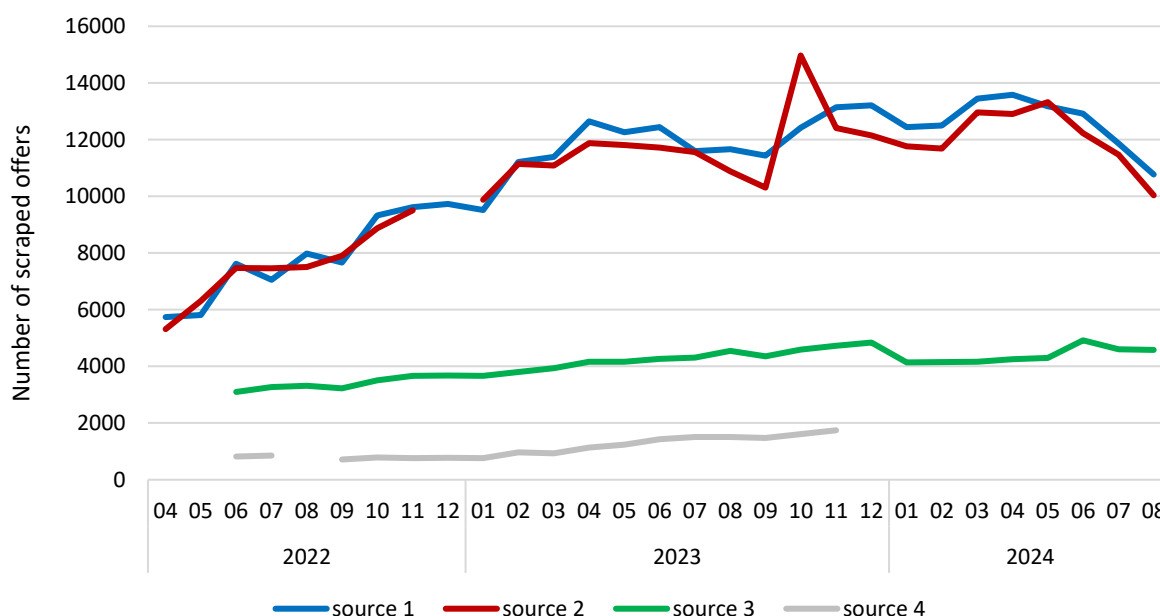


Figure 22. Number of offers for rent in Poland

Data completeness

One of the indicators for data quality assessment is completeness ratio. It allows to assess the scale of missing data, which may cause significant bias. The average monthly completeness ratio varies depending on the variables and sources. Source 3 appears to be the most complete data sources for offers for sale as completeness ratio for all mandatory variables is above 92%. Sources 1 and 2 have a high percentage of missing data on price, while the source 4 – number of rooms. The relatively high number of missing data on the offer price concerns primarily new buildings, where the price is disclosed only after contacting with the real estate agency. This problem does not occur in the case of other buildings and the rental market.

¹³ Tax data, for example, covers only the aggregated information on the rental income, which does not allow identifying any specific information about the apartments the tax corresponds to.

Table 23. The average monthly completeness ratio for selected variables¹⁴

| Data source | Price | Surface | Rooms | Location |
|-------------|-------|---------|-------|----------|
| Sell | | | | |
| Source 1 | 75.9% | 97.7% | 97.4% | 97.7% |
| Source 2 | 71.1% | 93.2% | 92.9% | 93.8% |
| Source 3 | 92.0% | 96.9% | 96.9% | 96.9% |
| Source 4 | 97.2% | 97.2% | 72.3% | 97.2% |
| Rent | | | | |
| Source 1 | 97.4% | 97.5% | 97.4% | 97.8% |
| Source 2 | 94.7% | 94.8% | 94.4% | 95.3% |
| Source 3 | 97.2% | 97.3% | 97.3% | 97.3% |
| Source 4 | 97.5% | 97.5% | 81.3% | 97.6% |

Comparison between web data and official statistics

The next section presents the comparison of the average price in PLN/m² based on web-scraped data with two official data sources:

- the Real Estate Price Register which is the main source for the real estate statistics published by GUS¹⁵,
- the survey among real estate agents conducted by the Central Bank of the Republic of Poland (NBP)¹⁶.

Comparison of the average price for apartments for sale, with the above mentioned sources indicated a high convergence of trends. The comparison was made for the largest cities in terms of population, and even the offer prices were higher than transaction prices, the trends were practically identical.

¹⁴ Percentage of non-missing values in a dataset or a specific column.

¹⁵ This is the public register kept by a local governments. It includes data on real estate prices given in notarial acts.

¹⁶ NBP publishes both transaction and offer prices of apartments.

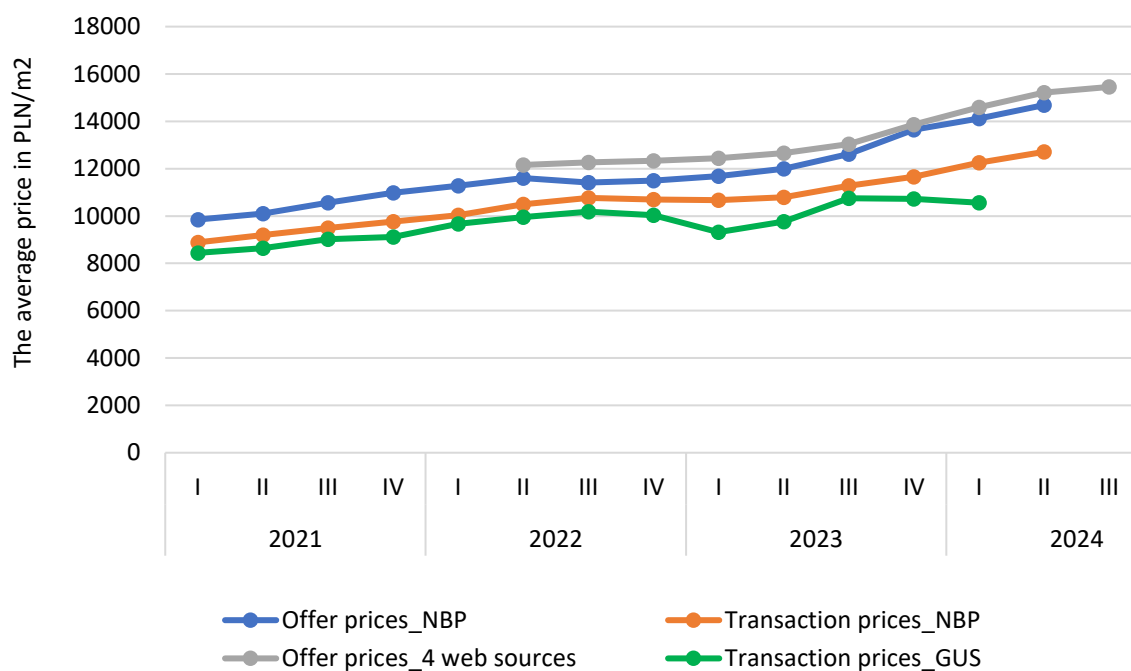


Figure 23. Comparison between web data and official statistics for the largest cities in Poland¹⁷

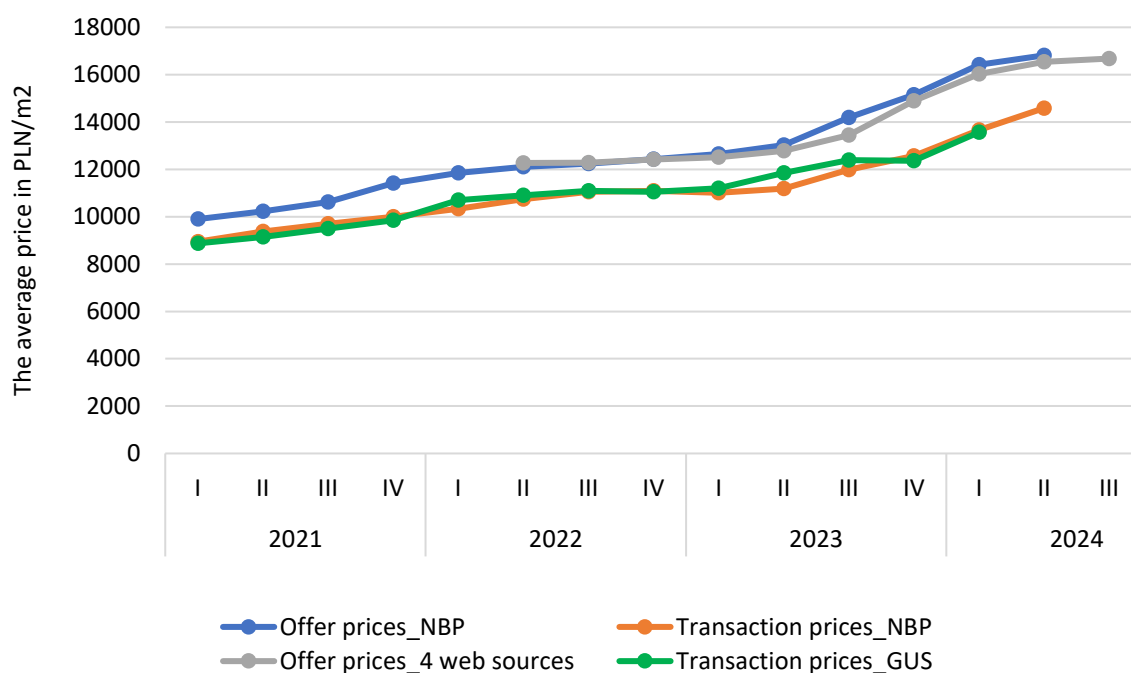


Figure 24. Comparison between web data and official statistics for Krakow

¹⁷ Gdańsk, Gdynia, Kraków, Łódź, Poznań, Warszawa, Wrocław.

This statement can be supported by statistical measures. The Pearson correlation coefficients between the web data sources and official data sources are very high and statistically significant. Only, transaction prices calculated on the basis of the Real Estate Price Register have lower correlation coefficient (and in the case of all large cities in Poland it is not statistically significant). It arises the relevant question on the quality of the register as such, especially since it is fed with data with a long delay.

Table 24. The Pearson correlation coefficient and p-value between the web and official data sources - large cities in Poland

| variable | correlation coefficient | p-value |
|-------------------------------------|-------------------------|----------|
| Large_cities_offer_prices_NBP | 0.988010 | 0.000004 |
| Large_cities_transaction_prices_NBP | 0.988083 | 0.000004 |
| Large_cities_transaction_prices_GUS | 0.650912 | 0.080451 |

Table 25. The Pearson correlation coefficient and p-value between the web and official data sources - Krakow

| variable | correlation coefficient | p-value |
|-------------------------------|-------------------------|----------|
| Krakow_offer_prices_NBP | 0.988486 | 0.000004 |
| Krakow_transaction_prices_NBP | 0.987838 | 0.000004 |
| Krakow_transaction_prices_GUS | 0.940380 | 0.000506 |

6.3. Germany – HSL

Figure 25 shows the total number of collected offers for newly constructed objects in 2023. At first, the different size and relevance of platforms becomes visible. In fact, only Portal 1 and Portal 2 are relevant in terms of size. Together, they make up over 93% of all collected offers (see Table 25).

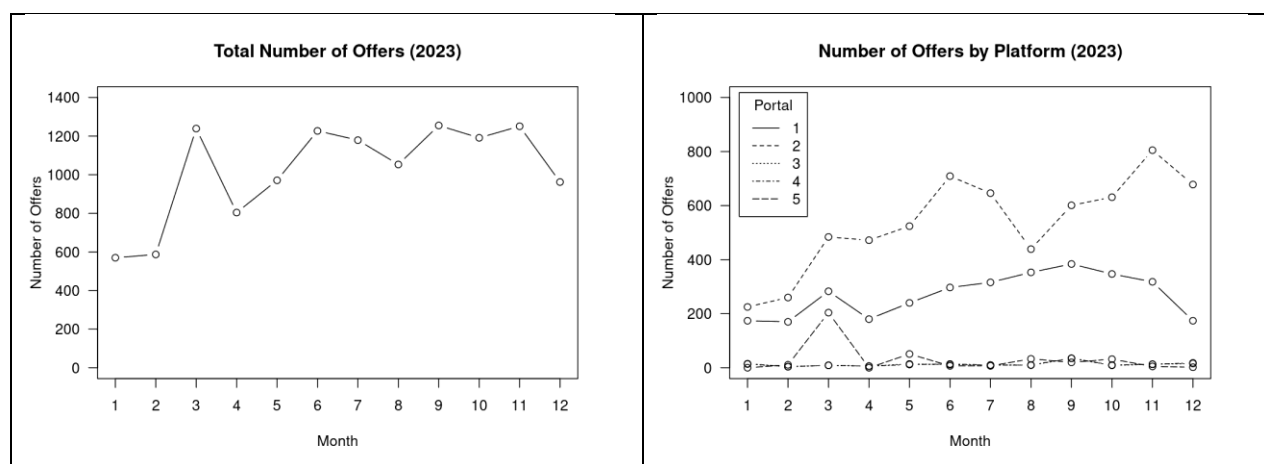


Figure 25. Number of offers for newly constructed objects (2023): total number by platform provider

Table 25 shows the number of collected offers for the years 2022 and 2023 for all portals. The smaller portals have been chosen in order to investigate some niches of the online real estate market. In 2022, the number of collected offers for Portal 2 is much lower than 2023 since not all months of 2022 are covered by this source.

Table 26. Number of offers for newly constructed objects in 2022 by portal

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

Typically, many offers describe objects in the urban Rhine-Main region around Frankfurt am Main and the capital city of Wiesbaden, whereas there only are fewer offers from the more rural NUTS3 regions in the north of Hesse (Figure 26).

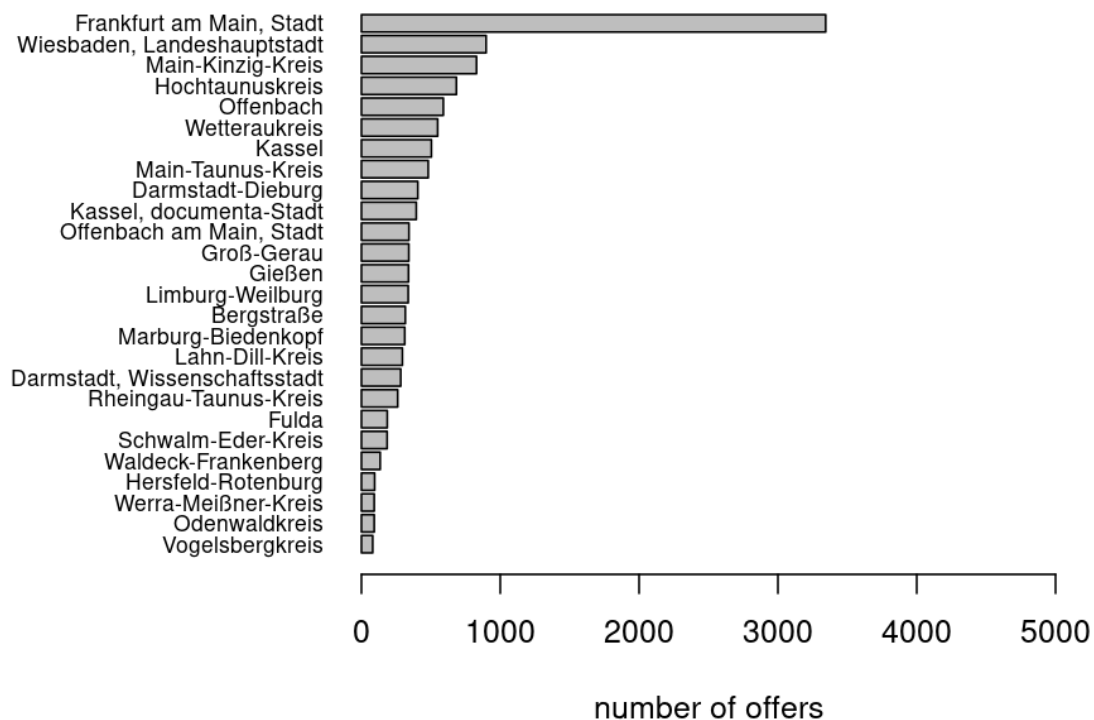


Figure 26. Number of offers for newly constructed objects in 2023 by NUTS3 region

Figure 27 shows the number of collected offers in 2023 by type of offer (object for rent or sale) and type of building (houses vs. apartments; category “other” is not shown): In 2023, there are more offers for objects to buy than to rent and the number of offered apartments is much higher than the number of offered houses.

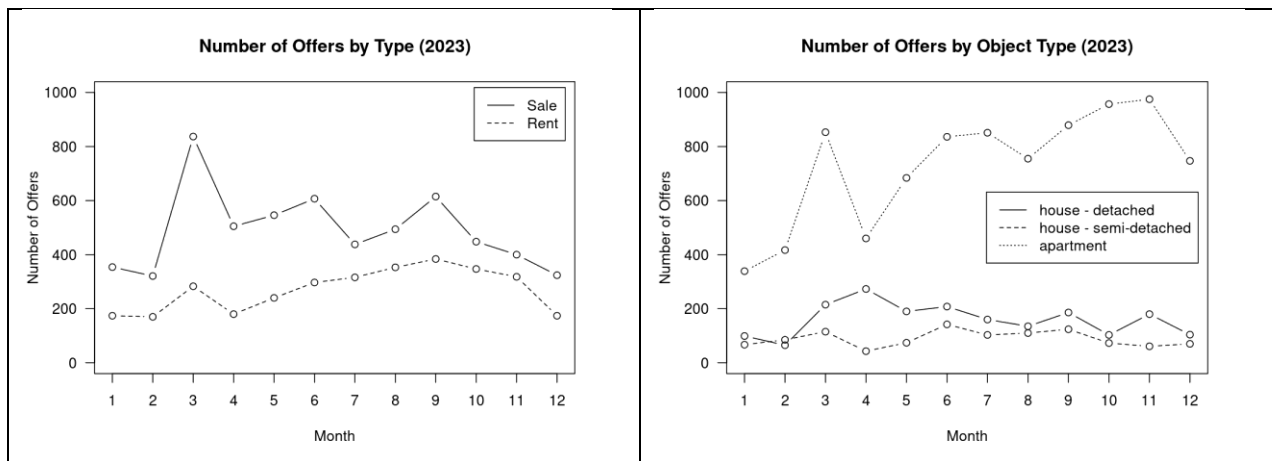


Figure 27. Number of offers (2023) for newly constructed objects by offer type and building type

Despite the large number of collected offers for both objects to buy and objects to rent, as well as houses and apartments, looking at the objects advertised does not cover objects, that are in fact newly constructed but not available at the market now (and may not for many years) since they are built by and for their owners use.

Use Case 2 did investigate coverage of all construction activities by using advertisements on real estate portals. In general, coverage again varies largely between different regions and is higher in more urban regions. And even when this kind of data not yet seems to be a good indicator to predict construction activities in general or total, measuring other trends or phenomena (e.g. looking at price development, sizes, special features and amenities, accessibility e.g. barrier-free access or bathrooms), using this data is possible.

6.4. Germany – SSI-BBB

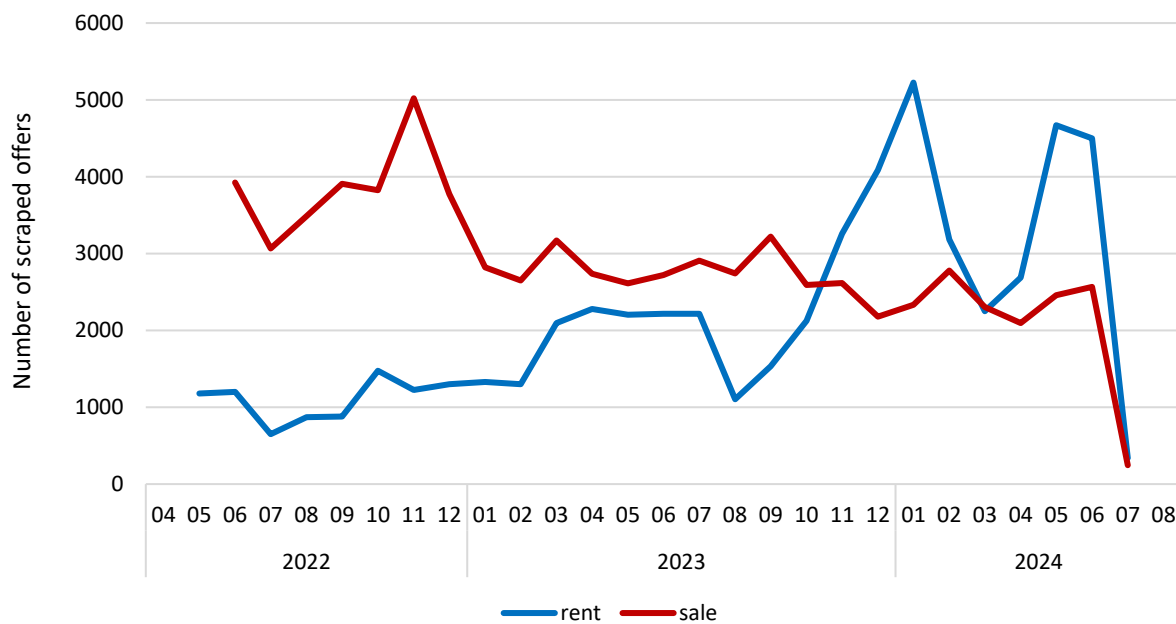


Figure 28. Number of offers for sale and rent in Berlin

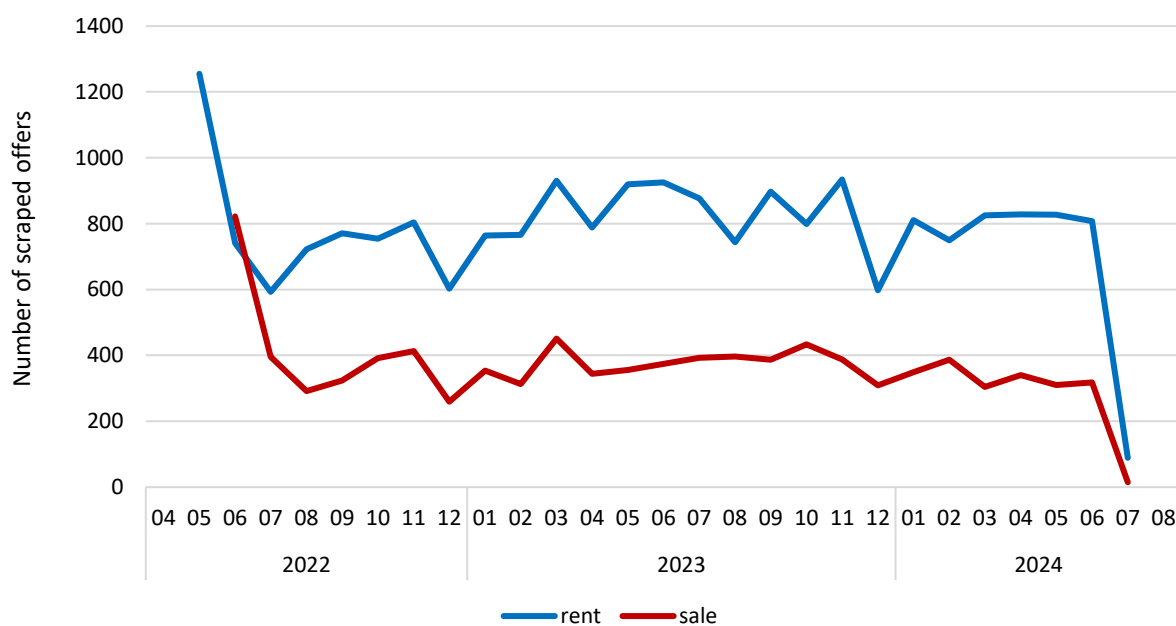


Figure 29. Number of offers for sale and rent in Brandenburg

The Berlin-Brandenburg metropolitan region offers a diverse array of properties for sale and to rent, with significant variations across different areas. The number of properties for sale in both states is very roughly in a 2:1 ratio to rental properties. There are considerably more offerings in major cities like Berlin, Potsdam, and Cottbus, compared to rural areas. In more remote rural regions, especially those far from Berlin, real estate portals list far fewer properties, especially rentals. This may be due to limited availability or

a preference for traditional, offline sales channels in these regions. Such disparity suggests that areas with sparse listings might not yield reliable data for the project's scope.

Remarkably, the data shows a relatively small number of rental properties in Berlin. This could be due to the extremely high demand for apartments, leading to properties being rented out before they are even listed online; this is an aspect that requires further investigation.

6.5. Finland

Stability over time

Oikotie data has good coverage and is stable over time. Yearly cycles in real estate sales market are visible in Figure 30 as there are less offers in each winter holiday seasons. The COVID pandemic in 2020 to 2021 decreased the numbers of sale offers considerably. The pre-pandemic levels of numbers were reached again in 2023.

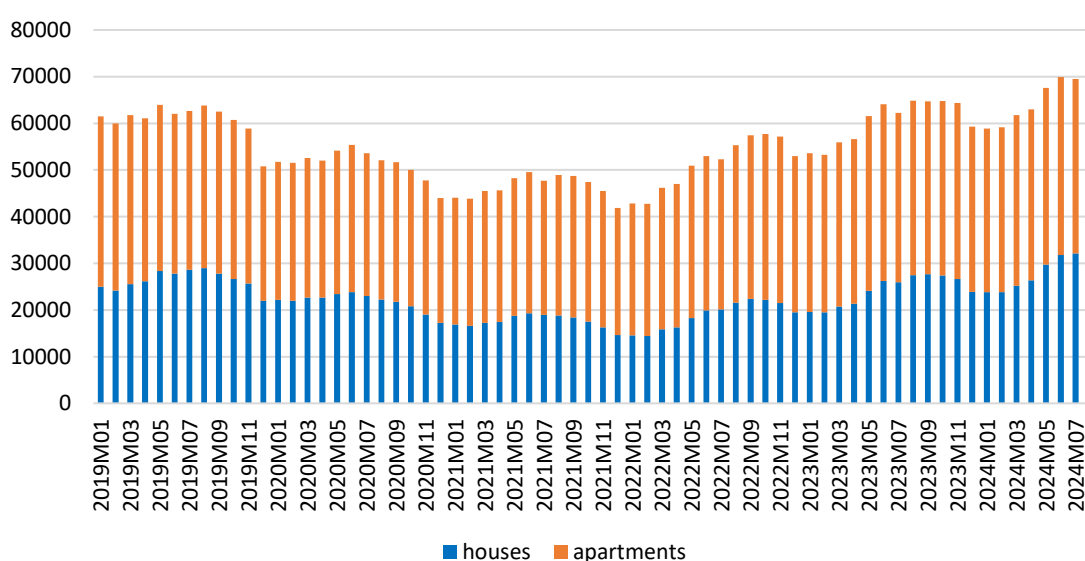


Figure 30. Number of offers for sale on Oikotie 2019M01 to 2024M08

Numbers of offers on rental market are also high and stable. Over the years there seems to be a growing trend in numbers in Figure 31. This was due to increased supply in rental apartments in the past years, especially in Greater Helsinki where there has been a construction boom. Most of the new apartments were owned by investors who then rented out the apartments.

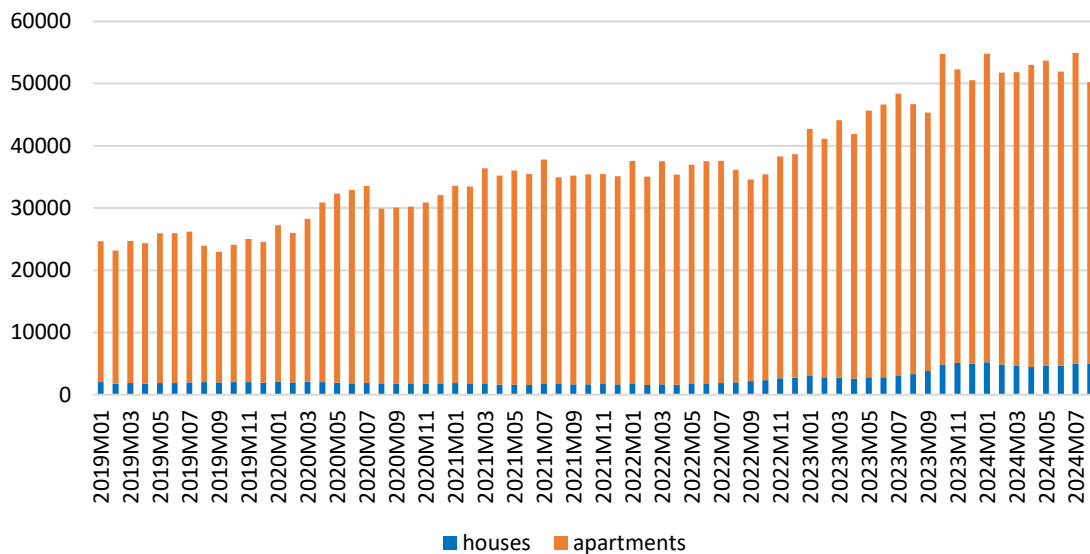


Figure 31. Number of offers for rent on Oikotie 2019M01 to 2024M08

Regarding the data agreement though, since the data was acquired by purchasing after yearly negotiations, budget constraints can become an issue in the future for Statistics Finland. This is a significant risk for stability and continuous production of these Oikotie statistics.

Comparison between web data and official statistics

Average prices and rents on Oikotie were compared with some official Finnish statistics. The official statistics have vast coverage: sales of dwellings and houses are based on nearly complete data on all Finnish transactions that are acquired from Tax Administration and National Land Survey.

Official rents are based on the Social Insurance Institution's register of housing allowances and on the data on rental housing companies included in the data collection. Tenancies in which the landlord is a private person, and the tenant does not get any housing allowance, are not covered in the official rent statistics.

Due to differences in some definitions, scope and calculation methods, the official statistics may not be fully comparable with Oikotie data. However, approximate conclusions can be drawn.

To make comparisons with the official statistics, averages of Oikotie data had to be calculated on quarterly time period. New developments had to be excluded from both sales and rents, so that it was possible to compare with official statistics regarding old dwellings and houses.

Firstly a comparison of numbers of observations between Oikotie and national statistics had been done. For this, only removed advertisements from Oikotie were chosen, assuming that most of them ended up being sold or rented.

Numbers of sale advertisements and official sales can be compared using monthly statistics (note that Statistics Finland publishes monthly experimental statistics from Oikotie data). However, national building classification was used, which consists of old dwellings in housing companies. This includes apartments and terraced houses which is not same as the building categories in this project. The comparative time series spans from 2020 to 2023 in Figure 32.

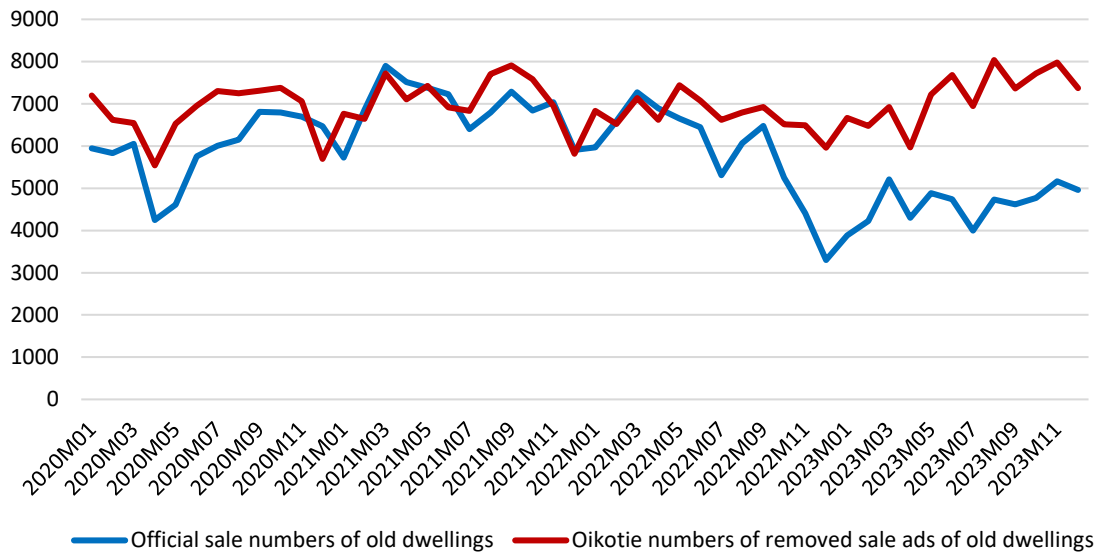


Figure 32. Comparison of numbers of sale observations between web data and official Finnish statistics 2020Q1 to 2023Q4

Most of the time, there are about 1000 observations more on Oikotie or the numbers are on similar level. However, since the late 2022 there have been noticeably less sale transactions than offers on Oikotie. This is probably due to the significant change in interest rates since 2022, when home loans became considerably more expensive after a long time of low interest rates. This decreased the demand for buying an own house.

In Figure 33, numbers of removed rent offers and official new rent tenancies were compared. Quarterly numbers of new tenancies according to Statistics Finland were about 20000 to 30000. Numbers of rent offers during same quarter was however somewhat higher, around 35000 to 65000 observations. Especially in the last few years, it seems that the higher supply of rental apartments may have increased the gap between advertisements and actual new tenancies.

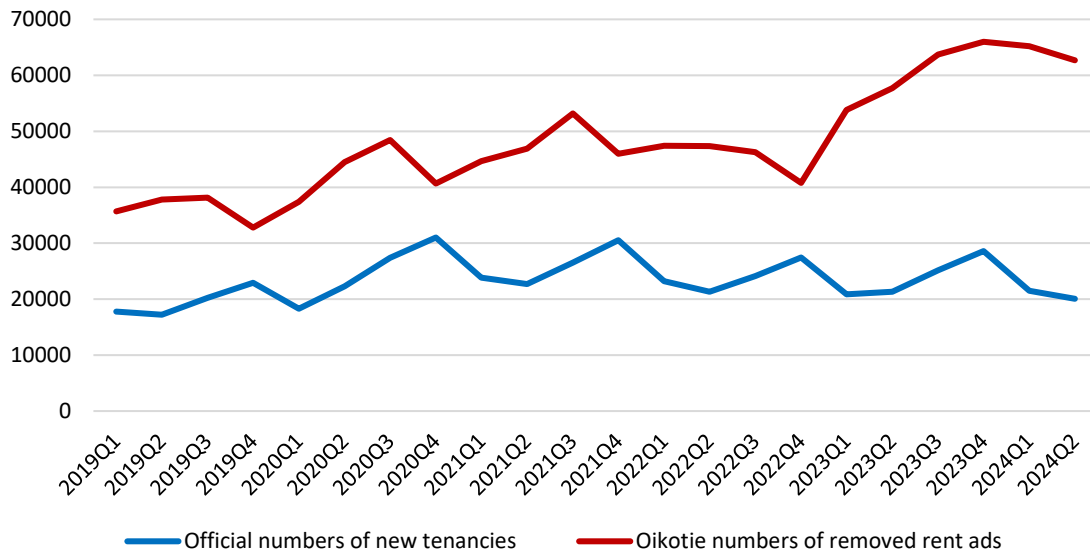


Figure 33. Comparison of numbers of rent observations between web data and official Finnish statistics 2019Q1 to 2024Q2

For price and rent comparisons, all Oikotie observations were used, not just the removed advertisements like in earlier figures regarding numbers.

Prices of old houses do not have a direct counterpart in Finnish statistics, because they are categorized differently than in this project. That is why, old houses on Oikotie (meaning they are not new development) were compared to both terraced houses (old dwellings in housing companies) and old single-family houses (detached real property).

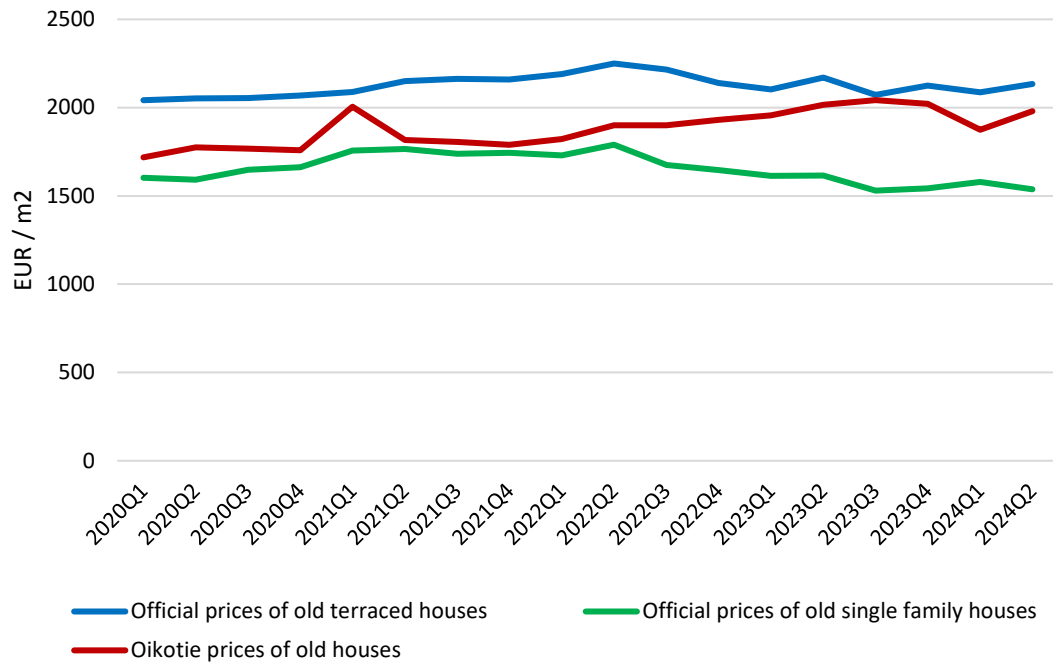


Figure 34. Comparison of house prices between web data and official Finnish statistics 2020Q1 to 2024Q2

In Figure 34 the prices on Oikotie fall between the two official statistics. It was expected since the categorization in this project includes both terraced houses and detached single-family houses. Prices on Oikotie are more volatile, ranging up and down between the two official price curves.

Prices of old Oikotie apartments were compared with the prices of old dwellings in blocks of flats. Here the prices develop in somewhat opposite directions in Figure 35. In 2020 to 2021 the asking prices in Oikotie were almost 500 euros lower per square metre. Then by the end of 2022 they surpass the level of actual official price statistics. Oikotie prices grew more in time whereas official prices stayed more stable at about 3000 euros per square metre. The decrease in actual prices in 2022 shows well the change in house market that was caused by increased interest rates.

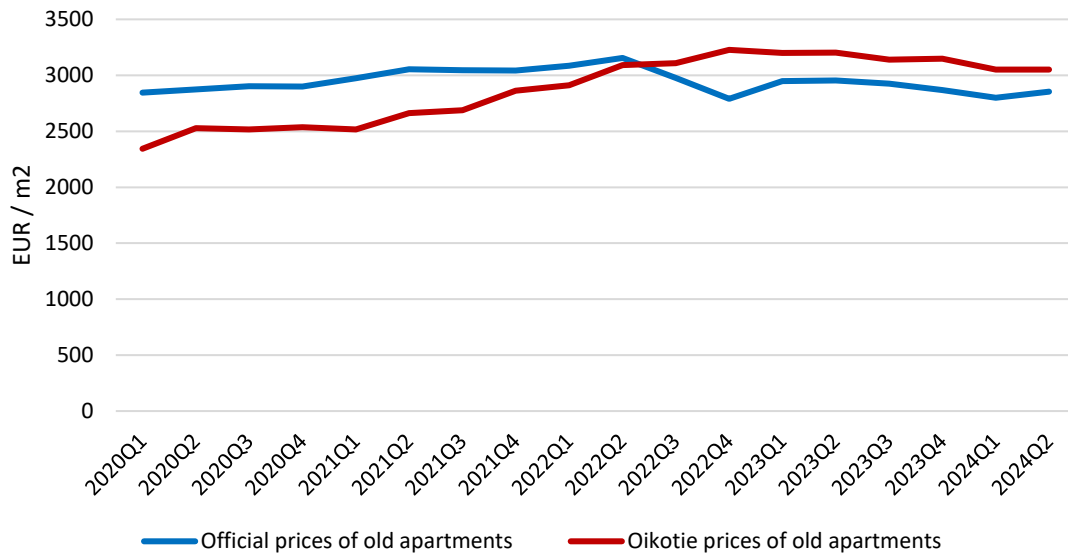


Figure 35. Comparison of apartment prices between web data and official Finnish statistics 2020Q1 to 2024Q2

For rents data, observations meant for short-term contracts and those that are rented furnished, were deleted from Oikotie data for this comparison. Total rents were changed into rents per square meter. Oikotie averages were calculated for all building types total, since the official rent statistic does not separate different building types.

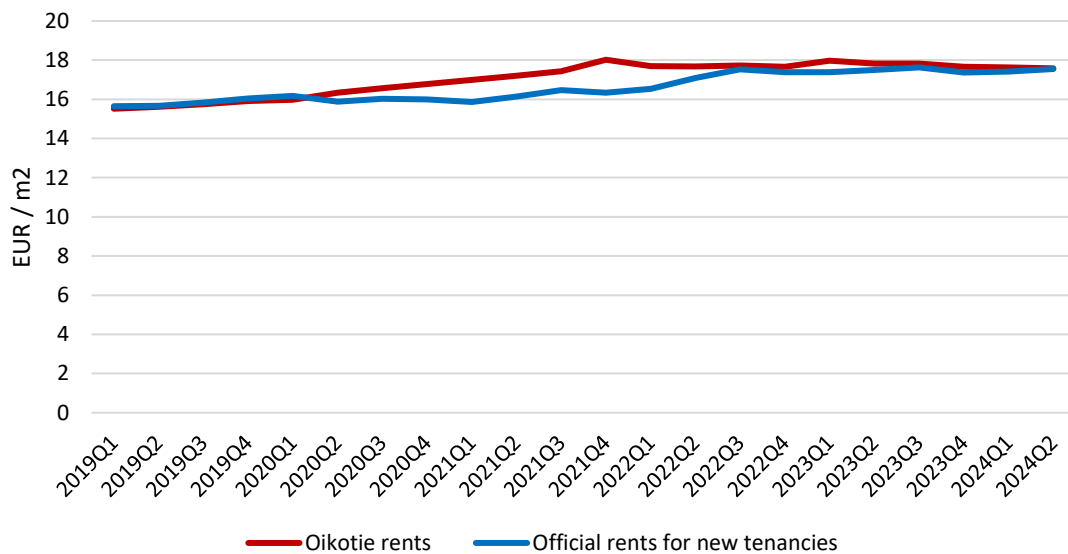


Figure 36. Comparison of rents between web data and official Finnish statistics 2019Q1 to 2024Q2

Average rents for new tenancy agreements in the official statistics were quite close to the asking rents in Oikotie advertisements in the whole timespan. As expected, asking rents were usually higher than the official rents. The biggest difference is 1.68 euros per square meter on the last quarter of 2021.

Missing values

The total number of missing values in mandatory variables in August 2024 was 10901 (3.0% of all mandatory values) for sale offers and 6107 (2.4% of all mandatory values) for rent offers.

Table 27. Monthly average number of missing data in mandatory variables in all collected data

| Data source | Total | Price | Surface area | Rooms | Floor | Location |
|---------------|-------|-------|--------------|-------|--------|----------|
| Oikotie sales | 2.84% | 0.53% | 0.20% | 1.06% | 12.42% | 0.00% |
| Oikotie rents | 2.21% | 0.36% | 0.03% | 1.38% | 9.25% | 0.00% |

6.6. France

The rationale behind acquiring web data on rents was to construct an alternative index for rents that would be available at a higher frequency than the current one (which is a quarterly index). As it is not possible to have longitudinal information for rented dwellings in the web data from SeLogger, the mere solution consists of estimating hedonic price index based on dwellings' characteristics. This entails several methodological choices that have been tested:

- a spatial stratification on which sub-indices are estimated, then aggregated at a higher (regional or national) level;
- a strategy for aggregating those sub-indices (weighted mean, arithmetic mean, geometric mean or median being some of those strategies that have been tested);
- variables to be included in the hedonic models.

Several experimental indices have been computed based on those methodological choices. Those indices can also be spatial-wise broken up, which will in theory increase the opportunities in terms of analysing local real estate markets. However, the possibility of zooming on local areas remain quite limited, especially in rural areas where the number of ads can be quite low, making the spatial aggregation for stratification rather wide.

7. Dissemination of the experimental statistics and results

This chapter presents the selected statistics and results of the project. It covers the basic information on the number of offers in time series and territorial disaggregation. The experimental statistics with all mandatory variables are presented on the [Wiki WIN](#)¹⁸.

Figures below present number of offers for sale and rent collected from the biggest online real estate platform in the selected countries¹⁹.

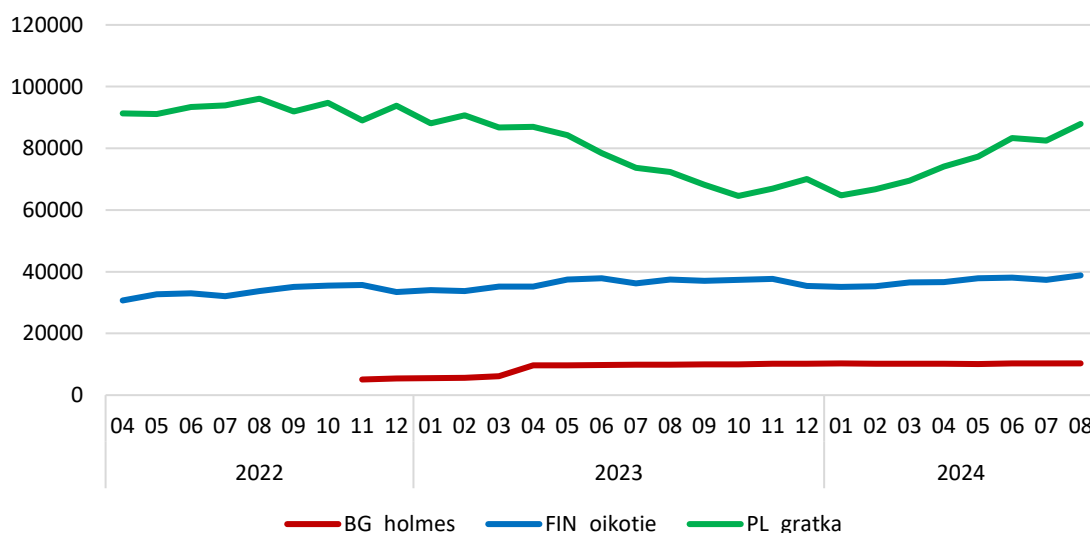


Figure 37. Number of offers for sale per country

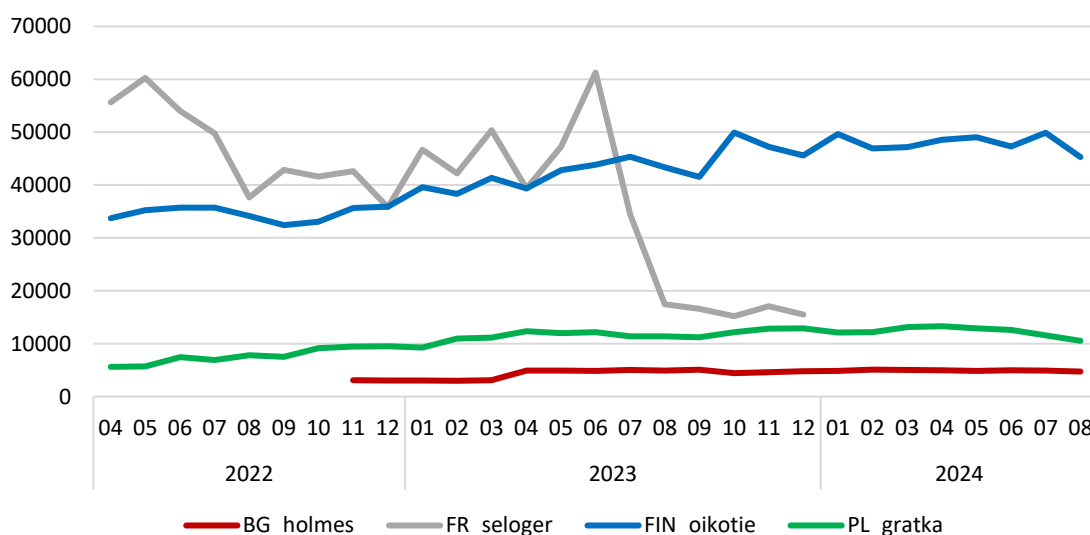


Figure 38. Number of offers for rent per country

¹⁸ Available only for the WIN and WISER members.

¹⁹ The figures do not include data from HSL and SSI-BBB, as well as for sale market in France.

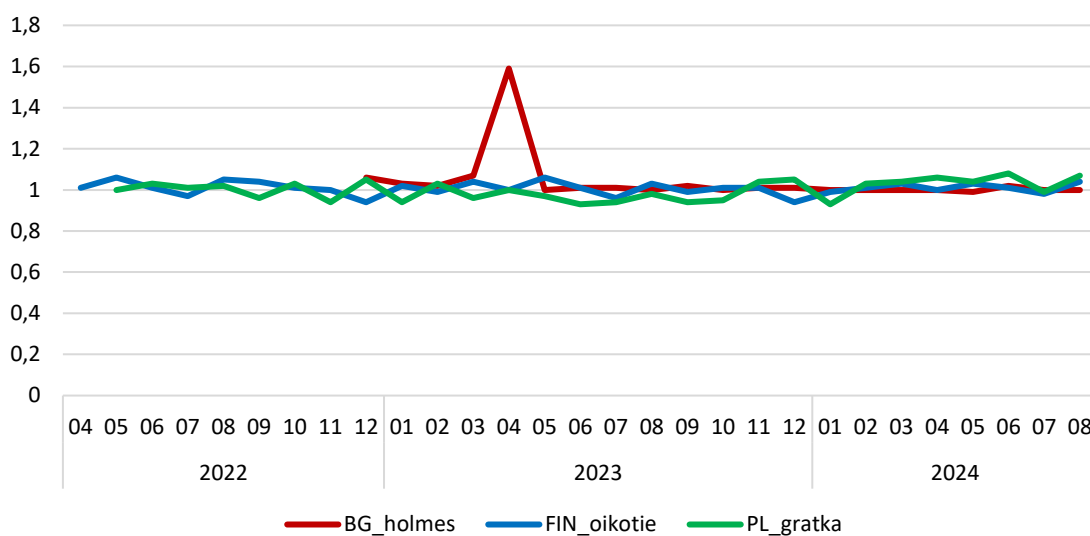


Figure 39. Changes of number of offers for sale per country m/m

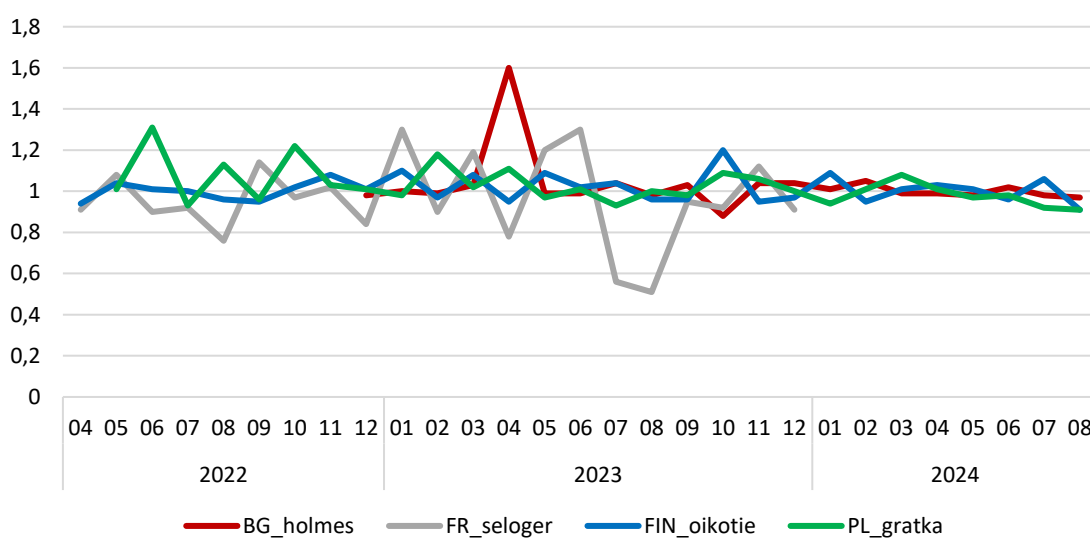


Figure 40. Changes of number of offers for sale per country m/m

In the most cases the number of offers is stable and/or reflects the changes on the real estate market. Only the number of offers for rent in France fluctuates significantly between subsequent periods, which will require further investigation.

Sales and rental offers in all countries are concentrated in the largest cities. A relatively high number of offers is also available in the attractive tourist regions, in coastal or mountain areas. Territorial distribution of offers with the information of shares of individual data sources is presented in the maps below.

7.1. Bulgaria

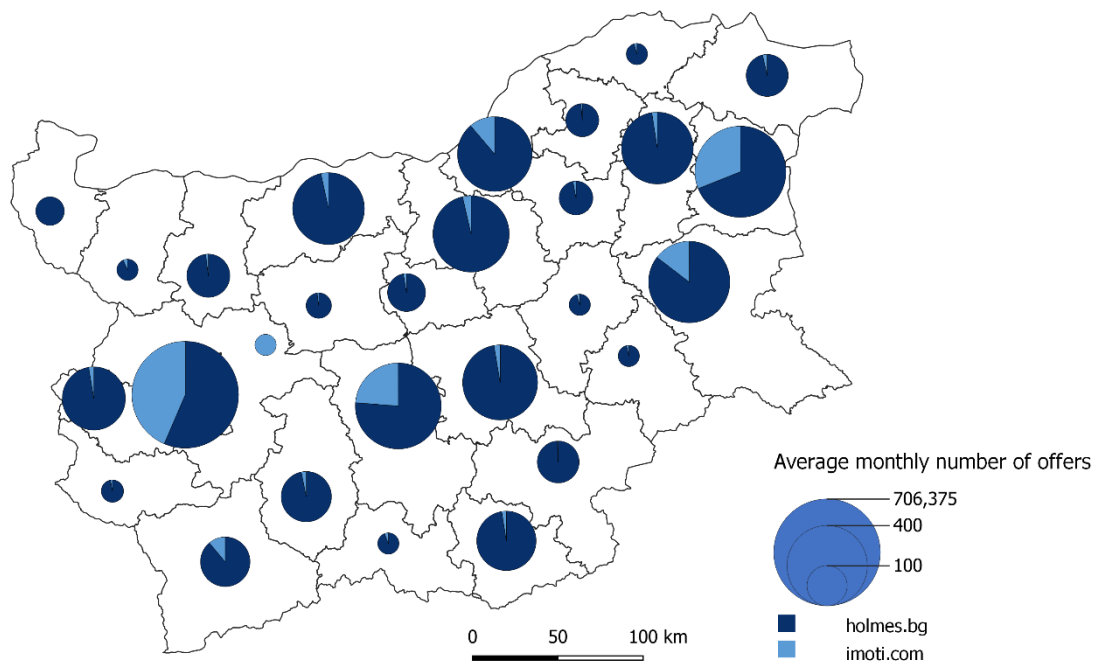


Figure 41. Average monthly number of offers for sale in Bulgaria in 2023 by source

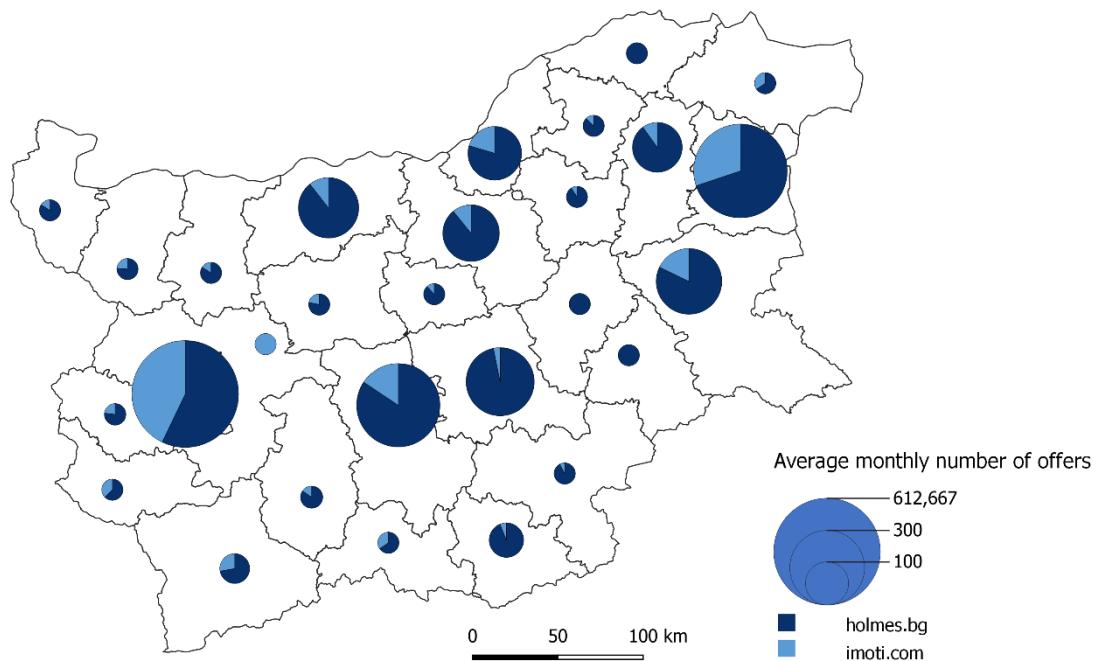


Figure 42. Average monthly number of offers for rent in Bulgaria in 2023 by source

7.2. Poland

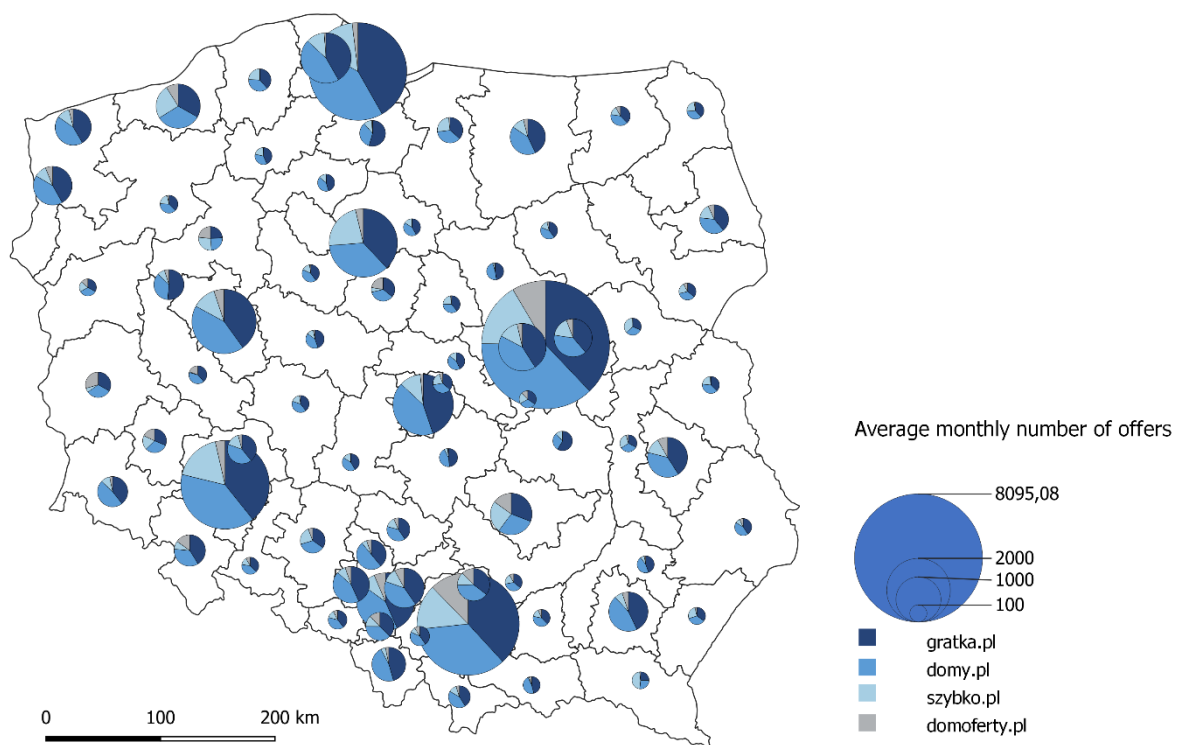


Figure 43. Average monthly number of offers for sale in Poland in 2023 by source

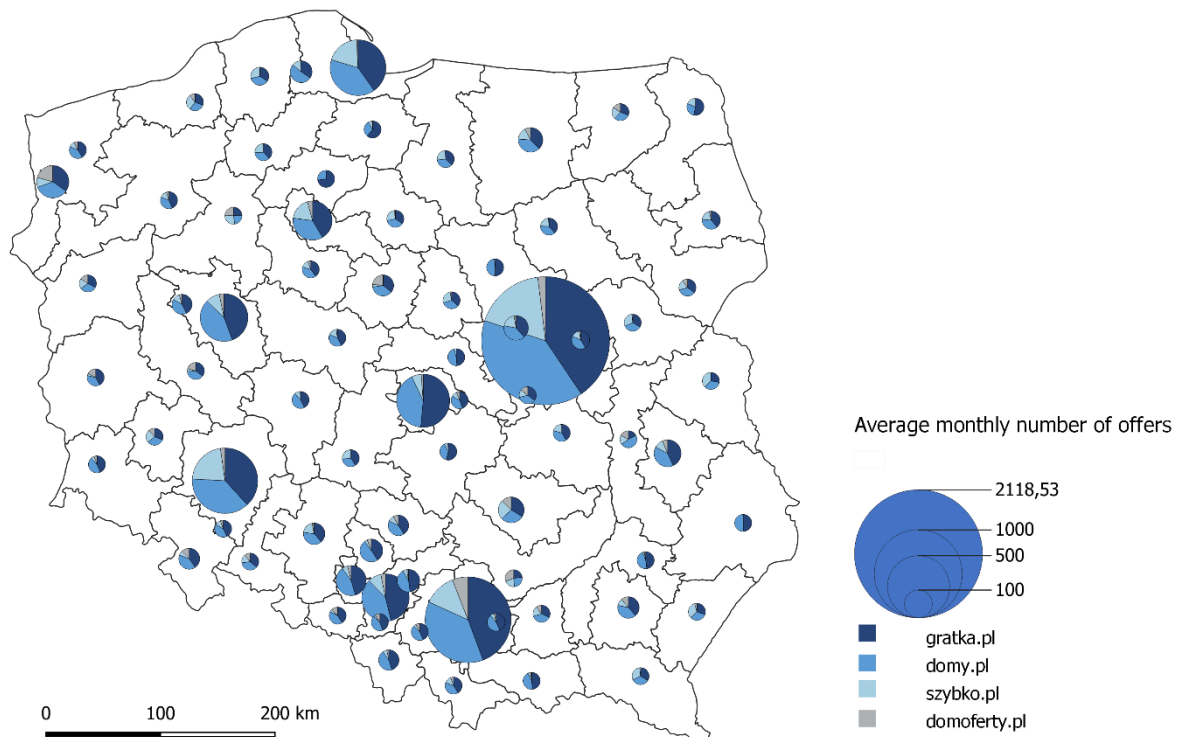


Figure 44. Average monthly number of offers for rent in Poland in 2023 by source

7.3. Germany – HSL

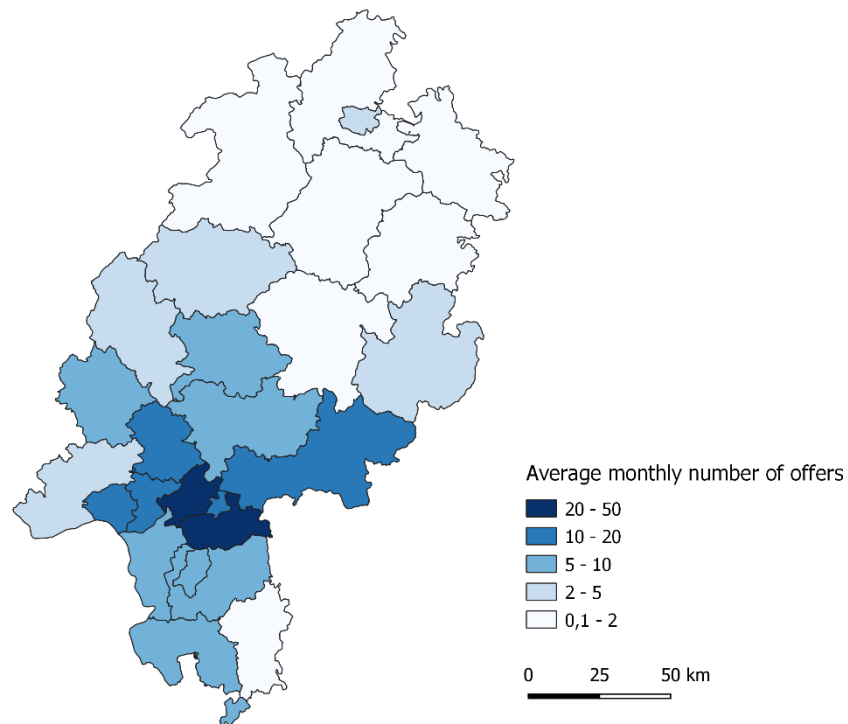


Figure 45. Average monthly number of offers for sale in Hesse in 2023

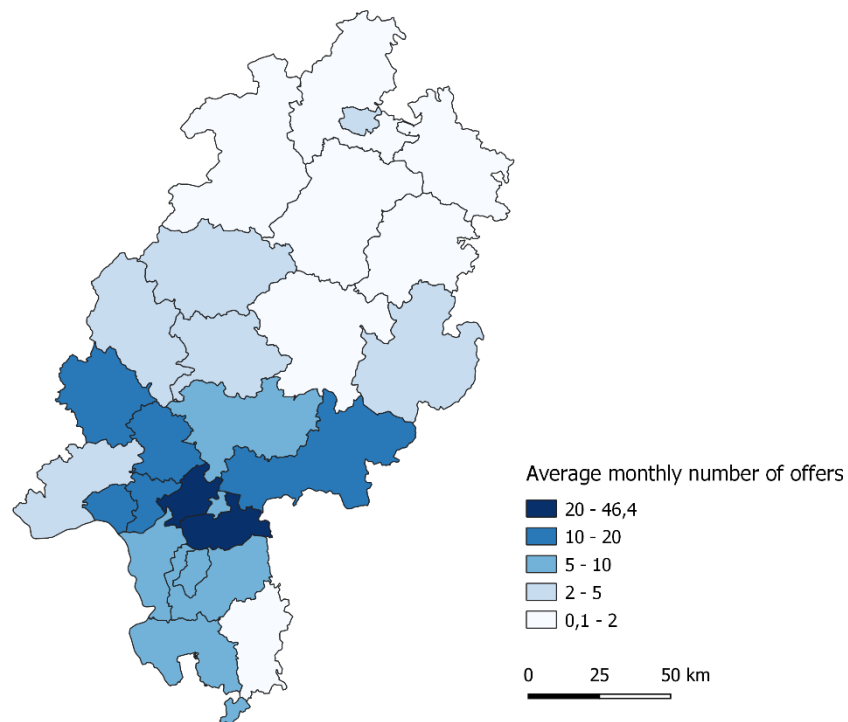


Figure 46. Average monthly number of offers for rent in Hesse in 2023

7.4. Germany – SSI-BBB

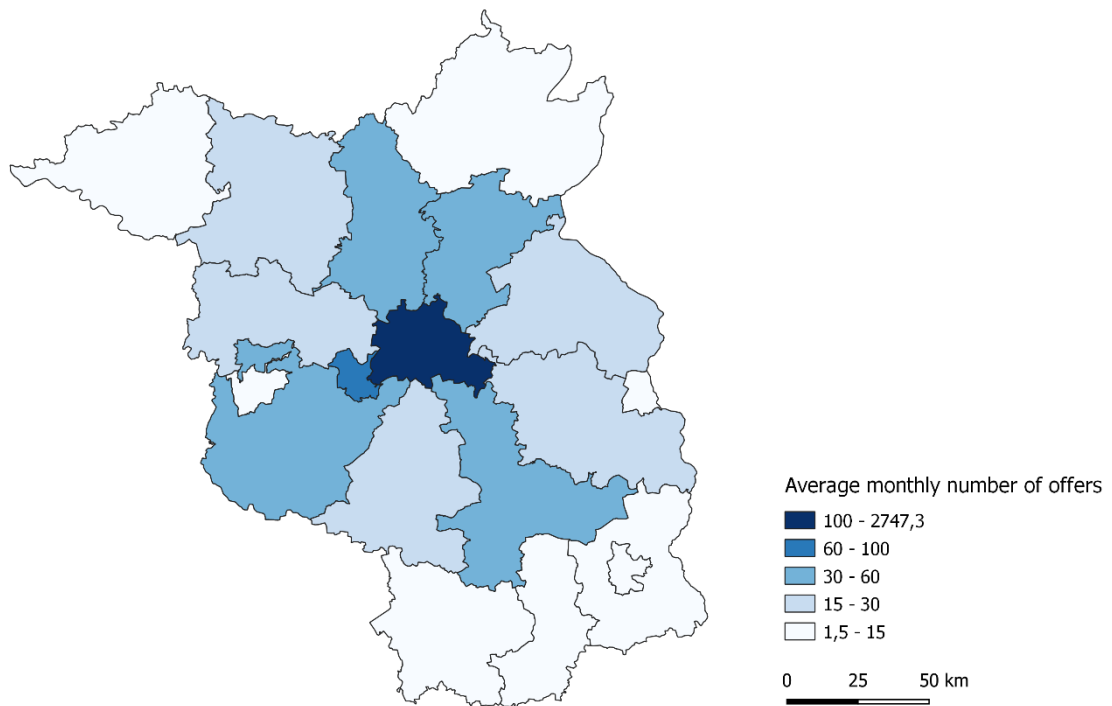


Figure 47. Average monthly number of offers for sale in Berlin-Brandenburg in 2023

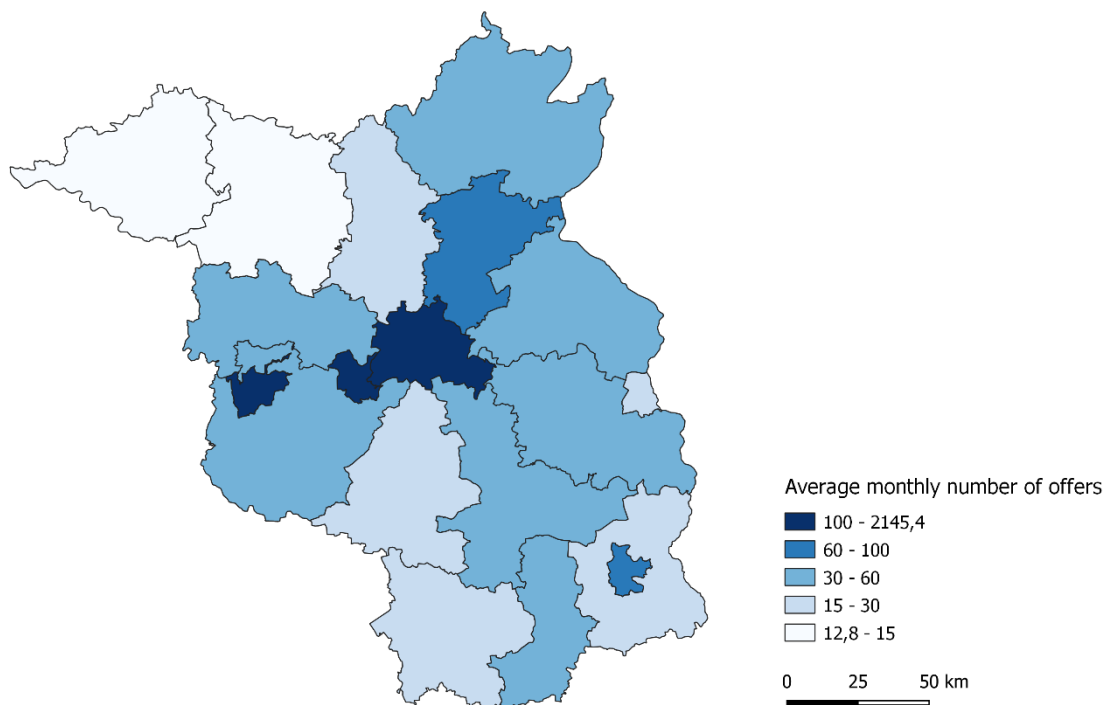


Figure 48. Average monthly number of offers for rent in Berlin-Brandenburg in 2023

7.5. Finland

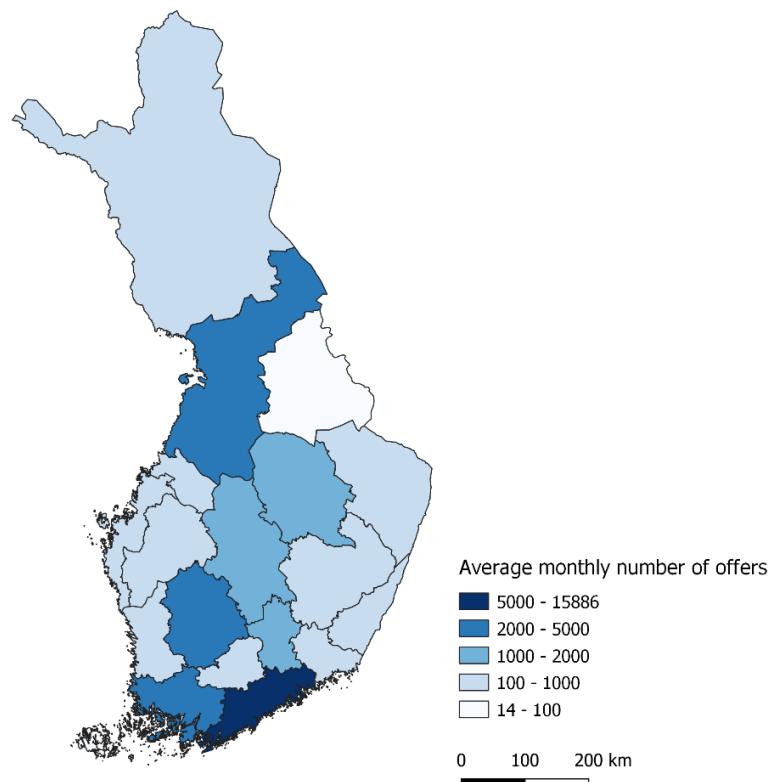


Figure 49. Average monthly number of offers for sale in Finland in 2023

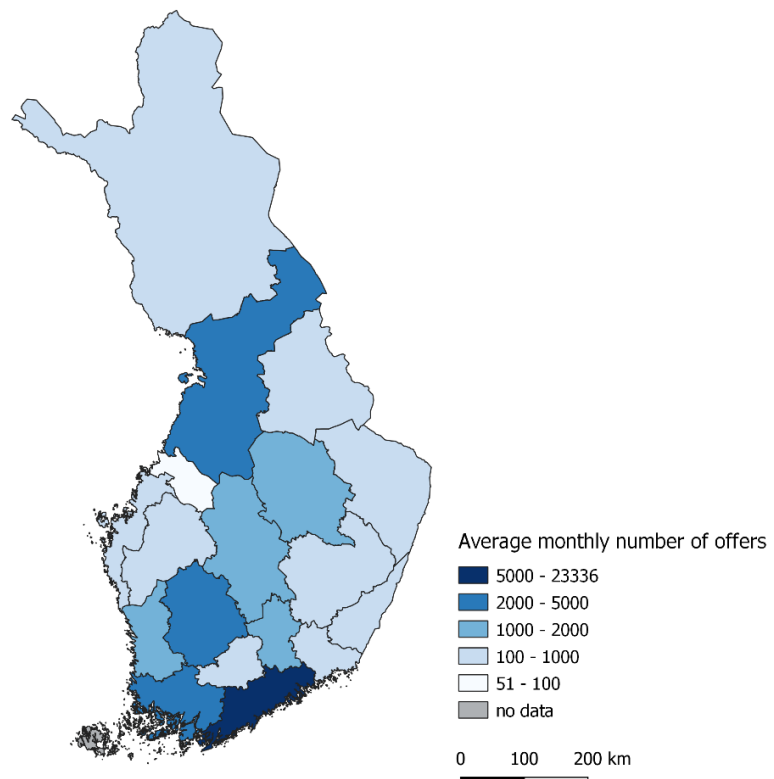


Figure 50. Average monthly number of offers for rent in Finland in 2023

7.6. France

Regarding dissemination of the results, indices obtained through those data are tentative and should be regarded as highly experimental. At this stage, it does not seem possible to use these data to produce rent trend indices. In fact, the results are fragile and the data reflect results on the evolution of market rents, but they do not carry information about the evolution of rents for tenants already in place.



Web Intelligence
Network



**Funded by
the European Union**

8. Conclusions

The use of web data in official statistics brings lots of challenges to the project. Researchers are confronted with problems that usually do not exist or have minimal impact with traditionally data acquisition methods. With traditional inputs a researcher has control on the rules and ways of collecting information from respondents. The definitions, metadata, chosen population, sample frames or data from other systems are governed by statisticians. The same applies to the tools for data acquisition: questionnaire with logically ordered questions, standardised sets of answers and well-defined classifications. All of this makes the acquired data standardised and bound within predetermined rules. However, in the case of web data, these preconditions are out of reach or hard to stick to. It may be impossible to define the specific surveyed population or match its desired definition, referring points, the possible values one variable may present, the lengths of variables, their definitions, how the data was pre-processed, deduplicated or recoded.

Although the data sources studied in this project were spread over different countries, different websites and were collected in different ways, there were similarities in the challenges faced at all these stages. This is promising because it gives the chance to develop a common methodology for obtaining and processing data in order to produce consistent information at an international level. However, the challenges faced narrow the possibilities of utilising the data, from landscaping, through data ingestion, methodology issues and calculating results. Below are presented main conclusions from the work conducted by the different partners, grouped by their topics.

Data acquisition / web scraping software

Firstly, it has to be noted that some organizations already had a pre-prepared infrastructure for web data acquisition, while others were building it from scratch. Therefore, for some countries this task consisted only on adjusting the existing IT environment/data labs to the purposes of the project, while others built not only the scrapers, but also databases. Sometimes this lack of expertise resulted in unexpected technical problems, such as server restarts during scraping, unforeseen authentication problems between different systems as well as character encoding differences. These issues were caused by internal IT environment settings and were later resolved.

During the data acquisition process organizations encountered several changes in page structures. Some of them were related to the HTML tag's location or XPATH expressions, others to major changes in a portal structure (including a complete redesign of the page). These modifications should be monitored not only by checking if specific variables have been collected, but also by checking that the values are correct. There may be a situation in which a variable, by changing the tag location or name on the page, is incorrectly completed with other data, while for the controlling process it may look correct. For instance, the number of rooms can easily be mistaken with the floor number as both pieces of information store integers with a fairly comparable range of values.

When using web data, it is also important to pay attention to the structure of individual pages within the source as they may differ from those used in the standard portal.

Methodology

The main methodology-related challenges identified within the project are: merging data from several sources, linkage with statistical and administrative sources, deduplication and compilation of aggregated data. These tasks are complex and largely depend on the quality and type of data collected.

Users of web data should be aware of the over-coverage problems and missing data in both, whole observations or single variables. Excluding the cases of duplicates, in some sources were observed records that did not fit into the scope of the project. The problem was partially overcome by implementing cleaning processes, but to fully prevent the final set from over-coverage a more expanded tool could be used. In this task, a good referring source would be helpful also for tracking outlying observations.

Moreover, web portals may use different definitions of an apartment, which affects the type of offers posted later in the category of 'apartments'. What can be more confusing is the way of counting rooms in different countries. In Germany, for example, some offers specify the number of rooms with an accuracy of half (e.g. 1.5 room, 2.5 rooms, 4.5 rooms), which has its origins in an outdated classification that counted smaller living rooms as halves, and is still used by some portals. In this particular project, this was not a problem as each room was counted as one.

When gathering data from websites one should be aware that work with unstructured data requires significant data cleaning and editing efforts. For instance, the number of rooms or the floor on which the object is located may need to be extracted from a string value. In some portals, information about the number of rooms or floor number was not stored as a raw number, but: '2 rooms', 'ground floor', '1st floor'. All of these cases can be easily resolved at the data cleaning process, but each new case will force the process to be adjusted. Same problem exists related to variables storing locations of the offers. Many city names were found to be called in multiple variations.

It also turned out that it is not possible to calculate some indicators due to the different way of presenting the information on web pages. For instance, the indicator 'average number of rooms' could not be calculated due to the different scale of measure used by portals. Number of rooms is not always present as numeric value, sometimes when the numbers gets higher the portal starts to group them in one category. In some cases it is '8 or more' or '5 or more'. What makes it impossible to calculate the average properly. Such problems are especially visible in combining data from multiple sources.

Trying to combine data from sources around various countries an issue arose in creating appropriate category levels in produced aggregated results. The problem is specifically visible on the variable Price. Maximum level in one country may be a medium price in other. Thus, trying to create a universal categories set for aggregation was problematic.

Missing data

Despite the fact that a mandatory variable is visible on the browsed offers, for some portals, not all mandatory variables were available after data collection or there was a high percentage of missing data. Dealing with missing data requires comprehensive case analysis and possibilities to impute data. For the purposes of the project, it was decided not to exclude incomplete cases, as this would mean a large loss of offers and could neglect the risk of bias. Therefore, all offers were included in the data analysis, even if there were missing data in the "mandatory variables" (e.g. price, area, number of rooms). Given the number of offers with missing data and the large variation in price differences (e.g. due to building type, size and especially address), implementing imputation techniques would require to increase the number of complete observations. This problem is particularly visible in the case of new buildings where often the price is not disclosed. It becomes an even bigger problem in rural areas, where there are fewer offers than in larger cities. Using an appropriate imputation technique for missing data, may prevent questioning the quality of the calculated results for such areas. Moreover, it is likely that due to the extremely high demand for apartments in the rental market, properties are advertised online only for a very short time (implying a scraping interval needs to be very short to cover this short-term offers) or even are being rented even

before they are listed online. Such disparities in various regions suggests that areas with sparse listings may also not provide reliable data for the project's goal.

Redundancy / duplicates

In the case of data collection based on non-standardized, privately-held data, the data provider sees its data as a set of offers, while the statistical offices are focused on obtaining information on unique dwellings. That leads to a major challenge which is deduplication, both within a given portal and between portals. Identifying duplicates and deciding whether two advertisements refer to the same object or different objects is not a trivial task, since different objects can have the same set of features, and the same object can be advertised more than once. Downloaded objects may have the same or very similar set of features - at the same or at a different time-period.

Even after collecting offers into possible groups of potential duplicates, it is also not easy to confirm their actual redundancy. Additionally, different sets of available information and/or missing data hamper identification of duplicates over time as well as across portals. Therefore, it is not possible to be certain whether data is duplicated or refers to separate objects, not to provide a general set of rules or guidelines for detecting duplicate data.

The state-of-the-art approach may involve using machine learning technics on images of the apartments. However, they were not collected during the project. Second of all, the photos published online do not always present actual apartments. In some cases, these may be computer-generated images from the planning stage typical for new apartments.

Data collected directly from the data provider

The main advantages of this type of data acquisition is obtaining more structured datasets. Sometimes they can even be adjusted to the needs of the statistical office and provided with additional information that is not visible to an ordinary user (e.g. actual location of the property, number of views of the offer, additional modification dates).

Another advantage is the greater stability of the data acquisition. Data is made available to an organisation on a regular schedule agreed with the data provider, so there are usually no unexpected delays which may occur while scraping information from the web. That kind of data acquisition is also insensitive to temporal HTML changes of the website. However, this can also be a disadvantage, as the data receiver is fully dependent on the data provider and the viability of its business. Statistical organisations need to develop good cooperation solutions that will allow for a quick response to existing problems or being informed in advance of planned changes.

Despite gathering more structured and wider scope of data, the acquired datasets indicate the same problems as scraped data regarding the linking process e.g. with tax administration's office data. This is due to the fact that information from the tax administration's office lack some locational variables such as apartments' addresses. From experiences of one of the partners, even imputing missing data by linking with data from building registers and location data pool there were still about 30 percent of house sales observations left without appropriate location information.

Another important issue is maintaining constant access to data from the private company. The company is not always interested in extending the agreement and any delay in contract negotiations may result in interruption of data series. Worth mentioning are also the costs and the limitations related to paying for data by data receivers. A mature organization that is able to obtain data from the network should have a unit that will supervise the process of negotiation or renegotiation of contracts.

Summary

Web data for real estate market that was analysed and used in the project allow us to draw conclusions about their potential use in official statistics. Not all of the data sets obtained create a possibility to allow to replace the statistics currently produced, but some countries have shown high stability level of web data and good level of reflecting the current market situation in real estate. The data that have been acquired can certainly be a supplement to the current real estate market monitoring system or be used to create leading indices where the data flow in classic researches is long. These data can also be used to build hedonic indices or models classifying real estate in new cross-sections. That is why after dealing with all above mentioned obstacles presented in this chapter, the data has high opportunity for using in real estate market studies.



Web Intelligence
Network



**Funded by
the European Union**

9. Annexes

Annex 1. Checklist for assessing web data sources ([version 2021](#))

Annex 2. Checklist for assessing web data sources ([version 2024](#))

Annex 3. Experimental statistics (available [here](#) only for the WIN and WISER members)

10. List of tables

| | |
|--|----|
| Table 1. Common criteria list for assessing utility of web data sources | 7 |
| Table 2. General concept and definitions for real estate market data | 9 |
| Table 3. Assessed real estate portals with highest score (maximum = 100) | 11 |
| Table 4. Assessed real estate portals with highest score (maximum = 100) | 12 |
| Table 5. Portals general information | 14 |
| Table 6. Assessed real estate portals with highest score (maximum = 100) | 14 |
| Table 7. Assessed real estate portals with highest score (maximum = 100) | 17 |
| Table 8. Information coverage by data source (<i>X, if the information exist in the source</i>) | 30 |
| Table 9. Information coverage in scraped datasets | 31 |
| Table 10. Number of advertisements per year | 35 |
| Table 11. Classification of mandatory variables | 36 |
| Table 12. Cleaning and editing rules | 39 |
| Table 13. Offers with same values, except floor number | 41 |
| Table 14. Offers with same values in all variables | 42 |
| Table 15. offers with exactly the same values in all variables | 42 |
| Table 16. group of offers that may contain duplicate information | 42 |
| Table 17. Offers with same values except Price | 43 |
| Table 18. Offers with very similar values | 43 |
| Table 19. Example of a duplicate offer within one data source | 44 |
| Table 20. Amount of missing information by source and variable for newly constructed offers in 2023 | 45 |
| Table 21. Number of observations available per indicator (Portal 2 and Portal 5) collected from April 2022 to January 2023 | 46 |
| Table 22. Comparison of price index in Île-de-France region | 54 |
| Table 23. The average monthly completeness ratio for selected variables | 58 |
| Table 24. The Pearson correlation coefficient and p-value between the web and official data sources - large cities in Poland | 60 |
| Table 25. The Pearson correlation coefficient and p-value between the web and official data sources - Krakow | 60 |
| Table 26. Number of offers for newly constructed objects in 2022 by portal | 61 |
| Table 27. Monthly average number of missing data in mandatory variables in all collected data | 70 |



11. List of figures

| | |
|--|----|
| Figure 1. Search result overview and specific advertisement..... | 16 |
| Figure 2. Data collection process steps | 23 |
| Figure 3. Windows Task Scheduler view | 23 |
| Figure 4. View of Bulgarian NSI's scraper log for monitoring collection of data from multiple sources | 24 |
| Figure 5. Histogram of online duration of Berlin's rental apartments in h (<1 week) | 34 |
| Figure 6. Histogram of online duration of Brandenburg's houses for sale in h (<1 week) | 34 |
| Figure 7. Extraction of predefined key words | 37 |
| Figure 8. Extraction of information by splitting the text | 38 |
| Figure 9. Script example for basic deduplication process | 38 |
| Figure 10. The iterative process of linkage dataset with the register of the territorial division (keys) | 40 |
| Figure 11. Outliers in objects for sale in Berlin | 46 |
| Figure 12. Proportion of scraped data for Sale or Rent on Portal 2 from April 2022 to July 2024, Categorised by Completeness of Address Information..... | 47 |
| Figure 13. 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 | 48 |
| Figure 14. Example of a construction project in Berlin, advertised on both Portal 5 and Portal 2 | 48 |
| Figure 15. Example of duplicates in different portals | 49 |
| Figure 16. Example of offers in a residential construction project | 50 |
| Figure 17. Example of previously purchased rental offer | 51 |
| Figure 18. Number of relocations of dwellings by the department number in Île-de-France region..... | 54 |
| Figure 19. Number of offers for sale in Bulgaria | 55 |
| Figure 20. Number of offers for rent in Bulgaria..... | 56 |
| Figure 21. Number of offers for sale in Poland | 56 |
| Figure 22. Number of offers for rent in Poland..... | 57 |
| Figure 23. Comparison between web data and official statistics for the largest cities in Poland | 59 |
| Figure 24. Comparison between web data and official statistics for Krakow | 59 |
| Figure 25. Number of offers for newly constructed objects (2023): total number by platform provider.... | 60 |
| Figure 26. Number of offers for newly constructed objects in 2023 by NUTS3 region | 61 |
| Figure 27. Number of offers (2023) for newly constructed objects by offer type and building type | 62 |
| Figure 28. Number of offers for sale and rent in Berlin | 63 |
| Figure 29. Number of offers for sale and rent in Brandenburg | 63 |
| Figure 30. Number of offers for sale on Oikotie 2019M01 to 2024M08 | 64 |
| Figure 31. Number of offers for rent on Oikotie 2019M01 to 2024M08..... | 65 |
| Figure 32. Comparison of numbers of sale observations between web data and official Finnish statistics 2020Q1 to 2023Q4 | 66 |
| Figure 33. Comparison of numbers of rent observations between web data and official Finnish statistics 2019Q1 to 2024Q2 | 67 |
| Figure 34. Comparison of house prices between web data and official Finnish statistics 2020Q1 to 2024Q2 | 68 |
| Figure 35. Comparison of apartment prices between web data and official Finnish statistics 2020Q1 to 2024Q2 | 69 |
| Figure 36. Comparison of rents between web data and official Finnish statistics 2019Q1 to 2024Q2 | 69 |
| Figure 37. Number of offers for sale per country | 71 |
| Figure 38. Number of offers for rent per country | 71 |
| Figure 39. Changes of number of offers for sale per country m/m | 72 |

| | |
|--|----|
| Figure 40. Changes of number of offers for sale per country m/m | 72 |
| Figure 41. Average monthly number of offers for sale in Bulgaria in 2023 by source..... | 73 |
| Figure 42. Average monthly number of offers for rent in Bulgaria in 2023 by source | 73 |
| Figure 43. Average monthly number of offers for sale in Poland in 2023 by source..... | 74 |
| Figure 44. Average monthly number of offers for rent in Poland in 2023 by source | 74 |
| Figure 45. Average monthly number of offers for sale in Hesse in 2023..... | 75 |
| Figure 46. Average monthly number of offers for rent in Hesse in 2023 | 75 |
| Figure 47. Average monthly number of offers for sale in Berlin-Brandenburg in 2023 | 76 |
| Figure 48. Average monthly number of offers for rent in Berlin-Brandenburg in 2023 | 76 |
| Figure 49. Average monthly number of offers for sale in Finland in 2023 | 77 |
| Figure 50. Average monthly number of offers for rent in Finland in 2023 | 77 |

