Chair of Econometrics and
Business Statistic
Faculty VII - Economics and Management
Technical University of Berlin

Seminar: Applied Environmental Econometrics in R

The Causal Effect on Urban Residential Property Prices of Proximity to Air-Emitting Production Facilities : Evidence from Germany

Submitted by: Maria Baschun
Matriculation number: 385296
Programme: Master of Science in Statistics
Examiner: Prof. Dr. Astrid Cullmann
Supervisor: PhD student Marco David Schmandt

Date of Submission: 27.11.2024

Declaration of Originality

Declaration of Originality I hereby declare that the present seminar paper and the work reported herein was composed by and originated entirely from me without any help. All sources used from published or unpublished work of others are reported in the list of references. All parts of my work that are based on others' work are cited as such. This seminar paper has not been submitted for any degree or other purposes, neither at the Technical University of Berlin nor at any other university or college.

Berlin, 27.11.24

Maria Baschun

[Name and surname] [Signature]

Abstract

Production facilities located near residential properties can negatively affect property values due to environmental externalities, such as pollution and hazard risks, as well as aesthetic externalities, such as visual disamenities. Focusing specifically on air-emitting production facilities, this study uses German emission data (2007–2022) from the PRTR and residential property sale data (2007–2022) from RWI for the 15 largest cities in Germany. The analysis concentrates on the 10-year period from 2012 to 2021 and examines the influence of production facility proximity on property values within the same 1-km^2 raster cell. By employing propensity score matching with a nearest-neighbor algorithm, a suitable control group was created, and the Average Treatment Effect on the Treated (ATT) was then estimated at -0.0242 through regression on the matched sample. This indicates that properties located within the same 1-km^2 raster cell as emitters are, on average, 2.42% cheaper due to their proximity to the emitter. These findings highlight the causal economic implications of environmental and aesthetic externalities on the housing market.

Contents

List of Tables

List of Figures

Abbreviations

Other abbreviations

ATT	Average Treatment Effect on the Treated
CIA	Conditional Independence Assumption
E-PRTR	European Pollutant Release and Transfer Register
K-NN	K-Nearest Neighbors
NO_2	Nitrogen Dioxide
PRTR	Pollutant Release and Transfer Register
PS	Propensity Score

Mathematical symbols

α	Alpha
β	Beta

1. Introduction

The proximity of residential properties to production facilities and industrial areas is an important topic in urban economics due to its potential impact on housing prices. Production facilities can reduce property values through visual disamenities, pollution, and hazard risks. Understanding these effects is critical, especially in urban environments where industrial and residential zones often coexist.

This seminar paper examines the extent to which proximity to production facilities affects residential property values in Germany's 15 largest cities, focusing specifically on air-emitting facilities. Using longitudinal data from the German Pollutant Release and Transfer Register (PRTR) and a dataset on residential property sales from the RWI Leibniz Institute for Economic Research, this study spans the 10-year period from 2012 to 2021. The analysis focuses on properties located in the same 1-km^2 raster cell as emitters and employs a robust methodology to estimate the causal impact of industrial proximity on housing prices.

A key challenge in this research arises from the non-random assignment of treatment. Apartments near production facilities systematically differ from those further away; they are, on average, older, smaller, less likely to have a balcony, in worse condition, and more frequently located in less central or more industrial areas. These differences introduce selection bias, making simple comparisons between treated and control groups unreliable. For example, untreated apartments—those not located in the same 1-km^2 raster cell as an emitter—were, on average, 9.37% more expensive, a difference that cannot solely be attributed to industrial proximity. To address this, propensity score matching is employed, using a logit model to account for covariates such as postcode-level location, year of construction, living area, apartment condition, and other factors influencing treatment status or the outcome variable, i.e., the log price per m^2. The model's predictive power was evaluated by assessing the share of correct predictions, which was 90% based on a cut-off using the share of treated observations. This approach ensures the creation of comparable groups, isolating the effect of proximity to production facilities from other confounding factors.

The findings of this study indicate that residential properties located in the same 1-km^2 raster cell as air-emitting production facilities are, on average, 2.42% cheaper. This reduction represents a causal effect of proximity to air-emitting facilities, isolated from confounding factors through the use of propensity score matching. These results highlight the economic implications of environmental externalities, providing valuable insights for urban planners, policymakers, and environmental economists.

2. Data

2.1. Data Cleaning

2.1.1. Real-Estate Dataset

The data comes from the RWI Leibniz Institute for Economic Research, specifically from the Panel Campus File `RWI-GEO-RED Panel` Version V3.1, and was kindly provided by the Department of Econometrics and Economic Statistics at TU Berlin. It is based solely on residential advertisements and covers the period from January 2007 to December 2021.

The Panel Campus File data is organized into three separate datasets: houses for sale, apartments for sale, and apartments for rent. Each dataset covers the 15 largest cities in Germany. Additionally, there is a small dataset containing the number of observations per city and dataset in the Panel Campus File. For the purposes of this seminar paper, the dataset on apartments for sale and the small dataset with the number of observations were used. The former includes regional information and a rich set of housing characteristics. Table 2.1 presents all the variables from the Panel Campus File considered in this seminar paper.
The raw dataset on apartments for sale contained 927,319 observations. First, we retained only those observations with an end date between 2012 and 2021. By considering this 10-year period, the number of observations dropped to 579,654. In the following description, we consider these 579,654 observations as 100%. The data were processed as follows:

- Objects that have the same object ID but are located in different 1-km^2 raster cells were removed, as the 1-km^2 raster cell is used to identify the group to which the object belongs, either in the treatment or control group. In this step, we lost 0.60% of the data.
- Only 15 out of the initial 72 variables were retained from the dataset on apartments for sale, specifically all variables represented in table 2.1, except for the city name, which comes from the smaller dataset containing the number of observations. The choice of variables will be justified in the Methods chapter.
- There were many observations with missing data labeled as "Other missing," which were removed with the exception of the floor variable. After removing missing values, we lost 15.05% of the data. The floor variable was renamed as "unknown" when an observation was missing. Otherwise, we would have lost 27.70% of the data alone by removing observations with missing floor values. We kept these observations, despite the missing data, because we did not use the floor variable for the estimation of the propensity score, but only as a control variable.

Name of the variable	Short description	Detailed description
obid	Object id	Each property is uniquely identified by an artificial ID number. IDs are property-specific and do not change over time even if the object is temporarily withdrawn from the pool of advised real estates and offered again at a later time.
plz	Postcode	The postal code the object is located in.
kaufpreis	Purchasing price in EUR	Price at which the owner advertises to sell the object, expressed in EUR.
baujahr	Year of construction	Year in which the object was built. Observations that lie in the future are not necessarily faulty entries, potentially indicating that an object is still under construction.
wohnflaeche	Living area in m^2	Living space in square meters.
zimmeranzahl	Number of rooms	Number of rooms, excluding kitchen, bath, or corridors.
blid	State	German federal state where the object is located in.
adat	Start date	This is a numerical variable that refers to the month and the year during which an object is first advertised.
edat[1]	End date	This numeric variable refers to the month and the year of the end of the advertisement.
gid2019[2]	Municipality identifier	This is the municipality identifier according to the German Official Municipality Key. It is based on the territorial definition of 2019 (end of year).
price_sqm	Price per m^2 in EUR	Calculated price per square meter by price and size of apartment.
balkon	Balcony	This variable indicates the presence of a balcony.
ergg_1km	1-km^2 raster cell	This variable indicates the grid cell of a 1-square-km raster of Germany according to the INSPIRE guideline. Addresses are matched to this raster based on their geocoded location.
objektzustand	Condition of object	Each property is assigned exactly to one of 11 possible conditions.
etage	Floor	Floor on which object is located.
city_name	City	Name of the city the object is located in.

Tab. 2.1.: Descriptions of Utilized Variables from the Panel Campus File

[1]Some apartments were advertised before they were built, and therefore the year of construction is later than the end date of advertising. [2]Municipality identifier is unique for each of the 15 biggest cities covered in the Panel Campus File.

- There were many observations with the same object ID, and the goal was to ensure that each apartment, i.e., each object ID, was present only once in the dataset. The following strategy was pursued: If two or more observations had the same values for the variables object ID, postcode, year of construction, living area, number of rooms, state, municipality identifier, balcony, 1-km^2 raster cell, condition of the object, and floor, but different or the same purchasing price, start date, end date, or price per square meter, we retained only the observation with the latest end date (see table 2.1). If two or more observations with the same object ID had different values for the postcode, year of construction, living area, number of rooms, state, balcony, 1-km^2 raster cell, municipality identifier, condition of the object, or floor, we deleted all observations with that object ID, since these characteristics are typically time-invariant for an object, except for the condition of the object. In this step, we lost another 99,354 observations, which corresponds to 17.14% of the initial dataset. It is important to note that only 13,714 unique object IDs were eventually removed. The remaining removed observations were either duplicates of the removed objects or duplicates of the kept objects. The new dataset contained 389,605 observations, with each object ID being unique.

- In the next step, we modified the variables (i.e., renamed, changed the type, merged new variables, and aggregated the values).
  -The balcony variable was converted from *Yes* and *No* to numeric values, with *Yes* becoming 1 and *No* becoming 0.
  -The type of all variables from table 2.1, except for the 1-km^2 raster cell, condition of the object, floor, city, and state, was changed to numeric.
  -The current dataset, which was previously composed only of data from the dataset on apartments for sale, was complemented with the variable city name by merging it with the dataset containing the number of observations. The merge was performed using the municipality identifier, which is present in both initial datasets.
  -Initially, 11 values of the object's condition were aggregated into 3 categories: *good*, *bad*, and *not specified*, with *not specified* being the chosen reference category (see the exact aggregation in the table A.1).
  -The initial 44 values of the floor variable were aggregated into 8 character-type values: *-1*, *1*, *2*, *3*, *4-6*, *7+*, *Implausible value*, and *Unknown*, with *Unknown* being the chosen reference category.
  -Two new variables were derived from the end date of the advertisement: the month and the year of the end of the advertisement.

- The last step in cleaning the Real-Estate data involved retaining only one of each pair of objects that were suspiciously similar, specifically keeping the one with the latest end date of the advertisement and removing the other. Two objects were considered "suspiciously similar" if the year of construction, living area in square meters, number of rooms, state,

municipality identifier, balcony value, 1-km^2 raster cell, condition of the object, purchasing price, price per square meter, city, floor, and the end year of the advertisement were the same, but the object ID and at least one of the following characteristics were different: start date of the advertisement, postcode[1], or the end month of the advertisement. Thus, it was more reasonable to assume that it was the same apartment rather than assuming the opposite. In this step, 15.02% of the original dataset was removed.

To conclude, after filtering the observations to include only those with an end date within the period from 2012 to 2021, we were left with 579,654 observations, which we consider as 100%. During the data cleaning process, we lost 47.81% of the data, and the further analysis will be conducted on a dataset of 302,438 observations, where each object ID is unique. It is important to note that the 47.81% loss refers to the number of removed observations only, regardless of whether these observations were duplicates of the kept ones or not. However, if we focus solely on unique object IDs in the initial dataset covering the years 2012 to 2021 and compare it with the cleaned data, we find that we lost only 32.79% of the unique object IDs and continue working with 67.21%.

2.1.2. Emissions Dataset

The data on emissions came from the PRTR and was made publicly accessible by the Umweltbundesamt on November 2, 2023. First, we downloaded the data `PRTR-Gesamtdatenbestand` as a ZIP archive. This ZIP archive contained four files with the PRTR data for releases, shipments with wastewater, disposal of hazardous waste, and disposal of non-hazardous waste. Each of the datasets covers the period from 2007 to 2022. We used the data on releases titled `2023-12-08_PRTR-Deutschland_Freisetzungen`. The variables used for this seminar paper are represented in table 2.2. There were also other variables, such as industry sector, type of pollutant, company activity, and others. The data description can be downloaded from the same link as the emissions data itself (Umweltbundesamt, 2023).

After downloading the raw dataset with 71,708 observations, we first filtered the data to include only air as the environmental compartment, reducing the number of observations to 42,792. The idea is to measure the optical and polluting effects caused by air emissions, as they are more perceptible than emissions in water or soil to people living near the facility. In the next step, using the `sgo` library in `R`, we converted the latitude and longitude coordinates into the equal area grid coordinates `ergg_1km` that are also present in the Real-Estate Dataset (see table 2.1). For all filtered observations, the reported annual release in kg/year was always greater than or equal to the pollutant threshold value, with the exception of 24 observations for which the pollutant threshold value was unavailable.

1 Take into account that the observations still share the same 1-km^2 raster cell.

Name of the variable	Short description	Detailed description
umweltkompartiment	The environmental compartment	The environmental compartment (water, air, or soil), into which the substance was released.
lat	Latitude	Y-coordinate of the facility's location in the GPS coordinate system.
lon	Longitude	X-coordinate of the facility's location in the GPS coordinate system.
betriebsname	Name of the company	Name of the emitting company.
jahr[1]	Year	Reporting year.
schadstoff_schwellenwer[2]	Pollutant threshold value	Minimum value, above which the released substance must be reported to the PRTR by the company.
jahresfracht_freisetzung	Annual release in kg/year	Amount of substance released this year in kg/year.

Tab. 2.2.: Descriptions of Utilized Variables from the Emissions Dataset

[1]The reporting year coincides with the year in which the emission took place. [2]The threshold values vary depending on the pollutant.

We used the latitude and longitude combination as a unique identifier of the emitting production facility. Sometimes, for each latitude and longitude combination, there were different observations with approximately the same name, e.g., "Kraftwerk Grenzach-Wyhlen" and "Kraftwerk Grenzach-Wyhlen KGW GmbH". However, since the decimal degree coordinates were provided with five digits after the decimal, maintaining 1-meter accuracy, we chose to use the unique identifier without considering the company name (Humboldt State University, 2018).

Most combinations of coordinates appeared several times in the dataset, e.g., due to different pollutants emitted from the same facility or different reporting years for the same facility. Thus, for each combination of latitude and longitude, we took all related observations and defined two new variables, `min_year` and `max_year`. The first gives the earliest reporting year, and the second gives the latest reporting year contained in the dataset for the unique combination of latitude and longitude.

Finally, we created the dataset where, for each unique combination of latitude and longitude, we added the following characteristics: 1-km^2 raster cell[2], earliest reporting year, latest

2 I thank Isabell Fetzer for providing the code to generate the ergg variable from longitude and latitude coordinates.

reporting year, and the name of the company. These characteristics were unique for each combination of latitude and longitude, with the exception of the company name. To generate a unique company name for each observation, we randomly selected the first name from the list of company names corresponding to the location in cases where several names were available. The created dataset contained 5,052 observations, which we used to create the treatment status for each of the apartments for sale.

2.1.3. Creation of Treatment Status

If a production facility's earliest emitting year is t and the latest emitting year is $t+n$, it is assumed that it has been emitting throughout the entire period from year t to $t+n$. For instance, if the earliest reporting year for a facility was 2010 and the latest reporting year was 2013, we assume that this production facility was emitting in 2010, 2011, 2012, and 2013.

For each year from 2007 to 2021, we created a subset of the emissions dataset. Specifically, for each year t, the subset `emissions_t` includes only those facilities where the earliest reporting year is less than or equal to t and the latest reporting year is greater than or equal to t. In other words, the subset `emissions_t` contains data for facilities that were operational in the year t. The subsets for different years are not disjoint; on the contrary, they have many observations in common. This is because a facility could be active in several years, which is the case in 3,672 out of 5,052 cases.

After building the subsets of operational facilities for each year t, we assigned the apartments for sale to either the control or treatment group. If an apartment with an advertisement ending in year t is in the same raster cell as at least one facility that was active that year[3], the apartment was assigned to the treatment group; otherwise, it was assigned to the control group. The new binary variable `emissions` was created with corresponding values 0 and 1.

Some facilities did not have any apartments for sale in the raster cell where they were located during the years they were active. These are referred to as "Non-affecting facilities" and are to be distinguished from "Affecting facilities", which were active and had apartments for sale nearby[4]. Ultimately, we want to measure how such an "Affecting facility" affects the price of apartments that are in the same 1-km^2 raster cell. In Figure 1, all facilities in 2020 are shown for the four cities in Germany where the most apartments for sale were treated that year. In 2020, the number of treated apartments for sale was 97 in Berlin, 62 in Düsseldorf and München, and 54 in Hannover.

3 I.e., a facility from the subset `emissions_t`.

4 Meaning in the same 1-km^2 raster cell.

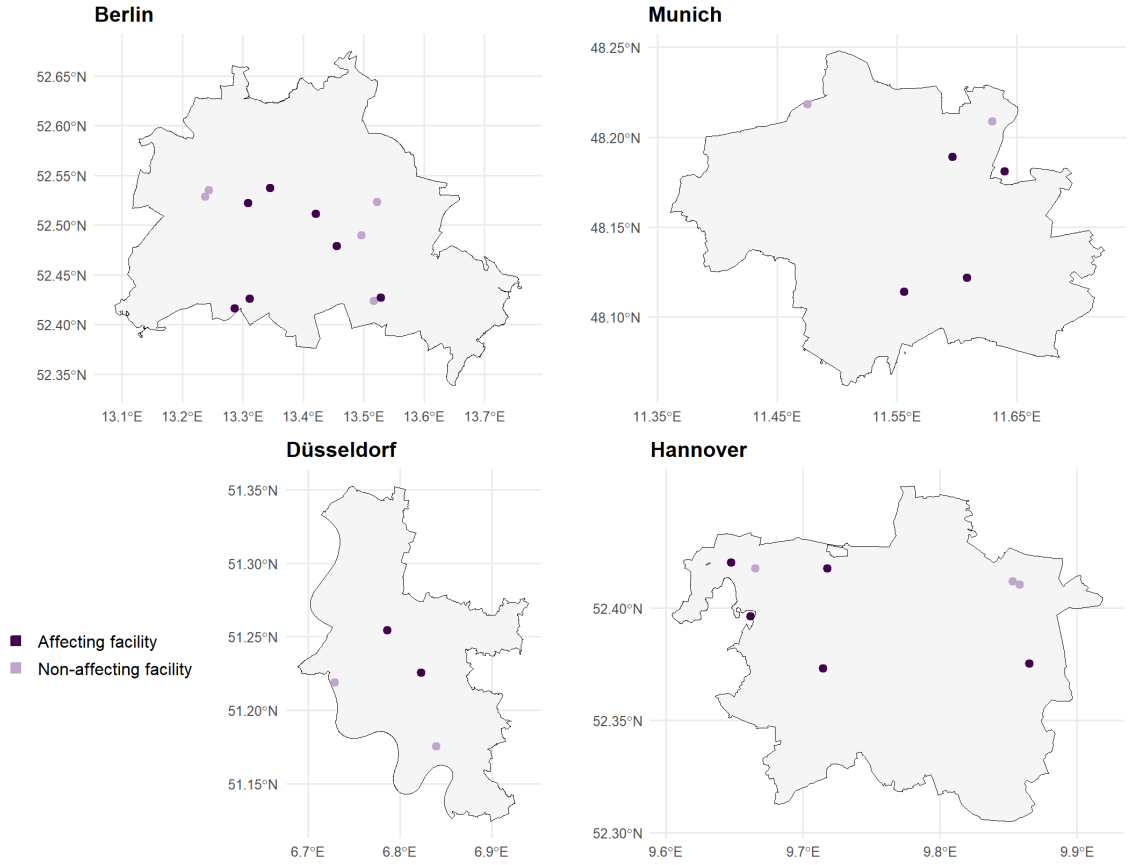


Fig. 2.1.: Affecting and Non-affecting Facilities in 2020

After defining the treatment status, we obtained the cleaned dataset that will be used for the estimation of the propensity score and subsequent matching. The summary statistics of the cleaned dataset are presented in 2.3.

Numerical Variables	Mean	Std. Deviation	Minimum	Maximum
Year of construction	1971	38.62	1500	2022
Living area in m^2	82.76	34.09	25.89	230.00
Number of rooms	2.84	1.06	1	10
Balcony	0.7785	0.4153	0	1
Purchasing price in EUR	328,922	242,796	16,800	1,990,000
Price per m^2	3,874.52	2,115.08	406.50	13,017.86
Advertisement end year	2017	2.86	2012	2021
Advertisement end month	6.67	3.49	1	12
Emissions	0.0177	0.1320	0	1
Number of observations	**Treatment group**	**Control group**	**Total**	
Absolute values	297,076	5,362	302,438	
Share	98.23%	1.77%	100%	

Tab. 2.3.: Summary Statistics for Cleaned Data

Other categorical variables from the cleaned dataset that are not represented in the table above include: state, municipality identifier, 1-km^2 raster cell, condition of object, object ID, city name, postcode, floor, and start and end date of the advertisement.

3. Methods

3.1. The Idea Behind Matching

In a randomized experiment, treatment and control groups are selected randomly, ensuring that they are balanced with regard to all observed and unobserved factors that could influence the outcome. This balance allows any differences in the outcome to be attributed solely to the intervention (Matteucci Gothe, 2023). However, in non-experimental data, such balance is not guaranteed, which can lead to selection bias (Dehejia & Wahba, 2002). Observational studies using existing data are often used to estimate the Average Treatment Effect on the Treated (ATT) when randomization is not possible or ethical (Rosenbaum, 2010).

The goal of this study is to measure the ATT of having a production facility located near an apartment for sale. The outcome variable is the log-transformed price per square meter of the apartment, which reflects the percentage decrease in price associated with the treatment. The production facility could affect apartment prices through both visual impacts and emissions effects.

Table 3.1 displays the mean differences in the unmatched dataset. These differences indicate that the covariates are unbalanced between the treatment and control groups, suggesting that the price per square meter would differ even in the absence of the treatment. This selection bias can lead to inaccurate estimates of the ATT if only simple mean differences between the treatment and control groups are used. Specifically, apartments in the treatment group are, on average, 9.37% cheaper, and this observed difference cannot be entirely attributed to the presence of the production facility. In this study, the assignment of apartments to the treatment group is not random. It is assumed that treated apartments are, on average, located in less central or more industrial areas, which could significantly influence their price. To isolate the effect of proximity to the production facility from location factors, the location of the apartment must be considered. Additionally, apartments in the treatment group are, on average, older, smaller, and less likely to have a balcony, which could be responsible for the lower average price per square meter. Conversely, these apartments have fewer rooms, which might increase the price per square meter, as the number of rooms has a significant negative effect on the price when regressing the log-price per square meter on all covariates from table 2.1.

The objective of matching is to balance relevant covariates between the treatment and control groups by combining comparable observations that are similar in all relevant characteristics except for treatment status, and excluding observations that could not be matched. This approach aims to provide an unbiased estimate of the ATT (Caliendo & Kopeinig, 2005).

Variables	Treatment	Control	Difference	t value	p value
Balcony	0.725	0.779	-0.054	-8.782	0.000
Year of construction	1969.801	1970.592	-0.791	-1.505	0.132
Purchasing price in EUR	287,539.275	329,668.766	-42,129.492	-13.926	0.000
Price per m^2 in EUR	3,617.455	3,878.144	-260.689	-9.209	0.000
Living area in m^2	77.915	82.843	-4.928	-10.684	0.000
Number of rooms	2.727	2.843	-0.116	-7.785	0.000
Propensity Score	0.259	0.013	0.246	111.241	0.000
Advertisement end year	2016.444	2016.524	-0.081	-2.101	0.036

Tab. 3.1.: Sample Means and Their Differences for Covariates in the Unmatched Dataset

A two-sided t-test on the difference in means was conducted with the null hypothesis $H_0 : mean_0 = mean_1$, taking into account unequal variances. Accordingly to the p-values, the null hypothesis H_0 would be rejected at a confidence level of $\alpha = 5\%$ for all covariates, except for the year of construction. A t-test cannot be performed for categorical variables with more than two levels. Therefore, the distribution of variables regarding the location and condition of the apartment will be examined in section 3.6 without performing a t-test.

3.2. Notation and Formal Definition of Parameter of Interest

Determining the effect of a treatment on an outcome requires considering what the outcome would have been if the object had not received the treatment. The potential outcome approach, as described in the Roy-Rubin model, is the standard framework used to formalize this concept (Caliendo & Kopeinig, 2005). Further notation in this section is consistent with the notation described by Caliendo and Kopeinig, 2005.

In our case, the outcome variable is the logarithmic price per m^2. Thus, the random variable $Y := \log(\, price\, per\, m^2)$ is the outcome variable, and D is the treatment status, where $D \in 0, 1$. We also know that Y_i is dependent on D_i because we expect that the price per m^2 differs if the apartment is located near the production facility. Therefore, we can say that Y_i is the outcome variable of an object $i = 1, \ldots, 302\,438$. The treatment effect for an object i can be written as:

$$Treatment\, Effect_i = Y_i(1) - Y_i(0) \tag{3.1}$$

That is, the price per m^2 of an apartment when treated minus the price per m^2 of the exact same apartment when not treated. In this way, we would receive the exact effect of the treatment on the price. The problem with this approach is that one of these terms is counterfactual and therefore unobservable, because the apartment is either located near the facility at the time it is being sold or it is not. Because our focus is on how the price of

apartments near a production facility is affected, we are considering the treatment effect on the treated. For the treated apartment i, it would be:

$$Treatment\ Effect_i = Y_i(1) - Y_i(0), \text{ given } D_i = 1 \tag{3.2}$$

We are interested in how the treatment affects the price per m^2 of a treated apartment on average, rather than its effect on a specific object i. Therefore, the parameter of interest is the Average Treatment Effect on the Treated. This can be written as:

$$ATT = E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1], \tag{3.3}$$

where $E[Y(1) \mid D = 1]$ is the expected outcome for treated apartments, and $E[Y(0) \mid D = 1]$ is the counterfactual expected outcome for the same apartments had they not been treated. The second term is unobservable and thus must be estimated. What we can observe is the logarithmic price per m^2 of the apartments from the control group when they are not treated, i.e., $E[Y(0) \mid D = 0]$. The difference between these two terms represents the selection bias, which can be formally expressed as:

$$Selection\ Bias = E[Y(0) \mid D = 1] - E[Y(0) \mid D = 0] \tag{3.4}$$

In non-experimental studies, certain identifying assumptions need to be made to circumvent the problem of selection bias (Caliendo & Kopeinig, 2005).

3.3. Assumptions

To use the propensity score (PS) and perform matching based on it, two assumptions must be made. The first assumption is the Conditional Independence Assumption (CIA). It states that, after accounting for certain observed characteristics X of an object, the treatment assignment D is independent of the potential outcomes $Y(0)$ and $Y(1)$ for all values of X. This implies that selection into the treatment depends only on observable characteristics X, and once X is given, the selection into treatment is random. The CIA is a strong assumption because it requires that all factors influencing both the treatment assignment and the outcomes are observed and included in the analysis. This assumption ensures that $E[Y(0)|D = 1, X] = E[Y(0)|D = 0, X]$ for all values of X, meaning the selection bias from equation (3.4) is zero given the observed characteristics. However, if this assumption is violated, the estimated ATT may be biased (Caliendo & Kopeinig, 2005).

It is important to note that the matching process becomes challenging when conditioning on all relevant covariates in a high-dimensional vector. This is due to the curse of dimensionality, where the number of potential combinations of covariates increases as more variables are

added, making it difficult to find matching objects with similar covariate profiles (Caliendo & Kopeinig, 2005). To address this issue, Rosenbaum and Rubin, 1983 introduced the concept of balancing scores. They demonstrated that if the CIA holds, meaning the potential outcomes are independent of the treatment when conditioned on covariates X, then they are also independent when conditioned on a balancing score $b(X)$. The PS, defined as $P(D = 1|X)$, or in short notation $P(X)$, represents the probability of an object receiving the treatment based on its observed characteristics X and serves as one example of a balancing score (Rosenbaum and Rubin, 1983, p.42). The PS is useful for dimensionality reduction, specifically reducing the problem to one dimension (Dehejia & Wahba, 2002). The CIA can be expressed in terms of the PS as follows (Caliendo & Kopeinig, 2005):

$$Y(0), Y(1) \perp\!\!\!\perp D \mid P(X), \text{for all values of } X \tag{3.5}$$

The second assumption is the common support or overlap condition. This condition prevents perfect predictability of the treatment assignment D based on the covariates X, formally stated $0 < P(D = 1 \mid X) < 1$. It ensures that objects with the same values of X could potentially be in the treatment and control groups (Caliendo & Kopeinig, 2005). Without this condition, we wouldn't be able to find a match from the other group, as the assignment of such objects would be predetermined, resulting in no overlap between the treatment and control groups for those X values. For this seminar paper, we will assume that the common support assumption, the CIA, and therefore equation (3.5) hold.[5]

3.4. Estimation of the Propensity Score

As mentioned before, the PS is the conditional probability of receiving the treatment, given characteristics X, regardless of actual treatment status, where X is a random vector with dimension[6] n. To estimate the PS, we first need to estimate the coefficients $\beta_0, \ldots, \beta_n$. Assuming that the conditional probabilities of being treated $P(D = 1|X)$ are based on the cumulative logistic distribution $F(z) = \frac{e^z}{1+e^z}$, $z \in \mathbb{R}$ the coefficients β_0 and $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_n)^\top$ can be iteratively estimated from $F(\beta_0 + \boldsymbol{\beta}^\top X)$ and sample data by using maximum likelihood method, as advised by Cunningham, 2021. After computing $\widehat{\beta_0}$ and $\widehat{\boldsymbol{\beta}}$, the PS for each object i can be calculated as the fitted values from the logistic regression model (see Cunningham, 2021 and Matteucci Gothe, 2023):

$$\widehat{P}(X_i) = \widehat{P}(D = 1|X_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}^\top X_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}^\top X_i}}, \; i = 1, \ldots, 302\,438 \tag{3.6}$$

5 Please note that instead of the CIA, to estimate the ATT, it would also be sufficient to assume that only $Y(0)$ is independent of the treatment D given $P(X)$ for all values of X (Caliendo & Kopeinig, 2005). This is a weaker assumption than the one stated in equation (3.5).

6 In this seminar paper, $n = 1590$, including 1584 dummy variables for the postcode, 2 dummy variables for the condition of the object, and 4 numerical variables.

An important aspect is the choice of variables included in the model. Some authors recommend including only variables that simultaneously influence both the treatment status D and the outcome Y. However, other authors argue that a variable should only be excluded from analysis if it is either unrelated to the outcome or not a suitable covariate (Caliendo & Kopeinig, 2005).The choice of variables for this seminar is based on the recommendations by Matteucci Gothe, 2023, who outlines which variables should be included and which should not.

The model should primarily include variables that affect both the outcome Y and the treatment status D. In our case, these variables are the year the advertisement ended and the postcode. The treatment status depends on the year the advertisement ended, as the proportion of treated apartments varies over time, making it more likely in certain years for apartments to be treated. For instance the share of objects receiving treatment was 2.2% in 2015 and only 1.5% in 2020. The location is also significant, as apartments for sale in specific postcodes, such as those in industrial or non-central areas, are more likely to be treated.[7] Price per m^2 in EUR also clearly depends on the year the apartment was offered for sale and on its location.

Confounders of the outcome variable Y should also be included. In our analysis, these variables are the year of construction, living area in square meters, presence of a balcony, and the condition of the property, as they all affect the price of the apartment. We did not include other variables that affect only Y and not D due to missing values and to avoid overfitting, as overfitting can lead to several issues. Such issues could be an exacerbated common support problem, where certain values of $P(X)$ are present in only one group, or increased variance of the estimates (Caliendo & Kopeinig, 2005).

Variables that perfectly predict the treatment assignment should be excluded. Therefore, we did not include the 1-km^2 raster cell variable, as it directly defines the treatment status, and instead considered the postcode. Finally, variables included in the model should not be influenced by the treatment itself, which is ensured by the selection of the mentioned variables.

In table 3.2, all variables used in the logit model for the estimation of the PS are presented, namely the characteristics X and the treatment variable D. The estimation of the PS was performed using R. The code can be found in Appendix B.1.

7 We assessed the share of correct predictions using the share of observations receiving treatment as a cut-off. If the estimated PS was below or equal to this share, we predicted that the object was from the control group, and from the treatment group otherwise. The share of correct predictions was 90% with the postcode variable included in the logit model, compared to only 56% without it. This confirms that location has strong explanatory power for the treatment status of apartments, and therefore, a model that accounts for it should be used for the estimation of the PS.

Variables	Variable	Description
D	Emissions	Treatment, binary variable
X_1	Advertisement end year	Numerical variable
X_2	Year of construction	Numerical variable
X_3	Living area in m^2	Numerical variable
X_4	Balcony	Binary variable
X_5	Bad condition of the property	Binary variable[1]
X_6	Good condition of the property	Binary variable[1]
$X_7 - X_{1590}$	Postcode	Binary variables[2]

Tab. 3.2.: Variables used in the Logit Model for the Estimation of the PS

[1]Not specified condition of the property is the excluded reference category. [2]A categorical variable with 1585 levels was converted into 1584 dummy variables with one reference postcode excluded. The handling of categorical variables and their transformation into dummy variables, with a focus on logistic regression, is described by Hosmer et al., 2013 on page 36.

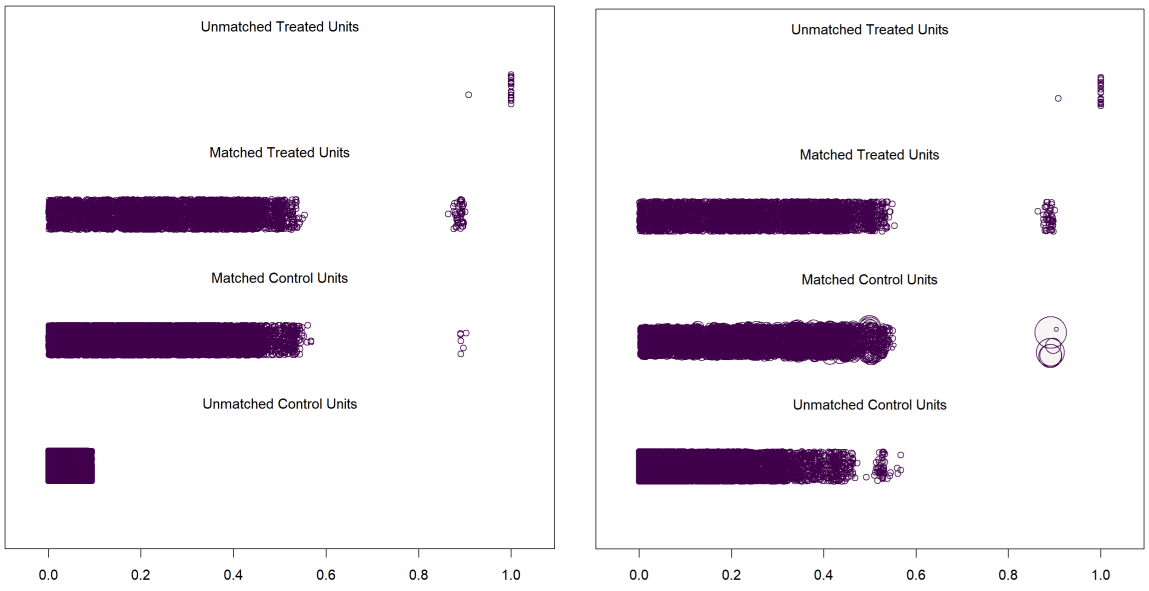
3.5. Matching Method and Model for ATT Estimation

Once the propensity score was generated, various matching algorithms could be applied to match comparable units (Caliendo and Kopeinig, 2005, p.8). In our analysis, we applied the 3-Nearest Neighbor (3-NN) matching method, both with and without replacement. Both procedures were carried out as follows: first, 76.21% of the units with propensity scores outside the common support range were discharged from the analysis, and then three apartments from the control group were matched to each treated apartment based on the closest propensity scores, respectively. This resulted in a balanced dataset, ensuring that, on average, the comparison units were more similar to the treated apartments in the characteristics described in table 3.2.

The choice of K in K-nearest neighbors (K-NN) involves a trade-off between bias and variance. A higher K increases bias because the quality of matches may decline, but it decreases variance by increasing the size of the matched sample (Caliendo & Kopeinig, 2005). It is generally recommended to use more than one nearest neighbor (Caliendo & Kopeinig, 2005). In our analysis, we selected 3-NN based on a visual examination of the distribution of propensity scores, as illustrated in figure 3.2. When we tested K=4 or higher order, we observed a noticeable decline in the quality of matching, as the distributions of PS in the treatment and control groups were less similar. This led us to compare K=3 and K=2, with the former being eventually preferred to maximize the number of observations used and, consequently, to increase the precision of the ATT estimator. Since the ratio of treated to controls in the unmatched sample was approximately 1:56, using three matches for each unit in the treatment group was feasible.

The decision to match with or without replacement also involves a trade-off between variance and bias. In cases of substantial overlap in the distribution of propensity scores between the treatment and comparison groups, most matching algorithms tend to produce similar results. However, when the propensity scores for treated and control units differ significantly, achieving appropriate matches without replacement can be challenging (Dehejia & Wahba, 2002). Specifically, once all the "good" matches are utilized, the remaining treated units must be paired with control units that are considerably dissimilar. In such situations, matching with replacement becomes the preferred approach (Dehejia & Wahba, 2002).

The distribution of propensity scores in our dataset for the different groups is illustrated in figure 3.1. On one hand, there is a significant overlap in the distributions of the treatment and control groups among the matched units. On the other hand, when examining the matched units with propensity scores higher than 0.8, it is evident that the number of these units in the control group is smaller than in the treatment group, which means that some treated units were matched with control units that had significantly lower propensity scores. To avoid obtaining "too different" matches, which could lead to biased estimates of the ATT, matching with replacement is advisable. As shown in figure 3.1, when allowing for replacement, matched units from the control group with propensity scores above 0.8 are matched multiple times, as indicated by the size of the unit circles. That increase the quality of matches and thus of the estimate.



(a) 3-NN without replacement **(b)** 3-NN with replacement

Fig. 3.1.: Distributions of Propensity Score Across Different Groups

Note: "unmatched" units in this figure also include the discharged units outside of the common support.

If we were able to perfectly match on all features that influence both the treatment status and the outcome variable, because equation (3.5) holds, a simple linear model would yield an unbiased estimate of the ATT, which would be equivalent to simply taking the mean

difference in logarithmic price per m^2 between the treatment and control groups in the matched sample. However, "perfect" matching on PS is not achievable in practice. Figure 3.2 shows the distribution of the propensity scores before and after matching.

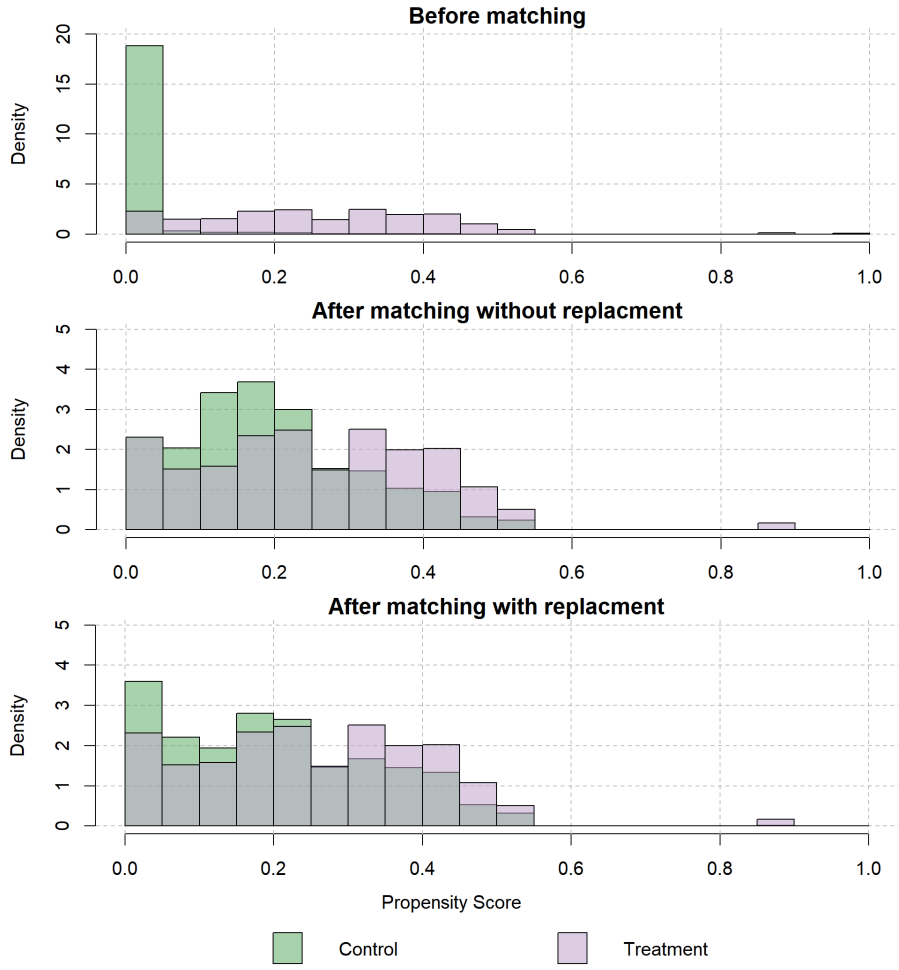


Fig. 3.2.: Distribution of Propensity Score Across Different Samples

Although the distributions are more similar after matching, they still differ, indicating that a simple linear model alone will not provide a correct estimate, and we need to control for covariates when estimating the ATT. Therefore, we use a multiple log-linear model with standard errors clustered at the postcode level, as follows[8] :

$$\begin{aligned}\log\left(Y_{\text{price/m}^2}\right) = \alpha &+ \beta_1 \, \mathrm{D}_{\text{emissions}} + \beta_2 \, \mathrm{X}_{\text{year built}} + \beta_3 \, \mathrm{X}_{\text{area in m}^2} + \beta_4 \, \mathrm{X}_{\text{number of rooms}} \\ &+ \beta_5 \, \mathrm{D}_{\text{balcony}} + \beta_6 \, \mathrm{D}_{\text{bad condition}} + \beta_7 \, \mathrm{D}_{\text{good condition}} + \beta_8 \, \mathrm{D}_{\text{floor -1}} + \beta_9 \, \mathrm{D}_{\text{floor 0}} + \cdots \\ &+ \beta_{15} \, \mathrm{D}_{\text{floor 7 or higher}} + \beta_{16} \, \mathrm{D}_{\text{ad end year 2013}} + \beta_{17} \, \mathrm{D}_{\text{ad end year 2014}} + \cdots \\ &+ \beta_{24} \, \mathrm{D}_{\text{ad end year 2021}} + \sum_{i \in \text{postcodes}} \beta_i \, \mathrm{D}_{\mathrm{i}}\end{aligned} \tag{3.7}$$

8 The excluded reference categories are: for advertisement end year, 2012; for condition of the property, 'Not specified'; for floor, 'Unknown'; and for postcode, the reference postcode varies between samples.

It is important to note that in regression using a matched sample with replacement, control units matched more than once are assigned proportionally greater weight.

3.6. Evaluating Quality of Matching

Before presenting and discussing the results, we need to evaluate whether matching led to a balanced dataset in terms of the covariates described in 3.1.[9]

For numerical variables, after matching, a two-sided t-test on the difference in means was conducted with the null hypothesis $H_0 : mean_0 = mean_1$, taking into account unequal variances. According to the p-values[10] shown in table 4.1, the null hypothesis H_0 could not be rejected at a confidence level of $\alpha = 5\%$ for all covariates, except for the end year of advertisement when matching without replacement, and for all covariates when allowing replacement. In contrast, for the t-test conducted on the unmatched sample, as described in table 3.1, the null hypothesis H_0 was rejected at a confidence level of $\alpha = 5\%$ for all covariates except for the year of construction, indicating that the sample means in the control and treatment groups were significantly different before matching. Thus, matching, particularly with replacement, led to a more balanced set of numerical covariates. Table 3.3 provides an overview of whether H_0 could not be rejected for different samples.

Variable	**Sample**		
	Unmatched	**Matched without replacment**	**Matched with replacment**
Balcony	✗	✓	✓
Year of construction	✓	✓	✓
Living area in m^2	✗	✓	✓
Number of rooms	✗	✓	✓
Propensity Score	✗	✗	✗
Advertisement end year	✗	✗	✓

Tab. 3.3.: Results of t-Tests for Equality of Means Across Variables and Samples

Note that ✓ indicates we cannot reject $H_0 : \text{mean}_0 = \text{mean}_1$ when performing a t-test on the difference in means at a confidence level of $\alpha = 5\%$.

As can be seen in figure 3.2, matching makes the distributions of PS more similar between the treatment and control groups, with matching with replacement delivering better results. Although we would always reject the hypothesis of on-average equal PS in the treatment and control groups, it can be observed that the mean difference decreases with matching without replacement and decreases even further when allowing for replacement (see table 4.1).

9 We are not concerned with the balancing of the purchasing price in EUR and the price per m^2, as these refer to the outcome variable.

10 p-values are given in parentheses.

When it comes to categorical variables with more than two categories, such as the location and condition of the apartment, we cannot perform a t-test. However, we can still evaluate whether the distribution of these variables becomes more similar after matching. By analyzing the difference in distributions of property conditions in table A.2, it can be observed that the distributions become more similar in the matched sample, with smaller differences when allowing for replacement. Regarding the location, comparing postcodes with 1,585 categories would be difficult to visualize graphically; however, we can at least compare the distributions at the city level. As shown in figure 3.3, after matching, the distribution of cities becomes more similar, and once again, matching with replacement shows the best performance.

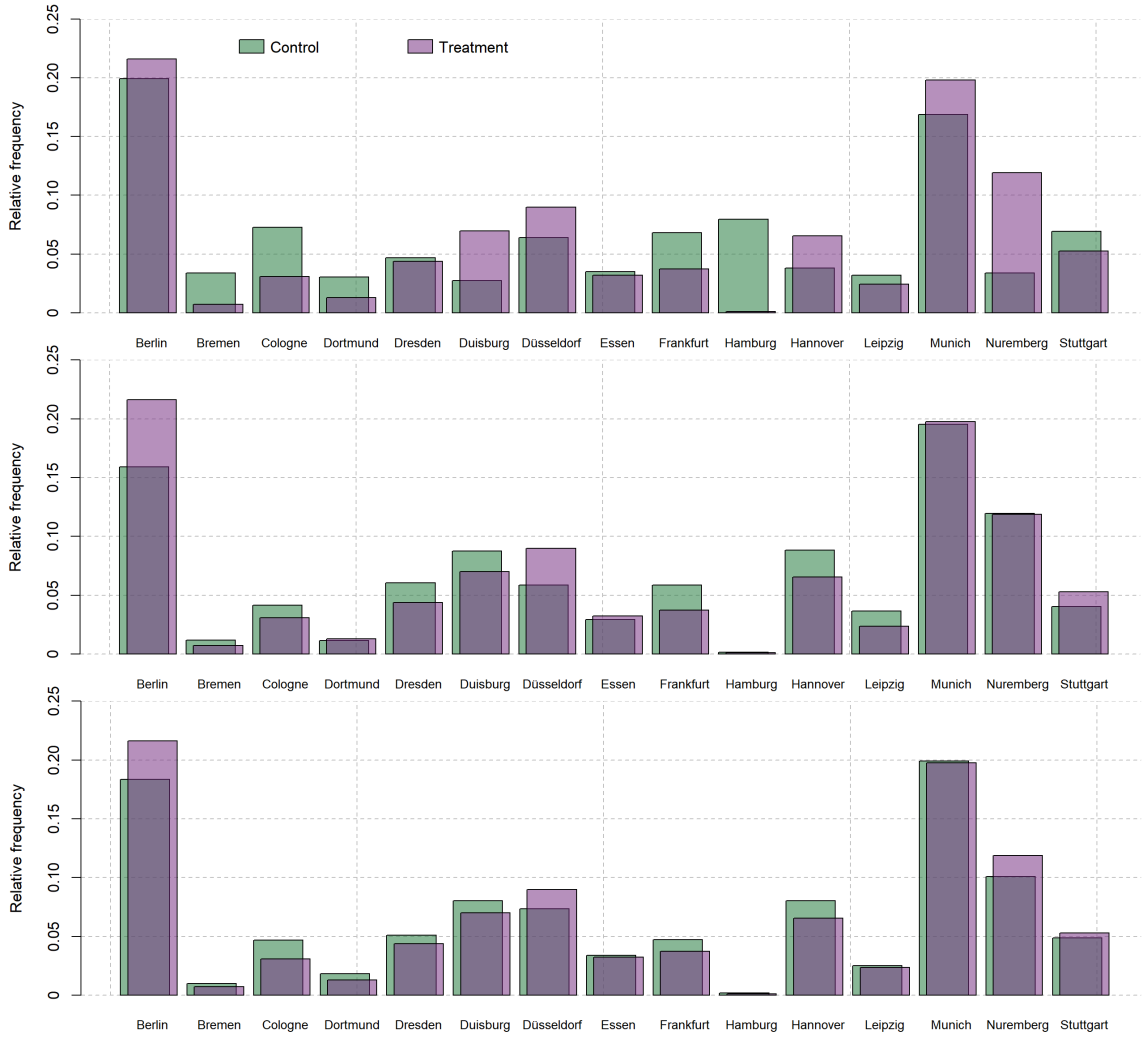


Fig. 3.3.: Distribution of Cities Across Samples: Above - Unmatched Sample, Middle - Matched Sample Without Replacement, Below - Matched Sample With Replacement

To conclude, in our case, matching with the 3-NN algorithm balanced the treatment and control groups in terms of covariates and propensity score, with matching with replacement delivering better balance.

4. Results

Tab. 4.1.: Mean Differences between Control and Treatment Groups and ATT Estimates for Different Samples

	$\mathbf{n_0}$	$\mathbf{n_1}$	Mean difference between control and treatment group[3]						ATT[1]	
Covariate			**Year of construction**	**Living area in m²**	**Balcony**	**Number of rooms**	**Advertisement end year**	**Propensity Score**	**Simple linear regression**	**Multiple log-linear regression**
Sample									$\hat{\beta}_1$	$\hat{\beta}_1$
Data without matching	297,076	5,362	-0.791 (0.132)	-4.928 (0.000)	-0.054 (0.000)	-0.116 (0.000)	-0.081 (0.036)	0.246 (0.000)	-0.0937*** (0.000)	-0.0300* (0.025)
3-NN without replacement	16,011	5,337	0.912 (0.128)	0.454 (0.387)	-0.005 (0.436)	0.000 (0.997)	-0.118 (0.008)	0.058 (0.000)	0.0194$^{\cdot}$ (0.071)	-0.0295* (0.032)
3-NN with replacement[2]	8,966	5,337	0.884 (0.177)	-0.562 (0.328)	-0.003 (0.676)	-0.019 (0.299)	-0.033 (0.505)	0.052 (0.000)	-0.0041 (0.719)	-0.0242$^{\cdot}$ (0.071)
Significance codes			(0) '***' (0.001) '**' (0.01) '*' (0.05) '$^{\cdot}$' (0.1) ' '							

[1]The ATT is the coefficient β_1, estimated from the regression equation (3.7) or from the simple log-linear regression $\log\left(Y_{\text{price/m}^2}\right) = \alpha + \beta_1 \, \mathrm{D}_{\text{emissions}}$, using different samples. [2]In the case of matching with replacement, apartments from the control group that were matched more than once received a proportionally higher weight in the regression, while treated apartments were matched only once and received the same weight. [3]The difference is computed as $mean_1 - mean_0$. All p-values are reported in parentheses. For the mean difference between control and treatment groups, p-values are derived from a two-sided t-test on the difference in means under the null hypothesis $H_0 : mean_0 = mean_1$, accounting for unequal variances. The p-values for the ATT results are based on standard t-tests for slope parameters from OLS regressions, with standard errors clustered at the postcode level in case of multiple regression. Note that because the propensity score and covariate distributions are not perfectly balanced after matching, coefficients from simple linear regression that corresponds to a simple mean comparison are not valid estimators of ATT and are presented for comparative purposes. R^2 values are not reported, as the primary focus is on the causal interpretation of ATT rather than model fit. However, depending on the sample, the adjusted R^2 for multiple regression is 81.11% or higher.

5. Discussion

By applying multiple log-linear regression on a matched sample, as described in equation (3.7), we obtained estimates of the ATT. Using the 3-NN matching algorithm, we found ATT = -0.0295, significant at the 5% level when matching without replacement, and ATT = -0.0242, significant at the 10% level when matching with replacement. As discussed in priveous section we can conclude that both matching with and without replacment are sutable for our case, however the estimate that is retrieved from matched sample with replacment is less biased but exhibits greater variability. When matching without replacement, the sample size was nearly double, resulting in smaller standard errors, larger |t|-values, and consequently smaller p-values, which explains smaller p-value of the first estimate. Matching with replacement achieved better balance of covariates and despite having a smaller sample size—nearly half—it still produced a significant result. Therefore, ATT = -0.0242 is considered the more accurate estimate.

This implies that when an apartment for sale is located in a 1-km^2 raster cell with a production facility, its value decreases by 2.42%, which is 0.58% less than the estimate obtained without matching. While a 2.42% price reduction may not appear substantial at first glance, it becomes significant on a larger scale. For instance, if a construction firm buys or sells two houses with 20 apartments each, valued at 0.5 million euros per apartment, this seemingly small percentage difference would result in a financial impact of 4.8 million euros. In our analysis, we controlled for location at the postcode level, which allowed us to at least partially isolate the effect of the facility's presence from being in a postcode or district with lower-priced apartments. It is also important to note that the effect of a facility's presence may stem from either visual disturbances or pollutants and hazards. In the context of this seminar paper, we did not differentiate between these two sources of externalities; therefore, the ATT reflects their combined impact. Future research could refine these findings by isolating the effects of specific externalities.

Cordera et al., 2019 examined how various factors, including the distance from a steel factory, nitrogen dioxide (NO_2) concentration levels, daily average particulate matter, and subjective indicators such as perceived pollution and noise, influenced real estate prices. The study was conducted in the Italian province of Taranto, home to ILVA, the largest steel factory in Europe. The data used in their study were collected in October 2012 from a real estate website. Similar to this seminar paper, the researchers considered structural characteristics of the properties, such as the number of rooms, living area, and apartment condition, as well as other factors. The dependent variable in their study was the logarithm of the asking price, while in this paper, we use the logarithm of the asking price per square meter. This difference does not affect the interpretation of the results. Their findings

showed that the distance from the factory was positively correlated with real estate prices, and among the pollution indicators, only high levels of (NO_2) had a negative effect. At a 90% confidence level, their estimation suggested a 0.4% increase in real estate values for every kilometer away from the industrial area. While these results cannot be directly compared with ours—since all the apartments considered in their study were at least 2.46 km from the steel mill, and the industrial sector they analyzed specifically focused on the steel industry and is larger than in our study—the direction of the effect aligns with our findings.

Another study examined the impact of oil and gas facilities on rural residential property values using data from Central Alberta, Canada (Boxall et al., 2005). This was the first academic investigation into the effects of oil and gas production facilities on property values. The study found that the presence of wells, particularly sour gas wells, tended to lower property values. However, the number of pipelines transporting sour gas did not show a significant effect. On average, property values within 4 km of industry facilities were estimated to be reduced by 4 to 8 percent. While the results are not directly comparable to our own study—since we did not focus exclusively on oil and gas production and used data from the 15 largest cities in Germany, rather than the rural areas studied by Boxall et al., 2005—the negative impact of nearby factories on property prices is consistent with our findings.

Von Graevenitz et al., 2018 studied the impact of the first wave of pollutant emission data release in 2009 by the E-PRTR on housing prices, using the same emissions data employed in our analysis. As an outcome variable, they analyzed quarterly housing prices at the German postal code level from 2007 to 2011.[11] Revealing that polluting facilities are non-randomly located, their study used propensity score matching to form adequate control groups, accounting for postal code characteristics such as land use, housing type distribution, and other relevant factors. After forming the treatment and control groups, they then used a differences-in-differences model to estimate the ATE of the emission-data release event and found no significant effect on house values in affected postcodes. Given the brief existence of the E-PRTR at the time, they recommended further research using long-term data, ideally with microdata, to assess potential long-term effects. While our research question differs, it follows the approach from von Graevenitz et al., 2018 by using PRTR data, matching, and their suggestion of using microdata, but it focuses on evaluating the impact of facility proximity on apartment values rather than the publication event itself.

11 It is important to note that in this study, the treated units are postal codes, not individual properties. Several definitions of "treated" postal codes are discussed in the paper. One defines a postal code area as treated if any part of its land lies within a 500-meter buffer distance from an emitter. Another defines treated postal code areas as those that had at least one report published in the E-PRTR register for the year 2007. Additional definitions are also considered.

6. Conclusion

This study analyzes the causal effect of air-emitting production facilities located near urban residential properties on their prices, based on a 10-year period from 2012 to 2021 and the 15 largest cities in Germany. As suggested in the literature, the analysis uses microdata. The real estate dataset includes detailed characteristics of each property for sale, including the asking price, while the emissions dataset provides the precise latitude and longitude of emitters.

The following methods were applied to estimate the Average Treatment Effect on the Treated (ATT): First, a logit model was used to estimate the propensity score for each apartment. This estimation indicated the probability of being treated—specifically, being located within the same 1-km^2 raster cell as an emitter. Using a cut-off based on the share of treated observations, the model correctly predicted approximately 90% of the treatment assignments. The location of the apartment at the postcode level played a significant role, as it influenced both the treatment assignment and the outcome variable (log price per m^2). Without this variable, the model's prediction accuracy dropped to 56%. Following the estimation of propensity scores, the 3-Nearest Neighbors matching algorithm was implemented both with and without replacement to create suitable control groups and thus balanced samples. Based on the matched samples, regression analysis was then conducted to estimate the ATT while controlling for structural property characteristics, temporal trends, and locational effects.

The estimates varied depending on whether matching was performed with or without replacement, reflecting the tradeoff between bias and variance. The ATT estimate obtained through matching without replacement was -0.0295, significant at the 5% level, with smaller standard errors and higher precision due to a larger matched sample. However, matching with replacement produced a more balanced covariate distribution, resulting in a less biased ATT estimate of -0.0242, significant at the 10% level, despite the smaller sample size and larger standard errors. This suggests that while matching with replacement increases variability, it provides a more accurate representation of the causal effect by ensuring better comparability between treatment and control groups.

In conclusion, apartments located within the same 1-km^2 raster cell as air-emitting facilities were estimated to be, on average, 2.42% cheaper due to their proximity to the emitter. This significant reduction highlights the combined effect of environmental and aesthetic externalities. Our findings align with prior studies that emphasize the negative impact of industrial proximity on property values, while contributing a broader perspective by focusing on all types of air-emitting facilities rather than a specific industry, as is commonly found in the literature. Additionally, this study provides evidence for urban areas, where industrial

facilities and residential zones often coexist in close proximity. These findings represent significant financial impacts on a larger scale and can be useful for construction companies in cost-profit estimations, as well as for urban planners and policymakers seeking to balance industrial development with residential welfare. Future research could refine these findings by isolating the effects of specific externalities, such as pollutants and hazards versus visual disamenities.

Bibliography

Boxall, P. C., Chan, W. H., & McMillan, M. L. (2005): The impact of oil and natural gas facilities on rural residential property values: A spatial hedonic analysis. *Resource and Energy Economics*, *27*(3), 248–269. https://doi.org/10.1016/j.reseneeco.2004.11.003

Caliendo, M., & Kopeinig, S. (2005, April): *Some practical guidance for the implementation of propensity score matching* (Discussion Papers). DIW Berlin. https://www.diw.de/de/diw_01.c.449235.de/publikationen/diskussionspapiere/2005_0485/some_practical_guidance_for_the_implementation_of_propensity_score_matching.html

Cordera, R., Chiarazzo, V., Ottomanelli, M., dell'Olio, L., & Ibeas, A. (2019): The impact of undesirable externalities on residential property values: Spatial regressive models and an empirical study. *Transport Policy*, *80*, 177–187. https://doi.org/10.1016/j.tranpol.2018.04.010

Cunningham, S. (2021): *Causal inference: The mixtape: Matching and subclassification*. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27

Dehejia, R. H., & Wahba, S. (2002): Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, *84*(1), 151–161. https://doi.org/10.1162/003465302317331982

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013): *Applied logistic regression* (Third edition). Wiley.

Humboldt State University. (2018): *Reporting geographic coordinates: Spatial precision in dd*. Retrieved July 11, 2024, from https://gsp.humboldt.edu/OLM/Lessons/GIS/01%20SphericalCoordinates/Reporting_Geographic_Coordinates.html

Matteucci Gothe, R. (2023): Die anwendung der propensity-score-methode. *Prävention und Gesundheitsförderung*. https://doi.org/10.1007/s11553-023-01050-7

Rosenbaum, P. R. (2010): *Design of observational studies* (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-46405-9_11

Rosenbaum, P. R., & Rubin, D. B. (1983): The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Umweltbundesamt. (2023, November 13): *PRTR-Gesamtdatenbestand 2007 bis 2022*. Thru.de. Retrieved January 11, 2024, from https://thru.de/downloads/

von Graevenitz, K., Römer, D., & Rohlf, A. (2018): The effect of emission information on housing prices: Quasi-experimental evidence from the european pollutant release and transfer register. *Environmental and Resource Economics*, *69*, 23–74. https://doi.org/10.1007/s10640-016-0065-8

A. Appendix-Tables

Initial value	Aggregated value
Completely renovated	Good
First occupancy after reconstruction	Good
Modernised	Good
Reconstructed	Good
First occupancy	Good
Like new	Good
Well kempt	Good
By arrangement	Bad
Needs renovation	Bad
Dilapidated	Bad
Not specified	Not specified

Tab. A.1.: Aggregation of Objects' Condition

Sample	Condition of object	Treatment	Control	Difference
Unmatched				
	Not specified	26.31 %	23.58 %	2.73 %
	Good	68.45 %	71.89 %	-3.44 %
	Bad	5.24 %	4.53 %	0.71 %
Matched without replacment				
	Not specified	27.56 %	26.34 %	1.22 %
	Good	67.49 %	68.47 %	-0.98 %
	Bad	4.96 %	5.19 %	-0.23 %
Matched with replacment				
	Not specified	27.2 %	26.34 %	0.86 %
	Good	67.66 %	68.47 %	-0.81 %
	Bad	5.14 %	5.19 %	-0.05 %

Tab. A.2.: Group Distributions of Property Conditions and Their Differences Before and After Matching

B. Appendix-Code

B.1. Estimation of Propensity Score

In the R code, the estimation of the propensity score was performed using the glm() and predict() functions and the cleaned dataset `wk` as follows: first, the coefficients were estimated via

```
glm(emissions ~ baujahr + wohnflaeche + balkon + objektzustand +
    end_year + as.factor(plz), data = wk, family = binomial(link =
    "logit"))
```

and then the propensity score was computed using

```
wk$pscore <- predict(logit, type = "response")
```

B.2. Multiple Log-Linear Model for ATT Estimation

The function feols() from the package fixest was used to compute the coefficients of the model described in equation 3.7, as follows:

```
model <-feols(log(price_sqm) ~ freisetzungen + baujahr +
    wohnflaeche + zimmeranzahl + balkon + objektzustand + etage +
    as.factor(sell_year) |as.factor(plz),
    cluster="plz",
    data = wk)
```

The dataset `wk` represents the unmatched data. Subsequently, when estimating the ATT using the matched sample, the `data` parameter was set to a different dataset, while all other arguments remained unchanged. When estimating the ATT with a matched sample using replacement, the argument `weights = matched_r_wk$weights` also needed to be included.

C. Digital Appendix

data_loading.R

data_cleaning_real_estate.R

data_cleaning_emissions.R

estimation.R

cities_bar_chart.R

cities_map.R

PS.rds