




Does noise affect housing prices? A case study in the urban area of Thessaloniki

Georgios Kamtziridis^{1*} , Dimitris Vrakas¹ and Grigorios Tsoumakas¹

*Correspondence:

georgekam96@gmail.com

¹Department of Informatics,
Aristotle University of Thessaloniki,
Thessaloniki, Greece

Abstract

Real estate markets depend on various methods to predict housing prices, including models that have been trained on datasets of residential or commercial properties. Most studies endeavor to create more accurate machine learning models by utilizing data such as basic property characteristics as well as urban features like distances from amenities and road accessibility. Even though environmental factors like noise pollution can potentially affect prices, the research around this topic is limited. One of the reasons is the lack of data. In this paper, we reconstruct and make publicly available a general purpose noise pollution dataset based on published studies conducted by the Hellenic Ministry of Environment and Energy for the city of Thessaloniki, Greece. Then, we train ensemble machine learning models, like XGBoost, on property data for different areas of Thessaloniki to investigate the way noise influences prices through interpretability evaluation techniques. Our study provides a new noise pollution dataset that not only demonstrates the impact noise has on housing prices, but also indicates that the influence of noise on prices significantly varies among different areas of the same city.

Keywords: Housing prices prediction; Noise pollution; Ensemble models; Interpretability

1 Introduction

The real estate market plays an important role in people's lives, from individuals and families, to small businesses and large corporations. The process of purchasing or renting a property, whether for residential or commercial purposes, mainly depends on the economic and financial planning of a family or a company. Additionally, it is strongly related to the macroeconomics and the financial stability of much larger groups of people such as countries. Any sign of inconsistency or fluctuation in the real estate market can provoke apprehension in the state, trigger an economic recession or, ultimately, even lead to financial crises through *housing bubble bursts*. The potential risks are well known to the concerned parties and more importantly to governments that monitor the market on a regular basis. Banks have also invested greatly in real estate in order to obtain accurate house pricing estimates for mortgages and housing loans. These organizations often need to estimate the value of a given property for auctions or damage control when clients are unable to pay their debts. Besides states and organizations, property owners and investors

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

should have the right to access valuable insights about the value of their properties too. This knowledge can increase the efficiency of managing assets or even help make profitable property investments.

Property estimations are performed by human experts like real estate brokers and engineers. This estimation process considers properties' features and amenities, as well as external factors such as bus station density or distances to city centers. These are combined with other metrics, like the House Price Index [1], which tracks the changes in property prices, to arrive at a price estimate. During this process, there is no way of quantifying the accuracy of prediction nor the importance of each component that was included in the task. Therefore, the absence of confidence increases the risk of the forthcoming decision, which can end up being financially harmful.

In the contemporary world, the real estate market is represented mostly through different web-based services. In each country, there are numerous websites with vast amounts of properties available for renting or buying. These data have been utilized in the past for different analyses, ranging from creating models capable of predicting house prices based on their features to estimating prices over time in order to understand their seasonality. There has been a lot of research on this topic over the years, with big real estate datasets containing hundreds of properties being used to train machine learning models with the ultimate goal of providing meaningful price estimates. These datasets contain basic property features that are specific to the building itself, such as location, size, floor level and heating type to name a few. Moreover, they can incorporate other features related to the surrounding area of the property, such as road network accessibility and distances from basic points of interest. All these features contribute to the urban profile of a neighborhood, which can directly or indirectly affect prices to a great extent. The importance of these features and their correlation to the price estimates have been validated in previous research [2–5].

Environmental factors have not been taken into consideration in the literature as much as they should have, despite their obvious role when selecting a property. The two most popular ones are the air quality index and the noise pollution. The first indicates the level of cleanliness in the air that influences the overall health of the population in a given area [6–8]. The second one is related to the actual noise caused by road traffic, crowds, aviation and other factors such as the presence of night clubs or manufacturing establishments. The influence of noise pollution on the health of citizens living in an urban environment is well-established. There are numerous cases in the research literature underlining the negative aspects of noise [9–12].

Environmental Noise Directive (END)¹ is the primary law in the European Union (EU) dealing with noise pollution affairs. One of its main goals is to inform the public about the environmental noise and its effects on people's health. Moreover, it requires from EU countries to provide noise maps and noise management plans on a regular basis.

Although noise pollution plays a major role in the nature of a neighborhood, research on its impact on house prices remains largely underexplored. To some extent, this is to be expected given the practical challenges of gathering environmental data, such as expensive measuring and monitoring tools, specialized software, and on-site orchestration of distributed sensors. In Greece, these studies are conducted by large corporations or state

¹<https://environment.ec.europa.eu/>.

departments that subsequently hold the data for internal use. Of course, there are some crowdsourced initiatives [13] that aim to collect noise data, but for small countries like Greece, these are usually inadequate.

The impact of the real estate market on a country, in addition to the innovations that can emerge through research in the field, highlights the potential profit of such work. Being able to generate valuable environmental features of an urban area and, then, use those in the housing price prediction problem can help individuals, small and medium-sized businesses, all the way to large corporations, banks and government experts make profitable decisions. Aside from profitability, it can shed light on the various factors that influence prices. Knowing if and how the environment affects housing prices can assist urban planners to design more functional and efficient cities.

In the first part of this paper, we extract environmental data, and more specifically noise pollution, from published scientific studies. We focus on studies performed by the Hellenic Ministry of Environment and Energy² for the urban area of Thessaloniki, Greece. The end results were published by the government with heat maps demonstrating the spatial distribution of noise across the city. However, none of the core noise measurements were made public, making any future use or contribution to the field difficult. We have managed to overcome this limitation by meticulously re-creating the sense of noise into a general-purpose and easy to use dataset.

In the second part of this work, we highlight the importance of noise in predicting house prices. To verify this, we have used the property database of Openhouse,³ which is a real estate platform operating in major cities of Greece and, mainly, in the area of Thessaloniki. Regarding the machine learning models, we choose to use *ensemble* methods that proved to work well in the research literature. The property and the noise data are used to create multiple models with distinct configurations, exploring different aspects of the same problem.

The main contributions of this work are:

- 1 A new general-purpose sense-of-noise dataset, as well as a new housing price dataset containing noise information for the area of Thessaloniki.⁴
- 2 An extensive experimental evaluation of the contribution of noise in the property price estimation process via ensemble models such as *XGBoost* [14] and *light gradient boosting* [15] models.

2 Related work

This section presents relevant research in the field of housing price prediction from a data perspective. It is important to discuss key relevant work in order to better understand the current state of the area, as well as to position this paper properly within the literature. We begin by outlining the most recent and best-performing solutions proposed for housing price estimates, considering basic property features. Then, we showcase approaches that incorporate various environmental features, with a specific focus on noise pollution. In both cases, we aim to investigate how the various features, especially environmental noise, affect prices.

²<https://ypen.gov.gr/>.

³<https://openhouse.gr>.

⁴<https://github.com/gkamtzir/housing-prices-and-noise-thessaloniki>.

Baldominos [3] studies the housing price prediction problem in the Salamanca district of Madrid. With a collection of 2266 properties from popular online sites containing the fundamental characteristics, they test the correlation between the features and the price to find out that size is the most important one. They use these data to construct various regression models of different specifications, such as support vector machines, multi-layer perceptrons and ensembles of regression trees, all trying to predict prices given the features. The final results showcase the superiority of the ensemble trees when compared to others. Imran [16] follows another approach for the capital of Pakistan, Islamabad. Alongside the basic property characteristics, they gather some features related to the surrounding area of a property. For instance, they attempt to include neighborhood related information through binary values (yes/no) indicating the existence of core amenities and services like hospitals, schools and entertainment. Although their experiments encapsulate many features, the results show that besides the total size, the number of bedrooms and bathrooms, also, radically influence the price, with support vector machines being the best performing model.

Truong [5] focuses on the Beijing area by using the “Housing Price in Beijing” dataset which contains more than 300,000 properties. Each property, apart from its standard attributes, has various spatial information like distance from the city center and subway accessibility. The exploratory analysis demonstrates direct correlation between the location and the property price, since each district has a different price range. Initially, random forest [17], XGBoost and lightweight gradient boosting models were used for training. Then, the authors combine these to build a stacked generalization model [18] by placing random forest and lightweight gradient boosting at the first level and XGBoost at the second one. This architecture outperforms any of the individual ones in terms of accuracy, with a much higher computational cost. Similarly, Xue [19] accumulates property data and urban details like bus and metro stations and routes, traffic and road network information for the city of Xi’an, China. The urban data are preprocessed and new meaningful indices are introduced. The property features and the new indices are utilized by ensemble models to highlight the fact that size is, again, the most influential factor in the matter of predicting prices. Additionally, they illustrate the importance of the neighborhood of a property, because the next most important group of features is related to the spatial indices. Along the same lines, Kang [20] engineers relevant features from more generic urban characteristics like human mobility patterns and socioeconomic data. They experiment with a gradient boosting ensemble in order to analyze features’ significance, where they come to the conclusion that some spatial features can play a more decisive role when it comes to predicting prices. For example, the prices of properties located near university campuses are mainly affected by the distance to the campus rather than their total size.

Environmental conditions can, also, act on prices. Chiarazzo [21] gathers property and air pollution data for the city of Taranto in Italy, which is marked as a high environmental risk area due to its heavy industry. With feature selection and an artificial neural network they put to the test the correlation of each feature through an one-by-one elimination process. Interestingly, they state that sulfur dioxide concentration, one of the five major air pollutants, is the most determinant with respect to price, ranking higher than other characteristics such as floor level and distance to the city center. Shanghai is another industrialized city, where Zou [22] evaluates the air pollution phenomenon in connection with property prices to quantify even more their relation. A total of 27,608 properties in

conjunction with air pollutants are used as training data in a gradient boosting model which it attributes 1.6% in terms of contribution. Under no circumstance, this percentage can be considered as minimal, since a reduction of $1 \mu\text{g}/\text{m}^3$ in nitrogen dioxide increases the price by roughly 278 Yuan per square meter.

Regarding noise pollution, there is much less research available attempting to correlate house prices to noise levels. In general, noise pollution is measured in decibels, where higher values suggest noisier environments. Blanco [23] uses hedonic models to analyze the connection between prices and noise levels in three different areas in the United Kingdom. They suggest that when evaluating properties with similar amenities the presence or absence of noise affects people's choices. In particular, the way noise impinges on prices differ depending on the area, where in some there is a positive correlation and in others a negative one. Brandt [24] investigates the same hypothesis in the city of Hamburg, Germany by combining multiple sources such as road, air and rail traffic noise pollution with hedonic models too. They highlight the non-linear relationship among noise and price by stating that price decreases significantly lower in areas with low levels of noise, as opposed to high noise level areas where the decrease is more remarkable. Contrary to Brandt's work, Szczepanska [25] study the noise effect on two rather dissimilar locations, with reference to noise, in the city of Olsztyn, Poland. They indicate the existence of linear correlation between prices and noise pollution which underlines the notion that location can influence the noise-price connection in great measure.

Tsao and Lu [26] collect property data from the Ministry of the Interior of Taiwan for the city of Taoyuan and enhances them with a five year period of noise pollution data from the international airport of Taoyuan. The authors investigate the way aviation noise impacts the real estate market of the city, due to heavy air traffic in lower altitudes, with hedonic models. The models indicate that as the number of flights increases on top of an area, which translates to more noisy conditions, the prices of the corresponding properties decrease noticeably. Moreover, they measure the rate of price decline in certain decibel ranges and conclude that for roughly 65 dB of noise due to air traffic the decrease in price can get to 2356USD, where for more polluted areas the decline reaches the amount of 3622USD. Similarly, Morano [27] study the area of Bari, Italy in order to link noise pollution to house prices, with a total of 200 properties and noise information from the Strategic Noise Map of Bari as well as perceptual views for the quality of an area with regards to noise from residents. To measure the effect of noise, they employ a variation of a data-driven technique known as Evolutionary Polynomial Regression, or ERP [28], referred to as ERP-MOGA [29] which utilize genetic algorithms. The final results outline the negative correlation between prices and noise levels, where highly polluted areas lead to cheaper housing.

The studies mentioned previously span across different cities, countries or, even, cultures. Even though cross-cultural validation [30] is out of the scope of the current paper, we think it's important to mention it since it can fuel future work around this topic.

The related work indicates that the forefront of housing price prediction has been dominated by machine learning approaches, demonstrating their effectiveness in capturing intricate relationships within diverse property features. However, in the realm of incorporating noise pollution as a crucial determinant, prevailing methodologies have largely relied on conventional hedonic regression models. In this study, we endeavor to utilize machine learning models, with a specific emphasis on noise pollution as a pivotal pre-

dictor of housing prices. Moreover, we leverage modern explainability techniques, which have demonstrated efficacy in prior research [31], to untangle the complex dynamics between noise levels and their impact on the real estate market. Through these efforts, we aim to provide a comprehensive and innovative perspective on the interplay between environmental factors and property valuation. These two focal points represent the primary distinctions between the current work and its counterparts in the related literature.

3 Noise data reconstruction

As previously stated, noise data are difficult to obtain because they require specialized equipment for precise measurements, as well as urban environmental specialists capable of completing a task of this complexity. These data must include geographical references in a form of a coordinate system, mapping points or blocks on a map to certain noise values in decibels. This process is usually done with Geographic Information System (GIS) software tools that try to model noise pollution [32, 33].

As far as we know, there is no such data openly available for the urban area of Thessaloniki, Greece. However, there are official studies of noise pollution for Thessaloniki orchestrated by the Hellenic Ministry of Environment and Energy.⁵ The studies were conducted in 2015 for three major municipalities of the urban area of Thessaloniki, namely Thessaloniki, Neapoli and Kalamaria, with specialized equipment capable of measuring ground sounds levels caused mainly, but not only, by factors like vehicles (local transportation), crowds and nightlife, while additionally calculating aviation sound produced by airplanes landing to or taking off at the nearby airport. These noise sources are considered to be the primary causes of noise pollution in urban environments [34]. The duration of the studies were set to 46 consecutive days, capturing noise pollution at least once every hour or, in cases, every 15 minutes.

The final results were illustrated on a heatmap, where discrete colors represent different noise ranges of 5 decibel intervals. For each municipality, the results are segmented into daytime and nighttime noise and, in both cases, the data accumulate the sound sources by taking into account both traffic and aviation disturbances. Additionally, for Kalamaria there is a separate heatmap representing only the aviation noise.

Even though the data were gathered in 2015 they can still be relevant today for the city of Thessaloniki for two reasons. The first one is due to published studies indicating that noise pollution in Thessaloniki remains the same along the years [35]. The second one is the fact that noise outliers, such as noise coming from construction sites or extreme weather conditions, were excluded from the official heatmaps, rendering the dataset more accurate and relatively timeless in terms of the actual noise.

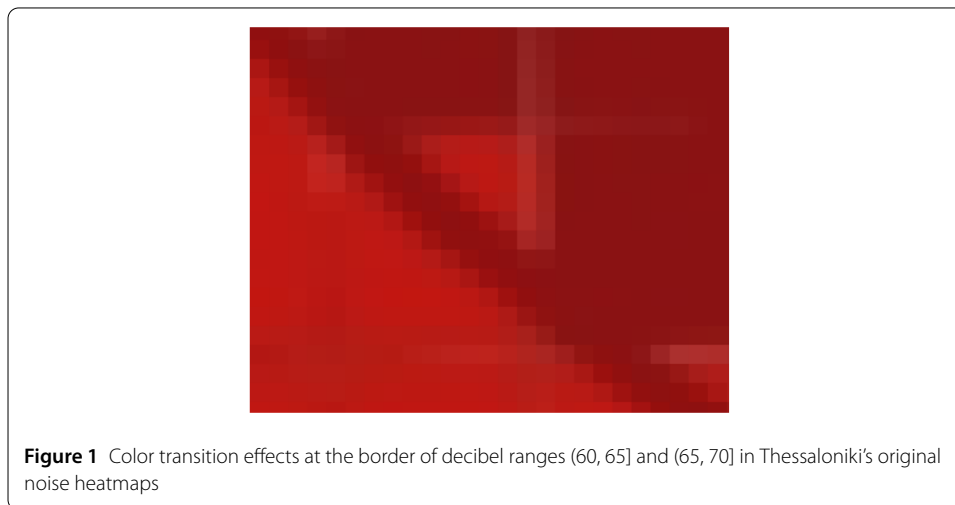
3.1 Idea and approach

The aforementioned studies did not make public the core measurement data that were used to create the provided heatmaps. To overcome this problem, we had to reconstruct these data with a small error. It is important to state that heatmaps used discrete colors mapped to specific small ranges of decibels as shown in Tables 1(a) and 1(b) (note that the ranges and the colors between the two tables are different). This means that each color represents the entire range without changing its tone. The ultimate goal is to be able to

⁵<https://ypen.gov.gr/perivallon/thoryvos-aktinovolies/chartografisi-thoryvou-poleodomikon-sygekrotimatou/>.

Table 1 The original mapping between the noise ranges and the corresponding colors (in RGB) for the area of Thessaloniki/Neapoli (left) and Kalamaria (right). Each noise range within an area is mapped to a different color, while the color mappings between the two areas differ

(a) Thessaloniki & Neapoli ranges		(b) Kalamaria ranges	
Range	Color (RGB)	Range	Color (RGB)
[40, 45) dB	[182, 254, 191]	[35, 40) dB	[80, 167, 50]
[45, 50) dB	[255, 255, 0]	[40, 45) dB	[14, 113, 49]
[50, 55) dB	[254, 196, 71]	[45, 50) dB	[255, 243, 59]
[55, 60) dB	[253, 103, 2]	[50, 55) dB	[172, 121, 78]
[60, 65) dB	[255, 51, 50]	[55, 60) dB	[255, 94, 55]
[65, 70) dB	[152, 0, 51]	[60, 65) dB	[192, 23, 18]
[70, 75) dB	[174, 155, 219]	[65, 70) dB	[138, 18, 19]
[75, 80) dB	[1, 0, 251]	[70, 75) dB	[144, 14, 102]
80+ dB	[1, 1, 65]	[75, 80) dB	[40, 115, 183]
		80+ dB	[10, 65, 121]



create the exact same maps by utilizing the reconstructed data. More specifically, the new dataset will contain the noise, in decibels, of a point given its latitude and longitude coordinates.

It is evident that an approximation of noise levels can be inferred from the colors on the maps. However, spatial information is insufficient to precisely map each pixel to its corresponding place on a geographic map. To address this, we employ a technique known as Georeferencing [36] in QGIS.⁶ This technique performs spatial interpolation by aligning the heatmaps of the noise studies with an actual map, thereby enhancing the heatmap with spatial characteristics (no upsampling was performed). Subsequently, we associate every pixel in the image with a noise value in decibels based on its color, using color mapping [37]. Since there are transitioning effects, as demonstrated in Fig. 1, we compute the difference between the colors in the heatmaps and the predefined color ranges [38]. When this difference is sufficiently small, we can assign noise values based on the corresponding color. To calculate the difference, taking into account human perception of colors, state-of-the-art solutions propose the ΔE^* method, which is based on the LAB color format [39].

⁶<https://www.qgis.org/en/site/>.

We used the most recent version of the ΔE^* method, called CIELAB2000 [40]. The color comparison result is a number, where 0 means a complete match and as the number increases, the difference between the colors increases too. To use this method, one must carefully select the threshold after which the two colors will be considered to be non-matching. After running some experiments, we set the threshold to 20. As a consequence of the discreteness of colors on the heatmaps, the threshold is not considered to be that crucial in our scenario, because the main goal is to differentiate between the predefined ranges. However, it is important to state that lower thresholds (stricter comparisons) were discarding valid noise locations, while higher thresholds (less strict comparisons) were introducing noise locations in places where there weren't any. Essentially, when the color of a pixel matches a color range, this pixel is assigned the corresponding noise value. Since our intention is to correlate housing prices with human perceived noise, we choose to represent each noise range with its arithmetic mean. So, if a pixel color matches the color of the 50-55 range, it will receive 52.5 as its noise value. The final result will be sufficient to describe the noise perception of an area if we take into account the "3 dB rule" in the field of Acoustics [41]. The rule states that during an increase of 3 decibels, the sound energy is doubled and, thus, it is accepted as the smallest difference that can be easily heard by most people. For instance, the average human will rarely notice a transition from 50 to 51 decibels or between 60 and 61.

It is clear that not all pixels are important due to the transitioning effects we mentioned earlier. For example, in cases where two ranges of radically different colors are adjacent on the map, the transitioning effect will add some pixels in between that probably will not match any color range. Additionally, there are cases where the initial studies could not accurately receive measurements, like the inside of buildings and at the sea. These pixels are not matched to any of the available noise ranges and, hence, are dropped to declutter the data. Table 2 gives the structure of the final dataset where latitude and longitude are expressed in WGS84 (World Geodetic System 1984) [42], also known as EPSG:4326.

These datasets can be used to create heatmaps that resemble the initial ones. Even though most parts of the images were removed in the process, the remaining locations are still great in number. This can be verified by considering the dataset size in terms of number of rows in the second column of Table 3. To plot that many points on a single map is exceedingly difficult due to memory constraints. At the same time, the datasets hold spatial information that is way too dense, making them really hard to work with. The dataset contains spatial information for the city of Thessaloniki ($latitude \in [40.56989, 40.678946]$ and $longitude \in [22.880402, 23.014126]$ ⁷) supporting an accuracy of at least 5 decimal points

Table 2 Final dataset structure. The feature names, data types and value ranges of the new noise dataset

Features	Type	Range
latitude	float	[-90, 90]
longitude	float	[-180, 180]
red	int	[0, 255]
green	int	[0, 255]
blue	int	[0, 255]
noise	float	[0, 85]

⁷This is a bounding box approximation for the city of Thessaloniki.

Table 3 The size of the dataset in regards to the number of rows underlining the reduction in size after tessellation

Dataset	# Rows	# Rows (tessellated)	Reduction
Thessaloniki & Neapoli (Day)	3,312,310	197,445	94%
Thessaloniki & Neapoli (Night)	3,157,730	189,046	94%
Kalamaria (Day)	21,606,947	109,245	99.4%
Kalamaria (Night)	20,355,609	104,070	99.4%
Kalamaria Aviation (Day)	21,843,537	110,111	99.4%
Kalamaria Aviation (Night)	21,831,157	109,736	99.4%

Table 4 The property related features (from Openhouse), their data type, the proportion of missing values, as well as how the imputation was done in each case

Feature	Type	Missing values	Imputation
Size	Float	0%	–
NumberOfRooms	Int	0%	–
Latitude	Float	0%	–
Longitude	Float	0%	–
EnergyEfficiencyId	Categorical	0%	–
ConstructionDate	Datetime	13.75%	Mean
SubTypeld	Categorical	0.17%	Mode
FloorLevellld	Categorical	0.30%	Rounded mean
BasicHeatingTypeld	Categorical	30.71%	Mode
DoorFrameTypeld	Categorical	31.77%	Mode

in terms of latitude and longitude. By taking this into consideration, Lambert's formula [43] translates the accuracy in actual distances to 1.11 meters, meaning that each pixel has spatial coverage of about 1.23 m². This level of detail is unnecessary and superfluous for the purposes of this work. To minimize the density of information to more practical levels, we utilize *tessellation*. Through this method, the map is segmented into separate same-sized squared tiles. We chose to tessellate the map by keeping only the four decimal points of the coordinates. Thus, the accuracy decreases to a resolution of 10 meters that is more manageable and adequate for our case. We group the points based on this rule and aggregate their noise using the arithmetic mean to create a representative indicator for the noise level of the given tile. This technique alters the shape of the dataset as shown in the third column of Table 3 and allows us to plot the results on a map.

4 Implementation and experimentation

4.1 Property data

Investigating the correlation and influence of noise in housing prices requires a real world housing prices dataset. For the purposes of this paper, we have utilized Openhouse.⁸ Openhouse is a real estate platform operating in major cities of Greece. It contains high quality information for a wide range of properties, considering multiple aspects of them. Since Openhouse is a data oriented platform, paying critical attention to their service, they have provided Thessaloniki's properties in order to experiment with the noise data reconstructed in the previous section. The data refer to residential properties offered for sale that were listed on the platform in October 2022. Each property has the features mentioned in Table 4.

⁸<https://openhouse.gr>.

The majority of the features are self-explanatory with the exception of ‘SubTypeId’ and ‘DoorFrameTypeId’. ‘SubTypeId’ refers to the structural subtype of the residential property receiving values like ‘apartment’ and ‘studio’ among others. ‘DoorFrameTypeId’ corresponds to the type of door frames a property has, such as ‘synthetics’ and ‘aluminum’ to name a few. We have performed an exploratory data analysis on the given dataset to locate potential outliers and verify the overall integrity. Outlier detection was done with the interquartile range (IQR) method. Using IQR in the ‘NumberOfRooms’ feature led to an upper limit of 7 rooms, which decreased the dataset size by no more than 1%. Similarly, in the ‘Size’ feature, the upper limit was 300 m², which consequently reduced the size by almost 8%. Additionally, price outliers were removed too, by forcing a price range between 10,000 and 500,000 euros. Eventually, the filtered set consists of 2014 properties. The missing values were filled according to Table 4, where different aggregations were used depending on the data type. It must be noted that although ‘DoorFrameTypeId’ and ‘BasicHeatingTypeId’ features are missing approximately 30% of their values, they are considered of significant importance in the housing price prediction process based on the domain knowledge provided by Openhouse. Therefore, we decided to fill these too, and check their influence in practice. As far as the encoding of features, the ‘EnergyEfficiencyId’ and ‘FloorLevelId’ were encoded using incremental indices because they are ordinal categorical features. The other categorical features are nominal so one-hot and binary encoding [44] were used and compared. The one-hot encoding achieved better results and, thus, used in the following experiments.

4.2 Experiments

To investigate the correlation between housing prices and noise we utilize tree-based models that perform well in similar cases [5, 20, 22]. In particular, we use decision trees, random forest, XGBoost and light gradient boosting models. To verify the impact of noise we employ standard interpretability methods like feature importance, partial dependence [45, 46] and permutation importance [47] plots. To shed even more light on interpretability, we employ other advanced techniques such as local interpretable model-agnostic explanations, or LIME, [48] and Shapley additive explanations, or SHAP, [49]. The hyperparameter tuning for each model was accomplished with Bayesian optimization [50], which outperformed grid search, and 5-fold cross-validation. For the evaluation metric, we’ve used the mean absolute error.

The experiments were structured in three different axes. The first one corresponds to the procedure followed to assign the appropriate noise value to each property of the dataset. We choose to average the noise within a certain radius around each property, where the actual radius distance is manually set to 50 and 100 meters. The reasoning behind the selected distances is based on the inverse square law of noise modeling [51] which dictates that for each doubling of the distance from the source of noise, the intensity of the noise is decreased by roughly 6 dB. For example, a typical car (700-1300 cm³) has an average noise level of 82 dB [34]. If a person is exposed to such noise at a distance of 1 meter, at 50 meters the noise levels will attenuate to 48 dB and at 100 meters to 42 dB. For reference, the noise level during a normal conversation is approximately at 55 dB [52]. With that being said, selecting a radius larger than 100 meters will capture noise that will be most likely imperceptible to humans. For the sake of completeness, we should mention that the inverse square law holds true in open fields. In urban environments, noise does not follow exactly the inverse square law [53, 54], but it is still a good approximation.

The second one refers to the main noise characteristics we can use when assigning a noise value to a property. These characteristics are the following:

- One feature for the average day noise and one for the average night noise (I)
- One feature which averages both day and night noise (II)
- One feature for the average day noise (III)
- One feature for the average night noise (IV)
- No features for noise in the baseline model (-)

The baseline model uses the same features as in cases I, II, III, and IV, without accounting for noise pollution characteristics. This configuration enables a more direct comparison of how noise features may affect housing prices.

The third and last experimental component is the area where we examine the effect of noise in pricing. The presence of noise can be translated differently depending on the urban attributes of each part of a city [30, 55]. Good examples that demonstrate this behavior are city centers, where the noise levels are usually increased compared to other places in the same city as a consequence of the high road and pedestrian traffic. In turn, the traffic is caused by the commercial nature of the center since most of the provided services and amenities are located there. In these areas, properties with high noise pollution may command higher prices compared to properties with lower levels of environmental noise. However, this pattern does not hold true for other parts of the city. For instance, in the suburbs, where there are mostly residential properties of families, the absence of noise is generally considered to be a positive factor that can raise the prices. Taking these into consideration, we focus on three different areas of Thessaloniki with contrasting urban features: the city center (A), Triandria, Toumpa and Harilaou areas (B) and Kalamaria area (C), as they are depicted in Fig. 2.

Based on the Hellenic Statistical Authority (ELSTAT)⁹ these areas have approximately the same population (around 90,000 to 100,000). However, area C has relatively lower population density compared to the other areas highlighting its suburb-like characteristics. As for the actual properties, in each of the three areas the number of properties is, again, of the same order. More specifically there are 481,358 and 472 properties in areas A, B and

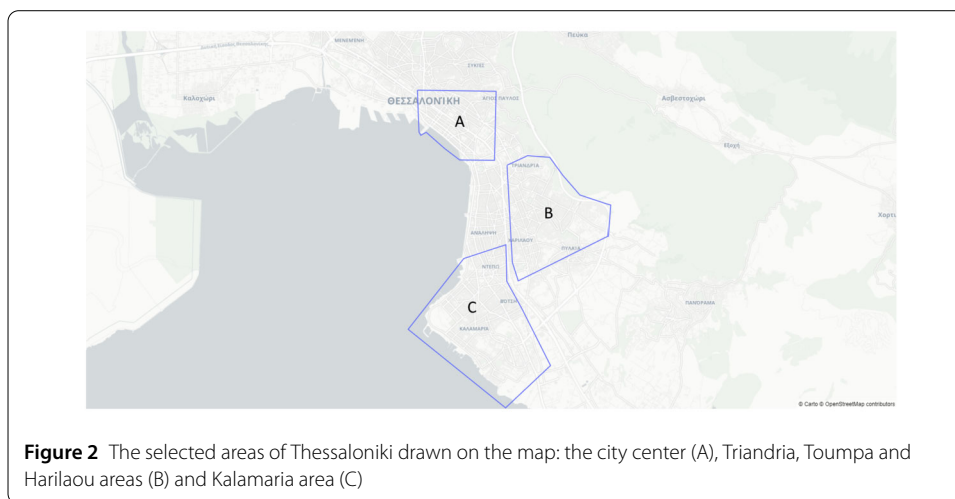


Figure 2 The selected areas of Thessaloniki drawn on the map: the city center (A), Triandria, Toumpa and Harilaou areas (B) and Kalamaria area (C)

⁹<https://www.statistics.gr/en/2021-census-res-pop-results>.

Table 5 The R-Squared (R^2) values between the noise in decibels and the price per m^2 for each area, showcasing that areas are influenced differently by noise pollution

Area	R^2
Entire	-0.026
A	0.1214
B	-0.069
C	-0.1410

C respectively. From a price perspective, which is the target variable of the models, Open-House data suggests that area C is more expensive, having an average of 2370€ per m^2 , while areas A and B range close to 2134€ per m^2 and 2136€ per m^2 respectively. This fact underlines that area C is considered to be more valuable and desirable than the other two. Across all areas, the average price for sale is set to 202,412.14€ with a standard deviation of 113,179.89€.

Regarding the public transport, Urban Transport Organization of Thessaloniki (OASTH) is the only operator in the area.¹⁰ The latest OpenStreetMap data showcase that areas A and C have similar access to public transport of about 187 and 170 bus stops respectively, while area B has significantly lower access to public transport with only 110 bus stops.

Another reason why we chose these areas is their difference in terms of price-noise correlation. This is illustrated in Fig. 9, where the correlation between price per m^2 and noise is plotted for the entire area of interest as well as each individual area. While it is challenging to discern any significant correlation across the entire area of Thessaloniki, focusing on areas A and C reveals a subtle contrast in trends. This observation spurred us to embark on a more thorough examination of these regions. Area B adheres to the pattern depicted in Fig. 9a and was selected as a representative subset of the entire area. The strength of the correlation can be quantified by calculating the R-Squared values between noise levels and the price per square meter, as illustrated in Table 5. Once again, despite the modest R-Squared values indicating a limited correlation, the divergence in tendencies between areas A and C piqued our interest for further exploration. Furthermore, in order to gain a deeper understanding of the data distribution within each area, we have compiled statistical tables for the fundamental features, available in Appendix D in Tables 8, 9, 10, 11, 12 and 13.

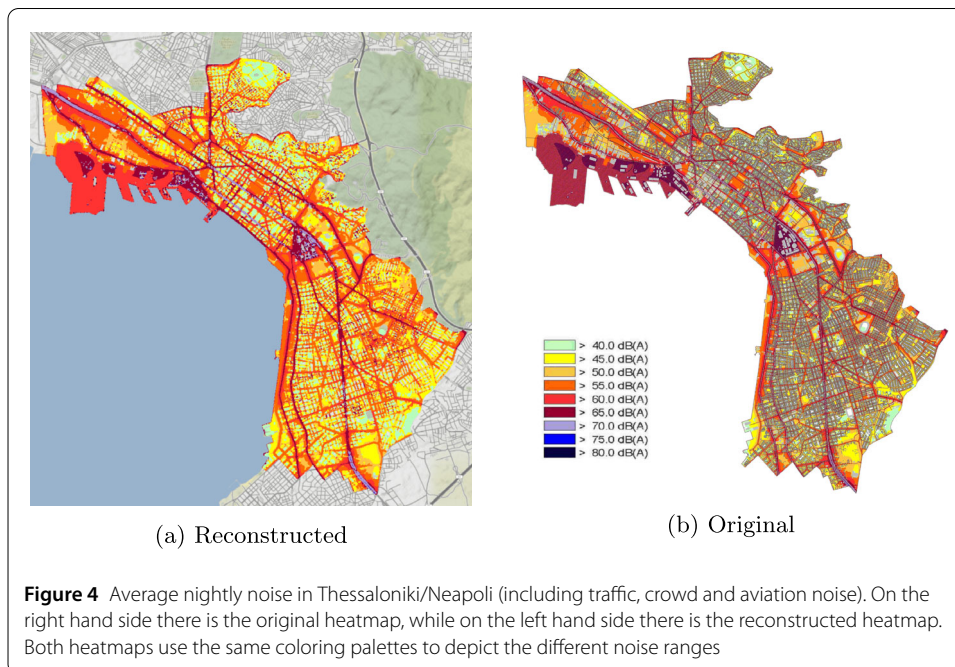
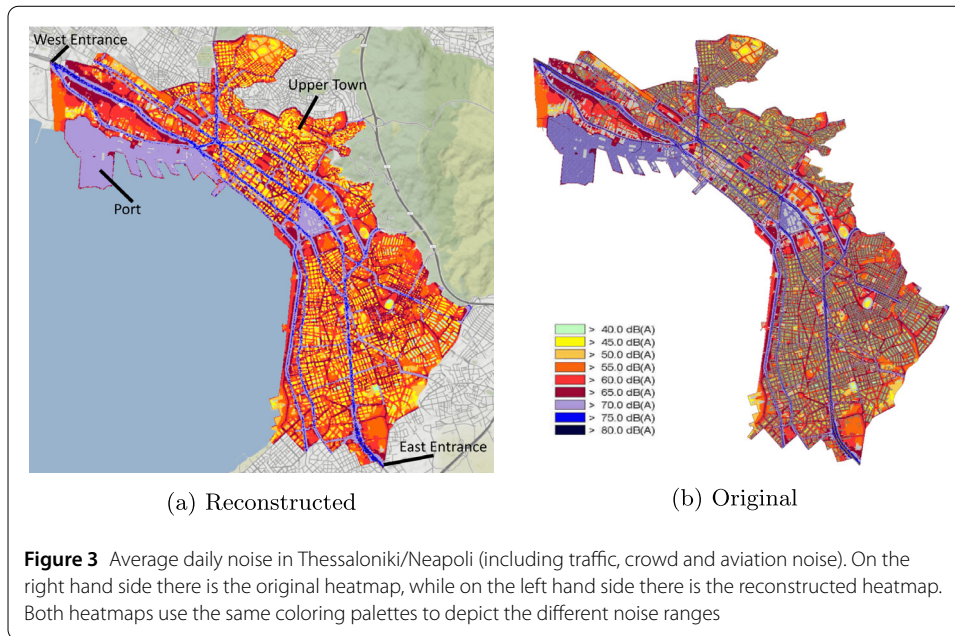
5 Results and discussion

In this section the first two subsections are dedicated to the results and discussion of the noise reconstruction process that was previously explained for the two municipalities of Thessaloniki/Neapoli and Kalamaria. Then, the experimental results and discussion are presented initially with general comments followed by specific ones focusing on each of the three selected areas of the previous section.

5.1 Noise reconstruction results for Thessaloniki and Neapoli

The results of the noise reconstruction process for the areas of Thessaloniki and Neapoli are showcased in Figs. 3 and 4. Figure 3 shows the average daily noise, ranging from 40 dB to almost 85 dB, for both the original and the reconstructed versions. The noisiest parts

¹⁰Public transport in Thessaloniki is solely based on buses.



are the main roads and the intersections that can accommodate large numbers of vehicles. The two most distinguishable examples are the East and West entrances of the city where the noise can reach a level of 80 dB. Also, it is visible the way that the noise spreads almost equally around these highly polluted spots, which, in fact, increase the noise pollution of the surrounding area. Besides the road network, one more part of the city that is apparently noisy is the port, which is very big in size and greatly active during both daytime and nighttime. Furthermore, the correlation between road size, which leads to high traffic, and the noise pollution can be validated in urban areas with narrow streets. A very useful example is the area of “Upper Town” marked in Fig. 3, which is one of the oldest parts of

the city where due to the increased elevation and the rough terrain the roads are extremely narrow. This fact, except the restrictions it imposes on the number of vehicles that can pass simultaneously, makes access difficult and not appealing to drivers. This is one reason why it is one of the quietest places in Thessaloniki. Figure 4 shows the average nightly noise in the same area in which, although the noisiest and quietest places remain the same, the noise pollution levels are much lower.

5.2 Noise reconstruction results for Kalamaria

As in the previous subsection, Fig. 5 shows the average daily noise for the area of Kalamaria. Once again, the noisiest places are the main roads, while the quietest are those surrounded by low traffic streets. The yellow color indicates regions with the maximum noise levels such as the core intersections. Contrary to the heatmaps of the other two municipalities where the noise was almost entirely driven by the road network, in Kalamaria there are certain zones with little or no road network that are very noisy. This is caused due to air traffic, since the airline routes pass over the vicinity in relatively low altitude and the turbines generate noise that can reach over 100 dB [56]. This effect is more recognizable at night (see Fig. 6). Despite the fact that the road network has minimal traffic, some areas are noisier compared to others. The noise generated by airplanes is shown in Fig. 7 and 8. These figures are zoomed in a bit to improve readability and distinguish the street layout. The aviation data can be of great interest both in research and in industry, so in this paper, we provide a separate dataset for the aviation noise.

5.3 Experimental results

The experimental results are organized into two different groups based on the noise radius that was used. For each group all four models were trained on the three areas of Thessaloniki for all four noise characteristics described in the previous section. Due to the area segmentation, the number of properties has declined, leading to a concern about the sufficiency of the training set. To make sure the data were enough to be able to make valid conclusions, we plotted the learning curves of each model and verified that the curves reach a plateau. Essentially, we've trained the models with an increasing number of samples while keeping a hold-out set fixed. For area A, the training curve converged after incorporating approximately 300 properties, while for the areas B and C, after 230 and 250 properties respectively.

Also, because of the large number of different experimental combinations based on the experimentation axes we mentioned previously, we decided to omit showcasing every examination of the noise characteristics and keep, only, the one that performs the best. However, for the sake of completeness, we have included the detailed results in Tables 14 and 15 in Appendix E. We should point out that when changing noise radius there are circumstances where a property can end up without a noise value, especially when the radius decreases. In such cases, these properties are removed from the dataset and this is why there seems to be inconsistencies in the results when switching from one radius to another, even without incorporating the noise data.

The results of Table 6, where the radius is set to 100 meters, indicate a clear dominance of the XGBoost model in terms of both mean absolute error (MAE) and mean absolute percentage error (MAPE) values when compared to the baseline model. The performance gain in each area varies as well as the noise characteristics that are used. More precisely,

Table 6 Results for radius set to 100 m using 5-fold cross-validation for areas A, B and C. The results are presented in terms of the mean absolute error (MAE) and mean absolute percentage error (MAPE). Bold text marks the best score across all models for a given area. The dagger symbol indicates that noise pollution was included in the experiment. The “Noise” column refers to the different noise characteristics: one feature for the average day noise and one for the average night noise (I), one feature which averages both day and night noise (II), one feature for the average day noise (III), one feature for the average night noise (IV) and no features for noise in the baseline model (-)

Model	A			B			C		
	MAE	MAPE	Noise	MAE	MAPE	Noise	MAE	MAPE	Noise
XGBoost	28,919	0.223	-	22,504	0.15	-	31,511	0.138	-
XGBoost †	28,888	0.233	II	19,189	0.144	I	30,141	0.128	IV
LGBM	32,572	0.258	-	21,618	0.158	-	32,752	0.151	-
LGBM †	31,477	0.258	II	22,715	0.173	II	31,139	0.138	IV
RF	32,519	0.267	-	24,259	0.181	-	38,922	0.165	-
RF †	31,759	0.259	II	24,481	0.182	II	40,389	0.173	II
DT	35,264	0.277	-	28,966	0.238	-	48,802	0.209	-
DT †	31,771	0.271	III	30,671	0.25	II	52,077	0.225	III

Table 7 Results for radius set to 50 m using 5-fold cross-validation for areas A, B and C. The results are presented in terms of the mean absolute error (MAE) and mean absolute percentage error (MAPE). Bold text marks the best score across all models for a given area. The dagger symbol indicates that noise pollution was included in the experiment. The “Noise” column refers to the different noise characteristics: one feature for the average day noise and one for the average night noise (I), one feature which averages both day and night noise (II), one feature for the average day noise (III), one feature for the average night noise (IV) and no features for noise in the baseline model (-)

Model	A			B			C		
	MAE	MAPE	Noise	MAE	MAPE	Noise	MAE	MAPE	Noise
XGBoost	28,919	0.223	-	22,504	0.15	-	31,511	0.138	-
XGBoost †	28,001	0.229	I	20,858	0.15	I	31,370	0.132	III
LGBM	32,572	0.258	-	21,618	0.158	-	32,752	0.151	-
LGBM †	30,216	0.241	III	22,408	0.161	IV	29,872	0.13	III
RF	31,785	0.256	-	24,224	0.182	-	38,886	0.165	-
RF †	31,380	0.254	II	24,028	0.183	I	39,626	0.168	IV
DT	35,319	0.279	-	33,561	0.27	-	48,802	0.209	-
DT †	35,453	0.271	III	27,756	0.191	III	50,290	0.209	II

in area A there is no significant improvement, while in the other two areas noise improves both scores radically. The LGBM model benefits from noise only in area C. The random forest and decision tree models are unable to make use of noise with the exception of area A where both are boosted. When the radius is set to 50 meters in Table 7, we observe the same pattern where the hierarchy between the models remains the same. The main differences appear to be the LGBM model that achieves finer results than XGBoost in area C and, also, the decision tree which is crucially improved with the use of noise in area B. Regarding the best performing models, even though setting the radius to 50 meters can reduce the MAE in areas A and C, the MAPE does not change remarkably. Furthermore, decreasing the radius exacerbates the results in area B, so the radius switch does not necessarily enhance the overall performance of the model. The hyperparameter configuration can be found in the Appendix F.

To measure the extent by which noise increases model performance and investigate the correlation between noise and price through interpretability evaluation methods, for the best performing models of each area, we plot the feature importance, permutation importance and partial dependence plots together with LIME and SHAP plots. We must

mention that in permutation importance plots the measure of importance in XGBoost refers to the average gain across all splits a feature is used in, while in LGBM refers to the number of times a feature is used to split the data.

5.3.1 Area A

In the central area of Thessaloniki, XGBoost outperforms all other tested models in terms of MAE when the radius is set to 50 meters. However, the improvement compared to the baseline model is marginal, approximately 3%. In this model, both average day and night noise are used as features in the training. The average day noise is ranked in the feature importance plot of Fig. 10 almost as high as the construction date, while the night noise is located at a couple of ranks below. In the same plot, the 'SubTypeId_4', which denotes properties classified as studios, is marked as the most important feature. The partial dependence plot in Fig. 11a shows that property prices increase as the noise increases, which confirms the initial claim that city centers evaluate noise positively, which most probably occurs due to their commerciality. This can be verified by the LIME weights in Fig. 12 where high noise values correspond to bigger weights. SHAP values in the beeswarm of Fig. 13 highlights this relationship too, since the left hand-side is mostly colored with blue (low values), while the right hand-side with red (high values). At last, in Fig. 11b the night noise does not appear to act on prices at the quieter areas. However, as we progress to noisier parts, night noise has a negative impact on pricing. This is not strange because during night time the commerciality factor is not that crucial. The final predictions in this area demonstrate that one of the main factors that increases MAE is the property size. As the size increases, the model performance slightly decreases.

5.3.2 Area B

Once again, XGBoost with a noise radius of 100 meters demonstrates the best results for the Triandria, Toumpa, and Harilaou areas, reducing the MAE by 14.7% compared to the baseline. Additionally, in this area, the model performs exceptionally better in terms of absolute MAE values compared to the other areas. One plausible reason is the lower standard deviation of price per square meter, as shown in Appendix D. Another contributing factor might be the varying property types in each area. For instance, in area A, 58% of properties are apartments, while 34% are studios. In contrast, in area B, 76% of properties are apartments, and only 12% are studios. As for the factors adversely affecting the model's performance, the predictions emphasize that energy efficiency plays a major role. As the property's energy efficiency increases, the performance tends to decrease.

It should be noted that area B is the only area where setting the radius to 100 meters leads to better results when compared to setting it to 50 meters. Even though the reasoning behind this remains unknown, there are some factors that might be responsible. One of them is the density of the housing units in each area. In high density areas, choosing a larger area might improve the model since there are more neighboring houses within a certain radius. Also, the various topographical and geographical features can affect how sound propagates. For instance, area B is the only one located far from the coastline, while A and C are both seaside areas.

As in the previous area, this model utilizes both average day and night noise values. The night noise has similar importance to features such as the location and the heating type as it is depicted in Fig. 14. In the same figure, the permutation importance plot showcases

that the overall noise affects at some degree the accuracy of the model. Even though, at first, day and night noise do not seem to influence prices, after a certain threshold in decibels they do have a negative effect on prices which contradicts the results of area A. One of the possible reasons why noise does not cause price changes in the initial decibel ranges is the fact that some parts of area B are close to the city center and, hence, noise is not directly considered as a bad attribute. Once more Figs. 15, 16 and 17 reinforces the previous findings about the generally negative correlation between noise and price.

5.3.3 Area C

In the Kalamaria area, the LGBM model when trained with a noise radius of 50 meters while taking into consideration only the average day noise achieves the best scores. Compared to the baseline model, there is an approximate performance gain of 2880 euros, representing a reduction of more than 9% in terms of MAE. Additionally, for MAPE, the improvement is marginally over 2%. When assessing factors that negatively influence the model's performance, the predictions emphasize the significant role of the construction date. Specifically, the model tends to have higher prediction errors for older properties.

As for the features, the noise is ranked almost as high as 'Size' with regard to importance in Fig. 18. This area is located far from the center and as a consequence the noise appears to influence price negatively at most noise ranges. In Fig. 19a, the price declines almost linearly as we move to more noisy parts of the area, while LIME weights in Fig. 19b indicate the preference of the model to assign higher prices to properties with relatively low surrounding noise. Once more, the SHAP values of Fig. 20 confirm the aforementioned observations, where high average day noise values cause price drops and low average noises escalate prices. Concerning the noise characteristic used, one plausible reason why the model chooses to incorporate only the day noise is that contrary to the previous areas, Kalamaria includes also the aviation noise. As it can be seen by the corresponding heatmaps, aviation noise during night increases the overall noise which at some extent narrows the gap between day and night noise. This means that the two noise features are more correlated and, thus, one of them can potentially be redundant.

6 Conclusion

The main goal of this paper was to investigate how urban noise impacts residential property prices in the area of Thessaloniki. Currently, there is no publicly available spatial data regarding noise for the area of interest. Therefore, the first part of this work attempts to create a general purpose dataset indicating the sense of noise based on coordinates by taking advantage of official and public studies conducted by the Hellenic Ministry of Environment and Energy.

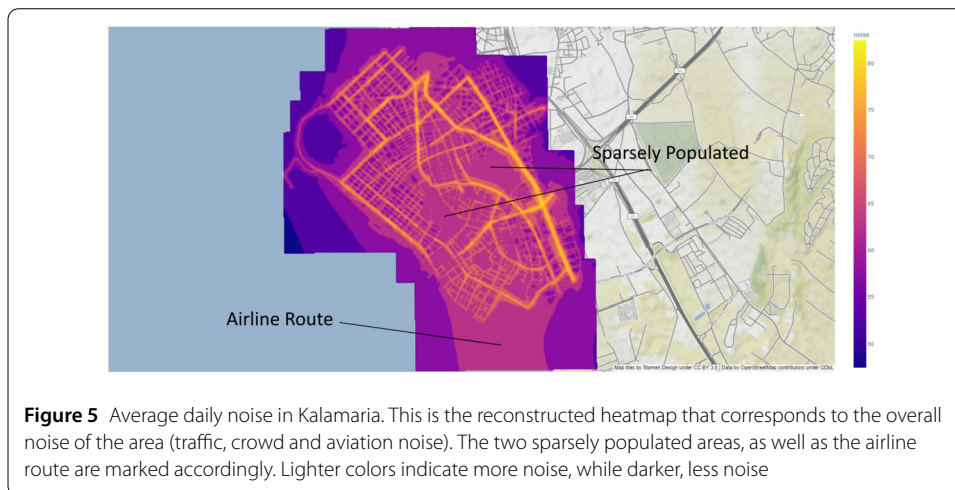
This new dataset is combined with the properties of the Openhouse platform to train tree-based machine learning models in order to verify the importance of noise in housing price estimates. The assumption that noise might be translated differently depending on the location of the property led us to focus the experiments on three separate regions of Thessaloniki with dissimilar characteristics. XGBoost and LGBM models attain the best results which first of all confirm that noise, as a matter of fact, influences prices and, secondly, it can affect some locations positively while others negatively. More specifically, property prices in the city center as well as locations in its vicinity, do increase as noise increases, which is probably the aftereffect of the overall commerciality of the area. In contrast, properties located far from the center are impacted negatively by noise. This makes

sense considering that in decentralized areas, such as suburbs, there are mainly houses of families where quietness is more appreciated.

While this study provides valuable insights into how noise influences property prices, it is important to acknowledge its limitations. To begin with, the noise data used to train the models were sourced from the Hellenic Ministry of Environment and Energy and, then, reconstructed into a general purpose dataset. While we took measures to ensure data integrity, there may still be inherent limitations in the accuracy and completeness of the dataset. Furthermore, although the method employed to calculate noise pollution for each property is generally accepted, factoring in the surrounding buildings and accounting for elevation differences may lead to further refinement of the results. Lastly, despite the current sample size of housing properties appearing adequate for training, augmenting the dataset with additional samples could potentially enhance the model's robustness.

As previously demonstrated, noise is a unique factor that can impact prices differently even within the same city. With this in mind, we strongly encourage the community to delve deeper into this subject, exploring various models, property types, and features. The newly reconstructed noise dataset can play a pivotal role in this endeavor, as its value extends far beyond real estate applications. Its versatility makes it an invaluable resource for a wide array of commercial projects and research pursuits, reaching even beyond fields directly associated with real estate.

Appendix A: Reconstructed heatmaps



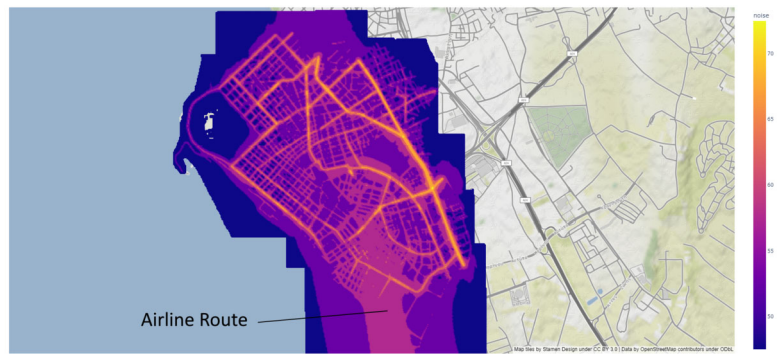


Figure 6 Average nightly noise in Kalamaria. This is the reconstructed heatmap that corresponds to the overall noise of the area (traffic, crowd and aviation noise), with the airline route marked on the map. Lighter colors indicate more noise, while darker, less noise

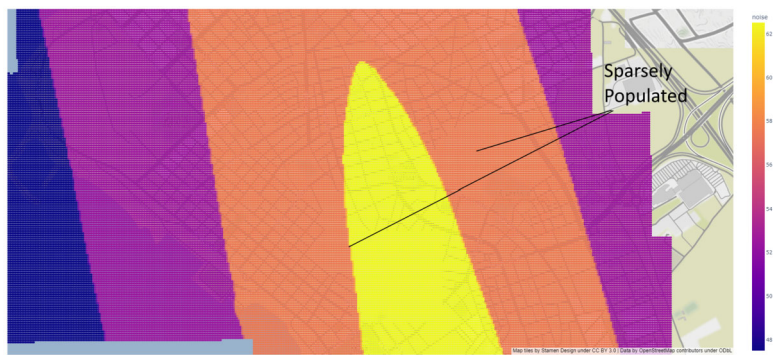
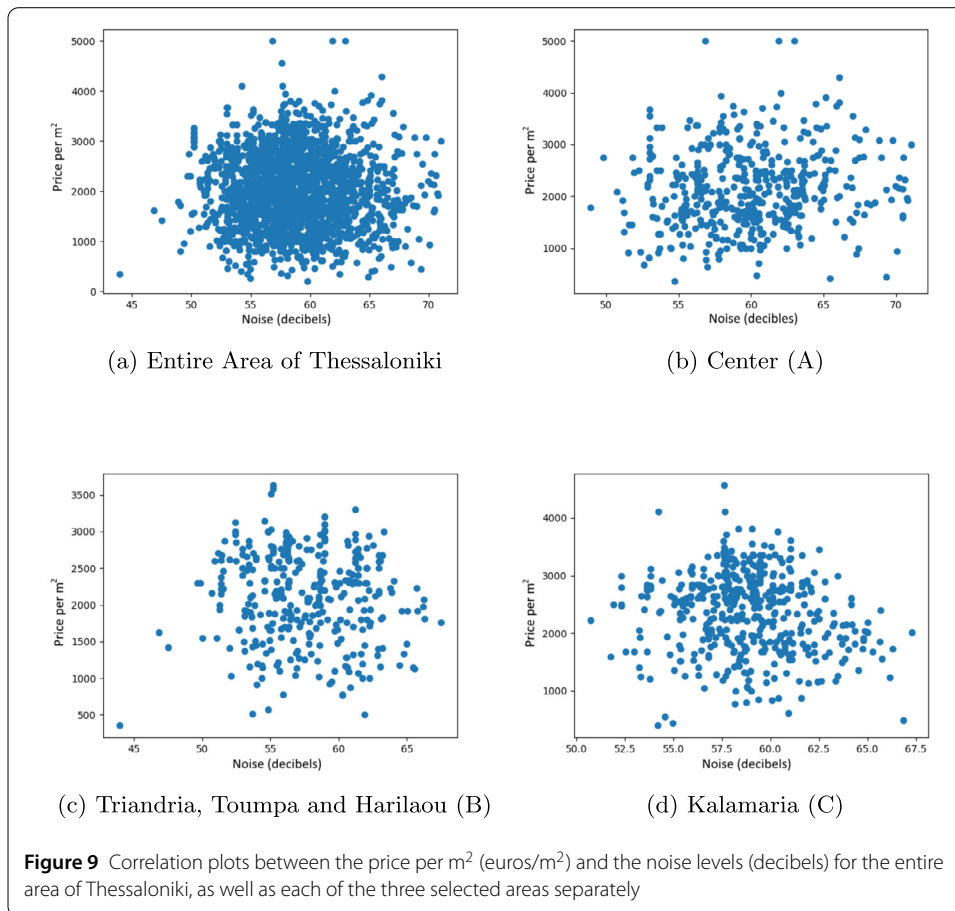


Figure 7 Average daily aviation noise in Kalamaria. This is the reconstructed heatmap that corresponds solely to the aviation noise. The two sparsely populated areas are marked accordingly. Lighter colors indicate more noise, while darker, less noise

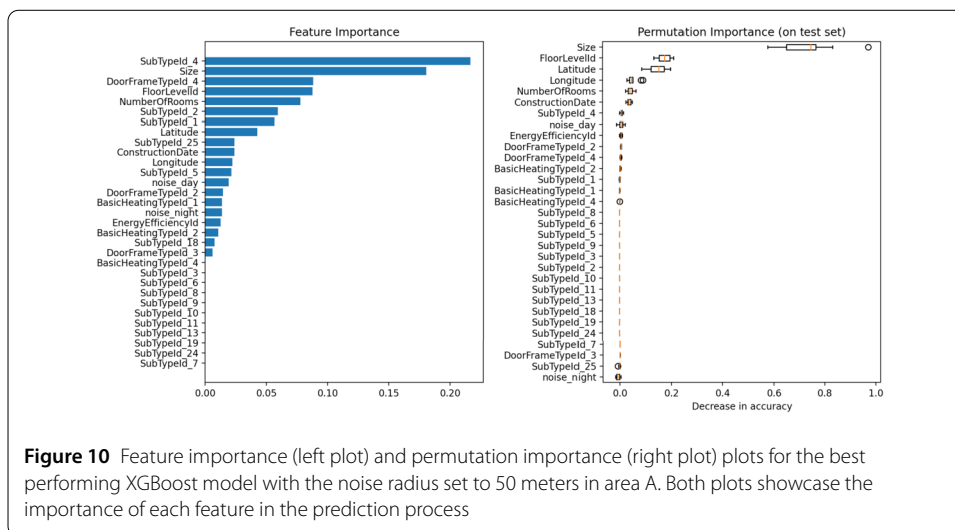


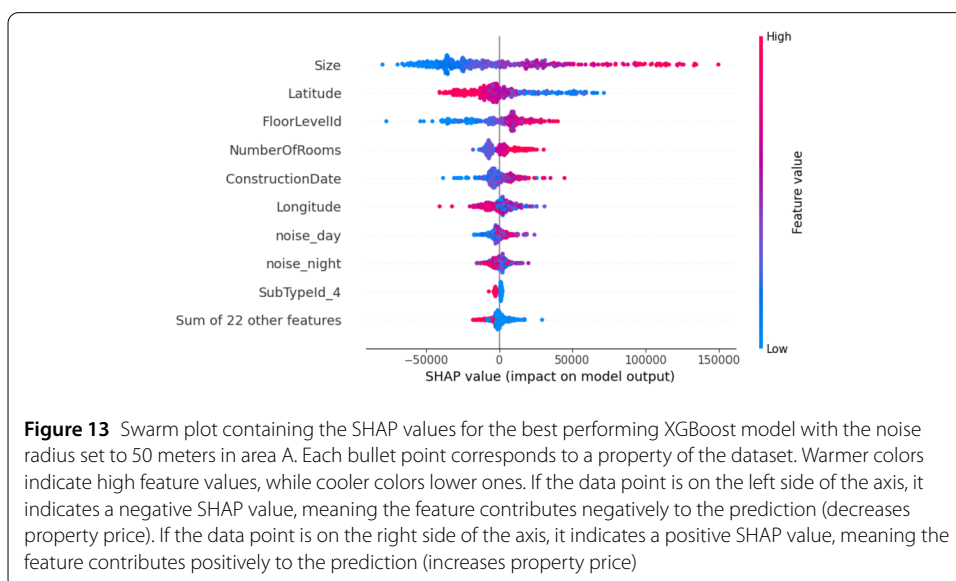
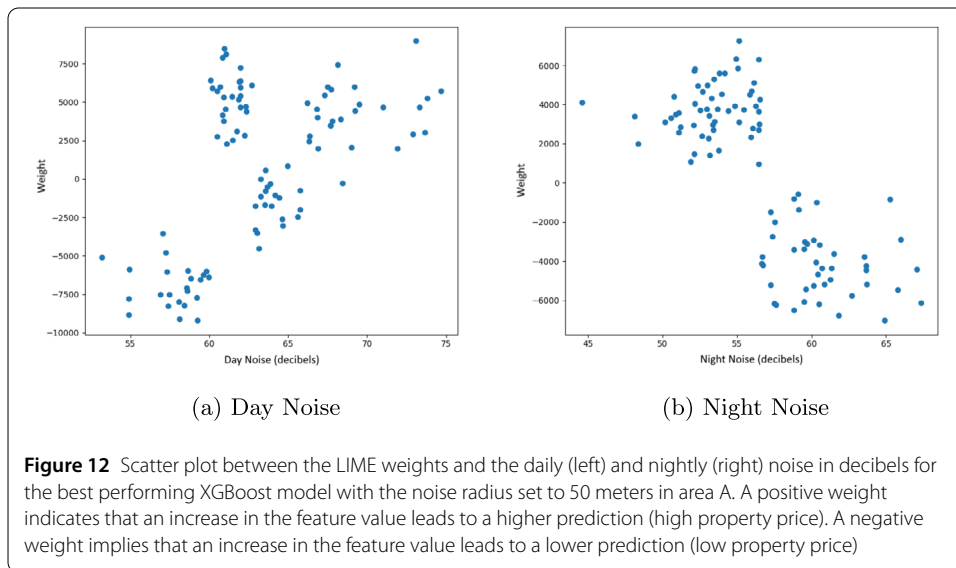
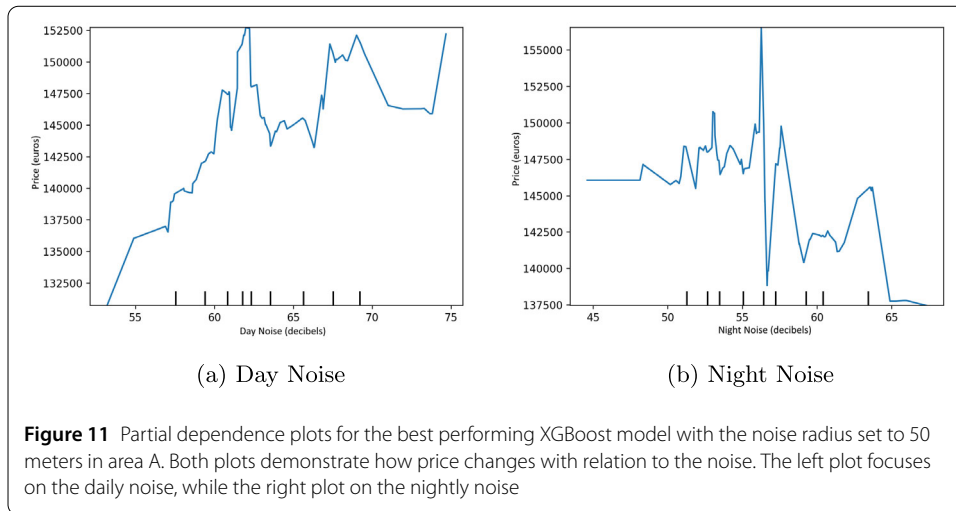
Figure 8 Average nightly aviation noise in Kalamaria. This is the reconstructed heatmap that corresponds solely to the aviation noise. The two sparsely populated areas are marked accordingly. Lighter colors indicate more noise, while darker, less noise

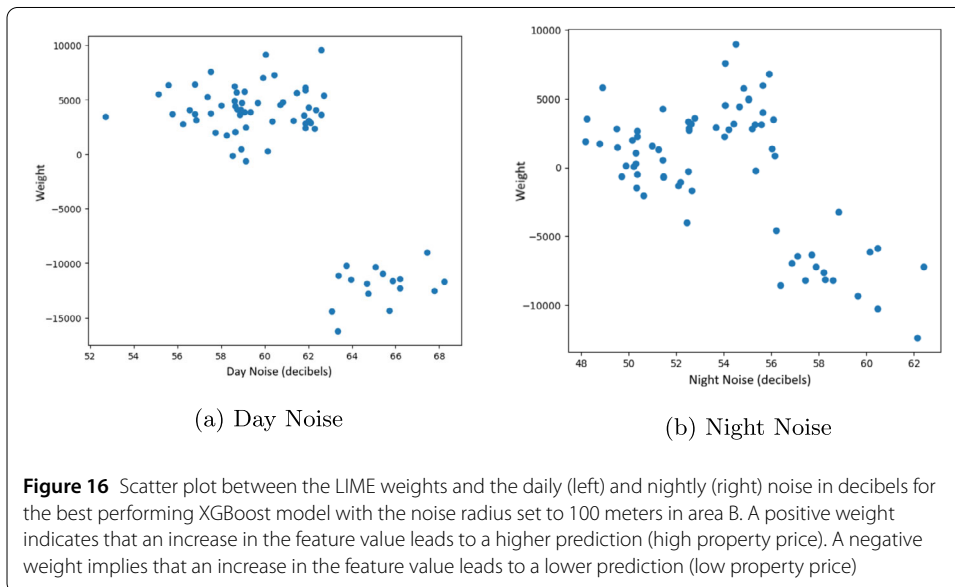
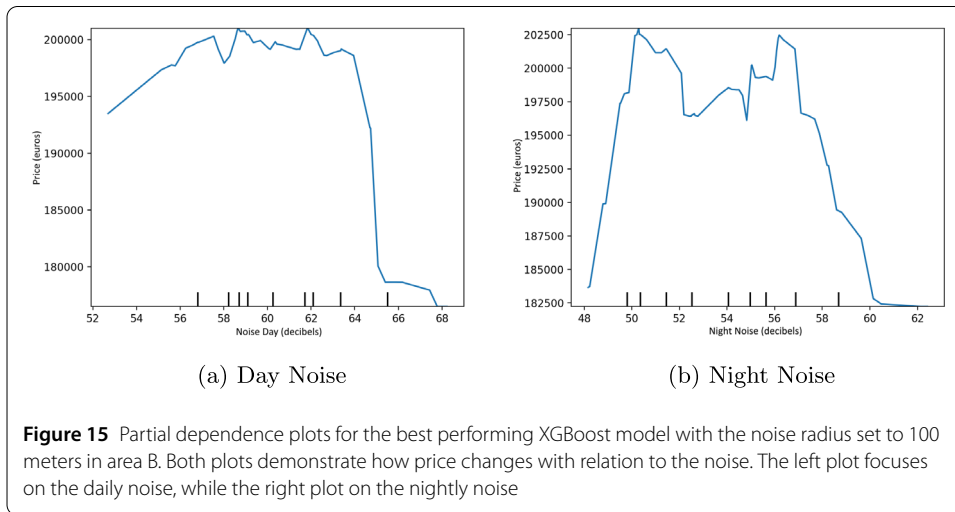
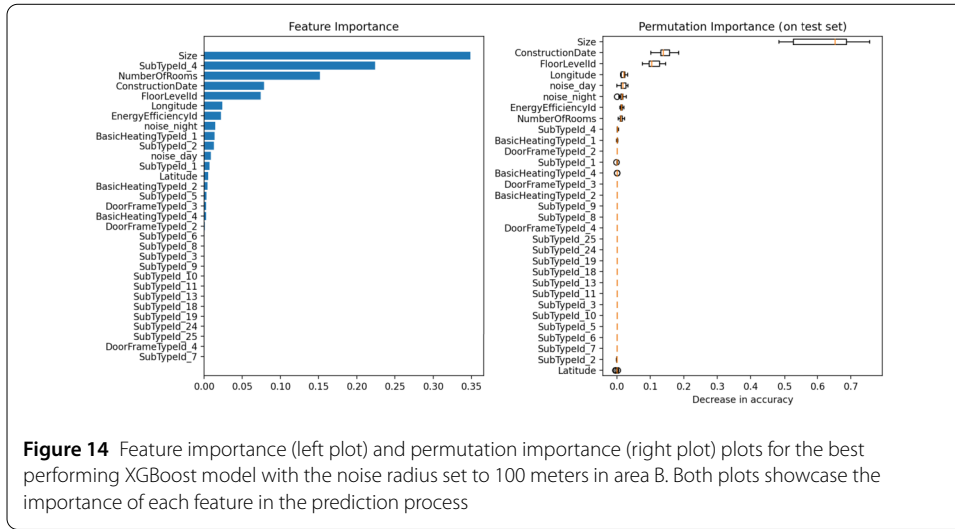
Appendix B: Correlation plots



Appendix C: Result plots







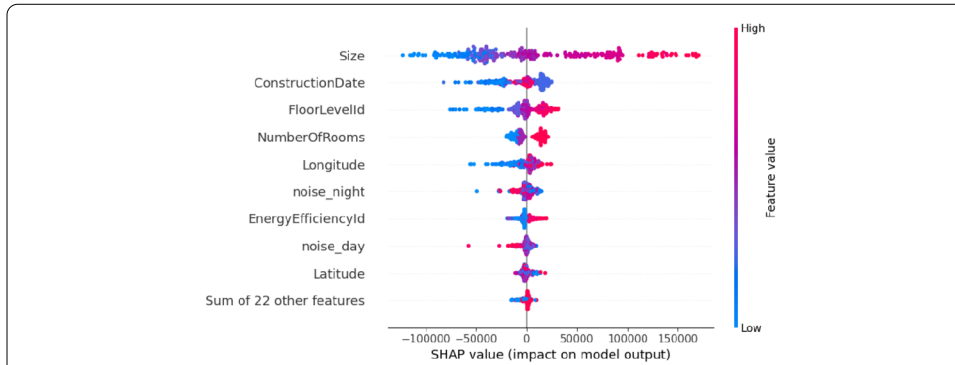


Figure 17 Swarm plot containing the SHAP values for the best performing XGBoost model with the noise radius set to 100 meters in area B. Each bullet point corresponds to a property of the dataset. Warmer colors indicate high feature values, while cooler colors lower ones. If the data point is on the left side of the axis, it indicates a negative SHAP value, meaning the feature contributes negatively to the prediction (decreases property price). If the data point is on the right side of the axis, it indicates a positive SHAP value, meaning the feature contributes positively to the prediction (increases property price)

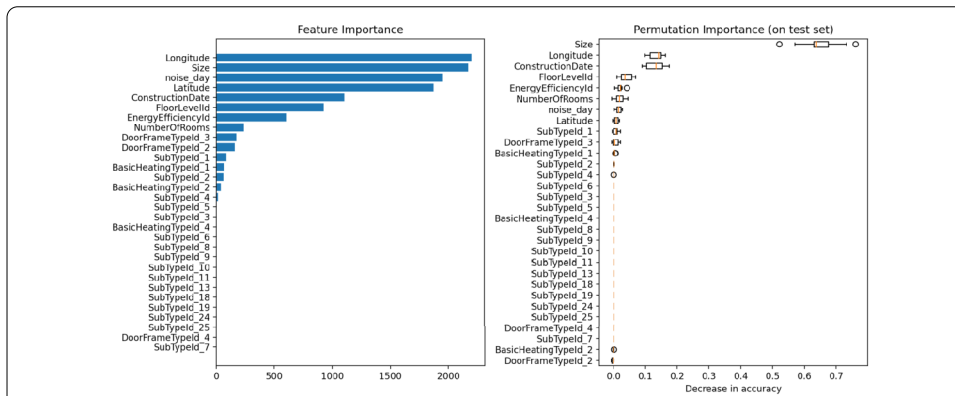


Figure 18 Feature importance (left plot) and permutation importance (right plot) plots for the best performing LGBM model with the noise radius set to 50 meters in area C. Both plots showcase the importance of each feature in the prediction process

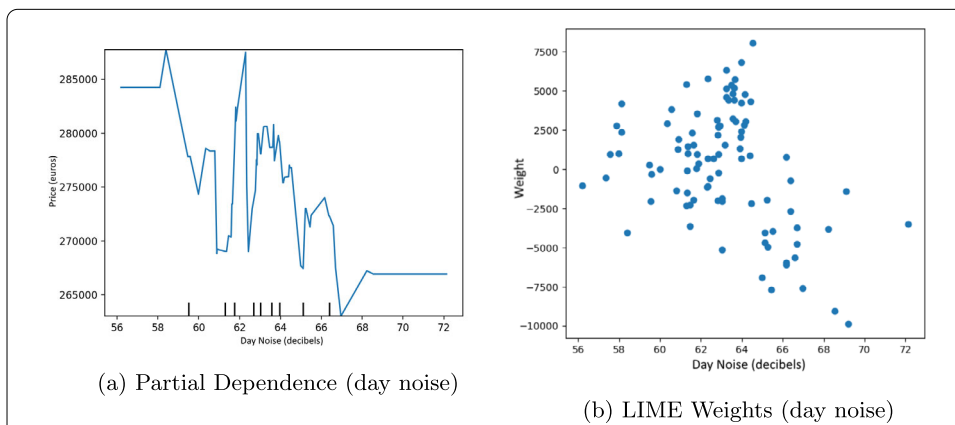


Figure 19 Partial dependence plot (left) and scatter plot between the LIME weights and the daily noise in decibels (right) for the best performing LGBM model with the noise radius set to 50 meters in area C. The left plot demonstrates how price changes with relation to the noise. In the right plot, a positive weight indicates that an increase in the feature value leads to a higher prediction (high property price). A negative weight implies that an increase in the feature value leads to a lower prediction (low property price)

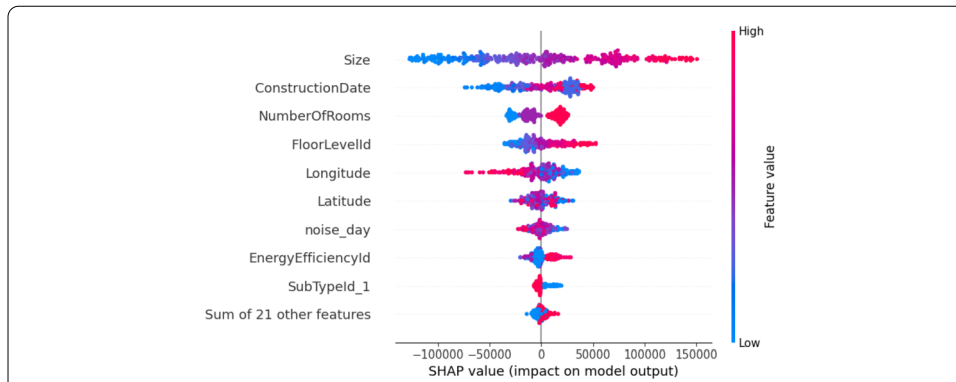


Figure 20 Swarm plot containing the SHAP values for the best performing LGBM model with the noise radius set to 50 meters in area C. Each bullet point corresponds to a property of the dataset. Warmer colors indicate high feature values, while cooler colors lower ones. If the data point is on the left side of the axis, it indicates a negative SHAP value, meaning the feature contributes negatively to the prediction (decreases property price). If the data point is on the right side of the axis, it indicates a positive SHAP value, meaning the feature contributes positively to the prediction (increases property price)

Appendix D: Area statistics

Table 8 Statistics for the basic features of area A with a radius set to 100 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value.

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	71.64	35.67	15	45	63	88	270
Price/m ²	2134.23	756.33	352.37	1627.27	2090.90	2605.26	5000
Rooms	1.63	1.03	0	1	2	2	6
Construction Year	1971.46	17.90	1912	1964	1970	1978	2020
Floor Level	6.47	2.42	1	5	6	8	12
Day Noise (dB)	63.30	3.18	55.75	61.17	63.61	65.42	72.63
Night Noise (dB)	56.74	2.91	48.62	54.89	56.86	58.50	66.08

Table 9 Statistics for the basic features of area B with a radius set to 100 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value.

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	93.48	36.99	18	69	91	120	220
Price/m ²	2136.48	617.59	352.92	1703.06	2200	2621.35	3629.62
Rooms	2.12	0.93	0	2	2	3	5
Construction Year	1985.79	15.66	1955	1978	1980	1988.5	2022
Floor Level	6.84	2.25	2	5	7	9	14
Day Noise (dB)	60.79	3.06	51.86	58.64	60.71	62.791	68.22
Night Noise (dB)	53.89	3.26	42.5	51.42	54.03	56.14	62.41

Table 10 Statistics for the basic features of area C with a radius set to 100 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	107.75	40.27	27	80	109.5	130	280
Price/m ²	2370.76	694.75	400	1851.38	2488.53	2867.39	4565.21
Rooms	2.40	0.85	0	2	3	3	5
Construction Year	1989.16	17.42	1927	1980	1980	2002	2022
Floor Level	6.48	2.06	2	5	6	8	14
Day Noise (dB)	63.37	2.24	56.83	62.04	63.82	64.95	68.46
Night Noise (dB)	54.69	2.22	48.39	53.22	55.07	56.18	60.04

Table 11 Statistics for the basic features of area A with a radius set to 50 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	71.64	35.67	15	45	63	88	270
Price/m ²	2134.23	756.33	352.37	1627.27	2090.90	2605.26	5000
Rooms	1.63	1.03	0	1	2	2	6
Construction Year	1971.46	17.90	1912	1964	1970	1978	2020
Floor Level	6.47	2.42	1	5	6	8	12
Day Noise (dB)	63.11	4.49	53.20	60.04	62.70	66.20	74.83
Night Noise (dB)	56.61	4.30	44.58	53.32	56.49	59.48	67.76

Table 12 Statistics for the basic features of area B with a radius set to 50 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	93.48	36.99	18	69	91	120	220
Price/m ²	2136.48	617.59	352.92	1703.06	2200	2621.35	3629.62
Rooms	2.12	0.93	0	2	2	3	5
Construction Year	1985.79	15.66	1955	1978	1980	1988.5	2022
Floor Level	6.84	2.25	2	5	7	9	14
Day Noise (dB)	61.03	3.95	44.83	58.69	61.11	64.14	71.18
Night Noise (dB)	54.06	4.01	42.5	50.96	53.79	57.13	64.12

Table 13 Statistics for the basic features of area C with a radius set to 50 m. These are: mean value, standard deviation, minimum value, 25%, 50% and 75% percentiles and maximum value

	Mean	Std	Min	25%	50%	75%	Max
Size (m ²)	107.75	40.27	27	80	109.5	130	280
Price/m ²	2370.76	694.75	400	1851.38	2488.53	2867.39	4565.21
Rooms	2.40	0.85	0	2	3	3	5
Construction Year	1989.16	17.42	1927	1980	1980	2002	2022
Floor Level	6.48	2.06	2	5	6	8	14
Day Noise (dB)	63.24	2.89	54.15	61.77	63.09	64.79	72.24
Night Noise (dB)	54.47	2.76	47.28	52.87	54.42	56.08	63.40

Appendix E: Result tables

Table 14 Complete set of results with a radius set to 100 m using 5-fold cross-validation for areas A, B and C. The results are presented in terms of the mean absolute error (MAE) and mean absolute percentage error (MAPE). Bold text marks the best score across all models for a given area. The dagger symbol indicates that noise pollution was included in the experiment. The “Noise” column refers to the different noise characteristics: one feature for the average day noise and one for the average night noise (I), one feature which averages both day and night noise (II), one feature for the average day noise (III), one feature for the average night noise (IV) and no features for noise in the baseline model (-)

Model	A			B			C		
	MAE	MAPE	Noise	MAE	MAPE	Noise	MAE	MAPE	Noise
XGBoost	28919	0.223	-	22504	0.15	-	31,511	0.138	-
XGBoost †	29,698.5	0.241	I	19,189	0.144	I	30,749.4	0.136	I
XGBoost †	28,888	0.233	II	20,605	0.147	II	33,436.3	0.143	II
XGBoost †	28,735.5	0.236	III	21,040.7	0.152	III	30,185.5	0.134	III
XGBoost †	29,662.7	0.235	IV	19,762.4	0.146	IV	30,141	0.128	IV
LGBM	32,572	0.258	-	21,618	0.158	-	32,752	0.151	-
LGBM †	31,629.7	0.266	I	23,702	0.183	I	31,812.2	0.138	I
LGBM †	31,477	0.258	II	22,715	0.173	II	31,448.3	0.143	II
LGBM †	32,180.6	0.263	III	23,218.7	0.175	III	33,286.6	0.145	III
LGBM †	31,423.7	0.272	IV	25,358.3	0.199	IV	31,139	0.138	IV
RF	32,519	0.267	-	24,259	0.181	-	38,922	0.1655	-
RF †	32,383	0.262	I	25,574.9	0.186	I	41,436.5	0.174	I
RF †	31,759	0.259	II	24,481	0.182	II	40,389	0.173	II
RF †	32,054.2	0.264	III	24,698.3	0.185	III	40,335.9	0.175	III
RF †	31,863.9	0.269	IV	24,868.8	0.196	IV	42,175.9	0.182	IV
DT	35,264	0.277	-	28,966	0.238	-	48,802	0.209	-
DT †	36,449.4	0.284	I	30,932.4	0.26	I	54,755.2	0.234	I
DT †	31,771.3	0.271	II	30,671	0.25	II	52,368.8	0.226	II
DT †	31,771	0.271	III	31,862.5	0.212	III	52,077	0.225	III
DT †	36,316.5	0.284	IV	31,694.2	0.264	IV	52,355	0.226	IV

Table 15 Complete set of results with a radius set to 50 m using 5-fold cross-validation for areas A, B and C. The results are presented in terms of the mean absolute error (MAE) and mean absolute percentage error (MAPE). Bold text marks the best score across all models for a given area. The dagger symbol indicates that noise pollution was included in the experiment. The “Noise” column refers to the different noise characteristics: one feature for the average day noise and one for the average night noise (I), one feature which averages both day and night noise (II), one feature for the average day noise (III), one feature for the average night noise (IV) and no features for noise in the baseline model (-)

Model	A			B			C		
	MAE	MAPE	Noise	MAE	MAPE	Noise	MAE	MAPE	Noise
XGBoost	28,919	0.223	-	22,504	0.15	-	31,511	0.138	-
XGBoost †	28,001	0.229	I	20,858	0.15	I	33,215.9	0.147	I
XGBoost †	30,694.4	0.241	II	22,179.5	0.16	II	33,146.6	0.143	II
XGBoost †	28,753.3	0.222	III	21,009.4	0.157	III	31,370	0.132	III
XGBoost †	28,873.4	0.241	IV	22,577.1	0.15	IV	33,255.8	0.147	IV
LGBM	32,572	0.258	-	21,618	0.158	-	32,752	0.151	-
LGBM †	30,956.4	0.252	I	22,384.6	0.17	I	30,774.6	0.14	I
LGBM †	31,171.7	0.249	II	24,053.8	0.165	II	32,324	0.138	II
LGBM †	30,216	0.241	III	23,651.4	0.175	III	29,872	0.13	III
LGBM †	31,687.7	0.264	IV	22,408	0.161	IV	30,624.6	0.138	IV
RF	31,785	0.256	-	24,224	0.182	-	38,886	0.165	-
RF †	31,944.6	0.257	I	24,028	0.183	I	40,999.5	0.173	I
RF †	31,380	0.254	II	24,875.3	0.187	II	41,596.8	0.179	II
RF †	31,691	0.268	III	24,718.8	0.183	III	41,719	0.177	III
RF †	32,528.9	0.268	IV	24,885.7	0.184	IV	39,626	0.168	IV
DT	35,319	0.279	-	33,561	0.27	-	48,802	0.209	-
DT †	35,866.8	0.281	I	31,669	0.215	I	57,638.8	0.247	I
DT †	36,438.4	0.279	II	31,539.9	0.21	II	50,290	0.209	II
DT †	35,453	0.271	III	27,756	0.191	III	54,599.1	0.225	III
DT †	35,976.7	0.28	IV	34,275.4	0.274	IV	55,202.1	0.243	IV

Appendix F: Hyperparameter configuration

The hyperparameters of the best performing models can be found in the paper's Github repository. In this Appendix, we outline some of the basic hyperparameters for each area based on the scikit-learn library [57].

F.1 Area A

The best performing model was XGBoost with the following hyperparameters: radius=50, colsample_bytree=0.6754824399235599, learning_rate=0.041788775351237435, max_depth=5, n_estimators=1000.

F.2 Area B

The best performing model was XGBoost with the following hyperparameters: radius=100, colsample_bytree=0.6812720467459926, learning_rate=0.0436683325289631, max_depth=7, n_estimators=279.

F.3 Area C

The best performing model was LGBM with the following hyperparameters: radius=50, colsample_bytree=0.6609840999909818, learning_rate=0.07216196668844649, max_depth=10, n_estimators=877, num_leaves=120.

Acknowledgements

The authors thank to Anna-Maria Feneri for her helpful comments.

Abbreviations

DT, Decision Trees; GIS, Geographic Information System; IQR, Interquartile Range; LGBM, Light Gradient Boosting Models; LIME, Local Interpretable Model-Agnostic Explanations; MAPE, Mean Absolute Percentage Error; MAE, Mean Absolute Error; RF, Random Forest; SHAP, Shapley Additive Explanations; XGBoost, Extreme Gradient Boosting.

Availability of data and materials

All data generated or analyzed during this study are included in this published article at the link: <https://drive.google.com/drive/folders/142-YkH6WtpKnRS0YuA8rglowneBqvTML>

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contributions

Conceptualization: GK, GT; Data collection: GK; Formal analysis: GK, Investigation: GK, Writing original manuscript: GK, GT, DV. All authors read and approved the final manuscript.

Received: 5 February 2023 Accepted: 3 October 2023 Published online: 17 October 2023

References

1. Truong Q, Nguyen M, Dang H, Mei B (2020) Housing price prediction via improved machine learning techniques. *Proc Comput Sci* 174:433–442. <https://doi.org/10.1016/j.procs.2020.06.111>
2. Nadai MD, Lepri B (2018) The economic value of neighborhoods: predicting real estate prices from the urban environment. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), pp 323–330
3. Baldominos A, Blanco I, Moreno A, Iturrarte R, Bernárdez Ó, Afonso C (2018) Identifying real estate opportunities using machine learning. *Appl Sci* 8(11):2321. <https://doi.org/10.3390/app8112321>
4. Park B, Bae JK (2015) Using machine learning algorithms for housing price prediction: the case of Fairfax county, Virginia housing data. *Expert Syst Appl* 42(6):2928–2934. <https://doi.org/10.1016/j.eswa.2014.11.040>
5. Truong Q, Nguyen M, Dang H, Mei B (2020) Housing price prediction via improved machine learning techniques. *Proc Comput Sci* 174:433–442. <https://doi.org/10.1016/j.procs.2020.06.111>
6. Ren C, Tong S (2008) Health effects of ambient air pollution - recent research development and contemporary methodological challenges. *Environ Health* 7:56. <https://doi.org/10.1186/1476-069X-7-56>
7. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and health impacts of air pollution: a review. *Front Public Health* 8:14. <https://doi.org/10.3389/fpubh.2020.00014>
8. Keswani A, Akselrod H, Anenberg SC (2022) Health and clinical impacts of air pollution and linkages with climate change. *NEJM Evid* 1(7):2200068. <https://doi.org/10.1056/EVIDra2200068>. <https://evidence.nejm.org/doi/pdf/10.1056/EVIDra2200068>
9. de Paiva Vianna KM, Cardoso MRA, Rodrigues R (2015) Noise pollution and annoyance: an urban soundscapes study. *Noise Health* 17:125–133

10. Koprowska K, Łaskiewicz E, Kronenberg J, Marcińczak S (2018) Subjective perception of noise exposure in relation to urban green space availability. *Urban For Urban Greening* 31:93–102. <https://doi.org/10.1016/j.ufug.2018.01.018>
11. Aletta F, De Coensel B, Lindborg P (2021) Editorial: human perception of environmental sounds. *Front Psychol* 12:714591. <https://doi.org/10.3389/fpsyg.2021.714591>
12. Popescu D (2020) Case study of the environmental noise and its perception in the city of Cluj-Napoca, Romania. *Arch Acoust* 45(4):625–631
13. Mitchell A, Oberman T, Aletta F, Erfanian M, Kachlicka M, Lionello M, Kang J (2022) The international soundscape database: an integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information. <https://doi.org/10.5281/zenodo.6331810>
14. Chen T, Guestrin C (2016) XGBoost. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York. <https://doi.org/10.1145/2939672.2939785>
15. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, vol 30, pp 3146–3154
16. Imran ZU, Waqar M, Zaman A (2021) Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Fundam Inform* 1:11–23. <https://doi.org/10.22995/scmi.2021.1.1.03>
17. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
18. Wolpert DH (1992) Stacked generalization. *Neural Netw* 5(2):241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
19. Xue C, Ju Y, Li S, Zhou Q, Liu Q (2020) Research on accurate house price analysis by using GIS technology and transport accessibility: a case study of Xi'an, China. *Symmetry* 12(8):1329. <https://doi.org/10.3390/sym12081329>
20. Kang Y, Zhang F, Peng W, Gao S, Rao J, Duarte F, Ratti C (2021) Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy* 111:104919. <https://doi.org/10.1016/j.landusepol.2020.104919>
21. Chiarazzo V, Caggiani L, Marinelli M, Ottomanelli M (2014) A neural network based model for real estate price estimation considering environmental quality of property location. *Transp Res Proc* 3:810–817. <https://doi.org/10.1016/j.trpro.2014.10.067>
22. Zou G, Lai Z, Li Y, Liu X, Li W (2022) Exploring the nonlinear impact of air pollution on housing prices: a machine learning approach. *Econ Transp* 31:100272. <https://doi.org/10.1016/j.ecotra.2022.100272>
23. Blanco JC, Flindell I (2011) Property prices in urban areas affected by road traffic noise. *Appl Acoust* 72(4):133–141. <https://doi.org/10.1016/j.apacoust.2010.11.004>
24. Brandt S, Maennig W (2011) Road noise exposure and residential property prices: evidence from Hamburg. *Transp Res, Part D, Transp Environ* 16(1):23–30. <https://doi.org/10.1016/j.trd.2010.07.008>
25. Szczepańska A, Senetra A, Wasilewicz-Pszczółkowska M (2015) The effect of road traffic noise on the prices of residential property – a case study of the Polish city of Olsztyn. *Transp Res, Part D, Transp Environ* 36:167–177. <https://doi.org/10.1016/j.trd.2015.02.011>
26. Tsao H-C, Lu C-J (2022) Assessing the impact of aviation noise on housing prices using new estimated noise value: the case of Taiwan Taoyuan international airport. *Sustainability* 14(3):1713. <https://doi.org/10.3390/su14031713>
27. Morano P, Tajani F, Di Liddo F, Darò M (2021) Economic evaluation of the indoor environmental quality of buildings: the noise pollution effects on housing prices in the city of Bari (Italy). *Build* 11(5):213. <https://doi.org/10.3390/buildings11050213>
28. Bruno DE, Barca E, Goncalves RM, de Araujo Queiroz HA, Berardi L, Passarella G (2018) Linear and evolutionary polynomial regression models to forecast coastal dynamics: comparison and reliability assessment. *Geomorphology* 300:128–140. <https://doi.org/10.1016/j.geomorph.2017.10.012>
29. Giustolisi O, Savic D (2009) Advances in data-driven analyses and modelling using epr-moga. *J Hydroinform* 11:225–236. <https://doi.org/10.2166/hydro.2009.017>
30. Chiarini B, D'Agostino A, Marzano E, Regoli A (2020) The perception of air pollution and noise in urban environments: a subjective indicator across European countries. *J Environ Manag* 263:110272. <https://doi.org/10.1016/j.jenvman.2020.110272>
31. Rico-Juan JR, Taltavull de La Paz P (2021) Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Syst Appl* 171:114590. <https://doi.org/10.1016/j.eswa.2021.114590>
32. Farcaş F, Sivertun Å (2012) Road traffic noise: GIS tools for noise mapping and a case study for Skane region. In: International archives of the photogrammetry, remote sensing and spatial information sciences, vol 34
33. Bocher E, Guillaume G, Picaut J, Petit G, Fortin N (2019) Noisemodelling: an open source GIS based tool to produce environmental noise maps. *ISPRS Intl J Geo-Inf* 8(3):130. <https://doi.org/10.3390/ijgi8030130>
34. Grubesa S, Suhanek M (2020) Traffic noise. In: Siano D, González AE (eds) Noise and environment. IntechOpen, Rijeka. Chapter 5. <https://doi.org/10.5772/intechopen.92892>
35. Begou P, Kassomenos P, Kelessis A (2020) Dataset on the road traffic noise measurements in the municipality of Thessaloniki, Greece. *Data Brief* 29:105214. <https://doi.org/10.1016/j.dib.2020.105214>
36. Yao XA (2020) Georeferencing and geocoding. In: Kobayashi A (ed) International encyclopedia of human geography, 2nd edn. Elsevier, Oxford, pp 111–117. <https://doi.org/10.1016/B978-0-08-102295-5.10548-7>. <https://www.sciencedirect.com/science/article/pii/B9780081022955105487>
37. Faridul H, Pouli T, Chamaret C, Stauder J, Reinhard E, Kuzovkin D, Treméau A (2015) Color mapping: a review of recent methods, extensions, and applications. *Comput Graph Forum* 35:59–88. <https://doi.org/10.1111/cgf.12671>
38. Sharma G, Wu W, Dalal EN (2005) The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *Color Res Appl* 30:21–30
39. Mokrzycki W, Tatol M (2009) Perceptual difference in $l^*a^*b^*$ color space as the base for object colour identification. <https://doi.org/10.13140/2.1.1160.2241>
40. Luo M, Cui G, Rigg B (2001) The development of the CIE 2000 colour-difference formula: Ciede2000. *Color Res Appl* 26:340–350. <https://doi.org/10.1002/col.1049>
41. Long M (2014) 3 - human perception and reaction to sound. In: Long M (ed) Architectural acoustics, 2nd edn. Academic Press, Boston, pp 81–127. <https://doi.org/10.1016/B978-0-12-398258-2.00003-9>. <https://www.sciencedirect.com/science/article/pii/B9780123982582000039>

42. Agency, DM (1991) Department of defense world geodetic system 1984: its definition and relationships with local geodetic systems. Defense Technical Information Center
43. Lambert JH (2022) In: Caddeo R, Papadopoulos A (eds) Notes and comments on the composition of terrestrial and celestial maps. Springer, Cham, pp 367–422
44. Potdar K, Pardawala TS, Pai CD (2017) A comparative study of categorical variable encoding techniques for neural network classifiers. *Int J Comput Appl* 175(4):7–9
45. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer series in statistics. Springer, New York
46. Greenwell BM, Boehmke BC, McCarthy AJ (2018) A simple and effective model-based variable importance measure. [arXiv:1805.04755](https://arxiv.org/abs/1805.04755)
47. Altmann A, Tolosi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* 26:1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>
48. Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?”: explaining the predictions of any classifier. [arXiv:1602.04938](https://arxiv.org/abs/1602.04938)
49. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. [arXiv:1705.07874](https://arxiv.org/abs/1705.07874)
50. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. [arXiv:1206.2944](https://arxiv.org/abs/1206.2944)
51. Zafar MI, Dubey R, Bharadwaj S, Kumar A, Paswan KK, Srivastava A, Tiwary SK, Biswas S (2023) GIS based road traffic noise mapping and assessment of health hazards for a developing urban intersection. *Acoust* 5(1):87–119. <https://doi.org/10.3390/acoustics5010006>
52. Weisser A, Buchholz JM (2019) Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. *J Acoust Soc Am* 145(1):349
53. McAlexander T, Gershon R, Neitzel R (2015) Street-level noise in an urban setting: assessment and contribution to personal exposure. *Environ Health* 14:18. <https://doi.org/10.1186/s12940-015-0006-y>
54. Niesten J, Tenpierik M, Krimm J (2022) Sound predictions in an urban context. *Build Acoust* 29(1):27–52. <https://doi.org/10.1177/1351010X211034665>
55. Liu F, Jiang S, Kang J, Wu Y, Yang D, Meng Q, Wang C (2022) On the definition of noise. *Humanit Soc Sci Commun* 9(1):406. <https://doi.org/10.1057/s41599-022-01431-x>
56. Konopka W, Pawlaczyk-Luszczynska M, Śliwińska-Kowalska M (2014) The influence of jet engine noise on hearing of technical staff. *Med Pr* 65:583–592. <https://doi.org/10.13075/mp.5893.00045>
57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
