Causal discovery on geospatial data: Investigating the impact of amenity proximity on residential property values

Shalaleh Hillebrand[1,2], Arjen Hommersom[1], Jesse Heyninck[1], and Natasha Alechina[1]

[1] Department of Computer Science, Open University of the Netherlands
shalaleh.hillebrand@ou.nl
[2] a.s.r. real estate

Abstract. This study investigates the causal effects of proximity to amenities such as schools and childcare centers on local housing prices. Understanding this relationship is critical for urban planning and policy decisions. We apply causal discovery algorithms to both synthetic and real-world housing datasets to identify potential causal relationships. These types of data provide a set of challenging properties for applying such algorithms. In particular, this study emphasizes the importance of using location as an effect modifier when dealing with clustered geospatial data. Our analysis reveals that proximity to educational facilities has a significant impact on housing prices, with variations across different postcode areas, however, this impact is not always positive. In fact the identified causal structure, indicates a negative impact of proximity of many amenities on the house prices. We conclude that causal discovery algorithms have the potential to provide novel insights into the determinants of house prices.

1 Introduction

The residential property tax in the Netherlands is based on the estimated value of the properties which is carried out yearly by the municipalities (known as WOZ value[1]). The WOZ value of the residential properties is calculated based on the market analysis of the residential property transactions (provided by The Netherlands' Cadastre, Land Registry and Mapping Agency). The WOZ value also serves as a crucial metric for investors, providing insights into emerging neighborhoods with strong growth potential. In this study, our objective is to identify the causal structure underlying WOZ value. Such a structure allows us to determine which factors have a causal effect on WOZ value, and further, to evaluate how interventions on these factors would influence the outcome which can help the investors and municipalities to pinpoint the potential areas of co-operation and making informed policy or investment decisions.

[1] De Wet Waardering Onroerende Zaken or WOZ is the market value of a property at a fixed reference date.

It should be noted that the purpose of this study is not to compare forecasting performance. Instead, this work specifically focuses on estimating the capitalization of public amenities and identifying the causal structure through which these amenities – particularly educational facilities – influence the WOZ value. Understanding which neighborhood amenities positively influence property values (WOZ value) creates an opportunity for collaboration between real estate investors and local governments. These amenities not only improve the quality of life for residents, but also contribute to the economic development of the area [25, 13]. For investors, this means that strategic interventions can help increase the value of properties while also generating positive social impact. This approach supports both financial returns and community well-being, making it a valuable strategy for sustainable urban development.

One approach to identify the causal structure is through randomized experiments, which, however, is not feasible in many fields, including the application at hand. An alternative method of identifying causal structure is by applying causal discovery algorithms on observational data [19], which analyzes the data to uncover causal relationships using statistical dependencies that are produced by a causal mechanism [23]. Such algorithms are strongly rooted in AI [16], and have gained popularity in recent years in many application areas such as medicine, recommendation, economics and other domains [26]. While quite some benchmark datasets have been used in the past for evaluating causal discovery algorithms, such datasets typically have a fairly simple structure, with a few thousand samples and a limited number of features (see e.g. [15]). This study contributes to the evaluation of the suitability of causal discovery algorithms in a challenging practical application.

The main challenge of dealing with housing data is the fact that it contains spatial heterogeneity, with clusters of cities, rural areas and uninhabited regions such as agricultural land, forests, or bodies of water, which results in a non-random spatial patterns. For example, urban and rural areas differ significantly in terms of distances to public amenities and the housing market. Furthermore, within cities, adjacent neighborhoods tend to share similar characteristics with one another. However, most causal discovery algorithms assume that the data are independent and identically distributed (IID). This IID assumption is not necessarily valid when working with geospatial data, because such data involve observations which could be related to one another based on their locations. In addition to spatial characteristics, the data follows a non-Gaussian distribution which adds to the complexity of the causal discovery process. The rather large available dataset on housing data (with more than 350.000 entries) and the large number of potential variables also pose a computational challenge. In this paper, we show all of these challenges can be addressed to obtain interesting insights into the problem at hand.

This paper is organized as follows. In Section 2, we review several studies which examine the effects of proximity to educational and childcare facilities on housing prices. In Section 3, we describe different causal discovery algorithms and their applicability to the type of data used in this case study. In Section 4, we present the conducted experiments beginning with experiments on synthetic data resembling the characteristics of the case study (Section 4.1), we identify the causal discovery algorithms that are applicable to our setting. We use a spatially aware analysis in which location will be used as an effect modifier and aggregation of the variables in adjacent nodes will be added to the analysis as a new variable [14]. In Section 4.2 we present preliminary results on the effects of proximity to public amenities on WOZ values and we discuss our findings, limitations and suggestions for future studies in Section 5 and Section 6.

2 Related Works

The impact of schools and childcare facilities on the residential market has been studied for many years. Studies have shown the interconnected impact of supply, demand and various other factors on the capitalization of these types of amenities. In this section, we discuss some important findings.

Proximity to educational facilities is believed to play a significant role in influencing residential property values. Several studies have shown that being near schools and kindergartens generally increases house prices. For example, Bergantino et al. (2022) [2] find a positive effect of proximity to private kindergartens and schools on neighborhood house prices, using longitudinal data from 2010 to 2017. Their analysis assumes that the structural characteristics of the kindergartens and municipalities do not change during the period of study.

However, the impact of proximity may depend on other factors as well. For instance Theisen and Emblem (2018) [24] emphasize the importance of considering new kindergartens that are opened during the study period. Their spatial econometric model shows a positive effect of proximity to kindergartens on the house prices, however according to this model the house prices decline in the immediate vicinity of kindergartens. Their empirical model on the other hand shows the positive effect of proximity to kindergartens especially for single family homes. The house prices according to the empirical model do not decline in the immediate vicinity of the kindergartens.

Another key driver of this capitalization is school quality. This capitalization effect is stronger in areas with limited developable land and higher home ownership [9]. In markets where housing supply is constrained (e.g. due to zoning regulations) the value of school quality is more strongly reflected in prices. On the other hand, in areas with more elastic supply, the same demand for school quality results in less price change [3].

Policy mechanisms also influence how educational access is capitalized. Reback (2005) [17] examines the impact of policies that allow students to attend schools in districts other than the one they reside. Historically, the residents pay a premium to purchase residential properties in the areas with good public schools. Their study demonstrates the freedom to enroll in the schools in other districts increases property values in less popular districts and a decrease in popular districts with incoming transfer students. The time period for the capitalization is said to be around eight years.

Many of the quantitative works mentioned above follow a hedonic price model [18] to quantify the capitalization of proximity to amenities. While the hedonic price model is widely used in real estate research to analyze house prices, it is not designed to identify causal relationships. However, to understand the potential effects of interventions, we need causal inference, which requires knowledge of the underlying causal structure and relationships. Therefore, identifying the causal structure is very crucial, which we will discuss further in the next section.

3 Causal Discovery

In this section, we discuss various causal discovery algorithms with a particular focus on those that are used in this paper. We highlight one of the main assumptions of these algorithms, that the data are independent and identically distributed (IID). Finally, we discuss the applicability of these algorithms to non-Gaussian data.

3.1 Causal discovery algorithms

Causal discovery can be categorized into three main families of algorithms: Constraint-based algorithms, Score-based algorithms, and Functional causal models.

Constraint-based algorithms search for the Markov equivalence class of graphs which match the conditional independence over the variables in the population. A description of the algorithms from this family, which are used in the present work, is provided below.

PC Algorithm [23] This algorithm starts from the complete undirected graph with an edge between each pair of variables. The edges between two variables X and Y are eliminated if they are unconditionally independent. For every three variables that share an edge between them, the edge between two of the three is removed if they are independent of each other conditional on the third variable. This procedure is repeated with increasing the number of subset variables that is conditioned on. In the later steps the direction of the edges are determined using a set of orientations rules, resulting in a partially directed acyclic graph (PDAG). Many different statistical procedures could be applied to this algorithm to investigate the dependence of the variables [6].

Fast Causal Inference or FCI algorithm[23] FCI is a variation of the PC algorithm which can take unobserved confounders into account and detect hidden confounding. Like PC, it starts with a fully connected graph. Then it calculates the marginal dependency between each pair of variables, if the pair are marginally independent, the edge between them is removed. For all the pairs with a shared edge in between, the conditional independency is calculated where size of the conditioning set is increased each time. Due to causal insufficiency some conditional independencies might have not correctly detected. During the second phase, a broader set of conditioning variables is used to detect the missed independencies. In the final phase, the FCI algorithm applies orientation rules to produce a partial ancestral graph (PAG).

Really Fast Causal Inference or RFCI [5] This is a variation of FCI which is computationally less expensive and is suitable when dealing with large datasets.

Score-based algorithms (e.g. Chickering 2002 [4]) use goodness of fit score instead of independence tests. We will not further use these family of algorithms in this work, as these algorithms typically assume linear-Gaussian distributions, which does not match the data at hand.

Functional causal models are based on the asymmetry between the cause and effect through taking the data generating process into account. In this work we only consider one algorithm of this family: the Linear Non-Gaussian Acyclic Model (ICA-LiNGAM) [21]. This algorithm enables causal discovery in data which is non-Gaussian with linear relationships between the variables. Through the use of statistical properties of non-Gaussian distributions, it can identify a full causal model. A key assumption of ICA-LiNGAM is causal sufficiency, which requires that all confounders are observed. This algorithm further requires the data to be continuous.

Many causal discovery algorithms result in a graph that represents a Markov equivalence class, where some of the directions of the edges are not defined. In such cases, one can consider some other sources of information such as temporal order or domain knowledge as well as past empirical findings.

3.2 Causal discovery with non-IID data

Most causal discovery algorithms rely on two key assumptions: the faithfulness and Markov assumptions [23]. According to the faithfulness assumption the conditional independencies present in the probability distribution come from the structure of the underlying causal graph. That implies that if two variables in an acyclic directed graph are conditionally independent, then they are d-separated [23]. The Markov assumption states that each variable is independent of its non-effects (non-descendants) given its direct causes (parents) in the causal graph [23].

One requirement to satisfy these two assumptions is that the data are independent and identically distributed (IID), which guarantees that the observations are drawn from the same probability distribution and are independent from each other. Geospatial data involve observations which may be spatially correlated due to their geographical proximity. This leads to spatial dependencies. Many causal discovery algorithms may not perform well under these conditions. Adaptations of algorithms like the Fast Causal Inference (FCI) algorithm have been proposed to handle spatial dependencies by incorporating background knowledge of the spatial structure into the causal discovery process. For example, in a work by Mielke et al. [14], where the spatially hierarchical case of directional flow in rivers is studied, the authors introduce a new category of variables to their data points by aggregating the variables upstream of the river that influence the variables at an observation point. By introducing this new category of variables, the problem turns into a spatially-aware case.

3.3 Causal discovery with non-Gaussian data

Constraint-based algorithms perform a conditional independence test to construct the Markov-equivalent graph. However, many of these tests assume that the data follows a Gaussian distribution. This assumption often fails to hold in real-world scenarios. The same is the case for the data in this study. Although many traditional independence tests can only detect linear relationships, kernel independence tests are applicable on non-Gaussian data with complex relationships [28]. The Hilbert–Schmidt Independence Criterion (HSIC), a kernel independence test used in the analysis in Section 4, measures the distance between the joint distribution of the variables and the product of the marginal distribution of the variables. A distance of zero indicates the independence between the variables.

4 Experiments

In this section we consider the effectiveness of different causal discovery strategies in our geospatial case involving clustered data (and thus to adjust for the confounding effects of location). In Section 4.1, we describe experiments using synthetic data for which we know the true causal graph. The synthetic data is generated in a way to reflect the dependency structure observed in our case study. By performing experiments on synthetic data, for which the true causal graph is known, we can evaluate the performance of different algorithms and strategies and identify the most effective ones. From the results, we then move to a case study with real-world data, which is described in Section 4.2.

4.1 Experiments with synthetic data

We perform three sets of experiments on synthetic data. In Experiment I, the data is organized into clusters representing distinct geographic regions—such as cities or rural areas—where variables within each cluster exhibit intra-cluster dependency, simulating shared regional characteristics. Our strategy in this case involves adjusting for clusters with small inter-cluster distances to improve causal discovery. In Experiment II, an additional layer of dependency is introduced among data points within each cluster to reflect spatial adjacency, such as adjacent neighborhoods within a city. These dependency structures allow us to model both macro-level regional variation and micro-level spatial interactions, which are critical in real estate related analyses. Finally, in Experiment III, we consider the computational cost of various algorithms.

Experiment I: Macro-Level Simulation To create the synthetic dataset which simulates cluster-level dependency, we generate data with three variables (X_1, X_2, and X_3), for which we draw samples from a non-Gaussian distribution (i.e. a uniform distribution). Clusters are formed by introducing an upper-bound of the uniform distribution that varies between clusters.

If there are K clusters indexed by $k = 1, ..., K$, for each cluster, we introduce upper-bounds $c_{1k},\ c_{2k},\ c_{3k} > 0$. The three variables are then generated as follows:

$$X_1^{(k)} \sim \mathit{Unif}(0, c_{1k}) \tag{4.1}$$

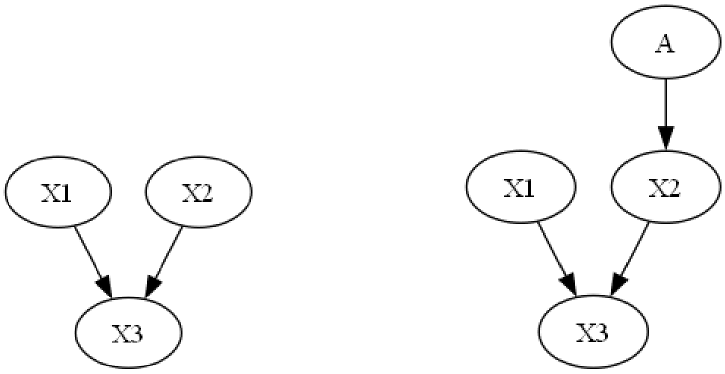
$$X_2^{(k)} \sim \mathit{Unif}(0, c_{2k}) \tag{4.2}$$

$$X_3^{(k)} \ = \ X_1^{(k)} + X_2^{(k)} \ + \ \epsilon \tag{4.3}$$

where $\epsilon \sim \mathit{Unif}(0, c_{3k})$.

The upper-bounds c_{ik} are defined as a function of k in order to control the scale of variables; $c_{1k} = 1 + 0.5 \cdot k$, while $c_{2k} = 1 + 0.3 \cdot k$ and $c_{3k} = 1 + 0.1 \cdot k$ have similar functional forms. The experiment was conducted using a maximum of 15 clusters.

The true graph for the data generated according to Equations 4.1–4.3 has no edge between X_1 and X_2, and X_3 is a collider with edges from X_1 and X_2 toward it (see Figure 1a).



(a)

(b)

Fig. 1: (a) The true graph generated according to Equations 4.1–4.3. (b) The true graph generated using Equations 4.1, 4.3, and 4.4.

We perform the causal discovery analysis using four algorithms: FCI, RFCI and PC and ICA-LiNGAM. Since the data is non-Gaussian a kernel-based conditional independence test is used for constraint based algorithms (PC, FCI, RFCI). When using the data from all clusters, a dependency between the two variables X_1 and X_2 is detected for all four algorithms. When we repeat the analysis separately for group of clusters that are similar in characteristics, all four algorithms result in the correct causal graph.

These results show that when working with clustered data—such as cities vs. rural areas—it is important to analyze different types of clusters separately. However, it is not necessary to analyze each individual cluster on its own. Instead, clusters with similar characteristics can be grouped together and analyzed as one set. What is not recommended is combining all clusters into a single analysis, as this can lead to incorrect conclusions about causal relationships.

Experiment II: Micro-Level Simulation In a subsequent set of tests we simulate localized interaction between adjacent objects within individual clusters. To achieve this, we randomly generate an adjacency matrix for each cluster which defines pairwise proximity among nodes. We then alter the generation of X_2 to reflect the effect of adjacency by introducing an extra term which is the average value of X_1 from adjacent nodes.

More formally, the data generating process is as follows. Let nodes be denoted as a pair (i,k) where node i belongs to cluster k. Let $\mathcal{V}_k$ be the set of nodes in cluster k. Given a node (i,k), its neighbors are denoted by $\mathcal{N}(i) \subseteq \mathcal{V}_k$. Then:

$$X_2^{(i,k)} \;=\; \mathit{Unif}(0, c_{2k}) \;+\; A_i \tag{4.4}$$

where A_i is the neighbor aggregate of node i and is defined as:

$$A_i \coloneqq \frac{1}{n_i} \sum_{j \in \mathcal{N}(i)} X_1^{(j)} \tag{4.5}$$

The true causal graph for experiment II is shown in Figure 1b.

In this experiment, we want to investigate whether ignoring the information that is shared between adjacent objects affects the results of causal discovery and leads to false relationships. To verify this, we first perform experiments without including variable A in the causal discovery analysis. The results—using only variables X_1, X_2, and X_3 —are the same as in experiment I: if we do not adjust for similar clusters, the causal graph is incorrect. But when we do adjust for similar clusters, the correct causal graph is identified. This means that including or excluding variable A does not change the final result, as long as we adjust for cluster similarity. In other words, localized dependencies do not affect the outcome if they are not considered—at least in the way the data was generated in this experiment.

When we include variable A in the analysis, PC, FCI and RFCI detect edges from A toward all three variables (X_1, X_2 and X_3) while the only direct edge should be toward X_2. ICA-LiNGAM detects the correct causal graph in some iterations however the results are not stable through iterations. This is because the ICA component in ICA-LiNGAM uses random initialization, and therefore results can vary between runs and may sometimes converge to local optima. Adjusting for clusters improves the performance of all four algorithms, with each producing the correct causal graph.

Based on these two experiments, we conclude that it is essential to adjust for similar clusters when dealing with both inter-cluster and intra-cluster dependencies. In the experiment using real-world data described in Section 4.2, we use data from specific urbanity levels only. The characteristics of these urbanity levels are described in detail within that section.

Experiment III: Computational Cost Many constraint-based algorithms suffer from very high computational cost as the number of samples and variables increases. This arises from the growth of the conditioning sets considered in conditional independence tests. In the worst case for a dense graphs the number of conditional independence tests considered by e.g. PC algorithm is $N_{\text{tests}}(p,k) = \frac{p(p-1)}{2} \sum_{s=0}^{\min(k,p-2)} \binom{p-2}{s}$ where p is the number of variables and k is the maximum conditioning set size [27]. When using kernel HSIC, each conditional independence test costs $O(n^2)$ where n is the number of samples. Together this will result in a worst-case runtime $O\big(N_{\text{tests}}(p,k) \cdot n^2\big)$.

In contrast, ICA-LiNGAM does not face this limitation and has a complexity of $O(np^3 + p^4)$ [20]. Figure 2 presents the computation time across different algorithms and varying sample sizes and number of variables.

In our case study, discussed in Section 4.2, we deal with tens of thousands of samples and many variables. Consequently, it is necessary to be selective in choosing the algorithms to apply. Since ICA-LiNGAM, after adjusting for clusters, produces a causal graph that resembles the true graph, and since its computation time is on the order of seconds (i.e., less than a second to a few seconds depending on the number of variables), we chose to use ICA-LiNGAM in the case study described next.

4.2 Case Study

As it is mentioned in the introduction, we aim to identify the causal relationship between the vicinity to amenities and the WOZ value. The data that is used in this study is gathered from publicly available source (`https://www.cbs.nl`). CBS is an institution of Dutch government which collects and publishes data and statistics on a wide range of societal topics.

The data that we have used in this study are from 2017 and include specifically the following datasets:

- Key Statistics for Districts and Neighborhoods: This dataset includes many variables among which are the WOZ value, number of residents in a neighborhood and the subdivision (age groups, education level, marital status etc), Income level and Housing stock.
- Proximity to Facilities and Distance to Location for districts, and neighborhood: This dataset contains the following variables among others: Distance to supermarkets, schools, or parks, hospitals, restaurants and cafes.

The data are available at different granularity levels. For this work we use the data with granularity level of 6-digits postcode or PC6[2]. To give an idea of the

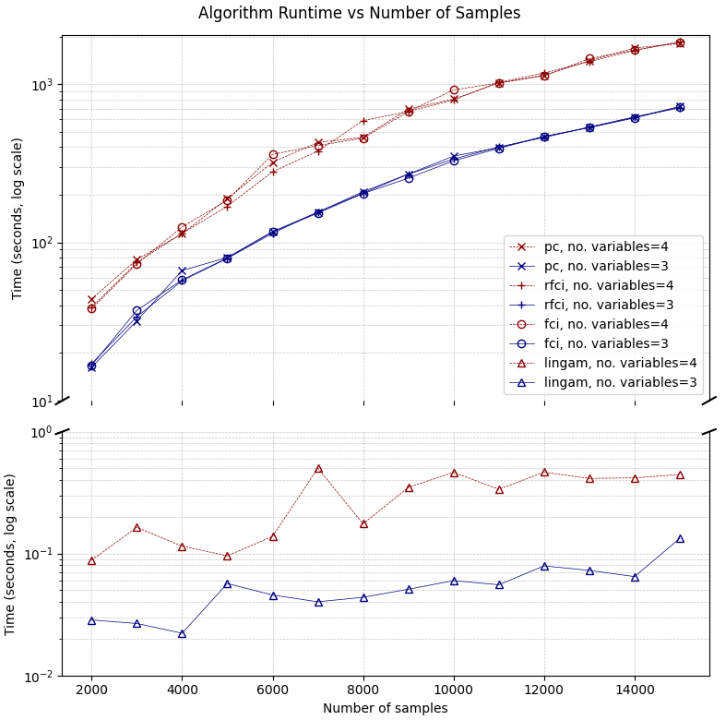


Fig. 2: Computation time (in seconds, log scale) of four causal discovery algorithms (PC, RFCI, FCI, ICA-LiNGAM) as a function of sample size and number of variables. The four algorithms were implemented using the R package pcalg [11] [8]. All analyses were conducted on a personal computer.

size of such areas, Figure 3 shows the division of PC6 areas in Amsterdam, focusing on the south region.

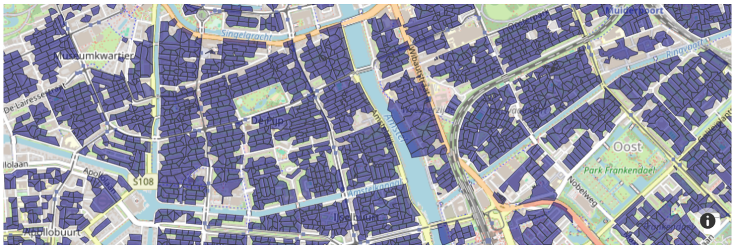


Fig. 3: An example of PC6 zones with residential function in Amsterdam South

The PC6 areas are not isolated points. They are spatially connected and they share similar characteristics with the adjacent areas. Therefore we aim to introduce the propagated information due to spatial connectivity as a new feature (similar to [14]). Different aggregations can be assumed to take the information propagated from neighboring points into consideration. In this work we use the average of the WOZ value at the adjacent PC6 zones as new variable. The size of PC6 areas varies, where urban areas tend to have smaller PC6 zones while those located in less densely populated areas appear larger. The influence of a single PC6 area on its adjacent counterparts is therefore less significant in rural regions compared to those in urban areas. In cities the PC6 areas are smaller and more tightly packed, where for example a single street might contain several PC6 zones. This emphasizes the importance of analyzing rural and urban areas separately. In the Netherlands, a standard measure of urbanity is the average surrounding address density, which classifies areas into following five levels:

- Very highly urbanized: an average density of 2,500 or more addresses per km^2.
- Highly urbanized: an average density of 1,500 to 2,500 addresses per km^2.
- Moderately urbanized: an average density of 1,000 to 1,500 addresses per km^2.
- Slightly urbanized: an average density of 500 to 1,000 addresses per km^2.
- Non-urbanized: an average density of fewer than 500 addresses per km^2.

In this work we use PC6 areas with an urbanity level of 2,3 and 4 which contains more than 150000 entries. The variables included in this preliminary analysis are as follows: the number of various amenities within a 1 km radial distance (primary schools, kindergartens, afterschool daycares, and supermarkets); the number of secondary schools within a 3 km radius; the distance to

[2] 6-digit postcode where the first 4 digits represent, city and part of the city, while the 2 letters at the end narrows it down to specific street or a part of a street.

main roads, train stations, and the nearest transfer stations; the WOZ value of the corresponding PC6 zone; and the average WOZ value of its adjacent zones.

As discussed in Section 4.1, we employed the ICA-LiNGAM algorithm for this analysis due to its low computational complexity compared to other causal discovery methods. In the first analysis, all the available PC6 objects in the dataset are incorporated. The resulting causal graph, is presented in Figure 4.

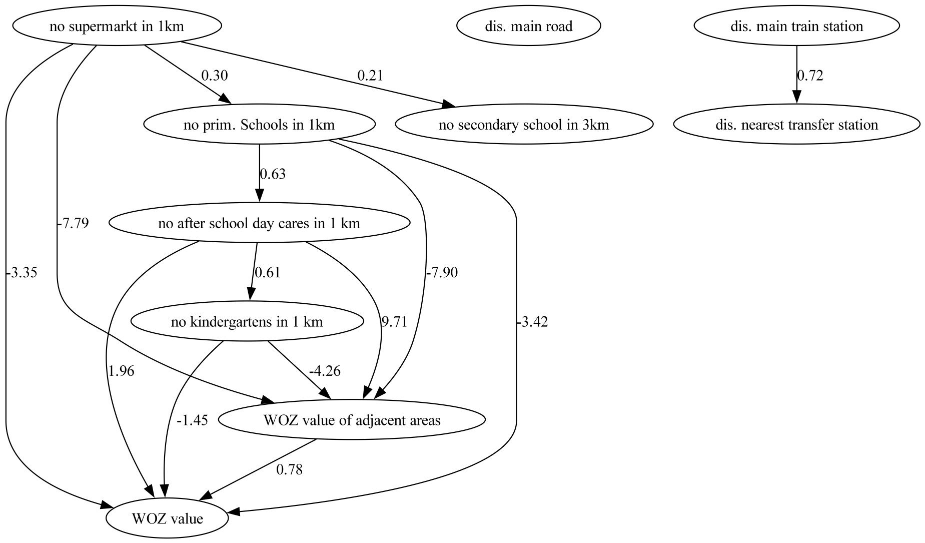


Fig. 4: Identified causal structure with ICA-LiNGAM using all the data

When we examine the weights on the edges, we observe that many of the amenities have a negative effect on the WOZ value of residential properties. For example, the results show that the number of supermarkets, primary schools and kindergartens within a 1 km radius have a negative influence on the WOZ value. This finding can be partially explained by urban development patterns. In newly developed neighborhoods, housing stock is more modern and energy-efficient; however, educational amenities are often still under development [1]. Although these areas have limited access to certain amenities, property values tend to be relatively high as homebuyers are generally willing to pay a premium for newly constructed homes [7]. Additionally, urban planning policies often concentrate supermarkets in densely populated areas characterized by smaller housing units [12], which are typically associated with lower WOZ values. Another interpretation of this finding could be that the close proximity to those amenities may lead to excessive noise. For instance as the number of supermarkets in 1 km distance to a PC6 zone increases, this may potentially lead to heavier traffic in an area and thus to an increased level of congestion and noise. Also in less densely populated rural neighborhoods with villas, the immediate vicinity to educational amenities is not desired.

For our second experiment we focus on the residential properties with a WOZ value less than 320k (corresponding to the mean plus one standard deviation). In Figure 5 the causal graph identified with this set of data is shown.

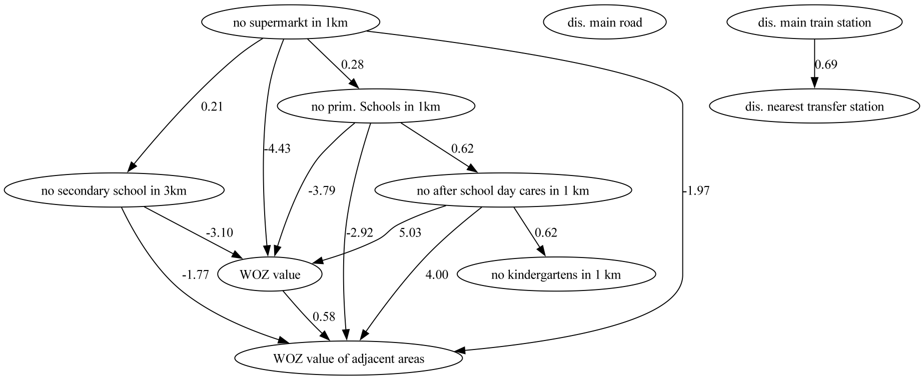


Fig. 5: Identified causal structure with ICA-LiNGAM using the data with WOZ value less than 320k.

The reversed edge from WOZ value toward the WOZ value of adjacent areas is one of the key differences when comparing the two graphs. The overall structure of the identified causal graph remains largely consistent with the one derived from the full dataset. In both graphs, we observe that the distances to major roads, train stations, and transfer stations do not exhibit any significant causal effect on the WOZ value.

5 Discussion

We have made several methodological and analytical choices and certain assumptions throughout this study. While these decisions provide a good foundation for this research, it also in some cases introduces some limitations. In this section we are discussing some of these choices in more details.

The Impact of Spatial Granularity When data is aggregated at the neighborhood level, the granularity is too coarse, which results in a smoothing effect, and therefore localized patterns are no longer available. This reduces the sensitivity of the variables to changes in the others. Consequently, causal discovery methods applied to such data yield inaccurate causal structures. To address this limitation, we employ a finer spatial resolution, namely PC6 zones, which preserves local heterogeneity.

Area vs. Point data One question which might arise is whether to use 2D features in the analysis due to the spatial nature of our study case. The PC6 areas however are not natural spatial features but rather administrative ones. A PC6 zone contains few addresses each with specific characteristics, but the reported key figures only contain the aggregated information on those addresses. Given that the information is already aggregated at the area level, treating each PC6 area as a single data point is a justifiable simplification.

Non-Gaussian data The non-Gaussian dataset increases the complexity of the problem since many algorithms require a Gaussian data type. For instance, in the PC algorithm, the Gaussianity assumption helps identifying several graphs (Markov equivalent) that meet the relevant conditional independency. Dealing with non-Gaussian data is an important challenge in many applications. One approach is to use a Gaussian copula framework [22], which preserves the dependencies between variables. This property is particularly important since calculating the dependencies (e.g. correlation) between different variables is used to infer causal relationship. Another approach to dealing with non-Gaussian data is to use kernel-based conditional independence tests (KCIT) [28]. In this work we adopted the second method, where we have employed KCIT for constraint based algorithms. However we observed that the computation time was quite high as a result of using this conditional independence test even when applied to the synthetic data with limited number of variables. For this reason, we did not use constraint-based algorithms in the case study. This is a limitation of our current work, which we plan to address in the future, either through parallelization or other efficiency improvements.

Linearity assumption One of the necessary assumptions in a wide range of functional algorithms is that the relationships among variables are linear. The pairwise comparison of the data in our study however shows a non-linear relationship among many variables. The non-linear additive noise models [10] are suggested when dealing with non-linear data generating processes however the computation costs of such algorithms could potentially be quite high comparing with linear models.

Continuous data type Although we are using the variables such as number of school in 1 km distance from the PC6 area which are inherently discrete, these values are not reported as exact counts. Because these variables in our dataset are produced through spatial smoothing, they should be regarded as continuous rather than discrete.

6 Conclusions

In this study, we employed causal discovery techniques (specifically the ICA-LiNGAM algorithm) as an alternative to traditional econometric approaches such as the hedonic pricing model. This allowed us to investigate the complex relationships among the various determinants of housing prices. Unlike models that rely on predefined assumptions about variable interactions, causal discovery enables a data-driven investigation of underlying structures.

Our findings highlight the importance of including more variables which might be the missing confounders in the preliminary analyses. For instance, the causal graphs suggest that certain amenities, such as supermarkets and educational facilities, may have unexpected effects on property values depending on urban planning strategies and neighborhood development stages. Many macroeconomic factors also influence housing prices. For instance, the average education level within a neighborhood could act as a confounder for the effect of distance to educational facilities on the WOZ value.

In this paper, we have presented a causal graph based on housing data using the ICA-LiNGAM algorithm. In future work, we plan to extend the analysis by applying other causal discovery algorithms. Furthermore, this study investigates only the instantaneous effects of proximity of amenities on the WOZ value, while the impact of such factors on house prices might take several years to fully realize, e.g., Reback et al. 2005 [17] uses an eight year gap between the introduction of a policy in school choice program and its effects on property values. Moreover, this study has focused only on amenities in a close proximity (e.g. number of amenities within 1 km radius). This close proximity however might lead to increased level of noise and other discomforts. Homeowners are sometimes willing to travel longer distance if it means their exposure to public disturbances at their property location is reduced.

Overall, this approach provides a better understanding of housing market dynamics and creates opportunities for further research into the various determinants of housing prices.

References

1. Aksoy, E.S., Venverloo, T., Benson, T., Duarte, F.: Evaluating amenity access of new and repurposed housing within the 15-Minute City framework in Amsterdam. Discov Cities **2**(1), 47 (2025). https://doi.org/10.1007/s44327-025-00087-x
2. Bergantino, A.S., Biscione, A., De Felice, A., Porcelli, F., Zagaria, R.: Kindergarten Proximity and the Housing Market Price in Italy. Economies **10**(9), 222 (2022). https://doi.org/10.3390/economies10090222
3. Cheshire, P., Sheppard, S.: Capitalising the Value of Free Schools: The Impact of Supply Characteristics and Uncertainty. The Economic Journal **114**(499), F397–F424 (2004). https://doi.org/10.1111/j.1468-0297.2004.00252.x
4. Chickering, D.M.: Optimal Structure Identification With Greedy Search. (2002). https://doi.org/10.1162/153244303321897717
5. Colombo, D., Maathuis, M.H., Kalisch, M., Richardson, T.S.: Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann. Statist. **40**(1) (2012). https://doi.org/10.1214/11-AOS940. arXiv: 1104.5617[stat]
6. Glymour, C., Zhang, K., Spirtes, P.: Review of Causal Discovery Methods Based on Graphical Models. Front. Genet. **10**, 524 (2019). https://doi.org/10.3389/fgene.2019.00524

7. Gordon, B.L., Winkler, D.T.: New House Premiums, Market Conditions, and the Decision to Purchase a New Versus Existing House. Journal of Real Estate Research **41**(3), 379–410 (2019). https://doi.org/10.22300/0896-5803.41.3.379
8. Hauser, A., Bühlmann, P.: Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research **13**, 2409–2464 (2012)
9. Hilber, C.A., Mayer, C.: Why do households without children support local public schools? Linking house price capitalization to school spending. Journal of Urban Economics **65**(1), 74–90 (2009). https://doi.org/10.1016/j.jue.2008.09.001
10. Hoyer, P.O., Janzing, D., Mooij, J., Peters, J., Scholkopf, B.: Nonlinear causal discovery with additive noise models
11. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software **47**(11), 1–26 (2012). https://doi.org/10.18637/jss.v047.i11
12. Kesarovski, T., Hernández-Palacio, F.: Time, the other dimension of urban form: Measuring the relationship between urban density and accessibility to grocery shops in the 10-minute city. Environment and Planning B: Urban Analytics and City Science **50**(1), 44–59 (2023). https://doi.org/10.1177/23998083221103259
13. Marquez, J., Casas, F., Taylor, L., De Neve, J.-E.: Economic Development and Adolescent Wellbeing in 139 Countries. Child Ind Res **17**(4), 1405–1442 (2024). https://doi.org/10.1007/s12187-024-10131-8
14. Mielke, K.P., Schipper, A.M., Heskes, T., Zijp, M.C., Posthuma, L., Huijbregts, M.A.J., Claassen, T.: Discovering Ecological Relationships in Flowing Freshwater Ecosystems. Front. Ecol. Evol. **9**, 782554 (2022). https://doi.org/10.3389/fevo.2021.782554
15. Nogueira, A.R., Pugnana, A., Ruggieri, S., Pedreschi, D., Gama, J.: Methods and tools for causal discovery and causal inference. WIREs Data Min & Knowl **12**(2), e1449 (2022). https://doi.org/10.1002/widm.1449
16. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2009)
17. Reback, R.: House prices and the provision of local public services: capitalization under school choice programs. Journal of Urban Economics **57**(2), 275–301 (2005). https://doi.org/10.1016/j.jue.2004.10.005
18. Rosen, S.: Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy **82**(1), 34–55 (1974). https://doi.org/10.1086/260169
19. Rubin, D.B.: Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005). https://doi.org/10.1198/016214504000001880
20. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model, (2011). https://doi.org/10.48550/arXiv.1101.2489. arXiv: 1101.2489[stat].
21. Shimizu, S., Jp, I.A., Hoyer, P.O., Hoyer, P., Hyvarinen, A., Hyvarinen, A., Kerminen, A., Kerminen, A.: A Linear Non-Gaussian Acyclic Model for Causal Discovery. (2006)
22. Sokolova, E., Von Rhein, D., Naaijen, J., Groot, P., Claassen, T., Buitelaar, J., Heskes, T.: Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD. Int J Data Sci Anal **3**(2), 105–119 (2017). https://doi.org/10.1007/s41060-016-0034-x

23. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer New York, New York, NY (1993)
24. Theisen, T., Emblem, A.W.: House prices and proximity to kindergarten – costs of distance and external effects? Journal of Property Research **35**(4), 321–343 (2018). `https://doi.org/10.1080/09599916.2018.1513057`
25. Wong, C.: The Relationship Between Quality of Life and Local Economic Development: An Empirical Study of Local Authority Areas in England. Cities **18**(1), 25–32 (2001). `https://doi.org/10.1016/S0264-2751(00)00051-2`
26. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: A Survey on Causal Inference. ACM Trans. Knowl. Discov. Data **15**(5), 1–46 (2021). `https://doi.org/10.1145/3444944`
27. Zhang, K., Tian, C., Zhang, K., Johnson, T., Jiang, X.: A Fast PC Algorithm with Reversed-order Pruning and A Parallelization Strategy, (2021). `https://doi.org/10.48550/arXiv.2109.04626`. arXiv: `2109.04626[cs]`.
28. Zhang, K., Peters, J., Janzing, D., Scholkopf, B.: Kernel-based Conditional Independence Test and Application in Causal Discovery. In: UAI 2011, Barcelona, Spain, (2011). `https://doi.org/10.48550/arXiv.1202.3775`