HUMAN-MACHINE SYNERGY IN REAL ESTATE SIMILARITY CONCEPT

Małgorzata Renigier – Biłozor¹, Artur Janowski²

¹ Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn, ul. Michała Oczapowskiego 2, 10-719 Olsztyn, Poland, e-mail: malgorzata.renigier@uwm.edu.pl, ORCID: 0000-0002-4630-75647
² Institute of Geodesy and Construction, Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn, ul. Michała Oczapowskiego 2, 10-719 Olsztyn, Poland, e-mail: artur.janowski@uwm.edu.pl, ORCID: 0000-0002-5535-408X

* Corresponding author

ARTICLE INFO

Keywords: human-machine similarity analysis, real estate market, Property Cognitive Information System (PCIS), artificial intelligence, synergistic data processing

JEL Classification: C45, C60, D91, R30

ABSTRACT

The issue of similarity in the real estate market is a widely recognized aspect of analysis, yet it remains underexplored in scientific research. This study aims to address this gap by introducing the concept of a Property Cognitive Information System (PCIS), which offers an innovative approach to analyzing similarity in the real estate market. The PCIS introduces non-classical and alternative solutions, departing from the conventional data analysis practices commonly employed in the real estate market. Moreover, the study delves into the integration of artificial intelligence (AI) in the PCIS. The paper highlights the value added by the PCIS, specifically discussing the validity of using automatic ML-based solutions to objectify the results of synergistic data processing in the real estate market. Furthermore, the article establishes a set of essential assumptions and recommendations that contribute to a well-defined and interpretable notion of similarity in the context of human-machine analyses. By exploring the intricacies of similarity in the real estate market through the innovative PCIS and AI-based solutions, this research seeks to broaden the understanding and applicability of data analysis techniques in this domain.

1. Introduction

Real estate market analysis provides guidance for the many decision-makers involved in real estate development. The goal of market analysis is to minimize risks and maximize opportunities for developers and investors by providing analysis that is as timely and accurate as possible (Bełej et al., 2016). One of the most difficult issues in property market analysis is the heterogeneity of this phenomenon, which varies depending on the behaviors and actions of many entities. Moreover, real estate markets are sensitive to social, demographic, political, and economic changes (Rącka, 2017) and they are in a state of permanent imbalance (Kucharska-Stasiak et al., 2012). According to Brzezicka and Wiśniewski (2013), the real estate market can be considered a special type of market, with its own rules and far from the definition given by mainstream economics. On the other hand (Kucharska, 2022) pointed out that the complexity of this category implies that, if valuation practitioners perceive it as poorly defined, one can hardly expect the value to accurately reflect the market.

It is somewhat paradoxical that, on one hand, we use very simple tools and very simple models to analyze the market, while the property market is one of the most complex economic and social phenomena and one of the most important assets in everyone’s life. The function of models is to simplify complex reality so that humans can better interpret it. Over-simplification of reality often leads to unrealistic representations of reality and, therefore, to wrong conclusions. Therefore, it is essential to seek solutions that closely approximate the analyzed market reality. This involves replacing highly simplified structures and commonly understood assumptions and parameters with solutions derived from the criterion of...
maximizing the potential for reflecting the analyzed reality.

The issue of similarity in the real estate market is the most common aspect of all analyses, yet one of the least conceptually considered phenomena in scientific research. This could be due to the complexity of the subject matter, including similarities, and the need to apply highly simplifying assumptions to maintain practical and applicable valuation practices. From the perspective of property valuation and appraisers, who expect fast and simple solutions, this is understandable. However, the lack of deeper understanding of the essence of the real estate market and the human factor as the main reason for its existence and functioning results in the perpetuation of stereotypes and the use of methodologies in pursuit of the assumption: simpler solutions are better, even if they produce erroneous results and force one to manipulate the data and bend the picture of reality in order to confirm and accept the result, e.g. a set valuation value.

Therefore, the inspiration for this research is a discussion within the scientific community regarding the foundations of the assumptions used in real estate market analysis, specifically in determining the similarity of phenomena and objects in the real estate market. The authors' focus is on exploring non-classical and alternative solutions compared to those commonly used in data analysis practice in the real estate market.

Artificial intelligence is among significant contemporary topics that seem to find application in almost all areas of human functioning. There are controversies and emotions surrounding the topic, as it appears to be slipping out of our control. In this context, we have two choices: either close our eyes and pretend it does not exist or understand the functioning details of AI-based solutions and try to build solutions that support human decision-making processes. Knowledge about their functioning and participation in building systems based on AI will also allow us to control the possible targets of its use.

Thus, the aim of the paper is to explore the potential (in terms of reflecting reality or human perception) of technologically cutting-edge solutions for assessing similarity in the real estate market. Therefore, the thesis was formulated: a cognitive system based on ML technology for comparing (defining/selecting) similar properties is an effective alternative to currently used methods based on the ceteris paribus assumption.

The paper presents the concept of the property cognitive information system (PCIS) developed by the authors, which is useful for analyzing similarity in the real estate market. The main assumption of this system is to preserve the original data structure describing properties and their synergy as a complete image of the described objects. The added value of the proposed PCIS system is a discussion on the validity of using automatic ML-based solutions in the context of objectifying the results of synergistic data processing. Additionally, the article provides a set of assumptions and recommendations concerning the definition and interpretation of similarity in the field of human-machine analyses.

2. Literature review
2.1. Definition of similarity

It is a natural human tendency to seek the causes behind the formation of phenomena and the relationships between influencing factors. An important aspect of such activities is the comparison and matching of similar objects. Defining similarity is closely related to the characteristics of a given object and the purpose of the analysis. The specifics of this classification in terms of attributes, size, or values belong to the domain of a specific field of research, and the priority should be to maximize the reflection of their real relationships. Subsequent generalizations are usually the result of achieving a compromise between the analytical goals and time-financial effectiveness.

The concept of “similarity” is encountered numerous times in our lives, and its definition is intuitively sensed by every individual. However, the most significant question is, what does it mean for an object to be “similar”? Most of us would answer: having few differences and sharing key characteristics. One thing we know for sure from this response is the vague and fuzzy nature of the assessment of “similarity.” In a general context, similarity means that two objects have certain common features that allow for comparison, indicating that they are similar in some respect. In establishing what similarity means, a crucial role is played by precisely defining its criteria, measurement methods, and research methodology (Makowska, 2016; Walesiak, 2016). The definition of similarity can be diverse and depends on the domain or context in which it is used. In biology, similarity can refer to genetic similarities between organisms or similar physical traits between species. In psychology, it relates to similarities in behavior, personality,
preferences, or emotional reactions between individuals. In technology, similarity may involve similarities in features, functions, or appearance between different objects.

As seen, similarity can be characterized and defined in various ways. Defining and subsequently evaluating it can be an exceptionally challenging task, especially requiring the determination of similarity criteria, which may vary depending on the research needs. It could be asserted, as stated by (Walesiak, 2016) that two objects are more similar when they differ less from each other. The procedure for examining similarity is most frequently carried out using clustering, classification, delimitation methods, employing different types of measurements and similarity metrics. According to Zyga (2011), similarity can be defined as a quantity that reflects the strength and number of relations between two objects and their characteristics, allowing for an objective and numerical measurement. Some proponents of the definition of similarity suggest that objects belonging to the same set should be as similar as possible (the principle of internal consistency), while those belonging to different sets should be as different as possible from each other (the principle of external isolation) (Makowska, 2016). Expanding on this approach, as proposed by Makowska (2016) it assumes certain categorical assumptions where classification should meet three basic clustering principles: completeness, disjointness, and nonemptiness. These principles respectively signify that all objects within a given set should be assigned to some group (completeness), one object cannot belong to more than one group (disjointness), and there cannot be a group without any elements (nonemptiness).

Another question is: is a “similar” object the same as a “homogeneous” one? The term “homogeneous” refers to entities forming a coherent whole composed of elements that are indistinguishable from each other (Glosbe, 2023). In everyday language, the terms “similar” and “homogeneous” are often used interchangeably, considering them as synonyms. However, they have distinct definitions, and their meanings can differ depending on the context. “Similar” refers to characteristics of elements or units that share some common aspects but may differ to some extent. Similar elements are more diverse than homogeneous ones but still possess common features that allow for comparison or contrasted (Nowiński & Kowalski, 2018). “Homogeneous” refers to characteristics of elements or units that are almost identical or very similar in terms of specific features or parameters (Ślężak & Zgrzywa, 2019). In general, “similar” means that two or more things have certain common features, enabling a comparison between them. “Homogeneous,” on the other hand, indicates objects that are practically identical, with little difference, often implying that they have the same specificity, characteristics, parameters, or features, making it difficult (or even impossible) to distinguish significant differences between them. Distinguishing between the above terms and their meanings in practice is quite apparent when dealing with unambiguous and highly precise objects. However, the situation becomes more complicated when the obtained data and available information are characterized by high generality, ambiguity, and imprecision, as is the case in the real estate market. In this context, theories and methods based on AI can be helpful. For instance, according to the theory of rough sets, a definition of indistinguishability for objects that are not precisely identical was introduced, taking into account the upper/lower approximately of the indiscernibility. This suggests that an object (consisting of features) is not identical to the one being compared, but they are similar to some extent. This approach allows for the inclusion of the object within the analyzed decisional rule based on a certain threshold of similarity (Chmielewska et al., 2022).

2.2. Concept of Similarity in the Real Estate Market

Assessing real estate properties based on their similarity is an exceptionally challenging topic. Not only are there no two identical properties, but their diversity within even a single segment is unusually high. In the Polish legal system, there is no precisely defined category of “similar property” in the context of real estate valuation. According to the basic legal act regulating issues related to real estate valuation in Poland, paragraph 16 of article 4 of the Real Estate Management Act (Ustawa o gospodarce, 1997) defines “similar property” as a property comparable to the property being appraised, in terms of location, legal status, purpose, manner of use, and other characteristics affecting its value. The term “similar property” was also used in the regulation on real estate valuation (Rozporządzenie w sprawie wyceny, 2004), where knowledge of transaction prices of properties similar to the property being appraised and influencing their price level is necessary. Apart from the statutory characteristics, there are other
distinguishing features of properties that need to be determined by a property appraiser. These features are differentiating and determine the similarity or dissimilarity of properties. Additionally, the interpretative note on the application of the comparative approach to real estate valuation (PKZW, 2009) indicates that similar properties are those that were subject to sale in the nearest period preceding the valuation date (but not longer than two years), taking into account the appropriate market of similar properties and market features significantly affecting price differences.

The lack of a concrete definition of similar properties and a procedure for selecting similar properties forces the person conducting the selection, such as comparable properties, to rely on subjective interpretation. One simplification in dealing with the complexity of this issue is the principle of ceteris paribus, which attempts to neutralize the fact that comparable objects are not identical by assuming that pairs of properties differ in terms of only one feature. The difficulty arising from such assumptions is evident when starting any procedure for selecting similar properties and defining related characteristics. The consequences of this fact are listed in various publications (e.g. Frukacz et al., 2011; Gaca & Sawilow, 2015, etc.) and result in a very high risk of "value falsification."

In property valuation, the comparative method is often used, which assumes that the more similar the properties are in terms of the analyzed characteristics to the property being assessed, the more accurate the valuation will be. It is worth mentioning another possible criterion of similarity, that a property can be considered in terms of an alternative or competing good for the potential buyer, which means two similar properties located in similar markets, although from the point of view of valuation principles, they should usually have the same location. In this context, it is worth mentioning the judgment of the Supreme Administrative Court (NSA) stating that "similarity, as understood in article 4 point 16 of the Real Estate Management Act, does not mean identity of parameters. It means, however, the commonality of essential market characteristics that have a fundamental impact on the property value. It is thus a bond based on similarity, not identity" (wyrok NSA, 2020). Moreover, according to NSA judgment from 2021, “the property appraiser determines the market value of the property and in accordance with article 154 paragraph 1 of the Real Estate Management Act, has the guaranteed freedom to choose the appropriate approach, method, and techniques for property valuation” (wyrok NSA, 2021), which also applies to the features of properties affecting their value and similar properties themselves.

Therefore, one of the most complex and difficult stages in any real estate market analysis is defining similar properties and their selection. This topic is the subject of consideration by many researchers, both from a methodological and technological perspective, but not often from a substantive perspective. According to McCluskey & Borst (2017), the high impact of uncertainties in the property market is related to information availability, market conditions, and specific inputs for the subject property. They claimed that the criteria for selecting comparables from the entire set of sales transactions must be defined to ensure that sufficient comparables, "similar" to the subject property, are extracted. This indicates that the most crucial stage in such analyses is to define the concept of similarity itself. Zyga (2011) defined similarity in terms of the similarity of comparative properties in relation to the object of assessment/valuation, as well as the similarity of comparative properties in relation to each other. He postulates the elimination of interpretative freedom in the definition of similar properties, which is harmful to the profession of property appraiser, and proposes changing the definition to: "similar property is a property with certain characteristics consistent with the characteristics of the property being valued; resembling the property in some respects, of the same type or kind, almost identical to it". According to Barańska (2010), the similarity of two properties can be defined as the degree to which the properties are similar to each other. Forsys (2016) on the other hand, argues that the criterion of similarity is imprecise: "the adopted attributes are in many cases influenced by the specific purpose of the acquired information. Due to the nature of the real estate market, there is no final and enduring set of attributes for individual property types that could be used to create future-proof price indices." Baranska (2010) claimed that considering the importance of selecting objects similar to the appraised one, as they form the basis for the entire further appraisal process in a comparative approach, it is worth examining the effect of the similarity assessment method on the final appraisal result, i.e., on the real estate value as well as on the accuracy of its determination. A detailed definition of similarity is crucial for distinguishing property
submarkets. According to Renigier-Biłozor et al. (2022), “homogeneous” transactions are not a strict set (precise/exact) but are a rough (approximate) set consistent with the adopted definition of similarity, which is the result of defining the comparable area unit, defining the features to be taken into account, elaborating the methodology, and selecting suitable methods, as well as verifying the obtained results.

To facilitate real estate market analyses regarding the selection of similar properties, it is essential to choose appropriate methods, procedures, and technologies according to the principle (e.g. d’Amato & Kauko, 2017) of selecting appropriate methods for the specifics of available information, rather than, as often observed in practice, adapting existing information to popular analytical methods. Zyga (2016) argues that the use of statistical methods in real estate valuation should not be questioned but is often applied without the necessary knowledge, leading to drawing conclusions based on false premises. Many researchers focus on details related to the potential application of hedonic modeling in market analysis, where determining the similarity of the analyzed objects, whether properties themselves or entire markets, is crucial (e.g. Belniak & Głuszak, 2011; Beracha, 2013; Bitner, 2007; Cellmer et al., 2020; Doszyn, 2020; Royuela & Duque, 2013; Sawilow, 2011). When studying the real estate market, it is essential to consider that the relationships on it can be purely random, and an additional difficulty is their dynamically changing nature. Rapid technological development allows to look optimistically at aspects related to solving complex and multidimensional problems, so many researchers focus on the application of alternative solutions or tools. For example, Gnat proposed the use of a modified measure of entropy for identifying the homogeneity of areas in terms of specific property features, or the XGBoost algorithm for mass valuation (Gnat, 2019). Dittmann (2013) examined the phenomenon of convergence and divergence in local markets in terms of their similarities within offers and transaction prices. Heckman (2008) examined the possibility of using non-linear causality in multivariable fractional polynomials (MFPs) to indicate homogenous markets. Belej and Kulesza (2014) analyzed the similarities of housing prices using multidimensional analysis and a damped harmonic oscillator model. Many authors have considered the application of technologies and methods based on artificial intelligence assumptions, such as genetic algorithms (GA) proposed by Del Giudice et al. (2017), fuzzy clustering by Hwang and Thill (Hwang & Thill, 2009), non-parametric smoothing and spline functions used by Pavlov (Pavlov, 2000), rough sets by d’Amato (Kauko & d’Amato, 2008), rough set theory combined with entropy theory by Renigier-Biłozor et al. (2022). Neural networks are a popular method used in real estate analyses among AI researchers: e.g., Ćetković et al., 2018; Choy et al., 2023; Demetriou, 2016; McCluskey et al., 2014; Wang et al., 2014; Zhou et al., 2018.

3. Material and methods

The research methodology for real estate decision making, augmented by AI, will be based on the following assumptions and technologies: human perception in decision-making, cognitive science, and machine learning - specifically soft computing technology.

3.1 Understanding human decision-making processes

The real estate market involves human decision-makers who are influenced by motives, emotions, and subsequent reactions. Hence, the selection of appropriate research methods should consider these human aspects and attempt to emulate human behaviors. However, decision-making in the real estate market is challenging due to the heterogeneity of analyzed objects (properties and the market) and the homogeneity assumption present in many commonly used methods for property market analysis. The primary challenge lies in the behavioral factor, which is often overlooked in existing methodological solutions. Humans have limited knowledge of their surroundings and can process only a finite amount of information. When making decisions, humans tend to simplify cognitive processes and prioritize solutions that ensure survival rather than maximizing profits, especially financial gains. Understanding human decision-making processes necessitates comprehending the mind’s capabilities and limitations and the specific context in which decisions are made. These assumptions have laid the foundation for a field of knowledge called cognitive science and a technology known as soft computing. Cognitive science explores the human mind, senses, and brain function, seeking to understand what the mind is and how it operates (Miller, 2003). On the other hand, soft computing is an approach to computer science that considers the human mind’s ability to reason and learn. Unlike hard computing, soft computing tolerates inaccuracies, uncertainties, partial truths, and
approximations (Ning et al., 2013). Both concepts are heavily reliant on understanding the phenomenon of human perception. Perception, according to the Oxford language dictionary, is a “complex cognitive process that leads a human to perceive various phenomena or processes resulting from specific stimuli acting on sensory organs (Oxford Learner’s Dictionaries, 2023). It involves the subjective reflection of objects, phenomena, and processes by humans. In the context of the real estate market and similarity assessment, vision is the leading sense used. Visual perception can be divided into two phases (Maruszewski, 2016; Merleau-Ponty, 2001). The first phase is the early stage of visual data processing, where the human eye registers light reflected from a specific visual scene through photoreceptors. Subsequently, when the photon beam enters the eye, neural processing occurs based on personal understanding, interpretation, and emotions. Humans seek categories that best fit the incoming stimuli, and this process is sequential. The more stimuli a person considers during the comparison process, the more time it takes. In the analysis of property transaction participants’ motives, research shows that their perception is limited to only a few attributes, especially if they are imposed to analyze phenomena using a ceteris paribus-based approach. According to analyses conducted at Stanford University, an average person cannot make a decision based on more than seven criteria or rank attributes according to their significance (Ries & Trout, 1986). It is worth noting that research (e.g., Carbon, 2011; Willis & Todorov, 2006) indicates that we tend to perceive what we are familiar with. If we lack prior knowledge about certain things, we might even overlook essential details in a pattern because we do not have a strong association with something meaningful. The processing of sensory input data through human semantic networks allows us to recognize familiar objects within milliseconds. One could argue that applying classical current market analysis methods complicates the analysis of real estate for humans. The foundation of classically used methods and approaches, even in the context of selecting and evaluating property attributes, is the controversial assumption that the analyzed reality is a world consisting of split elements. This directly justifies and reinforces the use of the ceteris paribus principle, based on simplification by extracting individual elements.

### 3.2 Unveiling cognitive systems

Cognitive systems were defined by Ogiela (2017) as systems that: “describe intelligent information systems designed to conduct in-depth data analyses based on the semantic content of the data. Semantic analyses are carried out using algorithms that describe this data based on processed expert information (such as knowledge bases) and the processes of machine (computer) perception and understanding of data, often performed using mathematical linguistics” (Ogiela, 2013, 2014).

Thus, cognitive systems are utilized to semantically interpret and analyze various data sets, provided these sets contain layers of semantic information. The semantic context should be understood as the entire information content contained in the analyzed datasets, as well as external information that directly refers to it. All information concerning the causes of phenomena, situations, the determinants of future behavior, and the course of phenomena can constitute this semantic context. Cognitive data analysis adopts the methods of description and interpretation from the syntactic approach. Data interpretation processes aimed at its cognitive analysis are more complex than data recognition processes. In this regard, the analysis of data sets cannot focus solely and exclusively on the information collected in these datasets but must primarily be based on information outside of them (Branquinho, 2001; Ogiela, 2014, 2017).

Cognitive information systems are used to analyze various datasets (Grossberg, 2013). Due to the different nature of the sets analyzed, cognitive systems have been classified into six basic types of understanding-based systems:

- Decision Support Systems,
- Image Analysis Systems,
- Management Support Systems,
- Person Authentication Systems,
- Signal Analysis Systems,
- Automatic Control Systems.

All the types of cognitive information systems mentioned above have broad applications, from management and economics, sociology and philosophy, to technical, military, and defense sciences, medicine, as well as natural and biological sciences (Anderson, 2018; Bañares et al., 2016; Ogiela, 2013; Ziolkowski et al., 2019).
3.3 Machine learning similarity in real estate analyses

The definition of similarity is also an essential component of the machine learning process. It often serves as its main objective. It typically indicates the closeness in the ML model space (in the domain of ML model) between two entities. When determining similarity, the attributes of these entities are taken into account. An entity here refers to a unique unit or object (in the real estate market, it could be a single property, transaction, owner, etc.). Once the similarity, also referred to as distance, is defined or computed, it can be utilized to construct subsets of entities. These subsets will represent groups of entities that are most similar to each other.

Similarity (or the requirements concerning its calculations) is explicitly defined by specifying the structure of the model and subjecting it to a multi-stage estimation based on multiple input data. In this context, one can look at machine learning algorithms aimed at modeling similarity in the real estate market - finding objects in this market that are similar to each other or assigning them to specific groups, i.e., classification. Similarity becomes the sole criterion and the primary goal of machine learning algorithms, making it their leading objective. With a properly trained machine learning model, it is possible to predict the classification of new entities (e.g., specific properties) and the required attribute values for objects intended to be appropriately classified. For example, assuming that each individual entity is described by values of certain number of attributes and is assigned to a specific group of entities allows, in situations of uncertainty regarding the values of individual entity attributes, to determine their values based on the range of characteristic values of the group to which they belong.

It is worth noting that in machine learning, there are numerous methods for defining and computing similarity between objects, which find or may find applications in the real estate market. These methods include, for example:

- **Geometric methods** – the use of these methods should be preceded by the removal of differentiation in the scales used to express individual real estate attributes through the use of normalization and then dimensionality reduction using one of the chosen methods e.g., PCA and Sparse PCA (Principal Component Analysis) (George, 2012; Halko, 2011), Randomized PCA, t-SNE (t-Distributed Stochastic Neighbor Embedding) (Belkina, 2019), UMAP (Uniform Manifold Approximation and Projection), LDA (Linear Discriminant Analysis) (Li, 2023), MDS (Multidimensional Scaling) (Little, 2022), Isomap, Factor Analysis, NMF (Non-negative Matrix Factorization) (Aonishi, 2022), ICA (Independent Component Analysis) (Howlander, 2022), etc., to facilitate interpretation and improve the efficiency of the estimation of similarity models obtained. Geometric similarity models result directly from geometric assumptions, where they are most often defined as Euclidean distances in the coordinate space determined by entity attributes (the positions of the entities are determined by their normalized attribute values). Other distances used to determine similarity include the Manhattan, Minkowski, Chebyshev methods, or in more sophisticated tasks, even the Hamming method. One of the best known and easiest models for visual interpretation of geometric similarity is SOM (Self Organized Maps) or Kohonen networks. In the case of the real estate market, they allow visualizing categories of properties with similar market value, surface area, functionality, etc.

- **Statistical methods** - in ML, they define the most classical similarity and are most commonly used. They use Pearson's correlations, Jaccard's coefficient, Mahalanobis distance. During the analysis of similarity of property test descriptions, Levenshtein's distance method is useful (referring to the number of operations transforming one text string into another). These methods use techniques derived from probability theory and statistics to model and measure similarity. In the context of real estate, these methods may include techniques such as regression, classification, clustering, PCA and are usually used for modeling and forecasting real estate prices, but can also be used to assess similarity in the real estate market.

- **Heuristic methods** - are a way to provide effective empirical solutions for complex multicriteria optimization problems, including the assessment of similarity in the real estate market. Their effectiveness compared to traditional solutions often comes down to finding a local rather than a global solution to the task, but close enough to the expected solution to accept it in relation to the costs (time, procedural, financial) of the similarity determination process. Their application to determine similarity between data (objects) is one of the main goals and can be implemented in the real estate market for example by heuristics:

  1. Geometric - properties are similar if they have
similar floor area, number of rooms, shape etc.

2. Type of property - properties are similar when they are of the same type - land, premises, etc.

3. Time - properties similar when they have similar construction time, transaction date, etc.

4. Spatial - similar properties will be located nearby, or in a location with a similar environment, etc.

5. Other: physical, legal, economic features, etc.

**Kernel methods** - are useful in situations of comparing objects that are difficult to represent as points in Euclidean space, i.e., when it is difficult to define deterministic and linear cause-and-effect relations (between variables) or when such relations do not occur at all. There are at least several groups of such methods for determining similarity:

- **a)** Gaussian kernels in the RBF (Radial Basis Function) approach - the kernel similarity is presented by the formula:

  \[ K(x, y) = \exp\left(-\gamma \|x - y\|^2\right) \]  

  where:

  \( x, y \) are real estate attributes, \( \gamma \) is a parameter that controls the "width" of the kernel, controlling how quickly similarity value decreases.

- **Polynomial kernels** - the similarity kernel is presented by the formula:

  \[ K(x, y) = (x \cdot y + c)^d \]  

  where: \( x, y \) is real estate attributes, \( c \) is a constant added to the scalar product, \( d \) is the degree of the polynomial.

- **b)** kernels built on the basis of a definable similarity function - in general, they return a higher value the more similar the compared features are. The most popular and classic ones include:

  - Cosine similarity function presented by the formula:

    \[ \text{cosine similarity}(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \]  

  - Pearson similarity function (Mana & Sasipraba, 2021) presented by formula:

    \[ \text{Pearson similarity}(x, y) = \frac{\sum_{i=1}^{n}(x_i-x)(y_i-y)}{\sqrt{\sum_{i=1}^{n}(x_i-x)^2 \sum_{i=1}^{n}(y_i-y)^2}} \]  

  where: \( n \) is the number of elements in vectors \( x \) and \( y \), \( x_i, y_i \) is respectively the i-th elements of vectors \( x \) and \( y \), \( \bar{x}, \bar{y} \) are the average values of vectors \( x \) and \( y \).

- **Hamming similarity function** (McKeown & Buchanan, 2023)

  \[ (x, y) = \frac{\text{number of positions where } x_i \neq y_i}{n} \]  

- **Jaccard similarity function** (Zhang, 2023).

\[ \text{Jaccard similarity}(A, B) = \frac{|A \cap B|}{|A \cup B|} \]  

where: \( |A \cap B| \) is the number of elements belonging simultaneously to set A and B, \( |A \cup B| \) is the number of elements in the set that contains elements of both set A and set B.

**Methods based on deep learning:**

- **Convolutional Neural Network (CNN)** (Lee, 2023; Ziolkowski et al., 2021): strictly used for image processing. Their various architectures and modifications enable the extraction of elements presented in the images. By having extracted image features (real estates), it becomes possible to compare them, defining similarity based on factors such as for example the building age. CNN models can be trained to detect characteristic features related to construction periods and utilize these features to classify buildings based on their age (construction time).

- **Autoencoder Networks** - used in unsupervised learning where the model learns to best reconstruct memorized (Mohan & Giridhar, 2022). Among tasks such as dimensionality reduction and anomaly detection in datasets, one of their uses is data augmentation (when there is a scarcity of data for model training).

- **Generative Adversarial Network (GAN)** (Nelson et al., 2021): deep learning models composed of two sub-models: the generator (creating fake entities) and the discriminator (distinguishing between the generated fake entities and real occurrences). The role of the discriminator is to determine the probability of the authenticity of data passed by the user. GANs are primarily used for generating generative sets, creating artificial data. In the context of the real estate market, they find applications in generating simulated room layouts, creating synthetic property images (matching given assumptions), and predicting property prices.

- **Memory-Based Neural Networks (MBNN)** (Qiu et al., 2020): particularly effective for sequential data records, where the sequence is ordered by time (time series). LSTM (Long Short Term Memory) or GRU (Gated Recurrent Unit)
networks, which employ recursion, are used for processing such data. The output from one cell of their structure is used as input for the next, allowing them to retain and process information from the past as far as the network depth. For instance, having access to the history of property transactions described by attributes enables a unified description of properties in the time domain - embedding property features as functions of time, thus facilitating their comparison regardless of the time of transaction data acquisition.

4. Results
Considering the assumptions presented above, the authors propose to develop a flexible and compliant cognitive information model for the real estate market, specifically designed for conducting similarity analyses. The cognitive computer models consist of the following modules/components: machine learning/artificial intelligence, cognitive science, algorithm application, and solution interpretation and verification. The proposed concept of model is depicted in Figure 1.

Fig. 1. The concept of a Property Cognitive Information System (PCIS) based on Machine Learning application. Source: own elaboration.

The motivation to develop the proposed system stems from the lack of data synergy consideration in the property market. Cognitive information systems are designed using structural reasoning techniques to comprehend the patterns defined within the system's knowledge bases. The concept of the presented PCIS facilitates the collection, processing, analysis, and inference from diverse combined data. These tasks are achieved by establishing sequences of derivative rules that generate these patterns.

Contemporary real estate data can be generalized into three groups in terms of type:
- image data,
- text data,
- numeric data.

Each of these groups requires a different approach for extracting relevant descriptive features and then inputting this information into a common concatenated feature vector.

1. **Image data** allows for rapid detection of visual details of properties. Their use enables the collection of comprehensive information regarding the number of characteristic objects represented, their belonging to specific subgroups (e.g., Group 1 - windows, Subgroup 1b - plastic windows, Subgroup 1b1 - rectangular plastic windows, etc.), and their mutual
relationships in direct proximity. During training, augmentation of the acquired images is performed to increase the flexibility of the constructed model. A sub-model of the cognitive system is a pre-trained (transfer learning) CNN model such as YOLO, ReNnet, Faster R-CNN, or others, used for detecting features in images. Transfer learning speeds up the process of achieving the desired convergence of the built model during the process of obtaining its optimal performance. The utilization of multiple property images taken separately or, even more so, in the form of video material may lead to redundancy in the number of detected individual object classes, which must be removed (redundancy removal).

2. **Numeric data**, which includes information such as area, number of floors, number of rooms, construction dates, etc., must be normalized to remove initial improper weighting caused by different scales in which they are expressed. In the final stage, a mapped feature vector is obtained from the numeric data.

3. **Descriptive data** consists of sequences of characters found in property descriptions, which may include expert descriptions, user ratings from real estate portals, auction sites, notarial deeds, etc. This data needs to be categorized by understanding sentence contexts, phrases, or words. Hence, it is divided into tokens to which meanings are assigned. Then, the tokens are arranged using embedding to create vectors that map the text being analyzed into a multidimensional space (tokens with similar meanings are located closely in this space). Trained Natural Language Processing (NLP) models such as BERT (Google) should also be employed for text handling. Descriptions will likely contain redundant content, which must be removed, resulting in a unified feature vector as the output.

All three data groups generate their feature vectors, which may overlap or even exclude each other in terms of semantic value. For example, of an image segmentation analysis might indicate two horizontal rows of windows as the result, numeric data may show that the building has three floors, and the text description of the property may include information that it is a two-story building with a usable attic (Janowski et al., 2021). Prioritization relations, such as considering similar information (choosing the appropriate information or averaging them), are another element of the model found in the concatenation and normalization layer. From the feature vector obtained by merging the text, image, and descriptive path vectors, a Feature Learning vector is obtained, and from it, in subsequent transformations, a fingerprint of the property data is derived. The fingerprint is a synthetic vector of features describing the property after discarding irrelevant or least important information passed through the previous layers of the model. The obtained features are not the real features of the property but synergistic features that encompass all the dependencies detected by the model within the information about a specific property.

The ultimate goal of the presented model is to determine the degree of similarity between two analyzed properties. Hence, the authors employed an unsupervised learning process (used in learning where there are no a priori assumptions defining the accuracy of the results). In each epoch of model tuning, data about property pairs is required following the scheme presented in Fig 2. After obtaining the fingerprints for these two properties, they are compared based on a selected similarity function. The final step is the interpretation of the obtained result, which can take the form of a continuous value (percentage range from 0 to 100%) or defined intervals in a discrete form (e.g., homogeneous, very similar, similar, weakly similar, dissimilar).

The concept of data flow presented in this way allows for a refined definition of similarity - the similarity of property objects based on cognitive information systems. Using machine learning in artificial intelligence, the similarity of property objects is measured by a function that determines how similar or indiscernible the analyzed objects are. This is achieved through similarity learning, using a pseudo-metric function \( S \) that follows the conditions defined by the authors: non-negativity, indiscernibility, symmetry, subadditivity (or the triangle inequality), relative comparisons, and ranking learning. These requirements are strictly relevant to a set of properties \( (R_1...R_n) \), as illustrated in Fig.3.

The empirical use of the system was presented using an example of a property pair described in Fig 4. The property characteristics were limited to a finite number of selected features to enable the presentation of the computational process.
Fig. 2. Data flow within the cognitive system for property similarity analysis based on machine learning. Source: own elaboration.
In line with the initial assumptions, it is necessary to analyze 3 types of data and reduce their informational content to a common feature vector. The data analysis process has been presented on reduced-size vectors to simplify its description.

Text processing: firstly, entire strings of text are tokenized, then transformed into vectors using a pre-trained language model like BERT, which could result in high-dimensional vectors. For simplicity, it is assumed that these vectors are n-dimensional. After dimensionality reduction using, e.g.: PCA, 10-dimensional vectors are obtained:

<table>
<thead>
<tr>
<th>GROUP OF DATA</th>
<th>DESCRIPTION</th>
<th>PROPERTY NO 1</th>
<th>PROPERTY NO 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE DATA</td>
<td>PICTURES</td>
<td>![Picture 1]</td>
<td>![Picture 2]</td>
</tr>
<tr>
<td>NUMERICAL DATA</td>
<td>NUMBER OF BEDROOMS (number)</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>DISTANCE TO THE CITY CENTER (mins)</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>PLOT AREA (m²)</td>
<td>2000</td>
<td>2500</td>
</tr>
<tr>
<td></td>
<td>AGE OF CONSTRUCTION (years)</td>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>OPINIONS, EMOTION RECOGNITION (%)</td>
<td>70</td>
<td>40</td>
</tr>
<tr>
<td>DESCRIPTIVE DATA</td>
<td>ENVIRONMENT</td>
<td>Safe neighborhood, fenced housing estate, full social infrastructure</td>
<td>Safe neighbourhood, access to kindergarten, schools, shops, and gas station</td>
</tr>
<tr>
<td></td>
<td>STANDARD</td>
<td>Renovated, updated infrastructure, high-quality finishing products</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>DESCRIPTION OF THE PLOT</td>
<td>Decorative garden with small architectural elements</td>
<td>Small architecture</td>
</tr>
<tr>
<td></td>
<td>TERMS OF THE TRANSACTION</td>
<td>Free market</td>
<td>Free market</td>
</tr>
<tr>
<td>POSITION COORDINATES</td>
<td>LATITUDE, LONGITUDE (degrees)</td>
<td>52.1101</td>
<td>21.9876</td>
</tr>
<tr>
<td></td>
<td>COMPARISON RESULT</td>
<td>74% - SIMILARITY SCORE</td>
<td></td>
</tr>
</tbody>
</table>
for instance:
- property 1 text vector: \([0.2, 0.3, 0.2, 0.1, 0.0, 0.2, 0.2, 0.3, 0.2, 0.1]\),
- property 2 text vector: \([0.2, 0.3, 0.2, 0.1, 0.0, 0.2, 0.3, 0.2, 0.1, 0.0]\).

The numerical processing: data is normalized to form vectors. In this example, these are simple 3D vectors representing the normalized values of area, number of rooms, and year of construction:

for instance:
- property 1 numerical vector: \([0.5, 0.6, 0.0]\),
- property 2 numerical vector: \([0.6, 0.4, 0.6]\).

Image Processing: utilizing a CNN trained to identify key features, detected objects in images can be processed and represented as vectors. Detected objects for property 1 include a stone façade, 7 windows, roof, 2 chimneys, 4 trees, hedge, sidewalk, stairs, 2 lamps. Property 2 includes a concrete façade, 8 windows, roof, 1 chimney, 1 tree, 4 shrubs, sidewalk, stairs, 1 lamp. An m-dimensional vector with numbers of objects of classes is obtained, which is simplified to a smaller vector, such as a 3-dimensional vector.

for instance:
- property 1 image vector: \([0.4, 0.5, 0.3]\),
- property 2 image vector: \([0.4, 0.6, 0.4]\).

All obtained vectors for each property are concatenated, resulting in:
- property 1 concatenated vector: \([0.2, 0.3, 0.2, 0.1, 0.0, 0.2, 0.2, 0.3, 0.2, 0.1, 0.5, 0.6, 0.0, 0.4, 0.5, 0.3]\),
- property 2 concatenated vector: \([0.2, 0.3, 0.2, 0.1, 0.0, 0.2, 0.3, 0.2, 0.1, 0.0, 0.6, 0.4, 0.4, 0.6, 0.4, 0.6, 0.4]\).

Next, feature learning is performed using an autoencoder or similar unsupervised learning model to extract the most important features from the 16-dimensional vectors. The autoencoder comprises two components: an encoder and a decoder. The encoder’s task is to transform, to map the input vector \(v\) into a lower-dimensional vector \(u\). The decoder’s task is to reconstruct the vector \(v\) based on the vector \(u\). The network is trained by examining the error of this reconstruction. This yields, for example, 4-dimensional “fingerprint” vectors:

for instance:
- property 1 fingerprint vector: \([0.5, 0.7, 0.1, 0.2]\),
- property 2 fingerprint vector: \([0.7, 0.8, 0.5, 0.7]\).

Finally, the cosine similarity between these two “fingerprint” vectors is calculated. A cosine similarity of 1 indicates that the two fingerprints are identical, while a value close to 0 suggests that the properties are dissimilar. The result for this example is 0.74, meaning that these two properties share similarities in approximately 74% of their respective fingerprint representations according to the used model. In other words, the model suggests that the two properties are relatively similar according to the features and attributes used for comparison.

5. Discussion

In computerized processes of data analysis and decision-making, it is beneficial to use perceptual models characteristic of the human mind. The essence of this solution lies in the utilization of algorithms that can be transferred to systematic solutions in automatic data analysis. The more complete the acquisition of information on how specific cognitive tasks are performed, the more accurately the computer system imitates the human mind. It is challenging to determine unequivocally how closely they converge with the functioning of the human mind, but these two analysis processes originating from different worlds are becoming more similar, thus becoming more effective and efficient in reflecting reality.

The application of technologically pioneering AI solutions in the real estate market can provide a more comprehensive understanding of its functioning, as they attempt to emulate human reactions and behaviors that shape this market. The methodology based on ML (AI) allows for a compromise between the reality shaped by an “imperfect” and emotionally ambiguous human and the science represented by “perfect” and precise mathematical relationships. It enables the creation of a model conglomerate of features, phenomena, and processes similar to human perceptual abilities. The motivation behind creating the cognitive information system for real estate market analysis is to include all necessary sources of information in one model. The sources of information related to property market analysis are basically unlimited, especially from the assumption’s point of view or at the initial stage. Unlimited data refers to the variability and non-constancy in terms of the number and types of information sources, data accuracy, scope, trustworthiness, timeliness, various formats of data representation and presentation, etc.

The answer to this kind of problem in analysis is to find model solutions that are highly flexible in terms of the variability of the spectrum and characteristics of
input data, as well as having a high probability of reflecting the analyzed reality in terms of:

- **optimization** - concerns the analysis and assessment of all possible solutions, even those that are not obvious/unexpected, but highly probable,
- **decision-taking speed** - relates to the minimal time it takes to execute the decision-making process,
- **economy** - taking into account: human resources, hardware, software resources, and addressing the complexity and the level of difficulty of cognitive operations performed as part of the decision-making process.

The most important role/function of the proposed system is to improve decision-making in the real estate market, based on the following decision-making strategies (criteria) of the algorithm/system (distinguishing it from classical solutions not based on AI):

- **criterium of cognitive representation** - making decisions based on a cognitive representation of the analyzed decision-making task. According to the flexibility of the human mind, the generated cognitive representation of the decision-making problem may change partially or completely as a result of the system's fine-tuning, allowing for broader context comprehension of the considered decision-making process/phenomenon.
- **prioritization criterion** - selecting the most important conditions in a sensitivity and specificity mode according to assumed conditions, e.g., accuracy, training time, inference time, robustness, etc.
- **comparative optimization criterion** - comparing different approaches and solutions in terms of efficiency, performance, accuracy, speed, costs, and other relevant criteria that best meet the assumed objectives.

### 6. Conclusions

There is a clear gap between the big data provided by the external world and human strictly limited ability to process it. This gap widens even further when we consider that we not only need to process the data but ultimately give clear meaning to the essence of a given situation. The goal is to make a single decision based on a clear interpretation of that situation to take appropriate action. To achieve a clear interpretation of the situation, we need a mental model of the external world that is very clear and devoid of ambiguity and uncertainty. In a general situation, the model is a kind of caricature of the physical reality (Carbon, 2014). In summary, it can be stated that systems based on cognitive systems utilizing ML technologies can provide significant support in enhancing the efficiency of analyses in the real estate market. However, a prerequisite for this is knowledge about their functioning and the development of systems directly dedicated to the given problem. They can be particularly useful in cases where there is high complexity in the analyzed reality, which often leads to ambiguity in defining specific analytical issues, such as determining the similarity of objects.

The presented solution, which is based on the emulation of cognitive processes used in data analysis, has the main task of improving (legitimizing) decision-making. In the case of cognitive systems, decision-making processes are understood as a type of processes that facilitate making optimal decisions based on a thorough analysis of the problem, i.e., "Data Understanding."

The greatest value in cognitive systems based on ML lies in understanding the data iteratively, with the possibility of expanding knowledge and gaining new insights. The specifics of the model are characterized by flexibility, universality of form and type of data use, enabling inferences that consider a complete conglomeration of object (property) characteristics, considering it as a defined function of a certain feature (location, property description, time and transaction conditions, opinions/emotions, etc.). The model's receptiveness to diverse data and information sources, for example, enables more efficient (what can be understood that the obtained result of comparing objects is based on a model that does not need to be modified according to different types of data) an expert valuer work in demanding tasks such as arbitration tasks (Sokół & Sobolewska-Mikulska, 2023). The proposed system (PCIS) has an original added value in real estate analyses, such as structural reasoning rules definition based on synergic data processing. Analyzing the presented system, one should take into account that it has both a number of advantages and limitations. Below, in Figure 5, the most important ones are shown.
### ADVANTAGES:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automation</td>
<td>Automation: the system automatically finds similar properties, significantly speeding up the process and reducing the required manual effort.</td>
</tr>
<tr>
<td>Scalability</td>
<td>Most machine learning algorithms are scalable – an increasing amount of data to process does not result in a significant increase in computational power demand.</td>
</tr>
<tr>
<td>Unsupervised learning</td>
<td>Suitable for detecting relationships that are difficult to define a priori.</td>
</tr>
<tr>
<td>Incomplete data</td>
<td>It can handle incomplete or inconsistently recorded data records.</td>
</tr>
<tr>
<td>Multiple tested solutions</td>
<td>Performing a multi-threaded and multi-epoch sequence of searches for an optimal solution.</td>
</tr>
<tr>
<td>Accuracy control</td>
<td>Strict accuracy control: a properly trained machine learning model can be precise, leading to more accurate comparisons.</td>
</tr>
<tr>
<td>Flexibility of the system’s structure</td>
<td>The system automatically adjusts the parameter values of the model to the data or the purpose of the analysis.</td>
</tr>
<tr>
<td>Multiple models</td>
<td>The system analyzes multiple models and their parameters in the process of searching for the optimal solution.</td>
</tr>
<tr>
<td>Transfer learning</td>
<td>Retraining models. Data evaluated by the model in subsequent epochs of the system’s operation can also feed it, providing an opportunity for continuous updating – adaptation to new requirements, new data, and changes in temporal trends.</td>
</tr>
</tbody>
</table>

### LIMITATIONS:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model complexity</td>
<td>Building, training, and fine-tuning machine learning models can be complicated and require specialized knowledge.</td>
</tr>
<tr>
<td>Data requirements</td>
<td>The need for a large amount of semantically diversified input data to construct a flexible and appropriately precise model. Machine learning systems typically require a large amount of data for training, which can be challenging, especially when data is difficult to acquire.</td>
</tr>
<tr>
<td>Risk of overfitting</td>
<td>There is a risk that the model will be overly fitted to the training data. The consequence of this is a low ability to generalize, meaning weak prediction of outcomes, high sensitivity to noise and nuances present in the training data, lack of model flexibility, and optimization capabilities.</td>
</tr>
<tr>
<td>Difficulty of interpretation</td>
<td>Machine learning models, especially more complex ones, can be challenging to interpret. Explaining the source of a decision made by the model can be difficult.</td>
</tr>
<tr>
<td>Requirements</td>
<td>Training a model and ensuring its effective functioning may require significant computational resources and time.</td>
</tr>
</tbody>
</table>

![Fig 5. Advantages and Limitations of the PCIS Concept. Source: own elaboration.](image-url)

An additional barrier of such solutions is the current awareness and perception of AI-based solutions. Systems based on AI technology raise many controversies, and it is hard to find a domain of life where they are not utilized. AI is often attributed not only human intelligence but even superintelligence, a so-called “super-entity,” surpassing the human mind. However, it turns out that this is not yet a fact, and it is uncertain if it will ever happen. Current systems like Bing AI (utilizing Prometheus technology), Google Bart based on LaMDA technology, or perhaps the most popular ChatGPT based on the GPT algorithm are indeed very helpful solutions, but they are still far from human intelligence and drawing correct conclusions. The problem arises when they face a series of logical tasks, especially when not provided with data describing the specific problem. It is also worth mentioning their rapid obsolescence of knowledge and their usage of users as cyber-biomass feeding them.

**Acknowledgements**
The publication was written as a result of the authors; internship in the School of Government at the University of North Carolina, USA, co-financed by the European Union under the European Social Fund (Operational Program Knowledge Education Development), carried out in the project Development Program at the University of Warmia and Mazury in Olsztyn (POWR.03.05.00-00-Z310/17).

References


Gaca, R., & Sawiłow, E. (2015). Wykorzystanie metody korygowania stanów cech nieruchomości do staniu „ceteris paribus” dla określenia zmiany poziomu cen nieruchomości wskutek upływu czasu (Using the method of correcting the states of real estate features to the state of “ceteris paribus” to determine the change in the level of real estate prices due to the passage of time). *Rzeczoznawca Mątkażowy, 87*.


Rozporządzenie Rady Ministrów z dnia 21 września 2004 r. w sprawie wyceny nieruchomości i sporządzania operatu szacunkowego (The Regulation of the Council of Ministers of September 21, 2004, on the valuation of real estate and the preparation of the appraisal report).


Ustawa z dnia 21 sierpnia 1997 r. o gospodarce nieruchomościami (Act of real estate management) (Dz. U. z 2023 r. poz. 344, 1113, 1463.).


Wyrok NSA z 20 maja 2021, I OSK 4117/18, CBOSA. (2021). (Supreme Administrative Court judgment)

Wyrok NSA z 4 sierpnia 2020 r., I OSK 3563/18. (2020). (Supreme Administrative Court judgment)


