land

MDPI

Article

System Design for Detecting Real Estate Speculation Abusing Inside Information: For the Fair Reallocation of Land

Yeon-Jin Sim [1,†], Jeongmin Kim [1,†], Jaehyeon Choi [2] and Jun-Ho Huh [1,2,*]

[1] Department of Data Science, (National) Korea Maritime and Ocean University, Busan 49112, Korea; ajrrns@g.kmou.ac.kr (Y.-J.S.); kjmin@g.kmou.ac.kr (J.K.)
[2] Department of Data Informatics, (National) Korea Maritime and Ocean University, Busan 49112, Korea; jener05458@g.kmou.ac.kr
* Correspondence: 72networks@kmou.ac.kr
† These authors contributed equally to this work.

Abstract: In March 2021, a case of speculation that abused private internal information came to light, which involved a group of public officials from the Korea Land and Housing Corporation (LH), and has since been labeled the 'LH Scandal'. In this scandal, land was misappropriated as a means of creating fraudulent values, instead of returning it to marginalized people in real need of residential space. As a result of this, preventive measures for similar cases have become warranted. Consequently, related laws have been passed, but this is only expected to show its effect as a follow-up response, therefore requiring a preemptive response plan. In this paper, we will propose a conceptual framework that can detect speculation that abuses private internal information, enabling a preemptive response, utilizing outlier detection and Latent Dirichlet Allocation (LDA) methods. The system is designed to create a database (DB) with private inside real estate information, which is linked to another DB with a list of outlier-detected areas that can potentially indicate speculation, and then the system confirms any speculation by comparing the two DBs accordingly. Once a speculation case is confirmed, the system automatically reports the case to the investigative agency. By using this system, we expect to detect hidden speculation cases already committed, as well as speculation cases in real-time. Ultimately, we hope to protect the original purpose of redevelopment and the construction of new towns (housing/retail mixed-use zones), redistributing available land on behalf of marginalized people, who are lacking in residential space, by raising the utility of land.

Keywords: speculation detection; outlier detection; Latent Dirichlet Allocation; private internal information; real estate speculation; land solution

Citation: Sim, Y.-J.; Kim, J.; Choi, J.; Huh, J.-H. System Design for Detecting Real Estate Speculation Abusing Inside Information: For the Fair Reallocation of Land. *Land* **2022**, *11*, 565. https://doi.org/10.3390/land11040565

Academic Editor: Piyush Tiwari

Received: 28 February 2022
Accepted: 7 April 2022
Published: 11 April 2022

1. Introduction

In many countries, real estate speculation does not have a negative meaning. In those countries, real estate speculation denotes purchasing real estate and selling it when its price escalates through redevelopment or city development projects [1]. However, in the Republic of Korea (ROK), real estate speculation has a negative meaning. This is because speculation in the ROK is mainly carried out with information leaked from an insider who oversees redevelopment projects [2]. In this context, speculation in the ROK has a similar meaning to illegal insider trading abroad, rather than investment.

Speculation in the ROK is mainly carried out by purchasing redevelopment-planned real estate using the name of an insider and their family, then using various tricks to gain turnover profit and compensation [3]. This kind of speculation results, not only in the opposite effect of the main purpose of redevelopment, resolving social polarization, but also highlights the matter of public service ethics and the professionalism of public officials or related workers, who should pursue the public interest rather than look for their own private interest.

In March 2021, a case of speculation that abused private internal information involved a group of public officials from the Korea Land and Housing Corporation (LH), and has been labeled the 'LH Scandal' [4]. In this scandal, the land was misappropriated as a means of creating fraudulent values, instead of returning it to marginalized people in real need of residential space. Due to this scandal, LH has been criticized nationally because the redevelopment, which was intended as a project for the reallocation of land to solve social polarization, actually resulted in the opposite because the related workers pursued their own private interests. Therefore, voices urging strong punishment and the building of a system for detecting real estate speculation have been raised.

Despite this public opinion spreading across the country, concern that appropriate punishments would be impossible due to a lack of legal basis has also been raised. In fact, one year after the incident, out of the tens of thousands of people involved in the LH scandal, there has only been one person convicted, which resulted in a very low sentence [5,6]. This is not because the seriousness of this matter is small, but because a legal basis to punish real estate speculation that abuses private internal information has not been sufficiently established. With the LH incident as an example, public opinion has risen concerning strengthening real estate speculation penalties in the ROK. As a result, five laws were promoted, and three of them have been proposed and enacted.

As the laws were enacted, it became possible to monitor the property details of employees and related people of public institutions related to redevelopment. Additionally, unlike previous laws that could only punish the employees directly in charge of the redevelopment project, it is now possible to hold accountable everyone suspected of involvement in real estate speculation using undisclosed internal information. The level of legal punishment, which was previously insufficient, has also been further strengthened [7].

The ROK government decided to widen the use of an existing public official society audit system, the Public Ethics Total Information System (PETI), to monitor changes in real estate-related workers' property. However, the existing system has its limits because it is expected to involve simple monitoring, which is considered as not being sufficient to detect real estate speculation. Due to the nature of redevelopment projects, which take place over a long period of time, the purchase of land for speculation and the use of various tricks can be carried out secretly long before the redevelopment agenda becomes visible. Therefore, it is difficult to determine the true purpose of a land purchase, which plays an important part in determining whether it is speculation or not, if the person carrying out the monitoring is not simultaneously taking part in the redevelopment project. Therefore, it is difficult to expect that monitoring would work when applied to the case of speculation detection, unlike the monitoring method that showed its effect when it was being used to audit the public official society of the ROK.

As we can see in the case of the LH scandal, the way of dealing with the speculation after it occurred, and not detecting it beforehand, cannot eradicate the speculation matter completely. In this paper, we suggest a conceptual framework that can detect real estate speculation before it happens. Through this study, we propose a supportive system to enable fair reallocation of land, which is the true purpose of redevelopment.

This study deals with real estate speculation in the ROK, which is not a general phenomenon abroad. However, even in countries other than the ROK, similar cases can happen, where those involved in redevelopment or in a city development project do not act professionally and do not take the public interest into consideration. Current study on insider trading detection is only focused on the stock market, and real estate speculation is not considered due to the distinctiveness of the situation. Thus, we expect this proposed framework to provide motivation in terms of serious real estate speculation abroad, and related research will then proceed.

The structure of this paper is as follows. Section 2 describes the related studies, such as outlier detection and LDA, which is used in the module that makes up the proposed framework. Section 3 compares speculation in the ROK with insider trading and describes elements of speculation and techniques in the ROK. Section 4 describes the framework we

propose. The Discussion then presents possible constraints and limitations in applying our proposed framework to actual situations, and the technical, legal, and institutional efforts to overcome them. Lastly, the Conclusion summarizes this paper and explains the expected impacts.

2. Related Studies

2.1. Public Official Audit System in ROK

In the ROK, the first public official ethics law was enacted in 1981. Since then, the law has been amended 10 times, and through this, a legal basis for building the audit system for the public official society and enabling proper punishment has been partly established [8]. As a result, in the ROK, a self-audit system to detect signs of corruption and administrative errors, called 'Cheongbaek-e', has been developed and used since 2015 [9]. The Cheongbaek-e system secures local administrative information system data such as finance, architecture, personnel, welfare, etc., and credit card company approval data, and gathers all those data to build a DB known as the 'preventive administrative scenario'. With the implementation of Cheongbaek-e, more than 100,000 corruption cases have been detected as of 2016, and millions of dollars in tax omissions and hundreds of thousands of dollars in corruption have been detected and recovered.

Accordingly, the Ministry of Government Administration and Home Affairs of the ROK confirmed the effectiveness of the Cheongbaek-e system and decided to proceed with the expansion of the system by widening the audit area through expanding the linkage between related institutions and financial institutions and using a scenario pattern with the GIS to deal with the new methods of corruption.

However, it has been judged that detecting real estate speculation through this system is not possible due to the specificity of real estate speculation, considering that the system is specialized in detecting unnatural flows of funds. For this reason, no attempt has been made to detect real estate speculation through this system. Instead, attempts have been made to utilize PETI, another existing audit system, to detect real estate speculation.

The PETI is a system that allows the property information of a specific public official to be registered and reported, and then reviewed by the competent public officials ethics committee [10]. According to the amended law after the LH scandal, all public officials related to the real estate project, including LH employees, are designated as subjects of property registration by PETI, in order to detect the speculation [11]. To do this, the PETI system has also been improved to ensure that it can monitor when a related worker purchases new real estate, so the process of the formation of real estate is clearly seen.

However, because sanctions related to real estate are only imposed when inside information is abused, a case where the related worker acquires the real estate when the redevelopment project is verbally discussed, with the characteristics of a long-term redevelopment project, cannot be detected. In addition, in the process of PETI's audit of real estate acquisition, there can be cases in which the status of related redevelopment projects may not be clear, therefore speculation cannot be properly detected.

2.2. Sepculation and Insider Trading

As mentioned earlier, in the ROK, the term 'speculation' has a similar meaning to 'illegal insider trading' in other countries. This section deals with research related to speculation and insider trading in countries other than the ROK.

The U.S. Securities and Exchange Commission (SEC), which conducts insider trading investigations by interviewing witnesses, examining trading records and data, and subpoenaing phone records, has defined insider trading as follows: (1) Corporate insiders who have traded the company's securities after learning of significant, confidential developments; (2) insiders' friends and family, as well as other recipients of tips who have traded securities after receiving such information; (3) employees of service firms, such as law firms, banks, and brokerages, and printing companies, who have come across nonpublic

information on companies and traded on it; (4) government employees who have obtained inside information through their jobs.

Moreover, the SEC defines illegal insider trading as 'buying or selling securities in breach of fiduciary duty or other relationship of trust and confidence being in possession of material, nonpublic information about those securities', while arguing that the prevalence of illegal insider trading can undermine public trust in the capital markets and their functioning [12].

Jennifer Moore defined insider trading as trading based on non-public, private information and said it is known through illegal or unethical means [13]. John, A. Ryan looked at speculation from three perspectives: corporate, social, and moral, in order to answer whether speculation is wrong. The author says speculation is wrong for all three perspectives [14]. Patricia H. Werhane has stated that insider trading is often a great risk and may not be beneficial if the information is not complete or accurate. Additionally, engaging in insider trading is unfair because it uses privileged information [15].

Bainbridge, S. M states that, in terms of regulation of insider trading, the ambiguous and unclear concept of insider trading does not provide justification for regulation [16]. Therefore, before regulating insider trading, the conceptual and legal definition of insider trading should take precedence. Zhenxi Chen et al., have investigated and evaluated the effectiveness of government policies on speculation, divided into housing prices, relaxed policies, and pressure policies in China [17]. Nuno Fernandes et al., have investigated the impact of the application of insider trading laws and concluded that the implementation of insider trading laws could have different effects on stock price information across the world [18]. Kwabi, F. O. et al., said that if the insider trading law is supplemented and implemented, it will have a positive effect on the market [19].

Xiadong Du et al., revealed that speculation is one of the important factors of crude oil and agricultural price volatility, using the Bayesian Markov chain Monte Carlo method [20], while Atuf Mian and Sufi Amir conducted research on credit supply and housing speculation. As a result, it was shown that a speculator's housing purchases through new credit transactions do contribute to rising housing prices and increasing construction activities in nearby areas [21]. Azmi Shabestari et al., analyzed abnormal profits from insider trading for each period by dividing the transaction period into white windows, black windows, and blackest, based on the asymmetry of information [22]. Stephen Malpezzi et al., proposed and simulated their model to determine whether speculation is the cause or the result of a real estate price cycle. In this regard, it has been shown that a simple model simulating price changes for supply delays due to speculation is sufficient to generate a real estate price cycle [23].

Many studies have also been conducted to detect insider trading by analyzing the effects of insider trading and the patterns that occur. Esther B. Del Brio et al., tried to detect insider trading using a portfolio test that pinpoints abnormal returns characterized as insiders earning abnormal profits using privileged information [24]. Dallin M. Alldredge et al., focused on whether public information for a particular customer can predict insider trading activities using the difference between actual and expected returns, and they showed that some insider trading transactions generate revenues due to public information [25]. Steven Huddart et al., used the windows method to prove that insider trading using privileged information does affect stock prices and trading timing using $FREQ_p$, a frequency-based indicator, and $VALUE_p$, a price-based indicator. As a result, insider traders have shown a tendency to actively benefit when the risk seems low [26]. Steve Donoho conducted a study to detect insider trading early on, using options, stock trading, and news data. In addition, options and news data were synthesized, and data mining techniques were applied to predict future news [27].

2.3. Outlier Detection

2.3.1. Outlier Detection in Real Estate

There are also many studies that apply outlier detecting methods to the real estate market. Most studies on outlier detection in the real estate market have been conducted that focus on transaction amounts, namely housing prices, to detect outliers.

Chao Wang et al., introduced a normalized upper and lower interval regression model for multiple independent and single interval dependent variables and proposed dual and relaxation approaches to detect the outliers in each model. Using these models, they conducted a study to remove the outliers by applying them to housing prices [28]. Vilius Kontrimas et al., applied various outlier detection methods, such as closest distance to the center (CDC), Cook's distance, and the Kernel method, and compared their performance to detect real estate transactions with large differences in tax evasion. The authors showed that multilayer perceptrons perform well among outlier detection methods [29]. Pierluigi Morano et al., proposed an outlier detection method that uses the least median of squared residuals (LMS) method instead of the least square method, which is heavily affected if an outlier exists. Using this method, they applied it to the history of housing transactions in the Italian city of Bari to remove the outliers [30]. Anna Baránsska et al., studied statistical methods for detecting outliers in determining the value of real estate. The study proposed a method to detect outliers using Cook's distance and Pope's method using data on land real estate containing information on local planning, real estate purpose, real estate type, etc. [31].

In addition to the study of transaction amounts, an outlier detection study related to the online real estate market has been conducted. Rıza Özçelik et al., used five outlier detection algorithms (neighborhood, distance to end value, distance to average, Tukey's test, Thompson's modified tau test) to detect false advertising with category errors or outliers on real estate sales sites, and then implemented a voting method [32]. Arcchaporn Choukuljaratsiri et al., explained the need for automated alarm systems to detect outliers using user log files in the online real estate market and suggested that the number of seasonal user views can be identified week-to-week using the SARIMA model [33].

2.3.2. Outlier Detection and LOF Algorithm

In 1980, D. M. Hawkins defined the term outlier as an observation that deviates so much from the other observations so as to arouse suspicions that it was generated by a different mechanism [34]. Additionally, in statistics, Grubbs defined an outlier as a data point that appears to deviate markedly from other members of the sample in which it occurs [35].

Later, in 1987, Dorothy. E. Denning proposed the first version of outlier detection in the process of designing the intrusion detection system (IDS), naming it the 'Anomaly-Record Rule', which pinpoints outliers by comparing the profile patterns and event patterns of users with the previous patterns [36].

According to the type of dataset and purpose of data processing, a significant amount of detection methods, such as distance-based detection and density-based detection, have appeared. Edwin M. Knorr et al., proposed distance-based detection algorithms that deal with k-dimensional datasets and are advanced versions of existing methods dealing with two-dimensional datasets. Various algorithms have been proposed according to the dimension of datasets, such as cell-based algorithms, the nested-loop algorithm, and algorithm NL, and their effectiveness has been demonstrated as they add the results of anomaly detection through those algorithms using the actual data of NHL players' statistics, video surveillance systems, and workers' compensation claims [37].

One example of density-based detection is the local outlier factor (LOF) algorithm proposed by Markus M. Breunig et al. The LOF algorithm deviates from the global detection algorithms in that it uses the local reachability density of each observation to make an LOF value. An observation with a lower LRD takes a higher LOF by the definition of Markus M. Breunig et al., inferring that the observation is an outlier [38]. Aleksandar Lazarevic et al.,

presented a method that combines the LOF results in high dimensions to obtain diversified observations, thereby improving the quality of detection [39].

In 2008, Fei Tony Liu et al., suggested the Isolation Forest, another density-based detection method, which detects anomalies based on a combination of a decision tree and ensemble method [40]. Ke Zhang et al., proposed LDOF, which uses the relative location of an object to its neighbors to determine the degree to which the object deviates from its neighborhood [41]. Sahand Hariri et al., solved the existing problem of an Isolation Forest by Extended Isolation Forest (EIF), which uses a heat map [42].

2.4. Latent Dirichlet Allocation (LDA)

LDA is a probabilistic model of text corpus proposed by David M. Blei et al., in 2003. LDA stems from the idea that documents are represented by a mixture of latent topics, each of which is represented by a distribution over words [43]. LDA has been studied in a variety of fields, including topic models, such as information retrieval and documentation. Meanwhile, models that complement LDA shortcomings or perform better in certain areas are also being studied.

Daniel Maier et al., selected (a) pre-processing, (b) selecting model parameters, (c) evaluating the model's reliability, and (d) validating the result topics for LDA applications, and have presented a users' guide for applying the LDA topic model [44].

Based on the difficulty of controlling the number of topics in topic models, Juan Cao et al., studied the essential associations between topic distances and LDA as well as best topic structure, and then proposed a method of adaptively selecting the LDA model according to density. Experiments have demonstrated that the LDA model performs best when the average cosine distance of the subject reaches its minimum, thereby achieving the highest-level performance in LDA without manually adjusting the number of subjects [45].

Michal Rosen-Zvi et al., expanded LDA to introduce 'the author-topic model', a document generation model that includes author information. Each author is linked to a polynomial distribution for a topic, and each topic is linked to a polynomial distribution for the related word. Documents with authors are modeled as distributions for topics with mixed distributions related to the authors [46]. Rachit Arora et al., proposed an extraction-based multidocument summarization algorithm that uses a weight mechanism to select sentences in documents and combine them into summaries. The proposed method is differentiated from the approaches of existing studies in that it uses a mixed model to find topics and pick out sentences without considering the details of grammar and the structure of documents [47].

David Andrzejewski et al., proposed a mechanism for partial supervision called topic-in-set knowledge for latent topic modeling. This can help recover items that are more relevant to modeling the user aims [48]. Daniel Ramage et al., proposed label LDA, a topic model that limits LDA by defining a one-to-one correspondence between LDA's latent topics and user tags. Accordingly, the labeled LDA may directly learn the word tag correspondence relationships [49].

Ralf Krestel et al., proposed an LDA-based approach that recommends resource's tags to improve search results. Through this, the search function was improved by recommending tags belonging to the topic to a new resource [50]. Samaneh Mohaddam et al., discussed the LDA-based models used for opinion mining in customer reviews and presented guidelines for LDA models in terms of opinion mining and directions for future studies [51]. Xing Wei et al., conducted a study on how to efficiently improve ad-hoc retrieval using LDA. Through research, an LDA-based document model was proposed, and it was shown that clustering-based models can improve the search function [52].

3. Speculative Cases

3.1. LH Scandal

As mentioned earlier, land speculation abroad generally refers to investing in the value of land with the expectation of a monetary gain, and simply stating that the loss from speculation is acceptable does not necessarily mean that speculation is negative.

However, because speculation in the ROK uses undisclosed inside information, which stems from a type of illegality, it is more like having a definite benefit rather than taking the risk of a loss. Therefore, the word 'speculation' in ROK is similar to 'illegal insider trading' abroad.

Looking at the case of speculation crackdown by the Korean National Police Agency, employees who were in charge of development at the LH business headquarters, and who purchased 181,011 ft of land, and employees who used development information while working at LH to purchase 14,233 ft of land, were arrested. In addition, a lawmaker who bought 24,837 ft of land using housing complex development project information obtained through the county council and a governor who purchased land on the planned development site were also arrested. Additionally, two representatives of companies who stole KRW 370 billion in real estate were also arrested.

By type, speculation in which farmland was purchased by giving the impression that there was an intention to farm was the most common, followed by housing speculation involving fraudulent subscription and planned real estate.

Kang Kang-yoon, a member of the National Assembly, was prosecuted after it was revealed that 2500 trees had been planted, but the city paid an additional compensation for 500 trees.

The police acknowledged that it did not meet the public's expectations and said that, due to the nature of real estate speculation, investigative agencies have no choice but to move through reporting or accusations and urge the enactment of laws and countermeasures against real estate speculation.

After the LH scandal, the government is eager to take measures to eradicate illegal real estate speculation, but as real estate speculation continues to occur, the public's trust in the public service is declining day by day.

Although intensive crackdowns have been carried out since the LH scandal, the lack of legal basis for punishment has been raised because only a few people, including national congressman and high-ranking public officials, have been accused of real estate speculation.

3.2. Speculation Schemes

In the ROK, compensation for residents or real estate owners of redevelopment zones is specified by law. Many speculation fraud schemes have emerged that target such compensations. In this section, these speculation fraud schemes shall be introduced, especially focusing on those that appeared in the LH Scandal.

Firstly, cases targeting money for compensation will be introduced. When the government purchases a plot of land for a state-run enterprise, such as in redevelopment or new town buildings, it is supposed to compensate the owner, considering the number and species of trees planted on that land. In that case, compensation is generally in the form of money. Knowing this, some LH workers purchased land where redevelopment or new town buildings had already been decided internally. They then abnormally and densely planted trees with the highest compensation. They also purchased vast plots of land and divided them into 1000 m^2, which is the maximum area that one can own without there being legal consequences. They used their families' and relatives' names to register themselves as the owner of the land. In some cases, the land they bought did not have any access to a road, which basically proved their intent of speculation, because such land is rarely sold due to its low usability. There was even a case using an anonymous person's grave, where the government compensates the relatives of a grave site in case it is on a redevelopment plot and must be moved. LH workers found several anonymous persons' graves and supplemented their families' names to receive the compensation money.

Secondly, the residents of redevelopment or new town residential zones receive land for living when the construction is completed, and money for moving-in support. However, LH workers built concrete cells in clusters without water pipes and electric lines, thus no one could actually live in them, while they falsely pretended to live there.

Finally, they purchased land in those plots and constructed buildings quickly to be sold at a high cost for profit, thus abusing inside information. The chart below briefly summarizes those speculation fraud schemes. Table 1 shows the ROK's common method of speculation.

Table 1. ROK's Common Methods of Speculation.

Item(s)	Trick(s)
Trees	Abnormally and densely planting species of trees with the highest compensation
Land	Purchasing large amounts of land and registering 1000 m^2 per person using family or relatives' names
Field does not border the road	Purchasing land that does not border the road and so has no merit of usage
Building	Quickly constructing buildings on purchased land and overcharging for them
Concrete cells	Building concrete cells with no water pipe or electric lines
Anonymous grave	Receiving compensation claiming an anonymous person's grave as their own families'.

4. Conceptual Framework Design

In this paper, we suggest a conceptual framework that detects real estate speculation using big data. Additionally, the presented framework operates on an app. The framework consists of three modules, and the function of each module is as follows:

1. Through comparison of internal information and public information, an undisclosed internal information DB is created.
2. A DB is created using existing real estate information, real estate transactions in real-time, and applications for compensation. Outliers are then detected to find suspected speculation areas.
3. The release data of internal information is compared with the date of real estate transactions for confirming real estate speculation.

Figure 1 shows the entire framework, which consists of three modules. The detailed contents of each module will be described in Sections 4.1–4.3.

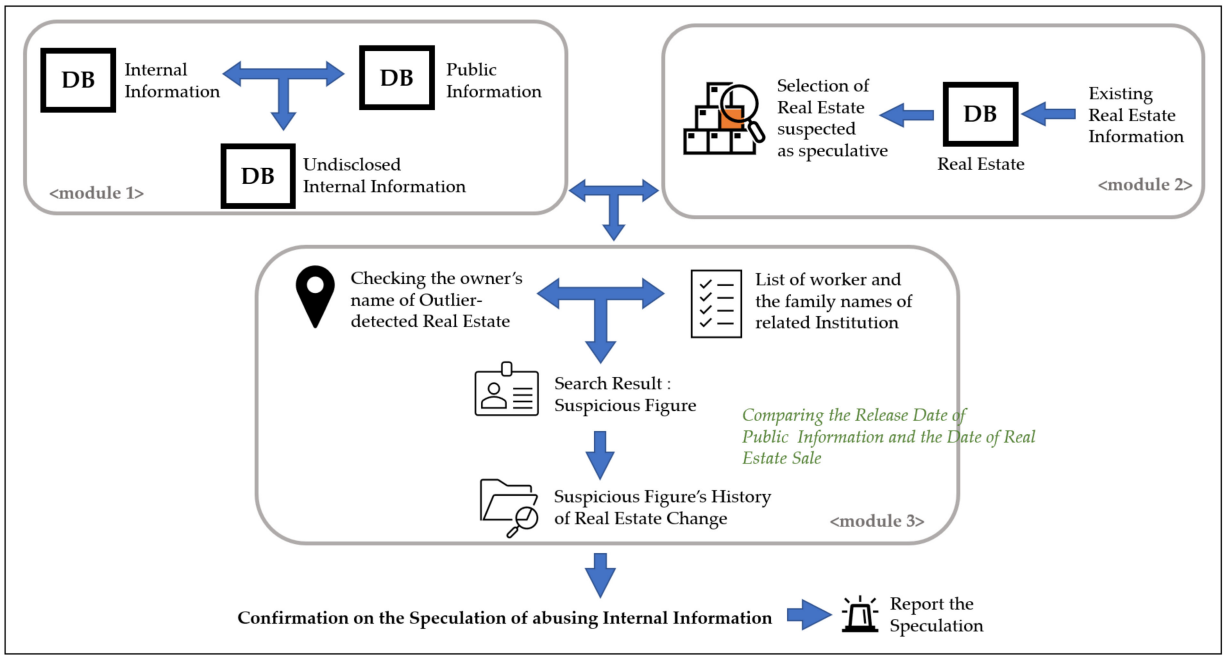

Figure 1. The Whole System of Each Module.

Figure 2 shows the flow chart for the entire framework process. Three different colors are used for each module. Orange, blue, and green colors denote Modules 1, 2, and 3, respectively. Modules 1 and 2 operate separately, and Module 3 operates by receiving the results of Modules 1 and 2.

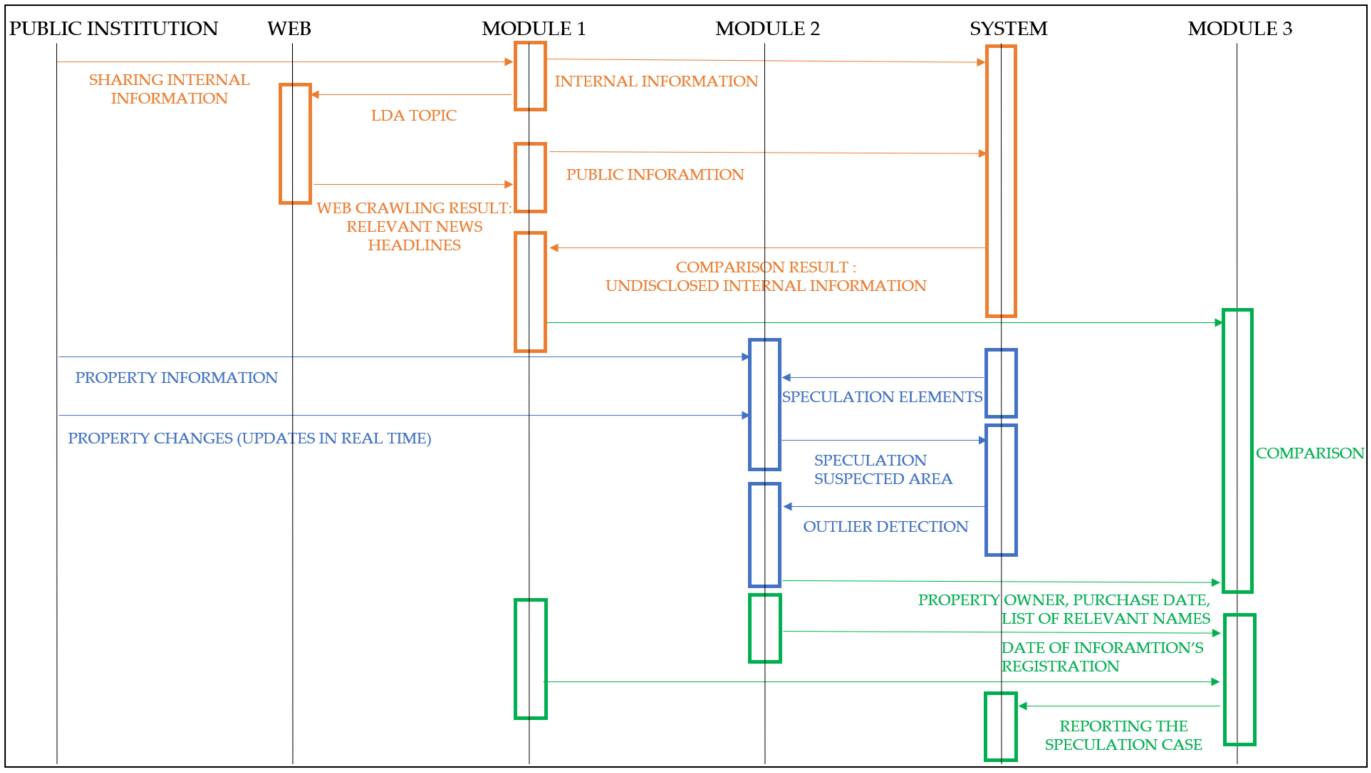

Figure 2. Flowchart showing how each module operates in the overall framework.

4.1. Creating an Undisclosed Internal Information DB

Prior to creating an undisclosed internal information DB in Module 1 of the framework, we need to create two different DBs. Firstly, to gather the internal information on ongoing projects, Module 1 is going to comprise an internal information DB with the official documents and reports provided by land authority state agencies. Once all the data is uploaded to the DB, Module 1 performs topic modeling using the LDA method.

Using LDA requires a number of preliminary tasks. First, documents in DB go through preprocessing. During the preprocessing, tokenization, stopword removal, and headword extraction are performed. Tokenization is the operation of separating the contents of a document by words, and stopword removal is the operation of removing words that have little contribution to actual semantic analysis, such as particles, articles, and conjunctions. Headword extraction is a method of reducing the total number of words by finding the root word. For example, 'am', 'are', and 'is' are different words, but the root word is 'be'. In this case, the headword of these words is 'be'.

A Term Frequency—Inverse Document Frequency (TF-IDF) matrix is created using the preprocessed words. TF-IDF is a method of calculating the importance of each word using term frequency and inverse document frequency. TF-IDF judges that a word that appears frequently in all documents has low importance and determines that a word that appears frequently in a specific document has high importance. When the value of TF-IDF is low, the importance is low, and when the value is large, the importance is high. A TF-IDF matrix is created by sorting based on the TF-IDF value, and this is used to perform LDA.

Secondly, Module 1 will create a public information DB with the results of web crawling. The topics of documents, gained by the LDA, will be form 'real estate keywords', and these will be used as the web crawling search word. With these words, the headlines

that the public information DB updates can be collected, which can prove whether internal information is open to the public or not.

Various libraries are used for web crawling. In this study, the request library, the Beautifulsoup library, and the KoNLPy library were used. The request library serves to receive the initial HTML from the portal site, and the Beautifulsoup library is a Python open source library used to scrape the received HTML web data.

KoNLPy is a Python library derived from NLP and is used to divide and process complex Korean texts by morphemes. In summary, a public information DB was created by collecting news article headlines from portal sites with Beautifulsoup and extracting only nouns from the headlines using KoNLPy.

Lastly, Module 1 will compare an internal information DB with a public information DB in order to create an undisclosed internal information DB. Local keyword is a subset of real estate keyword and is a set of words that refer to a specific area. Local keyword from a public information DB will be the search keyword to an internal information DB. Because being on a public information DB means the information is already open to the public, an undisclosed internal information DB only contains the information from such a DB that does not simultaneously belong to a public information DB.

The relationship between these three DBs can be seen in the simple diagram of Figure 3. When an investment that can doubt the awareness of the undisclosed internal information DB occurs, it could be a red flag indicating possible speculation. Figure 3 shows the process of generating an undisclosed internal information DB with a diagram that can help to understand the relationships between the three DBs.

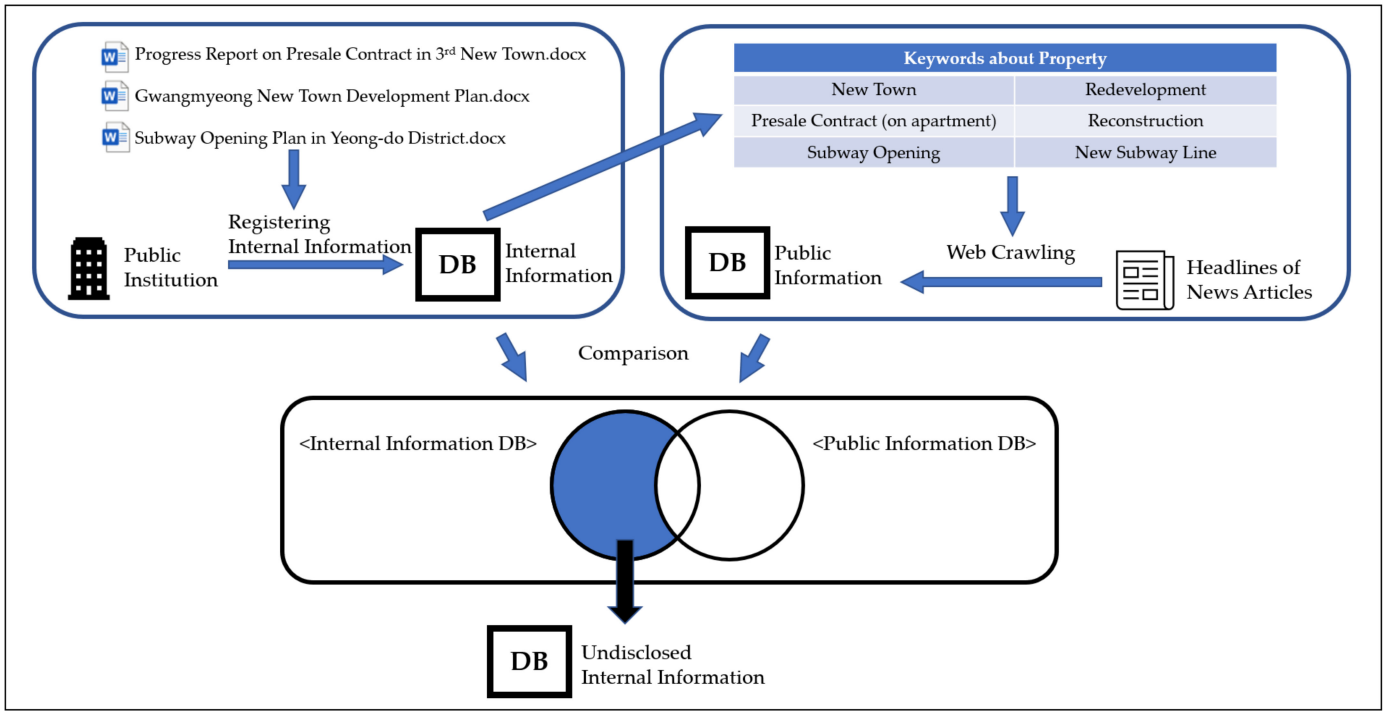

Figure 3. Process of Generating an undisclosed internal information DB.

Figure 4 shows the results of extracting 1500 news article headlines with the real estate keyword from Google and Korean portal sites Naver and Daum through a web crawler, removing duplicate articles, and visualizing the top 50 words with barplot.

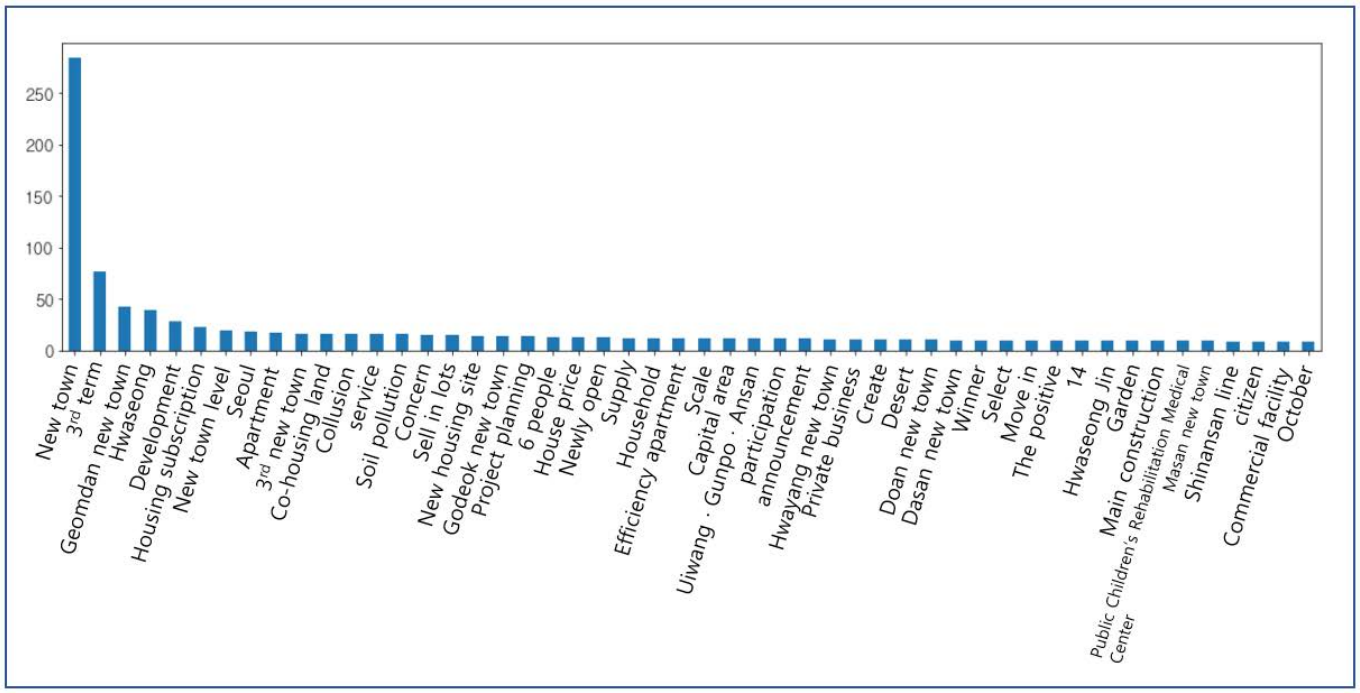

Figure 4. Barplot showing the 50 most frequent words in the results of web crawling as 'New Town'.

The words 'redevelopment', 'housing subscription', 'reconstruction', 'subway opening', and 'new route' in the Figure are the results of a web crawler, and one of the results, the Korean region 'Hwaseong', is also shown as the forth most frequent word in the graph.

Starting from the left, it shows the most tallies in the order of 'new town', '3rd term', 'Geomdan new town', and 'Hwaseong'. The extracted words are compared to region names, and the names Hwaseong, Seoul, Uiwang, Gunpo, Ansan, etc., have been selected as 'local keywords'.

4.2. Detecting Outliers to Find Suspected Speculation Areas

Module 2 will create a real estate DB with the data provided by a land authority state agency. App users will process their paperwork regarding real estate sales with the app, and the data will be collected in Module 2. Collected data will be used to update the real estate DB in real time.

Before performing outlier detection, Module 2 should perform data preprocessing to acquire the proper form of data for detection. For this, Module 2 collects the compensation application, real estate registration papers, and more. Real estate information from the collected papers goes through data preprocessing, generating visualized data that is specialized for outlier detection by the LOF algorithm.

The generation of the GIS visualization proceeds as follows: Module 2 classifies the collected papers by the speculation factors, and then marks the distribution or size of the factor on the virtual country map. For example, in the case of trees, Module 2 can mark the dots of trees in a specified area, and in the case of real estate, it can mark the land boundaries, add the information of the owner, and more. Maps for each factor become the visualized data.

Module 2 then performs the outlier detection on that data. The LOF algorithm helps to find the real estate containing the relevant factor that is frequently abused for speculation. For example, 25 evergreen trees planted in 1 sq. meter, which applies to the actual speculation cases involving LH workers, is abnormally dense. Therefore, it could mean the area has been used for another purpose. The LOF algorithm uses the low-density factor as an outlier, but in this paper, a factor with abnormally high density is the outlier. Considering the specific species of trees and land area at the same time would show whether the case

falls under a suspected speculation case. In a similar way, concrete cells can be detected by the LOF algorithm.

Other factors have other ways of being dealt with, and in this regard, land registered with a borrowed name can be detected by its size. If the area reaches 1000 sq. meters, it can be suspected as speculation, which also applies to the actual case with LH workers. The specific size of the land can be adjusted due to the laws of each country, and if the land also does not border a road, it raises a red flag of possible speculation. Falsifying information regarding an anonymous person's grave can be found by tallying the number of compensation applications per person, because it is possible that the fraudster will have repeated the falsification. These cases have not involved the LOF algorithm, but it has significant outliers in different ways.

After detecting all the outliers, Module 2 upgrades the relevant real estate with outliers to a 'suspected speculation area'. The duplication of speculation factors raises a red flag of possible speculation.

Figure 5 briefly shows the entire process of Module 2. The suspected speculation area finally selected in Module 2 is transferred to Module 3.

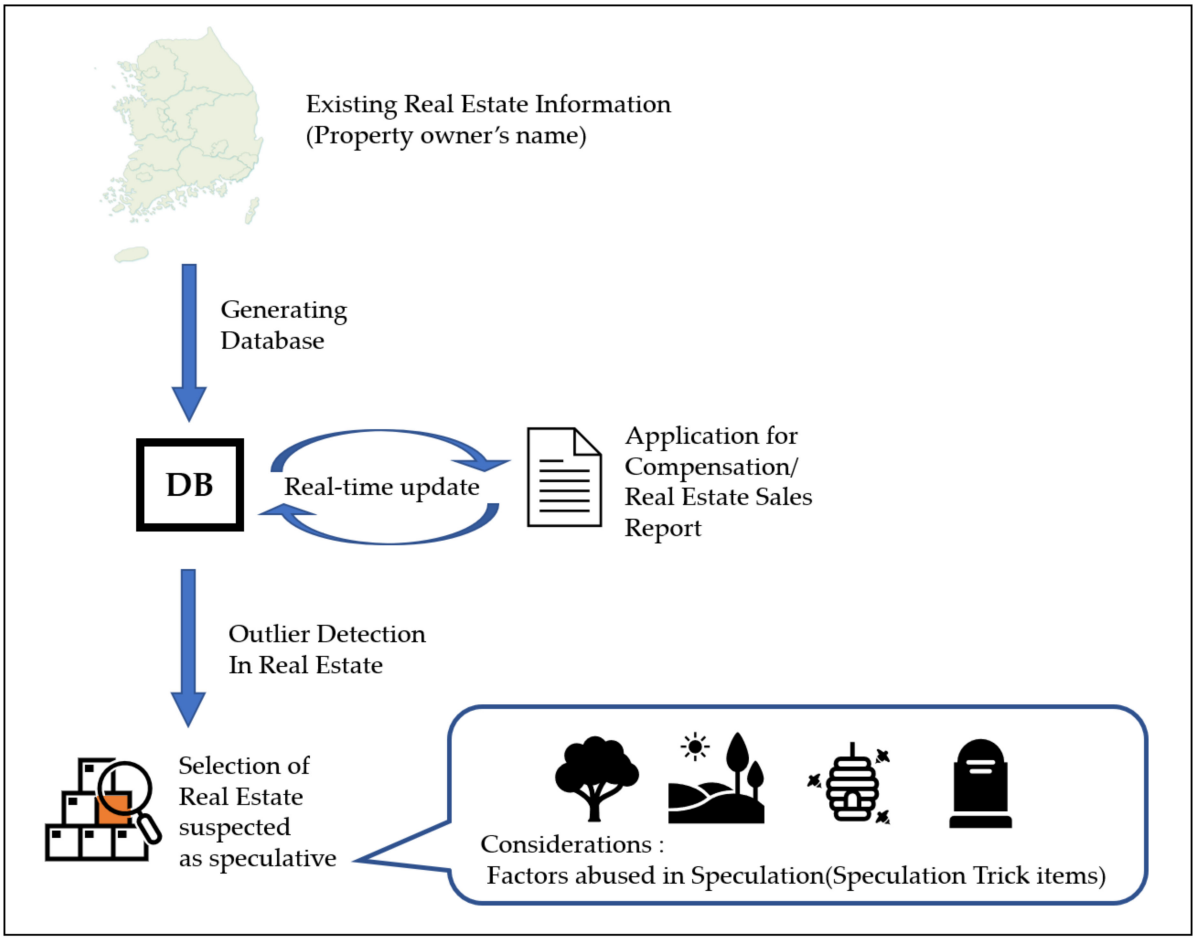

Figure 5. Outlier Detection for Selecting a Suspected Speculation Area.

4.3. Confirming the Speculation

Through Modules 1 and 2, an undisclosed internal information DB is developed, along with any suspected speculation areas. Module 3 will help in confirming a suspected speculation.

Module 3 compares the region of a suspected speculation area from Module 2 with the undisclosed internal information DB from Module 1. If the search result returns any matches, this is an indication that the owner of the land might have carried out speculation

using private internal information. Thus, Module 3 will run a comparison process with the related real estate in order to uncover the human route of a private internal information leakage. Module 3 will also compare the temporal data to ensure that no innocent person, who bought land without inside information, is accused.

However, there can also be a vulnerability with aiming for this blind spot, as follows. If some of the workers intend to speculate, they can delay uploading the official document to the DB and then purchase any related land for speculation. To prevent this, a suspected speculation area will remain in Module 3 and will be searched periodically until it shows any results. Figure 6 shows the process of confirming a suspected speculation.

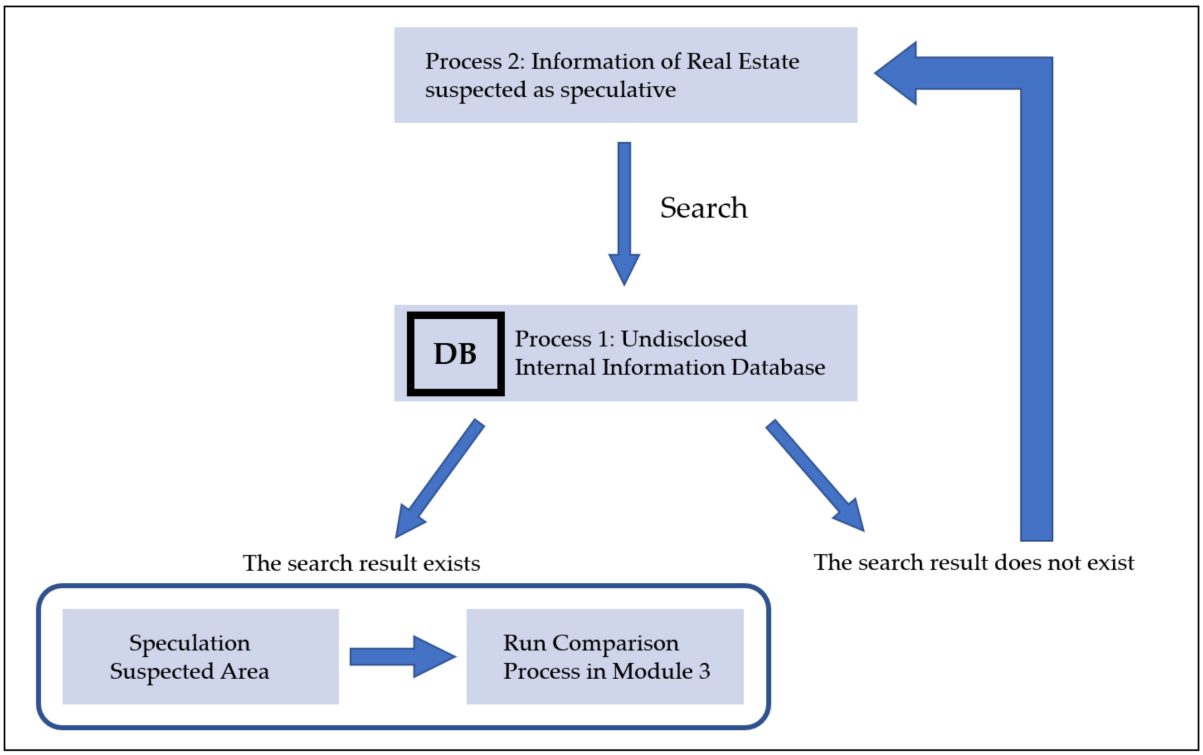

Figure 6. The Process of Confirming Speculation.

For the final step before the report, a comparison process will be undertaken. In the comparison process, the name of the owner of a suspected speculation area and a list of names from the authority state agencies will be compared. The list must include the names the of the workers' families and relatives for cases of 'borrowing names'. If Module 3 discovers an identical individual, Module 3 compares the release date of inside information and the date of the real estate sale.

There are four such cases, as follows:

Case (1) Release date of public information precedes the real estate sale.
Case (2) Release date of public information immediately precedes real estate sale.
Case (3) Release date of public information immediately follows real estate sale.
Case (4) Release date of public information follows the real estate sale.

It is hard to say whether Cases (1) and (4) are clearly speculation, considering such a coincidence. However, Cases (2) and (3) can be confirmed as speculation cases. Therefore, if Cases (2) and (3) do occur, Module 3 will report the speculation to the investigation department of the authority state agency or investigation agency.

Figure 7 shows the comparison process. When Module 3 discovers identical individuals through a comparison process, it will compare the release data of public information and the data of the real estate sale provided by the app.

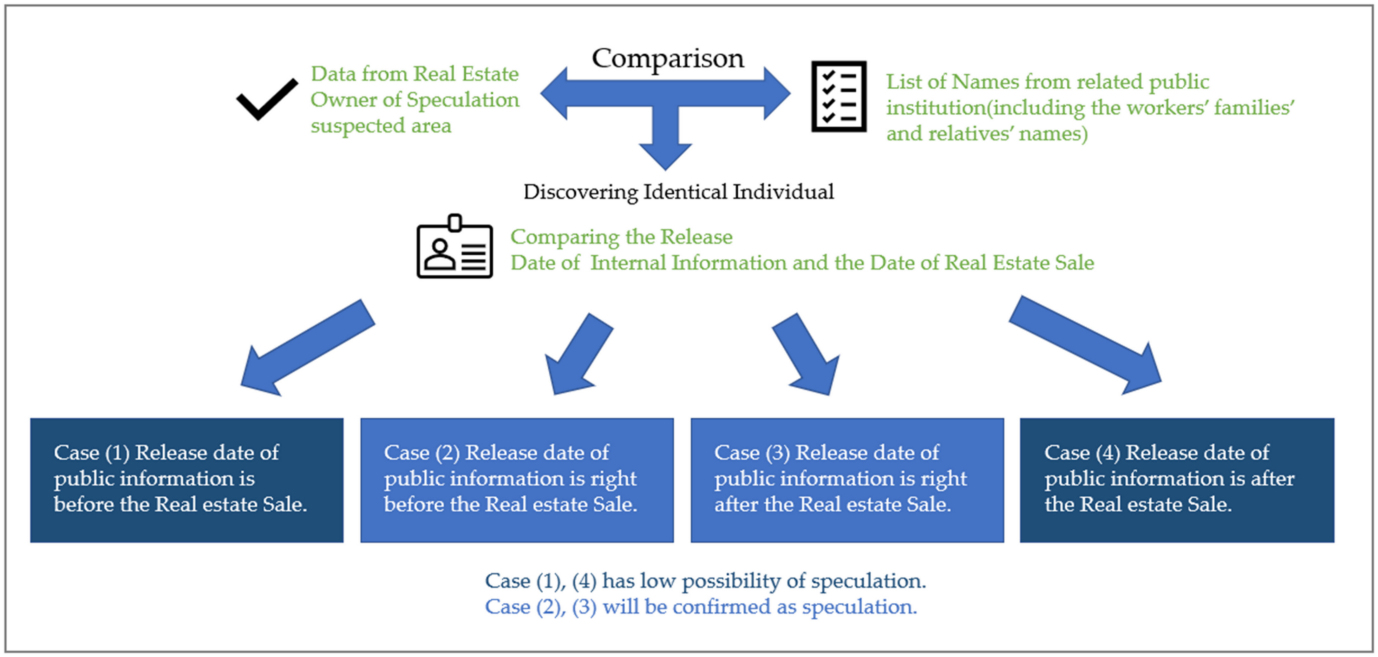

Figure 7. The Comparison Process.

5. Discussion

There are several constraints and limitations in applying our proposed framework to actual situations. In this section, the constraints and limitations are discussed separately for each module. The technical, legal, and institutional efforts to overcome these constraints and limitations are then also discussed.

The GIS visualization generated in Module 2 is defined in 3D spherical space based on latitude and longitude coordinates, rather than in vector space where LOF is often applied. Because of this, in order to apply the LOF, a process of projecting data into a vector space or modifying the LOF is required. In addition, the performance of the LOF may not be guaranteed. The biggest problem in Module 3 is that expert help is needed to ensure the performance of the proposed framework. In the process of comparing the release data of internal information and the data of real estate transactions, it is classified into four types. Among the types, a temporal difference is used to divide Case (1)–Case (2) and Case (3)–Case (4), which is quite ambiguous. In order to solve the problems in each of these modules, various technical efforts are required. In order to solve the problem of Module 2, a future study is needed to prove that the performance of the LOF can be guaranteed in 3D spherical space. For the problem of Module 3, data domain experts and actual speculative investigation experience are needed to formulate a specific temporal difference.

Regarding the legal aspect, while pointing out that the existing laws are not being activated, research is underway to find fundamental measures to prevent and eradicate real estate speculation. At the same time, related laws are enacted as the voices for legal improvement are growing [53]. As mentioned in the introduction, five laws were promoted to establish the legal basis for speculation punishment, and three of them have been proposed and enacted, therefore expanding the scope of punishment targets and enabling the monitoring. In addition, it has become possible to impose heavy fines and imprisonment when the speculation is confirmed.

The Korea Research Institute for Human Settlements is conducting research to suggest a way to prevent speculation, arguing that it is necessary to block speculation-centered compensation that violates the law or compensation received using gaps in the law [54]. However, there is a deficiency in this because two of the proposed five bills, which enable the monitoring of real estate disruptive behavior and insider trading, have still not been passed.

Because the three bills that have been passed are not retroactive, even when the system detects past speculation, punishment is still not possible. This must also be discussed and improved.

Additionally, assuming a case where a similar situation occurs abroad, and this system is introduced, legal support or modification would be required because this system needs a legal basis to gain private information and the information from public institutions, along with a bill for punishing speculation, to operate.

To confirm the speculation, the system needs all related internal information in the form of documents, produced during the period when the redevelopment project was being discussed. This is because related workers with the intent of speculation can carry out discussions on a redevelopment project verbally, purchase the real estate for speculation, and then register the related documents at different intervals of time. They can also hand over the undisclosed internal information to the media right before they purchase the real estate, therefore making the system misunderstand the case as 'purchasing the land with public information', not the undisclosed internal information.

6. Conclusions

In the ROK, there is an authority state agency called LH that handles real estate-related transactions and affairs. Here, the term 'speculation' is legally stipulated as being illegal. Nevertheless, some LH employees were caught committing speculation, which resulted in the destruction of public trust. In order to prevent this from reoccurring, we have attempted to design a conceptual framework to assist in the detection of real estate speculation. For this framework design, various methods have been applied to the real estate field, such as LDA and LOF.

The framework is designed to load existing data into a DB and upload newly generated data in real time. By using the proposed system, it is expected that any suspected speculation areas can be indicated, and such areas with a higher possibility of speculation due to their comparison results can be confirmed as a speculation case.

Although there are many limitations, as described in Section 5, this paper has presented the outline of a system for detecting speculative cases through the detection of outliers in real estate, and it is expected that this outline will contribute greatly to future research.

Author Contributions: Conceptualization, Y.-J.S., J.K., J.C. and J.-H.H.; data curation, Y.-J.S., J.K. and J.C.; formal analysis, Y.-J.S., J.K., J.C. and J.-H.H.; funding acquisition, J.-H.H.; investigation, J.K. and J.C.; methodology, Y.-J.S., J.K. and J.C.; resources, Y.-J.S. and J.C.; software, Y.-J.S., J.K., J.C. and J.-H.H.; supervision, J.C. and J.-H.H.; validation, J.-H.H.; visualization, J.K. and J.-H.H.; writing—original draft, Y.-J.S., J.K., J.C. and J.-H.H.; writing—review and editing, J.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

LH	Land and Housing Corporation
LDA	Latent Dirichlet Allocation
DB	Database
ROK	Republic of Korea
PETI	Public Ethics Total Information System
CDC	Closest Distance to the Center
LMS	Least Median of Squared residuals

IDS	Intrusion Detection System
LOF	Local Outlier Factor
EIF	Extended Isolation Forest
NLP	Natural Language Processing
TF-IDF	Term Frequency—Inverse Document Frequency

References

1. What Is Land Speculation? A Real Estate Investor's Guide. Millionacres. Available online: https://www.millionacres.com/real-estate-market/homebuying/what-is-land-speculation-a-real-estate-investors-guide/ (accessed on 27 March 2022).
2. Property Speculation Scandal Sweeps Korean Politics. The Korea Times. Available online: https://www.koreatimes.co.kr/www/nation/2021/03/113_305649.html (accessed on 27 March 2022).
3. Signal. Available online: https://signal.sedaily.com/NewsView/22JTY24DJE/GB04 (accessed on 8 November 2021).
4. The Den of Thieves': South Koreans Are Furious over Housing Scandal. The New York Times. Available online: https://www.nytimes.com/2021/03/23/world/asia/korea-housing-lh-scandal-moon-election.html (accessed on 27 March 2022).
5. MK. Available online: https://www.mk.co.kr/news/realestate/view/2021/03/241954/ (accessed on 27 March 2022).
6. MK. Available online: https://www.mk.co.kr/news/realestate/view/2021/10/986375/ (accessed on 27 March 2022).
7. THE300. Available online: https://the300.mt.co.kr/newsView.html?no=2021032414310183216 (accessed on 27 March 2022).
8. PETI. Available online: https://www.peti.go.kr/peSsmHist.do (accessed on 4 April 2022).
9. The Ministry of Public Administration and Security. Available online: https://mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=58056 (accessed on 4 April 2022).
10. PETI. Available online: https://www.peti.go.kr/pldtEtBsnOpn.do (accessed on 4 April 2022).
11. FNNEWS. Available online: https://www.fnnews.com/news/202109141003462350 (accessed on 4 April 2022).
12. How the SEC Tracks Insider Trading. Investopedia. Available online: https://www.investopedia.com/articles/investing/021815/how-sec-tracks-insider-trading.asp (accessed on 22 November 2021).
13. Moore, J. What is really unethical about insider trading? *J. Bus. Ethics* **1990**, *9*, 171–182. [CrossRef]
14. Ryan, J.A. The ethics of speculation. *Int. J. Ethics* **1902**, *12*, 335–347. [CrossRef]
15. Werhane, P.H. The indefensibility of insider trading. *J. Bus. Ethics* **1991**, *10*, 723–731. [CrossRef]
16. Bainbridge, S.M. An overview of insider trading law and policy: An introduction to the insider trading research handbook. In *Research Handbook on Insider Trading, Stephen Bainbridge*; Edward Elgar Publishing Ltd.: Cheltenham, UK, 2013; pp. 28–29.
17. Chen, Z.; Wang, C. Effects of intervention policies on speculation in housing market: Evidence from China. *J. Manag. Sci. Eng.* **2021**. [CrossRef]
18. Fernandes, N.; Ferreira, M. Insider Trading Laws and Stock Price Informativeness. *Rev. Financ. Stud.* **2009**, *22*, 1845–1887. [CrossRef]
19. Kwabi, F.O.; Boateng, A. The effect of insider trading laws and enforcement on stock market transaction cost. *Rev. Quant. Financ. Account.* **2021**, *56*, 939–964. [CrossRef]
20. Du, X.; Cindy, L.Y.; Hayes, D.J. Speculation and volatility spillover in the crude oil and agricultural commodity markets: A Bayesian analysis. *Energy Econ.* **2011**, *33*, 497–503. [CrossRef]
21. Mian, A.; Amir, S. Credit supply and housing speculation. No. w24823. *Natl. Bur. Econ. Res.* **2018**, *35*, 680–719.
22. Shabestari, M.A.; Cao, M.; Sarath, B. Patterns of Insider Trading: It Is Not All Black and White. *J. Account. Audit. Financ.* **2021**, *36*, 695–722. [CrossRef]
23. Malpezzi, S.; Wachter, S. The Role of Speculation in Real Estate Cycles. *J. Real Estate Lit.* **2005**, *13*, 141–164. [CrossRef]
24. Esther, B.; Brio, D.; Minguel, A.; Perote, J. An investigation of insider trading profits in the Spanish stock market. *Q. Rev. Econ. Financ.* **2002**, *42*, 73–94.
25. Dallin, M.; Alldredge, D.; Cicero, C. Attentive insider trading. *J. Financ. Econ.* **2015**, *115*, 84–101.
26. Huddart, S.; Ke, B.; Shi, C. Jeopardy, non-public information, and insider trading around SEC 10-K and 10-Q filings. *J. Account. Econ.* **2007**, *43*, 3–36. [CrossRef]
27. Donoho, S. Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 22–25 August 2004; pp. 420–429.
28. Wang, C.; Li, J.; Guo, P. The normalized interval regression model with outlier detection and its real-world application to house pricing problems. *Fuzzy Sets Syst.* **2015**, *274*, 109–123. [CrossRef]
29. Kontrimas, V.; Verikas, A. Tracking of doubtful real estate transactions by outlier detection methods: A comparative study. *Inf. Technol. Control* **2006**, *35*, 94–105.
30. Morano, P.; De Mare, G.; Tajani, F. LMS for Outliers Detection in the Analysis of a Real Estate Segment of Bari. In Proceedings of the International Conference on Computational Science and Its Applications, Ho Chi Minh City, Vietnam, 24–27 June 2013; pp. 457–472.
31. Barańska, A.; Śpiewak, B. The Influence of Chosen Statistical Methods of Detecting Outliers on Property Valuation Result. *Real Estate Manag. Valuat.* **2021**, *29*, 87–97. [CrossRef]

32. Özçelik, R.; Bayar, S. Outlier Detection Based on Majority Voting: A Case Study on Real Estate Prices. In Proceedings of the 2018 IEEE 12th International Conference on Application of Information and Communication Technologies, Almaty, Kazakhstan, 17–19 October 2018.
33. Choukuljaratsiri, A.; Lertwongkhanakool, N.; Thienprapasith, P.; Pratanwanich, N.; Chuangsuwanich, E. Anomaly Detection for Online Visiting Traffic as a Real-Estate Indicator: The Case of HomeBuyer. *Behav. Predict. Modeling Econ.* **2021**, *897*, 291–301.
34. Hawkins, D.M. *Identification of Outliers*; Springer: Berlin/Heidelberg, Germany, 1980.
35. Grubbs, F.E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, *11*, 1–21. [CrossRef]
36. Denning, D.E. An Intrusion-Detection Model. *IEEE Trans. Softw. Eng.* **1987**, *SE-13*, 2. [CrossRef]
37. Knorr, E.M.; Raymond, T.N.; Tucakov, V. Distance-based outliers: Algorithms and applications. *VLDB J.* **2000**, *8*, 237–253. [CrossRef]
38. Breunig, M.M.; Hans-Peter, K.; Raymond, T.N.; Sander, J. LOF: Identifying Density-Based Local Outliers. *Sigmod. Record* **2000**, *29*, 93–104. [CrossRef]
39. Lazarevic, A.; Kumar, V. Feature bagging for outlier detection. In Proceedings of the Eleventh ACM SIGKDD International Conference: Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; pp. 157–166.
40. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008.
41. Zhang, K.; Hutter, M.; Jin, H. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. In Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009), Bangkok, Thailand, 27–30 April 2009; pp. 813–822.
42. Hariri, S.; Kind, M.C.; Brunner, J.R. Extended Isolation Forest. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 4. [CrossRef]
43. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
44. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]
45. Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. A density-based method for adaptive LDA model selection. *Neurocomputing* **2009**, *72*, 1775–1781. [CrossRef]
46. Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; Smyth, P. The Author-Topic Model for Authors and Documents. *arXiv* **2012**, arXiv:1207.4169.
47. Arora, R.; Ravindran, B. Latent Dirichlet Allocation Based Multi-Document Summarization. In Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, Singapore, 24 July 2008; pp. 91–97.
48. Andrzejewski, D.; Zhu, X. Latent Dirichlet Allocation with Topic in-Set Knowledge. In proceeding of the NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing, Boulder, CO, USA, 22 June 2009; pp. 43–48.
49. Ramage, D.; Hall, D.; Nallapati, R.; Manning, C.D. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 248–256.
50. Krestel, R.; Fankhauser, P.; Nejdl, W. Latent dirichlet allocation for tag recommendation. In Proceedings of the Third ACM Conference on Recommender System, New York, NY, USA, 23–25 October 2009; pp. 61–68.
51. Moghaddam, S.; Ester, M. On the design of LDA models for aspect-based opinion mining. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012; pp. 803–812.
52. Wei, X.; Croft, W.B. LDA-Based Document Models for Ad-hoc Retrieval. In proceeding of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 6–11 August 2006; pp. 178–185.
53. Yul, P.C.; Jun, P.J.; Gum, S.S. A Study on the Revitalization of the Land Compensation System to Prevent Real Estate Speculation—Focusing on the need to revise the Land Compensation Act. *Korean Assoc. Cadastre Inf.* **2021**, *23*, 38–62.
54. Jong, K.S.; Seung, S.H.; Young, S.H. *Policy Directions for the Prevention of the Speculation in the Compensation*; Korea Research Institute for Human Settlements (KRIHS): Anyang, Korea, 2020; ISBN 9791158986100.