Article

A Machine Learning Approach to the Residential Relocation Distance of Households Living in the Seoul Metropolitan Region

Changhyo Yi [1,*] and Kijung Kim [2]

[1] Department of Urban Engineering, Hanbat National University, Daejeon 34158, Republic of Korea; yich@hanbat.ac.kr

[2] Department of Urban Planning and Design, University of Seoul, Seoul 02504, Republic of Korea; kimkj87@uos.ac.kr

* Correspondence: yich@hanbat.ac.kr; Tel.: +82-42-821-1194

Abstract: This study aimed to ascertain the applicability of a machine learning approach to the description of residential mobility patterns of households in the Seoul metropolitan region (SMR). The spatial range and temporal scope of the empirical study were set to 2015 to review the most recent residential mobility patterns in the SMR. The analysis data used in this study involve the microdata of Internal Migration Statistics provided by the Microdata Integrated Service of Statistics Korea. We analysed the residential relocation distance of households in the SMR by using machine learning techniques such as ordinary least squares regression and decision tree regression. The results of this study showed that a decision tree model can be more advantageous than ordinary least squares regression in terms of the explanatory power and estimation of moving distance. A large number of residential movements are mainly related to the accessibility to employment markets and some household characteristics. The shortest movements occur when households with two or more members move into densely populated districts. In contrast, job-based residential movements have relatively longer distance. Furthermore, we derived knowledge on residential relocation distance, which can provide significant information on the urban management of metropolitan residential districts and the construction of reasonable housing policies.

Keywords: residential relocation distance; residential movement; machine learning; decision tree regression; Seoul metropolitan region

1. Introduction

A large number of households experience multiple residential movements during their lifetime, although some people continue their lives in only one location. Residential movements have been researched in residential choices and preferences as a searching process of appropriate location and dwelling with respect to individual characteristics. However, residential choices and preferences should be clearly distinguished. Residential choices indicate the actual behaviour related to residential movement, and residential preference is related to the relative attractiveness of housing and residential environment that affect movers [1]. Residential mobility can be represented by the spatial moving pattern based on actual behaviours of the movers. Previous studies of spatial patterns of residential relocations focused on the conventional research topics: frequency, direction, and distance of residential mobility. The life-cycle model [2], sector model [3], and Ravenstein's Laws [4] are representative research achievements, as well as theories related to these subjects. However, the relevant empirical studies for household units have paid relatively scant attention to the topics of direction and distance of residential mobility. This could be because of the excessive complexity of influencing factors, lack of computing power to handle large volumes of data on household movements, and absence of appropriate analytical model [5].

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

After economic achievement and quantitative growth [6], the Korean housing market has experienced structural changes in terms of both supply and demand. The housing shortage problem of the Korean society is considered to have been resolved with the housing supply ratio exceeding 100% in the early 21st century, and a fundamental change in the nature of household [2], which is a basic unit of residential mobility and location change, is in progress. Representative phenomena involve the reduction in the household size and aging, indicating the emergence of a new demand class and the change of characteristics in the core demand groups. These situations are summarised as the transitioning from a supply-based housing market to a demand-driven market [7]. Regarding the demand, with the slowdown in population growth, the flow management of residential mobility, considering the relocations within the metropolitan region, is gaining more importance than the response to new demand caused by the increased population in the metropolitan region. Previous studies [8,9] confirmed that the frequency of residential relocations of the Korean households is relatively high among the Organization for Economic Cooperation and Development (OECD) countries. In addition, recent studies [10,11] have shown that residential relocation distance could be differentiated by household size and age of householder in the Seoul metropolitan region (SMR), which is the most representative and largest metropolitan region in Korea. These phenomena could be changed by the reduction of household size and aging trends.

An empirical understanding of spatial patterns and characteristics related to residential relocation is important for the establishment of an in-depth housing policy. In addition, considering the growing socio-economic complexity, residential mobility research using spatial Big Data is more advantageous than research using only aggregated data. The study based on actual moving data of households is more meaningful, as it can identify the practical residential moving patterns, considering the conditions of a household rather than the ideal pursuit of a specific household. In a continually changing housing market, such as the Korean housing market, the outcomes of such a study could be applied to build a simulation model for forecasting future residential relocations. Accordingly, academic reviews and empirical studies must attempt to apply new analytical methods such as machine learning, which is used to derive meaningful knowledge from Big Data in the housing and residential research fields. If such an attempt is successful, in the long run, it can be used to construct a sustainable-housing-market-management system from the socio-economic aspect.

In this context, this study aimed to ascertain the applicability of a machine learning approach to the description of the residential mobility patterns of households in the SMR. In particular, this study focuses on the residential relocation distances of households, which is one of the main topics representing residential spatial patterns and has not been focused upon in previous empirical studies on household units, because residential relocation distance is a key factor in determining the spatial extent of the housing (sub)market. In this paper, we first review literature on patterns and influencing factors of residential relocation and examined the relocation characteristics of the SMR in Korea. Next, we conducted empirical studies analysing the determinants in residential relocation distance by using a machine learning approach. Finally, we conclude by summarising the outcomes of this study and ascertaining the applicability of the machine learning method in estimating or forecasting studies of housing and residential research.

2. Literature Review

Residential mobility is defined as a process of adjusting location to better meet the needs and demands of a household [7,12,13]. Residential mobility can be divided into residential relocation and urban migration. Residential relocation implies moving a residence within an urban living region, and urban migration refers to moving out of that region. While urban migration mostly results from changes in urbanisation and industrial structure [14], residential relocation is influenced by internal and external factors of a household, such as income, composition, housing preference, and residential environments. Residential movement occurs based on not only dissatisfaction with current location, but also attractiveness to the new location [15–17]. Previous studies, which examined the influencing factors of residential mobility, assumed a household-based decision-making mechanism. These representative studies considered various household characteristics, such as composition of

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

household members [18], age and income [19], education level [20], and marriage duration [21]. However, these studies mainly focused on analysing the residential mobility.

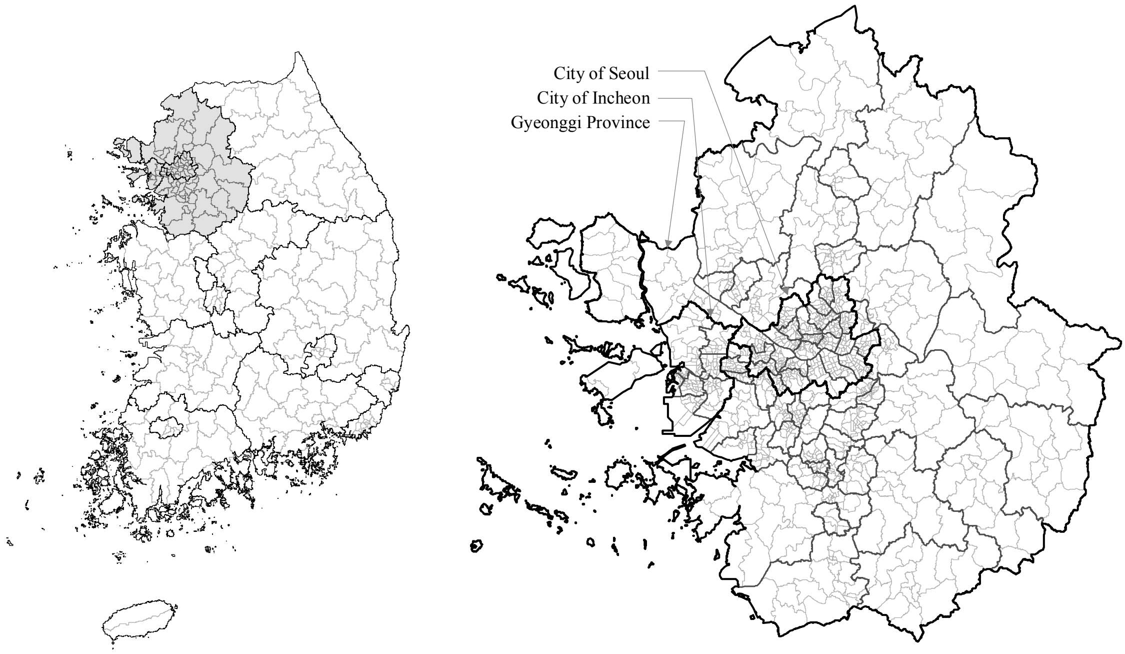
Recently, not only the amount of flow but also the residential mobility patterns have gained interest in terms of suggesting implications to spatial planning and housing policy [11,22–25]. The moving patterns of households can be explained using the frequency, direction, and distance of residential mobility. In terms of the frequency of residential movements, the main reasons of residential mobility are the characteristics of the household and the changes in the life cycle of the household. In particular, the life cycle is a series of processes that human beings experience in their lives, resulting in a change in needs and demands for the living space according to each stage [2,26,27]. According to the life-cycle model, the changes in the characteristics of the frequency of movements depend on family events, such as marriage (formation), birth of children (expansion), moving out (contraction), and divorce or death of a spouse (dissolution) [2]. As the characteristics of the household change according to the life-cycle stage, many researchers have studied the probability of residential mobility affected by a stage. The previous studies show various empirical results in consideration of birth, childcare, marital age, and income with respect to individual households [18–21,26,28,29].

In terms of residential mobility direction, Hoyt's sector theory, which states “High grade residential growth tends to proceed from the given point of origin, along established lines of travel or toward another existing nucleus of buildings or trading centers” [3], is the initial theory in this research field. This theory suggests that the direction of residential mobility is due to the difference in rents generated in urban space. In the empirical study related to this theory, Burnley et al. [30] found that most of the residential mobilities in Australia are biased toward the outward direction from an urban center. Furthermore, Yang [31] reported that 26% of households moved to the outskirts of the city from the urban center, while only 9% of the households moved in the opposite direction. Regarding the distance of residential movement, the widely known Ravenstein’s Laws suggests that most migration occurs over short distances [4]. The main research topics covered by related studies were concentrated to the quantity of flow of residential movements between origin and destination based on the gravity model. That is, the results of previous studies show a lack of in-depth research on the spatial patterns of residential moving distance. The short distance of residential movements is related to the existence of local housing markets (or housing submarkets) [32]. This study is a basic model that explains the residential relocation distance and links residential mobility to the local housing market. However, these previous studies did not consider demographic and socio-economic changes of modern society. In addition, while the studies on the frequency of residential movement considered the various characteristics of households, some studies on residential relocation distance and moving direction only considered the household characteristics.

Several studies have determined that the residential relocation distance differs according to household size, home ownership, job change, and parental status [10,11,29,33]. However, these studies compared and analysed the residential moving data aggregated by household characteristic. In addition, the model for estimating the moving distance of each household has not been developed yet. This is mostly because of the difficulty in obtaining the data of moved households and the lack of an analytical method for large volume data [5]. Nowadays, a large amount of residential relocation data of individual households is being provided by the Korean government agency and various analysing methods are being developed for Big Data. Especially in the Korean housing market, which is experiencing a rapid demographic change, the understanding of the spatial patterns of residential movements is gaining increasing importance because the housing demand and the behaviour of housing movement gradually change based on the household type. Therefore, this study focused on the application of a new approach that uses machine learning, which is advantageous for Big Data analysis, in order to empirically identify the impact of the household attributes and the location characteristics on the residential relocation distance in Korea.

3. Characteristics of Residential Relocation Distance in SMR

The main spatial range of this study was the SMR, which is a representative metropolitan region located in northwestern South Korea and includes the cities of Seoul and Incheon and the Gyeonggi province, and the temporal scope was the year 2015. The spatial unit of the present empirical analysis involved the administrative district (Eup, Myeon, and Dong), which is a minimum-sized administrative area-level unit in the SMR. The total area of the SMR is 11,828 km^2, with a population of 23.906 million people living in 9.519 million households. In addition, the SMR contains two metropolitan cities (Seoul and Incheon), one province (Gyeonggi-do), 28 cities (Si), 5 counties (Gun), and 53 boroughs (Gu), which comprise 1,133 small administrative areas (Eup, Myeon, and Dong) (see Figure 1 and Table 1).



(**a**) (**b**)

Figure 1. (**a**) Location of SMR in Korea; (**b**) Components of SMR.

Table 1. Seoul metropolitan region characteristics

Item		Total	Seoul	Incheon	Gyeonggi
Population (million people)		23.906	9.395	2.767	11.744
Household (million households)		9.519	3.915	1.066	4.538
Area (km^2)		11,828	605	1,048	10,175
City, county and borough level	Si	28	-	-	28
	Gun	5	-	2	3
	Gu	53	25	8	20
Minimum-sized administrative area level	Eup	34	-	1	33
	Myeon	127	-	19	108
	Dong	972	424	129	419

Source: Statistics of Urban Planning in 2015, 2015 Census in Korea.

The microdata of Internal Migration Statistics of Korea were used to analyse the spatial characteristics of residential relocation. Internal Migration Statistics includes information of Korean migrants from/to the smallest administrative areas of Eup, Myeon, and Dong obtained through using the migrant's moving-in notifications. First, in the data analysis collected in 2015, the total number of residential movements of households in Korea exceeds 6 million (6,098,915), of which approximately 3.1 million occurred in the SMR. The share of residential relocations within the SMR was 88.4%, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

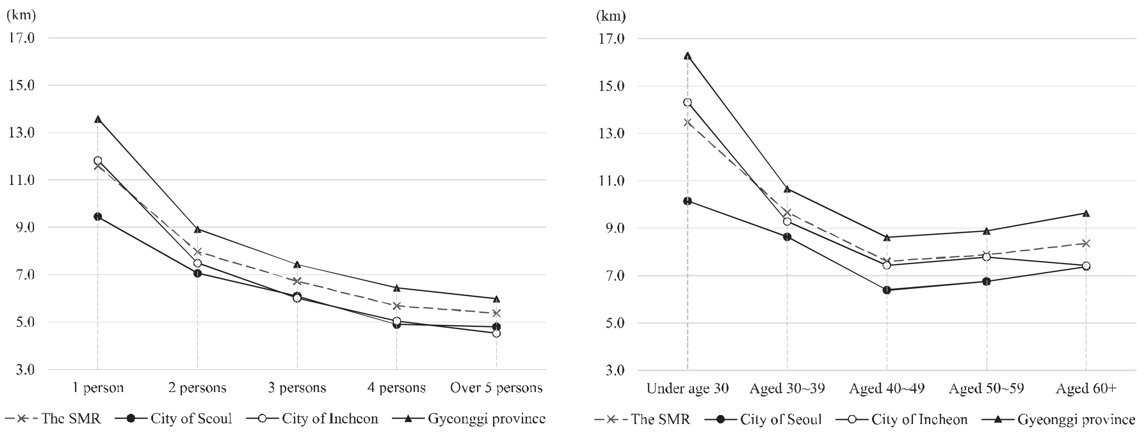
Peer-reviewed version available at *Sustainability* **2018**, *10*, 2996; doi:10.3390/su10092996

occupied the majority of residential mobility in the metropolitan region. The share of residential mobility in the metropolitan region was differentiated from the movement toward the inside and outside by the municipality. The rates of residential relocations within the area were relatively low in the metropolitan cities, such as Seoul and Incheon, and approximately 30% of residential movements were confirmed to move beyond the boundaries of each municipality. The number of residential movements per household was 0.326 in 2015, and the difference by area was not significant. Second, the average residential relocation distance was 9.123 km in the SMR. As expected, the average distance of residential movements from Seoul was the shortest (7.753 km), and that from Gyeonggi province was the longest (10.391). However, the moving-out beyond the boundary of Incheon city with the longest distance (29.112 km) was an unexpected outcome. This result is presumed to be caused by the difference between the characteristics of the moving-out households (refer to Table 2).

Table 2. Frequency and distance of residential movements

Item		SMR	Seoul	Incheon	Gyeonggi
Frequency of residential movements	Total	3,107,134 (100.0%)	1,287,379 (100.0%)	352,488 (100.0%)	1,467,267 (100.0%)
	Inside	2,747,380 (88.4%)	882,299 (68.5%)	247,760 (70.3%)	1,081,897 (73.7%)
	Outside	359,754 (11.6%)	405,080 (31.5%)	104,728 (29.7%)	385,370 (26.3%)
	Movement per household	0.326	0.329	0.331	0.323
Residential relocation distance [1] (km)	Total	9.123	7.753	8.894	10.391
	Inside	-	3.940	4.304	7.965
	Outside	-	23.909	29.112	25.412

[1] The average Euclidian distance calculated using 10% randomly sampled data from the raw data.



(**a**) (**b**)

Figure 2. Relocation Distance based on (**a**) Family Size and (**b**) Age Group of Householder

Figure 2 shows the difference among residential moving distances by household types according to characteristics of the household. In terms of household size, households with more members represented a shorter moving distance. The households with three or more people in the metropolitan cities (Seoul and Incheon) moved a similar distance, while the relocation distance of one-person households showed a significant difference among municipalities. In addition, the age of a householder is considered as a critical factor affecting the residential relocation distance of households, and this result is identical with previously confirmed outcomes. The longest relocation distance of households occurs for householders under age 30, and decreases to age range 40-49. Then,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

the distance of residential movements gradually increases with age. This phenomenon agrees with the results of previous literature [10,11] in Korea.

The estimated results of residential relocation distance in the SMR has the following implications. First, the moving distance with respect to a household could vary according to the area in which the household is located. Second, depending on the characteristics of the household members, there could be differences in the moving distance. These outcomes imply that characteristics of households and their location features should be considered in the construction of a model for estimating residential relocation distance.

4. Methods and Materials

4.1. Decision Tree Using Machine Learning

The main analytical methodology of this study is machine learning, which is an efficient tool in automatically detecting patterns of data and extracting information from large datasets [34]. Machine learning differs from conventional statistics in that it is more concerned with making estimations or predictions by using a model and formulating the generalisation process as a search through hypotheses. In contrast, conventional statistics is more concerned with testing hypotheses [35]. Machine learning focuses on estimation or prediction by considering an optimal model, while the latter concentrates on understanding the relationships between data. Recently, a few related studies applying a machine learning-based method have been reported in various research fields, such as environmental science, geomatics, and social science [36–39].

Decision trees in machine learning techniques are widely used for classification or regression problems and generate the result in a tree form, which can be interpreted relatively easily compared to the results of other techniques [40,41]. Thus, decision trees are known as a white-box model in the software engineering field. Decision trees are classified into classification and regression trees, which are constructed by repeatedly splitting the data. Each branch of a regression tree is partitioned according to the homogeneity of two resulting groups; the homogeneity is maximised according to the response variable. This method does not assume the relationship between the response and predictors, unlike the conventional statistical model in which the relationship independent and dependent variables is predefined and verified [42]. Therefore, the decision-tree regression method has more advantages than the conventional statistical models with respect to fitting and estimation using extremely complex data and structure. Therefore, in this study, the residential relocation distance of each household in the SMR was analysed using decision-tree regression, which can be regarded as the most appropriate model for analysis and estimation, considering rapidly changing demographic transitions and household characteristics in the Korean housing market.

4.2. Selection of Explanatory Variables and Generation of Analysing Data

The estimated residential relocation distance is the dependent variable for conducting empirical analysis by using a decision tree. The microdata obtained from the Internal Migration Statistics in this study provides information of the smallest administrative district (Eup, Myeon, and Dong), which is the same as a small-sized traffic analysis zone (TAZ), for the point of departure and destination of each household's residential movement. Therefore, we estimated the moving distance between the departure and destination based on the administrative center points by applying the Euclidian distance calculation method. The cases for which the point of departure and destination are the same, the following formula was applied to estimate the moving distance:

$$A = \pi r^2 \iff r = \sqrt{A/\pi}, \quad (1)$$

where A is the area of the administrative district and r is the radius that assumes an irregularly shaped administrative district as a circle.

The explanatory variables affecting the residential relocation distance of households moving within the SMR were selected based on the results of previous studies and the hypothesis of the present study. In this empirical analysis, not only the household attributes, but also the location characteristics were selected considering the results from previous related research, for example, life-cycle stages, residential mobility, and residential location choices. The variables contained in the household attributes group were available from the microdata of Internal Migration Statistics.

In Table 3, the explanatory variables are classified into household attributes and location characteristics. First, variables related to the attributes of household are *moving reason*, which include job, house, and education; *age*; *gender*; *members*; *elderly people*; *children*; and *proportion of males* in the household; these were collected from the microdata of Internal Migration Statistics in 2015. The three nominal variables labeled as *moving reason* were coded as 1 if each moving reason was job, house, or education, and 0 otherwise. These variables were selected to identify the influence of specific mobility reasons of households on the moving distance. *Age* is defined as the age of the householder. *Gender* is a nominal variable equal to 1 if the householder is male and 0 otherwise. *Member*, *elderly people,* and *children* are variables related to the household structure; these are measured as the number of corresponding members of each household. Finally, *proportion of male* is defined as the proportion of male among total household members; it is measured at a ratio.

Table 3. Explanatory Variables Applied in the Empirical Analysis

Variable		Description	Unit	Source
Household attributes	Moving reason	Major reasons for residential relocation; Job/House/Education	-	The microdata of Internal Migration Statistics
	Age	Age of householder	Year	
	Gender	Male and female	-	
	Members	Number of household members	People	
	Elderly people	Number of elderly household members	People	
	Children	Number of school-aged children; Primary/Secondary	People	
	Proportion of males	Proportion of male household members	%	
Location characteristics [1]	Accessibility	Accessibility to employment market	-	Census on Establishments
	Density	Population density	People/ha	Population Census
	New building	Proportion of new building; 1 year/5 years	%	Housing Census
	Housing ownership	Ratio of owner-occupied housing	%	Population Census
	Rail availability	Ratio of rail catchment area	%	Korea Transport Database
	Bus availability	Number of metropolitan bus routes	EA	

[1] The variables contained in the domain of location characteristics were calculated for both origin and destination locations.

Second, the location variables include *accessibility*, *density*, *new building*, *housing ownership*, *rail availability*, and *bus availability*, which were calculated with respect to both the departure and destination positions of each household's residential movement. *Accessibility* was selected as an explanatory variable measuring how the location advantage of employment opportunities affects the moving distance of households. The accessibility to employment market was calculated using the methodology representing location attraction, as mentioned by Hansen [43] and Wilson [44]:

$$Acc_i = \ln \sum_j Job_j \times \alpha\left(d_{ij}^{\beta}\right) \times exp\left(\gamma d_{ij}\right), \tag{2}$$

where Acc_i is the accessibility of administrative district i; Job_j is the number of jobs in potential destination administrative district j; d_{ij} represents the Euclidian distance between administrative districts i and j; and α, β, γ are the parameters. The parameters obtained from the analysis of commuting patterns in the SMR in 2015 (the Metropolitan Transport Association) were applied in the empirical analysis: 0.421 (α), 0.276 (β), −0.082 (γ). *Density* is defined as the population density based on administrative district, and *new building* is represented by the proportion of new buildings, that is, the ratio of buildings that were constructed within the past 1 year (or 5 years). *Housing ownership*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

is defined as the ratio of owner-occupied housing. These explanatory variables were selected to reflect the influence of residential environments and housing conditions on the relocation distance of households. In addition, two variables related to the availability of metropolitan transportation were selected in this study. *Rail availability* is represented by the ratio of the catchment area within 500 m from the metropolitan railway stations, and *bus availability* is defined as the number of metropolitan bus routes operating in each administrative district.

4.3. Descriptive Statistics

This empirical analysis contains 209,252 residential movement data samples which is randomly sampled 10% of the raw data including the householder information. The descriptive statistics for the selected and estimated variables are listed in Table 4.

In the dataset, the average moving distance of households is 9.12 km, and the range of distance is 0.24–267.31 km. Regarding the household attributes, 19% of the entire residential movements were caused by a job. In addition, 60% and 2% of the residential relocations were due to housing replacement and educational environment, respectively. Although the reasons for the residential movements were numerous, housing replacement accounted for more than half. The average age of householders was approximately 44.32, and the dataset consisted of 66% male and 34% female population. The number of household members ranged from 1 to 9, with an average value of 2.1. On average, the households included 0.14 elderly people, 0.12 primary school-aged children, and 0.14 secondary school-aged children. The proportion of males among household members was 53%.

Table 4. Descriptive Statistics

	Variable	Unit	Average	SD	Minimum	Maximum
-	Relocation distance	km	9.12	13.66	0.24	267.31
Household attributes	Moving reason: Job [1]	-	0.19	0.39	0.00	1.00
	Moving reason: House [1]	-	0.60	0.49	0.00	1.00
	Moving reason: Education [1]	-	0.02	0.14	0.00	1.00
	Age	Years	44.32	13.79	0.00	103.00
	Gender: Male [1]	-	0.66	0.47	0.00	1.00
	Members	People	2.10	1.30	1.00	9.00
	Elderly people	People	0.14	0.41	0.00	4.00
	Children: Primary	People	0.12	0.40	0.00	4.00
	Children: Secondary	People	0.14	0.42	0.00	7.00
	Proportion of males	%	53.52	38.27	0.00	100.00
Location characteristics in origin	Accessibility	-	14.26	0.53	6.40	14.79
	Density	People/ha	174.25	129.14	0.00	550.00
	New building: 1 year	%	2.93	2.46	0.27	17.36
	New building: 5 years	%	13.69	6.12	1.83	34.89
	Housing ownership	%	48.55	8.45	29.38	79.26
	Rail availability	%	25.39	27.76	0.00	100.00
	Bus availability	EA	7.54	10.11	0.00	71.00
Location characteristics in destination	Accessibility	-	14.23	0.54	6.40	14.79
	Density	People/ha	167.34	129.29	0.00	550.00
	New building: 1 year	%	3.11	2.74	0.27	17.36
	New building: 5 years	%	14.04	6.26	1.83	34.89
	Housing ownership	%	48.81	8.39	29.38	79.26
	Rail availability	%	24.46	27.58	0.00	100.00
	Bus availability	EA	7.66	10.28	0.00	71.00

[1] A reference of nominal variables.

As the residential relocation of household has a departure point and an arrival point, the location characteristics were classified into not only the origin, but also the destination domains. As the

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

location characteristics of administrative districts are assigned to individual households, the minimum and maximum values of characteristics at the origin and destination are the same. In contrast, the differences in the averages and standard deviations are due to the number of households included in each administrative district. The average values of accessibilities to origin and destination were 14.26 and 14.23, respectively. In addition, the population density at the origin location (174.25 people/ha) was larger than that of the destination (167.34 people/ha). These results indicate that the households moved out to less densely-populated districts. At the origin location, the proportion of newly-constructed buildings within a year was 2.93% and that within five years was 13.69%. Moreover, at the destination location, the proportion of newly constructed buildings within a year was 3.11% and that within five years was 14.04%. These outcomes imply that the households moved out to the districts with more new buildings in 2015. The ratio of rail catchment area in the origin districts was 25.39% on average, which is larger than that in the destination districts (24.46%). Moreover, the average number of bus routes was 7.66 at the destination and 7.54 at the origin location.

5. Results and Discussion

The analytical dataset was composed of 209,252 samples of residential households that moved in 2015. In a machine learning approach, the analytical dataset is randomly split into training and testing subsets. Generally, the former consists of 75% of the entire dataset and the latter consists of the remaining 25%.

5.1. Comparison of the Empirical Results between Ordinary Least Squares and Decision Tree Regressions

In this study, the empirical analysis on residential relocation distance in the SMR included the application of ordinary least squares regression and decision tree regression using a machine learning approach. The results of the empirical analysis are summarised in Table 5.

First, the training and test R-squared values in the ordinary least squares regression model were 0.180 and 0.190, respectively, showing low explanatory power. In the household attributes domain, among the residential moving reasons, *house* was an influencing factor that shortened the moving distance of households by about 2 km compared to other reasons. This can be interpreted as a result of the existence and influence of the housing sub-market in the SMR. On the other hand, *job* and *education* were significant factors—these were significant factors affecting residential mobility in previous studies [7,45]—in increasing the distance of residential movement of households over 5 km compared to other causes. *Age* and *squared age* were significant variables, and the residential relocation distance of households was the minimum at the householder age of approximately 59, which is similar to the residential mobility of the life-cycle model. For the explanatory variables that represent composition of a household, the number of household members and the number of children had negative coefficients at the 99% level. These outcomes are similar to previous literature related to mobility based on residential duration [46], which can be understood that households with more members have more complex decision-making system for their residential relocation and there is a tendency to maintain their community that was formed in the previous location. Whereas, *gender* was a positive determinant at the 99% level. This result can be interpreted as the relatively low resistance to residential moving distance in the households with a male householder or the long-distance residential movements due to changes in the workplace of the male householder.

In the location characteristics domain, the most important explanatory variable was *accessibility* to employment markets in the both the origin and destination residential locations. *Accessibility* variables had negative coefficients, which implies the importance of proximity to employment centers affecting residential location choice of household in previous studies [7,47–49]. *Density* and proportions of *new buildings* within one year or five years also had significant coefficients, but their signs showed opposite values in origin and destination locations of the residential movements. High population density is considered as a negative determinant in residential environment [50], whereas newly constructed houses are seen as a positive one. Since the former and the latter are a push factor and a pull factor [51], respectively, the difference of the distance, as well as the migration flow of intra-urban residential mobility can be generated. *Housing ownership* had negative coefficients in both

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

the origin and destination locations. These results can be interpreted as a relatively short movement of residents living in the stabilised settlements based on the high proportion of housing ownership. Moreover, the coefficient of *bus availability*, which is the number of inter-regional bus routes by administrative district, showed a significant positive sign in only destination residential location. This outcome means that even though it is located far away, the district with a large number of bus routes with relatively high inter-regional mobility can be a residential moving destination.

Table 5. Results of the Empirical Analysis Using Machine Learning Models

Variable (Feature)			Ordinary Least Squares Regression			Decision Tree Regression	
			β	Std. β	Sig.	Importance	Rank
	(Constant)		136.3587		0.000 **		
Household attributes	X(0)	Moving reason: Job	5.6836	2.2401	0.000 **	0.13180	3
	X(1)	Moving reason: House	-1.9648	-0.9630	0.000 **	-	-
	X(2)	Moving reason: Education	5.4827	0.7688	0.000 **	0.00289	8
	X(3)	Age	-0.2362	-3.2602	0.000 **	0.00114	9
	X(4)	Squared Age	0.0020	2.7813	0.000 **	-	-
	X(5)	Gender: Male	0.5935	0.2809	0.000 **	-	-
	X(6)	Members	-1.0025	-1.3052	0.000 **	0.01246	6
	X(7)	Elderly people	0.1631	0.0668	0.140	-	-
	X(8)	Children: Primary	-0.5795	-0.2338	0.000 **	-	-
	X(9)	Children: Secondary	-0.8717	-0.3697	0.000 **	-	-
	X(10)	Proportion of males	0.0019	0.0739	0.154	-	-
Location characteristics in origin	X(11)	Accessibility	-2.0368	-1.0766	0.000 **	0.57976	1
	X(12)	Density	0.0008	0.1081	0.015 *	0.01450	5
	X(13)	New building: 1 year	-0.1787	-0.4411	0.000 **	-	-
	X(14)	New building: 5 years	-0.0461	-0.2824	0.000 **	0.00434	7
	X(15)	Housing ownership	-0.0417	-0.3523	0.000 **	0.00001	12
	X(16)	Rail availability	0.0014	0.0392	0.352	-	-
	X(17)	Bus availability	0.0055	0.0553	0.138	-	-
Location characteristics in destination	X(18)	Accessibility	-6.1138	-3.2953	0.000 **	0.23433	2
	X(19)	Density	-0.0026	-0.3358	0.000 **	0.01749	4
	X(20)	New building: 1 year	0.1147	0.3156	0.000 **	-	-
	X(21)	New building: 5 years	0.0434	0.2719	0.000 **	-	-
	X(22)	Housing ownership	-0.0271	-0.2277	0.000 **	0.00039	11
	X(23)	Rail availability	0.0019	0.0526	0.219	-	-
	X(24)	Bus availability	0.0434	0.4457	0.000 **	0.00090	10
Explanatory Power			Training R^2: 0.180 Test R^2: 0.190			Training R^2: 0.512 Test R^2: 0.504	

* p-value < .05; ** p-value < .01.

Second, in decision trees, the complex tree constructed using the training dataset generally has an overfitting problem. Therefore, by setting the parameters for maximum depth and the leaf node minimum sample value, an early stopping method was applied to terminate the learning algorithm before the tree became too complex [52]. The application of early stopping has advantages of not only mitigating the overfitting problem, but also interpreting the derived tree structure. A trial and error method was applied to set the appropriate parameter values: the maximum depth is 6, and the leaf node minimum sample value is 10 (refer to Appendix A).

In the model applying decision tree regression, the explanatory powers of the final derived model showed a remarkable improvement over the ordinary least squares regression model. The training R-squared value was 0.512, and the test one was 0.504. Twelve features were contained in the derived decision tree. The importance of features reflects the contribution each variable makes in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

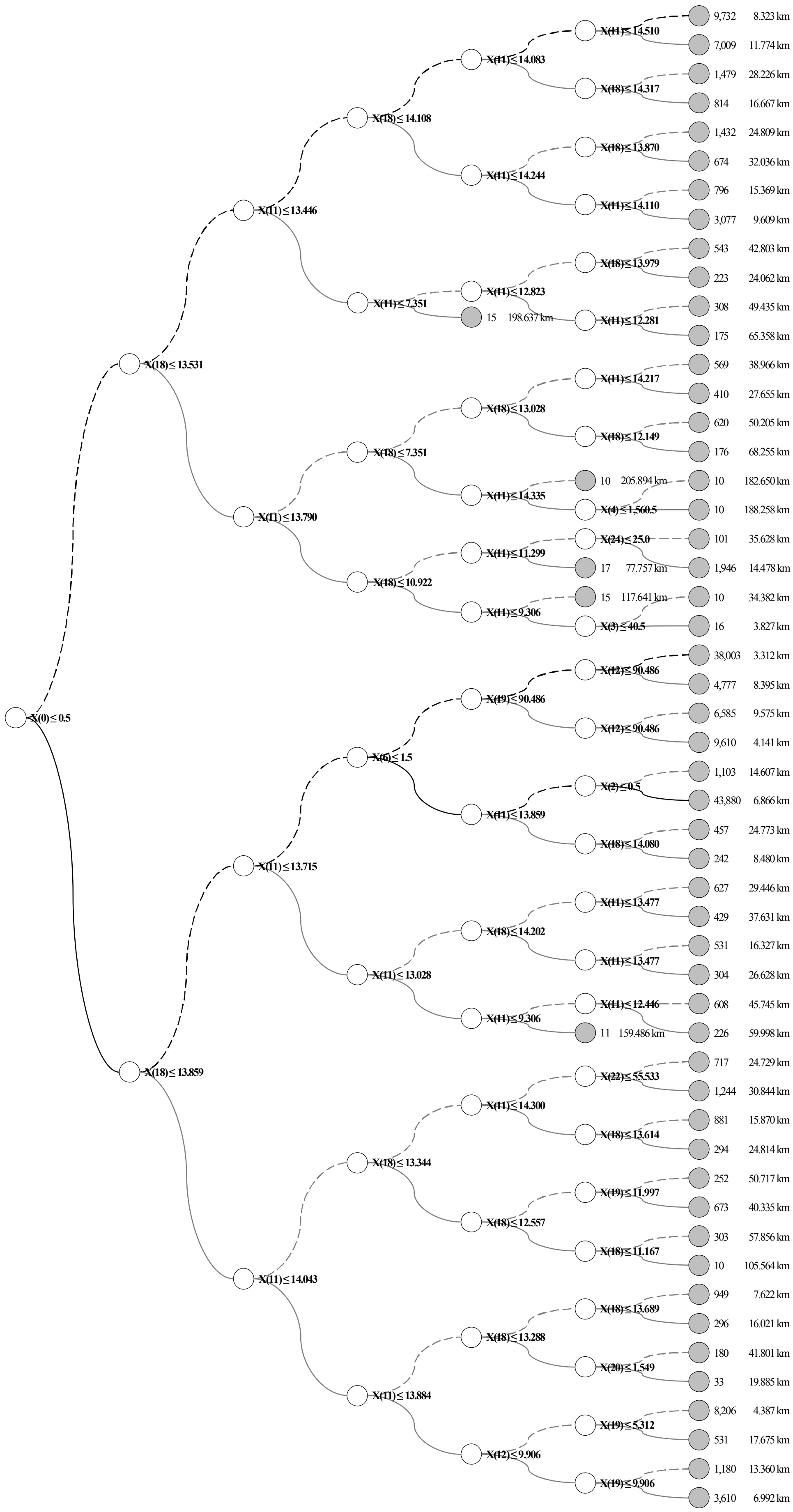
estimating the target variable, which is the residential relocation distance of each household in this study. The importance of a feature is estimated as the normalised total reduction of the criterion caused by the feature. In Table 5, two of the most important features were *accessibility* to employment markets in the locations before and after the residential movement. Among the residential moving reasons, *job* was ranked as the third most important feature. The importance of these three features accounted for approximately 95% of the total importance. In addition, in terms of importance, the following features were ranked: *density* of population in destination and origin locations, *members*, *new buildings* within five years in origin residential location, *moving reason: education*, *age* of householder, *bus availability* in destination, and *housing ownership* in destination and origin residential locations. These importance values are different from the standardised beta values that indicate the relative influence of explanatory variables on the results of the ordinary least squares regression model.

Decision trees, while not as powerful from a pure machine learning standpoint, are still one of the canonical examples of an understandable machine learning algorithm. That is, the structure of the derived decision tree can be represented, as shown in Figure 3. In the figure, the gray circle indicates a leaf node which is composed of 57 nodes, and intermediate nodes are represented using 56 white circles. Among them, the leftmost white circle is called the root node. In this study, the derived decision tree structure can be traced back to the splits from the training dataset starting with 156,939 samples at the root node. Moreover, in the tree structure, the solid lines mean that an observation goes to the lower branch if the condition shown at the intermediate node is satisfied, whereas the broken lines indicate that an observation goes to the upper branch if it is not satisfied. The equation presented on the right side of the intermediate node is a condition splitting the assigned samples of each node. Of the two numbers located to the right of the leaf nodes, the first and second numbers are the number of samples and the average relocation distance of the assigned samples to the nodes.

The black solid and broken lines represent branch paths that can reach the leaf nodes to which the top three most-allocated samples are assigned. In a decision tree model, describing the entire tree structure is not only extremely complex, but also inefficient. Therefore, the most important top three assigned leaf nodes and their assigned paths are described in this paper. First, the leaf node with the largest number of samples contains 43,880 households (27.96%) with an average residential moving distance of 6.866 km. The features affecting the path of branches to the leaf node were residential mobility caused by factors other than job or education (X(0) and X(2)), higher potential accessibility to employment markets from origin to destination residential location (X(11) and X(18)), and one-person household (X(6)). This result can be summarised as the pattern of general residential movements based on the employment market in the one-person household group. Second, the leaf node with the second largest number of samples includes 38,003 households (24.22%), with an average residential relocation distance of 3.312 km, which is the shortest distance leaf node in this derived tree. The features related to the leaf node were residential mobility caused by factors other than job (X(0)), higher potential accessibility to employment markets from origin to destination residential location (X(11) and X(18)), densely populated origin and destination locations (X(12) and X(19)), and households with more than two people (X(6)). This path can be understood as the shortest residential moving pattern of households with more than two members based on accessibility to employment markets, which are carried out among densely populated districts, for purposes other than a job. Finally, the path related to the third largest leaf node includes 9,732 households (6.20%) with an average moving distance of 8.323 km, which were affected by the features including residential mobility caused by job (X(0)) and lower potential accessibility to employment markets from origin to destination location (X(11) and X(18)). This result can be interpreted as the relatively longer residential moving distance of households caused by job or employment except other residential conditions. In addition, Figure 3 describes that there are a lot of paths based on the decision tree of households related to the residential relocation distance in the SMR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996



409 **Figure 3.** Decision Tree for Residential Relocation Distance of the Households in SMR

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

5.2. Application of Ordinary Least Squares Regression and Decision Tree Regression Models

This study focuses on not only the identification of the features and their structures affecting residential relocation distance but also on the applicability of the machine learning approach to residential mobile pattern analysis. Therefore, the application results of the previously constructed regression models and the actual moving distance values were directly compared.

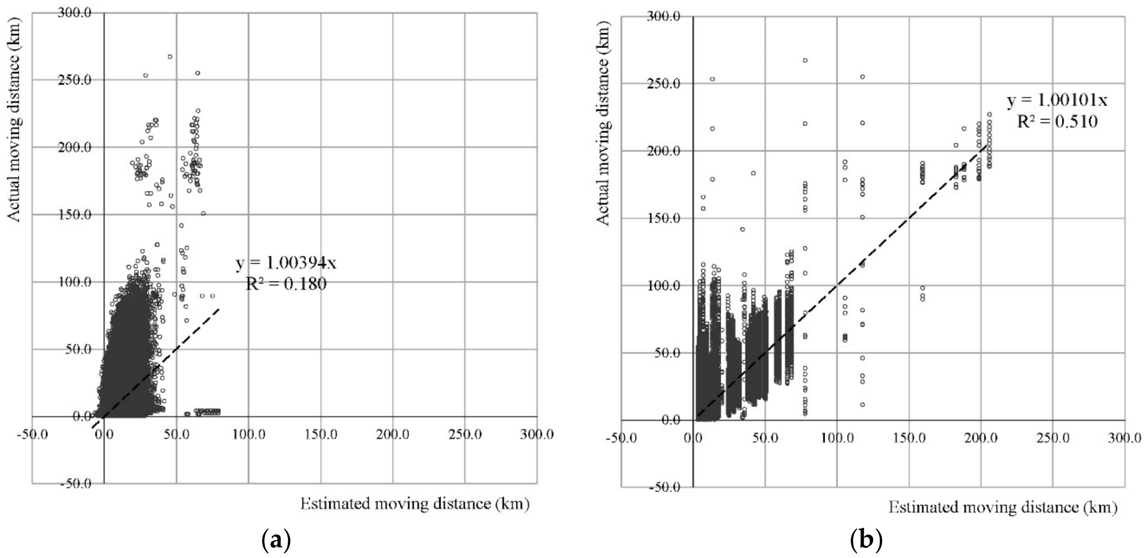


Figure 4. (**a**) Result of Applying the Ordinary Least Squares Regression Model; (**b**) Result of Applying the Decision Tree Regression Model

Figure 4 shows the comparison between the application results. The figure on the left is the result of comparing the actual moving distances of all samples and the estimated moving distance using the ordinary least squares regression model. The figure on the right is the comparison of the actual distance and the estimated distance by the decision tree regression model. As expected, the decision tree regression model results were relatively better. The application of the ordinary least squares regression model showed a large number of underestimated values and a large number of unrealistic residential moving distances, such as values less than zero. Whereas, the results of the decision tree application presented relatively few underestimated values, and there were no unrealistic estimates of the residential relocation distance. Thus, in the latter, the regression coefficient (1.00101) and the R-squared value (0.510) were also better.

6. Summary and Concluding Remarks

In the rapidly changing Korean housing market, from both supply and demand perspectives, understanding the spatial patterns of residential relocation is a meaningful task. This paper focused on the structure among determinants affecting residential relocation distance and the applicability of a new approach using spatial Big Data and a machine learning methodology. The results of the empirical analysis on residential relocation distance in the SMR by using ordinary least squares and decision tree regressions can be summarised as follows.

In terms of explanatory power, the decision tree regression model showed better performance than the ordinary least squares regression model. Twenty variables were significant in the ordinary least squares regression, whereas only twelve features were applied in the decision tree regression model, although the model had relatively complicated structures. As a result of the ordinary least squares regression, residential movements for housing-related reasons were shorter than the distance of residential movements caused by job or education. Households with a householder over 60-years-old or male householders showed longer residential relocation distance. On the other hand, households with a householder less than 60-years-old, households with multiple members, and households with school-aged children moved to a relatively close residential districts. In terms of the location characteristics in the origin and destination, accessibility to employment markets and housing ownership were the factors that shortened the household residential relocation distance. In the origin, the high population density led to longer residential movements, and the variables associated with the proportion of new buildings were factors that shortened the residential moving distance. However, those in the destination had the opposite effects.

To summarise the main outcomes of the decision tree regression, the most important features that determined the residential relocation distance were migration caused by a job and accessibility to employment markets, although a large number of residential relocations occurred for reasons other than a job. Additionally, this empirical study showed many residential movements to the districts with good access to employment. The shortest moving distance was found when the household with more than two people moved among densely-populated districts, whereas residential movements caused by job had a relatively longer moving distance.

Moreover, the ordinary least squares regression and the decision tree regression models were applied to compare their estimated values and the actual measurements based on the geographic data using the microdata of the Internal Migration Statistics. The estimated distances using the decision tree regression model were more realistic, with the estimated moving distances not containing values less than zero and there were few underestimated values. Its explanatory power was higher than that of the ordinary least squares regression model.

Thus, this study reviewed the applicability of the machine learning method using spatial Big Data, which is a focus in the urban planning and management field. In particular, this article attempted to overcome the limitation of conventional statistical models—low explanatory power and a lot of rigid constraints—using an interpretable and understandable machine learning model, the decision tree regression model. The results of this study have the following implications. First, the result of the decision tree regression model (the training R-squared: 0.512) showed a significant improvement in the explanatory power compared to that of the ordinary least squares regression model (the training R-squared: 0.180), which is similar to a conventional linear regression model. Second, the derived decision tree presented not only the diversity of structures that determine the residential relocation distance, but also the main features, such as movement caused by jobs and accessibility to employment markets, which form the structures. Finally, for the residential moving pattern, we found that the machine learning approach, such as decision trees, can estimate more realistic results than conventional methodologies.

The development of the forecasting model beyond the empirical analysis of the decision structures for the residential relocation distance and the inclusion of several explanatory variables that were not contained in the model require further research. In spite of these future tasks, this study presents a case using the machine learning approach with spatial bigdata in the urban planning and management field. Moreover, the outcomes of this research provide significant information about the sustainable urban management of metropolitan residential districts and the construction of reasonable housing policies, and it is expected to be the basis of further studies on spatial patterns of residential relocation in the future.

Acknowledgments: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03930624).

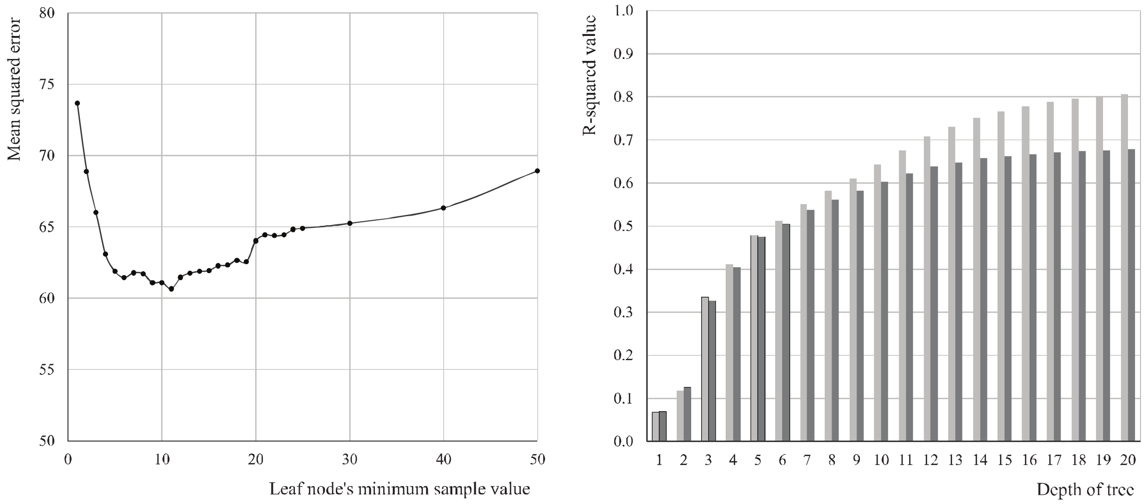
Appendix A

Regarding the selection of the appropriate parameters of the decision tree for controlling the overfitting problem, a trial and error method was applied in this study. In Figure A, the figure on the left represents the variation of the mean squared error value (MSE) according to the leaf node minimum sample value. The variation in the MSEs according to the leaf node minimum sample value without setting the depth of the tree shows the lowest level from 9 to 12 samples. The figure on the right indicates the variation of the R-squared values according to depth of tree after the leaf node minimum sample value is set to 10. The R-squared values, which mean the explanatory powers of the models applied to the training dataset and the test dataset, show the gaps of less than 1% up to a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 July 2018 doi:10.20944/preprints201807.0409.v1

Peer-reviewed version available at Sustainability 2018, 10, 2996; doi:10.3390/su10092996

depth of 6. Therefore, the appropriate parameter values of this decision tree regression model were selected as follows: the leaf node's minimum sample value is 10 and the maximum depth is 6.



(**a**) (**b**)

Figure A. (**a**) Variation of the mean squared errors according to the leaf node minimum sample value; (**b**) Variation of the R-squared values according to depth of tree

References

1. Seo, D., Kwon, Y. In-migration and housing choice in Ho Chi Minh City: Toward sustainable housing development in Vietnam. *Sustainability* **2017**, *9*, 1–17, doi:10.3390/su9101738.
2. Rossi, P. H. *Why families move: A study in the social psychology of urban residential mobility*; Free Press: New York, 1955;
3. Hoyt, H. *The Structure and Growth of Residential Neighborhoods in American Cities*; Washington, 1939;
4. Ravenstein, E. G. The Laws of Migration. *Journal of the Statistical Society of London* **1885**, *48*, 167–235.
5. Park, E. .; Lee, J. W. A study on policy literacy and public attitudes toward government innovation-focusing on Government 3.0 in South Korea. *Journal of Open Innovation: Technology, Market, and Complexity* **2015**, *1*, 1–13, doi:10.1186/s40852-015-0027-3.
6. Nho, H. J. Research ethics education in Korea for overcoming culture and value system differences. *Journal of Open Innovation: Technology, Market, and Complexity* **2016**, *2*, 1–11, doi:10.1186/s40852-016-0030-3.
7. Yi, C.; Lee, S. An empirical analysis of the characteristics of residential location choice in the rapidly changing Korean housing market. *Cities* **2014**, *39*, 156–163, doi:10.1016/j.cities.2014.03.002.
8. OECD *OECD Regions at a Glance 2016*; 2016; ISBN 9789264252097.
9. Kwon, W. Y. An examination of residential location behavior in the Seoul metropolitan area. *The Annals of Regional Science* **1984**, *18*, 33–48.
10. The Seoul Institute. *Seoul Statistical Series Section 04. Housing*; Seoul, 2014;
11. Hong, S., Lee, Y. Limitation of Residential Mobility Distance in Seoul Metropolitan Area -Focused on Migration Region and Family Size-. *Korea Real Estate Academy Review* **2015**, *60*, 115–126.
12. Brown, L. A.; Moore, E. G. The inter-urban migration process: A perspective. *Geografiska Annaler Series B, Human Geography* **1970**, *52*, 1–13.
13. Chun, H. S. The characteristics of housing mobility of the residents' in new town areas. *Gyeonggi Forum* **2004**, *6*, 91–111.

14. Wu, F. Intraurban residential relocation in Shanghai: Modes and stratification. *Environment and Planning A* **2004**, *36*, 7–25, doi:10.1068/a35177.
15. Moore, E. G. *Residential mobility in the city*; 1970;
16. Loren, D. L.; Eva, K.; Boaz, K. Residential relocation of amenity migrants to Florida: “Unpacking” post-amenity moves. *Journal of Aging and Health* **2010**, *22*, 1001–1028, doi:10.1177/0898264310374507.
17. Lee, B. H. Y.; Paul, W. Residential mobility and location choice: A nested logit model with sampling of alternatives. *Transportation* **2010**, *37*, 587–601, doi:10.1007/s11116-010-9270-4.
18. Stapleton, C. M. Reformulation of the family life-cycle concept: implications for residential mobility. *Environment and Planning A* **1980**, *12*, 1103–1118, doi:10.1068/a121103.
19. Pickvance, C. G. Life Cycle, Housing Tenure and Residential Mobility: A Path Analytic Approach. *Urban Studies* **1974**, *11*, 171–188, doi:10.1080/00420987420080331.
20. Morris, E. W.; Crull, S. R.; Winter, M. Housing Norms , Housing Satisfaction and the Propensity to Move. *Journal of Marriage and Family* **1976**, *38*, 309–320.
21. Chevan, A. Family growth, household density, and moving. *Demography* **1971**, *8*, 451–458, doi:10.2307/2060682.
22. Eluru, N.; Sener, I. N.; Bhat, C. R.; Pendyala, R. M.; Axhausen, K. W. Understanding Residential Mobility: Joint Model of the Reason for Residential Relocation and Stay Duration. *Transportation Research Record* **2009**, *2133*, 64–74, doi:10.3141/2133-07.
23. Zanganeh, Y.; Hamidian, A.; Karimi, H. The Analysis of Factors Affecting the Residential Mobility of Afghan Immigrants Residing in Mashhad (Case Study : Municipality Regions 4 , 5 and 6). *Asian Social Science* **2016**, *12*, 61–69, doi:10.5539/ass.v12n6p61.
24. Ha, S. K. *Housing policy and practice in Korea*; 3rd ed.; Pakyoungsa: Seoul, 2006;
25. Min, B.; Byun, M. Residential Mobility of the Population of Seoul: Spatial Analysis and the Classification of Residential Mobiltiy. *Seoul Studies* **2017**, *18*, 850102.
26. Clark, W. A. V.; Onaka, J. L. Life Cycle and Housing Adjustment as Explanations of Residential Mobility. *Urban Studies* **1983**, *20*, 47–57.
27. Yi, C.; Lee, S. I. Analyzing the Factors on Residential Mobility According to the Household Member’s Change - In consideration of residential duration of the households in the Seoul Metropolitan Area. *Journal of Korea Planning Association* **2012**, *47*, 205–217.
28. Yee, W.; Van Arsdol, M. D. J. Residential Mobility , Age , and the Life Cycle 1. *Journal of Gerontology* **1977**, *32*, 211–221.
29. Clark, W. A. V. Life course events and residential change : unpacking age effects on the probability of moving. *Journal of Population Research* **2013**, *30*, 319–334, doi:10.1007/s12546-013-9116-y.
30. Burnley, I. H.; Murphy, P. A.; Jenner, A. Selecting Suburbia : Residential Relocation to Outer Sydney. *Urban Studies* **1997**, *34*, 1109–1127.
31. Yang, J. Transportation Implications of Land Development in a Transitional Economy: Evidence from Housing Relocation in Beijing. *Transportation Research Record* **1996**, *1954*, 7–14.
32. Clark, W. A. V.; Dieleman, F. M. *Households and housing: choice and outcomes in the housing market*; New Brunswick: New Jersey, 1996;
33. Dieleman, F. M. Modelling residential mobility ; a review of recent trends in research. *Journal of Housing and the Built Environment* **2001**, *16*, 249–265.
34. Shalev-Shwartz, S.; Ben-david, S. *Understanding Machine Learning : From Theory to Algorithms*; 1st ed.; Cambridge University Press: New York, 2014; ISBN 9781107057135.

35. Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; 2nd ed.; Morgan Kaufmann: San Francisco, 2005; ISBN 0120884070.
36. Zhang, H.; Wu, P.; Yin, A.; Yang, X.; Zhang, M.; Gao, C. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Science of the Total Environment* **2017**, *592*, 704–713, doi:10.1016/j.scitotenv.2017.02.146.
37. Estelles-lopez, L.; Ropodi, A.; Pavlidis, D.; Fotopoulou, J.; Gkousari, C.; Peyrodie, A.; Panagou, E.; Nychas, G. J.; Mohareb, F. An automated ranking platform for machine learning regression models for meat spoilage prediction using multi-spectral imaging and metabolic profiling. *Food Research International* **2017**, *99*, 206–215, doi:10.1016/j.foodres.2017.05.013.
38. Chagas, C.; Junior, W.; Bhering, S.; Filho, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* **2016**, *139*, 232–240, doi:10.1016/j.catena.2016.01.001.
39. Oliveira, M.; Gama, J. An overview of social network analysis. *WIREs: Data Mining and Knowledge Discovery* **2012**, *2*, 99–105.
40. Tso, G. K. F.; Yau, K. K. W. Predicting electricity energy consumption: A comparison of regression analysis , decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768, doi:10.1016/j.energy.2006.11.010.
41. Xu, M.; Watanachaturaporn, P.; Varshney, P. K.; Arora, M. K. Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment* **2005**, *97*, 322–336, doi:10.1016/j.rse.2005.05.008.
42. Prasad, A. M.; Iverson, L. R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181–199, doi:10.1007/s10021-005-0054-1.
43. Hansen, W. G. How accessibility shapes land use. *Journal of the American Institute of Planners* **1959**, *25*, 73–76.
44. Wilson, A. G. *Entropy in Urban and Regional Modeling*; Pion: London, 1970;
45. Yi, Y.; Kim, E.; Choi, E. Linkage among school performance, housing prices, and residential mobility. *Sustainability (Switzerland)* **2017**, *9*, 1–18, doi:10.3390/su9061075.
46. Choi, Y.; Yim, H. K. Determinants of the Residents' Settlements Employing Poission Regression. *The Korea Spatial Planning Review* **2005**, *46*, 99–114.
47. Pagliara, F.; Simmonds, D. Conclusions. In *Residential location choice-models and applications*; Pagliara, F., Preston, J., Simmonds, D., Eds.; Springer, Verlag: Berlin, Heidelberg, 2010; pp. 243–248.
48. Simmonds, D. The DELTA residential location model. In *Residential location choice-models and applications2*; Pagliara, F., Preston, J., Simmonds, D., Eds.; Springer, Verlag: Berlin, Heidelberg, 2010; pp. 77–97.
49. Waddell, P. Modelling residential location in UrbanSim. In *Residential location choice-models and applications*; Pagliara, F., Preston, J., Simmonds, D., Eds.; Springer, Verlag: Berlin, Heidelberg, 2010; pp. 165–180.
50. Cadwallader, M. *Migration and Residential Mobility: Macro and Micro Approaches*; University of Wisconsin Press: Madision, 1992;
51. Cluttons *Residential Mobility in London: Unlocking Migration Patterns*; 2017;

52. Mueller, A. C.; Guido, S. *Introduction to Machine learning with python*; O'Reilly: Sebastopol, 2016; ISBN 9781491917213.