Małgorzata Renigier-Biłozor

Supplementing incomplete databases on the real estate market with the use of the rough set theory

Acta Scientiarum Polonorum. Administratio Locorum 9/3, 107-115

2010

ACTA SCIENTIARUM POLONORUM
Acta Sci. Pol., Administratio Locorum 9(3) 2010, 107–115

SUPPLEMENTING INCOMPLETE DATABASES ON THE REAL ESTATE MARKET WITH THE USE OF THE ROUGH SET THEORY

Małgorzata Renigier-Biłozor
University of Warmia and Mazury in Olsztyn

Abstract. This paper investigates the use of Rough Set Theory for supplementing databases on the real estate market. The proposed simplified procedure may pose an alternative for statistical methods, and it produces reliable results over a short period of time. The procedure of supplementing incomplete data has been developed based on the principles of Rough Set Theory and the valued tolerance relation. The above combination produces optimal results because it accounts for varied methods of entering property attributes.

Key words: rough set theory, real estate market, supplementing incomplete databases

INTRODUCTION

The real estate market is a highly complex system, and the selection of the appropriate analytic methods and procedures often poses a problem. The property market is characterized by numerous attributes, including varied quantities of data subject to the type of the analyzed market (region), complex data description methods (choice of various scales for entering attributes) as the same attribute can be described in a number of ways on grading scales with a different spread of rating points, a property's unique characteristics (no two properties are identical), its multi--criteria designation (every property can be used and managed in a variety of ways), incomplete information (the absence of standardized property data systems generates limited and incomplete information on the property and its attributes), imprecise and "fuzzy" property information (resulting from, among others, stochastic factors which are an expression of random processes that do not conform to the generally observed cause-and-effect relationships on the market), various functional dependencies

Corresponding author – Adres do korespondencji: Małgorzata Renigier-Biłozor, Katedra Gospodarki Nieruchomościami i Rozwoju Regionalnego, Uniwersytet Warmińsko-Mazurski w Olsztynie, ul. Prawocheńskiego 15, 10-720 Olsztyn, e-mail: malgorzata.reniger@uwm.edu.pl

between property attributes and the decision attribute represented by the property's value, function or management method.

The decision-making process in property management, including on the real estate market, is subject to limitations imposed by the analytical methods applied to determine optimal land use and property value, as well as by the quality of the available information. Owing to a broad variety of attributes (property features) and data, the process of evaluating and forecasting real estate value and planning property use and management is complex, time consuming and burdened with high risk. If decision uncertainty results from various elements of the decision-making process relating to, for example, data imprecision (data spreads, measurement errors), uncertainty (whether data are correct), lack of knowledge (lack of awareness that the relevant data exist) and incompleteness (absence of data), the preferred solution would be to deploy analytical methods based on artificial intelligence theories, in this case – the rough set theory. The rough set theory, developed by a Polish professor of information technology, Zdzisław Pawlak, is used to investigate imprecision, generalization and uncertainty in data analysis, i.e. a common set of qualities on the real estate market.

Although developed relatively recently, the rough set theory has found many applications in a large number of scientific disciplines [including in the works of: Deja 2000, Komorowski et al. 1999, Mrózek i Płonka 1999, Nowicki 2009, Polkowski and Skowron 1998a, 1998b, Pawlak 1997, Słowiński 1992]. The following authors have relied on the rough set theory in their studies of property management and the real estate market: d'Amato [2004, 2006, 2007, 2008], Kotkowski and Ratajczak [2002], Renigier [2006], Renigier-Biłozor [2008a, 2008b, 2009a, 2009b], Renigier-Biłozor and Biłozor [2007, 2008].

The classical rough set theory is used to supplement incomplete data [Adamus 2008, Stefanowski 2001], but in view of the specific nature of data on the real estate market, the knowledge generated by the rough set theory and the fuzzy set theory (valued tolerance relation) delivers the most satisfactory results.

SUPPLEMENTING DATA IN DECISION TABLES ON THE REAL ESTATE MARKET WITH THE USE OF THE ROUGH SET THEORY

There are many reasons for the presence of incomplete data on the real estate market, including technical, accidental or planned factors. Incomplete data may be handled in a variety of ways, including by:

- removing properties with incomplete data from the database,
- applying methods that tolerate "defective" data,
- supplementing data.

This paper focuses on the last variant, mainly the supplementation of data with the involvement of a procedure that relies on the rough set theory and the valued tolerance relation. In the proposed solution, data are supplemented in reference to an existing dataset.

The method of supplementing data with the use of the rough set theory has been presented on a randomly selected set of 10 apartment sale transactions conducted in Olsztyn in 2009. In the decision table (Table 1), the features of the analyzed properties are marked successively c_1, c_2, c_3, c_4, (Table 2) as conditional attributes, and property price d is the decision attribute. The decision table contains four properties, no. 1, 3, 6 and 10, with incomplete data.

Table 1. Decision table for transacted property
Tabela 1. Tablica decyzyjna nieruchomości transakcyjnych

No. Lp.	Price Cena	Usable area Pow. użytkowa	Standard	Storey Piętro	Location Lokalizacja
1	6100	50	1	**x**	**x**
2	5710	35	1	3	2
3	5833	**x**	1	3	1
4	6600	25	1	2	2
5	4319	60	2	2	2
6	4870	92	**x**	3	3
7	6006	50	1	2	2
8	4250	88	1	3	3
9	4958	78	1	3	1
10	4485	52	**x**	2	2

Table 2. List of analyzed attributes
Tabela 2. Zestawienie atrybutów przyjętych do badań

Conditional attributes Atrybuty warunkowe				Decision attribute Atrybut decyzyjny
c_1	c_2	c_3	c_4	d
Usable floor area Powierzchnia użytkowa	Standard Standard	Storey Położenie na piętrze	Location Lokalizacja	Price Cena

Source: own study
Źródło: opracowanie własne

Every attribute was assigned a domain in line with the preset requirements:
c_1 – property's usable floor area – in m^2
c_2 – standard: 1 – high, 2 – average, 3 – low
c_3 – storey: 1 – 1st floor, 2 – 2nd and 3rd floor, 3 – ground floor and above 3rd floor
c_4 – property location coded according to the following criteria: 1 – prime, 2 – average, 3 – poor
d – property price in PLN/m^2

Following the determination of attribute domains, the values of property attributes were grouped based on their degree of indiscernibility, in accordance with the rough set theory [Pawlak 1982, 1991]. During an analysis of a unique set of property data with various scales (including the ratio scale, ordinal scale, interval scale and nominal scale) for describing property attributes, the classic rough set theory has been enhanced with the valued tolerance relation formula. This

formula has been developed and discussed by Stefanowski and Tsoukias [2000] and Stefanowski [2001], and it was deployed in real estate market analyses by d'Amato [2006, 2007, 2008], Renigier-Biłozor [2008a, 2008b, 2009a, 2009b] and Renigier-Biłozor, Biłozor [2007, 2008].

A classical rough set theory relies on the indiscernibility relation concept as a crisp equivalence relation, namely that two properties will be indiscernible only if they have identical attributes. By introducing a valued tolerance relation into the rough set theory, the upper and lower approximation of the dataset can be determined with different degrees of indiscernibility. The above relation can be formally expressed with the below equation:

$$R_j(x, y) = \frac{\max(0, \min c_j(x), c_j(y)) + k - \max(c_j(x), c_j(y)))}{k} \quad (1)$$

where:

$R_j(x, y)$ – relation between two sets with membership function [0,1]
$c_j(x)$, $c_j(y)$ – variable of the analyzed property
k – coefficient adopted for a given property attribute

The above formula is used to compare two sets of data, in this case – two properties, and the obtained result in the 0–1 range determines the degree of indiscernibility. If coefficient k represents standard deviation for various attributes of the analyzed set, as per Table 3 (alternatively, standard deviation can be adopted for a collection of universal data for the analyzed set, e.g. a set of transactions conducted throughout the entire real estate market over a longer period of time), similarity (indiscernibility) matrices relative to coefficient k are identified separately for every property attribute. A sample matrix for the usable floor area attribute is displayed in Table 4.

Table 3. Coefficient k
Tabela 3. Wyznaczony współczynnik k

Conditional attribute Atrybut warunkowy	d	c_1	c_2	c_3	c_4
Coefficient k Współczynnik k	837	23	0.74	0.73	0.70

Source: own study
Źródło: opracowanie własne

In the next step of the procedure, the results produced by the above matrix were summed up, and the sum matrix was determined based on the below formula:

$$R_j(x, p) = \max\left(\sum_{j=1}^{n} R_j(x, p)\right) \quad (2)$$

where R_j is the valued tolerance relation, x is the analyzed property's attribute, p is the attribute in the conditional segment of the investigated decision rule, and n is the number of property attributes in the conditional segment of the decision rule.

Table 4. Matrix of the valued tolerance relation for the usable floor area attribute
Tabela 4. Macierz wartościowanej relacji tolerancji dla atrybutu – powierzchnia użytkowa nieruchomości

Number of decision rule Numer reguły decyzyjnej	1	2	3	4	5	6	7	8	9	10
1	1	0.35	0.00	0.00	0.57	0.00	1	0.00	0.00	0.91
2	0.35	1	0.00	0.57	0.00	0.00	0.35	0.00	0.00	0.26
3	0.00	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.57	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00
5	0.57	0.00	0.00	0.00	1	0.00	0.57	0.00	0.22	0.65
6	0.00	0.00	0.00	0.00	0.00	1	0.00	0.83	0.39	0.00
7	1	0.35	0.00	0.00	0.57	0.00	1	0.00	0.00	0.91
8	0.00	0.00	0.00	0.00	0.00	0.83	0.00	1	0.57	0.00
9	0.00	0.00	0.00	0.00	0.22	0.39	0.00	0.57	1	0.00
10	0.91	0.26	0.00	0.00	0.65	0.00	0.91	0.00	0.00	1

Source: own study
Źródło: opracowanie własne

A sample sum matrix is presented in Table 5. Due to the regular entry of decision attributes (price), the number of decision rules in the analyzed example will be equal to the number of properties, i.e. 10. In view of the above, the price can be included in the determination of the overall sum matrix based on the valued tolerance relation to maximize the probability that the missing property attribute is correctly determined based on the approximate tolerance relation.

Table 5. Sum matrix determined based on the matrix of the valued tolerance relation from each attribute
Tabela 5. Macierz sumy wyznaczona na podstawie macierzy wartościowanej relacji tolerancji z poszczególnych atrybutów

Number of decision rule Numer reguły decyzyjnej	1	2	3	4	5	6	7	8	9	10
1	5.00	1.88	1.68	0.40	0.57	0.00	2.89	1.00	1.00	0.91
2	1.88	5.00	2.85	1.57	1.00	1.00	2.99	2.00	2.10	1.26
3	1.68	2.85	5.00	0.08	0.00	1.00	1.79	2.00	3.00	0.00
4	0.40	1.57	0.08	5.00	1.00	0.00	1.29	0.00	0.00	1.00
5	0.57	1.00	0.00	1.00	5.00	0.34	2.57	0.92	0.45	**3.45**
6	0.00	1.00	1.00	0.00	0.34	5.00	0.00	3.09	2.29	1.54
7	2.89	2.99	1.79	1.29	2.57	0.00	5.00	1.00	1.00	2.91
8	1.00	2.00	2.00	0.00	0.92	**3.09**	1.00	5.00	2.72	0.72
9	1.00	2.10	**3.00**	0.00	0.45	2.29	1.00	2.72	5.00	0.43
10	0.91	1.26	0.00	1.00	3.45	1.54	2.91	0.72	0.43	5.00

Source: own study
Źródło: opracowanie własne

The most approximate attribute values were determined for the property with incomplete data. The results are presented in Table 6. The sum matrix indicates that property no. 1 is most similar to property no. 7 (2.89), property no. 3 – to property no. 9 (3), property no. 6 – to property no. 8 (3.09), and property no. 10 – to property no. 5 (3.45). The above suggests that incomplete attributes will take on the values indicated in Table 6 in line with the preset approximate decision rules.

Table 6. Approximate attribute values for property with incomplete data
Tabela 6. Wyniki przybliżonych wartości atrybutów nieruchomości dla brakujących danych

Property with incomplete attributes (number of decision rule) Nieruchomość z brakującymi atrybutami (numer reguły decyzyjnej)	Approximated property (number of decision rule) Nieruchomość przybliżona (numer reguły decyzyjnej)	Incomplete data values Wartości brakujących danych			
		C_1	C_2	C_3	C_4
1	7			2	2
3	9	78			
6	8		1		
10	5		2		

Source: own study
Źródło: opracowanie własne

The quality of approximation was determined in view of the quantity of incomplete data and the total number of attributes. The results are presented in Table 7.

Table 7. Approximation quality of data with incomplete attributes
Tabela. 7. Jakość aproksymacji klasyfikacji danych z brakującymi atrybutami

Number of decision attribute Numer atrybutu decyzyjnego	Total number of attributes Liczba wszystkich atrybutów	Number of incomplete attributes Liczba brakujących atrybutów	Number of known attributes Liczba wiadomych atrybutów	Value from the sum matrix of the approximate rule Wartość z macierzy sum reguły przybliżonej	Approximation quality Jakość aproksymacji
1	5	2	3	2.89	0.96 (2.89/3)
3	5	1	4	3.00	0.75
6	5	1	4	3.09	0.77
10	5	1	4	3.45	0.86

Source: own study
Źródło: opracowanie własne

Approximation quality has been determined based on the number of known attributes and the value of the approximate decision rule attribute from the sum matrix, indicating the significance of the supplemented attribute. In view of the specific features of the real estate market, the number of analyzed objects and the varied methodology of entering attributes, the results can be regarded as satisfactory with the lowest degree of probability reaching 75%.

CONCLUSIONS

The exploration of property market data poses numerous problems for a variety of reasons, the key obstacle being the incompleteness or unavailability of the relevant data. Various qualitative and quantitative methods have been proposed to deal with this problem. Selected methods have been discussed by Uden van 2009 (Harker, Shiraishi, Kwiesielewicz methods); Hoffman and Jasiński [2009] (k-nearest neighbor algorithms), Stefanowski [2001] (approximate sets), Adamus [2008] (approximate sets).

In the simplest solution, transactions containing empty data records are rejected. This solution would be effective if it were not for the fact that incompleteness is a wide-spread problem in databases on the real estate market. The above is due to technical reasons (difficulty of attribute ranking), human error (entry omission), organizational reasons (data requiring detailed field inspections) and economic reasons (varied data requires costly and time-consuming procedures).

The presented simplified procedure for supplementing incomplete data poses an alternative to statistical methods. It may be applied when data need to be quickly supplemented, including in small sets of transactional data, without preliminary analyses which are required in statistical methods. A combination of the rough set theory and the valued tolerance relation produces optimal results in the analysis of data on the real estate market because it accounts for diverse methods of entering attributes.

REFERENCES

Adamus E., 2008. Kierunkowe zbiory podobieństwa a problem niekompletności danych. Metody informatyki Stosowanej. Wyd. Kwartalnik Komisji Informatyki PAN, Oddział Gdańsk.

d'Amato M., 2004. A comparison between MRA and Rough Set Theory for mass appraisal. A case in Bari. International Journal of Strategic Property Management, 8(4), 205–218.

d'Amato M., 2006. Rough Set Theory as Automated Valuation Methodology. The Whole Story. International seminar about Advances in Mass Appraisal in Delft.

d'Amato M., 2007. Comparing rough set theory with multiple regression analysis as automated valuation methodologies. International Real Estate Review (in corso di pubblicazione), 10(2), 42–65.

d'Amato M., 2008. Rough set theory as property valuation methodology. The whole story. [W:] Mass Appraisal Methods. An international perspective for property valuers. Red. T. Kauko, M. d'Amato. Blackwell Publishing, Oxford. RICS Research.

Deja R., 2000. Zastosowanie teorii zbiorów przybliżonych w analizie konfliktów (praca doktorska). Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Hoffman S. Jasiński R., 2009. Uzupełnianie brakujących danych w systemach monitoringu powietrza. Wyd. Wydawnictwo Politechniki Częstochowskiej, Częstochowa.

Komorowski J.; Pawlak Z.; Polkowski L.; Skowron A., 1999. Rough sets: A tutorial. [W:] Rough fuzzy hybridization: A new trend in decision making. Red. S.K. Pal, A. Skowron. Springer-Verlag, Singapore, 3–98.

Kotkowski B., Ratajczak. W., 2002. Zbiory przybliżone w analizie danych geograficznych. [W:] Możliwości i ograniczenia zastosowań metod badawczych w geografii społeczno-ekonomicznej i gospodarce przestrzennej. Red. H. Rogacki. Bogucki Wydawnictwo Naukowe, Poznań, 35–44.

Mrózek A., Płonka L., 1999. Analiza danych metodą zbiorów przybliżonych. Akademicka Oficyna Wydawnicza PLJ, Warszawa.

Nowicki R., 2009. Rozmyte systemy decyzyjne w zadaniach z ograniczoną wiedzą. Wyd. Akademicka oficyna EXIT, Warszawa.

Pawlak Z., 1982. Rough sets. International Journal of Information and Computer Science, 11, 341.

Pawlak Z., 1991. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Press, Dordrecht.

Pawlak Z., 1997. Rough sets and their applications. Seminar Department of Computing – Macquarie University.

Renigier M., 2006. Zastosowanie analizy danych metodą zbiorów przybliżonych do zarządzania zasobami nieruchomości. Studia i Materiały Towarzystwa Naukowego Nieruchomości, 14(1).

Renigier-Biłozor M., 2008a. Problematyka teorii zbiorów przybliżonych w gospodarce nieruchomościami. Referat opublikowany w „Studiach i materiałach Towarzystwa Naukowego Nieruchomości, Olsztyn.

Renigier-Biłozor M., 2008b. Zastosowanie teorii zbiorów przybliżonych do masowej wyceny nieruchomości na małych rynkach. Acta Sci. Pol., Administratio Locorum 7(3), 35–51.

Renigier-Biłozor M., Biłozor A. 2007. Application of the rough set theory and the fuzzy set theory in land management. Referat wygłoszony 28 czerwca. Annual conference The European Real Estate Society – ERES, Londyn.

Renigier-Biłozor M., Biłozor A., 2008. Aspekty i możliwości zastosowań teorii zbiorów przybliżonych i teorii zbiorów rozmytych w gospodarce przestrzennej. Referat recenzowany i opublikowany w materiałach konferencyjnych „O nowy kształt badań regionalnych w geografii i gospodarce przestrzennej", Poznań.

Renigier-Biłozor M., Biłozor A., 2009a. Procedura określania istotności wpływu atrybutów nieruchomości z wykorzystaniem teorii zbiorów przybliżonych. Przegląd Geodezyjny 6, 3–7.

Renigier-Biłozor M., Biłozor A., 2009b. The significance of real estate attributes in the process of determining land function with the use of the rough set theory. Scientific Monograph. Value in the process of real estate management and land administration. Olsztyn, 91–102.

Rough Sets in Knowledge Discovery. 1. Methodology and Applications. Red. L. Polkowski, A. Skowron. 1998a. Physica-Verlag, Heidelberg.

Rough Sets in Knowledge Discovery. 2. Applications, Case Studies and Software Systems. Red. L. Polkowski, A. Skowron. 1998b. Physica-Verlag, Heidelberg.

Słowiński R., 1992 Intelligent decision support. Handbook of Applications and advances of the rough sets theory. Kluwer Academic Publishers, Dordrecht.

Stefanowski J., 2001. Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy. Rozprawa habilitacyjna. Wydawnictwo Politechniki Poznańskiej, Poznań.

Stefanowski J., Tsoukias A., 2000. Valued Tolerance and Decision Rules [W:] Proceedings of the RSCTC 2000 Conference. Red. W. Ziarko Y. Yao. Banff.

Uden E. van, 2009. Uzupełnianie brakujących danych w macierzach porównań parami. http://www.pg.gda.pl/~mkwies/dyd/msi/bo_pdf/warsz1.pdf, dostęp: 03.03.2010 r.

UZUPEŁNIANIE BRAKUJĄCYCH DANYCH NA RYNKU NIERUCHOMOŚCI Z WYKORZYSTANIEM TEORII ZBIORÓW PRZYBLIŻONYCH

Streszczenie. W artykule zaprezentowano możliwość wykorzystania teorii zbiorów przybliżonych do uzupełniania bazy danych na rynku nieruchomości. Zaproponowana uproszczona procedura może stanowić alternatywę dla metod statycznych. Daje wiarygodne wyniki w krótkim czasie. Opracowując procedurę uzupełniania brakujących danych, wykorzystano założenia teorii zbiorów przybliżonych w połączeniu z wartościowaną relacją tolerancji. Połączenie to daje możliwie najlepsze wyniki z uwagi na uwzględnianie różnorodności sposobu zapisu atrybutów nieruchomości.

Słowa kluczowe: teoria zbiorów przybliżonych, rynek nieruchomości, uzupełnianie brakujących danych

Zaakceptowano do druku – Accepted for print: 9.08.2010