dr hab. Agnieszka Bitner
Rzeczoznawca majątkowy

Statystyka opisowa w wycenie nieruchomości
Część II – graficzna prezentacja danych

WPROWADZENIE

Wykresy statystyczne są jedną z form uporządkowanego i zwięzłego przedstawiania danych. Graficzna prezentacja danych umożliwia o wiele szybsze, w porównaniu z tabelaryczną, przekazywanie informacji o badanym zbiorze, ułatwia wstępną diagnostykę badanego zjawiska. Wykresy są niezbędne do zaobserwowania, a następnie zbadania i określenia struktury zbioru danych. Analizując duże zbiory danych napotykamy na podstawowy problem analizy danych oraz ich prezentacji graficznej jakim jest nadmiar informacji. W celu scharakteryzowania i poznania struktury badanego zbioru dane porządkuje się i dzieli na grupy. Sposób grupowania danych zależy od rodzaju cechy statystycznej (cechy dyskretne i ciągłe), celu badania statystycznego (np. badanie rozkładu cechy, badanie zjawisk zachodzących w czasie) czy liczby danych. W praktyce, jeśli cecha przyjmuje dużo wartości, traktuje się ją jako ciągłą niezależnie od tego, czy z definicji jest ciągła, czy dyskretna. W badaniach rynku nieruchomości przykładem cechy dyskretnej jest kondygnacja, na której znajduje się lokal mieszkalny, cechy ciągłej – powierzchnia nieruchomości. Przykładem badania zjawiska zachodzącego w czasie jest wyznaczenie trendu zmian cen czy zbadanie dynamiki transakcji w danym okresie czasu. Grupy mogą zawierać dane o tej samej wartości (np. numer kondygnacji) lub dane o różnych wartościach (np. różne powierzchnie mieszkań).

Szczególną rolę w analizach rynku nieruchomości zajmuje badanie rozkładu danej cechy, na przykład rozkładu pól powierzchni działek katastralnych [Bitner 2010] czy badanie rozkładu cen [Bitner 2009]. Z kształtu rozkładu cen możemy wnioskować o ich rozproszeniu, asymetrii rozkładu czy wartościach dominujących. Znajomość rozkładu cen jest potrzebna rzeczoznawcom, ponieważ stosowane w wycenie testy statystyczne zakładają często rozkład normalny cen. Zbadanie rozkładu pól powierzchni działek katastralnych umożliwia określenie poziomu zurbanizowania terenu oraz delimitacji obszaru o danym poziomie zurbanizowania. Badania te są prowadzone na dużych zbiorach danych, które wymagają grupowania.

Celem artykułu jest przedstawienie dwóch podstawowych sposobów graficznej prezentacji danych statystycznych pogrupowanych, diagramu i histogramu. Artykuł rozpocznie się od przybliżenia pojęcia szeregu rozdzielczego. Następnie zostanie opisane tworzenie diagramów i histogramów z wykorzystaniem arkusza kalkulacyjnego Excel. Na zakończenie podane zostaną przykłady histogramów i diagramów dotyczące rynku nieruchomości.

GRAFICZNA PREZENTACJA DANYCH STATYSTYCZNYCH

Do najbardziej popularnych form graficznych, za pomocą których przedstawia się dane pogrupowane należą histogram i diagram. Zanim wykresy zostaną narysowane dane należy pogrupować w postaci szeregu rozdzielczego.

Szereg rozdzielczy powstaje, gdy zakres badanego zbioru danych zostaje podzielony na przedziały i wyznaczona została liczba danych w poszczególnych przedziałach. Przedziały te nazywamy **przedziałami klasowymi**. Dane, których wartości mieszczą się w granicach ustalonego przedziału klasowego tworzą **klasę**. Podział na przedziały klasowe powinien być rozłączny – jedna dana może trafić tylko do jednego przedziału, oraz wyczerpujący – wszystkie dane są objęte klasyfikacją. W miarę możliwości należy unikać podziału, w którym występują przedziały puste, o zerowej liczebności. Najczęściej stosuje się przedziały o równej długości, co ułatwia analizę. Niekiedy charakter badanej cechy wskazuje potrzebę zastosowania przedziałów o różnej długości. Reprezentantem danego przedziału klasowego jest najczęściej jego środek. Z wielkości tej korzysta się przy obliczaniu miar opisujących zbiór danych pogrupowanych, czy przy tworzeniu diagramu.

Liczebność to liczba danych należących do danej klasy. Zazwyczaj przyjmuje się prawostronnie domknięte przedziały klasowe. Liczebność może być wyrażona w liczbach bezwzględnych (absolutnych) – **częstość absolutna (częstość, liczebność)**, lub liczbach względnych, określanych w stosunku do liczebności całego zbioru danych – **częstość względna**. Częstości względne powstają z wydzielenia częstości absolutnych przez liczebność całego badanego zbioru. Suma częstości względnych jest równa jedności. Częstości względne pokazują udział poszczególnych klas w całym zbiorze danych, dlatego są one często przedstawiane w procentach. Szereg rozdzielczy obejmujący przedziały klasowe wraz z odpowiadającymi im częstościami przedstawia **rozkład częstości**. W przypadku badania konkretnej cechy, na przykład pola powierzchni nieruchomości gruntowych, szereg rozdzielczy określa **empiryczny rozkład cechy**. Empiryczny rozkład cechy stanowi podstawę analiz statystycznych dotyczących badanej cechy.

Tworzenie szeregu rozdzielczego składa się z następujących etapów:

1. Określenie liczby i zakresu danych
2. Ustalenie liczby przedziałów klasowych
3. Ustalenie długości przedziałów klasowych
4. Ustalenie dolnej granicy pierwszego przedziału klasowego i pozostałych granic przedziałów klasowych
5. Wyznaczenie liczebności klas

Przykład

Tworzenie szeregu rozdzielczego w celu zbadania rozkładu danej cechy.

Rozważmy zbiór dwudziestu danych analizowanych w I części artykułu opublikowanego w poprzednim numerze Rzeczoznawcy Majątkowego. Przyjmijmy podział na cztery przedziały klasowe.

Szereg prosty uporządkowany:
12, 13, 17, 21, 24, 24, 26, 27, 28, 31, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
Liczba danych: 20
Rozstęp: R = 58 – 12 = 46
Liczba klas: 4
Długość przedziału klasowego: 11,6 (46/4=11,5)
Przedziały klasowe: (11,8; 23,4], (23,4; 35], (35; 46,6], (46,6; 58,2]
Liczebności klas: 4, 8, 6, 2

Liczbę klas oraz długości przedziałów klasowych ustala się arbitralnie. W dalszej części artykułu podane zostaną wskazówki dotyczące określania liczby klas. Przyjęta w przykładzie długość przedziału równa 11,6, a nie 11,5 wynika z dostosowania granic przedziałów klasowych do algorytmu arkusza kalkulacyjnego Excel, co zostanie szczegółowo wyjaśnione w następnym przykładzie. Jeśli do utworzenia szeregu rozdzielczego nie wykorzystamy arkusza Excel, to podział na przedziały klasowe może być następujący: [12; 23,5], (23,5; 35], (35; 46,5], (46,5;58]. Oczywiście oba podziały są poprawne pod względem statystycznym.

Szereg rozdzielczy

Przedziały klasowe	Częstość	Częstość względna	Częstość względna (%)
11,8 – 23,4	4	0,2	20
23,4 – 35	8	0,4	40
35 – 46,6	6	0,3	30
46,6 – 58,2	2	0,1	10
Suma	20	1	100

W analizie danych często wyznacza się również **szeregi rozdzielcze skumulowane**, które podają ile danych w badanym zbiorze, przyjmuje wartości mniejsze bądź równe górnej granicy kolejnych przedziałów klasowych. Poszczególne liczby tych danych nazywamy **częstościami skumulowanymi**.

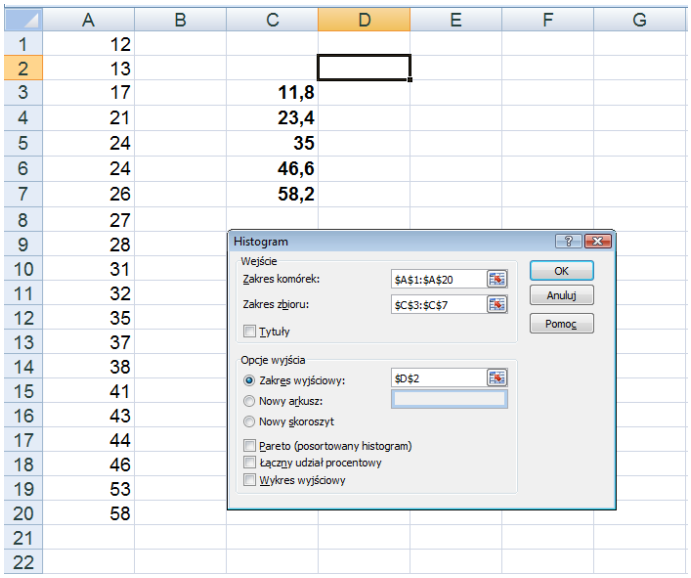
Szereg rozdzielczy skumulowany

Przedziały klasowe	Częstość skumulowana	Częstość względna skumulowana	Częstość względna skumulowana (%)
11,8 – 23,4	4	0,2	20
23,4 – 35	12	0,6	60
35 – 46,6	18	0,9	90
46,6 – 58,2	20	1	100

Przykład
Tworzenie szeregu rozdzielczego za pomocą arkusza kalkulacyjnego Excel dla danych:
12, 13, 17, 21, 24, 24, 26, 27, 28, 31, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Kolejne etapy tworzenia szeregu rozdzielczego przedstawia poniższy opis oraz rysunki 1 i 2.

1. Wpisanie w kolumnie arkusza zbioru danych: komórki A1:A20
2. Wyznaczenie liczby danych: 20
3. Wyznaczenie zakresu danych: od 12 do 58
4. Ustalenie liczby przedziałów klasowych: 4
5. Ustalenie długości przedziałów klasowych: 11,6 (46/4=11,5)
6. Ustalenie prawych krańców przedziałów: 11,8; 23,4; 35; 46,6; 58,2
7. Wpisanie w kolumnie prawych krańców przedziałów: komórki C3:C7
8. Wybranie opcji: *Dane/Analiza danych/Histogram*



Rys. 1

Zakres komórek to zbiór danych umieszczony w kolumnie, w przykładzie zakres A1:A20 (rys. 1). Zakres zbioru to prawe krańce przedziałów, w przykładzie zakres C3:C7. Zakres wyjściowy to miejsce w arkuszu, w którym chcemy, żeby został umieszczony szereg rozdzielczy. Standardowo szereg rozdzielczy jest umieszczany w nowym arkuszu. Jeśli chcemy go umieścić w bieżącym arkuszu, obok zbioru danych, wystarczy wpisać w zakresie wyjściowym adres pierwszej komórki bloku komórek, w którym chcemy umieścić szereg rozdzielczy, czyli adres komórki D2. Następnie wybieramy *OK*, w wyniku otrzymamy poniższy szereg rozdzielczy (rys. 2).

	A	B	C	D	E	F	G
1	12						
2	13			*Zbiór danych (koszyk)*	*Częstość*		
3	17		**11,8**	11,8	0		
4	21		**23,4**	23,4	4		
5	24		**35**	35	8		
6	24		**46,6**	46,6	6		
7	26		**58,2**	58,2	2		
8	27			Więcej	0		
9	28						
10	31						
11	32						
12	35						
13	37						
14	38						
15	41						
16	43						
17	44						
18	46						
19	53						
20	58						
21							
22							

Rys. 2

Uwaga:
Korzystając z arkusza kalkulacyjnego Excel, w komórkach C3:C7 zostały wpisane prawe krańce przedziałów. Liczba oznaczająca prawy kraniec pierwszego przedziału, czyli 11,8, musi być mniejsza od najmniejszej wartości badanej cechy w zbiorze danych, mniejsza od 12. Przedziały w algorytmie arkusza Excel są prawostronnie domknięte, czyli liczby w komórkach C3:C7 podzieliły zbiór liczb rzeczywistych na następujące przedziały: $(-\infty; 11{,}8]$, $(11{,}8; 23{,}4]$, $(23{,}4; 35]$, $(35; 46{,}6]$, $(46{,}6; 58{,}2]$, $(58{,}2; +\infty)$. Ostatni przedział tego podziału jest nazywany *Więcej*. Jeśli przyjęlibyśmy, że prawy kraniec pierwszego przedziału jest równy najmniejszej wartości badanej cechy, czyli 12, wówczas liczebność tego przedziału zawsze byłaby równa liczebności tej najmniejszej wartości, w przykładzie równa 1. Dlatego została przyjęta nieco szersza długość przedziału klasowego równa 11,6, żeby w sumie przedziały pokrywały z nadmiarem badany zbiór. Warunki poprawności zastosowania algorytmu programu Excel spełniałby również podział: (11,99; 23,5], (23,5; 35], (35; 46,5], (46,5;58].

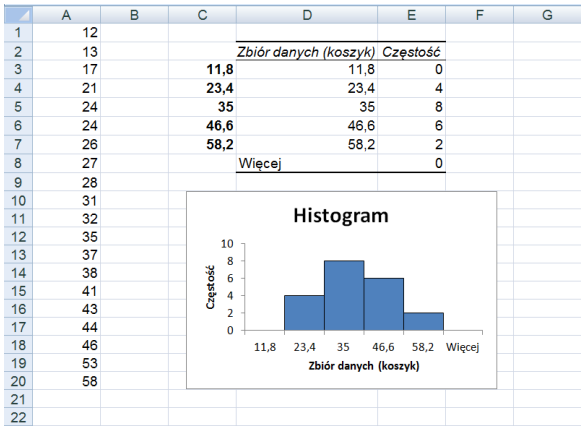
Histogram

Dane pogrupowane w klasy wraz z rozkładem częstości w klasach możemy przedstawić na wykresie zwanym histogramem. **Histogram** to zbiór słupków, których podstawy są równe długościom przedziałów klasowych i znajdują się na osi odciętych. Wysokości słupków określają częstości odpowiadające poszczególnym klasom. Sąsiednie słupki przylegają do siebie. Warto dodać, że kształt histogramu częstości absolutnych oraz częstości względnych jest taki sam. Zmienia się jedynie skala na osi rzędnych.

Przykład
Tworzenie histogramu za pomocą arkusza Excel

Histogram w arkuszu Excel tworzymy korzystając ze znanej z poprzedniego przykładu opcji *Dane/Analiza danych/Histogram*. Wystarczy w oknie dialogowym, przedstawionym na rysunku 1, wybrać dodatkowo opcję *Wykres wyjściowy*. Otrzymamy wówczas szereg rozdzielczy wraz z histogramem, co pokazano na rysunku 3. Należy dodać, że standardowo słupki histogramu nie przylegają do siebie. Histogram zgodny z definicją, która została

podana w niniejszym artykule, można otrzymać dzięki opcji *Formatuj serię danych*. Opcja ta ukaże się na ekranie, jeśli klikniemy prawym klawiszem myszy w dowolny słupek histogramu.

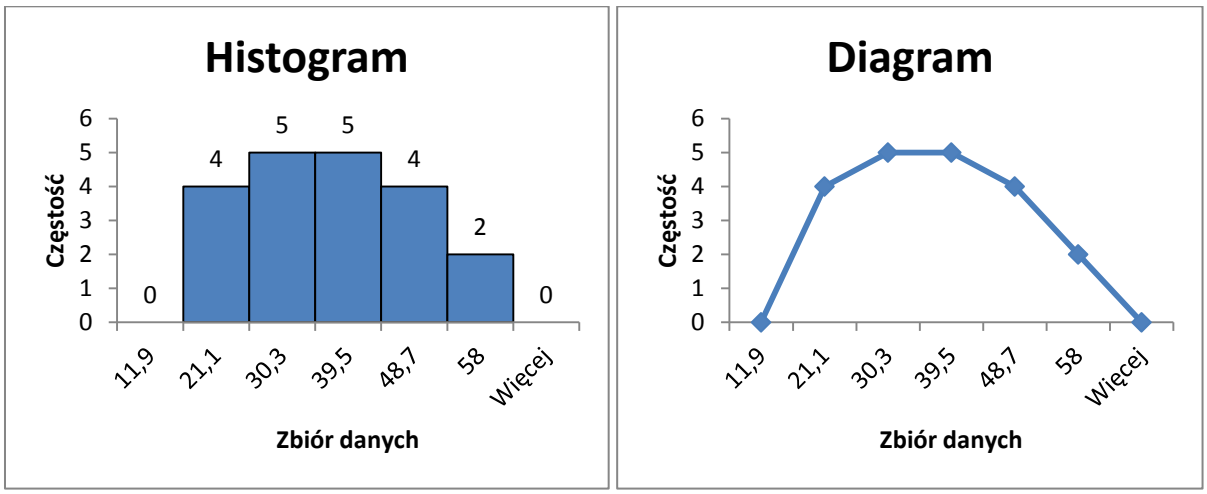


	A	C	D	E
1	12			
2	13		*Zbiór danych (koszyk)*	*Częstość*
3	17	**11,8**	11,8	0
4	21	**23,4**	23,4	4
5	24	**35**	35	8
6	24	**46,6**	46,6	6
7	26	**58,2**	58,2	2
8	27		Więcej	0
9	28			
10	31			
11	32			
12	35			
13	37			
14	38			
15	41			
16	43			
17	44			
18	46			
19	53			
20	58			
21				
22				

Rys. 3

Diagram (wielobok częstości)

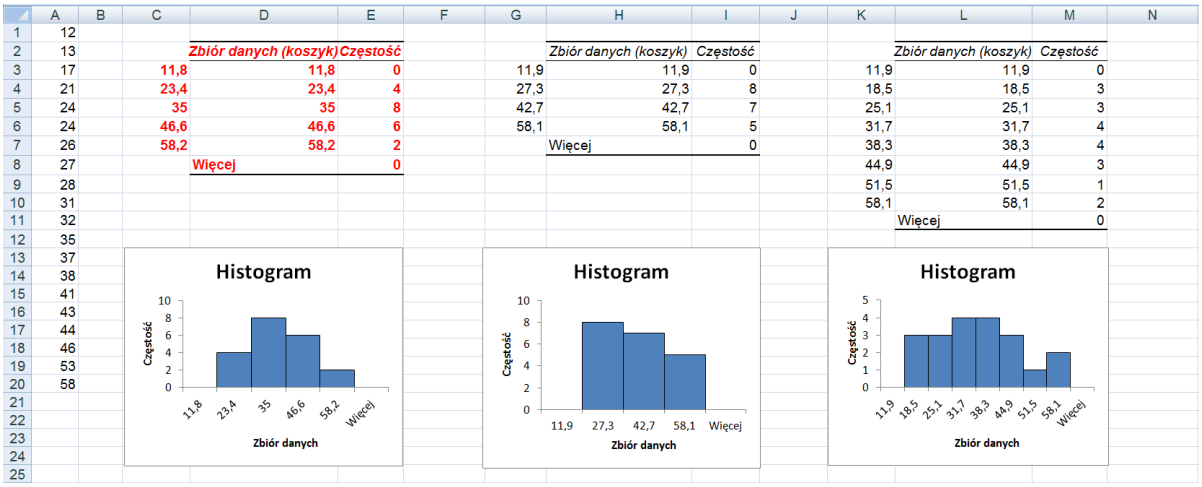
Środki górnych boków słupków histogramu połączone linią łamaną tworzą **diagram**. W celu lepszej prezentacji danych przyjmuje się często jako początek łamanej środek pustego przedziału poprzedzającego pierwszy słupek, jako koniec łamanej środek pustego przedziału następującego po ostatnim słupku histogramu. Rysunek 4 przedstawia histogram oraz diagram przy podziale na pięć przedziałów klasowych. Diagram w arkuszy Excel tworzymy korzystając z opcji *Wstawianie/Wykresy Punktowy z prostymi liniami i znacznikami*, dla wcześniej obliczonych w dwóch kolumnach wartości środków przedziałów klasowych i odpowiadających im liczebności.



Rys. 4

Powróćmy jeszcze do problemu zasygnalizowanego na początku artykułu, mianowicie do ustalenia liczby przedziałów klasowych. Jak już wspomniano, podział na klasy jest arbitralny. Postać szeregu rozdzielczego określa osoba wykonująca analizę. Poprawność dokonanego grupowania danych zależy zatem w dużym stopniu od doświadczenia i znajomości reguł statystycznych tej osoby. Do ostatecznej postaci szeregu rozdzielczego

dochodzi się metodą prób i błędów dokonując różnych podziałów na klasy. Dzięki komputerom możemy szybko utworzyć histogram i ocenić otrzymany rozkład. Badając podziały na różną liczbę klas ostatecznie wybieramy tę wersję, która najlepiej pokazuje strukturę danych, i która w sposób najbardziej zbliżony spełnia cele badania statystycznego. Nie ma jednoznacznych reguł tworzenia podziału na klasy. Można jedynie stwierdzić, że liczba przedziałów klasowych nie powinna być ani zbyt duża (następuje wtedy „rozmycie" struktury zjawiska), ani też zbyt mała (następuje wtedy „zatarcie" struktury badanego zjawiska) [Luszniewicz i Słaby 1996]. Podstawowym problemem przy tworzeniu szeregu rozdzielczego jest zatem ustalenie odpowiedniej liczby przedziałów klasowych. Problem ten ilustruje rysunek 5.



Rys. 5

Ustalając intuicyjnie liczbę przedziałów klasowych widzimy, że najwięcej informacji o rozkładzie częstości przekazuje histogram pierwszy z podziałem na cztery klasy. Podział na trzy klasy powoduje zbyt dużą utratę informacji. Podział na siedem klas pokazuje wciąż zbyt rozmytą strukturę badanej cechy. Najbardziej odpowiednim jest zatem podział na cztery klasy.

Jak już wspomniano, nie ma ściśle określonej reguły na ile klas rozdzielić zbiór danych. Na szczęście istnieje kilka wskazówek, które sugerują liczbę klas, na jakie najlepiej podzielić badany zbiór danych. Poniżej podano trzy z nich.

$$k = 1 + \log_2 n \qquad (1)$$

$$k \leq 5 \ln n \qquad (2)$$

$$k = \sqrt{n} \qquad (3)$$

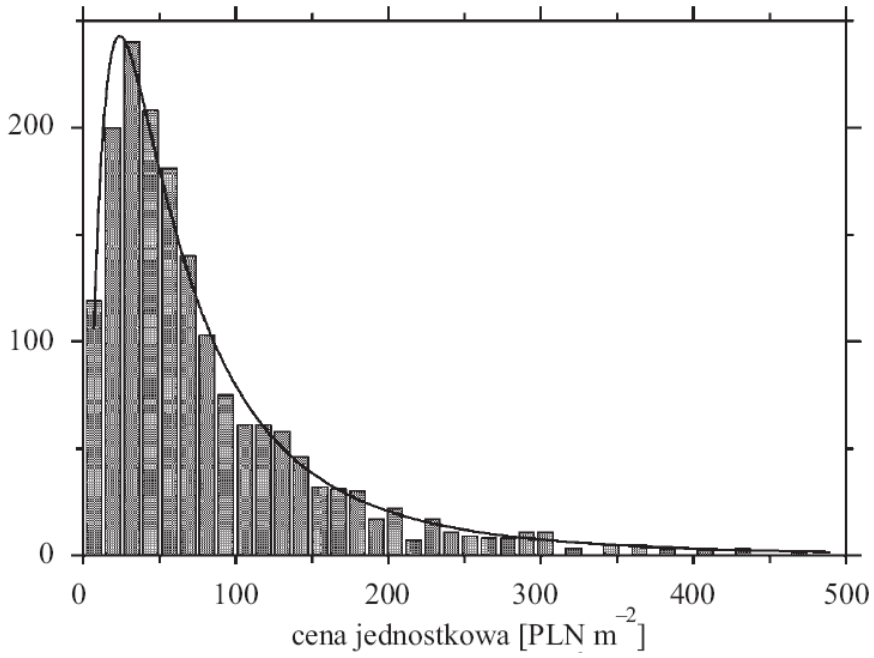
gdzie:
n – liczba wszystkich danych,
k – liczba klas,

Dla liczby danych $n = 20$, liczba przedziałów klasowych wyznaczona według wzorów (1), (2) i (3) wynosi odpowiednio $k_{(1)} = 5{,}3219$, $k_{(2)}$ jest mniejsza lub równa 14,9787, $k_{(3)} = 4{,}4721$. Po zaokrągleniu do całości, sugerowana liczba przedziałów klasowych wynosi 5 lub 4. Rysunki 4 i 5 potwierdzają trafność tej sugestii.

PRZYKŁADY HISTOGRAMÓW I DIAGRAMÓW W BADANIACH RYNKU NIERUCHOMOŚCI

1) Rozkład cen nieruchomości gruntowych.

Dane transakcyjne pochodzą z aktów notarialnych dotyczących umów kupna-sprzedaży nieruchomości gruntowych niezabudowanych położonych w granicach administracyjnych miasta Krakowa zawartych w latach 1996 – 1999. W niniejszej analizie uwzględniono jedynie grunty o przeznaczeniu M4 pod niską zabudowę mieszkaniową. Analizą objęto 1777 danych transakcyjnych. Pokazano, że na bardzo wysokim poziomie istotności rozkład cen jednostkowych jest zgodny z rozkładem logarytmiczno-normalnym (linia ciągła na wykresie), (rys. 6).



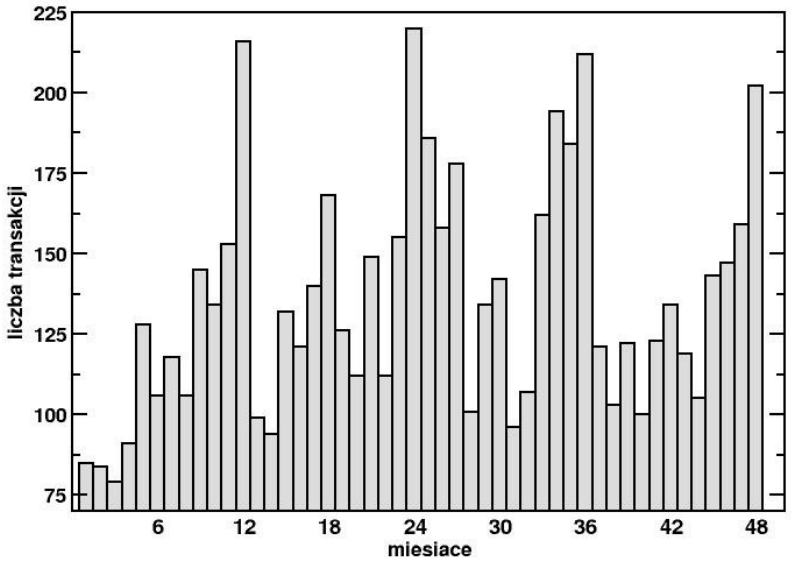
Rys. 6

Źródło: [Bitner 2009]

2) Dynamika transakcji w latach od 1996 do 1999 roku

Rysunek 7 przedstawia dynamikę sprzedaży nieruchomości gruntowych w okresie od stycznia 1996 do grudnia 1999 roku. Charakteryzuje ją pewna cykliczność. Najwięcej transakcji było zawieranych pod koniec każdego roku, przy czym grudzień był zawsze miesiącem dominujący. Miesiące styczeń i luty charakteryzowały się natomiast najmniejszą liczbą transakcji. Spadek liczby transakcji notuje się również w miesiącach wakacyjnych. łącznie, w pierwszej połowie każdego roku zawierano mniej transakcji niż

w drugim półroczu. Znajomość dynamiki sprzedaży może ułatwić prognozę czasu niezbędnego do wyeksponowania nieruchomości na rynku w celu jej sprzedaży, (rys. 7).

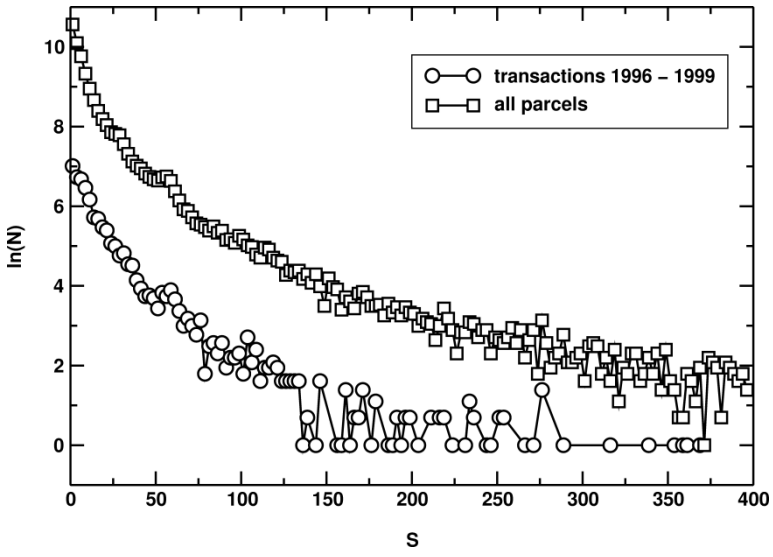


Rys. 7

Źródło: Badania własne.

3) Reprezentatywność bazy nieruchomości gruntowych niezabudowanych transakcyjnych ze względu na rozkład pól powierzchni działek.

Rysunek 8 przedstawia porównanie rozkładów pól powierzchni nieruchomości gruntowych transakcyjnych z lat 1996 – 1999 z całym zasobem działek ewidencyjnych w granicach administracyjnych miasta Krakowa, (rys. 8).

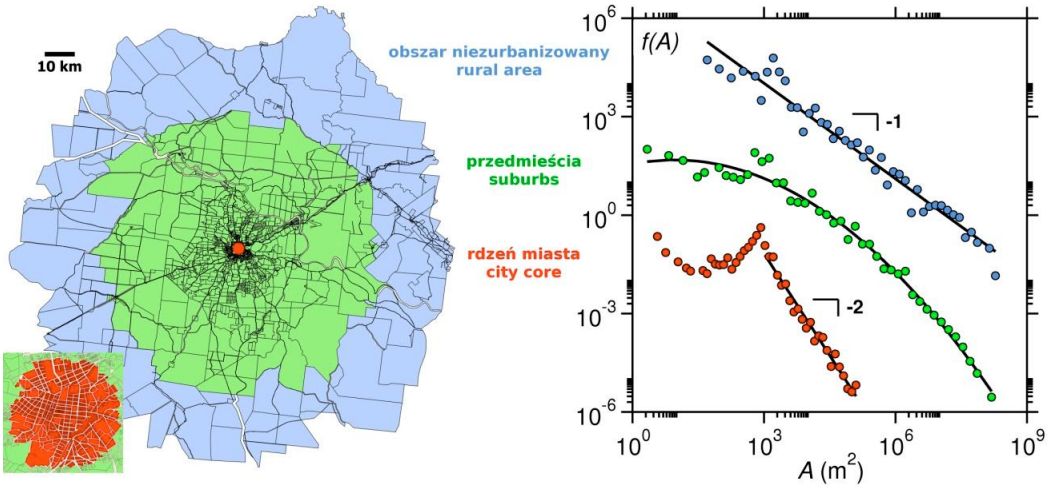


Rys. 8

Źródło: [Bitner 2002]

4) Rozkład pól powierzchni działek katastralnych

Mapa australijskiego miasta Charters Towers i jego okolic. Czarne linie oznaczają granice działek. Trzy kolory, którymi zaznaczono działki wyznaczają trzy obszaru o różnych poziomach urbanizacji: (1) rdzeń miasta – obszar najsilniej zurbanizowany, (2) przedmieścia – średni poziom urbanizacji, (3) teren niezurbanizowany. W lewym dolnym rogu umieszczono powiększenie rdzenia miasta. Wykres przedstawia rozkład pól powierzchni działek, *f(A)*, dla odpowiednich poziomów zurbanizowania. Osie wykresu są w skali logarytmicznej. Dla zwiększenia przejrzystości wykresu rozkłady zostały przesunięte w kierunku pionowym poprzez przemnożenie wartości dystrybucji przez pewne stałe, (rys. 9).



Rys. 9

Źródło: [Bitner 2010]

PODSUMOWANIE

W artykule opisane zostały dwa podstawowe sposoby prezentacji danych statystycznych, diagram i histogram. Przed narysowaniem tego typu wykresów zbiór danych należy przedstawić w postaci szeregu rozdzielczego, prezentującego uporządkowane i pogrupowane dane. W przypadku badania konkretnej cechy, szereg rozdzielczy określa empiryczny rozkład cechy, stanowiący podstawę dalszych analiz statystycznych. Przykładami rozkładów wykorzystywanych w wycenie nieruchomości są: rozkład cen danego typu nieruchomości czy rozkład konkretnej cechy. Korzystając z rozkładu cen przedstawionego za pomocą diagramu lub histogramu możemy wnioskować o ich rozproszeniu, asymetrii

rozkładu, wartościach dominujących. Możemy ponadto określić postać funkcyjną rozkładu, w celu stosowania bardziej zaawansowanych metod statystycznym – testów statystycznych.

Literatura:

Aczel A.D. Statystyka w zarządzaniu. PWN, Warszawa 2000

Bitner A., Nowa metoda określania poziomu zurbanizowania obszaru na podstawie morfologii podziału gruntu na działki. Infrastruktura i ekologia terenów wiejskich. Komisja Technicznej Infrastruktury Wsi PAN w Krakowie. 3, 164-179, 2010

Bitner A., The issue of the representativeness of random samples in the context of parcel field areas. Proceedings of the Geodesy and Environment Engineering Commission of the Polish Academy of Sciences - The Cracow Section, Geodesy, 39, 87-92, 2002

Bitner A., Rozkład jednostkowych cen nieruchomości gruntowych. Acta Scientiarum Polonorum - Administratio Locorum, 8(4), 41-50, 2009

Jóźwiak J., Podgórski J. Statystyka od podstaw. PWE, Warszawa 2000

Luszniewicz A., Słaby T. Statystyka stosowana. PWE, Warszawa 1996.

Parlińska M., Parliński J. Statystyczna analiza danych z Excelem. Wydawnictwo SGGW, Warszawa 2011