METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXI/4, 2020, s. 224 – 234

Received: 01.12.2020
Accepted: 28.12.2020

WIZUALIZACJA JAKO NARZĘDZIE WSPOMAGAJĄCE ANALIZĘ DANYCH W PROCESIE DYDAKTYCZNYM

Monika Zielińska-Sitkiewicz https://orcid.org/0000-0003-4829-3239
Mariola Chrzanowska https://orcid.org/0000-0002-8743-7437
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: monika_zielinska_sitkiewicz@sggw.edu.pl
mariola_chrzanowska@sggw.edu.pl

Streszczenie Prezentacja informacji w postaci graficznej jest jedną z podstawowych form wizualizacji danych. Taka forma przedstawienia informacji jest dużym wsparciem zarówno podczas wstępnej, jak i dalszej analizy. Jednakże niewłaściwa forma graficzna może prowadzić do mistyfikacji, a w konsekwencji do błędnych wniosków. W niniejszej pracy zaprezentowano wybrane przykłady graficznej prezentacji danych zaczerpnięte z praktyki dydaktycznej autorek. Ponadto omówiono przypadki poprawnej oraz błędnej prezentacji danych.

Słowa kluczowe: prezentacja graficzna, szeregi czasowe, praktyka dydaktyczna

JEL classification: C19, Y10

WPROWADZENIE

Graficzna prezentacja danych i wyników analiz pełni bardzo ważną rolę w procesie dydaktycznym. Jest ona istotna zarówno na wczesnym etapie, czyli podczas poznawania i interpretowania zależności opisywanych przez dane, jak i później, czyli podczas prezentacji tych zależności innym osobom.

Jak dowodzi P. Biecek, dobra grafika statystyczna powinna pokazywać informację zawartą w danych liczbowych. Powinna to robić w taki sposób, by łatwo było odczytać i zrozumieć związek pomiędzy informacją a danymi. Obrazować, jak duże są pewne wielkości, jak ryzykowne są pewne rozwiązania, jak wyglądają zależności pomiędzy zjawiskami [Biecek 2014].

Dość trudnym zadaniem dydaktycznym w procesie kształcenia studentów jest przekazanie im wiedzy na temat badania i oceny różnego rodzaju danych

https://doi.org/10.22630/MIBE.2020.21.4.20

ilościowych. Jednym ze sposobów prezentacji analizy jest jej opracowanie w formie graficznej. Pozwala to w przystępny sposób zaprezentować zależności zachodzące pomiędzy mierzonymi bądź obserwowanymi wielkościami [Lenik i in. 2007].

Dla człowieka najbardziej naturalnym i najlepiej rozwiniętym źródłem informacji o obserwowanych obiektach jest zmysł wzroku. Według E. Dale'a[1], który opracował w 1946 r. „Cone of Experience" (Stożek Doświadczeń), ludzie uczący się mogą znacznie poprawić zdolność zapamiętywania, przyswajając wiedzę w oparciu o formy audio-wizualne oraz korzystając z doświadczenia [Thalheimer 2006].

Ponadto liczne badania potwierdziły, że jeśli proces dydaktyczny zostanie uzupełniony metodami interaktywnymi, z odpowiednio opracowaną grafiką, to korzyści poznawcze mogą być dużo wyższe. Jedno z takich badań przeprowadzone przez G. L. Adamsa wykazało następujące zyski:

- skuteczność nauczania może być większa o 56%;
- zrozumienie tematu może wzrosnąć od 56% do 60%;
- oszczędność czasu może wynieść od 38% do70%;
- zakres przyswojonej wiedzy może być o 25% – 50% szerszy (por. [Adams 1992]).

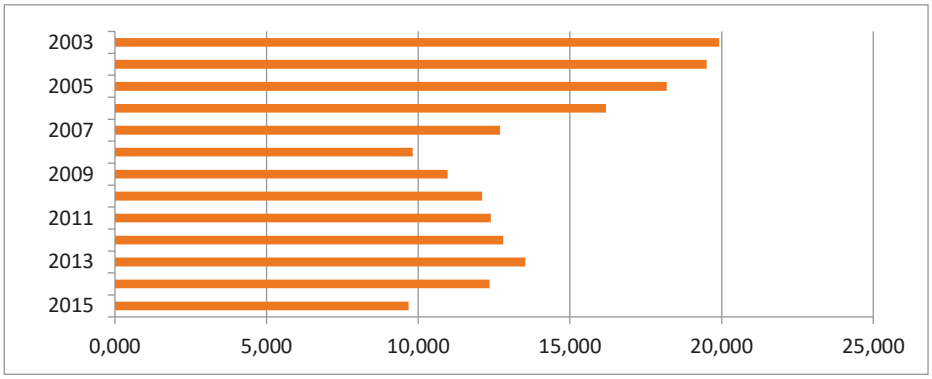
Zatem zwizualizowanie prezentowanego zagadnienia przyczynia się do lepszego zapamiętywania oraz budowania systemów skojarzeń, czyli instalowania tzw. haków pamięciowych.

Głównym celem pracy jest podkreślenie pomocniczej roli wizualizacji danych w kolejnych etapach analizy danych ze szczególnym uwzględnieniem danych czasowych. Cel został zrealizowany za pomocą kilku przykładów, które pozwolą bezpośrednio uzasadnić stosowanie metod wizualizacji podczas przeprowadzonych analiz. Każdy z przedstawionych przykładów jest propozycją wykorzystania metod wizualizacji danych na jednym z etapów eksploracji szeregu czasowego. Stanowią one również pewien zbiór dydaktycznych doświadczeń przedyskutowanych przez autorki ze studentami na różnych etapach ich edukacji.

Przykład z projektu studenckiego realizowanego na 1 roku studiów na kierunku Informatyka i Ekonometria (por. rysunek 1) obrazuje, jak bardzo potrzebne jest kształcenie w kierunku prawidłowego tworzenia i interpretacji grafiki danych czasowych. Uproszczona automatyzacja wielu narzędzi do tworzenia grafiki w popularnych programach komputerowych (np. MS Excel) powoduje, że autorzy nie przywiązują należytej uwagi do kontroli uzyskanych efektów wizualnych i nie dostrzegają błędów merytorycznych.

[1] Dale E. (1946, 1954, 1969). Audio-visual methods in teaching. New York: Dryden.

Rysunek 1. Grafika obrazująca stopę bezrobocia, wykonana przez studenta 1 roku, kierunku Informatyka i Ekonometria



Źródło: projekt studenta. Stopa bezrobocia (dane GUS)

W trakcie procesu dydaktycznego należy zatem zwracać szczególną uwagę na prawidłową i przemyślaną formę prezentacji danych.

WYKRYWANIE DANYCH ODSTAJĄCYCH

Analiza pierwotnego materiału statystycznego pozwala wyodrębniać jednostki, dla których wartości zmiennej znacznie odbiegają od pozostałych, czyli tzw. jednostki (obserwacje) odstające. Nietypowość informacji spowodowana jest na ogół niejednorodnością zbiorowości statystycznej, z której została pobrana próba, czy też nieoczekiwanymi zmianami, jakie zaszły w badanym zjawisku. Przyczyną jej powstania może być również błąd popełniony podczas pomiaru lub zapisywania wyników. Jak podaje P. Dittman, obserwacjami nietypowymi (odstającymi, odizolowanymi) nazywane są obserwacje, które znacząco różnią się od całego zbioru danych. Z tego powodu mogą one wywierać większy wpływ na oszacowania parametrów modelu niż pozostałe obserwacje[2] [Dittman 2000].

Obserwacje odstające wyraźnie różnią się wartością od pozostałych. Może to być efekt występujących w badanej zmiennej silnych wahań o charakterze losowym lub błędnego zapisu. Najprostszym sposobem wykrywania danych nietypowych dla danych jednowymiarowych jest ich wizualizacja na wykresie liniowym (rysunki 2 i 3) oraz wykresie pudełkowym ramka-wąsy (rysunek 4).

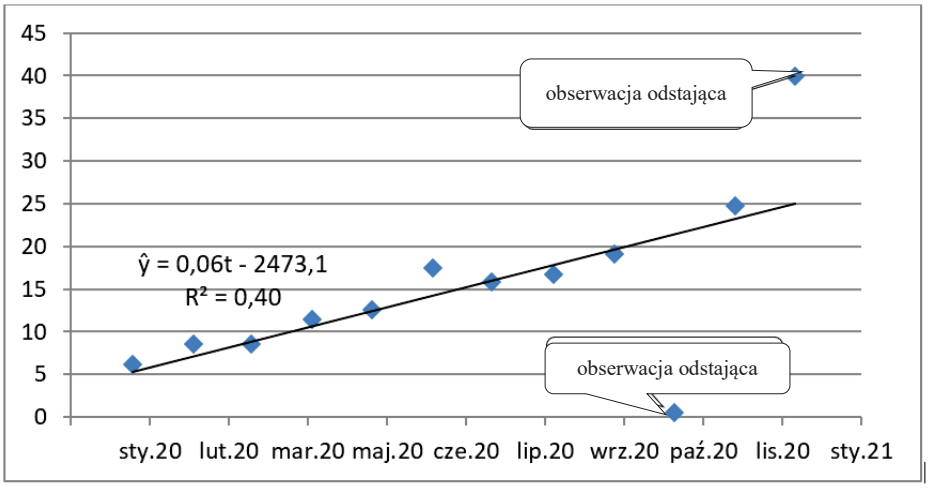
[2] W analizie zjawisk ekonomicznych informacja o pojawieniu się obserwacji nietypowych może wskazywać na zmianę tendencji analizowanego zjawiska, a w przypadku wartości ekstremalnych na występowanie tzw. punktów zwrotnych w analizie badanego zjawiska.

Tabela 1. Miesięczna sprzedaż (w tys. sztuk) zabawek typu Auto w 2020 roku

sty-20	lut-20	mar-20	kwi-20	maj-20	cze-20
6,17	8,57	8,61	11,47	12,54	17,49
lip-20	sie-20	wrz-20	paź-20	lis-20	gru-20
15,90	16,78	19,07	0,50	24,71	40,00

Źródło: dane umowne

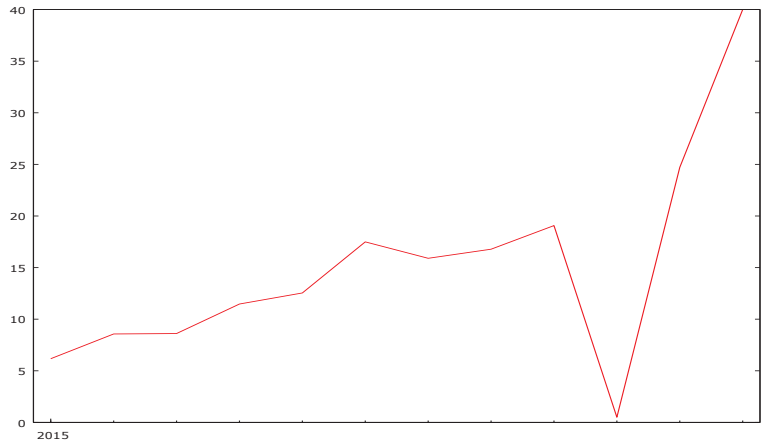
Rysunek 2. Prezentacja miesięcznej sprzedaży zabawek typu Auto w 2020 roku (tys. szt.)



Źródło: dane umowne (ilustracja MS Excel)

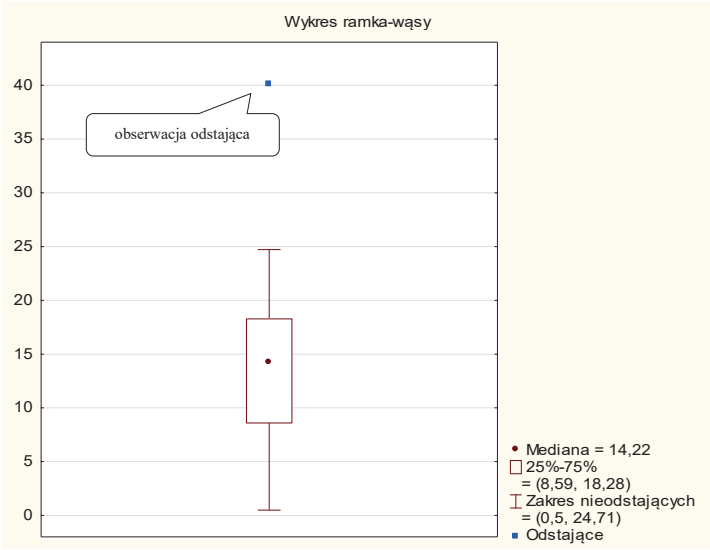
Rysunki 2 i 3 zawierają prezentacje miesięcznej sprzedaży zabawek typu Auto. Oba wykresy zostały zbudowane w dwóch programach powszechnie wykorzystywanych podczas zajęć praktycznych. Automatyzacja procedur tworzenia grafiki może jednak prowadzić do błędnych interpretacji ze względu na przypadkowe zniekształcenie percepcji danych, co obrazuje rysunek 3 – komunikat graficzny nie sugeruje w sposób czytelny występowania obserwacji nietypowych.

Rysunek 3. Prezentacja miesięcznej sprzedaży zabawek typu Auto w 2020 roku (tys. szt.)



Źródło: dane umowne (ilustracja pakiet R)

Rysunek 4. Wykres pudełkowy dla miesięcznej sprzedaży zabawek typu Auto w 2020 roku (tys. szt.)



Źródło: dane umowne (ilustracja pakiet Statistica).

Z kolei zaproponowany przez J. Tukeya [Tukey 1977] wykres pudełkowy ramka-wąsy ma kształt prostokąta z dołączonymi po bokach „wąsami". Skala przedstawiająca zakres danej zmiennej jest umieszczona równolegle do boku prostokąta i „wąsów". Obserwacje odstające to takie, których odległość przekracza

1,5 wysokości pudełka[3]. Obserwacje, których odległość od mediany jest wyższa niż trzykrotna długość prostokąta nazywane są obserwacjami ekstremalnymi. Na rysunku 4 przedstawiono klasyczny wykres pudełkowy dla miesięcznej sprzedaży zabawek typu Auto.

Zauważmy, że student może zwizualizować dane na kilka sposobów w zależności od oprogramowania używanego na zajęciach. Należy jednak podkreślić za P. Bieckiem, że nie jest ważne, czy na wykresie przedstawione są dobre wielkości lub zależności, lecz to, czy zostaną one prawidłowo z tego wykresu odczytane (por. P. Biecek [2014] s. 110). W omawianym przykładzie wykres pudełkowy ramka-wąsy jest najbardziej czytelny i spójny, stanowi bowiem dobry kompromis pomiędzy ilością informacji a ich zwięzłością. Ważnym celem nauczania wydaje się zatem zachęcenie słuchacza, by korzystał z wielu udostępnianych mu narzędzi służących do graficznej prezentacji danych. Istotne jest, by potrafił wybrać prawidłową formę ich ilustracji, z której wyciągnie poprawne wnioski.

PROSTE SPOSOBY IDENTYFIKACJI POSTACI FUNKCYJNEJ MODELU

Popularnym miernikiem sprawdzania jakości modelu zbudowanego Klasyczną Metodą Najmniejszych Kwadratów jest współczynnik determinacji R^2. Określa on poziom dopasowania zmiennej objaśnianej Y do danych empirycznych. Wiara studentów w wiarygodność tego miernika bywa bezwarunkowa i często jest to podstawowy sposób weryfikacji zbudowanego modelu.

Przykładem, który może podważyć taką opinię studentów jest tzw. kwartet Anscombe'a [Anscombe 1973]. Są to cztery zestawy danych specjalnie dobranych przez angielskiego statystyka F. Anscombe'a w taki sposób, aby w każdym z nich występowały identyczne mierniki statystyczne, takie jak średnia arytmetyczna, wariancja, współczynnik korelacji, równanie regresji liniowej czy współczynnik determinacji. Ten zbiór danych (por. tabela 2), opublikowany już w 1973 r., wskazuje, jak bardzo istotna jest prezentacja graficzna w procesie analizy statystycznej. Dla wszystkich par zmiennych ((*X1,Y1*), (*X2,Y2*), (*X3,Y3*) oraz (*X4,Y4*)) oszacowany za pomocą KMNK model ma postać: $\widehat{y}_i = 0{,}5 + 3{,}0x_i$, a współczynnik determinacji wynosi $R^2 = 0{,}67$.

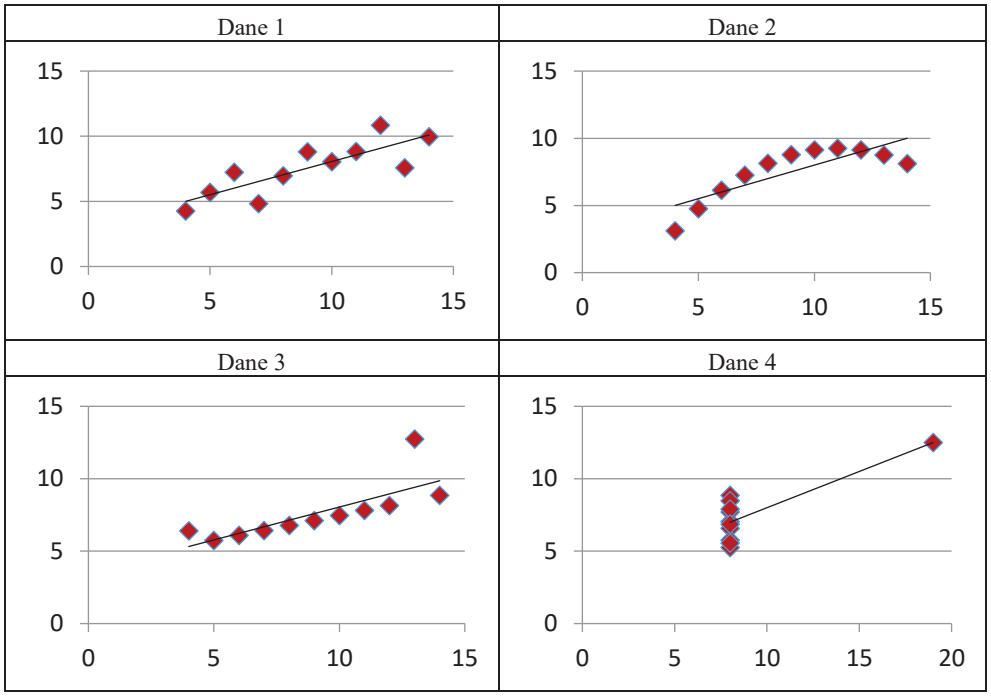
[3] Ta odległość jest zalecana w literaturze, w praktyce decyzja należy do badacza.

Tabela 2. Zbiór danych zaproponowany przez F. Anscombe'a

Dane 1	*X1*	10,00	8,00	13,00	9,00	11,00	14,00	6,00	4,00	12,00	7,00	5,00
	Y1	8,04	6,95	7,58	8,81	8,83	9,96	7,24	4,26	10,84	4,82	5,68
Dane 2	*X2*	10,00	8,00	13,00	9,00	11,00	14,00	6,00	4,00	12,00	7,00	5,00
	Y2	9,14	8,14	8,74	8,77	9,26	8,10	6,13	3,10	9,13	7,26	4,74
Dane 3	*X3*	10,00	8,00	13,00	9,00	11,00	14,00	6,00	4,00	12,00	7,00	5,00
	Y3	7,46	6,77	12,74	7,11	7,81	8,84	6,08	6,39	8,15	6,42	5,73
Dane 4	*X4*	8,00	8,00	8,00	8,00	8,00	8,00	8,00	19,00	8,00	8,00	8,00
	Y4	6,58	5,76	7,71	8,84	8,47	7,04	5,25	12,50	5,56	7,91	6,86

Źródło: Anscombe F. J. [1973] s. 17–21.

Rysunek 5. Graficzna prezentacja zbioru Anscombe'a



Źródło: opracowanie własne

W tym przypadku to komunikat graficzny pozwala szybko zauważyć kardynalne błędy, jakie można popełnić, stosując regresję liniową dla danych numer 2 i 4 (rysunek 5).

Należy podkreślić, że sprawność studentów w odczytywaniu i interpretowaniu danych z szeregów i tablic wzrasta wraz z rosnącą w trakcie studiów częstotliwością ich pracy na różnych zbiorach danych. W przedstawionym

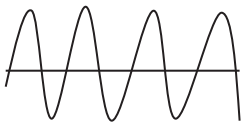
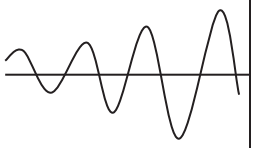
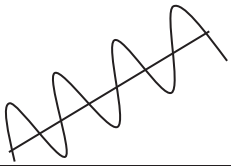
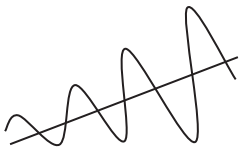
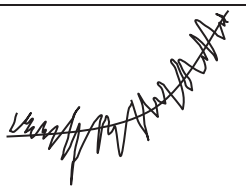
przykładzie „siła" przekazu prezentacji graficznej nie budzi wątpliwości i przypadek ten warto wprowadzać do zajęć praktycznych ze względu na eksponowane pułapki interpretacyjne.

DEKOMPOZYCJA SZEREGU CZASOWEGO

Celem analizy szeregów czasowych jest skonstruowanie modelu pewnego zjawiska w oparciu o obserwowane w czasie zmiany pewnych mierzalnych wielkości opisujących ten proces. Zgodnie z ogólnym założeniem, analizowany przebieg składa się z części systematycznej (trend, składowa stała, wahania sezonowe i cykliczne), w oparciu, o którą buduje się model, oraz części przypadkowej (szumu, wahań przypadkowych).

W analizie szeregów dąży się do wyodrębnienia i pomiaru czynników determinujących rozważane zjawisko, dokonując dekompozycji szeregu czasowego. Wykorzystując zbudowany model, można dokonywać predykcji (eksploracji) przebiegu szeregu lub jego składowych.

Tabela 3. Elementy składowej systematycznej

Sezonowość / Trend	Brak sezonowości	Sezonowość addytywna	Sezonowość multiplikatywna
Brak trendu			
Trend liniowy			
Trend nieliniowy			

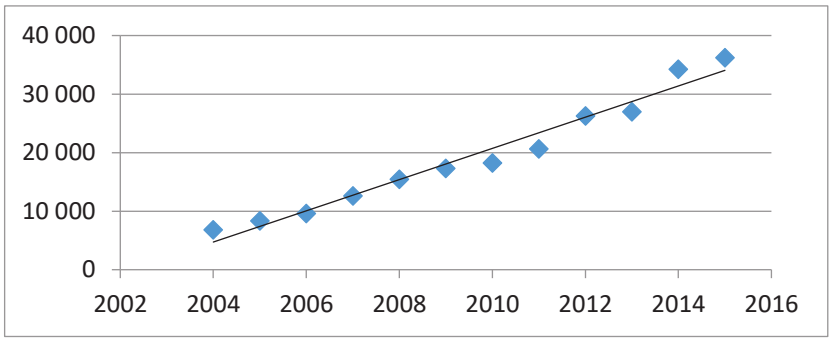
Źródło: ilustracja własna na podstawie M. Sobczyk [2021]

Czasami prezentacja graficzna badanego zjawiska umożliwia niemal natychmiastowe przeprowadzenie dekompozycji badanego szeregu. Analiza danych przedstawionych na rysunku 6 pozwala wyodrębnić stałą systematyczną (trend rosnący) oraz niewielkie wahania przypadkowe.

W praktyce jednak czasami trudno jednoznacznie stwierdzić, jakiego typu wahania występują przy eksploracji strumienia danych zawierających czas (rysunki 7 i 8).

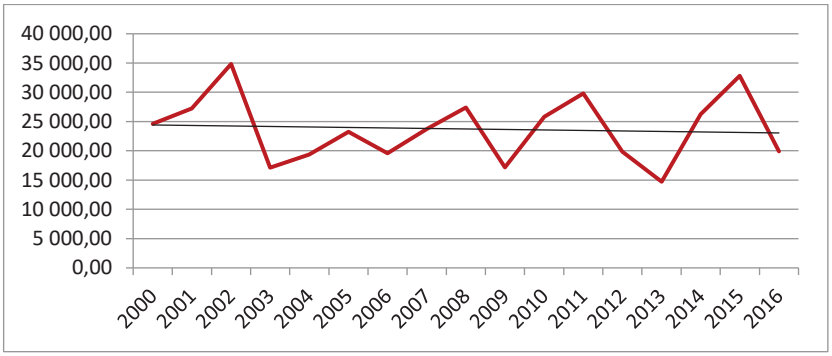
W szeregu prezentowanym na rysunku 7 nakreślona linia trendu pozwala stwierdzić występowanie nieznacznej tendencji spadkowej. W tym przypadku można również zaobserwować wyraźne wahania przypadkowe. Połączenie linią ciągłą surowych danych czasowych pozwala poprawić czytelność interpretacji wizualizacji. Ten komunikat graficzny był prawidłowo analizowany przez większość studentów w trakcie zajęć praktycznych.

Rysunek 6. Przykład 1 dekompozycji szeregu czasowego



Źródło: opracowanie własne na danych umownych

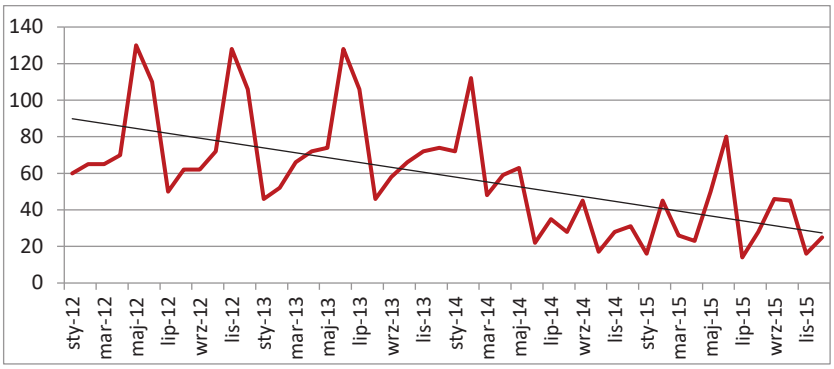
Rysunek 7. Przykład 2 dekompozycji szeregu czasowego



Źródło: opracowanie własne na danych umownych

Z kolei analizując wykres ilustrujący przykładowe dane przedstawione na rysunku 8, można zauważyć niestacjonarność wariancji tego szeregu. W okresie od stycznia 2012 do kwietnia 2014 występuje trend malejący oraz wahania sezonowe. W kolejnych miesiącach badane zjawisko nadal charakteryzuje tendencja spadkowa, ale bez wyraźnych wahań sezonowych, za to z wahaniami przypadkowymi.

Rysunek 8. Przykład 3 dekompozycji szeregu czasowego



Źródło: opracowanie własne na danych umownych

Eksploracja danych z przykładu 3 zestawionych jedynie w formie tabelarycznej sprawiała studentom trudność i skutkowała błędnymi wnioskami. W prawidłowo przeprowadzonej dekompozycji tego szeregu należy bowiem uwzględnić dwa różne okresy badawcze, co wyraźnie sugeruje komunikat graficzny.

WNIOSKI

Zaprezentowane przykłady z praktyki dydaktycznej pozwalają na sformułowanie wniosku o konieczności stosowania metod wizualizacji danych na każdym z etapów eksploracji. Ilustracja graficzna pozwala bowiem na wizualizację tego, czego nie można odczytać z tabeli zawierającej dane.

Należy uczyć studentów, że komunikat graficzny powinien być spójny i jak najbardziej czytelny dla odbiorcy oraz przedstawiać to, co istotne. Ponadto wykresy danych mogą stanowić również szczególnie użyteczne narzędzie podczas wyboru metod badawczych. Powinno się jednak pamiętać o krytycznej weryfikacji prezentacji graficznych pod kątem tego, co dokładnie na nich widać, a co może być zakłócone tym co chciałby zobaczyć badacz.

BIBLIOGRAFIA

Adams G. L. (1992) Why Interactive? Multimedia & Videodisc Monitor. Falls Church, Va, March 1992.

Anscombe F. J. (1973) Graphs in Statistical Analysis. American Statistician, s. 17-21.

Biecek P. (2014) Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce prezentowania danych. Fundacja Naukowa SmarterPoland, Warszawa.

Dittmann P. (2000) Metody prognozowania sprzedaży w przedsiębiorstwie. Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Lenik K., Dziedzic K., Czerkawska A. (2007) Wykorzystanie wybranych form graficznych do przedstawiania wyników badań doświadczalnych procesu tarcia i zużycia. Postępy Nauki i Techniki, 1, 25-32.

Tukey J. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Company.

Thalheimer W. (2006) People remember 10%, 20%...Oh Really?, www.willatworklearning.com (blog Willa'a Thalheimer'a Phd, Columbia University in the City of New York).

Sobczyk M. (2021) Statystyka. PWN, Warszawa.

VISUALIZATION AS AN INSTRUMENT OF SUPPORTING THE TEACHING PROCESS IN DATA ANALYSIS

Abstract Presentation of information in a graphical form is one of the basic forms of data presentation. It is a great support during both the preliminary and further analysis. However, an incorrect graphical form can lead to misinterpretation and, in consequence, to erroneous conclusions. This paper presents some examples of graphical data visualisation that come from authors' teaching experience. The article includes cases of both correct and incorrect data presentation.

Keywords: graphical data visualisation, time series, teaching experience

JEL classification: C19, Y10