Analiza próbek środowiskowych i przemysłowych

CHC015011l

(*opracował W. Zierkiewicz*)

Ćwiczenie: Zastosowanie macierzy współczynników korelacji liniowych oraz map ciepła do analizy wyników pomiarów.

1. WSTĘP

Większość zjawisk, np. przyrodniczych czy ekonomicznych, w otaczającym nas świecie występuje w różnorodnych związkach. Powiązania między nimi widać w prawach chemii, biologii, ekonomii czy innych dziedzin nauki. Statystyka dostarcza narzędzi, które pozwalają takie powiązania wykryć czy też zweryfikować. Jednym z takich narzędzi jest matematyczna teoria korelacji, która w sposób ilościowy określa stopień wzajemnego powiązania zmiennych. Współzależność pomiędzy zmiennymi może być funkcyjna (zmiana wartości jednej powoduje ściśle określoną zmianę wartości drugiej) lub stochastyczna (określonym wartościom jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennej). Korelacja jest miarą powiązania pomiędzy dwiema lub większą liczbą zmiennych. Współczynnik korelacji przyjmuje wartości z przedziału od -1 do 1. Wartość -1 świadczy o doskonałej korelacji ujemnej (zależności odwrotnie proporcjonalnej), natomiast wartość 1 oznacza doskonałą korelację dodatnią (zależność wprost proporcjonalną). Wartość 0 oznacza brak korelacji. Badanie zależności korelacyjnej ma sens jedynie wtedy, jeśli pomiędzy zmiennymi występuje związek przyczynowo-skutkowa, a rozkład wartości badanych zmiennych wykazuje cechy rozkładu normalnego. Zdarza się jednak, że nawet dla zmiennych silnie skorelowanych związek przyczynowo-skutkowy nie występuje (np. jednoczesny wzrost liczebności populacji bobrów europejskich i wzrost obrotu złotem w ostatnich latach, w Polsce). W takim przypadku mamy do czynienia ze współwystępowaniem zjawisk, a nie ze związkiem przyczynowo-skutkowym, zatem korelacja między tymi zmiennymi jest pozorna (iluzoryczna). Należy również pamiętać, że wartość korelacji liniowej bliska 0 nie musi oznaczać braku związku między zmiennymi, ponieważ może istnieć między nimi współzależność nieliniowa. Wskazane jest również sprawdzenie, czy badana zależność ma charakter liniowy. Obliczony współczynnik korelacji należy zweryfikować pod kątem jego istotności statystycznej. Weryfikacja polega na testowaniu hipotezy zerowej zakładającej brak korelacji liniowej pomiędzy badanymi zmiennymi, przy użyciu testu *t*-Studenta. W tym celu należy obliczyć wartość parametru statystycznego *t* (tzw. statystykę testową, zob. ćwiczenie 2.1.1.). Jeśli obliczona wartość bezwzględna *t* przewyższa wartość krytyczną odczytaną z tablic rozkładu *t*-Studenta (dla odpowiedniego stopnia swobody oraz poziomu istotności), wówczas hipotezę zerową należy odrzucić na rzecz hipotezy alternatywnej, co oznacza, że wykazany związek jest statystycznie istotny.

W przypadku dysponowania zbiorem wielu zmiennych można, korzystając np. z programu Excel i wbudowanego narzędzia do analizy danych, sporządzić macierz współczynników korelacji liniowej pomiędzy parami zmiennych, a następnie macierz zawierającą wartości *t* dla każdej pary zmiennych.

Dla większej klarowności interpretacji wyników taką macierz można przedstawić w formie mapy ciepła (heatmapy, heatplotu), gdzie kolor wskazuje - w przypadku macierzy współczynników korelacji liniowej - siłę i kierunek związku między zmiennymi, a w przypadku macierzy zawierającej wartości *t* - które z nich spełniają określony warunek (są większe od pewnej wartości krytycznej). Ponadto mapę ciepła można wykorzystać do wizualizacji danych przestrzennych lub czasowych, np. do przedstawienia zmiany stężenia zanieczyszczeń w powietrzu w cyklu dobowym.

2. ĆWICZENIA

2.1. Macierz współczynników korelacji liniowych pomiędzy wartościami wybranych parametrów opisujących jakość wody we Wrocławiu.

Jakość wody we Wrocławiu monitorowana jest przez Miejskie Przedsiębiorstwo Wodociągów i Kanalizacji (MPWiK), które regularnie publikuje dane o jakości wody z kilku punktów miasta. W Tabeli 1 zestawiono średnie wartości wybranych podstawowych parametrów jakości wody dostarczanej przez MPWiK we Wrocławiu, z trzech punktów pomiarowych, w okresie od I kwartału 2023 do I kwartału 2025 roku. Dane zostały zaczerpnięte ze strony MPWiK Wrocław (https://www.mpwik.wroc.pl/strefa-klienta/uslugi/uslugi-laboratoryjne/parametry-wody).

Tabela 1. Średnie wartości podstawowych parametrów jakości wody dostarczanej przez MPWiK Wrocław, z trzech punktów pomiarowych, w okresie od I kwartału 2023 do I kwartału 2025 roku.

Nr oznaczenia	Chlorki [mg/l]	Fluorki [mg/l]	Magnez [µg/l]	Odczyn pH	Przewodność ele. właściwa [µS/cm]	siarczany [mg/l]	Temperatura [°C]	Twardość ogólna [mg $CaCO_3$/l]	Wapń [mg/l]
	Cl	F	Mg	pH	pew	src	T	twr	Ca
1	49	0,19	13	7,5	625	93	11	270	86
2	45	0,18	14	7,7	592	86	4	288	91
3	79	0,22	22	7,6	851	149	10	343	100
4	47	0,20	13	7,5	622	84	11	253	80
5	57	0,20	16	7,5	741	110	6	381	126
6	91	0,26	25	7,6	1025	177	10,2	409	122
7	49,5	0,21	13	7,5	655	98	11	275	88
8	36	0,19	11	7,5	511	69	17	234	75
9	85	0,27	23	7,6	950	165	13	373	111
10	41,6	0,21	10	7,5	571	66	15	221	69
11	30	0,19	8	7,6	441	47	21	148	48
12	72	0,24	17	7,6	846	133	17	277	83
13	42,8	0,22	10	7,5	570	68	15	219	71

14	53	0,21	14	7,6	695	100	8	290	92
15	77	0,25	22	7,7	885	141	13	346	102
16	48	0,18	11	7,6	607	82	10	224	72
17	59	0,18	14	7,6	754	106	5	291	94
18	98	0,25	26	7,7	1073	187	10,6	406	119
19	51	0,21	12	7,5	655	98	10	253	81
20	43	0,20	13	7,5	590	84	15	268	86
21	90	0,29	27	7,6	1015	173	12	335	100
22	39	0,21	9	7,5	567	68	15	192	62
23	28	0,19	7	7,6	418	45	21	142	45
24	69	0,27	17	7,7	833	127	16	254	74
25	38	0,21	9	7,5	535	57	15	191	61
26	44	0,19	10	7,6	605	77	9	217	70
27	89	0,26	24	7,7	1001	168	13	372	108

2.1.1. Obliczenie współczynnika korelacji liniowej Pearsona dla pary parametrów: twardość ogólna oraz stężenie kationów wapnia.

Wiedząc o tym, że wartości parametrów: pew, src, T, twr oraz Ca, podanych w Tabeli 1 wykazują cechy rozkładu normalnego, dla dowolnych par tych parametrów można obliczyć współczynniki korelacji liniowej.

Oznaczając jako x twardość ogólną a jako y stężenie kationów wapnia obliczyć współczynnik korelacji liniowej Pearsona korzystając ze wzoru :

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

gdzie $\overline{x}$ i $\overline{y}$ oznaczają średnie arytmetyczne.

W kolejnym etapie należy sprawdzić czy obliczony współczynnik r jest istotny statystycznie dla założonego poziomu istotności (przyjąć wartość 0.05) oraz liczby oznaczeń (pomiarów). Polega to na weryfikacji hipotezy zerowej (H_0) zakładającej brak korelacji liniowej pomiędzy zmiennymi. W tym celu należy obliczyć z poniższego wzoru wartość statystyki testowej *t* :

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2}$$

gdzie n oznacza liczbę pomiarów.

Następnie należy porównać wartość bezwzględną wyliczonego wskaźnika *t* z wartością krytyczną znalezioną w tablicy rozkładu *t*-Studenta (dla założonego poziomu istotności oraz odpowiedniej liczby stopni swobody). Jeżeli wartość bezwzględna *t* jest większa od wartości krytycznej oznacza to, że dla założonego poziomu istotności obliczony współczynnik korelacji liniowej jest istotny statystycznie (co jest jednoznaczne z odrzuceniem hipotezy H_0). Na zakończenie sporządzić wykres

zależności twardość ogólnej wody od stężenia kationów wapnia (wykres punktowy), tak aby upewnić się że badana zależność ma charakter liniowy.

2.1.2. Sporządzenie macierzy współczynników korelacji liniowych dla par wybranych parametrów podanych w Tabeli 1.

Korzystając z programu Excel i wbudowanego narzędzia do analizy danych należy sporządzić macierz współczynników korelacji liniowych pomiędzy wybranymi wartościami parametrów jakości wody podanymi w Tabeli 1. Otrzymana macierz może wyglądać jak na Rys. 1.

	pew	src	T	twr
src	0.991	1.000		
T	-0.307	-0.306	1.000	
twr	0.890	0.907	-0.534	1.000
Ca	0.823	0.839	-0.611	0.987

Rys. 1. Macierz współczynników korelacji liniowych pomiędzy wartościami wybranych parametrów wody podanymi w Tabeli 1.

W kolejnym kroku należy sprawdzić, które z par parametrów wykazują istotną statystycznie korelację liniową. Zatem należy obliczyć wartości wskaźników *t* dla każdej z 10 par zmiennych. Nowo otrzymaną macierz (zawierającą wartości *t*) można poddać dalszej obróbce polegającej na zastosowaniu formatowania warunkowego, które pozwala wyróżnić te, które są większe od wartości *t* krytycznego (2,0595), a zatem są istotne statystycznie. Tak sporządzoną macierz ilustruje Rys. 2. Jest to przykład mapy ciepła (heatplotu), na której kolorem zielonym zaznaczono wartości większe od *t* krytycznego.

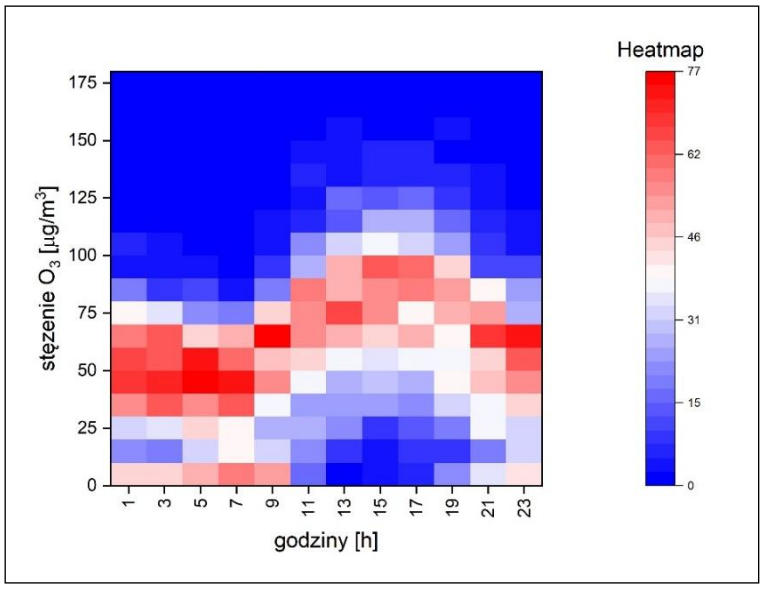
	pew	*src*	*T*	*twr*
src	36.535			
T	1.612	1.608		
twr	9.772	10.740	3.156	
Ca	7.250	7.715	3.862	30.780

Rys. 2. Macierz współczynników *t* każdej z par parametrów jakości wody podanych w Tabeli 1.

Na podstawie obu macierzy (Rys. 1 oraz 2) można wyciągnąć pewne wnioski np. taki, że pomiędzy twardością wody a stężeniem jonów wapnia istnieje bardzo silna dodatnia korelacja wynosząca 0.987, oraz taki, że część parametrów (pew oraz src) nie wykazuje istotnego związku statystycznego z temperaturą. Interpretując wyniki należy jednak pamiętać, że korelacja nie zawsze musi oznaczać, że pomiędzy zmiennymi występuje zależność przyczynowo-skutkowa.

2.2. Wykonanie mapy ciepła ilustrującej zmiany dobowe stężeń ozonu.

Monitoring jakości powietrza w ramach Państwowego Monitoringu Środowiska (PMŚ) obejmuje zadania związane z badaniem i oceną stanu zanieczyszczenia powietrza, w tym pomiary i oceny jakości powietrza. Na terenie Wrocławia zlokalizowanych jest kilka stacji pomiarowych. Jedną z nich jest stacja pomiarowa o nazwie Wrocław, wyb. Conrada-Korzeniowskiego. Na stacji prowadzone są pomiary automatyczne (1-godzinne) następujących zanieczyszczeń: pyły PM10 i PM2,5, ozon, oraz ditlenek siarki. Do przestawienia zmiany stężeń zanieczyszczeń w czasie może posłużyć mapa ciepła (heatplot). Przy jej pomocy można np. wykryć cykle dobowe lub sezonowe. W pliku o nazwie heatplot_cwiczenie_2_2.csv zostały zgromadzone wartości stężenia ozonu w powietrzu, na podstawie danych ze wspomnianej stacji pomiarowej, zmierzone w okresie od 01.01.2025 do 27.07.2025 w odstępach godzinnych **(**Źródło danych: GIOŚ – EKOINFONET, powietrze.gios.gov.pl). Dysponując danymi (w tym przypadku około 5000 pomiarów) należy wykonać analizę zmian dobowych stężeń ozonu w postaci mapy ciepła (Rys. 3). Do wykonania takiej mapy można posłużyć się programem Origin 2025.



Rys. 3. Mapa ciepła przedstawiająca rozkład czasowy wartości stężenia ozonu w powietrzu na podstawie pomiarów z okresu od 01.01.2025 do 27.07.2025
(Źródło danych: GIOŚ – EKOINFONET, powietrze.gios.gov.pl).

3. Zagadnienia

Zagadnienie 1.

W Tabeli 1 zaznaczono na szaro wiersze, w których podano wartości parametrów wody dla wybranego punktu pomiarowego (Zakład Produkcji Wody na „Grobli"). Dla tych wartość należy sporządzić macierz korelacji współczynników linowych a następnie macierz współczynników *t*, analogicznie jak to zostało omówione w ćwiczeniu 2.1.2 (dla tych samych parametrów: pew, src, T, twr oraz Ca). Proszę porównać wartości otrzymanych współczynników korelacji (istotnych statystycznie) z analogicznymi otrzymanymi w ćwiczeniu 2.1.2 (czyli dla wszystkich trzech punktów pomiarowych). Dla dwóch par parametrów, dla których wartość bezwzględna współczynnika korelacji liniowej jest największa oraz najmniejsza należy sporządzić wykresy punktowe.

Zagadnienie 2.

Na podstawie danych zgromadzonych w Tabeli 2 proszę sporządzić macierz korelacji pomiędzy parametrami powietrza. Czy wszystkie korelacje są statystycznie istotne? Proszę zaproponować interpretację dla jednej wybranej korelacji?

Tabela 2. Wartości wybranych parametrów powietrza na podstawie danych ze stacji pomiarowej: Wrocław, wyb. Conrada-Korzeniowskiego, zebranych w dniu 26.07.2025 (Źródło danych: GIOŚ – EKOINFONET, powietrze.gios.gov.pl**)**

Zanieczyszczenie	**PM10**	**PM2.5**	**O_3**	**SO_2**
Jednostka	**µg/m³**	**µg/m³**	**µg/m³**	**µg/m³**
26.07.2025. 01:00	22,3	15,5	39,3	13,2
26.07.2025. 02:00	23,0	16,5	18,8	12,6
26.07.2025. 03:00	26,0	19,2	11,8	11,7
26.07.2025. 04:00	22,2	17,1	21,0	13,5
26.07.2025. 05:00	18,9	14,5	39,8	13,5
26.07.2025. 06:00	23,6	17,1	33,2	13,5
26.07.2025. 07:00	27,4	19,8	36,3	11,8
26.07.2025. 08:00	22,5	16,1	49,1	12,8
26.07.2025. 09:00	19,9	14,9	63,1	13,4
26.07.2025. 10:00	21,2	14,4	72,6	13,9
26.07.2025. 11:00	19,8	14,1	77,6	14,1
26.07.2025. 12:00	22,2	14,3	83,6	14,7
26.07.2025. 13:00	19,6	13,8	86,0	15,7
26.07.2025. 14:00	16,6	13,6	88,9	14,6
26.07.2025. 15:00	16,3	13,0	93,6	15,8
26.07.2025. 16:00	14,0	11,8	97,9	16,0
26.07.2025. 17:00	15,1	11,8	95,6	16,4
26.07.2025. 18:00	18,1	13,4	94,0	15,8
26.07.2025. 19:00	18,4	14,0	93,3	15,7
26.07.2025. 20:00	21,3	14,8	78,7	15,6
26.07.2025. 21:00	23,6	16,3	64,4	15,2
26.07.2025. 22:00	24,2	16,4	67,0	14,7
26.07.2025. 23:00	24,2	15,0	65,1	14,4

Zagadnienie 3.

Proszę sporządzić mapę ciepła przedstawiająca rozkład czasowy wartości stężenia NO_2 w powietrzu korzystając z wyników pomiarów zgromadzonych przez stację pomiarową Wrocław, wyb. Conrada-Korzeniowskiego w okresie od 01.01.2025 do 27.07.2025 (odpowiednie dane do pobrania znajdują się w pliku o nazwie heatplot_zagadnienie_3.csv). Jakie można wyciągnąć wnioski na podstawie otrzymanej mapy?

Źródła.

1. T. Górecki, Podstawy statystyki z przykładami w R, Wyd. 1. Wydawnictwo BTC, Legionowo, 2011.

2. Strona internetowa Miejskiego Przedsiębiorstwa Wodociągów i Kanalizacji S.A. we Wrocławiu, https://www.mpwik.wroc.pl

3. Strona internetowa Głównego Inspektoratu Ochrony Środowiska, https://www.gov.pl/web/gios , Źródło danych: Dane gromadzone w systemie EKOINFONET GIOŚ

SPRAWOZDANIE

Sprawozdanie powinno zawierać:

1. Wstęp teoretyczny (do 1 strony) dotyczący zagadnień związanych z tematem ćwiczenia.
2. Przedyskutowane zagadnienia zaproponowane w instrukcji.
3. Krytyczna analiza uzyskanych wyników.
4. Podsumowanie.

ZAGADNIENIA DO KARTKÓWKI

1. Współzależność pomiędzy zmiennymi: funkcyjna, stochastyczna.
2. Współczynnik korelacji liniowej Pearsona.
3. Test *t*-Studenta, obliczanie wartość statystyki testowej *t*.
4. Korelacja liniowa przyczynowo-skutkowa oraz pozorna.
5. Liniowa korelacja dodatnia i ujemna.
6. Hipoteza zerowa.