

Data oddania: _____

Ocena: _____

Mateusz Walczak 216911

Konrad Kajszczyk 216790

Zadanie 2: Lingwistyczne podsumowania baz danych*

1. Cel

Celem zadania było zaprojektowanie aplikacji desktopowej, służącej do generowania podsumowań lingwistycznych oraz obliczania ich miar jakości dla wybranej bazy danych. Dodatkowo, aplikacja powinna posiadać graficzny interfejs użytkownika, który umożliwi intuicyjne korzystanie z programu.

2. Wprowadzenie

Rozważania we wprowadzeniu rozpoczniemy od *zbioru rozmytego*, czyli najbardziej podstawowego pojęcia, bez którego analiza działania lingwistycznych podsumowań baz danych nie byłaby możliwa. Przytoczmy zatem definicję *zbioru rozmytego*:

Definicja 1. Zbiór rozmyty A w niepustej przestrzeni X [5]

$$A = \{\langle x, \mu_A(x) \rangle : x \in X\}$$

gdzie $\mu_A(x) : \mathcal{X} \rightarrow [0, 1]$ nazywamy *funkcją przynależności do zbioru rozmytego A* .

W związku z faktem, iż pojęcie *funkcji przynależności* wystąpiło w powyższej definicji, w następnym podrozdziale zajmiemy się opisem tego rodzaju funkcji, wykorzystywanych przez nas.

* SVN: <https://github.com/Walducha1908/KSR2>

2.1. Funkcje przynależności

Funkcja przynależności określa w jakim stopniu dany element przynależy do zbioru. W zbiorach rozmytych zakres wartości jakie może ona przyjmować jest rozszerzony do przedziału $[0,1]$.

W naszym programie, posłużyliśmy się dwoma rodzajami funkcji przynależności:

- funkcją trójkątną opisaną wzorem:

$$f_{troj}(x) = \begin{cases} \frac{x-a}{b-a} & \text{jeśli } a \leq x < b \\ 1 & \text{jeśli } x = b \\ \frac{c-x}{c-b} & \text{jeśli } b < x \leq c \\ 0 & \text{w przeciwnym wypadku.} \end{cases} \quad (1)$$

- oraz funkcją trapezoidalną opisaną wzorem:

$$f_{trap}(x) = \begin{cases} \frac{x-a}{b-a} & \text{jeśli } a \leq x < b \\ 1 & \text{jeśli } b \leq x \leq c \\ \frac{d-x}{d-c} & \text{jeśli } c < x \leq d \\ 0 & \text{w przeciwnym wypadku.} \end{cases} \quad (2)$$

gdzie a, b, c oraz d są parametrami funkcji przynależności - wierzchołkami trójkąta lub trapezu na wykresie.

2.2. Podsumowania lingwistyczne

W tym rozdziale skoncentrujemy się na wyjaśnieniu czym są podsumowania lingwistyczne, które stanowią główny przedmiot rozważań tego zadania.

Ogólna postać lingwistycznego podsumowania bazy danych, prezentuje się następująco [4]:

$$Q \ P \text{ jest/są } S_j \ [T] \quad (3)$$

Omówmy poszczególne elementy powyższego wzoru:

- Q stanowi kwantyfikator, kwantyfikatory mogą być względne (np. *większość, prawie wszystkie, około połowa*) lub bezwzględne (np. *mniej niż 100, około 500*),
- P jest podmiotem podsumowania lingwistycznego, zestawem obiektów reprezentowanym przez krotki w bazie danych (np. *pomiary pogodowe, ludzie, samochody*),
- S_j jest sumaryzatorem, stanowi zbiór rozmyty na zbiorze wartości przyjmowanych przez j -tą kolumnę w bazie danych (np. *wysoka temperatura* w dziedzinie V_j $[-20, 40]$),
- T to stopień prawdziwości podsumowania (więcej informacji zamieszczono w rozdziale 2.3 dotyczącym miar jakości).

Zaprezentujemy przykład, dla omówionej postaci podsumowania lingwistycznego: *"Prawie wszyscy programiści zarabiają ponad 2000 złotych [0.88]"*, gdzie: *"Prawie wszyscy"* to kwantyfikator, *"programiści"* to podmiot lingwistyczny, *"zarabiają ponad 2000 złotych"* to sumaryzator, a *"[0.88]"* to stopień

prawdziwości podsumowania.

Aby zwiększyć stopień skomplikowania podsumowania lingwistycznego, można skorzystać ze złożonego sumaryzatora. Wtedy podsumowanie lingwistyczne przybiera postać:

$$Q \text{ } P \text{ jest/są } S_1 \text{ i/lub } S_2 \text{ i/lub } \dots \text{ i/lub } S_n \text{ } [T] \quad (4)$$

gdzie za pomocą słowa *lub* wyrażamy sumę sumaryzatorów zaś z wykorzystaniem słowa *i* - ich iloczyn.

Rozwińmy podany wcześniej przykład, tak aby wykorzystywał on sumaryzator złożony: *"Prawie wszyscy programiści zarabiają ponad 2000 złotych i mają bardzo drogi samochód [0.54]"*, sumaryzatorem złożonym w tym przykładzie jest *"zarabiają ponad 2000 złotych i mają bardzo drogi samochód"*.

Ostatnim etapem w naszej wędrówce po rozważaniach dotyczących podsumowań lingwistycznych, będzie podsumowanie z kwalifikatorem, które ma postać:

$$Q \text{ } P \text{ będących/mających własność } W \text{ jest/są } S_j \text{ } [T] \quad (5)$$

gdzie *W* jest kwalifikatorem - dodatkową właściwością (lub zestawem właściwości) posiadaną przez analizowane obiekty (np. *"[programiści] mający około 30 lat"*).

Kończąc nasze rozważania, zaprezentujemy jeszcze przytaczany wcześniej przykład, z wykorzystaniem kwalifikatora i sumaryzatora złożonego: *"Prawie wszyscy programiści mający około 30 lat zarabiają ponad 2000 złotych i mają bardzo drogi samochód [0.54]"*.

2.3. Miary jakości dla podsumowań lingwistycznych

Aby określić jakość naszych podsumowań zaimplementowaliśmy 11 miar jakości od T_1 do T_{11} . Poniżej zestawimy nazwy miar jakości wraz z oznaczeniami matematycznymi, którymi będziemy się posługiwać w następnych rozdziałach:

- T_1 - stopień prawdziwości,
- T_2 - stopień nieprecyzyjności,
- T_3 - stopień pokrycia,
- T_4 - stopień trafności,
- T_5 - długość podsumowania,
- T_6 - stopień nieprecyzyjności kwantyfikatora,
- T_7 - stopień licznosci kwantyfikatora,
- T_8 - stopień licznosci sumaryzator,
- T_9 - stopień nieprecyzyjności kwalifikatora,
- T_{10} - stopień licznosci kwalifikatora,
- T_{11} - długość kwalifikatora.

Dokładny opis miar i wzorów można znaleźć w rozdziałach 8.3 i 8.4 w monografii *Methods for the linguistic summarization of data - applications of fuzzy sets and their extensions*, Adam Niewiadomski [4].

3. Opis implementacji

Program został napisany w języku Java z wykorzystaniem narzędzia Maven [2], służącego do automatyzacji budowy oprogramowania. Graficzny interfejs użytkownika został zbudowany w oparciu o bibliotekę JavaFX [3].

W opisie implementacji skoncentrujemy się na jej najważniejszej części, czyli logice aplikacji.

W celu zbudowania struktury klas reprezentujących typy, parametry i własności zbiorów rozmytych oraz operacji na nich, zaimplementowaliśmy własną bibliotekę w formie pakietu w naszej aplikacji o nazwie *FuzzyLib*.

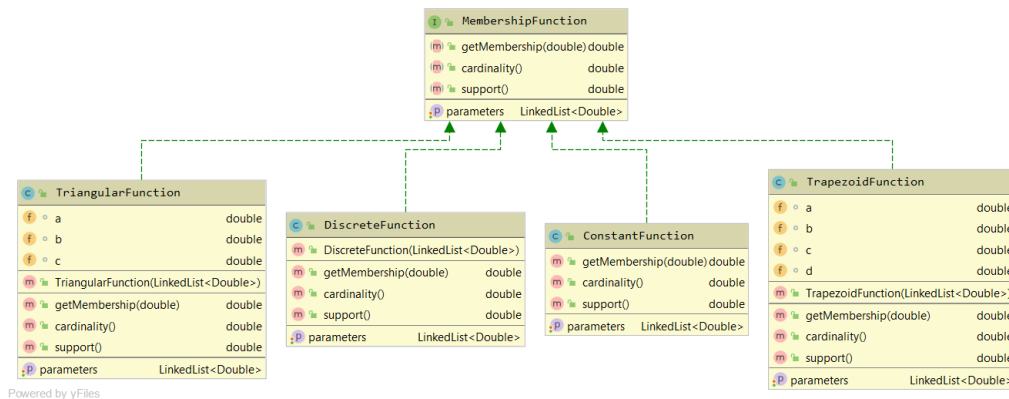
Pakiet *FuzzyLib* został podzielony na następujące podpakiety:

- *Membership*,
- *Logic*,
- *Containers*,
- *Summaries*.

W tym rozdziale omówione zostaną wszystkie wyżej wymienione podpakiety. Przedstawimy diagramy UML każdego z podpakietów a także omówimy zastosowanie poszczególnych klas.

3.1. Podpakiet Membership

Podpakiet *Membership* zawiera implementacje funkcji przynależności. Każda klasa tego podpakietu implementuje interfejs *MembershipFunction*.



Rysunek 1. Diagram UML dla podpakietu *Membership*

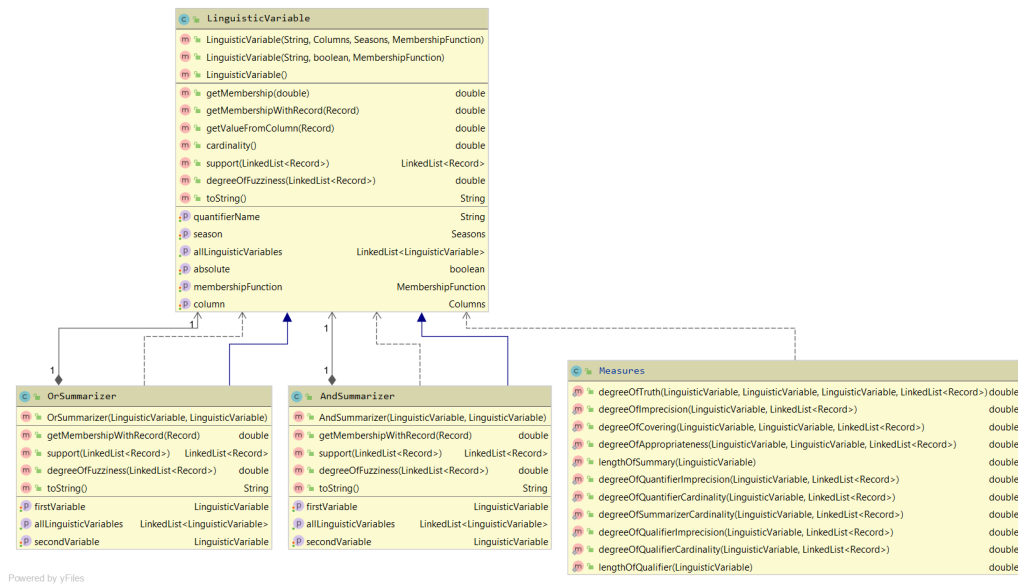
Klasy podpakietu *Membership* są następujące:

- *MembershipFunction* - interfejs implementowany przez wszystkie klasy należące do tego podpakietu,
- *TriangularFunction* - klasa implementująca trójkątną funkcję przynależności,

- *DiscreteFunction* - klasa implementująca dyskretną funkcję przynależności,
- *ConstantFunction* - klasa implementująca stałą funkcję przynależności,
- *TrapezoidFunction* - klasa implementująca trapezoidalną funkcję przynależności.

3.2. Podpakiet Logic

Podpakiet *Logic* odpowiada za implementacje zmiennych lingwistycznych, operacji sumy i iloczynu zbiorów rozmytych a także miar jakości podsumowań lingwistycznych.



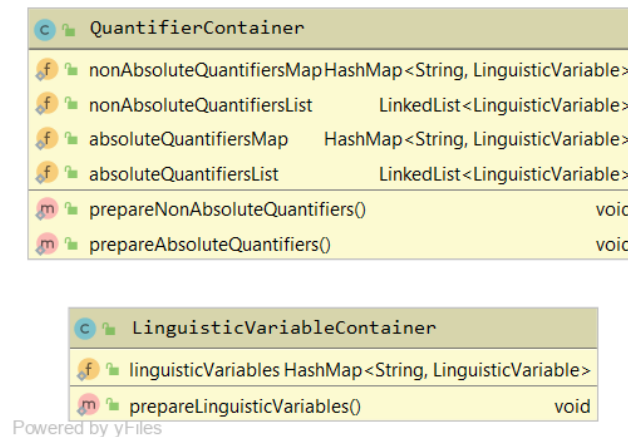
Rysunek 2. Diagram UML dla podpakietu *Logic*

Klasy podpakietu *Logic* są następujące:

- *LinguisticVariable* - klasa implementująca zmienną lingwistyczną,
- *AndSummarizer* - klasa implementująca operacje iloczynu, dziedziczy z klasy *LinguisticVariable*,
- *OrSummarizer* - klasa implementująca operacje sumy, dziedziczy z klasy *LinguisticVariable*,
- *Measures* - klasa implementująca miary jakości podsumowań lingwistycznych.

3.3. Podpakiet Containers

Podpakiet *Containers* zawiera klasy kontenerowe, tworzące i przetrzymujące zmienne lingwistyczne i kwantyfikatory.



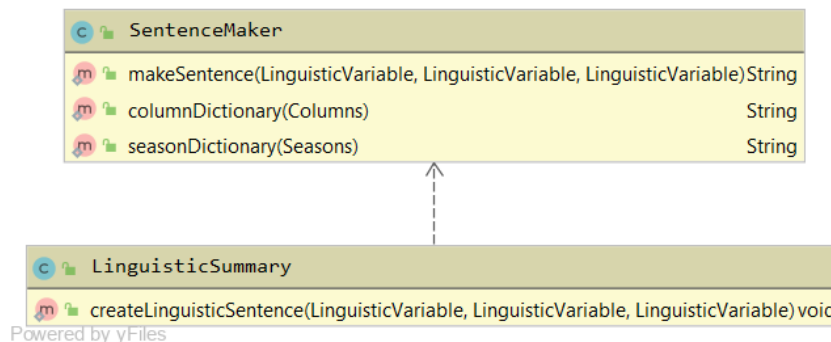
Rysunek 3. Diagram UML dla podpakietu *Containers*

Klasy podpakietu *Containers* są następujące:

- *LinguisticVariableContainer* - klasa kontenerowa dla zmiennych lingwistycznych,
- *QuantifierContainer* - klasa kontenerowa dla kwantyfikatorów.

3.4. Podpakiet Summaries

Podpakiet *Summaries* zawiera dwie klasy odpowiadające za budowę zdań podsumowania lingwistycznego w języku ludzkim (angielskim) oraz obliczenie wszystkich miar jakości dla danego podsumowania lingwistycznego - wywołanie statycznych metod klasy *Measures* podpakietu *Logic*.



Rysunek 4. Diagram UML dla podpakietu *Summaries*

Klasy podpakietu *Summaries* są następujące:

- *LinguisticSummary* - klasa odpowiadająca za obliczenie wszystkich miar jakości dla danego podsumowania,
- *SentenceMaker* - klasa odpowiadająca za budowanie zdań w języku ludzkim.

4. Materiały i metody

Wybrana przez nas baza danych zawiera historyczne pomiary pogodowe z Holandii [1]. Dane zostały zgromadzone przez KNMI (*Dutch weather institute* - Holenderski instytut pogody) na przestrzeni lat 1901-2018 i pochodzą z 50 różnych stacji pogody znajdujących się na terenie całego kraju.

Ze względu na fakt, iż oryginalna baza danych składa się z 804099 krotek, postanowiliśmy wybrać tylko niewielką część z dostępnych danych. Zdecydowaliśmy się na najnowsze dane pomiarowe - z lat 2016-2018. W ten sposób ograniczyliśmy liczbę wykorzystywanych krotek do 17000.

4.1. Wybór kolumn

W celu analizy bazy danych i tworzenia jej lingwistycznych podsumowań wybraliśmy 10 kolumn z danymi liczbowymi.

	A	B	C	D	E	F	G	H	I	J	K	L
1	STN	YYYYMMDD	FG	FHX	FHN	FXX	TG	TN	TX	T10N	Q	RH
2	380	20181231	23	30	10	60	83	74	91	70	96	19
3	370	20181231	27	40	20	70	89	73	99	72	115	5
4	350	20181231	28	40	20	80	89	75	98	72	132	5
5	375	20181231	28	50	20	90	90	73	98	71	105	1
6	290	20181231	36	60	20	90	88	74	99	73	126	2
7	275	20181231	30	50	20	90	86	72	98	70	132	1
8	279	20181231	40	60	20	100	86	78	97	77	150	-1
9	260	20181231	27	40	20	100	89	77	100	76	137	1
10	269	20181231	36	50	20	80	87	78	100	75	155	-1
11	280	20181231	40	60	20	110	88	80	99	77	174	5
12	240	20181231	45	70	30	110	90	83	109	76	205	-1
13	344	20181231	35	50	20	90	91	79	104	77	207	-1
14	215	20181231	42	60	20	90	89	82	97	80	212	0
15	235	20181231	51	80	40	110	90	83	98	77	229	-1
16	270	20181231	50	80	30	120	87	82	97	78	236	-1
17	310	20181231	48	60	30	90	88	73	102	71	335	3
18	375	20181230	32	40	20	90	79	53	92	42	137	31
19	350	20181230	29	50	20	80	82	56	93	49	139	14
20	260	20181230	26	50	20	100	85	62	95	52	112	10
21	370	20181230	33	50	20	90	80	51	91	45	156	11
22	269	20181230	33	60	20	110	82	63	92	54	119	8
23	344	20181230	36	60	20	120	86	50	97	36	141	4
24	215	20181230	43	60	20	100	87	57	99	48	127	7
25	275	20181230	35	70	20	120	77	55	90	48	138	24
26	279	20181230	39	80	20	160	76	67	84	62	148	10

Rysunek 5. Fragment widoku bazy w formacie *xlsx*

Wybrane kolumny są następujące:

- FG - średnia prędkość wiatru przez cały dzień [$0.1 \frac{m}{s}$].
- FHX - najwyższa średnia prędkość wiatru w ciągu jednej godziny [$0.1 \frac{m}{s}$].
- FHN - najniższa średnia prędkość wiatru w ciągu jednej godziny [$0.1 \frac{m}{s}$].
- FXX - najszybszy podmuch wiatru w ciągu całego dnia [$0.1 \frac{m}{s}$].
- TG - średnia dzienna temperatura [$0.1^{\circ}C$].
- TN - minimalna dzienna temperatura [$0.1^{\circ}C$].
- TX - maksymalna dzienna temperatura [$0.1^{\circ}C$].
- T10N - minimalna dzienna temperatura na wysokości 10 cm od poziomu gruntu [$0.1^{\circ}C$].

- Q - nasłonecznienie, energia słoneczna przypadająca na powierzchnię [$\frac{J}{cm^2}$].
- RH - suma opadów atmosferycznych w ciągu całego dnia [$0.1mm$].

Oprócz wyżej opisanych danych liczbowych, w naszej bazie znajdują się także dwie dodatkowe kolumny, służące do identyfikacji pomiaru:

- STN - numer stacji badawczej wykonującej pomiar.
- YYYYMMDD - data pomiaru w formacie opisanym przez nazwę kolumny.

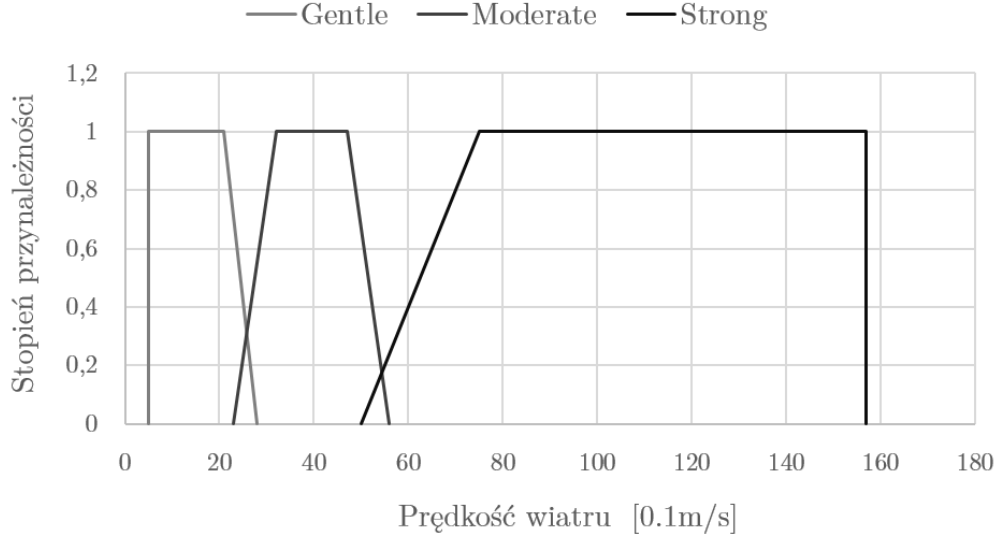
4.2. Zmienne lingwistyczne

W tym rozdziale przedstawimy wzory i wykresy opisujące wybrane, zaproponowane przez nas zmienne lingwistyczne. We wszystkich przypadkach, wykorzystywanymi przez nas funkcjami przynależności są funkcje trapezoidalne.¹

¹ Wartości prezentowane w tabelach są tylko propozycjami. Autorzy sprawozdania zastrzegają sobie możliwość do ich późniejszej modyfikacji

4.2.1. Kolumna FG

Wykres opisujący zmienną lingwistyczną dla kolumny zawierającej wartości średniej prędkości wiatru przez cały dzień (FG), zamieszczono poniżej.



Rysunek 6. Wykres opisujący zmienną lingwistyczną dla kolumny FG.

Wzory opisujące przynależność do poszczególnych etykiet prezentują się następująco.

Dla etykiety *Gentle*:

$$FG_{GENTLE}(x) = \begin{cases} 1 & \text{jeśli } 5 \leq x \leq 21 \\ \frac{28-x}{7} & \text{jeśli } 21 < x \leq 28 \end{cases} \quad (6)$$

Dla etykiety *Moderate*:

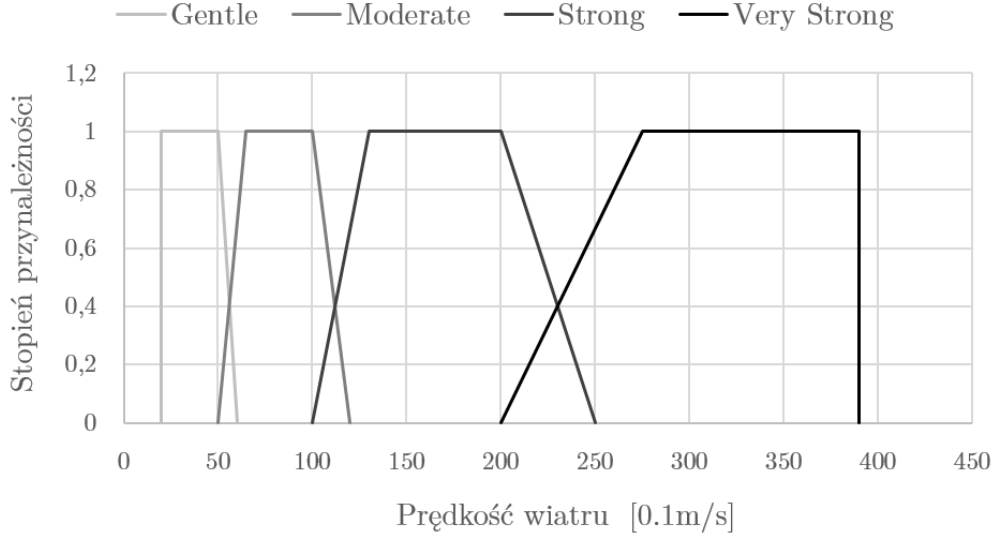
$$FG_{MODERATE}(x) = \begin{cases} \frac{x-23}{9} & \text{jeśli } 23 \leq x < 32 \\ 1 & \text{jeśli } 32 \leq x \leq 47 \\ \frac{56-x}{9} & \text{jeśli } 47 < x \leq 56 \end{cases} \quad (7)$$

Dla etykiety *Strong*:

$$FG_{STRONG}(x) = \begin{cases} \frac{x-50}{25} & \text{jeśli } 50 \leq x < 75 \\ 1 & \text{jeśli } 75 \leq x \leq 157 \end{cases} \quad (8)$$

4.2.2. Kolumna FXX

Wykres opisujący zmienną lingwistyczną dla kolumny zawierającej naj-
silniejszy powiew wiatru (FXX), zamieszczono poniżej.



Rysunek 7. Wykres opisujący zmienną lingwistyczną dla kolumny FXX.

Wzory opisujące przynależność do poszczególnych etykiet prezentują się następująco.

Dla etykiety *Gentle*:

$$FXX_{GENTLE}(x) = \begin{cases} 1 & \text{jeśli } 20 \leq x \leq 50 \\ \frac{60-x}{10} & \text{jeśli } 50 < x \leq 60 \end{cases} \quad (9)$$

Dla etykiety *Moderate*:

$$FXX_{MODERATE}(x) = \begin{cases} \frac{x-50}{15} & \text{jeśli } 50 \leq x < 65 \\ 1 & \text{jeśli } 65 \leq x \leq 100 \\ \frac{120-x}{20} & \text{jeśli } 100 < x \leq 120 \end{cases} \quad (10)$$

Dla etykiety *Strong*:

$$FXX_{STRONG}(x) = \begin{cases} \frac{x-100}{30} & \text{jeśli } 100 \leq x < 130 \\ 1 & \text{jeśli } 130 \leq x \leq 200 \\ \frac{250-x}{50} & \text{jeśli } 200 < x \leq 250 \end{cases} \quad (11)$$

Dla etykiety *Very Strong*:

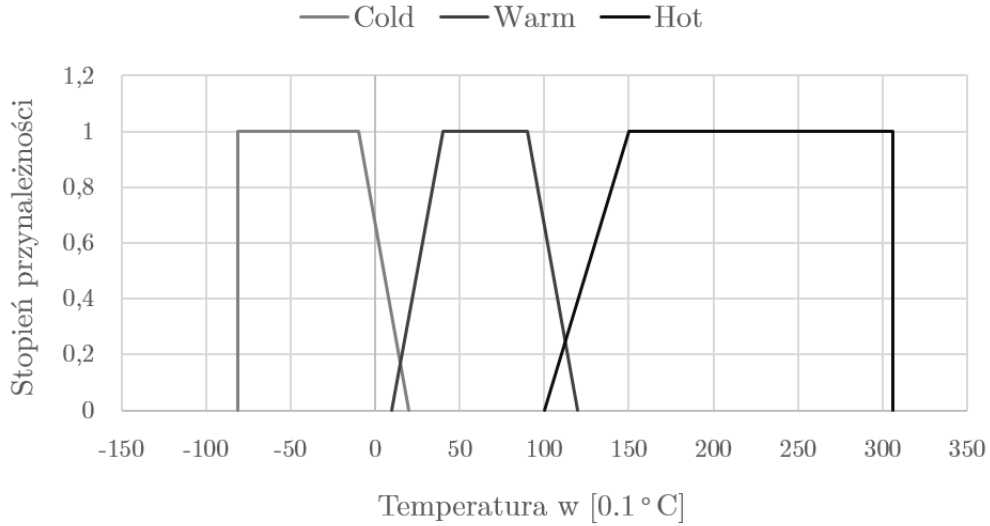
$$FXX_{VERYSTRONG}(x) = \begin{cases} \frac{x-200}{75} & \text{jeśli } 200 \leq x < 275 \\ 1 & \text{jeśli } 275 \leq x \leq 390 \end{cases} \quad (12)$$

4.2.3. Kolumna TG

W przypadku średniej dziennej temperatury (TG) oraz innych kolumn związanych z temperaturą (TN, TX oraz T10N), zdecydowaliśmy się podzielić nasze rozważania ze względu na pory roku. Dlatego też przyjęliśmy trzy różne warianty zmiennej lingwistycznej dla kolumny TG:

- TGW - dla pomiarów uzyskanych podczas astronomicznej zimy (litera *W* od *Winter*),
- TGSA - dla pomiarów uzyskanych podczas astronomicznej wiosny lub jesieni (*S* od *Spring*, *A* od *Autumn*),
- TGS - dla pomiarów uzyskanych podczas astronomicznego lata (litera *S* od *Summer*).

Rozpocznijmy od zmiennej lingwistycznej TGW.



Rysunek 8. Wykres opisujący zmienną lingwistyczną dla kolumny TG dla pomiarów wykonanych astronomiczną zimą.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TGW prezentują się następująco.

Dla etykiety *Cold*:

$$TGW_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -81 \leq x \leq -10 \\ \frac{20-x}{30} & \text{jeśli } -10 < x \leq 20 \end{cases} \quad (13)$$

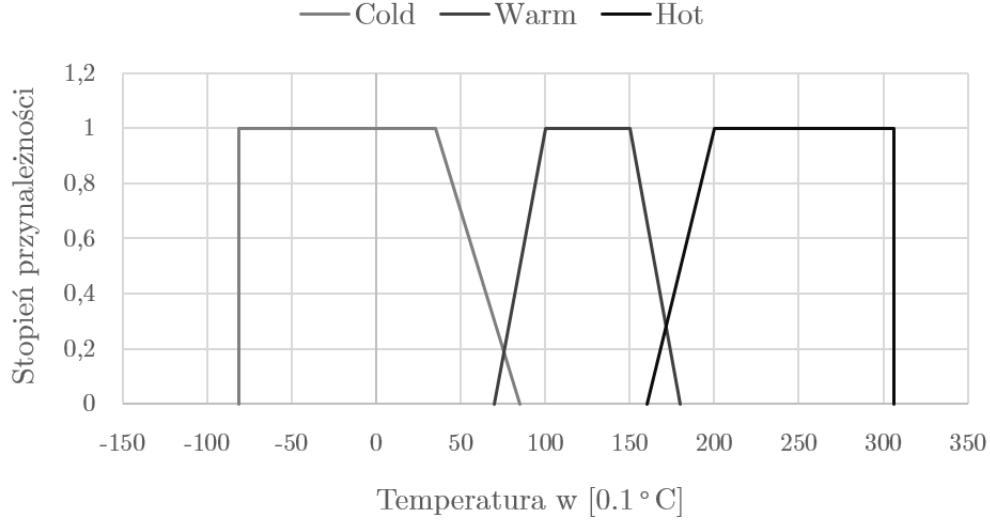
Dla etykiety *Warm*:

$$TGW_{WARM}(x) = \begin{cases} \frac{x-10}{30} & \text{jeśli } 10 \leq x < 40 \\ 1 & \text{jeśli } 40 \leq x \leq 90 \\ \frac{120-x}{30} & \text{jeśli } 90 < x \leq 120 \end{cases} \quad (14)$$

Dla etykiety *Hot*:

$$TGW_{HOT}(x) = \begin{cases} \frac{x-100}{50} & \text{jeśli } 100 \leq x < 150 \\ 1 & \text{jeśli } 150 \leq x \leq 306 \end{cases} \quad (15)$$

Następną prezentowaną zmienną, będzie zmienna lingwistyczna TGSA.



Rysunek 9. Wykres opisujący zmienną lingwistyczną dla kolumny TG dla pomiarów wykonanych astronomiczną wiosną i jesienią.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TGSA prezentują się następująco.

Dla etykiety *Cold*:

$$TGSA_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -81 \leq x \leq 35 \\ \frac{85-x}{50} & \text{jeśli } 35 < x \leq 85 \end{cases} \quad (16)$$

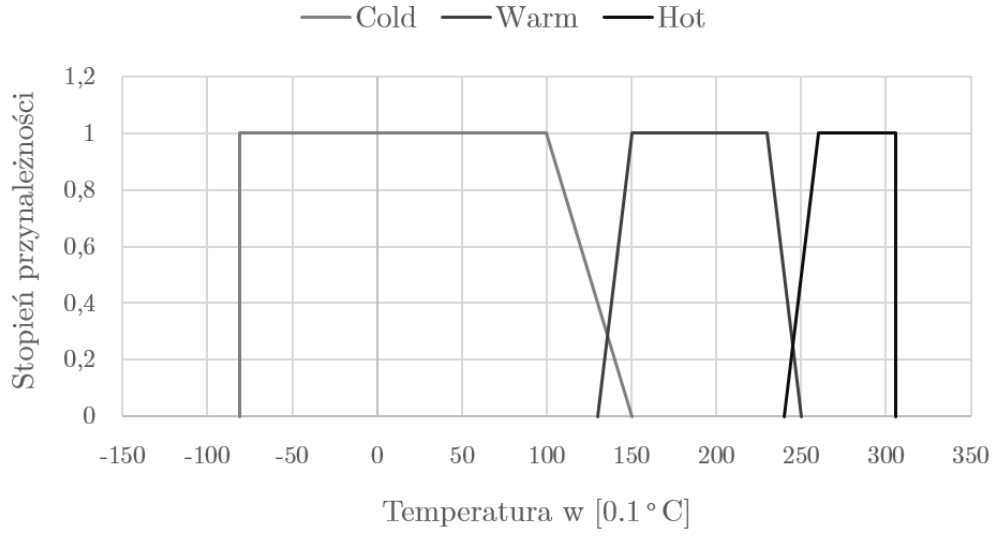
Dla etykiety *Warm*:

$$TGSA_{WARM}(x) = \begin{cases} \frac{x-70}{30} & \text{jeśli } 70 \leq x < 100 \\ 1 & \text{jeśli } 100 \leq x \leq 150 \\ \frac{180-x}{30} & \text{jeśli } 150 < x \leq 180 \end{cases} \quad (17)$$

Dla etykiety *Hot*:

$$TGSA_{HOT}(x) = \begin{cases} \frac{x-160}{40} & \text{jeśli } 160 \leq x < 200 \\ 1 & \text{jeśli } 200 \leq x \leq 306 \end{cases} \quad (18)$$

Ostatnią zmienną dla kolumny TG będzie zmienna dotycząca pomiarów letnich - TGS.



Rysunek 10. Wykres opisujący zmienną lingwistyczną dla kolumny TG dla pomiarów wykonanych astronomicznym latem.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TGS prezentują się następująco.

Dla etykiety *Cold*:

$$TGS_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -81 \leq x \leq 100 \\ \frac{150-x}{50} & \text{jeśli } 100 < x \leq 150 \end{cases} \quad (19)$$

Dla etykiety *Warm*:

$$TGS_{WARM}(x) = \begin{cases} \frac{x-130}{20} & \text{jeśli } 130 \leq x < 150 \\ 1 & \text{jeśli } 150 \leq x \leq 230 \\ \frac{250-x}{20} & \text{jeśli } 230 < x \leq 250 \end{cases} \quad (20)$$

Dla etykiety *Hot*:

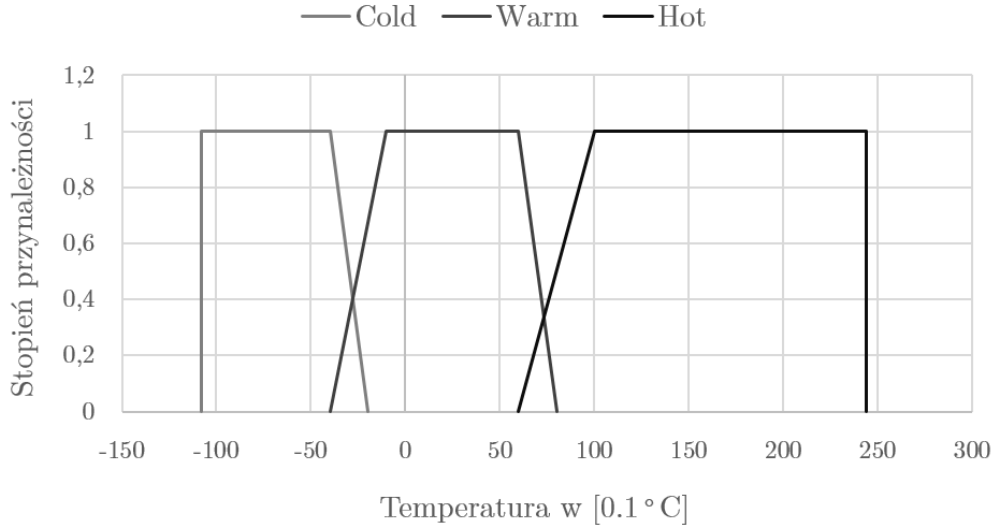
$$TGS_{HOT}(x) = \begin{cases} \frac{x-240}{20} & \text{jeśli } 240 \leq x < 260 \\ 1 & \text{jeśli } 260 \leq x \leq 306 \end{cases} \quad (21)$$

4.2.4. Kolumna TN

Kolumna TN zawiera najniższą temperaturę powietrza w ciągu dnia. Wszystkie trzy warianty zmiennej lingwistycznej dla kolumny TN zaprezentowano poniżej:

- TNW - dla pomiarów uzyskanych podczas astronomicznej zimy,
- TNSA - dla pomiarów uzyskanych podczas astronomicznej wiosny lub jesieni,
- TNS - dla pomiarów uzyskanych podczas astronomicznego lata.

Pomiary zimowe - zmienna TNW.



Rysunek 11. Wykres opisujący zmienną lingwistyczną dla kolumny TN dla pomiarów wykonanych astronomiczną zimą.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TNW prezentują się następująco.

Dla etykiety *Cold*:

$$TNW_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -108 \leq x \leq -40 \\ \frac{-20-x}{20} & \text{jeśli } -40 < x \leq -20 \end{cases} \quad (22)$$

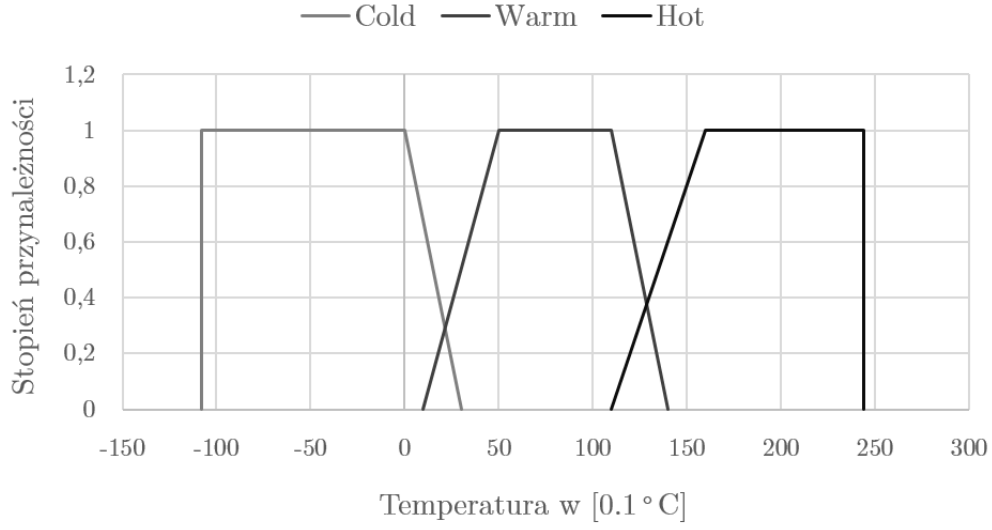
Dla etykiety *Warm*:

$$TNW_{WARM}(x) = \begin{cases} \frac{x+40}{30} & \text{jeśli } -40 \leq x < -10 \\ 1 & \text{jeśli } -10 \leq x \leq 60 \\ \frac{80-x}{20} & \text{jeśli } 60 < x \leq 80 \end{cases} \quad (23)$$

Dla etykiety *Hot*:

$$TNW_{HOT}(x) = \begin{cases} \frac{x-60}{40} & \text{jeśli } 60 \leq x < 100 \\ 1 & \text{jeśli } 100 \leq x \leq 244 \end{cases} \quad (24)$$

Pomiary wiosenne i jesienne - zmienna TNSA.



Rysunek 12. Wykres opisujący zmienną lingwistyczną dla kolumny TN dla pomiarów wykonanych astronomiczną wiosną i jesienią.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TNSA prezentują się następująco.

Dla etykiety *Cold*:

$$TNSA_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -108 \leq x \leq 0 \\ \frac{30-x}{30} & \text{jeśli } 0 < x \leq 30 \end{cases} \quad (25)$$

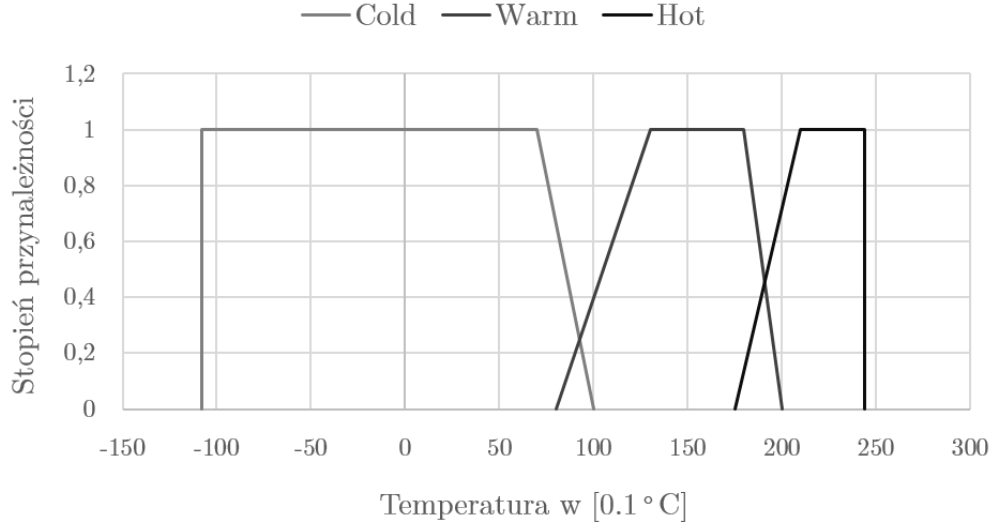
Dla etykiety *Warm*:

$$TNSA_{WARM}(x) = \begin{cases} \frac{x-10}{40} & \text{jeśli } 10 \leq x < 50 \\ 1 & \text{jeśli } 50 \leq x \leq 110 \\ \frac{140-x}{30} & \text{jeśli } 110 < x \leq 140 \end{cases} \quad (26)$$

Dla etykiety *Hot*:

$$TNSA_{HOT}(x) = \begin{cases} \frac{x-110}{50} & \text{jeśli } 110 \leq x < 160 \\ 1 & \text{jeśli } 160 \leq x \leq 244 \end{cases} \quad (27)$$

Pomiary letnie - zmienna TNS.



Rysunek 13. Wykres opisujący zmienną lingwistyczną dla kolumny TN dla pomiarów wykonanych astronomicznym latem.

Wzory opisujące przynależność do poszczególnych etykiet zmiennej lingwistycznej TNS prezentują się następująco.

Dla etykiety *Cold*:

$$TNS_{COLD}(x) = \begin{cases} 1 & \text{jeśli } -108 \leq x \leq 70 \\ \frac{100-x}{30} & \text{jeśli } 70 < x \leq 100 \end{cases} \quad (28)$$

Dla etykiety *Warm*:

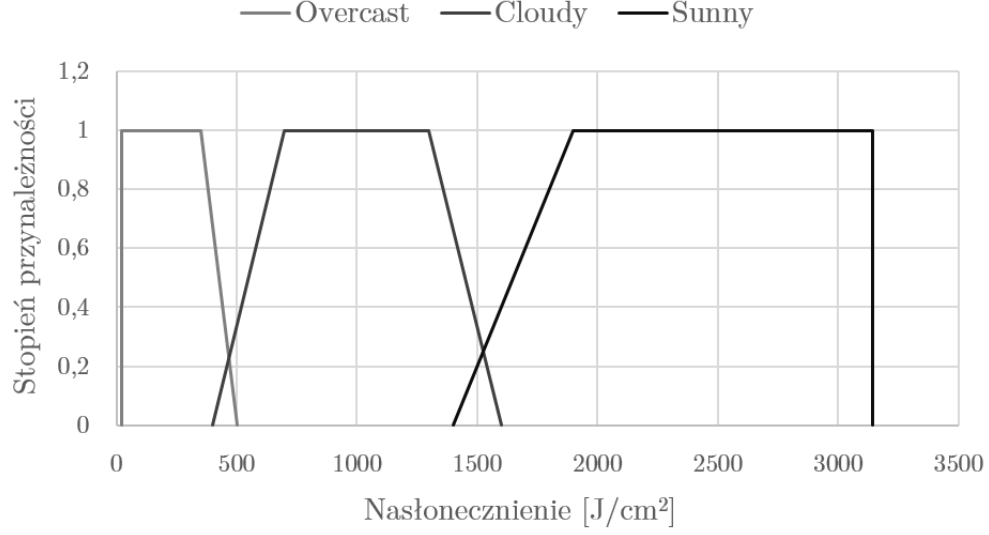
$$TNS_{WARM}(x) = \begin{cases} \frac{x-80}{50} & \text{jeśli } 80 \leq x < 130 \\ 1 & \text{jeśli } 130 \leq x \leq 180 \\ \frac{200-x}{20} & \text{jeśli } 180 < x \leq 200 \end{cases} \quad (29)$$

Dla etykiety *Hot*:

$$TNS_{HOT}(x) = \begin{cases} \frac{x-175}{35} & \text{jeśli } 175 \leq x < 210 \\ 1 & \text{jeśli } 210 \leq x \leq 244 \end{cases} \quad (30)$$

4.2.5. Kolumna Q

Wykres opisujący zmienną lingwistyczną dla kolumny zawierającej wartości nasłonecznienia (Q), zamieszczono poniżej.



Rysunek 14. Wykres opisujący zmienną lingwistyczną dla kolumny Q

Dla etykiety *Overcast*:

$$Q_{OVERCAST}(x) = \begin{cases} 1 & \text{jeśli } 24 \leq x \leq 350 \\ \frac{500-x}{150} & \text{jeśli } 350 < x \leq 500 \end{cases} \quad (31)$$

Dla etykiety *Cloudy*:

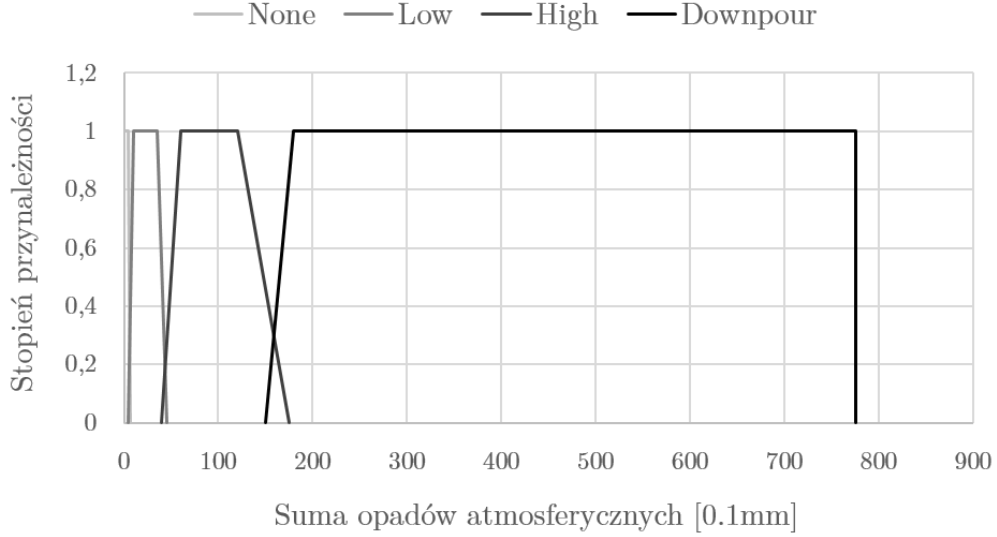
$$Q_{CLOUDY}(x) = \begin{cases} \frac{x-400}{300} & \text{jeśli } 400 \leq x < 700 \\ 1 & \text{jeśli } 700 \leq x \leq 1300 \\ \frac{1600-x}{300} & \text{jeśli } 1300 < x \leq 1600 \end{cases} \quad (32)$$

Dla etykiety *Sunny*:

$$Q_{SUNNY}(x) = \begin{cases} \frac{x-1400}{500} & \text{jeśli } 1400 \leq x < 1900 \\ 1 & \text{jeśli } 1900 \leq x \leq 3145 \end{cases} \quad (33)$$

4.2.6. Kolumna RH

Wykres opisujący zmienną lingwistyczną dla kolumny zawierającej sumę opadów atmosferycznych w ciągu całego dnia (RH), zamieszczono poniżej.



Rysunek 15. Wykres opisujący zmienną lingwistyczną dla kolumny RH

Wzory opisujące przynależność do poszczególnych etykiet prezentują się następująco.

Dla etykiety *None*:

$$RH_{NONE}(x) = \begin{cases} 1 & \text{jeśli } -1 \leq x \leq 5 \\ \frac{7-x}{2} & \text{jeśli } 5 < x \leq 7 \end{cases} \quad (34)$$

Dla etykiety *Low*:

$$RH_{LOW}(x) = \begin{cases} \frac{x-5}{5} & \text{jeśli } 5 \leq x < 10 \\ 1 & \text{jeśli } 10 \leq x \leq 35 \\ \frac{45-x}{10} & \text{jeśli } 35 < x \leq 45 \end{cases} \quad (35)$$

Dla etykiety *High*:

$$RH_{HIGH}(x) = \begin{cases} \frac{x-40}{20} & \text{jeśli } 40 \leq x < 60 \\ 1 & \text{jeśli } 60 \leq x \leq 120 \\ \frac{175-x}{55} & \text{jeśli } 120 < x \leq 175 \end{cases} \quad (36)$$

Dla etykiety *Downpour*:

$$RH_{DOWNPOUR}(x) = \begin{cases} \frac{x-150}{30} & \text{jeśli } 150 \leq x < 180 \\ 1 & \text{jeśli } 180 \leq x \leq 776 \end{cases} \quad (37)$$

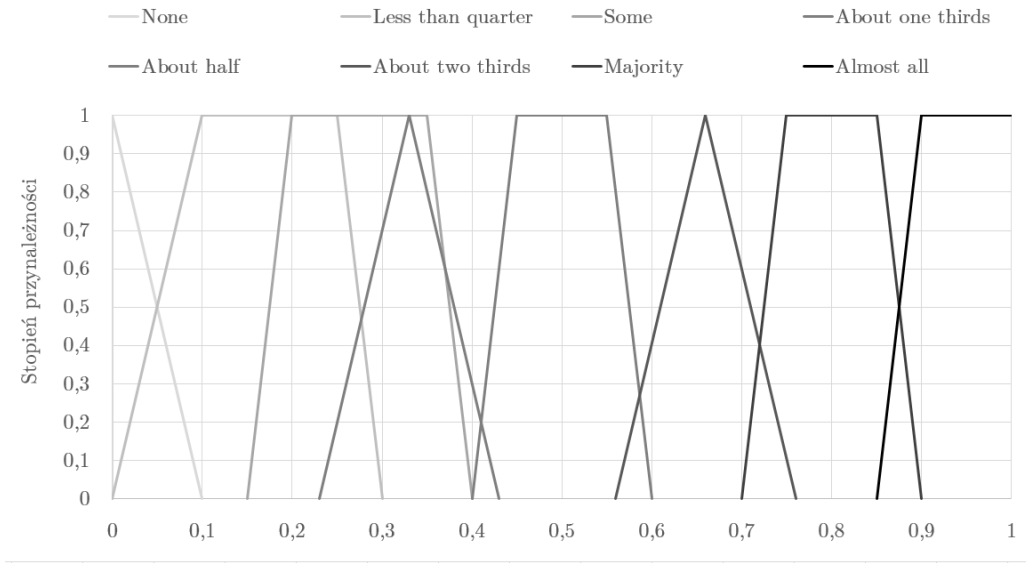
4.3. Kwantyfikatory

W tym rozdziale skoncentrujemy się na zaproponowanych przez nas kwantyfikatorach. Prezentację rozpoczniemy od kwantyfikatorów względnych, aby następnie omówić kwantyfikatory bezwzględne.

W przypadku kwantyfikatorów, wykorzystywanymi przez nas funkcjami przynależności są zarówno funkcje trapezoidalne jak i funkcje trójkątne. Wzory funkcji przynależności kwantyfikatorów zamieszczono pod wykresami.

4.3.1. Kwantyfikatory względne

Wykres ilustrujący wszystkie kwantyfikatory względne, zamieszczono poniżej.



Rysunek 16. Kwantyfikatory względne

Kwantyfikator *None* - funkcja trójkątna:

$$Qt_{NONE}(x) = \begin{cases} \frac{0.1-x}{0.1} & \text{jeśli } 0 \leq x \leq 0.1 \end{cases} \quad (38)$$

Kwantyfikator *Less than quarter* - funkcja trapezoidalna:

$$Qt_{LESS\ THAN\ QUARTER}(x) = \begin{cases} \frac{x}{0.1} & \text{jeśli } 0 \leq x < 0.1 \\ 1 & \text{jeśli } 0.1 \leq x \leq 0.25 \\ \frac{0.3-x}{0.05} & \text{jeśli } 0.25 < x \leq 0.3 \end{cases} \quad (39)$$

Kwantyfikator *Some* - funkcja trapezoidalna:

$$Qt_{SOME}(x) = \begin{cases} \frac{x-0.15}{0.05} & \text{jeśli } 0.15 \leq x < 0.2 \\ 1 & \text{jeśli } 0.2 \leq x \leq 0.35 \\ \frac{0.4-x}{0.05} & \text{jeśli } 0.35 < x \leq 0.4 \end{cases} \quad (40)$$

Kwantyfikator *About one thirds* - funkcja trójkątna:

$$Qt_{ABOUT\ ONE\ THIRDS}(x) = \begin{cases} \frac{x-0.23}{0.1} & \text{jeśli } 0.23 \leq x \leq 0.33 \\ \frac{0.43-x}{0.1} & \text{jeśli } 0.33 \leq x \leq 0.43 \end{cases} \quad (41)$$

Kwantyfikator *About half* - funkcja trapezoidalna:

$$Qt_{ABOUT HALF}(x) = \begin{cases} \frac{x-0.4}{0.05} & \text{jeśli } 0.4 \leq x < 0.45 \\ 1 & \text{jeśli } 0.45 \leq x \leq 0.55 \\ \frac{0.6-x}{0.05} & \text{jeśli } 0.55 < x \leq 0.6 \end{cases} \quad (42)$$

Kwantyfikator *About two thirds* - funkcja trójkątna:

$$Qt_{ABOUT TWO THIRDS}(x) = \begin{cases} \frac{x-0.56}{0.1} & \text{jeśli } 0.56 \leq x \leq 0.66 \\ \frac{0.76-x}{0.1} & \text{jeśli } 0.66 \leq x \leq 0.76 \end{cases} \quad (43)$$

Kwantyfikator *Majority* - funkcja trapezoidalna:

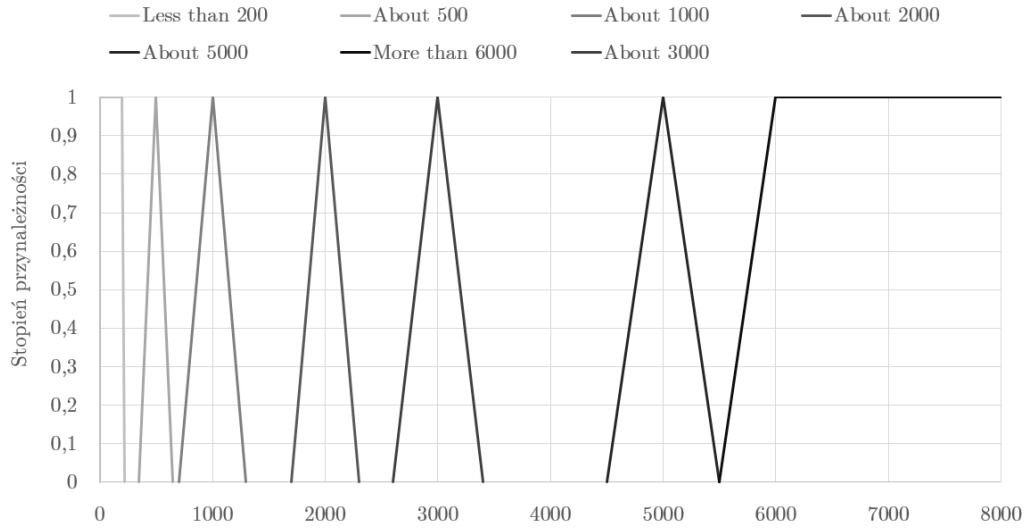
$$Qt_{MAJORITY}(x) = \begin{cases} \frac{x-0.7}{0.05} & \text{jeśli } 0.7 \leq x < 0.75 \\ 1 & \text{jeśli } 0.75 \leq x \leq 0.85 \\ \frac{0.9-x}{0.05} & \text{jeśli } 0.85 < x \leq 0.9 \end{cases} \quad (44)$$

Kwantyfikator *Almost all* - funkcja trapezoidalna:

$$Qt_{ALMOST ALL}(x) = \begin{cases} \frac{x-0.85}{0.05} & \text{jeśli } 0.85 \leq x < 0.9 \\ 1 & \text{jeśli } 0.9 \leq x \leq 1.0 \end{cases} \quad (45)$$

4.3.2. Kwantyfikatory bezwzględne

Wykres ilustrujący wszystkie kwantyfikatory bezwzględne, zamieszczono poniżej.



Rysunek 17. Kwantyfikatory bezwzględne

Kwantyfikator *Less than 200* - funkcja trapezoidalna:

$$Qt_{LESS THAN 200}(x) = \begin{cases} 1 & \text{jeśli } 0 \leq x \leq 200 \\ \frac{220-x}{20} & \text{jeśli } 200 < x \leq 220 \end{cases} \quad (46)$$

Kwantyfikator *About 500* - funkcja trójkątna:

$$Qt_{ABOUT 500}(x) = \begin{cases} \frac{x-350}{150} & \text{jeśli } 350 \leq x \leq 500 \\ \frac{650-x}{150} & \text{jeśli } 500 \leq x \leq 650 \end{cases} \quad (47)$$

Kwantyfikikator *About 1000* - funkcja trójkątna:

$$Qt_{ABOUT\ 1000}(x) = \begin{cases} \frac{x-700}{300} & \text{jeśli } 700 \leq x \leq 1000 \\ \frac{1300-x}{300} & \text{jeśli } 1000 \leq x \leq 1300 \end{cases} \quad (48)$$

Kwantyfikikator *About 2000* - funkcja trójkątna:

$$Qt_{ABOUT\ 2000}(x) = \begin{cases} \frac{x-1700}{300} & \text{jeśli } 1700 \leq x \leq 2000 \\ \frac{2300-x}{300} & \text{jeśli } 2000 \leq x \leq 2300 \end{cases} \quad (49)$$

Kwantyfikikator *About 3000* - funkcja trójkątna:

$$Qt_{ABOUT\ 3000}(x) = \begin{cases} \frac{x-2600}{400} & \text{jeśli } 2600 \leq x \leq 3000 \\ \frac{3400-x}{400} & \text{jeśli } 3000 \leq x \leq 3400 \end{cases} \quad (50)$$

Kwantyfikikator *About 5000* - funkcja trójkątna:

$$Qt_{ABOUT\ 5000}(x) = \begin{cases} \frac{x-4500}{500} & \text{jeśli } 4500 \leq x \leq 5000 \\ \frac{5500-x}{500} & \text{jeśli } 5000 \leq x \leq 5500 \end{cases} \quad (51)$$

Kwantyfikikator *More than 6000* - funkcja trapezoidalna:

$$Qt_{MORE\ THAN\ 6000}(x) = \begin{cases} \frac{x-5500}{500} & \text{jeśli } 5500 \leq x \leq 6000 \\ 1 & \text{jeśli } 6000 \leq x \leq 17000 \end{cases} \quad (52)$$

5. Wyniki

Przeprowadzone eksperymenty zostały podzielone na trzy etapy, w których porównywać będziemy różne zależności:

- W pierwszym etapie sprawdzimy jak zmieniać się będą miary podsumowań dla poszczególnych kwantyfikatorów.
- W drugim etapie przeanalizujemy zdania bardziej złożone - sumaryzatory wykorzystujące spójniki *AND* oraz *OR*.
- W trzecim etapie skoncentrujemy się na różnicy w rezultatach uzyskiwanych w wyniku wykorzystania kwalifikatora w opozycji do zdań, w których kwalifikator nie jest używany.

5.1. Wpływ wyboru kwantyfikatora na jakość podsumowania

W pierwszym etapie, wygenerowano następujące zdania:

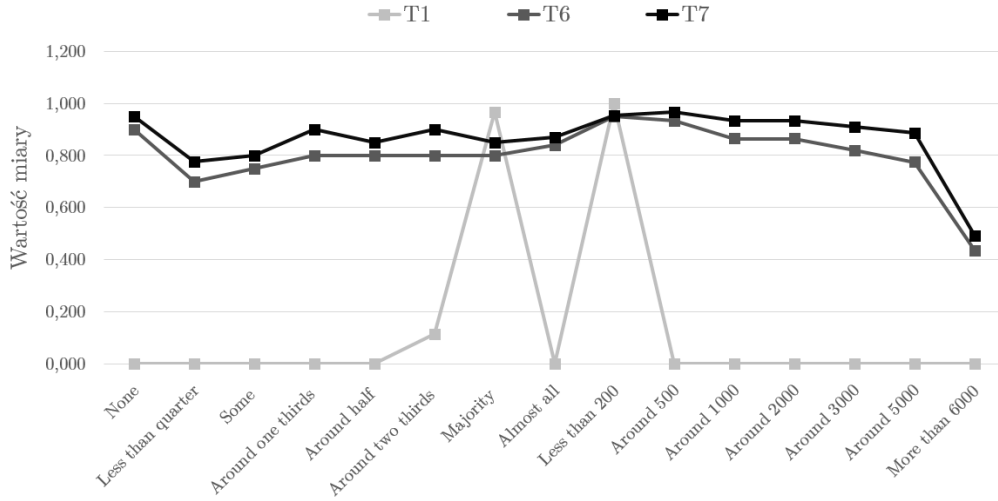
- Zdanie pierwsze: *X of summer measures with cold daily average temperature have cloudy insolation* - wykorzystujące zmienną lingwistyczną TGS jako kwalifikator (etykieta *Cold*) oraz zmienną Q jako sumaryzator (etykieta *Cloudy*).
- Zdanie drugie: *X of measures with very strong strongest wind blow have overcast insolation* - wykorzystujące zmienną lingwistyczną FXX jako kwalifikator (etykieta *Very strong*) oraz zmienną Q jako sumaryzator (etykieta *Overcast*).
- Zdanie trzecie: *X of measures with none precipitation have gentle daily wind speed average.* - wykorzystujące zmienną lingwistyczną RH jako kwalifikator (etykieta *None*) oraz zmienną FG jako sumaryzator (etykieta *Gentle*).

gdzie *X* jest jednym ze zdefiniowanych przez nas kwantyfikatorów.

Miary $T_2 - T_5$ oraz $T_8 - T_{11}$ osiągały stałe wartości w obrębie każdego ze zdań, dlatego też ich wartości zostały zamieszczone poza tabelami, w których zestawiono tylko te miary, których wartości były zmienne wewnątrz jednego eksperymentu.

5.1.1. Zdanie pierwsze

Wartości miar, które były jednakowe dla każdego kwantyfikatora, prezentują się następująco: $T_2 = 0,537$, $T_3 = 0,726$, $T_4 = 0,374$, $T_5 = 1$, $T_8 = 0,797$, $T_9 = 0,844$, $T_{10} = 0,954$, $T_{11} = 1$.



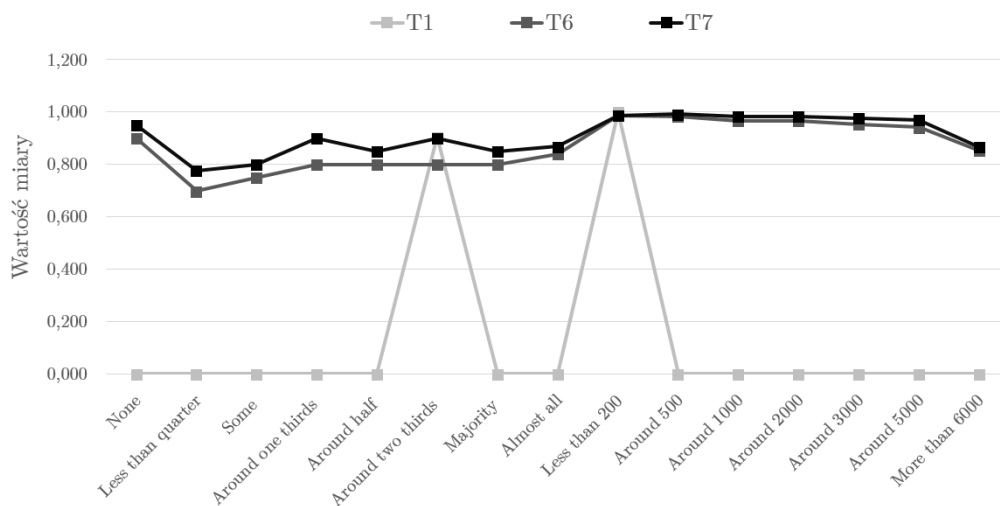
Rysunek 18. Wykres przedstawiający wyniki eksperymentu dla zdania pierwszego

Kwantyfikator	T_1	T_6	T_7
None	0,000	0,900	0,950
Less than quarter	0,000	0,700	0,775
Some	0,000	0,750	0,800
Around one thirds	0,000	0,800	0,900
Around half	0,000	0,800	0,850
Around two thirds	0,116	0,800	0,900
Majority	0,968	0,800	0,850
Almost all	0,000	0,840	0,870
Less than 200	1,000	0,950	0,953
Around 500	0,000	0,932	0,966
Around 1000	0,000	0,865	0,932
Around 2000	0,000	0,865	0,932
Around 3000	0,000	0,820	0,910
Around 5000	0,000	0,775	0,887
More than 6000	0,000	0,437	0,493

Tabela 1. Tabela przedstawiająca wyniki eksperymentu dla zdania pierwszego

5.1.2. Zdanie drugie

Wartości miar, które były jednakowe dla każdego kwantyfikatora, prezentują się następująco: $T_2 = 0,690$, $T_3 = 0,633$, $T_4 = 0,358$, $T_5 = 1$, $T_8 = 0,976$, $T_9 = 0,979$, $T_{10} = 0,991$, $T_{11} = 1$.



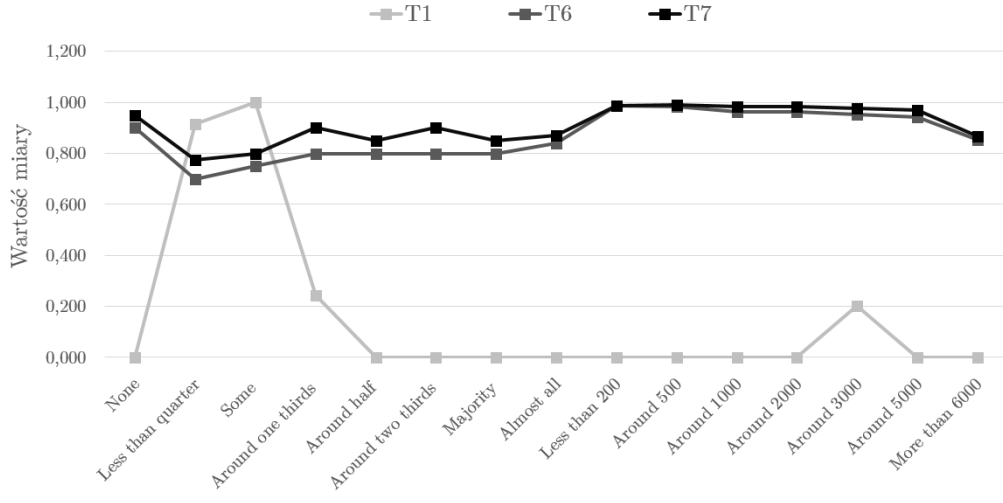
Rysunek 19. Wykres przedstawiający wyniki eksperymentu dla zdania drugiego

Kwantyfikator	T_1	T_6	T_7
None	0,000	0,900	0,950
Less than quarter	0,000	0,700	0,775
Some	0,000	0,750	0,800
Around one thirds	0,000	0,800	0,900
Around half	0,000	0,800	0,850
Around two thirds	0,905	0,800	0,900
Majority	0,000	0,800	0,850
Almost all	0,000	0,840	0,870
Less than 200	1,000	0,987	0,988
Around 500	0,000	0,982	0,991
Around 1000	0,000	0,965	0,982
Around 2000	0,000	0,965	0,982
Around 3000	0,000	0,953	0,976
Around 5000	0,000	0,941	0,971
More than 6000	0,000	0,853	0,868

Tabela 2. Tabela przedstawiająca wyniki eksperymentu dla zdania drugiego

5.1.3. Zdanie trzecie

Wartości miar, które były jednakowe dla każdego kwantyfikatora, prezentują się następująco: $T_2 = 0,742$, $T_3 = 0,332$, $T_4 = 0,141$, $T_5 = 1$, $T_8 = 0,999$, $T_9 = 0,373$, $T_{10} = 1$, $T_{11} = 1$.



Rysunek 20. Wykres przedstawiający wyniki eksperymentu dla zdania trzeciego

Kwantyfikator	T_1	T_6	T_7
None	0,000	0,900	0,950
Less than quarter	0,915	0,700	0,775
Some	1,000	0,750	0,800
Around one thirds	0,242	0,800	0,900
Around half	0,000	0,800	0,850
Around two thirds	0,000	0,800	0,900
Majority	0,000	0,800	0,850
Almost all	0,000	0,840	0,870
Less than 200	0,000	0,987	0,988
Around 500	0,000	0,982	0,991
Around 1000	0,000	0,965	0,982
Around 2000	0,000	0,965	0,982
Around 3000	0,201	0,953	0,976
Around 5000	0,000	0,941	0,971
More than 6000	0,000	0,853	0,868

Tabela 3. Tabela przedstawiająca wyniki eksperymentu dla zdania trzeciego

5.2. Wpływ wykorzystania operacji sumy i iloczynu na jakość podsumowania

Praca w toku

5.3. Wpływ wykorzystania kwalifikatora na jakość podsumowania

Praca w toku

6. Dyskusja

Praca w toku

7. Wnioski

Praca w toku

Literatura

- [1] Baza danych - "*Historical weather in the Netherlands 1901-2018*"
- [2] Narzędzie Maven
<https://maven.apache.org/>.
- [3] Biblioteka JavaFX
<https://openjfx.io/>
- [4] Methods for the linguistic summarization of data - applications of fuzzy sets and their extensions, Adam Niewiadomski, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008
- [5] Zadeh, L. A.: 1965, 'Fuzzy sets'. *Inf. and Control* 8, 338–353.