CASE STUDY – predykcja PM10 w Krakowie z wykorzystaniem sieci neuronowych

Natalia Kubańska 411933 13

Wstęp

Celem projektu była predykcja wartości stężenia PM10 w Krakowie. Analizowane dane składają się z dwóch plików xls: Krakow_pm10_2013.xls, Krakow_pm10_2014.xls. Ramka zawiera uśrednioną dzienną wartość dla 13 lokalizacji z czego dwie odpowiadają miastu Kraków. Ze względu na brak wskazań projektowych przyjęto koncepcje, w której w przedstawionej analizie skupiono się wyłącznie na wartości w jednym z punktów pomiarowych: Kraków, ul. Bujaka.

Pył zawieszony PM10; okres uśredniania wyników pomiarów 24-godz. (D_{24}=50 µg/m^3)						pomiary ciągle	bd	brak danych
Nazwa stacji / Data pomiaru	Kraków, ul. Bujaka	Kraków, ul. Bulwarowa	Tarnów, ul. Bitwy pod Studziankami	Nowy Sącz, ul. Nadbrzeżna	Bochnia, ul. Konfederatów Barskich	Gorlice, ul. Krasinskiego	Niepołomice, ul. 3 Maja	Proszowice, ul. Królewska
01.01.2013	142	138	25	112	21	30	bd	112
02.01.2013	82	64	56	150	78	65	bd	69
03.01.2013	24	25	22	53	16	27	bd	24
04.01.2013	22	22	26	18	24	14	20	22
05.01.2013	16	15	17	25	21	11	13	9
06.01.2013	27	31	24	25	33	18	32	25
07.01.2013	21	21	18	34	39	20	24	15
08.01.2013	90	76	34	110	56	33	88	77
09.01.2013	156	130	70	139	86	53	117	118
10.01.2013	63	57	28	74	24	47	38	35
11.01.2013	28	34	30	78	37	44	32	30
12.01.2013	23	48	24	39	24	27	20	28
13.01.2013	56	61	55	115	59	71	78	84
14.01.2013	94	96	86	151	232	81	89	91
15.01.2013	144	145	128	133	199	60	140	132
16.01.2013	152	124	136	134	127	53	105	89
17.01.2013	53	51	46	49	123	40	53	41
18.01.2013	34	45	38	38	38	46	48	35
19.01.2013	52	50	54	56	56	42	57	44
20.01.2013	102	122	111	225	123	57	112	105
21.01.2013	53	64	47	136	67	48	66	55
22.01.2013	68	57	44	46	56	42	72	51
23.01.2013	121	114	90	103	98	97	135	109
24.01.2013	215	226	153	206	168	150	255	233
25.01.2013	71	77	62	70	76	68	94	66

Figure 1: Fragment arkusza prezentujący strukturę danych

Preprocessing i eksploracyjna analiza danych

Oba z plików wczytano do R i sklejono w jedną ramkę danych. Kolumna Date przechowuje datę pomiaru, a PM10 wartość stężenia w jednostce µg/m3. W całej ramce dostępnych jest 730 obserwacji, 365 dla każdego roku.

	Date	PM10
1	01.01.2013	142
2	02.01.2013	82
3	03.01.2013	24
4	04.01.2013	22
5	05.01.2013	16
6	06.01.2013	27
7	07.01.2013	21
8	08.01.2013	90
9	09.01.2013	156

Figure 2: Wejściowe dane po sklejeniu

Kolejno sprawdzono typy kolumn i z typu character przekonwertowano je odpowiednio na format date oraz numeric. Po ponownym sprawdzeniu struktury zaobserwowano braki danych stanowiące 6% obserwacji, które oryginalnie były oznaczone jako „bd”.

```
      Date                 PM10       
 Min.   :2013-01-01   Min.   :  9.00  
 1st Qu.:2013-07-02   1st Qu.: 21.00  
 Median :2013-12-31   Median : 33.00  
 Mean   :2013-12-31   Mean   : 45.32  
 3rd Qu.:2014-07-01   3rd Qu.: 57.00  
 Max.   :2014-12-31   Max.   :228.00  
                      NA's   :45      
```

Figure 3: Podstawowe statystyki opisowe danych wejściowych

Po dokładniejszej analizie zauważono, że większość braków które się pojawia stanowi ciąg kolejno występujących po sobie dni. Przykładami są 2013-06-11 – 2013-06-13, 2013-08-07 - 2013-08-19, 2014-07-13 - 2014-07-21. Taki wzorzec może sugerować, że powodem braków mogą być jakieś powody techniczne jakimi są awaria czujników bądź rozładowanie baterii, która je zasila. Takie braki mogą zostać zaklasyfikowane jako missing not at random (brak danych zależy od jakiś nieznanych czynników). Konsekwencją tej sekwencji jest również dobór odpowiedniej metody imputacji brakujących wartości. Zdecydowano się na użycie średniej ruchomej z 14 dniowym oknem, które pozwoli na uzupełnienie wszystkich wartości i zachowanie zmienności danych.

Figure 4: Zmiana wartości PM10 z brakującymi danymi

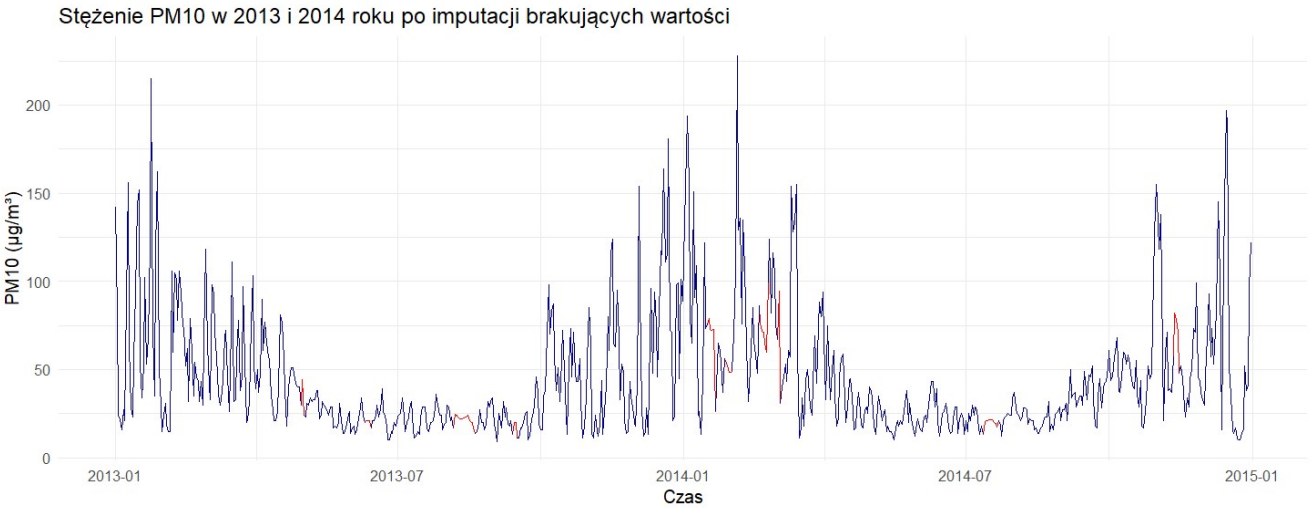


Figure 5: Zmiana wartości PM10 po uzupełnieniu braków

W nawiązaniu do poprzednio zaprezentowanych statystyk opisowych, dane nie przyjmują wartości, które mogłyby wskazywać na błędy pomiarowe i ich wartości zawierają się w sugerowanych zakresach PM10. Widoczne są dni, dla których wartość PM10 przekracza 200 µg/m3 jednak jest ich niewiele i wszystkie pojawiają się w okresie zimowym. Zauważono tendencję do wzrostu wartości stężenia w miesiącach zimowych i do spadku w letnich.

Indeks jakości powietrza	PM10 [µg/m^3]
Bardzo dobry	0 - 20
Dobry	20,1 - 50
Umiarkowany	50,1 - 80
Dostateczny	80,1 - 110
Zły	110,1 - 150
Bardzo zły	> 150

Figure 6: Indeks jakości powietrza

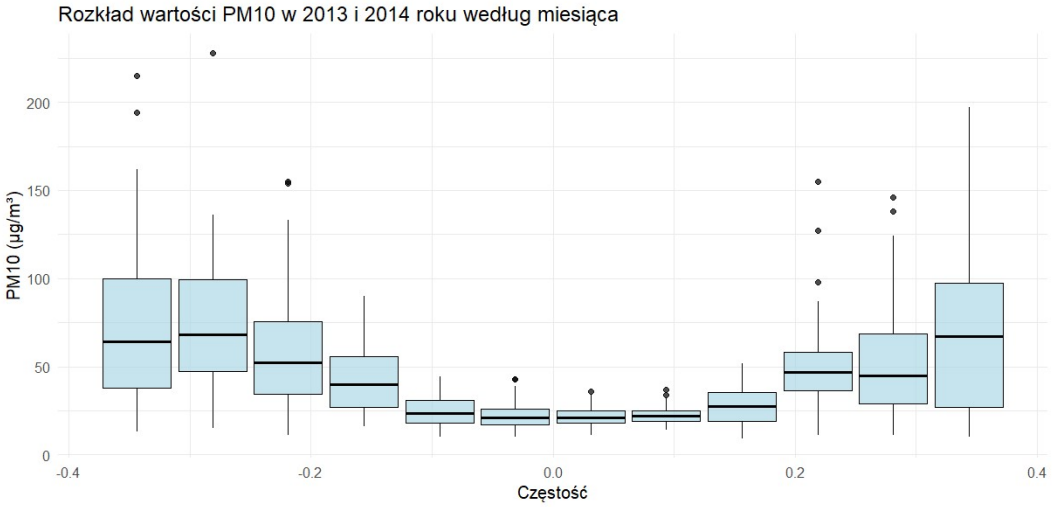


Figure 7: Boxploty dla każdego z miesięcy

Zarówno dla roku 2013 jak i 2014 rozkłady wartości są prawoskośne oraz wzajemnie do siebie podobne. Dla roku 2013 więcej obserwacji skupia się wokół wartości 25, ale poza tym rozkłady są do siebie bardzo podobne i nie wyróżnia się silnie tendencja wzrostowa/spadkowa pomiędzy latami. Widać, że wcześniej wspomniana większa ilość wartości w 2013, w 2014 rozkłada się przy tych wyższych wartościach, jednak zakres 2 lat nie jest wystarczający do sformułowania jednoznacznych wniosków.

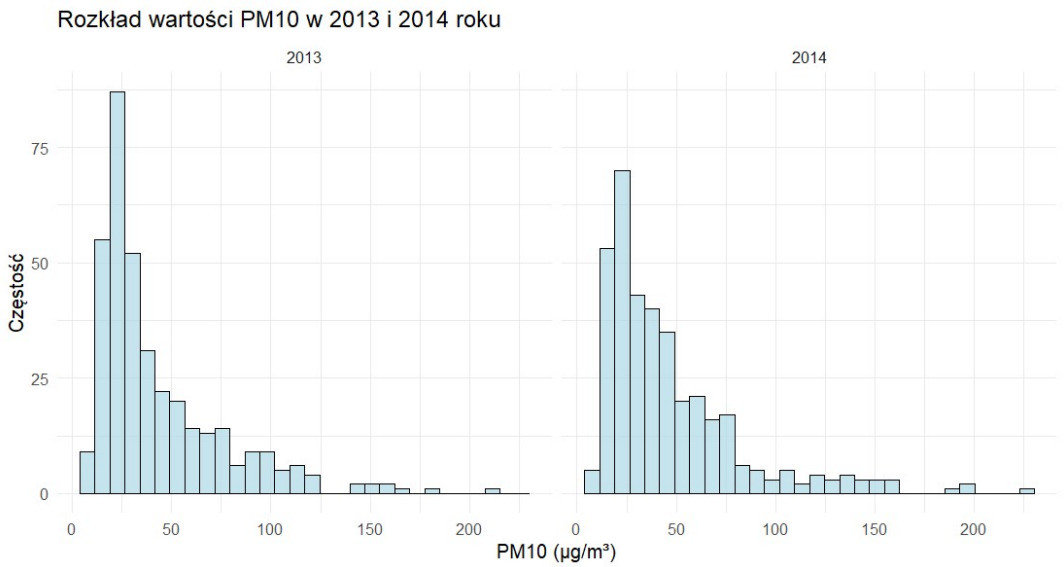


Figure 8: Histogramy PM10

Przeprowadzono rozszerzony test Dickeya-Fullera (ADF) i na podstawie jego wyników z p-value wynoszącym 0.01 odrzucono hipotezę zerową, że szereg nie jest stacjonarny.

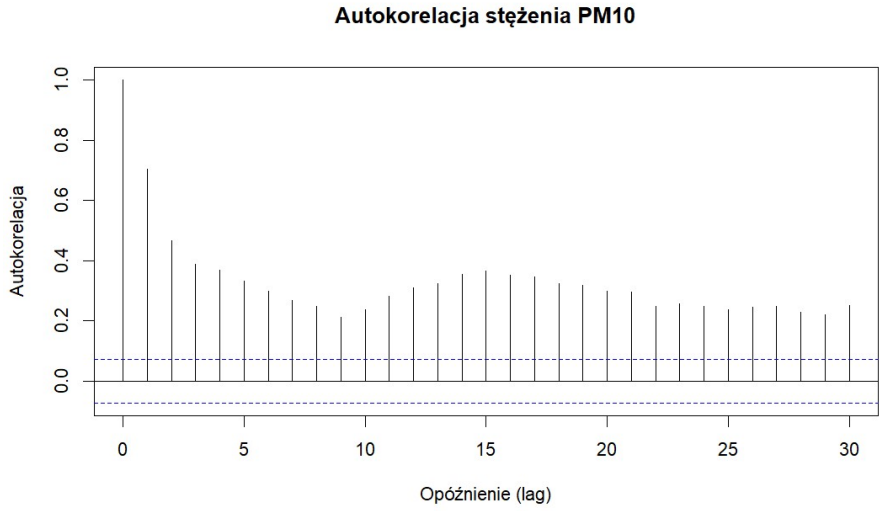


Figure 9: Wykres funkcji autokorelacji

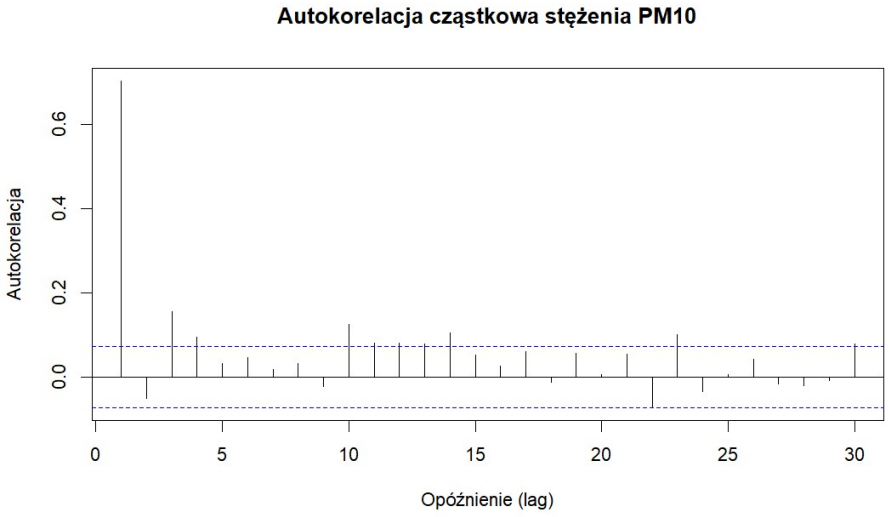


Figure 10: Wykres funkcji cząstkowej autokorelacji

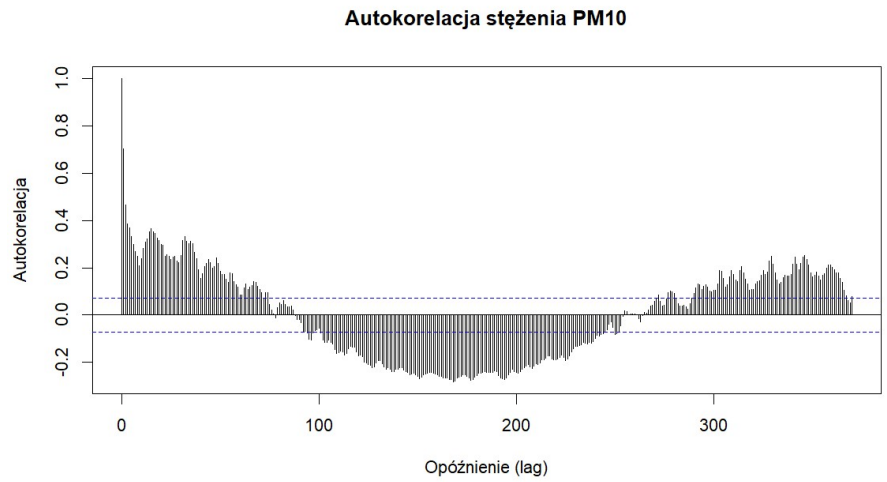


Figure 11: Wykres funkcji autokorelacji – maksymalne opóźnienie 370

Dla pierwszego opóźnienia wartość acf wynosi 0.702, a później stopniowo spada. W celu lepszego zrozumienia zależności i wcześniejszych podejrzeń co do wzrostu PM10 w sezonie jesienno-zimowym zwizualizowano również wartości dla maksymalnego opóźnienia 370 co potwierdza wcześniejsze obserwacje. Wartość funkcji autokorelacji spada i około 180 lagu po osiągnięciu swojego minimum zaczyna ponownie rosnąć. Obecna jest tutaj odwrotna zależność.

Wykres pacf obrazuje, że po uwzględnieniu zależności pierwszego opóźnienia wartość spada i nie ma stotnych zależności. Dla opóźnienia 3, 10 wartość funkcji przyjmuje również wartości powyżej ustawionego progu.

Tworzenie modeli predykcyjnych

Wylosowany model to sieci neuronowe. Takie modele przy uwzględnieniu odpowiedniej architektury mogą być używane do problemów regresyjnych, klasyfikacyjnych, klasteryzacyjnych, a także jak w tym przypadku do szeregów czasowych. Ze względu na brak wskazań zdecydowano się pierwszy model stworzyć przy wykorzystaniu wielowarstwowego perceptronu z pakietu *neuralnet,* a jako drugi model wykorzystać pakiet *keras* pozwalający na budowanie bardziej złożonych modeli, w tym

przypadku LSTM, którego architektura jest ukierunkowana na problemy sekwencyjne jak właśnie predykcja szeregów czasowych.

Na podstawie wcześniejszych wniosków jak zbiór cech, które mogą zostać użyte jako zmienne niezależne wybrano: lag1, lag3, lag10 oraz miesiąc reprezentowany przez zmienną numeryczną.

W celu normalizacji danych wykorzystano metodę min-max do której napisano swoją funkcję. Dane zostały podzielone na zbiór treningowy oraz testowy w proporcji 80:20. Dodatkowo utworzono dodatkowych zbiór jak zostało wskazane w poleceniu: zbiór testowy do predykcji, który składa się z 5 losowo wybranych obserwacji ze zbioru testowego.

Model 1

Do znalezienia odpowiednich cech, liczby neuronów oraz warstw, napisano funkcję która pozwala optymalnie przetestować różne parametry i zwraca wartość MAPE, który został wybrany na tym etapie jako błąd porównawczy.

```
[1] "MAPE, 1 : 0.344870273286545"
[1] "MAPE, 2 : 0.34530339335971"
[1] "MAPE, 3 : 0.345648359724326"
[1] "MAPE, 4 : 0.358701326475935"
[1] "MAPE, 5 : 0.346651629965908"
[1] "MAPE, 6 : 0.345740432016349"
[1] "MAPE, 7 : 0.346162156726555"
[1] "MAPE, 8 : 0.344562594531163"
[1] "MAPE, 9 : 0.345378836996884"
[1] "MAPE, 10 : 0.345697403596004"
[1] "MAPE, 11 : 0.345157577484652"
[1] "MAPE, 12 : 0.344480589911291"
[1] "MAPE, 13 : 0.364856758663"
[1] "MAPE, 14 : 0.346164354945102"
[1] "MAPE, 15 : 0.345246438457391"
[1] "MAPE, 16 : 0.345276484179451"
[1] "MAPE, 17 : 0.34593777303871"
[1] "MAPE, 18 : 0.344709614500451"
[1] "MAPE, 19 : 0.344980955328774"
[1] "MAPE, 20 : 0.345479341963111"
[1] "MAPE, 21 : 0.344394831973997"
[1] "MAPE, 22 : 0.345617717537353"
[1] "MAPE, 23 : 0.345543522387029"
[1] "MAPE, 24 : 0.345074736215536"
[1] "MAPE, 25 : 0.34491023816148"
[1] "MAPE, 26 : 0.345057578679841"
[1] "MAPE, 27 : 0.344767522917001"
[1] "MAPE, 28 : 0.345166346355503"
[1] "MAPE, 29 : 0.345016156823247"
[1] "MAPE, 30 : 0.345211421444649"
```

Figure 12: Przykładowe wartości MAPE dla różnej liczby neuronów modelu od lag1

```
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 8-11 MAPE: 0.341967993606608"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 1-4 MAPE: 0.344068166538634"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 5-11 MAPE: 0.342106577796618"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 16-15 MAPE: 0.342443873438351"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 14-1 MAPE: 0.344932805541499"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 8-6 MAPE: 0.340513531561991"
"Liczba warstw: 2 Liczba neuronów w każdej warstwie: 8-2 MAPE: 0.342607427510319"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 1-1-1 MAPE: 0.346723132220212"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 1-2-1 MAPE: 0.344731383601257"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 3-1-2 MAPE: 0.350846942790784"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 2-3-4 MAPE: 0.34598331461274"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 4-5-5 MAPE: 0.365064118343068"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 5-5-6 MAPE: 0.357244460056803"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 1-6-6 MAPE: 0.3468025855413"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 8-7-5 MAPE: 0.344371632247037"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 8-5-4 MAPE: 0.343646078902955"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 7-6-10 MAPE: 0.342853247453683"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 8-9-5 MAPE: 0.34497325795884"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 7-12-9 MAPE: 0.343138750862822"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 1-1-12 MAPE: 0.346769962219555"
"Liczba warstw: 3 Liczba neuronów w każdej warstwie: 7-3-8 MAPE: 0.345250338629428"
```

Figure 13: Przykładowe wartości MAPE dla różnej liczby neuronów i warstw modelu od lag1

```
[1] "MAPE, 1 : 0.352449367902843"
[1] "MAPE, 2 : 0.349391343006332"
[1] "MAPE, 3 : 0.347469050255317"
[1] "MAPE, 4 : 0.348555749610976"
[1] "MAPE, 5 : 0.35145563619494"
[1] "MAPE, 6 : 0.356191677672684"
[1] "MAPE, 7 : 0.349415604576356"
[1] "MAPE, 8 : 0.350976057559683"
[1] "MAPE, 9 : 0.353250956117116"
[1] "MAPE, 10 : 0.362630209313844"
[1] "MAPE, 11 : 0.352482103450756"
[1] "MAPE, 12 : 0.354321453418337"
[1] "MAPE, 13 : 0.356776678817619"
[1] "MAPE, 14 : 0.351240514995849"
[1] "MAPE, 15 : 0.357066798501489"
[1] "MAPE, 16 : 0.354209997133582"
[1] "MAPE, 17 : 0.348465243961704"
[1] "MAPE, 18 : 0.344718684089964"
[1] "MAPE, 19 : 0.361660570471971"
[1] "MAPE, 20 : 0.355600508720027"
```

Figure 14: Przykładowe wartości MAPE dla różnej liczby neuronów modelu od lag1 + lag3

```
[1] "MAPE, 1 : 0.370528704282731"
[1] "MAPE, 2 : 0.375944715916415"
[1] "MAPE, 3 : 0.383382651940613"
[1] "MAPE, 4 : 0.39434380052467"
[1] "MAPE, 5 : 0.392227986015787"
[1] "MAPE, 6 : 0.398856471417435"
[1] "MAPE, 7 : 0.44657847864593"
[1] "MAPE, 8 : 0.404383380018993"
```

Figure 15: Przykładowe wartości MAPE dla różnej liczby neuronów modelu od lag1 + lag3 +lag10

```
[1] "MAPE, 1 : 0.321741591283735"
[1] "MAPE, 2 : 0.374688288210298"
[1] "MAPE, 3 : 0.381475156650492"
[1] "MAPE, 4 : 0.39278438682685"
[1] "MAPE, 5 : 0.402597368813441"
[1] "MAPE, 6 : 0.39879080280633"
[1] "MAPE, 7 : 0.393776731111095"
[1] "MAPE, 8 : 0.394711589878031"
[1] "MAPE, 9 : 0.391828654961668"
[1] "MAPE, 10 : 0.395877010454137"
[1] "MAPE, 11 : 0.381917229424221"
[1] "MAPE, 12 : 0.391952368184012"
[1] "MAPE, 13 : 0.388891203187272"
[1] "MAPE, 14 : 0.390949826661003"
```

Figure 16: Przykładowe wartości MAPE dla różnej liczby neuronów modelu od lag1 + Month

```
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 1-1 MAPE: 0.326933043037594"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 2-2 MAPE: 0.385902032700131"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 3-1 MAPE: 0.325786863259395"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 4-1 MAPE: 0.322143644857774"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 1-4 MAPE: 0.326840952698939"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 2-4 MAPE: 0.325852646682206"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 6-1 MAPE: 0.403361749905383"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 6-1 MAPE: 0.387214510081081"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 7-1 MAPE: 0.3735004584232"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 5-3 MAPE: 0.395308271226741"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 6-11 MAPE: 0.398803269365148"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 6-10 MAPE: 0.399686460973475"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 6-11 MAPE: 0.400036901147036"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 2-11 MAPE: 0.383822287709733"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 4-2 MAPE: 0.387983160510585"
[1] "Liczba warstw: 2 Liczba neuronów w każdej warstwie: 3-15 MAPE: 0.379677665553289"
```

Figure 17: Przykładowe wartości MAPE dla różnej liczby neuronów i warstw modelu od lag1 + Month

Po przetestowaniu wielu cech zauważono, że największy wpływ odgrywa lag1 co było również widoczne na wykresie autokorelacji. Dołożenie pozostałych cech opóźnień nie poprawia wielkości błędu, a czasem go jeszcze bardziej pogłębia. Najlepsze rezultaty były widoczne w modelu zależnym od lag1 oraz miesiąca przy dwóch warstwach i finalnie zdecydowano się na model z dwiema warstwami ukrytymi dla którego MAPE wyniosło około 0.32.

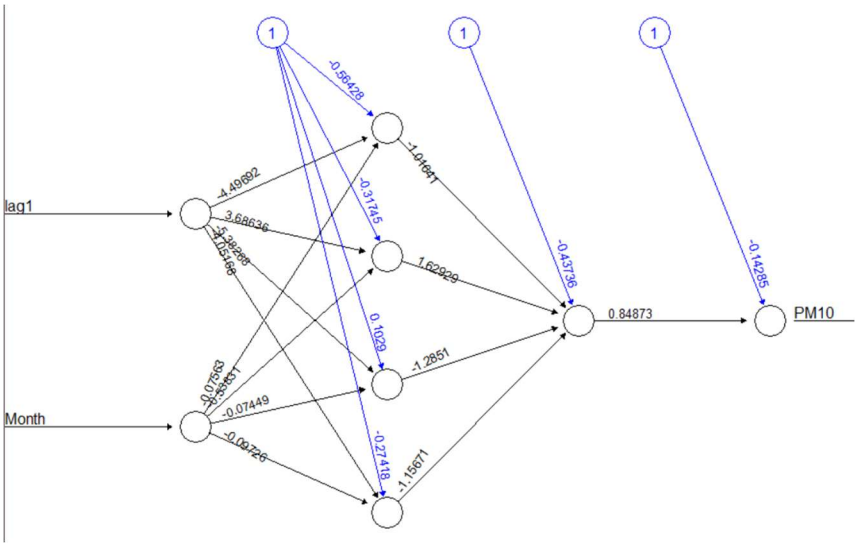


Figure 18: Architektura ostatecznego modelu

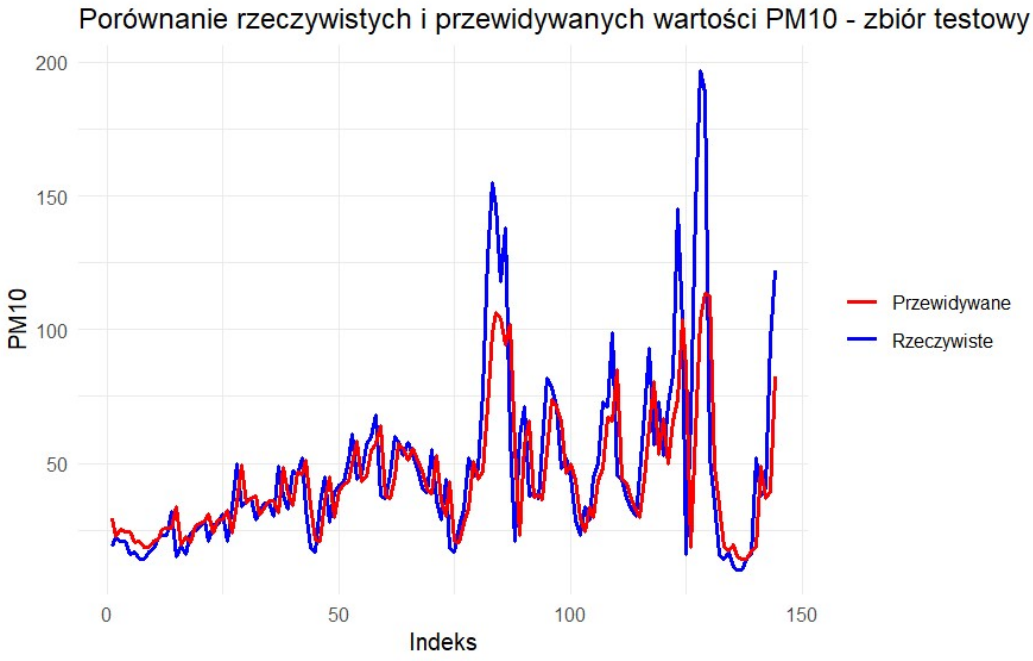
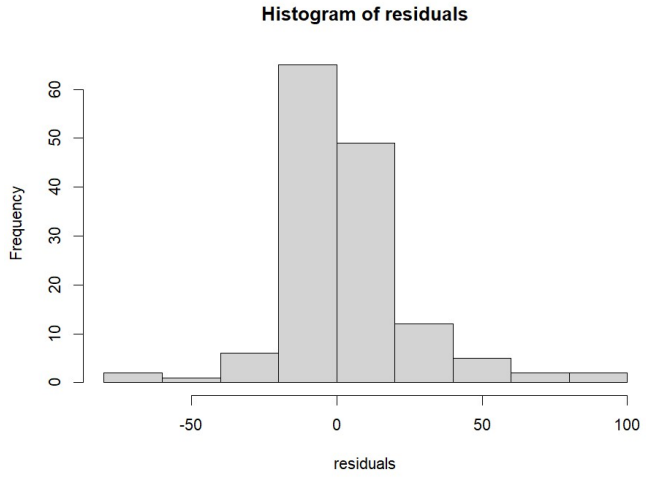


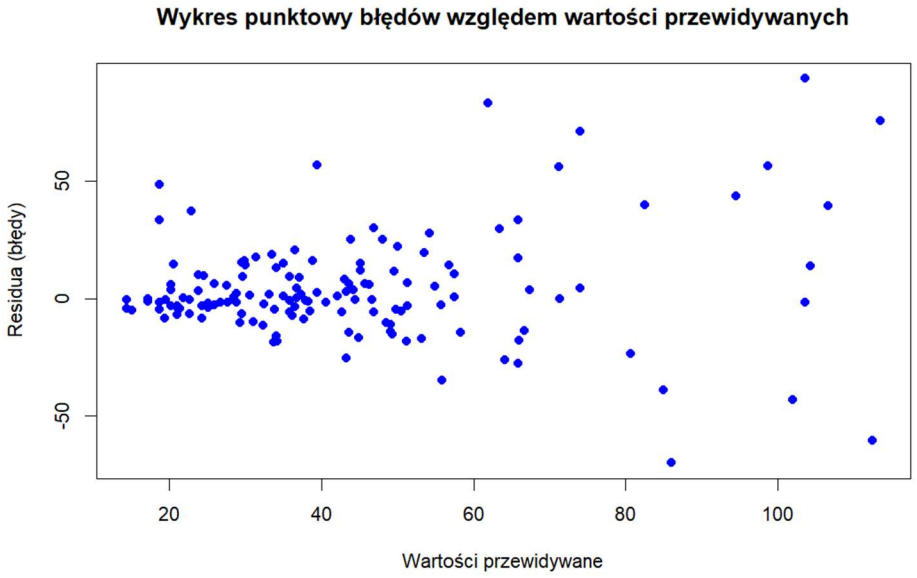
Figure 19: Wartości przewidywane vs rzeczywiste

Model dla całego zbioru testowego ma wartość MAPE: 0.323 i MAE: 14.6. Dla zbioru testowego do predykcji MAPE: 0.162 i MAE: 5.5. Jakość modelu jest zróżnicowana co jest widoczne również na wykresie przewidywane vs rzeczywiste. Model radzi sobie bardzo dobrze jednak, gdy pojawiają się nagłe wzrosty wartości to jego predykcje ją zaniżają. Model w takich sytuacjach ma również silny

trend wzrostowy, ale jego wartości nie są aż takie wysokie. Rozkład błędów ma rozkład normalny, a chmura punktów błędów również potwierdza tendencje do problemu z predykcją wysokich wartości.



Rysunek 20: Rozkład błędów



Rysunek 21: Rozkład błędów vs wartości przewidywane

```
Date       predicted_PM10_original[,1] PM10_original
 <date>                          <dbl>          <dbl>
2014-09-03                        28.9             31
2014-12-23                        15.4             10
2014-10-03                        43.4             46
2014-11-02                       105.              118
2014-09-23                        21.4             17
 mae(test_pred$predicted_PM10_original, test_pred$PM10_original)
1] 5.482987
 mape(test_pred$predicted_PM10_original, test_pred$PM10_original)
1] 0.1618225
```

Rysunek 22: Wylosowane wartości testowe

Model 2

Stworzono funkcję do tworzenia sekwencji danych o ustalonej długości. Przyjęto, że okno czasowe będzie wynosiło 10 dni, ponieważ było to maksymalne opóźnienie testowane w poprzednim modelu.

Sprawdzono modele z różną liczbą neuronów w warstwie, z jedną oraz dwoma warstwami LSTM oraz sprawdzono działanie funkcji aktywacji relu i tanh i najlepszą moc predykcyjną udało się uzyskać dla sieci z jedną warstwą LSTM ze 100 jednostkami, MAPE: 0.317. Kolejno tak samo jak poprzednio policzono MAE: 17.398. Dla zbioru testowego do predykcji MAPE: 0.276 i MAE: 10.8. Po zwizualizowaniu wartości przewidzianych względem faktycznych zauważono, że predykcje nie są najlepsze jakościowo. Model jest w stanie trafnie obrazować spadki i wzrosty jednak nie robi tego z dużą dokładnością. Rozkład błędów jest zbliżony do normalnego, ale widoczny jest prawy ogon.

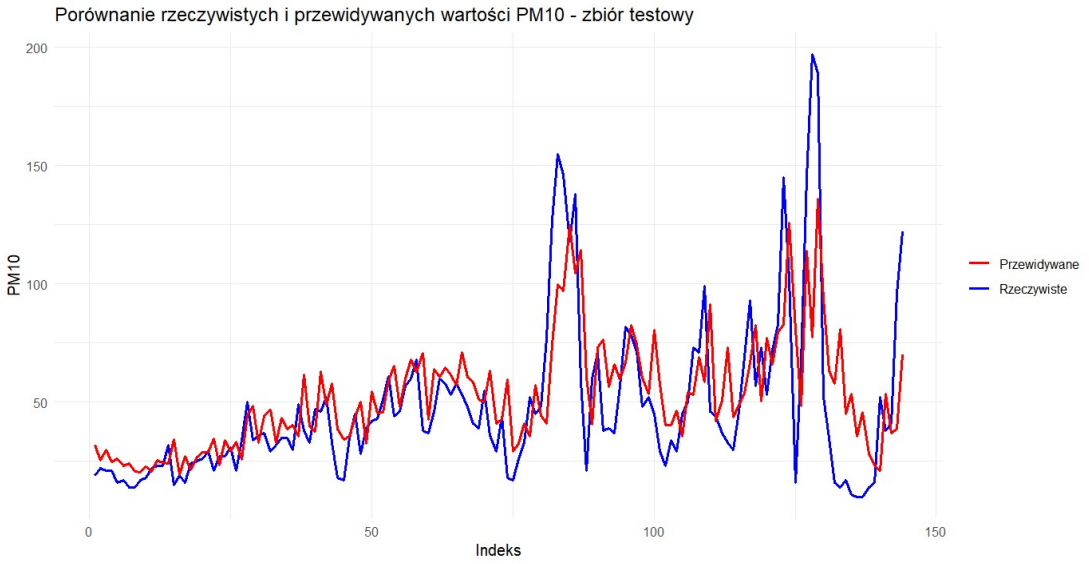
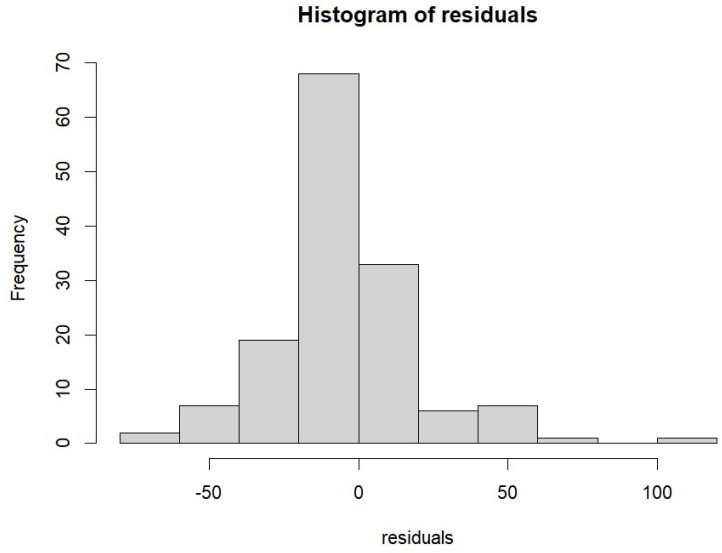
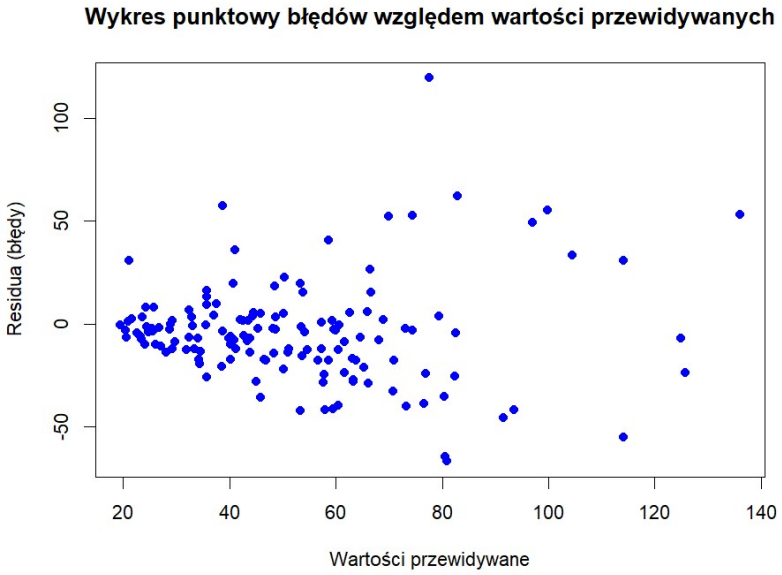


Figure 23: Wartości przewidywane vs rzeczywiste



Rysunek 24: Rozkład błędów



Rysunek 25: Rozkład błędów vs wartości przewidywane

```
> test_pred
          date predictions actual
25  2014-09-03    29.28648     31
136 2014-12-23    35.71506     10
55  2014-10-03    48.19532     46
85  2014-11-02   124.87376    118
45  2014-09-23    34.26474     17
```

Rysunek 26: Wylosowane wartości testowe

```
> mae(test_pred$predictions, test_pred$actual)
[1] 10.75248
> mape(test_pred$predictions, test_pred$actual)
[1] 0.2765949
```

Rysunek 27: Wylosowane wartości testowe

Podsumowanie

Pierwszy model poradził sobie dużo lepiej niż drugi, mimo nieco wyższego MAPE jakość jego predykcji jest wyższa, a reszty układają się w pożądany sposób. Model ma problem z predykcją wysokich wartości i mimo wielu testów z innymi parametrami wydaje się to być nieuniknione przy takim zbiorze cech na jakim pracowano. W celu znalezienia lepszego modelu istotne wydaje się dołożenie innych predykatorów, w szczególności danych pogodowych. Nagłe skoki PM10 mogą mieć związek z działaniem czynników atmosferycznych jak na przykład brak wiatru czy wysoka wilgotność powietrza. Korzystane dla uzyskania lepszego jakościowo modelu wydaje się również dołożenie danych ze wcześniejszych okresów co umożliwiłoby lepsze nauczenie się przez model trendów związanych z sezonowością. W przypadku okresu dwóch lat, gdzie zbiór testowy stanowił prawie pół roku z roku 2014 było to ograniczone. Oba modele wypadły dość słabo jakościowo w porównaniu ze standardowo oczekiwaną mocą predykcyjną jednak przy takim zbiorze zmiennych objaśniających rezultaty są zadawalające.