Przedstawienie zagadnienia

Przedmiotem analizy była pogoda, a dokładniej temperatura liczona w stopniach Celsjusza. Dane wykorzystane do analizy bayesowskiej pochodzą ze stacji meteorologicznych umieszczonych na terenie Polski. Analiza została przeprowadzona przy pomocy podejścia bayesowskiego do modelu regresji liniowej, rozkładu normalnego-gamma i rozwiązania analitycznego.

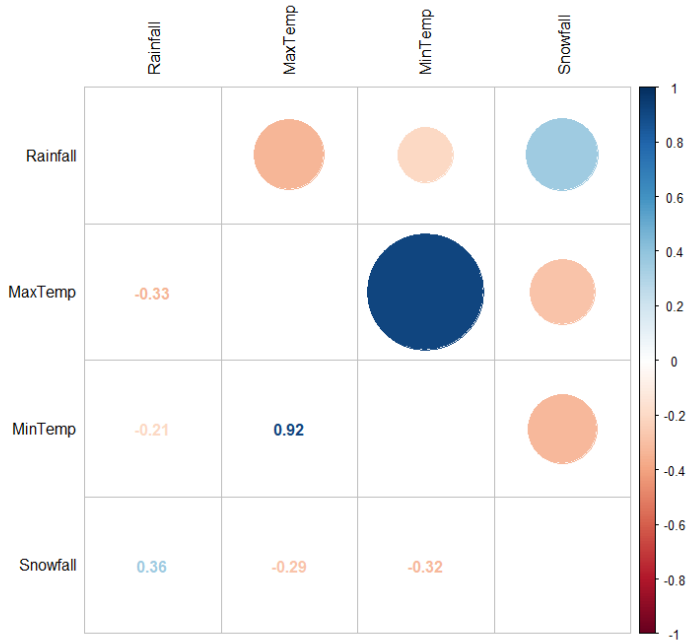
Użyte pomiary pochodzą z okresu od stycznia 2020 do lutego 2021. Dane zostały wybrane losowo, tak aby obserwacje nie pochodziły z tego samego dnia. W ten sposób mamy pewność, że zawarte zostały różnorodne zjawiska pogodowe. Zbiór danych zawiera następujące zmienne:

- nazwa stacji meteorologicznej,
- maksymalna temperatura w ciągu dnia,
- minimalna temperatura w ciągu dnia,
- dobowa suma opadów deszczu (w mm),
- dobowa suma opadów śniegu,
- minmalna temperatura dnia następnego.

Dane pochodzą ze strony Instytutu Meteorologii i Gospodarki Wodnej – Państwowy Instytut Badawczy (https://meteomodel.pl/dane/historyczne-dane-pomiarowe/). Liczba obserwacji w wylosowanym zbiorze wynosi 34.

Istnieje wiele badań potwierdzających przydatność metod data miningowych do prognozowania pogody Badacze używają m.in. regresji do predykcji różnych zjawisk pogodowych przy pomocy takich zmiennych jak minimalna, maksymalna temperatura powietrza czy suma opadów deszczu (Chauhan, Thakur, 2014, s. 2185).

W związku z bardzo wysoką korelacją pomiędzy zmienną minimalna temperatura a maksymalna temperatura należy pozbyć się współliniowości. W związku z tym, że zmienną objaśnianą w modelu jest minimalna temperatura, ze zbioru danych usunięto odczyty maksymalnej temperatury.



W warunkach współliniowości pomiędzy zmiennymi objaśniającymi, możemy mieć sytuację w której obliczona β wyjdzie poza przedział rozpinany pomiędzy odpowiednią wartość a priori, a odpowiednią wartość obliczoną z danych. Z tego właśnie powodu należy pozbyć się współliniowości.

Elicytacja parametrów a priori

Zestaw wartości parametrów a priori został wybrany przy pomocy oszacowania regresji na danych pochodzących z 1945 roku, ze stacji meteorologicznych położonych powyżej równoleżnika 50°. Pomiary pochodzą z lokalizacji, gdzie klimat jest w pewnym stopniu zbliżony do Polski. Z dostępnych danych wybrano z każdej stacji losowo po 34 obserwacje z lat 1944-1945. Wartości β a priori wyznaczone na podstawie modelu regresji liniowej są następujące:

Nazwa zmiennej	**Wartość parametru $\underline{\beta}$**
Wyraz wolny	0,136
Minimalna temperatura w ciągu dnia	0,98
Dobowa suma opadów deszczu (w mm)	-0,033
Dobowa suma opadów śniegu (w mm)	-0,027

Model wykorzystany do wyznaczenia parametrów a priori został oszacowany na 238 obserwacjach.

$\underline{v}$	238

Odchylenie standardowe w modelu liniowym wyniosło około 3,31, a więc parametr $\underline{s^{-2}}$ wynosi 0,09.

$\underline{s}^{-2}$	0,09

Od 1945 minęło dużo czasu, zaszły zmiany klimatyczne, a dodatkowo lokalizacje stacji meteorologicznych z danych a priori nie mają do końca takich samych uwarunkowań klimatycznych. W związku z tym zakładamy zerowe kowariancje pomiędzy zmiennymi - brak pewności czy rzeczywiście możemy wykorzystać kowariancje z macierzy wariancji-kowariancji z modelu pomocniczego. Dodatkowo wariancje z modelu pomocniczego zostały powiększone o 50%, aby zaaplikować niepewność.

$$\underline{U} = \begin{bmatrix} 0{,}102 & 0 & 0 & 0 \\ 0 & 0{,}0012 & 0 & 0 \\ 0 & 0 & 0{,}0029 & 0 \\ 0 & 0 & 0 & 0{,}0011 \end{bmatrix}$$

Wartość oczekiwana parametrów a posteriori

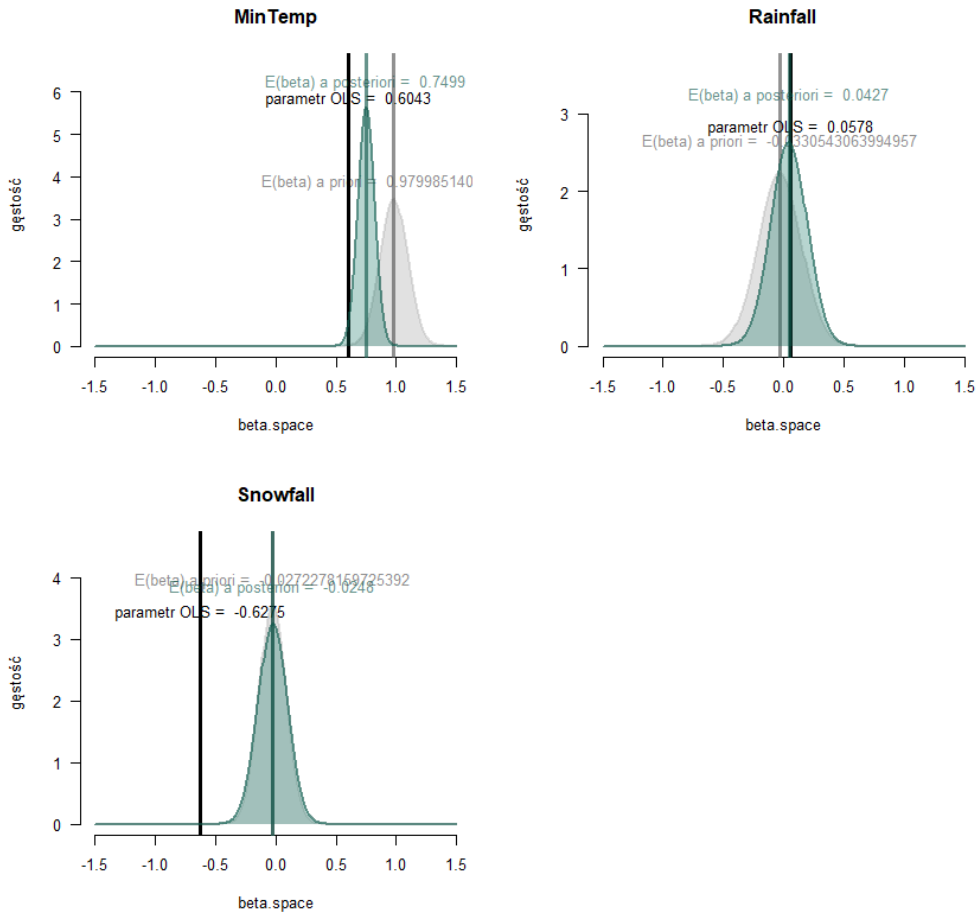
Nazwa zmiennej	**Wartość parametru $\overline{\beta}$**
Wyraz wolny	0,857
Minimalna temperatura w ciągu dnia	0,75

Dobowa suma opadów deszczu (w mm)	-0,043
Dobowa suma opadów śniegu (w mm)	-0.248

$$\overline{U} = \begin{bmatrix} 0{,}046 & -0.03 & -0.03 & -0.01 \\ -0.03 & 0{,}001 & 0 & 0 \\ -0.03 & 0 & 0{,}002 & 0 \\ -0.01 & 0 & 0 & 0{,}006 \end{bmatrix}$$

$\overline{\nu}$	272
$\underline{\nu s^2}$	3492,06
$\overline{s}^{-2}$	0,08

Rozkłady brzegowe parametrów a posteriori



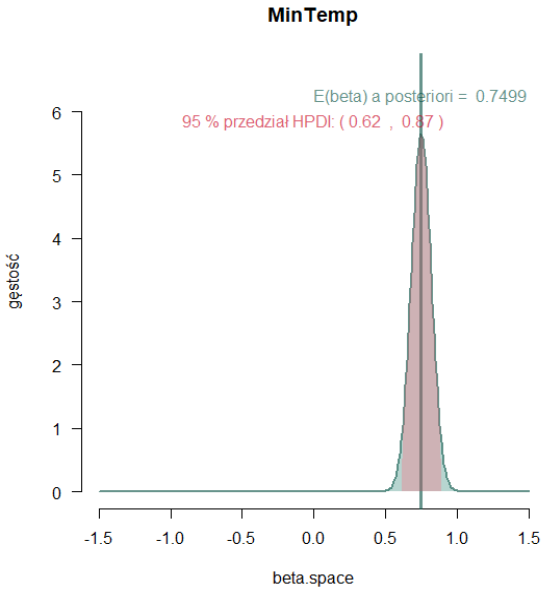
Wartości oczekiwane parametrów możemy traktować jak oszacowania punktowe. W przypadku zmiennej MinTemp wartość a posteriori zawiera się pomiędzy wartością a priori a oszacowaniem OLS. Dzięki podejściu bayesowskiemu uzyskaliśmy węższy rozkład brzegowy parametru β.

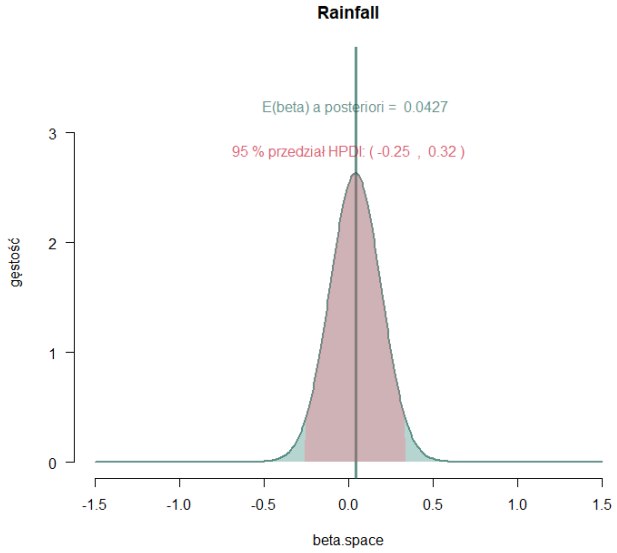
Rozkład brzegowy a posteriori parametru β zmiennej RainFall nie uległ znaczącej zmianie. Wartość oczekiwana parametru zmieniła znak, jednak spowodowane jest to faktem, że znaczna część masy

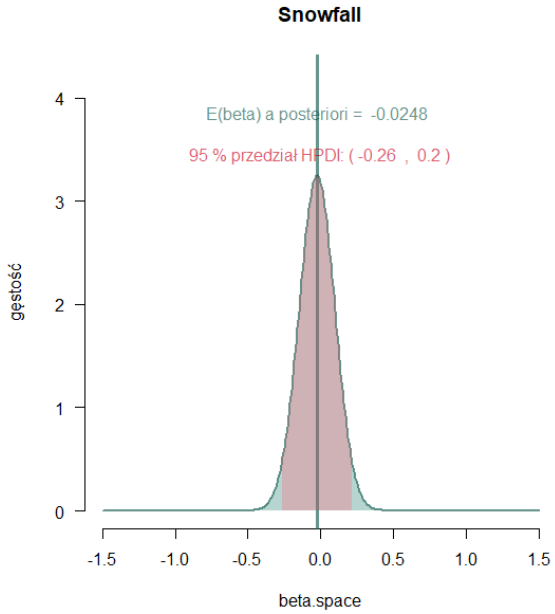
prawdopodobieństwa oscyluje wokół 0. Co za tym idzie wartość oczekiwana parametru β jest bardzo zbliżona do 0, a więc tego typu zmiana nie dziwi, ponieważ nominalnie parametr a priori różni się o bardzo niewiele od parametru a posteriori.

W przypadku zmiennej Snowfall informacja z modelu OLS ma niską wagę z powodu mniejszej ilości obserwacji w zbiorze wykorzystanym do modelu OLS. Ponadto wynik oszacowania odbiega znacznie od wiedzy a priori. Z tego względu informacja na temat wartości oczekiwanej parametru a posteriori pochodzi głównie z wiedzy a priori.

Ocena zmiennych







Znacza część masy prawdopodobieństwa rozkładu brzegowego parametru β dla zmiennej Rainfall i Snowfall oscyluje wokół 0. Oznacza to, że te zmienne mogą nie wnieść żadnej wartości dodanej do modelu.

W przypadku zmiennej MinTemp 95% rozkładu prawdopodobieństwa znajduje się w przedziale (0,62; 0,87). W tym rozkładzie wartość 0 występuje z prawdopodobieństwem niemal równym 0. Oznacza to, że zmienna odnosząca się do minimalnej temperatury w ciągu dnia ma duże znaczenie w modelowaniu zmiennej objaśnianej.

Zmienna	Czynnik Bayesa
MinTemp	5.23E+175
Rainfall	2.24E+151
Snowfall	3.24E+152

Uzyskane wyniki wskazują, że wszystkie zmienne w modelu są ważne w procesie modelowania minimalnej temperatury następnego dnia. Czynniki Bayesa nie potwierdziły wniosków wyciągniętych na podstawie wykresów gęstości rozkładów brzegowych parametrów.

Komentarze

1) Na potrzeby pracy domowej losowano obserwacje zarówno ze zbioru danych aktualnych (dotyczących Polski) jak i danych na podstawie których wnioskowano a priori (dane z okresu II wojny światowej). W przypadku dostępu do pełnych danych pogodowych Polski wyniki byłyby prawdopodobnie inne.
2) Przy liczeniu czynników Bayesa niezbędne było wykorzystanie kalkulatora o wysokiej precyzji. Mechanizm liczenia w R podczas kalkulacji niektórych komponentów zwracał wynik 0, gdy tak naprawdę była to bardzo mała liczba.

Literatura

Chauhan D., Thakur J., (2014), Data Mining Techniques for Weather Prediction: A Review.