Zaawansowana Ekonometria II - model I

Autorzy	Bartłomiej Kuźma, Maciej Odziemczyk
Który temat	Estymacja nieparametryczna

Informacje o artykule będącym inspiracją dla minimodelu

Tytuł	Modeling and Prediction of Rainfall in the U.K. Using Nonparametric Approach
Autorzy	A.H.M. Rahmatullah Imon, Noora N Saleh, Sirujam Munira Khan
Miejsce publikacji	Journal of Bangladesh Army University of Engineering & Technology, Volume 02, Issue 02
Rok	2020
Zakres stron	1-10
Tematyka, problemy i cele badawcze	Tematyka analizowanego artykułu obejmuje problem modelowania i predykcji wartości opadów w danym miesiącu w Wielkiej Brytanii. Głównym celem badania jest znalezienie najbardziej odpowiedniej metody do modelowania i predykcji wartości opadów deszczu, a także określenie najbardziej istotnych zmiennych objaśniających dla tego problemu.
Główne wnioski	Autorom udało się ustalić, że w przypadku danych pogodowych, o charakterze czasowym, modele MNK mają problem z brakiem normalności i heteroskedastycznością, przez co mają gorsze dopasowanie. Transformacja Boxa-Coxa nie daje znaczącej poprawy. Nieparametryczna metoda LOWESS osiąga dużo lepsze wyniki niż wcześniejsze modele. Produkuje też najlepsze prognozy. Nabardziej istotnymi dla UK okazały się godziny nasłonecznienia, dni z przymrozkami, pora roku i średnia temperatura.
Metodyka badawcza	Wyestymowano model MNK, następnie zastosowano transformację Boxa-Coxa i przeprowadzono diagnostykę otrzymanych modeli. w kolejnym kroku wyestymowano model LOWESS. Na koniec dokonano walidacji krzyżowej by ocenić, która metoda daje najlepsze rezulaty na analizowanym zbiorze danych
Dane	Dane pogodowe z 26 stacji meteorologicznych na terenie UK. Dane miesięczne, zawierają takie zmienne jak: suma opadów, maksymalna, minimalna i średnia temperatura, dni z przymrozkami i czas usłonecznienia. Zbiór danych pochodzi z lat 1886-2019.
Dlaczego wybrano ten artykuł	Kwestia przewidywania rozmiaru opadu ma ogromne znaczenie dla gospodarki. Dobry zbiór w rolnictwie wprost zależy od tego czy w danym okresie nie będzie suszy czy też podtopień, niszczących plony. Zważywszy na wagę tego tematu, a także zastosowanie nieparametrycznego modelu LOWESS zdecydowaliśmy się na wybranie tego artykułu.

Podstawowe informacje o minimodelu

Tytuł minimodelu	Modelowanie i predykcja opadów deszczu przy użyciu MNK i metod nieparametrycznych
Tematyka, problemy i cele badawcze	Projekt ma na celu sprawdzenie, czy za pomocą zwykłej Regresji Liniowej możliwe jest wyciągnięcie wniosków na temat ilościowych zależności zmiennych pogodowych. Istotnym punktem była również weryfikacja metod nieparametrycznych jako remedium na problemy pojawiające się przy modelowaniu dość restrykcyjnym MNK. Ponadto dużą wagę przywiązano do sensu i jakości predykcji w tego typu problemie. Projekt ma strukturę "akcja-reakcja" tj. rozpoczęto od klasycznej regresji i wraz z pojawianiem się problemów decydowano się na kolejne kroki mające je rozwiązać.
Metodyka badawcza	Oszacowanie modeli MNK (w wesji standardowej i przy użyciu tranformacji Boxa-Coxa), diagnostyka, oszacowanie modeli LOESS i regresji jądrowych, dokonanie predykcji, porównanie wyników metrykami MSE, RMSE, MAE i MAPE, próby uwzględnienia opóźnień, procesu AR(1) i Analizy Głównych. Składowych
Dane	Baza danych pochodzi ze strony IMGW i zawiera obserwacje dla 18 stacji meteorologicznych z terenu całej Polski. Badana próba ma charakter miesięczny i zawiera dane za okres 1961-2019. Zmienne użyte w badaniu korespondują z tymi, którymi posłużono sie w referencyjnym artykule, tj. suma opadów deszczu w ciągu miesiąca (w mm), liczba godzin nasłonecznienia w miesiącu, średnia temperatura miesięczna, liczba dni w miesiącu, w których wystąpiło zjawisko szronu (przybliżenie użytej przez autorów referencyjnego badania liczby dni z przymrozkami) oraz wygenerowaną zmienną określającą porę roku dla danej obserwacji.
Główne wnioski	Modele nieparametryczne są szybszym, bezpieczniejszym i konkurencyjnym, a nawet momentami lepszym rozwiązaniem niż modele Regresji Liniowej, ich głównym atutem jest mniejsze ryzyko dużych, jednostkowych pomyłek, jak również brak restrykcyjnych założeń. Ponadto złożoność procesu generującego opady wykracza poza możliwości modelowania za pomocą tak małego zbioru zmiennych objaśniających.

1 Wstęp

Ilościowe zdefiniowanie powiązań zmiennych klimatycznych pozwala na przewidywanie interesujących z jakiegoś punktu widzenia zjawisk, umiejętność ta wydaje się być niezwykle przydatna, zwłaszcza gdy przewidywanym zjawiskiem są opady deszczu. Wynikami ilościowych badań nad klimatem mogą być zainteresowani zarówno klimatologowie, ekonomiści jak i przedsiębiorcy, których główna działalność gospodarcza zależna jest w istotnym stopniu od opadów. Szczególnie zainteresowani postępami w tejże dziedzinie mogą być przede wszystkim rolnicy, organizatorzy imprez masowych czy też przedsiębiorstwa świadczące usługi, do których zrealizowania wymagane są odpowiednie warunki pogodowe. Śladem A.H.M. Rahmatullaha Imona, Noory N Saleh i Sirajumy Munira Khan, którzy w swoim artykule [1] zajęli się problemem modelowania i predykcji opadów deszczu w Wielkiej Brytanii, zdecydowano się w niniejszym projekcie na przeprowadzenie analogicznego badania dla Polski. Podobnie do pierwowzoru posłużono się modelami Regresji Liniowej w różnych wariantach oraz metodami nieparametrycznymi LOESS i regresjami jądrowymi z pakietu *np* w środowisku *R*. Dla modeli Regresji Liniowej przeprowadzono diagnostykę pod kątem spełnienia założeń estymatora Metody Najmniejszych Kwadratów, które jak się okazało w większości przypadków nie są spełniane. Pomimo ewentualnych obciążeń oszacowań parametrów, zdecydowano się na obliczenie metryk dopasowania do danych jak również wygenerowano i oceniono predykcje (obciążenia nie wykluczają bowiem zdolności predykcyjnych, bowiem istnieje szereg metod, w których parametry nie są identyfikowalne, a świetnie sprawdzają się w różnego rodzaju zadaniach). Modele oceniono za pomocą różnych metryk i porównano między sobą w celu wyłonienia najlepszego z nich. Badanie przeprowadzono dla 18 stacji ulokowanych na terenie Polski, jednakże w celu zachowania czytelności raportu szczegółową analizę zaprezentowano wyłącznie dla IMGW Stacji Meteorologiczno-Hydrologicznej w Suwałkach.

2 Dane

W niniejszym badaniu wyjściowy zbiór zmiennych objaśniających obejmował zmienne takie jak: średnia temperatura miesięczna (avg_month_temp), miesięczna suma godzin nasłonecznienia (sunhours_sum), liczba dni ze szronem (# days_foarfrost), jako reprezentant zmiennej "liczba dni z przymrozkami" z pierwowzoru oraz wygenerowane na podstawie dat zmienne reprezentujące porę roku - wiosna (spring), miesiące wchodzące w skład zmiennej: marzec, kwiecień, maj; lato (summer), miesiące wchodzące w skład zmiennej: czerwiec, lipiec, sierpień; zima (winter), miesiące wchodzące w skład zmiennej: grudzień, styczeń, luty. W odróżnieniu od badania dla Wielkiej Brytanii, zmienne sezonowe dla regresji MNK zostały zakodowane jako zmienne binarne - poziom bazowy to jesień (autumn) - podejście to jest odmienne w niż to zaprezentowane przez Imona i in. wydaje się jednak, że sezony należy traktować jako stany, zatem należy je kodować w oparciu o poziom bazowy. Zmienną objaśnianą we wszystkich modelach była miesięczna suma opadów w milimetrach (rainfall_sum). Transformacje, którym zostały poddane zmienne opisane zostały szerzej w częściach poświęconych metodologii i wynikom.

Analizie poddane zostały miesięczne dane synopytyczne dla 18 stacji meteorologicznych znajdujących się na terenie Polski: Zakopane, Kraków-Balice, Wrocław-Strachowice, Wieluń, Łódź-Lublinek, Lublin-Radawiec, Poznań-Ławica, Warszawa-Okęcie, Siedlce, Świnoujście, Szczecin, Mława, Kołobrzeg-Dźwirzyno, Ustka, Łeba, Hel, Elbląg-Milejewo i Suwałki, z czego dla tej ostatniej zaprezentowana została szczegółowa analiza. Wykorzystany został niemalże cały dostępny horyzont czasowy tj. od 1960 do 2019 roku. Stacje, które zostały poddane analizie wybrano na podstawie kryterium liczebności obserwacji (odrzucono wszystkie stacje dla których $N < 648$).

w toku analiz zauważono niepokojącą liczbę zerowych obserwacji zmiennej sunhours_sum. z reguły obserwowany był ciąg zer na początku horyzontu czasowego, wziąwszy pod uwagę notatkę zamieszczoną na stronie IMGW o wątpliwej jakości danych dla niektórych atrybutów do 1994 roku uznano te dane za nieprawidłowe, wydaje się wysoce nieprawdopodobne aby ciągiem, przez kilka lat suma godzin nasłonecznienia wynosiła zero. z uwagi na chęć zachowania ciągłości danych, zdecydowano się na usunięcie jedynie początkowego ciągu zer dla każdej analizowanej stacji.

3 Metodologia

W celu przeprowadzenia analizy w zakresie Regresji Liniowej zbudowano 7 różnych modeli:

- model ogólny - zawierający wszystkie zmienne zgodne z pierwowzorem (gen),
- model ogólny z lewostronnym przekształceniem Boxa-Coxa (gen-BC),
- model ogólny na zmiennych opóźnionych o 1 okres (gen-lag(1)),
- model ogólny na zmiennych opóźnionych o jeden okres z uwzględnieniem komponentu autoregresyjnego rzędu 1 (gen-lag(1)-ar(1)),
- model z restrykcjami (spec),
- model uwzględniający analizę głównych składowych (PCA),
- model uwzględniający analizę głównych składowych z lewostronnym przekształceniem Boxa-Coxa (PCA-BC).

Dla każdego z wyżej wymienionych modeli przeprowadzono diagnostykę w zakresie:

- występowania niedokładnej współliniowości zmiennych objaśniających - współczynniki inflacji wariancji (VIF),
- poprawności formy funkcyjnej - test RESET w różnych wariantach:
  - dla kwadratów zmiennych z modelu,
  - dla sześcianów zmiennych z modelu,
  - dla kwadratów transformacji liniowych PCA zmiennych z modelu,
  - dla sześcianów transformacji liniowych PCA zmiennych z modelu,
- homoskedastyczności reszt - test Breuscha-Pagana,
- autokorelacji reszt - test Breuscha-Godfreya (2 rząd),

- normalności reszt - test Jarque-Bera i test Przeskalowanych Momentów (Rescaled Moments).

Następnie oszacowano nieparametryczne modele:

- LOESS (różne zakresy szerokości pasma),
- estymatory jądrowe (porównanie różnych metod wyboru pasma, wykorzystany pakiet automatycznie dobiera odpowiednie jądro do typu zmiennej).

Dla każdego modelu obliczono metryki jakości dopasowania do danych:

- Błąd średniokwadratowy (MSE),
- spierwiastkowany błąd średniokwadratowy (RMSE),
- średni błąd absolutny (MAE),
- średni procentowy błąd absolutny (MAPE),

Te same metryki wykorzystano do zbadania jakości predykcji. Ponadto zdecydowano się również na graficzną analizę normalności i homoskedastyczności reszt z modelu ogólnego i z modelu z przekształceniem Boxa-Coxa, w celu głębszej analizy zasadności wykorzystania tego przekształcenia. Analizę graficzną poczyniono również w celu zbadania jakości predykcji, posłużono się również współczynnikiem korelacji rang Spearmana.

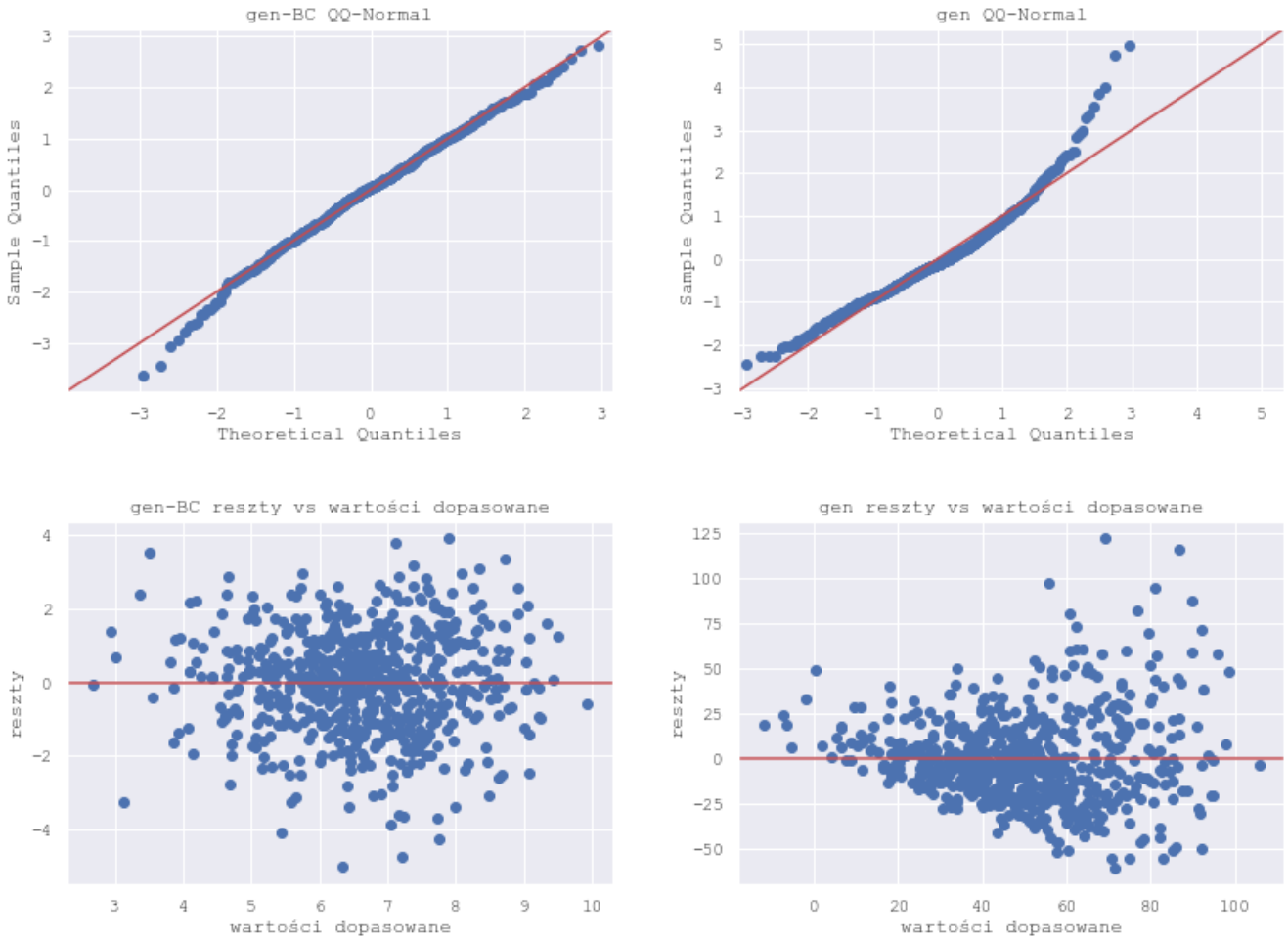
4 Wyniki

Osobliwym pomysłem wydała się predykcja oparta na zmiennych obserwowalnych równo ze zmienną zależną - co *de facto* miało miejsce w artykule Imona i in. z tego względu w niniejszym badaniu zdecydowano się w pierwszej kolejności na zbudowanie i sprawdzenie trzech modeli - model ogólny (gen) - zgodny z pierwowzorem oraz modele na zmiennych opóźnionych (gen-lag(1)) oraz model gen-lag(1) z komponentem autoregresyjnym rzędu pierwszego (gen-lag(1)-ar(1)). Modele oparte na zmiennych opóźnionych okazały się nieodpowiednie, bowiem statystycznie istotne na poziomie 5% okazały się jedynie zmienne avg_month_temp oraz spring, warto wyszczególnić nieistotność komponentu ar(1). Ponadto współczynniki determinacji w modelach na opóźnieniach okazały się niemalże dwukrotnie niższe od modelu podstawowego (gen), dużo niższe były również wyniki testu F na istotność regresji, ponadto modele te wypadały gorzej pod względem kryteriów informacyjnych (AIC) i (BIC). Zdecydowano się zatem porzucić nadzieje o predykcji z wyprzedzeniem na rzecz sprawdzenia generalizowania się modeli out-of-the sample - zastosowano podział na próbę in (90%) i out (10% ostatnich obserwacji) i w oparciu o taki podział prezentowane są wyniki w Tabeli 1. Fakt słabego dopasowania i nieistotności statystyczniej modeli na opóźnieniach można tłumaczyć niską częstotliwością danych (dane miesięczne). Oszacowania modelu ogólnego okazały się spójne zarówno co do znaku jak i rzędu wielkości z literaturą, z uwagi na odmienne kodowanie pór roku, nie należy tych dwóch badań porównywać ze sobą pod tym kątem. W Tabeli 1 poza oszacowaniami zamieszczono również wyniki testów diagnostycznych. Dla modelu gen hipoteza zerowa dla testu RESET została przyjęta tylko w wariancie alternatywy PCA^2, pozwala to wnioskować o niepoprawnej formie funkcyjnej modelu. Hipoteza zerowa odrzucona zostałą również w teście Breuscha-Pagana, Jarque-Bera i Rescaled Moments - na tej podstawie wywnioskowano o problemie heteroskedastyczności i braku normalności rozkładu reszt. Wynik testu Breuscha-Godfreya na autokorelację składnika resztowego pozwala przyjąć hipotezę zerową o jej braku, co również utwierdza w przekonaniu nieprzydatności modeli na opóźnieniach.

Wyniki testów diagnostycznych wskazują na naruszenie założeń estymatora MNK o poprawności i liniowości funkcyjnej, sferyczności błędów losowych oraz normalności ich rozkładu. Istnieje możliwość rozwiązania wszystkich problemów poprzez zastosowanie odpowiednich transformacji zmiennych; w pierwszej kolejności zdecydowano się na lewostronną transformację Boxa-Coxa. W przypadku stacji Suwałki, optymalna lambda wyniosła 0.275, oszacowany w ten sposób model nazwano gen-BC. Zastosowana transformacja pozwoliła na utrzymanie znaków i istotności oszacowań z modelu gen. Ponadto zaobserwowano poprawę w wynikach testów diagnostycznych - test RESET dla kwadratów zmiennych pomimo odrzucenia hipotezy zerowej zanotował wzrost wartości p do 0.038, odrzucona została również hipoteza zerowa w wariancie sześcianów głównych składowych, odrzucono alternatywy o sześcianie zmiennych z modelu i kwadracie głównych składowych. Warto zauważyć znaczącą poprawę w testach Breuscha-Pagana, Jarque-Bera i Rescaled Moments, mimo to na poziomie 5% nie można przyjąć hipotez zerowych tych testów. Poprawę w wynikach testów diagnostycznych w zakresie sferyczności reszt zdecydowano się zobrazować na wykresach, można na nich znacznie lepiej zobserwować poprawę. Wartym odnotowania jest również fakt

poprawy dopasowania po transforamcji BC - $\hat{R}^2$ wzrósł z 0.39 do 0.418. Modeli nie można porównywać pod względem kryteriów informacyjnych (inna postać zmiennej zależnej).

Rysunek 1: Diagnostyka reszt, porównanie modelu gen i gen-BC



Źródło: Opracowanie własne.

Tablica 1: Wyniki estymacji Regresji Liniowych

	gen	gen-BC	gen-lag(1)	gen-lag(1)-ar(1)	spec	PCA	PCA-BC
Intercept	59.5343***	7.2449***	33.8979***	33.4188***	47.4145***	50.0281***	6.7264***
	(3.1713)	(0.1786)	(3.6182)	(4.5292)	(3.3223)	(2.6908)	(0.1559)
avg_month_temp	3.0888***	0.2061***	1.5597***	1.5349***	1.0776***		
	(0.2974)	(0.0167)	(0.3392)	(0.3676)	(0.2708)		
sunhours_sum	-0.2838***	-0.0181***	0.0020	0.0043			
	(0.0238)	(0.0013)	(0.0272)	(0.0301)			
# days_hoarfrost	-1.3745***	-0.0849***	0.2322	0.2436	-1.2293***	-1.7092***	-0.1073***
	(0.3720)	(0.0209)	(0.4253)	(0.4305)	(0.4112)	(0.4010)	(0.0232)
spring	16.3469***	1.0516***	9.2474**	9.1157**	-8.7544***	2.0426	0.0976
	(3.4974)	(0.1969)	(3.9882)	(4.0607)	(3.0876)	(3.4803)	(0.2017)
summer	31.2581***	1.5847***	8.5367*	8.2868*	8.9171**	38.6108***	2.0751***
	(4.1971)	(0.2363)	(4.7862)	(4.9958)	(4.1530)	(4.4762)	(0.2594)
winter	-1.4106	-0.0020	3.3425	3.3563	-8.3380*	-27.9493***	-1.7720***
	(4.0175)	(0.2262)	(4.5857)	(4.5899)	(4.3967)	(3.3617)	(0.1948)
rainfall_sum_lag_1				0.0080			
				(0.0456)			
Temp_sun_PCA						-0.1423***	-0.0086***
						(0.0212)	(0.0012)
Observations	637	637	636	636	637	637	637
R^2	0.396	0.423	0.215	0.216	0.26	0.292	0.284
$\bar{R}^2$	0.390	0.418	0.208	0.207	0.254	0.286	0.278
AIC	5901.959	2236.674	6059.643	6061.611	6029.382	6001.263	2372.651
BIC	5933.156	2267.871	6090.829	6097.253	6056.123	6028.004	2399.392
F	68.769	77.003	28.786	24.640	44.252	51.945	49.948
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
RESET (sq.)	33.419	4.321	0.361	0.316	12.203	27.255	6.678
	(0.000)	(0.038)	(0.548)	(0.574)	(0.001)	(0.000)	(0.01)
RESET (cb.)	16.692	2.157	1.314	1.340	7.912	18.043	9.602
	(0.000)	(0.117)	(0.269)	(0.262)	(0.000)	(0.000)	(0.000)
RESET (PCA sq.)	1.009	0.662	2.159	2.548	10.854	3.824	0.214
	(0.316)	(0.416)	(0.142)	(0.111)	(0.001)	(0.051)	(0.644)
RESET (PCA cb.)	9.551	10.258	1.582	2.139	6.308	27.858	35.745
	(0.000)	(0.000)	(0.206)	(0.119)	(0.002)	(0.000)	(0.000)
Breusch-Pagan	59.670	13.092	63.089	65.592	59.813	46.687	13.436
	(0.000)	(0.042)	(0.000)	(0.000)	(0.000)	(0.000)	(0.02)
Breusch-Godfrey	1.689	1.061	0.748	0.706	0.116	0.646	1.433
	(0.43)	(0.589)	(0.688)	(0.703)	(0.944)	(0.724)	(0.489)
Jarque-Bera	253.276	8.778	284.137	285.483	294.622	237.504	3.784
	(0.000)	(0.012)	(0.000)	(0.000)	(0.000)	(0.000)	(0.151)
Rescaled Moments	263.575	9.098	295.592	298.684	304.941	245.537	3.897
	(0.000)	(0.011)	(0.000)	(0.000)	(0.000)	(0.000)	(0.143)

Źródło: Opracowanie własne.

Tablica 2: Współczynniki inflacji wariancji (VIF) - model gen

Zmienna	VIF
avg_month_temp	9.05
sunhours_sum	14.04
# days_hoarfrost	1.38
spring	3.03
summer	4.58
winter	1.89

Źródło: Opracowanie własne.

W tabeli 2 zamieszczono wartości współczynników inflacji wariancji dla modelu gen. Podwyższone wyniki dla zmiennych avg_month_temp i sunhours_sum oraz ich merytoryczne znaczenie pozwalają wyciągnąć wniosek o potencjalnej niedokładniej współliniowości. Hipotezę zdecydowano się zweryfikować przy pomocy wyresu rozrzutu oraz współczynnika korelacji rang Spearmana.

Rysunek 2: Rozrzut i korelacja avg_month_temp i sunhours_sum

Źródło: Opracowanie własne.

Zarówno wykres rozrzutu jak i istotny statystycznie współczynnik korelacji 0.81 (0.000) pozwala wnioskować o silnej korelacji zmiennych avg_month_temp i sunhours_sum. Próbę rozwiązania tego problemu podjęto poprzez oszacowanie modelu szczególnego (spec), w którym usunięto zmienną z najwyższym VIF tj. sunhours_sum. Jak można się było spodziewać, ze względu na istotność statystyczną nasłonecznienia wystąpił problem zmiennej pominiętej, oszacowanie spring zmieniło swój znak jednocześnie zachowując istotność, ponadto obciążone zostały wyniki pozostałych oszacowań - ściąganie w kierunku zera. Ponadto diametralnie spadły wartości $\hat{R}^2$, pogorszeniu uległy wartości kryteriów informacyjnych, a testy diagnostyczne odrzucają hipotezę zerową w każdym przypadku za wyjątkiem autokorelacji. Próbę rozwiązania nowopowstałych problemów podjęto poprzez zastosowanie Analizy Głównych Składowych nadmiernie skorelowanych zmiennych. Zdecydowano się na wykorzystanie tylko pierwszej składowej, bowiem wyjaśnia ona 99.7% wariancji analizowanych zmiennych - w ten sposób powstał model oznaczony jako PCA. Niestety model PCA nie rozwiązał problemu zmiennej pominiętej, w Tabeli 1 można zauważyć nieistotność spring i ogromne, istotne obciążenie winter oraz pomniejsze obciążenia pozostałych zmiennych. Na plus wychodzi natomiast poprawa współczynnika determinacji i kryteriów informacyjnych w stosunku do modelu szczególnego (spec), niestety model PCA nie przyczynił się do poprawy wyników testów diagnostycznych, wnioski z nich pozostały niezmienione. Ostatnią szansą na poprawę jakości Regresji Liniowej w analizie problemu było połączenie Analizy Głównych Składowych oraz lewostronnego przekształcenia Boxa-Coxa (PCA-BC). Optymalna lambda dla modelu PCA-BC podobnie jak dla modelu gen-BC wyniosła 0.275. Model PCA-BC co do znaków

oszacowań i ich istotności jest spójny z modelem PCA, notuje on natomiast obniżony współczynnik determinacji i przyjęcie hipotez zerowych testów na normalność składnika resztowego i testu RESET w wariancie z kwadratem głównych składowych. Zaskoczeniem mogą być natomiast znaki oszacowań głównych składowych, biorąc pod uwagę fakt wyższego co do wartości bezwzględnej oszacowania avg_month_temp niż sunhours_sum spodziewano się, że główna składowa przyjmie znak dodatni, tymczasem w Tabeli 1 widnieje minus. Model PCA-BC pod względem kryteriów informacyjnych możemy porównać jedynie z modelem gen-BC i wypada on pod tym względem gorzej. Stosunkowo niewielką zaletą w odniesieniu do wad modeli opartych o Analizę Głównych Składowych jest znaczne obniżenie VIFów - dla Temp_sun_PCA statystyka ta nie przekracza wartości 3.02.

Najlepszymi modelami regresji pod kątem dopasowania do danych okazały się modele gen i gen-BC, poniżej zamieszczono tabelę metryk dopasowania na zbiorze treningowym (in - sample).

Tablica 3: Metryki dopasowania dla modeli gen i gen-BC

	MSE	RMSE	MAE	MAPE
gen	605.095	24.6	18.307	56.297
gen-BC	597.947	24.452	17.637	50.322

Źródło: Opracowanie własne.

W Tabeli 3 zauważyć można, że model gen-BC jest lepszy nie tylko pod kątem $\bar{R}^2$, ale również w pozostałch metrykach. Tutaj warto wspomnieć, że Suwałki są wyjątkową stacją, z reguły obserwowano niższe MSE i RMSE dla modelu gen, natomiast MAE i MAPE niższe były dna gen-BC.

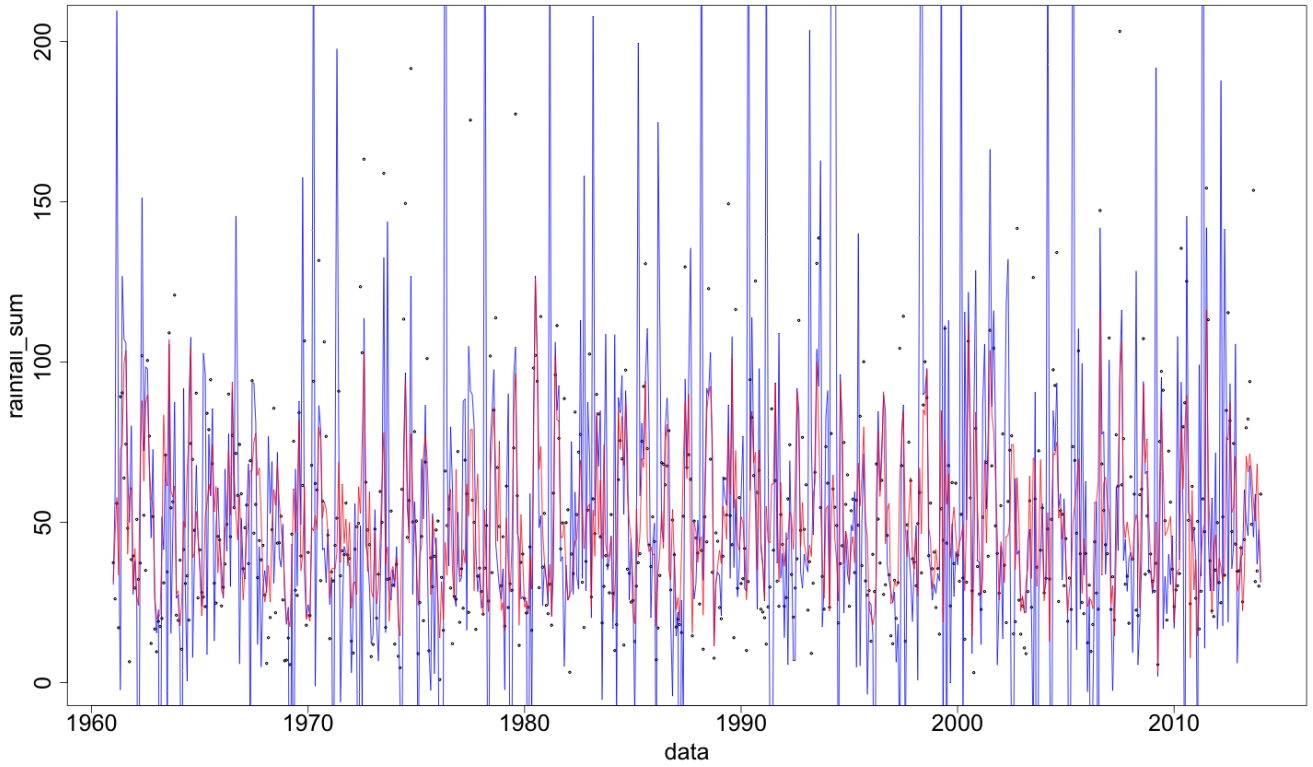
Podsumowując analiza Regresji Liniowej może ograniczyć się jedynie do predykcji, a raczej z uwagi na charakter problemu - dopasowanie modelu do danych (nieprzydatność opóźnień i ar(1)) i ewentualne wykorzystanie go w przypadku np. uzupełniania braków w bazach danych - wyniki out-of-sample dla modeli gen i gen-BC (najlepszych) zaprezentowane zostały w części poświęconej "predykcjom". Powodem, dla którego wnioskowanie statystyczne w tym przypadku nie ma sensu jest po pierwsze niespełnienie założeń estymatora MNK, a po drugie występowanie problemów wnioskowania przyczynowego; nawet w tym krótkim badaniu można było zauważyć problemy współliniowości czy zmiennych pominiętych ponadto modele, w których zastosowano przekształcenia Boxa-Coxa, ze względu na nietypową labmdę (0.275) tracą swoją interpretowalność, podobnie sytuacja wygląda w przypadku wykorzystania analizy głównych składowych. Proces kształtujący opady wydaje się nie ograniczać jedynie do zmiennych tutaj wykorzystanych. Biorąc pod uwagę uzyskane wyniki, szczególnie testy RESET tj. nieznajomość formy funkcyjnej, nawet w przypadku tak uproszczonej wersji zjawiska uzasadnione wydaje się być podjęcie próby estymacji nieparametrycznej, która pozwoli na oszacowanie funkcji nie znanej *a priori*.

Autorzy badania użyli nieparametrycznego modelu LOWESS (*Locally Weighted Scatterplot Smoothing*), czyli narzędzia służącego do graficznego wygładzania zależności miedzy zmiennymi objaśniającymi, a zmienną objaśnianą. Wartym podkreślenia jest natomiast fakt wykorzystywania modelu LOWESS zazwyczaj do badania zależności zmiennej objaśnianej z jednym regresorem. W opisanym tutaj badaniu postanowiono zatem wykorzystać model LOESS (*Locally Estimated Scatterplot Smoothing*), czyli wielowymiarową wersję LOWESSa.

W trakcie szacowania modeli zauważono, że w zbiorze danych znajduje się błędna obserwacja (czerwiec 2017 ID=12546). Obserwacja ta miała zerową wartość w kolumnie dot. liczby godzin słonecznych w ciągu miesiąca, co wydaje się niemożliwe i co generowało później braki danych w wartościach dopasowanych/predykcjach z modeli. By rozwiązać ten problem policzono wartość średnią dla wszystkich czerwcowych obserwacji w próbie i zastąpiono domniemany błąd (średnia dla czerwca w próbie wyniosła 242.7831). Pozwoliło to rozwiązać istniejący problem.

Estymacja modelu w pakiecie *R* pozwala na wybranie szerokości pasma, jako zmiennej, która wpływa na dopasowanie do danych. Na potrzeby tego badania ustawiono wartości argumentu *span* na 0.1, 0.25, 0.5, 0.75, 1 (wartość należy interpretować jako procent szerokości pasma). By uplastycznić na czym polega różnica pomiedzy szerokościami pasma na Rysunku 3. pokazano wygładzenie dla skrajnych badanych wartości szerokości pasma.

Rysunek 3: Wygładzenie zależności zmiennej rainfall_sum skrajnymi szerkościami pasma



Źródło: Opracowanie własne.

Jak widać przy wybraniu wąskiej szerokości (typu 0.1, niebieska linia) krzywa ma skłonności do nadmiernego dostosowywania się do danych, za to szerokie pasmo (w tym przypadku równe 1, czerwona linia), powoduje dużo większe wygładzenie, dlatego wykres ten w dużej mierze nie uwzględnia skrajnych wartości.

Następnie porównując metryki in-sample została wybrana wartość pasma, która będzie używana w dalszej części do predykcji i porównywania z innymi modelami.

Tablica 4: Metryki dopasowania dla różnych szerokości pasma w modelu LOESS

	MSE	RMSE	MAE	MAPE
loess_span_10	3341.481	57.806	35.769	1.664
loess_span_25	835.493	28.905	21.069	0.517
loess_span_50	585.361	24.194	18.238	0.464
loess_span_75	568.754	23.849	17.768	0.378
loess_span_100	566.436	23.8	17.625	0.364

Źródło: Opracowanie własne.

Jak widać w Tabeli 4, od pasma równego 0.5 wartości metryk nie zmieniają się już znacznie, a najlepszy wynik uzyskano dla najszerszego pasma i to model z taką właśnie szerokością zostanie użyty do predykcji w dalszej części.

Warto w tym miejscu odnotować pewną nieścisłość, autorzy referencyjnego artykułu podają wartość współczynnika R^2 równą około 0.57[1] jednak nie podają jak współczynnik ten został obliczony, możliwe że było to jakiegoś rodzaju *pseudo* R^2, które można traktować jako sprzeczne w idei z modelem LOWESS/LOESS, który

w założeniu służy eksploracji i idnetyfikacji wzorców i wygładzaniu danych. W niniejszym badaniu zdecydowano się na pominięcie tejże statystyki.

Zdecydowano się również skorzystać z jądrowej estymacji nieparametrycznej, za pomocą pakietu *np*[3], który w jakimś stopniu można okreslić jako odpowiednik pakietu *locpol* dla wielu zmiennych objaśniających, uwzględnia on także zarówno zmienne ciągłe jak i dyskretne, co nie jest w ogólności sprawą trywialną i oczywistą[2]. W pakiecie tym dla każdego typu zmiennych kernel używany przy regresji jądrowej jest wybierany automatycznie (system rozpoznaje czy zmienna jest ciągła, uporządkowana czy nieuporządkowana). W celu wykorzystania pakietu *np* należało stworzyć zmienną *season* by pozbyć się zmiennych zero-jedynkowych z modelu (działanie zalecane przez autorów pakietu *np*).

Pakiet *np* daje znacznie większe możliwości wyboru szerokości pasma, co nie jest bez znaczenia, zważywszy na to, że jest to czynność analogiczna do wyboru formy funkcyjnej w przypadku modeli parametrycznych, nie można tego kroku lekceważyć. W badaniu rozważano typ jądrowej regresji zastosowanej podczas estymacji (wybór między estymatorem Nadaraya-Watsona (estymator lokalnych stałych) **lc**, a estymatorem lokalnych funkcji liniowych **ll**). Dokonano też wyboru metody wybierania szerokości pasma (dostępne możliwości to walidacja krzyżowa Kullbacka-Leiblera **cv.aic** i walidacja krzyżowa najmniejszych kwadratów **cv.ls**). Ostatnim parametrem, który był testowany to typ pasma w przypadku zmiennych ciągłych (dostępne opcje to stały **fixed**, uogólniony najbliższych sąsiadów **g.nn** i adaptacyjny najbliższych sąsiadów **a.nn**). Wyniki metryk dla poszczególnych konfiguracji zostały przedstawione w Tabeli 5. poniżej.

Tablica 5: Metryki dopasowania dla różnych konfiguracji ustawień pasma w pakiecie *np*

				MSE	RMSE	MAE	MAPE
np_1	ll	cv.aic	fixed	566.185	23.795	17.692	0.375
np_2	lc	cv.aic	fixed	554.652	23.551	17.257	0.345
np_3	ll	cv.ls	fixed	548.245	23.415	17.321	0.351
np_4	lc	cv.ls	fixed	521.495	22.836	16.674	0.331
np_5	lc	cv.ls	a.nn	529.416	23.009	16.865	0.337
np_6	lc	cv.ls	g.nn	533.552	23.099	16.978	0.334

Źródło: Opracowanie własne.

Na podstawie wyników z Tabeli 5. wybrano model 4, który ma najniższe wartości wszystkich metryk. Model ten jest szacowany za pomocą estymatora Nadaraya-Watsona, pasmo jest wybierane przy użyciu walidacji krzyżowej najmniejszych kwadratów, a typ pasma w przypadku zmiennych ciągłych został ustalony na stały.

Pakiet *np* podaje również wartość współczynnika R^2 danego wzorem[3]:

$$R^2 = \frac{\left[\sum_n^{i=1}(y_i - \bar{y})(\hat{y_i} - \bar{y})\right]^2}{\sum_n^{i=1}(y_i - \bar{y})^2 \sum_n^{i=1}(\hat{y_i} - \bar{y})^2},$$

gdzie y_i to wartość obserwowana y dla obserwacji i,
a $\hat{y_i}$ to wartość dopasowane z modelu.

Miara ta przyjmuje zawsze wartości z przedziału $[0, 1]$ i jest identyczna ze standardową miarą R^2 w przypadku, kiedy model jest liniowy, dopasowany przy użyciu MNK i zawiera stałą. Wartość tego współczynnika dla modelu 4, który został wybrany do predykcji w dalsze części tego badania jest równa 0.4866, co jest wynikiem lepszym niż najlepszy z modeli MNK. Nie można ich jednak porównywać 1:1, dlatego w dalszej części modele porównywane będa przy użyciu metryk MSE, RMSE, MAE i MAPE.

5 Predykcje

W tej części zaprezentowane zostały wyniki out-of-sample dla najlepszych modeli Regresji Liniowej (gen i gen-BC) i modeli nieparametrycznych.

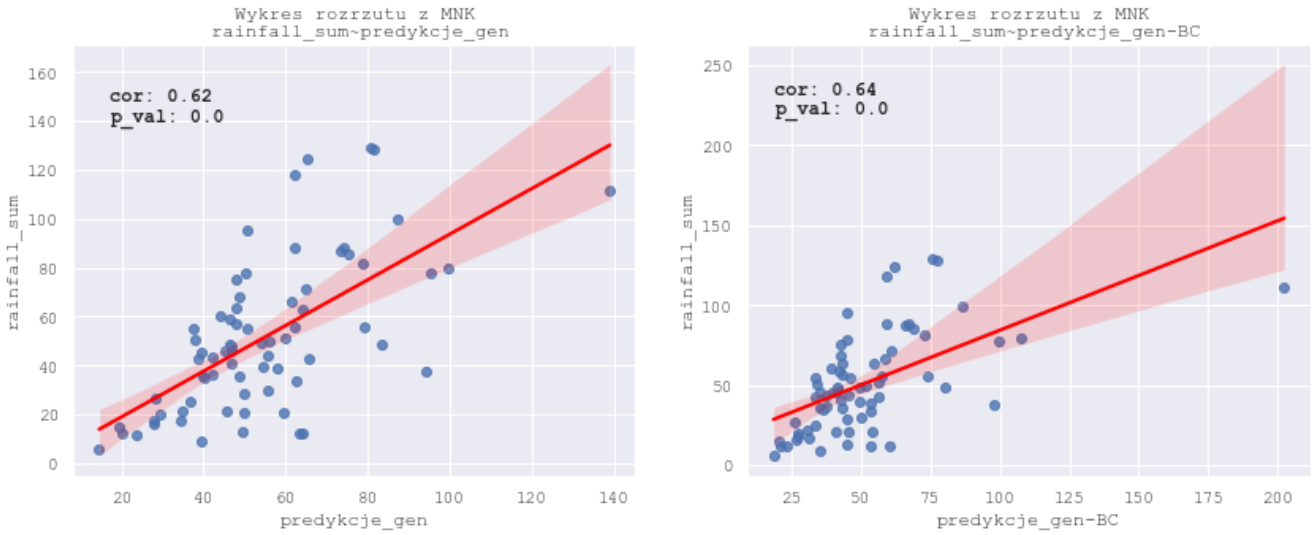
Tablica 6: Metryki dopasowania dla modeli gen i gen-BC

	MSE	RMSE	MAE	MAPE
gen	558.036	23.623	18.038	58.606
gen-BC	678.054	26.04	19.208	55.884

Źródło: Opracowanie własne.

W tabeli 6 zamieszczono wyniki out-of-sample metryk ewaluacyjnych dla modeli gen i gen-BC. w porównaniu z wynikami in-sample można wyciągnąć wniosek o lepszej generalizacji modelu gen, uzyskał on nawet niższe wartości metryk na nowych danych, jednocześnie model gen-BC można uznać za nadmiernie dopasowany - znacznie wyższe wartości wszystkich metryk. Jednak konstrukcja metryk opartych o kwadraty reszt - tj. duża kara za duże odchylenie prowokuje do postawienia pytania, czy Tabela 6 nie jest wynikiem maskującym występowanie outlierów - na to pytanie ma odpowiedzieć wykres predykcji zamieszczony poniżej.

Rysunek 4: Wykresy rozrzutu dla miesięcznej sumy opadów i predykcji modeli gen i gen-BC



Źródło: Opracowanie własne.

Na powyższych wykresach pierwszym co należy zauważyć jest jeden znaczący outlier będący efektem bardzo dużej reszty dla tej samej obserwacji w obydwu przypadkach. Model gen był bliższy jego wartości rzeczywistej (predykcja ok. 140 vs ok. 200 gen-BC, wartość obserwowana: poniżej 120). Jednocześnie ciężko określić, na którym wykresie punkty leżą bliżej wykreślonej linii regresji, istotny w obydwu przypadkach współczynnik korelacji Spearmana notuje jednak minimalnie wyższą wartość dla modelu gen-BC, wyniki wydają się jednak zachwiane występowaniem outliera, wydaje się, że model gen-BC z uwagi na lepsze wyniki metryk in-sample, testów diagnostycznych i korelację predykcji i wartości obserwowanych jest modelem lepszym. Wyniki testów diagnostycznych dla wszystkich stacji znajdują się w tabeli 8, a legenda kodowania stacji w tabeli 9.

Modele nieparametryczne wybrane wcześniej na podstawie wartości metryk w próbie *in-sample* wykorzystano do wygenerowania predykcji analogicznych do MNK. Wyniki zamieszczono w Tabeli 7.

Tablica 7: Metryki dopasowania dla modeli nieparametrycznych

	MSE	RMSE	MAE	MAPE
loess_span_100	653.425	25.562	19.780	0.372
np_4	617.927	24.858	19.753	0.380

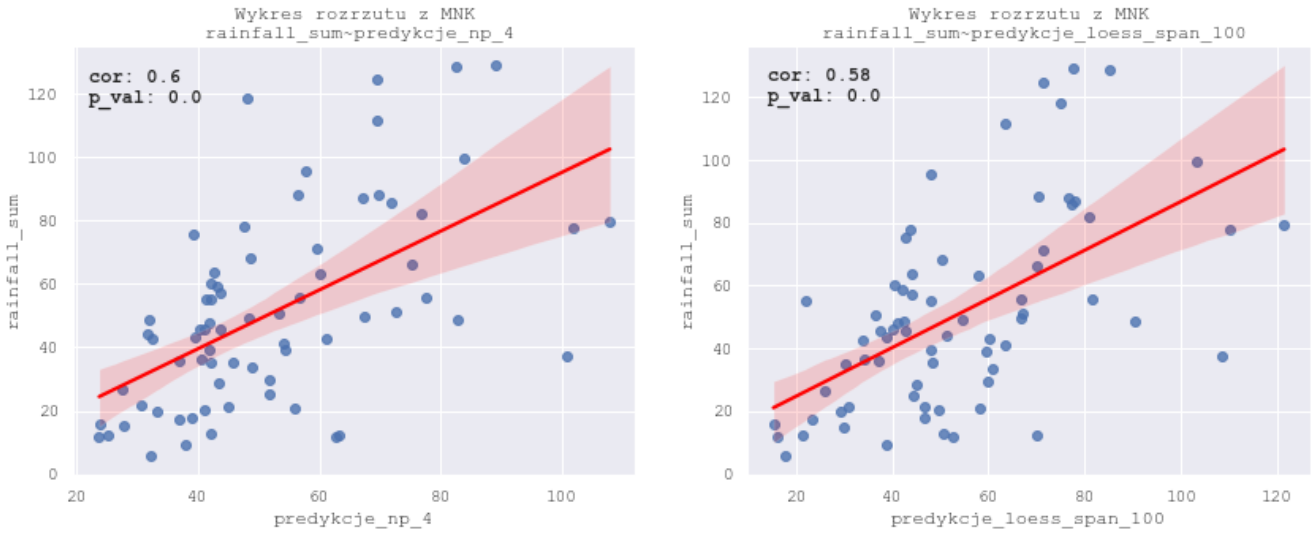
Źródło: Opracowanie własne.

Wyniki znacznie pogorszyły się w stosunku do próby *in-sample*. Najmniej zmieniła się watość metryki MAPE, która dalej jest na zadowalającym poziomie około 30%. Jak widać estymacja jądrowa osiąga korzystniejsze wartości metryk, co może świadczyć o lepszej generalizacji i mniejszym dopasowaniu do danych niż ma to miejsce w przypadku modelu LOESS.

Porównując wyniki modeli MNK i modeli nieparametrycznych można wnioskować, że patrząc na metrykę MAPE modele nieparametryczned deklasują modele parametryczne. Jeśli chodzi o pozostałe miary, estymacja jądrowa osiągna podobne wyniki co najlepszy model MNK. Co jest jednak problematyczne w przypadku modeli parametrycznych, czyli brak spełnienia założeń, nie jest już problem w przypadku modeli nieparametrycznych.

Na Wykresie 5. zaprezentowano rozrzut predykcji z obu modeli i miesięcznej sumy opadów, czyli zmiennej objaśnianej w modelu.

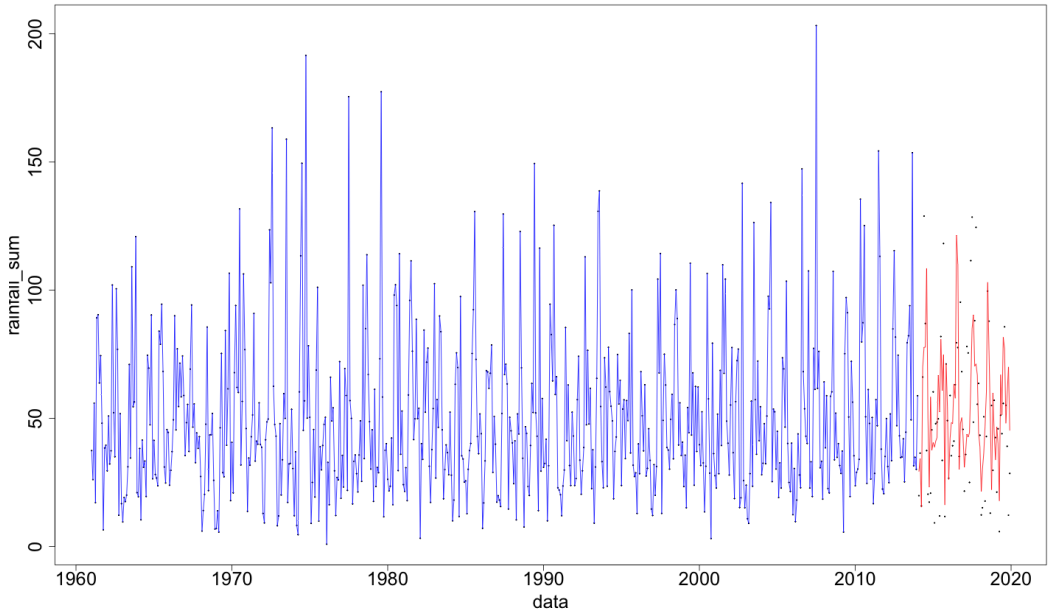
Rysunek 5: Wykresy rozrzutu dla miesięcznej sumy opadów i predykcji modeli nieparametrycznych



Źródło: Opracowanie własne.

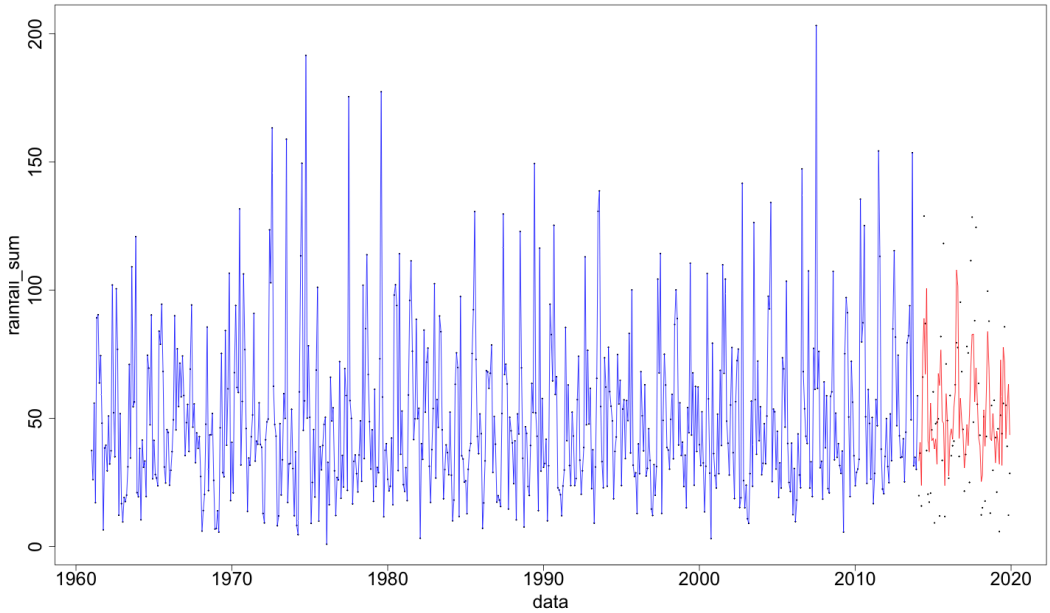
Na pierwszy rzut oka predykcje wygladają dość podobnie, brak jest znaczących outlierów, co jest niewątpliwym atutem w porównaniu do modeli parametrycznych. W przypadku obu nieparametrycznych podejść można zauważyć mniejsze wartości współczynnika korelacji rang Spearmana między zmienną objaśnianą, a predykcją, niż ma to miejsce w przypadku obu modeli MNK, jednak jak zostało to opisane powyżej, miara ta może cieprpieć z powodu występowania outliera.

Rysunek 6: Wartości obserwowane i predykcje zmiennej rainfall_sum, zmiany w czasie, model LOESS



Źródło: Opracowanie własne.

Rysunek 7: Wartości obserwowane i predykcje zmiennej rainfall_sum, zmiany w czasie, estymacja jądrowa



Źródło: Opracowanie własne.

Na rysunkach 6 i 7 przedstawiono, jak predykcje dla modeli nieparametrycznych wyglądają w porównaniu z obserwawanymi wartościami zmiennej rainfall_sum. Wykresy te pozwalają na obserwację, większego dopasowania się modelu LOESS do danych. Oba modele nie prognozują jednak tak wysokich skrajnych wartości jak ma to miejsce w naturze, ma to na wpływ dość wysoki zakres pasma, co powoduje większe wygładzenie zmiennych.

Na podstawie wykresów i statystyk dla obu nieparametrycznych podejść, można stwierdzić, że estymacja jądrowa lepiej sprawdza się w przypadku badanego zbioru danych. Jego wyniki są porównywalne z tymi uzyskanymi przy użyciu modelu MNK, ale nie są obarczone złamaniem założeń, dlatego używaniego go wydaje się bezpieczniejsze.

6 Podsumowanie

Przeprowadzone badanie miało na celu weryfikację przydatności klasycznej Regresji Liniowej i metod nieparametrycznych w modelowaniu i predykcji miesięcznej sumy opadów na terenie Polski. W toku analiz starano się rozwiązywać problemy powstałe przy korzystaniu z klasycznego podejścia MNK - niespełnianie założeń o liniowej zależności, sferyczności reszt czy tzw. "prawdziwości" modelu. Wyciągnięto również wniosek o złożoności modelowanego zjawiska wykraczającej poza wykorzystany zbiór zmiennych objaśniających co implikuje problem zmiennych pominiętych, będący kolejnym naruszeniem założeń estymatora MNK oraz powodującym w efekcie obciążenie oszacowań parametrów. Wydaje się również, że wnioski o złożoności można zastosować także w przypadku zmiennych objaśniających, co jest przesłanką do rozważania endogeniczności (zmienne objaśniające nie wydają się nie być pierwotnymi - niezależnymi od innych czynników). Pomimo braku zasadności wnioskowania statystycznego zdecydowano się na porównanie modeli pod kątem dopasowania do danych i generalizacji, celowo unika się na tym etapie słowa predykcja, bowiem zarówno lewa, jak i prawa strona równania obserwowalna jest w tym samym czasie, a próby modelowania na opóźnieniach zakończyły się niepowodzeniem. Pomijając wszelkie niedogodności, na szczególnym przypadku stacji w Suwałkach modele parametryczne i nieparametryczne cechowały się podobną jakością - w niektórych metrykach wygrywało jedno podejście, a w innych drugie. Z uwagi na pracę, jaką należy wykonać przy rozwiązywaniu problemów estymatora MNK i fakt, że nie zawsze praca ta daje oczekiwany efekt, można stwierdzić, że modele nieparametryczne wydają się po pierwsze bezpieczniejszym, po drugie szybszym i po trzecie konkurencyjnym, a nawet lepszym rozwiązaniem - nie popełniają tak dużych błędów jednostkowych.

Literatura

[1] Sirajum Munira Khan A.H.A. Rahmatullah Imon, Noora N Saleh. Modelling and prediction of rainfall in the u.k. using nonparametric approach. *Bangladesh Army University of Engineering Technology*, 2(2):1–10, 2020.

[2] Qi Lib Jeff Racine. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004.

[3] Jeffrey S. Racine Tristen Hayfield. The np package. *https://cran.r-project.org.*

7 Załącznik

Tablica 8: Testy diagnostyczne modeli gen i gen-BC dla wszystkich stacji biorących udział w badaniu (p_value)

	model	N	R	F	lambda	RESET (sq.)	RESET (cb.)	RESET PCA (sq.)	RESET PCA (cb.)	BP	BG	JB	RM
1	gen	381.0	0.456	0.0	0.0	0.137	0.0	0.5	0.001	0.0	0.49	0.0	0.0
	gen-BC	381.0	0.544	0.0	0.186	0.285	0.043	0.905	0.057	0.0	0.53	0.19	0.17
2	gen	599.0	0.303	0.0	0.0	0.003	0.0	0.085	0.015	0.0	0.49	0.0	0.0
	gen-BC	599.0	0.314	0.0	0.343	0.0	0.0	0.07	0.05	0.0	0.91	0.0	0.0
3	gen	502.0	0.36	0.0	0.0	0.968	0.0	0.511	0.002	0.0	0.02	0.0	0.0
	gen-BC	502.0	0.385	0.0	0.286	0.002	0.001	0.494	0.241	0.02	0.03	0.0	0.0
4	gen	584.0	0.368	0.0	0.0	0.0	0.0	0.087	0.004	0.0	0.0	0.0	0.0
	gen-BC	584.0	0.423	0.0	0.298	0.66	0.703	0.001	0.005	0.02	0.0	0.01	0.01
5	gen	584.0	0.327	0.0	0.0	0.001	0.0	0.226	0.003	0.0	0.46	0.0	0.0
	gen-BC	584.0	0.345	0.0	0.316	0.0	0.0	0.776	0.484	0.0	0.43	0.0	0.0
6	gen	472.0	0.347	0.0	0.0	0.0	0.0	0.621	0.0	0.0	0.23	0.0	0.0
	gen-BC	472.0	0.373	0.0	0.283	0.16	0.0	0.299	0.002	0.01	0.07	0.0	0.0
7	gen	599.0	0.314	0.0	0.0	0.801	0.081	0.311	0.105	0.0	0.66	0.0	0.0
	gen-BC	599.0	0.33	0.0	0.25	0.006	0.006	0.496	0.792	0.03	0.16	0.25	0.24
8	gen	393.0	0.275	0.0	0.0	0.999	0.071	0.187	0.273	0.0	0.88	0.0	0.0
	gen-BC	393.0	0.31	0.0	0.286	0.08	0.052	0.038	0.117	0.01	0.41	0.13	0.11
9	gen	395.0	0.383	0.0	0.0	0.001	0.0	0.793	0.0	0.0	0.08	0.0	0.0
	gen-BC	395.0	0.434	0.0	0.266	0.465	0.0	0.413	0.034	0.18	0.14	0.0	0.0
10	gen	439.0	0.121	0.0	0.0	0.074	0.034	0.019	0.03	0.0	0.58	0.0	0.0
	gen-BC	439.0	0.11	0.0	0.436	0.016	0.001	0.002	0.004	0.1	0.52	0.0	0.0
11	gen	584.0	0.307	0.0	0.0	0.167	0.151	0.615	0.23	0.0	0.11	0.0	0.0
	gen-BC	584.0	0.326	0.0	0.416	0.398	0.333	0.425	0.675	0.09	0.09	0.42	0.4
12	gen	509.0	0.355	0.0	0.0	0.01	0.004	0.827	0.002	0.0	0.32	0.0	0.0
	gen-BC	509.0	0.38	0.0	0.257	0.776	0.114	0.089	0.045	0.02	0.3	0.48	0.46
13	gen	639.0	0.367	0.0	0.0	0.0	0.0	0.631	0.001	0.0	0.38	0.0	0.0
	gen-BC	639.0	0.384	0.0	0.411	0.112	0.162	0.328	0.002	0.02	0.27	0.25	0.24
14	gen	515.0	0.18	0.0	0.0	0.079	0.101	0.262	0.048	0.0	0.69	0.0	0.0
	gen-BC	515.0	0.19	0.0	0.328	0.197	0.215	0.16	0.046	0.15	0.47	0.06	0.05
15	gen	330.0	0.337	0.0	0.0	0.547	0.0	0.79	0.056	0.0	0.82	0.0	0.0
	gen-BC	330.0	0.398	0.0	0.332	0.097	0.003	0.636	0.049	0.02	0.87	0.09	0.07
16	gen	584.0	0.219	0.0	0.0	0.661	0.026	0.034	0.006	0.0	0.58	0.0	0.0
	gen-BC	584.0	0.224	0.0	0.356	0.136	0.024	0.02	0.004	0.0	0.46	0.01	0.01
17	gen	506.0	0.319	0.0	0.0	0.42	0.001	0.029	0.0	0.0	0.42	0.0	0.0
	gen-BC	506.0	0.313	0.0	0.34	0.231	0.013	0.387	0.003	0.01	0.73	0.02	0.01
18	gen	637.0	0.39	0.0	0.0	0.0	0.0	0.316	0.0	0.0	0.43	0.0	0.0
	gen-BC	637.0	0.418	0.0	0.275	0.038	0.117	0.416	0.0	0.04	0.59	0.01	0.01

Źródło: Opracowanie własne.

Tablica 9: Legenda do Tabeli 5

numer	stacja
1	ZAKOPANE
2	KRAKÓW-BALICE
3	WROCŁAW-STRACHOWICE
4	WIELUŃ
5	ŁÓDŹ-LUBLINEK
6	LUBLIN-RADAWIEC
7	POZNAŃ-ŁAWICA
8	WARSZAWA-OKĘCIE
9	SIEDLCE
10	ŚWINOUJŚCIE
11	SZCZECIN
12	MŁAWA
13	KOŁOBRZEG-DŹWIRZYNO
14	USTKA
15	ŁEBA
16	HEL
17	ELBLĄG-MILEJEWO
18	SUWAŁKI

Źródło: Opracowanie własne.