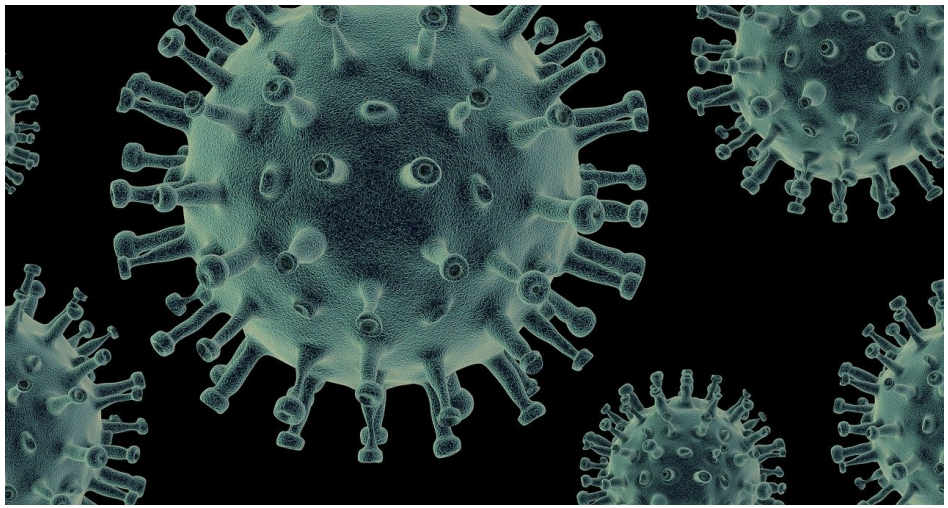
Gruntowna analiza szeregów czasowych - przykład.

31 maja 2021



Przykład - Temperatura w Tokio 1875-2010








Przykład

Dysponujemy danymi dotyczęcymi rocznej temperatury w Tokio w latach 1875-2020.

- **Dopasowanie modelu do danych**. Bazując na danych w latach 1875-2010 znajdziemy model w postaci:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t,$$

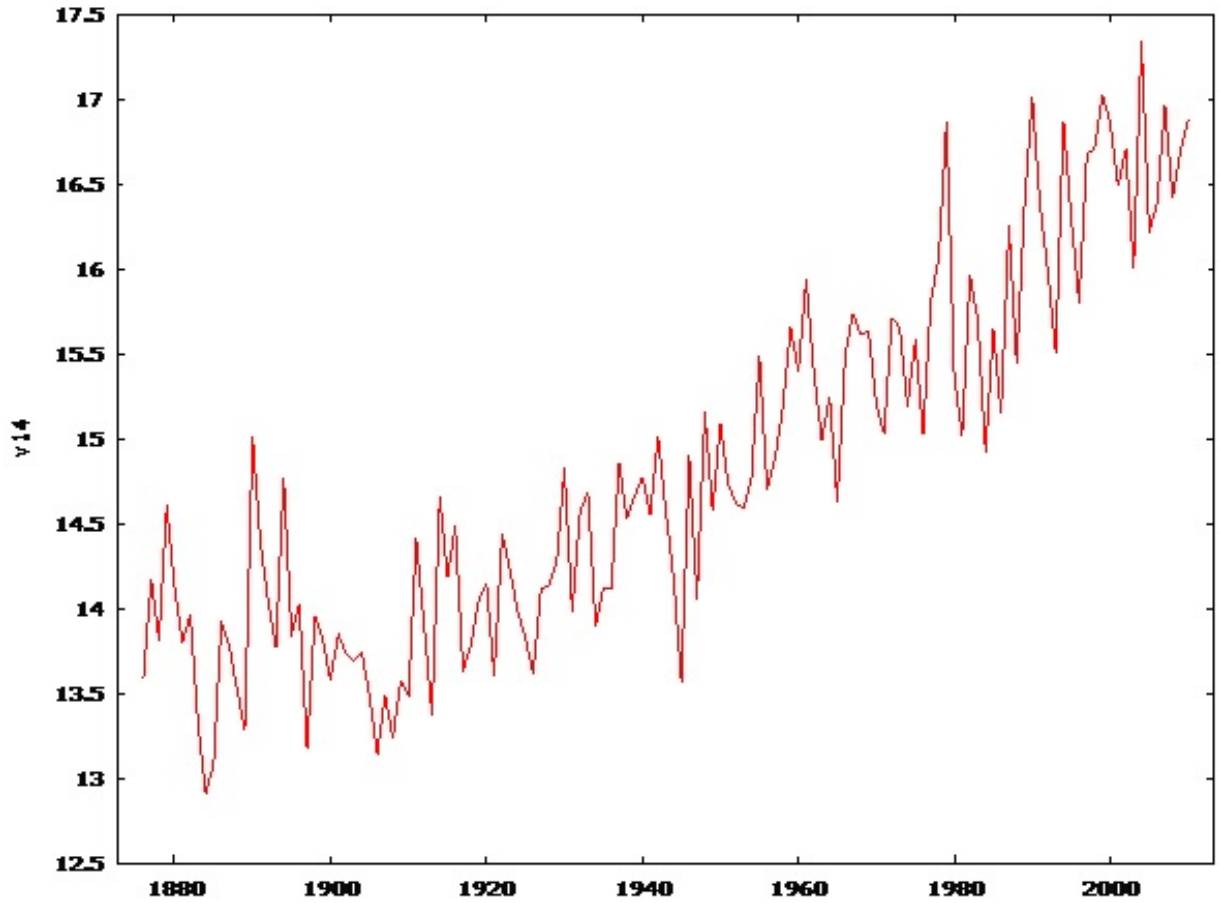
- **Analiza reszt**. Po dopasowaniu modelu do danych zastosujemy testy weryfikujące model.
- **Prognozowanie**. Na podstawie modelu opracujemy prognozy na lata 2011-2020.
- **Porównanie prognoz i przyszłych danych**. Porównujemy prognozy na lata 2011-2020 z rzeczywistymi danymi.

Procedury analiz reszt

1. Testy normalności (Jarque-Bera, Shapiro-Wilk, chi-kwadrat zgodności, graficzna ilustracja);
2. Testy autokorelacji (Durbin-Watson), w razie potrzeby procedura Cochrane'a-Orcutta;
3. Test homoskedastyczności (White);

Dane

Poniższy wykres przedstawia roczną temperturę w Tokio w latach 1875-2010.



Dopasowanie modelu - metoda najmniejszych kwadratów

Dopasowujemy model postaci

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t,$$

gdzie współczynniki obliczamy metodą najmniejszych kwadratów. Tabela przedstawia wyniki.

	współczynnik	$S_{\hat{\beta}_i}$	$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$	p-value
β_0	13.9397	0.156263	89.21	$3.64*10^{-119}$
β_1	-0.0185648	0.00991388	-1.873	$0.0634 > 0.05$
β_2	0.000521191	0.000169052	3.083	0.0025
β_3	$-1.68769 * 10^{-6}$	8.1726610^{-7}	-2.065	0.0409

Współczynnik β_1 jest nieistotny, więc rozważamy prostszy model:

$$Y_t = \beta_0 + \beta_1 t^2 + \beta_2 t^3 + \epsilon_t.$$

Dopasowanie modelu - metoda najmniejszych kwadratów

	współczynnik	$S_{\hat{\beta}_i}$	$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$	p-value
β_0	13.6844	0.0771209	177.4	<0.0001
β_1	0.000214423	$4.21307 * 10^{-5}$	5.089	< 0.0001
β_2	$-2.82426 * 10^{-07}$	3.26740e-07	-0.8644	$0.3890 > 0.05$

Ponieważ współczynnik β_2 jest nieistotny rozważamy prostszy model

$$Y_t = \beta_0 + \beta_1 * t^2 + \epsilon_t.$$

Dopasowanie modelu - metoda najmniejszych kwadratów

Kolejne wyniki można znaleźć w tabeli.

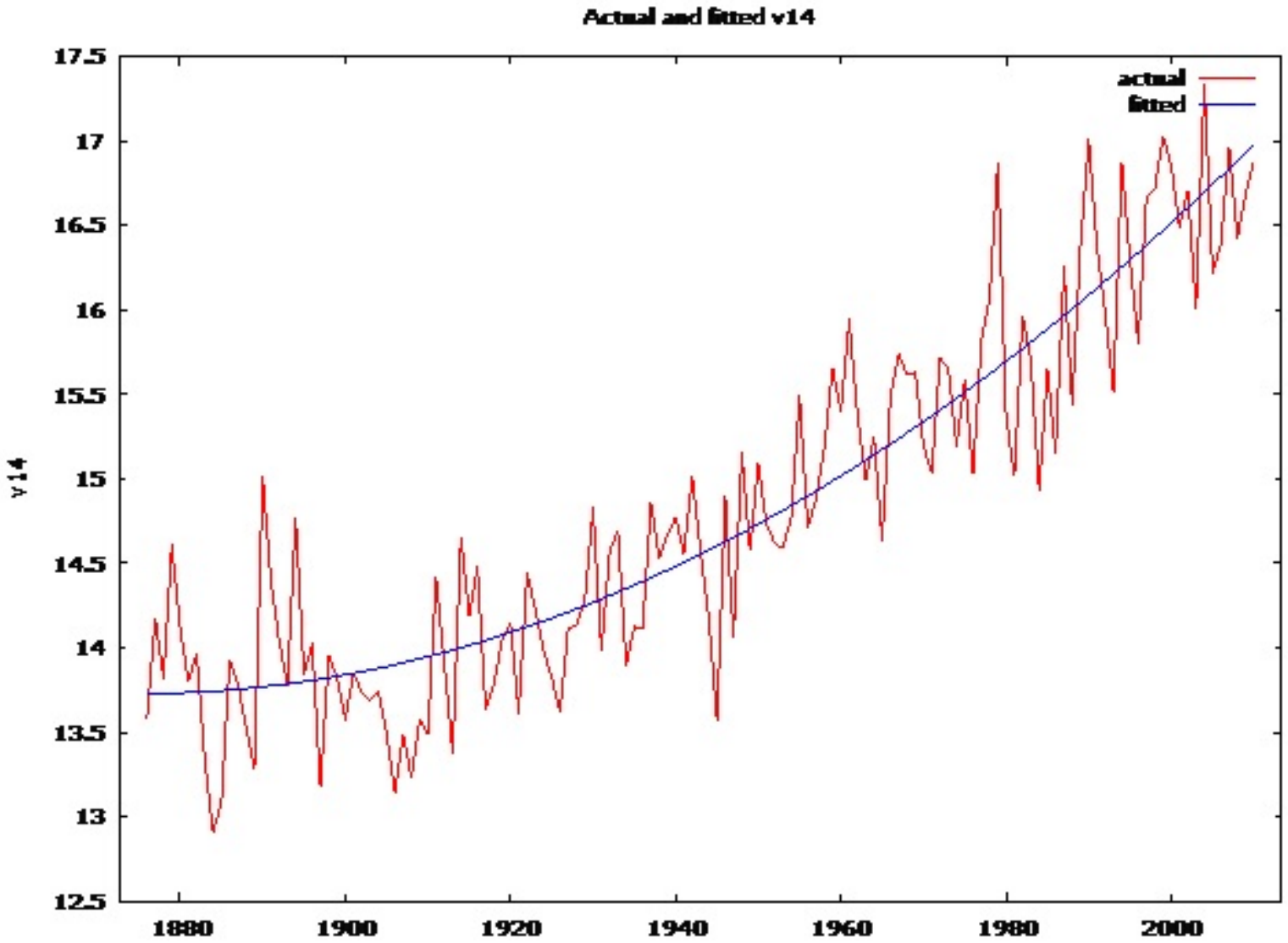
	współczynnik	$S_{\hat{\beta}_i}$	$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$	p-value
β_0	13.7287	0.0575963	238.4	$6.86 * 10^{-177}$
β_1	0.000178514	$7.00177 * 10^{-6}$	25.50	$4.78 * 10^{-53}$

Ponieważ oba współczynniki są istotne nie możemy zredukować modelu. Nasz model ma postać

$$Y_t = 13.7287 + 0.000178514 * t^2 + \epsilon_t.$$

Wykres danych i wielomianu regresji można znaleźć na następnym slajdzie.

Dopasowanie modelu - metoda najmniejszych kwadratów



Analiza reszt

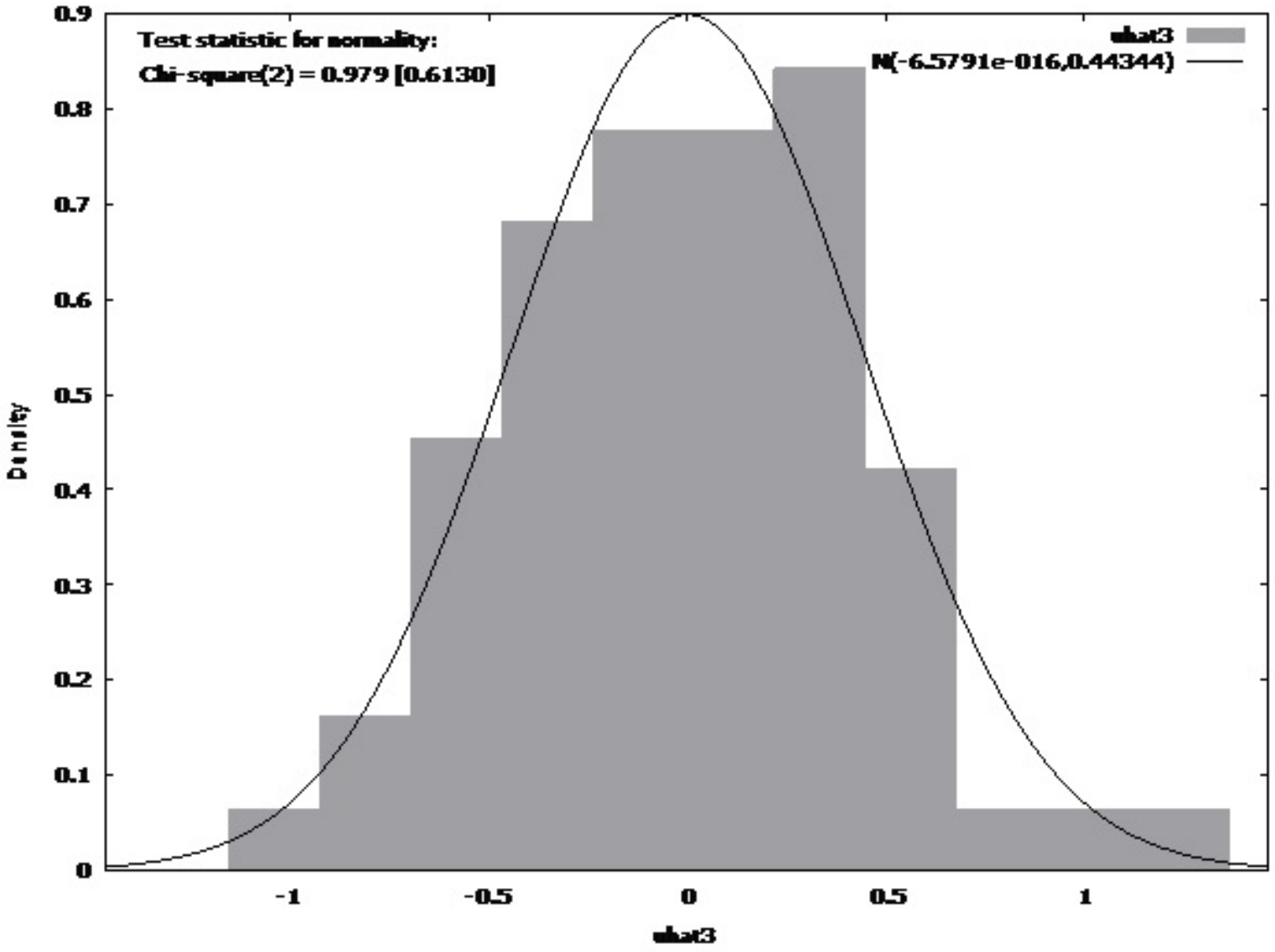
Wyznaczamy reszty z modelu:

$$\hat{\epsilon}_t = Y_t - 13.7287 - 0.000178514 * t^2.$$

Wykonujemy następujące kroki

- testy normalności rozkładu reszt:
  - graficzna ilustracja: histogram i wykres kwantylowy (ang. QQ-plot);
  - testy formalne;
- testy autokorelacji;
- testy homoskedastyczności.

Analiza reszt - ilustracja graficzna (histogram)

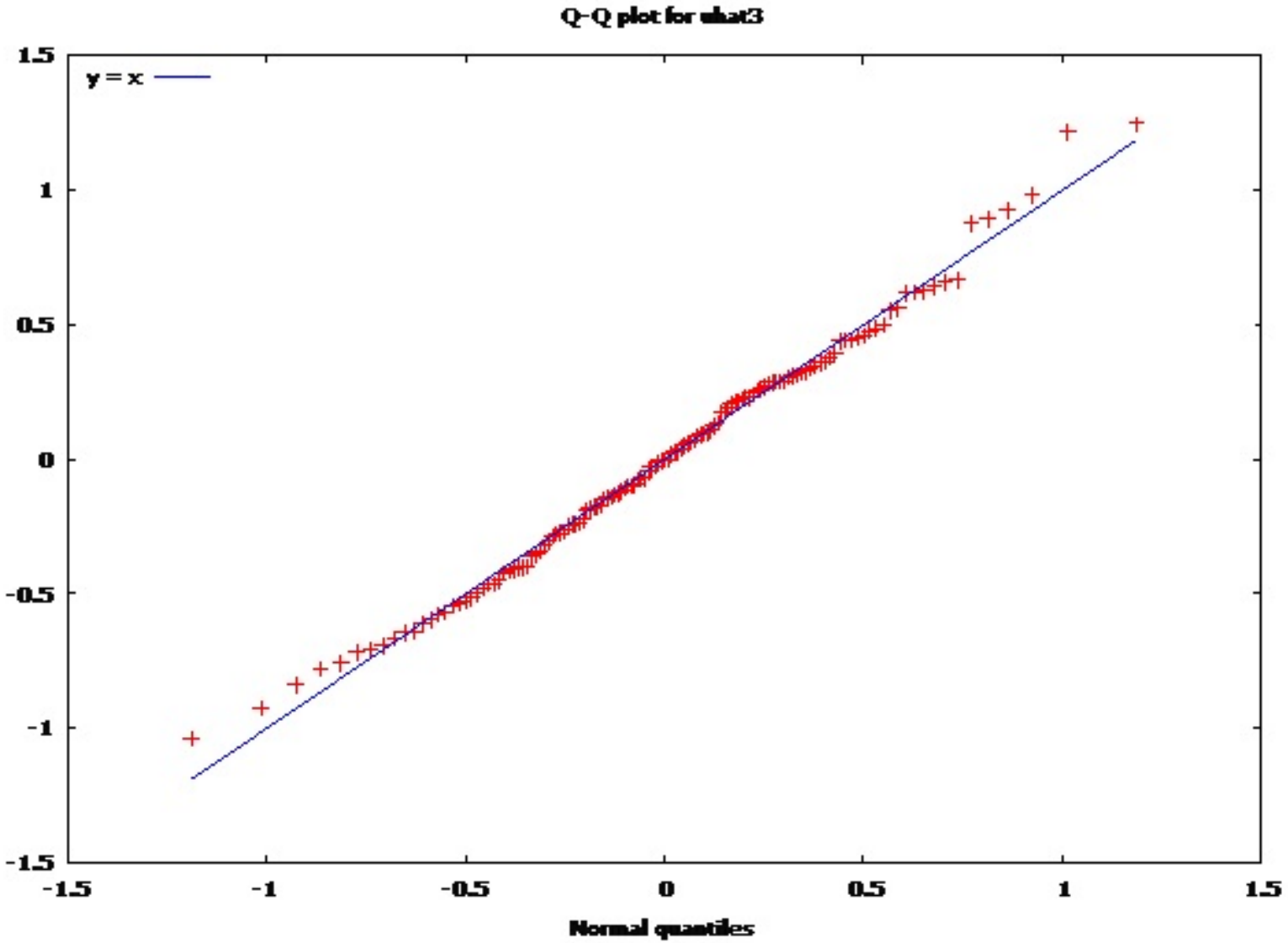


Analiza reszt - ilustracja graficzna (histogram)

Interpretacja histogramu

- Najbliższy rozkład normalny dla rozkładu reszt to $\mathcal{N}(0, 0.44244)$;
- Na pierwszy rzut oka histogram i gęstość rozkładu normalnego są stosunkowo blisko siebie;
- p-value dla testu chi-kwadrat zgodności z rozkładem normalnym wynosi $0.6130 > 0.05$;
- Test **potwierdza hipotezę o normalności rozkładu reszt**.

Analiza reszt - ilustracja graficzna (wykres kwantylowy)



Analiza reszt - ilustracja graficzna (wykres kwantylowy)

Interpretacja wykresu kwantylowego

- Na pierwszy rzut oka wykres kwantylowy leży na przekątnej $y = x$;
- Wykres kwantylowy potwierdza hipotezę, że reszty pochodzą z rozkładu normalnego.

Analiza reszt – formalne testy normalności

Aby potwierdzić lub zaprzeczyć hipotezie o normalności reszt zastosujemy następujące testy normalności.

Test	Statistyka	p-value
Doornik-Hansen	0.978766	0.613004
Shapiro-Wilk	0.99238	0.682274
Lilliefors	0.0358577	≈ 1
Jarque-Bera	0.946798	0.622881

Wszystkie testy potwierdzają hipotezę o normalności reszt, ponieważ każda *p-value* jest większa niż 0.05.

Wniosek

Hipoteza, że ϵ_t ma rozkład normalny została eksperymentalnie potwierdzona. Najbliższy rozkład normalny to $\mathcal{N}(0, 0.44344)$.

Analiza reszt - testy autokorelacji rozkładu

- Statystyka Durbina-Watson osiąga wartość

$$DW = 1.780623.$$

- Ponieważ $DW < 2$, stąd testujemy

$$H_0 : \rho = 0 \quad \text{przeciwko} \quad H_1 : \rho > 0;$$

- Mamy $n = 2010 - 1875 = 135$ obserwacji, oraz liczbę parametrów $k = 1$, zatem dolne i górne wartści krytyczne mają wartośi:

$$d_L = 1.704 \quad \text{oraz} \quad d_U = 1.7338.$$

- Mamy więc

$$DW = 1.780623 > 1.7338 = d_U,$$

zatem test Durbina-Watsona **przyjmuje hipotezę** $\rho = 0$;

- Innymi słowy test Durbina-Watsona potwierdza hipotezę o braku autokorelacji reszt.

Analiza reszt-test homoskedastyczności rozkładu

- Ostatnią potencjalną przeszkodą dla uznania modelu jako dobrze dopasowanego jest test White'a badający homoskedastycznośc reszt;
- Statistyka testu White'a ma postać $W = 0.172723$, a p-value $= 0.917263$;
- Zatem test White potwierdza hipotezę o normalności reszt ϵ_t;

Wniosek końcowy

Wszystkie testy normalności, jak również testy Durbina-Watsona oraz White'a, a także histogram i wykres kwantylowy wskazują na fakt, że nie ma podstaw, aby kwestionować model

$$Y_t = 13.7287 + 0.000178514 * t^2 + \epsilon_t.$$

Dokładniej nie mamy podstaw aby odrzucić hipotezę, że ϵ_t jest ciągiem niezależnych zmiennych losowych o stałym rozkładzie normalnym. Najbliższy rozkład normalny to $\mathcal{N}(0, 0.44344)$.

Prognozowanie

Model jest zbudowany na podstawie danych z okresu 1875 – 2010;

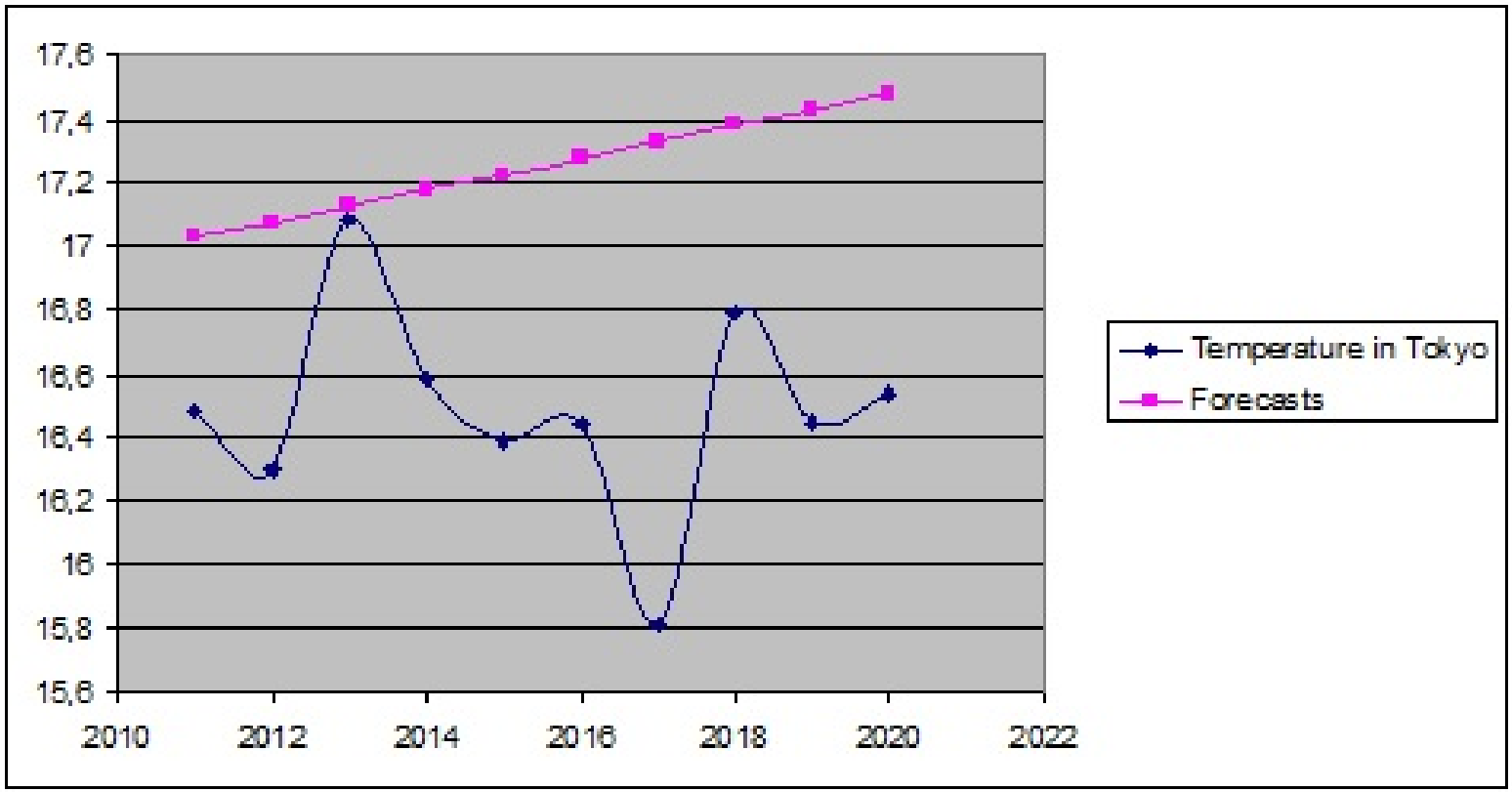
- Obliczymy prognozy na lata 2011 – 2020;
- Następnie porównamy prognozy z prawdziwymi wartościami temperatur w latach 2011 – 2020;
- W celu podsumowania wielkości błędów prognoz wykorzystami wskaźniki *ex-post*.

Prognozowanie i porównanie prognoz

Rok	Prognoza	Prawdziwa wartość w ^oC
2011	17,03049494	16,48333333
2012	17,07922927	16,3
2013	17,12832062	17,08333333
2014	17,17776899	16,58333333
2015	17,2275744	16,38333333
2016	17,27773683	16,44166667
2017	17,3282563	15,80833333
2018	17,37913279	16,79166667
2019	17,4303663	16,45
2020	17,48195685	16,53333333

Porównanie prognoz

Kolejny wykres wskazuje, że prawdziwe wartości zostały drastycznie przeszacowane przez model.



Porównanie prognoz

Wskaźniki *ex-post*.

Polska nazwa	Angielska nazwa	Wartość	Komentarz
Średni błąd	Mean Error (ME)	$-0,768250396$	Prognozy przeszacowały wartość
Średni bezwzględny błąd	Mean Absolute Error (MAE)	0,768250396	
Średni bezwzględny błąd procentowy	Mean Absolute Percentage Error (MAPE)	$4,7\%$	Średni błąd to ok. $4,7\%$ prawdziwych wartości
Współczynnik Theila	Theil idex I^2	0,002644554	

Porównanie prognoz

Rozkład współczynnika Theila wskazuje na udział poszczególnych źródeł błędów.

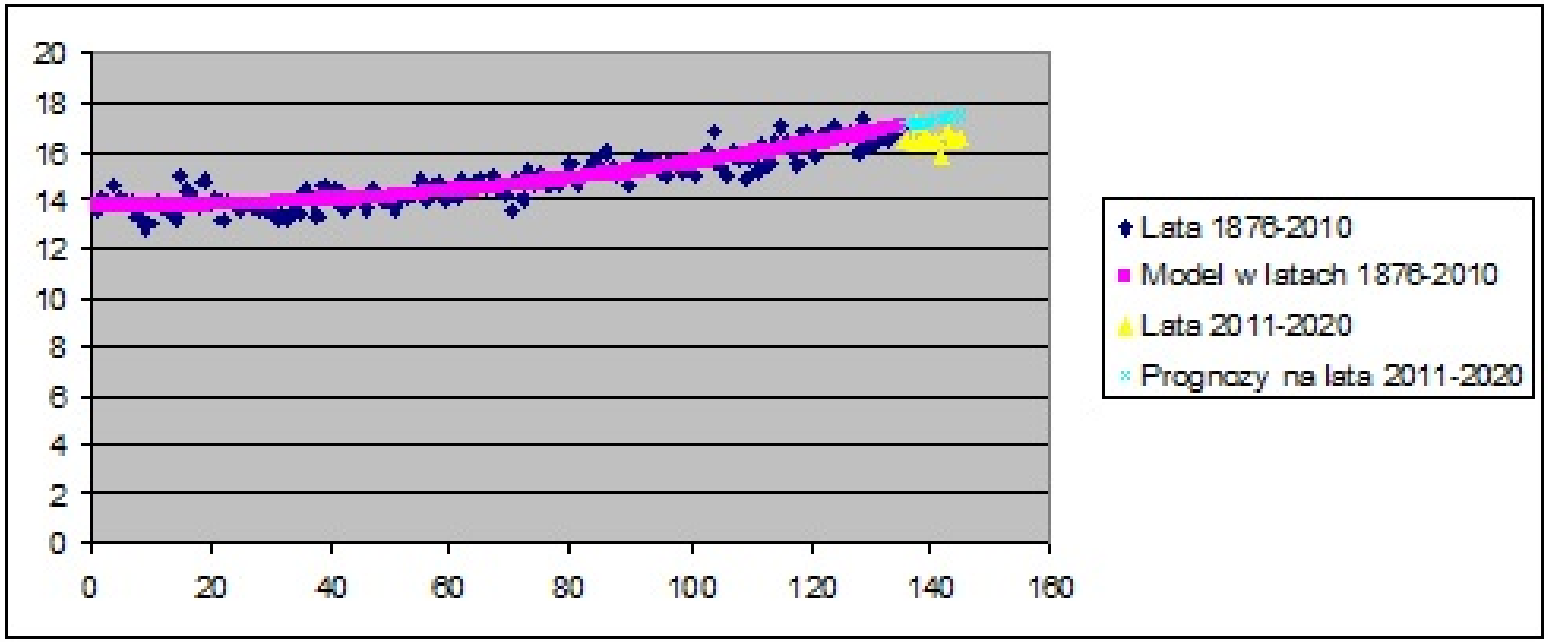
	Obciążoność	Brak elastyczności	Inne źródła
Czego dotyczy błąd?	średniej	zmienności	losowości, źródło nie do opanowania
Udział	$\tilde{I}_1^2 = 82.09\%$	$\tilde{I}_2^2 = 4.80\%$	$\tilde{I}_3^2 = 13.11\%$

Porównanie prognoz

Źródła błędów prognoz wg współczynnika Theila

- **Obciążoność**: średnie arytmetyczne prognoz i dokładności różnią się o zbyt dużą wartość (82.09%);
- **Brak elastyczności**: wahania prognoz i dokładnych wartości różnią się o zbyt duże wartości (4.8%);
- **Inne źródła błędów**: (13.11%);

Prognozowanie i porównanie



Podsumowanie

Model temperatury w Tokio ma postać:

$$Y_t = 13.7287 + 0.000178514 * t^2,$$

gdzie t jest indeksem roku licząc od roku 1875. Analiza reszt wskazuje, że model jest dobrze zweryfikowany na podstawie danych z lat $1875 - 2010$. Jednak nie do końca odzwierciedliło dokładności prognoz, które znacznie przeszacowały dokładnie wartości w latach $2011 - 2020$. Świadczy o tym ujemna wartość MA. Współczynnik Theila wskazuje na to, że przeszacowanie średniej stanowi 82% udziału błędów prognoz. Znacznie mniejszy udział w błędach prognoz 4.8% stanowi brak elastyczności, czyli nieprawidłowe określenie wahań przyszłych wartości. Ponieważ inne źródła błędów stanowią 13.11%, model jest daleki od doskonałego.