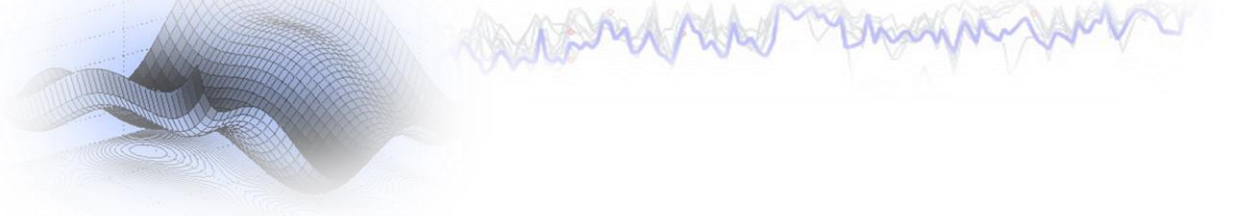


STATISTIČKA ANALIZA I PREDOBRAĐA PODATAKA. VREDNOVANJE REZULTATA



Željka Ujević Andrijić
zujevic@fkit.unizg.hr



SADRŽAJ

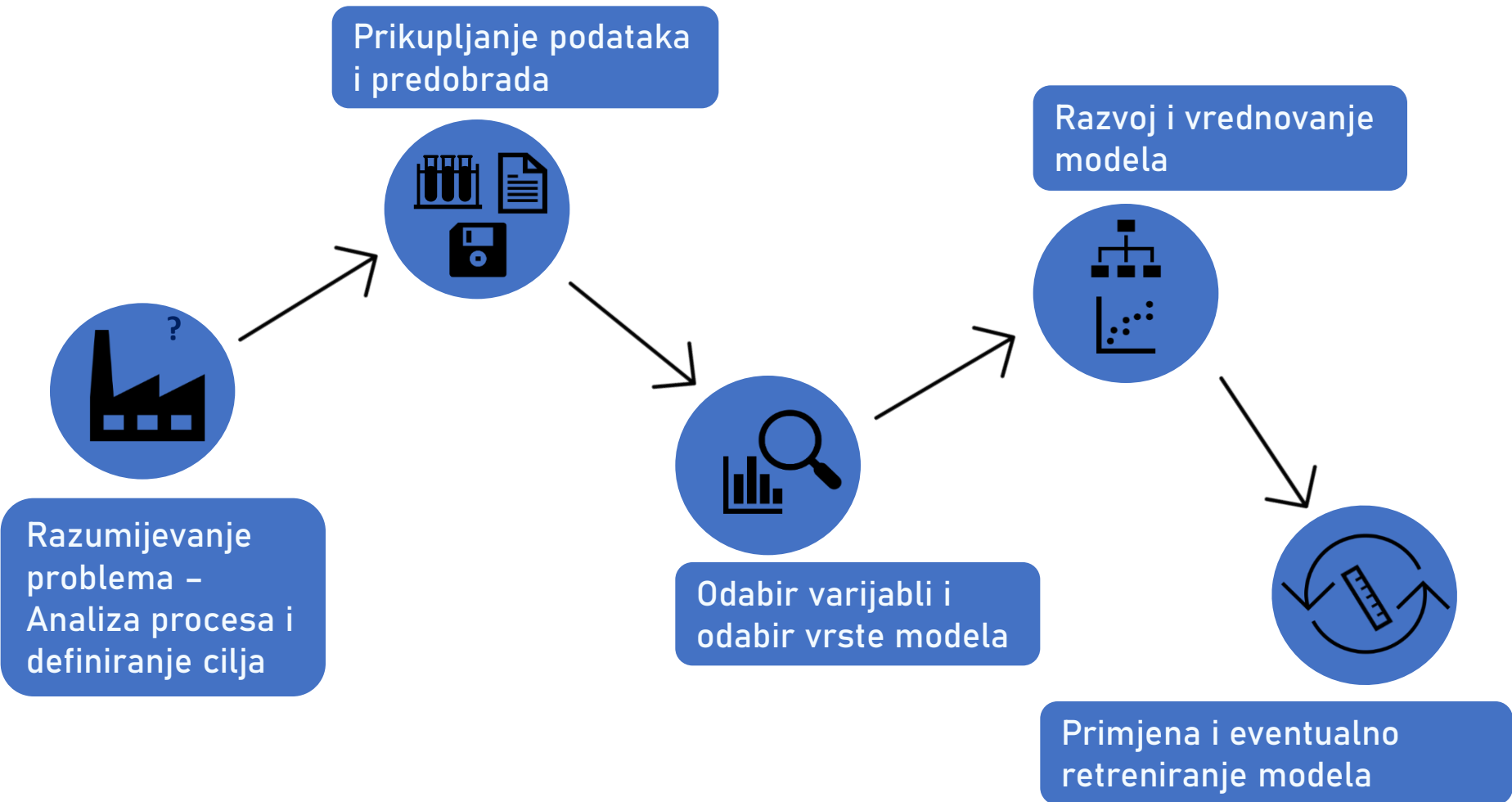
DESKRIPTIVNA STATISTIKA

PREDOBRADA PODATAKA (*outlieri*, filtriranje, skaliranje,...)

KRITERIJI VREDNOVANJA MODELA

FUNKCIJE U PYTHON-u

POSTUPAK RAZVOJA MODELA STR



*“Inferring models from observations and studying their properties is **really** what science is about.”*

L. Ljung

DESKRIPTIVNA STATISTIKA

Deskriptivna (opisna) statistika

- Prikupljanje, uređivanje i grupiranje podataka
- Tabelarno i grafičko prikazivanje
- Brojčani pokazatelji osnovnih karakteristika promatrane pojave
- Deskriptivna statistika analiziranog skupa podataka
- Prvi korak u statističkoj analizi ili dio složenije analize
- Može se koristiti i kod analize rezultata modela

DESKRIPTIVNA STATISTIKA

Metode (mjere) deskriptivne statistike:

- **Mjera centralne tendencije**

(aritmetička sredina, centralna vrijednost, dominantna vrijednost)

- **Kvartili**

- **Mjera raspršenja (disperzije) rezultata**

(standardna devijacija, varijanca, minimalna i maksimalna vrijednost podataka, raspon podataka, pogreška aritmetičke sredine)

- Određivanje indeksa **sploštenosti i asimetrije** distribucije

MJERE CENTRALNE TENDENCIJE – Aritmetička sredina i medijan

- **Aritmetička sredina** - mjera „centralne tendencije“ varijable

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad n - \text{broj podataka}$$

- Ako u podacima postoje ekstremne vrijednosti ili ako je distribucija asimetrična primjenjuje se **centralna vrijednost (medijan)**.

Ako je broj podataka **neparan**, medijan je vrijednost varijable središnjeg člana niza prema veličini.

Ako je broj podataka **paran**, medijan je jednak poluzbroju vrijednosti središnjih dvaju članova niza.

- Kod normalne raspodjele srednja vrijednost i medijan su identični.

MJERE CENTRALNE TENDENCIJE – KVANTILI, *MOD*

- Medijan se ubraja u kvantile.
- **Kvantil** - vrijednosti numeričke varijable koja uređen numerički ili redoslijedni niz dijele na jednakobrojne dijelove.
- **Kvartil** dijeli skup podataka na 4 jednaka dijela.

Gornji kvartil (*UQ*) - vrijednost od koje je **75%** podataka manje

Donji kvartil (*LQ*) - vrijednost od koje je **25%** podataka manje

- **Dominantna vrijednost (*mod*)**
najčešće postignuta vrijednost u nizu mjerenja

Primjer:

Niz: 10 14 7 12 1 5 3 7 2

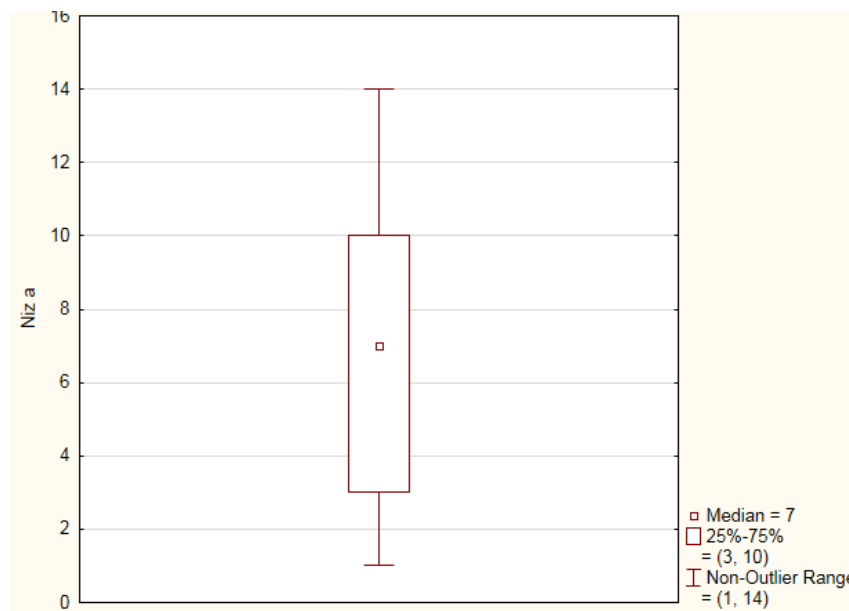
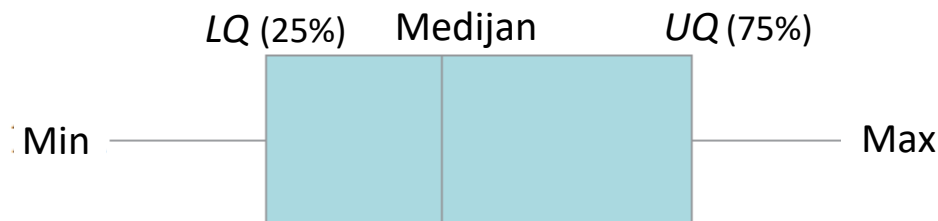
Prije određivanja gornjeg i donjeg kvartila nužno je urediti podatke prema veličini:

Uređeni niz: 1 2 3 5 7 7 10 12 14

UQ = 10; *LQ* = 3; *Mod* = 7

MJERE DISPERZIJE – RASPON, INTERKVARTIL

- **Raspon** rezultata
Razlika između najvećeg i najmanjeg rezultata.
- **Interkvartil** (engl. *interquartile range*)
Kvartilni raspon rezultata središnjih 50% članova niza uređenih podataka po veličini.
Razlika gornjeg i donjeg kvartila: $I_Q = UQ - LQ$
- Dijagram s pravokutnikom – **Box-Plot**



MJERE RASIPANJA (STANDARDNA DEVIJACIJA)

- **Standardna devijacija**

govori koliko su vrijednosti uzorka raspršene oko aritmetičke sredine (prosječno kvadrirano odstupanje od aritmetičke sredine):

$$s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_i - \bar{x})^2}$$

σ^2 – **Varijanca**

- **Koeficijent varijacije (V)** ili **relativna standardna devijacija**

omjer standardne devijacije i aritmetičke sredine.

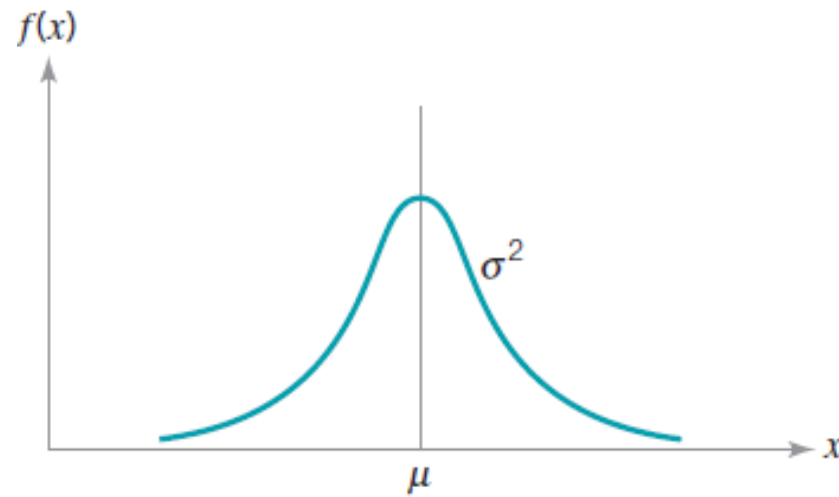
- Služi za **usporedbu varijabilnosti** mjernih rezultata čije su vrijednosti različitog reda veličine.

$$V = \frac{s}{\bar{x}} * 100\%$$

>30 % nije dobro,
10-30% prihvatljivo,
<10 % jako dobro

NORMALNA RASPODJELA (Gaussova distribucija)

- Mjereći različite pojave i karakteristike u prirodi i tehnici utvrđeno je da se podaci većinom **grupiraju** oko **centralne (srednje)** vrijednosti;
- Kad se grafički prikažu (podaci na horizontalnoj osi, broj podataka na vertikalnoj osi) tvore krivulju zvonolikog oblika poznatiju kao **normalna razdioba**;
- Normalna raspodjela je **simetrična** s vršnom vrijednosti kod srednje vrijednosti (50% podataka je manje od srednje vrijednosti, a 50% je veće od srednje vrijednosti).



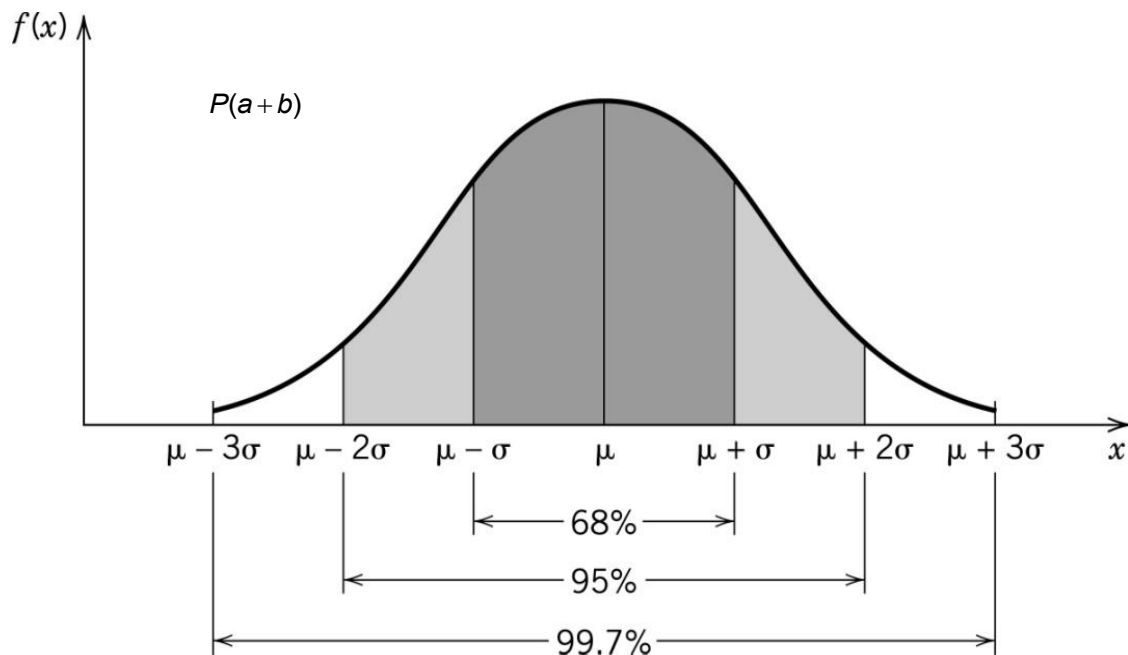
NORMALNA RASPODJELA

$$P(a < x < b) = \int_a^b f(x) dx$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \bar{x})^2}{2\sigma^2}\right]$$

a, b - konstante

$f(x)$ - funkcija gustoće vjerojatnosti



- Kada je zvonolika krivulja uska (podaci su koncentriraniji), σ je mala.
- Ako su podaci dosta raspršeni, a zvonolika krivulja plosnata, onda je σ relativno velika.
- Interval od **1 σ** udaljene od srednje vrijednosti u oba smjera obuhvaća oko **68%** razdiobe.
- Interval od **2 σ** udaljene od srednje vrijednosti obuhvaća oko **95%** razdiobe,
- Interval od **3 σ** devijacije obuhvaća **99.7%** razdiobe.

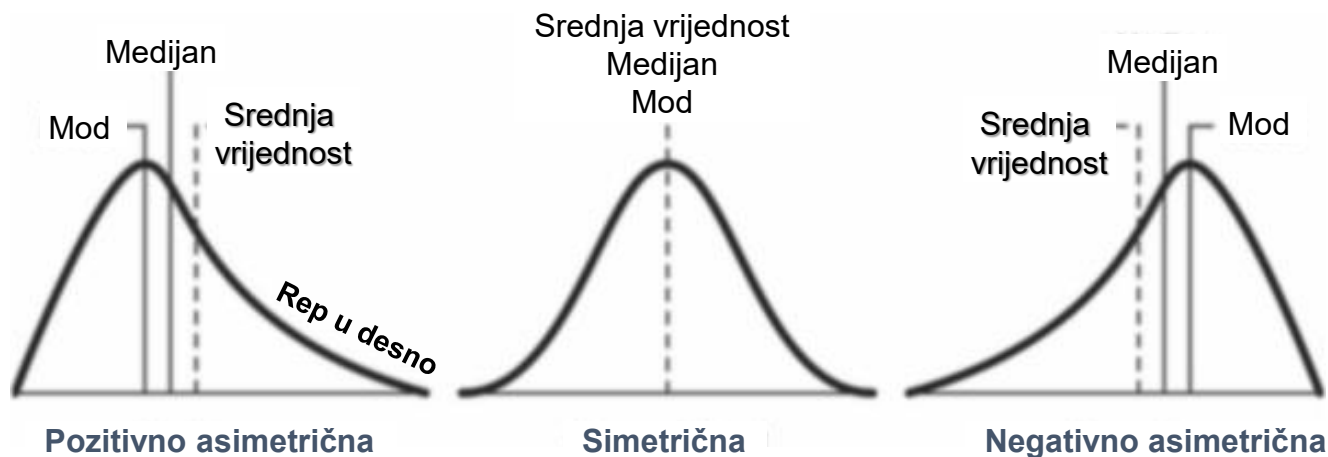
NORMALNA RASPODJELA (simetričnost)

- **Simetričnost distribucije** (engl. *Skewness*)

Indeks asimetrije:

$$\alpha = \frac{3(\bar{x} - M)}{\sigma}$$

$\alpha = 0 \rightarrow$ u potpunosti simetrična distribucija



NORMALNA RASPODJELA (zakrivljenost)

- Zakrivljenost (spljoštenost) distribucije** (engl. *Kurtosis*)

Ako je distribucija normalna, vrijednost je bliže nuli (od -1 do +1).

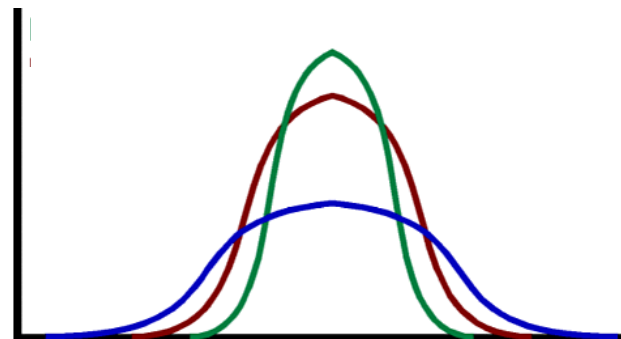
- Pozitivne vrijednosti - uska i visoka distribucija,
- Negativne - spljoštena i široka distribucija.

$$\text{Zakrivljenost} = [n \cdot (n+1) \cdot M_4 - 3 \cdot M_2^2 \cdot (n-1)] / [(n-1) \cdot (n-2) \cdot (n-3) \cdot \sigma^4]$$

$$M_j = \Sigma (x_i - \bar{x})^j$$

n – broj mjerenja

σ^4 - standardna devijacija na 4. potenciju



HISTOGRAMI

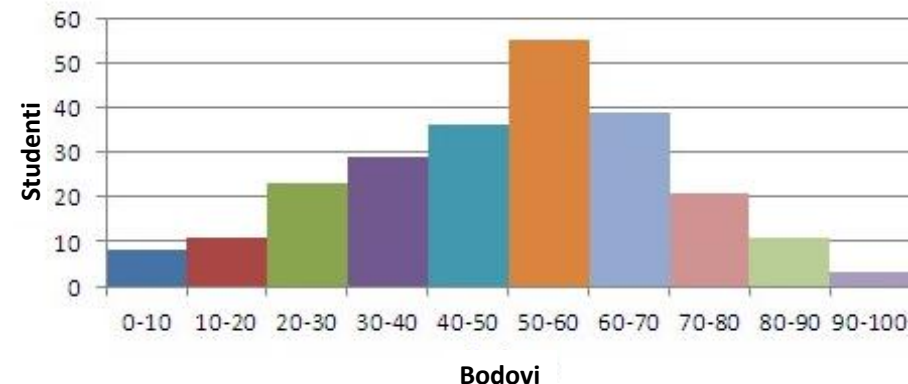
- ⊙ Histogram prikazuje razdiobu podataka, tj. učestalost pojavljivanja pojedinih vrijednosti.
- ⊙ Sastoji se od frekvencija prikazanih nizom pravokutnika. Visina pravokutnika jednaka je gustoći frekvencije pojedinog intervala, odnosno frekvenciji podijeljenoj s širinom intervala.
- ⊙ Histogram prikazuje broj mjerenja svrstanih u razdvojene kategorije (razrede).
- ⊙ U praksi se često broj razreda odabire kao kvadratni korijen iz broja mjerenja.
- ⊙ Uglavnom je potrebno eksperimentiranje za određivanje prikladne širine razreda h .

Što vidimo na histogramu?

- **Lokaciju podataka** – Jesu li podaci pravilno centrirani?
- **Širinu raspršenja podataka**
- **Oblik raspodjele podataka**
(simetričan, nesimetričan, dugačak rep, dva pika)
- **Ekstremno male ili velike vrijednosti**

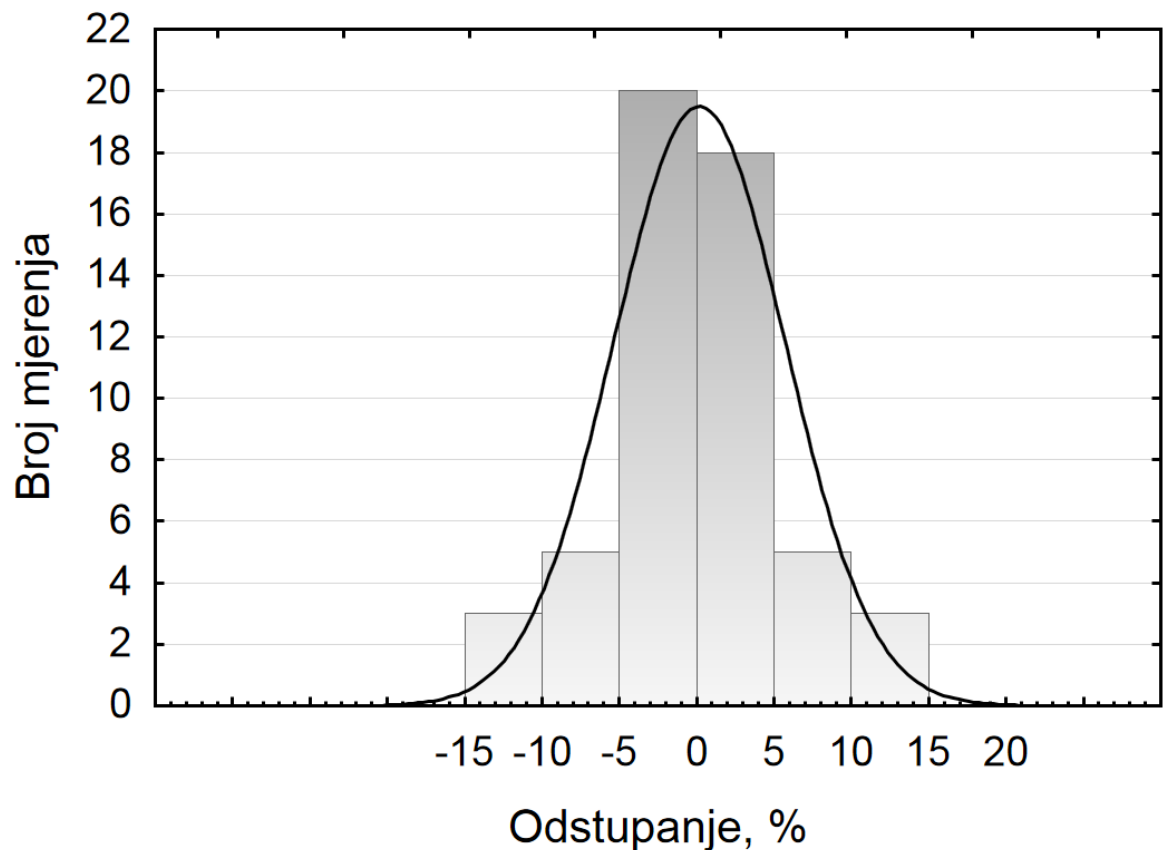
Na histogramu ne vidimo promjenu podataka u vremenu.

Primjer histograma - Rezultati ispita



HISTOGRAMI – Primjer prikaza odstupanja modela od eksperimentalne vrijednosti

Odstupanje konverzije izračunate prema modelu od eksperimentalne vrijednosti



Frekvencija odstupanja

Interval	Frekvencija
0 do 5	18
5 do 10	5
10 do 15	3
0 do -5	20
-5 do -10	5
-10 do -15	3

Knjižnice (*library*) za predobradu podataka u Pythonu



<https://pandas.pydata.org/>



SciPy

<https://www.scipy.org/>



NumPy

<https://numpy.org/>

Vizualizacija podataka:



<https://matplotlib.org/>



<https://scikit-learn.org/>

Funkcije u Pythonu za deskriptivnu statistiku



SciPy



pandas

Niz_podataka = [3, 5, 7, 9, 11]

Aritmetička sredina

Raspon

Minimum

Maksimum

Simetričnost distribucije

Zakrivljenost distribucije

Standardna devijacija

Varijanca

Metoda `stats.describe()` iz knjižnice **SciPy** daje **numerički sažetak osnovnih statističkih pokazatelja** za niz podataka.

```
1 from scipy import stats
2 stats.describe(Niz_podataka)
```

```
1 import pandas as pd
2 N = pd.Niz_brojeva
3 N.describe
```

Pandas metoda `.describe()` daje **tablični sažetak statistika** o podacima, što je vrlo praktično kod rada s tablicama ili serijama.

Napomena: Ovisno o verzijama knjižnica i/ili modula sintaksa može varirati. Provjeriti važeću dokumentaciju.

Funkcije u Pythonu za deskriptivnu statistiku



Aritmetička sredina

Medijan

Mod

Kvartili

Interkvartil

Standardna devijacija

Koeficijent varijacije

```

1 import numpy as np
2 from scipy import stats
3
4 aritmeticka_sredina = np.mean(Niz_podataka)
5 medijan = np.median(Niz_podataka)
6 mod = stats.mode(Niz_podataka)
7
8 Q1 = np.quantile(Niz_podataka, 0.25)
9 interkvartil = stats.iqr(Niz_podataka)
10
11 StandardnaDevijacija = np.std(Niz_podataka)
12
13 KoeficijentVarijacije = stats.variation(Niz_podataka)

```

Funkcije (grafičke) u Pythonu za deskriptivnu statistiku



Histogram

Box plot dijagram

Dijagram raspršenja (x-y plot ili *scatter plot*)

```

1  import matplotlib.pyplot as plt
2
3  plt.hist(Niz_podataka)
4
5  plt.boxplot(Niz_podataka)
6
7  plt.scatter(Niz_podataka1, Niz_podataka2)
8

```

Važnost deskriptivne analize

1. Razumijevanje podataka prije modeliranja (EDA – *Exploratory Data Analysis*)

- koje su **srednje vrijednosti, rasponi i odstupanja**,
- ima li **praznih podataka (NaN)** ili **outlier-a**,
- kakva je **raspodjela varijabli** (normalna, asimetrična, itd.).

2. Odabir i priprema značajki

Statističke mjere kao što su **varijanca, korelacija, standardna devijacija** pomažu:

- u **odabiru najvažnijih značajki**,
- u **uklanjanju redundantnih varijabli**,
- u **skaliranju podataka** (npr. *StandardScaler* koristi sredinu i devijaciju).

3. Provjera pretpostavki modela

Neki modeli (npr. linearna regresija) pretpostavljaju:

- linearnu vezu između varijabli,
- normalnu raspodjelu pogrešaka,
- nepostojanje multikolinearnosti.

4. Vrednovanje modela

statističke metrike (npr. R^2 , *MAE*, *RMSE*, preciznost, F1-score) služe za:

- **kvantitativno ocjenjivanje** uspješnosti modela,
- **usporedbu različitih modela**.

PREDOBRADA PODATAKA

Prikupljanje podataka iz baze podataka
postrojenja ili laboratorija

- Što više kvalitetnih (varijabilnih) podataka

Određivanje vremena uzorkovanja
podataka

- *Resempliranje* (kako bi se smanjio broj uzoraka, a zadržala informativnost signala).

Otkrivanje i uklanjanje ekstremnih
vrijednosti (engl. *outlier-a*)

- 3-sigma metoda;
- Hampel identifikator

Zamjena manjeg broja nedostajućih
vrijednosti

- Linearna, *Spline* (kubna) interpolacija, regresijske metode

Filtriranje podataka

- Filtar Loess, Lowes, Savitzky-Golay,...

Generiranje dodatnih izlaza

- *Spline* (kubni)
- MAR *Spline* algoritam

Odabir utjecajnih varijabli

- Koreliranost varijabli; PCA, PLS,...
- Konzultacije s tehnolozima

Detrendiranje

- Uklanjanje srednje vrijednosti (engl. *Remove means*)
- Uklanjanje linearnog trenda (engl. *Remove trends*)

Skaliranje podataka

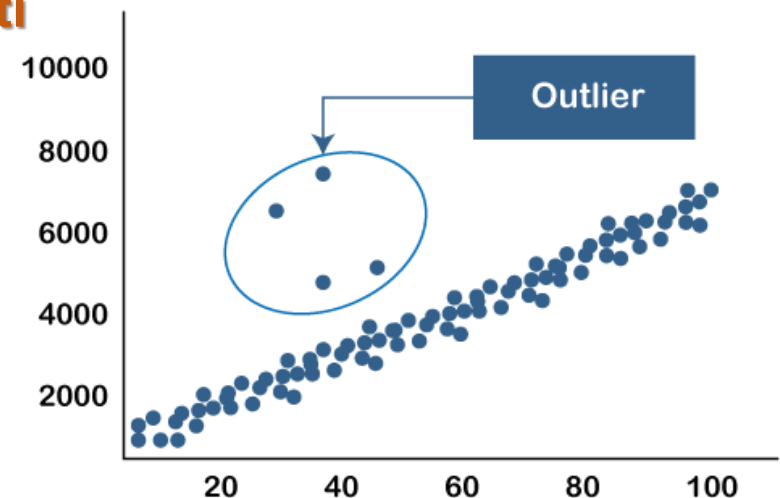
- Normiranje podataka u granicama ± 1
- Normiranje podataka tako da im je sr. vr. = 0, st. dev. = 1

Ekstremne vrijednosti (engl. *Outliers*)

- Ekstremne vrijednosti znatno odstupaju od okolnih podataka.
- Uzrok: kvar mjerne opreme, poteškoće u prijenosu signala, netočna očitavanja,...
- Izolirane ekstremne vrijednosti obično se zamjenjuju interpolacijom susjednih podataka.

Kriteriji otkrivanja ekstremnih vrijednosti

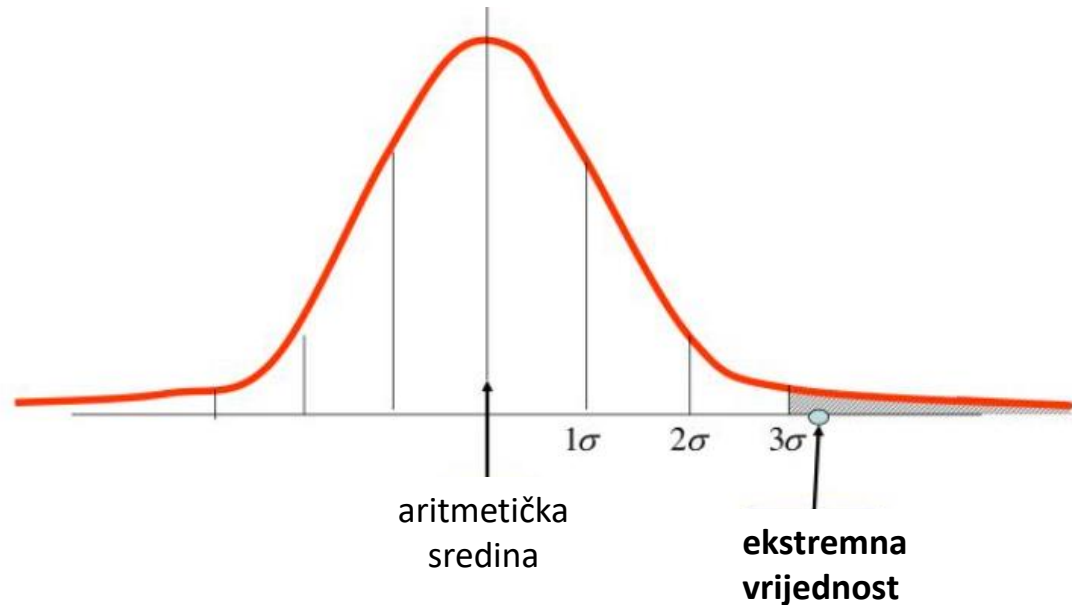
- 3σ pravilo,
- Hampel identifikator,
- Jolliffe metoda (preko PCA/PLS metode),
- Analiza reziduala kod linearne regresije,
- Mahalanobisova udaljenost



Ekstremne vrijednosti

3σ pravilo

$$d_i = \frac{x_i - \bar{x}}{\sigma_x}$$



d_i - normalizirana udaljenost svakog uzorka od srednje vrijednosti

x_i - i -ti uzorak

$\bar{x}$ - aritmetička sredina skupa uzoraka,

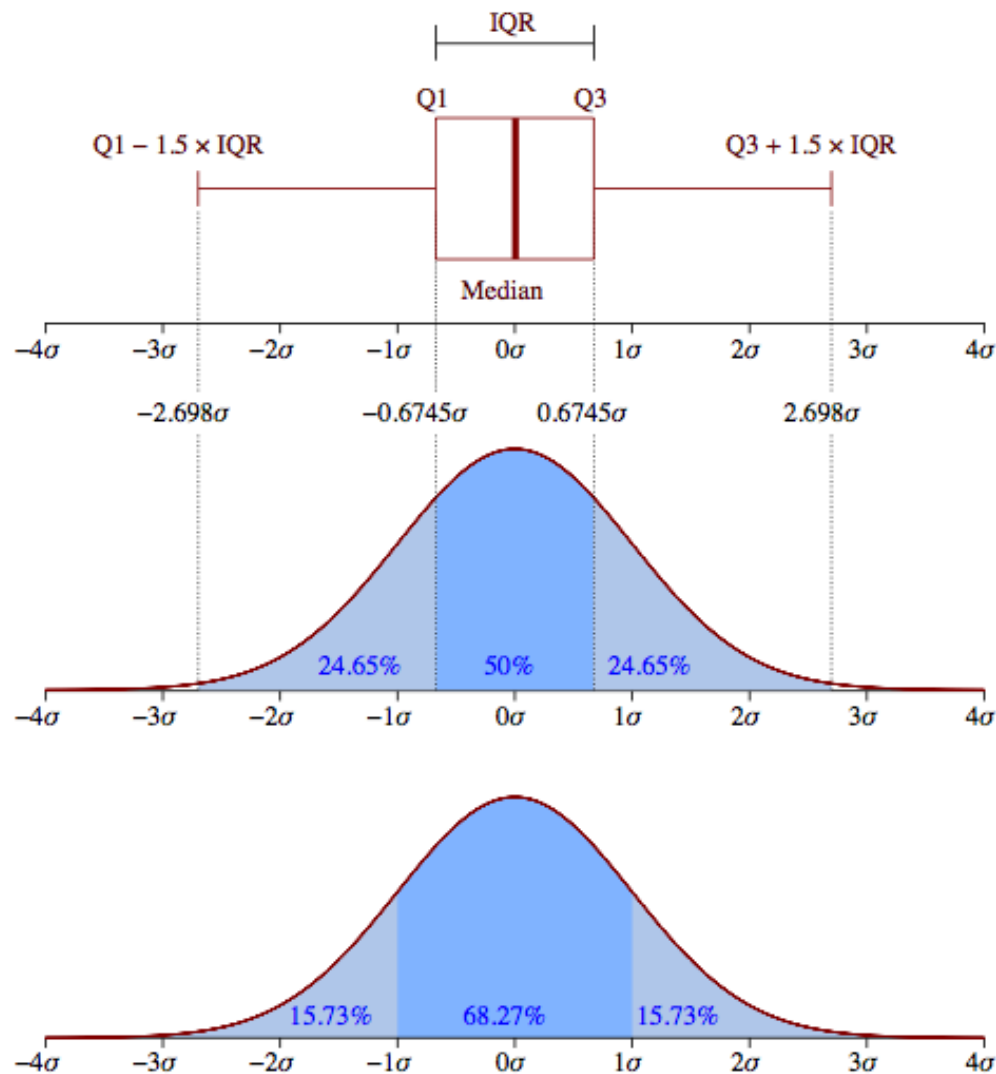
σ_x - standardna devijacija skupa uzoraka

Postoje li ekstremne vrijednosti može se i vizualno utvrditi **Box-plot** dijagramom.

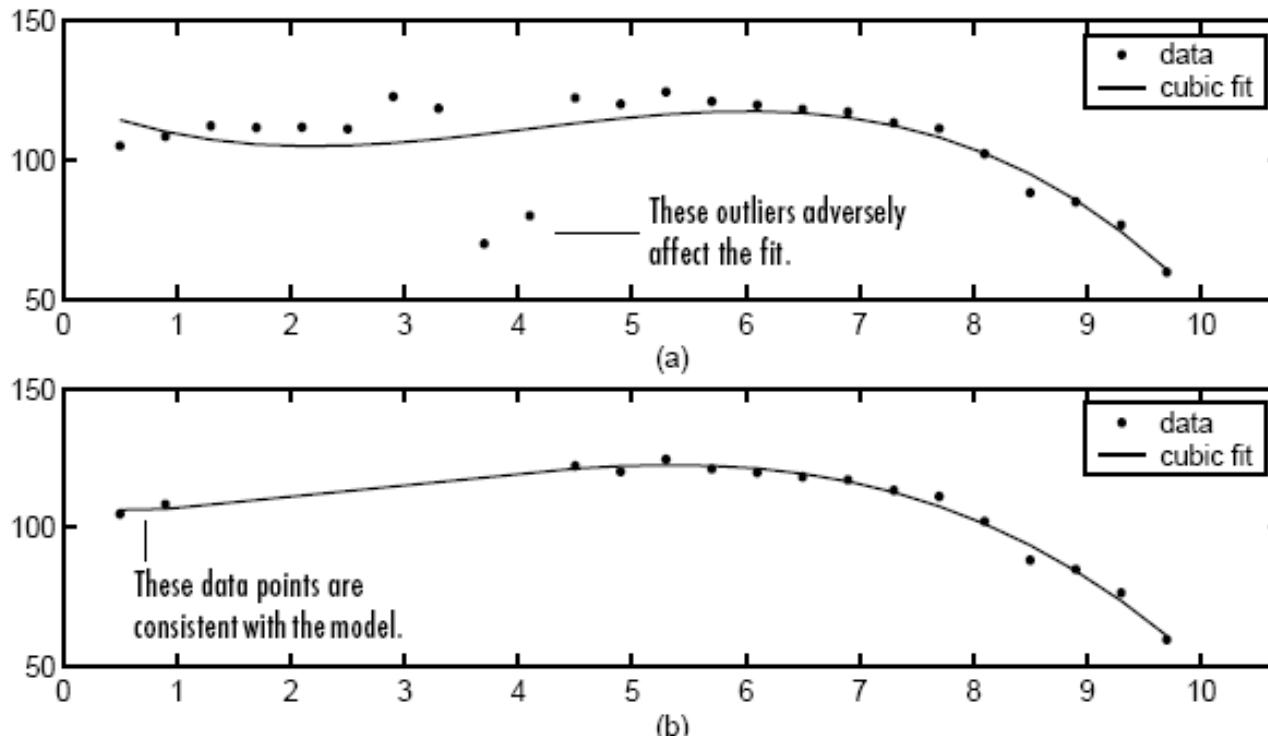
Ekstremne vrijednosti

Kada se želi smanjiti utjecaj višestrukih odstupajućih vrijednosti na procijenjenu srednju vrijednost i standardnu devijaciju varijable, srednja vrijednost se može zamijeniti **medijanom** od podataka, a standardna devijacija s medijanom apsolutnog odstupanja ulaznih podataka od njihovog medijana.

3 σ pravilo s ovakvim robusnijim skaliranjem - **Hampel identifikator**.



Ekstremne vrijednosti – utjecaj na model

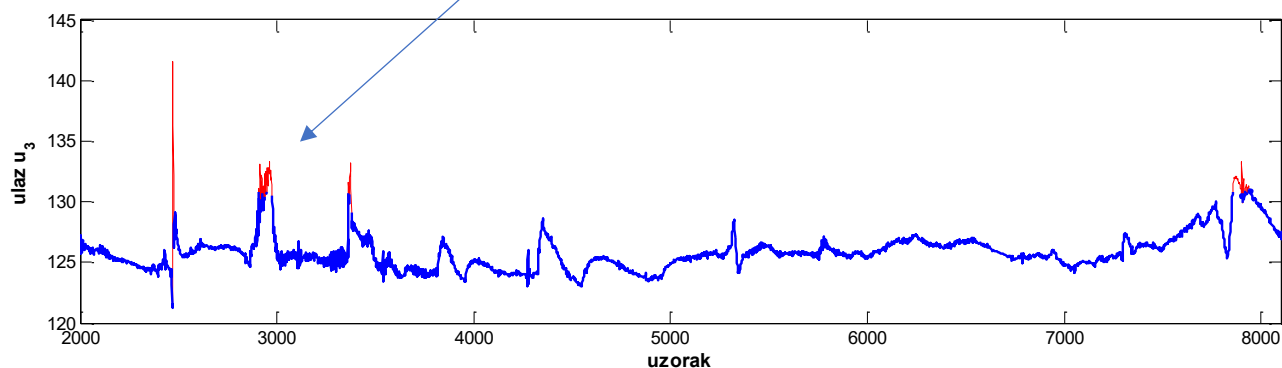


- (a) dva *outlier*-a bitno utječu na razvoj modela
- (b) dva podatka koja su konzistentna s modelom

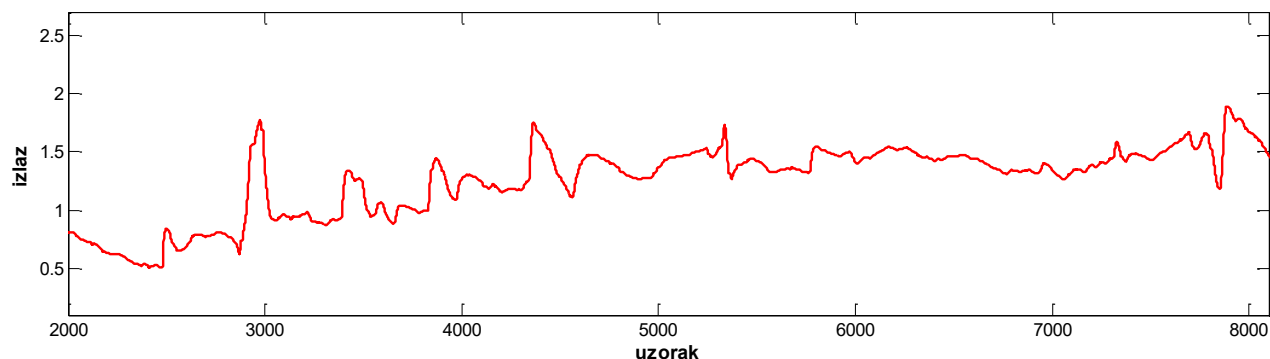
Primjer ekstremnih vrijednosti

Neuobičajeni podaci mogu ponekad predstavljati važnu karakteristiku dinamičkog vladanja procesa te je dobro osim automatskih postupaka podatke temeljito pregledati.

poremećaj u procesu, a ne *outlier*

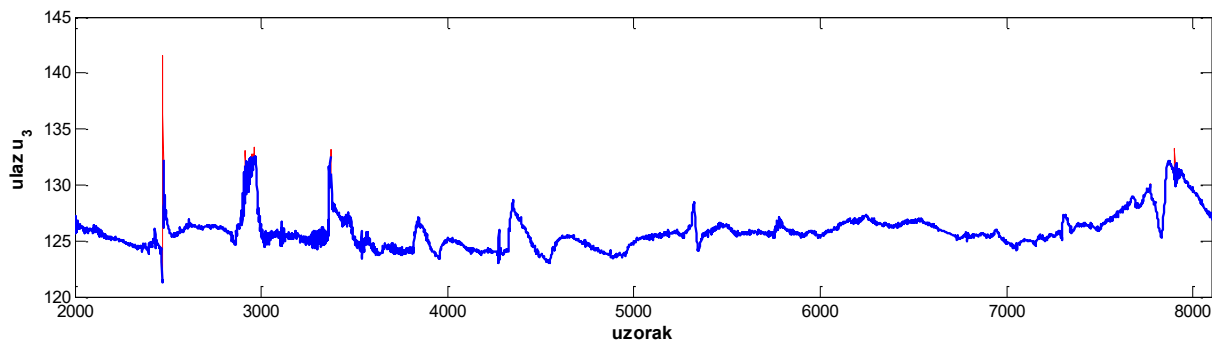


*ekstremne vrijednosti
crvena boja)*

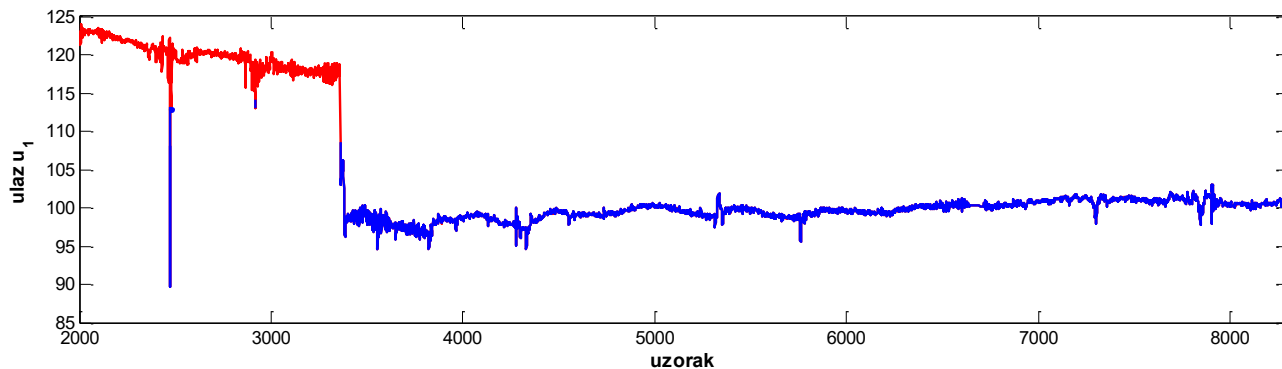


*Prikaz ulaza s otkrivenim ekstremnim vrijednostima (gornji graf)
pomoću **3-sigma** metode te izlaza modela (donji graf)*

Primjer ekstremnih vrijednosti



*Primjer prikaza ulaza s otkrivenim ekstremnim vrijednostima
pomoću **Hampel** identifikatora*



*Primjer prikaza ulaza 1 s otkrivenim ekstremnim vrijednostima koristeći
Hampel identifikator → trebalo bi podijeliti podatke na dva skupa!*

Funkcije u Pythonu za otkrivanje outlier-a



```
from scipy.stats import sigmaclip
```

3 Sigma pravilo

```
sigmaclip(Niz_podataka, l, h) #argumenti: l - broj STD ispod srednje  
vrijednosti, h - broj STD iznad srednje vrijednosti
```

Hampel filter

```
from hampel import hampel  
hampel_obj = hampel(Niz_podataka, window_size, n_sigmas) #argumenti: n_sigmas  
- definira prag detekcije  
Filtrirani_niz = hampel_obj.filtered_data # naredba filtered_data sprema  
filtrirani niz u zadanu varijablu # mogućnosti hampel_obj: filtered_data,  
outlier_indices, medians, median_absolute_deviations, thresholds
```

Napomena: Ovisno o verzijama knjižnica i/ili modula sintaksa može varirati. Provjeriti važeću dokumentaciju.

FILTRIRANJE PODATAKA

- Visokofrekventne te niskofrekventne smetnje trebaju se ukloniti iz podataka, jer mogu rezultirati nestabilnim i nedovoljno točnim modelom.
- **Filtriranje** je postupak kojim se nastoji **smanjiti mjerni šum** koji može nastati kod mjernih pretvornika i tijekom obrade ili prijenosa signala.
- Svrha filtriranja je "izgladiti" podatke te olakšati razvoj modela, tj. onemogućiti modeliranje šuma što rezultira nepotrebnom parametrizacijom modela.
- Filtriranje treba provoditi **s mjerom** jer u obrađenim podacima treba ostaviti dovoljno informacija.

FILTRIRANJE PODATAKA

- **Visokofrekventne smetnje** – **brze i kratkotrajne** promjene u signalu koje se javljaju na visokim frekvencijama. → uzrokovane raznim faktorima: električnim smetnjama, mehaničkim vibracijama, ili kratkim fluktuacijama u mjernim procesima.
- Ove smetnje mogu otežati analizu i modeliranje podataka jer mogu stvoriti "lažne" obrasce u podacima, što može dovesti do pogrešnih zaključaka ili precjenjivanja složenosti modela.
- **Niskofrekventne smetnje** - **polagane, dugotrajne** promjene u signalu koje se javljaju na niskim frekvencijama. → sezonske varijacije ili dugoročni trendovi u podacima.
- Niskofrekventni šum može biti uzrokovan promjenama u vanjskim uvjetima ili dugoročnim trendovima (npr. klimatske promjene, trošenje opreme).
- Mogu uzrokovati pogrešne interpretacije dugoročnih obrazaca (trendova).

Filtriranje podataka – Izgladivanje

- Kod svake metode definira se određeni **raspon** (prozor, **window** ili **span**) – područje susjednih točaka koje se uključuju u proračun nove vrijednosti.
- Ovaj raspon se pomiče duž podataka korak po korak za svaku novu točku;
 - **Veliki raspon** povećava glatkoću podataka, ali **smanjuje rezoluciju** (može izravnati važne lokalne promjene);
 - **Mali raspon** smanjuje glatkost, ali **povećava rezoluciju** podataka (bolje prati lokalne promjene u podacima)
- Optimalni raspon ovisi o skupu podatka, metodi izgladivanja i obično zahtijeva ponešto eksperimentiranja.

Moving average filtriranje

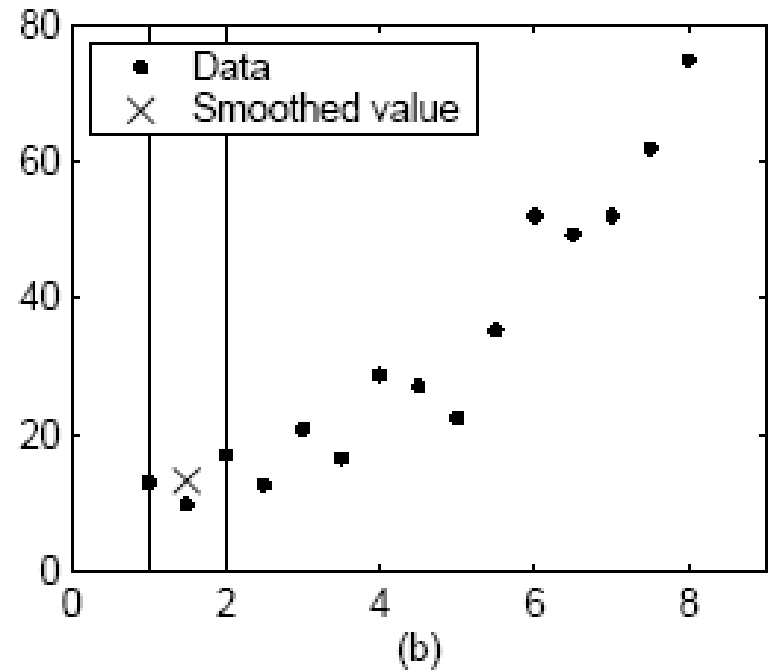
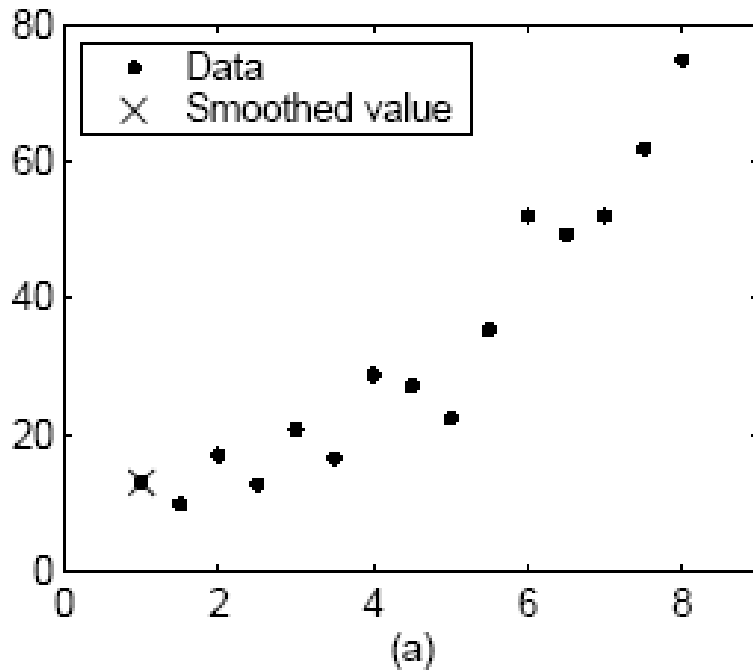
Moving average filtriranje (pokretni prosjek)

- Filtar koji **propušta niske frekvencije** (uklanja visokofrekventni šum);
- Izgladuje podatke tako da **svaki podatak zamjenjuje srednjom vrijednošću susjednih točaka unutar** definiranog raspona (prozora);
- Proračun je definiran s jednadžbom :

$$y_s(i) = \frac{1}{2N+1} (y(i+N) + y(i+N-1) + \dots + y(i-N))$$

$y_s(i)$	izgladena vrijednost i -tog podatka
N	broj susjednih podataka s obje strane
$2N+1$	širina prozora (raspon)

Moving average filtriranje



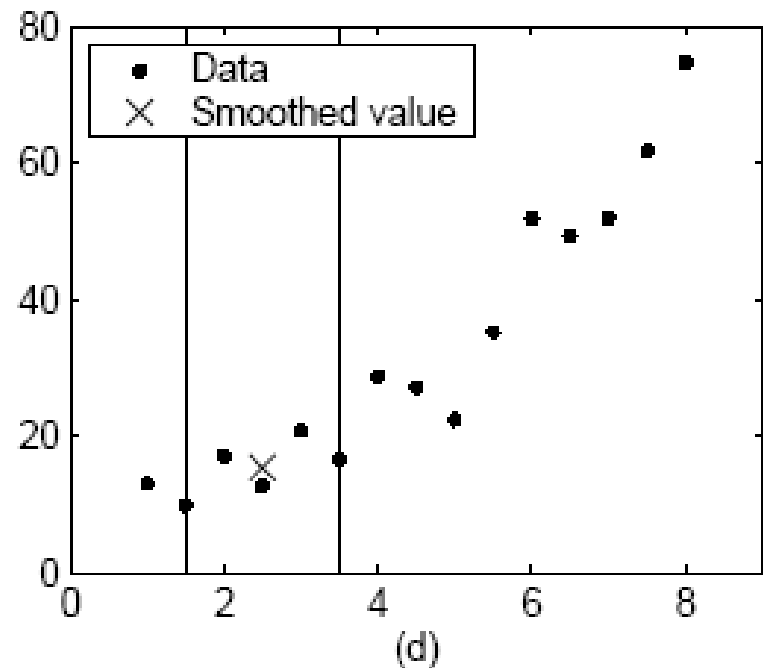
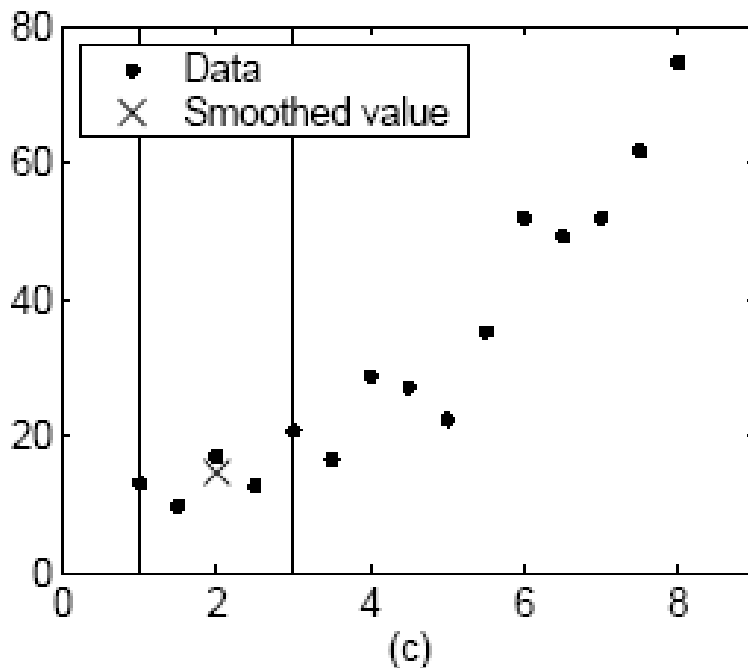
(a) Prva točka nije izgladana jer ne postoji raspon (koristi se samo postojeći podatak)

(b) Druga točka je izgladana primjenom raspona od tri podataka

(a) $y_s(1) = y(1)$

(b) $y_s(2) = (y(1) + y(2) + y(3)) / 3$

Moving average filtriranje (klizni prozor)



(c) i (d) za proračun izgladene vrijednosti primjenjuje se raspon od 5 točaka

$$y_s(3) = (y(1) + y(2) + y(3) + y(4) + y(5)) / 5$$

$$y_s(4) = (y(2) + y(3) + y(4) + y(5) + y(6)) / 5$$

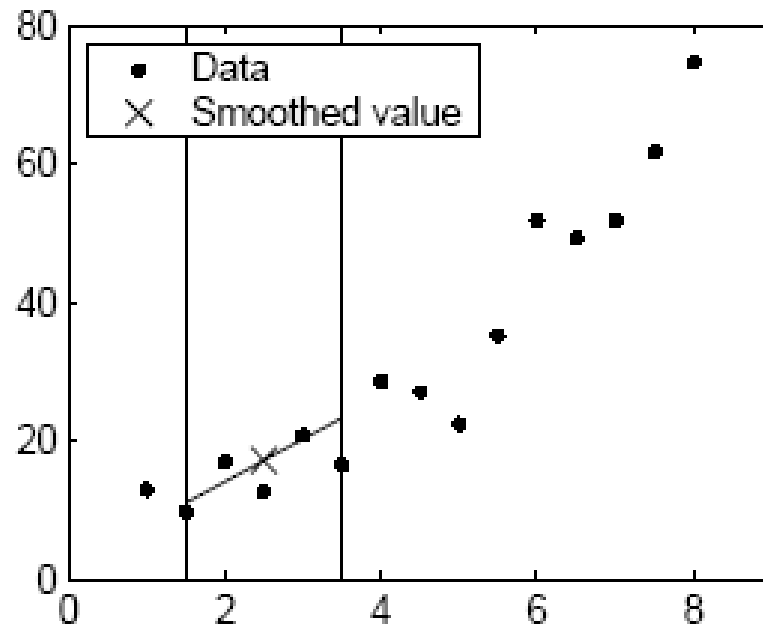
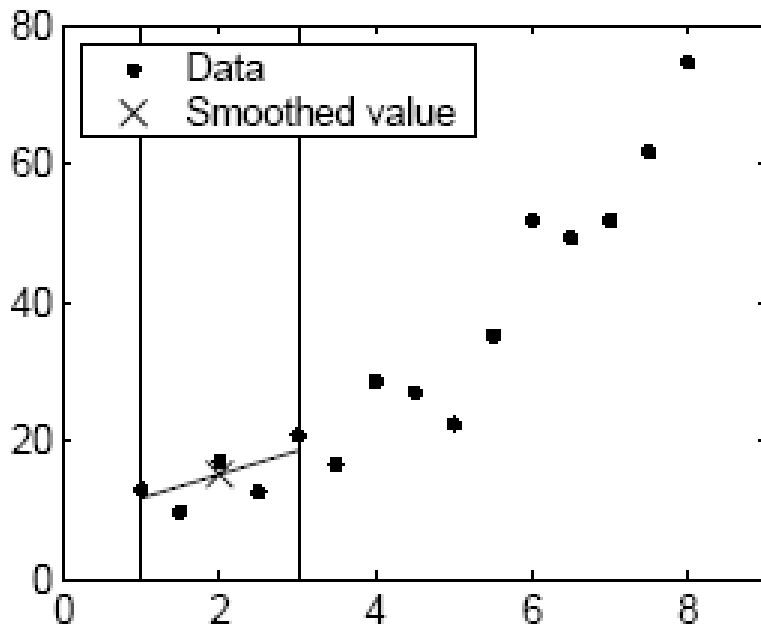
prozor širine **5 točaka** obuhvaća podatke od $y(1)$ do $y(5)$.

LOWESS i LOESS – izgladivanje lokalnom regresijom

Lowess, Loess – “*LOcally WEighted Scatter plot Smooth*”

- Obje metode primjenjuju **lokalnu** linearnu regresiju;
- Smatra se “**lokalnom**” jer se vrijednost svake točke određuje na temelju **susjednih točaka** definiranih rasponom (*span*);
- Podaci imaju svoje težinske vrijednosti. Može se primijeniti i robusna težinska funkcija kako bi se smanjio utjecaj ekstremnih vrijednosti.
- Ovisno o izabranoj “težini” filtriranja, Loess više ili manje zanemaruje podatke koji odskaku od lokalnog trenda te tako nastaju novi podaci koji imaju manje šuma.
- **Lowess** – primjena linearnog polinoma 1. reda
- **Loess** – primjena kvadratnog polinoma 2. reda

LOWESS i LOESS – izgladivanje lokalnom regresijom



- Raspon je stalan, a postupak izgladivanja provodi se od točke do točke.
- Ovisno o broju najbližih susjeda i položaju točke, težinske funkcije mogu i ne moraju biti **simetrične** oko točke oko koje se izgladuje.
- „Težina“ filtriranja omogućava da se odluči koliko podaci s određenim karakteristikama (npr. blizina) utječu na izračun.
- Težinski koeficijenti se dodjeljuju svakom podatku (x_j, y_j) na osnovu njihove udaljenosti od točke x_i → Blježe točke imaju veći utjecaj

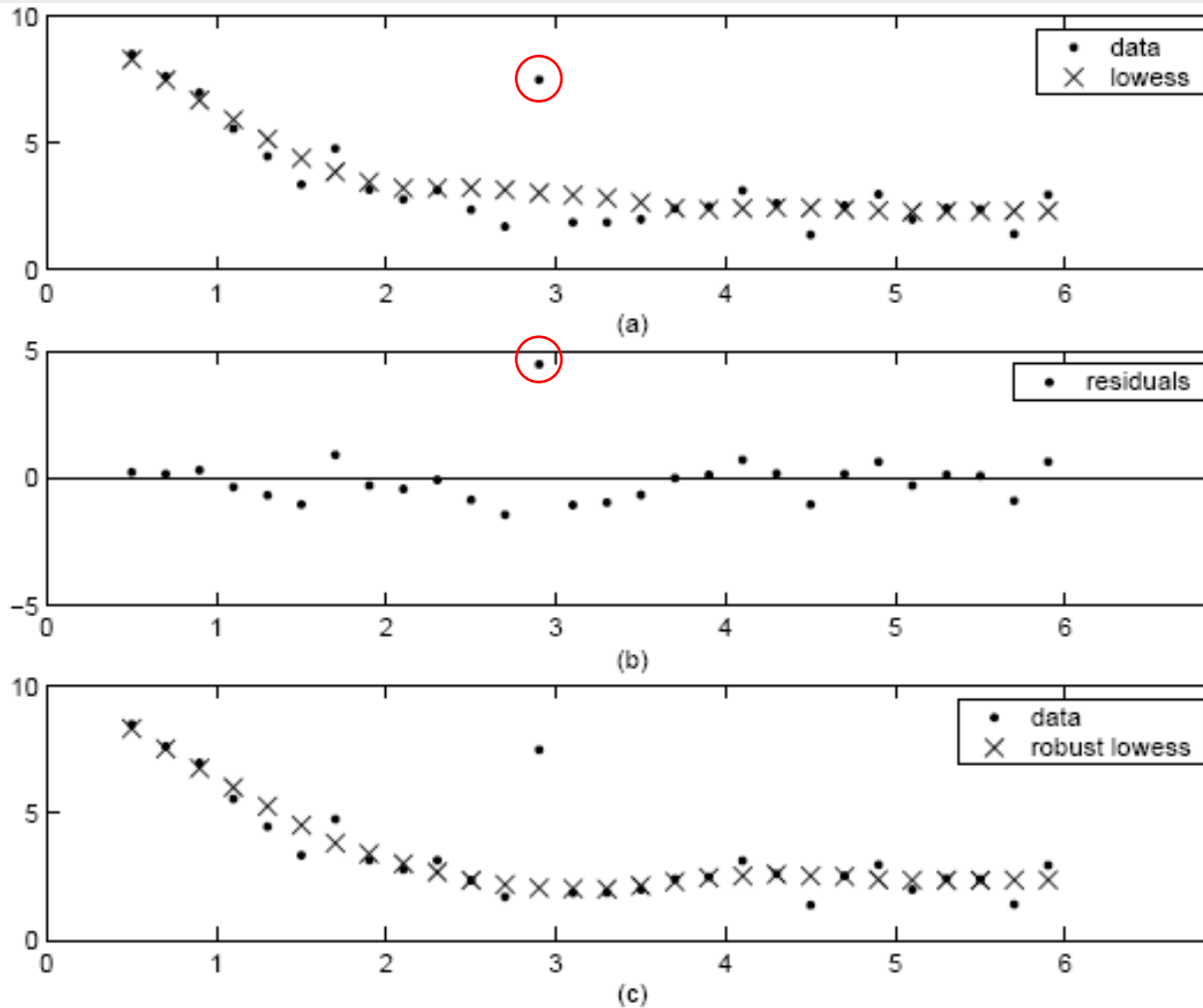
Težinski koeficijent $w(x_j)$

LOWESS

$$\hat{y}_i = \frac{\sum_{j=1}^n w(x_j)(y_j)}{\sum_{j=1}^n w(x_j)}$$

predviđena (izgladena) vrijednost za x_i

LOWESS i LOESS – izgladivanje lokalnom regresijom

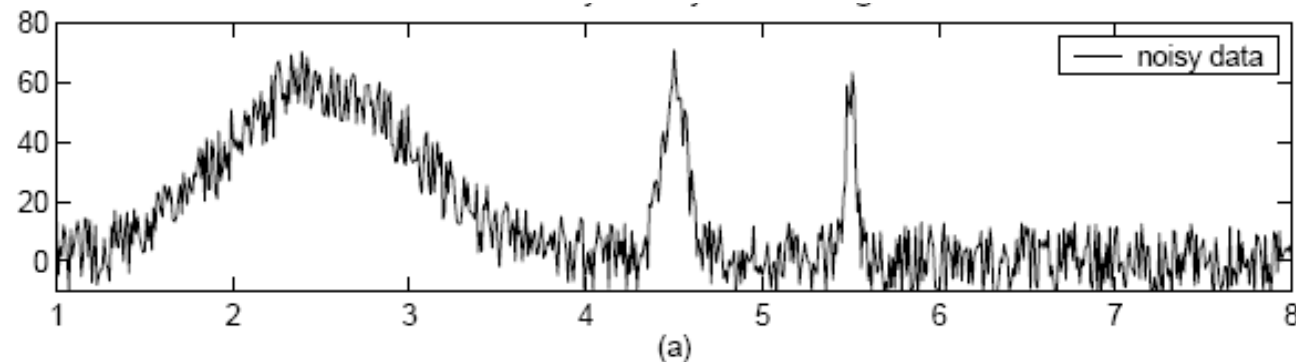


- (a) *Outlier* utječe na izgladenu vrijednost za nekoliko susjednih točaka
- (b) Prikaz ostataka – veći su od 6 medijana apsolutnih odstupanja
- (c) Točke koje jako odstupaju → **manja težina**; Izgladene vrijednosti oko *outlier*-a odražavaju većinu podataka

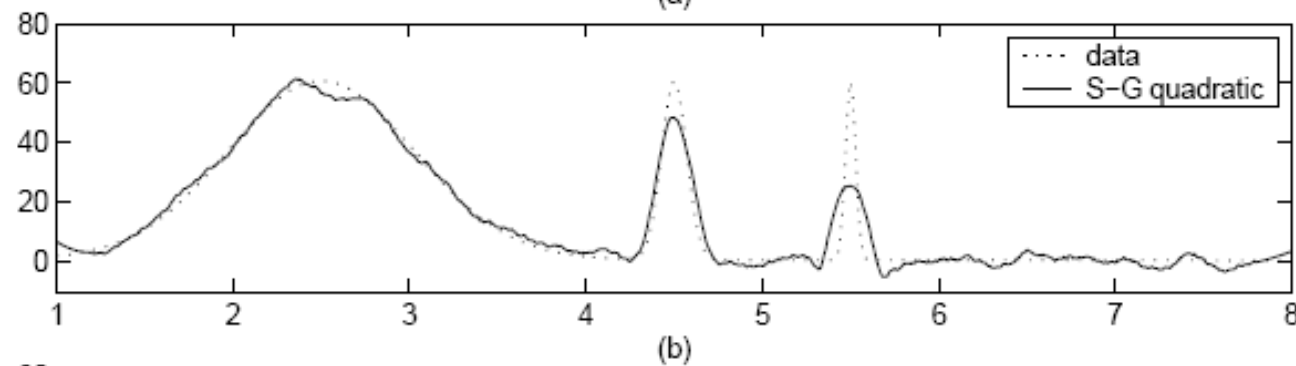
SAVITZKY-GOLAY FILTAR

- Poopćeni “*moving average*” postupak pri čemu se određuje **koeficijent filtra** provedbom netežinskog podešavanja linearnom metodom najmanjih kvadrata primjenom **polinoma određenog stupnja** (*digital smoothing polynomial filter* ili *least squares smoothing filter*)
- Višim redom polinoma može se postignuti visoka razina izgladivanja bez prigušenja podataka
- Često se koristi pri obradi frekvencijskih ili **spektroskopskih** podataka (s pikovima)!
- Kod frekvencijske analize djelotvoran je za očuvanje visokofrekventnih komponenti signala
- Kod spektroskopske analize dobar je za očuvanje vrhova pikova

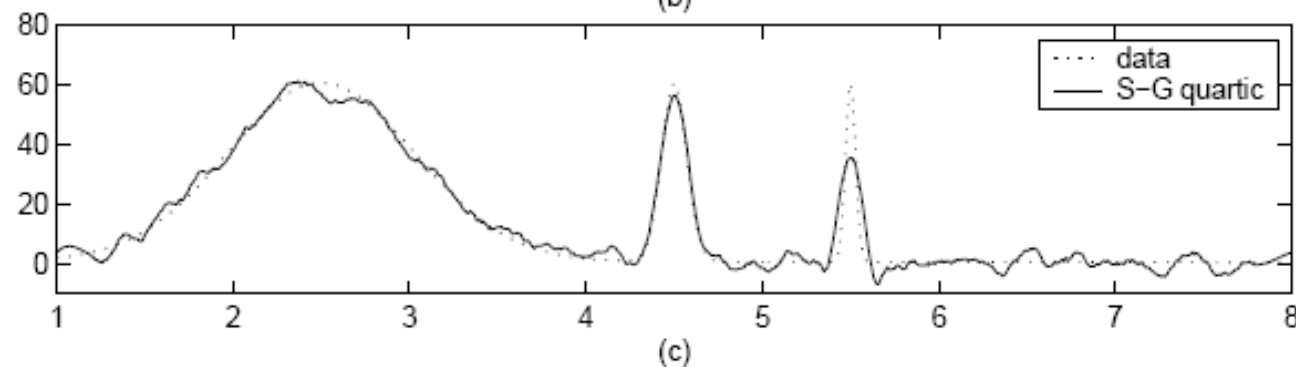
SAVITZKY-GOLAY FILTER



Podaci opterećeni **šumom**



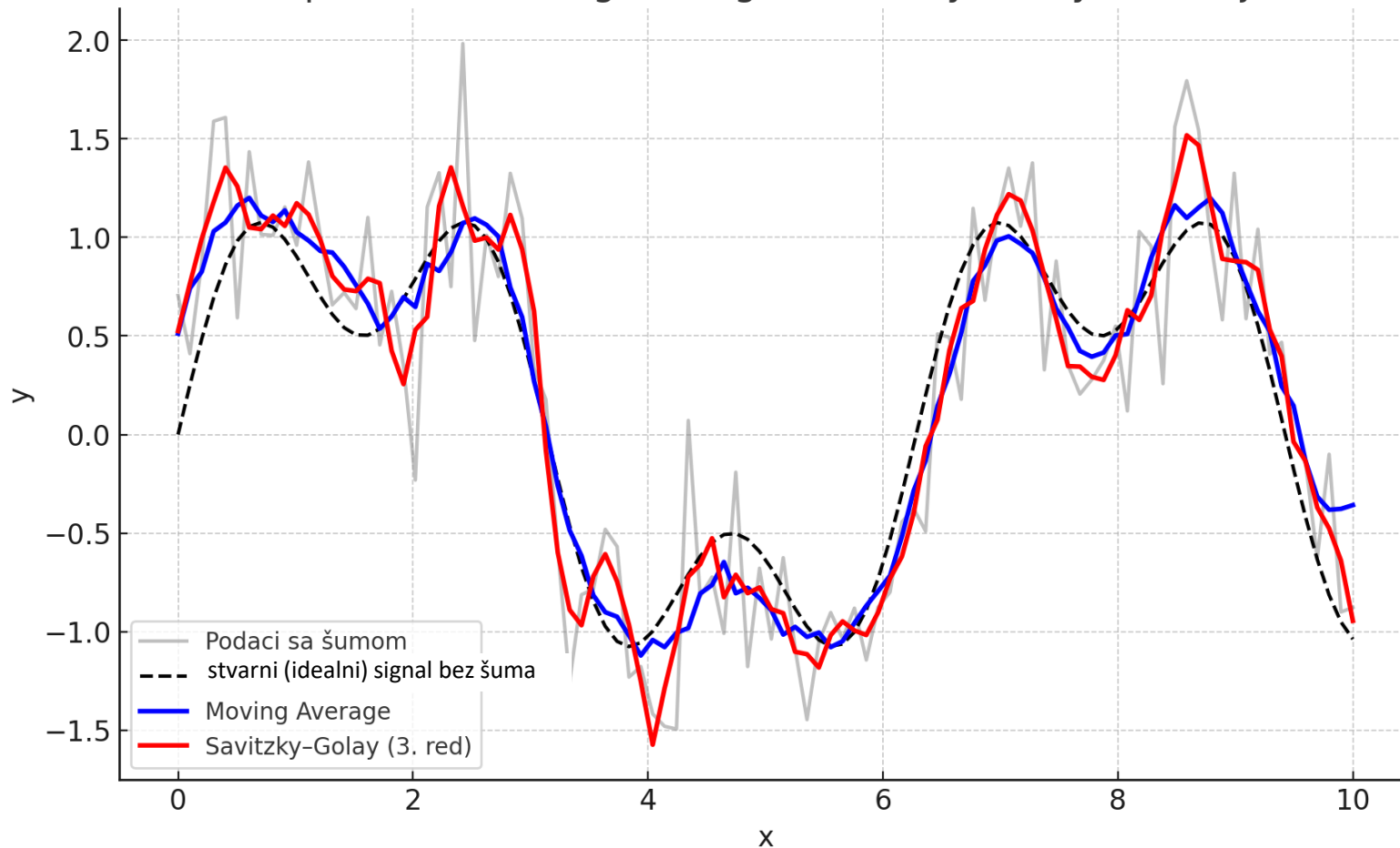
Podaci **bez šuma** –
izgladivanje kvadratnim
polinomom, ima problema
s uskim pikovima



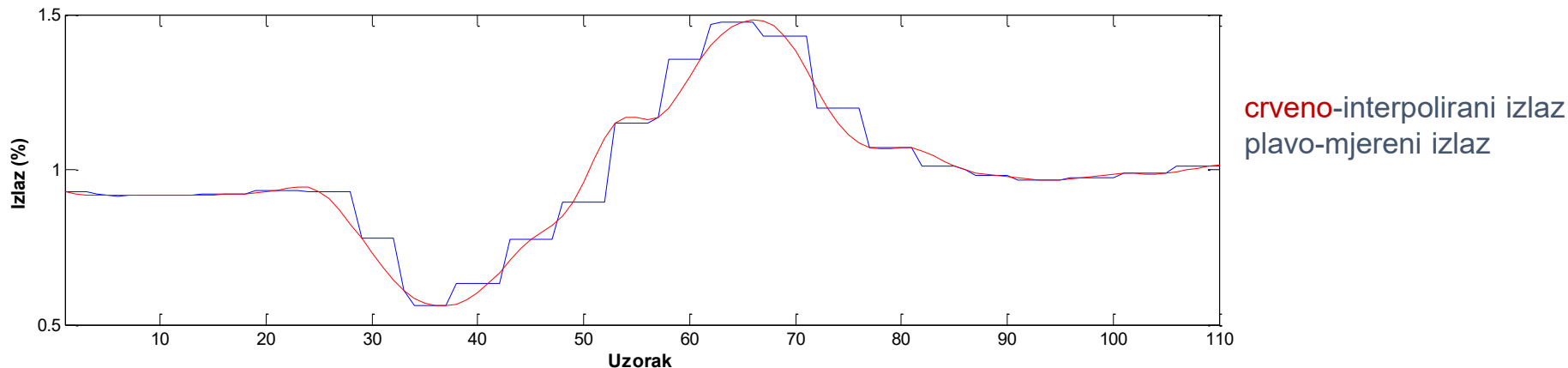
Podaci **bez šuma** –
izgladivanje Savitzky-Golay
polinomom; općenito, što
je veći red bolje se
“hvataju” uski pikovi, ali
slabije širi

Grafička usporedba filtera

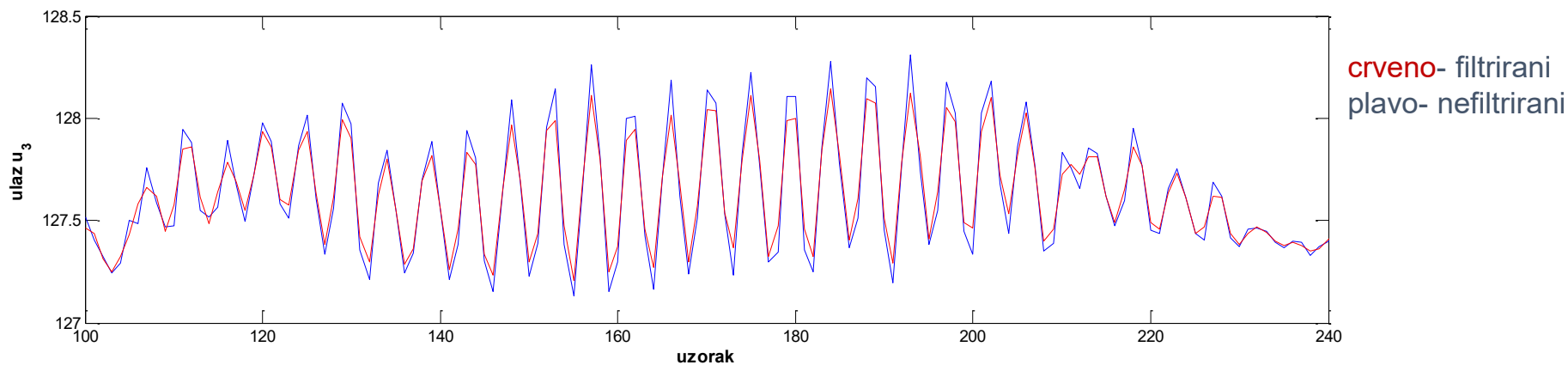
Usporedba: Moving Average i Savitzky-Golay filtriranja



PRIMJERI - Nadomještanje podataka izlaza; Filtriranje podataka



Prikaz interpoliranih vrijednosti dijela izlaznog signala pomoću kubnog spline-a



Uvećani dio usporedbe filtriranih i nefiltriranih podataka

Funkcije u Pythonu za filtriranje

Lowess smoothing filtar

```
from statsmodels.nonparametric.smoothers_lowess import lowess
Zagladeni_Niz = lowess(Niz_podataka, x, frac=0.3) #argumenti: frac - udio
podataka za lokalno zaglađivanje (možemo zamisliti kao svojevrtni window_size)
```

postotak podataka koji se koristi za lokalno izgladivanje. Veći frac = više izgladivanja

Savitzky-Golay filtar

```
from scipy.signal import savgol_filter
filtered_data = savgol_filter(Niz_podataka, window_size, polyorder)
```

broj točaka koje se uzimaju za računanje jedne izgladene vrijednosti (neparan broj)

red polinoma koji se koristi za aproksimaciju



Skaliranje podataka

Podaci mogu imati različite numeričke iznose, ovisno o fizikalnim jedinicama i prirodi procesa. To može povećati značaj brojčano većih veličina nad manjim tijekom postupka razvoja modela. Zbog toga se provodi **skaliranje podataka**.

"Min-max" normalizacija:

$$x_{\text{scaled}} = a + \frac{(x_i - x_{\min})(b - a)}{x_{\max} - x_{\min}}$$

Ova formula pretvara vrijednosti u raspon od **0 do 1** ili neki drugi **-1 do 1**.

x – neskaliрана varijabla
 x_{scaled} – skalirana varijabla
 a - donja granica raspona
 b - donja granica raspona

"Z-score" normalizacija (standardizacija):

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

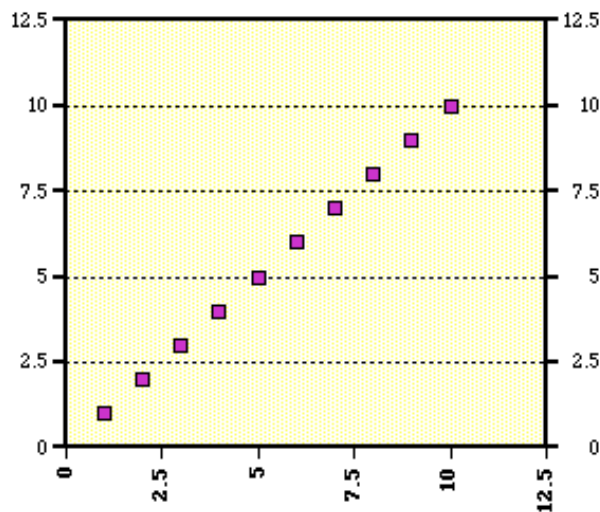
Ova formula normalizira vrijednosti tako da imaju srednju vrijednost 0 i standardnu devijaciju 1.

DIJAGRAM RASPRŠENJA (engl. *Scatterplot*)

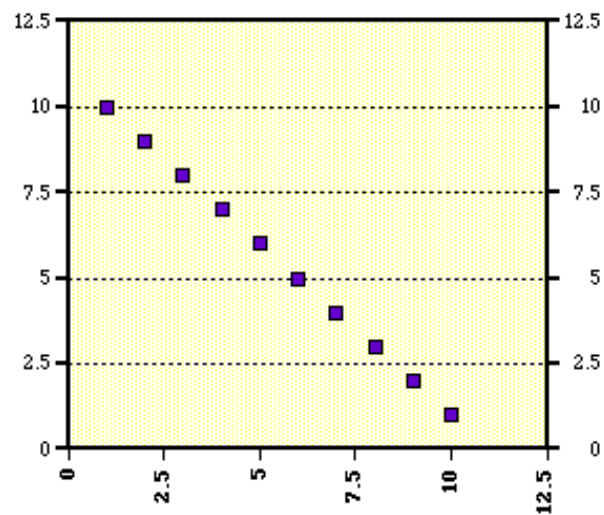
- Često opažamo kako dvije pojave pokazuju međusobnu zavisnost ili određeni stupanj povezanosti;
- Standardizirana mjera jakosti statističke veze između pojava predločenih dvjema kvantitativnim varijablama jest **koeficijent korelacije (R)**.
- **Dijagrami raspršenja** pokazuju koliko jedna varijabla utječe na drugu;
- Što su podaci bliže pravcu to je jača **linearna korelacija** između dvije varijable;
- Ako točke tvore pravac koji ide iz ishodišta do visoke x- i y- vrijednosti, onda varijable imaju **pozitivnu korelaciju**, a ako pravac ide od visoke vrijednosti na y-osi do visoke vrijednosti na x-osi, onda varijable imaju **negativnu korelaciju**.

DIJAGRAM RASPRŠENJA

Idealna pozitivna korelacija ($R=1$)



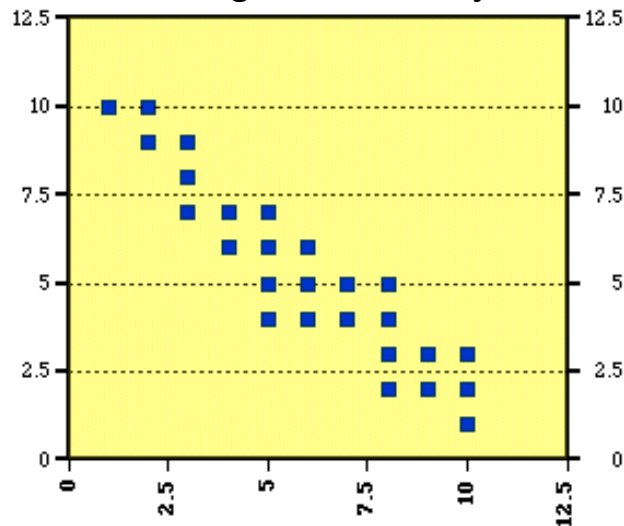
Idealna negativna korelacija ($R=-1$)



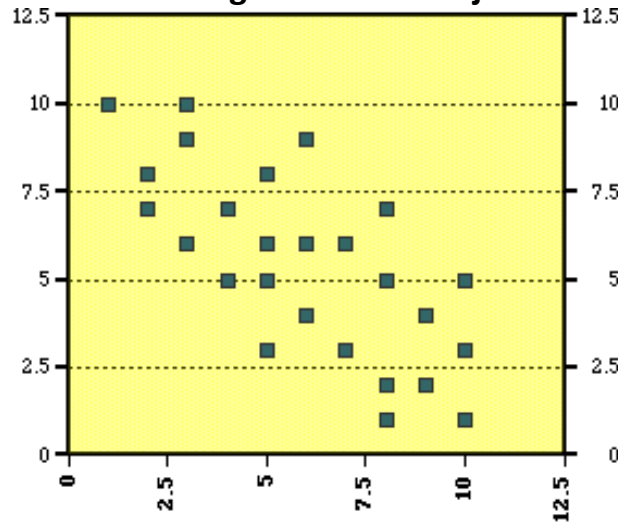
Što je korelacija bliža 1 ili -1, to je ona jača, a što je bliže 0, to je slabija korelacija.

DIJAGRAM RASPRŠENJA

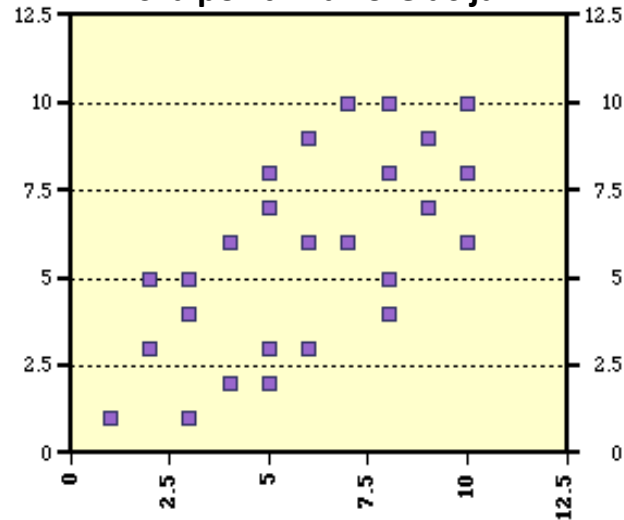
Velika negativna korelacija



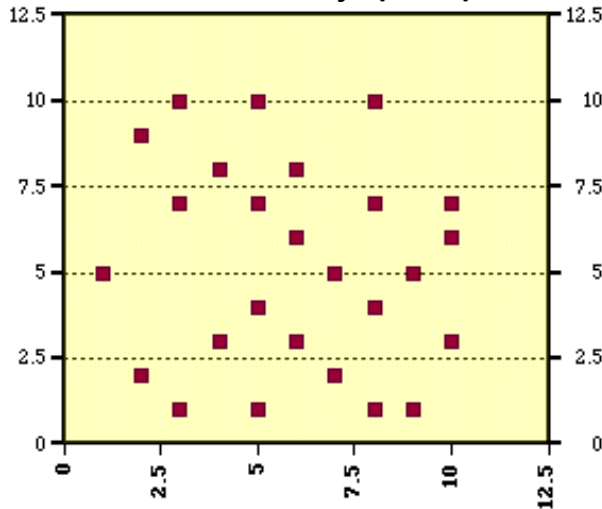
Mala negativna korelacija



Niska pozitivna korelacija



Nema korelacije ($R = 0$)



R od 0 do $\pm 0,2$

nikakva ili neznatna povezanost

R od $\pm 0,2$ do $\pm 0,4$

manja povezanost

R od $\pm 0,4$ do $\pm 0,7$

značajna povezanost

R od $\pm 0,7$ do $\pm 1,0$

visoka ili vrlo visoka povezanost

Regresijska analiza

-statistička metoda određivanja
jednadžbe koja najbolje prikazuje
ovisnost zavisne varijable o
nezavisnoj.

-Jednadžba regresije daje veličinu
promjene izlaznih veličina
uzrokovanih promjenama ulaznih
veličina, pa može služiti za
predviđanje događaja (omogućuje
predviđanje jedne varijable na
osnovu druge).

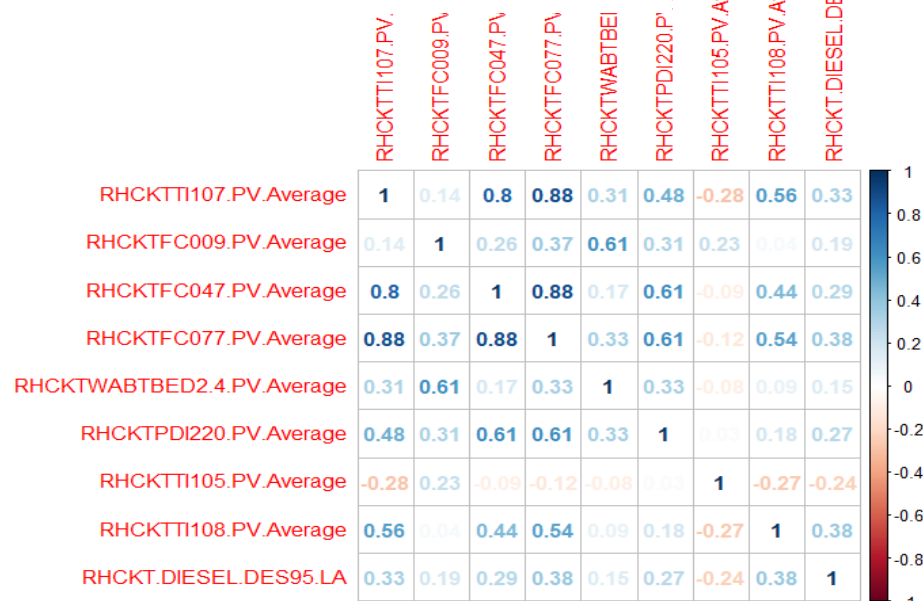
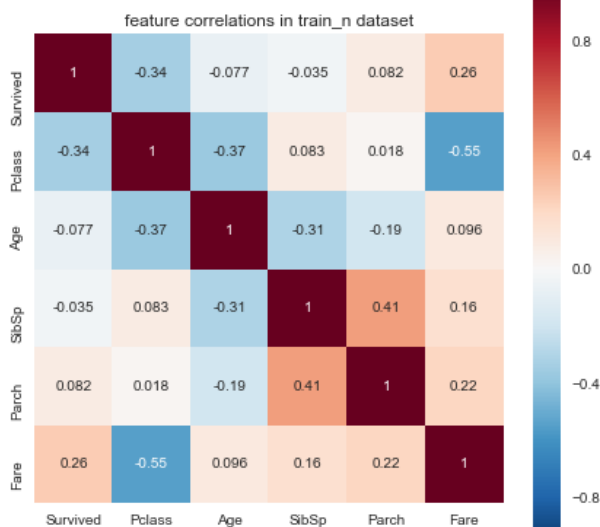
PEARSON-ov KOEFICIJENT KORELACIJE

- Mjera linearne povezanosti dvije normalno distribuirane varijable.
- Ako varijable ne ovise jedna o drugoj, onda je $R = 0$, a ako ovise onda se R nalazi u rasponu od minus 1 do 1.
- Koristi se kod **vrednovanja** modela te kod **odabira ulaznih varijabli** u model - multikolinearnost.
- Izračunava se kao omjer kovarijance između dviju varijabli (y_{exp} i y) i umnoška njihovih standardnih devijacija (omogućava normalizaciju kovarijance, što rezultira korelacijskim koeficijentom koji je u rasponu od -1 do 1).

$$R = \frac{n \left(\sum_{i=1}^n y_{\text{exp},i} \cdot \hat{y}_i \right) - \left(\sum_{i=1}^n y_{\text{exp},i} \right) \cdot \left(\sum_{i=1}^n \hat{y}_i \right)}{\sqrt{\left[n \sum_{i=1}^n y_{\text{exp},i}^2 - \left(\sum_{i=1}^n y_{\text{exp},i} \right)^2 \right] \cdot \left[n \sum_{i=1}^n \hat{y}_i^2 - \left(\sum_{i=1}^n \hat{y}_i \right)^2 \right]}}$$

PEARSON-ov KOEFICIJENT KORELACIJE (Primjeri)

	ulaz 1	ulaz 2	ulaz 3	ulaz 4	ulaz 5	ulaz 6	ulaz 7	ulaz 8	ulaz 9	ulaz 10	izlaz
ulaz 1	1,00	-0,58	0,37	0,77	-0,57	0,59	0,56	0,85	-0,64	-0,46	-0,73
ulaz 2	-	1,00	-0,31	-0,42	0,36	-0,39	-0,10	-0,51	0,38	0,28	0,38
ulaz 3	-	-	1,00	0,31	-0,14	0,40	-0,41	0,33	-0,16	-0,34	0,23
ulaz 4	-	-	-	1,00	-0,60	0,87	0,42	0,73	-0,56	-0,45	-0,71
ulaz 5	-	-	-	-	1,00	-0,69	-0,32	-0,85	0,42	0,49	0,66
ulaz 6	-	-	-	-	-	1,00	0,12	0,67	-0,37	-0,50	-0,53
ulaz 7	-	-	-	-	-	-	1,00	0,44	-0,36	0,09	-0,81
ulaz 8	-	-	-	-	-	-	-	1,00	-0,61	-0,56	-0,73
ulaz 9	-	-	-	-	-	-	-	-	1,00	0,56	0,55
ulaz 10	-	-	-	-	-	-	-	-	-	1,00	0,33
izlaz	-	-	-	-	-	-	-	-	-	-	1,00



SPEARMAN-ov KOEFICIJENT KORELACIJE

Pearsonov koeficijent se koristi za mjerenje linearne korelacije između dviju varijable koje su kontinuirane i normalno distribuirane.

Spearmanov koeficijent korelacije

- koristi se za mjerenje monotonih (ne nužno linearnih) odnosa između dviju varijabli. Monotoni odnos je povezanost između dviju varijabli u kojoj se, s promjenom jedne varijable, druga varijabla sistematski povećava ili smanjuje, bez obzira na to je li ta promjena linearna ili nelinearna. Može se primijeniti na ordinalne, intervalne ili kvotne podatke (kategorijske varijable koje imaju redoslijed).
- Spearmanov koeficijent koristi rangiranje podataka. Otporniji je na utjecaj ekstremnih vrijednosti jer se koristi rangiranje podataka.
- **Spearmanov koeficijent** se koristi kada podaci nisu normalno distribuirani ili kada se želi procijeniti monotona povezanost između varijabli.

$$R_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i : Razlika između rangova za svaki podatak u skupu od n podataka.

Npr., ako je za jedan podatak vrijednost prve varijable rangirana kao 2, a druge kao 5, tada je $d_i = 2 - 5 = -3$

Kriteriji vrednovanja modela

- Ocjenjuju podudarnost vladanja modela s vladanjem stvarnog procesa unutar postavljenih radnih uvjeta na neovisnom skupu podataka
- Pored koeficijenta korelacije R i R_s , u svrhu ocjene valjanosti modela primjenjuju se brojni statistički kriteriji.

$$R^2 = \frac{\sum_{i=1}^n \left(\hat{y}_i - \bar{y} \right)^2}{\sum_{i=1}^n \left(y_i - \bar{y} \right)^2}, \quad 0 \leq R^2 \leq 1$$

Koeficijent višestruke determinacije - R^2

Omjer zbroja kvadrata odstupanja protumačenog modelom i zbroja kvadrata odstupanja eksperimentalnih podataka.

$$\bar{R}^2 = 1 - \frac{n-1}{n-(K+1)} \cdot (1 - R^2), \quad \bar{R}^2 \leq R^2$$

Korigirani koeficijent determinacije

uzima u obzir broj stupnjeva slobode, koji za fiksno n ovisi o broju nezavisnih varijabli (K , prediktora) u modelu.

KRITERIJI VREDNOVANJA MODELA

- Standardni kriteriji:
 - Korijen iz srednje kvadratne pogreške (*Root Mean Square Error*)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

- Srednja apsolutna pogreška (*Mean Absolute Error*)

$$|\bar{e}| = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

y	izmjerene vrijednosti
$\hat{y}$	modelom procijenjena vrijednost
$\bar{y}$	srednja vrijednosti mjerene veličine
n	broj podataka

KRITERIJI VREDNOVANJA MODELA - *FIT*

- Kriterij slaganja modela (*FIT*)**

$$FIT = \left[1 - \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} \right] \cdot 100$$

y	Izmjerene vrijednosti
$\hat{y}$	modelom procijenjena vrijednost
$\bar{y}$	srednja vrijednosti izmjerenih vrijednosti
N	broj podataka

- Vrijednost *FIT* kriterija kreće se u rasponu od 0 do 100%.
0% → **minimalno** slaganje izmjerene vrijednosti i procjene modela.
100% → **savršeno** slaganje mjerenja i procjene modela.

KONAČNA POGREŠKA PREDVIĐANJA

- Funkcija gubitka (***V - Loss Function***)

$$V = \frac{1}{N} \sum_{k=1}^n (y(k) - \hat{y}(k, \Theta))^2$$

S povećanjem broja parametara smanjuje se vrijednost funkcije gubitka. No, to nije nužno slučaj i s pogreškom procijenjenih parametara modela.

Bolji kriterij je **konačna pogreška predviđanja**
(***FPE - Final Prediction Error***)

- modificira funkciju gubitka V tako što **"kažnjava" kompleksnost modela** (izraženu brojem parametara modela d) u odnosu na vrijednost pogreške predikcije

$$FPE = V \left(1 + \frac{2d}{N}\right)$$

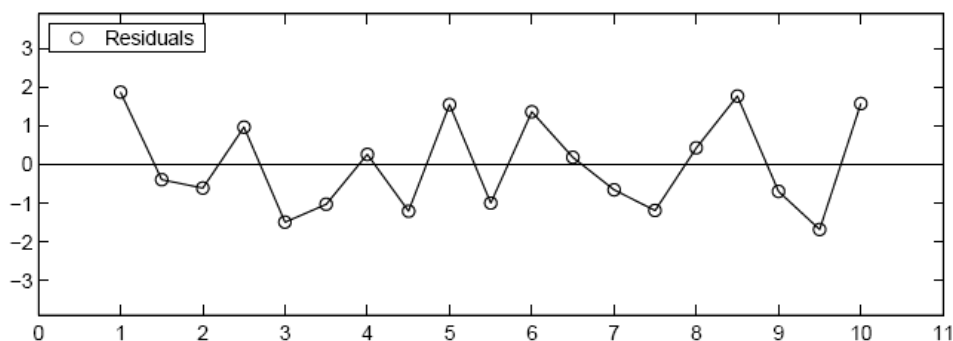
y	mjerena veličina
$\hat{y}$	modelom procijenjena vrijednost
N	broj podataka
Θ	procijenjeni parametri
d	broj procijenjenih parametara

Rezidualna analiza

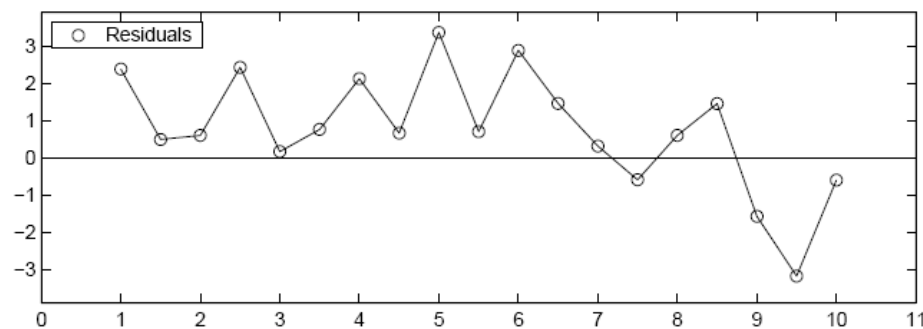
- **Praćenje trenda pogreške (reziduala)** - Reziduali se često grafički prikazuju zajedno s odgovarajućim intervalom pouzdanosti (obično 95%).
- **Histogram pogreške**

$$r = y - \hat{y}$$

Rezidual - Razlika između stvarne vrijednosti i vrijednosti modela.



Reziduali su mali i **slučajno** raspodijeljeni oko nule što znači da model dobro opisuje podatke

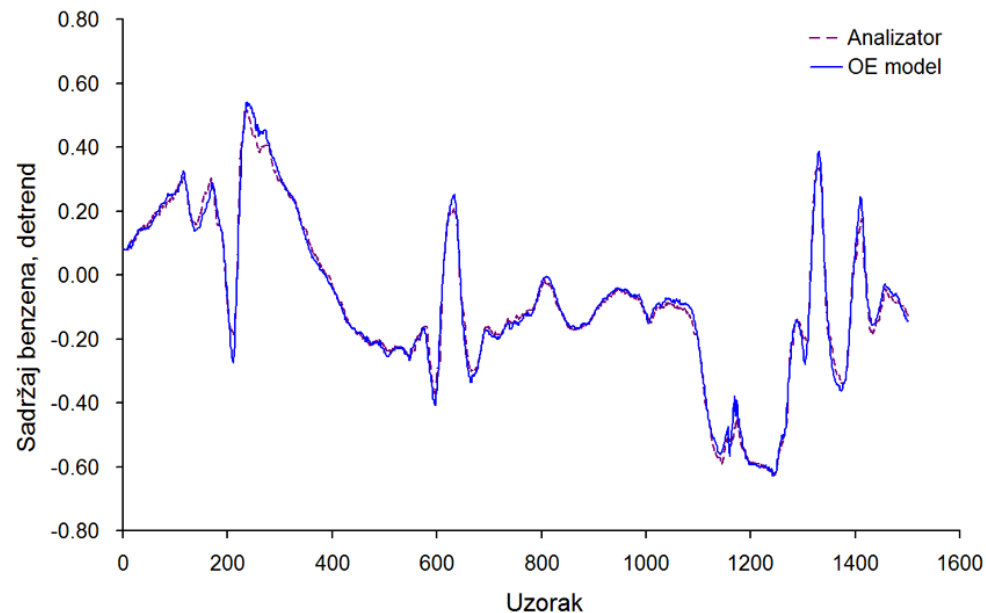


Reziduali su pozitivni za većinu podataka što znači da model ne opisuje dobro podatke (sustavna pogreška)

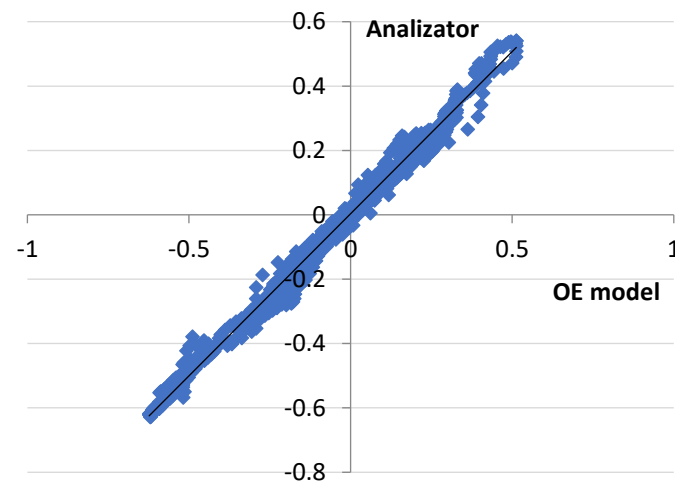
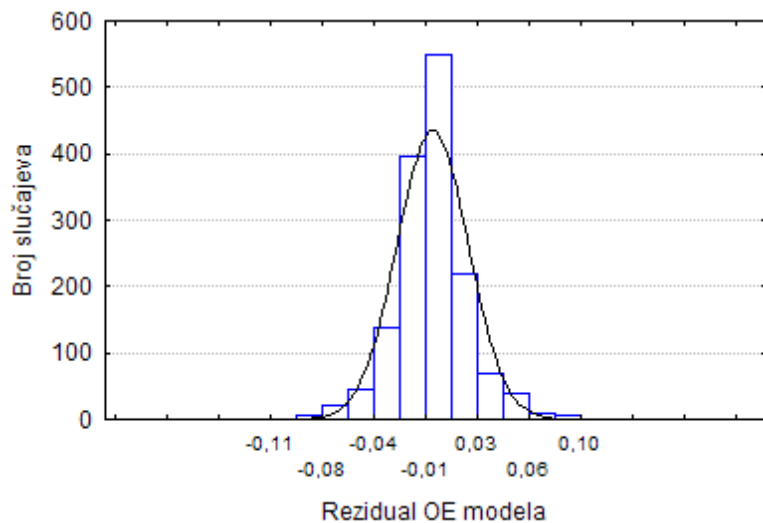
Primjer vrednovanja modela

Sadržaj benzena u
lakom reformatu

Kriterij	
FIT	90,267
FPE	0,0023
RMS	0,0239
e_{MAE}	0,0171



Usporedba sadržaja benzena određenog analizatorom i modelom



Funkcije u Pythonu za vrednovanje modela



Srednja kvadratna pogreška

Srednja apsolutna pogreška

Korijen iz sr. kvadratne pogreške

```
from sklearn.metrics import mean_squared_error, mean_absolute_error
```

```
# Srednja kvadratna pogreška
mse = mean_squared_error(Niz_podataka1, Niz_podataka2)
```

```
mae = mean_absolute_error(Niz_podataka1, Niz_podataka2)
```

```
rmse = mean_squared_error(Niz_podataka1, Niz_podataka2, squared=False)
#argumenti: squared = False omogućuje ispis korjena srednje vrijednosti
```

Skaliranje podataka

Kubni spline

Pearsonov koef. korelacije

R^2

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler
scaler = StandardScaler().fit(Niz_podataka)
skalirani_niz = scaler.transform(Niz_podataka)
```

```
from scipy.interpolate import CubicSpline
CS = CubicSpline(Niz_podataka1, Niz_podataka2)
```

```
Pearsonov_kor_koef = np.corrcoef(Niz_podataka1, Niz_podataka2)
```

```
from scipy import metrics
R_kv = metrics.r2_score(Niz_podataka1, Niz_podataka2)
```

Napomena: Ovisno o verzijama knjižnica i/ili modula sintaksa može varirati. Provjeriti važeću dokumentaciju.

Kriteriji vrednovanja klasifikacijskih modela

Kriterij	Formula
Točnost (engl. <i>Accuracy</i>)	$Acc = (TP + TN) / (TP + FP + FN + TN)$
Preciznost (engl. <i>Precision</i> , <i>P</i>)	$P = TP / (TP + FP)$
Odziv (Osjetljivost, engl. <i>Recall</i> , <i>R</i>)	$R = TP / (TP + FN)$
Specifičnost (engl. <i>Specificity</i>)	$S = TN / (TN + FP)$
Mjera F1 (engl. <i>F1 Score</i>)	$F1 = 2 * (P * R) / (P + R)$

Matrica zabune
(engl. ***Confusion matrix***)

TP (engl. *True Positive*) je broj točno predviđenih pozitivnih primjera

TN (engl. *True Negative*) je broj točno predviđenih negativnih primjera.

FP (engl. *False Positive*) je broj netočno predviđenih pozitivnih primjera (pozitivno predviđenih, ali stvarno negativnih),

FN (engl. *False Negative*) je broj netočno predviđenih negativnih primjera (negativno predviđenih, ali stvarno pozitivnih).

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Kriteriji vrednovanja klasifikacijskih modela

Matrica zabune

(engl. ***Confusion matrix***)

TP - broj točno predviđenih pozitivnih primjera

TN - broj točno predviđenih negativnih primjera.

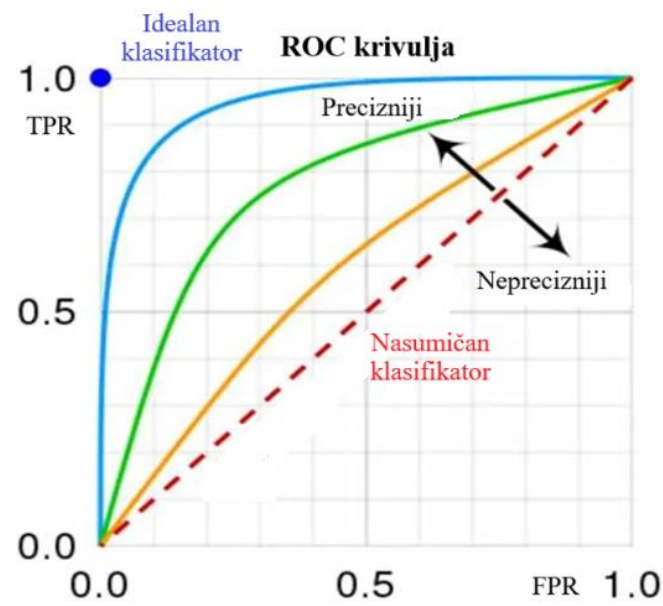
FP - broj netočno predviđenih pozitivnih primjera
(pozitivno predviđenih, ali stvarno negativnih),

FN - broj netočno predviđenih negativnih primjera
(negativno predviđenih, ali stvarno pozitivnih).

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Kriteriji vrednovanja klasifikacijskih modela

- **ROC krivulja** (engl. *Receiver Operating Characteristic curve*) grafički je prikaz performansi klasifikacijskog modela za sve moguće pragove klasifikacije.
- Prikaz kompromisa između osjetljivosti (engl. *True Positive Rate*, TPR) i specifičnosti (odnosno, $1 - \text{False Positive Rate}$, FPR). Na horizontalnoj osi prikazuje se FPR (udio negativnih uzoraka koji su pogrešno klasificirani kao pozitivni), dok se na vertikalnoj osi prikazuje TPR (udio ispravno klasificiranih pozitivnih uzoraka). Svaka točka na ROC krivulji odgovara određenom pragu klasifikacije modela.
- Idealna ROC krivulja brzo se uzdiže prema gornjem lijevom kutu dijagrama, što ukazuje na visoku osjetljivost i nisku stopu lažno pozitivnih rezultata.
- Dijagonalna linija od (0,0) do (1,1) predstavlja performansu nasumičnog klasifikatora, a sve što se nalazi iznad te linije ukazuje na učinkovitost modela bolju od slučajne.



Kriteriji vrednovanja klasifikacijskih modela

- Jedna od najčešće korištenih mjera povezanih s ROC krivuljom je **AUC** (engl. *Area Under the Curve*), odnosno površina ispod ROC krivulje.
- AUC vrijednost može se interpretirati kao vjerojatnost da će model ispravno rangirati nasumično odabrani pozitivni uzorak iznad nasumično odabranog negativnog.
- AUC vrijednost iznosa 1 označava savršen model, dok vrijednost od 0,5 ukazuje na model koji ne razlikuje pozitivne i negativne primjere bolje od slučajnog pogađanja.