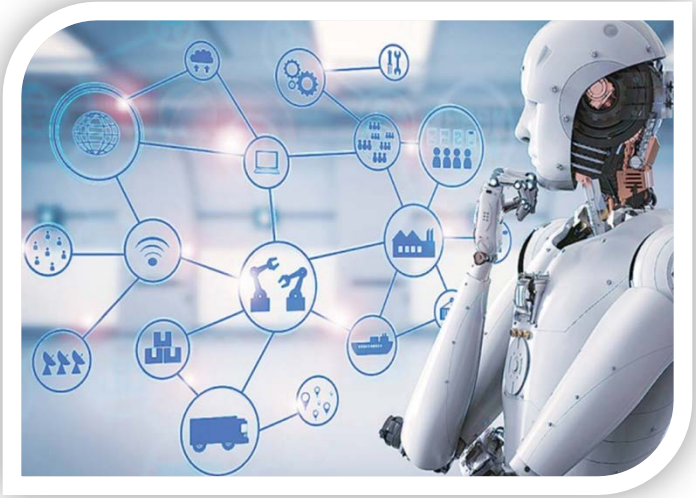
STATISTIČKA ANALIZA I PREDOBRADA PODATAKA. VREDNOVANJE REZULTATA



Željka Ujević Andrijić
zujevic@fkit.unizg.hr

SADRŽAJ

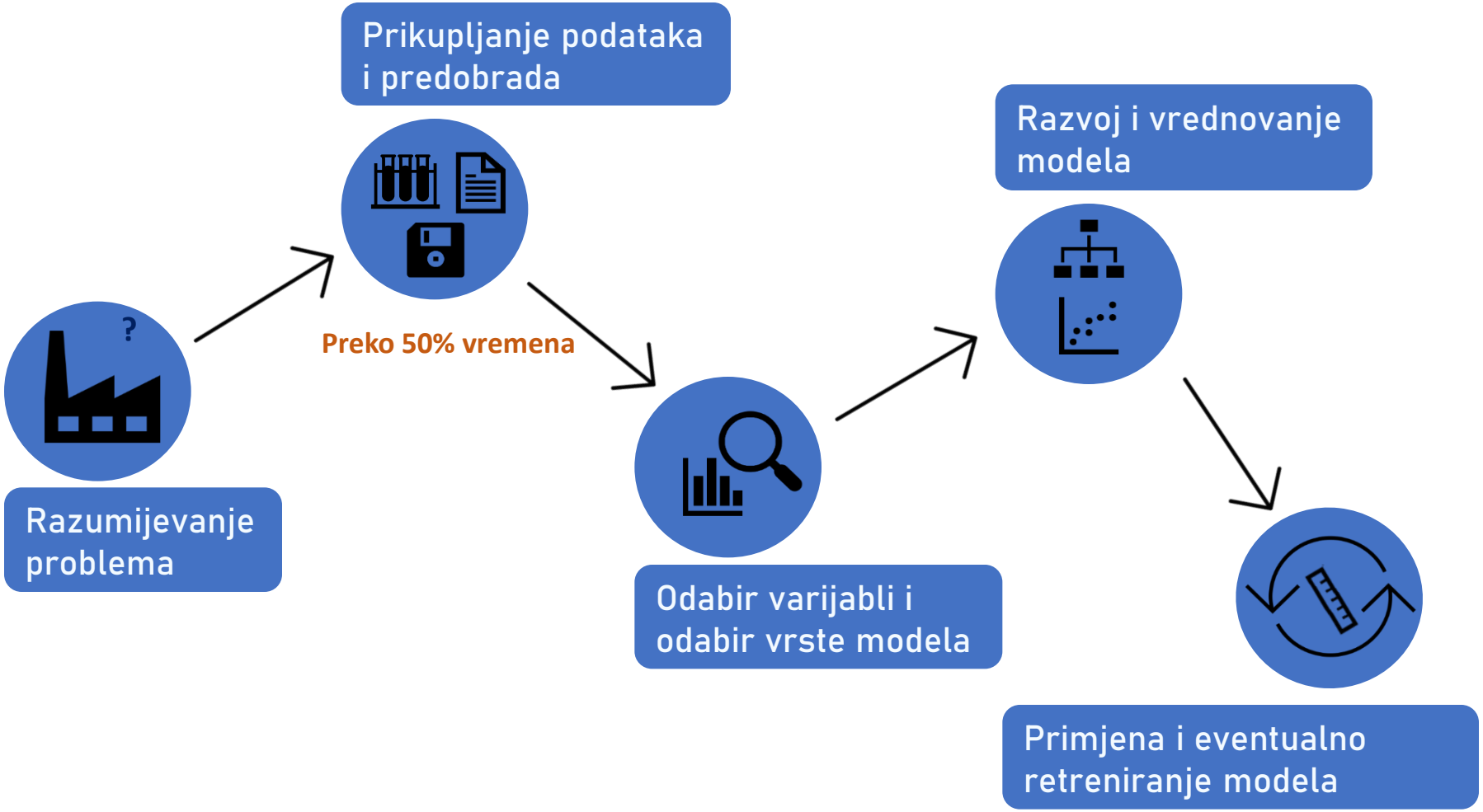
DESKRIPTIVNA STATISTIKA

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

KRITERIJI VREDNOVANJA MODELA

FUNKCIJE U PYTHON-u

POSTUPAK RAZVOJA MODELA STROJNOG UČENJA



"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 **spljoštenosti i asimetrije** distribucije

MJERE CENTRALNE TENDENCIJE – Aritmetička sredina i medijan

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

n – 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

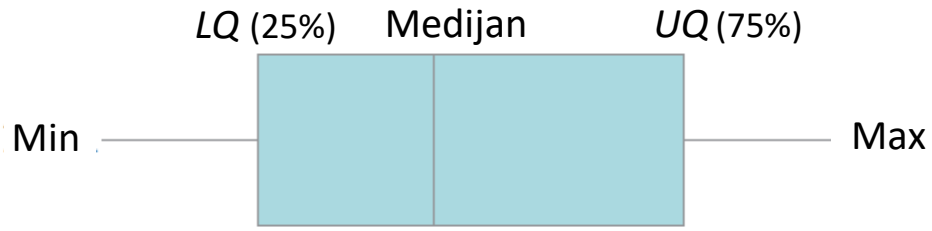
Prije određivanja gornjeg i donjeg kvartila nužno je urediti podatke prema veličini:

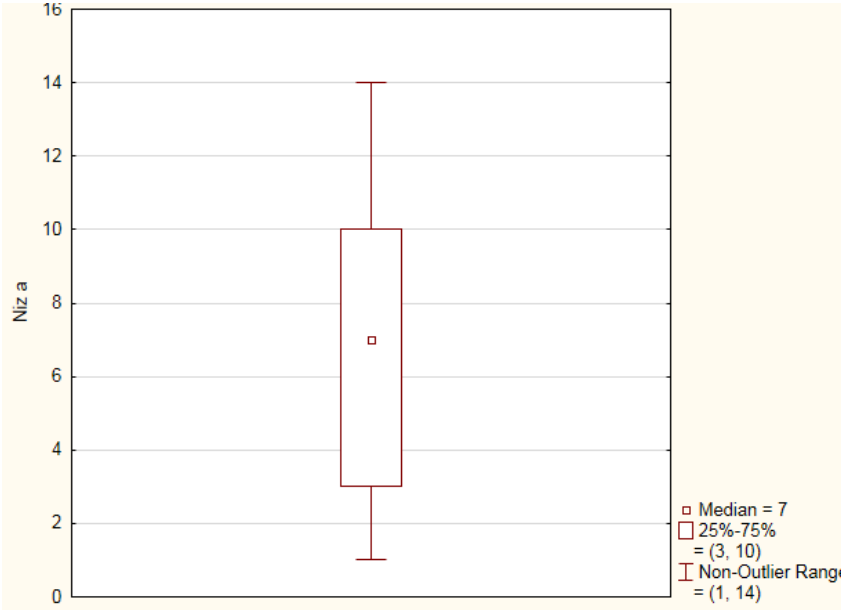
Uređeni niz: 1 2 3 5 7 7 10 12 14

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.

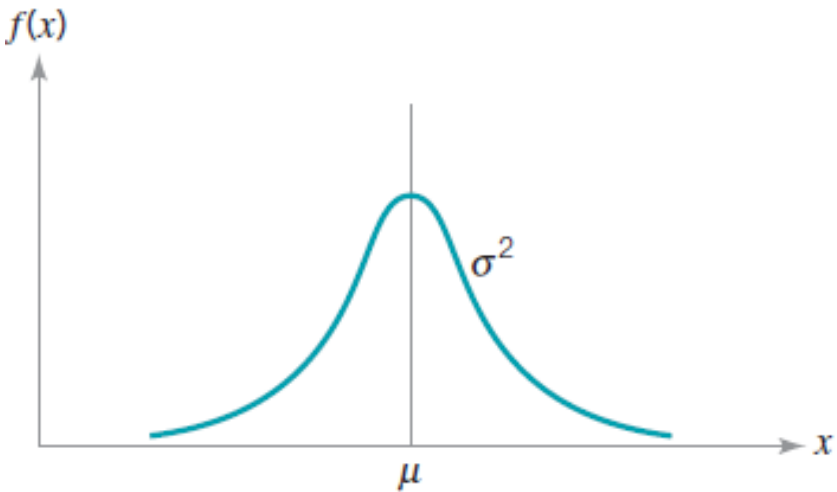
- Služi za **usporedbu varijabilnosti** mjernih rezultata čije su vrijednosti različitog reda veličine.

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

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

NORMALNA RASPODJELA (Gaussova distribucija)

- Statistička distribucija pojavljuje se kod većine pojava u prirodi i tehnici (visina učenika, težina beba, krvni tlak, itd.);
- Mjereći različite pojave i karakteristike 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).

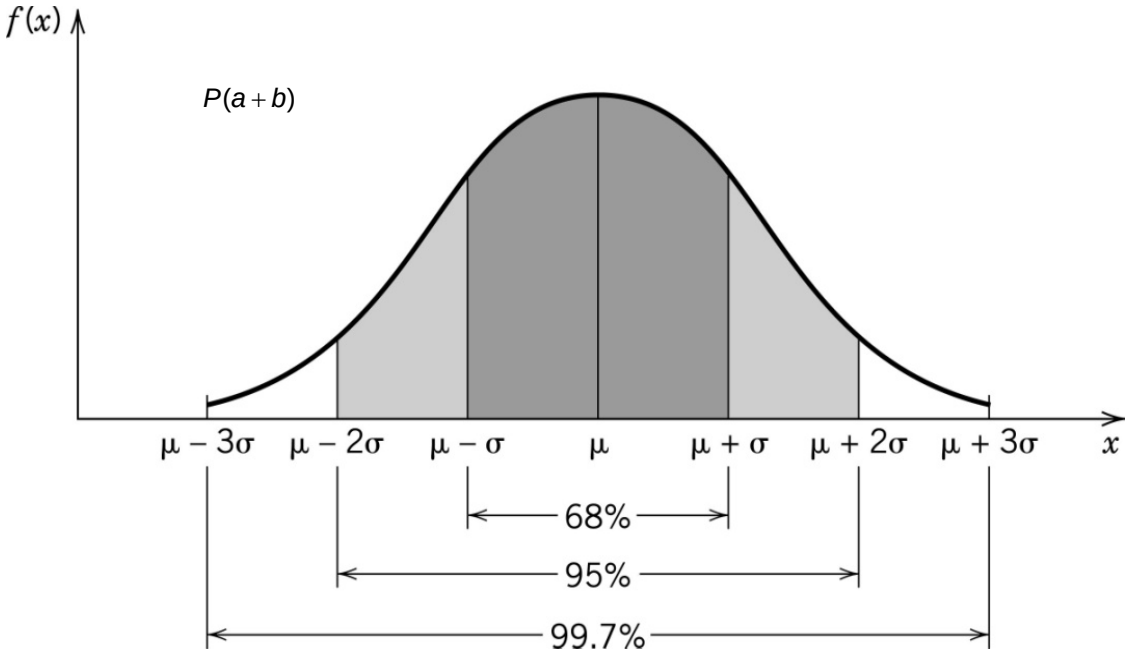


NORMALNA RASPODJELA

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

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

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.

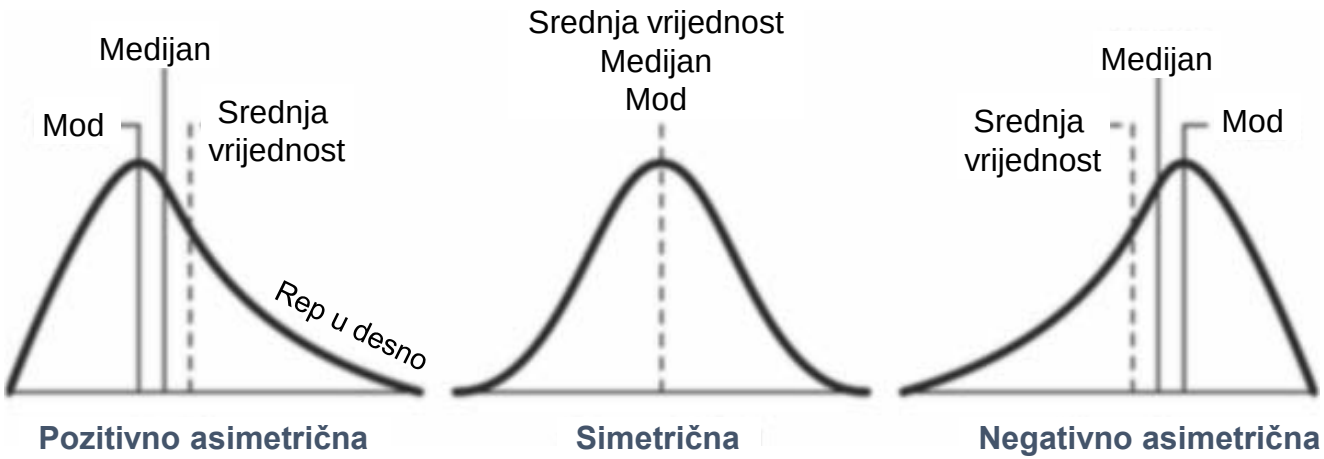
NORMALNA RASPODJELA (simetričnost)

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

Indeks asimetrije:

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

$\alpha = 0$ → u potpunosti simetrična distribucija



NORMALNA RASPODJELA (zakrivljenost)

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

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

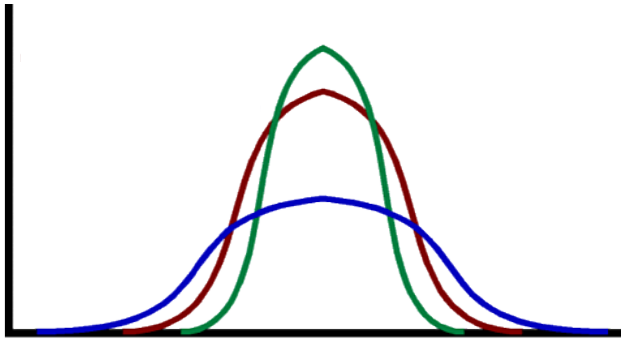
- Pozitivne vrijednosti - uska i visoka distribucija,
- Negativne - spljoštena i široka distribucija.

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

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

n – broj mjerenja

σ^4 - standardna devijacija na 4. potenciju



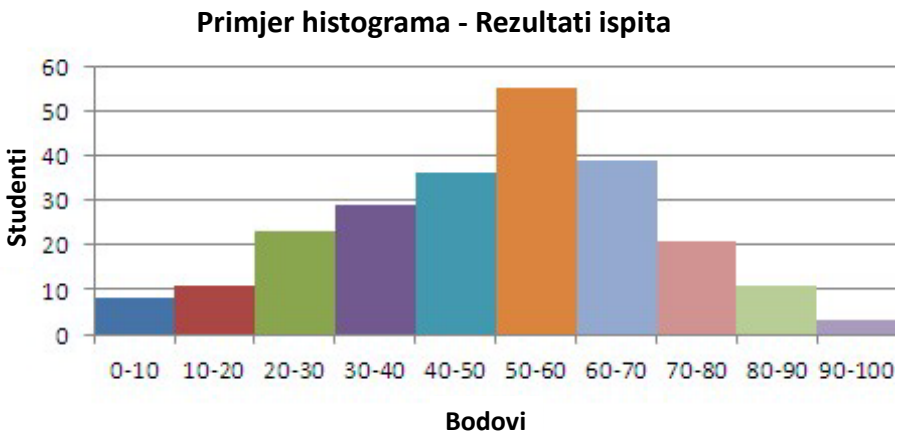
HISTOGRAMI

- 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.
- Ako je duljina intervala na *x*-osi jednaka 1, onda se histogram naziva **relativni prikaz frekvencija** (engl. *relative frequency plot*).

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.



HISTOGRAMI

- Histogram prikazuje broj mjerenja svrstanih u razdvojene kategorije (razrede).
- Ako je n = ukupni broj mjerenja, a k = ukupni broj razreda, histogram m_i se matematički može definirati kao:

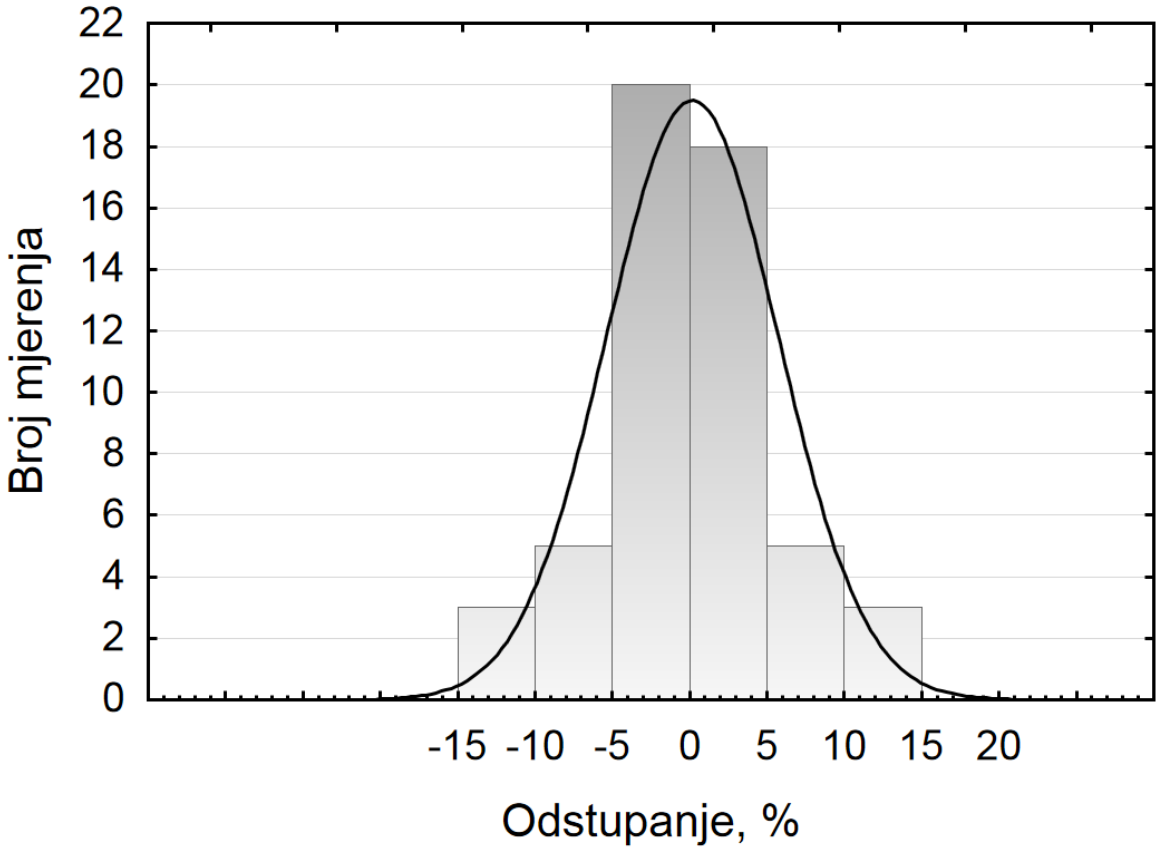
$$n = \sum_{i=1}^{k} m_i$$

- Ne postoji „najbolji” broj razreda.
- U praksi se često **broj razreda** odabire kao kvadratni korijen iz broja mjerenja.
- Ovisno o podacima distribucije i ciljevima analize uglavnom je potrebno eksperimentiranje za određivanje prikladne širine razreda h.
- Broj razreda k može se dodijeliti i direktno ili se može izračunati iz preporučene širine razreda h:

$$k = \frac{\max x - \min x}{h}$$

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

PREDOBRADA PODATAKA

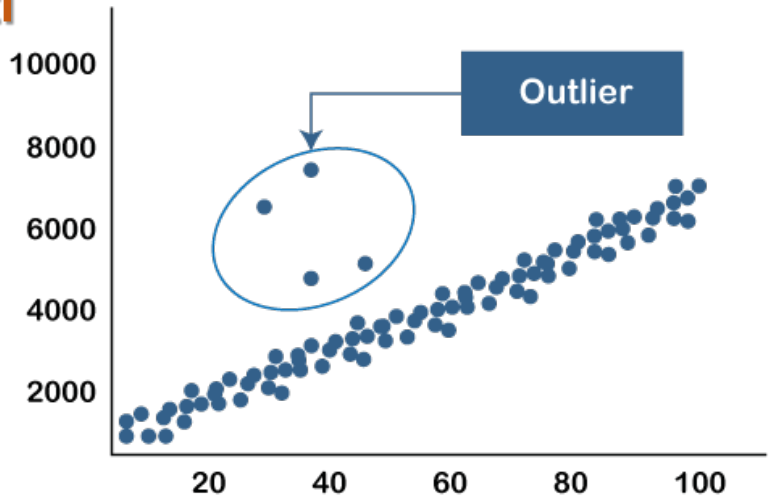
Prikupljanje podataka iz baze podataka postrojenja ili laboratorija.	• **PHD (*Process history database*), *laboratorij***
Određivanje vremena uzorkovanja podataka	• *Resempliranje* (kako bi se smanjio broj uzoraka, a zadržala informativnost signala).
Otkrivanje i uklanjanje ekstremnih vrijednosti (engl. *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	• MAR *Spline* algoritam • *Spline* (kubni)
Odabir utjecajnih varijabli	• Koreliranost varijabli; PCA, PLS,... • Konzultacije s tehnolozima/operaterima
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 očitanja,…
- Izolirane ekstremne vrijednosti obično se zamjenjuju interpolacijom susjednih podataka.

Kriteriji otkrivanja ekstremnih vrijednosti

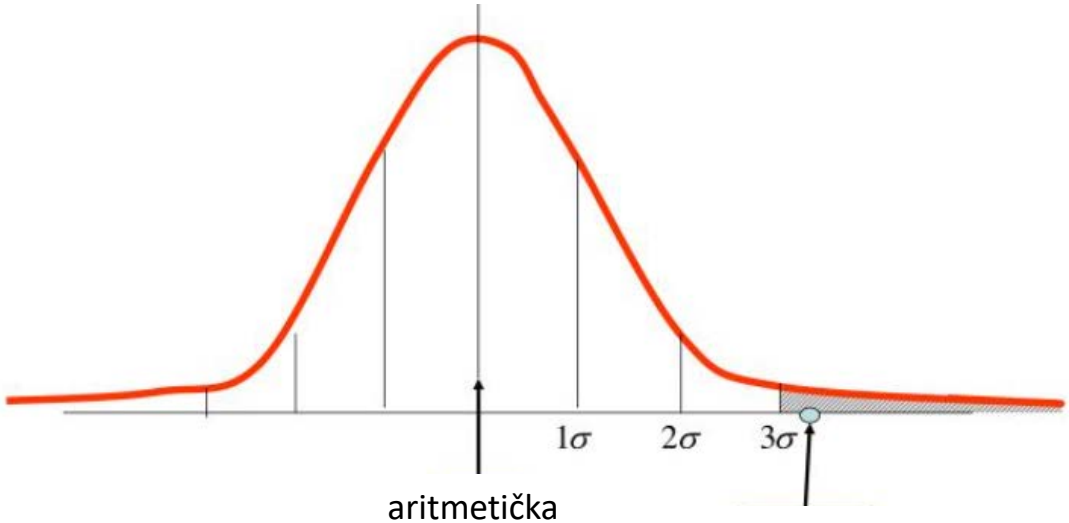
- 3σ pravilo,
- Hampel identifikator,
- Jolliffe metoda (preko PCA/PLS metode),
- Analiza reziduala kod linearne regresije,
- Mahalanobisova udaljenost



Ekstremne vrijednosti

3σ pravilo

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



d_i - normalizirana udaljenost svakog uzorka od srednje vrijednosti

x_i - *i*-ti uzorak

x - aritmetička sredina skupa uzoraka,

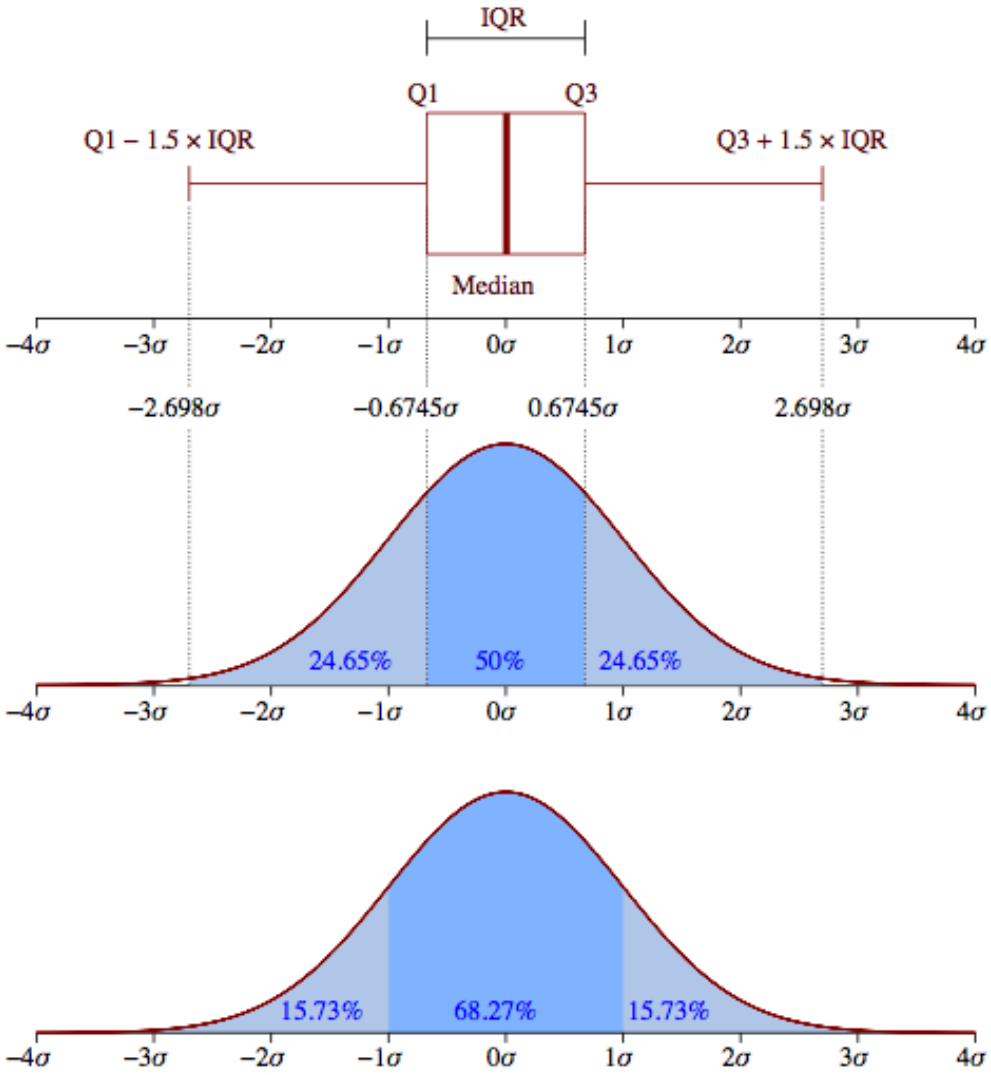
σ_x – standardna devijacija skupa uzoraka

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

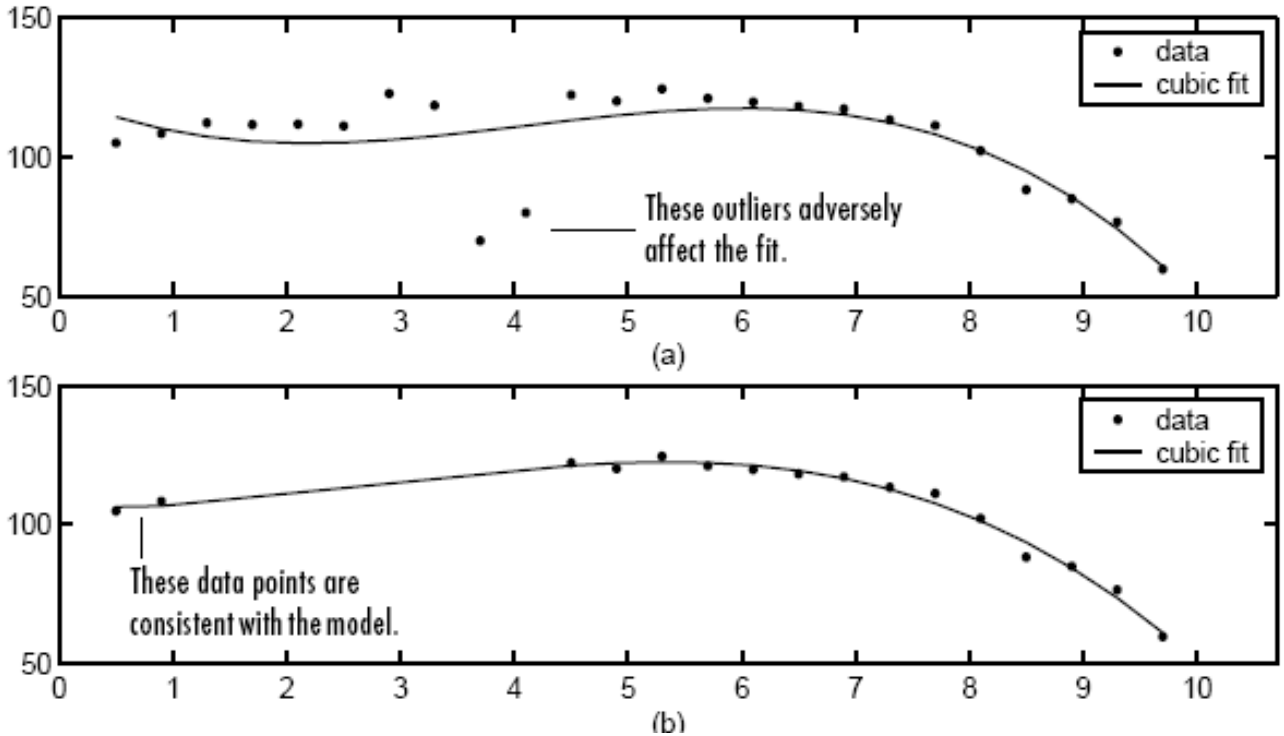
Ekstremne vrijednosti

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.

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



Ekstremne vrijednosti – utjecaj na model



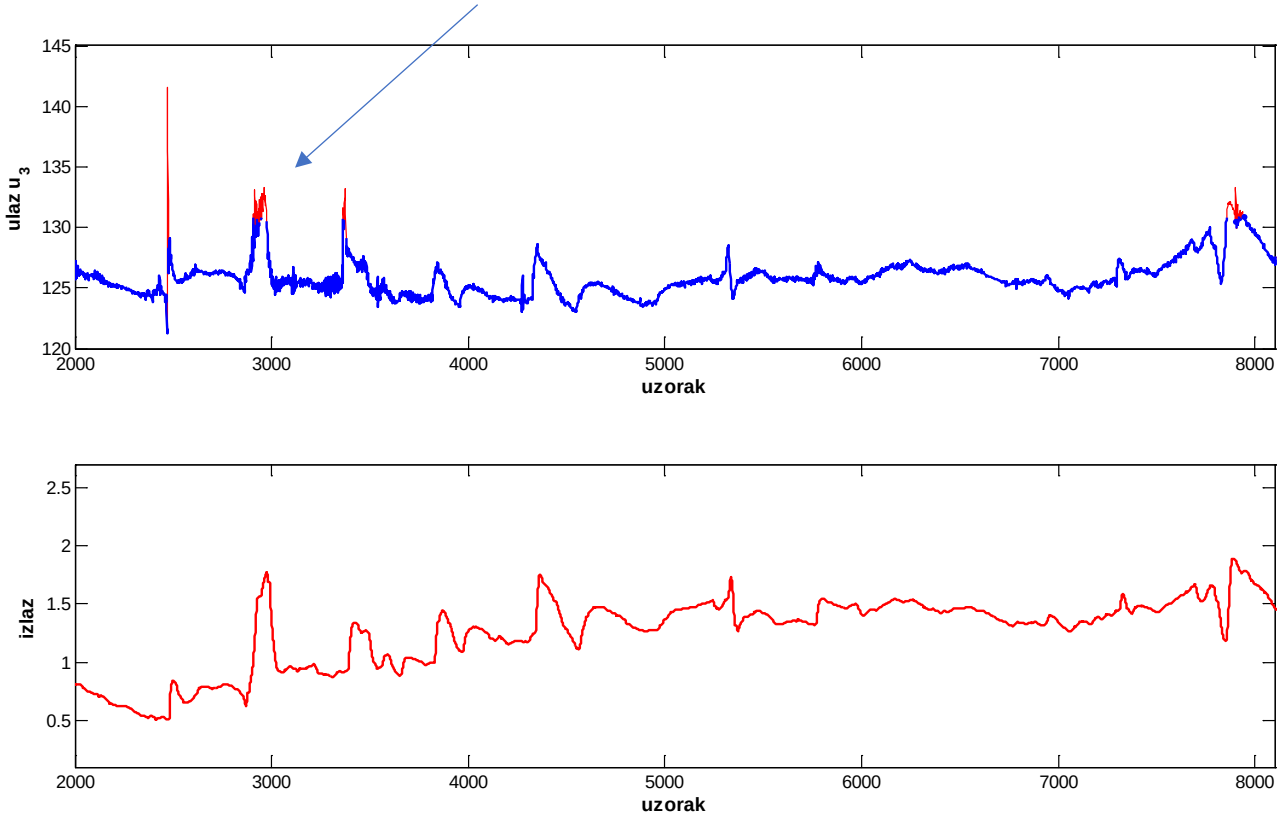
(a) dva *outlier*-a bitno utječu na razvoj modela

(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.

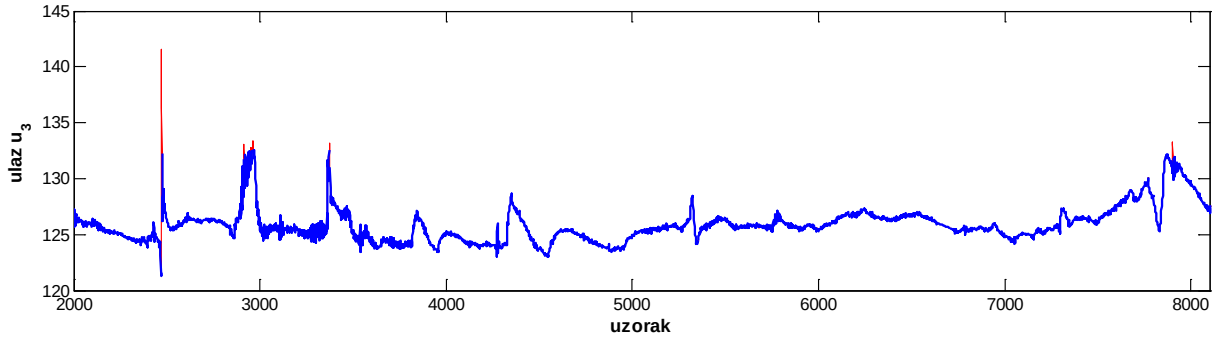
poremećaj u procesu, a ne *outlier*



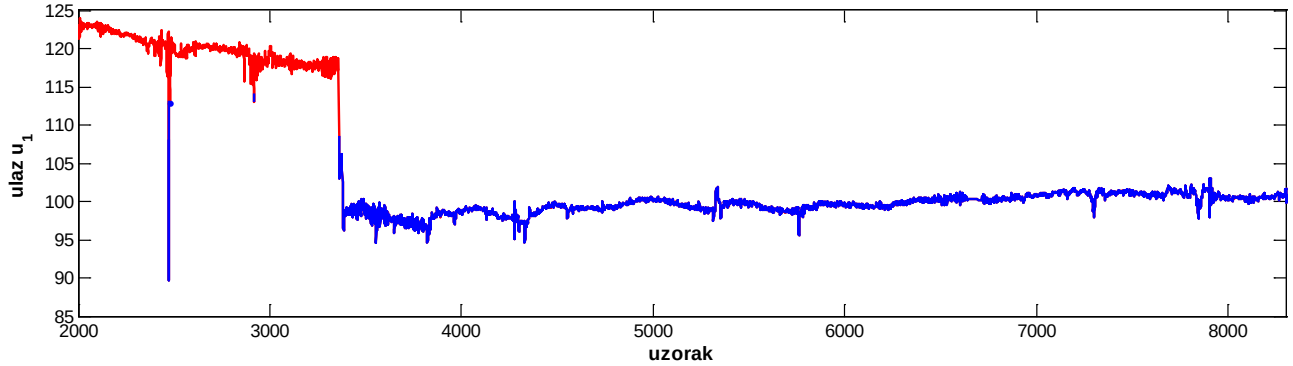
ekstremne vrijednosti crvena boja)

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!*

FILTRIRANJE PODATAKA

- Visokofrekventne te niskofrekventne smetnje trebaju se ukloniti, jer mogu rezultirati nestabilnim i nedovoljno točnim modelom.
- **Filtriranje** je postupak kojim se nastoji **smanjiti mjerni šum** prisutan kod mjernih pretvornika. 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 su obično brze, kratkotrajne promjene u signalu koje se događaju na visokim frekvencijama. Ovaj tip šuma može biti uzrokovan raznim faktorima, uključujući električne smetnje, vibracije, ili kratke fluktuacije u mjernim procesima.

- U industrijskim mjerenjima, visokofrekventni šum može nastati od elektromagnetskih interferencija ili od mehaničkih vibracija u okruženju.
- 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 su polagane, dugotrajne promjene koje se javljaju na niskim frekvencijama. Ovaj tip šuma može uključivati sezonske varijacije ili trendove u podacima.

- Niskofrekventni šum može biti uzrokovan promjenama u vanjskim uvjetima ili dugoročnim trendovima (npr. klimatske promjene).
- Niskofrekventne smetnje mogu uzrokovati pogrešne interpretacije dugoročnih obrazaca (trendova).

Filtriranje podataka – Izglađivanje

- Kod svake metode definira se određeni **raspon** (***span***) – područje susjednih točaka koje se uključuju u proračun nove točke;

- Ovaj raspon se pomiče duž podataka korak po korak za svaku novu vrijednost prediktora;

  - **Veliki raspon** povećava glatkoću, ali **smanjuje rezoluciju** izglađenih podataka (može izravnati važne lokalne promjene);

  - **Mali raspon** smanjuje glatkost, ali **povećava rezoluciju** izglađenih podataka (bolje prati lokalne promjene)

- Optimalni raspon ovisi o skupu podatka, metodi izglađivanja i obično zahtijeva ponešto eksperimentiranja.

Filtriranje podataka – Izglađivanje

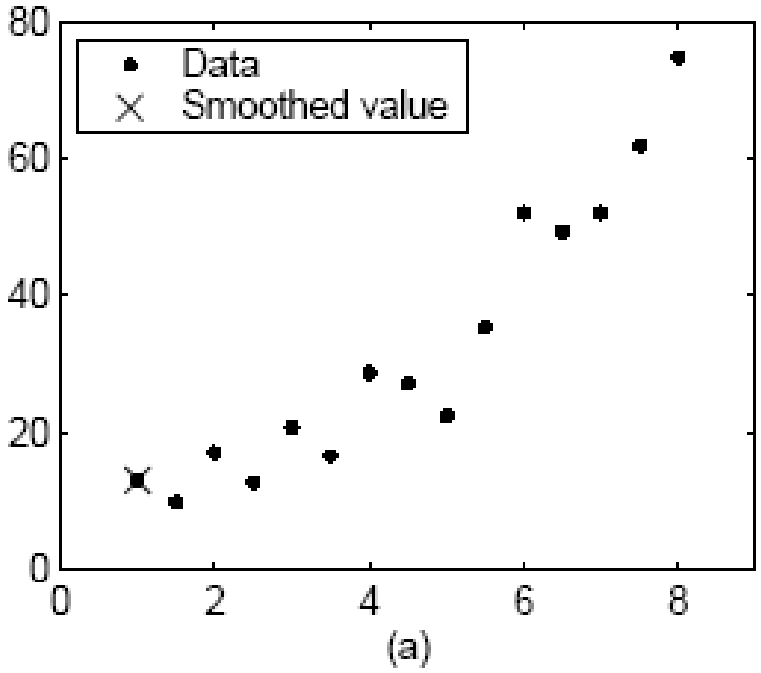
Moving average filtriranje

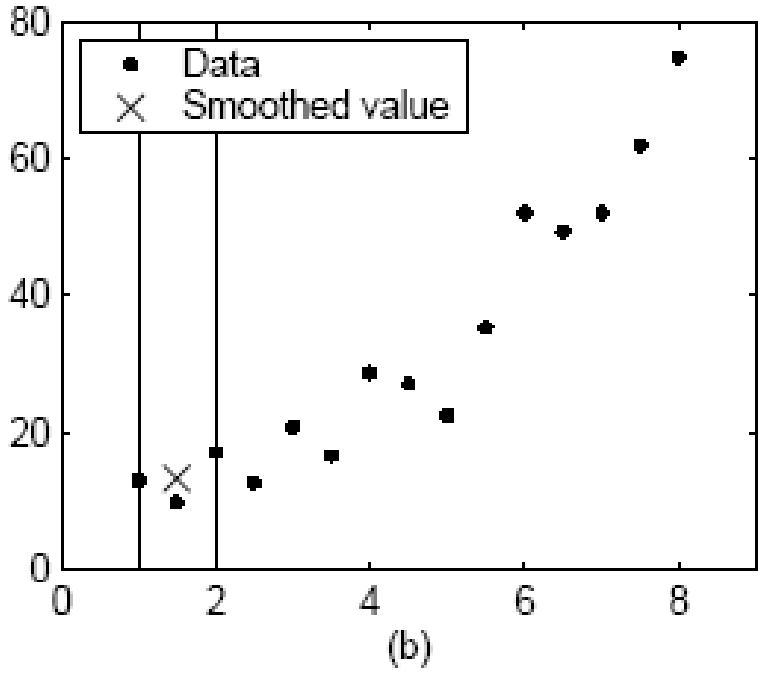
- Filtar koji **propušta niske frekvencije** i uzima sredinu od susjednih podataka;
- Izglađuje podatke tako da **svaki podatak zamjenjuje sa srednjom vrijednosti susjednih točaka** definiranih rasponom;
- Proračun je definiran s jednadžbom razlika:

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

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

Moving average filtriranje



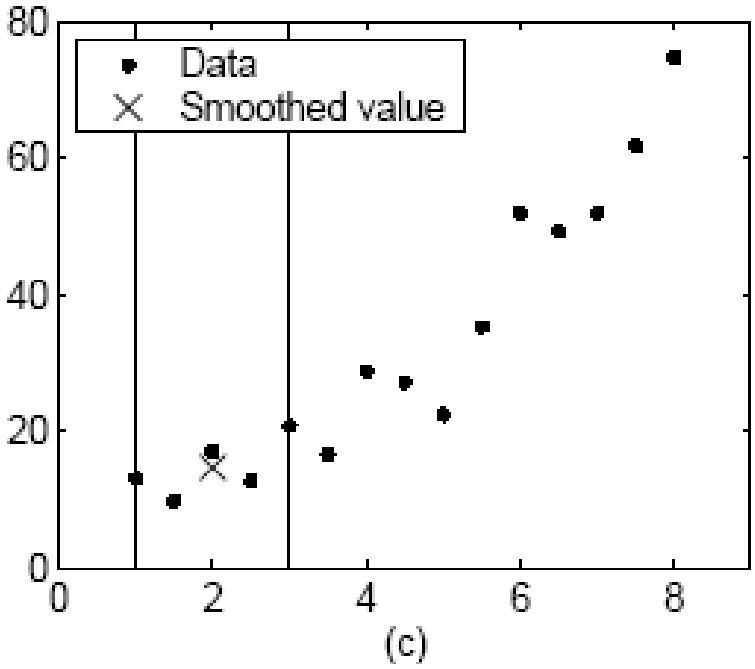


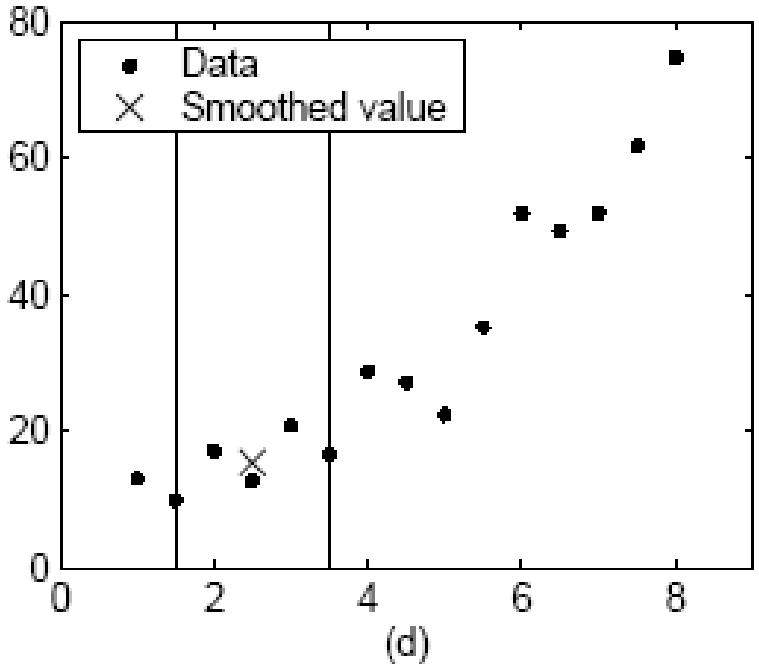
(a) Prva točka nije izglađena jer ne postoji raspon

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

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

Moving average filtriranje





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

$$y_s(3) = (y(1)+y(2)+y(3)+y(4)+y(5))/5$$
$$y_s(4) = (y(2)+y(3)+y(4)+y(5)+y(6))/5$$

LOWESS i LOESS – izglađivanje lokalnom regresijom

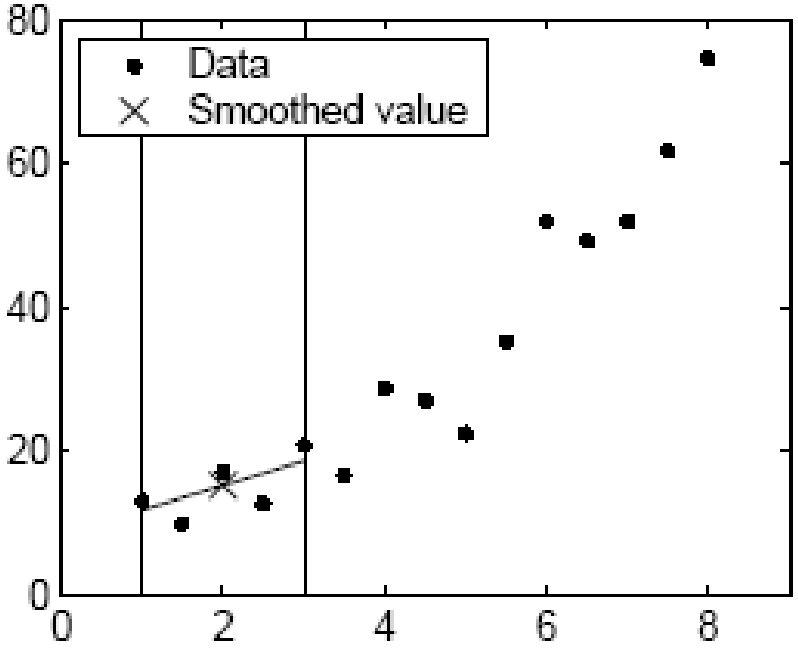
Lowess, **Loess** – "***LOcally WEighted Scatter plot Smooth***"

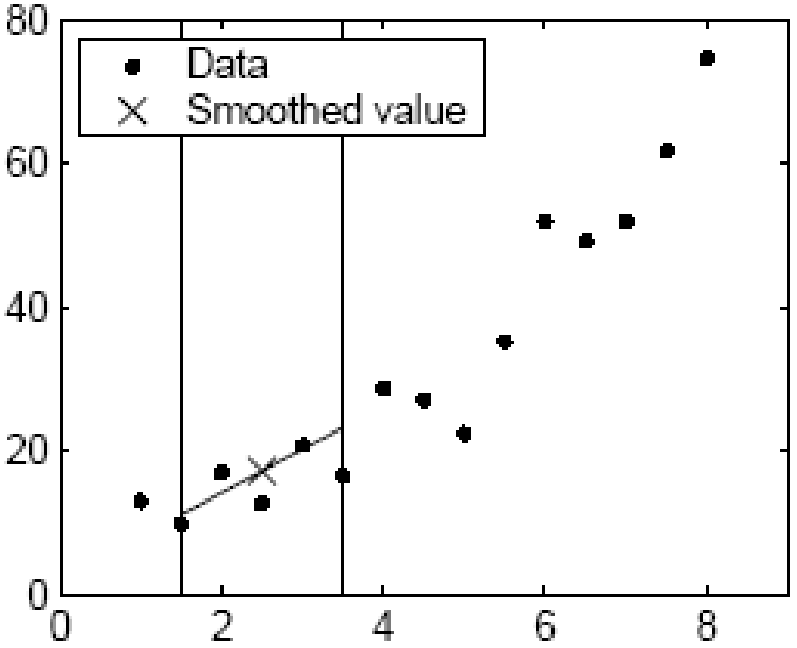
- Obje metode primjenjuju **lokalnu težinsku** linearnu regresiju;
- Smatra se "**lokalnom**" jer se vrijednost određuje na temelju **susjednih točaka** definiranih rasponom;
- Podaci imaju svoje težinske vrijednosti, a može se primijeniti i robusna težinska funkcija kako bi se isključile ekstremne vrijednosti.
- Ovisno o izabranoj "težini" filtriranja, Loess više ili manje zanemaruje podatke koji odskaču 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 – izglađivanje lokalnom regresijom

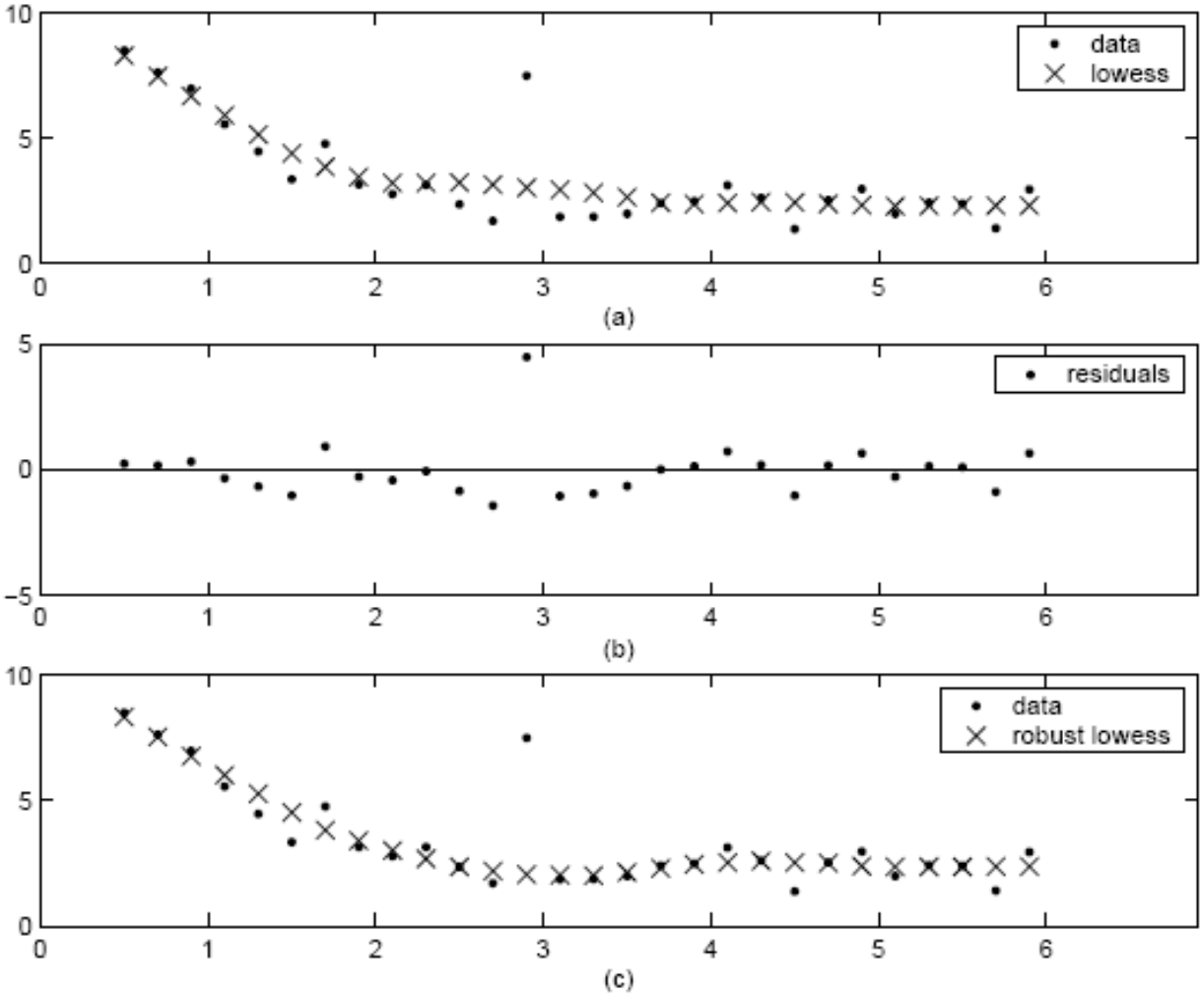




Raspon je stalan, a postupak izglađivanja provodi se od točke do točke.

Ovisno o broju najbližih susjeda, težinske funkcije mogu i ne moraju biti **simetrične** oko točke oko koje se izglađuje.

LOWESS i LOESS – izglađivanje lokalnom regresijom



(a) *Outlier* utječe na izglađenu vrijednost za nekoliko susjednih točaka

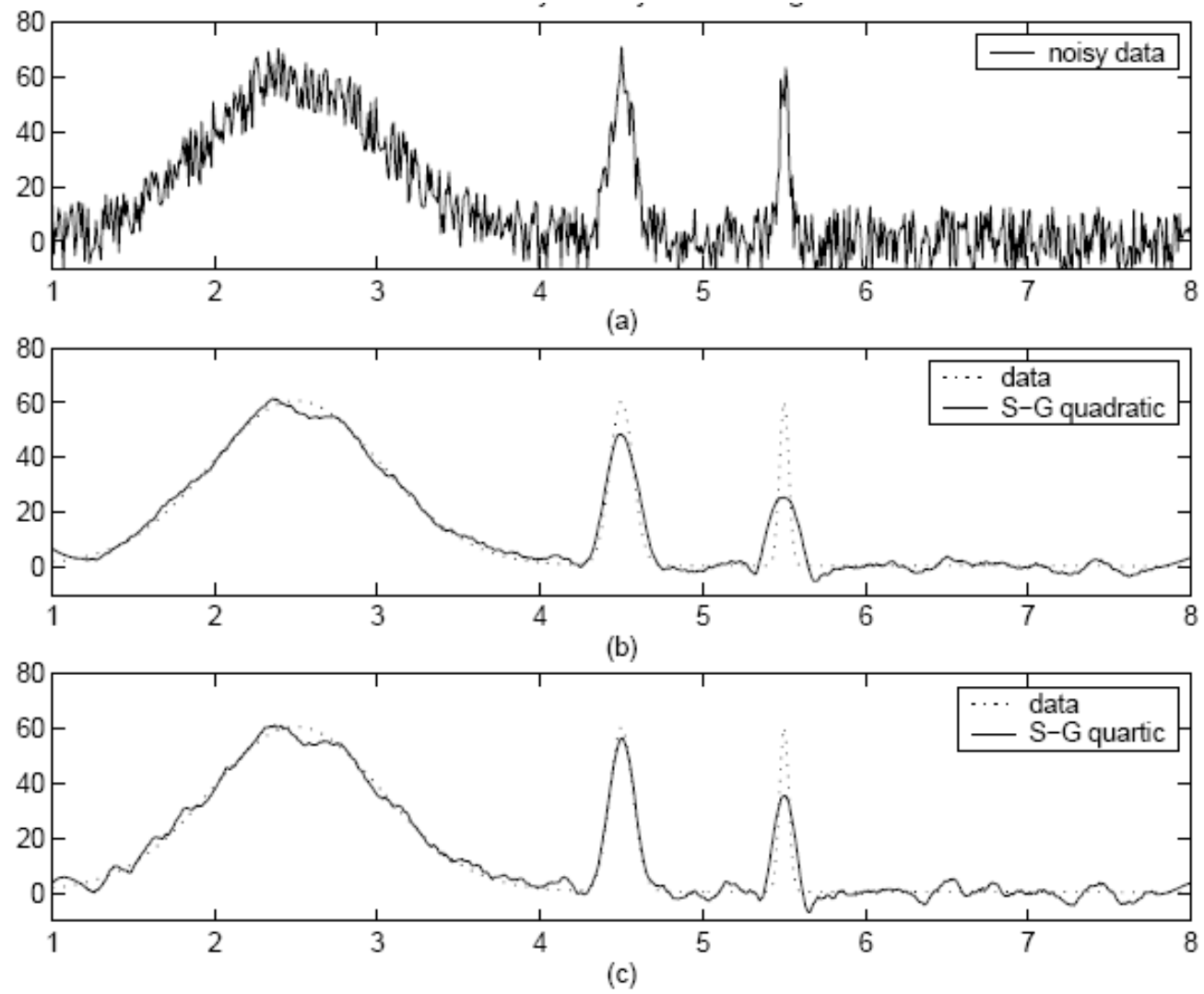
(b) Prikaz ostataka – veći su od 6 medijana apsolutnih odstupanja

(c) Izglađene 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 najmanjeg kvadrata primjenom polinoma određenog stupnja (***digital smoothing polynomial filter*** ili ***least squares smoothing filter***)
- Višim redom polinoma može se postignuti visoka razina izglađivanja 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
- Za usporedbu, MA filtrira veliki dio visokofrekventnog sadržaja, a SG je manje uspješan od MA kod skidanja šuma

SAVITZKY-GOLAY FILTAR

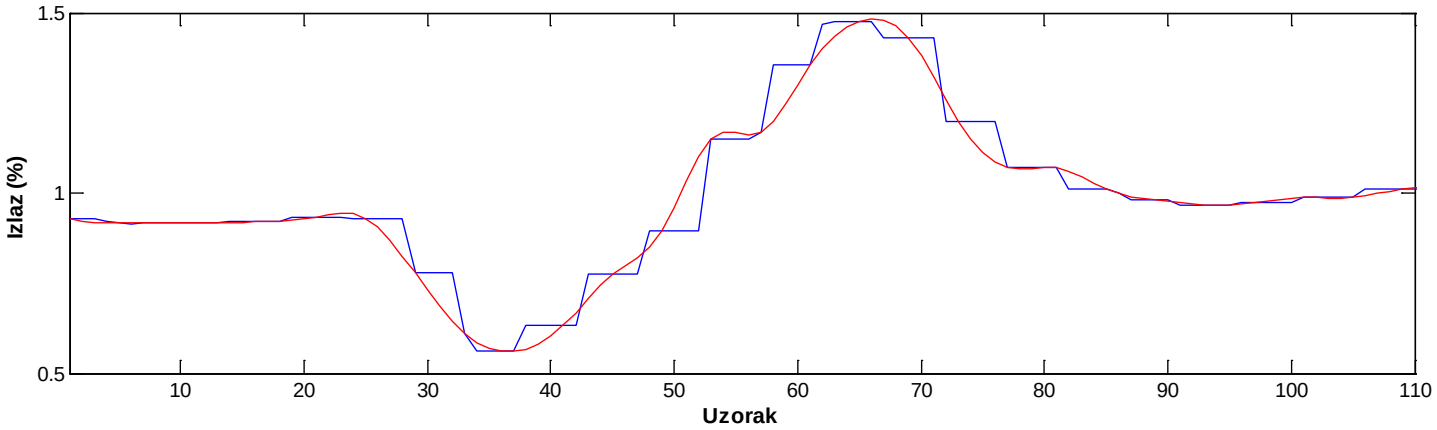


Podaci opterećeni **šumom**

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

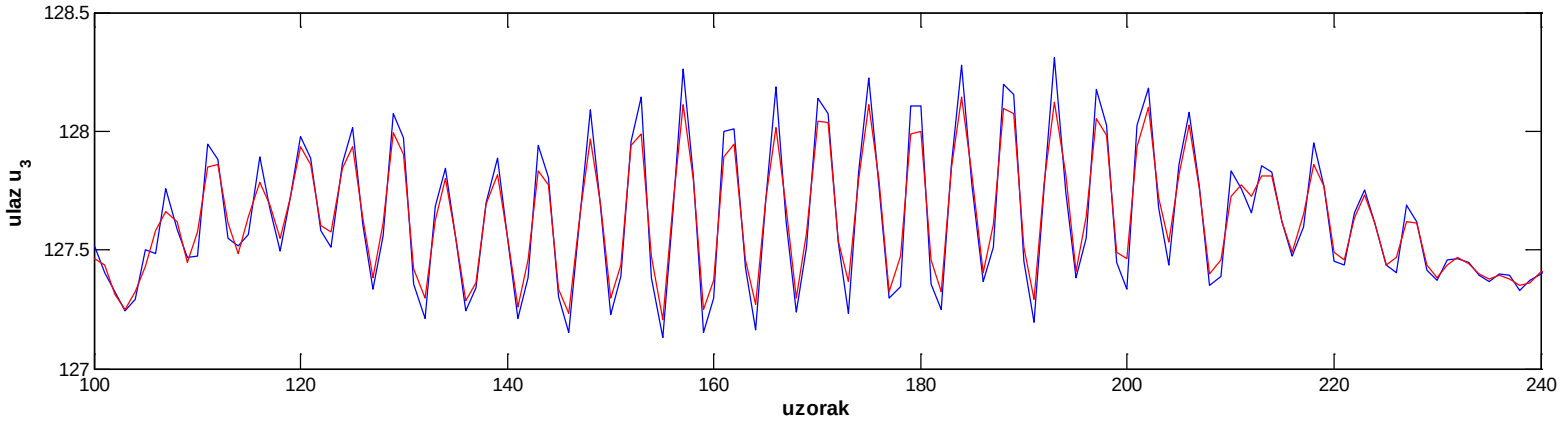
Podaci **bez šuma** – izglađivanje Savitzky-Golay polinomom; općenito, što je veći red bolje se "hvataju" uski pikovi, ali slabije širi

PRIMJERI - Nadomještanje podataka izlaza; Filtriranje podataka



crveno-interpolirani izlaz
plavo-mjereni izlaz

Prikaz interpoliranih vrijednosti dijela izlaznog signala pomoću kubnog spline-a



crveno- filtrirani
plavo- nefiltrirani

Uvećani dio usporedbe filtriranih i nefiltriranih podataka

$\overline{x}$

Skaliranje podataka

Podaci mogu imati različite 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:

$$x^i_{norm} = \frac{x^i - x_{\min}}{x_{\max} - x_{\min}}$$

Ova formula pretvara vrijednosti u raspon od 0 do 1

x – neskalirana varijabla
x_{norm} – skalirana varijabla

"Z-score" normalizacija (standardizacija):

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

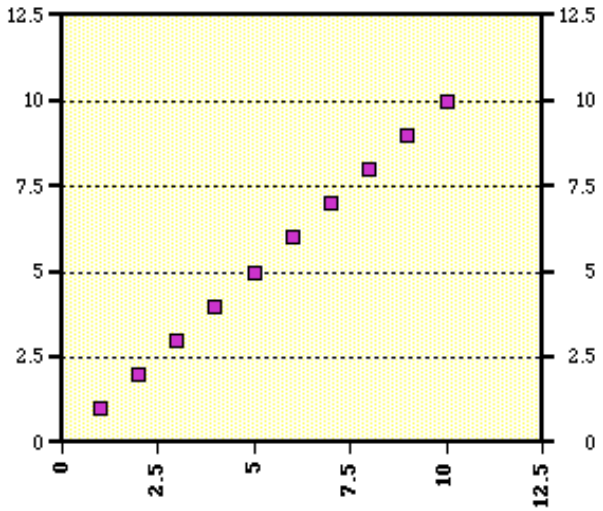
Ova formula normalizira vrijednosti tako da imaju srednju vrijednost 0 i standardnu devijaciju 1.

DIJAGRAM RASPRŠENJA (engl. *Scatterplot*)

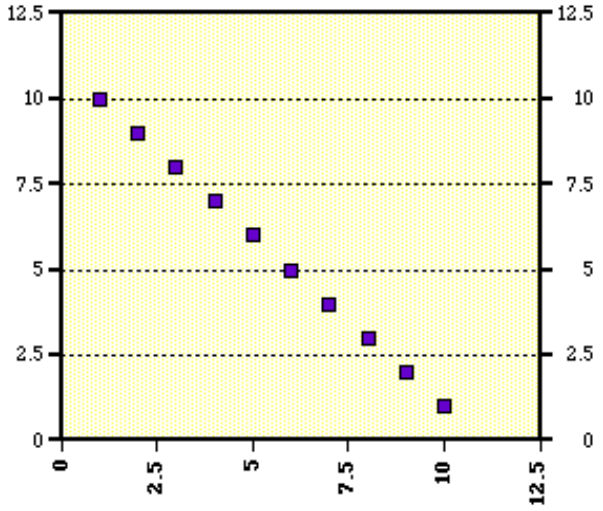
- Često opažamo kako dvije pojave pokazuju međusobnu zavisnost ili određeni stupanj povezanosti;
- Standardizirana mjera jakosti statističke veze između pojava predočenih dvjema kvantitativnim varijablama jest **koeficijent korelacije (*R*)**.
- **Dijagrami raspršenja** pokazuju koliko jedna varijabla 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)



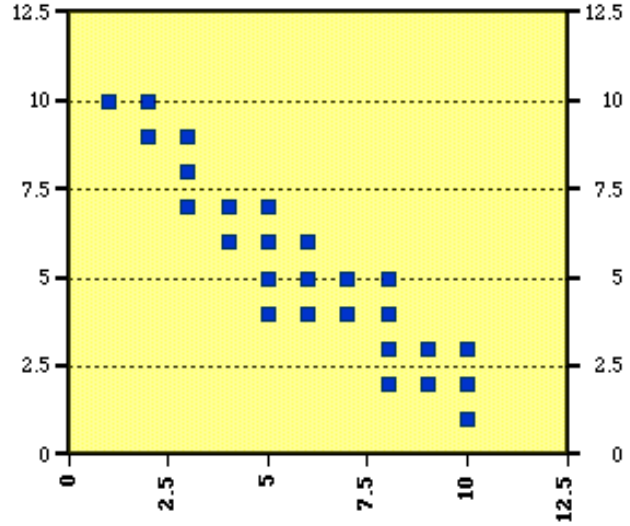
Idealna negativna korelacija (*R*=-1)



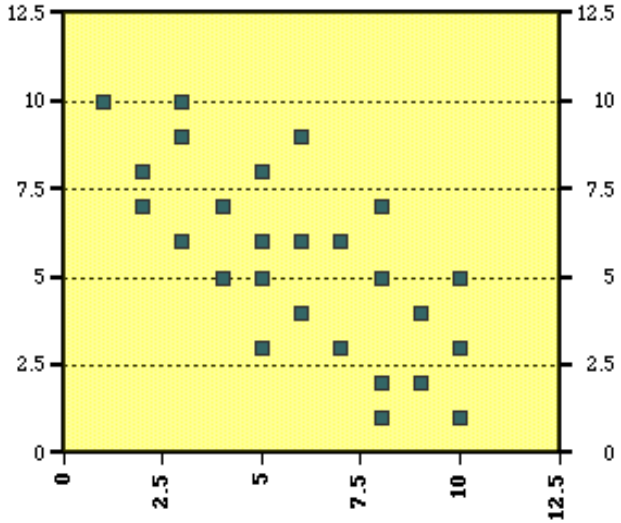
Što je korelacija bliža 1 ili -1, to je ona jača, a što je bliže 0, to je slabija korelacija.

DIJAGRAM RASPRŠENJA

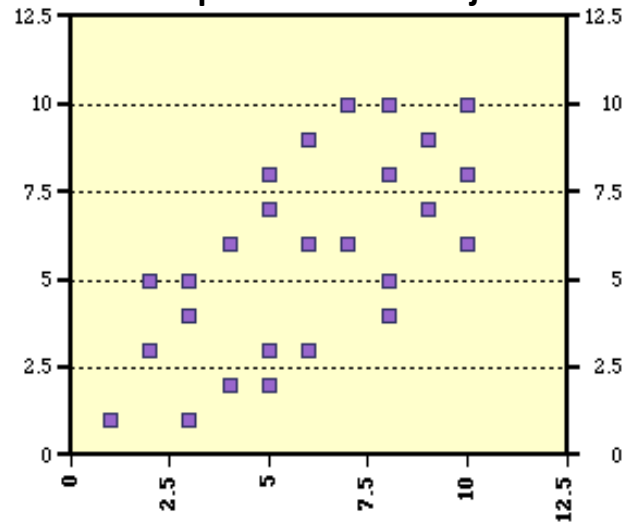
Velika negativna korelacija



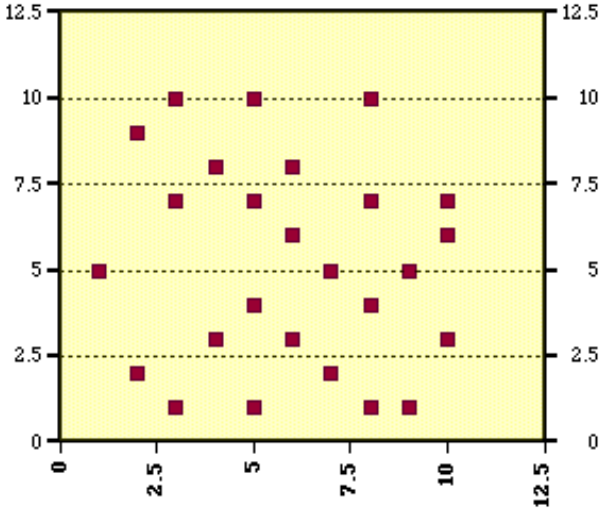
Mala negativna korelacija



Niska pozitivna korelacija



Nema korelacije (*R* = 0)



***R* od 0 do ± 0,2**
nikakva ili neznatna povezanost

***R* od ± 0,2 do ± 0,4**
manja povezanost

***R* od ± 0,4 do ± 0,7**
značajna povezanost

***R* od ± 0,7 do ± 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).

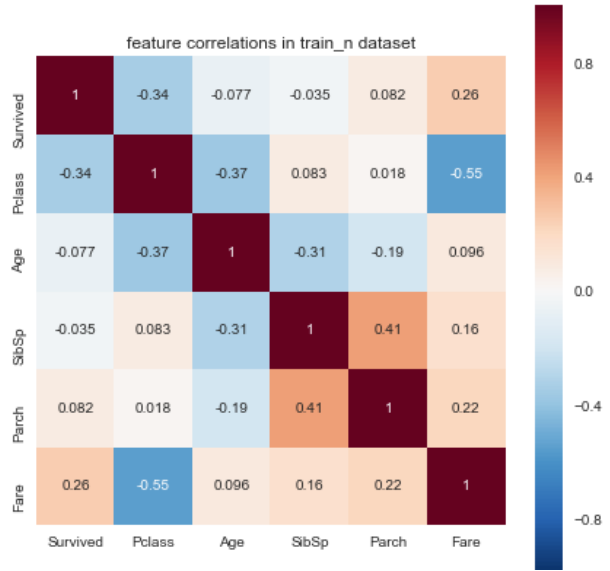
PEARSON-ov KOEFICIJENT KORELACIJE

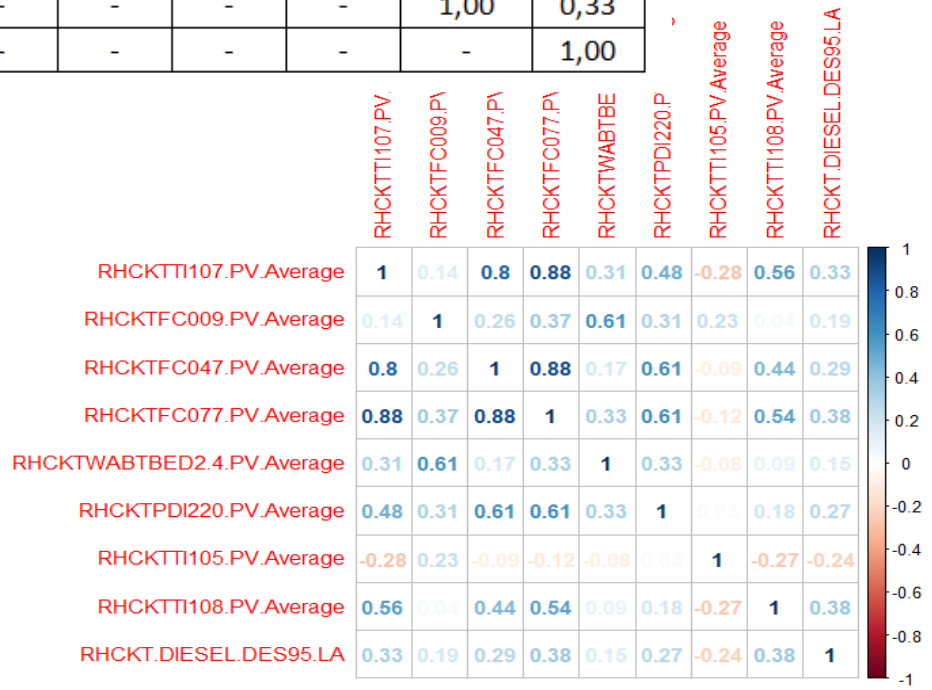
- Mjera linearne povezanosti dvije normalno distribuirane varijable.
- Ako varijable ne ovise jedna o drugoj, onda je $R = 0$, a ako ovise onda se R nalazi u rasponu od minus 1 do 1.
- Koristi se kod **vrednovanja** modela te kod **odabira ulaznih varijabli** u model - multikolinearnost.
- Izračunava 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).

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

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, što smanjuje utjecaj *outlier*-a na rezultat.
- **Spearmanov koeficijent** se koristi kada podaci nisu normalno distribuirani ili kada se želi procijeniti monotona povezanost između varijabli.

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

$\boldsymbol{d_i}$: Razlika između rangova za svaki podatak. Npr., ako je za jedan podatak vrijednost prve varijable rangirana kao 2, a druge kao 5, tada je $\boldsymbol{d_i} = 2-5 = -3$. *d*i se računa za svaki podatak u skupu podataka.
$\boldsymbol{n}$: Broj podataka

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.

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

Koeficijent višestruke determinacije - R^2

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

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

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*)**

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

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

KRITERIJI VREDNOVANJA MODELA - *FIT*

- **Kriterij slaganja modela** (*FIT*)

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

y	Izmjerene vrijednosti
$\hat{y}$	modelom procijenjena vrijednost
$\overline{y}$	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

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

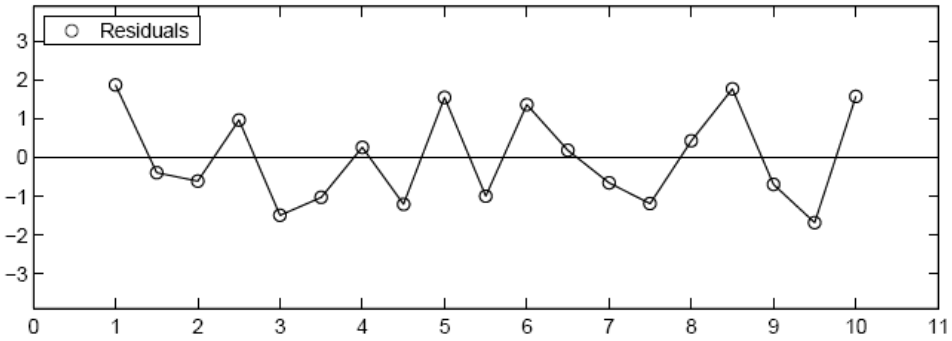
y	mjerena veličina
$\hat{y}$	modelom procijenjena vrijednost
N	broj podataka
Θ	procjenjeni parametri
d	broj procjenjenih parametara

Rezidualna analiza

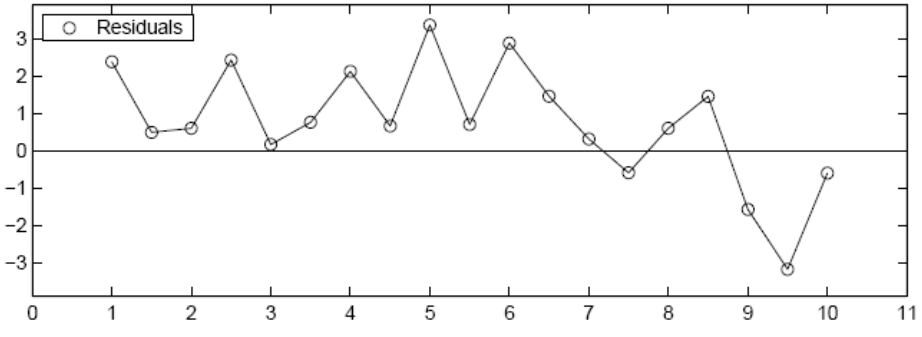
- **Praćenje trenda pogreške (reziduala)** - Reziduali se često grafički prikazuju zajedno s odgovarajućim intervalom pouzdanosti (obično 95%).
- **Histogram pogreške**

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

Rezidual - Razlika između stvarne vrijednosti i vrijednosti modela.



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

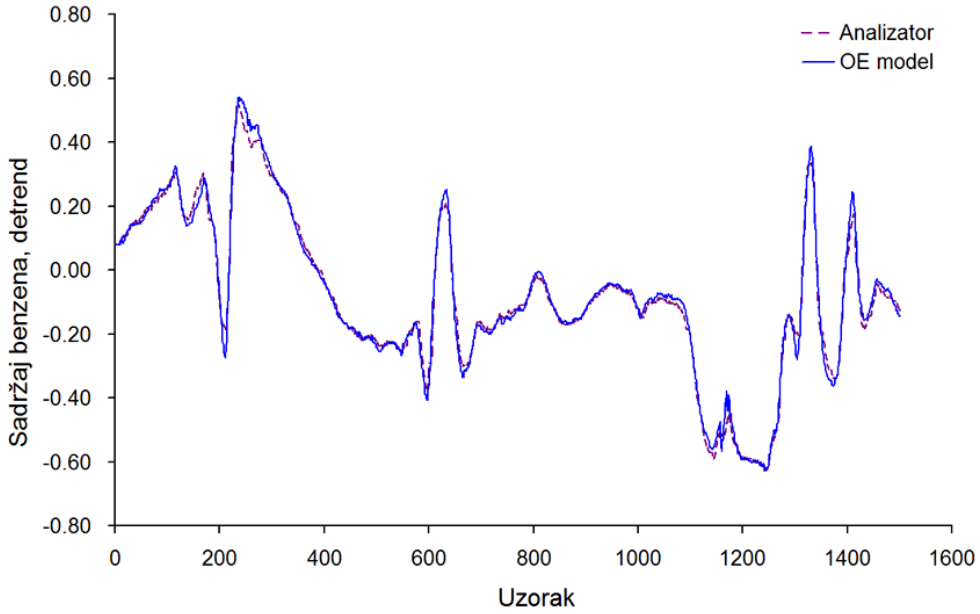


Reziduali su pozitivni za većinu podataka što znači da model ne opisuje dobro podatke (sustavna pogreška)

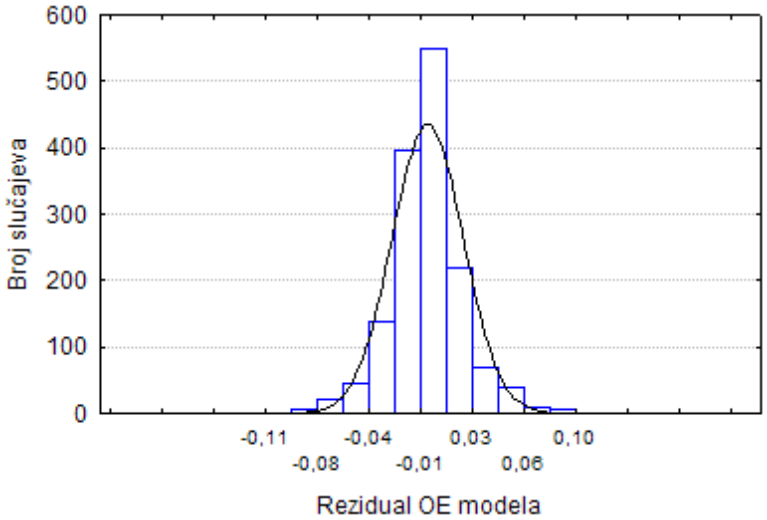
Primjer vrednovanja modela

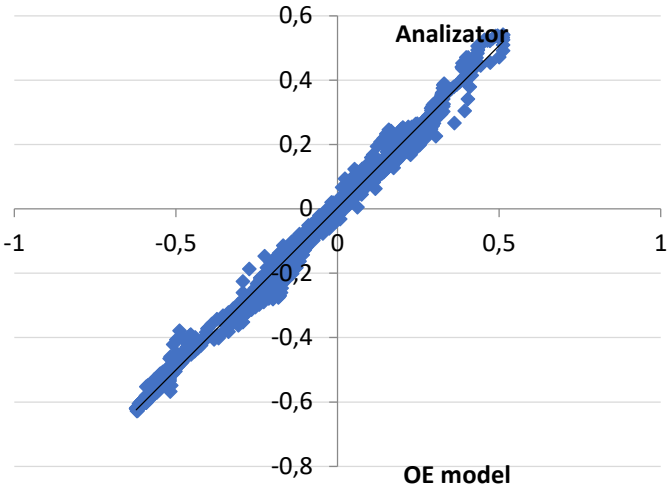
Sadržaj benzena u lakom reformatu

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



Usporedba sadržaja benzena određenog analizatorom i modelom





Knjižnice (*library*) za predobradu podataka u Pythonu

python™

pandas

https://pandas.pydata.org/

NumPy

https://numpy.org/

SciPy

https://www.scipy.org/

Vizualizacija podataka:

scikit learn

https://scikit-learn.org/

matplotlib

https://matplotlib.org/

Funkcije u Pythonu za predobradu podataka

Aritmetička sredina

Raspon

Minimum

Maksimum

Simetričnost distribucije

Zakrivljenost distribucije

Standardna devijacija

Varijanca

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

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

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

Funkcije u Pythonu za predobradu podataka

python™

Aritmetička sredina

Medijan

Mod

Kvartili

Interkvartil

Standardna devijacija

Koeficijent varijacije

Skaliranje podataka

Kubni spline

Pearsonov koef. korelacije

R^2 (*R square*)

```
import numpy as np
from scipy import stats

aritmeticka_sredina = np.mean(Niz_podataka)
medijan = np.median(Niz_podataka)
mod = stats.mode(Niz_podataka)

Q1 = np.quantile(Niz_podataka,0.25)
interkvartil = stats.iqr(Niz_podataka)

StandardnaDevijacija = np.std(Niz_podataka)

KoeficijentVarijacije = stats.variation(Niz_podataka)

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_podatakal,Niz_podataka2)

Pearsonov_kor_koef= np.corrcoef(Niz_podatakal,Niz_podataka2)

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

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

Funkcije u Pythonu za predobradu podataka

3 Sigma pravilo

```
from scipy.stats import sigmaclip

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

Hampel filtar

```
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 hample_obj: filtered_data, outlier_indices, medians, medain_absolute_deviations, thresholds
```

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 (možemp zamisliti kao svojevrsni window_size)
```

Savitzky-Golay filtar

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

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.
```

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

Funkcije u Pythonu za predobradu podataka

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
```

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