L4-5

Grigore Stamatescu
grigore.stamatescu@upb.ro

Realizarea unei histograme în MATLAB folosind generatorul de numere aleatoare

În cazul în care nu aveţi la dispoziţie date experimentale pentru realizarea unei histograme în MATLAB, puteţi folosi funcţia *random* pentru a genera numere aleatoare cu o distribuţie de probabilitate dată. Acest mod de lucru este util în special pentru a realiza simulări. În acest caz vom simula un set de date pentru a exersa realizarea histogramelor.

Sintaxa pentru apelarea funcţiei random este:

- random(NUME,A) returnează un vector de numere aleatoare, alese din distribuţia de probabilitate cu un parametru, specificată de NUME, cu valoarea parametrului A.
- random(NUME,A,B) sau random(NUME,A,B,C) returnează un vector de numere aleatoare, alese dintr-o distribuţie de probabilitate cu valorile parametrilor A, B (şi C).

Există multe tipuri de distribuţii ce pot fi alese la utilizarea acestei funcţii. Introduceţi *help random* în linia de comandă pentru a afişa o listă completă cu numele şi parametrii de intrare.

Distribuţie	Parametrul de intrare A	Parametrul de intrare B
'bino' sau ' Binomial'	n: numărul de încercări	p: probabilitatea de succes pentru fiecare încercare
'exp' sau 'Exponential'	μ: media	-
'norm' sau 'Normal'	μ: media	σ: deviaţia standard
'unif' sau 'Uniform'	a: punct inferior (minim)	b: punct superior (maxim)

Pentru a crea un vector sau o matrice cu numere aleatoare, folositi functiile de mai sus, urmate de dimensiunile matricei. De exemplu:

- random('norm',mu,sigma,1,N) va returna o selecţie aleatoare de N valori dintr-o distribuţie normală şi va plasa valorile într-un vector uni-dimensional de lungime N.
- random('bino',n,p,M,N) va returna o selecţie aleatoare de valori dintr-o distribuţie binomială şi va plasa valorile într-o matrice MxN.

Odată creat setul de date cu ajutorul generatorului de numere aletoare, puteţi reprezenta grafic datele, folosind funcţia histogramă.

Exemplu Realizaţi o rutină MATLAB pentru a genera o selecţie aleatoare de 1000 de puncte de date dintr-o distribuţie Gaussiană cu $\mu = 1$ şi $\sigma = 0.5$

```
media=1;
dispersia=0.5;
N=1000;

date=random('Norm',media,dispersia,1,N);

figure(1);
hist(date); \% valoarea standard este de 10 clase
title('Histograma cu 10 clase');
```

Exerciţiu Modificaţi μ, σ, N şi numărul de clase. Cum afectează aceste modificări forma histogramei?

În multe aplicații apare necesitatea ca pe baza datelor preluate prin experiment direct şi pentru care se remarcă dependența cauzală între factori,să se determine în formă explicită dependența dintre mărimea (mărimile) cauză pe care o vom defini *mărime predictor* şi mărimea efect pe care o definim ca *mărime prezisă***.**

În acest sens un prim scop al lucrării propuse este de a prezenta modul de impunere a unei asemenea aplicații, posibilități de soluționare şi evaluare a soluțiilor impuse.

4.1.2 Descrierea lucrării

Considerăm un obiect oricare supus unui factor extern măsurabil şi eventual controlabil şi care generează un efect evidențiat printr-un al doilea factor măsurabil.

Asupra obiectului acționează din exterior o serie de mărimi cu acțiune perturbatoare pe care le considerăm nemăsurabile şi necontrolabile. Mărimile cu acțiune perturbatoare sunt mărimi aleatoare (eventual caracterizate din punct de vedere statistic).

Mărimile cunoscute care constituie date de intrare sunt $\{x_i, y_i\}, i \in 1,2,....n$, deci eşantioane corelate între mărimile cauză şi efect obținute prin măsurare directă în prezența factorilor aleatori.

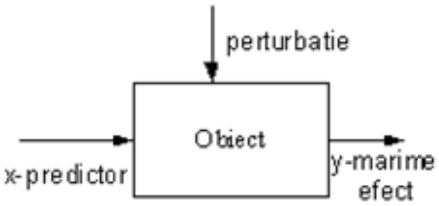


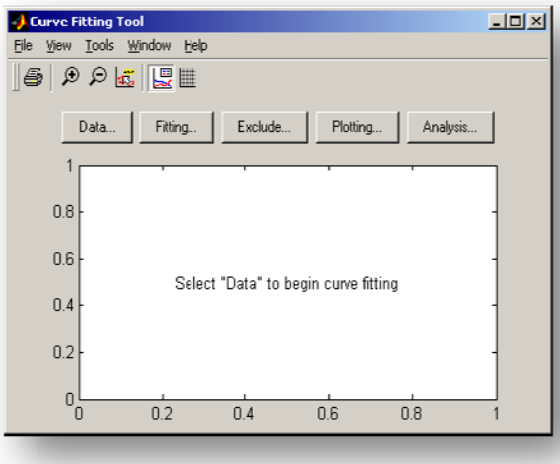
Fig.4.1. Exemplificarea modului de abordare a lucrării

Urmărim să obținem pe baza datelor de intrare o caracterizare completă, cantitativă a dependenței acceptate în forma: $y = f(x)$.

Impusă în această formă problema este deosebit de dificil de abordat. Dificultățile de abordare sunt cauzate de:

1. Nu este cunoscută dependența de tip funcție $f(\cdot)$
2. Mărimile măsurate sunt afectate de perturbațiile de natură aleatoare.

Problema poate fi simplificată considerabil dacă impunem forma funcției $f(\cdot)$ dar pentru care rămâne să stabilim o serie de parametri ce caracterizează funcția. Introducem în acest mod un grad de aproximare în construcția propusă care în final va trebui caracterizată. În maniera propusă evaluarea aproximativă va fi de forma:



$$y_i = f(x_i) + \varepsilon_i$$

relație care pune în evidență eroarea reziduală:

$$\varepsilon_i = y_i - p_i = y_i - f(x_i)$$

Eroarea reziduală este dependentă de impunerea

forțată a funcției $f(\cdot)$ precum şi de influența mărimilor perturbatoare.

În fereastra de mai sus sunt fixate butoanele de comandă:

- *Data* **.** Permite importul de date pentru asigurarea datelor primare.
  *Fitting***.** Permite alegerea tipului de funcție de predicție şi stabilirea parametrilor ce caracterizează această funcție.
- *Exclude***.** Permite eliminarea de date vizibil afectate de perturbații.
- *Plotting***.** Setează şi afişează graficul dependențelor interesante în elaborarea rezultatelor.
- *Analysis***.** Prezintă indicatorii sintetici necesari evaluării predicției

Construcția funcției de predicție începe prin fixarea datelor primare. Selectăm comanda *Data* care lansează fereastra de lucru prezentată în figura de mai jos.

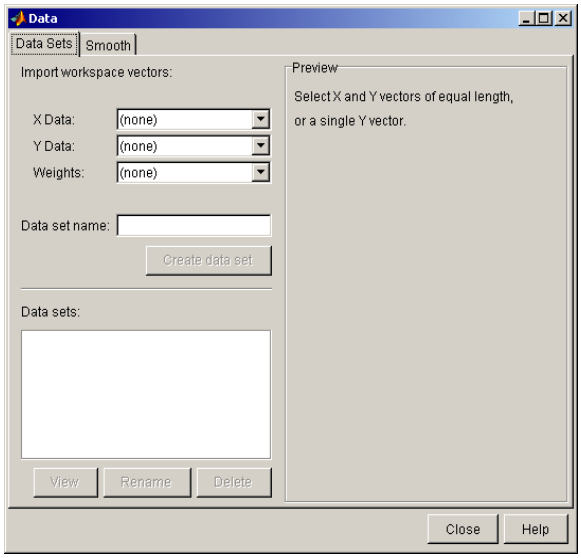


Fig.4.3. Fereastra pentru introducerea datelor primare

În această fereastră de lucru distingem două opțiuni:

1. *Data Sets***.** Fereastră care permite importul datelor primare.
2. *Smooth***.** Fereastra de lucru pentru setarea algoritmilor de filtrare.

Fereastra de lucru pentru introducerea datelor primare, prezentată în fig.4.3, permite importul din spațiu de lucru vectorului mărimilor predictor (*X-Data*), a vectorului mărimilor efect (*Y-Data*), şi a ponderilor - *Weight*- (în lucrare nu utilizăm o astfel de conjunctură). În această fereastră vom defini ansamblul mărimilor cu care vom opera în continuare (*Data set name*) . Fereastra permite prezentarea graficului datelor pe măsura introducerii acestora (*Preview***)**.

Opțiunea *Smooth* deschide fereastra de lucru prezentată în figura din stanga. Este necesar să introducem: datele inițiale ce urmează a fi filtrate (*Original data set***),** denumirea setului de date obținut după operația de filtrare (*Smoothed data set*), metoda de filtrare (*Method*), parametrii impuşi metodei de filtrare aleasă (*Span,Degree*). Filtrarea se realizează activând butonul *Create smoothed data set***.**

Modul *Fitting* reprezintă partea principală a aplicației. Fereastra de lucru apelată este prezentată în figura de mai jos.

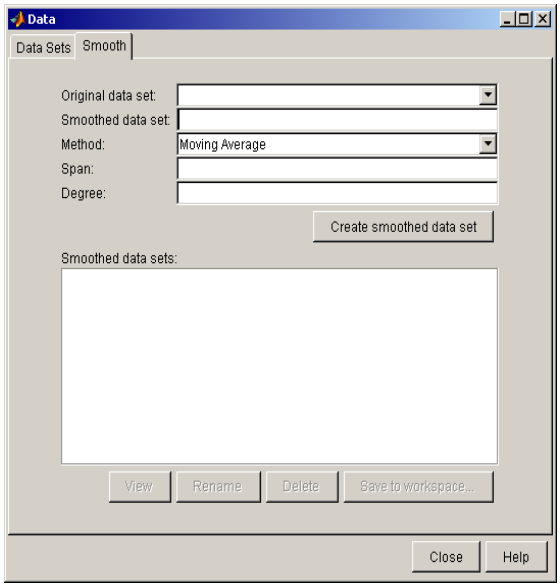


Fig. 4.4. Fereastra pentru filtrarea datelor

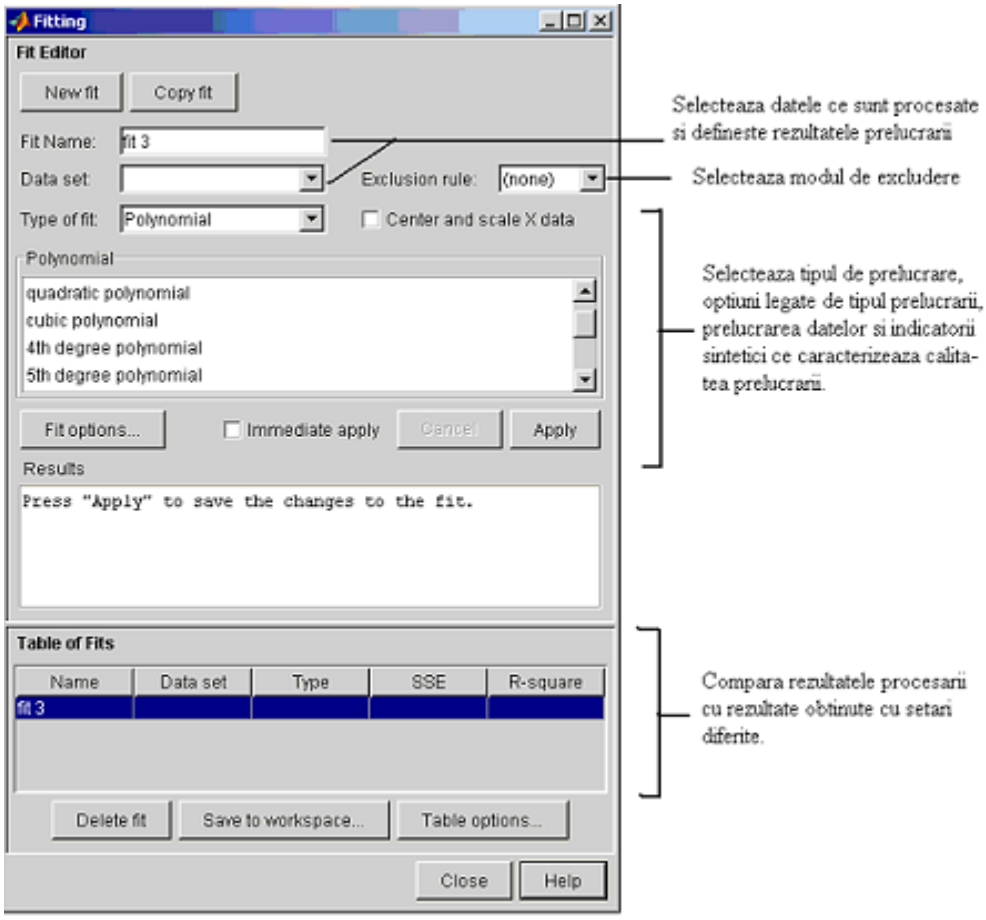


Fig.4.5. Fereastra de setare a parametrilor şi metodei de analiză

Este necesar să precizăm setul de date pe baza căruia construim funcția de predicție (*Fit name, Data set*) precum şi tipul funcției de predicție (*Type of fit*) , parametrii funcției de predicție precum şi indicatorii sintetici ce trebuie calculați în urma predicției.

În final, după rularea programului, în parte de jos a ferestrei sunt prezentați indicatorii sintetici pe baza cărora putem evalua calitatea predicției.

Opțiunile *Analysis* , *Exclude* , *Plotting* nu sunt utilizate în această lucrare.

În continuare vom prezenta pe scurt semnificația indicatorilor sintetici ce se constituie în indicatori de calitate pentru evaluarea funcției de predicție.

După rularea programului legat de stabilirea parametrilor funcției de predicție impusă este necesar să evaluăm în ce măsură predicția este conformă datelor reale. Facilitățile oferite de programul prezentat pot fi evidențiate grafic sau numeric şi se referă fie la evaluarea erorilor reziduale, fie la erori legate de utilizarea funcției de predicție în evaluarea datelor pentru un nivel de încredere fixat.

Erorile reziduale au fost definite anterior prin relația de calcul

$$\varepsilon_i = y_i - p_i$$

relație în care y_i reprezintă valoarea reală măsurată, iar $p_i = f(x_i)$ valoarea predictată.

Putem face o evaluare a calității predicției prin simpla vizualizare a erorilor reziduale. Dacă erorile ε_i au o distribuție aleatoare certifică corecta alegere a tipului funcției de predicție. Într-o astfel de situație erorile reziduale sunt cauzate de perturbațiile de natură aleatoare. Dacă erorile reziduale sunt distribuite preferențial către valori pozitive sau negative, înseamnă că funcția de predicție nu este corect aleasă şi impune alegerea unei alte funcții. În general o apreciere a unei astfel de tendințe poate fi făcută prin evaluarea mediei erorilor reziduale. Dacă media este nulă funcția a fost corect aleasă. În caz contrar se impune alegerea unei alte funcții.

Indicatorii sintetici ce pot fi evaluați în prin rularea programului de procesare şi care sunt utilizați în această lucrare sunt:

- *SSE-(Sum of Squares Error)* **.** Suma pătratelor erorilor este calculată în forma:

$$SSE = \sum_{i=1}^{n} w_i (y_i - p_i)^2 \qquad (4.4)$$

unde w_i reprezintă ponderile alese în procesul de predicție. Evident cu cât valoarea acestui indicator este mai apropiată de zero cu atât predicția este mai bună.

- *R-square* **.** Este un indicator sintetic care indică în ce măsură datele predictate pot urmări variația funcției. Indicatorul este definit în forma:

$$R - square = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (4.5)$$

în care

$$SSR = \sum_{i=1}^{n} w_i \left(p_i - \bar{y}\right)^2 \qquad (4.6)$$

şi

$$SST = \sum_{i=1}^{n} w_i \left(y_i - \bar{y}\right)^2 \qquad (4.7)$$

Indicatorul ia valori cuprinse între 0 şi 1 reliefând o predicție corectă pentru valori apropiate de unitate.

4.1.3 Modul de lucru

Pentru o mai bună înțelegere a modului de operare prezentăm în cele ce urmează o aplicație de procesare numerică în ideea construcției funcției de predicție. Un prim set de date are un caracter determinist nefiind perturbat. Acest set de date este construit astfel:

$$y_i = 0.0001 \cdot (2x_i^3 + 14x_i^2) + 20$$

Cel de al doilea set de date este construit pe baza primului set peste care suprapunem o componentă aleatoare. Secvența de linii de comandă necesară introducerii datelor primare este prezentată mai jos:

```
X=30*randn(1,100)+50;↓
>> Y1=0.0001*(2*X.^3+14*X.^2)+20;↓
>> Y=0.0001*(2*X.^3+14*X.^2)+20+10*20*randn(1,100);↓
```

Începem cu prima secvență. Introducem de la tastatură:

cftool↓ - comandă care lansează fereastra grafică. Selectăm *Data* şi în fereastra grafică deschisă introducem variabilele primare.

Înscriem predictorul X în *XData* şi vectorul mărimilor efect Y1 în *YData*. Ansamblul mărimilor primare este denumit *Test1*.

Mărimile introduse au un caracter determinist şi nu este necesară o preprocesare. Cazul în care peste semnalul util cu caracter determinist se suprapune aditiv un semnal aleatoriu, şi pentru care este necesară o preprocesare va fi prezentat în §.4.2.

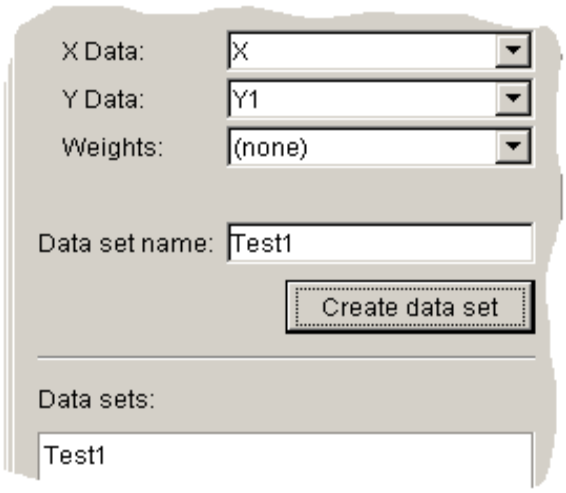
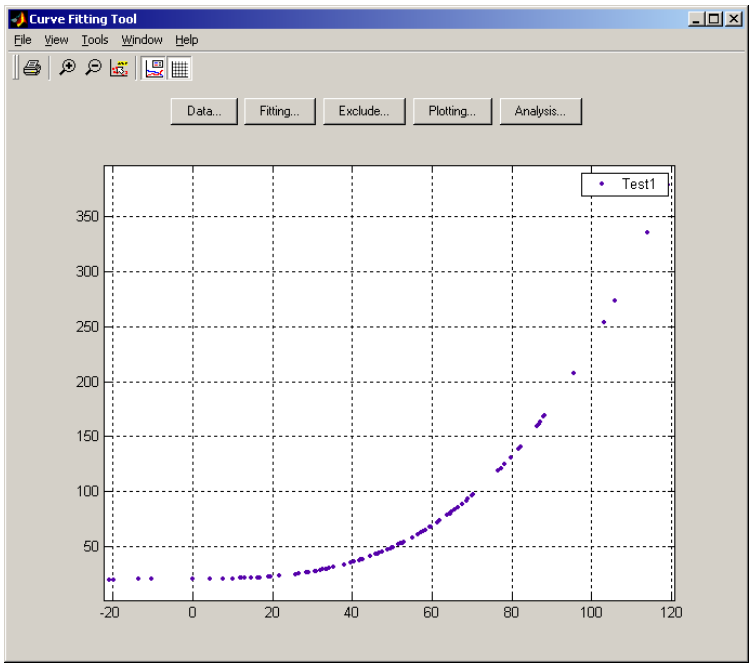


Fig 4.6. Fereastra de setare curentă

Închiderea ferestrei de introducere a datelor lansează automat o fereastră de prezentare a graficului datelor de intrare.

Fig.4.7. Graficul datelor de intrare



Apelăm în continuare opțiunea *Fitting* care deschide fereastra de lucru *Fit Editor* (fig.4.8).

Implicit se inițializează *Fit Name* cu datele conținute în *Test1*. Selectăm opțiunea *Polynomial [linear polynomial* pentru *Type of fit***.** Programul se lansează în execuție acționând *Apply***.**

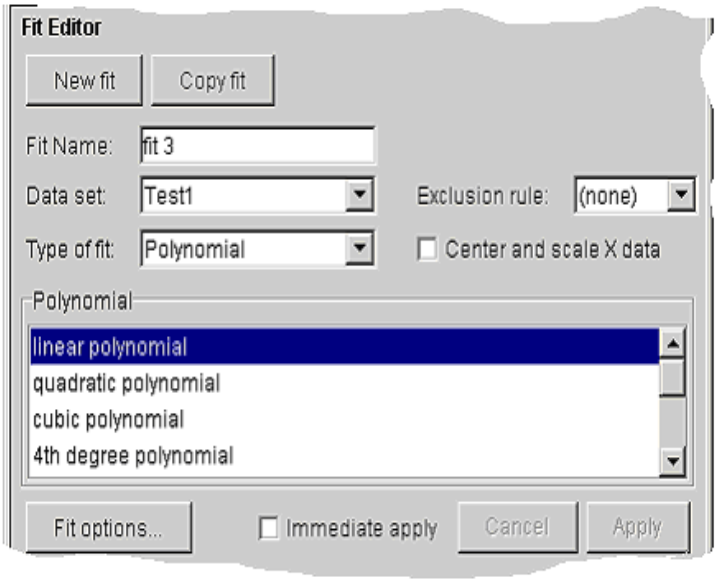


Fig 4.8. Setarea tipului funcției de predicție

La încheierea programului se afişează în partea inferioară a ferestrei *Fit Editor* rezultatele în urma procesării datelor (*Results*) precum şi indicatorii statistici de calitate (*Tables of Fits*) .Rezultatele sunt prezentate în figura 4.9.

Simultan în fereastra *Curve Fitting Tool* este prezentat graficul de variație a curbei predictate pentru a putea fi comparată cu curba de variație reală. Se continuă cu secvența *New Fit – Polynomial-Quadratic Polynomial-Apply, New Fit –Polynomial-Cubic Polynomial-Apply. New Fit –Polynomial-4-th Degree Polynomial-Apply.* În general putem evalua comparativ şi cu alte tipuri de funcții conținute în biblioteca programului sau care pot fi imouse prin realizarea unor subrutine.

Rezultatele sunt prezentate în ultima figura din acest tutorial.

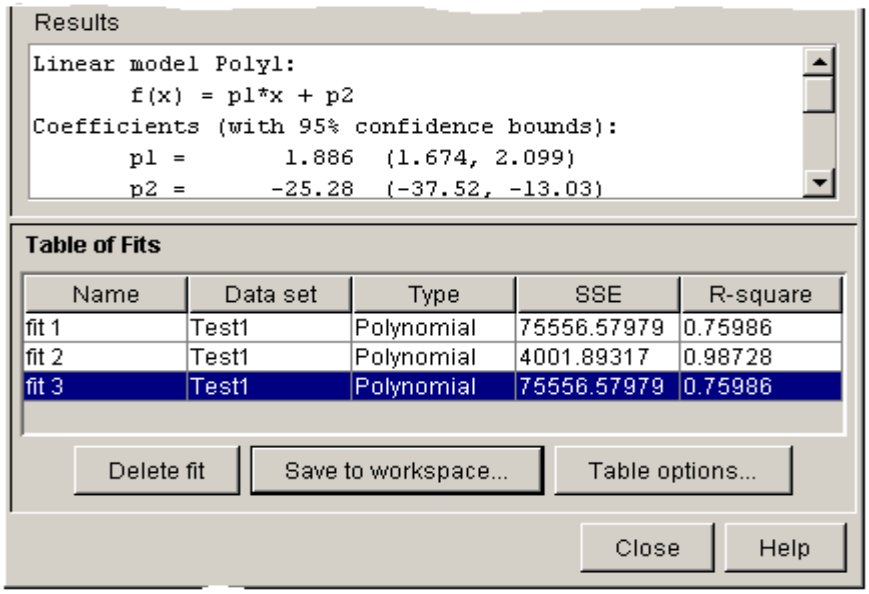
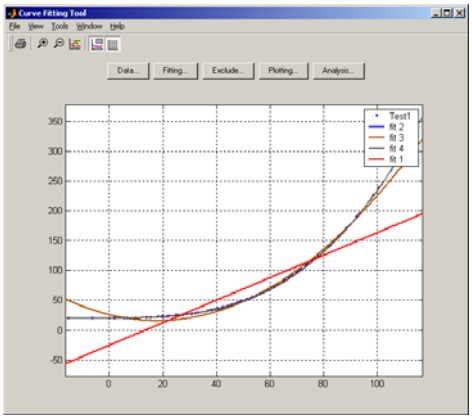


Fig 4.9. Fereastra de prezentare a rezultatelor finale



Graficul de prezentare a rezultatelor finale

Metode şi algoritmi pentru studiul erorilor reziduale

Scopul lucrării

Lucrarea pe care o prezentăm în cele ce urmează este continuarea celei anterioare descrise pentru cazul în care datele utile sunt puternic afectate de o perturbație ce se aplică aditiv. Într-o astfel de situație studiul erorilor reziduale devine deosebit de important.

Descrierea lucrării

Se consideră un obiect supus unor mărimi de excitație externe $x_i \; i \in 1,2,...N$ şi care formează vectorul mărimilor de intrare. Răspunsul obiectului considerat la o excitație x_i se consideră y_i. Între cele două mărimi admitem că există o dependență biunivocă de forma:

$$y_i = f(x_i)$$

pentru care nu cunoaştem forma funcției. Mai mult, în procesul de evaluare a valorilor y_i peste semnalul util se suprapune un semnal de natură aleatoare. Problema propusă (la fel ca şi în lucrarea precedentă) este ca pe baza datelor de măsurare reale să se determine funcția de interdependență între variabilele $x_i \; si \; y_i$. În majoritatea cazurilor problema se soluționează astfel:

- Alegem tipul de funcție care să permită o bună aproximare a curbei de variație reală, care conține o serie de parametri necunoscuți;
- Printr-o prelucrare a datelor reale obținem soluția de optim, deci setul de parametri care asigură o cât mai bună aproximare a dependenței considerate.

Modelul de abordare a problemei propuse este prezentat în fig.4.11.

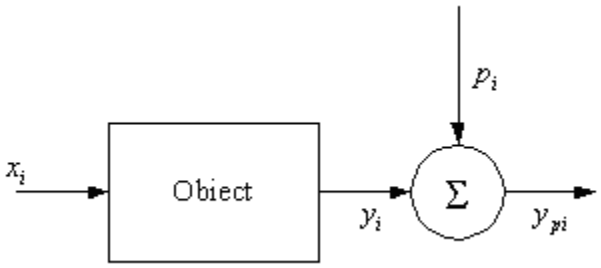


Fig 4.11. Modelul de abordare a problemei

Considerăm funcția de aproximare în forma:

$$y_i \approx f(x_i, m_1, m_2, ...m_q) \qquad (4.9)$$

şi prin urmare eroarea reziduală se obține în forma:

$$\varepsilon_i = y_{pi} - f(x_i, m_1, ...m_n) \qquad (4.10)$$

Impunem criteriul de calitate:

$$J(m_1, m_2, ...m_n) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_{pi} - f(x_i, m_1, ...m_n \right)^2 \qquad (4.11)$$

Condițiile necesare de minim pentru criteriul propus sunt:

$$\frac{\partial}{\partial m_i} \left(J(m_1, m_2, ...m_n) \right) = 0 \qquad (4.12)$$

Soluționând sistemul de ecuații astfel obținut obținem parametrii ce caracterizează complet funcția de predicție. Programul *cftool* realizează automat determinarea parametrilor optimi pentru o alegere prealabilă a tipului de funcție de predicție. De asemenea avem posibilitatea evidențierii erorilor reziduale în formă grafică.

4.2.3 Modul de lucru

Considerăm că semnalul de intrare este un semnal aleatoriu, cu distribuție normală şi medie nenulă. Semnalul real măsurat conține o componentă în legătură direct cauzală cu mărimea de intrare, peste care se aplică aditiv o mărime aleatoare cu distribuție normală şi medie nenulă. În acest scop introducem de la tastatură liniile de comandă:

```
>> X=30*randn(1,100)+50;
>> Y=0.0001*(2*X.^3+14*X.^2)+20+10*20*randn(1,100);
```

Cu comanda *cftool* deschidem fereastra de lucru şi procedăm la introducerea datelor inițiale exact ca în *lucrarea 1*.

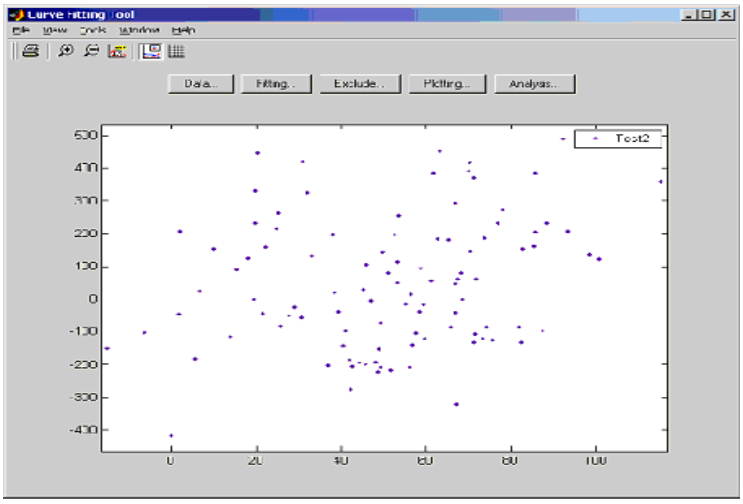


Fig 4.12. Graficul de distribuție a datelor inițiale

În fereastra de lucru *Curve Fitting Tool* este prezentat graficul de împrăştiere a datelor de intrare. Revenind în fereastra *Data* selectăm *Smooth* pentru a realiza o preprocesare a datelor de intrare. În această fereastră selectăm:

- *Original Data Set— Test2* (selectare implicită în ipoteza în care testul original a fost definit *Test2)*
- *Smoothed Data Set— Test3*
- *Method— Moving Average* (se impune metoda mediilor alunecătoare ca metodă de filtrare).
- *Span— 3* (procesul de mediere se face pe ultimele trei date).

Fereastra de lucru astfel completată este prezentată în fig.4.13.

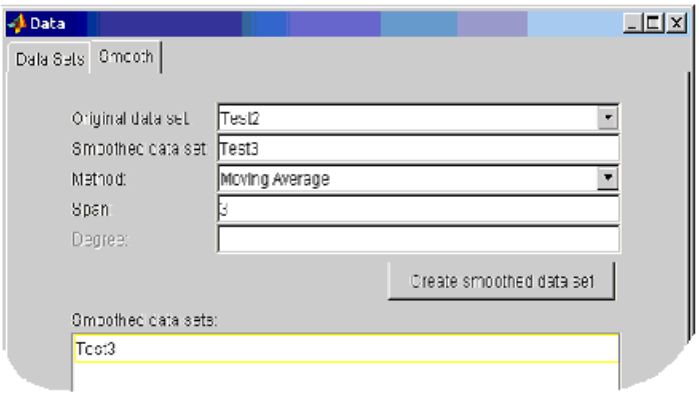


Fig.4.13. Fereastra de setare a datelor de preprocesare

Comanda *Create smoothed data set* lansează algoritmul de filtrare şi deschide fereastra grafică *Curve Fitting Tool* în care

sunt prezentate grafic atât datele inițiale cât şi datele preprocesate

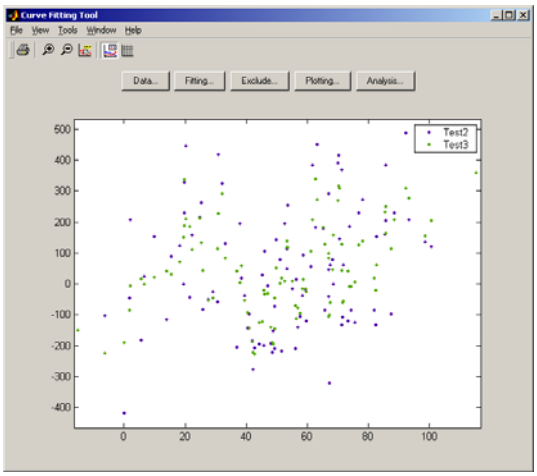


Fig.4.14. Graficul datelor preprocesate

În fereastra *Curve Fitting Tool* selectăm *Fitting* lansând fereastra de procesare *Fitting* prezentată în fig.4.15.

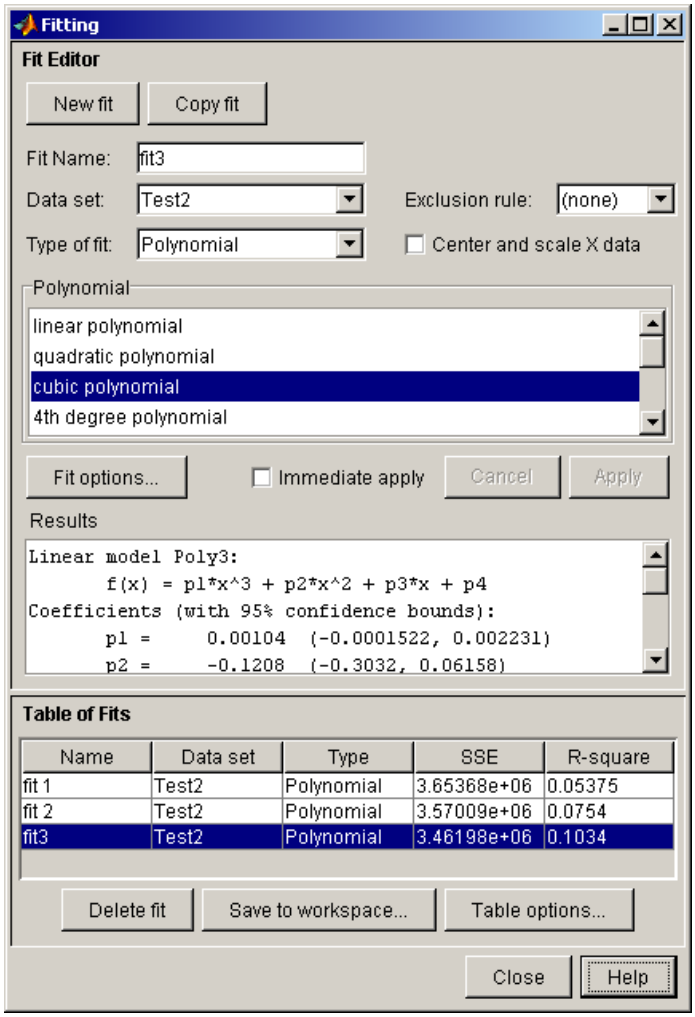


Fig 4.15. Fereastra de setare a parametrilor funcției de predicție

În această fereastră impunem tipul de funcție ales pentru predicție şi lansăm procesarea selectând *Apply*. . Pentru aplicația considerată am ales:

- *Fit New —Polynomial—linear polynomial*
- *Fit New —Polynomial—quadratic polynomial*
- *Fit New —Polynomial—cubic polynomial*

În fereastra de lucru *Curve Fitting Tool* selectăm *View—Residuals—Line Plot* pentru vizualizarea erorilor reziduale.

Închiderea ferestrei de procesare *Fitting* permite evaluarea grafică a datelor inițiale, a datelor preprocesate, a curbelor de variație pentru diferitele funcții de predicție precum şi a erorilor reziduale (vezi fig.4.16).

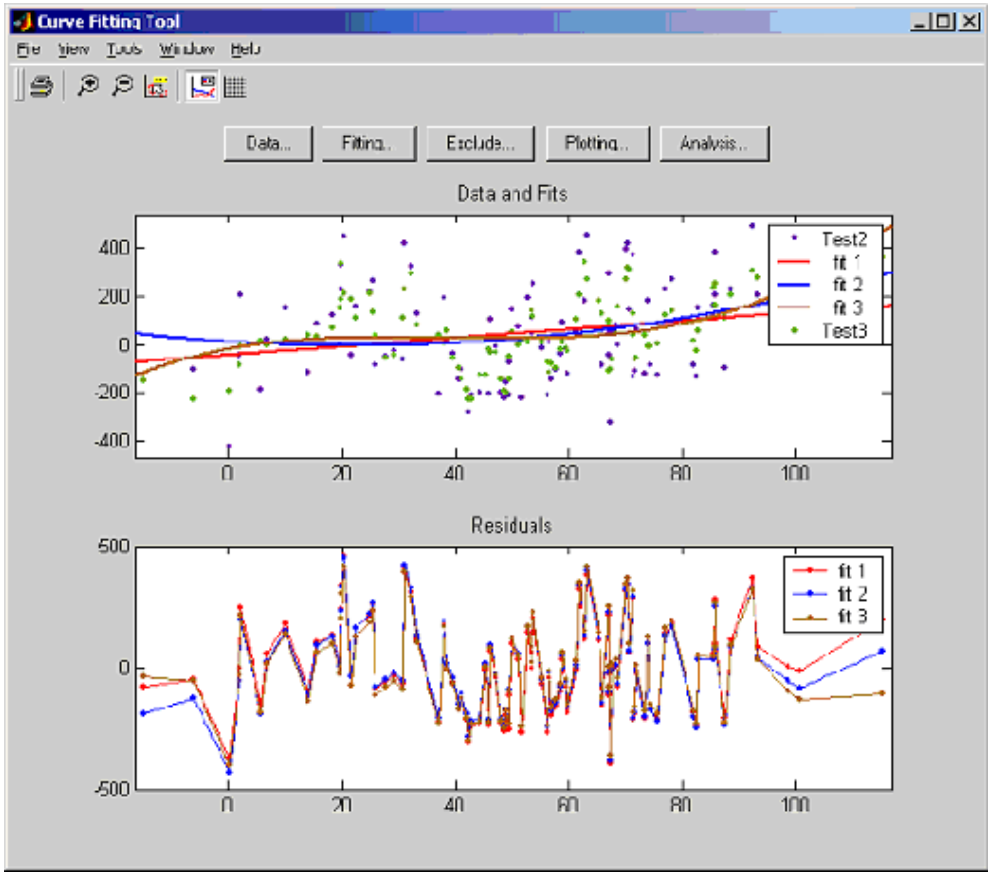


Fig .4.16. Fereastra de prezentare a rezultatelor finale

Pe lângă coeficienții optimi pentru diverse funcții de predicție în partea inferioară a ferestrei de lucru sunt prezentați indicatori statistici asociați fiecărei predicții pe baza cărora pot face selecția unei cele mai bune predicții.

Chestiuni de studiat

Se cere stabilirea unei cele mai bune funcții de predicție asociată datelor impuse, justificând cantitativ alegerea făcută.