PRELUCRARE DE DATE

ANALIZA STISTICA

Mircea V. Rusu 2005

Analiza statistica a datelor

- Cele mai simple mărimi statistice asociate unor măsurători succesive asupra unui sistem dat si care se fac *empiric*, neţinând cont de nici o informaţie asupra datelor in afara presupunerii ca aceste date se refera la un sistem care este staţionar, adică nu se modifica in intervalul efectuării celor *n* măsurători, sunt:
- *Valoarea medie*:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3)$$

- *Abaterea medie absoluta* care poate fi doar pozitiva:

$$AMA = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n} \qquad (4)$$

- *Abaterea medie* care poate fi pozitiva sau negativa:

$$AM = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} \qquad (5)$$

- *Abaterea medie pătratica* sau *varianta*:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (6)$$

- *Abaterea standard:*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (7)$$

introducând valoarea medie din (3) obţinem:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i)^2 - \sum_{i=1}^{n}2x_i\bar{x} + \sum_{i=1}^{n}(\bar{x})^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}(\bar{x})^2\right]$$

$$\sigma^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] \qquad (8)$$

Aceasta ultima relaţie este convenabila deoarece sumele si pot fi calculate simplu cu un calculator de buzunar, sau de câtre programe de tip tabelar (*workshit*) cum ar fi EXCEL si apoi introduse direct in formula finala.

Masuratori

- Daca presupunem ca erorile efectuate la măsurători sunt pur aleatoare, ne având si componente sistematice, atunci cea mai simpla presupunere este ca ele se distribuie aleator in jurul unei valori medii (aritmetice) si ca ele sunt distribuite normal, Gaussian.
- Teoria lui Gauss privind “erorile” de măsurătoare presupune ca in condiţii normale valorile obţinute din determinări se distribuie egal probabil in jurul unei valori medii care se calculeaza din media aritmetica a datelor experimentale iar probabilitatea ca la o măsurătoare sa obţinem o valoare diferita de cea medie este cu atât mai puţin probabila cu cat valoarea ei diferă mai mult de valoarea medie.

Clopotul lui Gauss

- Gauss a dedus ca lege de distribuţie a probabilităţilor erorilor este simetrica fata de o valoare medie (*m*) si are forma denumita *clopotul lui Gauss*, care are expresia matematica:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (9)$$

unde s este abaterea standard, care poate fi calculata cu relaţia (8) iar *m* - media, este calculabila cu formula (3) daca măsurătorile efectuate corespund distribuţiei Gauss (9), care se mai numeşte si *distribuţie normala*. Aceasta corespunde unei distribuţii *continue* a *variabilei aleatoare x*. Variabila aleatoare poate fi si *discreta* atunci când se referea la valori exprimate de exemplu prin numere întregi (valorile care pot ieşi la aruncarea unui zar, la numerele de la Loto, sau la distribuţia numărului indivizilor intr-o populaţie data).

Regresia

- Daca acum se presupune ca datele trebuiesc corespund unui fenomen descris de o lege presupusa corecta (model sau teorie) si care se considera ca reprezintă bine fenomenele, atunci datele experimentale trebuie sa fie *corelate* astfel încât ele sa descrie funcţia teoretica considerata ca model. De exemplu daca datele trebuie sa fie descrise printr-o linie dreapta atunci alegem ecuaţia unei drepte ca fiind forma analitica care sa fie satisfăcuta cat mai bine de câtre datele experimentale. Aceasta se numeşte *fitare liniara*, sau *regresie liniara.*

- Termenul de regresie a fost introdus pentru prima data de catre **Francis Galton** in 1886, cel care este considerat unul dintre creatorii statisticii aplicate. El a observat statistic ca părinţii mai înalţi decât media tind sa aibă copii mai scunzi iar cei mai mici ca statura decât media, tind sa aibă copii de statura mai mare. Observaţia a fost interpretata ca un fel de “regresie” către medie si de aici termenul a fost folosit pentru găsirea funcţiei care sa descrie modul de variaţie, sau tendinţa ce se poate observa intr-un set de măsurători.

Regresia

- Daca curba de regresie este un polinom de grad mai mare ca unu sau o funcţie neliniara (trigonometrica, logaritmica, exponenţiala, etc) atunci procedeul de aflare a coeficienţilor prin minimizarea erorii globale se numeşte *regresie neliniara.* Daca funcţia care trebuie sa fie găsita este dependenta de mai multe variabile independente, atunci ea se numeşte *regresie multipla.*
- Minimizarea erorii globale prin alegerea potrivita a coeficienţilor de regresie se realizeaza printr-un procedeu ce utilizează proprietatea care spune ca o funcţie prezintă un maxim sau un minim (un *extremum*) daca in acele punct derivata se anulează.

Cazul general.

Este cazul in care se căuta sa se minimizeze eroarea aflării coeficienţilor unei funcţii de mai multe variabile de tipul:

$$f_i = \varphi_i(x_1, x_2, x_3, x_4, ..., x_n) \quad \text{cu } i = 1,2,3,4, \quad ,m, \qquad (1)$$

când se cunosc o serie de valori (experimentale) ale variabilelor independente $\{x_i\}$.

Daca funcţia este de forma polinomiala:

$$\varphi_i(x_1, x_2, x_3, x_4, ..., x_n) = a_i x_1 + b_i x_2 + c_i x_3 + d_i x_4 + ... + l_i x_n \qquad (2)$$

coeficienţii care se căuta sunt . Forma polinomiala (2) arata ca funcţia este de gradul întâi si deci regresia este *liniara* dar multipla; ea nu este *neliniara*, caz in care variabilele independente ce intervin ar fi la o putere mai mare ca 1.

Acest sistem de *m* ecuaţii este **supradeterminat** atunci când *m* > *n*. Când *m* = *n* avem un sistem de ecuaţii obişnuite, iar daca *m* < *n* el este **subdeterminat** si nu are o soluţie.

Pentru un sistem supradeterminat ce porneşte de la nişte date experimentale, este aproape sigur ca nu exista o soluţie unica pentru ecuaţiile (1). Se cauta deci un criteriu si o metoda care sa permită determinarea *celor mai „buni" coeficienţi* pentru (2). Ce înseamnă „cei mai buni", rezulta din condiţiile pe care noi le luam in consideraţie, sau le acceptam. Deci pot exista mai multe metode de determinare care sa nu fie echivalente si sa nu conducă la un set identic de coeficienţi.

Regresia liniara

Ca prim exemplu vom descrie *regresia liniara* dând toate relaţiile corespunzătoare de calcul.

Pentru un set de *N* date experimentale de forma yi={xi} care se presupun a fi liniar corelate se poate căuta o funcţie de fitare de forma:

$$y(x)=a + bx \qquad (10)$$

Pentru a obţine valoarile cele mai bune pentru coeficientii regresiei, se foloseste functia funcţia $\chi 2$ care este considerata o masura a erorii globale fata de o funcţie de referinţa, in cazul de fata ecuaţia de gradul întâi (10):

$$\chi^2 = \sum_i \left[\frac{1}{\sigma_i^2} (y_i - a - bx_i)^2 \right] \qquad (11)$$

Procedura de fitare presupune minimizarea funcţiei $\chi 2$ simultan in raport cu fiecare coeficient. Coeficienţii se afla din rezolvarea sistemului rezultat si au expresiile:

Aflarea coeficienţilor

- Procedura de fitare presupune minimizarea funcţiei χ^2 simultan in raport cu fiecare coeficient. Coeficienţii se afla din rezolvarea sistemului rezultat si au expresiile:

$$a = \frac{1}{\Delta}\left(\sum\frac{x_i^2}{\sigma_i^2}\sum\frac{y_i}{\sigma_i^2} - \sum\frac{x_i}{\sigma_i^2}\sum\frac{x_i\,y_i}{\sigma_i^2}\right)$$

$$b = \frac{1}{\Delta}\left(\sum\frac{1}{\sigma_i^2}\sum\frac{x_i\,y_i}{\sigma_i^2} - \sum\frac{x_i^2}{\sigma_i^2}\sum\frac{y_i}{\sigma_i^2}\right) \qquad (12)$$

$$\Delta = \left(\sum\frac{1}{\sigma_i^2}\sum\frac{x_i^2}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right)$$

Erori ale coeficientilor

Estimarea deviatei standard se face conform relaţiei

$$\sigma^2 \cong s^2 = \frac{1}{N-2}\sum (y_i - a - bx_i)^2 \qquad (13)$$

Estimarea erorilor coeficienţilor de regresie se face cu formulele:

$$\sigma_a^2 \cong \frac{\sigma^2}{\Delta}\sum x_i^2 \qquad (14)$$

$$\sigma_b^2 \cong \frac{N}{\Delta}\sigma^2 \qquad (15)$$

P.R.Bevington, "Data reduction and error analysis for the physical sciences", McGraw-Hill, 1969;

Coeficientul de corelaţie

Gradul de a acestor date experimentale se poate estima folosind expresia:

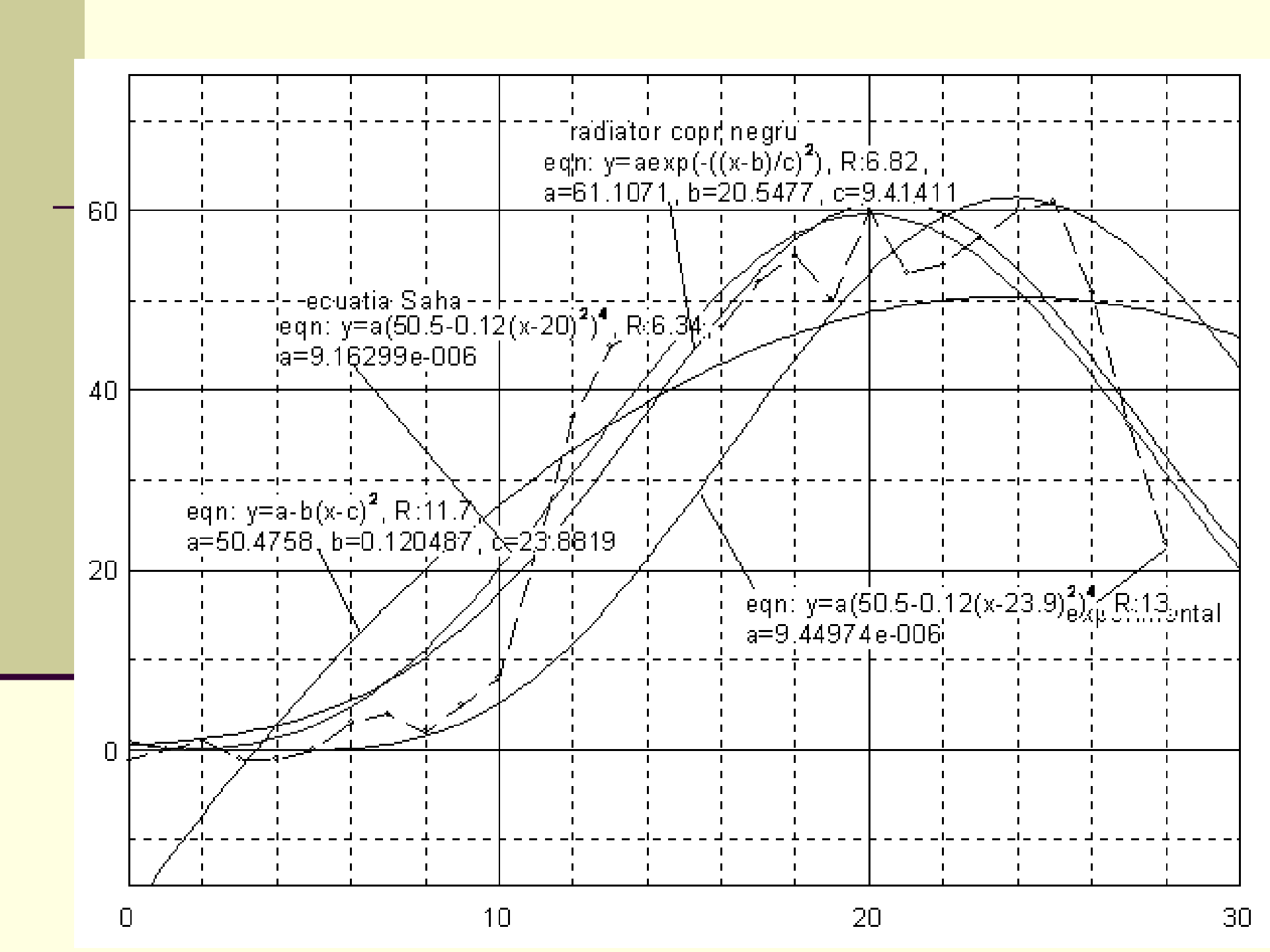
$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\left[N\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}\left[N\sum y_i^2 - \left(\sum y_i\right)^2\right]^{1/2}}$$

(16)

Alungirea unui elastic

Transmisia caldurii

Analiza unei linii spectrale


radiator copr negru
eqn: y=aexp(-((x-b)/c)^2), R:6.82,
a=61.1071, b=20.5477, c=9.41411
ecuatia Saha
eqn: y=a(50.5-0.12(x-20)^2)^4, R:6.34,
a=9.16299e-006
eqn: y=a-b(x-c)^2, R:11.7,
a=50.4758, b=0.120487, c=23.8819
eqn: y=a(50.5-0.12(x-23.9)^2)^4, R:13
a=9.44974e-006
60
40
20
0
0
10
20
30

Legea lui Kepler, sistem solar

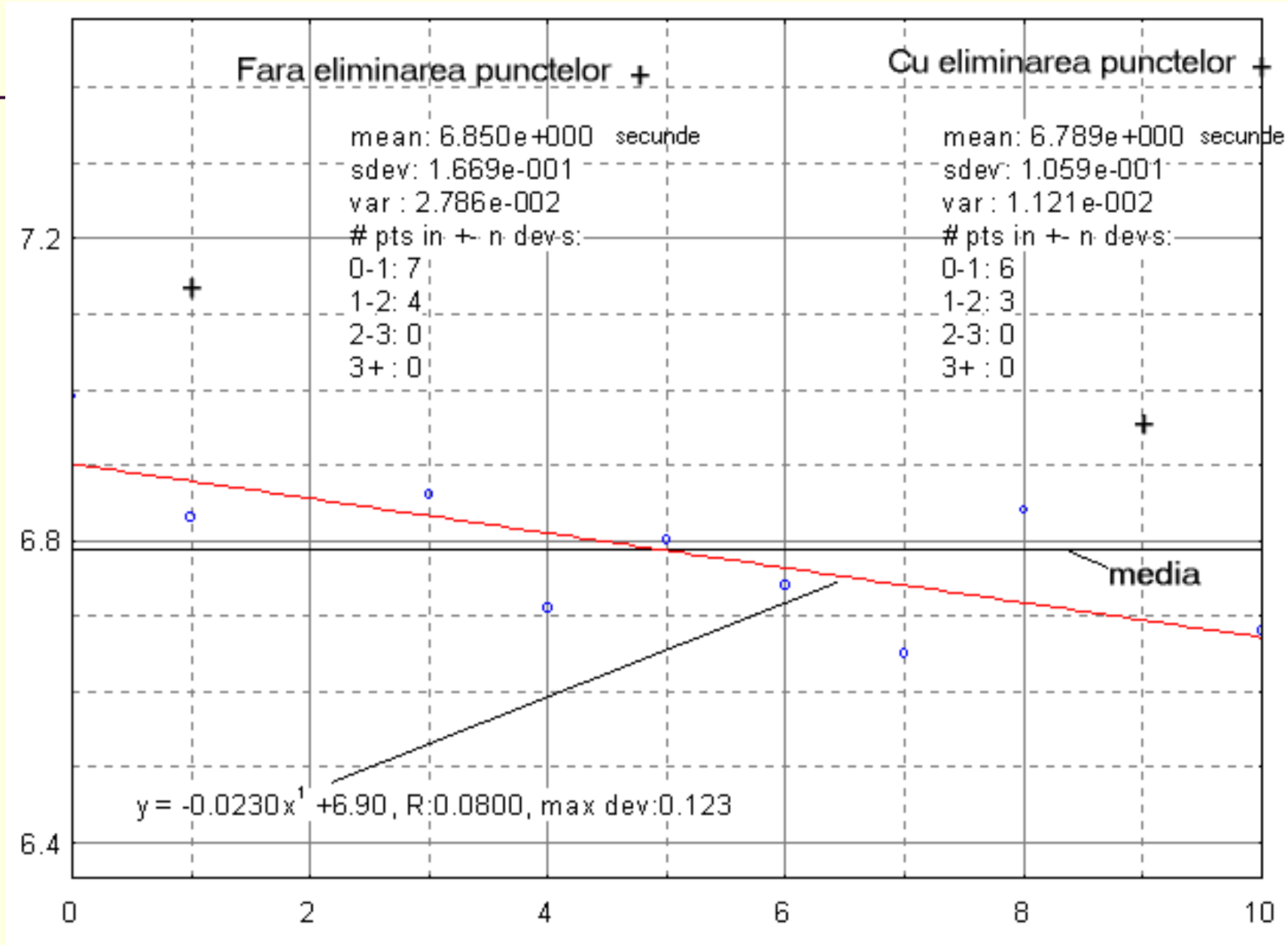
Curbe de lumina asteroizi

Analiza de semnal

Difractie

Ocultatie stea dubla

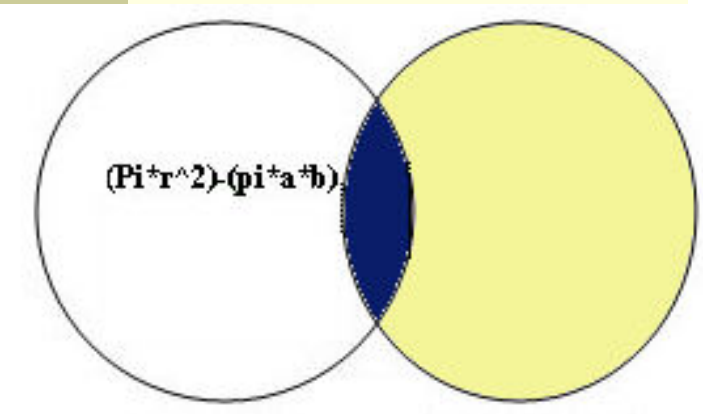
Anaiza datelor semnificative

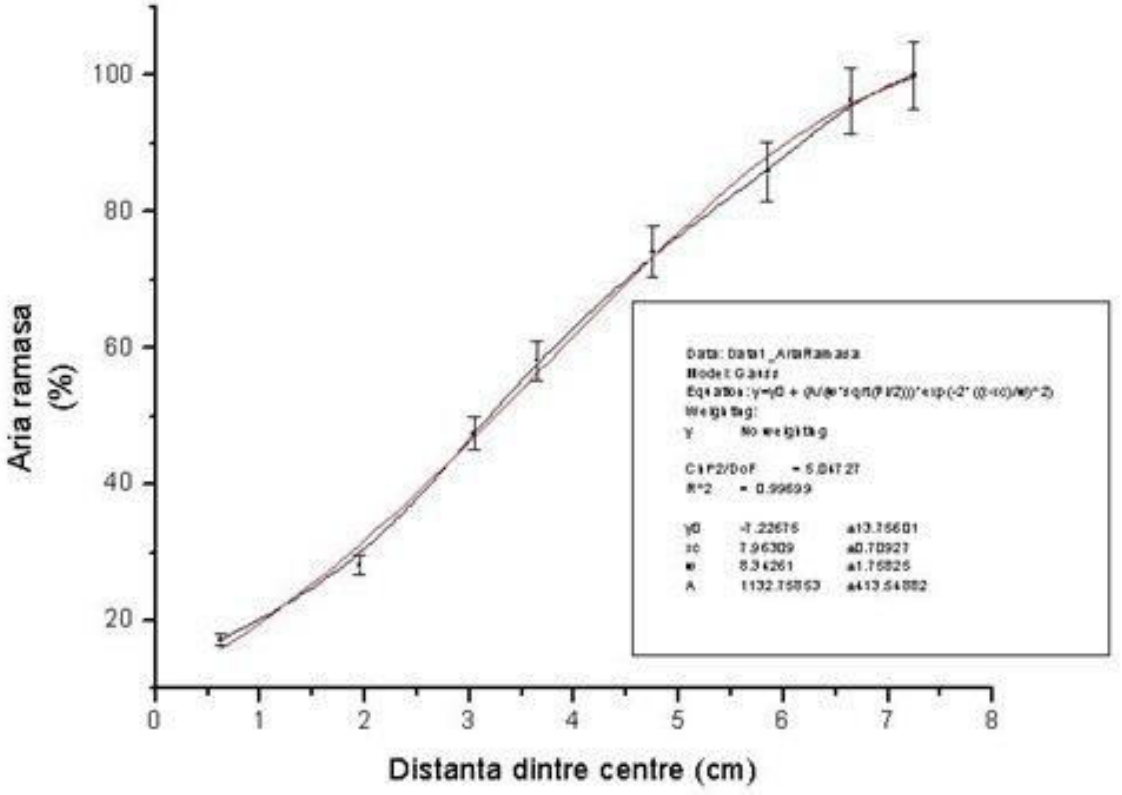


Anailza coroanei la eclipsa totala

Aproximatii

Distanta dintre centrele cercurilor (cm)	Aria neacoperita (%)	Eroare pentru aria neacoperita(%)	Aria acoperita	Eroare pentru aria acoperita
7,25	100	5	0	0
6,65	96,27	4,8135	3,73	0,1865
5,85	85,86	4,293	14,14	0,707
4,75	74,06	3,703	25,94	1,297
3,65	58,11	2,9055	41,89	2,0945
3,05	47,43	2,3715	52,57	2,6285
1,95	28,08	1,404	71,92	3,596
0,625	17,06	0,853	82,4	4,12





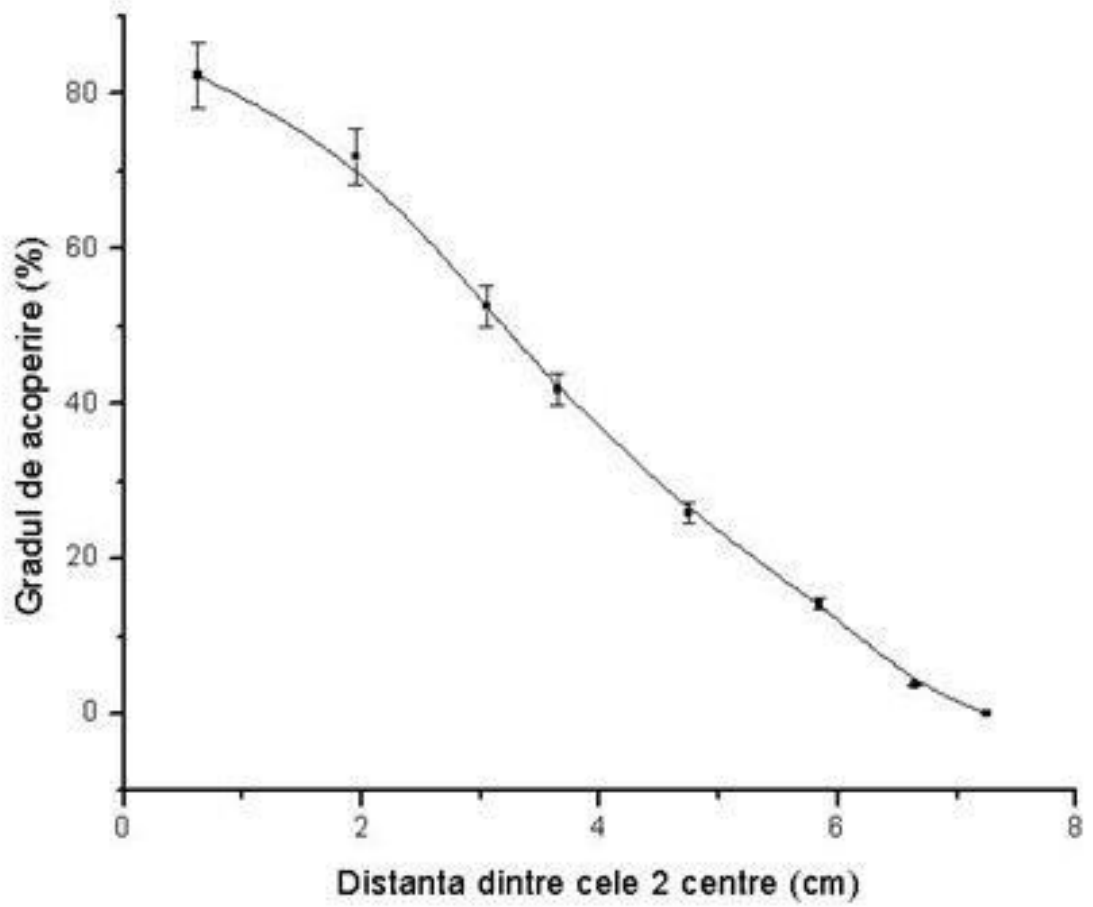
Aproximatii

Model: **Gauss**

Equation: **y=y0 + (A/(w*sqrt(PI/2)))*exp(-2*((x-xc)/w)^2)**

Chi^2/DoF = 5.04727
R^2 = 0.99699

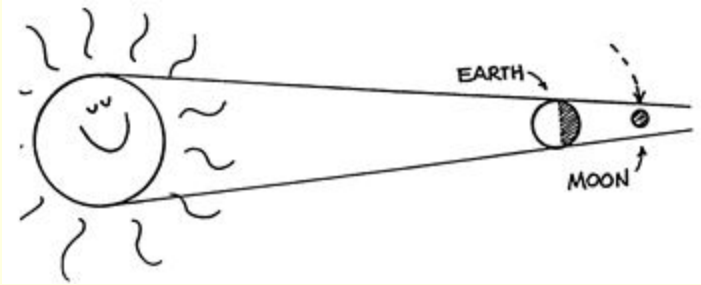
y0	-7.22675	13.75601
xc	7.96309	0.70927
w	8.34261	1.75825
A	1132.75853	413.54882

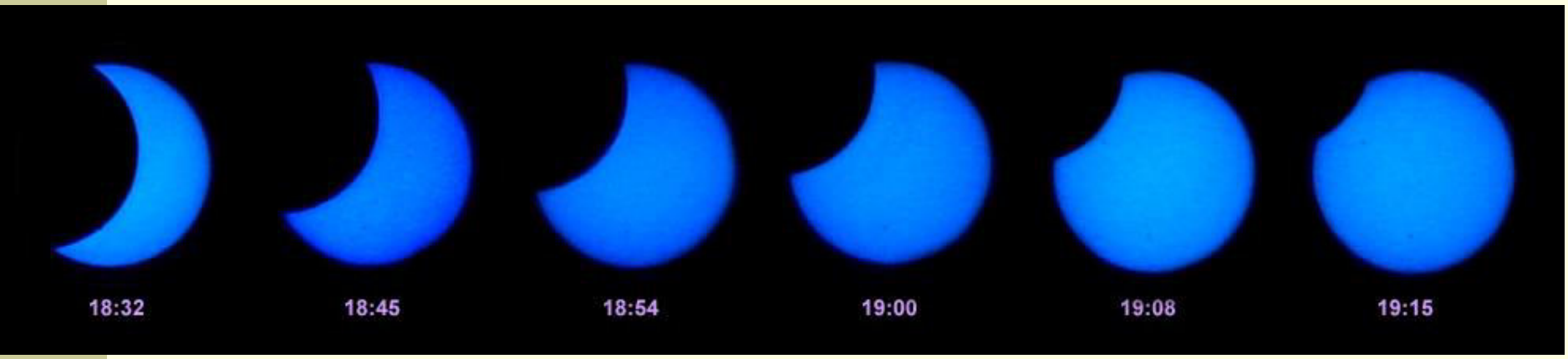


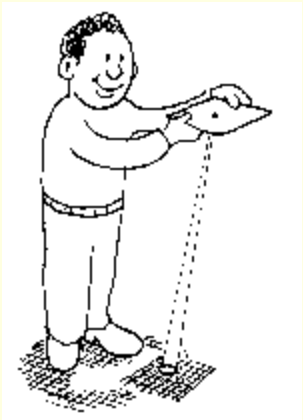
Exemplu: datele privind eclipsa de Soare din 2006

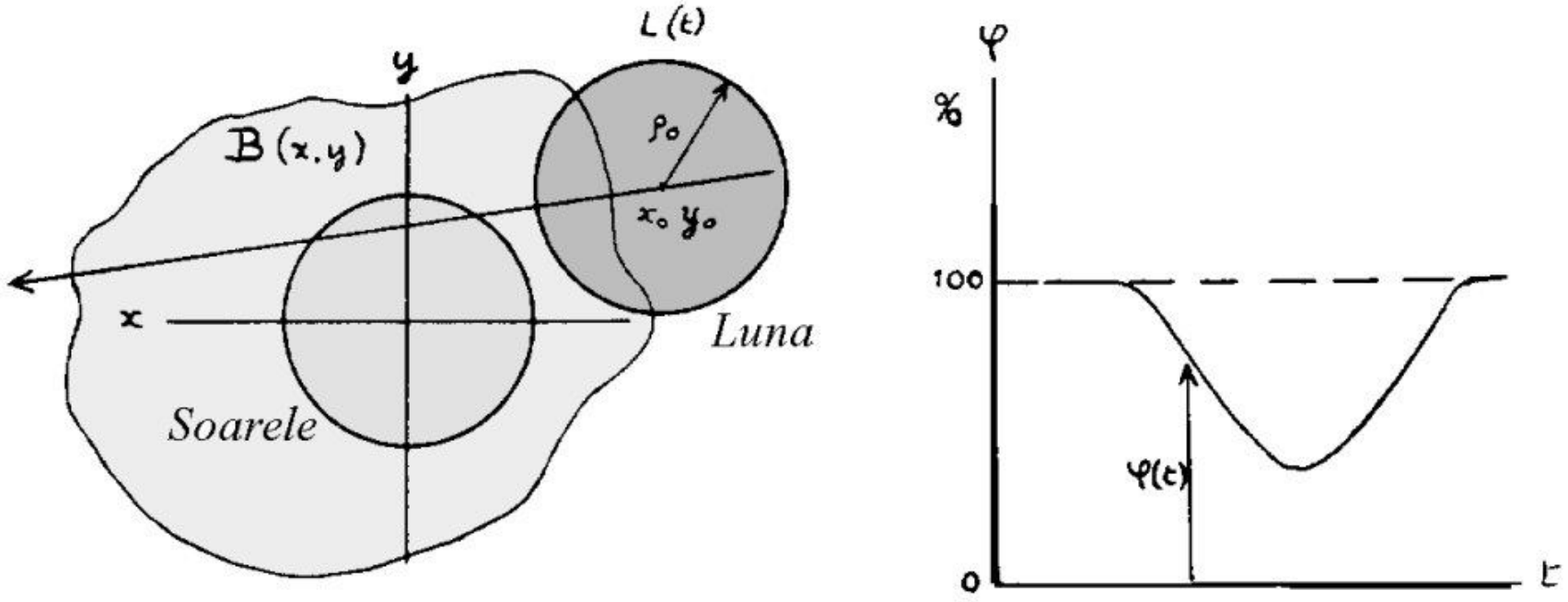


Eclipsele



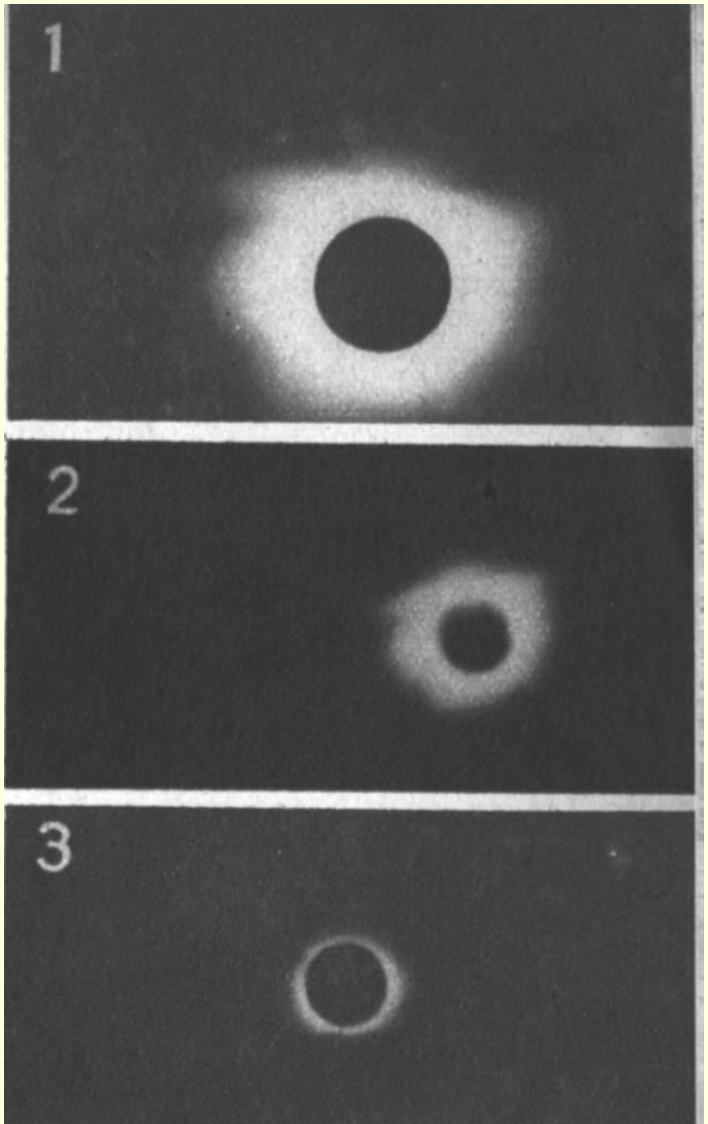


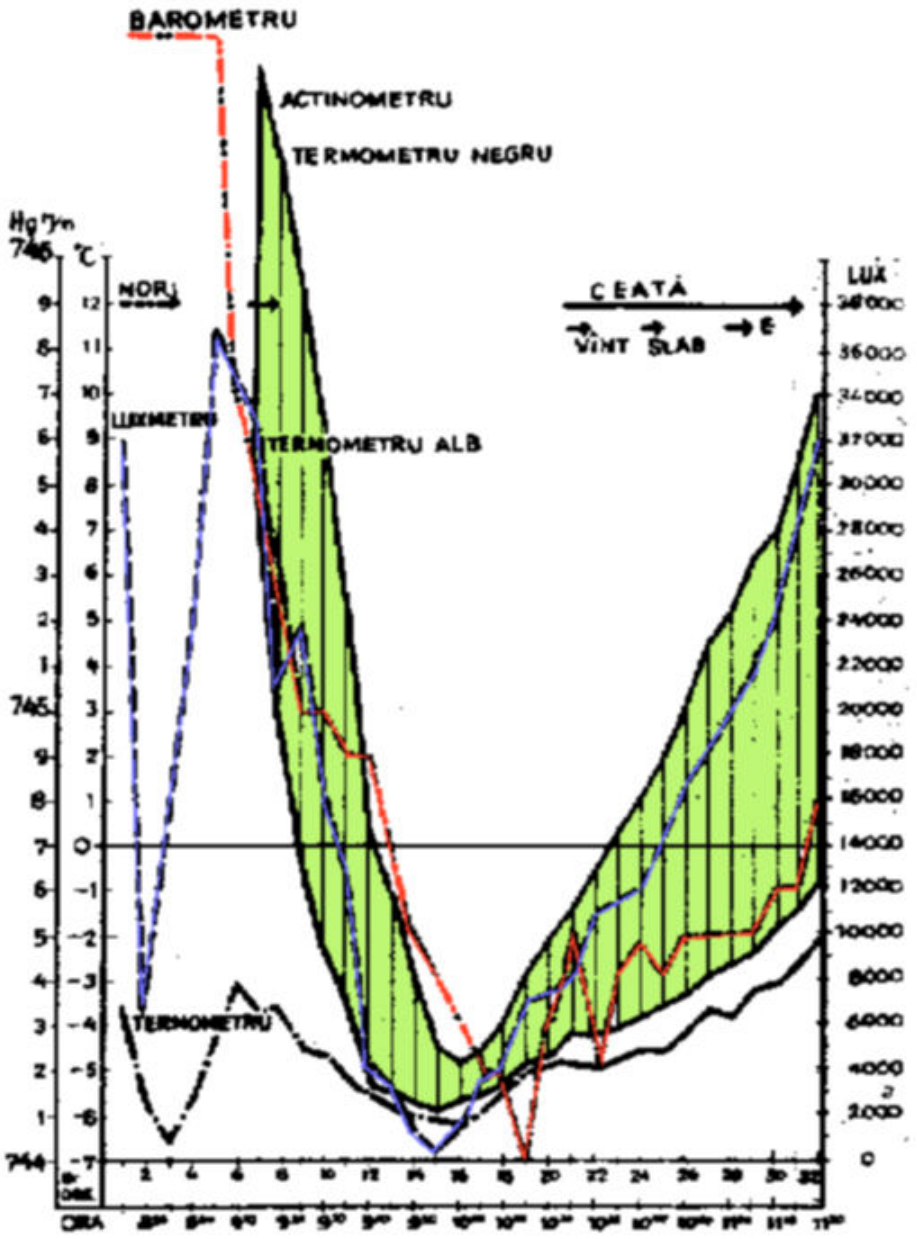




Eclipsa de Soare și curba de ocultație

1961
Bucuresti





1998 Caraibe








13:47:09


14:03:57


14:08:45


14:10:37


14:12:07


14:12:50


14:12:54

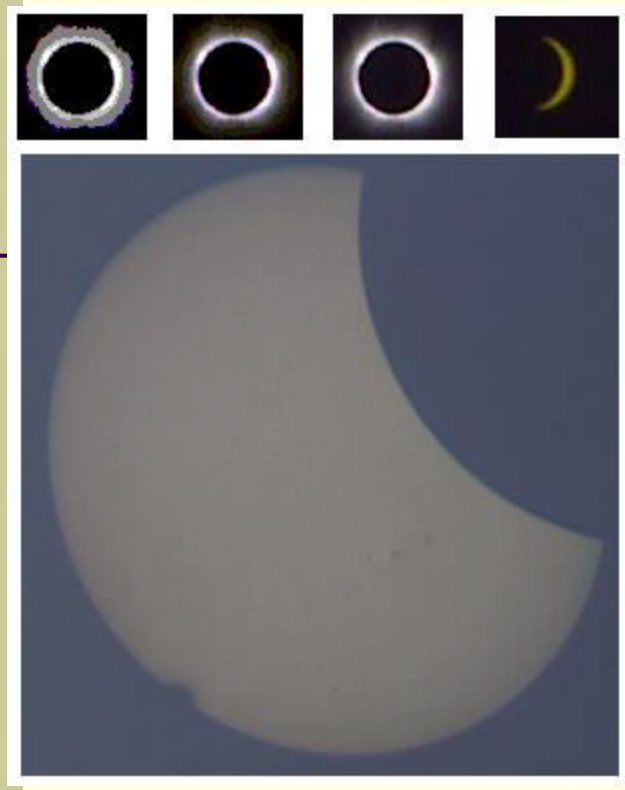

14:13:24





1999 Bucuresti

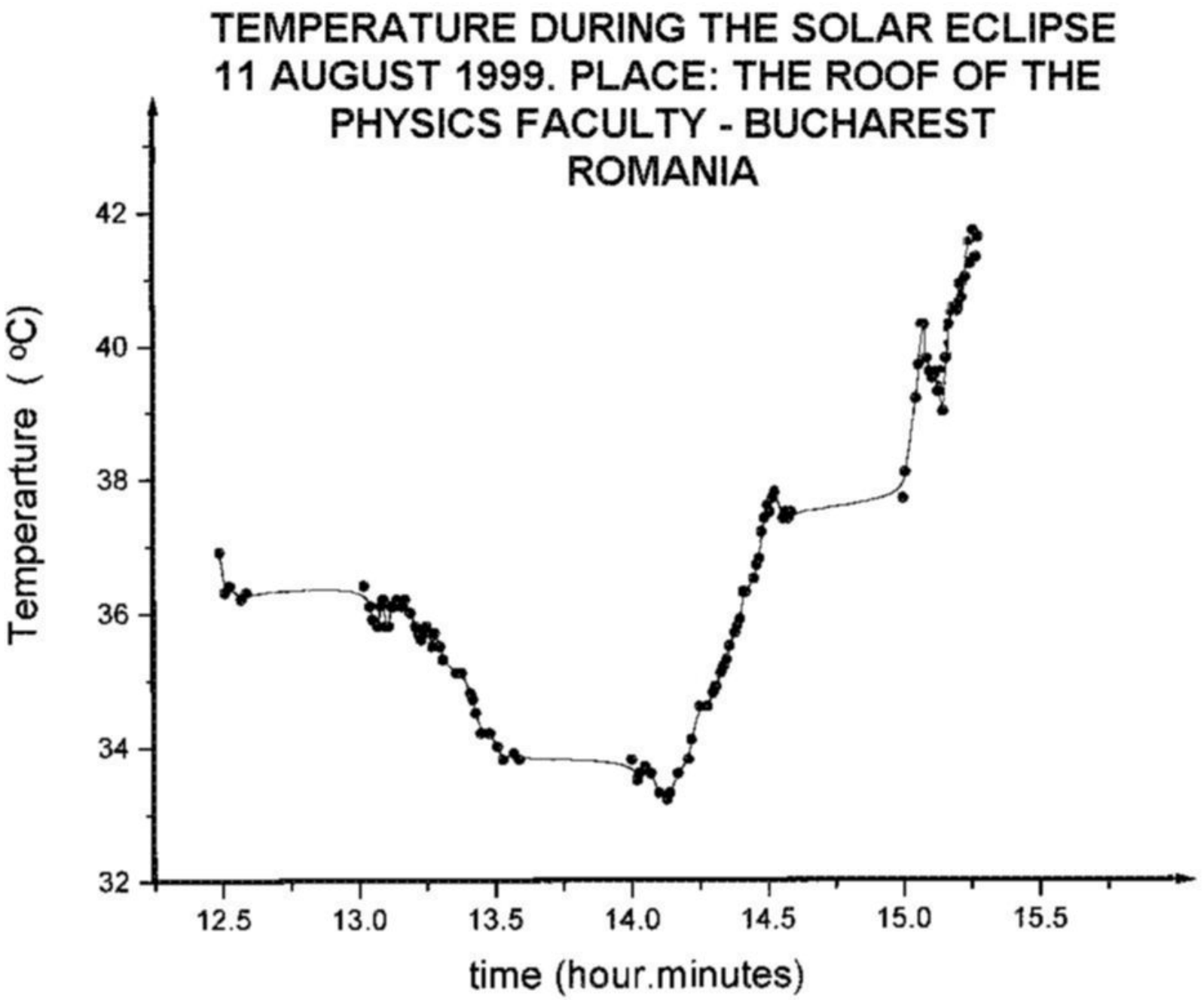








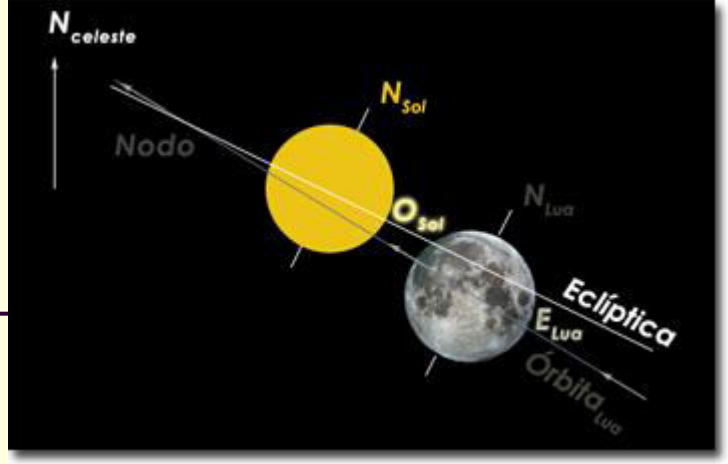

EARTH

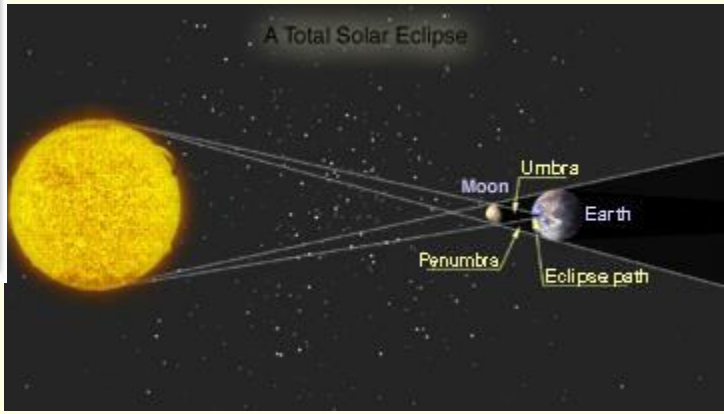

TEMPERATURE DURING THE SOLAR ECLIPSE
11 AUGUST 1999. PLACE: THE ROOF OF THE
PHYSICS FACULTY - BUCHAREST
ROMANIA
Temperarture (oC)
42
40
38
36
34
32
12.5
13.0
13.5
14.0
14.5
15.0
15.5
time (hour.minutes)

Ray Pashakoff 1999



2006


N celeste
N Sol
Nodo
O Sol
N Lua
Ecliptica
E Lua
Órbita Lua

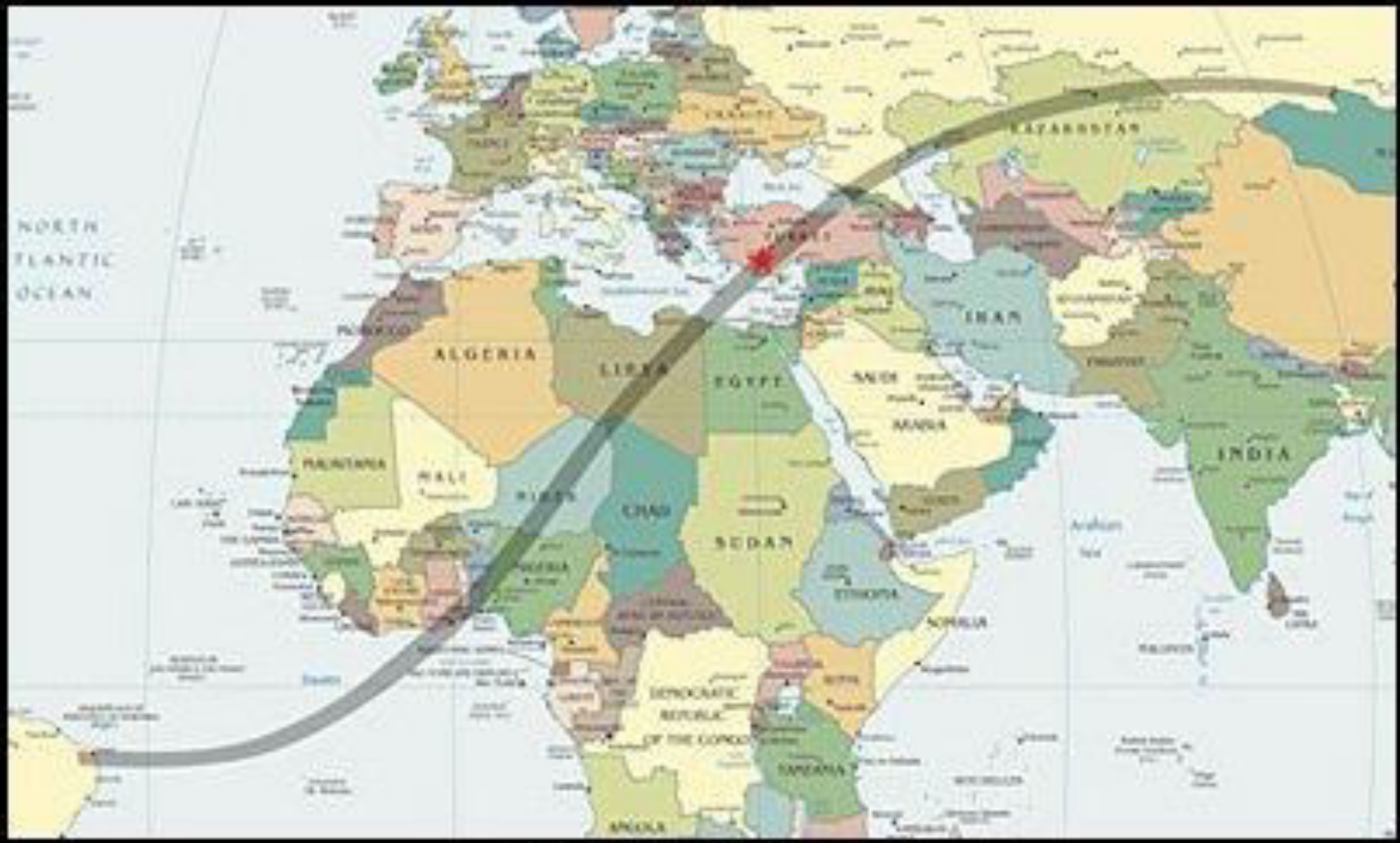

A Total Solar Eclipse
Umbra
Moon
Earth
Penumbra
Eclipse path





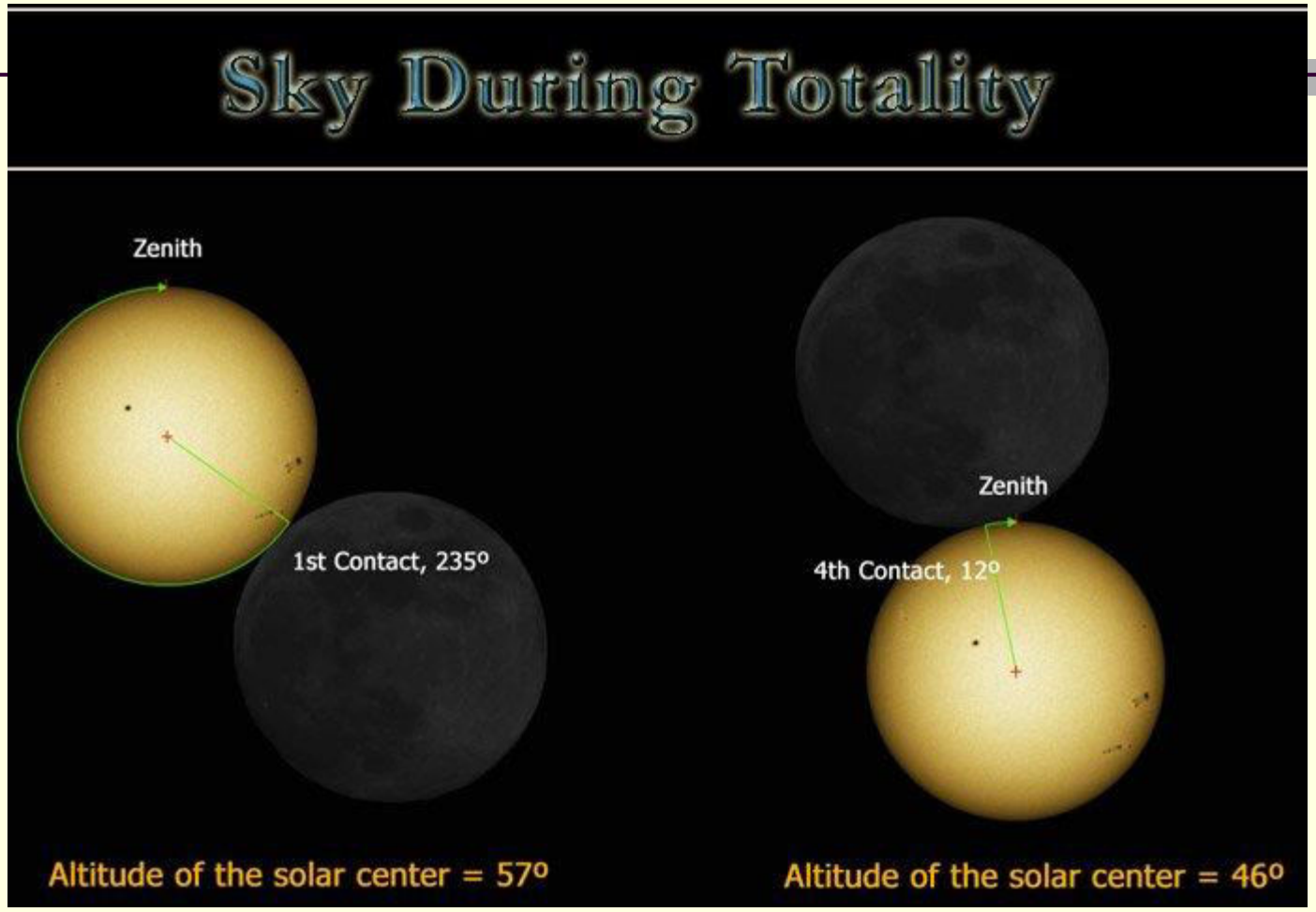

UKRAINE
Kharkiv
Volgograd
Vienna
Bratislava
Budapest
Chisinau
Rostov
Ljubljana
Zagreb
Belgrade
ROMANIA
Bucharest
ITALY
Sarajevo
Sofia
BULGARIA
Black Sea
Caspian Sea
Tbilisi
Rome
Naples
Skopje
Ankara
Yerevan
Baku
TURKEY
Palermo
Athens
Izmir
Adana
Tabriz
Valletta
Aleppo
Kirkuk
Tehran
Mediterranean Sea
Nicosia
SYRIA
IRAQ
Damascus
Beirut
Baghdad
Tripoli
Benghazi
Alexandria
Amman
JORDAN
Cairo
Al Basrah
Kuwait
Shiraz
LIBYA
EGYPT
SAUDI
Manama
Riyadh
Doha
Abu Dhabi
QATAR

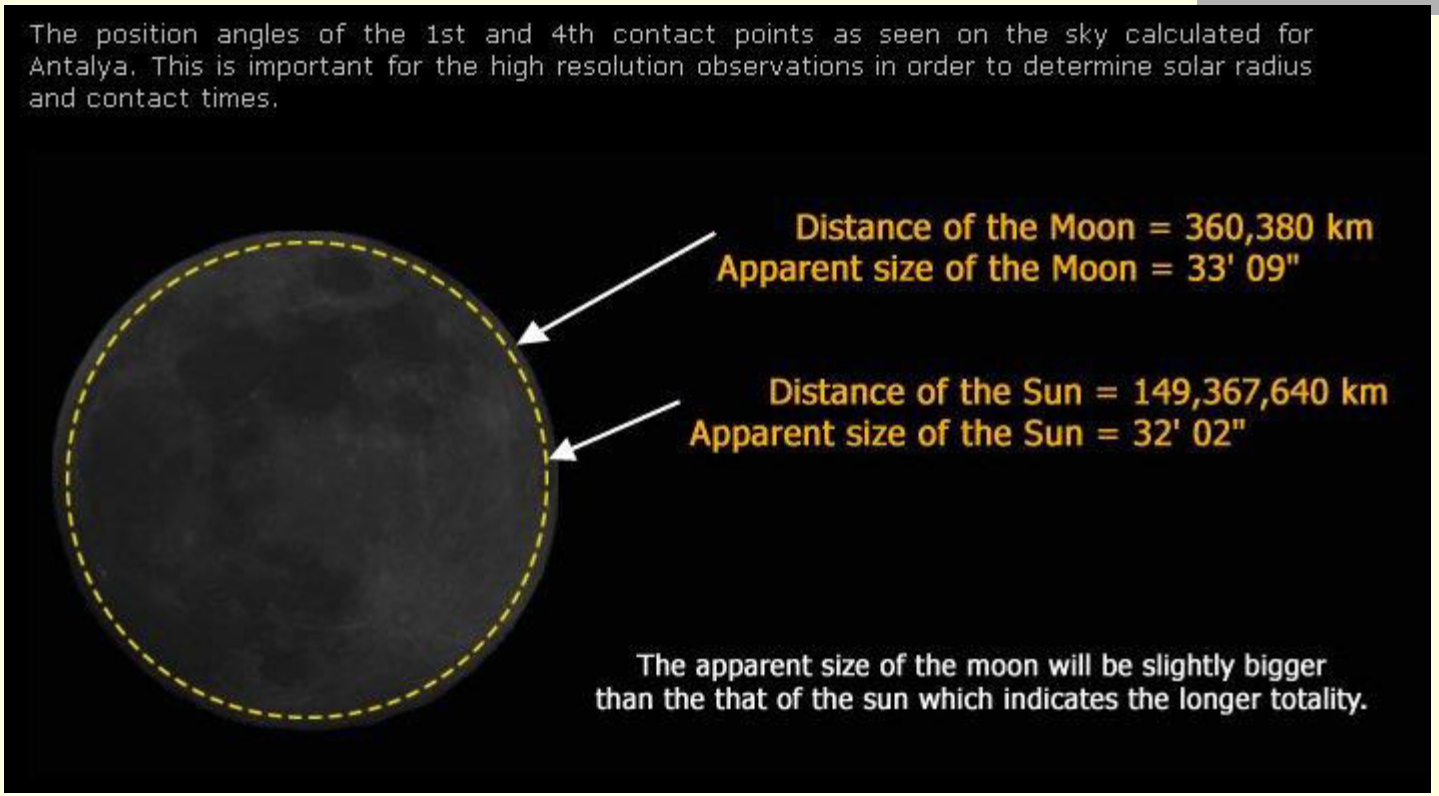
2006



The Path of Totality

Sky During Totality




The position angles of the 1st and 4th contact points as seen on the sky calculated for Antalya. This is important for the high resolution observations in order to determine solar radius and contact times.
Distance of the Moon = 360,380 km
Apparent size of the Moon = 33' 09"
Distance of the Sun = 149,367,640 km
Apparent size of the Sun = 32' 02"
The apparent size of the moon will be slightly bigger than the that of the sun which indicates the longer totality.

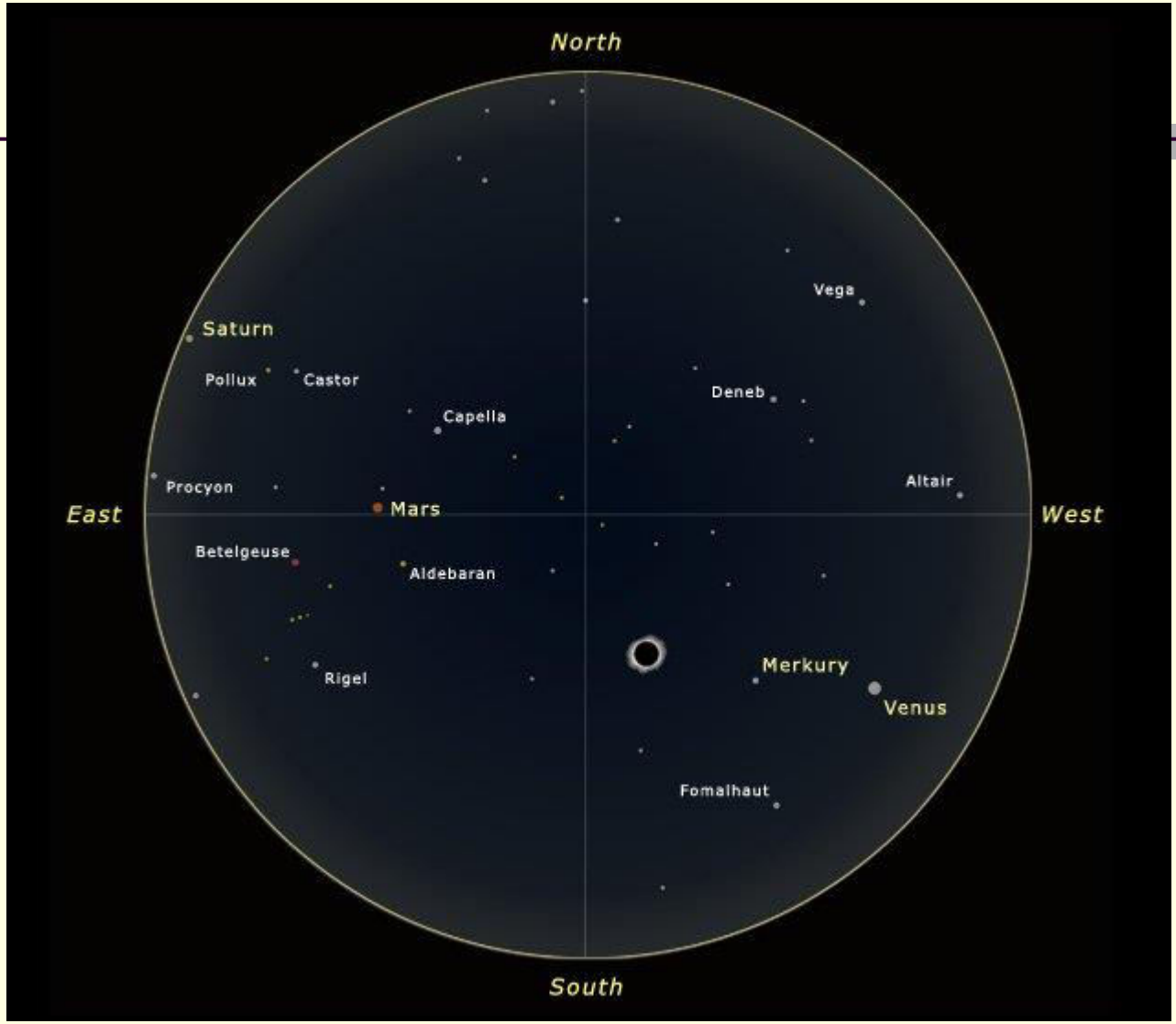

North
East
West
South
Saturn
Pollux
Castor
Capella
Procyon
Mars
Betelgeuse
Aldebaran
Rigel
Vega
Deneb
Altair
Merkury
Venus
Fomalhaut

Table 20 - Solar Eclipses of Saros Series 139

First Eclipse: 1501 May 17 Duration of Series: 1262.1 yrs.
Last Eclipse: 2763 Jul 03 Number of Eclipses: 71

Saros Summary: Partial: 16 Annular: 0 Total: 43 Hybrid: 12

Date	Eclipse Type	Gamma	Mag./ Width	Center Durat.
1501 May 17	Pb	1.500	0.091	
1519 May 28	P	1.418	0.235	
1537 Jun 07	P	1.337	0.380	
1555 Jun 19	P	1.254	0.530	
1573 Jun 29	P	1.172	0.678	
1591 Jul 20	P	1.091	0.826	
1609 Jul 30	P	1.014	0.966	
1627 Aug 11	H	0.940	2	00m01s
1645 Aug 21	H	0.871	28	00m16s
1663 Sep 01	H	0.807	38	00m29s
1681 Sep 12	H	0.750	43	00m40s
1699 Sep 23	H	0.700	46	00m49s
1717 Oct 04	H	0.656	47	00m56s
1735 Oct 16	H	0.620	48	01m02s
1753 Oct 26	H	0.591	49	01m08s
1771 Nov 06	H	0.567	50	01m13s
1789 Nov 17	H	0.550	52	01m19s
1807 Nov 29	H	0.538	55	01m26s
1825 Dec 09	H	0.530	60	01m34s
1843 Dec 21	T	0.523	66	01m43s
1861 Dec 31	T	0.519	74	01m55s
1880 Jan 11	T	0.514	84	02m07s
1898 Jan 22	T	0.508	96	02m21s
1916 Feb 03	T	0.499	108	02m36s
1934 Feb 14	T	0.487	123	02m52s
1952 Feb 25	T	0.470	138	03m09s
1970 Mar 07	T	0.447	153	03m28s
1988 Mar 18	T	0.419	169	03m46s
2006 Mar 29	T	0.384	183	04m07s
2024 Apr 08	T	0.343	197	04m28s
2042 Apr 20	T	0.296	210	04m51s
2060 Apr 30	T	0.242	222	05m15s
2078 May 11	T	0.184	232	05m40s
2096 May 22	T	0.120	241	06m06s
2114 Jun 03	T	0.053	248	06m32s
2132 Jun 13	Tm	-0.018	255	06m55s
2150 Jun 25	T	-0.091	260	07m14s
2168 Jul 05	T	-0.166	264	07m26s
2186 Jul 16	T	-0.239	267	07m29s
2204 Jul 27	T	-0.313	269	07m22s
2222 Aug 08	T	-0.383	270	07m06s
2240 Aug 18	T	-0.452	270	06m40s
2258 Aug 29	T	-0.516	269	06m09s
2276 Sep 09	T	-0.575	266	05m33s
2294 Sep 20	T	-0.630	263	04m57s
2312 Oct 01	T	-0.678	257	04m20s
2330 Oct 13	T	-0.720	251	03m46s
2348 Oct 23	T	-0.756	242	03m14s
2366 Nov 03	T	-0.786	231	02m47s
2384 Nov 14	T	-0.810	217	02m22s

Eclipse Type:
- P - Partial
- A- - Non-central Annular
- T - Total
- H - Hybrid (Annular/Total)
- Pb - Partial Eclipse (Saros Series Begins)
- Pe - Partial Eclipse (Saros Series Ends)
- Tm - Middle Eclipse of Saros series.
- Tn - Total Eclipse (no northern limit).

Note: Mag./Width column gives either the eclipse magnitude (for partial eclipses) or the umbral path width in kilometers (for total and annular eclipses).

Re



Picture from the "Book of the Dead". One of the two figures with an orange disk is Re-Harakhti, the other may be Re. The other two figures are pharoahs.

Eclipse.exe