5. Патент Великобритании. №1042183. Кл. H1 D

6. *Козин Л.Ф., Нигметова Р.Ш., Дергачова М.Б.* Термодинамика бинарных амальгамных систем Алма-Ата, 1977.

7. *Уэймаус Д.* Газоразрядные лампы. М., 1977.

8. *Федоренко А.С.* Экспериментальные исследования, расчетное моделирование плазмы положительного столба и создание перспективных конструкторских и технологических решений в области люминесцентных ламп низкого давления: Дис. докт. техн. наук. Саранск, Всесоюзн. научно-иссл., проектно-констр. и технологич. инст. им. А.Н. Лодыгина, 1990.

Поступила 25.01.2000

Summary

The usage of Indium amalgam and special mixture of noble gases makes possible to increase the power of the low-pressure mercury discharge up to 200 watts with high efficiency. The dependence of the efficiency of the lamp with such discharge is weak in the wide range of the lamp wall temperatures. It's shown the possibility to ignite 180-watt amalgam lamp in starter scheme with glow discharge starter.

C. Dochitoiu, N. Jula

A STATISTICAL WAY TO FIND CAUSAL FACTORS

Technical Military Academy,
Bd. George Cosbuc, 83–85, sector 5, Bucharest, Romania

Introduction

There are many cases, the influence of one or more factors on a certain process is obviously perceptible but this influence can't be found out on a deductive basis. The statistical tool provides various types of experimental data processing, which not only notice such influences but also even evaluate them.

Such an approach is very successful in many areas, for example the military, economy, sociology, health, ecology, etc.

The statistical tool used in this article is the correlation coefficient applied to empirical data considered as random sample variables. The elaborated method allows pointing out the cumulative influence of more factors and their weight using some correlation matrix properties.

Correlation coefficient

For every random variable X we note (as in [1], [3]) $M(X)$ and $\sigma(X)$ their mean value and standard deviation of X and Y respectively.

For the variables X ,Y the number $\rho(X,Y)=\dfrac{M[(X-M(X))(Y-M(Y))]}{\sigma(X).\sigma(Y)}$ is *the correlation coefficient* of X and Y . It has the property: $-1 \le \rho(X.Y) \le 1$ and $\rho(X,Y) = \pm 1$ if and only if the relation $Y-M(Y)=\pm\dfrac{\sigma(Y)}{\sigma(X)}\left(X-M(X)\right)$ is a sure event. Generally we think every relation between random variables as a sure event.

In other words, if $\rho(X, Y)$ reaches values ± 1 it may be concluded the linear dependence of "factors" X and Y and moreover it is provided the factor $\pm\dfrac{\sigma(Y)}{\sigma(X)}$ of this dependence.

© Dochitoiu C., Jula N., Электронная обработка материалов, 2000, № 2, С. 51–55.

Generally the experimental data are not in the case *ρ(X, Y) = ± 1* but rather *ρ(X, Y)* have values more or less close to *± 1*. For summary conclusions on the linear dependence of factors one considers the following three cases: a) $|\rho(X,Y)|$ ≥ *0,66*. In this case *X* and *Y* are considered strong correlated; b) *O, 33* < $|\rho(X,Y)|$ < *0,66*. The *X* and *Y* are said correlated; c) $|\rho(X,Y)|$ ≤ *0,33*. In this case *X* and *Y* are considered weakly correlated or rather uncorrelated.

Correlation matrix

Let $X_1, X_2, ..., X_n$ be random variables, $\rho_{ij} = \rho(X_i, X_j)$ and $A = (\rho_{ij})$ their correlation matrix. It is a symmetric matrix consisting of number in *[-1, 1]*.

Using the correlation coefficient properties, [2] is shown that if all numbers in *A* are close to *1* the deviation of the variables are proportional. As a result, fixing one of them, namely X_k, all the other variables have the deviations proportional to deviations of X_k:

$$X_i - M(X_i) = \frac{\sigma(X_i)}{\sigma(X_k)}(X_k - M(X_k)). \tag{3}$$

Statistically every variable may be taken as X_k, i. e. may be considered as representing the causal factor for the others. It is to discern which is the true causal factor.

In this case the correlation matrix rank is one and it is concluded that a single factor, the X_k, is the causal, all the other being determined by X_k, with the weights $\frac{\sigma(X_i)}{\sigma(X_k)}$.

The investigation is extended, further, to the case the matrix *A* have the rank equal to *r ≤ n*. As a result, *r* among the *n* factor will be find as statistically independent and causal for the other. The deviation of all the other variables may be written as linear combinations of causal factor's deviations. As well as in the case *r = 1*, from statistical point of view, there are many sets of *r* causal factors, the especially knowledge is necessary to find the true.

Linear dependence of random variables

For every other random variable Y we note $\rho_i = \rho(Y, X_i)$ and $\sigma_i = \sigma(X_i)$. Obviously,

$$M[(X_i - M(X_i))(X_j - M(X_j))] = \sigma_i \sigma_j \rho_{ij}. \tag{4}$$

The following proposition shows the conditions that *Y* to be written as a linear combination of $X_1, X_2, ..., X_n$.

Proportion. If Y – *M (Y)* = $\sum_{i=1}^{n} \alpha_i (X_i - M(X_i))$ then the numbers $\beta_1, \beta_2, ..., \beta_n$ defined by: $\beta_j = \frac{\alpha_j \sigma_j}{\sigma(Y)}$ fulfil the conditions: $\rho_i = \sum_{j=1}^{n} \beta_j \rho_{ij}$ and $\sum_{i,j=1}^{n} \beta_i \beta_j \rho_{ij} = 1$. For all *i = 1, 2, ..., n.*

Conversely, if there are the numbers $\beta_1, \beta_2, ..., \beta_n$ satisfying the relations: $\rho_i = \sum_{j=1}^{n} \beta_j \rho_{ij}$ for all *i* and $\sum_{i,j=1}^{n} \beta_i \beta_j \rho_{ij} = 1$ then: *Y* – *M (Y)* = $\sum_{i=1}^{n} \alpha_i (X_i - M(X_i))$ where $\alpha_i = \frac{\sigma(Y)\beta_i}{\sigma_i}$.

Proof. Let $Y - M(Y) = \sum_{i=1}^{n} \alpha_i (X_i - M(X_i))$ be a sure event. Then:

$$\rho_i = \rho(Y, X_i) = \frac{1}{\sigma(Y)\sigma(X_i)} M[(Y - M(Y))(X_i - M(X_i))] =$$

$$\frac{1}{\sigma(Y)\sigma(X_i)} M[(\sum_{j=1}^{n} \alpha_j (X_j - M(X_j)))(X_i - M(X_i))] =$$

$$\frac{1}{\sigma(Y)\sigma_i} \sum_{j=1}^{n} \alpha_j M[(X_j - M(X_j))(X_i - M(X_i))] = \frac{1}{\sigma(Y)\sigma_i} \sum_{j=1}^{n} \alpha_j \sigma_i \sigma_j \rho_{ij} = \sum_{j=1}^{n} \frac{\alpha_j \sigma_j}{\sigma(Y)} \rho_{ij}.$$

Taking $\beta_j = \frac{\alpha_j \sigma_j}{\sigma(Y)}$ we have: $\rho_i = \sum_{j=1}^{n} \beta_j \rho_{ij}$.

On the other hand,

$$1 = \rho(Y, Y) = \frac{1}{\sigma^2(Y)} M[(Y - M(Y))(Y - M(Y))] =$$

$$\frac{1}{\sigma^2(Y)} M\left[\left(\sum_{i=1}^{n} \alpha_i (X_i - M(X_i))\right)\left(\sum_{ji=1}^{n} \alpha_j (X_j - M(X_{ji}))\right)\right] =$$

$$\frac{1}{\sigma^2(Y)} \sum_{i,j=1}^{n} \alpha_i \alpha_j M[(X_i - M(X_i))(X_j - M(X_j))] = \sum_{i,j=1}^{n} \frac{\alpha_i \alpha_j}{\sigma^2(Y)} \sigma_i \sigma_j \rho_{ij} = \sum_{i,j=1}^{n} \beta_i \beta_j \rho_{ij}$$

Conversely, let $\beta_1, \beta_2, \ldots, \beta_n$ be some real numbers so that:

$$\sum_{i,j=1}^{n} \beta_i \beta_j \rho_{ij} = 1 \text{ and } \rho_i = \sum_{j=1}^{n} \beta_j \rho_{ij} \text{ for all } i = 1, 2, \ldots, n$$

and denote $\alpha_i = \frac{\sigma(Y)\beta_i}{\sigma_i}$. Then

$$\sigma^2\left(\sum_{i=1}^{n} \alpha_i X_i\right) = M\left[\left(\sum_{i=1}^{n} \alpha_i (X_i - M(X_i))\right)^2\right] = M\left(\sum_{i,j=1}^{n} \alpha_i \alpha_j (X_i - M(X_i))(X_j - M(X_j))\right) =$$

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j M[(X_i - M(X_i))(X_j - M(X_j))] = \sum_{i,j=1}^{n} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij} =$$

$$\sum_{i,j=1}^{n} \frac{\sigma(Y)\beta_i}{\sigma_i} \frac{\sigma(Y)\beta_j}{\sigma_j} \sigma_i \sigma_j \rho_{ij} = \sigma^2(Y) \sum_{i,j=1}^{n} \beta_i \beta_j \rho_{ij} = \sigma^2(Y),$$

so that $\sigma\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sigma(Y)$.

Furthermore

$$\rho\left(Y, \sum_{i=1}^{n} \alpha_i X_i\right) = \frac{1}{\sigma(Y)\sigma(\sum_{i=1}^{n} \alpha_i X_i)} M\left[(Y - M(Y))\left(\sum_{i=1}^{n} \alpha_i (X_i - M(X_i))\right)\right] =$$

$$\frac{1}{\sigma^2(Y)} \sum_{i=1}^{n} \alpha_i M[(Y - M(Y))(X_i - M(X_i))] = \frac{1}{\sigma^2(Y)} \sum_{i=1}^{n} \frac{\sigma(Y)\beta_i}{\sigma_i} \rho_i \sigma(Y) \sigma(X_i) =$$

$$\sum_{i=1}^{n} \beta_i \rho_i = \sum_{i=1}^{n} \beta_i \sum_{j=1}^{n} \beta_j \rho_{ij} = \sum \beta_i \beta_j \rho_{ij} = 1.$$

The correlation coefficient property for two random variables leads us to:

$$Y - M(Y) = \frac{\sigma(Y)}{\sigma\left(\sum_{i=1}^{n} \alpha_i X_i\right)} \sum_{i=1}^{n} \alpha_i (X_i - M(X_i)) = \sum_{i=1}^{n} \alpha_i (X_i - M(X_i)). \text{ Q.E.D.}$$

The rank of the correlation matrix

Let $A = (\rho_{ij})$ be the correlation matrix of random variables $X_1, X_2, ..., X_n$, i. e. $\rho(X_i, X_j) = \rho_{ij}$ and $r = rankA$. Renumbering eventually the variables we can suppose the first r lines in A are independent.

Theorem. For every $k > r$ the variable $X_k - M(X_k)$ can be written as a linear combination of Xi – $M(X_i)$; $i = 1, 2,..., n$.

Proof. There are real numbers $\beta_1, \beta_2, \beta_r$ so that $\rho_{ki} = \sum_{j=1}^{r} \beta_j \rho_{ji}$ for every $i = 1, 2,..., n$. The linear independence of the first r lines in A allows us to say that the β_i are determined. Also the sum can be extended to n taking $\beta_{r+1} = 0, \beta_{r+2} = 0,..., \beta_n = 0$.

Obviously, $\rho(X_k,X_k)=1$. On the other hand, $\rho(X_k, X_k) = \sum_{j=1}^{n} \beta_j \rho_{jk} = \sum \beta_j \rho_{kj} =$ $\sum_{j=1}^{r} \beta_j \sum_{h=1}^{r} \beta_h \rho_{hj} = \sum_{i,j=1}^{r} \beta_j \beta_h \rho_{hj} = \sum_{i,j=1}^{n} \beta_j \beta_h \rho_{hj}$ so that $\sum_{i,j=1}^{n} \beta_j \beta_h \rho_{hj} = 1$.

The preceding proposition can be applied by taking $Y = X_k$ and we get:

$$X_k - M(X_k) = \sum_{i=1}^{n} \alpha_i \left(X_i - M(X_i)\right); \text{ where } \alpha_i = \frac{\sigma(Y)\beta_i}{\sigma_i} = \frac{\sigma(X_k)\beta_i}{\sigma_i} = \frac{\sigma_k \beta_i}{\sigma_i}.$$

Because $\beta_i = 0$ for $i > r$ we obtain $X_k - M(X_k) = \sum_{i=1}^{r} \alpha_i \left(X_i - M(X_i)\right)$. Q.E.D.

Example

Consider five factors each represented by a set of *50* numerical data in Table 1. The number sets *A, B, C* are taken at random using the function RAND in *Excel*. The number sets *D* and *E* are taken so that:

$$D_i = 2A_i + 3B_i - 4C_i, E_i = 4A_i - 3B_i + 2C_i. \quad (15)$$

The coefficients in (15) expressing linear dependence between the data sets (considered as representing factors which suitable causal dependencies) will be find using the method elaborated before. This proves the consistency of the

Owing to the linearity in (15), the same relations are valid for the suitable deviations (3).

The standard deviations $\sigma_A, \sigma_B, \sigma_C, \sigma_D, \sigma_E$ are provided by the function STDEV from *Excel* and the correlation matrix by the function CORREL. These values are in the Tables 1 and 2.

Table 1. Experimental data

NR.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	17	10	97	96	51	13	54	74	62	99	83	14	82	0	63	41	0	69	20	30	60	12	69	31	75
B	82	63	71	78	71	47	17	42	63	53	47	18	3	77	48	7	88	91	7	18	13	49	53	30	13
C	25	62	92	77	76	71	49	27	31	12	69	17	5	15	39	88	13	63	81	32	84	8	1	92	21
D	182	-40	38	117	11	-116	-38	166	189	311	31	13	154	173	114	-249	210	161	-266	-15	-180	136	294	-216	108
E	-129	-24	359	305	142	51	265	223	120	259	329	35	328	-201	187	316	-236	131	222	130	369	-80	117	218	304

NR.	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
A	72	100	51	38	20	71	6	52	38	40	7	23	75	88	73	78	83	78	57	9	39	74	23	32	11
B	34	78	45	74	63	34	66	35	19	7	91	14	0	12	1	68	70	82	68	38	49	61	24	63	21
C	48	24	3	52	7	27	39	61	45	73	3	83	84	63	8	33	32	10	91	9	99	14	42	81	9
D	56	337	225	91	203	136	53	-35	-47	-192	274	-245	-184	-41	117	231	246	363	-49	93	-174	272	-50	-71	49
E	283	212	77	32	-98	239	-98	224	186	284	-239	216	467	444	307	173	185	86	205	-57	209	142	104	101	-1

Table 2. Standard deviations

A	B	C	D	E
30,0444	27,3844	30,7896	166,0148	167,6445

Table 3. Correlation matrix

	A	B	C	D	E
A	1	-0,01994	0,07261	0,298214	0,753305
B	-0,01994	1	-0,16331	0,608788	-0,564328
C	0,07261	-0,16331	1	-0,79639	0,4994005
D	0,298214	0,608788	-0,79639	1	-0,377084
E	0,753305	-0,56433	0,4994	-0,37708	1

One can verify that the rank of the correlation matrix is three and the first three lines are independent. The forth and fifth lines, a_{4*} and a_{5*}, can be written as linear combinations of $a_{1*}, a_{2*}, a3_{*}$:

$$a_{4*} = \beta_1 a_{1*} + \beta_2 a_{2*} + \beta_3 a_{3*};\quad a_{5*} = \overline{\beta}_1 a_{1*} + \overline{\beta}_2 a_{2*} + \overline{\beta}_3 a_{3*}. \qquad (16)$$

After solving the systems (16) one obtained:

$\beta_1 = 0{,}3619; \beta_2 = 0{,}4948; \beta_3 = -0{,}7418$ and: $\overline{\beta}_1 = 0{,}7168; \overline{\beta}_2 = -0{,}4900; \overline{\beta}_3 = 0{,}3673$

Proceeding from the theorem,

$$D_i = \alpha_1 A_i + \alpha_2 B_i + \alpha_3 C_i \ ; \ E_i = \overline{\alpha}_1 A_i + \overline{\alpha}_2 B_i + \overline{\alpha}_3 C_i$$

where: $\alpha_1 = \sigma_D \beta_1 / \sigma_A = 2$; $\alpha_2 = \sigma_D \beta_2 / \sigma_B = 3$; $\alpha_3 = \sigma_D \beta_3 / \sigma_C = -4$ and: $\overline{\alpha}_1 = \sigma_E \overline{\beta}_1 / \sigma_A = 4$, $\overline{\alpha}_2 = \sigma_E \overline{\beta}_2 / \sigma_B = -3$; $\overline{\alpha}_3 = \sigma_E \overline{\beta}_3 / \sigma_C = 2$ which are exactly the coefficients (15).

The solutions of systems (16) have lead to linear dependence of *D* and *E* of *A, B* and *C*, so that the last can be taken as causal factors.

If the (16) haven't any solution, (for example if the rank of correlation matrix is five) possibly for other number sets, one deducts the independence of suitable factors. This information may be also important.

In the above exempla are considered five factors, each represented by *50* numerical data. Practically there is no limit for the number of factors and the numerical data volume. The example is however notices the consistency of the method.

Conclusions

This method, based on the rank of correlation matrix is an efficient proceeding to find causal links.

The method is easy to apply, using functions in the program *Excel*.

The example contains in itself the abstraction, i. e. it is no restriction about the nature of the factors *A,B*.

The results obtained using the method, and the knowledge of researcher in the studied area may be important information in order to elaborate suitable strategies.

REFERENCES

1. *Iosifescu M., Mihoc Gh., Theodorescu R.* Teoria probabilităţilor şi statistica matematică. Bucureşti, 1966.
2. *Arnold O. Allen.* Probability, Statistics, and Queuing theory, Academic Press, New York, 1978.
3. *Chris Spatz, James O. Johnston,* Basic Statistics, Pittsburgh, U.S.A., 1990.

Received 17.12.99

Summary

In this paper we present a statistical method for the identification of causal factors in a complex process. The method is based on the use of the correlation coefficient and of the correlation matrix rank, in the goal to determine linear relations between huge volume data sets. The above mentioned relations can drive to the verification of some hypothesis or to determine new connections or links between different phenomenon. The paper also presents an example for the method. The interested domains, from this method point of view, are diverse and important: military, social, medical, economic, etc.