Piotr JANKOWSKI

Gdynia Maritime University

doi:10.15199/48.2016.04.05

About the lack of convergence in an environment with limited representation of the number

***Abstract**. The article presents cases of lack of convergence of transitional problems solutions both for the linear and non-linear examples because of using IEEE-754 standard. In a linear case, electric circuit solution using state variables was presented. As a non-linear case, ferroresonance system solved by various numerical procedures was shown. To solve the proposed problems, the 64-bit Mathcad Prime 3.0 environment was used.*

Streszczenie.** W artykule przedstawiono przykłady braku zbieżności w rozwiązaniu liniowego obwodu metodą zmiennych stanu oraz obwodu ferrrorezonasowego w 64 bitowym programie Mathcad Prime 3.0 stosującym obowiązujący standard IEEE-754. **Przykłady braku zbieżności w rozwiązaniu liniowego obwodu metodą zmiennych stanu

Keywords: Mathcad, convergence, transitional matrix.
Słowa kluczowe: Mathcad, macierz przejściowa

Introduction

Nowadays, very often to solve technical matters described by mathematical equations, ready applications such as Matlab, Mathcad, or Mathematica are used. The programs contain numerous preset procedures allowing to change various parameters which also have influence on accuracy. It is obvious that the value of many of these parameters has its limitations. Such a limitation is the maximum number of digits representing the number in these environments (17digits). It is caused by a still used internal way of representing floating point numbers using the BFP record (binary floating point). In order to standardize rules on floating point operations the IEEE-754 standard [4] was developed which defines the BFP formats available for binary system that is binary_32 called single precision, binary_64 - double precision and binary_128 - quadruple precision. However, in most of the packages in both numeric (as Matlab) and universal as (Mathcad) they use numbers record in double precision. Their record is composed of 64 bits, where 11 bits fall for the exponent and 52 bits for the mantissa. Such a method of recording the number gives a range from about ±2.2 10^{-308} to ±1.8 10^{308}. For example, the expression $2K!\ /\ K! = 2$ will not be calculated for K> 170 because the program will attempt to determine separately the value of a counter and afterwards a denominator thus exceeding the limit values understood by the computer. In addition to limiting the representation of numbers due to the length of computer words it is worth emphasizing the need of rounding the approximate irrational numbers.

Additionally, one should note that for many rational decimals there is no exact binary representation for example: $0.2_{10}=0.(0011)_2$. For this reason, one can encounter technical issues whose mathematical description gives a solution proven as convergent, and which using the environment such as, for example Mathcad turns out not to give approximate results even for well-conditioned problems. As examples for such problems the solutions of the classical tasks of electrotechnics theory such as transients for linear and non-linear circuit will be presented in the paper. Following example the problem of achieving convergence solutions in Mathcad environment was shown.

Example of linear circuit analysis

To solve the linear circuit (Fig.1) the equation of state variables was applied (1). In turn, to determine the transition matrix the Sylvester method, and further the method of Taylor series developing were used (5). When using a computer it should be emphasized that the method of Taylor is offered in references as effective for any non-singular matrix $\boldsymbol{A}$, and for any time [1].

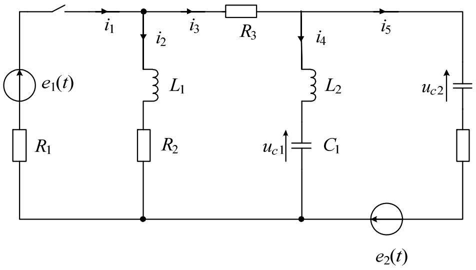

Fig.1. Tested linear circuit, $e_i(t)=E_{mi}\sin(\omega t)$

Table 1 Circuit parameters from Fig.1

E_{m1}[V]	E_{m2}[V]	ω[rad/s]	L_1[H]	L_2[H]	R_1[Ω]
15	18	314	0.3	0.1	2
R_2[Ω]	R_3[Ω]	R_4[Ω]	C_1[μF]	C_2[μF]	
5	3	4	15	40	

(1) $$\dot{\boldsymbol{x}} = \boldsymbol{Ax} + \boldsymbol{Be}(t)$$

To obtain the equation of state (1) in the normal form we formulate the Kirchhoff`s equations obtaining the following matrix form:

(2) $$\boldsymbol{H}\dot{\boldsymbol{x}} = \boldsymbol{Cx} + \boldsymbol{De}(t)$$

where: $\boldsymbol{x} = [u_{c1}(t), u_{c2}(t), i_2(t), i_4(t)]^{\mathrm{T}}$

Then multiplying equation (2) by $\boldsymbol{H}^{-1}$ we obtain:

(3) $$\dot{\boldsymbol{x}} = \boldsymbol{H}^{-1}\boldsymbol{Cx} + \boldsymbol{H}^{-1}\boldsymbol{De}(t) = \boldsymbol{Ax} + \boldsymbol{Be}(t)$$

In the considered example matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in Mathcad are determined as follows:

$$\boldsymbol{H}^{-1}\boldsymbol{C} = \boldsymbol{A} = \begin{pmatrix} -R_1C_1 & -R_1C_2 & -L_1 & 0 \\ -R_3C_1 & -R_3C_2 & L_1 & -L_2 \\ 0 & -R_4C_2 & 0 & L_2 \\ C_1 & 0 & 0 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 0 & R_1+R_1 & 0 \\ 1 & 0 & -R_2 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{H}^{-1}\boldsymbol{D} = \boldsymbol{B} = \begin{pmatrix} -R_1C_1 & -R_1C_2 & -L_1 & 0 \\ -R_3C_1 & -R_3C_2 & L_1 & -L_2 \\ 0 & -R_4C_2 & 0 & L_2 \\ C_1 & 0 & 0 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Because the solution of the inhomogeneous equation (1) can be represented as a formula (4):

$$(4)\qquad \boldsymbol{x}(t)=\boldsymbol{e}^{\boldsymbol{A}t}\boldsymbol{x_0}+\int_{0}^{t}\boldsymbol{e}^{\boldsymbol{A}(t-\tau)}\boldsymbol{Be}(\tau)\mathrm{d}\tau$$

hence fundamental to the application of the method of state variables is the determination of a transitional matrix (matrix exponentials) defined as the following series:

$$(5)\qquad e^{At}=A_{Taylor}(t)=\sum_{K}\frac{(\boldsymbol{A}\cdot t)^{K}}{K!}$$

When the matrix $\boldsymbol{A}$ of n-th order has n different eigenvalues, the transition matrix (the matrix exponential function) may be determined by a closed Sylvester formula [7]:

$$(6)\quad e^{\mathbf{A}t}=Asylvest(t)=\sum_{i=1}^{n}\exp(\lambda_i t)\frac{\prod_{s\neq i}(\lambda_s\cdot\mathbf{1}-\boldsymbol{A})}{\prod_{s\neq i}(\lambda_s-\lambda_i)}$$

Using formula (6) an analytic solution of considered circuit was obtained in Mathcad. Additionally to validate solutions the symbolic processor of this environment was used allowing the use of Laplace method. Examples of voltage and current waveforms showing their compatibility are presented in Fig.2.3.

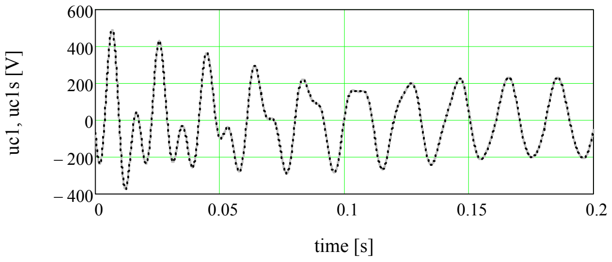

Fig.2. The waveforms of the voltages for the solution of the inhomogeneous equation by Sylvester and Laplace method

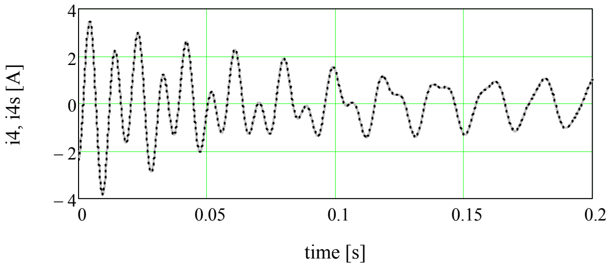

Fig.3. The waveforms of the currents for the solution of the inhomogeneous equation by Sylvester and Laplace method

It so happens that the matrix $\boldsymbol{A}$ in the equation of state (1) has multiple roots of the characteristic equation and then you can not apply the Sylvester method to determine transitional matrix. Additionally in the literature an alternative calculation of this matrix by the formula (5) is reported [1, 2, 3, 6, 7, 9]. Therefore, it has been attempted to solve the same example using Taylor method to develop a transitional matrix. In the Fig. 4.5 there are examples of solutions of the same example but with use of (5). Comparing these solutions, it was found that after a short time, the bifurcation of these solutions was started, although in Taylor series as many as K = 100 components were used. In addition, it must be admitted that the time of the solutions of inhomogeneous state equation using the formula (4) by using (5) increased by several times compared with the use of (6). Therefore, in further simulations a superposition of states using a simple symbolic solution of steady state for sine extortion was employed. The results coincided with the results using (6).

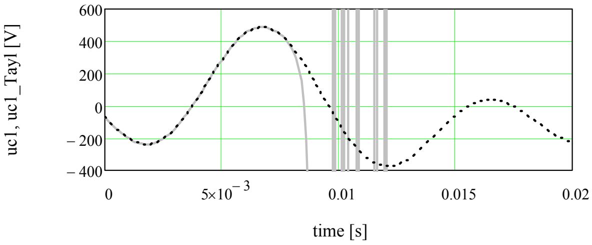

Fig.4. The voltages waveforms for the solution of the inhomogeneous equation by Sylvester and Taylor method for K=100

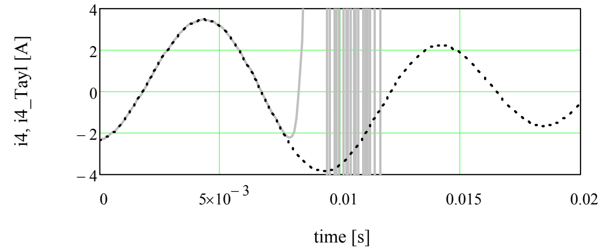

Fig.5. The currents waveforms for the solution of the inhomogeneous equation by Sylvester and Taylor method for K=100

Reason for lack of convergence of waveforms with Fig.4,5 is obvious if one compares the transitional matrices as determined by Sylvester and Taylor patterns:

$$\mathrm{Asylves}(0.023)=\begin{pmatrix}-0.645 & 0.048 & -2.39 & 22.557\\ 0.018 & 2.833\times10^{-3} & -1.141 & 2.899\\ 1.195\times10^{-4} & 1.521\times10^{-4} & 0.585 & -2.611\times10^{-3}\\ -2.256\times10^{-3} & -7.731\times10^{-4} & -5.221\times10^{-3} & -0.659\end{pmatrix}$$

$$\mathrm{A_{Taylor}}(0.023)=\begin{pmatrix}-1.314\times10^{8} & -6.382\times10^{9} & 8.846\times10^{6} & -2.659\times10^{10}\\ -4.019\times10^{9} & -1.987\times10^{11} & -7.449\times10^{10} & -5.746\times10^{11}\\ 9.389\times10^{5} & 1.677\times10^{7} & 3.663\times10^{7} & 2.052\times10^{8}\\ 4.5\times10^{6} & 1.258\times10^{8} & -5.676\times10^{7} & 1.192\times10^{9}\end{pmatrix}$$

The situation in this case is improved (but still for a unsatisfactory range of less than 15 ms) by an increased number of components of the series (5) to K = 170 that is the maximum value interpreted in Mathcad for the expression K! The results for this case are shown in Fig.6 and Fig.7.

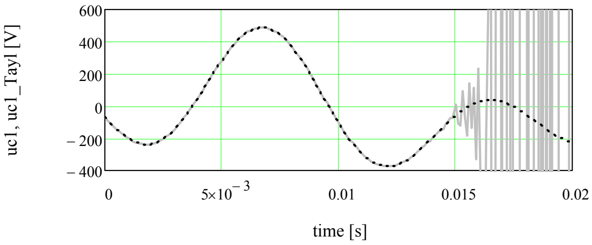

Fig.6. The voltages waveforms for the solution of the inhomogeneous equation by Sylvester and Taylor method for K=170

It turns out that in environments such as Mathcad one can not obtain satisfactory accuracy of transitional matrix determined on the basis of Taylor series development even using a large number of components. In this case, for K>170 Mathcad generates a message: *Found a number with a magnitude greater than 10^307 while trying to evaluate this expression.*

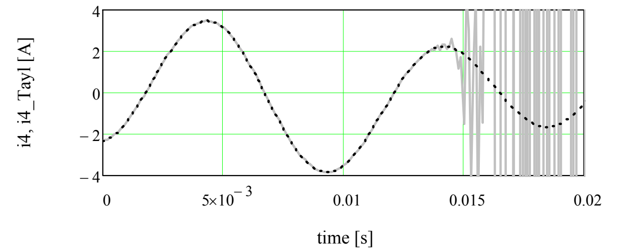

Fig.7. The voltages waveforms for the solution of the inhomogeneous equation by Sylvester and Taylor method for *K*=170

In [2] pattern (7) is formulated which determines the maximum error made in calculating the matrix function according to the formula (5). This ensures that the rest of the series (above *K* components) is convergent, if: $\|AT\| \le 1$

$$r_{i,j} \le \frac{\|AT\|^{K+1}}{(K+1)!} \cdot \frac{1}{1-\|AT\|} \tag{7}$$

The following is a fragment of simulation in Mathcad environment which computes the norms of AT for the $T1$ and $T2$ as well as the maximum error based on (7):

$\mathrm{T1} := 15\cdot10^{-3}$ $\quad \mathrm{T2} := 15\cdot10^{-6}$ $\quad \mathrm{normi}(\mathrm{A}\cdot\mathrm{T1}) = 1\times10^{3}$ $\quad \mathrm{normi}(\mathrm{A}\cdot\mathrm{T2}) = 1$

$\mathrm{K} := 100$ $\quad r_{ij} := \dfrac{\mathrm{normi}(\mathrm{A}\cdot\mathrm{T1})^{(\mathrm{K}+1)}}{(\mathrm{K}+1)!}\cdot\dfrac{1}{1-\mathrm{normi}(\mathrm{A}\cdot\mathrm{T1})}$ $\quad r_{ij} = -1.062\times10^{140}$

As you can see from the above simulation the norm of AT assumes a value of 1 already for the time 15 μs while the error for $T1$ (where the norm amounts to 1000) is enormous and rises with increasing K. On the other hand, below is shown a simulation of a transitional matrix for the time $T1$ to which waveforms of the determined quantities were still overlapping Fig.6,7.

$$\mathrm{A_{Taylor}}(0.01) = \begin{pmatrix} 0.79 & 0.018 & 0.01 & 33.261 \\ 6.775\times10^{-3} & -5.767\times10^{-3} & -1.595 & -4.107 \\ -5.093\times10^{-7} & 2.127\times10^{-4} & 0.792 & -2.233\times10^{-3} \\ -3.326\times10^{-3} & 1.095\times10^{-3} & -4.466\times10^{-3} & 0.779 \end{pmatrix}$$

$$\mathrm{Asylves}(0.01) = \begin{pmatrix} 0.79 & 0.018 & 0.01 & 33.261 \\ 6.775\times10^{-3} & -5.881\times10^{-3} & -1.595 & -4.107 \\ -5.093\times10^{-7} & 2.127\times10^{-4} & 0.792 & -2.233\times10^{-3} \\ -3.326\times10^{-3} & 1.095\times10^{-3} & -4.465\times10^{-3} & 0.779 \end{pmatrix}$$

Assuming that the values of the elements of the matrix from Sylvester method are accurate a matrix of errors was also defined:

$$\mathrm{Asylves}(0.01) - \mathrm{A_{Taylor}}(0.01) = \begin{pmatrix} -6.135\times10^{-9} & -2.084\times10^{-6} & -6.356\times10^{-6} & 5.204\times10^{-6} \\ 7.555\times10^{-8} & -1.149\times10^{-4} & -1.92\times10^{-4} & 2.145\times10^{-4} \\ 1.875\times10^{-11} & 2.093\times10^{-8} & 4.967\times10^{-8} & -3.196\times10^{-8} \\ -1.6\times10^{-10} & 4.518\times10^{-8} & 3.323\times10^{-7} & -1.96\times10^{-7} \end{pmatrix}$$

As you can see the values of matrix elements are indeed lower than the error estimated on the basis of (7), but the estimate does not constitute reliable information about the error of method. Of course, in the case of application of Taylor method one can use the discrete (iterative) method [2] where the matrix exp(AT) is determined once for the satisfying low-value T.

Example of nonlinear circuit analysis

The second considered example is a circuit with the highly non-linear coil (Fig.8). The circuit was a simplified model of the system power transformer [8]. Characteristics *i(Ψ)* was approximated by a polynomial of 11-th degree (8).

$$i = a\Psi + b\Psi^{11} \tag{8}$$

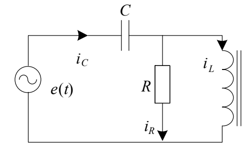

Fig.8. Tested nonlinear circuit

Table 2 Circuit parameters from Fig.8

$e(t)=E_m\sin(\omega t)$	a[A/Wb]	b[A/Wb^{11}]	C[nF]	R[Ω]
E_m=13.5kV	2.8x10^{-3}	7.2x10^{-3}	1250.77	10^{20}

Most of the procedures of Mathcad environment requires the normal form of equations, which for the circuit of Fig.8 is as follows:

$$\dot{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{x} = \begin{bmatrix} -x_1 + E_m\cdot\sin(\omega t) \\ \dfrac{1}{RC}\cdot(-x_1 + E_m\cdot\sin(\omega t)) + \dfrac{a}{C}\cdot x_0 + \dfrac{b}{C}\cdot x_0^{11} \end{bmatrix} \tag{9}$$

where: $\boldsymbol{x} = [\Psi(t), u_c(t)]^{\mathrm{T}}$

In turn, reduction of above system in the flux function leads to the well-known second-order equation called nonlinear damped oscillator:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}t^2} + \frac{1}{RC}\frac{\mathrm{d}\Psi}{\mathrm{d}t^2} + \frac{a}{C}\Psi + \frac{b}{C}\Psi^{11} = \omega E_{\mathrm{m}}\cos(\omega t) \tag{10}$$

In the first approach to the solution of the system (9) the Runge-Kutta procedure (in Mathcad – rkfixed) was used, wherein to determine the step leading to the convergence, iterative simulations were carried out reducing the step by half in each iteration. Figure 9 shows the course of the flux for a pre-100-fold lower voltage $E_{\mathrm{m}} = 135\mathrm{V}$ where convergence was achieved very quickly even for as wide a time range as more than 20 seconds.

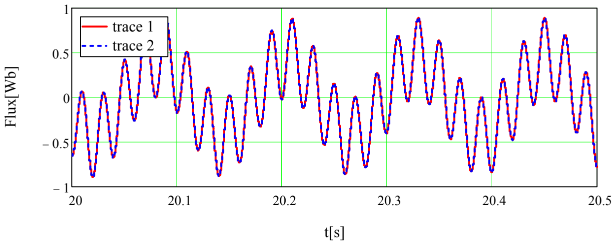

Fig.9. Flux for rkfixed (for h step -trace 1, for 0.5h step -trace 2) for E_{m}=135V

Further simulations were carried out already for a given voltage $E_{\mathrm{m}} = 13.5\mathrm{kV}$ (Tab. 2). This time by increasing the number of steps to a maximum value accepted by the Mathcad environment (89x10^{6}) one failed to achieve the expected convergence. As you can see in Fig 10, at a borderline small step one has managed to retain convergence only for a small time range, i.e. a little over 0.15 seconds. Since the matrix $\boldsymbol{A}$ of the system (9) is not best conditioned, also the procedure dedicated for badly conditioned systems (in Mathcad - Stiffr procedure) was applied. Unfortunately one also failed to achieve convergence for any step. In Fig.11 flux waveforms obtained by the Stiffr and Runge-Kutta procedure for the same small step were compared. As expected waveforms coincided only for a small time range. It should be emphasized that the simulations were carried out for the case of both a dimensionless and dimensional form but unfortunately it did not improve the convergence. The Fig.12 shows the waveforms of relative errors for both of them in the functions of iterations which means step decreasing by half.

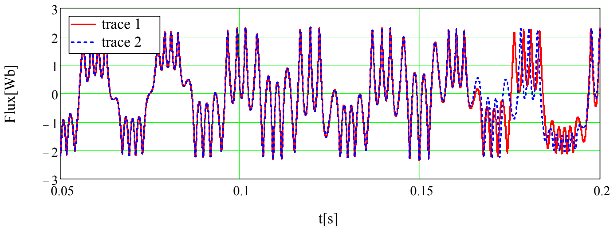

Fig.10. Flux for rkfixed (for h step -trace 1, for 0.5h step -trace 2)

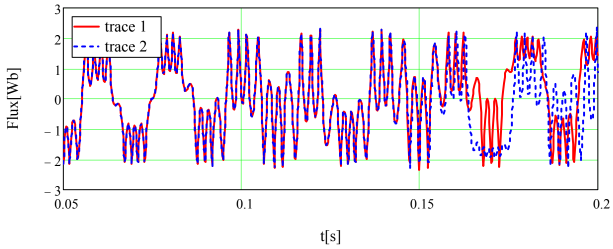

Fig.11. Flux for rkfixed and Stiffr (for h step -trace 1, for 0.5h step -trace 2)

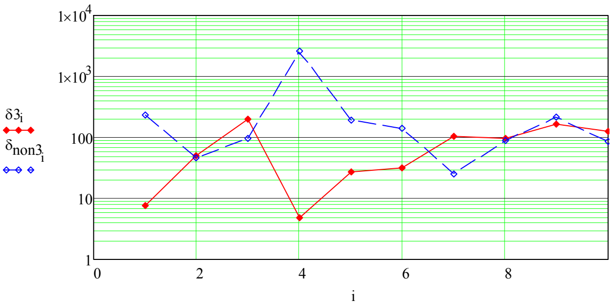

Fig. 12. Flux relative error between i+1 and i-th iteration for increasing number of steps for dimensional and nondimensional form for *t*=0.2s.

Using all family of procedures of Mathcad environment one did not get convergence for any of the desired time ranges. What is worse, Mathcad should display a message: *not converging*, and in the considered case it does not appear [5].

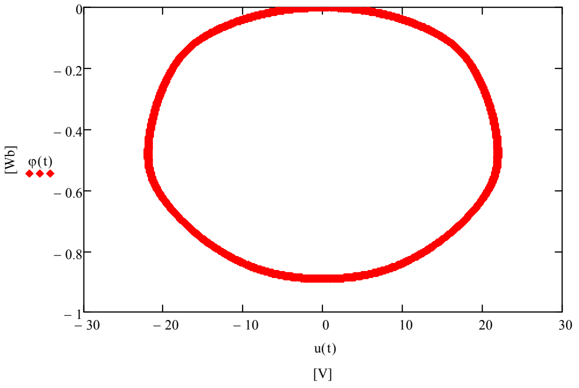

Fig.13. Poincare map for E_m=135V

Because the tested issue was considered for chaotic behavior [8] one carried out additional simulations for different voltages E_m determining the Poincare maps which means phasing portraits for times distant by a period of excitation (20ms). It was found that these maps for small voltages for which one managed to achieve waveforms convergence and their periodicity took the form of closed trajectories Fig.13. Otherwise, they had the form of chaotic sets Fig.14.

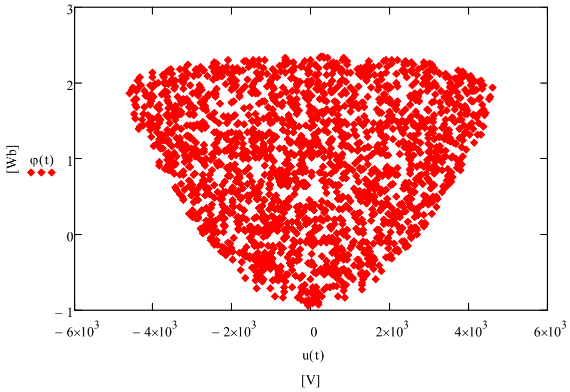

Fig.14. Poincare map for E_m=13.5kV

Conclusion

The presented results lead to the fundamental conclusion that for certain parameters known mathematical formulas (despite dedicating them to computing [1,2,3,6]) may not be effective in many currently popular numeric or multi-tasking programs, such as Mathcad or Matlab due to the binding IEEE-754 standard. At the same time whereas in the first example, the inaccuracy of the determination of transitional matrix based on the Taylor series was clearly visible then in the case of the nonlinear example the user can assume that if the program does not alert of the lack of convergence or exceeding limit values it means that it has obtained a reliable solution. Of course, there are specialized environments based on the size of binary128 (quadruple precision), which would undoubtedly improve the results and it seems that it is a matter of time when popular calculation tools mentioned in the article increase the precision of number representation.

Author*: Piotr Jankowski Department of Marine Electrical Engineering, Gdynia Maritime University*
Email:keopiotr@am.gdynia.pl

REFERENCES

[1] Bolkowski S., Teoria Obwodów Elektrycznych, Warszawa WNT 2005 p-343
[2] Chua L.O.,Lin P.M., Komputerowa Analiza układów elektronicznych, algorytmy i metody obliczeniowe WNT Warszawa 1979 .
[3] Direktor S., Rohrer A., Introduction To System . Mc-Graw-Hill 1972.
[4] IEEE 754-2008 - Standard for Floating-Point Arithmetic. DOI 10.1109/IEEESTD.2008.4610935, 2008.
[5] Jankowski P., Wybrane Zagadnienia Elektrotechniki w środowisku Mathcad Wyd. A.M. Gdynia 2011.
[6] Krakowski M., Elektrotechnika teoretyczna. Obwody liniowe i nieliniowe. PWN Warszawa 1995.
[7] Noble B., Applied Linear Algebra, Englewood Cliffs, N.J.:Prentice Hall,Inc.,1969.
[8] Marti J. R., Soudack A. C., Ferrorezonans in power systems: Fundamental solutions. *IEEE proceedings-C* Vol. 138, No.4. p-321-329, July 1991
[9] Osiowski J., Szabatin J., Podstawy teorii obwodów, t. III, WNT, Warszawa, 1995.