13. Chizhenkova R.A. Pulse flows of populations of cortical neurons under microwave exposure: the number of burst activity // Radiational biology. Radioecology. - 2010. - V. 50. - No. 2. - P. 201-210 (in Russian).
14. Chizhenkova R.A., Safroshkina A.A. Effect of low-intensity microwaves on the behavior of cortical neurons // Bioelectrochemistry and Bioenergetics. - 1993. - V. 30. - No. 1. - P. 287-391.
15. Chizhenkova R.A., Safroshkina A.A. Electrical reactions of the brain to microwave irradiation // Electro- and Magnetobiology. - 1996. - V. 15. - No. 3. - P. 253-258.
16. Chizhenkova R.A., Safroshkina A.A., Slashcheva N.A., Chernukhin V.Yu. Bibliometrical analysis of neurophysiological aspects of action of non-ionized radiation // Uspekhi sovremennoy biologii. - 2004. - V. 124. - No. 5. - P. 472-479 (in Russian).

THE OPTIMIZATION ENGINEERING COMPUTATIONS IN MICROSOFT OFFICE: REGRESSION ANALYSIS WITH UNIVERSAL SELECTION METHOD

S. Khomenko

National technical university of Ukraine "Kiev Polytechnic Institute", faculty of informatics and computer technics, Department of Computer Science, 37 Prospect Peremogy, Kiev 03056, Ukraine, khomenko@brainscode.com

The subject of the article is methods of optimization engineering computations in Microsoft Office. An algorithm for optimization, which is based on moving main calculations into compiled dynamic library, is displayed here. Also math methods for optimizations calculation regression analysis in way of using universal calculation method are considered here. In this way we can significantly reduce the number of calculations and increase speed of algorithm. Quantitative results of the coefficient usage effectiveness are given for the described method.

Microsoft Excel is widely used spreadsheet program. In the Excel we can use special macro language, Visual Basic for Applications (VBA) (1,2). VBA is an implementation of Microsoft's event-driven programming language Visual Basic 6, and associated integrated development environment (IDE), which is built into most Microsoft Office applications. That macro language provides methods and services which can be used for engineering computations. But time of calculation is too long for big task of statistics problems, because VBA is an interpreted programming language.

For engineering computations we need more strong compile language which will be compiled and will be able to calculate it quickly. In our statistics problem it's very important, because we have to calculate a great amounts of data. For example, in our article we also consider statistics problem of analysis linear regression models. And speed of calculations isn't enough for needed number of criterion.

The software way of decreasing time of calculations

One of the variants of increasing speed of computations is using external DLLs (dynamic link libraries) and moving all difficult calculation into DLL which will be coded in compiled programming language, for example, C++. Rosen and Partin (3) have described how to use VBA computations for task of chemical engineering. In that work we can read about calling external Fortran procedures for calculation VBA tasks. One of the advantages of using DLL's models is a possibility to use different programming languages in our VBA computations. In this way we can use all programmes which were coded before in different languages. We have only one condition for those languages: they must support compiling into DLL. In our time it's not difficult and the most common languages support this possibility. During this research we have tested communication between VBA and C/C++, Pascal, Object Ada.

VBA also can directly utilize functions and subroutines written in the popular C++ programming language (4). C++ language provides ways to build DLLs (4,5). In this way we can decrease time of calculation and use all of the advantages of an object-oriented programming paradigm. For a future possibility of increasing speed calculation we can use parallel and distributed programming, cluster technologies for calculations.

To confirm our theory we have completed a few tests. In that test there are results of the calculations of the same tasks, but which have been coded in different languages: VBA calculations with direct access to data(all calculations were calculated with cells in excel), optimized VBA calculations(developed special function for getting data from spreadsheet and calculation direct with structures of VBA), C++ DLL.

Basing on these data, we can conclude that the using compiled programming languages inside external DLL allows significantly reduce the time of calculation. In our example there is more than 138 times in comparison with direct access Excel solutions and more than 5 times in comparison with optimized Excel solutions. But it does not limit what we can enlarge speed with using parallel and distributed technologies.

Meanwhile, we have a possibility to use an object-oriented programming paradigm and all advantages of this paradigm.

Table 1

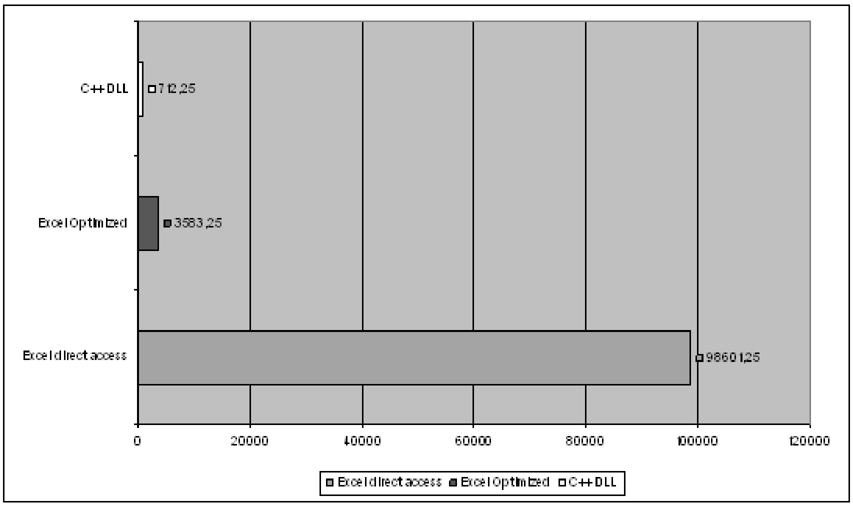
Time of working solutions based on each programming language

	Excel direct access	Excel Optimized	C++ DLL
1	98485	3601	718
2	98606	3578	719
3	98735	3592	703
4	98579	3562	709
avg.	98601,25	3583,25	712,25

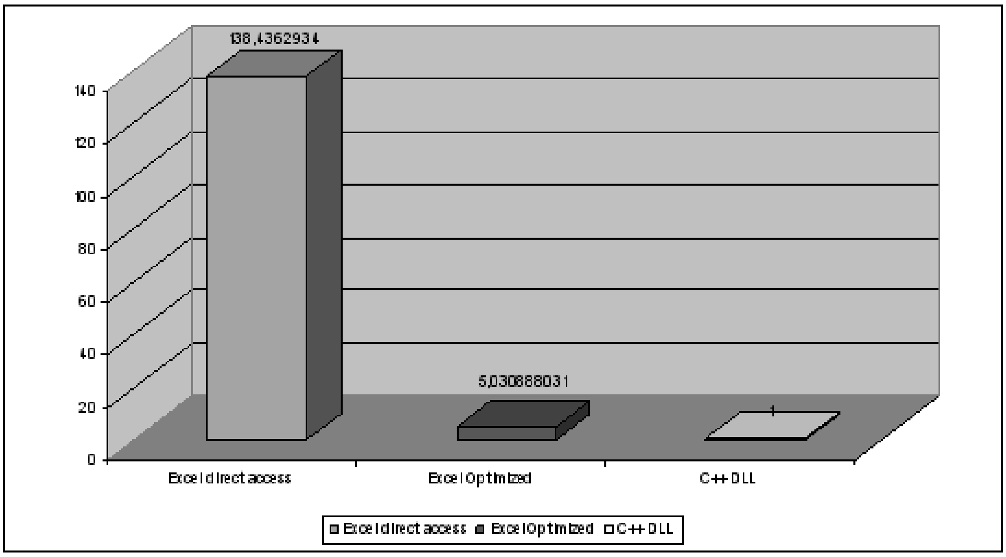
Table 2

The ratio of time computations in comparison with the time computations based on C++ DLL solutions

	Excel direct access	Excel Optimized	C++ DLL
1	137,1657382	5,015320334	1
2	137,1432545	4,97635605	1
3	140,4480797	5,109530583	1
4	139,0394922	5,023977433	1
avg.	138,4362934	5,030888031	1



Pics. 1 — time of computations solutions based on each programming language



Pic. 2 — the ratio of time computations in comparison with the time computations based on C++ DLL solutions

The theoretical way of decreasing number of calculations

Least square estimation using QR-method model:

$$Y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i$$

and in vector-matrix form:

$$\overrightarrow{Y} = X\overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

$\varepsilon_1,...,\varepsilon_n$ are independent random variables with $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$

Here $X \in R^{n,k}$ $(Rank(X)=k)$ is the design-matrix and contains the values of the regressor variables. The QR-factorisation (via Householder or Givens rotations) leads to

$$X = QR$$

where $Q = R^{n,n}$ is an orthogonal matrix $Q^{-1}=Q^T$, and $R \in R^{n,k}$ is a upper triangular matrix: $r_{i,j}=0$ for $j<i$.Since orthogonal transformations do not change the norm of vectors, we have

$$\begin{aligned}\|\overrightarrow{Y}-X\overrightarrow{\beta}\|^2 &= \|\overrightarrow{Y}-QR\overrightarrow{\beta}\|^2\\ &= \|Q^T(\overrightarrow{Y}-QR\overrightarrow{\beta})\|^2\\ &= \|Q^T\overrightarrow{Y}-R\overrightarrow{\beta}\|^2\end{aligned}$$

We split the vectors $Q^T\overrightarrow{Y}$ and $R\overrightarrow{\beta}$ into two parts:

$$Q^T\overrightarrow{Y} = \begin{pmatrix}\vec{a}\\ \vec{b}\end{pmatrix} \ with\ \vec{a}\in R^k,\ \vec{b}\in R^{n-k},$$

$$R\overrightarrow{\beta} = \begin{pmatrix}\vec{c}\\ \vec{o}\end{pmatrix} \in R^n,\ \vec{c}\in R^k$$

Let $\tilde{R}\in R^{k,k}$ an upper triangular matrix with the same nonzero entries as R: $\tilde{r}=r_{ij}\ for\ i,j=1...k$. Now $\vec{c}=\tilde{R}\vec{\beta}$. Hence

$$\|\vec{Y}-X\vec{\beta}\|^2 = \|\vec{a}-\tilde{R}\vec{\beta}\|^2 + \|\vec{b}\|^2$$

We achieve the minimum of $\|\vec{Y}-X\vec{\beta}\|^2$ for

$$\vec{a} = \tilde{R}\vec{\beta}$$

This system can be solved directly in inverse order, first for β_k then for $\beta_{k-1},...$ The solution gives the estimators $\hat{\vec{\beta}}$ for the model parameters. Moreover,

$$T_n = \min_{\beta} \|\vec{Y}-X\vec{\beta}\|^2 = \|\vec{b}\|^2$$

The aim is to consider all submodels and their minimum residual sum of squares. We introduce a submodel

$$\vec{Y} = \tilde{X}\vec{\gamma}+\vec{\varepsilon},\ \ \tilde{X}=XD,\ \ D=(\vec{e}_{i_1},\vec{e}_{i_2},...,\vec{e}_{i_l})\in R^{k,l},$$
$$1\le i_1 < i_2 < ... < i_l \le k$$

.

This submodel has $l=l(\nu)$ parameters $\gamma_1=\beta_{i_1},...,\gamma_l=\beta_{i_l}$. $\overrightarrow{e_i}=(0,...,1_i,0,...)^T\in R^k$ is the $i-th$ unit vector. We assign to the submodel a number $\nu=\sum_{j=1}^{l}2^{i_j-1}=\sum_{i\in I(\nu)}2^i$

with $I(\nu) =: \{i : \beta_i \text{ is in the model}\}, \nu \in \{0,1,...,2^k - 2\}$ The full model has the number $2^k - 1$ To each submodel we assign a complexity number $d(\nu)$ We have

$$\begin{aligned} \| \vec{Y} - XD\vec{\gamma} \|^2 &= \| \vec{Y} - QRD\vec{\gamma} \|^2 \\ &= \| Q^T(\vec{Y} - QRD\vec{\gamma}) \|^2 \\ &= \| Q^T \vec{Y} - RD\vec{\gamma} \|^2 \\ &= \| \vec{a} - \tilde{R}D\vec{\gamma} \|^2 + \| \vec{b} \|^2 \end{aligned}$$

We consider the special case where $i_j = j + k - l$. Then

$$\begin{aligned} S_\nu = S(i_1,...,i_i) &= \min_{\vec{\gamma}} \| \vec{Y} - XD\vec{\gamma} \|^2 \\ &= \min_{\vec{\gamma}} \| \vec{a} - \tilde{R}D\vec{\gamma} \|^2 + \| \vec{b} \|^2 \\ &= \min_{\vec{\gamma}} \| Q_\nu^T (\vec{a} - Q_\nu R_\nu \vec{\gamma}) \|^2 + \| \vec{b} \|^2 \\ &= \min_{\vec{\gamma}} \| \vec{a}_\nu - R_\nu \vec{\gamma} \|^2 + \| \vec{b}_\nu \|^2 + \| \vec{b} \|^2 = \| \vec{b}_\nu \|^2 + \| \vec{b} \|^2 \end{aligned}$$

where $\tilde{R}D = Q_\nu R_\nu$ is the QR-decomposition, $Q_\nu \in R^{k,k}$, Q_ν is orthogonal, $R_\nu \in R^{k,l(\nu)}$ is an upper triangular matrix,

$$Q_\nu^T = \begin{pmatrix} \vec{a}_\nu \\ \vec{b}_\nu \end{pmatrix} \in R^k, \ \vec{a}_\nu \in R^l, \ \vec{b}_\nu \in R^{k-l}$$

The solution $\vec{\gamma}$ of

$$\vec{a}_\nu = R_\nu \vec{\gamma}$$

gives the estimator $\hat{\vec{\gamma}}$ of the submodel.

Lemma 1:

$$\begin{aligned} & S(1,...,k-1) = \| \vec{b} \|^2 + a_k^2, \ S(1,...,k-2) \\ & = \| \vec{b} \|^2 + a_k^2 + a_{k-1}^2, ..., S(1) = \| \vec{b} \|^2 + a_k^2 + a_{k-1}^2 ... + a_2^2 \\ & \vec{a} = (a_1,...,a_m)^T \end{aligned}$$

Proof: Let

$$\begin{aligned} & \vec{g} = (a_1,...,a_m)^T, \vec{h} = (a_{m+1},...,a_k)^T, \\ & \tilde{R} = (r_{i,j}) \ i = 1..m, \ j = 1..k \\ & S(1,..,m) = \| \vec{b} \|^2 + \| \vec{h} \|^2 + \min_{\vec{\gamma}} \| \vec{g} - \tilde{R}D\vec{\gamma} \|^2 = \| \vec{b} \|^2 + \| \vec{h} \|^2 \end{aligned}$$

Idea for excluding submodels

The following inequalities are equivalent:

$$\overline{M_n}(\nu) = \frac{S_\nu - T_n}{\frac{1}{n-l} S_\nu} > \psi$$

$$S_\nu > T_n \cdot \frac{n-l}{n-l-\psi}$$

On the other hand $S_m > S_\nu$ for submodels m if submodel m has only a subset of parameters of that of submodel ν If for a submodel the inequality $S_\nu > T_n \cdot \dfrac{n-l}{n-l-\psi_n(d,l)}$ holds, then a submodel m with with a subset of l parameters and complexity d can be excluded because it could not be acceptable.

Define

$$\psi_n(d,l) = \chi^2_{k-l}(1 - \alpha_n(d)).$$

Rule for the selection:

Select a model ν^* such that

$$d(\nu^*) =$$
$$\min\{d(\nu): 0 \le \nu \le 2^k - 1, \overline{M}_n(\nu) < \psi_n(d(\nu), l(\nu))\}$$
$$and$$
$$\overline{M}_n(\nu^*) = \min\{\overline{M}_n(\nu): 0 \le \nu \le 2^k - 1, d(\nu) = d(\nu^*)\}$$

A submodel is called an acceptable one if $\overline{M}_n(\nu) < \psi_n(d(\nu), l(\nu))$ The central idea is to prefer any admissible model with lower complexity. If there is more than one admissible model with the same minimum complexity, then we take the model with minimium value of $\overline{M}_n(\nu)$

Algorithm of a universal selection method in linear regression models

1. Compute QR-decompsition, $\vec{a}, \tilde{R}$

$$T_n = \|\vec{b}\|^2$$

2. For all submodels ν do

2.1. Compute $S_\nu = \|\vec{b}_\nu\|^2 + \|\vec{b}\|^2$

$$\overline{M}_n(\nu) = \frac{S_\nu - T_n}{\frac{1}{n - l(\nu)} S_\nu}$$

2.2. Let $l = l(\nu)$ be the number of parameters in the submodel $\nu, d = d(\nu)$ be the complexity of submodel ν Decide whether $\overline{M}_n(\nu) < \psi_n(d, l).$. Then the model is acceptable.

2.3. Use Lemma 1 to compute the residual sum of squares S_m for submodels m with fewer variables in that cases where they are not computed in an earlier step. For these models, decide whether

$$\overline{M}_n(m) < \psi_n(d(m), l(m)).$$

Then the corresponding model is acceptable.

3. Among all acceptable models, search for a submodel ν^* with a minimum complexity number $d(\nu^*)$ If there is more than one admissible model with the same minimum complexity, then we take the model with minimium value of $\overline{M}_n(\nu)$.

4. Compute the estimators of the best submodel ν^*

$$\hat{\vec{\beta}} = \tilde{R}^{-1}\vec{a} \quad or \quad \hat{\vec{\lambda}} = \tilde{R}_\nu^{-1}\vec{a}_\nu$$

Conclusions

In this article we reviewed methods and ways of calculation an engineering task in Microsoft Excel. Among the methods we considered: calculations based on Excel calculations, calculations based on optimized Excel calculations and calculations based on calculations inside C+ DLL. From the results of testing can be inferred about increasing the speed of calculation. In our test it was 138 times less in comparison with the direct access Excel solutions and more than 5 times in comparison with optimized Excel solutions.

Also the article shows the technique of analytics reducing the computational complexity of selection method in linear regression. For decreasing number of operations we have used QR-decompositions. It allows to decrease an amount of data for calculations. Basing on lemma about high order number of submodels we are able to calculate parameters only for the highest number of set. And do calculations for smaller order in the analytics way.

In the future we can increase the calculation speed and thus increase the number of criteria that can be used in the model. One of the way of increasing speed of the calculations can be the using technologies of parallel and distributed programming, as we now use C++ DLL for main calculations it will be not difficult, because C++ language includes such technologies. Another way is optimizing theoretical ways of selection optional submodels in the model.

REFERENCES

1. Rob Bovey, Dennis Wallentin, Stephen Bullen, John Green, Professional Excel Development: The Definitive Guide to Developing Applications Using Microsoft Excel, VBA, and .NET (2nd Edition), Addison-Wesley Professional, 2nd edition, Boston (2009)
2. John Walkenbach, Excel 2010 Power Programming with VBA, Wiley, New York, (2010)
3. Rosen, E. M. and L. R. Partin "A Perspective: The Use of the Spreadsheet for Chemical Engineering Computations" , I&EC Chemistry Research, (2000)
4. D. S. Malik, C++ Programming: From Problem Analysis to Program Design, Course Technology, 5 edition, Stamford, (2010)
5. D. Ryan Stephens, Christopher Diggins, Jonathan Turkanis, Jeff Cogswell, C++ Cookbook, O'Reilly Media, (2005)
6. Akaike, Hirotugu (1974). A new look at the statistical model identification, IEEE Transactions on Automatic Control 19, 716–723.
7. F Bunea, M Wegkamp, A Auguste (2006), Consistent variable selection in high dimensional regression via multiple testing, Journal of Statistical Planning and Inference, 136, 4349-4346
8. Gatu, Cristian, Yanev, Petko I., Kontoghiorghes, Erricos J., (2007), A graph approach to generate all possible regression submodels, Computational Statistics & Data Analysis, 52, No. 2, 799-815
9. Hannan, E. J. and Quinn, B. G. 1979. The Determination of the Order of an Autoregression, Journal of Royal Statistical Society, B 41 (2), 190 – 195

ORGANIZATION METHODS OF INFORMATION PRODUCTS IN DATASPACE

N. Schakhovska,

PhD, Associate Professor, Lviv Polytechnic National University, Ukraine, natalya233@gmail.com

The physical system is dynamic and its elements have evolved at different rates. It complicates the collection and processing of information on elements of such a system. To work with different types of information from different sources, we can apply dataspace. One element of the dataspace is an information product.

Information Product (IP) is a documented information resource, prepared according to the needs of users and submitted as product. Information product can be software, text files, web pages, spreadsheets, xml-files, databases, datawarehouses, etc.

Catalogue of Information Product – metadata about information products – describes the location of an information product, its structure, methods of access to the information resource, etc.

Traditionally, experts used usual for them sources of information for solving tasks [1,4,5]. Apparently, this approach has incomplete information, which is processed. Many sources of data and services that exist on the Internet are causing the need for a radical change in the methods of getting data. This change is a task, which is formulated independently of existing data sources. After its formulation the identification of relevant sources, bringing data to the appropriate type, integration, identification services which allow solving a separate part of tasks should be carried out. The adoption of adequate solution require the data, coming from different sources to satisfy the following requirements: be complete, consistent and received on time; be informative, because they should be applied for decision support; be of uniform structure for the opportunity of being downloaded in single datawarehouse and analyzed; kept in uniform models of data and be independent of the development platform for the opportunity of using this data in other means. But today there are no data processing methods that would satisfy all of the requirements for data processing [2,3].

ALGEBRAIC SYSTEM OF DATASPACE

There are methods of data processing from sources with different data structures. (Table 1).

Dataspace is a set of all information product domain *DS*= **<DB, DW, Wb, Nd, Gr>,** where **DB, DW, Wb, Nd, Gr** are information products that submit a set of databases**,** datawarehouse, web pages, text files, spreadsheets, image data respectively. In an energy system databases, datawarehouses, text files, spreadsheets, which are described in different formats are used.

Consolidated data is derived from multiple sources and systematically integrated heterogeneous information resources, which together are have such features as completeness, integrity, consistency and adequacy/ This consolidated information is model of the subject area for its analysis and processing efficiency in the processes of decision making.

Information products describe the specific subject area, and consolidated data constitute the data space. One of the problems that persist in the process of consolidation is the uncertainty of data, the result of duplication, inaccuracies, data absence, contradictions of the data (Fig. 1). Also one of the areas which are