77th European Study Group with Industry

27th September – 1st October 2010

Stefan Banach International Mathematical Center, Warsaw, Poland

Organizers:

Systems Research Institute
Polish Academy of Sciences

Institute of Mathematics
Polish Academy of Sciences

OCCAM
Oxford Centre for Collaborative Applied Mathematics

RB/4/2011

Models and measures to evaluate the effectiveness of funds utilization for scientific research and development of advanced technologies

REPORT ON THE PROBLEM

Honorary patronage:

British Embassy

Sponsors:

Sygnity
competence on

Industrial Development Agency
Joint Stock Company

Problem founded by

Information Processing Centre

Report authors

Jakub Lengiewicz (Institute of Fundamental Technological Research PAS)
Krzysztof Turek (Jagiellonian University)
Jacek Lewkowicz (University of Warsaw)

Contributors

Poul Hjorth (Technical University of Denmark)
Agnieszka Kaszkowiak (University of Warsaw)
Anna Kortyka (Institute of Physics PAS)
Mariusz Marczewski (Commercial Discoveries Processing)
John Ockendon (University of Oxford)
Wojciech Okrasiński (Wrocław University of Technology)
Zbigniew Peradzyński (University of Warsaw)
Piotr Wojdyłło (Institute of Mathematics PAS)
Karolina Wojtasik (University of Silesia)
Maciej Żmuda (Wrocław University of Economics)

ESGI77 was jointly organised by

System Research Institute of the Polish Academy of Sciences
Institute of Mathematics of the Polish Academy of Sciences
Oxford Centre for Collaborative Applied Mathematics

and it was supported by

Sygnity S.A.
Industrial Development Agency Joint Stock Company

under the honorary patronage of

The British Embassy in Poland

Executive Summary

The purpose of this report was to construct some alternative methods to estimate the effectiveness of investments in scientific research and development of advanced technologies, especially their long-term effects.

Study Group decided to focus on the sub-problem of finding the relation between the spending on science and the quality of science itself. As a result, we have developed two independent methodologies. The most promising one is based on the theory of time-delay systems, which allows capturing effects of the time-lag between the use of funds and the results related to scientific work. Moreover, the methodology gives an opportunity to seek the optimal spending scenario that would fulfill some prescribed constraints (e.g. it would minimize costs and at the same time remain above a desired level of quality of science).

The second methodology is premised on Stochastic Frontier Analysis and it can be applied to determine the form of relation between the amount of financing and the results of scientific work. It offers considerable advantages for analyses of several forms of relation at once (production functions) and for a suitable choice of the best one.

Both methods are promising, however, additional work is necessary to apply them successfully to some real-life problems.

Contents

1 Introduction

1.1 Problem description

(1.1.1) There is a common belief that channeling funds for scientific research and advanced technologies is one of the most efficient ways to make long-term investments. Despite the multitude of reports and publications on the subject, it is still unclear whether more money allocated to scientific research brings about desired effects. Thus, long-term funding plans submitted by politicians are more and more often brought into question.

(1.1.2) Existing models that describe the correlations between funds spent on scientific research & advanced technologies and prosperity indicators (e.g. GDP) tend to be inaccurate for several reasons. For instance, significant difficulties arise when providing comparable measurement conditions or statistical insignificance of input data. On the other hand, a comprehensive analysis cannot be performed by means of tools like Science, Technology & Innovation Indicators, as well as various econometric, nonparametric or scoring methods, even after adequate modifications. The study should also take into account some long-term effects of investments in scientific research, like a number of implementations, financial profits, impact on economy or quality of life. Including long-term effects in the analysis and predicting the outcome would be a breakthrough, which would allow managing funds in a more effective way.

(1.1.3) **Main challenge**

The purpose of this project is to construct some alternative methods to estimate the effectiveness of investments in scientific research and development of advanced technologies, with particular emphasis placed on the long-term effects.

1.2 Problem breakdown

(1.2.1) It would be a difficult task to analyze a direct impact of financing of scientific research and advanced technologies by means of some global prosperity indicators (like GDP). This stems from the fact that such global indicators depend strongly on a large number of various factors (e.g. monetary policy, cyclic fluctuations in the economy, political situation) which are barely related to the direct results of e.g. scientific research. Thus, any such modeling would need to describe the economy of the whole country comprehensively (or maybe even the World Economy) in order to extract exclusively the desired effects (correspondence).

(1.2.2) Keeping this in mind, the Study Group focused on describing the relation between the funds used for scientific research and some direct results of scientific work. We are interested in results that comprise among others indicators of the quality of science or indicators concerning the forms of the application of scientific work.

(1.2.3) The two main approaches are considered in the report. The first one is based on the concept of time-delay systems, which allows modelling the time-lag between spending money and resulting changes. The second approach is grounded on one of the methods of economic modelling – the Stochastic Frontier Analysis.

1.3 Input data

(1.3.1) Usefulness of the proposed models can be achieved by operating on some well defined instruments (e.g. collective indicators) derived from some measurable indicators. The Study Group proposed the set of such basic, measurable indicators (which may be further extended, if necessary):

$S_1(t)$ - number of trained scientists,

$S_2(t)$ - number of PhD students ,

$S_3(t)$ - number of PhDs working in science,

$S_4(t)$ - value of scientific infrastructure,

$S_5(t)$ - number of PhDs in industry,

$S_6(t)$ - number of industrial research centers,

$S_7(t)$ - number of patents,

$S_8(t)$ - total maintenance costs of scientific infrastructure.

Note that indicators 1-4 describe the quality of science itself, whereas indicators 5-7 relate to the application of science and indicator 8 may be understood as a fraction of a science budget (spending).

(1.3.2) It is crucial to provide the appropriate (desired) formula for constructing collective indicators (e.g. the quality of science indicator $P(t)$, used in Chapter 2), however, the choice of such a formula remains arbitrary, unless some additional information is given. For example, during the process of constructing a collective indicator it might turn out that economists or politicians decide which basic indicators are more important. Therefore, we find this problem to be out of the scope of the report.

2 Time-delay system approach

2.1 Basic modelling

(2.1.1) Following Pitcher [1] and referenced literature (especially Middleton 2006) we assume that the evolution of quality of science $P(t)$ depends on both the current values of the total budget $B(t)$ and the fraction allocated to science and on their values at earlier times. We divide the science budget into the education budget $B_e(t)$ and the research budget $B_r(t)$ due to different 'delay times' between allocation and measurable impact. These time-delays are denoted by τ_e and τ_r respectively (say 5 and 10 years).

(2.1.2) It is convenient to work with the fractions $U_e = \frac{B_e}{B}$, $U_r = \frac{B_r}{B}$. We shall now model the evolution of $P(t)$ by

$$\begin{aligned}\frac{dP(t)}{dt} = {} & \alpha_e \cdot U_e(t-\tau_e)\cdot B(t-\tau_e) + \beta_e \cdot [\delta_e + U_e(t)] \cdot B(t) + \\ & \alpha_r \cdot U_r(t-\tau_r)\cdot B(t-\tau_r) + \beta_r \cdot [\delta_r + U_r(t)] \cdot B(t) - \gamma \cdot P(t)\end{aligned}, \qquad (1)$$

where $\alpha_e, \alpha_r, \beta_e, \beta_r, \delta_e, \delta_r, \gamma$ are positive parameters which will be briefly discussed. We can now perceive the equation as a model for P given the total budget B with the allocations U_e and U_r as control variables at the discretion of the Ministry of Science. In this simple modelling, only government funding is included, however, there is no obstacle to extending the model.

(2.1.3) The right-hand-side of (1) consists of 5 terms, and describes the evolution of the quality of science at time t. The first term indicates how the spending on education at time $t-\tau_e$ influences the present increase in the quality. The second term describes the dependence on current expenses for education. Note that parameter δ_e is used to prevent the artificial effect of complete deterioration of the quality of science when no funding is provided, however further study is needed to better understand its influence on the solution. The third term and the fourth one regard the research and their meaning is analogous to the respective terms described above. The last one simulates the spontaneous deterioration of the quality of science (depending on the definition of $P(t)$, cf. (1.3.2)), due to e.g. corruption of scientific infrastructure or drop of 'attractiveness' of knowledge (if something was invented a long time ago, it has probably been already exploited).

(2.1.4) Estimation of the model parameters is based on the historical data concerning some discrete moments in time. Subsequently, the model (1) has to be transformed into a time-discretized counterpart e.g. by substituting $dP(t)/dt$ with $(P(t+\Delta t)-P(t))/\Delta t$. We assume that the values of $B(t)$, $U_e(t)$ and $U_r(t)$ can be provided for a sufficiently large number of time instants in the history.

(2.1.5) During the estimation process, we can keep some prescribed values of time delays τ_e and τ_r, and fit the model with regard to the following parameters: $\alpha_e, \alpha_r, \beta_e, \beta_r, \delta_e, \delta_r, \gamma$. On the other hand, we can include also τ_e and τ_r as estimated parameters. This allows also adjusting time delays, which makes the estimation possibly more accurate; however, in such a case, the problem of model identification becomes a discrete optimization problem, which is more difficult to solve.

2.2 Budget optimization

(2.2.1) Having estimated the model parameters, one can use the model to predict the future values of $P(t)$ for some given control variables $B(t)$, $U_e(t)$ and $U_r(t)$. This kind of a case study for different controllers may be an interesting task *per se*, however, it is far more interesting and useful to find the values for control variables for which some additional constraints, besides Eqn. (1), are fulfilled. This leads to the problem of optimal control.

(2.2.2) General, discrete optimal control problem

Minimize the sum

$$\sum_{i=1}^{N} \Phi_i\left(P(t_i), B(t_i), U_e(t_i), U_r(t_i), t_i\right), \tag{2}$$

subject to the discrete version of Eqn. (1)

$$\begin{aligned}\frac{P(t_{i+1}) - P(t_i)}{t_{i+1} - t_i} = {} & \alpha_e \cdot U_e(t_i - \tau_e) \cdot B(t_i - \tau_e) + \beta_e \cdot [\delta_e + U_e(t_i)] \cdot B(t_i) + \\ & \alpha_r \cdot U_r(t_i - \tau_r) \cdot B(t_i - \tau_r) + \beta_r \cdot [\delta_r + U_r(t_i)] \cdot B(t_i) - \gamma \cdot P(t_i)\end{aligned}, \tag{3}$$

the set of algebraic path constraints

$$b\left(P(t_i), B(t_i), U_e(t_i), U_r(t_i), t_i\right) \le 0, \tag{4}$$

and initial conditions accounting for time-delay requirements (depending on values of τ_e and τ_r).

Note that, although we have proposed the time-discretized version of the optimal control problem, it transforms straightforwardly into its continual counterpart.

(2.2.3) Example 1

We are about to put forward one of the possibly useful specifications of the problem (2.2.2). We assume that the total budget $B(t)$ for subsequent K years has already been agreed (it is out of control). The Ministry of Science and Higher Education tries to minimise funds for science. However, at the same time, it wants to meet some minimal requirements about the quality of science (given by a set of waypoints $\overline{P}(t_i)$ for subsequent years). Therefore, the problem (2.2.2) will become the minimisation of

$$\sum_{i=1}^{K} B(t_i) \cdot [U_e(t_i) + U_r(t_i)], \tag{5}$$

subject to (3) and constraints

$$\overline{P}(t_i) - P(t_i) \le 0 \qquad \mathit{for} \quad i = 1 \ldots K. \tag{6}$$

(2.2.4) Example 2

The second example describes the situation in which some total K-year budget $\overline{C}$ for science is reserved, and all we have to do is to optimize the spending in subsequent years in such a way that the quality of science would be maximized. We may write it down as maximization of

$$\sum_{i=1}^{K} P(t_i), \tag{7}$$

subject to (3) and constraint

$$\sum_{i=1}^{K} B(t_i) \cdot [U_e(t_i) + U_r(t_i)] = \overline{C}. \tag{8}$$

(2.2.5) Note that not all specifications of the problem (2.2.2) make sense. For example, if we want to maximize the quality of science with some upper limits on funds, then the optimal solution will always reach upper limits, which stays in accordance with real-life experience and the tendency according to which giving more money improves the quality.

2.3 More advanced modelling

(2.3.1) The idea proposed in this chapter is based on the fact that scientific workforce is a crucial factor – having no workers means producing no effects. There are some assumptions about the modeling. We surmise that the number and the quality of scientific workforce depend only on population and funds spent on education (we do not generally model educational system). Undoubtedly, the reduction of investments in education to zero does not mean that there will be no scientists at all (e.g. a flux of specialists from industry, immigrants will still remain etc.) – that is why δ_e appears in Eqn. (7). Scientific workforce is also prone to degradation (drop in the quality due to age, retirement, emigration, deaths etc.) and this effect is denoted by the term γ_1 below.

(2.3.2) The model is given by the set of delay differential equations:

$$\begin{cases} \dfrac{dN(t)}{dt} = \alpha_e \cdot M(t-\tau_1) \cdot U_e(t-\tau_1) \cdot B(t-\tau_1) + \beta_e \cdot (\delta_e + U_e(t)) \cdot B(t) - \gamma_1 \cdot N(t) \\ \dfrac{dP(t)}{dt} = \alpha_r \cdot N(t-\tau_2) \cdot U_r(t-\tau_2) \cdot B(t-\tau_2) + \beta_r \cdot N(t) \cdot (\delta + U_r(t)) \cdot B(t) - \gamma_2 \cdot P(t) \end{cases} \tag{9}$$

where:

$N(t)$ - potential of scientific workforce (number of PhDs and their quality),

$M(t)$ - population of people around 25 years of age (potential PhDs),

$U_r(t)$ - fraction of budget spending on research,

$U_e(t)$ - fraction of budget spending on education (especially higher education),

τ_1 - delay of the entrance of PhDs on the labour market associated with the cost of education,

τ_2 - delay of effects of research.

3 Data-based modelling

(3.1.1) The methodology proposed in this chapter aims to determine the form of dependency between the amount of financing and the results of scientific work. The method is based strongly on a given set of (historical) data – from various available forms of dependency, one needs to choose the one that in certain sense fits the data best (more precise description below).

(3.1.2) The main idea is to distinguish three sets of indicators for each research centre (institute): first, they should describe the quality of science *per se* (indicators of quality of pure science: PS_i^A), such as:

$PS_1^A(t) = S_1^A(t)$ - number of trained scientists

$PS_2^A(t) = S_2^A(t)$ - number of PhD students

$PS_3^A(t) = S_3^A(t)$ - number of PhDs

$PS_4^A(t) = S_4^A(t)$ - value of infrastructure

indicators describing the usefulness of scientific work for society/economy (indicators of application of science: AS_j^A), such as:

$AS_1^A(t) = S_5^A(t)$ - number of PhDs in industry

$AS_2^A(t) = S_6^A(t)$ - number of industrial research centers

$AS_3^A(t) = S_7^A(t)$ - number of patents

indicators describing the financing of science (indicators of costs: C_k^A), such as:

$C_1^A(t) = S_8^A(t)$ - total maintenance costs,

where integer t is a respective time period and A indicates a research institute.

(3.1.3) The analysis is performed in two steps. The first one consists in finding the relation between the indicators describing the quality of pure science and the indicators of application of science. In this stage we select for further analysis only these quality indicators, which are important in a certain sense (given below). Subsequently, we look for the relation between the financing of science and the previously selected subset of important indicators of the quality of pure science.

(3.1.4) For the purpose of modelling dependency between the indicators applied in the first phase, we use the Stochastic Frontier Analysis (SFA) [1] . The main stages of the SFA are as follows:

1. Given the form of production volume for a given research institute A:

$$Y^A(t) = f(\{PS_i^A(t)\}, \{b_r\}, t) \cdot E \cdot e^v \tag{10}$$

where

$Y^A(t)$ - output indicator (one of the applications of science indicator)

$f(\bullet)$ - form of production function

$\{PS_i^A(t)\}$ - set of input indicators (subset of pure science quality indicators)

$\{b_r\}$ - vector of parameters of the model (where b_0 is a scaling parameter)

E - effectiveness of production (random factor); it has the same type of distribution for each research institute, e.g. log-normal, variation and mean are estimated later on.

v - random error, the same distribution for every research institute (only one type is given, mean and variation will be estimated).

2. Choose one of available forms of production function for each product $R \in \{AS_i\}$, e.g. Cobb-Douglas production function [3]:

$$f_R(\{PS_i(t)\}, \{b_{R,r}\}, t) = e^{b_{R,0}} \prod_{i=1}^{|PS^A|} PS_i^{b_{R,i}} \tag{11}$$

3. Take a subset of indicators $\{\overline{PS}_j^A(t)\} \subseteq \{PS_i^A(t)\}$ (the same types of indicators for each A) and adequate subset of parameters $\{\bar{b}_{R,r}\} \subseteq \{b_{R,r}\}$. Then for each type of product $R \in \{AS_i\}$ define the Stochastic Frontier model:
$$\overline{Y}_R^A(t) = f_R^A(\{\overline{PS}_j^A(t)\},\{\bar{b}_{R,r}\},t) \cdot E_R^A \cdot e^{v_R} \tag{12}$$
4. For each type of product $R \in \{AS_i\}$ perform the simultaneous estimation of parameters $\bar{b}_{R,s}, E_R, v_R$ (using maximum likelihood estimators [2], Bayesian analysis [4] or any other technique) to fit the data for all research institutes:
$$R^A \approx f_R^A(\{\overline{PS}_j^A(t)\},\{\bar{b}_{R,r}\},t) \cdot E_R^A \cdot e^{v_R} \text{ for every } A \tag{13}$$
5. If estimated random error v_R is sufficiently small for each $R \in \{AS_i\}$ (meaning that the model describes the reality well), then we select the subset of indicators $\{\overline{\overline{PS}}_j^A(t)\} \subseteq \{\overline{PS}_i^A(t)\}$ that have sufficiently high values of corresponding weights $\{\bar{\bar{b}}_{R,j}\} \subseteq \{\bar{b}_{R,i}\}$.
6. One may repeat the procedure starting with step 3 to find even better subset of indicators.
7. One may repeat the procedure starting with step 2 to find the model that fits the data even better.

(3.1.5) In order to find the relation between financing in science and important indicators of the quality of pure science we use the SFA as earlier. We repeat also the whole procedure of testing for the best form of cost function and the best subset of quality of science indicators. It is very important to find a model that would fit well to reality and have relatively small number of input indicators. The main effect of this algorithm is the model (the form of the model and the corresponding sets of indicators).

The SFA model of costs is as follows:
$$Y_C(t) = f_C\left(\left\{\overline{\overline{PS}}_j^A(t)\right\},\{b_c\}\right) \cdot E_C \cdot e^{v_C} \tag{14}$$
where

$Y_C(t)$ - output indicator (total cost of maintenance)

$f_C(\bullet)$ - form of cost function

$\{\overline{\overline{PS}}_j^A(t)\}$ - set of input indicators (a subset of important pure science indicators)

$\{b_c\}$ - vector of parameters of the model (where b_0 is a scaling parameter)

$E_C^A \in [0,1]$ - effectiveness of production (random factor); it has the same type of distribution, variation and mean are estimated later on.

v_c - random error; it has the same distribution for every research institute (only one type is given, mean and variation will be estimated).

As earlier, we estimate all parameters and random factors of the model.

If some of important indicators of pure science quality $\overline{\overline{PS}}_{j_0}^{A}(t)$ are connected only with parameters $\{b_c\}$ with very low value in the best fitting model, then we eliminate $\overline{\overline{PS}}_{j_0}^{A}(t)$.

(3.1.6) We assume that the best prize for scientific work of each quality is its value, thus, E_A^C is effectiveness of institute.

Summary

(3.1.7) The crux of the study lies in the fact that we have proposed the method of choosing an optimal set of indicators of the quality of pure science by using the SFA. The estimation of frontier costs is a standard method of measuring efficiency for units with multiple outputs. However, this methodology might be too simple because there is always a problem of too many indicators (it is hard to choose the right value of parameters, because indicators are mutually intertwined) or, alternatively, too few indicators (a poor description of reality).

(3.1.8) Elimination of redundant indicators is important, otherwise it might lead to the reduction of random error without significant rise in explanatory power. This results from a high dependency between input indicators and facts. Additionally, if we take input indicator independent of output indicators, we will almost always have non zero weights connected with them in SFA model.

(3.1.9) First, it has to be indicated that the second stage of the procedure does not take into account parameters from the first one (they are of measure importance in relation to input indicators). This problem is quite complicated, because if we just take each parameter to power sum (or weighted average) of its weights, then after estimation we will achieve the same result as without powers. The simplest solution is to assume that cost function exhibits constant returns to scale (so the sum of parameters without rescaling parameters equals one) and take indicators to proper power as input indicators. This solution can greatly increase a random error of the model.

(3.1.10) The next problem is connected with different forms of the production function in the first step. The solution is simple – if some of the parameters describing effects of science on the economy/population are modelled well only by Transcendental Logarithm [3], then we should create additional artificial parameters – one indicator to exponentiate the log of another one, in the case of linear production form we take exponents of input and output indicators.

(3.1.11) There is a risk that the actual quality of science depends on some other indicators, but this dependency takes other forms (not tested in our model). Unfortunately, it is impossible to take into account all substantial factors with the right form of dependency.

(3.1.11) Should any initial grouping of parameters be introduced? If we have e.g. indicators of the number and the quality of PhDs, why not to multiply the former by the latter or to exponentiate them? These questions are essential, but cannot be answered without testing the data.

4 Conclusion and further research

4.1 Conclusion

(4.1.1) The concept of time-delay systems has been applied with the aim of modelling the relation between financing scientific research and the quality of science. The specific form of Equation (1) has been put forward and the problem of estimation of model parameters has been discussed.

(4.1.2) The optimal control problem has been posed as a problem of finding the optimal strategy subject to some given constraints. We have also proposed two examples of such an optimal control problem, which showed the capability of the method for the purpose of rationalizing funds on scientific research..

(4.1.3) Introduced in Chapter 2.3, the more advanced model includes the effect of the evolution of scientific workforce and its influence on the evolution of quality of science. This modelling can be further extended, e.g. combining with the idea described in (4.2.2).

(4.1.4) Proposed was a competitive method based on the Stochastic Frontier Analysis. This technique might allow choosing the appropriate model for a given data set. Nevertheless, it has many drawbacks in a present form, yet, they might be overcome further on.

4.2 Further research

(4.2.1) We suggest further exploration of both proposed methodologies, still, we believe that the method described in Chapter 2 is more promising, since it gives the opportunity to seek for some optimal funding scenarios. Application of the models to the real data would conclusively show the usefulness of each method and possible directions of its development.

(4.2.2) One of the ways to model the influence of the research on the widely understood economy is to monitor the transfer of human capital between these two branches. There is a possibility to measure the respective fractions of PhD holders and delays in years between their graduation and the moment they undertake R&D projects in industry. Since the data on PhD graduates and R&D projects in industry is gathered in the OPI databases, the processes can be given a specific and quantitative meaning. Additionally, the level of finances for “granty celowe” (special purpose grant)/technology transfer grants can serve as a measure of the influence of research on industry. Incorporation of these two measures may contribute to the further analyses on the considered topic.

Bibliography

[1] Pitcher, A., *Mathematical modelling and optimal control of constrained systems*, PhD Thesis, University of Oxford, 2009

[2] Greene, W.H., Maximum Likelihood Estimation of Econometric Frontier Functions, *Journal of Econometrics*, Vol. 13(1), pp. 27-56, 1980

[3] Baltagi, B., *A Companion to Theoretical Econometrics*, Blackwell Publishing Ltd., 2001

[4] Van den Broeck, J., et al., Stochastic frontier models: A Bayesian perspective, *Journal of Econometrics*, Vol. 61(2), pp. 273-303, 1994