International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 6 Issue 4 Ver. II ǁ 2018 ǁ PP. 10-19

The Universal Law Of The Scientific Productivity Distribution In Academic Institutions

JiříSouček, Martin Souček
Charles University, Faculty of ArtsU Krize 8, Praha 5, 15800, Czech Republic
Charles University, Faculty of ArtsU Krize 8, Praha 5, 15800, Czech Republic
Corresponding Author: JiříSouček

Abstract:*In this paper we report on the discovery of the universal law of the scientific productivity distribution in academic institutions.We are giving a mathematical model for the function describing the (aggregated) scientific productivity of a research institution. The mathematical model uses the function based on the modified ellipse. The empirical data used in the paper are taken from a very high quality database run by the Czech government describing the scientific productivity distributions in Charles university in Prague and in their faculties. These data are used to show a high accuracy coincidence between empirical data and their mathematical model (correlation bigger than 0.99). We think that this is (probably) the first case of a high-accuracy social law and we believe that this discovery could be a starting point for the discovery of other high-accuracy social laws.*

Date of Submission: 06-04-2018 Date of acceptance: 21-04-2018

I. The Introduction

In information science, there are several basic bibliometric laws, e.g., the Benford law, the Zipf law, or the Pareto distribution (see [1] – [5]). These rules have a character of universal laws, since they describe to a certain extent a nature of relations reflecting specific social situations.

A general aim of this paper is to study the distribution of the research productivity[1] of scientists at some research institutions (universities or their faculties, research institutes, or other scientific institutions). In particular, we are going to study the distribution of research productivity distribution at Charles university in Prague (UK) and some of their faculties.

We would like to understand how the number of research results created by x percent of the most productive scientists of a given research institution depends on the value of x and to create a mathematical model for this dependence. Based on empirical data extracted from Czech government database in the field of science and research, we have discovered such a model and we are showing below that the mathematical model has a very high correlation with the given experimental data.

The inspiration is coming from the Pareto law claiming that roughly 20 % of the most productive scientists of a given research institution produce 80 % of results. Looking to empirical data, we have found the relation which is similar to the Pareto's law, 20 % of the most productive scientisists create roughly 75 % of results.

In general, it is possible to ask question whether the law of scientific productivity has a property of scale invariance. Based on empirical data studied in the paper, the unambiguous answer is **no**. The law of scientific productivity does not belong to the group of power laws ($y = x^{\alpha}$).

The empirical data used in the paper are describing results for Charles university and for its faculties. [2]

[1] We use the term "research productivity" in the sense "research production". Thus perhaps the better expression would be "the research production" but we feel this expression as rather strange. So we shall use the more common term "the research productivity" even if this tterm is slightly imprecise.

[2] We plan to extend the comparision between empirical data and the mathematical model to include further universities in Czech republic, institutes of Academy of Sciences of Czech Republic, and further scientific and research institutions.

We have discovered the mathematical model which successfully describethe empirical dependence of the aggregated productivity depending on the variable xin the form of a function of the type of modified ellipse. The function depends on two parameters α and β(which often belongs to the interval between 2.3 and 2.7)

$$F_{\alpha,\beta}(x) = (1 - (1 - x)^{\alpha})^{1/\beta} \quad .$$

At this first stage of our study we assume that the parameters are equal α = β, so that we shall use the function

$$F_{\alpha}(x) = (1 - (1 - x)^{\alpha})^{1/\alpha} \quad .$$

We shall first find the optimal value of the parameter α for the whole Charles University, and then for its individual faculties. There will be only a very small difference between results for various faculties, hence the mathematical model described in the paper is universal (with the varying constants α, β).

Our **general conjecture**(for a general institution) is the following:

For each institution the aggregated empirical distribution of the productivity of its workers can be modelled by the function $F_{\alpha,\beta}(x)$ for appropriate parameters α, β .

We propose that this conjecture is a general high-precison social law of the productivity.

We shall test this conjecture on academic institutions, in fact on the faculties of the Charles university and on a University as a whole.

The **main result** of the paper is the **confirmation** of the very high correlation between empirical data and the suggested mathematical model with the **level of correlation higher than 0.994.**

We have found that the empirical distribution of the scientific productivity is basically almost the same for different faculties and it always hasa character of a function $F_{\alpha}(x)$. This is quite remarkable, because it implies that there should be a general mechanism leading to such a behaviour. (We have tested very different types of faculties – from natural sciences to humanities.)

We do not have any knowledge about the essence of this mechanism. Our results shows that there should exists some general sociological rules (social forces) behind the fact that for a majority of institutions such as faculties of a university, the distribution of the scientific productivity of scientists is to a high extent uniform. Details of the distribution can (to a certain level) depend on a type of a faculty or on professional orientation of the institution (natural sciences, technical sciences, medicine, humanities, social sciences, etc.) and also on the type of the used evalution metrics. These details are studied here and they will be studied in future papers.

There is a question how the resulting empirical distribution depends on the metric used in the evaluation of results. This is an important question which will be studied in the following studies. Our conjecture is that the general form of the dependence $F_{\alpha,\beta}$ will be the same as above but the values of α, β will depend on the metrics used in the evaluation.

II. The Description Of Sets Of Data

To start the investigation, it is necessary to decribe precisely data used in the study. This is the content of this section. We shall consider certain research institution having sciences and research in its agenda (typically faculties, schools, or institutes of a university, institutes of Academy of Sciences, or another research organizations).

Every such institution contains a group of scientists producing research results. In the paper, we shall study a set of data obtained from the open source from the Research and Development Information System of the Government Office of the Czech Republic (IS VaVaI, see [7]).

Our first task is to define precisely who is considered to be a research worker in a given institution. Our definition is: a research worker is a worker of the institution who gained in a given

period of time (standardly a 5-years period) a number of points bigger or equal to a chosen positive lower bound in the evalution of the research work (more details on the evaluation will be given below). It can be shown that results do not depend significantly on the value chosen for the positive lower bound for number of points in the definition above. So we shall work with the lower bound equal to 1.

To go further, it is necessary to choose a suitable evaluation of research results. In our approach, we are using the evaluation of the company Scimetrics Praha, Ltd., which is not significantly different (at the level of aggregation used in this paper) from the evaluation RIV used by the Government Office of the Czech Republic.

In principle, it can be true that the results of the study are robust (i.e. almost independent of the evaluation method) but further empirical study of such **conjecture** is postponed to another paper.

Research productivity of a given author is a sum of evaluations of all research publications of the author (always normalized, i.e., divided by number of authors). Research productivity is usually expressed in points (it depends on methodology used in the evaluation), We are using the methodology developed by thecompany Scimetrics Praha, Ltd. mentioned above.

The next thing to be specified is the time interval used in the study. We are using 5 years time window (2010-2014). It is clear that the law studied in this paper can depend on the length of the time interval. It is also possible that the law could depend on the choice of the time window with a given length. We postpone a study of both these questions to future papers. The law itself may depend also on a character of the given institution. We shall give an example of such dependence below.

So let us consider a group of n_{max}scientists ordered into a sequence with the decreasing research productivity (production). Suppose that the scientist on the position nin the sequence has the research productivity equal top_n, where n =1, ... , n_{max}. Choose a positive integer n between 1 andn_{max}. Let us create the aggregated productivity for the group of scientists with indices 1, …, n, .., n_{max}using the formula

$$Y_n = p_1 + \dots + p_n , \quad \text{where} \qquad n = 1, \dots , n_{max}.$$

The value$Y_n$describes the aggregated productivity of the subsequence of scientists with indices running from 1 to n (expressed in points). We replace the set of (absolute) indices $n = 1, \dots , n_{max}$ by its normalized version

$$x_n = n / n_{max} \ , \ (x_n>0, x_n\leq 1).$$

It is also useful to normalize variableY_n. It is clear that the value of Y_n isincreasing with respect to the variable x. Hence the maximum $Y_{max}$is the value ofYreached for x=1. The normalized variabley_n is then defined by

$$y_n = Y_n / Y_{max}.$$

In such a way, we can translate the available empirical data to a sequence of pairs (x_n ,y_n) for all values n= 1 , …, n_{max}. The main goal of the paper is to find a mathematical model for such data, i.e., to find a functiony = f(x) on the interval (0,1) with the property that values f (x_n) approximate values y_n for n= 1 , …, n_{max}in a best possible way. (We shall see that the mathematical model depends mildly on the character of the institution under the study.)

III. The Mathematical Model.

A study of available experimental data leads us to suggest the following mathematical model given by the functioncalled the modified ellipse. It can be described as follows. The left upper quadrant of the standard ellipse with the center in the point (n_{max}, 0) and with half-axesn_{max}andY_{max} is described by the equation (with axis n and Y)

$$(1 - n/n_{max})^2 + (Y/Y_{max})^2 = 1 .$$

After normalization, the equation of the ellipse above has the form of the circle with the center at the point (1,0):

$$(1 - x)^2 + y^2 = 1 .$$

Normalized form of the modified ellipse with the center at the point (1,0) is given by the equation

$$(1 - x)^{\alpha} + y^{\beta} = 1 .$$

We shall obtain the function (by calculating the dependence of y on x)

$$y = F_{\alpha,\beta}(x) = (1 - (1 - x)^{\alpha})^{1/\beta} .$$

We shall consider in this paper only the case where $\alpha = \beta$ .The values of the parameter α are in general greater then1.

The equation for the modified ellipse implies the following relation

$$y = F_{\alpha}(x), \qquad 0 < x < 1, \text{ where} F_{\alpha}(x) = (1 - (1 - x)^{\alpha})^{1/\alpha} .$$

We shall use this function $y = F_{\alpha}(x)$ to model the way how the (normalized) aggregated productivity of the first x % of scientists of the given institution dependson the value of x. It is a curve of a type of a deformation of an ellipse. We are getting different curves for different values of the parameter α.

We illustrate their behaviour below for $\alpha = 5$, $\alpha = 2.7$, $\alpha = 2.5$, $\alpha = 2$, $\alpha = 1$. The curves are drawn in turn from top to bottom, the last one corresponding to the value $\alpha = 1$ is just a line. The second curve (for $\alpha = 2.7$) describes the model curve for the case of Charles university (UK), the third curve (for $\alpha = 2.5$) describes the model curve of an average faculty of UK. The forth curve ($\alpha = 2$) is just an ellipse. For the last case ($\alpha = 1$), we get a linear function (it corresponds to the case when all scientists contribute to the aggregated value by the same amount,i.e., the productivity p is a constant function of the variable x).

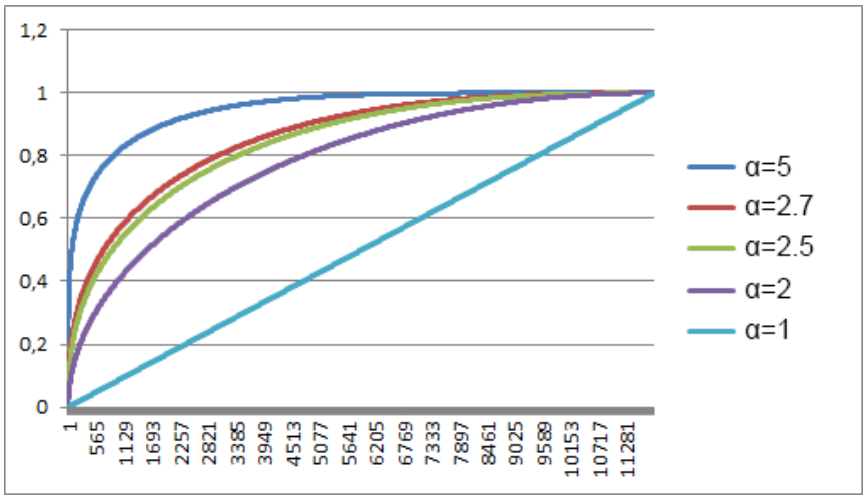


Chart 1- Different curves for different alpha parameter value

IV. The Correlation Between Empirical Data And The Proposed Mathematical Model

First we are going to study the correlation between the empirical dependence y = f (x) and the model functiony = F_{α} (x) for the whole Charles University. We shall optimize the correlation with respect to the value of the parameter α and we shall consider the optimal value of the parameter as a characteristic property of the given institution. Based on detailed tests of various possible choices, the resulting optimal value of the parameter is

$$\alpha_{\text{Charles University}} = 2.7,$$

which is well visible from the graph given below:

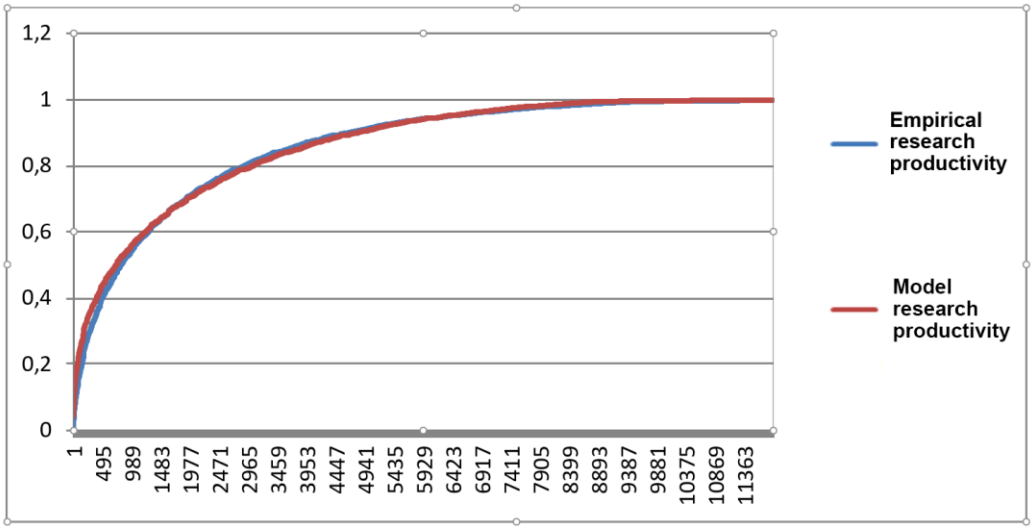


Chart 2-empirical data for UK (blue), mathematicalmodel for data for UK (red)

On the Chart 2 we can see blue curve - empirical data and red curve –the mathematical model $F_{2.7.}$ The x axis is parametrized using the ordinal number of a scientist (ordered from the most productive to less productive ones).

The two obtained curves have a very high measure of correlation (computed using the statisticalsoftware) given bycorrelation = 0,998424,which is also nicely illustrated in the graph above. The value of the correlation obtained and the graph reproduced above shows that there is a very high coincidence of the empiric data and the suggested mathematical model.

The whole situation can also been interpreted as an example of a **quite unusual measure of accuracy** in the modelling of social processes.A real cause of the distribution of scientific productivity inside a given institution is aresult of productivity in the whole society as well as social forces influencing the choice of working positions of scientists. But the full explanation of the observed distribution is not known to us.

There is a question, if the law of scientific productivity has a property of scale invariance. The present study shows without any doubts that the law of scientific productivity **does not have this property**, it cannot be described by a power law ($y = x^{\alpha}$).

A comparision of the empirical data with a possible approximation using the power law is shown in the graph below, where it is possible to find both empirical data and the graph of a power function for an exponent giving the best correlation – it is the power law $y = x^{0.19}$.

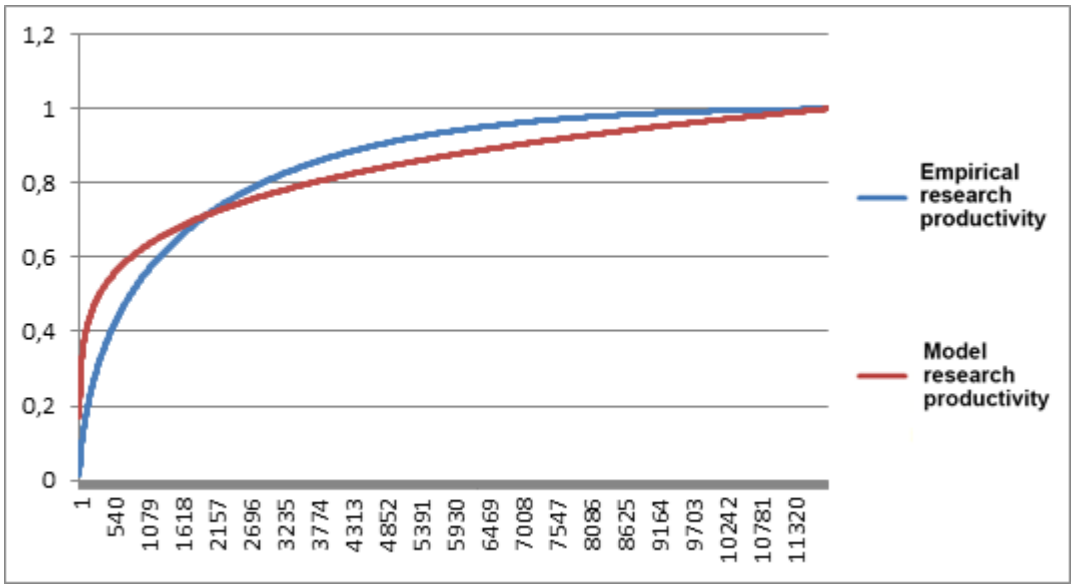


Chart 3- empirical data for UK (blue), model data –the best power function (red)

It is clearly visible that the approximation using the power function is not good, the correlation is quite low. So we can deduce that the studied empiric data **cannot be modelled by a power function**. In the next part, we are going to discuss the distribution of scientific productivity at individual faculties of Charles university.

V. Further Data

Now we are going to study empirical behaviour of scientific productivity for individual faculties of Charles university and to describe their behaviour using the model based on the modified ellipse for a suitable value of the parameter α. For each faculty, we are going to find the optimal value of the parameterα.

We have chosen nine big faculties of Charles university and we summarize in the table below (optimal) values of the parameter α as well as values describing the correlation between empirical data and their mathematical model.

Code	Abbrev.	Alfa	correlation	Full name
11110	1LF	2.4	0.998	Charles university / Firstfaculty of medicine
11120	3LF	2.6	0.999	Charles university / Third faculty of medicine
11130	2LF	2.4	0.998	Charles university / Second faculty of medicine
11210	FF	2.3	0.996	Charles university / Faculty of arts
11220	PF	2.5	0.994	Charles university / Faculty of law
11230	FSV	2.4	0.997	Charles university / Faculty of social sciences
11240	FHS	2.6	0.994	Charles university / Faculty of humanities
11310	PrF	2.6	0.996	Charles university / Faculty of natural sciences
11320	MFF	2.5	0.998	Charles university / Faculty of mathematics and physics
	UK_OSO	2.7	0.998	Charles university

The graphs for all nine faculties are presented in Appendix.

All correlations are higher then 0.994. One can see that the lowest correlation 0.994 holds for Faculty of law and for Fakulty of humanities. Four faculties and the university have the correlation higher than 0.998. The necessary condition for such a high correlation is the high-precision of the empirical data.

VI. Consequences

The table summarizing data for the graph above showing the distribution of scientific productivity for Charles university implies the following modifications of the Pareto'srule:
20-75 ,i.e., 20% of scientists create 75% of the overall scientific production,
25-80 ,i.e., 25% of scientists create 80% of the overall scientific production,
50-94 ,i.e., 50% of scientists create 94% of the overall scientific production,
75-98 ,i.e., 75% of scientists create 98% of the overall scientific production.

So it is possible to say that the Pareto's rule remains (in principle) valid. The distribution of scientific productivity, however, is not described by a power law **but by the law of the modified ellipse.**

Distributions obtained for individual faculties differ only mildly, differences are not significant. The value of the parameter α for individual faculties is between 2.3 and 2.6, while for the whole university, the value is $\alpha = 2.7$.

The average value for faculties of UK is 2.5, which is smaller by 0.2 than the value for the whole Charles University. To understand the reason for this difference is the theme for a next study. On the other hand, there could be a systematic difference related to the size of the institution).

The difference between $F_{2.7}$ (UK) and $F_{2.5}$ (faculties of UK) can be seen when comparing the second and the third curve in the Fig. 1.

VII. Conclusions

Based on analysis of available empirical data for Charles University, we showed that the aggregated distribution of the scientific productivity of Charles University and of its faculties is with a high accuracy (correlation bigger than 0.994) described by the curve called the modified ellipse.

We analyzed data from nine biggest faculties of Charles University and we showed that the corresponding disctributionsare very similar for individual faculties as well as for the university.

We plan to study data for scientific productivity from different universities and institutes of Academy of Science of Czech Republic.

We conjecture that results will be very similar to those presented in the paper and that the law of scientific productivity formulated above can be considered as the universal law.

Our general conjecture was formulated in the introduction. We have successfully tested this conjecture on the empirical data from faculties of the Charles university at Prague.

Our results were based on very precise open sources data in the field of research and development in Czech Republic (the information system IS VaVaI run by the Government Office for Research, Development and Innovation) and on computations of evaluations of results made by the company Scimetrics Praha, Ltd.

The high quality of the data used in the paper is due in particular to the fact that the open sources data mentioned above contains very precise and (almost absolutely) unique identificators for persons, institutions and results involved.

Without such first-rate identificators, any analysis could only be rough and imprecise, while having high-quality data, it was possible to deduce quite precise results.

Note that deficiency in identificators for persons and institutions in the databases Web of Science and Scopus make impossible to use their data to analyse the scientific productivity. It is clear that our results are open to a lot of new questions, we shall consider some of them in future publications.

VIII. Appendix – Data For Faculties Of Charles University

Here we present graphs describing scientific productivity for nine (big) faculties of Charles University including also the web address where primary data for graphs can be found.
In the graphs below:

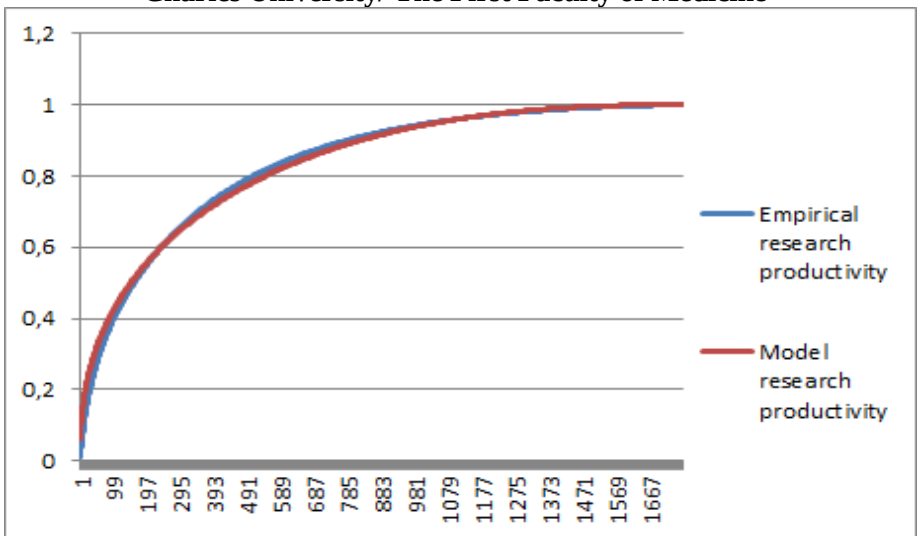
series– blue: represents empirical data,

series– red: represents model data (the curve F_α) for agiven faculty

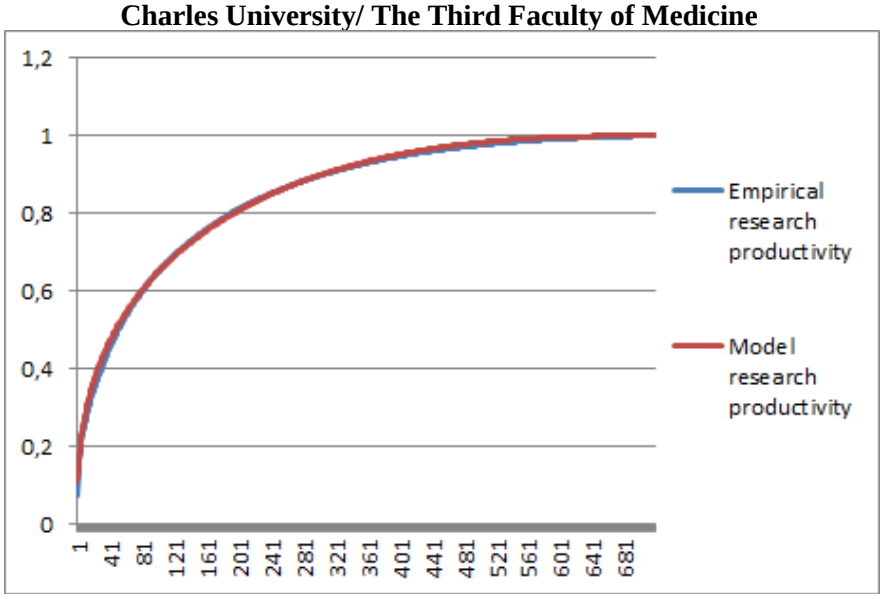
Coincidence of empirical and model data is hence represented by the coincidence between the first (blue) and the second (red) curve.

An approximate coincidence of empirical data for the whole Charles University and for their individual faculties is represented by (approximate) coincidence between the first (blue) curve for the given faculty and the third (green) curve for the whole university.

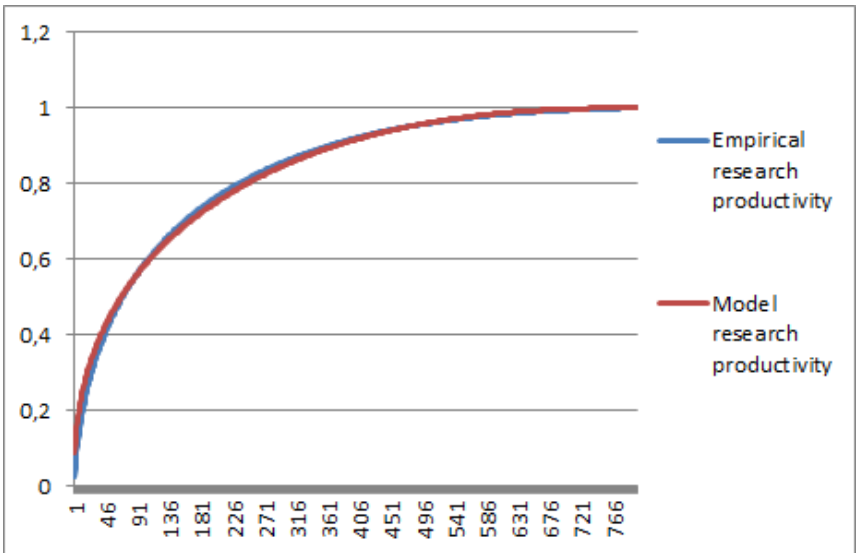
Charles University/ The First Faculty of Medicine



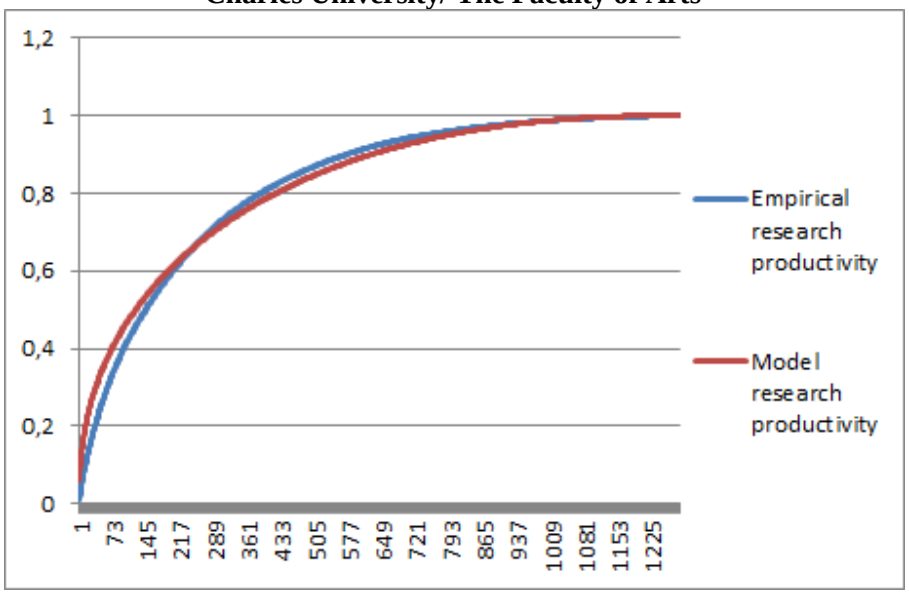
Charles University/ The Third Faculty of Medicine



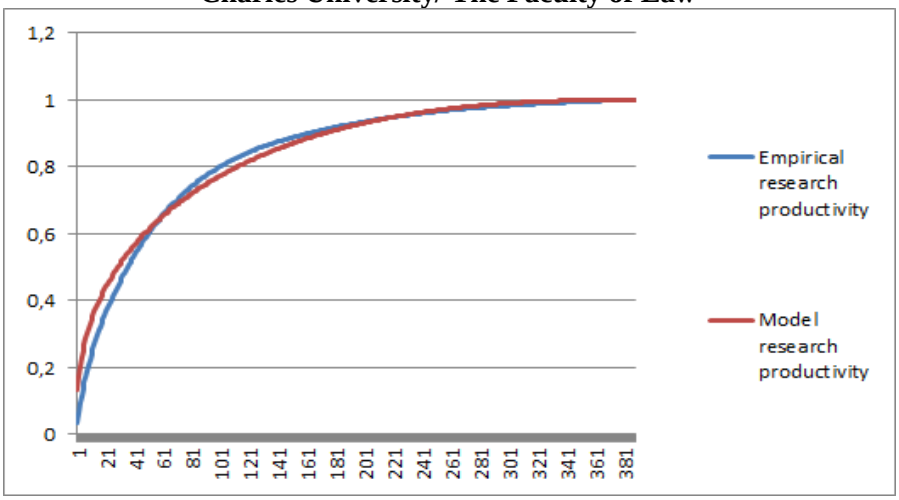
Charles University/ The Second Faculty of Medicine


1,2
1
0,8
0,6
0,4
0,2
0
1
46
91
136
181
226
271
316
361
406
451
496
541
586
631
676
721
766
Empirical research productivity
Model research productivity

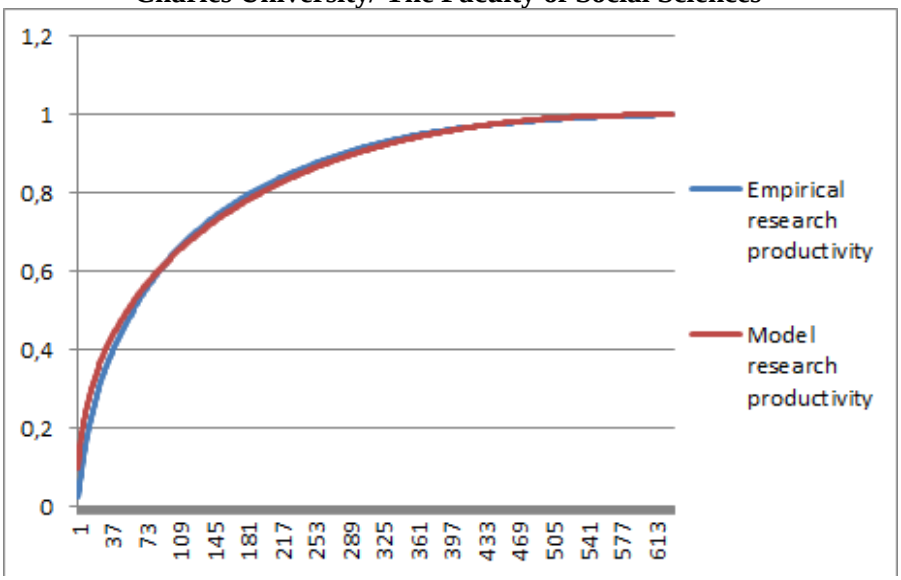
Charles University/ The Faculty of Arts



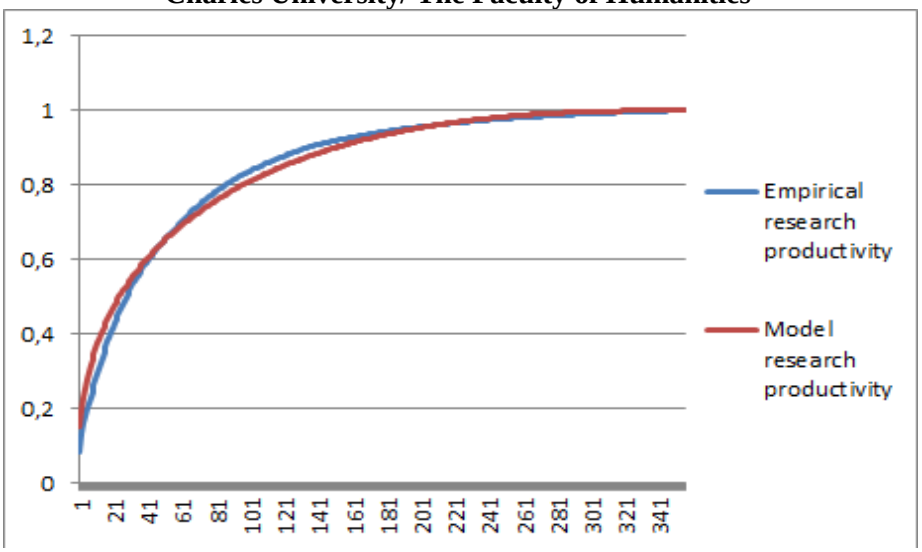
Charles University/ The Faculty of Law



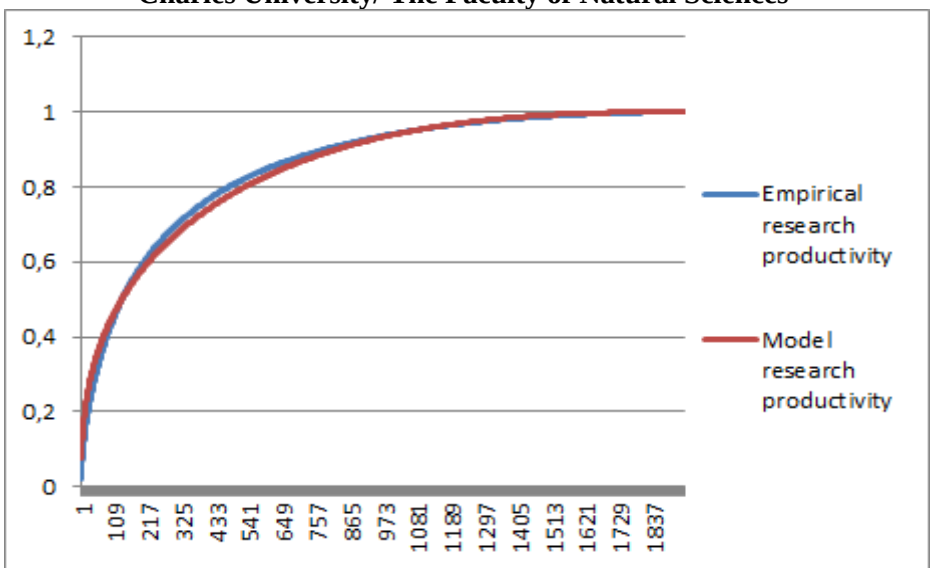
Charles University/ The Faculty of Social Sciences



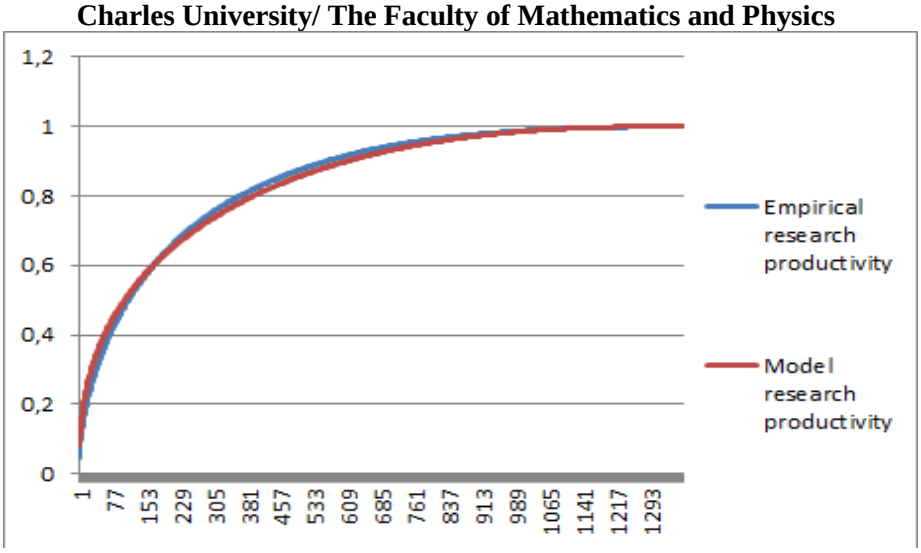
Charles University/ The Faculty of Humanities



Charles University/ The Faculty of Natural Sciences



Charles University/ The Faculty of Mathematics and Physics



The primary data for the graphs presented above can be found on the web address given in [6].

References

[1]. Percy Deift, "Universality for physical and dynamical systems", International Congress of Mathematicians, Vol. I, 125-152, Eur. Math. Soc., Zurich, 2007

[2]. Theodore P. Hill, "Base-invariance implies Benford's law", Proc. Amer. Math. Soc. **123** (1995), 887-895

[3]. MEJ Newman, "Power laws, Pareto distributions and Zipf's law", Contemporary physics, 2005 - Taylor & Francis

[4]. WJ Reed, "The Pareto, Zipf and other power laws", Economics Letters, Volume 74, Issue 1, 20 December 2001, Pages 15–19

[5]. **Aaron Clauset**, **CosmaRohillaShalizi** and **M. E. J. Newman** , "Power-Law Distributions in Empirical Data", *SIAM Rev.*, 51(4), 661–703. (43 pages)

[6]. The primary data for our study can be found athttps://goo.gl/Vg8fi1

[7]. Vyzkum.cz: Výzkum a vývoj v ČR. [online]. 2018 [cit. 2018-03-27]. Accessible from: www.vyzkum.cz

JiříSouček."The Universal Law Of The Scientific Productivity Distribution In Academic Institutions." International Journal of Research in Engineering and Science (IJRES), vol. 06, no. 04, 2018, pp. 10–19.