Qualitative and Quantitative Analysis of Scientific and Scholarly Communication

Series editors

Wolfgang Glänzel, Katholieke Univeristeit Leuven, Leuven, Belgium
Andras Schubert, Hungarian Academy of Sciences, Budapest, Hungary

More information about this series at http://www.springer.com/series/13902

Nikolay K. Vitanov

Science Dynamics and Research Production

Indicators, Indexes, Statistical Laws and Mathematical Models

Springer

Nikolay K. Vitanov
Institute of Mechanics
Sofia
Bulgaria

and

Max-Planck Institute for the Physics of
Complex Systems
Dresden
Germany

ISSN 2365-8371 ISSN 2365-838X (electronic)
Qualitative and Quantitative Analysis of Scientific and Scholarly Communication
ISBN 978-3-319-41629-8 ISBN 978-3-319-41631-1 (eBook)
DOI 10.1007/978-3-319-41631-1

Library of Congress Control Number: 2016944335

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my parents and teachers, who helped me
to find my way through the mountains
and valleys of life.

Preface

He who sees things grow from the beginning will have the best view of them

Aristotle

There is a variety of books on the topic of the "science of science," books, that are devoted to the social and economic aspects of science [1–8]; books devoted to innovation and technological change [9–11]; books devoted to the study of models of science dynamics [12–14]; books devoted to studies in the area of scientometrics, bibliometrics, informetrics, webometrics, scientometric indicators and their applications [15–36]; and especially books devoted to citations and citation analysis [37, 38]. The goal of this book is different from those of most of the books mentioned above, because this book is designed as an introductory textbook with elements of a handbook. Its goal is to introduce the reader to two selected areas of the science of science: (i) indicators and indexes for assessment of research production and (ii) statistical laws and mathematical models connected to science dynamics and research production. The introduction is from the point of view of applied mathematics (i.e., no proofs of theorems are presented).

In the course of time, science becomes more and more costly to produce, and because of this, the dynamics of research organizations and assessment of research production are receiving increasing attention. As a consequence of the increasing costs, many national funding authorities are pressed by the governments for better assessment of the results of their investment in scientific research. And this pressure tends to increase. Because of this, interest in objectively addressing the quality of scientific research has increased greatly in recent years. One observes an increase in the frequency of the formation and action of various groups for quality assessment of scientific research of individuals, departments, universities, systems of institutes, and even nations.

Mathematics may provide considerable help in the assessment of complex research organizations. Numerous indicators and indexes for the measurement of performance of researchers, research groups, research institutes, etc. have been developed. Numerous models and statistical laws inform us about specific modalities of the evolution of scientific fields and research organizations. We shall discuss below some of these indicators, indexes, statistical laws, and mathematical models.

Let us consider the potential readers of this book from the point of view of their knowledge about science dynamics and the tools for evaluation of research production. We shall see in Chap. 4 that rankings often lead to a power-law distribution and to an effect called the concentration–dispersion effect: If we have components of some organization, and these components own units, then often large numbers of units are concentrated in a small percentage of the components (concentration), and the remaining units are dispersed among the remaining larger number of components (dispersion). Let us assume that this effect is valid for the readers of this book (the components) with respect to their knowledge about science dynamics (measured in units of research articles read on this subject). Then there may be a concentration of much knowledge about dynamics of science and features of research production in a small group of highly competent readers. The concentration–dispersion effect helps us to identify target groups of readers as follows.

- **Target group 1**: *Readers who want to understand the dynamics of research organizations and assessment of research production but don't have knowledge about the dynamics of such organizations and/or about the tools for assessment of research production.*
  This group is very important, since every researcher and every manager of a research organization was a member of this group at least at the beginning of his/her career. In order to make this book more valuable for this group of readers, we discuss a large number of topics on a small number of pages, and the level of mathematical difficulty is kept low. The presence of numerous references allows us to achieve this degree of compactness.
- **Target group 2**: *Readers who (i) have some knowledge in the area of theory of science dynamics, (ii) have some practice in the assessment of research, and (iii) want to increase their knowledge about science dynamics and assessment of research.*
  This group of intermediate size is quite important, since large number of researchers and managers belong to it. I hope that the part of the book devoted to models will be of interest to the practitioners, and that the discussions of concepts and results from their practical implementation will be of interest to theoreticians.
- **Group 3**: *Very experienced researchers and practitioners in the areas of science dynamics and assessment of research production.*
  This relatively small group of researchers is very competent and has much knowledge. I hope, however, that this book will also be of interest to such readers as a collection of tools and concepts about the evaluation of research production and the dynamics of research organizations, and as an applied mathematics point of view on the features of such organizations.

The positioning of this book as an introduction to the large field of the mathematical description of science dynamics and to quantitative assessment of research production determined the choice of the concepts and models discussed and led to the following features:

- A relatively large number of mathematical models, concepts, and tools are discussed. The goal of this is to provide the reader with an impression and basic knowledge about the huge field of models of science dynamics and about the even larger field of research on indicators and indexes for assessment of research production. Nevertheless, the number of discussed models is small in comparison to the number of existing models. Thus many classes of models, e.g., network models of research structures, are not discussed in detail. This is compensated by numerous references.
- The focus of the book is on the quantitative description of science dynamics and on the quantitative tools for assessment of research production. Because of this, a significant mathematical arsenal, especially from the area of probability theory and the theory of stochastic systems, was used. Nevertheless, many complicated mathematical models were omitted, but after studying the material of the book, the interested reader should have no difficulty in understanding even the most complicated models.
- About 1,200 references are included in the book. This allowed me to keep the size of the book compact, using the feature of references as a compressed form of research information. By means of the numerous references, the reader may quickly obtain a large quantity of additional information about the corresponding topic of interest directly from sources that represent the original points of view of experienced researchers.

The book consists of three parts. The first part of the book is devoted to a brief introduction to the complexity of science and to some of its features. The triple helix model of a knowledge-based economy is described, and scientific competition among nations is discussed from the point of view of the academic diamond. The importance of scientometrics and bibliometrics is emphasized, and different features of research production and its evaluation are discussed. A mathematical model for quantification of research performance is described.

The second part of the book contains a discussion of the indicators and indexes of research production of individual researchers and groups of researchers. It is hard to find an alternative to peer review if one wants to evaluate the quality of a paper or the quality of scientific work of a single researcher. But if one has to evaluate the research work of collectives of researchers from some department or institute, then one may need additional methodology, such as a methodology for analysis of citations and publications. The building blocks of such methodology as well as selected indicators and indexes are described in this book, and many examples for the calculation of corresponding indexes are presented. In such a way, the reader may observe the indexes “in action,” and he/she can get a good impression of their strengths and weaknesses. An important goal of this part is to serve as a handbook of useful indicators and indexes. Nevertheless, some discussion about features of indexes is presented. Special attention is devoted to the Lorenz curve and to the definition of sizes of different scientific elites on the basis of this curve.

The third part of the book is devoted to the statistical laws and mathematical models connected to research organizations, and the focus is on the models of research production connected to the units of information (such as research publication) and to units of importance of this information (such as citations of research publications). Numerous non-Gaussian statistical power laws of research production and other features of science are discussed. Special attention is devoted to the application of statistical distributions (such as the Yule distribution, Waring distribution, Poisson distribution, negative binomial distribution) to modeling features connected to the dynamics of research publications and their citations. In addition, deterministic models of science dynamics (such as models based on concepts of epidemics and other Lotka–Volterra models) and models based on the reproduction–transport equation and on a master equation, etc., are discussed.

Several concluding remarks are summarized in the last chapter of the book.

In the process of writing of a book, every author uses some resources and discusses different aspects of the text with colleagues. I would like to thank the Max-Planck Institute for the Physics of Complex Systems in Dresden, Germany, where I was able to use the scientific resources of the Max-Planck Society. In fact, two-thirds of the book was written in Dresden. I would like to thank personally Prof. Holger Kantz, of MPIPKS, for his extensive support during the writing of the book, as well as Prof. Peter Fulde for extensive advice about practical aspects of science dynamics and research management. I would like to thank also two COST Actions: TD1210 "Analyzing the dynamics of information and knowledge landscapes—KNOWeSCAPE" and TD1306 "PEERE" for the possibility of numerous discussions with leading scientists in the area of scientometrics and evaluation of scientific performance. I would like thank Dr. Zlatinka Dimitrova and Kaloyan Vitanov for countless discussions on different questions connected to the book and for their help in the preparation of the manuscript. Many thanks to the Springer team and especially to Dr. Claus Ascheron for their excellent work in the process of preparation of the book. Finally, I would like to thank the (wise) anonymous reviewer, who advised me on how to arrange the text. That was useful indeed.

Sofia and Dresden Nikolay K. Vitanov

References

1. J.D. Bernal, *The Social Function of Science* (The MIT Press, Cambridge, MA, 1939)
2. V.V. Nalimov, *Faces of Science* (ISI Press, Philadelphia, 1981)
3. G. Böhme, N. Stehr (eds.), *The Knowledge Society* (Springer, Netherlands, 1986)
4. M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Throw, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* (Sage Publications, London, 1994)

5. E. Mansfield, *Industrial Research and Technological Innovation: An Econometric Analysis* (Norton, New York, 1968)
6. W. Krohn, E.T. Layton, Jr., P. Weingart, *The Dynamics of Science and Technology* (Reidel, Dordrecht, 1978)
7. M. Hirooka. *Innovation Dynamism and Economic Growth. A Nonlinear Perspective* (Edward Elgar Publishing, Cheltenham, UK, 2006)
8. P.A.A. van den Besselaar, L.A. Leydesdorff, *Evolutionary Economics and Chaos Theory: New Directions in Technology Studies* (Frances Pinter Publishers, 1994)
9. H. Grupp (ed.), *Dynamics of Science-Based Innovation* (Springer, Berlin, 1992)
10. L. Girifalco, *Dynamics of Technological Change* (Van Nostrand Reinhold, New York, 1991)
11. H. Etzkowitz, *The Triple Helix: University-Industry-Government Innovation in Action* (Routledge, New York, 2008)
12. A.I. Yablonskii, *Mathematical Methods in the Study of Science* (Nauka, Moscow, 1986) (in Russian)
13. H. Small, *Bibliometrics of Basic Research* (National Technical Information Service, 1990)
14. A. Scharnhorst, K. Börner, P. van den Besselaar (eds.), *Models for Science Dynamics* (Springer, Berlin, 2012)
15. L. Leydesdorff, *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications* (DSWO Press, Leiden, 1995)
16. E. Garfield, *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (Willey, New York, 1979)
17. D. de Solla Price, *Little Science, Big Science* (Columbia University Press, New York, 1963)
18. A. Andres, *Measuring Academic Research. How to Undertake a Bibliometric Study* (Chandos, Oxford, 2009)
19. S.D. Haitun, *Scientometrics: State and Perspectives* (Nauka, Moscow, 1983) (in Russian)
20. S.D. Haitun, *Quantitative Analysis of Social Phenomena* (URSS, Moscow, 2005) (in Russian)
21. I.K. Ravichandra Rao, *Quantitative Methods for Library and Information Science* (Wiley-Eastern. New Delhi, 1983)
22. A.F.J. van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology* (North-Holland, Amsterdam, 1988)
23. Y. Ding, R. Rousseau, D. Wolfram (eds.), *Measuring Scholarly Impact* (Springer, Cham, 2014)
24. L. Egghe, R. Rousseau, *Introduction to Informetrics: Quantitative Methods in Library, Documentation, and Information Science* (Elsevier, Amsterdam, 1980)
25. M. Callon, J. Law, A. Rip, *Mapping of the Dynamics of Science and Technology* (McMillan, London, 1986)
26. L. Egghe, *Power Laws in the Information Production Process: Lotkaian Informetrics* (Elsevier, Amsterdam, 2005)
27. D. Wolfram, *Applied Informatics for Information Retrieval Research* (Libraries Unlimited, Westport, CT, 2003)
28. M. Thelwall, *Introduction to Webometrics: Quantitative Web Research for the Social Sciences* (Morgan & Claypool, San Rafael, CA, 2009)
29. K. Fisher, *Changing Landscapes of Nuclear Physics: A Scientometric Study* (Springer, Berlin, 1993)
30. T. Braun, E. Bujdoso, A. Schubert, *Literature of Analytical Chemistry: A Scientometric Evaluation* (CRC Press, Boca Raton, FL, 1987)
31. P. Ingwersen, *Scientometric Indicators and Webometrics and the Polyrepresentation Principle in Information Retrieval* (ESS Publications, New Delhi Bangalore, India, 2012)
32. B. Cronin, C.R. Sugimoto, *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (MIT Press, Cambridge, MA, 2014)
33. T. Braun, W. Glänzel, A. Schubert, *Scientometrics Indicators. A 32 Country Comparison of Publication Productivity and Citation Impact* (World Scientific, London, 1985)

34. H.F. Moed, W. Glänzel, U. Schmoch (eds.), *Handbook of Quantitative Science and Technology Research* (Springer Netherlands, 2005)
35. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
36. M.A. Akoev, V.A. Markusova, O.V. Moskaleva, V.V. Pislyakov, *Handbook of Scientometrics: Indicators for Development of Science and Technology* (University of Ural Publishing, Ekaterinburg, 2014) (in Russian)
37. B. Cronin, *The Citation Process. The Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984)
38. H. Moed, *Citation Analysis in Research Evaluation*. (Springer, Netherlands, 2005)

Contents

Part I Science and Society. Research Organizations and Assessment of Research

1 Science and Society. Assessment of Research 3
1.1 Introductory Remarks 4
1.2 Science, Technology, and Society 5
1.3 Remarks on Dissipativity and the Structure of Science Systems 7
1.3.1 Financial, Material, and Human Resource Flows Keep Science in an Organized State 7
1.3.2 Levels, Characteristic Features, and Evolution of Scientific Structures 8
1.4 Triple Helix Model of the Knowledge-Based Economy 10
1.5 Scientific Competition Among Nations: The Academic Diamond 11
1.6 Assessment of Research: The Role of Research Publications 12
1.7 Quality and Performance: Processes and Process Indicators 13
1.8 Latent Variables, Measurement Scales, and Kinds of Measurements 14
1.9 Notes on Differences in Statistical Characteristics of Processes in Nature and Society 17
1.10 Several Notes on Scientometrics, Bibliometrics, Webometrics, and Informetrics 20
1.10.1 Examples of Quantities that May Be Analyzed in the Process of the Study of Research Dynamics 21
1.10.2 Inequality of Scientific Achievements 23
1.10.3 Knowledge Landscapes 24
1.11 Notes on Research Production and Research Productivity 25
1.12 Notes on the Methods of Research Assessment 29
1.12.1 Method of Expert Evaluation 29
1.12.2 Assessment of Basic Research 31

1.12.3 Evaluation of Research Organizations and Groups of Research Organizations 33
1.13 Mathematics and Quantification of Research Performance. English–Czerwon Method 34
1.13.1 Weighting Without Accounting for the Current Performance 34
1.13.2 Weighting with Accounting for the Current Performance 35
1.13.3 How to Determine the Values of Parameters 36
1.14 Concluding Remarks 37
References 37

Part II Indicators and Indexes for Assessment of Research Production

2 Commonly Used Indexes for Assessment of Research Production 55
2.1 Introductory Remarks 55
2.2 Peer Review and Assessment by Indicators and Indexes 58
2.3 Several General Remarks About Indicators and Indexes 58
2.4 Additional Discussion on Citations as a Measure of Reception, Impact, and Quality of Research 61
2.5 The h-Index of Hirsch 63
2.5.1 Advantages and Disadvantages of the h-Index 64
2.5.2 Normalized h-Index 66
2.5.3 Tapered h-Index 67
2.5.4 Temporally Bounded h-Index. Age-Dependent h-Index 68
2.5.5 The Problem of Multiple Authorship. $\overline{h}$-Index of Hirsch and gh-Index of Galam 68
2.5.6 The m-Index 71
2.5.7 h-Like Indexes and Indexes Complementary to the Hirsch Index 72
2.6 The g-Index of Egghe 76
2.7 The i_n-Index 77
2.8 p-Index. IQ_p-Index 78
2.9 A-Index and R-Index 80
2.10 More Indexes for Quantification of Research Production 82
2.10.1 Indexes Based on Normalization Mechanisms 82
2.10.2 PI-Indexes 83
2.10.3 Indexes for Personal Success of a Researcher 84
2.10.4 Indexes for Characterization of Research Networks 87
2.11 Concluding Remarks 88
References 89

3 Additional Indexes and Indicators for Assessment of Research Production . . . 101
3.1 Introductory Remarks . . . 101
3.2 Simple Indexes . . . 103
3.2.1 A Simple Index of Quality of Scientific Output Based on the Publications in Major Journals . . . 103
3.2.2 Actual Use of Information Published Earlier: Annual Impact Index . . . 105
3.2.3 MAPR-Index, T-Index, and RPG-Index . . . 105
3.2.4 Total Publication Productivity, Total Institutional Authorship . . . 108
3.3 Indexes for Deviation from a Single Tendency . . . 108
3.3.1 Schutz Coefficient of Inequality . . . 109
3.3.2 Wilcox Deviation from the Mode (from the Maximum Percentage) . . . 109
3.3.3 Nagel's Index of Equality . . . 110
3.3.4 Coefficient of Variation . . . 111
3.4 Indexes for Differences Between Components . . . 111
3.4.1 Gini's Mean Relative Difference . . . 111
3.4.2 Gini's Coefficient of Inequality . . . 112
3.5 Indexes of Concentration, Dissimilarity, Coherence, and Diversity . . . 113
3.5.1 Herfindahl–Hirschmann Index of Concentration . . . 113
3.5.2 Horvath's Index of Concentration . . . 114
3.5.3 RTS-Index of Concentration . . . 115
3.5.4 Diversity Index of Lieberson . . . 115
3.5.5 Second Index of Diversity of Lieberson . . . 116
3.5.6 Generalized Stirling Diversity Index . . . 117
3.5.7 Index of Dissimilarity . . . 118
3.5.8 Generalized Coherence Index . . . 118
3.6 Indexes of Imbalance and Fragmentation . . . 119
3.6.1 Index of Imbalance of Taagepera . . . 119
3.6.2 RT-Index of Fragmentation . . . 119
3.7 Indexes Based on the Concept of Entropy . . . 120
3.7.1 Theil's Index of Entropy . . . 121
3.7.2 Redundancy Index of Theil . . . 122
3.7.3 Negative Entropy Index . . . 122
3.7.4 Expected Information Content of Theil . . . 123
3.8 The Lorenz Curve and Associated Indexes . . . 123
3.8.1 Lorenz Curve . . . 123
3.8.2 The Index of Gini from the Point of View of the Lorenz Curve . . . 124
3.8.3 Index of Kuznets . . . 125
3.8.4 Pareto Diagram (Pareto Chart) . . . 125

3.9 Indexes for the Case of Stratified Data 126
3.10 Indexes of Inequality and Advantage 127
3.10.1 Index of Net Difference of Lieberson 127
3.10.2 Index of Average Relative Advantage. 128
3.10.3 Index of Inequity of Coulter 129
3.10.4 Proportionality Index of Nagel. 129
3.11 The RELEV Method for Assessment of Scientific Research Performance in Public Institutes 130
3.12 Comparison Among Scientific Communities in Different Countries 131
3.13 Efficiency of Research Production from the Point of View of Publications and Patents. 134
3.14 Indicators for Leadership 135
3.15 Additional Characteristics of Scientific Production of a Nation 136
3.16 Brief Remarks on Journal Citation Measures. 141
3.17 Scientific Elites. Geometric Tool for Detection of Elites 144
3.17.1 Size of Elite, Superelite, Hyperelite, ... 145
3.17.2 Strength of Elite 147
References 149

Part III Statistical Laws and Selected Models

4 Frequency and Rank Approaches to Research Production. Classical Statistical Laws 157
4.1 Introductory Remarks. 158
4.2 Publications and Assessment of Research 158
4.3 Frequency Approach and Rank Approach: General Remarks 161
4.4 The Status of the Zipf Distribution in the World of Non-Gaussian Distributions. 163
4.5 Stable Non-Gaussian Distributions and the Organization of Science. 165
4.6 How to Recognize the Gaussian or Non-Gaussian Nature of Distributions and Populations 166
4.7 Frequency Approach. Law of Lotka for Scientific Publications 168
4.7.1 Presence of Extremely Productive Scientists: $i_{max} \to \infty$ 169
4.7.2 i_{max} Finite: The Most Productive Scientist Has Finite Productivity. Scientific Elite According to Price. 170
4.7.3 The Exponent α as a Measure of Inequality. Concentration–Dispersion Effect. Ortega Hypothesis. . . . 172
4.7.4 The Continuous Limit: From the Law of Lotka to the Distribution of Pareto. Pareto II Distribution 174

4.8 Rank Approach 176
4.8.1 Law of Zipf 176
4.8.2 Zipf–Mandelbrot Law 177
4.8.3 Law of Bradford for Scientific Journals 178
4.9 Matthew Effect in Science 180
4.10 Additional Remarks on the Relationships Among Statistical Laws 182
4.11 On Power Laws as Informetric Distributions 184
References 189

5 Selected Models for Dynamics of Research Organizations and Research Production 195
5.1 Introductory Remarks 196
5.2 Deterministic Models Connected to Research Publications 197
5.2.1 Simple Models. Logistic Curve and Other Models of Growth 197
5.2.2 Epidemic Models 200
5.2.3 Change in the Number of Publications in a Research Field. SI (Susceptibles–Infectives) Model of Change in The Number of Researchers Working in a Field 201
5.2.4 Goffman–Newill Continuous Model for the Dynamics of Populations of Scientists and Publications 202
5.2.5 Price Model of Knowledge Growth. Cycles of Growth of Knowledge 204
5.3 A Deterministic Model Connected to Dynamics of Citations 205
5.4 Deterministic Models Connected to Research Dynamics 207
5.4.1 Continuous Model of Competition Between Systems of Ideas 207
5.4.2 Reproduction–Transport Equation Model of the Evolution of Scientific Subfields 210
5.4.3 Deterministic Model of Science as a Component of the Economic Growth of a Country 211
5.5 Several General Remarks About Probability Models and Corresponding Processes 214
5.6 Probability Model for Research Publications. Yule Process 217
5.6.1 Definition, Initial Conditions, and Differential Equations for the Process 218
5.6.2 How a Yule Process Occurs 218
5.6.3 Properties of Research Production According to the Model 219
5.7 Probability Models Connected to Dynamics of Citations 221
5.7.1 Poisson Model of Citations Dynamics of a Set of Articles Published at the Same Time 221

5.7.2 Mixed Poisson Model of Papers Published in a Journal Volume 224
5.8 Aging of Scientific Information 226
5.8.1 Death Stochastic Process Model of Aging of Scientific Information 226
5.8.2 Inhomogeneous Birth Process Model of Aging of Scientific Information. Waring Distribution 227
5.8.3 Quantities Connected to the Age of Citations 240
5.9 Probability Models Connected to Research Dynamics 241
5.9.1 Variation Approach to Scientific Production 241
5.9.2 Modeling Production/Citation Process 245
5.9.3 The GIGP (Generalized Inverse Gaussian–Poisson Distribution): Model Distribution for Bibliometric Data. Relation to Other Bibliometric Distributions 250
5.9.4 Master Equation Model of Scientific Productivity 252
5.10 Probability Model for Importance of the Human Factor in Science 255
5.10.1 The Effective Solutions of Research Problems Depend on the Size of the Corresponding Research Community 255
5.10.2 Increasing Complexity of Problems Requires Increase of the Size of Group of Researchers that Has to Solve Them 256
5.11 Concluding Remarks 257
References 261

6 Concluding Remarks 269
6.1 Science, Society, Public Funding, and Research 269
6.2 Assessment of Research Systems. Indicators and Indexes of Research Production 271
6.3 Frequency and Rank Approaches to Scientific Production. Importance of the Zipf Distribution 272
6.4 Deterministic and Probability Models of Science Dynamics and Research Production 273
6.5 Remarks on Application of Mathematics 274
6.6 Several Very Final Remarks 276
References 277

Index 281