Wiadomości Statystyczne. The Polish Statistician, 2024, *69*(8), 71–84 https://doi.org/10.59139/ws.2024.08.5
30 lat *Fundamentalnych zasad statystyki publicznej* / 30 years of the Fundamental Principles of Official Statistics

Unlocking insights: mastering quantitative methods for understanding social dynamics

Uwolnić potencjał. Doskonalenie metod ilościowych w celu zrozumienia dynamiki społecznej

1. Introduction

The Fundamental Principles of Official Statistics (FPOS), adopted by the United Nations Statistical Commission in 1994 and reaffirmed by the United Nations General Assembly in 2014, serve as a guiding framework for the production and dissemination of official statistics worldwide. Some of the main issues related to these principles include:

- *Relevance:* official statistics should meet the needs of their users, including policymakers, researchers and the public. Ensuring that statistical data are relevant to the current socio-economic context and address key societal issues is essential;
- *Impartiality and objectivity:* official statistics must be produced and disseminated in an impartial and objective manner, free from political influence or bias. This ensures the credibility and trustworthiness of statistical information;
- *Professionalism:* statistical agencies should adhere to high professional standards in collecting, processing and analysing data. This includes employing qualified staff, using sound methodologies and maintaining the confidentiality of individual responses;
- *Transparency and accountability:* statistical agencies should be transparent about their methods, sources and the limitations of data. They should also be accountable to the public and stakeholders for the quality and accuracy of their statistical outputs;
- *Confidentiality:* protecting the confidentiality of individual responses is crucial for maintaining public trust and encouraging participation in statistical surveys. Statistical agencies must ensure that data are used only for statistical purposes and are not disclosed in a manner that could identify individual respondents;
- *Sound methodology:* official statistics should be based on sound statistical methodologies that are appropriate for the data being collected and analysed. This includes using standardised procedures for sampling, data collection and estimation to produce reliable and comparable results;

- *Timeliness and punctuality:* statistical agencies should strive to produce and disseminate data in a timely manner to meet the current needs of users. Timeliness ensures that statistical information remains relevant and useful for decision-making;
- *Cost-effectiveness:* the efficient use of resources is important in the production of official statistics. Statistical agencies should strive to minimise costs while maintaining the quality and accuracy of data.

In general, the adherence to the FPOS ensures the integrity, credibility and usefulness of statistical information, which is essential for decision-making, policy formulation and monitoring progress towards Sustainable Development Goals.

Quantitative and statistical methods, evolving over time alongside digital technologies, big data and artificial intelligence, are conventionally perceived as fitting instruments for elucidating social phenomena, making predictions and delineating trends. However, it appears imperative to reassess and reorient these methods to accommodate the demands of the contemporary era.

The article delves into the theoretical and methodological aspects surrounding the utilisation of statistics in the social sciences. It aims to identify any gaps, limitations and overlooked determinants pertinent to the explanation of social phenomena, based on the organisation and analysis of statistical data collected by various institutions, organisations and statistical offices.

2. Eurostat: challenges and trends

Eurostat (the statistical office of the EU) national statistical institutes (NSIs) and other data-producing organisations, which together constitute the European Statistical System, collaborate with each other to ensure coherence and comparability of statistical data across the EU. This collaboration involves harmonising methodologies, exchanging best practices and conducting quality assessments to maintain high standards in statistical production.

While efforts are made to uphold these principles, practical challenges may arise in the process, such as ensuring the full independence of statistical offices from political interference, the necessity to address resource constraints and adapt to the evolving data needs and technological advancements. Continuous monitoring, peer reviews and international cooperation play important roles in promoting adherence to these principles and enhancing the quality and reliability of official statistics in EU member states.

Eurostat has faced various challenges and issues over the years, including those related to:

- *Data quality and accuracy:* ensuring high quality and accuracy of statistical data is a constant challenge for Eurostat. Issues such as data inconsistencies, errors and discrepancies among data collected from member states can arise, impacting the reliability and comparability of European statistics;
- *Data timeliness:* timely availability of statistical data is crucial for decision-making and policy formulation. Eurostat faces challenges in ensuring that data are collected, processed and disseminated in a timely manner, especially when dealing with diverse national statistical systems and reporting practices;
- *Data gaps and coverage:* despite efforts made to harmonise statistical methodologies and improve data collection, there are still gaps and limitations in coverage across different statistical domains. Closing these gaps and ensuring comprehensive coverage of relevant topics remain ongoing challenges for Eurostat;
- *Resource constraints:* Eurostat operates within a limited budget and resources, which can limit its capacity to fulfil its mandates effectively. Adequate funding and staffing are essential to maintaining high quality, timeliness and sufficient breadth of statistical outputs;
- *Complexity of the ESS:* the ESS is a complex, decentralised network. Coordinating and harmonising statistical activities within such framework can be challenging, particularly in ensuring consistency and comparability of data;
- *Legal and institutional framework:* Eurostat operates within the legal and institutional framework of the EU, which may create challenges related to governance, accountability and decision-making processes. Ensuring independence, transparency and accountability in statistical activities is essential for maintaining public trust;
- *Technological advancements:* embracing technological advancements and innovations in data collection, processing and dissemination creates not only opportunities, but also challenges for Eurostat. Keeping pace with rapid developments in data science, digitalisation and data privacy regulations requires continuous investment in technological infrastructure and expertise;
- *Public perception and trust:* maintaining public trust and confidence in European statistics is crucial for the credibility and legitimacy of Eurostat. Addressing issues related to data transparency, reliability and relevance is essential for fostering trust among policymakers, researchers and the public.

Eurostat continues to work towards overcoming these challenges through ongoing collaboration with member states, stakeholders and international partners, as well as through the continuous improvement of its methodologies, processes and communication strategies.

The statistical offices of EU member states are generally expected to adhere to the FPOS, because, as mentioned before, they provide a globally recognised framework for the production and dissemination of statistical information. However, the extent to which these principles are followed may vary among individual countries and statistical offices due to factors such as legal frameworks, institutional arrangements, resource constraints and cultural factors.

In general, EU member states are bound by the EU regulations and directives related to statistical practices, which often align with international standards and principles. For example, Regulation (EC) No 223/2009 on European statistics sets out common rules and guidelines for the production of European statistics, including the requirement of impartiality, reliability, relevance, confidentiality and transparency.

3. Statistical analysis: challenges and problems

Statistical analysis faces a range of challenges and problems, particularly concerning models, data collection and other related issues. Some of the main problems in statistical analysis include:

- *Model selection and complexity:* selecting an appropriate statistical model can be challenging, especially when dealing with complex data or phenomena. Choosing between different modelling techniques (e.g. linear regression, logistic regression, machine learning algorithms) and determining the appropriate level of model complexity (e.g. overfitting vs. underfitting) requires careful consideration and expertise;
- *Assumptions violation:* many statistical models rely on specific assumptions about the data, such as normality, independence and homoscedasticity. Violations of these assumptions can lead to biased estimates, incorrect inferences and unreliable predictions. Addressing or accommodating violations of model assumptions is a common challenge in statistical analysis;
- *Missing data:* missing data are a common problem in statistical analysis and can occur for various reasons, including non-response, data entry errors and study design issues. The appropriate handling of missing data is essential to avoid bias and maintain the validity of statistical results. However, selecting and implementing suitable techniques for dealing with missing data (e.g. imputation, modelling missingness mechanisms) can be complex;
- *Data quality and measurement errors:* statistical analysis relies on the quality and accuracy of the underlying data. Issues such as measurement errors, data inconsistencies and outliers can affect the reliability and validity of statistical conclusions. Identifying and addressing data quality issues is essential for producing meaningful and actionable insights;

- *Sample size and power:* the size of the sample used in statistical analysis can impact the reliability and precision of the estimates, and the statistical power to detect effects of interest. Determining an appropriate sample size to achieve sufficient statistical power while considering practical constraints (e.g. time, cost) poses a major challenge in study design and data collection;
- *Bias and confounding:* bias and confounding factors can distort statistical results and lead to incorrect conclusions. Identifying and controlling for potential sources of bias and confounding (e.g. selection bias, confounding variables) is essential for obtaining accurate and unbiased estimates of the relationships between variables;
- *Data collection methods and sources:* choosing the most appropriate data collection methods and sources is crucial for obtaining high-quality data. Issues such as sampling bias, non-random sampling and data source limitations can introduce bias and affect the generalisability of statistical findings;
- *Interpretation and communication:* communicating statistical findings effectively and accurately to stakeholders, policymakers and the public is a critical aspect of statistical analysis. However, interpreting complex statistical results in a meaningful and accessible way, while avoiding misinterpretation and misunderstanding, can be challenging.

Addressing these problems requires a combination of statistical expertise, domain knowledge and methodological rigor. Collaborative efforts between statisticians, subject matter experts and data scientists are essential for overcoming these challenges and producing reliable and actionable insights from statistical analysis.

4. Research rationale

According to multiple sources on the Internet, last decades saw a vast expansion of econometric and quantitative-modelling techniques in social sciences. The evolution of models has seen the proliferation of various frameworks across different domains. Originally rooted in the rational-choice framework of neo-classical economics, these models have expanded to encompass diverse areas, such as mathematical models of risk-analysis and game theory models.

Today, these frameworks have permeated into the digital realms of managerial processes, state formations and ethno-linguistic identities, reflecting the increasing digitisation and globalisation of modern society. This integration has given rise to a multidisciplinary field that blends elements of economics, statistics and quantitative methods.

This interdisciplinary field is not only the cornerstone of scholarly inquiry but also plays a significant role in shaping policy options for global institutions. The insights yielded by these models enable policymakers to make informed decisions and devise strategies to address complex socio-economic challenges on a global scale. Thus, the evolution of models has not only broadened their scope, but also increased their relevance and impact in contemporary discourse and policymaking.

The fundamental objection to the current state of quantitative methods (QM) and econometrics lies in their reliance on underlying assumptions and post-hoc hypotheses, which often go unexamined in a systematic manner.

In recent years, the state-of-the-art development theory has taken a multifaceted approach. Firstly, it emphasises the mobilisation of researchers to distil complex social phenomena into quantifiable and comparable datasets. However, this process of reduction is not free from challenges, as it frequently involves implicit value judgments that are rarely subjected to scrutiny. Secondly, the focus shifts towards the models themselves. While these models serve as valuable tools for understanding and predicting socio-economic dynamics, they are not immune to criticism. Issues such as oversimplification, lack of robustness and inherent biases can undermine the validity and reliability of their predictions.

Therefore, it becomes imperative to critically examine the assumptions, hypotheses and value judgments underlying QM, econometrics, statistics and contemporary development theories. By doing so, we can enhance the rigor and credibility of our analytical frameworks and pave the way for more nuanced and contextually sensitive approaches to understanding and addressing complex social phenomena.

Critical attention is directed towards two key aspects: the theoretical assumptions that underlie the 'hypothetical leap' made between statistical results, and the researcher's ultimate interpretation and explanation of those results.

In the realm of social sciences, the reduction of complex social, historical and political determinants to numerical data marks the beginning of the analytical process. However, at the endpoint of this process, these determinants often re-emerge in a disembedded form or are only incorporated into the researcher's interpretation through the lens of common sense, which, in essence, reflects ideological assumptions.

This phenomenon underscores the intricate relationship between data analysis and broader socio-political contexts. It highlights the importance of a critical examination not only of the statistical methodologies employed but also the underlying assumptions and biases inherent in the researcher's interpretation. By interrogating these elements, we can strive for a more nuanced understanding of social phenomena that transcends simplistic numerical representations and acknowledges the complexities of the human experience.

5. Use of quantitative measures and methodologies in social sciences

Econometrics and statistics play a vital role in analysing and interpreting data, but it is also essential to recognise their limitations as lower-order tools. While they can generate correlations or discrepancies, the true understanding of these findings often requires more sophisticated and nuanced forms of inquiry that are both conceptually and empirically developed.

QM and econometric analysis are frequently utilised beyond their intended scope. Individuals may manipulate data to suit their agendas, unveiling their potential for misuse. However, when employed correctly and presented transparently, they can serve as valuable tools for decision-making and policy formulation.

It is crucial to view QM, statistical and econometric methods as neutral tools of analysis, as inherently neither positive nor negative. Rather, their effectiveness depends on how they are applied and interpreted.

Alternatively, there are other techniques available for analysis, such as market analysis, experimental methods and social models applied to groups. Each of these approaches offers its own set of advantages and limitations, and their selection depends on the specific context and objectives of the analysis. Only through the careful consideration of the strengths and weaknesses of different analytical approaches can researchers enhance the robustness and reliability of their findings.

6. Laws and limits of QM, statistics and econometrics

The econometric approach, while valuable, is not without shortcomings. Drawing from a template in sociology, six laws have been formulated to delineate the mainstream activities of econometrics and elucidate their scientific boundaries. These laws encapsulate the inherent challenges and constraints faced within the discipline.

Moreover, proximity theorems have been developed to quantify explicitly the boundaries of our ability to approximate the generating mechanism of data and optimise forecasts of future observations. Phillips (1995) delves into these theorems, shedding light on the finite nature of observations and the consequent limitations on predictive accuracy.

The magnitude of these boundaries is contingent upon the characteristics of the model and the trajectory of the data. As we contemplate the future of econometrics, there is a growing recognition of the need to harness advanced econometric methods interactively, leveraging the capabilities of modern technology, such as web browsers.

A fundamental issue permeating all practical economic analyses is the extent to which we can truly comprehend economic phenomena. This process typically involves developing theories, gathering observations and fitting models. However, a pressing question arises whether there are inherent limits to our ability to predict future observations using empirical models derived from such processes.

These inquiries underscore the ongoing quest to extend the boundaries of econometrics while acknowledging and navigating its inherent limitations. As researchers continue to grapple with these challenges, the evolution of econometric methods promises to enhance our understanding of complex economic dynamics and inform more robust decision-making processes.

The essence of econometric modelling lies in navigating the uncertainties inherent in data analysis. One fundamental truth is that the appropriate model for any given dataset remains elusive and unknown. Nevertheless, econometricians strive to formulate models that capture underlying relationships accurately. However, even the most meticulously crafted model is contingent upon estimating parameters from the available data, entailing an element of uncertainty.

This challenge is compounded by the scarcity of data relative to the multitude of parameters that necessitate estimation. This imbalance requires the use of nonparametric or semiparametric methods, which offer flexibility in modelling functional relationships without imposing overly restrictive assumptions.

Moreover, the empirical limitations on modelling extend beyond traditional finite parameter models. As econometricians encounter these complexities, they are also confronted with the reality that perfect modelling is unattainable, and the quest for accurate representation must navigate the inherent limitations.

A central tenet of econometrics is that all models are inherently flawed representations of reality. Economic theories serve as metaphors of reality, offering simplified frameworks that often diverge from empirical observations. Despite their imperfections, these models persist, driven by the perception that they encapsulate essential economic principles or 'laws'.

Indeed, the decision to employ a particular model often hinges on its perceived kernel of truth, even if only an approximate one. This strategic utilisation of economic insights in crafting empirical models is deemed advantageous, offering a pragmatic approach to navigating the complexities of real-world data analysis.

In essence, the pragmatic use of imperfect models, guided by economic intuition, is preferable to either an entirely unrestricted or arbitrarily restricted modelling approach. By embracing the imperfections inherent in economic modelling, econometricians can extract valuable insights that support decision-making processes.

These fundamental principles underscore the ongoing evolution of econometrics and its enduring relevance in modern economic analysis. As scholars continue to grapple with the complexities of econometric modelling, Sargan (2008) offers invaluable insights into the guiding 'laws' of econometrics.

Through critical examination and refinement, econometricians strive to navigate the intricate interplay between theory and empirical reality, extending the boundaries of economic inquiry and fostering a deeper understanding of the forces shaping our economic landscape.

It is crucial to recognise that the laws of econometrics are not immutable truths but rather attempts to encapsulate the essence of econometric practices and elucidate the challenges inherent in explaining and predicting economic phenomena. In this vein, scholars such as Cartwright (2009a, 2009b) and Gaddis (2001) have proposed related perspectives on modelling.

Cartwright (2009a, 2009b) suggests that models can be conceptualised as machines that generate laws, akin to nomological machines. These laws, resembling the morals drawn from fables, emerge from the modelling process and provide insights into economic relationships.

Gaddis (2001), on the other hand, emphasises the utility of economic modelling in shedding light on empirical relationships. He claims that formal laws in economics may not necessarily advance understanding, as even accumulated falsifications or anomalies do not prompt scientists to abandon an approach, unless a better alternative is available.

Rissanen (1985, 1986a, 1986b, 1987, 1989, 2006) challenges the notion of a true model, instead viewing statistics as a language for articulating regular features of data. He introduces proximity theorems that measure how closely an empirical model aligns with the true model within a parametric family, offering insights into the limits of modelling accuracy.

Moon et al. (2007) and Ploberger and Phillips (2008, 2012) investigate the intricacies of proximity theorems, highlighting their dependence on both the data and the model being employed. They observe that these bounds are greater for trending data compared to stationary data, and they allow for finite parameter families as well as families with local misspecifications.

Furthermore, modelling algorithms accommodate gross misspecification within family groups, offering flexibility in modelling approaches. These theorems also extend to prediction, quantifying limits on empirical forecasting capability in situations where the model specification is questionable.

Collectively, these perspectives contribute to a nuanced understanding of the laws and limits of econometrics, highlighting the complexities inherent in modelling economic phenomena and the challenges associated with predicting future outcomes.

7. Six laws of econometrics

7.1. Laws and limits of econometrics

The six laws of econometrics do not claim to be universal scientific truths; they rather serve to characterise the activities and constraints inherent in econometric analysis. Let us examine each of these laws to gain a deeper understanding of the complexities within the discipline.

- *First law: some methods work, some do not*
  Econometrics has long been engaged in developing statistical tools tailored to economic models and data. This ongoing process illustrates the dynamic nature of econometric practice, wherein the effectiveness of statistical methods varies. While some methods prove robust, others falter, prompting continuous refinement and evolution in modelling practices.
- *Second law: it is different on infinite dimensional spaces*
  Econometrics strives for generality in modelling, especially in areas where prior knowledge is limited. Yet, where models align closely with the underlying economic hypotheses, specificity through direct parameterisation is sought. This duality has spurred extensive exploration into nonparametric and semiparametric estimation methods.
- *Third law: unit roots always cause trouble*
  Unit roots, akin to the 'hill people' of econometrics, present persistent challenges due to their nonstandard limit distributions. Their disruptive influence is compounded by discontinuities in limit theory, particularly as the autoregressive parameter approaches unity.
- *Fourth law: cross-section dependence also causes trouble*
  Cross-section dependence poses significant challenges in panel modelling, particularly in microeconomic and macroeconomic applications. Despite being a burgeoning field of study, unresolved difficulties persist, partly stemming from the absence of a natural ordering of cross-sectional data.
- *Fifth law: no one understands trends*
  Despite their significance in macroeconomic research, trends remain enigmatic, with the economic theory offering limited guidance. Common trend formulations, such as polynomial time trends or stochastic trends, reflect this ambiguity and are subject to empirical scrutiny.
- *Sixth law: spurious regression and trend representation*
  Spurious regression, a ubiquitous issue, underscores the importance of validating relationships through rigorous testing. Deterministic trend functions and continuous stochastic processes offer frameworks for understanding trend behaviour, though challenges persist in terms of representing trending data accurately.

By scrutinising these laws, we gain insight into the complexities and nuances of econometric analysis, guiding researchers in navigating its inherent challenges and advancing the field's methodological frontiers.

7.2. Initial challenges in big data

The advent of 'big data' has thrust statistics into the limelight, offering numerous opportunities for individuals trained in statistical analysis. However, along with its promise, big data present a host of challenges, particularly concerning hardware infrastructure and algorithm design.

One significant issue is the emergence of what can be termed as 'small big data'. This refers to datasets collected by individuals whose data collection skills surpass their ability to analyse the data effectively. These individuals often lack the qualifications or preparation to conduct meaningful analysis. Moreover, insufficient budget allocation for data analysis, a common occurrence in grant-funded projects, exacerbates this problem, leaving researchers stranded without the means to analyse the collected data.

7.3. Statistical data editing

Statistical data editing (SDE) serves as a critical tool in survey processing, spanning various phases from frame development to quality assurance. The overarching aim of SDE is to enhance procedures and automate processes to ensure the accuracy of published estimates and microdata.

SDE encompasses two primary subcategories: the Fellegi-Holt (FH) Methods and Systems (Scholtus, 2014) and General Methods and Systems. The former are rooted in the Fellegi-Holt model of editing, and they employ if-then-else rules developed by subject-matter specialists for data editing. While FH systems offer formal development methods and customisable edit restraints, errors may arise if specialists fail to devise comprehensive logic for edit rules. The General Methods and Systems encompasses all other editing methods beyond the FH model (Winkler & Petkunas, 1997). It offers flexibility in edit rule development, but may lack the structured approach of FH systems.

7.4. Misspecifications in data gathering

Recent years have seen a surge in research areas heavily reliant on record linkage techniques. Three prominent areas include:

- *Microdata confidentiality and re-identification methods,* addressing concerns around data privacy and re-identification risks;
- *Analytic linking,* which involves merging and analysing quantitative and discrete data from multiple sources to adjust for biases due to linkage errors;

- *Information retrieval and machine learning*, which are leveraging techniques from computer science, such as information retrieval and machine learning, in web search engines and data mining applications.

Navigating these challenges and leveraging the potential of big data requires concerted efforts in data collection, analysis and methodology development, ensuring the reliability and validity of insights derived from large-scale datasets.

8. Two sets of data: basic and advanced research problems

Basic research problems are:

- *Improving frequency-based matching*: a longstanding question in record linkage research when frequency-based matching offers advantages over simple agree/disagree matching. Progress has been made since the seminal works of Fellegi and Sunter (1969) and Newcomb et al. (1958), yet challenges persist in determining the optimal utilisation of identifying information in matching processes;
- *Parameter estimation under conditional independence*: another fundamental challenge is determining the most effective method for estimating parameters under conditional independence. Researchers have made efforts to address this issue, as reflected in the works of Herzog et al. (2007), Nigam et al. (1999) and Winkler (1988).

Advanced research problems include:

- *Confidentiality of microdata:* with the increasing need to provide researchers with access to large public-use files while safeguarding individual privacy, evaluating re-identification risks becomes crucial. Record linkage methods play a vital role in assessing such risks, especially concerning quantitative data in public-use files;
- *Analytic linking methodology:* analytic linking, pioneered by Scheuren and Winkler (1997), enables the merging and analysis of non-comparable data items, thereby improving the accuracy of matching and mitigating the impact of matching errors on analyses;
- *Data mining and information retrieval:* Bayesian networks, widely employed in data mining and information retrieval tasks, offer valuable tools for classifying documents and enhancing the functionalities of web search engines;
- *Analytic linking challenges:* researchers often face challenges when analysing merged administrative data files lacking unique identifiers. Adjusting regression analyses for biases arising from linkage errors, as demonstrated by Lahiri and Larsen (2005), remains a critical area of investigation;
- *Data mining applications:* machine learning algorithms leveraging Bayesian networks are increasingly applied to text classification and information retrieval tasks, including web search engine algorithms, highlighting their versatility in data mining contexts.

Efforts made in the area of both basic and advanced research problems are essential for advancing the field of record linkage, ensuring the accuracy, confidentiality and utility of linked datasets across various applications and disciplines.

9. Conclusions

In conclusion, it is imperative to acknowledge and address the inherent problems and limitations in quantitative methods, econometrics and statistics. Employing the most appropriate methodology tailored to the specific context of the study is essential for producing reliable results. Comparing and referencing similar studies can provide valuable insights and enhance the credibility of research findings.

The interpretation and explanation of results must be made with caution to avoid drawing unwarranted conclusions. It is crucial to recognise that every scientific endeavour has its grey areas and uncertainties, and embracing simplicity can often lead to elegant solutions.

The process of data collection poses significant challenges and requires careful planning and execution. Increasing the involvement of statisticians in both data analysis and study design is essential for ensuring the quality and usefulness of the collected data. However, the demand for statisticians often exceeds their supply, highlighting the need for innovative approaches to skills dissemination.

Exploring alternative methods to impart statistical skills without solely relying on physical presence is essential for meeting the growing demand for expertise in the field. By fostering creativity and leveraging technological advancements, we can expand the reach of statistical knowledge and empower more individuals to contribute meaningfully to research and decision-making processes.

References

Cartwright, N. (2009a). Causal Laws, Policy Predictions, and the Need for Genuine Powers. In T. Handfield (Ed.), *Dispositions and Causes* (pp. 127–157). Oxford University Press. https://doi.org/10.1093/oso/9780199558933.003.0005.

Cartwright, N. (2009b). Causality, Invariance, and Policy. In H. Kincaid & D. Ross (Eds.), *The Oxford Handbook of Philosophy of Economics* (pp. 410–423). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195189254.003.0015.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049.

Gaddis, J. L. (2001). *Strategies of Containment, Past and Future*. Hoover Digest.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer. https://doi.org/10.1007/0-387-69505-2.

Lahiri, P., & Larsen, M. D. (2005). Regression Analysis With Linked Data. *Journal of the American ;Statistical Association*, *100*(469), 222–230. https://doi.org/10.1198/016214504000001277.

Moon, H. R., Perron, B., & Phillips, P. C. B. (2007). Incidental trends and the power of panel unit root tests. *Journal of Econometrics*, *141*(2), 416–459. https://doi.org/10.1016/j.jeconom.2006.10.003.

Newcomb, T. M., Maccoby, E. E., & Hartley, E. L. (Eds.). (1958). *Readings in Social Psychology* (3rd edition). Henry Holt.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using Maximum Entropy for Text Classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, *1*, 61-67. http://www.kamalnigam.com/papers/maxent-ijcaiws99.pdf.

Phillips, P. C. B. (1995). *The dramatic expansion of econometric and quantitative-modelling techniques* (Cowles Foundation Paper No. 1081). Cowles Foundation for Research in Economics at Yale University.

Phillips, P. C. B. (2003). Laws and Limits of Econometrics. *The Economic Journal*, *113*(486), C26–C52.

Ploberger, W., & Phillips, P. C. B. (2008). An Introduction to Best Empirical Models when the Parameter Space is Infinite Dimensional. *Oxford Bulletin of Economics and Statistics*, *65*(1), 877–890. https://doi.org/10.1046/j.0305-9049.2003.00089.x.

Ploberger, W., & Phillips, P. C. B. (2012). Optimal Estimation under Nonstandard Conditions. *Journal of Econometrics*, *169*(2), 258–265. https://doi.org/10.1016/j.jeconom.2012.01.025.

Rissanen, J. (1985). Minimum Description Length Principle. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (5th edition, pp. 523–527). John Wiley and Sons.

Rissanen, J. (1986a). Stochastic Complexity and Modeling. *Annals of Statistics*, *14*(3), 1080–1100. https://doi.org/10.1214/aos/1176350051.

Rissanen, J. (1986b). A Predictive Least-Squares Principle. *IMA Journal of Mathematical Control and Information*, *3*(2–3), 211–222. https://doi.org/10.1093/imamci/3.2-3.211.

Rissanen, J. (1987). Stochastic Complexity. *Journal of the Royal Statistical Society. Series B (Methodological)*, *49*(3), 223–265.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry* (2nd edition). World Scientific Publishing.

Rissanen, J. (2006). *Information and Complexity in Statistical Modeling*. Springer-Verlag. https://doi.org/10.1007/978-0-387-68812-1.

Sargan, J. D. (2008). *The Sargan Lecture*. https://res.org.uk/sargan-lecture/.

Scheuren, F., & Winkler, E. W. (1997). Regression Analysis of Data Files that are Computer Matched – Part II. *Survey Methodology*, *23*(2), 126–138.

Scholtus, S. (2014, 28–30 April). *A generalised Fellegi-Holt paradigm for automatic editing* [paper presentation]. Conference of European Statisticians, Paris. https://doi.org/10.13140/2.1.2211.7446.

Winkler, W. E. (1988). *The State of Record Linkage and Current Research Problems*. U.S. Bureau of the Census.

Winkler, W. E., & Petkunas, T. F. (1997). *The Discrete Edit System*. U.S. Bureau of the Census.

Eleftherios Thalassinos

University of Malta, Faculty of Economics, Management & Accountancy, Malta;
Retired Professor of University of Piraeus, School of Maritime and Industrial Studies, Greece.
ORCID: https://orcid.org/0000-0003-3526-4930. E-mail: ethalassinos@gmail.com