# Research Ideas for the Journal of Informatics and Data Mining: Opinion*

**Editor-in-Chief**

**Michael McAleer**

Department of Quantitative Finance
National Tsing Hua University
Taiwan
and
Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
And
Tinbergen Institute
The Netherlands
And
Department of Quantitative Economics
Complutense University of Madrid
Spain

EI2015-26

September 2015

## Abstract

The purpose of this Opinion article is to discuss some ideas that might lead to papers that are suitable for publication in the Journal of Informatics and Data Mining. The suggestions include the analysis of citations databases, PI-BETA (Papers Ignored – By Even The Authors), model specification and testing, pre-test bias and data mining, international rankings of academic journals based on citations, international rankings of academic institutions based on citations and other factors, and case studies in numerous disciplines in the sciences and social sciences.

**Keywords:** Citations databases, model specification and testing, pre-test bias, international rankings of journals and institutions, case studies.

**JEL:** B23, C55, C81, C82, C87, C88.

*"The code is more what you'd call 'guidelines' than actual rules."*

**Captain Hector Barbossa**
**Pirates of the Caribbean: The Curse of the Black Pearl (2003)**

## 1. Code: Rules or Guidelines?

1. *Research papers are written to be published in academic journals.*

2. *Journal publications should be cited.*

3. *A researcher's academic impact is based on journal publications and citations.*

4. *A journal's academic impact is based on citations.*

5. *Many journal publications are not cited.*

6. *Journal publications that are not cited should not have been published.*

## 2. Introduction

In order to meet the growing needs of academic researchers and practitioners in informatics and data mining, in 2015 a new Open Access international publication in the area was established, namely the Journal of Informatics and Data Mining (JIDM).

JIDM is intended as a generalist outlet for high quality articles in a wide range of alternative methods of computer science, measurement, and data mining.

The intention of JIDM is to publish theoretical and applied papers, including case studies that will enable the portability of methods and techniques, on a wide range of topics in both informatics, which encompasses the science of information and the practice of information processing, and the correspondingly essential techniques associated with data mining, which includes the measurement of publications and citations.

Informatics and data mining appeal to both academic researchers and practitioners because of the direct and immediate applicability of established and newly developed theories, as well as

the availability and accessibility of large data sets, including panel, cross section and times series data.

Academic and practical research papers and case studies that might typically be considered under the related disciplines of bibliometrics, scientometrics, informetrics and webometrics are eminently suitable for JIDM.

JIDM is an international journal with the goal of advancing the knowledge and understanding of informatics and data mining using rigorous and powerful mathematical, statistical and econometric methods in data mining to test theoretical models and empirical regularities in informatics.

Some of the topics that might be considered include rigorous technical, theoretical and applied research in informatics and data mining that includes, but is not restricted to: acquisition and storage, alternative metrics, bibliographic and bibliometric databases, complex information systems, computer hardware and software, computer and information science, cross section data, diagnostic methods and testing, experimental data, high frequency time series data, information processing, knowledge discovery and management, latent variables, machine learning, measurement errors, measurement systems, methods and techniques, model specification and misspecification, optimal use of information, quantitative methods, rankings of individuals, journals and institutions, scientific impact, time series data, univariate and multivariate models, and ultra-high frequency time series data.

JHME seeks academically rigorous papers that will appeal to theoreticians and will also have direct relevance to practitioners in informatics and data mining.

Research papers that would be of interest to JIDM should be based on sound theory and practice in informatics and data mining. Technically rigorous papers that are based on mathematical, econometric and statistical methods in the analysis and evaluation of theoretical models in informatics and empirical regularities in data mining are strongly encouraged. Case studies that will enable portability of the theoretical and practical findings to other data sets are also warmly welcome.

The remainder of the paper is as follows: Section 3 discusses some ideas and suggestions that might lead to papers that are suitable for publication in the Journal of Informatics and Data Mining, including the analysis of citations databases, PI-BETA (Papers Ignored – By Even The Authors), model specification and testing, pre-test bias and data mining, international rankings of academic journals based on citations, international rankings of academic institutions based on citations and other factors, and case studies in numerous disciplines in the sciences and social sciences. Section 4 provides an encouragement to submit papers to JIDM.

## 3. Research Suggestions

Some research ideas that are pertinent and of substantial interest to JIDM include, but are not restricted to, the following topics:

1. **Analysis of citations databases:**

   Academic journals are ranked almost entirely according to citations, whereas individual academic researchers are ranked according to publications and citations. There are many citations databases. Some of the more widely used across most, if not all, academic disciplines in the Sciences and Social sciences include Thomson Reuters ISI, Google Scholar, Scopus, Microsoft Academic Search, and ResearchGate. There do not seem to be as many discipline-specific databases. The Social Science Research Network (SSRN) is widely used for the Social Sciences, and Research Papers in Economics (RePEc) is widely used in Economics, Finance, Accounting, Statistics, and related disciplines. Numerous variations of the functions of the citations data are available, and form an important part of inforamtics and data mining, and so are most definitely suitable for JIDM.

2. **PI-BETA (Papers Ignored – By Even The Authors):**

   *"All citations rankings are useful, but some are more useful than others."*
   Chang and McAleer (2015b)

   Chang et al. (2011a) argue that the lack of citations of published papers, especially if they are not recent publications, reflects on journal quality by exposing editorial mistakes in publishing papers that are subsequently not cited. PI-BETA was developed

by Chang et al. (2011b) as an indication of a journal's mistakes in publishing a paper that the international academic community, including the authors, do not take seriously. Many journals have high citation rates, despite having high PI-BETA values, which emphasizes that the reputation of such journals are based on the very night numbers of citations of a small proportion of the published papers. This bibliometric measure should always be considered in evaluating the quality and influence of academic journals.

## 3. Model specification and testing:

*"Essentially, all models are wrong, but some are useful."*
 Box and Draper (1987, p. 424)

The above statement is a well-known definitional fact of models. As all models are based on sets of assumptions, with all assumptions being false, it follows that all models are false, such false models can lead to biased and inconsistent parameter estimates, as well as a loss of efficiency. Consequently, model specification tests and diagnostic checks for, among others, incorrectly omitted variables, extraneous inclusion of variables, incorrect functional form, causality, endogeneity and exogeneity, measurement errors, weak instruments, omitted equations, sensitivity analysis, robustness, valid inferences, implied, conditional, stochastic and realized volatility, asymptotic theory, accommodating theory and data, and re-evaluation and reformulation of theories, using the most powerful statistical and econometric methods and techniques available, are strongly encouraged for JIDM.

## 4. Pre-test bias and data mining:

Pre-test bias involves statistical testing of various null hypotheses, and subsequent re-estimation and testing without appropriate allowance being made for the underlying probability of a type one errors, namely when the significance levels are incorrect, and hence can and do lead to inappropriate statistical inferences. Consequently, pre-testing is widely interpreted as involving estimation rather than statistical testing, and has also been referred to pejoratively, especially in theoretical and applied econometrics, as data mining. The use of alternative data sets has been advised as an appropriate testing approach. Pre-testing is widely ignored, especially when there are many observations,

such as in empirical investment finance, where high frequency and ultra-high frequency data are available, such as nano data and time series data at the frequency of seconds, minutes, hours and days.

5. **International rankings of academic journals based on citations:**

Such rankings are primarily based on citations and functions thereof. Chang and McAleer (2015a, p. 120) argue that "The gold standard for bibliometric rankings based on citations data is the widely-used Thomson Reuters Web of Science (2014) citations database, which publishes, among others, the celebrated Impact Factor." They present, define and compare the 16 most well-known Thomson Reuters bibliometric measures that are based on citations data. Many more bibliometric measures can be developed using different variations of the citations data, as well as indexes, or weighted measures, based on one of the three Pythagorean means, namely the arithmetic, geometric and harmonic means. Numerous such bibliometric measures are directly related to the interesting theoretical and practical topics covered directly by JIDM.

6. **International rankings of academic institutions based on citations and other factors:**

Universities worldwide have been ranked using a wide range of arbitrary factors, including research quality and quantity, as well as journal citations. The three main world rankings based on different criteria are Shanghai Academic Ranking of World Universities (ARWU) (first reported by Shanghai Jiaotong University in 2003, and inaugurated as ARWU in 2011), Times Higher Education (THE) - Quacquarelli Symonds (QS) World University rankings inaugurated in 2004, which subsequently separated into THE World University rankings (inaugurated in 2011), and QS World University rankings (inaugurated as 2012). The Centre for Science and Technology Studies (CWTS) Leiden rankings differ from the above three world rankings in measuring the scientific performance and scientific collaboration of universities. Such rankings, and other that can be developed using citations data and other important factors, are in the realm of JIDM.

7. **Case studies in numerous disciplines in the sciences and social sciences**:

These areas would include altmetrics, article-level metrics, article downloads and views, article influence, artificial intelligence, author-level metrics, automated information systems, big data, bioinformatics, biological mechanisms, biomedical informatics, biometrics, business informatics, chemoinformatics, clinical informatics, communications technology, computational theory and tools, computational tools, computer hardware, creditmetrics, criminometrics, cybermetrics, data analytics and processing, database management, decision making and support systems, digital communications, eigenfactor, environmetrics, epidemiology, functionalism of new technologies, generalized metrics, genetic algorithms, health informatics, impact factor, article and journal influence, information analysis and communication technologies, information production processes and systems, informetrics, infrastructure, internet informatics, investment metrics, journal-level metrics, marketing metrics, medical informatics, methodology, nanoinformatics, networks, neuroinformatics, non-traditional metrics, organizational informatics, pattern recognition, pharmacoepidemiology, portfolios, prediction, productivity, psychometrics, recorded information, risk factors, risk metrics, science of information, scientific communication and information, scientometrics, social informatics, social mechanisms, sociometrics, computer software, source code repositories, strategies, structure, technologies, technometrics, valuation methods, and webometrics.

## 4. Encouragement to submit papers to JIDM

There are numerous exciting and novel topics that would be of interest to JIDM, some of which have been discussed above. These are personal opinions, and talented researchers worldwide are the best judges of what might be of interest in both informatics and data mining.

Academic, theoretical and practical researchers will undoubtedly be able to develop exciting, novel and interesting research ideas that will use rigorous mathematical, statistical and econometric methods and techniques to test established theories and evaluate empirical regularities in informatics and data mining.

# References

Box, G.E.P., and N.R. Draper (1987), *Empirical Model Building and Response Surfaces*, Wiley, New York.

Chang, C.-L. and M. McAleer (2015a), "Bibliometric rankings of journals based on the Thomson Reuters citations database", *Journal of Reviews on Global Economics*, 4, 120-125.

Chang, C.-L. and M. McAleer (2015b), "Quality weighted citations versus total citations in finance and accounting", to appear in *Managerial Finance*.

Chang, C.-L., M. McAleer and L. Oxley (2011a), "Great expectatrics: Great papers, great journals, great econometrics", *Econometric Reviews*, 30(6), 583-619.

Chang, C.-L., M. McAleer and L. Oxley (2011b), "What makes a great journal great in the sciences? Which came first, the chicken or the egg?", *Scientometrics*, 87(1), 17-40.

Thomson Reuters Web of Science (2014), *Journal Citation Reports, Essential Science Indicators*, Thomson Reuters.