PAPER • OPEN ACCESS

Application of econometric and ecology analysis methods in physics software

View the article online for updates and enhancements.

Application of econometric and ecology analysis methods in physics software

Min Cheol Han[1], Gabriela Hoff[2], Chan Hyeong Kim[1], Sung Hun Kim[1], Maria Grazia Pia[3], Elisabetta Ronchieri[4], Paolo Saracco[3]

[1] Hanyang University, Seoul 04763, Korea
[2] Instituto Politecnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, RJ 28630-050, Brazil
[3] INFN, Sezione di Genova, Genoa 16146, Italy
[4] INFN CNAF, Bologna 40126, Italy

E-mail: mariagrazia.pia@ge.infn.it,elisabetta.ronchieri@cnaf.infn.it

Abstract.
  Some data analysis methods typically used in econometric studies and in ecology have been evaluated and applied in physics software environments. They concern the evolution of observables through objective identification of change points and trends, and measurements of inequality, diversity and evenness across a data set. Within each analysis area, various statistical tests and measures have been examined. This conference paper summarizes a brief overview of some of these methods.

1. Introduction
This conference paper reports an exploration of the applicability of data analysis methods derived from disciplines other than physics, namely economy and ecology, to physics software problems.

The investigation concerns measurements of inequality, diversity and concentration, trend analysis and the identification of change points. These analysis methods originate from econometrics (e.g. the analysis of inequality) or ecology (e.g. the analysis of diversity); some of them are typically used both in economics and environmental sciences (e.g. trend analysis).

The exploration of these analysis methods is motivated by requirements of ongoing projects concerning the validation of Geant4 [1, 2] physics and the assessment of its maintainability; nevertheless, these concepts and techniques have a wider scope of application in physics analysis.

This conference paper illustrates some examples of application of these methods to ongoing studies concerning Geant4. It is wothwhile to stress that these examples are shown here only for the purpose of elucidating the statistical techniques that are investigated; as these examples are taken out of the full context of complex studies to which they pertain, any extrapolation to make an assessment of Geant4 capabilities out of the small set of statistical results reported here would be inappropriate. Both the statistical methods and the results of their application that are briefly illustrated in this conference overview will be extensively documented in dedicated publications in scholarly journals.
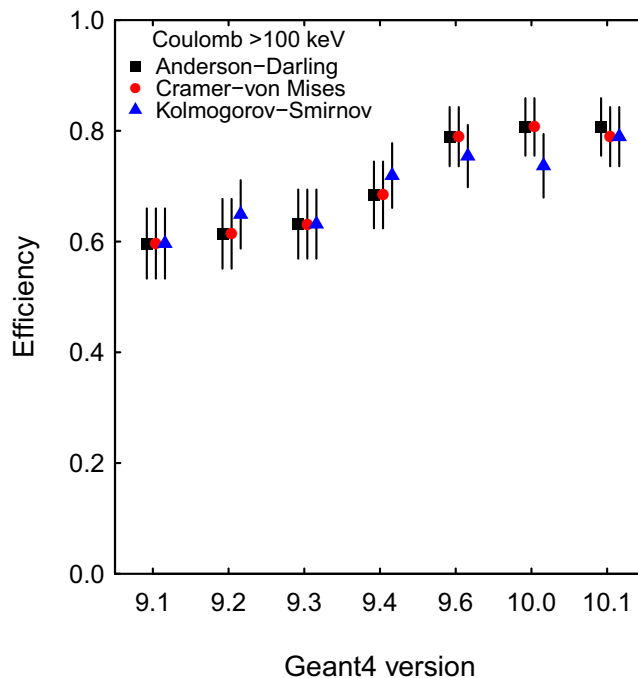
Figure 1.   Evolution of compatibility with experiment ("efficiency") of a simulation configuration including single electron Coulomb scattering versus Geant4 version, concerning primary electrons with energy above 100 keV. Similar efficiency is observed when compatibility with experiment is determined using different goodness-of-fit tests (Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov).

2. Trend analysis

Trend analysis is commonly applied in economics and finance, as well as in disciplines related to environmental sciences. It exploits statistical methods to identify patterns in a series of data, usually temporally ordered, with the purpose of distinguishing them from randomness. Although this kind of analysis does not predict the future, sometimes it is used for financial forecasts; a word of warning should be issued in this respect regarding the use of trend analysis techniques beyond their proper scope of applicability.

The need of performing a trend analysis may also arise in physics software scenarios. As an example, one can consider the study of the evolution of the efficiency of a software model, a reconstruction algorithm or a detection instrument. One may want to know whether observed changes of efficiency as a function of time are compatible with random fluctuations or are statistically significant, i.e. whether they are concrete manifestations of improvement or degradation of performance that deserve attention. The observation of statistically significant trends could highlight the need of taking corrective actions, e.g. to restore acceptable performance or to account for modified operation in dependent parts of the system.

An example of the identification of statistically significant trends in efficiency is illustrated in Figure 1. In this case the plotted efficiency represents the evolution of compatibility with experiment of a simulated observable (the fraction of backscattered electrons), obtained with the same user application code [3–5] and the same set of reference experimental measurements, but using different Geant4 versions. Since the user application code remained unchanged, significant differences in compatibility with experiment are likely to derive from changes to the underlying particle transport system used by the user application.

Figure 1 qualitatively hints to some apparent upward trend, which is quantified through a

statistical test. Various parametric and non-parametric tests are available for statistical inference regarding the presence of a trend; one of them is the Mann-Kendall test [6,7], which tests a null hypothesis H_0 of no monotonic trend with respect to an alternative hypothesis H_1 corresponding to the presence of a monotonic trend, e.g. an upward trend in the case of Figure 1). It is worthwhile to note that, unlike qualitative visual evaluations of plots, the detection of a trend based on a statistical test has an objective mathematical foundation. Moreover, the use of statistical methods allows one to automate the test process; this is an essential requirement when the analysis encompasses a large number of test cases.

In the scenario of Figure 1 the p-value resulting from the Mann-Kendall test is 0.003; therefore the null hypothesis that the observed pattern is consistent with randomness is rejected with 0.01 significance in favour of the alternative hypothesis of improved compatibility with experiment. It is worthwhile to note that the same trend test detected a statistically significant degradation of compatibility with experiment for another physics configuration used in the same application [8], for which the null hypothesis was rejected with 0.01 significance in favour of the alternative hypothesis of a downward trend in the efficiency resulting from that configuration.

Trend analysis is also performed in the context of an ongoing project that evaluates the quality of Geant4 software, based on a set of established software quality metrics [9]. In that environment it identifies objectively whether some characteristics of the software have evolved towards significantly better or worse quality as a function of time, i.e. of Geant4 version.

Some examples of trends analysis concerning the evolution of software quality metrics are illustrated in Figure 2 and 3, which concern the coupling between objects and the number of member methods in base classes of the Geant4 electromagnetic *utils* and geometry *management* packages, respectively. Evolution towards greater coupling between objects is observed in Figure 2, while a decreasing trend in the number of member methods is observed in Figure 3. The hypothesis of randomness is rejected by the Mann-Kendall test in both cases with 0.01 significance, in favour of the alternative hypothesis of upward and downward trend, respectively.

Since increased coupling between object is not desirable in object oriented design, the observation of a statistically significant trend in that direction could motivate a software design review to improve an apparently deteriorating design, while a significant trend towards the reduction of the number of member methods hints to leaner classes, which would be a desirable design feature.

3. Inequality analysis

The need to estimate inequality within a data sample may arise in a variety of physics scenarios, where one would like to appraise how some property of interest is distributed within the elements of that sample. In the evaluation of software quality metrics [9], the analysis of inequality allows one to aggregate sparse information pertaining to individual elements of a software package (e.g. files, classes, functions) into a single variable, which summarizes the distribution of the measurements.

Various statistical methods are used in the context of econometrics to identify and quantify inequality [10]: typical applications are the evaluation of the distribution of resources within a country, or in general within a group of individuals. For instance, the most common measure of inequality, the Gini index, is often reported in the press and in mainstream media to express the income distribution among the citizens of a country. It is a number between 0 and 1; its calculation is based on residents' net income and the resulting value measures the gap between the rich and the poor. Perfect equality among the inhabitants corresponds to a Gini index of zero, while Gini index 1 represents perfect inequality, i.e. a single individual concentrating the whole income.

Other measures of inequality are the Ricci-Schutz coefficient (also known as Pietra index), Theil's entropy measure and Atkinson's measure. Other measures of inequality are related to
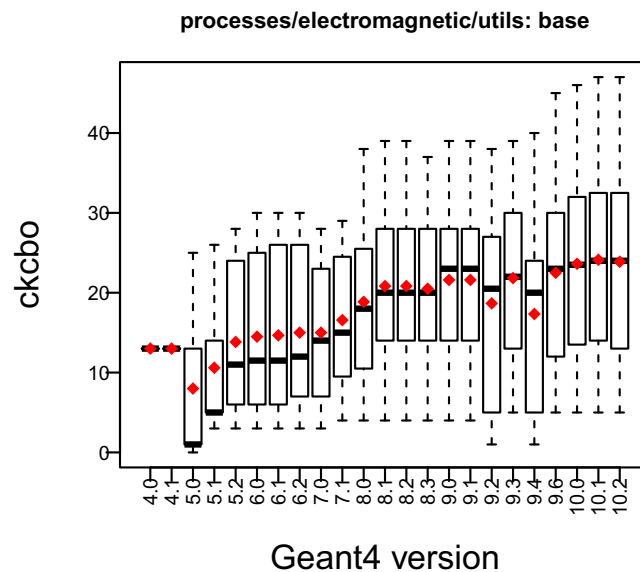
processes/electromagnetic/utils: base

Figure 2. Evolution of the Coupling Between Objects software metric of base classes in the *utils* package of Geant4 electromagnetic physics versus Geant4 version.
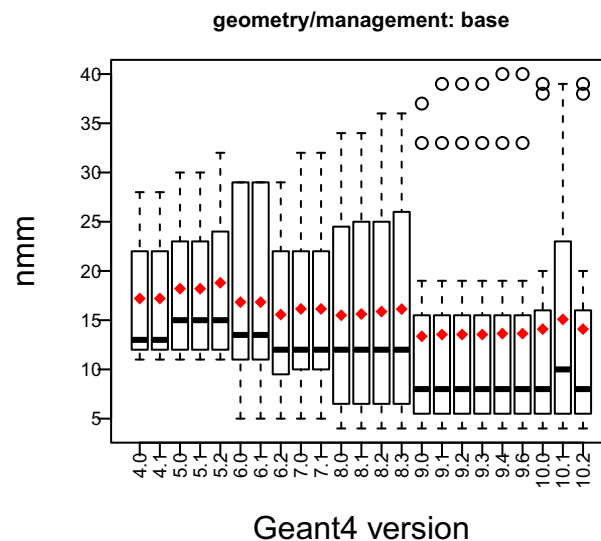
geometry/management: base

Figure 3. Evolution of the number of member methods in the base classes of the Geant4 geometry *management* package as a function of Geant4 version.

the concept of entropy in the context of information theory. Their specific features and their role in the analysis of Geant4 software quality will be extensively discussed in a full paper currently in preparation for publication in a scholarly journal.

An example of inequality analysis is shown in Figs. 4-7, which concern the Halstead Program Volume metric [11] calculated over the *solids* package of Geant4. This metric, along with others developed by Halstead, is a measure of complexity of the software. The plots are the result of a preliminary analysis; they exhibit similarities across the various inequality mesures, although
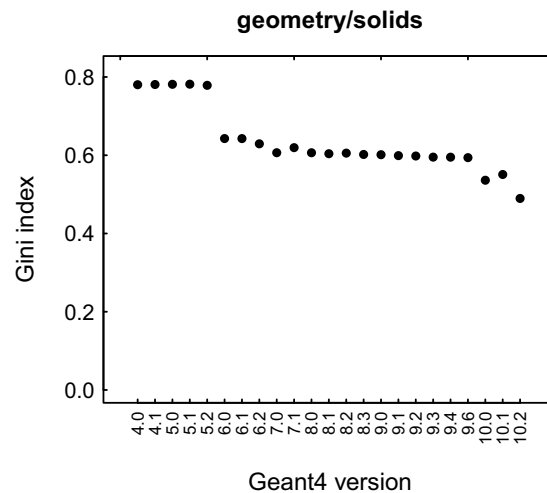
Figure 4. The Gini index calculated over Halstead Program Volume measure in Geant4 solids package, as a function of Geant4 version.
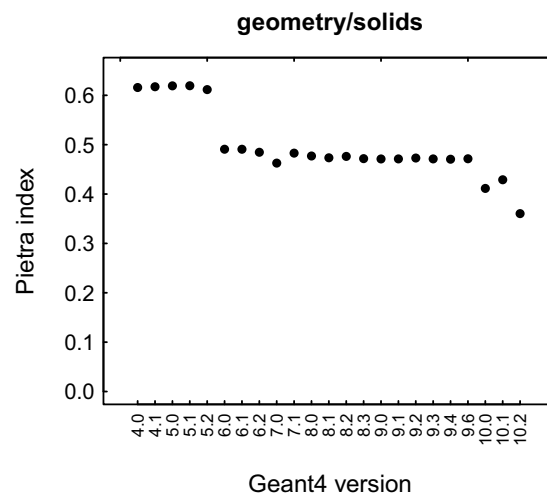

Figure 5. The Pietra index calculated over Halstead Program Volume measure in Geant4 solids package, as a function of Geant4 version.

their numerical values are different.

4. Conclusion
Data analysis methods pertinent to disciplines other than physics are powerful instruments also in the context of physics software. This conference paper summarizes a brief overview of innovative applications of these techniques physics scenarios, such as the validation of simulation models and the analysis of the quality of scientific software. A wider report on the investigation of analysis methods derived from econometrics and ecology, along with more extensive results of their application, will be documented in a forthcoming publication in a scholarly journal.
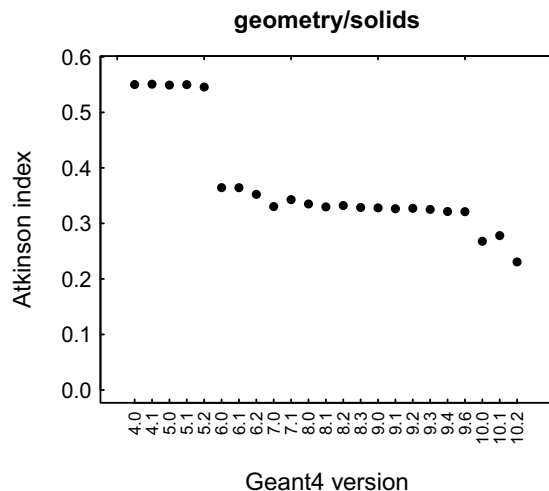
Figure 6. The Atkinson index calculated over Halstead Program Volume measure in Geant4 solids package, as a function of Geant4 version.
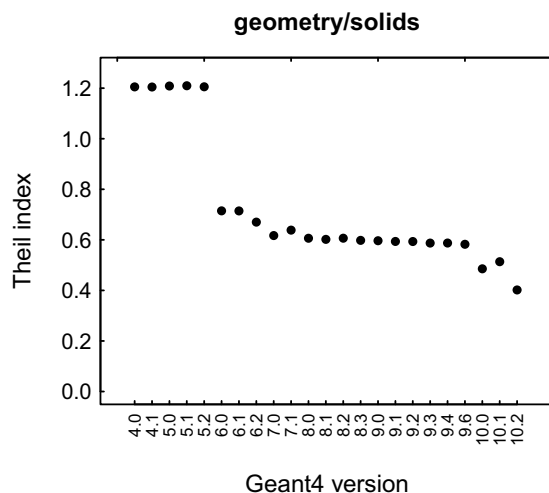

Figure 7. The Theil index calculated over Halstead Program Volume measure in Geant4 solids package, as a function of Geant4 version.

Acknowledgment

[1] Agostinelli S *et al* 2003 *Nucl. Instrum. Meth. A* **506** 250
[2] Allison J *et al* 2006 *IEEE Trans. Nucl. Sci.* **53** 270
[3] Kim S H *et al* 2015 *IEEE Trans. Nucl. Sci.* **62** 451
[4] Basaglia T *et al* 2015 *IEEE Trans. Nucl. Sci.* **62** 1812
[5] Basaglia T *et al* 2016 *IEEE Trans. Nucl. Sci.* **63** 2849
[6] Mann H B 1945 *Econometrica* **13** 163
[7] Kendall M G 1975 *Rank Correlation Methods* 4th edition, Charles Griffin, London
[8] Pia M G *et al* 2016 *Proc. IEEE Nucl. Sci. Symp.*, in press
[9] Ronchieri E *et al* 2015 *J. Phys. Conf. Series* **664** 062053
[10] Cowell F A 2015 *Handbook of Income Distribution*, Ed. A. B. Atkinson, F. Bourguignon, North Holland
[11] Halstead, M H 1977 *Elements of Software Science* Elsevier North-Holland