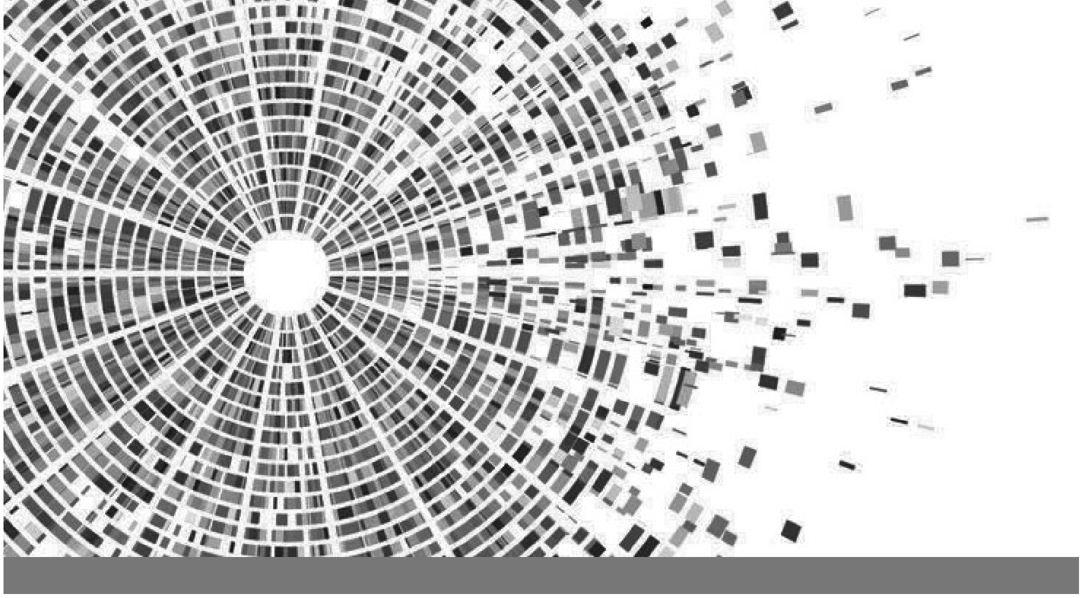

CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Pearson

SCIENTIFIC PROGRAM COMMITTEE

Carla Rampichini (chair, University of Florence - Italy)

Claudio Agostinelli (University of Trento - Italy)
Michela Battauz (University of Udine - Italy)
Antonio Canale (University of Padua - Italy)
Carlo Cavicchia (Erasmus University Rotterdam - Netherlands)
Claudio Conversano (University of Cagliari - Italy)
Eustasio del Barrio (University of Valladolid - Spain)
Roberto Di Mari (University of Catania - Italy)
Stefania Fensore (University of "G. d'Annunzio" - Italy)
Nial Friel (University College Dublin - Ireland)
Maria Giovanna Ranalli (University of Perugia - Italy)
Leonardo Grilli (University of Firenze - Italy)
Luigi Grossi (University of Padua - Italy)
Christian Hennig (University of Bologna - Italy)
Mia Hubert (KU Leuven - Belgium)
Alfonso Iodice D'Enza (University of Naples "Federico II" - Italy)
Julien Jacques (University of Lyon - France)
José Joaquim Dias Curto (ISCTE-Instituto Universitário de Lisboa- Portugal)
Michele La Rocca (University of Salerno - Italy)
Silvia Montagna (University of Turin - Italy)
Barbara Pawelek (University of Cracow - Poland)
Fulvia Pennoni (University of Milano-Bicocca - Italy)
Mario Rosario Guarracino (University of Cassino - Italy)
Katrijn Van Deun (University of Tilburg - Netherlands)
Simone Vantini (Politecnico di Milano - Italy)
Donatella Vicari (Sapienza University of Rome - Italy)
Helga Wagner (Johannes Kepler University Linz - Austria)
Hiroshi Yadohisa (Doshisha University - Japan)

LOCAL PROGRAM COMMITTEE

Michele La Rocca (chair, University of Salerno - Italy)

Pietro Coretto (University of Salerno - Italy)
Giuseppe Giordano (University of Salerno - Italy)
Paolo Rocca Comite Mascambruno (University of Salerno - Italy)
Marcella Niglio (University of Salerno - Italy)
Maria Lucia Parrella (University of Salerno - Italy)
Marialuisa Restaino (University of Salerno - Italy)
Domenico Vistocco (University of Naples "Federico II" - Italy)
Maria Prosperina Vitale (University of Salerno - Italy)

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:
14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13, 2023
edited by Carla Rampichini, Michele La Rocca, Pietro Coretto, Giuseppe Giordano, Maria Lucia Parrella

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,
© Tartila / Shutterstock

Published by Pearson Education Resources, Italia
www.pearson.it
ISBN: 9788891935632

Stratified Sampling on Data Nuggets: a strategy for Data Reduction

Ravi Kumar Gangadharan, Vanessa Petrarca, Maria Chiara Pagliarella, Giovanni Porzio

Department of Economics and Law, University of Cassino and Southern Lazio,
(e-mail: mc.pagliarella@unicas.it, porzio@unicas.it)

Abstract: The increased volume and velocity of data production has been causing a growing cost in storing and analysing data. Thus, due to this continuously increasing phenomenon, the urgency of data reduction technique arises.

Data reduction aims at decreasing storage and computational costs for data analysis. In order to tackle with this very large and complex issue, many techniques have been developed and employed (such as clustering, principal points, support points, prototypes, etc.). Among the many, this work focuses on a recently introduced specific type of data reduction method which has been called Data Nuggets.

Data Nuggets reduces huge datasets and compresses the observations into few points, by saving essential information on the data structure. In parallel with standard classic procedures, Data Nuggets splits a dataset in several subsets (called Nuggets) which are defined by three main components: a Center, a Weight (representing the number of observations within each subset), and a Scale (representing the average Nugget within variance).

Particularly, our work aims at investigating to what extent Data Nuggets can be used as a tool to obtain stratified samples from large datasets so that some computational cost can be gained. A comparison in terms of efficiency with respect to statistical techniques applied to a simple random sample drawn from the same large dataset will be provided.

Keywords: Large datasets, computational effort, data partition.

References

Beavers, T., Cabrera, J., & Lubomirski, M. 2020. datanugget: Create, Refine, and Cluster Data Nuggets. *R package version 1.0.0, https://CRAN.R-project.org/package=datanugget*.

Cherasia K.E., Cabrera J., Fernholz L.T., & Fernholz R. 2023. Data Nuggets in Supervised Learning. In M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*, Springer, https://doi.org/10.1007/978-3-031-22687-8_20.