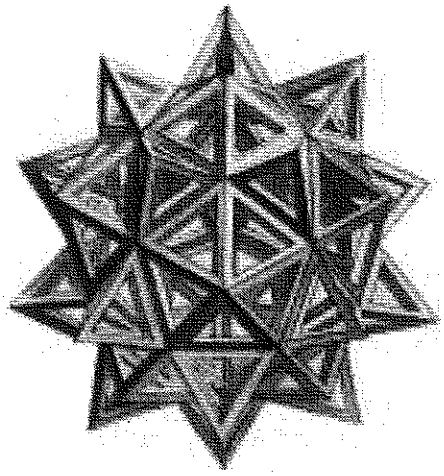
CLADAG2017



Book of Short Papers

Editors: Francesca Greselin,
Francesco Mola and Mariangela Zenga

via Sottoriva, 9 - 46100 Mantova (MN)
P. IVA 02346110204
E-book (PDF version) published in September 2017
ISBN 978-88-99459-71-0

This book is the collection of the Abstract / Short Papers submitted by the authors of the International Conference of The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS), held in Milan (Italy), University of Milano-Bicocca, September 13-15, 2017.

Euro 9,00

Keynotes

Statistical models for complex extremes

Antony Davison,
Institute of Mathematics,
Ecole Polytechnique Federale de Lausanne, Switzerland

Classified Mixed Model Prediction

J. Sunil Rao,
Division of Biostatistics,
Department of Public Health Sciences, University of Miami, Florida

An URV approach to cluster ordinal data

Roberto Rocci,
Dipartimento di Economia e Finanza,
Università degli studi di Tor Vergata, Rome, Italy

Contributed sessions

Classification of Multiway and Functional Data

A generalized Mahalanobis distance for the classification of functional data
Andrea Ghiglietti, Francesca Ieva, Anna Maria Paganoni

Classification methods for multivariate functional data with applications to biomedical signals
Andrea Martino, Andrea Ghiglietti, Anna M. Paganoni

A new Biclustering method for functional data: theory and applications
Jacopo Di Iorio, Simone Vantini

A leap into functional Hilbert spaces with Harold Hotelling
Alessia Pini, Aymeric Stamm, *Simone Vantini*

Sampling Designs and Stochastic models

Statistical matching under informative probability sampling
Daniela Marella, Danny Pfeffermann

Goodness-of-fit test for discrete distributions under complex sampling design
Pier Luigi Conti

Structural learning for complex survey data
Daniela Marella, *Paola Vicard*

The size distribution of Italian firms: an empirical analysis
Anna Maria Fiori, Anna Motta

Robust statistical methods

New proposal for clustering based on trimming and restrictions
Luis Angel Garcìa Escudero, Francesca Greselin, *Agustin Mayo Iscar*

Wine authenticity assessed via trimming
Andrea Cappozzo, Francesca Greselin

Robust and sparse clustering for high-dimensional data
Sarka Brodinova, Peter Filzmoser, Thomas Ortner, Maia Zaharieva, Christian Breiteneder

M-quantile regression for multivariate longitudinal data
Marco Alfo', *Maria Francesca Marino,* Maria Giovanna Ranalli, Nicola Salvati, Nikos Tzavidis

New proposals in Clustering methods

Reduced K-means Principal Component Multinomial Regression for studying the relationships between spectrometry and soil texture
Pietro Amenta, *Antonio Lucadamo*, Antonio Pasquale Leone

Comparing clusterings by copula information based distance
Marta Nai Ruscone

Fuzzy methods for the analysis of psychometric data
Isabella Morlini

Inverse clustering: the paradigm, its meaning, and illustrative examples
Jan W. Owsinski, Jaroslaw Stanczak, Karol Opara, Slawomir Zadrozny

Big data mining and classification

The importance of the minorities' viewpoints: Rare Event Sampling Technique on Sentiment analysis supervised algorithm
Marika Arena, *Anna Calissano,* Simone Vantini

A generalized K-means algorithm for multivariate big data with correlated components
Giacomo Aletti, *Alessandra Micheletti*

Big data process analysis: from data mining to process mining
Massimiliano Giacalone, Carlo Cusatelli, Roberto Casadei, Angelo Romano, Vito Santarcangelo

Semiparametric estimation of large conditional variance-covariance and correlation matrices with an application to financial data
Claudio Morana

Advances in model-based clustering

Probabilistic Distance Algorithm generalization to Student's t mixtures
Christopher Rainey, Cristina Tortora, *Francesco Palumbo*

Model-based Clustering of Data with Measurement Errors
Michael Fop, Thomas Brendan Murphy, Lorraine Hanlon

Gaussian Mixture Modeling Under Measurement Uncertainty
Volodymyr Melnykov, Shuchismita Sarkar, Rong Zhengi

A dynamic model-based approach to detect the trend of Statistics from 1970 to 2015
Laura Anderlucci, Angela Montanari, Cinzia Viroli

Bayesian methods and networks

Non parametric Bayesian Networks for measurement error detection
Daniela Marella, Paola Vicard, *Vincenzina Vitale*

Sparse Naïve Bayes Classification
Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo, *M. Remedios Sillero-Denamiel*

A Constraint-based Algorithm for Nonparanormal Data
Flaminia Musella, *Paola Vicard,* Vincenzina Vitale

Interventional data and Markov equivalence classes of DAGs
Federico Castelletti, Guido Consonni

Categorical data analysis

Study of context-specific independencies through Chain Stratified Graph Models for categorical variables
Federica Nicolussi, Manuela Cazzaro

Redundancy Analysis Models with Categorical Endogenous Variables: A New Estimation Technique
Gianmarco Vaccà

Mixture of copulae based approach for defining the subjects distance in cluster analysis
Andrea Bonanomi, Marta Nai Ruscone, Silvia Angela Osmetti

Dissimilarity profile analysis for assessing the quality of imputation in cardiovascular risk studies
Nadia Solaro

Data Analysis

Measuring vulnerability: a Structural Equation Modelling approach
Ambra Altimari, *Simona Balzano*, Gennaro Zezza

On the turning point detection in financial time series
Riccardo Bramante, *Silvia Facchinetti*

Optimization of the Listwise Deletion Method
Graziano Vernizzi, Miki Nakai

Discretization of measures: an IRT approach
Silvia Golia

Mixture and Latent Class Models for Clustering

Analysis of university teaching quality merging student ratings with professor characteristics and opinions
Francesca Bassi, Leonardo Grilli, Omar Paccagnella, *Carla Rampichini,* Roberta Varriale

Clustering technique for grouped survival data with a nonparametric frailty term
Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples

A latent trajectory model for migrants' remittances: an application to the German Socio-Economic Panel data
Silvia Bacci, Francesco Bartolucci, Giulia Bettin, *Claudia Pigini*

Stepwise latent Markov modelling with covariates in presence of direct effects
Roberto Di Mari, Zsuzsa Bakk

Network analysis

Non-parametric inference for network-valued data
Ilenia Lovato, Alessia Pini, Aymeric Stamm, Simone Vantini

Applying network analysis to online news big data
Giovanni Giuffrida, Simona Gozzo and Francesco Mazzeo Rinaldi, *Venera Tomaselli*

Interval Regression Analysis for the Representation of the Core-Periphery Structure on Large Networks
Carlo Drago

A Latent Space Model for Multidimensional Networks
Silvia D'Angelo, Thomas Brendan Murphy and Marco Alfò

Advances in LMs, GLMs and PCA

Bootstrap prediction intervals in linear models
Davide Passaretti, Domenico Vistocco

Bayesian Variable Selection in Linear Regression Models with non-normal Errors
Saverio Ranciati, Giuliano Galimberti and Gabriele Soffritti

Principal Component Analysis: the Gini Approach
Stéphane Mussard, *Téa Laurent Jérome Akeywidi Ouraga*

On Proportional Odds Modelling and Marginal Effects of Sardinian Hotels data
Giulia Contu, Claudio Conversano, Thomas W. Yee

Advances in Classification

Using PAM and DTW for time series classification
Ilaria Lucrezia Amerise

On Support Vector Machines under multiple-cost scenario
Sandra Benítez-Peña, Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo

Macroeconomic forecasting: a non-standard optimisation approach to the calibration of dynamic factor models
Fabio Della Marra

Classification of Textual Data

Measuring popularity from Twitter
Farideh Tavazoee, Claudio Conversano, Francesco Mola

A Gamification Approach to Text Classification in R
Giorgio Maria Di Nunzio

From unstructured data and word vectorization to meaning: Text mining in Insurance
Mattia Borrelli, Diego Zappa

Gamlss for Big Data: ROC curve prediction using Twitter data
Paolo Mariani, *Andrea Marletta,* Mariangela Sciandra

Evaluation in Education

Nonparametric mixed-effects model for unsupervised classification in the Italian education system
Chiara Masci, Francesca Ieva, Anna Maria Paganoni, Tommaso Agasisti

Multivariate mixed models for assessing equity and efficacy in education. An analysis over time using EU15 PISA data
Isabella Sulis, Francesca Giambona, Mariano Porcu

A zero-inflated beta regression model for predicting first-year performance in university career
Matilde Bini, Lucio Masserini

Students' satisfaction in higher education: how to identify courses with low-quality teaching
Marco Guerra, *Francesca Bassi,* José G. Dias

Statistical models for complex data

Spatial Survival Models for Analysis of Exocytotic Events on Human beta-cells Recorded by TIRF Imaging
Thi Huong Phan, Giuliana Cortese

Testing different structures of spatial dynamic panel data models
Francesco Giordano, Massimo Pacella, *Maria Lucia Parrella*

Identification of earthquake clusters through a new space-time-magnitude metric
Renata Rotondi, Antonella Peresan, Stefania Gentili, Elisa Varin

A circular density strip plot
Davide Buttarazzi, Giovanni Camillo Porzio

Mixture Models

A special Dirichlet mixture model for multivariate bounded responses
Agnese Maria Di Brisco, Sonia Migliorati

Cluster-Weighted Beta Regression
Marco Alfò, *Luciano Nieddu,* Cecilia Vitiello

A Special Dirichlet Mixture Model in a Bayesian Perspective
Roberto Ascari, Sonia Migliorati, Andrea Ongaro

Advances in data Analysis

Assessing Heterogeneity in a Matching Estimation of Endogenous Treatment Effect
Maria Gabriella Campolo, *Antonino Di Pino,* Edoardo Otranto

Template matching for hospital comparison: an application to birth event data in Italy
Massimo Cannas, Paolo Berta, Francesco Mola

On variability analysis of evolutionary algorithm-based estimation
Manuel Rizzo

BIG DATA PROCESS ANALYSIS : FROM DATA MINING TO PROCESS MINING

Massimiliano Giacalone [1] , Carlo Cusatelli[2], Roberto Casadei[3], Angelo Romano[3], and last Vito Santarcangelo[4]

[1] Department of Economics and Statistics, University of Naples 'Federico II', (e-mail: massimiliano.giacalone@unina.it)

[2] Ionian Department, University of Bari 'Aldo Moro'

[3] iInformatica S.r.l.s., Corso Italia 77, Trapani

[4] Department of Mathematics and Computer Science, University of Catania

ABSTRACT: Process mining is the approach that extracts real workflows by database of events (logs) and compares them to the predefined procedures estimating the process gap for process improvement. It is a different overture from Data Mining that extracts hidden information and relations from data, with whom it is often confused. The most important tool for the development of process mining is ProM, an open source suite which implements a lot of technical approaches for process mining.
This paper aims to present Process Mining approach showing the differences from Data Mining, and the implementation by ProM on real logs of an Italian company, comparing the extracted workflows to ISO9001 predefined procedures.

KEYWORDS: Big Data, Process Mining, Process Tool, Data Mining.

1 Introduction

Data analysis for clustering and extracting patterns from a big group of data is an old issue.
Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s) (Pelloni, 1987).
During the whole 20th century, the fields that involved more the development of this branches of statistics were the finance and (of course) the birth of computer science. Finding a valid way to understand the relationship among data is essential in finance analysis when predicting future trends.
Especially computer science-after the launch of web-lead the problem to a higher level. Infact, the dimension of random data that can be extracted from the web is huge. This is the reason why, since the 90's, Data Mining became

the heart of various branches identified as Business intelligence and Big Data analysis played also an important rule for DNA studies, data warehouse and for meta data business intelligence (Giacalone, Scippacercola, 2016).

2 Data Mining and Process Mining comparison

Data Mining and Process Mining can be categorised as Business Intelligence that refers to techniques and tools used to analyse large amounts of digital data and retrieve valuable business knowledge out of them. This purpose is as true for data mining techniques as process mining techniques, even if with different perspectives on the analysis and the results they produce. Both techniques are used to analyse large amounts of data, that it would be impossible to analyse manually and they produce information that can be used in business decisions (D'Alessandro, et al, 2015).

Data Mining techniques are primarily used to find patterns in a large data sets. With data mining techniques it may be possible to find that certain categories of customers demand a certain product, or to find that the customers who most frequently buy product A are also the ones who just as often buy product B, or that the products placed on a specific location in the shop are also the ones that sell the best. Or in a medical analysis that patients that smoke are the most related to develop lung cancer, or that a large consume of alcohol increase the amount of depressed people. In this way is possible to understand important relationships to improve a business to plan more awareness against cancer.

Process mining is not used to find relationship data patterns, but rather to find process relationships among data. Process relationship among data tries also to analyse the relationship between causes and effects among the data in a certain process. The input to the process mining analysis are event logs, audit trails, events. So, the analysis provides an overview of processes and activities. Process mining's perspective is not on patterns in the data but in the process events (Trnka, 2010).

Process Mining is the 'missing link' between data mining and traditional BPM (Business Process Management).

Data Mining provides valuable insights through analysis of data, but it does not generally concern processes. The scope of the two branches is to give a powerful instrument in order to better understand data and process and then to find a way to underline the data relationship of pattern and process to find out the weakness and try to improve the business.

3 Process Mining for the Process Gap Analysis

One of the most important tools used to conduct real Process Mining from logs is ProM, an extensible framework, written in Java, that supports a wide variety of process mining techniques in the form of plug-ins. The input of ProM is represented by logs, characterized by events, concepts and timestamps. Considering an OLTP log, events are represented by the tasks of the operators (concepts) and timestamps are the date and time records of the operations.
The choice of the algorithm of process mining determines the different representation of the process analyzed. In the following images there are the workflows obtained through the use of inductive visual Miner and Petri Net. This is an important choice to focus the attention on the Process Gap Analysis.

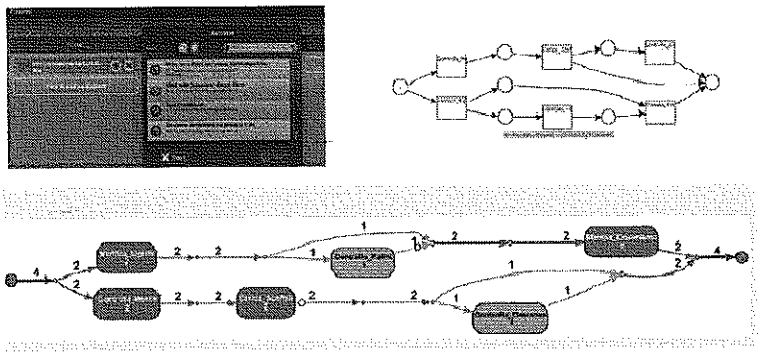


Figure 1. *ProM mining*

The evaluation of the process gap is the estimation of the distance between the actual process model and the expected process model. The last is a process known as conformance checking (Van der Aalst, et al, 2012).
These models differ by nature as the former is merely descriptive whereas the latter is essentially prescriptive. In addition, they are usually developed with both different conceptual and practical tools and, as a consequence, they may also be represented in different ways and formats. In general, moreover, the models can be encoded in multiple representations to serve different goals. When comparing two models using a practical computer-based procedure, obvious prerequisites include the comparability of their (ultimate) representations – which must be digital, formal, unambiguous – and the comparability at the conceptual meta-level. In this sense, the flexibility of ProM for what concerns the output format represents a valuable feature for a tool which aimes

to support automated process analysis.The intended process model is created in order to document how the actual business process should be carried out. To fulfill its prescriptive role, a hard or soft copy of the procedural description of the process is typically handed out to the operational stakeholders. Such a procedural specification is often represented in natural language – which is inherently subject to ambiguity – possibly accompanied by diagrams expressed in informal or semi-formal notations (e.g., UML activity diagrams or BPMN (White, Stephen, 2004)).
In more concrete terms, the key question is the following one: how can we create – e.g., from process descriptions expressed in natural language – a model that can be used to produce representations that can be effectively compared against the actual process description – e.g., as produced by a tool such as ProM? It seems reasonable to have the chosen approach provide for i) a common semantic layer to give name and meaning to process elements, ii) a well-defined notation (comprehensible and/or usable by business experts) for describing processes with clear semantic links, and iii) tools to analyse and compare process descriptions according to proper semantic rules.
In practice, the choice of the modeling language is not easy because a tension exists between expressivity and analysability. For example, a notation such as BPMN, while suitable for modeling, tends to produce diagrams that are not amenable to analysis – unless considering a proper BPMN subset or transforming BPMN diagrams to Petri nets (Kalenkova, et al, 2015).
By the way, keeping a distinction between 'external' models (employed for process specification and human communication) and 'internal' models (used for analysis) may be valuable. Conformance checking is commonly implemented by replaying history (i.e., event logs) on the expected process model, which is typically represented as a transition system such as a Petri net. However, the initial model representation may be different. For instance, it is possible to load a BPMN diagram into ProM, which results in a BPMN-to-Petri-Net conversion, and then use the tool to analyse and enrich the model with conformance information (Kalenkova, et al, 2015).

4 Process Mining on ISO 9001 process

An interesting application of Process Mining is related to the control of processes based on ISO STANDARDs, as the ISO9001 (International Organization for Standardization, 2015).
ISO 9001:2015 sets out the criteria for a quality management system and it

is a certifiable standard. It can be used by any organization, large or small, regardless of its field of activity. In fact, there are over one million companies and organizations in over 170 countries certified.

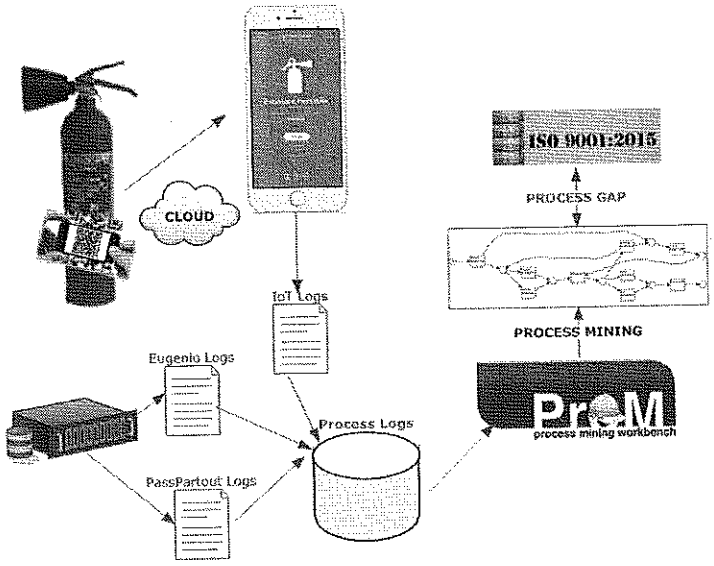


Figure 2. *Process Mining on ISO 9001 Maintenance Extinguishers process*

As a case study application of process mining, we show the process mining applied to an ISO 9001 Maintenance Extinguishers process of an Italian company. Process Mining requires logs about the process, then to avoid this task we have considered the log of Eugenio (fire extinguishers management tool), PassPartout (accounting tool) and the log of the Internet of Things scenario of security objects.
The IoT scenario has been made through the use of unique QRCODE on each devices that produces an innovative interactive network of objects, producing also OLTP logs thanks to the interactions of the users (maintainers and clients) with the single objects. All these logs have been analyzed through ProM suite and compared with the models obtained from ISO 9001 Procedures.
The process gap obtained underlined the underestimation of the Work Plans imparted by the specialist coordinator of maintainers and the incorrect completion of the work plans by the maintainers that made the intervention. After one year, the process mining analyis has led to a renovation of some workflows with noted improvement of efficiency and employees' awareness.

5 Final Remarks

We have created a dataset named "ProLantincendio" that collects data from 3 years (2014, 2015 and 2016) for all logs and data from Eugenio software, PassPartout, and Internet of Things software. In fact, this data warehouse has as its common elements the individual protection devices and the individual operators, present in all 3 software collections that come into the information dataset which our business started from. This data warehouse allows to proceed with 2 types of analysis, Process Mining on Transaction Data (Log) and Data Mining on non-transactional data (Sales, Estimates, Contracts, Customers, Maintenance, Timeline, Operator Performance). Process Mining from process evidence, modeling its real flow and allowing comparison with the process defined by procedures (the company is ISO 9001 certified). Data Mining does not extract knowledge from logs, but extracts hidden know-how within heterogeneous sources. Our experiments focused essentially on the use of algorithm J48 (implementation of C4.5 decision trees) with the extraction of evidence between operator performances, time of year, localization of the customer's business and maintainer's experience.

References

D'Alessandro, M.T., Santarcangelo, V. et al. (2015). Process Mining : review and case study. Choice and preference analysis for quality improvement and seminar on experimentation. ASA 2015 Conference, Bari.

Giacalone M., Scippacercola S. (2016). Big Data: Issues and an Overview in Some Strategic Sectors, Journal of Applied Quantitative Methods , Issue 3, vol. 11 pp. 1 -17.

International Organization for Standardization (2015). ISO 9001:2015 Quality Management Systems.

Kalenkova, A., Van der Aalst, W. et al. (2015). Process mining using BPMN: relating event logs and process models, Software & Systems Modeling.

Pelloni, G. (1987). A note on friedman and the neo-bayesian approach, The Manchester School, Volume 55, Issue 4, pp. 407-418.

Trnka, A. (2010). Market Basket Analysis with Data Mining methods, International Conference on Networking and Information Technology.

Van der Aalst, W. et al. (2012). Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Wiley Online Library.

White, Stephen A (2004). Introduction to BPMN, IBM Cooperation.