A (very short) introductory journey through data science

2022 International Workshop in Applied Statistics and Data Science (Cartagena – Colombia)

June 29 – July 1, 2022

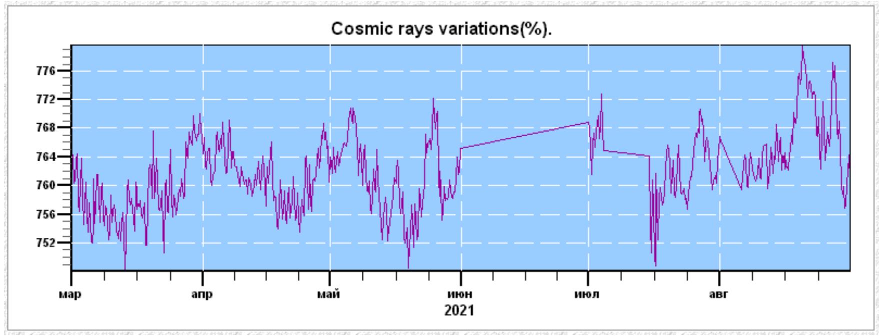
D. Sierra Porta
Faculty of Basic Sciences
Universidad Tecnológica de Bolivar
dporta@utb.edu.co

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

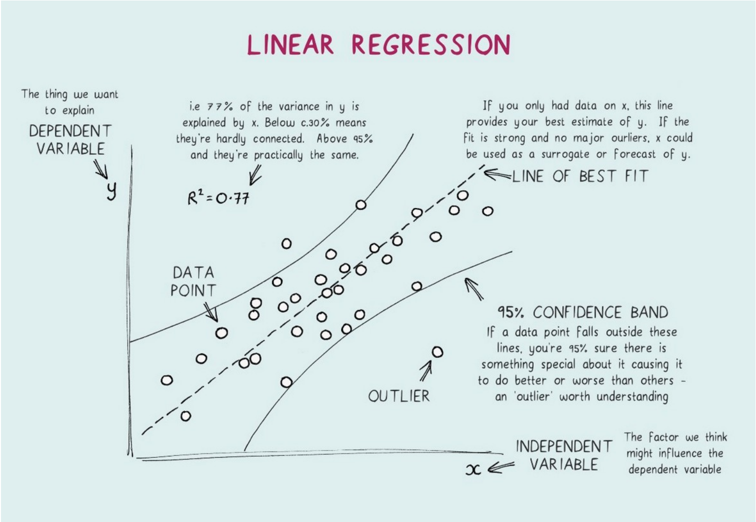
Contents

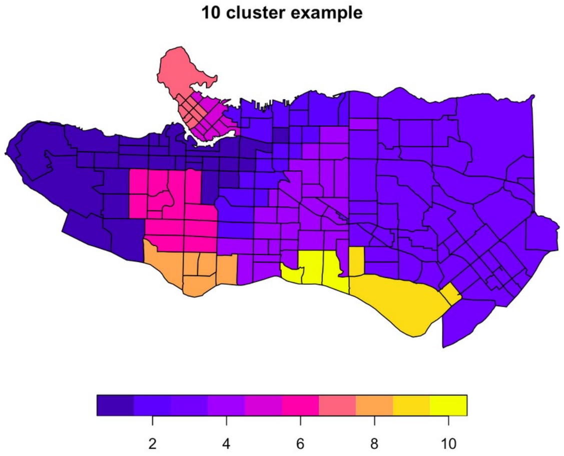
We are talking about three important topics (or common task) in Data Sciene:

Imputation methods

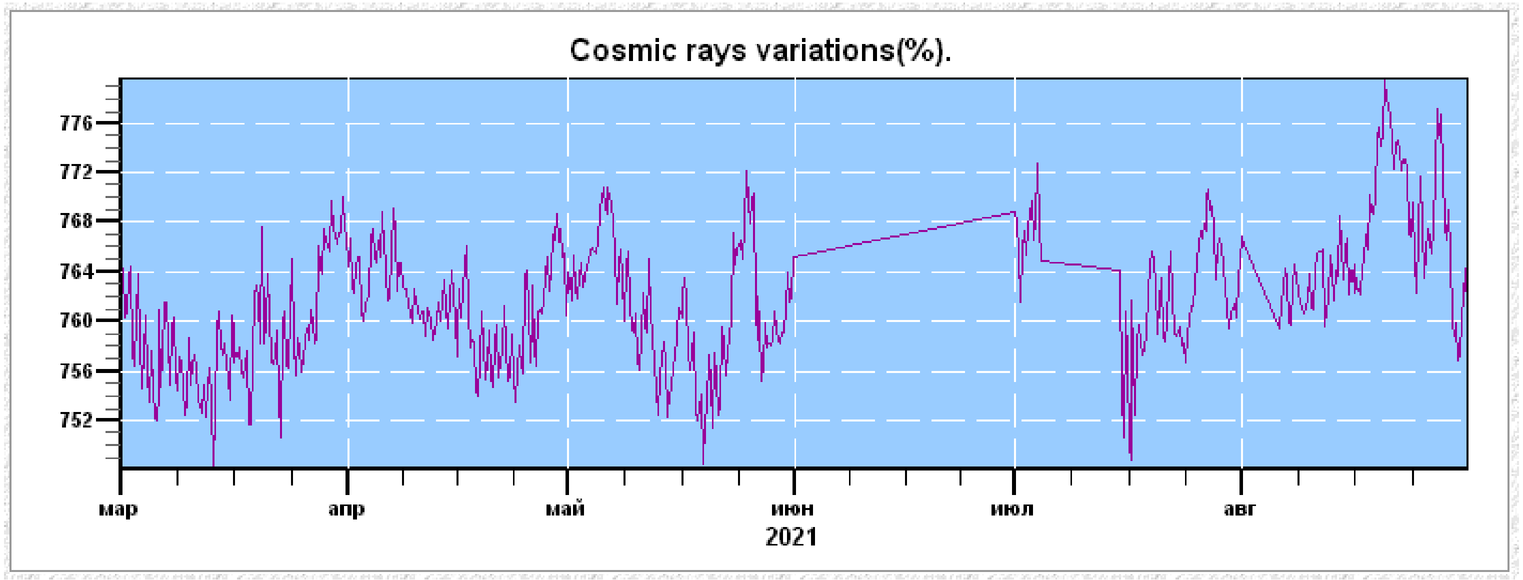


Regression methods





Clustering methods



Tutorial: Introduction to Missing Data Imputation

D. Sierra Porta
Faculty of Basic Sciences
Universidad Tecnológica de Bolivar
dporta@utb.edu.co

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Think fast 1...

131	→	12
241	→	24
162	→	36
343	→	48
146	→	**??**

Think fast 2...

111	→	12
231	→	24
324	→	36
453	→	48
542	→	**??**

Think fast 1..:

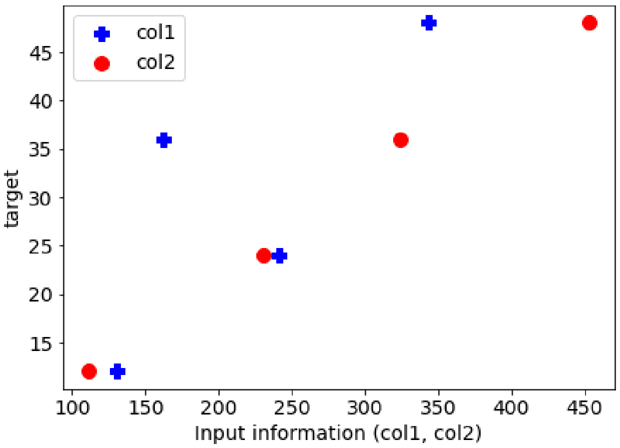
131 → 12
241 → 24
162 → 36
343 → 48
146 → ??

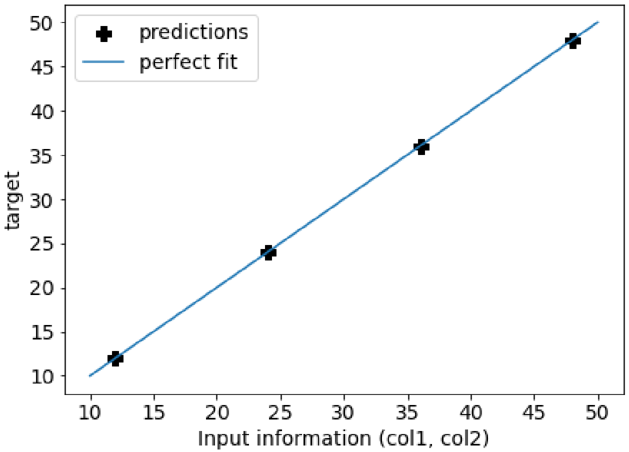
	col1	col2	targ
0	131	111	12
1	241	231	24
2	162	324	36
3	343	453	48
4	146	542	??

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   targ   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.319e+05
Date:                Mon, 27 Jun 2022   Prob (F-statistic):            0.00195
Time:                        08:56:49   Log-Likelihood:                 8.9039
No. Observations:                   4   AIC:                            -11.81
Df Residuals:                       1   BIC:                            -13.65
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2792      0.075     17.113      0.037       0.329       2.229
col1          -0.0160      0.001    -30.284      0.021      -0.023      -0.009
col2           0.1152      0.000    333.744      0.002       0.111       0.120
==============================================================================
```

Think fast 2..:

111 → 12
231 → 24
324 → 36
453 → 48
542 → ??



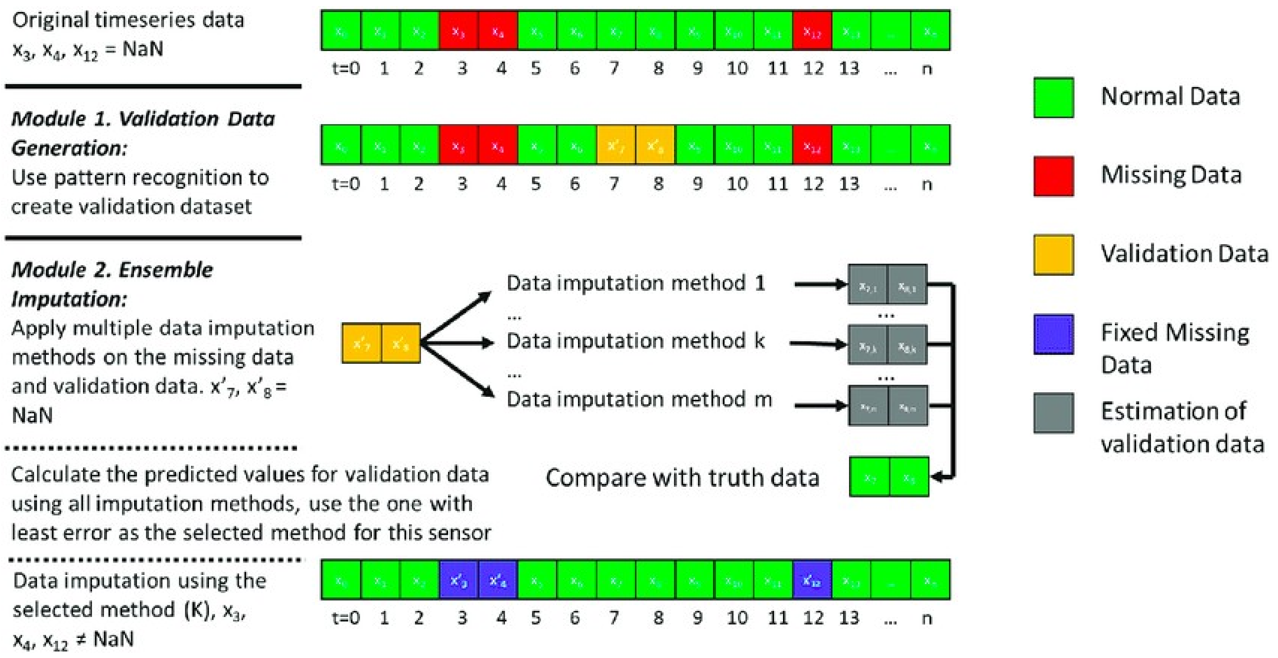


Model: targ ~ col1 + col2, *target(predcition)=**61***

Missing data mechanisms

The study of missing data was formalized by Donald Rubin (see [6], [5]) with the concept of missing mechanism in which missing-data indicators are random variables and assigned a distribution. Missing data mechanism describes the underlying mechanism that generates missing data and can be categorized into three types — missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Informally speaking, MCAR means that the occurrence of missing values is completely at random, not related to any variable. MAR implies that the missingness only relate to the observed data and NMAR refers to the case that the missing values are related to both observed and unobserved variable and the missing mechanism cannot be ignored.

Missing data is a common problem in practical data analysis. They are simply observations that we intend to make but did not. In datasets, missing values could be represented as ‘?’, ‘nan’, ’N/A’, blank cell, or sometimes ‘-999’, ’inf’, ‘-inf’. The aim of this tutorial is to provide an introduction of missing data and describe some basic methods on how to handle them.



https://zenodo.org/record/4731169#.YrmoADXMJ8s

zenodo

Search | Upload | Communities | Log in | Sign up

April 30, 2021 | Dataset | Open Access

Water-quality data imputation with a high percentage of missing values: a machine learning approach

Rafael Rodríguez; Marcos Pastorini; Lorena Etcheverry; Christian Chreties; Mónica Fossati; Alberto Castro; Angela Gorgoglione

The monitoring of surface-water quality followed by water-quality modeling and analysis is essential for generating effective strategies in water resource management. However, water-quality studies are limited by the lack of complete and reliable data sets on surface-water-quality variables. These deficiencies are particularly noticeable in developing countries.

This work focuses on surface-water-quality data from Santa Lucía Chico river (Uruguay), a mixed lotic and lentic river system. Data collected at six monitoring stations are publicly available at https://www.dinama.gub.uy/oan/datos-abiertos/calidad-agua/. The high temporal and spatial variability that characterizes water-quality variables and the high rate of missing values (between 50% and 70%) raises significant challenges.

To deal with missing values, we applied several statistical and machine-learning imputation methods. The competing algorithms implemented belonged to both univariate and multivariate imputation methods (inverse distance weighting (IDW), Random Forest Regressor (RFR), Ridge (R), Bayesian Ridge (BR), AdaBoost (AB), Huber Regressor (HR), Support Vector Regressor (SVR), and K-nearest neighbors Regressor (KNNR)).

IDW outperformed the others, achieving a very good performance (NSE greater than 0.8) in most cases.

In this dataset, we include the original and imputed values for the following variables:

- Water temperature (*Tw*)
- Dissolved oxygen (*DO*)
- Electrical conductivity (EC)
- pH
- Turbidity (*Turb*)
- Nitrite (*NO2-*)
- Nitrate (*NO3-*)
- Total Nitrogen (*TN*)

Publication date: April 30, 2021

DOI: DOI 10.5281/zenodo.4731169

Keyword(s): data scarcity | water quality | missing data | univariate imputation | multivariate imputation

Related identifiers:
Derived from
10.20944/preprints202105.0105.v1 (Preprint)
10.3390/su13116318 (Journal article)

Versions

https://www.mdpi.com/2071-1050/13/11/6318

Open Access | Article

Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach

by Rafael Rodríguez [1], Marcos Pastorini [2], Lorena Etcheverry [2], Christian Chreties [1], Mónica Fossati [1], Alberto Castro [2] and Angela Gorgoglione [1,*]

[1] Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA), Facultad de Ingeniería, Universidad de la República, Montevideo 11300, Uruguay

[2] Instituto de Computación (InCo), Facultad de Ingeniería, Universidad de la República, Montevideo 11300, Uruguay

[*] Author to whom correspondence should be addressed.

Academic Editor: Ashwani Kumar Tiwari

Sustainability **2021**, *13*(11), 6318; https://doi.org/10.3390/su13116318

Received: 3 May 2021 / Revised: 30 May 2021 / Accepted: 1 June 2021 / Published: 2 June 2021

(This article belongs to the Special Issue Water Quality: Current State and Future Trends)

View Full-Text | Download PDF | Browse Figures | Citation Export

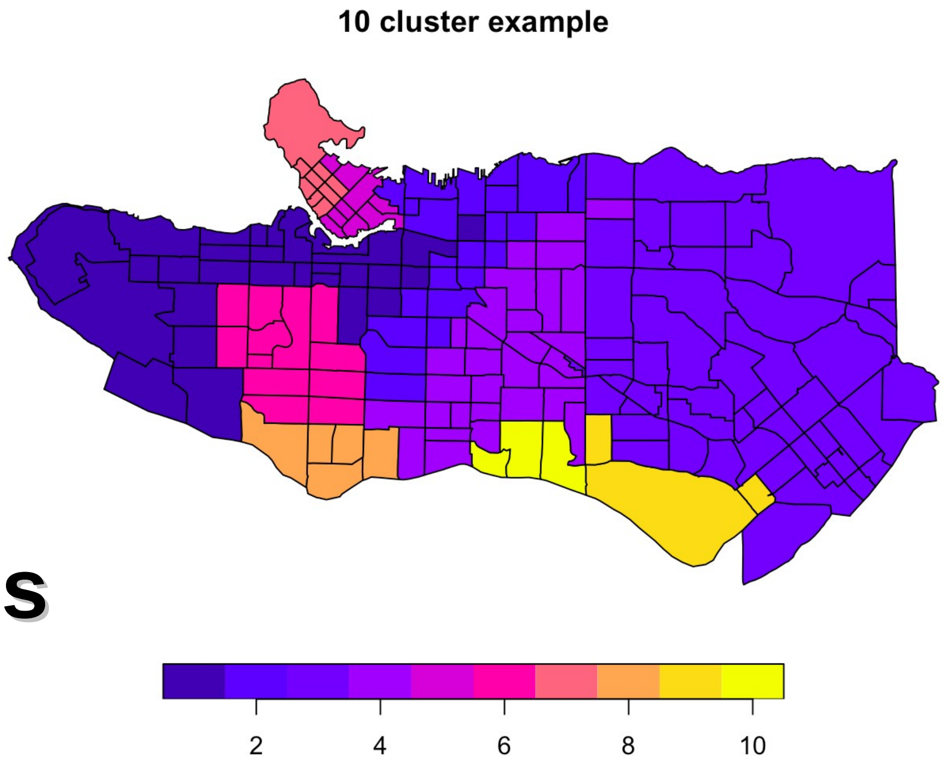
Abstract

The monitoring of surface-water quality followed by water-quality modeling and analysis are essential for generating effective strategies in surface-water-resource management. However, worldwide, particularly in developing countries, water-quality studies are limited due to the lack of a complete and reliable dataset of surface-water-quality variables. In this context, several statistical and machine-learning models were assessed for imputing water-quality data at six monitoring stations located in the Santa Lucía Chico river (Uruguay), a mixed lotic and lentic river system. The challenge of this study is represented by the high percentage of missing data (between 50% and 70%) and the high temporal and spatial variability that characterizes the water-quality variables. The competing algorithms implement univariate and multivariate imputation methods (inverse distance weighting (IDW), Random Forest Regressor (RFR), Ridge (R), Bayesian Ridge (BR), AdaBoost (AB), Hubber Regressor (HR), Support Vector Regressor (SVR) and K-nearest neighbors Regressor (KNNR)). According to the results, more than 76% of the imputation outcomes are considered "satisfactory" (NSE > 0.45). The imputation performance shows better results at the monitoring stations located inside the reservoir than those positioned along the mainstream. IDW was the model with the best imputation results, followed by RFR, HR and SVR. The approach proposed in this study is expected to aid water-resource researchers and managers in augmenting water-quality datasets and overcoming the missing data issue to increase the number of future studies related to the water-quality matter. View Full-Text

Keywords: data scarcity; water quality; missing data; univariate imputation; multivariate imputation; machine learning; hydroinformatics

▼ Show Figures

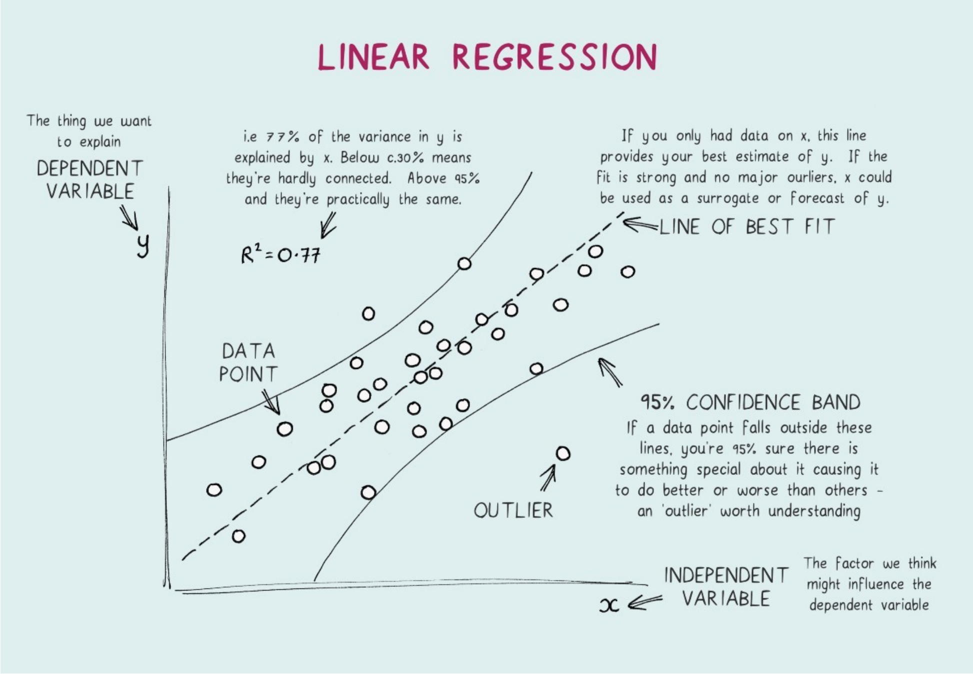
...let's get to the data and get to work…



Clustering methods

D. Sierra Porta
Faculty of Basic Sciences
Universidad Tecnológica de Bolivar
dporta@utb.edu.co

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS



Regression methods

D. Sierra Porta
Faculty of Basic Sciences
Universidad Tecnológica de Bolivar
dporta@utb.edu.co

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS