Chemometrics and Intelligent Laboratory Systems 134 (2014) 23–33

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets

M.P. Gómez-Carracedo [a], J.M. Andrade [a,*], P. López-Mahía [a], S. Muniategui [b], D. Prada [b]

[a] Department of Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña, Spain
[b] Institute of Environmental Sciences, University of A Coruña, Pazo de Lóngora, Liáns, E-15179, Oleiros, Spain

ARTICLE INFO

Article history:
Received 4 December 2013
Received in revised form 20 February 2014
Accepted 23 February 2014
Available online 27 February 2014

Keywords:
Missing data
Single imputation
Expectation–maximization
Multiple imputation
Air quality

ABSTRACT

Datasets with missing data ratios ranging from 24% to 4%, corresponding to three air quality monitoring studies, were used to ascertain whether major differences occur when five currently used imputation methods are applied (four single imputation methods and a multiple imputation one). Unrotated and Varimax-rotated factor analyses performed on the imputed datasets were compared. All methods performed similarly, although multiple imputation yielded more disperse imputed values. Main differences occurred when a variable with missing values correlated poorly to the other features and when a variable had relevant loadings in several unrotated factors, which sometimes changed the order of the rotated factors.

1. Introduction

Incomplete data matrices in pollution studies constitute an insidious, recurrent problem which appears almost always when they are extended on time. Data (isolated or constituting 'gaps') can be missed because of too many uncontrollable situations: malfunctioning of the instruments, maintenance and/or repairing, calibration, etc. Automated immission stations to monitor air quality require scheduled shutdowns for maintenance but also unscheduled ones when unexpected values are recorded. For instance, many air analyzers need 1 or 2 h each fortnight to assess (correct) the zero value and the input flow of air, and a full calibration/maintenance every 6 months (i.e., a full working day shutdown). In addition, remote stations suffer rather long interruptions due to faults on power supply, problems with air aspiration pumps; or malfunctioning of electronic processing cards.

Missing data are problematic because many common chemometric methods cannot handle them and so conclusions cannot be derived. Imputation consists of substituting reasonable estimates for the missing values [1,2]. Conceptually, every specimen-sample in a randomly chosen collection of specimen-samples can be replaced by a new individual that is randomly chosen from the same source population as the original specimen, without compromising the conclusions [3].

A large number of imputation methods were developed [e.g. 4–8], although so far pretty less works focused on comparing their performance in real datasets (some are summarized in the next section). A reason is that most developments must be exemplified using simulated missing data. However, environmental scientists are challenged by datasets with missing data whose values cannot be known at all, with a complex distribution and, so, the performance of the different methods cannot be evaluated on theoretical grounds. As imputation methods rely on a different basis there is a need to compare their performance from a pragmatic viewpoint because not every method can be tested each time imputation is required. Applied method-comparison reports can aid scientists in making their choice because, although the conclusions derived from particular studies are limited, scientists face similar problems frequently. It is worth understanding that:

1) for those working in environmental studies imputation is not an end in itself [2]. There is no way to retrieve the 'true' values of the missing data; the best we can do is to preserve the distributional features that will be used for chemometric studies [9].
2) there is not *a golden* method. Sophisticated approaches excel sometimes although they can be worst than simpler ones in other occasions [2]. Therefore, care is required when using too sophisticated *ad-hoc* techniques ('*in real-life applications where missing data are a nuisance rather than a major focus of scientific enquiry, a readily available, approximate solution with good properties can be preferable to one that is more efficient but problem-specific and complicated to implement*' [1]).
3) basic principles of the imputation techniques may be implausible in real datasets. *'Information is not being invented with multiple*

[*] Corresponding author. Fax: +34 981167065.
E-mail address: andrade@udc.es (J.M. Andrade).

http://dx.doi.org/10.1016/j.chemolab.2014.02.007
0169-7439/© 2014 Elsevier B.V. All rights reserved.

imputation any more than with expectation-maximization or other well accepted likelyhood-based methods […] rather than by simulation' [9].

The scarcity of pragmatic comparative studies to address missing data on real datasets impelled us to select five imputation methods of common use (with widely different bases and increased computation complexity) to study their performance on three air quality monitoring datasets. As the true data are unknown the feasibility of the imputed values was studied by means of factor analysis, a common step in most chemometric studies.

Section 2 presents the datasets, along with an overview of the imputation methods. Section 3 compares the results yielded by each method on each dataset. Finally, some conclusions are given.

2. Experimental

2.1. Samples

An automatic immission station located at a suburban, coastal area close to the city of A Coruña (43° 20′ 14″ N, 8° 21′ 7″ O; NW Spain) was considered. It is 9 km off the centre of the city although within its hinterland (ca. 500,000 inhabitants). Eight gaseous pollutants were measured real-time following international protocols (Table 1). Physical models governing the dispersion of pollutants and their on-site concentration or satellite observations were not available. Daily averages of hourly-measured pollutants will be used throughout.

Three situations in air pollution monitoring were studied, with high (23.5%) to low (3.9%) overall percentages of missing data (Table 1). The highest ratios corresponded to 2006 when the immission station started routine functioning and several sensors and systems had occasional failures. In 2009 and 2010 maintenance and calibration tasks were scheduled to reduce the number of shutdowns and simultaneously control different systems. Thus, although 2006 had pretty more missing data, than 2009 and 2010, the number of days without any value was higher in the latter years.

Despite some percentages of missing data are very high (e.g., 43.5% for O_3 in 2006), this mere number is not of concern for the imputation methods, but their structure and the 'quality' of the information in the measured data are [8]. Gaps due to calibration, repairing, technical problems and maintenance yield complex patterns of voids (Fig. 1) and several variables may have missing values when a sensor fails (as for the nitrogen oxides).

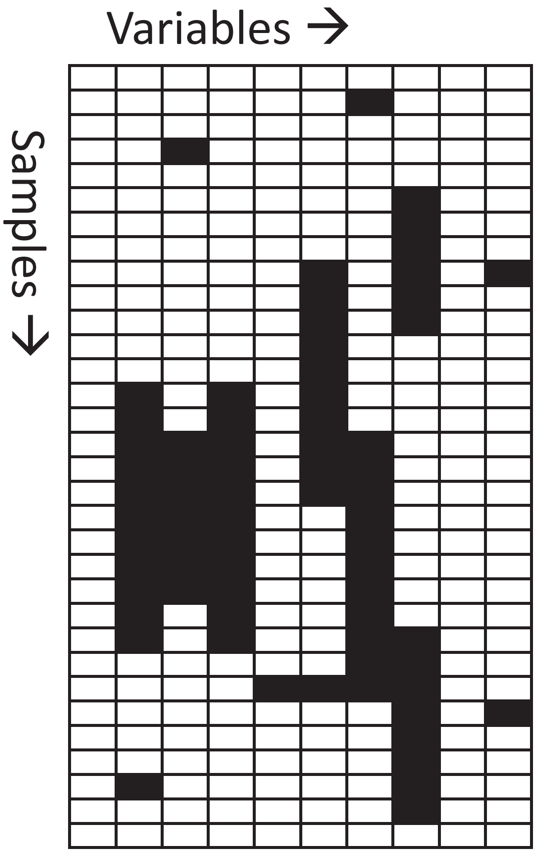

Fig. 1. Appearance of the pattern of missing data in the datasets.

Table 1
Details of the three original datasets, PM = particulate matter (the numbers indicate the average maximum size). UNE-EN stands for a European Analytical Method.

	Univariate statistics								
Year		NO	NO_2	NO_x	CO	O_3	PM_{10}	$PM_{2.5}$	PM_1
2010	Mean	4.28	9.90	16.08	0.08	58.54	18.88	16.48	12.49
	Median	1.98	9.11	13.84	0.07	57.78	17.66	15.61	11.58
	Variance	16.38	23.19	99.13	0.00	429.80	38.52	28.07	24.69
	Skewness	7.98	4.58	5.66	21.70	9.65	17.20	20.72	21.77
	Kurtosis	1.10	0.55	0.59	62.27	21.58	47.03	64.46	68.20
2009	Mean	4.67	15.68	22.61	0.10	48.46	18.09	15.04	11.06
	Median	2.98	13.43	17.87	0.09	48.07	16.02	13.10	9.17
	Variance	20.26	82.75	241.52	0.00	351.32	53.03	37.68	33.66
	Skewness	13.57	9.55	10.25	15.61	1.96	13.82	17.90	18.64
	Kurosis	16.64	9.91	8.65	25.87	0.38	19.26	32.47	33.54
2006	Mean	3.51	6.31	11.03	0.41	53.71	12.77	9.63	7.48
	Median	1.48	4.98	6.86	0.41	55.01	11.16	7.89	5.74
	Variance	18.97	17.25	112.86	0.04	443.87	42.13	26.38	20.95
	Skewness	14.99	10.32	12.66	7.35	1.10	9.78	9.16	10.35
	Kurtosis	17.09	13.38	12.83	14.73	−0.36	7.98	4.99	6.24
	% Missing data								
	Overall	NO	NO_2	NO_x	CO	O_3	PM_{10}	$PM_{2.5}$	PM_1
2010	3.85	5.35	5.35	5.35	6.92	7.86	0	0	0
2009	11.95	16.82	16.82	16.82	27.41	0	5.92	5.92	5.92
2006	23.52	32.23	32.23	32.23	36.36	43.53	3.86	3.86	3.86
		Analytical methodology							
UNE-EN		Chemiluminescence 14211:2006			IR spectrometry 14626:2006	UV–VIS 14625:2005	Gravimetry 12341:1999/14907:2006		