Moshenberg *et al. Environ Syst Res (2015) 4:26*
DOI 10.1186/s40068-015-0052-z

Environmental Systems Research
a SpringerOpen Journal

RESEARCH Open Access

Spectral methods for imputation of missing air quality data

Shai Moshenberg, Uri Lerner and Barak Fishbain*

Abstract

Background: Air quality is well recognized as a contributing factor for various physical phenomena and as a public health risk factor. Consequently, there is a need for an accurate way to measure the level of exposure to various pollutants. Longitudinal continuous monitoring however, is often incomplete due to measurement errors, hardware problems or insufficient sampling frequency. In this paper we introduce the discrete sampling theorem for the task of imputing missing data in longitudinal air-quality time series. Within the context of the discrete sampling theorem, two spectral schemes for filling missing values are presented—a Discrete Cosine Transform (DCT) and Clustering Single Variable Decomposition (K-SVD) based methods.

Results: The evaluation of the suggested methods in terms of *accuracy* and *robustness* showed that the spectral methods are comparable to the state of the art when the data is missing at random and do have the upper hand when data is missing in big chunks. The *accuracy* was evaluated using a complete very long air pollutants time series. Previous studies used incomplete shorter series, altering the results. The *robustness* of the imputation method was evaluated by examining its performance with increasing portions of missing data.

Conclusions: Spectral methods are a great option for air quality data imputation, which should be considered especially when the missing data patterns are unknown.

Keywords: Missing data, Air quality, Univariate, Imputing, Spectral methods, Discrete sampling theorem, Sparse coding, K-SVD

Background

Air quality has a profound effect on our physical and economic health (Künzli et al. 2000; Kampa and Castanas 2008; Laumbach and Kipen 2012). Air pollution is originated either from natural phenomenon or from anthropogenic activity (Cullis and Hirschler 1980; Robinson and Robbins 1970). Regardless of its sources, air pollution undergoes a set of chemical processes in the atmosphere, depending on initial concentration and ambient conditions. The large number of sources and the complexity of the chemical processes lead to the creation of complex scenarios with highly variable spatial and temporal pollution patterns. Thus, the analysis of air-pollution and its effects is a challenging task (Nazaroff and Alvarez-Cohen 2001; Levy et al. 2014; Moltchanov et al. 2015; Lerner et al. 2015).

One of the primary tools to assess air-pollution patterns is through continuous monitoring of pollutants ambient levels. To accomplish this, numerous physico-chemical methods have been developed and Air quality monitoring (AQM) station networks have been deployed all around the world. However, any longitudinal data acquisitioning system suffers from economical constraints, measurement errors, routine downtime due to maintenance and technical malfunctions, which result in missing data points. Data can be missing in long chunks due to a critical failure or in short intervals due to, for example, calibration or a temporary power outage. To cope with this inherent problem, many imputation methods have been proposed. The length of the missing interval and the kind of study conducted, are important in determining the best method for interpolation.

*Correspondence: fishbain@technion.ac.il
The Technion Center of Excellence in Exposure Science and Environmental Health (TCEEH), Faculty of Civil and Environmental Engineering, Technion-Israel Institute of Technology, Haifa 3200003, Israel

Regardless of the method used for assigning the missing values, one can compute one value per missing sample, i.e., single imputation, or a few, which are drawn from a prior distribution—multiple imputation (Little and Rubin 2002). The latter has shown promising results in surveys (Rubin 2004; Su et al. 2011), where the distribution to draw the imputed values from is known or can be assessed from the available samples. Air pollution time series are time variant with rapid, sometimes large changes, which would limit the use of multiple imputation methods. Therefore the focus here is on single imputation methods.

Physical imputation models estimate missing values by utilizing environmental conditions and air quality measurements acquired in other sites before, at and after the fact, and measurements acquired at the same location before and after the fact (Hopke 1991). This approach works if the physical laws governing the different phenomena are well known and relatively simple. However, generally, the nature of the entire physical and chemical processes which govern the observed phenomenon, are either unknown or too complex to be described by an analytical model, thereby rendering the physical model approach unsuitable.

Data driven models, typically, do not assume any physical regime governing the observed phenomenon. These methods fill data gaps by using patterns and relations that are observed in the available data (Solomatine et al. 2008). Data driven methods are either based on a single variable or on multi-variable imputation. Single-variable methods estimate missing values through available measurements of the same environmental variable (e.g., NO_2, CO or O_3). Prominent single variable methods are replacing missing values with the available samples' mean, nearest neighbor (NN), linear interpolation and spline (Junninen et al. 2004). Multi-variable imputation techniques calculate missing samples using data of more than one variable, exploiting relationships between different variables that manifest themselves in the data [e.g., NO_2 versus O_3 presence (Lee et al. 2002; Haagen-Smit et al. 1953)]. All the aforementioned methods, however, are local methods either in space or in time; meaning missing data is recovered by using data from preceding and succeeding available samples (i.e., locality in time) or adjacent stations (i.e., locality in space). Thus, these methods are mostly effective for cases with a relatively low number of missing data points, they are easy to compute but quickly become less accurate as the amount of missing data increases.

Spectral representation of a signal refers to its analysis with respect to frequency, rather than time (Hamilton 1994). Frequency representation of a signal correspond to how much of the signal lies within each given frequency band over a range of frequencies. A signal can be converted between the time and frequency domains through transformations that project the signal onto a set of basis-functions which differ in their change rates, i.e., frequencies. The Fourier transform (Bracewell 1965), for example, projects the time series onto a set of sine waves of different frequencies, each of which represents a frequency component. Similarly, the cosine transform projects the signal on a set of cosine functions oscillating at different frequencies. As the AQM acquired air-pollution time-series, which are inherently discrete, the discrete forms of these transformations—the Discrete Fourier Transform (DFT) (Bracewell 1965) and the Discrete Cosine Transform (DCT) (Rao et al. 1990) can be used. The justification of signals' spectra for analysis is twofold. First, spectral methods are global, i.e. they use the complete signal for computation, not just local extremes, similar sub-sequences or areas near the missing data. Second, as atmospheric composition changes over a finite length of time, ambient pollutants levels and meteorological variables (i.e., temperature and wind) can be viewed as a data signal with a low rate of change. Hence, the signal can be represented by a small number of coefficients that correspond to the low frequencies (Varotsos et al. 2005; Marr and Harley 2002; Chellali et al. 2010).

A formal mathematical framework for recovering missing signal's samples in the frequency domain, the discrete sampling theorem, was presented by Yaroslavsky et. al (2009). The discrete sampling theorem states the terms and conditions a band-limited frequency representation of a signal with missing samples must fulfill so the signal can be fully recovered, given it is narrow banded in any spectral domain. The theorem constitutes the new data imputation scheme presented here. Within its context two spectral signal representations are considered: The DCT (Rao et al. 1990; Yaroslavsky et al. 2009) and the *sparse coding* K-Cluster Single Variable Decomposition (K-SVD) (Aharon et al. 2006). The application of the suggested methods show that they are comparable to the state-of-the-art when imputing short missing sequences and do hold the upper hand when larger chunks of subsequent data are missing.

Prior art

Several mathematical models have been suggested for air-pollution data imputation (Junninen et al. 2004; Plaia and Bondi 2006; Schneider 2001). These methods include local methods, such as Nearest Neighbor (NN), mean, linear interpolation, spline and Expectation Maximization (EM). All these methods are thoroughly described in the literature and are recapitulated here for the sake of completeness.

Simple local methods for data imputation such as NN, Mean and Linear Interpolation were shown to

be effective, especially when signal's average levels are estimated (Junninen et al. 2004). NN fills missing samples using the value of its nearest known neighbor. Linear Interpolation infers the missing values based on a weighted average of the neighboring known samples based on the temporal distance. Mean interpolation replaces missing values with the average of the set of known samples within a temporal window around the missing sample.

A computational intense local imputation method is the Spline interpolation. Spline describes the signal between the available samples through a set of continuous functions. It can be thought of setting a rope through the available *k* known samples (nodes). The signal is broken into *k-1* segments, each represented by a third degree polynomial function:

$$f_i(x_i) = a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i \tag{1}$$

For the *k−1* segments, Spline will set *k−1* piecewise functions, composing a total of *4(k−1)* unknown parameters—$\{[a_i, b_i, c_i, d_i]_{i\in[1,k-1]}\}$. In order to compute these *4(k−1)* unknowns, it is imposed that the first and second derivatives are equal at each node:

$$\begin{aligned} \frac{df_i(x)}{dx} &= \frac{df_{i-1}(x)}{dx} \\ \frac{d^2 f_i(x_i)}{d^2 x} &= \frac{d^2 f_{i-1}(x)}{d^2 x} \end{aligned} \tag{2}$$

This results in $4 \cdot (k-1) + 2$ equations, i.e., 2 equations more than the number of unknowns. To deal with the extra 2 equations, the second derivatives at the end points are set to 0.

The advantage of this method is that while being relatively simple to calculate, the smooth function achieved with continuity in the first and second derivatives, better describes the changing nature of physical phenomenon over time. This method was shown to work well with short intervals of missing data points (Junninen et al. 2004).

Expectation Maximization (EM) algorithm is often used for filling in missing data using available data from the entire time series (Junninen et al. 2004; Dempster et al. 1977). The main assumption is that the missing data has linear relation with available data. To exploit that the data is split into a set of equal length vectors, $\{d_{(k)}\} \in D$, e.g., daily, weekly or monthly sequees. Then the missing samples are assigned with an initial guess of the missing values (i.e., zeros or, for each vector, the average of its available samples). The missing data points in one vector are computed by a linear combination of the vectors with non-missing corresponding data points. The covariance between the vectors is used as a way to determine how dominant a particular vector will be in the proposed linear combination.

Formally, let *A* be a matrix of $G \times P$ data points, where *G* is the number of time periods evaluated (e.g., a week or a day) and *P* is the number of records per the above time period (e.g., samples per week or day). Let $\{a\} \subseteq A$ be the set of *available* data and $\{m\} \subseteq A$ be the set of missing samples. For a given column *c*, let $\{a_a^c\}$ and $\{a_m^c\}$ be the sets of available and missing data in *c* respectively and μ_c is the mean value of $\{a_a^c\}$. Finally, $A \backslash c$ is matrix *A* excluding column *c*. Using the notation above, missing values of *A* are estimated through the following linear regression model:

$$\{a_m^c\} = \mu_c + (\{a\}_a^{A\backslash c} - \mu_{A\backslash c})\mathrm{B} + \mathrm{e} \tag{3}$$

e is the residual matrix assumed to have a zero mean and *B* is the matrix of the regression parameters to be calculated using the covariance matrix, Σ:

$$B = \Sigma_{aa}^{-1} \cdot \Sigma_{am} \tag{4}$$

where, Σ_{aa} denotes the sub-convergence matrix of columns of the available values with the columns of the available values. Σ_{am} denotes the sub-convergence matrix of columns of the available values with the columns of the missing values.

Applying Eq. 3 results in filling the missing data. Having the data in hand, a new mean and covariance matrix are calculated. Using the new *B* and Σ the process is repeated for all originally missing samples. The process is repeated until convergence.

The Regulated EM algorithm (Smith et al. 2003) presents a slight modification in the EM method—the sub-convergence matrix Σ_{aa}^{-1} is replaced by the following equation:

$$\Sigma_{aa}^{-1} \leftarrow \left(\Sigma_{aa} + h^2 D\right)^{-1} \tag{5}$$

where D is the diagonal of Σ_{aa} and h is a scalar regulation parameter. This modification ensures that the matrix is positive definite, invertible and converges faster, while artificially makes the variance of each vector more dominant with respect to its covariance with the rest of the columns.

The EM method may lead to better results especially if the missing segments are part of a recurring pattern. But if the pattern is not recurring in a set rhythm, this method may not work. Further, this is an iterative method which will lead to a greater computational costs compared to the local methods described above.

All the aforementioned methods for air quality missing data imputation have been well documented. However, all these methods are not sufficiently accurate when

longer segments of data are missing or in the event that the relationship between the data segments is not linear. Spectral methods consider the entire signal in their evaluation of missing data, which in return present better results when large chucks of data are missing.

A quasi-spectral method for air-quality data imputation, which uses information from the air monitoring stations array is Site-Dependent Effect Method (SDEM) (Plaia and Bondi 2006). The SDEM assumes that there are similarities in air quality sequences throughout the week, as well as between a given day of the week e.g. Sunday or Monday, and given hour of day. The missing data is then imputed by taking the mean value of all the non-missing measurements from the other stations at the missing time point, and modifying it based on the week, day and hour effect of the given station. This method is similar to spectral methods in the way it utilizes intuitive cycles i.e. hours, days and weeks, but it misses less obvious cycles, that may contain a lot of information such as local-specific phenomena of limited regions. In addition, all the weights in this method are arbitrarily set and it stands to reason that some cycles have a more profound effect than others and thus different weight values may produce more accurate results.

Results and discussion

For evaluating the imputation methods and their suitability for different scenarios and loss patterns, one must use a complete dataset and impose different data loss patterns and portions. The data was acquired from a standard AQM station, maintained by the Haifa District Municipal Association for Environmental Protection (HDMAE).[1] The station is situated on the roof of the HDMAE headquarters building, located at the center of the Haifa Bay industrial- commercial area. The station is ~12 m above ground level and reports every 30 min the average temperature, wind speed and direction, $PM_{2.5}$ and PM_{10} levels, O_3, NOx, NO_2, SO_2, and CO. In this study two long *complete* time series of SO_2 and NO_2 were used. SO_2 data was acquired using a pulsed fluorescence analyzer, over 167 days from Decmber 31st, 2006 until July 18th, 2006—a total of 8016 half hour average samples. NO_2 levels were recorded using a chemiluminescence analyzer from January 27th, 2008 to June 4th, 2008—a total of 138 days with 6240 samples. For the EM computation the data was divided into 24 h sequences, each consists of 48 measurements constructing 167 SO_2 sequences and 138 NO_2 sequences.

Previous data imputation studies (Junninen et al. 2004; Plaia and Bondi 2006; Schneider 2001) used shorter time series with gaps of missing samples in the original data. To cope with that, the gaps in the original time series were imputed as a preprocessing phase. After filling these gaps, a deliberate omission of data was executed and the omitted data was recovered. The error between the values of omitted and recovered samples was reported. Thus, the imputed data in the preprocessing phase was regarded as ground truth. All imputation methods that do not relay on physical models must base their missing data estimates on the signals' characteristics and behavior. Working with time series that has imputed data, alter these characteristics and thus hamper the results. In this study, working with a complete longitudinal datasets has mitigated these biases.

Epidemiologic and exposure studies on health effects of air pollution look at long term chronic and short term acute health implications (Lebowitz 1996). Chronic studies look at the long term effects of pollutants. These studies mainly focus on cumulative exposure and less on sudden increase in the concentration of any one pollutant. Acute effects, on the other hand, are transient and are a result of time variant exposure (Peng and Dominici 2008). The evaluation criteria for data imputation must account for the different nature of these two classes of studies. When chronic exposure is sought, the reconstruction mean error, typically through the mean squared error (MSE), should be the performance measure. For acute exposure assessment studies, one should evaluate the maximum error in signal's values and behavior. As an assessment how well the reconstructed signal presents the original data's behavior, the difference in the second statistical moments of the original and reconstructed signals are computed. For practicality sake, the runtimes are also reported.

In order to evaluate imputation performance, two data omission mechanisms were used. The first mechanism is omission of data at completely random locations, i.e., data Missing Completely At Random (MCAR) (Little and Rubin 2002). This type of behavior was found to characterize air quality data standard AQM stations (Junninen et al. 2004; Rubin 1976). The second samples removal mechanism removes a single chunk of data starting at a completely random time period.

For both data omission mechanisms a series of tests were carried out, where at each test, an increasing portion of data was omitted. The data portion that was omitted ranged from one sample to 99 % of the entire dataset. At each test the number of omitted samples was increased by 1. This set of tests extends previous studies, which evaluated the methods for small sets (i.e., small number of tests) all limited to small portions of missing samples [3, 10 and 25 %—(Schneider 2001; Plaia and Bondi 2006; Junninen et al. 2004) respectively]. The location of the omitted samples is chosen at random at each

[1] http://www.envihaifa.org.il/eng.

run. In order to mitigate a possible bias due to a specific location of the omitted data, for each number of omitted samples, five random tests were carried out, so at each test different random samples were chosen. The average performance indicator's value over the five runs is reported.

In all cases and scenarios tested here, the spline method is never the method of choice. Hence, for all examined cases, the spline method was shown to be inferior. In some performance criteria, such as MSE, maximum error and standard deviation differences for batch omission, the error for spline is between five to seven orders of magnitudes larger than the rest of the methods, making it hard to be put on the same graph. Therefore, the spline results are omitted from Fig. 1 through Fig. 4.

Figure 1 depicts the reconstruction MSE for increasing portions of missing data for the NO_2 (Fig. 1a) and SO_2 (Fig. 1b) time series. The error for each portion of missing data is computed over five runs, for the two omission mechanisms described above—randomly scattered (Fig. 1a) and batch (Fig. 1b). When data is omitted at random, the local methods perform best; namely the Spline

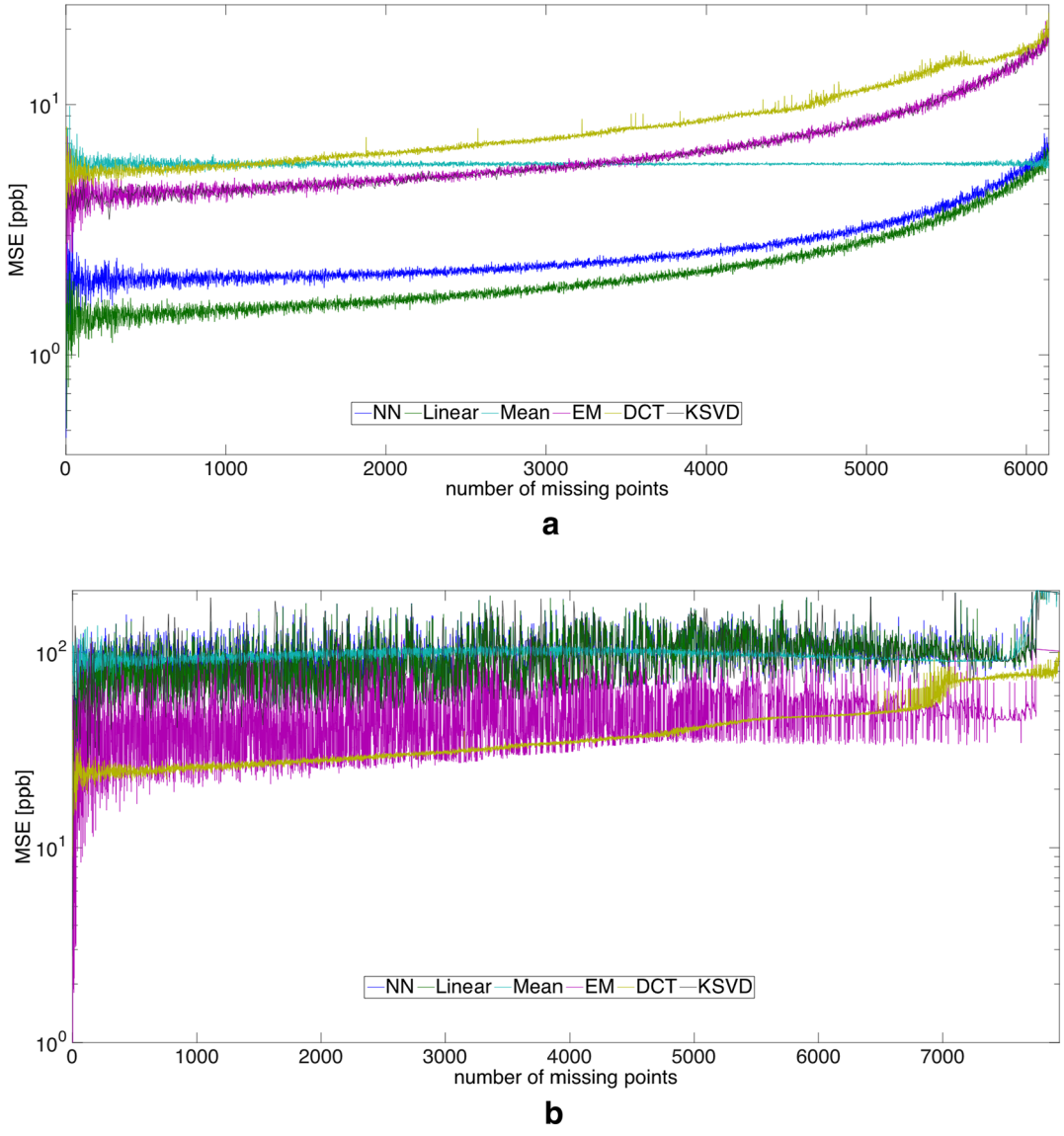

Fig. 1 Five runs averaged MSE as a function of number of omitted data samples reconstructing the signal with the following imputation methods NN (*blue*), Linear Interpolation (*green*), substituting missing data with the Mean (*cyan*), EM (*purple*), DCT (*yellow*) and K-SVD (*black*). The *color codes* are detailed in a panel. The missing data was omitted either in random (**a**) or in batches (**b**). **a** NO_2 (random locations). **b** SO_2 (continuous batch)

in a case of a very low corruption rate and linear interpolation as the rate of corruption increases. When the data is missing in segments, the EM, Discrete Cosine Transform (DCT) and K-SVD algorithms preform the best. In studies centered on chronic effects of air pollution, if the data is missing in short intervals, the linear interpolation is the best method for filling in the missing data. But in the event of a long missing sequence, the K-SVD should be the method of choice.

Figure 2, presents the maximum difference between the original signal and the reconstructed one. It can be seen that for randomly omitted data (Fig. 2a) the K-SVD and the DCT methods have the smallest deviation with a slight advantage for the K-SVD over the DCT. Consequentially, studies conducted in order to investigate the acute effect of air pollution should fill in missing data with the K-SVD method. When the data is missing in segments (Fig. 2b), for all methods the error is in the same magnitude of the signal. Therefore such studies should not use time series with long temporal windows missing.

Figure 3 presents the difference in the standard deviation between the reconstructed and the original signal. For both random and batch omitted data the DCT, K-SVD and EM are at par, outperforming the local

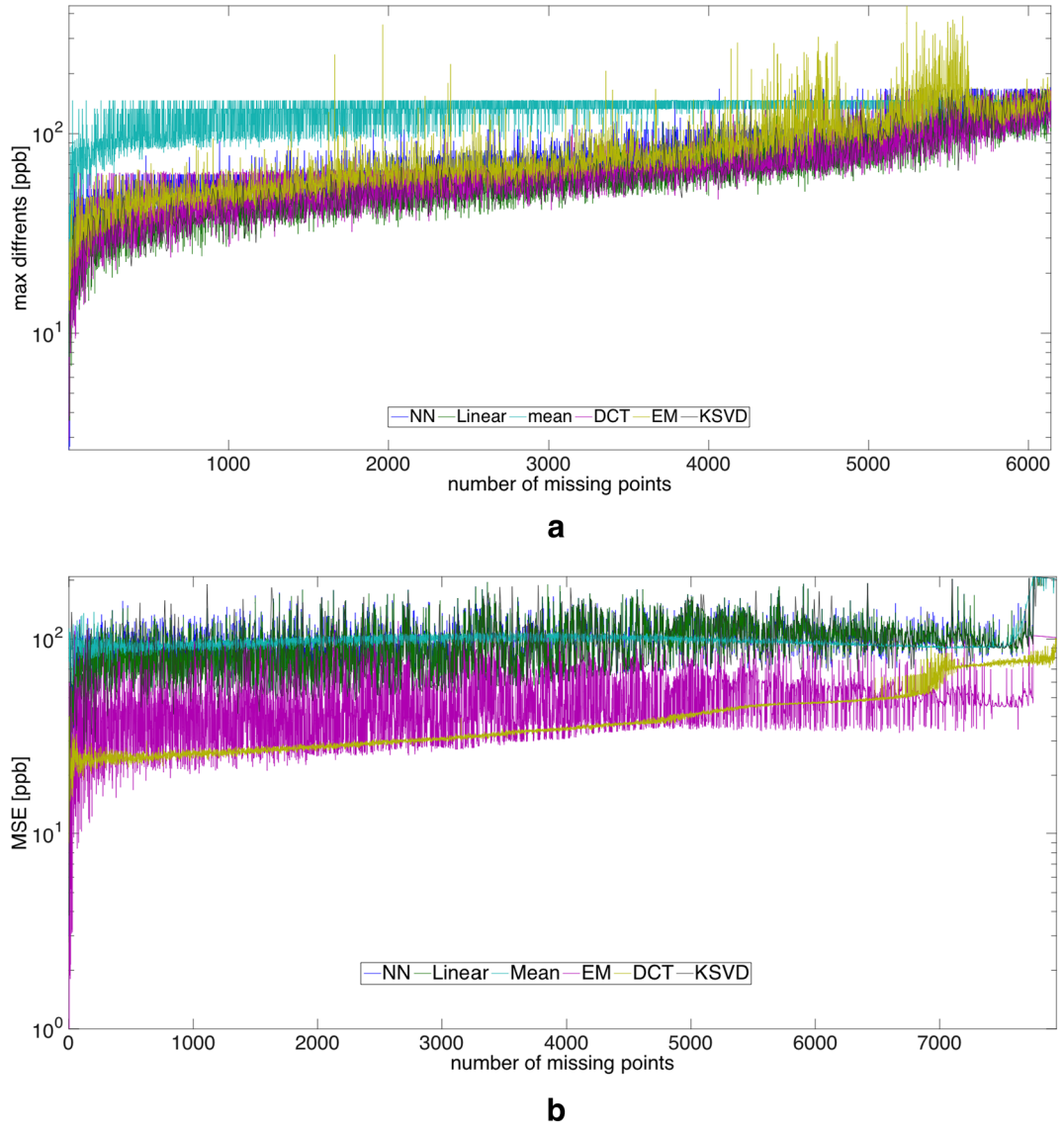

Fig. 2 Five runs averaged max difference as a function of number of omitted data samples reconstructing the signal with the following imputation methods NN (*blue*), Linear Interpolation (*green*), substituting missing data with the Mean (*cyan*), EM (*purple*), DCT (*yellow*) and K-SVD (*black*). The *color codes* are detailed in a panel. The missing data was omitted either in random (**a**) or in batches (**b**). **a** NO_2 (random locations). **b** SO_2 (continuous batch)

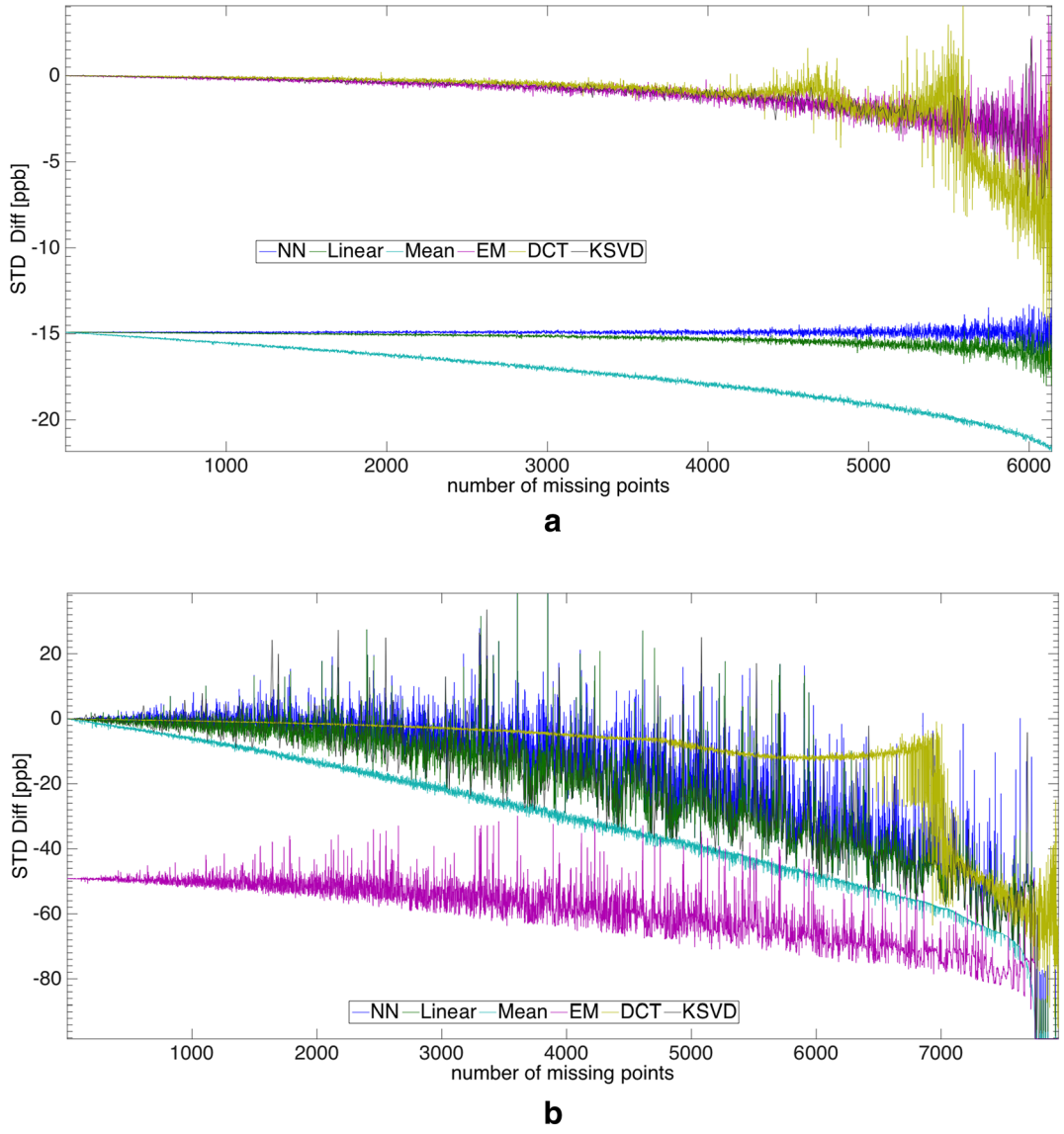

Fig. 3 Five runs averaged max difference as a function of number of omitted data samples reconstructing the signal with the following imputation methods NN (*blue*), Linear Interpolation (*green*), substituting missing data with the Mean (*cyan*), EM (*purple*), DCT (*yellow*) and K-SVD (*black*). The *color codes* are detailed in a panel. The missing data was omitted either in random (**a**) or in batches (**b**). **a** NO_2 (random locations). **b** SO_2 (continuous batch)

methods. The standard deviation difference between the original signal and the reconstructed one is smallest using the DCT method if the data is missing at random. The EM method is best if a long sequence is missing. Overall the DCT method reconstructs the missing data in a way that is more similar to the original signal in terms of STD compared to all the other methods.

The computation times of the various methods are presented in Fig. 4. As expected, the spectral methods i.e. K-SVD for signal recovery from sparse dimension, DCT and EM are much more costly in terms of computation times. In both omission mechanisms, the computation time decreases as the portion of missing data increases. Even though the spectral methods are more costly in terms of computation, one should note that the data being processed describes months' worth of data, when the computation times are in the order of minutes. Therefore the longer computation times should not prevent one form using these spectral methods for data imputation.

Conclusions

In this paper two spectral methods for data imputation, originating from the discrete sampling theorem, are

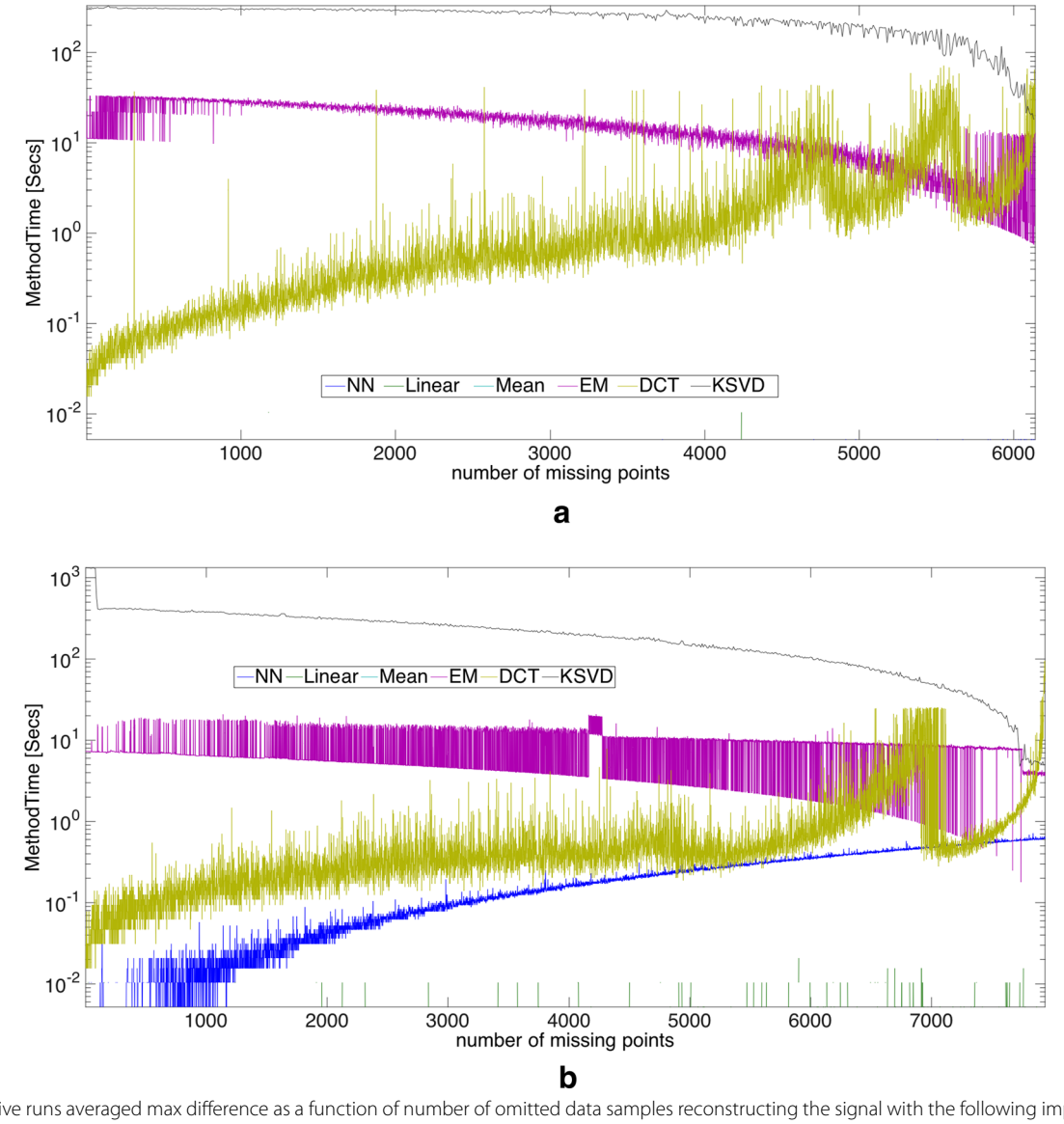

Fig. 4 Five runs averaged max difference as a function of number of omitted data samples reconstructing the signal with the following imputation methods NN (*blue*), Linear Interpolation (*green*), substituting missing data with the Mean (*cyan*), EM (*purple*), DCT (*yellow*) and K-SVD (*black*). The *color codes* are detailed in a panel. The missing data was omitted either in random (**a**) or in batches (**b**). **a** NO_2 (random locations). **b** SO_2 (continuous batch)

introduced to air quality time series with missing data. The methods are thoroughly evaluated with respect to the common practice and the state of the art missing data recovery methods. The evaluation of the methods here is much more comprehensive than previous studies, as it uses much longer air quality time series with no missing data, under significantly larger number missing data scenarios.

The evaluation results are summarized in Fig. 5 and Table 1 (for randomly omitted data) and Fig. 6 and Table 2 (for chunks removal). Figures 5a and 6a depict the average MSE for NO_2 (randomly omitted data) and SO_2 (chunks removal) time series. The average is computed over three sets of test runs. The first set, low signal degradation due to missing data, are all the tests with 1 sample missing up to 33 % of missing samples. This set is dubbed low and is marked in blue (Low-blue). The second set is all runs with 33 % samples omitted up to 66 % (Mid-green). The last set of runs present 66–99 % samples missing, i.e., high degradation (High-Yellow). For data missing at random, Fig. 5a and Table 1 (MSE), the simple imputation methods provide the best MSE results. However, the spectral methods

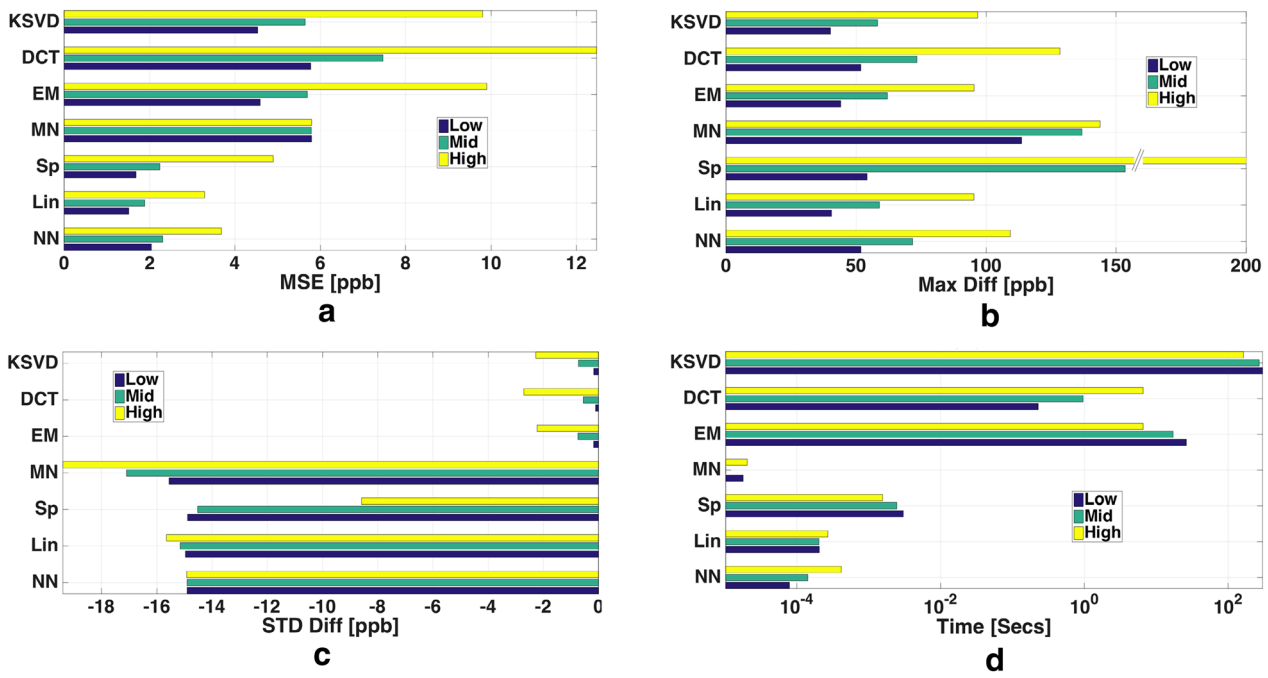

Fig. 5 Results summary—NO_2 data missing at random for low signal degradation, i.e. small portion of missing data, mid range signal degradation and high signal degradation. The *color codes* are detailed in panels within each image. **a** MSE, **b** Max difference, **c** STD difference, and **d** execution time

Table 1 Results summary—NO_2 data missing at random

	KSVD	DCT	EM	MN	Sp	Lin	NN
MSE (ppb)							
Low	4.54	5.77	4.60	5.79	1.68	*1.51*	2.04
Mid	5.64	7.47	5.69	5.79	2.24	*1.88*	2.31
High	9.80	12.48	9.89	5.79	4.9	*3.29*	3.69
Diff (ppb)							
Low	*40.12*	51.81	44.06	113.6	54.23	40.5	109.3
Mid	*58.26*	73.40	62.05	136.9	153.5	58.97	71.77
High	96.83	128.4	95.47	143.9	622.6	*95.39*	51.84
STD (ppb)							
Low	−0.17	*−0.11*	−0.18	−15.56	−14.89	−14.97	−14.91
Mid	−0.72	*−0.55*	−0.75	−17.09	−14.52	−15.16	−14.91
High	−2.28	−2.71	*−2.23*	−19.40	−8.59	−15.66	−14.92
Time (sec)							
Low	303	0.23	26.22	1.78×10^{-5}	3×10^{-3}	2×10^{-4}	7.89×10^{-5}
Mid	272	0.96	17.07	1.01×10^{-5}	2.5×10^{-3}	2×10^{-4}	1.4×10^{-4}
High	164	6.6	6.59	2.03×10^{-5}	1.5×10^{-3}	2.7×10^{-4}	4.1×10^{-4}

do not fall far behind. For data missing in chunks, Fig. 6a and Table 2 (MSE), the spectral methods, DCT and K-SVD, present the best performance MSE-wise.

Figure 5b and Table 1 (Diff) and Fig. 6b with Table 2 (Diff) present the max difference. For randomly missing data, the spectral methods, i.e., K-SVD and DCT,

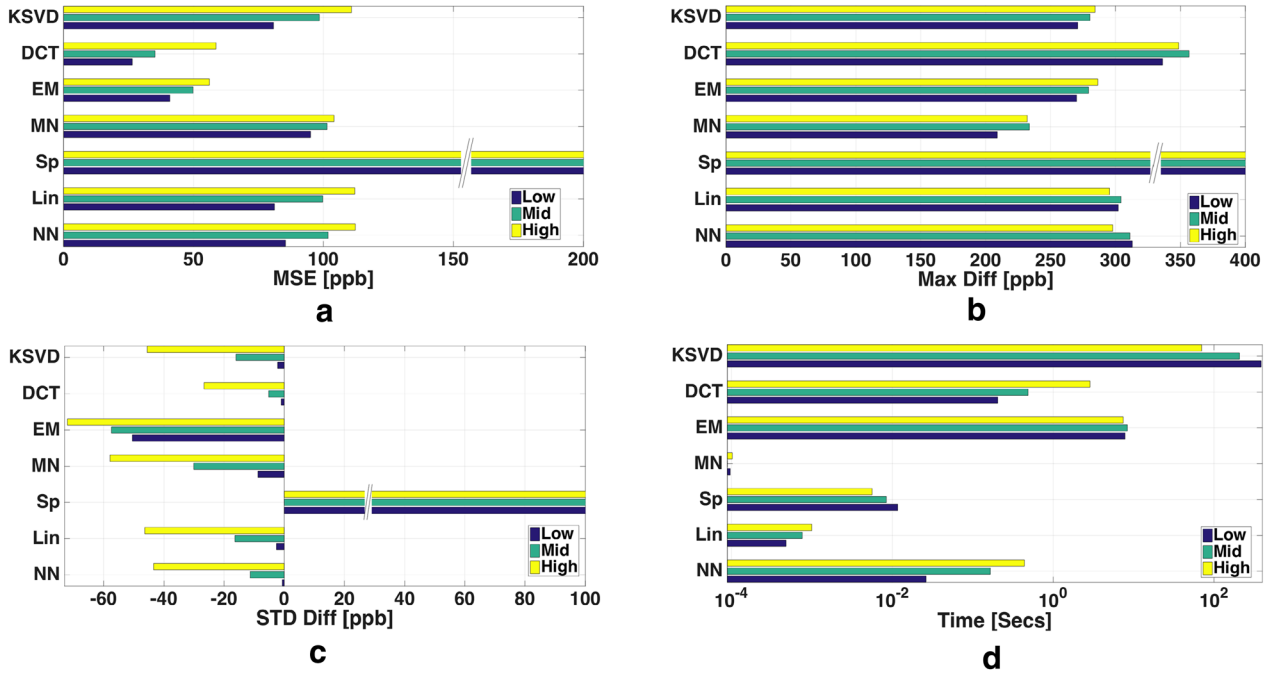

Fig. 6 Results summary—SO_2 data missing in batches for low signal degradation, i.e. small portion of missing data, mid range signal degradation and high signal degradation. The *color codes* are detailed in panels within each image. **a** MSE, **b** Max difference, **c** STD difference, and **d** execution time

Table 2 Results summary—SO_2 data missing in batches

	KSVD	DCT	EM	MN	Sp	Lin	NN
MSE (ppb)							
Low	80.76	*26.48*	40.82	95.06	3.89×10^{9}	81.17	85.39
Mid	98.45	*35.11*	49.74	101.4	7.9×10^{10}	99.77	101.8
High	110.7	58.59	*56.07*	104.1	3.29×10^{11}	112	112.2
Diff (ppb)							
Low	*270.9*	336.1	270	208.8	3.28×10^{10}	302.2	313
Mid	*280.3*	356.6	279.3	233.6	5.18×10^{11}	304.3	311.2
High	*284.3*	348.5	286.4	231.9	1.57×10^{12}	295.5	297.9
STD (ppb)							
Low	−2.19	−0.97	−50.42	−8.75	2.91×10^{9}	−2.59	*−0.64*
Mid	−16.06	*−5.17*	−57.39	−30.12	7.73×10^{10}	−16.42	−11.31
High	−45.48	*−26.69*	−72.05	−57.86	3.62×10^{11}	−46.28	−43.49
Time (sec)							
Low	387.1	0.2	7.86	9.64×10^{-5}	11.7×10^{-3}	4.78×10^{-4}	26.3×10^{-3}
Mid	207.7	0.49	8.4	8.89×10^{-5}	8.41×10^{-3}	7.61×10^{-4}	0.16
High	70.06	2.88	7.46	1.02×10^{-4}	5.59×10^{-3}	1×10^{-3}	0.44

have the upper hand. For missing chunks, the error produced by all methods is so large, that none of them are recommended for this case. Figure 5c, Table 1 (STD), Fig. 6c and Table 2 (STD) detail the difference in the second moment of the original and reconstructed signals. For both random and chunks missing data patterns, the K-SVD and DCT present the best results.

While the spectral methods do present higher computational times (Fig. 5d, Table 1 (Time), Fig. 6d and Table 2 (Time)), these times are still feasible. Moreover,

the spectral methods have the upper hand, MSE-wise, when the data is missing in chunks and when evaluating acute exposure, i.e., max-difference and signal behavior through its standard deviation. In the cases where the simple methods prevail, the spectral methods do not fall far behind. Therefore, we conclude that the spectral methods in general, K-SVD and DCT in particular, do present viable tool for data imputation and should be used as the tool of choice in general as it presents the overall best performance.

Both the KSVD (Aharon et al. 2006) and DCT (Yaroslavsky et al. 2009) methods assume band-limited signal, i.e., only a small portion of signal's spectral representation coefficients are non-zero. In most implementations the portion of non-zero coefficients is predetermined. Choosing larger portions of non-zero coefficient would result in longer execution times and may jeopardize the convergence of both the DCT and KSVD imputation methods. Smaller portions of non-zero coefficients decrease computation times and mitigate the risk of not converging, but may increase the output error. Therefore, for using these methods, one should carefully assess what is the correct portion of non-zero coefficients in the signals' spectra.

Methods

The discrete sampling theorem

Next we outline the discrete sampling theorem, which constitute the data imputation scheme presented here.

Let $a(t)$ be a continuous signal and $\{A^N\}$ the set of N measurements of the signal. These N samples constitute a uniform sampling grid and are acquired is such way that if all these N-samples are known, following the Nyquist–Shannon sampling theorem, (Unser 2000) they are sufficient for representing the continuous signal. Let $\{A^K\}$ be the set of K (out of N) available data points taken at irregular positions of the signal regular sampling grid. Due to data loss $K < N$. Note that the missing samples are $\{A^N\}$ excluding $\{A^K\}$, $\{A^M\} = \{A^N\} \backslash \{A^K\}$. The goal then, is to generate out of this incomplete set of K samples, a complete set of N signal samples that secures the most accurate, in a certain metrics—typically L_2, approximation. The discrete sampling theorem (Yaroslavsky et al. 2009) states the terms and conditions a signal must fulfills in the transform domain so its $\{A^N\}$ samples can be recovered from the available $\{A^K\}$ samples:

Theorem 1 *The Discrete Sampling Theorem* (Yaroslavsky et al. 2009)*—Any discrete signal of N samples defined by its $K \leq N$ sparse and not necessarily regularly arranged samples, and is known to have only $K \leq N$ non-zero transform coefficients for certain transform Φ_N (i.e., Φ_N- transform "band-limited" signal) can be fully recovered from exactly K of its samples provided positions of the samples secure the existence an inverse transform matrix, $\{\Phi_{K\ of\ N}\}^{-1}$, where $\Phi_{K\ of\ N}$ consists of K rows of the transform matrix Φ_N that correspond to the K samples positions. If the signal has more than K non-zero transform coefficients, the recovery process guarantees minimum reconstruction error.*

Theorem 1 implies that selecting a transform that features the best energy compaction with the smallest number of transform coefficients secures the best approximation of $\{A^N\}$ for a given subset $\{A^K\}$ of its samples. The recovery process is based on the following simple iterative procedure (Yaroslavsky et al. 2009):

Algorithm 1. Discrete sampling theorem imputation algorithm.

1) *Initialization* – Set the signal's initial estimate, $\{\widehat{A_0^N}\}$, so all known samples receives their measured value - $\widehat{A_0^N}(j) = A^K(j)\big|_{\forall j \in \{A^K\}}$ and all missing samples, $\widehat{A_0^N}(j)\big|_{\forall j \notin A_K}$, are set to an arbitrary value. This value can be zeros, the average value of the available samples, i.e., $mean\{A^K\}$, or the result of a bilinear interpolation of neighboring available samples.

For each iteration i do:

2) *Signal transform* – transform $\{\widehat{A_t^N}\}$ to a domain in which the signal's representation is known to be band limited. $\widehat{\alpha_t^N} = \Phi\widehat{A_t^N}$.
3) *Zeroing* – In the transform domain zero any coefficient outside the band limited area, R - $\widehat{\alpha_t^N}(j)\big|_{j \notin R} = 0$
4) *Inverse transform* - invers back to the signal's original domain: $\widehat{A_{t+1}^N} = \Phi^{-1}\widehat{\alpha_t^N}$
5) *Replacement* – reinstate the $\{A^K\}$ known samples back to the obtained signal- $\widehat{A_{t+1}^N}(j)\big|_{\forall j \in A_K} = A^K(j)$
6) *Stop condition* – Stop if $(i \geq M)$, number of iterations has reached its limit, M, or if $\sum_{j=1}^{N}\left(\widehat{A_{t+1}^N} - \widehat{A_t^N}\right)^2 < \varepsilon$, i.e., the squared difference between two iterations is below a threshold.
7) $\widehat{A_t^N} \leftarrow \widehat{A_{t+1}^N}$, go to 2.

The mean square error of this algorithm is calculated by:

$$MSE = A_i^N - \widehat{A_i^N}_2^2 = \Sigma_{j \notin R}\left|\widehat{\alpha_i^N}(j)\right|^2 \tag{6}$$

The transform of choice here is the discrete cosine transform (DCT), given by:

$$\alpha^N(k) = \Sigma_{j=0}^{N-1} A^N(j) \cos\left[\frac{\pi}{N}(j+0.5)k\right] k = 0, \ldots N-1 \tag{7}$$

DCT is a widely used function in the field of image processing and data compression because of its tendency to concentrate most of the energy from a signal in a narrow band (Wang et al. 2000).

Sparse coding data imputation

Sparse coding is an emerging spectral approach to data analysis (Elad 2010). While classical spectral methods utilize predetermined set of basis-functions (e.g., DFT,

DCT) for representing the signal, sparse sensing methods, compute the set of basis-functions, which results in the sparsest representation of the signal, i.e., most coefficients of the signal's representation are zeros. The set of basis-functions is referred to as *dictionary*, where each element in the dictionary is an *atom*. The process of computing the basis-functions is called *dictionary learning* (Kreutz-Delgado et al. 2003). K-SVD (Aharon et al. 2006) is a dictionary learning algorithm for creating a set of basis functions for sparse representations. K-SVD is a generalization of the k-means clustering method, and it works by iteratively alternating between sparse coding of the input data (based on the current dictionary), and updating the atoms in the dictionary to better fit the data.

The discrete sampling theorem suggests that the smaller the number of non-zero coefficients in the domain transform, the better the reconstruction is. Yet, the transform of choice, DCT, is known to have good energy compaction in general but it is not costumed nor guaranteed for the specific data in hand and may or may not yield a sparse representation (Elad 2010). To cope with this problem, building a custom transform or *dictionary* for sparse coding is suggested (Elad 2010; Aharon et al. 2006). The dictionary is an overcomplete matrix, $D \in \mathbb{R}^{N\times P}$, that consists of P atoms (with a length of N) and is designed so a signal A^N can be then represented by a sparse linear combination of these atoms. For finding D the K-cluster Single Value Decomposition (K-SVD) method (Aharon et al. 2006) is employed (for details see Section S1 in the Additional file 1), where the matrix A^N is utilized as the training set for the process.

Having the dictionary in hand, a sparse representation for A^N is sought. Given D, the dictionary, or basis functions, the sparse representation, x^N, aims at minimizing the error between the original signal A^N and the estimate of the signal using the basis function set, D:

$$\min_x \left\{ \left\| A^N - Dx^N \right\|_F^2 \right\} \quad S.T \ \left\| x^N \right\|_0 < K \tag{8}$$

where $\|\cdot\|_0$ is the ℓ^0-norm, counting the non-zero elements in a vector and $\|B\|_F$ is the Frobenius Norm: $\|B\|_F = \sqrt{\sum_{ij} \left(B_{ij}\right)^2}$.

Equation 8 was shown to be NP-hard, i.e., no efficient methodology is known for finding x^N. The approximation of the sparse representation can be obtained through the matching pursuit (MP) algorithm (Mallat and Zhang 1993), or through the K-SVD algorithm, which produces x^N as a byproduct. Note, that x^N is guaranteed to be both sparse (i.e., band limited) and zero in all coefficients outside the band limited area, R. Thus, unlike the DCT solution, there is no need to assume beforehand which coefficients are outside R and no need to zero them. Having both K-SVD and MP algorithms in hand, Algorithm 1 becomes:

Algorithm 2. K-SVD imputation algorithm

1) *Initialization* – Set the signal's initial estimate, $\widehat{A_0^N}$, so all known samples receives their measured value - $\widehat{A_0^N}(j) = A^K(j)\big|_{\forall j\in\{A^K\}}$ and all missing samples, $\widehat{A_0^N}(j)\big|_{\forall j\notin A_K}$, are set to $mean\left\{A^K(j)\big|_{\forall j\in\{A^K\}}\right\}$ (Yaroslavsky, et al., 2000).

For each iteration i do:

2) *Dictionary Learning* – Find the dictionary D^i that best sparsely represents $\left\{\widehat{A_i^N}\right\}$ (K-SVD).

3) *Sparse Representation* – Find a sparse representation, x_i^N for $\widehat{A_i^N}$ by optimizing Equation 8 (typically part of the K-SVD algorithm).

4) *Inverse Transform* – Reconstruct, $\widehat{A_{i+1}^N} = D^i x_i^N$.

5) *Replacement* – reinstate the $\{A^K\}$ known samples back to the obtained signal- $\widehat{A_{i+1}^N}(j)\big|_{\forall j\in A_K} = A^K(j)$

6) *Stop condition* – Stop if $(i \geq M)$, number of iterations has reached its limit, M, or if $\sum_{j=1}^{N}\left(\widehat{A_{i+1}^N} - \widehat{A_i^N}\right)^2 < \varepsilon$, i.e., the squared difference between two iterations is below a threshold.

The data omission, imputation and evaluation codes were written as Matlab scripts (Matlab R2012b) and are available for academic purposes from http://fishbain.net.technion.ac.il.

Additional file

Additional file 1. K-SVD Algorithm.

Authors' contributions

SM developed and implemented the methods, conducted the numerical experimentation, and led the drafting of the manuscript. UL designed the study, analyzed the results and took part in drafting the manuscript. BF advised and directed the presented research as well as contributed to drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was partially supported by the 7th European Framework Program (FP7) ENV.2012.6.5-1, Grant Agreement No. 308524 (CITI-SENSE), the Technion Center of Excellence in Exposure Science and Environ- mental Health (TCEEH), the New-York Metropolitan Research Fund, and the Environmental Health Foundation (EHF). The contribution to this paper of Michael Elad's course on sparse and redundant representations and Elad's advises are acknowledged.

Competing interests

The authors declare that they have no competing interests.

Received: 2 October 2015 Accepted: 23 November 2015
Published online: 22 December 2015

References

Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. Signal Process IEEE Trans 54(11):4311–4322

Bracewell RN (1965) The Fourier transform and its applications. Mcgraw-Hill, New-York

Chellali F, Khellaf A, Belouchrani A (2010) Wavelet spectral analysis of the temperature and wind speed data at Adrar, Algeria. Renewable Energy 35(6):1214–1219
Cullis C, Hirschler M (1980) Atmospheric sulphur: natural and man-made sources. Atmos Environ 14(11):1263–1278
Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc 39(1):1–38
Elad M (2010) Sparse and redundant representations: from theory to applications in signal and image processing. Springer Science, Haifa
Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
Haagen-Smit A, Bradley C, Fox M (1953) Ozone formation in photochemical oxidation of organic substances. Ind Eng Chem 45(9):2086–2089
Hopke P (1991) Receptor modeling for air quality management, vol 7. Elsevier, Amsterdam, The Netherlands
Junninen H et al (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18):2895–2907
Kampa M, Castanas E (2008) Human health effects of air pollution. Environ Pollut 151(2):362–367
Kreutz-Delgado K et al (2003) Dictionary learning algorithms for sparse representation. Neural Comput 15(2):349–396
Künzli N et al (2000) Public-health impact of outdoor and traffic-related air pollution: a European assessment. Lancet 356(9232):795–801
Laumbach RJ, Kipen HM (2012) Respiratory health effects of air pollution: update on biomass smoke and traffic pollution. J Allergy Clin Immunol 129(1):3–12
Lebowitz M (1996) Epidemiological studies of the respiratory effects of air pollution. Euro Respir J 9(5):1029–1054
Little R, Rubin D (2002) Bayes and multiple imputation. In: Statistical analysis with missing data, 2nd edn. Wiley, Hoboken, New Jersey, pp 200–222
Lee K, Xue J, Geyh A, Ozkaynak H, Leaderer B, Weschler C, Spengler J (2002) Nitrous acid,nitrogen dioxide, and ozone concentrations in residential environments. Environ Health Perspect 110(2):145
Lerner U, Yacobi T, Levy I, Moltchanov S, Cole-Hunter T, Fishbain B (2015) The effect of egomotionon environmental monitoring. Sci Total Environ 533:8-16
Levy I, Mihele C, Lu G, Narayan J, Brook JR (2014) Evaluating multipollutant exposure and urban air quality: pollutant interrelationships, neighborhood variability, and nitrogen dioxide as a proxy pollutant. Environ Health Perspect 122(1):65-72
Marr L, Harley R (2002) Spectral analysis of weekday–weekend differences in ambient ozone, nitrogen oxide, and non-methane hydrocarbon time series in California. Atmos Environ 36(14):2327–2335
Mallat S, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. Signal Process IEEE Transa 41(12):3397–3415
Nazaroff W, Alvarez-Cohen L (2001) Environmental Engineering Science. John Wiley, New-York
Peng RD, Dominici F (2008) Statistical methods for environmental epidemiology with r: a case study in air pollution and health. Springer, Berlin
Plaia A, Bondi A (2006) Single imputation method of missing values in environmental pollution data sets. Atmos Environ 40(38):7316–7330
Rao KR, Yip P, Ramamohan Rao K (1990) Discrete cosine transform: algorithms, advantages, applications. Academic Press, Boston
Robinson E, Robbins RC (1970) Gaseous nitrogen compound pollutants from urban and natural sources. J Air Pollut Control Assoc 20(5):303–306
Rubin D (1976) Inference and missing data. Biometrika 65:581–592
Rubin D (2004) Multiple imputation for nonresponse in surveys, vol 81. Wiley, Hoboken
Schneider T (2001) Analysis of incomplete climate data : estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. Am Meteorol Soc 14:853–871
Smith R, Kolenikov S, Cox L (2003) Spatiotemporal modeling of PM2.5 data with missing values.J Geophys Res 108(D24):11-1–11-10
Solomatine D, See LM, Abrahart RJ (2008) Data-driven modelling: concepts, approaches and experiences. In: Practical hydroinformatics. Springer, Berlin, Heidelberg, pp 17–30
Su Y-S, Gelman A, Hill J, Yajima M (2011) Multiple Imputation with Diagnostics (mi) in R: opening windows into the black box. J Stat Softw 45(2):1–31
Unser M (2000) Sampling 50 years after Shannon. Proc IEEE 88(4):569–587
Varotsos C, Ondov J, Efstathiou M (2005) Scaling properties of air pollution in Athens, Greece and Baltimore, Maryland. Atmos Environ 39(22):4041-4047
Wang Y, Vilermo M, Yaroslavsky L (2000) Energy compaction property of the MDCT in comparison with other transforms. Los-Angeles, CA
Moltchanov S, Levy I, Etzion Y, Lerner U, Broday DM, Fishbain B (2015) On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Sci Total Environ 502:537–547
Yaroslavsky LP, Shabat G, Salomon BG, Ideses IA, Fishbain B (2009) Nonuniform sampling, image recovery from sparse data and the discrete sampling theorem. JOSA A 26(3):566-575