Ann. Geophys., 34, 437–449, 2016
www.ann-geophys.net/34/437/2016/
doi:10.5194/angeo-34-437-2016

Annales Geophysicae Open Access EGU

Effect of data gaps: comparison of different spectral analysis methods

Costel Munteanu[1,2,3], **Catalin Negrea**[1,4,5], **Marius Echim**[1,6], **and Kalevi Mursula**[2]

[1]Institute of Space Science, Magurele, Romania
[2]Astronomy and Space Physics Research Unit, University of Oulu, Oulu, Finland
[3]Department of Physics, University of Bucharest, Magurele, Romania
[4]Cooperative Institute for Research in Environmental Sciences, Univ. of Colorado, Boulder, Colorado, USA
[5]Space Weather Prediction Center, NOAA, Boulder, Colorado, USA
[6]Belgian Institute of Space Aeronomy, Brussels, Belgium

Correspondence to: Costel Munteanu (costelm@spacescience.ro)

Received: 31 August 2015 – Revised: 14 March 2016 – Accepted: 28 March 2016 – Published: 19 April 2016

Abstract. In this paper we investigate quantitatively the effect of data gaps for four methods of estimating the amplitude spectrum of a time series: fast Fourier transform (FFT), discrete Fourier transform (DFT), *Z* transform (ZTR) and the Lomb–Scargle algorithm (LST). We devise two tests: the single-large-gap test, which can probe the effect of a single data gap of varying size and the multiple-small-gaps test, used to study the effect of numerous small gaps of variable size distributed within the time series. The tests are applied on two data sets: a synthetic data set composed of a superposition of four sinusoidal modes, and one component of the magnetic field measured by the *Venus Express* (VEX) spacecraft in orbit around the planet Venus. For single data gaps, FFT and DFT give an amplitude monotonically decreasing with gap size. However, the shape of their amplitude spectrum remains unmodified even for a large data gap. On the other hand, ZTR and LST preserve the absolute level of amplitude but lead to greatly increased spectral noise for increasing gap size. For multiple small data gaps, DFT, ZTR and LST can, unlike FFT, find the correct amplitude of sinusoidal modes even for large data gap percentage. However, for in-situ data collected in a turbulent plasma environment, these three methods overestimate the high frequency part of the amplitude spectrum above a threshold depending on the maximum gap size, while FFT slightly underestimates it.

Keywords. Space plasma physics (instruments and techniques)

1 Introduction

Spectral analysis is a widely used tool in data analysis and processing in most fields of science. The technique became very popular with the introduction of the fast Fourier transform algorithm which allowed for an extremely rapid computation of the Fourier Transform. In the absence of modern supercomputers, this was not just useful, but also the only realistic solution for such calculations. This limitation is no longer relevant except for data sets of extremely large size. Still, FFT has remained the most popular tool for spectral analysis, because it is both easy to use and very fast. This makes it an extremely powerful tool and generally it is the first choice from the "toolbox" of spectral analysis methods. It is readily available in all programming languages of notice and it is accurate under perfect conditions. However, real data is rarely perfect. We address here the very common problem of data gaps.

Due to the wide usage of FFT, the literature on this subject is spread over a wide array of scientific disciplines: astronomy (Scargle, 1982, 1989), seismology (Baisch and Bokelmann, 1999), paleoclimatology (Heslop and Dekkers, 2002; Rehfeld et al., 2011), turbulence (Britz and Antonia, 1996; van Maanen and Oldenziel, 1998; Broersen et al., 2000; Harteveld et al., 2005), biomedical sciences (Schimmel, 2001). However, despite the vast amount of literature, studies on the effects of data gaps are few in number (Stahn and Gizon, 2008; Musial et al., 2011).

The problem of data gaps, i.e., occasionally missing observations from an otherwise continuous set of measurements, can be viewed as a special case of the more general problem of non-uniformly sampled data, i.e., measurements made at random time intervals throughout the data set. Although our study is mostly concerned with data gaps, all the methods used in our paper can also be straight-forwardly applied to problems related to non-uniform sampling. A recent review of the methods for spectral analysis of non-uniformly sampled data is presented in Babu and Stoica (2010).

Data gaps can occur in a variety of ways, from temporary malfunctions of the measuring instrument, to the inherent non-continuous nature of the physical phenomenon being studied. Pope et al. (2011), for example, describe the various challenges and difficulties in measuring the magnetic field environment around the planet Venus using the magnetically unclean spacecraft *Venus Express* (VEX). They determined that distinctive artificial patterns are introduced in the magnetic field data by the magnetic disturbances produced by different spacecraft and scientific instruments, and they tried to correct the data using advanced cleaning algorithms. The cleaning procedures, although successful in generating scientific data with a good data coverage, still left many gaps and other artifacts (various interpolation procedures were also used in the cleaning process) in the data. Also, the high-resolution measurements of the magnetic field are more severely affected, making them impossible to clean.

Many papers dealing with spectral analysis of real data often avoid the problems related to data gaps by subdividing the time series into smaller samples with continuous data coverage. One such paper is the study by Teodorescu et al. (2015), where the spectral properties of fast and slow solar wind are analyzed using tools based on FFT. They analyzed magnetic field data provided by the VEX satellite, which contains a large number of data gaps. They avoid the data gap problem by imposing a series of constraints regarding the maximum length and distribution of data gaps. This resulted in the loss of more than 80 % of their initial data set. Constraints and quality checks of this type are not uncommon in time series analysis.

In the case of multiple data gaps distributed throughout the data set and/or irregular time sampling, where avoiding data gaps is not an option, one can either use data reconstruction (interpolation) and FFT, or more advanced spectral analysis methods designed to handle non-uniform sampling. Reconstruction can be achieved with a simple linear interpolation across the gaps (which is one of the methods used in our study), or with more advanced reconstruction techniques, like the method based on singular-spectrum analysis (SSA) (Ghil et al., 2002; Kondrashov and Ghil, 2006).The SSA method was used by Kondrashov et al. (2010) and Kondrashov et al. (2014) to fill in gaps in solar wind data. Kondrashov et al. (2014) used the filled-in, continuous solar wind data as input into the TS05 empirical magnetic field model (Tsyganenko and Sitnov, 2005), and checked the reconstruction accuracy by comparing these results with GOES measurements at geostationary orbit. They found that the SSA gap-filling method improves the accuracy of the empirical magnetic field model, especially for large gaps.

We intend to quantify the distortions introduced by data gaps using four popular methods of estimating the frequency spectrum: fast Fourier transform, discrete Fourier transform, Z transform and Lomb–Scargle transform. The results are compared qualitatively and quantitatively using synthetic and real data sets.

We use a synthetic time series consisting of several periodic signals and added noise. Different, frequently encountered data gap configurations are applied to it and the amplitude spectra are calculated and compared to the known spectrum of the unaltered signal. Two gap configurations are used to highlight the effects: (a) a single large gap, where the original data set is altered by removing an increasing number of points from the central part, and (b) multiple small gaps, where we remove short series of consequent points whose length and precise location are randomly selected. The same methodology is also used to test the effect of data gaps on the amplitude spectra of magnetic field measurements made by the *Venus Express* spacecraft in orbit around the planet Venus. For these two tests, we determine practical thresholds where the use of the methods is no longer feasible.

The paper is structured as follows: a description of the general methodology and methods is presented in Sect. 2; Sect. 3 shows the results for a synthetic data set comprising of four sinusoidal signals, and Sect. 4 applies the same methodology on a real data set of *Venus Express* magnetic field measurements. Section 5 gives a brief summary and presents the main conclusions of our study.

2 Analysis methods

The fast Fourier transform is extremely fast to calculate, but requires strictly uniform sampling. It is by far the most popular method for computing the frequency spectrum. It is sometimes used on non-uniformly sampled data, first using linear interpolation to fill in the data gaps. Linear interpolation alters the signals, but the FFT is still able to capture an acceptable level of spectral details, depending on the size and number of gaps. This is demonstrated quantitatively and qualitatively in Sects. 3 and 4.

The effect of linear interpolation can be derived analytically from first principles. For a given signal $x(t)$, the Fourier transform $y(\omega)$ is defined as (see, e.g., Bloomfield, 2000):

$$y(\omega) = \int_{t_1}^{t_n} x(t) e^{-i\omega t} \mathrm{d}t. \quad (1)$$

If we assume a gap between t_a and t_b, we will have the following:

$$y(\omega) = \int_{t_1}^{t_a} x(t)e^{-i\omega t}\mathrm{d}t + g + \int_{t_b}^{t_n} x(t)e^{-i\omega t}\mathrm{d}t, \quad (2)$$

where

$$g = \int_{t_a}^{t_b} x(t)e^{-i\omega t}\mathrm{d}t \cong -g1 + g2, \quad (3)$$

with

$$g1 = i\left(x(t_a) - \frac{x(t_b) - x(t_a)}{t_b - t_a}t_a\right) \times \frac{e^{-i\omega t_a} - e^{-i\omega t_b}}{\omega}, \quad (4)$$

and

$$g2 = \frac{x(t_b) - x(t_a)}{t_b - t_a} \times \frac{e^{-i\omega t_b}(1 + i\omega t_b) - e^{-i\omega t_a}(1 + \omega t_a)}{\omega^2}, \quad (5)$$

where the data gap is replaced by a straight line. As the gap size $(t_b - t_a)$ is increased, both $g1$ and $g2$ will decrease, resulting in smaller Fourier amplitudes $y(\omega)$. This is also true for the dependence on ω: as we increase the frequency, the two terms corresponding to the data gap will decrease, resulting in decreased amplitudes. At low frequencies, and for large differences between the two end points of the gap $(x(t_b) - x(t_a) > 0)$, $g2$ can become larger than $g1$ leading to increased Fourier amplitudes. This simple analytical example shows that FFT, in case of linearly interpolated data gaps, can lead to an underestimation of high-frequency amplitudes and an overestimation of low-frequency amplitudes.

The discrete Fourier transform is a discretization of the Fourier integral of Eq. (1), which we chose to do using the trapezoidal method:

$$y(\omega) = \sum_{j=1}^{n} \frac{x(t_{j+1})e^{-i\omega t_{j+1}} + x(t_j)e^{-i\omega t_j}}{2} \times (t_{j+1} - t_j). \quad (6)$$

For a comprehensive description of the FFT, DFT, and Fourier analysis in general, the reader is invited to consult monographs such as those by Bath (1974) and Priestley (1981).

The z-transform is a generalization of the Fourier Transform for discrete series rather than for continuous functions (see, e.g., Jury, 1973; Weeks, 2011). By definition, the z-transform of a signal $x(t)$ is

$$y(z) = \sum_{j=1}^{n} x(t_j)z^{-t_j}, \quad (7)$$

where z is a complex number. Using the exponential notation, $z = re^{i\omega}$, and choosing $r = 1$, we get the following:

$$y(\omega) = \sum_{j=1}^{n} x(t_j)e^{-i\omega t_j}, \quad (8)$$

which is usually considered to be the formal definition for the discrete Fourier transform. In order to use both forms of the discrete Fourier transform, Eq. (6) will define the DFT method used throughout our paper, and Eq. (8) will be referred to as ZTR.

The Lomb-Scargle method performs a least squares fit of the data using a superposition of sinusoidal modes (Lomb, 1976; Scargle, 1982, 1989; Hocke and Kämpfer, 2009).

By applying the four methods above we obtain a complex amplitude spectrum $y(\omega)$ for each method, where $\omega = 2\pi f$, with the values for the frequency f uniformly distributed between $f_{\min} = f_s/L$ and $f_{\max} = fs/2$, $f_s = 1/\mathrm{d}t$ being the sampling frequency, $\mathrm{d}t$ the sampling interval, and L the number of points. We then compute the one-sided amplitude spectrum by calculating the complex magnitude (modulus) of $y(\omega)$, normalize it by the number of points L, and multiplying it by 2 (since we use only positive frequencies).

By squaring the one-sided amplitude spectrum one can obtain the periodogram, which is a non-parametric estimate of the power spectral density (PSD). Our methodology of systematically comparing the spectra for the signals with gaps with the spectrum for the original, uniformly sampled signal, allows us to use only the simple one-sided spectrum to study the effect of data gaps. While more comprehensive methodologies of estimating the PSD do exist, they are beyond the scope of this study. In the following, we will use the term amplitude spectrum to denote the one-sided amplitude spectrum.

While DFT, ZTR and LST are equivalent to FFT for uniform sampling, they can provide, as we will show later, very different results when analyzing time series with data gaps.

Note that the issues due to data gaps are not limited to a single method, but are a fundamental property of the resulting amplitude spectrum. For a uniformly sampled data set, there is an orthogonal set of frequencies for which the values of the Fourier coefficients are independent. In the case of non-uniformly sampled data, such an orthogonal set generally does not exist (Van Dongen et al., 1999), allowing for spectral leakage to occur. This cannot be avoided regardless of the method used since the problem does not derive from the algorithm. For this reason the problem cannot be entirely resolved. We intend to diagnose the extent by which the data gaps impact the results obtained with different methods and for different gap configurations.

3 Synthetic data tests

In order to determine the effect introduced by data gaps on the Fourier transform, we test the four above-mentioned spectral analysis methods on a synthetic signal with artificial gaps. The FFT method is applied to a signal where the gaps are linearly interpolated, while DFT, ZTR and LST are applied to a signal with no data interpolation. We test two configurations of gaps: (a) a single large gap (SLG), based

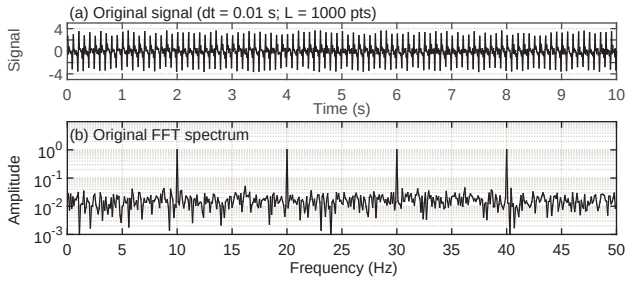

Figure 1. (a) Synthetic signal and **(b)** its associated FFT amplitude spectrum. The signal consists of four sinusoidal modes, with amplitude 1 and frequencies $f1 = 10$ Hz, $f2 = 20$ Hz, $f3 = 30$ Hz and $f4 = 40$ Hz, plus noise.

on the alteration of the original data set by removing an increasing number of points from the central part; the procedure is repeated until the central gap reaches 99.8 % of the total length of the signal, and (b) multiple small gaps (MSG), based on the alteration of the signal by removal of short series of consequent points whose length and precise location are randomly selected; the procedure is applied repeatedly for various distributions of random gaps. The two cases were studied first on a synthetic signal sampled uniformly.

The synthetic data set is a superposition of four sinusoidal modes with unit amplitude and the following frequencies: $f_1 = 10$ Hz, $f_2 = 20$ Hz, $f_3 = 30$ Hz and $f_4 = 40$ Hz. The signal is sampled at 100 samples per second with a total length of $L = 1000$ points. We also added a white noise with unit amplitude such that the synthetic signal $x(t)$ can be described by the following:

$$x(t) = \text{noise} + \sum_{i=1}^{4} \sin(2\pi f_i t), \tag{9}$$

and is illustrated in Fig. 1 together with its amplitude (FFT) spectrum. The latter serves as reference and will be compared to the spectra obtained by the four methods applied to various distributions of gaps.

3.1 Single-large-gap test applied to the synthetic data set

In this case the data gap is created by removing a number of points from the central part of the synthetic signal. We generate 100 signals from Eq. (9) on which we generate a single gap with size varying from 1 to 99.8 % of the length of the original time series. As an example, we illustrate in Fig. 2 the performance of the four methods on a signal whose gap is 50 % of the length of $y(t)$. Figure 2 shows that when the FFT analysis is applied to the interpolated signal, it provides a Fourier spectrum whose amplitude is half of the original spectrum at all four eigen-frequencies of the synthetic signal.

The accuracy of the amplitude spectrum computed with DFT is sensitive not only to the size of the central gap but also to the phase at the two end points of the gap. If at least one end point has a value different from the mean value of the signal (which is zero in the case of our synthetic signals) then the results are distorted. This distortion is also seen in Fig. 2, where the DFT amplitudes depict a very large background level (even above one), and the four signals barely rise above the background. This distortion is larger for large deviations of the two end points from the mean.

Since we are interested mainly in the effect of the gap size, we apply a Tukey (tapered cosine) window (Bloomfield, 2000) to the two parts of the signal around the gap, which cancels the offset on either side of the gap (as well as at start and end of the signal). In order to treat all four methods similarly, we apply the same windowing procedure to all the four methods. The results obtained after the windowing procedure are shown in Fig. 3, in the same format as Fig. 2. For FFT, the windowing procedure removes the high amplitudes at very low frequencies seen in Fig. 2b, which were due to the slope of the linear interpolation. DFT results are now similar to FFT, and show the same 50 % decrease in amplitude.

Figure 2 shows that the two other spectral analysis methods, ZTR and LST, provide very accurately the same amplitude level as the original, full data set, even when the gap is quite wide. We also see that the spectral background level ($\sim 3 \times 10^{-2}$) is larger for these two methods, compared to the original level of $\sim 10^{-2}$ depicted in Fig. 1b. Figure 3 shows that the windowing procedure hardly affects the ZTR and LST results. The increase of the spectral background will have important implications for the analysis of real data, as we will see in Sect. 4.

Figures 4 and 5 show how the amplitudes change when the length of the central gap is increased. Figure 4 shows the amplitude spectra obtained by the four methods in color coding, with the y-axis giving the total gap percentage (TGP) of the signal and the x axis representing the frequency. Figure 4 includes 100 spectra of the signal (Eq. 9) with the central gap increasing linearly from 1 to 99.8 % of the signal. Figure 4a shows that the amplitude spectrum calculated by the FFT transform decreases systematically with the increasing size of the central gap. The results of the FFT and the windowed DFT are very similar until the TGP exceeds about 80 %, when the DFT background noise level increases dramatically. The spectra corresponding to gap sizes larger than about 60 % exhibit a gradual appearance and broadening of a series of side lobes, leading to a palm-tree shape in the vicinity of the spectral peaks. The gradual broadening of the individual spectral peaks can be explained by the finite length of the original signal; the degradation of the signal by removal of an increasingly large central part decreases the number of sinusoidal peaks and leads to a broadening of the spectral peak.

The spectra obtained with ZTR and LST are quite different, but bear some similar features. As the size of the central gap increases, the signal amplitude remains almost con-

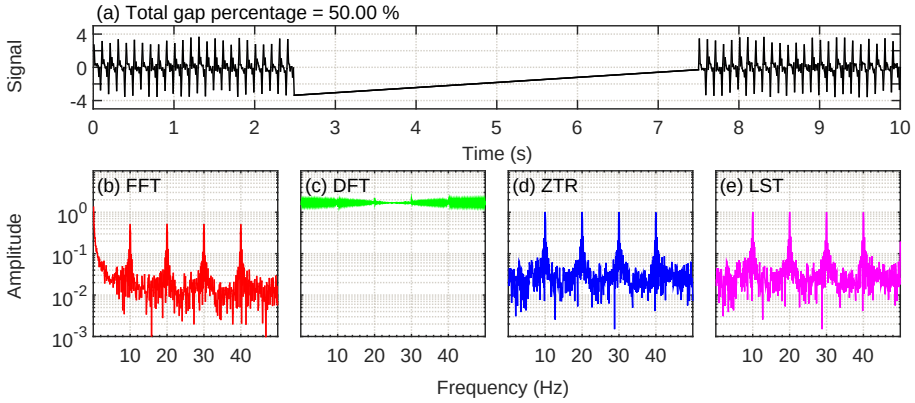

Figure 2. Case study for the SLG test applied to synthetic data. Panel **(a)** shows the non-windowed signal with total gap percentage (TGP) of 50 % and panels **(b)**, **(c)**, **(d)** and **(e)** show the corresponding amplitude spectra computed with the fast Fourier transform (FFT), the discrete Fourier transform (DFT), the Z transform (ZTR) and the Lomb–Scargle transform (LST). FFT is applied to the signal were the data gap was linearly interpolated, while the other three methods are applied to the signal containing the data gap. The TGP parameter represents the total number of points removed from the time series, and is defined as a percentage of the original length L.

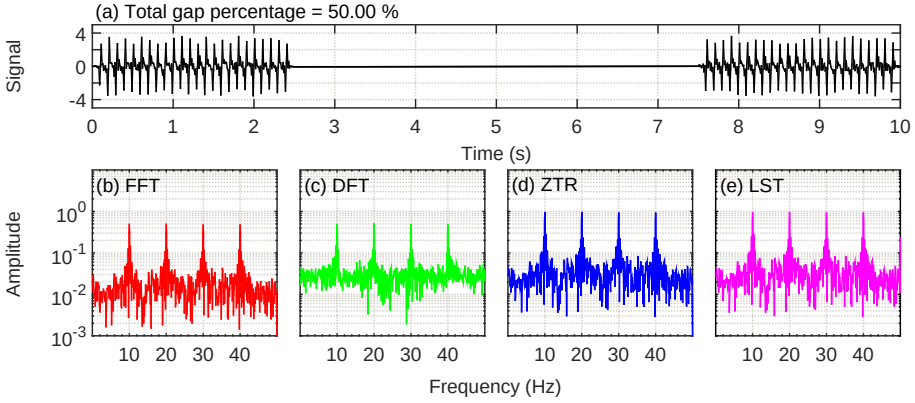

Figure 3. Case study for the SLG test applied to the synthetic data. Panel **(a)** shows the windowed signal with TGP of 50 % and the other panels are the same as in Fig. 2. The windowing procedure, using a Tukey window, is applied to the two parts of the signal around the gap, thus canceling the offset on either side of the gap.

stant for all the four spectral peaks. Moreover, Fig. 4 shows that the side lobes (the palm-tree) and the background noise have considerably larger amplitudes for ZTR and LST than for FFT and DFT.

In order to illustrate even more quantitatively the response of the four methods to the increasing size of the central gap we have studied the change for the first frequency, $f_1 = 10$ Hz, of the full signal. Figure 5 shows how the amplitude of the spectrum at f_1 varies with TGP. Up to TGP < 80 %, FFT and DFT show that the amplitude is monotonically decreasing with increasing TGP. Beyond TGP of about 80 %, the DFT amplitude increases rather erratically, indicating the increasing background level (see Fig. 2). On the other hand, the amplitude at f_1 obtained from ZTR and LST remains very close to 1 up to TGP of about 95 %. Beyond this value, both methods give increasingly disturbed amplitude levels.

Figure 6 shows the integral of the amplitude spectrum, i.e., the sum of all amplitudes, as a function of the TGP. In order to study the effect of spectral noise we calculated the integral not only for the synthetic signal given by Eq. (9), but also for a "clean" signal (the sum of sinusoids without noise) as well as for pure noise. We found that FFT and DFT behave in a similar way within the limit of small TGP, up to about 60 % (see earlier discussion). The integral corresponding to the clean signal has an almost constant value as we increase the gap size. For pure noise, the integral decreases with increasing gap size. For the signal contaminated with noise, the original signal given by Eq. (9), the result is a combination of these two cases, i.e., the integral slowly decreases with increasing TGP.

For the ZTR and LST methods, the integral increases as we increase the gap size not only for pure noise and the noisy signal, but also for the clean sum of sinusoids. We already

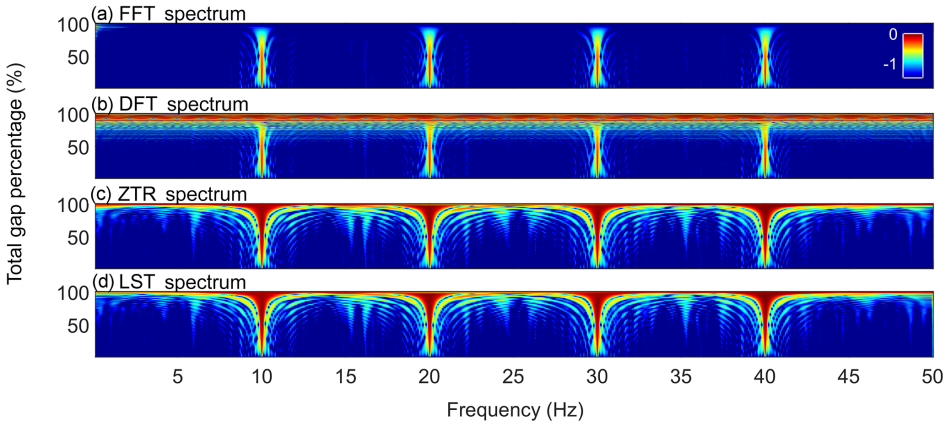

Figure 4. SLG test applied to the synthetic data: amplitude in color code as a function of TGP and frequency. For the SLG test, we generate 100 signals, indexed from 1 to 100, with signal 1 corresponding to the signal with the smallest TGP and signal 100 to the one with the highest TGP. For SLG, the TGP is a linear function of signal index. Panels **(a)**, **(b)**, **(c)** and **(d)** show the results for FFT, DFT, ZTR and LST. Color scale is logarithmic and the color bars denote the log10 of amplitude.

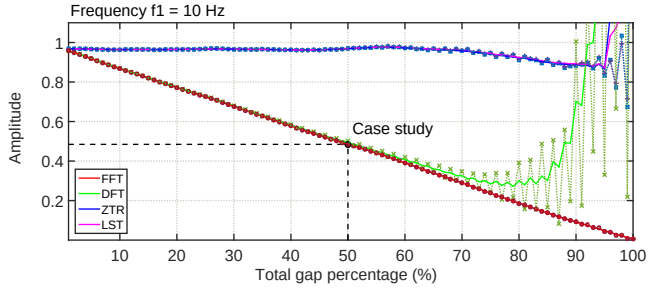

Figure 5. SLG test applied to the synthetic data: amplitude of the 10 Hz sinusoidal mode (f_1) as a function of TGP for: FFT (red line marked with circles), DFT (green with x), ZTR (blue and square) and LST (magenta and +). The case study for a TGP of 50 % (see Fig. 3) is noted as a vertical black line, and the horizontal black line gives the FFT amplitude for this case.

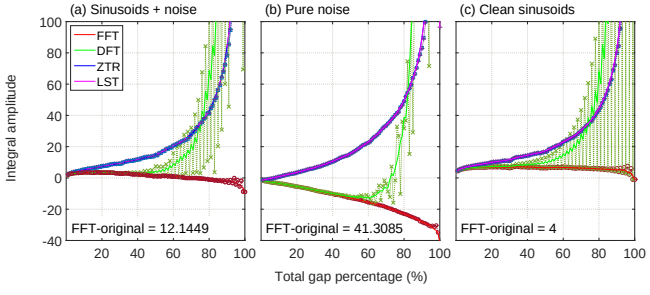

Figure 6. SLG test applied to the synthetic data: integral amplitudes as a function of TGP for the signal with noise (panel **a**), for pure noise (panel **b**) and for the clean signal (panel **c**). The clean signal is obtained by removing the noise from the original signal (see Fig. 1). Colors denote the four methods. The thick lines are the average values.

observed in Fig. 3 that, for the 50 % TGP, the spectral background was larger than the original background. We see here that the background level increases systematically as we increase the gap size, eventually dominating the integral for large TGP and explaining the overall increase for these two methods. This result will have an important impact on the analysis of solar wind data, where dominant harmonics are less frequent and the spectral noise dominates the integral.

3.2 Multiple-small-gaps test applied to the synthetic data set

The second test performed on the synthetic signal of Eq. (8) consists of removing a number of randomly distributed points from the original signal. This test mimics the situation often encountered in the experimental investigation of various geophysical or space systems (e.g., ground-based measurement of the geomagnetic field or satellite measurement of the plasma and field parameters of the solar wind), where randomly distributed data gaps are an inherent problem.

We choose the size distribution of the gaps using the gamma function, which can be described by two parameters: the shape parameter A_g and the scale parameter B_g. The probability density function (PDF) for the gamma distribution can be expressed in terms of A_g and B_g, as follows:

$$\mathrm{PDF_g}(x \mid A_g, B_g) = \frac{x^{A_g-1} \times e^{-x/B_g}}{B_g^{A_g} \times \Gamma(A_g)}. \quad (10)$$

We use this PDF to create statistical ensembles of gaps, with gap size probability being controlled by the mean ($M_g = A_g \times B_g$) and variance ($V_g = A_g \times B_g^2$) of the gamma distribution. In practice we choose a set of values for M_g and V_g and then compute a vector of gamma-distributed random numbers according to Eq. (9). The obtained vector comprises a set

of real numbers from ∼ zero to a positive value G_m depending on M_g and V_g. These numbers are rounded to the nearest integer value and thus we obtain the distribution of gap sizes, each integer giving the number of consecutive points to be removed from the uniformly sampled signal. The increasing degradation of the signal is achieved by increasing both M_g, which increases the size of the most probable gap, and V_g, which increases the probability of obtaining large gaps.

As in the case of SLG, we calculate the Fourier spectra and index them according to the selection of M_g and V_g values. We also define the TGP associated with each spectrum as the total percent of points removed from the signal, computed as the sum of all gaps.

Figure 7 shows an example of a distribution of gaps for $M_g = 2.1$ and $V_g = 1.1$, which removes 51 % of the points of the signal. In this example the gaps with small size (between 1 and 3 consecutive points) have high probability, and their cumulative effect is to remove more than 40 % of the points of the original signal, as indicated by Fig. 7b. The largest gap in this example has a size of 8 points; there is only one gap of this size.

The amplitude spectrum obtained with FFT shows that the amplitude of the sinusoidal modes decreases systematically with increasing frequency. The other three methods (DFT, ZTR and LST) are very robust for this configuration of gaps and show no major modification in the amplitude spectra compared with the original results (see Fig. 1).

The methodology outlined above was applied to an ensemble of 100 synthetic signals obtained by degrading the original signal by increasing the number of missing points according to the gamma distribution. The mean and variance of the distribution of gaps increase with signal index, and thereby, the total number of removed points increases, although not strictly linearly. Figure 8a shows the distribution of gap percentage as a function of signal index and gap size, and Fig. 8b shows for each of the 100 degraded signals the corresponding TGP. We note that this statistical ensemble of signals covers indeed a broad range of different possible configurations, relevant to investigate the response of the four spectral analysis methods.

Figure 9 depicts the amplitude spectra of the 100 signals described above and in Fig. 8 as a function of TGP and frequency, similar to the SLG test of Fig. 4. One can see that the FFT amplitude decreases with frequency and with increasing TGP. The results for DFT, ZTR and LST show little decrease in amplitude and no frequency dependence when TGP increases.

Figure 10 shows the FFT amplitudes as a function of TGP separately for the four frequencies f_1, f_2, f_3 and f_4. FFT amplitudes decrease rather systematically with TGP. However, there is some variation in amplitudes, especially for large TGP, which is due to the different effect of each individual gap sample. So, TGP is not the only factor affecting the amplitude, but also the distribution of gaps matters.

Figure 11 shows the analogue of Fig. 6 for the MSG test, i.e., the amplitude integral as a function of the TGP. As in Fig. 6, the amplitude integral for the original signal is removed in each case. There are interesting differences between Figs. 11 and 6 that are related to the different gap structure between the two cases, even for the same total TGP. In particular, there is a much larger integral amplitude in FFT, even for clean sinusoids, which is due to the increased background power. Figure 9 shows that the FFT amplitude at low frequencies increases with TGP. However, there is an increased background level even at lower TGP of about 40 % (see Fig. 9), where the integral amplitude attains its maximum, as seen in Fig. 11.

4 Satellite data tests

The solar wind is a supersonic, turbulent plasma stream released from the upper atmosphere of the Sun. It is often considered to be the ideal turbulence laboratory, due to its very large scale, compared to the usual Earth-based laboratories, and to the large fleet of spacecraft actively investigating its properties. Starting with the work by Coleman (1968), and many more similar studies since, we now know that the magnetic frequency spectrum of the solar wind, in a range of intermediate frequencies, roughly behaves as a power-law (see, e.g., Bruno and Carbone, 2013). In this regard, the solar wind magnetic field measurements used in our study constitute a highly representative sample. By comparing the results of the SLG and MSG tests for two time series with very different properties, one containing distinct spectral peaks, while the other one showing a power-law behavior of the amplitude spectrum, allows us to strengthen and extrapolate the results for the simple artificial data set to the much more complex real-world time series.

We now apply the above methods to the magnetic field data from the *Venus Express* (VEX) spacecraft (Zhang et al., 2006) in orbit around the planet Venus. The signal represents a sample of the x component of the solar wind magnetic field, measured by VEX on 17 January 2007. It includes 1000 data points sampled at 1 s time resolution, forming a time series of 16 min and 40 s without gaps. Figure 12 shows the signal and its FFT amplitude spectrum.

4.1 Single-large-gap test applied to the VEX data set

This testing method is identical to the one described above in Sect. 3.1. A case study obtained by removing 50 % of the central part of the original signal (TGP = 50 %) is shown in Fig. 13. Since the signal does not have significant peaks in the original spectrum (see Fig. 12), we study the performance of the four methods by comparing the average amplitude spectra. For a given spectrum, the average is computed using a moving average filter with a span of 100 points. Like in the corresponding synthetic test, we see that FFT and DFT per-

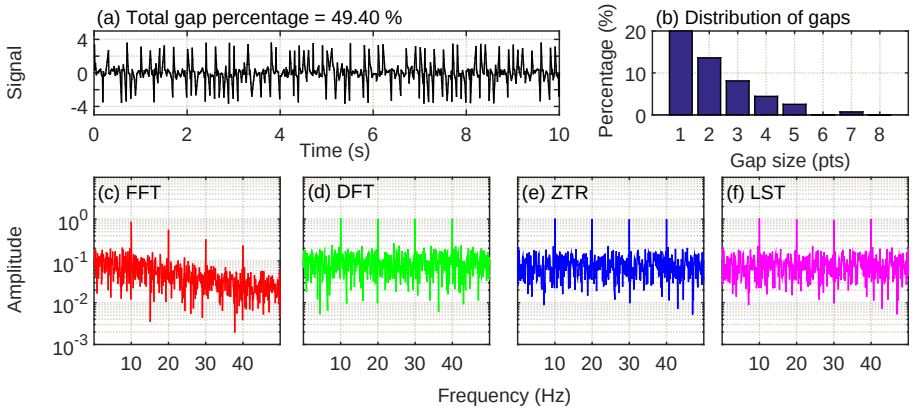

Figure 7. Case study for the MSG test applied to the synthetic data. The format is similar to Fig. 2 and 3, except for panel **(b)**, which shows the individual gap percentage as a function of gap size corresponding to this case study.

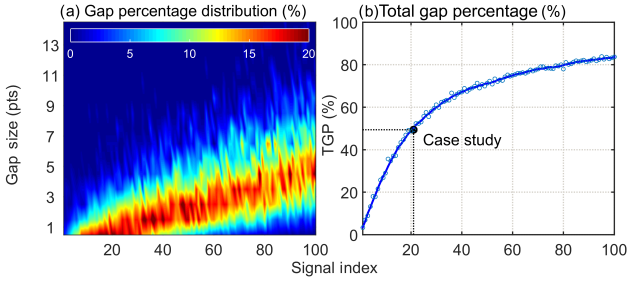

Figure 8. MSG test applied to the synthetic data: **(a)** distribution of gap percentage (color coded) as a function of signal index and gap size; **(b)** TGP as a function of signal index. Note that for the MSG test, the TGP is not a linear function of signal index.

form roughly in a similar way. The spectral amplitudes of the signal with TGP = 50 % are at a clearly lower level than the original spectrum over the whole frequency interval. For ZTR and LST, the overall average level of the spectrum is fairly similar to the original one.

The results for the ensemble of 100 signals with the TGP increasing linearly from 1 to 99.8 % (the procedure described in Sect. 3.1) are shown in Figs. 14 and 15 (analogues of Figs. 4 and 5). These figures illustrate the difference between the estimated spectra of the signal with gaps and the original spectrum. FFT and DFT show in Fig. 14 an overall decrease in amplitude with increasing gap size. However, due to the more complicated spectral content of the real signal, the results are not as clear as for the synthetic signal (Fig. 4). Also, due to the more complex nature of the signal, the windowing procedure is not very effective at large TGP values, and the edge effect will result in a large increase of DFT amplitudes for TGP above 80 %. ZTR and LST show an increase in amplitude with increasing gap size. Figure 14 shows the integral of the amplitude spectra as a function of TGP. We see here a pattern very similar to the one obtained for pure noise analysis (see Fig. 6b): a decrease in FFT and DFT (until about 60 % TGP) and an increase in ZTR and LST as we increase the gap size.

4.2 Multiple-small-gaps test applied to the VEX data set

A case study for the VEX signal from which we have removed 49.4 % of the points by introducing small gaps according to the gamma distribution is shown in Fig. 16. We see that this gap configuration produces excessive power over most of the frequency range for DFT, ZTR and LST. Only FFT produces an amplitude spectrum that is close to the original one. One can notice a threshold frequency at about $f_t = 10^{-1}$ Hz which is common for the three methods, above which the spectral slope departs strongly from the original one. This threshold value is related to the size of the largest gap. In this case study, the period corresponding to the threshold frequency is $p_t = 1/f_t = 10$ s, which is very close the largest gap of 6 points (i.e., 6 s) (see Fig. 16b).

Figures 17 and 18 show the results of the test for the ensemble of 100 signals with variable gap size distribution generated by the gamma function. Figure 17 shows the difference between the actual amplitude spectrum and the original spectrum. It gives us a synoptic view of the behavior of the four methods when degradation is increasing. Figure 17 shows that the FFT amplitude spectrum underestimates the original spectrum for frequencies higher than about 0.1 Hz.

On the other hand, DFT, ZTR and LST overestimate the spectrum over most of the frequency range. A significant increase is detected for frequencies larger than 0.1 Hz, in agreement with the case study depicted in Fig. 16. Interestingly the increase seen in these three methods is frequency dependent, and the affected range of frequencies seems to depend on the actual distribution of the gaps and the power of the signal. These results indicate that, at least when the data includes gaps, FFT is the best method to approximate reliably the spectral slope of a signal recorded in a turbulent environment. Figure 18 illustrates the integral spectral amplitude

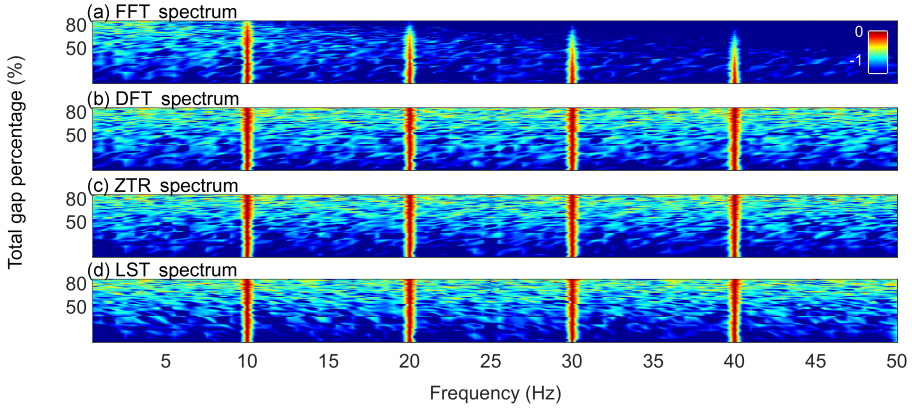

Figure 9. MSG test applied to the synthetic data: amplitude spectra in color code as a function of TGP and frequency for: FFT (panel **a**) DFT (panel **b**), ZTR (panel **c**) and LST (panel **d**). Color as in Fig. 4.

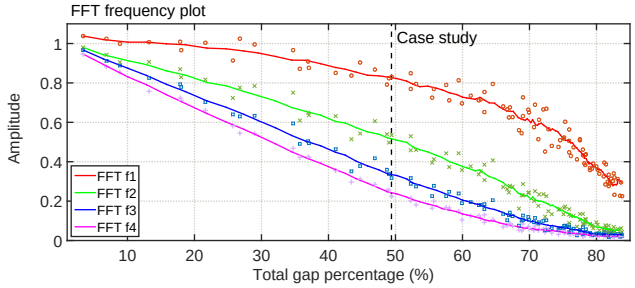

Figure 10. MSG test applied to the synthetic data: FFT amplitude as a function of TGP for the four frequencies of the synthetic signal: $f_1 = 10$ Hz (red), $f_2 = 20$ Hz (green), $f_3 = 30$ Hz (blue) and $f_4 = 40$ Hz (magenta). The thick lines are the average values, computed using a moving average filter with a span of 10 points.

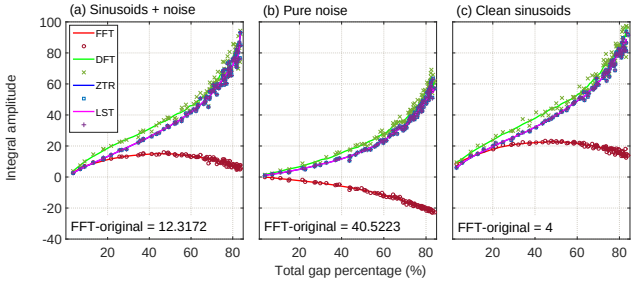

Figure 11. MSG test applied to the synthetic data: integral amplitudes as a function of TGP. The thick lines are the average values of the individual integral amplitudes. The format is identical to Fig. 6.

(as difference to the original) as a function of the TGP. The integral gives a global measure of the spectrally differentiated behavior seen in Fig. 17. Figure 18 shows quantitatively the better agreement of the FFT amplitude with the original spectrum than the three other methods, even for large TGP.

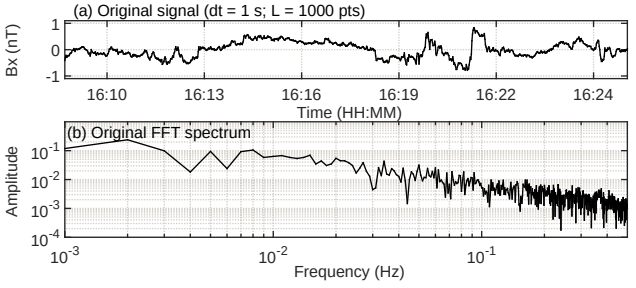

Figure 12. A *Venus Express* (VEX) magnetic field signal. Panel **(a)** shows the B_x component of the magnetic field as a function of time and panel **(b)** shows its FFT amplitude spectrum.

5 Discussion and conclusions

We have analyzed the effect of data gaps on four commonly used spectral analysis methods: the fast Fourier transform, the discrete Fourier transform, the Z transform and the Lomb–Scargle algorithm. FFT is extremely fast and readily available in all programming languages, and it is by far the most popular method of estimating the amplitude spectrum. It is often applied also to signals containing data gaps, using interpolation to compensate for the lack of data. The simple discretization of the Fourier integral using the trapezoidal method (DFT), can be used without interpolation even in the presence of data gaps. The Z transform, a generalization of the Fourier transform for discrete series and the Lomb-Scargle algorithm, a least squares fit of the data using a superposition of sinusoidal modes, are straight-forwardly applicable for time series with non-uniform sampling and/or data gaps.

In order to study the effect of data gaps and to mimic frequently encountered gap configurations, we devised two tests: the single-large-gap (SLG) test, which removes a number of consecutive points from the signal, and the multiple-small-gaps (MSG) test, which removes a number of ran-

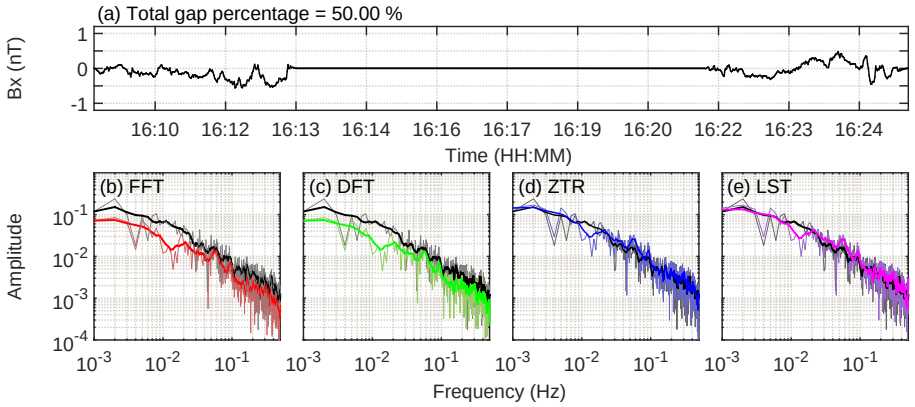

Figure 13. Case study for the SLG test applied to the VEX data set. The format is similar to Fig. 3. In addition, panels **(b)**–**(e)** also show the original FFT spectrum (black) and the average spectra for each method (thick lines).

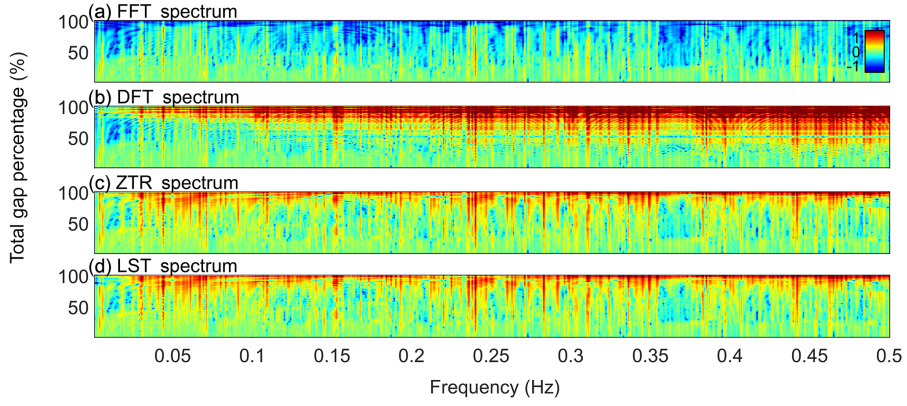

Figure 14. SLG test applied to the VEX data set. Represented are the difference of amplitude spectra (method – original) as a function of TGP and frequency for: **(a)** FFT, **(b)** DFT, **(c)** ZTR and **(d)** LST.

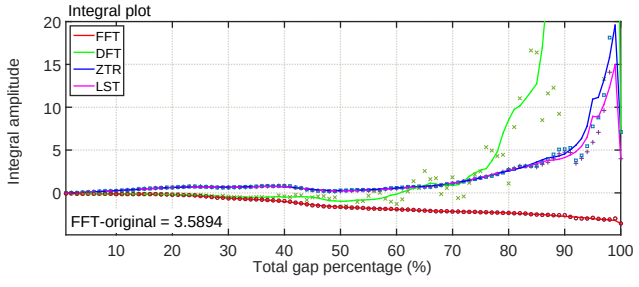

Figure 15. SLG test applied to the VEX data set: integral amplitudes as a function of TGP. Shown are the results for: FFT (red), DFT (green), ZTR (blue) and LST (magenta).

domly distributed gaps whose size was given by the gamma distribution. Both tests include an ensemble of 100 signals with gap percentage increasing from 1 % (4.3 %) to 99.8 % (83.6 %) in the case of the SLG test (MSG test, respectively). The tests are applied to two data sets: a simple noisy superposition of four sinusoidal modes and magnetic field measurements made by the *Venus Express* spacecraft in orbit around the planet Venus.

The DFT method is very sensitive to the phase at the two end points of the gap, giving very distorted results for large deviations of the end points from the mean value of the signal. To remove this distortion, the SLG test uses a windowing procedure, where we apply a window function on the two parts of the signal around the central gap, thus canceling the offset on either side of the gap (as well as at the start and end of the signal). The signal was windowed for all four methods, although the effect of windowing was minimal for the other three methods.

For FFT and DFT, the SLG test shows monotonically decreasing amplitudes of the sinusoidal modes, with increasing total gap percentage (TGP). For DFT this decrease continues only to the limit of TGP of about 80 %, where after the background level increases considerably because the windowing procedure becomes less effective (due to the small number of data points and related numerical noise). For the VEX data, used here as an example of a realistic data set

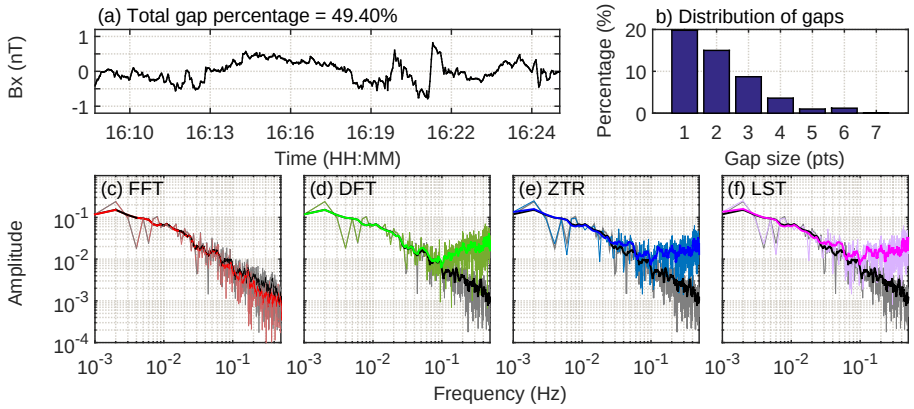

Figure 16. Case study for the MSG test applied to the VEX data set. The format is similar to Fig. 7. Panels **(c)**, **(d)**, **(e)** and **(f)** also show the original FFT spectrum (black) and the average spectra for each method (thick lines).

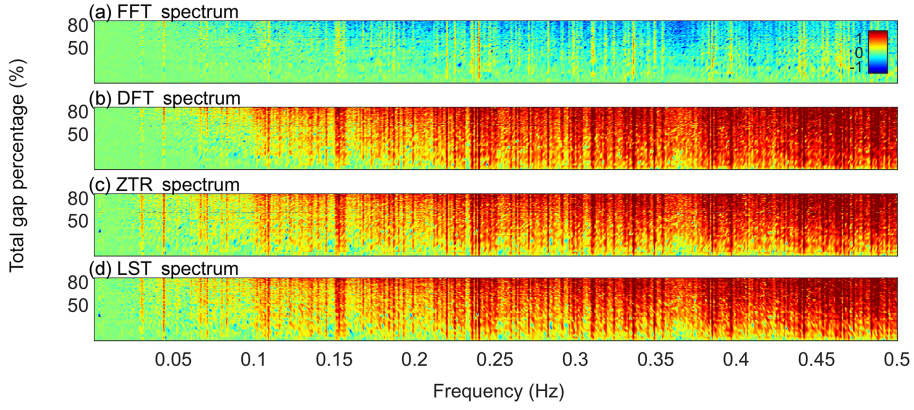

Figure 17. MSG test applied to the VEX data set. The format is similar to Fig. 14.

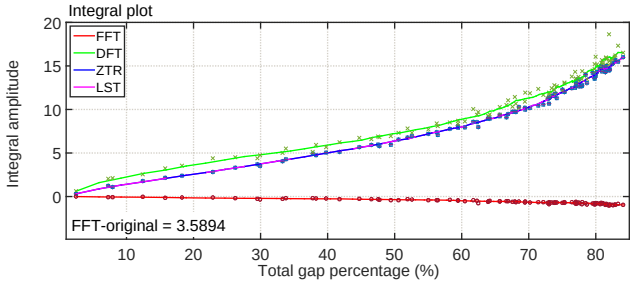

Figure 18. MSG test applied to the VEX data set: integral amplitudes as a function of TGP. The format is similar to Fig. 15.

with a more complex spectrum, the results are influenced by the non-stationarity of the time series, but, overall, we see a similar decrease in FFT and DFT amplitude when increasing the TGP, as for the synthetic data. For the synthetic data set, ZTR and LST show neither decrease in the amplitude of the sinusoidal modes nor frequency dependence, when the TGP is increased in the SLG test. However, for very large gaps, the spectral background becomes increasingly noisy, leading to an overall increase of the average level of the spectrum. For the VEX data, since we do not have clear spectral peaks, we only see the increase of the average level of the spectrum as TGP increases.

The MSG test probes the effect of small, randomly distributed, gaps within the time series. For the synthetic data set, FFT is the only method which is severely affected by such small gaps, showing decreased amplitude and systematic frequency dependence in amplitude reduction, with high frequencies being most affected. On the other hand, DFT, ZTR and LST are able to recover the amplitudes of the sinusoidal modes, but the spectral background becomes increasingly noisy when increasing the TGP. For the VEX data, the FFT shows a similar decrease in amplitude and frequency dependence as for the synthetic case. On the other hand, DFT, ZTR and LST seriously overestimate the high frequency part of the amplitude spectrum above a certain threshold frequency. Moreover, we found that this threshold is dependent on the distribution of the small data gaps, and is moving to a lower frequency as the gap size increases. Beyond this threshold the spectral amplitude is roughly constant since the

Table 1. Summary of the results of the comparative numerical studies with various types of gaps applied to synthetic and real data from the *Venus Express* satellite, respectively.

Data set	Single large gap	Multiple small gaps
Synthetic	FFT and DFT show monotonically decreasing amplitudes of the sinusoidal modes, with increasing TGP. ZTR and LST show neither decrease in the amplitude of the sinusoidal modes nor frequency dependence, when the TGP is increased.	FFT shows decreased amplitudes and a systematic frequency dependence in amplitude reduction, with high frequencies being most affected. DFT, ZTR and LST recover the amplitudes of the sinusoidal modes, but the spectral background becomes increasingly noisy when increasing the TGP.
Venus Express	FFT and DFT show a decrease in the average level of the spectrum when we increase the TGP. ZTR and LST show an overall increase of the average level of the spectrum as TGP increases.	FFT shows a decrease in amplitudes and a frequency dependence. DFT, ZTR and LST overestimate the high frequency part of the amplitude spectrum.

data gaps cover a large range of gap sizes corresponding to the frequency range above the threshold.

Table 1 summarizes the main conclusions of our study. The two columns show the results of the two tests (SLG – left and MSG – right) for the synthetic data set (row 1) and for the *Venus Express* data set (row 2).

Concluding, the FFT method can be used even for relatively large single data gaps, although the absolute value of the amplitude spectrum is systematically reduced with gap size. On the other hand, the ZTR and LST methods preserve the absolute level of the amplitude spectrum, but are more vulnerable to increasing spectral background arising from increasing the TGP. They are recommended for the analysis of signals with strong sinusoidal modes, giving robust results for the amplitude of sinusoidal modes. For more turbulent spectra, the appearance of side lobs and spectral noise makes the effect of data gaps more pronounced for these methods than for FFT. Thus, our results indicate that, at least when the data includes gaps, FFT is the best among the four tested methods to approximate reliably the spectral slope of a signal recorded in a turbulent environment.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement 313038 (STORM), and a grant of the Romanian Ministry of National Education, CNCSUEFISCDI, project No. PN-II-ID-PCE-2012-4-0418. We also acknowledge the financial support by the Academy of Finland to the ReSoLVE Center of Excellence (project 272157) and to project 264994.

The topical editor, G. Balasis, thanks Z. Voros and R. V. Donner for help in evaluating this paper.

References

Babu, P. and Stoica, P.: Spectral analysis of nonuniformly sampled data – a review, Digital Signal Processing, 20, 359–378, doi:10.1016/j.dsp.2009.06.019, 2010.

Baisch, S. and Bokelmann, G. H.: Spectral analysis with incomplete time series: an example from seismology, Comput. Geosci., 25, 739–750, doi:10.1016/S0098-3004(99)00026-6, 1999.

Bath, M.: Spectral Analysis in Geophysics, Elsevier Scientific Publishing Company, Amsterdam, ISBN: 0-444-41222-0, 1974.

Bloomfield, P.: Fourier Analysis of Time Series: An Introduction, 2nd Edition, A Wiley-Interscience Publication, http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471889482.html, ISBN: 0-471-88948-2, 2000.

Britz, D. and Antonia, R. A.: A comparison of methods of computing power spectra of LDA signals, Meas. Sci. Technol., 7, 1042, doi:10.1088/0957-0233/7/7/008, 1996.

Broersen, P., de Waele, S., and Bos, R.: Accuracy of Time Series Analysis for Laser-Doppler Velocimetry, in: Proceedings of the 10th International Symposium on Applications of Laser Techniques to Fluid Dynamics, Lisbon, Portugal, available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.473.3886, 2000.

Bruno, R. and Carbone, V.: The Solar Wind as a Turbulence Laboratory, Living Reviews in Solar Physics, 10, doi:10.12942/lrsp-2013-2, 2013.

Coleman, Jr., P. J.: Turbulence, Viscosity, and Dissipation in the Solar-Wind Plasma, Astrophys. J., 153, 371, doi:10.1086/149674, 1968.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P.: Advanced spectral methods for climatic time series, Rev. Geophys., 40, 3–1–3–41, doi:10.1029/2000RG000092, 2002.

Harteveld, W., Mudde, R., and den Akker, H. V.: Estimation of turbulence power spectra for bubbly flows from Laser Doppler Anemometry signals, Chem. Eng. Sci., 60, 6160–6168, doi:10.1016/j.ces.2005.03.037, 2005.

Heslop, D. and Dekkers, M.: Spectral analysis of unevenly spaced climatic time series using CLEAN: signal recovery and derivation of significance levels using a Monte Carlo simulation, Phys. Earth Planet. In., 130, 103–116, doi:10.1016/S0031-9201(01)00310-7, 2002.

Hocke, K. and Kämpfer, N.: Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram, Atmos. Chem. Phys., 9, 4197–4206, doi:10.5194/acp-9-4197-2009, 2009.

Jury, E. I.: Theory and Application of the Z-Transform Method, R. E. Krieger Publishing Company, ISBN: 9780882751221, 1973.

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151–159, doi:10.5194/npg-13-151-2006, 2006.

Kondrashov, D., Shprits, Y., and Ghil, M.: Gap filling of solar wind data by singular spectrum analysis, Geophys. Res. Lett., 37, L15101, doi:10.1029/2010GL044138, 2010.

Kondrashov, D., Denton, R., Shprits, Y. Y., and Singer, H. J.: Reconstruction of gaps in the past history of solar wind parameters, Geophys. Res. Lett., 41, 2702–2707, doi:10.1002/2014GL059741, 2014.

Lomb, N.: Least-squares frequency analysis of unequally spaced data, Astrophys. Space Sci., 39, 447–462, doi:10.1007/BF00648343, 1976.

Musial, J. P., Verstraete, M. M., and Gobron, N.: Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series, Atmos. Chem. Phys., 11, 7905–7923, doi:10.5194/acp-11-7905-2011, 2011.

Pope, S. A., Zhang, T. L., Balikhin, M. A., Delva, M., Hvizdos, L., Kudela, K., and Dimmock, A. P.: Exploring planetary magnetic environments using magnetically unclean spacecraft: a systems approach to VEX MAG data analysis, Ann. Geophys., 29, 639–647, doi:10.5194/angeo-29-639-2011, 2011.

Priestley, M. B.: Spectral Analysis and Time Series, Academic Press, London, ISBN: 0-12-564901-0, 1981.

Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J.: Comparison of correlation analysis techniques for irregularly sampled time series, Nonlin. Processes Geophys., 18, 389–404, doi:10.5194/npg-18-389-2011, 2011.

Scargle, J. D.: Studies in astronomical time series analysis. II – Statistical aspects of spectral analysis of unevenly spaced data, The Astrophysical Journal, 263, 835–853, doi:10.1086/160554, 1982.

Scargle, J. D.: Studies in astronomical time series analysis. III – Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data, The Astrophysical Journal, 343, 874–887, doi:10.1086/167757, 1989.

Schimmel, M.: Emphasizing Difficulties in the Detection of Rhythms with Lomb-Scargle Periodograms, Biol. Rhythm Res., 32, 341–346, doi:10.1076/brhm.32.3.341.1340, 2001.

Stahn, T. and Gizon, L.: Fourier Analysis of Gapped Time Series: Improved Estimates of Solar and Stellar Oscillation Parameters, Solar Phys., 251, 31, doi:10.1007/s11207-008-9181-0, 2008.

Teodorescu, E., Echim, M., Munteanu, C., Zhang, T., Bruno, R., and Kovacs, P.: Inertial range turbulence of fast and slow solar wind at 0.72 AU and solar minimum, Astrophys. J. Lett., 804, doi:10.1088/2041-8205/804/2/L41, 2015.

Tsyganenko, N. A. and Sitnov, M. I.: Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms, J. Geophys. Res. (Space Physics), 110, A03208, doi:10.1029/2004JA010798, 2005.

Van Dongen, H., Olofsen, E., Van Hartevelt, J., and Kruyt, E.: A Procedure of Multiple Period Searching in Unequally Spaced Time-Series with the Lomb-Scargle Method, Biol. Rhythm Res., 30, 149–177, doi:10.1076/brhm.30.2.149.1424, 1999.

van Maanen, H. R. E. and Oldenziel, A.: Estimation of turbulence power spectra from randomly sampled data by curve-fit to the autocorrelation function applied to laser-Doppler anemometry, Meas. Sci. Technol., 9, 458, doi:10.1088/0957-0233/9/3/021, 1998.

Weeks, M.: Digital Signal Processing Using MATLAB and Wavelets, Second Edition, Jones and Bartlett Publishers, Inc., 2nd edn., http://dl.acm.org/citation.cfm?id=1841667, 2011.

Zhang, T., Baumjohann, W., Delva, M., Auster, H.-U., Balogh, A., Russell, C., Barabash, S., Balikhin, M., Berghofer, G., Biernat, H., Lammer, H., Lichtenegger, H., Magnes, W., Nakamura, R., Penz, T., Schwingenschuh, K., Vörös, Z., Zambelli, W., Fornacon, K.-H., Glassmeier, K.-H., Richter, I., Carr, C., Kudela, K., Shi, J., Zhao, H., Motschmann, U., and Lebreton, J.-P.: Magnetic field investigation of the Venus plasma environment: Expected new results from Venus Express, Planet. Space Sci., 54, 1336–1343, doi:10.1016/j.pss.2006.04.018, 2006.