Effects of Inhomogeneous Data Coverage on Spectral Analysis

R. Pail, W.-D. Schuh

Mathematical Geodesy and Geoinformatics (MGGI)
Technical University Graz, Steyrergasse 30, A-8010 Graz, AUSTRIA

ABSTRACT: The present study investigates the effects of irregular data distribution on the corresponding frequency representation. For a basic understanding it starts with a simple 1D-model and provides a description of effects emerging from transition of the infinite continuous to the finite discrete case. Considering regular (gridded) and irregular data distributions and even gap regions without data coverage, the corresponding results in frequency domain and the numerical stability are analyzed using trigonometric functions as base functions.

1 Introduction

In most cases the spectral analysis is closely linked with continuous or discrete equispaced sampled functions. Each arbitrary continuous function $f(x)$ (signal) can be represented by an infinite series of $sin-$ and $cos-$functions (*Fourier series*),

$$f(x) = {\sum_{k=0}^{\infty}}' a_k \cos kx + \sum_{k=1}^{\infty} b_k \sin kx \qquad (1)$$

with an infinite set of parameters a_k and b_k [1]. The determination of these parameters a_k and b_k

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx \qquad (2)$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx \qquad (3)$$

results from the orthogonality of the base functions sin and cos with respect to the integration in the interval $[-\pi,\pi]$ (*Fourier integrals*).

Similar relations also hold if N equispaced discrete samples of the function $f(x_i)$ are known. In connection with a finite and discrete set of data locations N, the reconstruction of the maximum involved frequency is bounded to the Nyquist frequency $(N-1)/2$. The usual equispaced (complete) discretization performs an orthogonal and thus numerically stable reconstruction of the signal. The unknown coefficients a_k and b_k can be determined parameterwise by

$$a_k = \frac{2}{N}{\sum_{i=0}^{N-1}}' f(x_i)\cos ki\,\frac{2\pi}{N} \qquad (4)$$

$$b_k = \frac{2}{N}\sum_{i=1}^{N-1} f(x_i)\sin ki\,\frac{2\pi}{N}\ . \qquad (5)$$

Considering that the relations $\frac{2}{N} = \frac{\Delta x}{\pi}$ and $x_i = i\frac{2\pi}{N}$ hold, this formula can be interpreted as the discrete solution of the Fourier integrals (2), (3), resp. However, on the other side this computation can also be seen as inverse problem due to equation (1). Each data point contributes a linear equation (reproduction condition) and yields an orthogonal design matrix (cf. fig. 1a).

Most of the discussions on spectral analysis deal with these two cases, which are closely linked to the continuous and discrete *Fourier transform*, and some effects emerge from the transition from the continuous to the equispaced discrete case (aliasing, leakage) and the restriction to a finite or special interval (Gibbs-phenomena, edge-effects[2]). In practice only a finite number of generally not equispaced data is available. Therefore, we have to distinguish between a variety of different cases (cf. tab. 1).

Table 1 illustrates the large variety of spectral analysis techniques.

max. involved frequency L	number of data locations N	max. resolved frequency P	number of parameter $M=2P+1$	
∞	∞	∞	∞	cont. Fourier analysis
L	$N=2L+1$	L	$M=N$	classical sampling (disc. Fourier analysis)
	$N<2L+1$	**incomplete discretization - aliased problems**		
		$P=\frac{N-1}{2}$	$M=N$	definite aliased
		$P>\frac{N-1}{2}$	$M>N$	underdetermined aliased (Backus&Gilbert)
		$P<\frac{N-1}{2}$	$M<N$	overdetermined aliased (least squares adjust.)
	$N>2L+1$	**complete discretization - overdetermined problems**		
		$P=L$	$N>M$	well-balanced
		$\frac{N-1}{2}>P>L$	$N>M$	overparameterized
		$P<L$	$N>M$	underparameterized (spectral leakage)

Table 1: Spectral Analysis Techniques.

The most important item within the analysis represents the maximum frequency L inherent in the signal $f(x)$. A strict reconstruction of the signal requires a sufficient sampling (cf. tab. 1, complete discretization). Only a higher sampling rate gives access to higher frequencies, and as an extreme case an infinite number of measurements is necessary to

[1] The prime in the first sum refers to the convention that the term $k=0$ should be taken with *half weight*.

[2] A special form of leakage caused by an underrepresentation of long wavelengths.

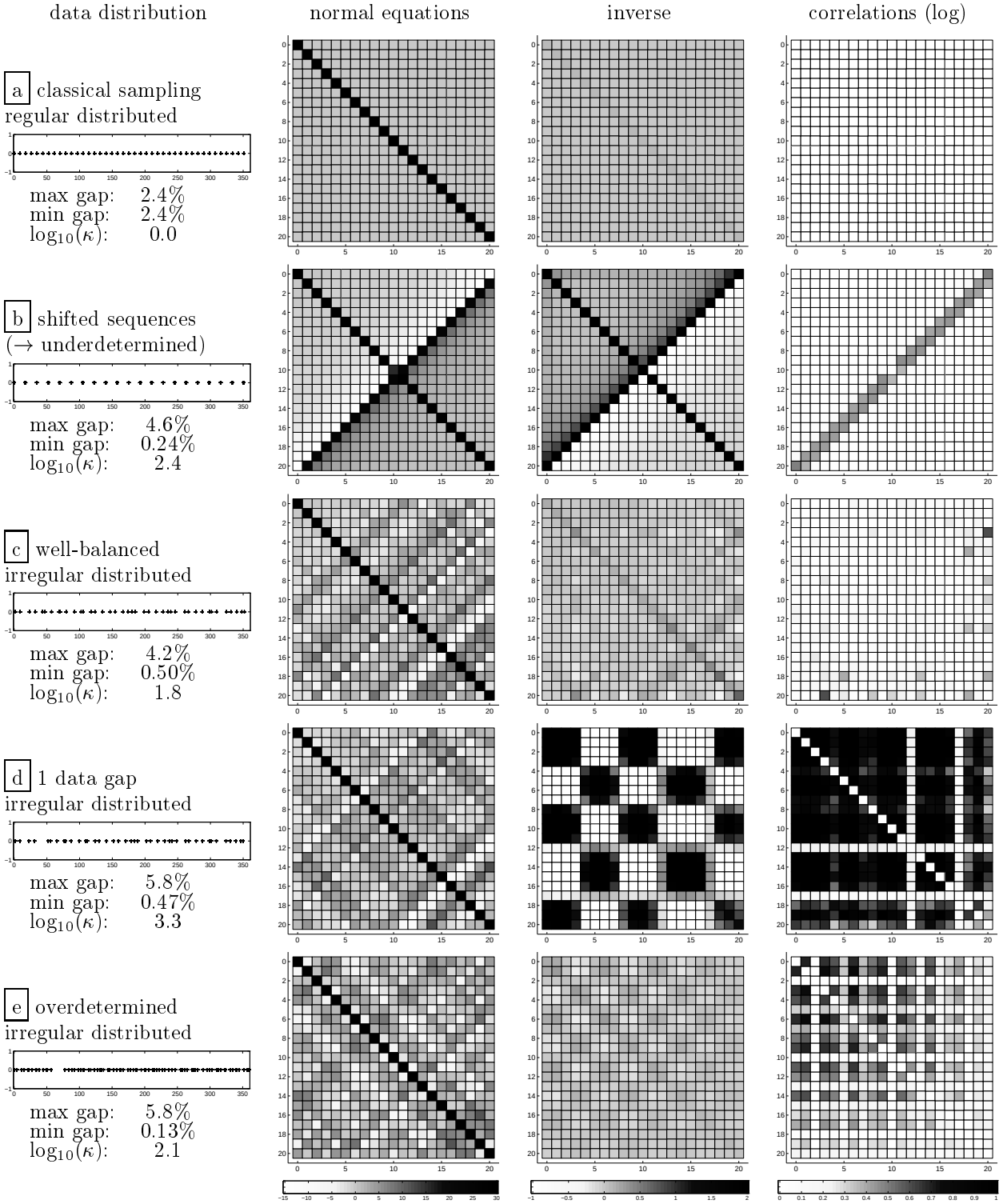

Figure 1: This figure shows the numerical behaviour of spectral analysis due to typical data distributions. In addition to some alphanumeric information on the left, the normal equations, the inverse of the normal equations and a rescaled correlation matrix ($\mathtt{log}_{10}(1 - \mathtt{abs}(\rho_{ij}))$) are plotted. Besides the classical equispaced sampling (fig. 1a), different types of data distributions are studied. Sampling by shifting equispaced sequences generally allows a stable reconstruction of the signal. In the normal equations this relative shift is reflected by non-zero entries in the second main diagonal, whose amplitude grows with increasing shift rate (fig. 1b). Simultaneously the condition number worsens. In the extreme case the two equispaced sequences coincide, resulting in an underdetermined and thus singular system.
Many test computations with irregular data coverage (fig. 1c) yield that the numerical behaviour is dominated by the size of the maximum data gap. This fact is underpinned by simulation fig. 1d, where the maximum data gap is enlarged compared with simulation fig. 1c, leading to an increase in the condition number representing the stability of the system. In general, the extension of the gap size influences the numerical behaviour much more extensively than a distance reduction between adjacent data locations. The stability can be improved by additional measurements, resulting in an overdetermined system demonstrated in simulation fig. 1e and consequently in a reduction of the condition number.

resolve a signal with infinite frequency content. Otherwise an underdetermined problem occurs (Backus & Gilbert problem, cf. [1] or [5]) and additional information (e.g. degree-variances, cf. [3]) is required to determine an unique set of parameters.

As mentioned above, the number of equispaced data locations steers the maximum resolvable frequency (Nyquist frequency). The same fact holds also for irregular distributed data locations. Each location x_i supplies an equation of type (1). The reconstruction of the parameters a_k and b_k can again be performed by a discrete version of the Fourier integrals or by the solution of the $N \times N$ equation system. However, if some disturbances due to higher frequencies inherent in the signal or some disturbances in the discrete function values occur, these two approaches pursue different principles. The integral approach tries to reconstruct an orthogonal base by a weighting strategy due to the data location. Therefore, the recreation of an individual frequency forms the main target. On the other side, the inversion (adjustment) procedure tries to find a best reconstruction of the signal, regardless of the recreation of each individual frequency.

Irregular data distributions lead to a loss of orthogonality and thus to non-diagonal systems of normal equations. The stability (condition number) of the design matrix on the one side and the frequency of the residuals on the other side form the main criteria. Fig. 1 reflects the behaviour of typical situations.

2 Data Gaps

Detailed investigations have been performed to explore the connection between data gaps and the numerical stability of the system of normal equations. Many test computations with irregular data coverage yield that its numerical behaviour is dominated by the size of the maximum data gap. Therefore we investigate in this section the simplified configuration of one growing data gap with equispaced sampling intervals beyond the gap region. The examples of section 3 will confirm that similar results can also be obtained from irregular data coverage.

Fig. 2 shows the stability of the corresponding system of normal equations, expressed by the decadic logarithm of the condition number κ, as a function of the relative gap size r, which is defined by

$$r = \frac{\text{absolute gap}}{\text{analysis interval}} \quad . \tag{6}$$

P and μ denote the maximum resolved frequency (cf. tab. 1) and a factor ($1 \leq \mu < \infty$) describing the redundancy of the system, respectively. Consequently, for the number of data follows

$$N = 2\mu P + 1 \; . \tag{7}$$

Fig. 2 yields an approximately linear relation between the relative gap size r and the logarithm of the condition number κ up to the numerically computable limit of about 15 digits and thus provides a first possibility to predict the numerical stability.

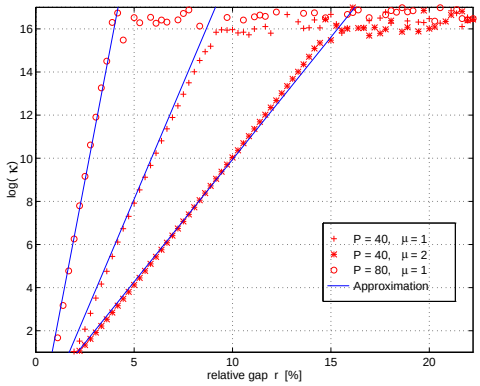

Figure 2: condition κ vs. relative gap size r, based on a given maximum resolved frequency P and a redundancy μ.

This nearly linear behaviour is used to normalize the logarithmic condition number $log_{10}(\kappa)$ by the relative gap size r. The resulting normalized quantity characterizes the slope of the curves shown in fig. 2. Plotting these slopes for various redundancies μ versus the maximum resolved frequency P, again an approximately linear behaviour becomes obvious.

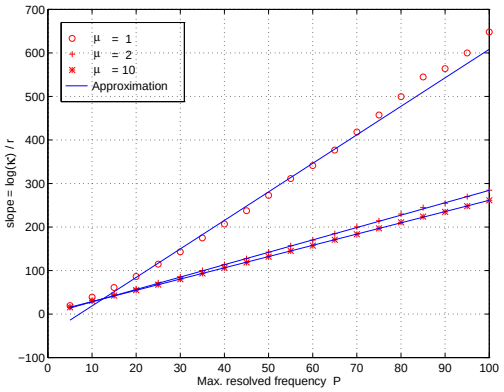

Figure 3: slope ($\alpha P + \beta$) vs. maximum resolved frequency P (redundancy μ=const.).

Finally, applying regression analysis, the linear regression parameters $\alpha(\mu)$ and $\beta(\mu)$, depending on the redundancy μ, can be estimated (cf. fig. 3), and consequently a relation between the maximum resolved frequency P, the redundancy μ (resp. number of data N), the gap size r, and the logarithm of condition number κ can be established[3]:

$$\log_{10}(\kappa) = (\alpha P + \beta) \left(r - \frac{1}{2P+1} \right) \tag{8}$$

with

$$\alpha = \frac{0.32}{\mu - 0.92} + 2.55 \; ; \quad \beta = \frac{-4.00}{\mu - 0.92} + 3.30 \; . \tag{9}$$

[3] The term $1/(2P+1)$ in (8) describes the fact that the curves in fig. 2 do not cross the origin of the coordinate system.

The solid lines in figs. 2 and 3, respectively, yield that this rule of thumb fits the strict solutions very well. Inversely to fig. 3, fig. 4 demonstrates that

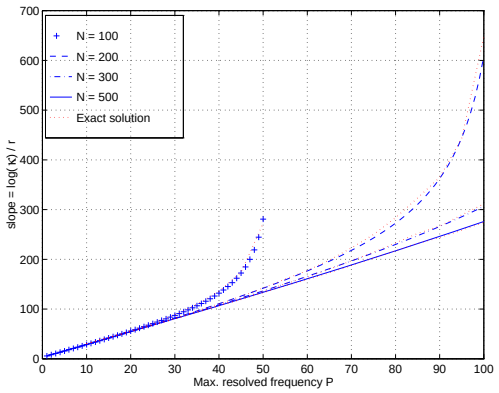

Figure 4: slope ($\alpha P+\beta$) vs. maximum resolved frequency P (number of data N=const). The approximation results are compared with the exact solutions (dotted curves).

for a given number of data points N within a fixed analysis interval, a reduction of the maximum resolved frequency P (corresponding to an increase of redundancy μ) leads to a marked stabilization of the system and consequently to a decrease of condition number κ. For example, concerning the dashed line ($N = 200$) in fig. 4, we can estimate a reduction of the slope from 600 for the max. resolved frequency $P = 100$ to about 140 for $P = 50$, corresponding to the redundancy of $\mu = 2$. Both relations can also be extracted from fig. 3. For the well-balanced case ($\mu = 1$) using the curve indicated by circles ($P = 100 \rightarrow slope \sim 600$) and the two-fold overdetermined case ($\mu = 2$) taking the crossed curve into account ($P = 50 \rightarrow slope \sim 140$). Both figures illustrate very impressively the marked improve of the numerical behaviour already for small redundancies up to two. A further redundancy increase improves the stability only slightly.

Consequently, the rule of thumb (8) as a function of κ, r, P, μ, allows the estimation of one parameter, if the other three parameters are fixed. This is demonstrated on the basis of a few examples in the following section.

3 Examples

Example 1 demonstrates the straightforward prediction of the condition number based on the given parameters P, r and μ. The last line provides the results for a strict computation with respect to regular and irregular data coverage beyond the gap.

IN	max. degree	P	40
	relative gap	r	5%
	redundancy	μ	1
OUT	condition	$\log_{10}(\kappa)$	8.1
STRICT	$\log_{10}(\kappa)$	regular	7.9
		irregular	8.4

Example 1.

This situation is also treated graphically by the crossed curve in fig. 2.

Example 2: Which additional number of data is required to remain under a given condition number $\kappa = 10^5$? The rule of thumb estimates a redundancy factor of 1.24. A recalculation using this configuration yields condition numbers of $10^{4.8}$ (regular case) and $10^{5.4}$ (irregular case), respectively.

IN	max. degree	P	40
	condition	$\log_{10}(\kappa)$	5
	relative gap	r	5%
OUT	redundancy	μ	1.24
STRICT	$\log_{10}(\kappa)$	regular	4.8
		irregular	5.4

Example 2.

Example 3: How many parameters (max. frequency) can be resolved if we postulate a condition $\kappa = 10^5$, gap size $r = 5\%$ and redundancy $\mu = 2$? Applying the resulting maximum frequency estimate $P = 45$, the recalculation reflects the postulated stability.

IN	redundancy	μ	2
	condition	$\log_{10}(\kappa)$	5
	relative gap	r	5%
OUT	max. degree	P	45
STRICT	$\log_{10}(\kappa)$	regular	4.9
		irregular	5.0

Example 3.

4 Two Data Gaps

In section 2 we mentioned that in the case of irregular data coverage the numerical behaviour is dominated by the size $r = r_1$ of the maximum data gap. In most cases the numerical stability worsens with a growing second gap r_2 and markedly depends on the relative position of the two gaps. As an example fig. 5

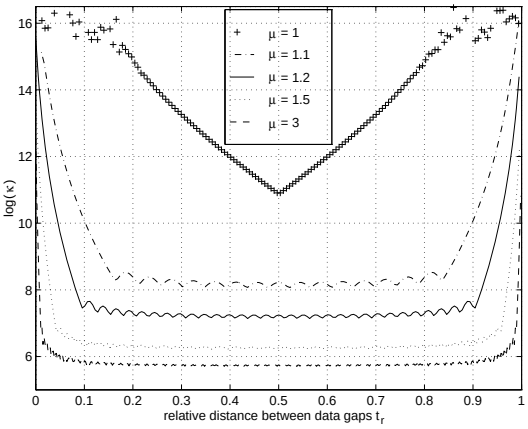

Figure 5: condition number κ vs. relative distance between gaps t_r.

demonstrates the numerical behaviour of a system of two data gaps of equal size and equispaced data

distribution in between as a function of the distance between the two gaps and for different redundancy factors.

In the *well-balanced case* ($\mu = 1$), the condition number grows with decreasing distance and in this example even reaches the numerical limit in the range of $\kappa = 10^{15}$. The minimum occurs for a periodic configuration. In this special case of two periodic gaps of equal size the condition number turns out to be lower by about 10 to 15 % compared with only one gap of the same size. As the contrary extreme case the two gaps unify, $r = 2\,r_1$, and thus the rule of thumb (8) provides the worst case estimate for two data gaps[4].

Please note the marked drop of the condition number in the case of *overdetermination* ($\mu > 1$). In this case a whole interval of distances yields approximately equal condition numbers, where this interval extends with increasing degree of redundancy. As a consequence, already a small number of additional measurements leads to an extensive improvement of the system's numerical stability.

5 Conclusions

Given a continuous function, its frequency properties - represented by the Fourier coefficients - can be uniquely determined by solving the Fourier integrals. In the case of the transition to a discrete, equidistantly sampled function, the reconstruction of coefficients can be performed either by a discrete analogon of the Fourier integrals or by the solution of an orthogonal system of equations. These two approaches differ only if certain disturbances are inherent in the signal.

Concerning irregular data coverage, numerical problems emerge due to the loss of orthogonality properties of the base functions, leading to non-diagonal normal equation systems, whose numerical behaviour turns out to be dominated by the size of the maximum data gap. Consequently, we propose a simple rule of thumb based on this one gap configuration, which gives a good estimate ($\pm 10\%$) for the numerical stability. Since this rule of thumb can also be applied in the presence of large data gaps, it may help to facilitate feasibility studies by estimating whether a certain data configuration is numerically stable, and how to modify the parameter model (max. resolved frequency, redundancy) in order to stabilize the system, respectively. The rule of thumb is applied in a selection of examples in order to demonstrate its practicability.

[4] An improved estimate for arbitrary ranges results from a linear approximation between the periodic case (rule of thumb value for data gap size r_1, diminished by about 10 %) and the case of gap unification (rule of thumb value for gap size $2\,r_1$), which finally reads

$$\log_{10}(\kappa) = (6.55P - 46.7)\left[2r - \frac{1}{2P+1} - \left(2.2r - \frac{0.2}{2P+1}\right) t_r\right] \quad (10)$$

with

$$t_r = \frac{\text{absolute gap distance}}{\text{analysis interval}}\ ; \quad t_r \in [0,\ 1/2]\ . \quad (11)$$

References

[1] BACKUS G., F. GILBERT (1970): Uniqueness in the Inversion of Inaccurate Gross Earth Data. Philos. Trans. R. Soc. London, 266, pp. 123-192.

[2] BRIGHAM E.Oran (1974): The Fast Fourier Transform. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[3] SCHWARZ K.P. (1976): Least Squares Collocation for Large Systems. Bollettino di Geodesia e Science Affini, Anno XXXV, N.3, pp. 309-324.

[4] SNIEDER R. (1990): Global inversions using normal modes and long-period surface waves. In: Seismic tomography, edited by H. M. Iyer and K. Hirahara, pp. 23-63, Prentice-Hall,London.

[5] TARANTOLA A. (1987): Inverse Problem Theory. Elsevier, Amsterdam.