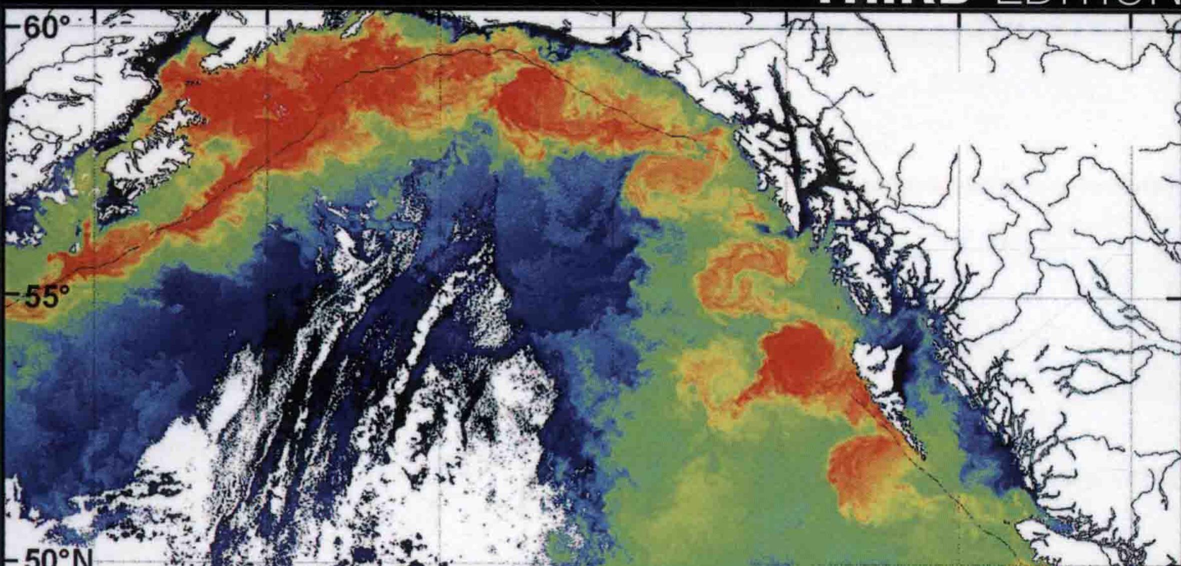
ELSEVIER

[redacted]

DATA ANALYSIS METHODS IN PHYSICAL OCEANOGRAPHY

THIRD EDITION



Richard E. Thomson
William J. Emery

DATA ANALYSIS METHODS IN PHYSICAL OCEANOGRAPHY

THIRD EDITION

RICHARD E. THOMSON
Fisheries and Oceans Canada
Institute of Ocean Sciences
Sidney, British Columbia
Canada

and

WILLIAM J. EMERY
University of Colorado
Aerospace Engineering Sciences Department
Boulder, CO
USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
225, Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-387782-6

For information on all **Elsevier** publications visit our web site at http://store.elsevier.com/

Printed in the United States of America

- Coastal photo courtesy of Dr. Audrey Dallimore, School of Environment and Sustainability, Royal Roads University, British Columbia, Canada
- Satellite image from Thomson, R.E., and J.F.R. Gower. 1998. A basin-scale oceanic instability event in the Gulf of Alaska. J. Geophys. Res., 103, 3033-3040

DATA ANALYSIS METHODS IN PHYSICAL OCEANOGRAPHY

THIRD EDITION

Dedication

Richard Thomson dedicates this book to his wife Irma, daughters Justine and Karen, and grandchildren Brenden and Nicholas.

Bill Emery dedicates this book to his wife Dora Emery, his children Alysse, Eric, and Micah, and to his grandchildren Margot and Elliot.

Preface

There have been numerous books written on data analysis methods in the physical sciences over the past several decades. Most of these books are heavily directed toward the more theoretical aspects of data processing or narrowly focus on one particular topic. Few books span the range from basic data sampling and statistical analysis to more modern techniques such as wavelet analysis, rotary spectral decomposition, Kalman filtering, and self-organizing maps. Texts that also provide detailed information on the sensor and instruments that collect the data are even more rare. In writing this book we saw a clear need for a practical reference volume for earth and ocean sciences that brings established and modern processing techniques together under a single cover. The text is intended for students and established scientists alike. For the most part, graduate programs in oceanography have some form of methods course in which students learn about the measurement, calibration, processing, and interpretation of geophysical data. The classes are intended to give the students needed experience in both the logistics of data collection and the practical problems of data processing and analysis. Because the class material generally is based on the experience of the faculty members giving the course, each class emphasizes different aspects of data collection and analysis. Formalism and presentation can differ widely. While it is valuable to learn from the first-hand experiences of the class instructor, it seemed to us important to have available a central reference text that could be used to provide some uniformity in the material being covered within the oceanographic community. This 3rd Edition provides a much needed update on oceanographic instrumentation and data processing methods that have become more widely available over the past decade.

Many of the data analysis techniques most useful to oceanographers can be found in books and journals covering a wide variety of topics. Much of the technical information on these techniques is detailed in texts on numerical methods, time series analysis, and statistical methods. In this book, we attempt to bring together many of the key data processing methods found in the literature, as well as add new information on spatial and temporal data analysis techniques that were not readily available in older texts. Chapter 1 also provides a description of most of the instruments used in physical oceanography today. This is not a straightforward task given the rapidly changing technology for both remote and in situ oceanic sensors, and the ever-accelerating rate of data collection and transmission. Our hope is that this book will provide instructional material for students in the marine sciences and serve as a general reference volume for those directly involved with oceanographic and other branches of geophysical research.

The broad scope and rapidly evolving nature of oceanographic sciences has meant that it has not been possible for us to fully detail all existing instrumentation or emerging data analysis methods. However, we believe that many of the methods and procedures outlined in this book will provide a basic understanding of the kinds of options available to the user for interpretation of data sets. Our intention is to describe general statistical and analytical methods that will be sufficiently fundamental to maintain a high level of utility over the years.

Finally, we also believe that the analysis procedures discussed in this book apply to a wide readership in the geophysical sciences. As with oceanographers, this wider community of scientists would likely benefit from a central source of information that encompasses not only a description of the mathematical methods, but also considers some of the practical aspects of data analyses. It is this synthesis between theoretical insight and the logistical limitations of real data measurement that is a primary goal of this text.

Richard E. Thomson and William J. Emery
North Saanich, British Columbia
and Boulder, Colorado

Acknowledgments

Many people have contributed to the three editions of this book over the years. Dudley Chelton of Oregon State University and Alexander Rabinovich of Moscow State University and the Institute of Ocean Sciences have helped with several chapters. Dudley proved to be a most impressive reviewer and Sasha has provided figures that have significantly improved the book. For this edition, we also thank friends and colleagues who took time from their research to review sections of the text or to provide suggestions for new material. There were others, far too numerous to mention, whose comments and words of advice have added to the usefulness of the text. We thank Andrew Bennett of Oregon State University for reviewing the section on inverse methods in Chapter 4, Brenda Burd of Ecostat Research for reviewing the bootstrap method in Chapter 3, and Steve Mihály of Ocean Networks Canada for assisting with the new section on self-organizing maps in Chapter 4. Roy Hourston and Maxim Krassovski of the Institute of Ocean Sciences helped generously with various sections of the book, including new sections on regime shifts and wavelet analysis in Chapter 5. The contributions to Chapter 1 from Tamás Juhász and David Spear, two accomplished technicians at the Institute of Ocean Sciences, are gratefully acknowledged. Patricia Kimber of Tango Design helped draft many of the figures.

Expert contributions to the third edition—including reports of errors and omissions in the previous editions—were also provided by Michael Foreman, Robie Macdonald, Isaac Fine, Joseph Linguanti, Ron Lindsay, Germaine Gatien, Steve Romaine, and Lucius Perreault (Institute of Ocean Sciences, Fisheries and Oceans Canada), Philip Woodworth (Permanent Service for Mean Sea Level, United Kingdom), William (Bill) Woodward (President and CEO of CLS America, Inc., USA), Richard Lumpkin (NOAA/AOML, USA), Laurence Breaker (California State University, USA), Jo Suijlen (National Institute for Coastal and Marine Management, The Netherlands), Guohong Fang (First Institute of Oceanography, China), Vlado Malačič (National Institute of Biology, Slovenia), Ųyvind Knutsen (University of Bergen, Norway), Parker MacCready (University of Washington, USA), Andrew Slater (University of Colorado, USA), David Dixon (Plymouth, United Kingdom), Drew Lucas (Scripps Institute of Oceanography, USA), Wayne Martin (University of Washington, USA), David Ciochetto (Dalhousie University, Canada), Alan Plueddemann (Woods Hole Oceanographic Institution, USA), Fabien Durand (IRD/LEGOS, Toulouse, France), Jack Harlan (NOAA, Boulder Colorado, USA), Denis Gilbert (The Maurice Lamontagne Institute, Fisheries and Oceans Canada), Igor Yashayaev (Bedford Institute of Oceanography, Fisheries and Oceans Canada), Ben Hamlington (University of Colorado, USA), John Hunter (University of Tasmania, Australia), Irene Alonso (Instituto de Cienciias Mrinas de Andalucia, Spain), Yonggang Liu (University of South Florida, USA), Gary Borstad (ASL Environmental Sciences Ltd., Canada), Earl Davis and Bob Meldrum (Pacific Geosciences Centre, Canada), and Mohammand Bahmanpour (University of Western Australia, Australia).

Contents

6. Digital Filters

CHAPTER

1

Data Acquisition and Recording

1.1 INTRODUCTION

Physical oceanography is an ever-evolving science in which the instruments, types of observations, and methods of analysis undergo continuous advancement and refinement. The changes that have occurred since we completed the 2nd Edition of this book over a decade ago have been impressive. Recent progress in oceanographic theory, instrumentation, sensor platforms, and software development has led to significant advances in marine science and the way that the findings are presented. The advent of digital computers has revolutionized data collection procedures and the way that data are reduced and analyzed. No longer is the individual scientist personally familiar with each data point and its contribution to his or her study. Instrumentation and data collection are moving out of direct application by the scientist and into the hands of skilled technicians who are becoming increasingly more specialized in the operation and maintenance of equipment. New electronic instruments operate at data rates and storage capacity not possible with earlier mechanical devices and produce volumes of information that can only be handled by high-speed computers. Most modern data collection systems transmit sensor data directly to computer-based data acquisition systems where they are stored in digital format on some type of electronic medium such as hard drives, flash cards, or optical disks. High-speed analog-to-digital converters and digital-signal-processors are now used to convert voltage or current signals from sensors to digital values. Increasing numbers of cabled observatories extending into the deep ocean through shore stations are now providing high-bandwidth data flow in near real time supported by previously impossible sustained levels of power and storage capacity. As funding for research vessels diminishes and existing fleets continue to age, open-ocean studies are gradually being assumed by satellites, gliders, pop-up drifters, and long-term moorings. The days of limited power supply, insufficient data storage space, and weeks at sea on ships collecting routine survey data may soon be a thing of the past. Ships will still be needed but their role will be more focused on process-related studies and the deployment, servicing, and recovery of oceanographic and meteorological equipment, including sensor packages incorporated in cabled observatory networks. All of these developments are moving physical oceanographers into analysts of what is becoming known as "big data". Faced with large volumes of information, the challenge to oceanographers is deciding

Data Analysis Methods in Physical Oceanography
http://dx.doi.org/10.1016/B978-0-12-387782-6.00001-6

how to approach these mega data and how to select the measurements and numerical simulations that are most relevant to the problems of interest. One of the goals of this book is to provide insight into the analyses of the ever-growing volume of oceanographic data in order to assist the practitioner in deciding where to invest his/her effort.

With the many technological advances taking place, it is important for marine scientists to be aware of both the capabilities and limitations of their sampling equipment. This requires a basic understanding of the sensors, the recording systems and the data processing tools. If these are known and the experiment carefully planned, many problems commonly encountered during the processing stage can be avoided. We cannot overemphasize the need for thoughtful experimental planning and proper calibration of all oceanographic sensors. If instruments are not in near-optimal locations or the researcher is unsure of the values coming out of the machines, then it will be difficult to believe the results gathered in the field. To be truly reliable, instruments should be calibrated on a regular basis at intervals determined by use and the susceptibility of the sensor to drift. More specifically, the output from all oceanic instruments such as thermometers, pressure sensors, dissolved oxygen probes, and fixed pathlength transmissometers drift with time and need to be calibrated before and after each field deployment. For example, the zero point for the Paroscientific Digiquartz (0–10,000 psi) pressure sensors used in the Hawaii Ocean Time-series at station "Aloha" 100 km north of Honolulu drifts about 4 dbar in three years. As a consequence, the sensors are calibrated about every six months against a Paroscientific laboratory standard, which is recalibrated periodically at special calibration facilities in the United States (Lukas, 1994). Even the most reliable platinum thermometers—the backbone of temperature measurement in marine sciences—can drift of order 0.001 °C over a year. Our shipboard experience also shows that opportunistic over-the-side field calibrations during oceanic surveys can be highly valuable to others in the science community regardless of whether the work is specific to one's own research program. As we discuss in the following chapters, there are a number of fundamental requirements to be considered when planning the collection of field records, including such basic considerations as the sampling interval, sampling duration, and sampling location.

It is the purpose of this chapter to review many of the standard instruments and measurement techniques used in physical oceanography in order to provide the reader with a common understanding of both the utility and limitations of the resulting measurements. The discussion is not intended to serve as a detailed "user's manual" nor as an "observer's handbook". Rather, our purpose is to describe the fundamentals of the instruments in order to give some insight into the data they collect. An understanding of the basic observational concepts, and their limitations, is a prerequisite for the development of methods, techniques, and procedures used to analyze and interpret the data that are collected.

Rather than treat each measurement tool individually, we have attempted to group them into generic classes and to limit our discussion to common features of the particular instruments and associated techniques. Specific references to particular company's products and the quotation of manufacturer's engineering specifications have been avoided whenever possible. Instead, we refer to published material addressing the measurement systems or the data recorded by them. Those studies that compare measurements made by similar instruments are particularly valuable. On the other hand, there are companies whose products have become the "gold standard" against which other manufacturers are compared. Reliability and service are critical factors in the choice of any instrument. The emphasis of the instrument review section is to give the reader a background in the collection of data in physical oceanography. For those readers interested in more complete information regarding a specific instrument or measurement technique, we refer to the references at the end of the book where we list the sources of the material quoted. We realize that, in terms of specific measurement systems, and their review, this text will be quickly dated as new and better systems evolve. Still, we hope that the general outline we present for accuracy, precision, and data coverage will serve as a useful guide to the employment of newer instruments and methods.

1.2 BASIC SAMPLING REQUIREMENTS

A primary concern in most observational work is the accuracy of the measurement device, a common performance statistic for the instrument. Absolute accuracy requires frequent instrument calibration to detect and correct for any shifts in behavior. The inconvenience of frequent calibration often causes the scientist to substitute instrument precision as the measurement capability of an instrument. Unlike absolute accuracy, precision is a relative term and simply represents the ability of the instrument to repeat the observation without deviation. Absolute accuracy further requires that the observation be consistent in magnitude with some universally accepted reference standard. In most cases, the user must be satisfied with having good precision and repeatability of the measurement rather than having absolute measurement accuracy. Any instrument that fails to maintain its precision, fails to provide data that can be handled in any meaningful statistical fashion. The best instruments are those that provide both high precision and defensible absolute accuracy. It is sometimes advantageous to measure simultaneously the same variable with more than one reliable instrument. However, if the instruments have the same precision but not the same absolute accuracy, we are reminded of the saying that "a man with two watches does not know the time".

Digital instrument resolution is measured in bits, where a resolution of N bits means that the full range of the sensor is partitioned into 2^N equal segments ($N=1, 2\ldots$). For example, eight-bit resolution means that the specified full-scale range of the sensor, say $V=10$ V, is divided into $2^8=256$ increments, with a bit resolution of $V/256=0.039$ V. Whether the instrument can actually measure to a resolution or accuracy of $V/2^N$ units is another matter. The sensor range can always be divided into an increasing number of smaller increments but eventually one reaches a point where the value of each bit is buried in the noise level of the sensor and is no longer significant.

1.2.1 Sampling Interval

Assuming the instrument selected can produce reliable and useful data, the next highest priority sampling requirement is that the measurements be collected often enough in space and time to resolve the phenomena of interest. For example, in the days when oceanographers were only interested in the mean stratification of the world ocean, water property profiles from discrete-level hydrographic (bottle) casts were adequate to resolve the general vertical density structure. On the other hand, these same discrete-level profiles failed to resolve the detailed structure associated with interleaving and mixing processes, including those associated with thermohaline staircases (salt fingering and diffusive convection), that now are resolved by the rapid vertical sampling provided by modern conductivity-temperature-depth (CTD) probes. The need for higher resolution assumes that the oceanographer has some prior knowledge of the process of interest. Often this prior knowledge has been collected with instruments incapable of resolving the true variability and may, therefore, only be suggested by highly aliased (distorted) data collected using earlier techniques. In addition, laboratory and theoretical studies may provide information on the scales that must be resolved by the measurement system.

For discrete digital data $x(t_i)$ measured at times t_i, the choice of the sampling increment Δt (or Δx in the case of spatial measurements) is the quantity of importance. In essence, we want to sample often enough that we can pick out the highest frequency component of interest in the time series but not oversample so that we fill up the data storage file, use up all the battery power, or become swamped with unnecessary data. In the case of real-time cabled observatories, it is also possible to sample so rapidly (hundreds of times per second) that inserting the essential time stamps in the data string can disrupt the cadence of the record. We might also want to sample at irregular intervals to avoid built-in bias in our sampling scheme. If the sampling interval is too large to resolve higher frequency components, it becomes necessary to suppress these components during sampling using a sensor whose response is limited to frequencies equal to that of the sampling frequency. As we discuss in our section on processing satellite-tracked drifter data, these lessons are often learned too late—after the buoys have been cast adrift in the sea.

The important aspect to keep in mind is that, for a given sampling interval Δt, the highest frequency we can hope to resolve is the *Nyquist* (or *folding*) *frequency*, f_N, defined as

$$f_N = 1/2\Delta t \qquad (1.1)$$

We cannot resolve any higher frequencies than this. For example, if we sample every 10 h, the highest frequency we can hope to see in the data is $f_N = 0.05$ cph (cycles per hour). Equation (1.1) states the obvious—that it takes at least two sampling intervals (or three data points) to resolve a sinusoidal-type oscillation with period $1/f_N$ (Figure 1.1). In practice, we need to contend with noise and sampling errors so that it takes something like three or more sampling increments (i.e., $\geq$ four data points) to accurately determine the highest observable frequency. Thus, f_N is an upper limit. The highest frequency we can resolve for a sampling of $\Delta t = 10$ h in Figure 1.1 is closer to $1/(3\Delta t) \approx 0.033$ cph. (Replacing Δt with Δx in the case of spatial sampling increments allows us to interpret these limitations in terms of the highest wavenumber (*Nyquist wavenumber*) the data are able to resolve.)

An important consequence of Eqn (1.1) is the problem *of aliasing*. In particular, if there is energy at frequencies $f > f_N$—which we obviously cannot resolve because of the Δt we picked—this energy gets folded back into the range of frequencies, $f < f_N$, which we are attempting to resolve (hence, the alternate name "folding frequency" for f_N). This unresolved energy does not disappear but gets redistributed within the frequency range of interest. To make matters worse, the folded-back energy is disguised (or aliased) within frequency components different from those of its origin. We cannot distinguish this folded-back energy from that which actually belongs to the lower frequencies. Thus, we end up with erroneous (aliased) estimates of the spectral energy variance over the resolvable range of frequencies. An example of highly aliased data would be current meter data collected using 13-h sampling in a region dominated by strong semidiurnal (12.42-h period) tidal currents. More will be said on this topic in Chapter 5.

As a general rule, one should plan a measurement program based on the frequencies and wavenumbers (estimated from the corresponding periods and wavelengths) of the parameters of interest over the study domain. This requirement may then dictate the selection of the measurement tool or technique. If the instrument cannot sample rapidly enough to resolve the frequencies of concern it should not be used. It should be emphasized that the Nyquist frequency concept applies to both time and space and the Nyquist wavenumber is a valid means

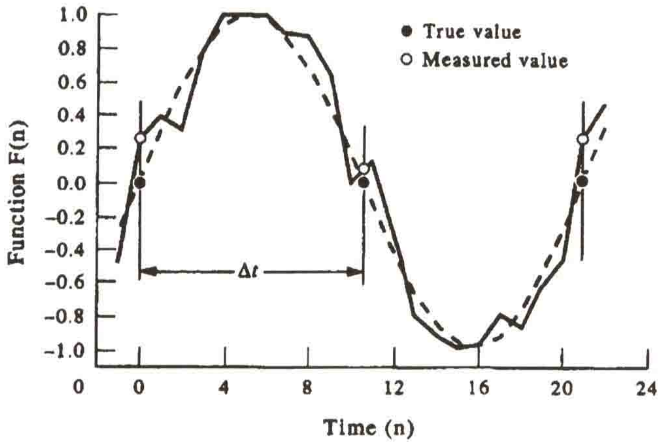


FIGURE 1.1 Plot of the function $F(n) = \sin(2\pi n/20 + \phi)$ where time is given by the integer $n = -1, 0, \ldots, 24$. The period $2\Delta t = 1/f_N$ is 20 units and ϕ is a random phase with a small magnitude in the range ± 0.1 radians. Open circles denote measured points and solid points the curve F(n). Noise makes it necessary to use more than three data values to accurately define the oscillation period.

of determining the fundamental wavelength that must be sampled.

1.2.2 Sampling Duration

The next concern is that one samples long enough to establish a statistically significant determination of the process being studied. For time-series measurements, this amounts to a requirement that the data be collected over a period sufficiently long that repeated cycles of the phenomenon are observed. This also applies to spatial sampling where statistical considerations require a large enough sample to define multiple cycles of the process being studied. Again, the requirement places basic limitations on the instrument selected for use. If the equipment cannot continuously collect the data needed for the length of time required to resolve repeated cycles of the process, it is not well suited to the measurement required.

Consider the duration of the sampling at time step Δt. The longer we make the record the better we are to resolve different frequency components in the data. In the case of spatially separated data, Δx, resolution increases with increased spatial coverage of the data. It is the total record length $T = N\Delta t$ obtained for N data samples that: (1) determines the lowest frequency *(the fundamental frequency)*

$$f_0 = 1/(N\Delta t) = 1/T \tag{1.2}$$

that can be extracted from the time-series record; (2) determines the frequency resolution or minimum difference in frequency $\Delta f = |f_2 - f_1| = 1/(N\Delta t)$ that can be resolved between adjoining frequency components, f_1 and f_2 (Figure 1.2); and (3) determines the amount of band averaging (averaging of adjacent frequency bands) that can be applied to enhance the statistical significance of individual spectral estimates. In Figure 1.2, the two separate waveforms of equal amplitude but different frequency produce a single spectrum. The two frequencies are well resolved for $\Delta f = 2/(N\Delta t)$ and $3/(2N\Delta t)$, just resolved for $\Delta f = 1/(N\Delta t)$, and not resolved for $\Delta f = 1/(2N\Delta t)$.

In theory, we should be able to resolve all frequency components, f, in the frequency range $f_0 \le f \le f_N$, where f_N and f_0 are defined by Eqns (1.1) and (1.2), respectively. Herein lies a classic sampling problem. In order to resolve the

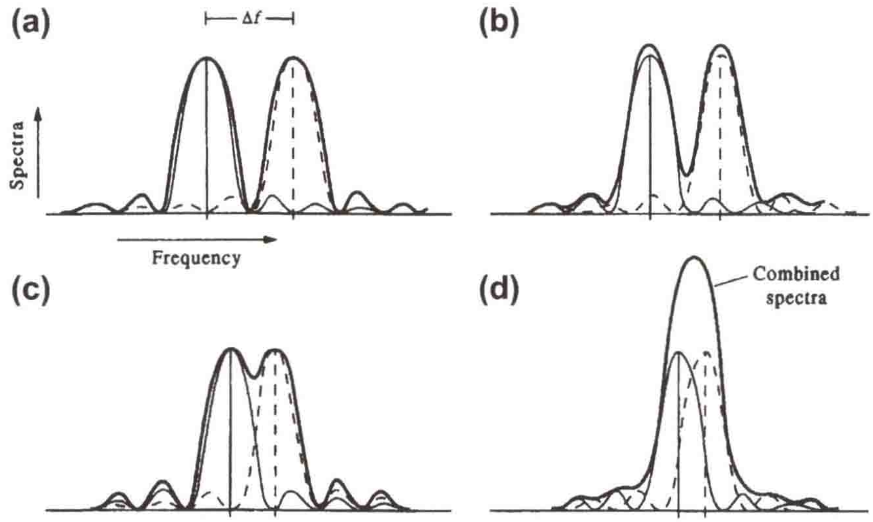


FIGURE 1.2 Spectral peaks of two separate waveforms of equal amplitude and frequencies f_1 and f_2 (dashed and thin line) together with the calculated spectrum (solid line). (a) and (b) are well-resolved spectra; (c) just resolved spectra; and (d) not resolved. Thick solid line is total spectrum for two underlying signals with slightly different peak frequencies.

frequencies of interest in a time series, we need to sample for a long time (T large) so that f_0 covers the low end of the frequency spectrum and Δf is small (frequency resolution is high). At the same time, we would like to sample sufficiently rapidly (Δt small) so that f_N extends beyond all frequency components with significant spectral energy. Unfortunately, the longer and more rapidly we want to sample, the more data we need to collect and store, the more power we need to provide, and the more time, effort, and money we need to put into the sensor design and sampling program.

Our ability to resolve frequency components follows from Rayleigh's criterion for the resolution of adjacent spectral peaks in light shone onto a diffraction grating. It states that two adjacent frequency components are just resolved when the peaks of the spectra are separated by frequency difference $\Delta f = f_0 = 1/(N\Delta t)$ (Figure 1.2). For example, to separate the spectral peak associated with the lunar–solar semidiurnal tidal component M_2 (frequency = 0.08051 cph) from that of the solar semidiurnal tidal component S_2 (0.08333 cph), for which $\Delta f = 0.00282$ cph, it requires $N = 355$ data points at a sampling interval $\Delta t = 1$ h or $N = 71$ data points at $\Delta t = 5$ h. Similarly, a total of 328 data values at 1-h sampling are needed to separate the two main diurnal constituents K_1 and O_1 ($\Delta f = 0.00305$ cph). Note that since f_N is the highest frequency we can measure and f_0 is the limit of our frequency resolution, then

$$f_N/f_0 = (1/2\Delta t)/(1/N\Delta t) = N/2 \qquad (1.3)$$

is the maximum number of Fourier components we can hope to estimate in any analysis.

1.2.3 Sampling Accuracy

According to the two previous sections, we need to sample long enough and often enough if we hope to resolve the range of scales of interest in the variables we are measuring. It is intuitively obvious that we also need to sample as accurately as possible—with the degree of recording accuracy determined by the response characteristics of the sensors, the number of