AD-A258 550

MENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1992	THESIS/XXXXXXXXXXX

4. TITLE AND SUBTITLE
Toward an Optimal Strategy for Sampling Chaotic Time Series Using the Samle-Williams Dynamical System

5. FUNDING NUMBERS

6. AUTHOR(S)
Mark Anthony Sellers, Captain

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFIT Student Attending: Pennsylvania State University

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/CI/CIA- 92-111

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFIT/CI
Wright-Patterson AFB OH 45433-6583

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release IAW 190-1
Distribution Unlimited
ERNEST A. HAYGOOD, Captain, USAF
Executive Officer

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS	15. NUMBER OF PAGES
	94
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

The Pennsylvania State University

The Graduate School

Department of Meteorology

TOWARD AN OPTIMAL STRATEGY FOR SAMPLING CHAOTIC TIME SERIES USING THE SMALE-WILLIAMS DYNAMICAL SYSTEM

A Thesis in

Meteorology

by

Mark Anthony Sellers

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 1992

I grant The Pennsylvania State University the nonexclusive right to use this work for the University's own purposes and to make single copies of the work available to the public on a not-for-profit basis if copies are not otherwise available.

Mark Anthony Sellers

We approve the thesis of Mark Anthony Sellers.

Date of Signature

Hampton N. Shirer — 26 Aug 92

Hampton N. Shirer

Associate Professor of Meteorology

Thesis Adviser

Robert Wells — 26 Aug 92

Robert Wells

Professor of Mathematics

Dennis W. Thomson — 26 August 92

Dennis W. Thomson

Professor of Meteorology

Head of the Department of Meteorology

ABSTRACT

Chaotic dynamical systems such as the atmosphere can be characterized by their associated geometrical structures in phase space that are known as strange attractors. Time series data obtained from these dynamical systems must be sampled properly in order to estimate the fractal dimensions of these strange attractors with the greatest possible accuracy. The Smale-Williams attractor is used here because the value of the correlation dimension for this attractor can be determined analytically. Two types of sampling strategies are considered to determine how a time series might be best used to estimate the characteristics of the attractor.

This attractor can be characterized by either a histogram of distances from the origin to each point on the attractor or by an estimate of its correlation dimension. Differences between histograms of a sufficiently long control series and subsets of this series are used to quantify the accuracy of representations of the attractor. Differences between analytic and estimated values of the correlation dimension are also used. Finally, the independence of the time series data is quantified in a new way and the relationship between the independence of sampled data and the accuracy of characterizations of the attractor by histograms of sampled data or by estimates of the correlation dimension is examined.

Data are sampled from a control series according to different strategies and the accuracies of the characterizations of the attractor by histograms and by estimated

correlation dimensions are assessed. Increasingly accurate characterizations of the attractor by histograms are generally obtained by increasing the number of points sampled rather than by varying the sampling strategy. If only a small number of points is sampled, however, then an optimal sampling strategy can be found that yields the most accurate possible characterization of the attractor. This optimal strategy is somewhat subjective and entails sampling groups of consecutive points that are separated by large gaps of unsampled data.

A correlation between the accuracy of characterizations of the attractor by histograms and the independence of sampled data is suggested by some, but not all, cases. In contrast, the accuracy of estimates of the correlation dimension of the attractor and the independence of sampled data are not found to be correlated; however, much smaller sample sizes are used than can be used to construct histograms. All the distances between sampled points must be determined to estimate the correlation dimension and this number of distances is proportional to the square of the number of points used. However, only the distances between sampled points and a single point must be determined to construct a histogram. It is therefore much more difficult to determine whether a relationship exists between the independence of sampled data and the accuracy of estimates of its correlation dimension. Consequently, no correlation between the accuracy of histogram representations and the accuracy of correlation dimension estimates is observed.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF TABLES

ACKNOWLEDGMENTS

An undertaking such as this could never be accomplished without the support and assistance of many individuals. I have been extremely fortunate in having been surrounded by some of the brightest and most giving people with whom a person could ever hope to work. I take great pleasure in acknowledging them here.

Major Mary E. Smith and her staff at the Air Force Institute of Technology Civilian Institutions Program were always most supportive in providing assistance whenever I needed it, in keeping the amount of administrative effort required of me to a minimum, and most importantly, in footing the bill for this project. My Air Force comrades in the meteorology department, in particular Captains Gregory Gust and David Valler and Major Mark Weadon, provided an *esprit de corps* that made this assignment the most enjoyable of my career. Major John Lanicci's pep talks did much to keep us all afloat through some of the more demanding episodes of this assignment and Captain Jeff Doran kept me buoyed up at the beginning of this project when I could not even spell FORTRAN, let alone program in it.

The greatest single fortune of my academic career has been the opportunity to work under the tutelage of Dr. Hampton N. Shirer and Dr. Robert Wells, the two giants on whose shoulders I have had the privilege of standing for the past two years. The experience of working with these gentlemen has been both edifying and liberating. Not only have I been offered a deeper understanding of the complexity and intricacy of the

global atmosphere, but I have also been provided a glimpse of a universe whose existence is not constrained to predetermined pedestrian motion in three integer dimensions. Dr. Shirer's ability to clarify the complicated details of the dynamics of atmospheric flow is perfectly complemented by Dr. Well's ability to render aspects of topology and statistics intelligible to the mathematically unwashed. As if revolutionizing my world view was not enough, Drs. Shirer and Wells also managed to make a demanding course of study a great deal of fun. I cannot remember a single time when they did not meet advances with great enthusiasm and setbacks with patience and assurance. One accomplishment to which I can never look forward will be thanking them enough for this experience.

The greatest single fortune of my life is to have met my wife, Sue. Her support through this endeavor has always been both strong and compassionate. She shared my excitement during moments of success and never let me become discouraged during times of frustration. By undertaking the task of learning about chaos and fractals, she was able not only to keep me motivated, but honest as well. Her endless fascination with the subject made it possible to share this academic experience, and thereby make it twice as enriching. I will never be able to thank her enough either, but at least I have my whole life to try.

Chapter 1

INTRODUCTION

Irregularity is an essential characteristic of many natural hydrodynamic flows, and also of many mathematical solutions to models of these flows. This irregularity appears in highly stressed (high Reynolds number) flows and is characterized by three fundamental properties: aperiodicity, instability, and the existence of strange attractors (Lorenz 1984). The states of irregular flows do not repeat themselves exactly over time, and it is this aperiodicity that most readily characterizes an irregular, or chaotic, flow. Much of the irregularity of atmospheric flow originates from the fact that future states of a flow depend on previous states in an unstable manner (Tsonis and Elsner 1989). Essentially, forcing adds energy to the flow while dissipation removes it, thereby ensuring that the energy cannot grow without bound. This competition between forcing and dissipation leads to fluctuations in the energy that are chaotic owing to the strongly nonlinear interactions inherent in the flow. Irregularity is therefore part of the intrinsic dynamics of the flow and is not caused by unpredictable outside influences (Baker and Gollub 1990).

1.1. Instability, Aperiodicity, and Strange Attractors

The first two fundamental properties of an irregular flow, aperiodicity and instability, characterize natural flows on the broadest range of scales, from the development of localized turbulence over seconds to the evolution of global climate systems over eons. If these flows were periodic and stable, then deriving from a time series of observations the equation that models a flow would be a relatively simple matter. It would be necessary only to sample certain variables of the flow, such as the velocity or temperature, at an adequately high rate to determine their amplitudes and the durations of their periods. These periods could then be incorporated into fairly simple linear models of the flow, and these models could be trusted to predict accurately future states of the flow. However, aperiodicity and irregularity render such a simple approach to modeling susceptible to erroneous predictions.

The irregularity inherent in the solution to a model of a hydrodynamic flow was first demonstrated by Lorenz (1963). In this seminal work, aperiodic solutions were obtained from numerical integrations of a deterministic system: a 3-component model of shallow cellular convection. Lorenz observed that typically two solutions to this model diverged from each other, even though they nearly coincided initially. This type of instability in the solution, which has come to be known as sensitive dependence on initial conditions, is the most fundamental property of irregular dynamical systems and does not originate from roundoff or truncation error; it is inherent in any approximation (Lorenz

1984). It is because of this property, and because models can never be initialized beyond a certain precision or certain temporal or spatial resolution, that future states of a natural hydrodynamic flow, such as the atmosphere, can never be forecast for periods longer than a certain bound because errors are guaranteed to grow and so to eventually overwhelm the results.

The third fundamental property of an irregular and transient-free flow is the existence of a strange attractor. Strange attractors are geometrical structures, constructed from time series data, that represent a flow in a phase space of coordinates given by the variables that describe the state of the flow, rather than in an Euclidean space given by the variables for ordinary physical space and time (Lorenz 1984). For all periodic and aperiodic flows, each point in the phase space represents the state of the flow at a given moment, and so a trajectory that connects successive points represents the evolution of the system (Tsonis and Elsner 1989). Depending on the nature of the dynamical system, different types of trajectories are possible. In stable regimes, solutions to the model will evolve either to a single state, which can be represented by a fixed point in phase space, or to an exactly periodic solution, which can be represented by a closed curve or closed orbit in phase space (Higgins 1987).

If the solution is aperiodic but bounded, then its trajectory will be closed only after an infinite amount of time because the solution repeats only after an infinite amount of time. The energy of a flow is constrained by the competition between forcing and dissipation, so states of the flow described by individual points in phase space are

restricted to a certain volume in that space. The set of points revisited at any scale by this infinitely long trajectory is known as a strange attractor (Tsonis and Elsner 1989). In the case of an atmospheric model, this finite region of space occupied by a strange attractor may be said to represent the climate of the atmosphere. The trajectory in phase space of an aperiodic flow cannot cross itself, because doing so would mean that the flow had returned to a previous state and could not therefore be aperiodic. The infinitely long trajectory usually does not exist in phase space as a topological manifold, but rather exists as a fractal set (Tsonis and Elsner 1989).

A strange attractor can be characterized by its fractal dimension. The creation of the fractal structure of a strange attractor may be imagined in terms of the stretching and folding of a volume in phase space by the forcing and dissipation of the system (Nese 1987). This "stretching and folding" gives the fine leaf-like structure of the fractal. For the calculation of the fractal dimension of a strange attractor from a sample of data to give the best possible estimate, it seems reasonable that the data should be sampled in such a way as to capture this fine structure. Fractal dimensions (noninteger values) have been sought for a variety of hydrodynamic systems as evidence for the existence of strange attractors. For at least one system, the global climate, the existence and value of a fractal dimension is debatable. Several authors (e.g. Nicolis and Nicolis 1984, Fraedrich 1987, Essex *et al.* 1987) have claimed its existence and calculated its value while others have failed to find such a dimension (e.g. Grassberger 1986, Krishna-Mohan *et al.* 1989). The latter two have cited deficiencies in the data used as arising from the manner in which the

data were sampled. This controversy underscores the necessity for developing appropriate sampling strategies.

1.2. Dimensions of Attractors of Modeled and Natural Systems

Different fractal dimensions of attractors generated by a variety of nonlinear models, laboratory flows, and historical data series have been calculated. For example, the value of the correlation dimension of the Lorenz shallow convection model has been estimated to be 2.06 (Sparrow 1982), while dimensions for the Lorenz system and 7- and 11-component spectral models have been found by Nese *et al.* (1987). Brandstäter *et al.* (1983) have determined the value of the correlation dimension for a turbulent Couette-Taylor laboratory flow to be between 4 and 5. The correlation dimension estimated for the flow within a rotating annulus in the regime of geostrophic turbulence was determined by Guckenheimer and Buzyna (1983) to be between 7 and 12. Henderson and Wells (1988) used 500-mb height index values (differences between the zonally averaged 500-mb heights at 60 °N and 30 °N) and found a value of the correlation dimension for this series to be approximately 6.5. The same authors also estimated the value of the correlation dimension of a time series of gust front vertical velocities to be approximately 4.0. Hense (1987) examined rainfall amounts and sea surface temperature data from the Pacific Ocean and found an attractor dimension for the Southern Oscillation to be between 2.5 and 6.0.

Values for correlation dimensions have also been calculated from climate data, although these values are not entirely unassailable. The value of the correlation dimension for oxygen isotope ratios from the equatorial Pacific deep sea core V28-238 for the past one to two million years has been calculated by Nicolis and Nicolis (1984) to be approximately 3.1. Because of this result, they conclude that only 4 independent variables would be required to model the global climate. Their estimate of this dimension is based on 500 data points separated by 2,000 yr. The raw data of this core consists of only 184 points, and the remainder of their data is interpolated from the raw data. Fraedrich (1987) also concluded that a climate attractor exists and that the approximate lower bound of the correlation dimension of this attractor is between 4.4 and 4.8. His estimate is based on 182 values of eastern equatorial Atlantic oxygen isotope ratios that extend over 775,000 years. Essex *et al.* (1987) examined 12,084 values of 500-mb geopotential height from 1946 to 1982 and found an attractor dimension of approximately 6.0. They claim this supports the existence of a climate attractor.

The results of Nicolis and Nicolis (1984) have been questioned by Grassberger (1986) who used the same V28-238 core data as Nicolis and Nicolis (1984), but only used 230 points separated by 5,000 yr in the calculation. Although data from the same core were used, Grassberger did not find any such indication that a climate attractor exists. To substantiate this conclusion, data from a second core, V28-239 were also used in an attempt to find a dimension for a climate attractor, and again none was found. An examination of tree-ring data covering the last 7,100 yr also failed to reveal any sign that

such a climatic attractor exists. Grassberger (1986) interprets the results of Nicolis and Nicolis (1984) as being due to the highly correlated nature of the interpolated data they used in their calculation. Krishna-Mohan *et al.* (1989) also examined the V28-238 core data and found no indication that a climatic attractor exists. They conclude that the data set used by Nicolis and Nicolis (1984) was too noisy and was not sampled frequently enough to draw such a conclusion.

The results of Grassberger (1986) and Krishna-Mohan *et al.* (1989) underscore the importance of choosing the proper sampling rate and strategy in selecting the data that can be used to make a meaningful estimate of the correlation dimension. The correlation of interpolated data used in the calculation of Nicolis and Nicolis (1984) enabled them to conclude that a climate attractor exists, while Grassberger found no indication for this attractor. Krishna-Mohan *et al.* concluded that the sampling rate used to calculate the dimension for a climate attractor was too low. Essex *et al.* (1987) caution that their results may not be globally valid because the data sampled were restricted to values from eastern North America. The short time scale (37 years) they used may also cast doubt on their conclusion that a climate attractor exists.

1.3. Sampling and Estimates of Attractor Dimensions

As is evident from the controversy surrounding the possible existence of a climate attractor, one of the central problems in quantifying an attractor, or deciding whether it

even exists, is to develop sampling strategies that optimize the quantity and quality of the data used in the estimation. Therefore, when sampling time series data from which a dimension will be calculated, a sampling strategy should be used that ensures that samples are uncorrelated but sampled in sufficient number. Choosing uncorrelated data may be necessary to guarantee that points do not contribute redundant information, and using large enough sample sizes may be necessary to capture as much detail of the attractor as is possible.

Calculations of dimension estimates are numerically intensive, and so there is an interest in keeping the number of points used as low as possible without sacrificing the quality of the estimate. Estimates of the dimensions of fractal structures entail the computation of distances between points. The number of distances increases as the square of the number of points, so the penalty for increasing the number of points is severe. Conversely, enough points must be sampled to estimate a dimension accurately. These two conflicting requirements can be ameliorated by a sampling strategy that optimizes the quality of the data used.

If the number of points to be used in estimating a dimension is restricted by the availability of data or by computational practicalities, then the pattern of sampling that should be used to approximate the attractor must optimize the estimate. For a periodic solution, sampling at a constant rate is sufficient for modeling the flow. It does not necessarily follow from this observation, however, that periodic sampling will be the best approach for reconstructing a strange attractor. If a flow is aperiodic, then many scales of

motion exist within the flow. Periodic sampling may capture the details of only certain scales while missing the details of flows that exist on scales related to higher frequencies than those that can be captured at the sampling rate used. If the number of points used is constrained for any reason, then the only alternative is to use a sampling strategy other than sampling at a simple fixed rate. The leaf-like structure of strange attractors suggests how these strategies can be developed. The points sampled should be distributed among the leaves so that each leaf is represented in as much detail as the fixed number of points will allow. Because the leaves are distributed within the attractor in a fractal pattern, the optimal sampling strategy must be related somehow to this fractal structure. The points are sampled from a *time* series, while the attractor represents this time series in a phase *space*. A temporal sampling strategy must be developed so as to capture this fractal spatial structure. To do so, the sampled points must be uncorrelated; by choosing uncorrelated points, as many leaves as possible will be represented because each point will represent a different leaf.

Having accurate estimates of attractor dimensions is necessary for a variety of reasons. If the dimension is estimated accurately and confidently, then the existence of a strange attractor for the hydrodynamic system may be accepted. Although the magnitude of this dimension may not be directly correlated with the number of independent variables necessary to model the system, its magnitude does give some indication of the relative complexity or simplicity of the system. The predictability of the system can be determined from its Lyapunov exponents, which can be estimated from the same sample of points as

those used to estimate dimension. Consequently, an accurate estimate of the dimension can be used to certify the quality of the data used to calculate the Lyapunov exponents, thereby leading to the most accurate estimates of the predictability of the system. The purpose of the research discussed in this thesis is to gain some insight into finding optimal strategies for sampling data used to estimate a fractal dimension. A simple dynamical system will be studied, the correlation dimension of which can be determined analytically and which can be used as a basis for assessing the accuracy of the dimension estimate and of the quality of the sampled data. It is hoped that the strategies developed and compared, if not directly applicable to other dynamical systems, will at least provide a basis for developing optimal strategies for other systems.

Chapter 2

METHODS OF ANALYSIS

The overall objective of this research is to determine the best approach for sampling a time series that approximates a chaotic attractor. It is not possible to generate the infinitely long series of points that determines an attractor; therefore, a sufficiently long control series of points must be generated to approximate the attractor. Subsets of points can be sampled from this control series using a variety of strategies, and estimates of the value of the correlation dimension ν of the attractor can be calculated from these subsets. If the dimension of the attractor can be known analytically, then values of the dimension calculated from the sampled subset can be compared with this known value. The magnitude of the correlation dimension error between that for the sampled data and the known value can be used to compare the effectiveness of different sampling strategies. To be useful in this study, other characterizations of model output, such as the dependence of samples and frequency distributions of distances between points representing the attractor, must also vary sensitively with sampling strategy.

An attractor whose correlation dimension can be determined analytically is the Smale-Williams hyperbolic attractor (Smale 1977). This is a discrete dynamical system that therefore has the additional advantage of not requiring the use of derivative

approximation techniques such as those required for solving systems of differential equations. Although it is not a simple model of a meteorological process such as convection or global circulation (e.g., Lorenz 1963, 1984), the Smale-Williams attractor satisfies all the above sensitivity requirements.

2.1. The Smale-Williams Attractor

The Smale-Williams attractor, or solenoid, is described by Smale (1977) and is illustrated schematically in Figure 2.1. This dynamical system is a map Φ that carries each point (x, y) on any transverse disc of an initial solid torus onto a point $(\hat{x}, \hat{y})$ on a thinner and longer embedded image torus by winding the initial torus an integral number n times around its center hole. (In Figure 2.1 this winding number n is 2, and below the case of winding number 3 is considered.) Because of this multiple winding, the intersection of each cross sectional transverse disc of the original torus and the image torus consists of n new disjoint discs, each with a radius equal to the radius of the initial disc shrunk by a factor β. Applying the map twice in succession to the original torus produces a new torus winding n times within the image torus and so n^2 times within the original torus. Repeating this procedure produces a family of nested tori, each winding n times within its predecessor.

With each iteration, the map Φ stretches the torus by a factor equal to n in one direction while contracting its radius by β. This stretching and contraction of the torus by

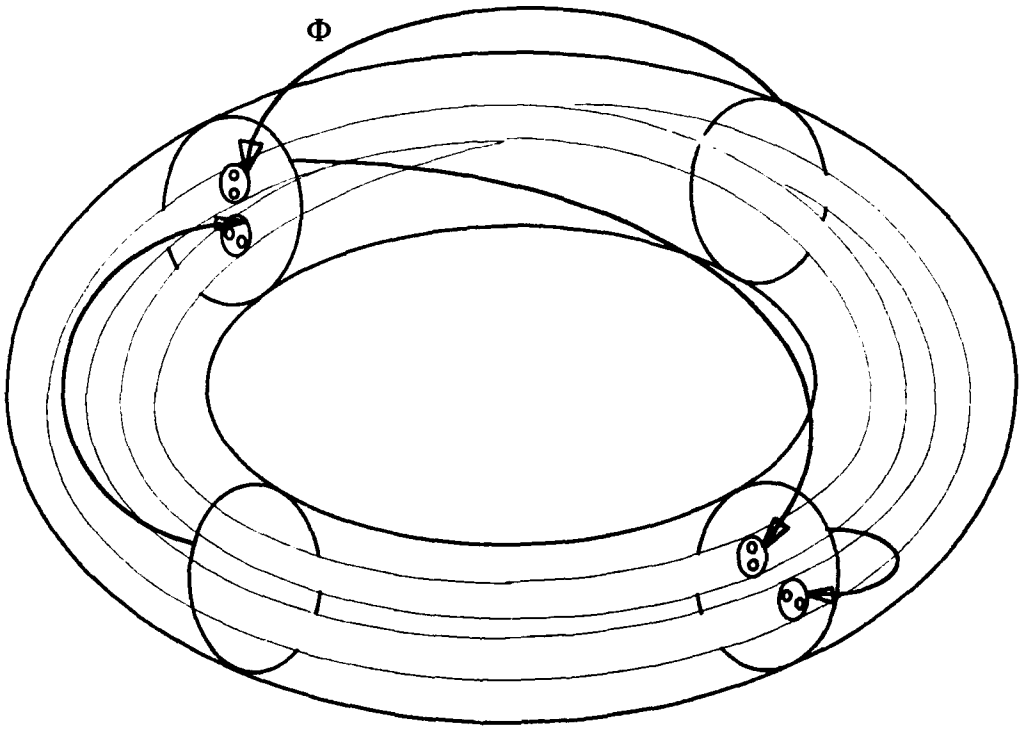


Figure 2.1. Schematic depiction of the Smale-Williams attractor showing four transverse discs of the initial torus and cross sectional discs resulting from two iterations of the map Φ. The lighter lines inside the initial torus show the overall effect by the map of winding that torus into an image torus. Curved arrows depict the first mapping of individual points on the transverse discs of the initial torus to points on the image torus.

winding leads to the expectation that in the limit the structure of the cross sectional discs should be fractal. The successive image tori may be regarded as higher and higher resolution representations of this attractor. Thus, the cross sections produced by this multiple winding leads to Cantor dust embedded in the corresponding transverse disc. This Cantor dust is unlike the standard Cantor set in that it does not lie in a single line, or even in a single smooth curve in the transverse disc. This Cantor dust is embedded in 2 dimensions although its intrinsic dimension is not 2. This intrinsic dimension is found in Section 2.2.

2.1.1. The Smale-Williams Attractor in Euclidean Space

The Smale Williams dynamical system can be identified with a standard torus embedded in Euclidean X-Y-Z space; that is, values of x, y, and θ from the map described below in Section 2.1.2 can be used to parameterize the standard torus. Figure 2.2 illustrates the use of these values for a point P in the construction of a standard torus. The standard torus is obtained by rotating a disc about the Z axis, where the disc is perpendicular to the X-Y plane and is at an angle θ from the X-axis. The center C of the disc is at a distance R from the origin O of the torus and has coordinates $(R\cos\theta, R\sin\theta, 0)$. The disc has radius r, with $0 < r < R$. An x'-y' coordinate system on the disc can be specified and the values of x and y can be used to parameterize this disc.

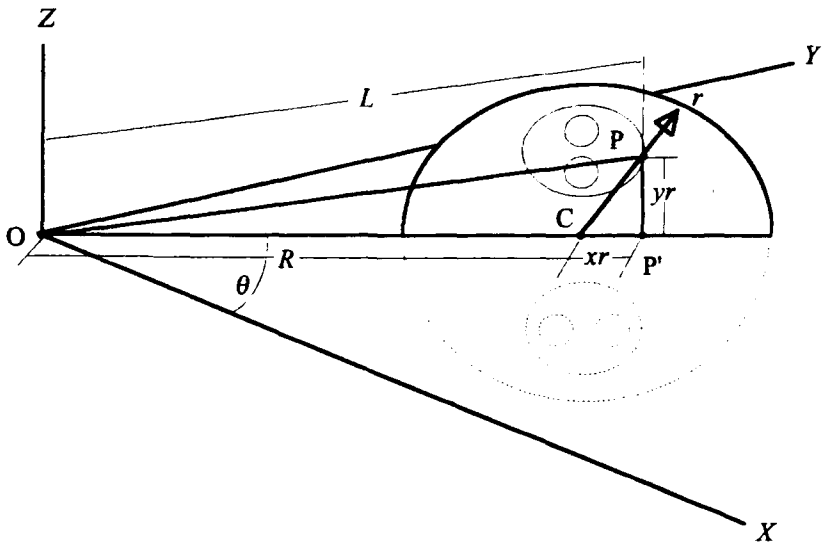


Figure 2.2. The coordinate system representing the Smale-Williams attractor on a standard torus. Here L, R, xr, yr, and θ are used to calculate the distance between point P on the standard torus and the center of the torus.

Because the inequality $x'^2 + y'^2 \le r^2$ holds, point P on the disc will have the coordinates $(x', y') = (xr, yr)$ with $x^2 + y^2 \le 1$. As can be seen in Figure 2.2, the projection P' onto the X-Y plane of point P is at a distance $R + xr$ from the origin of the torus, and the point P is at a distance yr above the X-Y plane. The coordinates of P in the Euclidean space are therefore given by $((R+xr)\cos\theta, (R+xr)\sin\theta, yr)$. The distance L of P from the origin is the hypotenuse of a right triangle whose length is therefore

$$L = \sqrt{(R+xr)^2 \cos^2\theta + (R+xr)^2 \sin^2\theta + (yr)^2}. \qquad (2.1.1)$$

This equation can be simplified to

$$L = \sqrt{(R + xr)^2 + y^2 r^2}\,. \qquad (2.1.2)$$

Histograms of this distance are discussed in Section 2.3 as providing a means for quantifying the attractor structure.

2.1.2. The Map Functions of the Smale-Williams Attractor

Any point on a disc can also be defined using coordinates (u, z), which are related to (x, y, θ) as shown below, subject to the conditions that u and z are complex, that $|u| = 1$, and that $|z| \leq 1$. The particular case of the map that is considered has the form

$$\Phi{:}(u, z) \mapsto (\hat{u}, \hat{z}) \qquad (2.1.3)$$

where $\hat{u} = u^3$ and $\hat{z} = \alpha u + \beta z$, under the constraints that $0 < \alpha$, $0 < \beta$, and $\alpha + \beta < 1$. Because $|u| = 1$ holds, mapping u to u^3 changes the complex value of this coordinate without changing its magnitude. The version of this attractor described above in Section 2.1.1 and in Smale (1977) uses a winding number of 2. Owing to the binary processing of data by the computer, this version of the attractor spuriously approaches a fixed point. Choosing a winding number of 3 yields a chaotic time series because the cube of a value can only be approximated numerically. Mapping z into $\alpha u + \beta z$, together with the constraints $0 < \alpha$, $0 < \beta$, and $\alpha + \beta < 1$, contracts the initial value of z by a factor β while adding a translational term.

The complex values of u and u^3 can be rewritten using Euler's relationship as

$$u = \cos\theta + i\sin\theta \tag{2.1.4}$$
$$u^3 = \cos\hat{\theta} + i\sin\hat{\theta} \tag{2.1.5}$$

where $\hat{\theta} = 3\theta$. The substitution of $3\theta - 2\pi n'$, where n' is an integer, for $\hat{\theta}$ yields the same values, as follows from application of the trigonometric identities $\cos\hat{\theta} = \cos\left(\hat{\theta} - 2\pi n'\right)$ and $\sin\hat{\theta} = \sin\left(\hat{\theta} - 2\pi n'\right)$. The appropriate value of n' here is given by $n' = \left[\hat{\theta}/2\pi\right]$, where the brackets denote the largest integer part. This trigonometric conversion of u can be further parameterized by allowing $\theta = 2\pi t$ and $\hat{\theta} = 2\pi\hat{t}$; thus (2.1.2) and (2.1.3) can be rewritten in terms of t and $\hat{t} = 3t$ as

$$u = \cos(2\pi t) + i\sin(2\pi t) \tag{2.1.6}$$
$$u^3 = \cos\left(2\pi\left(\hat{t} - \left[\hat{t}\right]\right)\right) + i\sin\left(2\pi\left(\hat{t} - \left[\hat{t}\right]\right)\right). \tag{2.1.7}$$

Therefore, if $\hat{u} = u_{n+1} = u_n^3$ then from (2.1.6) and (2.1.7) it is sufficient, although not necessary, to require that

$$t_{n+1} = 3t_n - \left[3t_n\right]. \tag{2.1.8}$$

The complex value of z can be expressed as $z = x + iy$, and so $\hat{z} = z_{n+1} = \alpha u + \beta z$ can be written using (2.1.4) as

$$\hat{z} = \alpha e^{i\theta} + \beta z = \alpha(\cos\theta + i\sin\theta) + \beta(x + iy). \tag{2.1.9}$$

If (2.1.9) is separated into real and imaginary components, then $\hat{z} = \hat{x} + i\hat{y}$ may be expressed in terms of x, y, and θ via

$$\hat{x} = \alpha\cos\theta + \beta x \tag{2.1.10}$$
$$\hat{y} = \alpha\sin\theta + \beta y. \tag{2.1.11}$$

It follows from (2.1.6), (2.1.8) and (2.1.9) that the map defined by setting $u_{n+1} = u_n^3$ and $z_{n+1} = \alpha u_n + \beta z_n$ may be represented by the equations

$$\begin{aligned} t_{n+1} &= 3t_n - [3t_n] \\ x_{n+1} &= \alpha\cos(2\pi t_n) + \beta x_n, \\ y_{n+1} &= \alpha\sin(2\pi t_n) + \beta y_n \end{aligned} \tag{2.1.12}$$

where $t = t_n$, $\theta = 2\pi t_n$, and $\hat{t} = t_{n+1}$; $x = x_n$ and $\hat{x} = x_{n+1}$; and $y = y_n$ and $\hat{y} = y_{n+1}$. It may appear that because the original two variables are complex, the final form of these functions should have produced a map in a 4-dimensional space. The constraint that $|u| = 1$ restricts its domain to the 1-dimensional domain of a circle, and so this map exists in a 3-dimensional space.

Values of x and y from (2.1.12) are shown in Figure 2.3 for $\alpha = 0.45$ and $\beta = 0.3$. Pairs of x and y are plotted without regard to their respective values of t, and so this figure represents a projection of all transverse discs of the torus onto a single cross section.

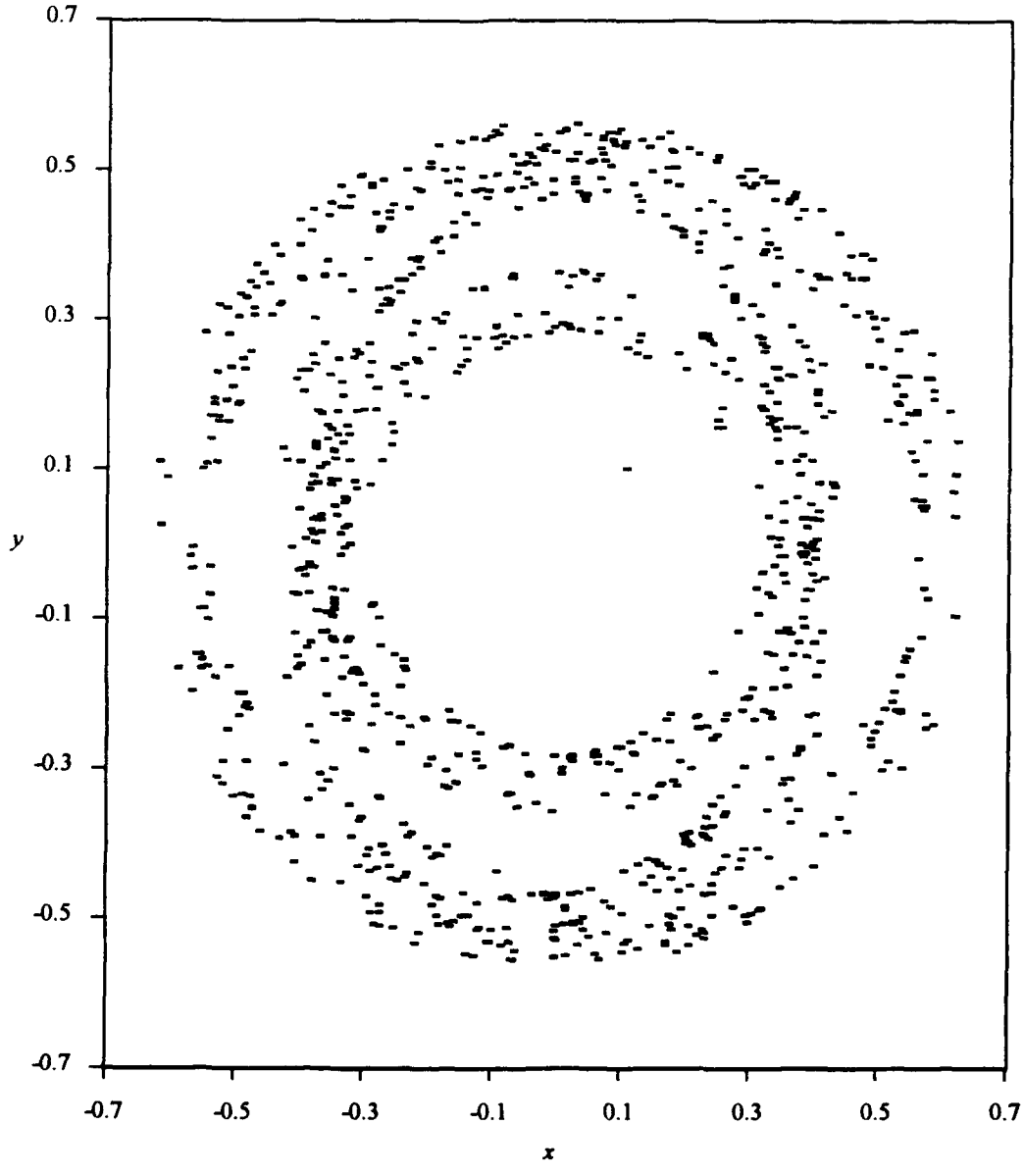


Figure 2.3. Values of x and y for 1000 points of the Smale-Williams attractor. Values of t are not considered in the selection of these points. The values of x and y shown represent points on 1000 different transverse discs of 1000 iterations of the map projected on to a single disc rather than points on a single transverse disc.

2.2. The Correlation Dimension of the Smale-Williams Attractor

The correlation dimension ν of the Smale-Williams attractor can be shown to be equal to its box-counting dimension (Mori and Fujisaka 1980). A heuristic approach can be employed to calculate the box-counting dimension d_B of this attractor. This argument is illustrated in Figure 2.4, and begins with a box $B(\varepsilon)$ that has sides parallel to the x, y and t axes and has length ε along each side. After one iteration $\Phi B(\varepsilon)$ of the map for which the winding number is 3, this box is stretched to a length 3ε along the t axis and contracted to lengths $\beta\varepsilon$ along the other two axes. The new box $\Phi B(\varepsilon)$ can be covered (approximately) with k boxes, each having sides $\beta\varepsilon$. The value of k is given by

$$k = \frac{3\beta^2\varepsilon^3}{(\beta\varepsilon)^3} = \frac{3}{\beta}. \qquad (2.2.1)$$

A measure of the mass density μ of points on the attractor contained by the boxes $B(\varepsilon)$ and $\Phi B(\varepsilon)$ can be related to d_B in the following manner:

$$\mu(B(\varepsilon)) = \varepsilon^{d_B} \qquad (2.2.2)$$

$$\mu(\varphi B(\varepsilon)) = k(\beta\varepsilon)^{d_B}. \qquad (2.2.3)$$

The number of points contained in $\Phi B(\varepsilon)$ is the same as that contained in $B(\varepsilon)$, and so $\varepsilon^{d_B} = k\beta^{d_B}\varepsilon^{d_B}$, or $1 = k\beta^{d_B}$. The value of d_B is found by substituting $3/\beta$ for k in

$1 = k\beta^{d_B}$. Taking logarithms of this result yields the analytical value of the box-counting dimension d_B of the Smale-Williams attractor, namely

$$d_B = 1 + \ln 3 / \ln(1/\beta). \tag{2.2.4}$$

The expression for the Hausdorff dimension d_H of the Smale-Williams attractor is determined by Mori and Fujisaka (1980) to be equivalent to (2.2.4). Falconer (1990) also shows that $d_B = d_H$ for this attractor.

The Smale-Williams attractor is a hyperbolic dynamical system and so it admits an invariant and ergodic mass distribution μ (Smale 1977). Therefore, it follows from elementary ergodic theory that the correlation dimension of a small subregion of the

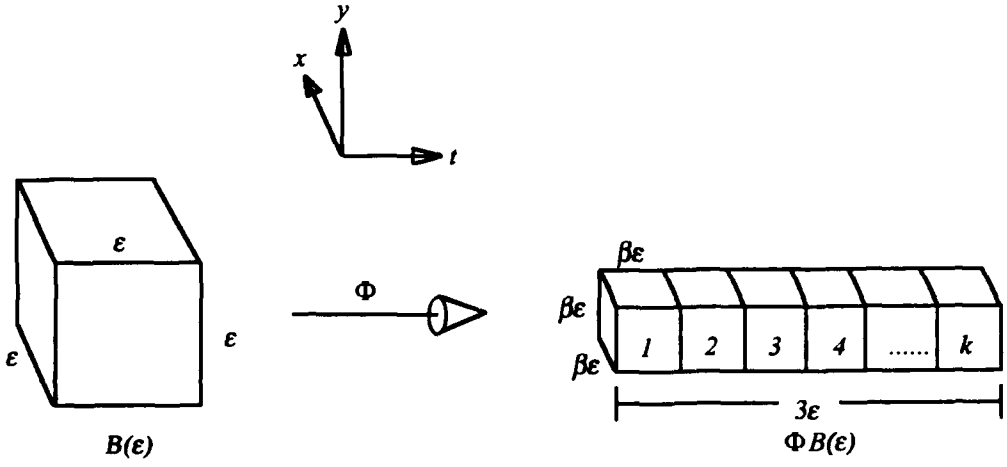


Figure 2.4. The change of a volume $B(\varepsilon)$ covering a portion of the Smale-Williams attractor after one application of the map Φ. The map takes this portion of the initial torus and stretches it to 3 times its length while contracting its cross sectional dimensions by β. The volume $\Phi B(\varepsilon)$ required to cover the points after one application of the map is given by k boxes of volume $(\beta\varepsilon)^3$.

attractor, *i.e.*, the local correlation dimension $\nu(x)$, is essentially constant with respect to mass density (Petersen 1971). If $\nu(x)$ is essentially constant, then it is equal to the Hausdorff dimension d_H (Falconer 1990) and is also equal to the correlation dimension ν (Pesin 1992). Clearly, because $\nu(x)$ is essentially constant, it follows for the Smale-Williams attractor that

$$d_B = d_H = \nu = 1 + \ln 3/\ln(1/\beta). \tag{2.2.5}$$

This expression can be used to determine a standard value against which estimates of the correlation dimension from a set of points representing the attractor can be assessed.

2.3. Mean Absolute Difference

A simple and inexpensive way of quantifying the structure of an attractor is to construct a histogram of distances between points on the attractor and a single specified point in the phase space (Doran 1991). The number of distances in individual intervals or bins of the histogram varies as the time series is lengthened or the sampling strategy is changed. Thus, differences in sampling strategies may be related to differences between histograms constructed from these data and the histogram of distances obtained from the control series.

The torus can be reconstructed using the values of x, y, and θ once scaling values of R and r are chosen so that the attractor can be visualized most clearly (Section 2.1.1).

Thus, values for R and r were specified as 4.0 and 1.0, respectively. Values for L (2.1.2) were then distributed into 128 separate bins, and their histograms plotted. The numbers of distances in each histogram of sampled data and the data of the control series were then normalized so that the total number of distances in all histograms would be equal and so that differences between histograms of different sample sizes could be compared. In this case the mean absolute difference D_{ab} between two histograms H_a and H_b is defined by

$$D_{ab} = \frac{1}{128}\sum_{i=1}^{128}\left[\left(H_a(L_i) - H_b(L_i)\right)^2\right]^{1/2} \qquad (2.3.1)$$

where $H_a(L_i)$ and $H_b(L_i)$ are the normalized numbers of distances L_i in each bin i of the two histograms. The histogram for a control series of 1,048,576 points is illustrated in Figure 2.5. Mean absolute differences D_{ab} between two histograms H_a and H_b have been used successfully to quantify differences in the structures of subsets of points from attractors (Doran 1991).

2.4. Independence of Sampled Data

In order to extract information most efficiently from a sampled data set, it is necessary that the individual sampled points be independent of the other sampled points. The rate at which points following a given point in a time series decorrelate therefore

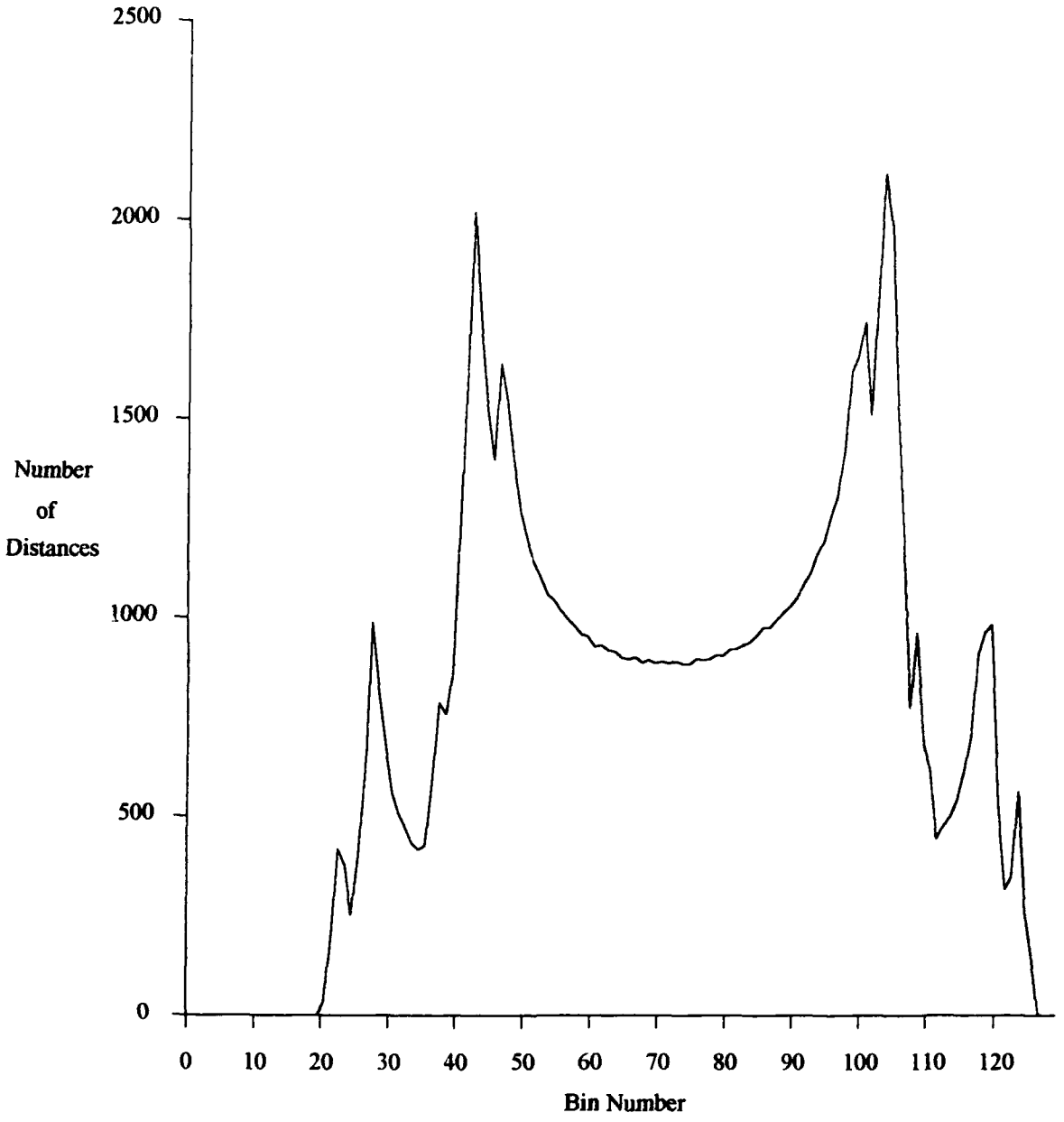


Figure 2.5. Histogram of 1,048,576 distances L (2.1.2) for the Smale-Williams attractor. The total number of distances in the histogram is normalized to 131,072, so that each distance on the ordinate actually represents eight distances in the torus.

influences the possible ways in which data may be sampled. If decorrelation takes a long time, then samples must be taken at longer intervals than if they require only a few iterations to decorrelate. Autocorrelation would be one method for determining how many time steps between retained points must occur before sampled data are independent of each other. Here is presented an alternative method for obtaining the most efficient sampling strategy.

The sampling method proposed here is based on the independence of groups of sampled model output from which histograms are constructed. This method is a much more expedient way of determining the independence of sampled data because the independence of groups of data from other groups is determined rather than the independence of sampled points from other sampled points. Fewer comparisons must be made because a smaller number of groups than of points is used. This grouped data can be sampled according to different strategies. The number of consecutive points within subsections, the gaps between points within subsections, and the gaps between subsections can be varied in any pattern to define individual strategies. Many more sampling strategies may be devised and tested if the method of assessing their efficiency is easily implemented.

A time series of distances $L_1, L_2, \cdots, L_N$ can be constructed for the N points of a sufficiently long control series using (2.1.2). This control series is divided into m subseries of I/m points each, and histograms of x bins are formed from each subseries. These subseries of I/m points are chosen according to different sampling strategies and comparisons of their associated histograms with each other can be used to assess each

strategy. For a given strategy, the histograms of sampled model data from each of the m non-overlapping subseries of the control series may be constructed where

$$\begin{aligned} H_1(x) &\equiv \text{Histogram of } L_1, L_2, \cdots, L_I \\ H_2(x) &\equiv \text{Histogram of } L_{G+1}, L_{G+2}, \cdots, L_{G+I} \\ &\vdots \\ H_m(x) &\equiv \text{Histogram of } L_{(m-1)G+1}, L_{(m-1)G+2}, \cdots, L_{(m-1)G+I} \end{aligned} \quad . \qquad (2.4.1)$$

In (2.4.1) G represents the gap between the initial points of the subseries of the control series and is the same for each sampling strategy. A sequence $S(m)$ of distances composed of all values of L_i used to create the histograms defined in (2.4.1) can be used to produce the average, or control, histogram of all the sampled data. An equivalent way to find this average histogram $H_{S(m)}(x)$ is simply to average the histograms H_i via

$$H_{S(m)}(x) \equiv \frac{1}{m}\sum_{i=1}^{m} H_i(x). \qquad (2.4.2)$$

This mean $H_{S(m)}(x)$ can be compared with the histogram $H_{S(m-1)}(x)$ using an integral form of the definition of D_{ab} (2.3.1)

$$D_{ab} = \int_0^1 \left(H_{S(m)}(x) - H_{S(m-1)}(x)\right)^2 dx. \qquad (2.4.3)$$

If the integrand of (2.4.3) is expanded, then this expression for D_{ab} becomes

$$D_{ab} = \left(\frac{1}{m(m-1)}\right)^2 \left[\int_0^1 \left\{\left(H_1(x) - H_{S(m)}(x)\right)^2 + \cdots + \left(H_{m-1}(x) - H_{S(m)}(x)\right)^2\right\} dx\right] + \frac{1}{m^2}\int_0^1 \left(H_m(x) - H_{S(m)}(x)\right)^2 dx \tag{2.4.4}$$

only if the cross terms disappear, i.e.

$$\int_0^1 \left(H_i(x) - H_{S(m)}(x)\right)\left(H_j(x) - H_{S(m)}(x)\right) dx = 0. \tag{2.4.5}$$

for all i, j such that $0 \le i < j \le m$. This latter condition (2.4.5) will be valid if the samples are independent. This condition is anticipated in the limit owing to the ergodic mixing that occurs in the Smale-Williams attractor (Smale 1977); however, this vanishing should never be expected to occur in practice. This expression can be used to define a dependence measure $|\xi|$, as is discussed below. Although this quantity would not be equal to zero, its value would be small and may be taken as a measure of the dependence of the subsets of sampled data. The larger the magnitude of ξ is, the more dependent the samples may be said to be, and so $|\xi|$ may be used as an index of dependence.

The mean squared difference between two samples will be small if the samples are large enough. To determine how large a sample needs to be, *a priori* error bounds can be

set, such that if D_{ab} is within these error bounds then the sampling can be said to be adequate. The variance of the sampled data in the m subsections can be used to establish these error bounds. Clearly, this variance is related to the independence of sampled data: the greater the independence of the sampled data, the smaller the variance.

The variance $V_{S(m)}(x)$ of the m histograms constructed from the data within each of the subsections of the control series is defined by

$$V_{S(m)}(x) = \frac{1}{m}\left[\left(H_1(x) - H_{S(m)}(x)\right)^2 + \cdots + \left(H_m(x) - H_{S(m)}(x)\right)^2\right] \qquad (2.4.6)$$

and the total variance is defined by

$$V_m^I = \int_0^1 V_{S(m)}(x)dx \qquad (2.4.7)$$

The total variance can be substituted into the first term of the right side of (2.4.4) to obtain an expression for D_{ab}:

$$D_{ab} = \frac{1}{m(m-1)^2}V_m^I + \left(\frac{1}{m^2} - \left(\frac{1}{m(m-1)}\right)^2\right)\int_0^1 \left(H_m(x) - H_{S(m)}(x)\right)^2 dx. \qquad (2.4.8)$$

From (2.4.6) it is apparent that

$$mV_{S(m)}(x) = \left[\left(H_1(x) - H_{S(m)}(x)\right)^2 + \cdots + \left(H_m(x) - H_{S(m)}(x)\right)^2\right] \geq \left(H_m(x) - H_{S(m)}(x)\right)^2 \qquad (2.4.9)$$

From this inequality and the definition of the total variance (2.4.7), the mean squared difference can be expressed as an inequality by substitution of these into (2.4.8):

$$D_{ab} \le \frac{1}{m(m-1)^2}\int_0^1 V_{S(m)}(x)dx + m\left[\frac{1}{m^2} - \left(\frac{1}{m(m-1)}\right)^2\right]\int_0^1 V_{S(m)}(x)dx \qquad (2.4.10)$$

Integration and cancellation of like coefficients in (2.4.10) leads to:

$$D_{ab}^2 \le \frac{1}{m} V_m^I \qquad (2.4.12)$$

From (2.4.11) it is apparent that if $m \to \infty$ then $D_{ab} \to 0$ if the samples are independent and if V_m^I is bounded. This argument is valid only if the data from which the histograms are constructed are independent and the cross terms vanish, *i.e.*, if (2.4.5) holds. Because this can never be expected to be strictly true, some measure of dependence is necessary to determine the degree to which (2.4.12) is valid. Minimizing the dependence measure suggested by (2.4.5) can reasonably be expected to minimize the mean squared difference.

The above argument can be extended to four histograms of samples i, j, k, and l via a proposed analogous expression for the dependence ξ

$$\int_0^1 \left[\left(H_i(x) - H_{S(m)}(x)\right) - \left(H_j(x) - H_{S(m)}(x)\right)\right]\left[\left(H_k(x) - H_{S(m)}(x)\right) - \left(H_l(x) - H_{S(m)}(x)\right)\right]dx = \xi$$

(2.4.12)

where $i \neq j \neq k \neq l$. This expression can be simplified and used to define the dependence measure ξ, defined by

$$\xi = \int_0^1 \big(H_i(x) - H_j(x)\big)\big(H_k(x) - H_l(x)\big)dx. \qquad (2.4.13)$$

Equation (2.4.13) provides a method for determining the dependence of data within samples by comparing all combinations of four subsets of the sampled data from all the subsets. The magnitude of ξ determines dependence and so $|\xi|$ is used. If the data are independent then $\xi = 0$, and they are nearly so when $|\xi|$ is small. No knowledge of the true histogram is now necessary.

2.5. Correlation Dimension Estimations from Sampled Data

The correlation dimension ν takes into account the number of pairwise correlations of points on the attractor with other points on the attractor that fall within a sphere of radius ε. The number of pairwise correlations of points within a distance ε of each other is specified by the correlation function $C(\varepsilon)$ (Grassberger and Procaccia 1983):

$$C(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i,j=1}^{N} \Theta\big(\varepsilon - |\mathbf{x}_i - \mathbf{x}_j|\big). \qquad (2.5.1)$$

In this expression, $|\mathbf{x}_i - \mathbf{x}_j|$ denotes a generic distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ on the attractor. The specific distance measure used to calculate the correlation function for the Smale-Williams attractor is defined below in Chapter 3. In (2.5.1) Θ is the Heaviside function given by $\Theta(y) = 1$ if $y \geq 0$ and $\Theta(y) = 0$ if $y < 0$. The 3-element column vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ specify the positions of the i^{th} and j^{th} attractor points in x-y-t space. Thus the correlation function yields the number N of distances between points separated by a distance less than ε, normalized by the total number of points included in the set that is used to represent the attractor. Values of ε up to the maximum distance between any two points on the attractor must be considered to determine the correlation function. The correlation dimension ν is assumed to be related exponentially to the correlation function $C(\varepsilon)$ via

$$C(\varepsilon) \propto \varepsilon^{\nu}. \tag{2.5.2}$$

The traditional approach for determining the correlation dimension is to plot the natural logarithm of $C(\varepsilon)$ against the natural logarithm of ε and then to estimate the slope of this line. As an alternative to this method, the following approach is used in this study. An arbitrary function $F(\varepsilon)$ can be included to express (2.5.2) as an equality

$$C(\varepsilon) = F(\varepsilon)\varepsilon^{\nu}. \tag{2.5.3}$$

Taking logarithms of this expression gives

$$\ln F(\varepsilon) = \ln C(\varepsilon) - \nu \ln(\varepsilon). \tag{2.5.4}$$

This expression can be used in either of two ways. If the correlation dimension is known analytically, then its value can be used in (2.5.4) to solve for $\ln F(\varepsilon)$. The value of $\ln F(\varepsilon)$ should be a constant equal to $\ln F[\varepsilon_1, \varepsilon_2]$ over the closed interval $[\varepsilon_1, \varepsilon_2]$, and so standard deviations of $\ln F(\varepsilon)$ over the interval $[\varepsilon_1, \varepsilon_2]$ for different characterizations by correlation dimension estimates can be used as a basis for comparing those characterizations. Alternatively, the value of v can be adjusted so as to minimize the variation of values of $\ln F(\varepsilon)$ within the interval $[\varepsilon_1, \varepsilon_2]$, thereby providing a way of estimating the correlation dimension if it is not known analytically. Although the second method is considered briefly, in this study the first approach is used because the analytical value of v is known (Section 2.2).

2.6. Higher Moments of the Correlation Function

A second method for estimating the value of the correlation dimension is described by Wells *et al.* (1992). This method incorporates higher moments p of the correlation function as independent sources of information. In this approach, higher moments of distances between points r that are within a distance ε of other points are normalized and summed to define the moment function $M(p, \varepsilon, N_T)$, given by

$$M(p, \varepsilon, N_T) = \frac{1}{N(\varepsilon)} \sum \left\{ \left(\frac{r_{i,j}}{\varepsilon} \right)^p \middle| r_{i,j} \le \varepsilon \text{ and } i, j \le N_T \right\} \qquad (2.6.1)$$

where N_T is the total number of points sampled and $N(\varepsilon)$ is defined as the number of distances $r_{i,j} = |\mathbf{x}_i - \mathbf{x}_j|$ between points $\mathbf{x}_i$ and $\mathbf{x}_j$ that are less than ε. The moment function $M(p, \varepsilon, N_T)$ can be used to determine the correlation dimension from the relation

$$\nu = \lim_{\varepsilon \to 0} \lim_{N_T \to \infty} \frac{pM(p, \varepsilon, N_T)}{1 - M(p, \varepsilon, N_T)}. \tag{2.6.2}$$

Values for ν calculated using (2.6.2) (without taking limits) are then plotted against values of ε. The forms of these curves are compared with the analytical value of the correlation dimension for data sets obtained using different sampling strategies.

Chapter 3

QUANTITATIVE AND QUALITATIVE SAMPLING EXPERIMENTS

A given set of points from an attractor may represent that attractor with greater or lesser accuracy. This accuracy depends on two general considerations: the number of points that are used and the manner in which these points are sampled from a control series of points. To assess how these two considerations affect the accuracy of samplings of the Smale-Williams attractor, the results of quantitative and qualitative sampling experiments are presented. To determine the improvement in accuracy with increasing sample size, mean absolute differences D_{ab} between normalized histograms for series of different lengths are examined. The dependence measure $|\xi|$ for sampled subseries of points within a control series is then considered as a means for evaluating different sampling strategies. Two general types of sampling strategies are examined. In Type I, the number of points to be sampled from the time series is fixed, but the distribution of windows within the time series from which the points are sampled is varied to create different strategies. In Type II, the distribution of windows within the time series remains fixed, while the number of points taken from each window is varied. Mean absolute differences D_{ab} between histograms of the sampled data and the control series are determined and then compared with the dependence measure $|\xi|$ for the sampled data.

3.1. Series Length and Mean Absolute Difference

The accuracy of the representation of an attractor by a series of points increases with the number of points. The fact that the mean absolute difference D_{ab} between histograms of lengths N and $N+\Delta N$, where ΔN is a fixed increment of points, decreases with increasing N can be used to determine the number of points needed to characterize the attractor accurately (Doran 1991). An initial experiment was designed to determine for the Smale-Williams attractor an appropriate number of points to be used to construct a control histogram for sampling strategy experiments, as well as to identify the nature of the rate of decrease of D_{ab} with increasing N. To determine this rate of decrease, time series of values of x, y, and t were generated for series having as many as five million points. For this portion of the study, the values of α and β were set to 0.45 and 0.30, respectively. All calculations were performed using double precision on an IBM 370 mainframe computer. The same initial conditions were used in all cases, namely $x_0 = 0.1$, $y_0 = 0.1$, and $t_0 = 0.2$. The first million points of these series were always discarded to ensure that any transient signal present in the series is negligible. The series used therefore ranged in length from $N = 50{,}000$ to $N = 4{,}000{,}000$ points in 50,000-point intervals. For each series, histograms based on the distance function L (2.1.2) were constructed and then normalized so that the distribution among bins is preserved and the area of the histogram is that which would be given by a histogram of 100,000 distances. Mean absolute differences D_{ab} between these normalized histograms were next calculated

according to (2.3.1). Values of ΔN from 50,000 to 1,000,000 in 50,000 point increments were used to determine the mean absolute difference D_{ab} between series of lengths N and $N + \Delta N$. The relationship between the magnitude of D_{ab} and the varying series length N, with ΔN held constant, is illustrated by plotting D_{ab} as a function of N. The resulting family of curves for all values of ΔN is shown in Figure 3.1.

To ascertain the rate at which D_{ab} decreases with increasing series length, linear regressions of D_{ab} on series length N were performed for each separation interval ΔN. Doran (1991) found that variations of D_{ab} with increasing N for the Lorenz attractor were best modeled by an exponential decrease of the form

$$D_{ab} = A\exp(-RN) + B. \tag{3.1.1}$$

For the Lorenz attractor he found that D_{ab} decreased to a positive floor value B rather than to zero as might be expected. An alternative model of the rate decrease would be a power law in which the D_{ab} between successive histograms is proportional to the series length raised to a negative exponent $-R$ whose magnitude R is greater than unity, *i.e.*,

$$D_{ab} = AN^{-R} + B, \quad R > 1 \tag{3.1.2}$$

In order to determine whether an exponential or a power law relation better models the rate of decrease of D_{ab}, both log-normal and log-log linear regressions of the data were

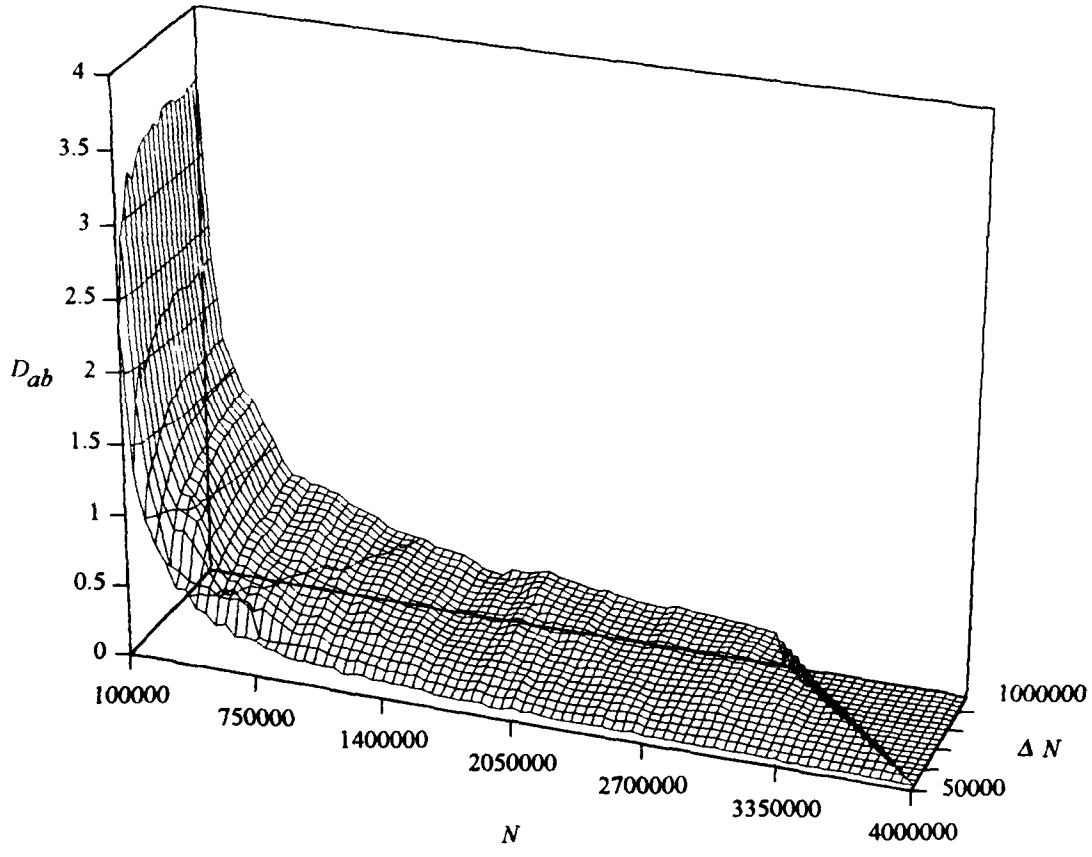


Figure 3.1. Mean absolute difference D_{ab} as a function of series length N and separation interval ΔN. Individual curves are for each separation interval ΔN, from 50,000 points (nearest curve) to one million points (farthest curve), increasing by 50,000 points for each curve. The sharp drop-off on the right side of the figure is artificial and caused by the lack of data for computing differences.

performed. The values of the slopes R, intercepts $\ln A$, and regression coefficients r^2 for the results of the most successful linear regressions are presented in Table 3.1.

Values of the regression coefficient r^2 can be used to determine which of the two models better accounts for the decrease of mean absolute differences D_{ab} with increasing values of N. For both the exponential model and the power law model, the best values of r^2 were obtained when the floor value B vanished. The range of values of r^2 for the exponential model increases from 0.81671 to 0.84535 with increasing values of ΔN. Results obtained for the power law model account for much more of the variation of D_{ab} with series length as indicated by consistently higher values of r^2, but these values of r^2 decrease from 0.99261 to 0.93445 as values of ΔN increase. Increasing the value of ΔN corresponds to *decreasing* the number of pairs of values of D_{ab} and N available for the calculation of regression statistics. For the case of $\Delta N = 50{,}000$, 40 pairs of values of D_{ab} and N are available, while for the case of $\Delta N = 1{,}000{,}000$, only 21 such pairs are available. Increasing the number of points in the regression increases the value of r^2 for the power law model, but decreases r^2 for the exponential model. Nevertheless for the fewest pairs of values, the value of r^2 for the power law model is still significantly greater than the value of r^2 for the exponential model. On this statistical basis, the power law model of the decrease of mean absolute difference D_{ab} with series length N is accepted for the Smale-Williams attractor.

It is easy to argue that the power law model is superior to an exponential model for small values of N. For a power law model in which the floor value B is zero, one

Table 3.1. Regression statistics for the power law (3.1.1) and exponential (3.1.2) models of the rate at which the mean absolute difference D_{ab} decreases with increasing series length N. The results presented in this table are for the most successful regressions for both models, those obtained for the case of $B = 0$. The number of pairs of values of D_{ab} and N are indicated in the far left column.

		Exponential $\ln D_{ab} = -RN + \ln A$			Power Law $\ln D_{ab} = -R \ln N + \ln A$		
Pairs	Interval ΔN	R $(\times 10^{-7})$	$\ln A$	Correlation Coefficient r^2	R	$\ln A$	Correlation Coefficient r^2
40	50000	6.89336	-0.77477	0.81671	1.04236	12.70119	0.99261
39	100000	6.68548	-0.47032	0.81777	1.06414	13.38043	0.98118
38	150000	6.60951	-0.26432	0.82611	1.10022	14.13660	0.98624
37	200000	6.50738	-0.13157	0.83057	1.12303	14.63363	0.98236
36	250000	6.39487	-0.03653	0.83227	1.14087	15.02271	0.97848
35	300000	6.33700	0.05036	0.83325	1.16469	15.47756	0.97372
34	350000	6.29164	0.12921	0.83331	1.18829	15.91817	0.96827
33	400000	6.23555	0.19690	0.83136	1.20883	16.30556	0.96252
32	450000	6.21451	0.26348	0.83306	1.23283	16.73387	0.95866
31	500000	6.18019	0.32251	0.83231	1.25380	17.11373	0.95373
30	550000	6.17865	0.38490	0.83335	1.27922	17.55397	0.94946
29	600000	6.19206	0.44771	0.83322	1.30771	18.03579	0.94521
28	650000	6.17908	0.50140	0.83244	1.33067	18.43439	0.94154
27	700000	6.17958	0.55570	0.83373	1.35546	18.85695	0.93981
26	750000	6.17373	0.60671	0.83526	1.37826	19.24890	0.93851
25	800000	6.19058	0.66125	0.83826	1.40531	19.70099	0.93853
24	850000	6.22235	0.71709	0.84011	1.43532	20.19421	0.93724
23	900000	6.26200	0.77423	0.84249	1.46685	20.70909	0.93650
22	950000	6.29008	0.82613	0.84448	1.49577	21.18358	0.93444
21	1000000	6.31363	0.87597	0.84535	1.52449	21.65354	0.93445

would expect by (3.1.2) that $D_{ab} \to \infty$ as $N \to 0$ and that $D_{ab} \to 0$ as $N \to \infty$. Both of these expectations can be explained heuristically in terms of the amount of information that is available in the time series to describe the attractor. If only $N = 1$ distance L is available, then infinitely more is known about the attractor than if $N = 0$ and so the mean absolute difference D_{ab} between the histograms of these two cases would be infinite. Increasing the number of available distances L from $N = 1$ to $N = 2$ doubles the amount of available information and would result in a finite nonnegative value for D_{ab}. Increasing the number of distances from $N = 2$ to $N = 3$ would increase the amount of information by 50% and would again reduce the decrease of the mean absolute difference. Finally, as $N \to \infty$, $N \approx N + 1$ and so $D_{ab} \to 0$ because the amount of information in the two histograms would be virtually the same. In contrast, the expectation that $D_{ab} \to 0$ as $N \to \infty$ would also follow from an exponential model, but the expectation that $D_{ab} \to \infty$ as $N \to 0$ would not. By (3.1.1), $D_{ab} = A + B$ for $N = 0$, and $A + B \neq \infty$. The heuristic argument that the mean absolute difference D_{ab} between the histogram for $N = 0$ and a histogram for some $N > 0$ would be infinite is contradicted by the exponential model.

Linear regressions of the values of the intercept $\ln A$ and the values of the rate of decrease R as functions of the increment values ΔN were also performed. Significant correlations between the values of both $\ln A$ and R and the value of ΔN were found. The relationship between the intercept $\ln A$ and ΔN is shown in Figure 3.2. For this linear relationship, the slope is 8.941×10^{-6} and the intercept is 12.6556. The rate of decrease

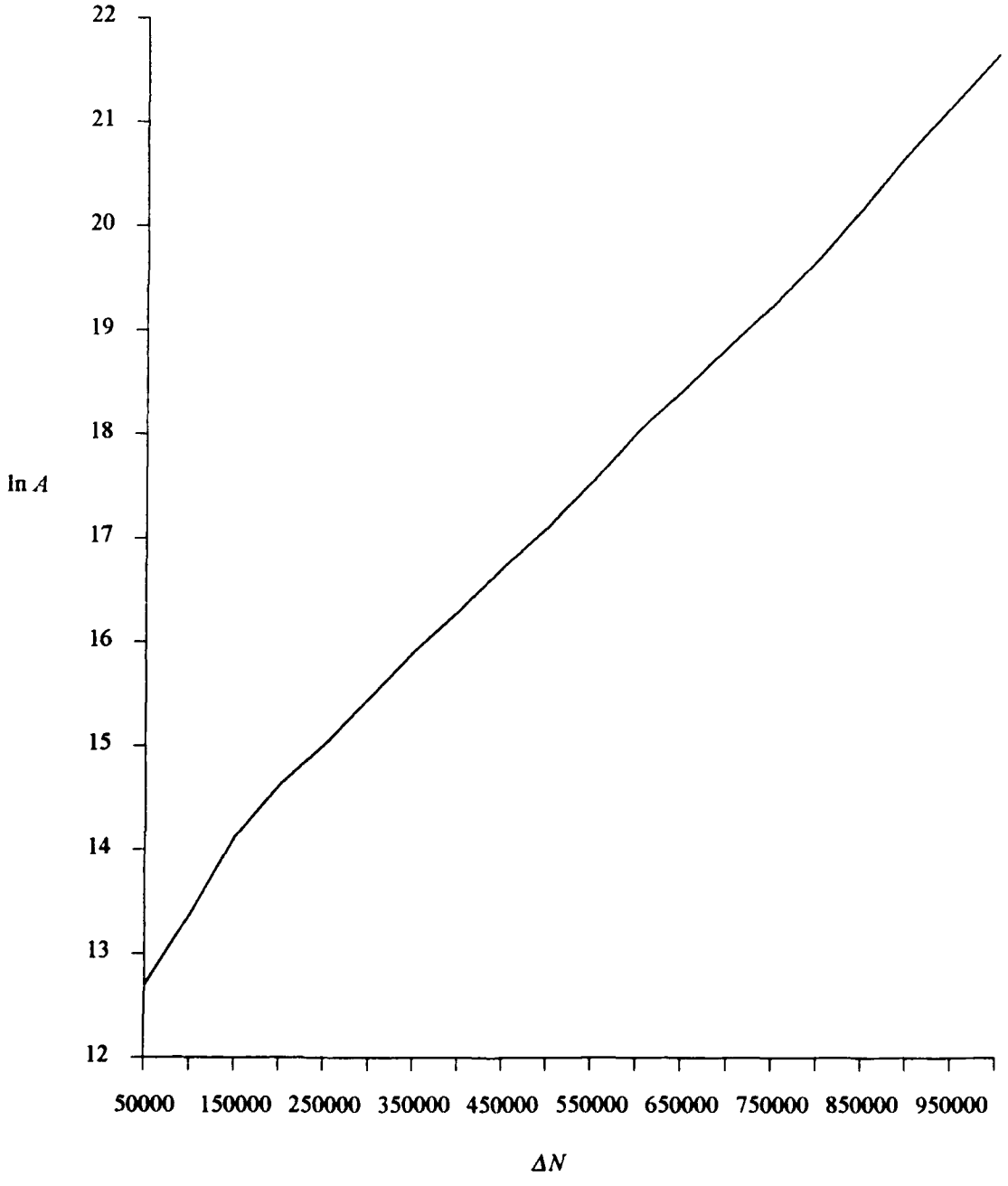


Figure 3.2. The linear relationship between the constant $\ln A$ and the increment ΔN separating series of lengths N and $N + \Delta N$ for which mean absolute differences D_{ab} were determined. The value of the correlation coefficient r^2 for this relationship is 0.99768.

R as a function of ΔN is shown in Figure 3.3. The slope of this line is -4.9500×10^{-7} and the intercept is -1.01503. The correlation coefficient r^2 for the regression of $\ln A$ on ΔN is 0.99768 and for the regression of R on ΔN it is 0.99740. The results of these analyses of the relationships between intercepts, rates of decrease, and increment length can be incorporated into the power law model (3.1.2) as

$$D_{ab} = \frac{\exp\left(\left(8.941 \times 10^{-6}\right)\Delta N + 12.6556\right)}{N^{\left(\left(4.9500 \times 10^{-7}\right)\Delta N + 1.01503\right)}} \tag{3.1.3}$$

where $A = \exp\left(\left(8.941 \times 10^{-6}\right)\Delta N + 12.6556\right)$, $R = \left(\left(4.95 \times 10^{-7}\right)\Delta N + 1.01503\right) > 1$, and $B = 0$ and where $N \le 4{,}000{,}000$ and $\Delta N \le 1{,}000{,}000$.

The interval ΔN represents the difference in the lengths of the time series for which mean absolute differences D_{ab} between histograms are determined, and the effect of increasing ΔN is to increase the magnitude of the coefficient A and to increase the magnitude of the rate of decrease R. If the value of ΔN is large then the representation of the attractor by $N + \Delta N$ points should be much better than that provided by N points, and D_{ab} will be greater than it would be for smaller values of ΔN. For a given value of N, mean absolute differences D_{ab} between histograms will be greater for larger ΔN, so the coefficient A should be greater. Consequently for any specific value of N, a more rapid decrease in the mean absolute difference D_{ab} is observed for larger values of ΔN because all curves converge on the same floor value $B = 0$. The constant 1.01503 in the

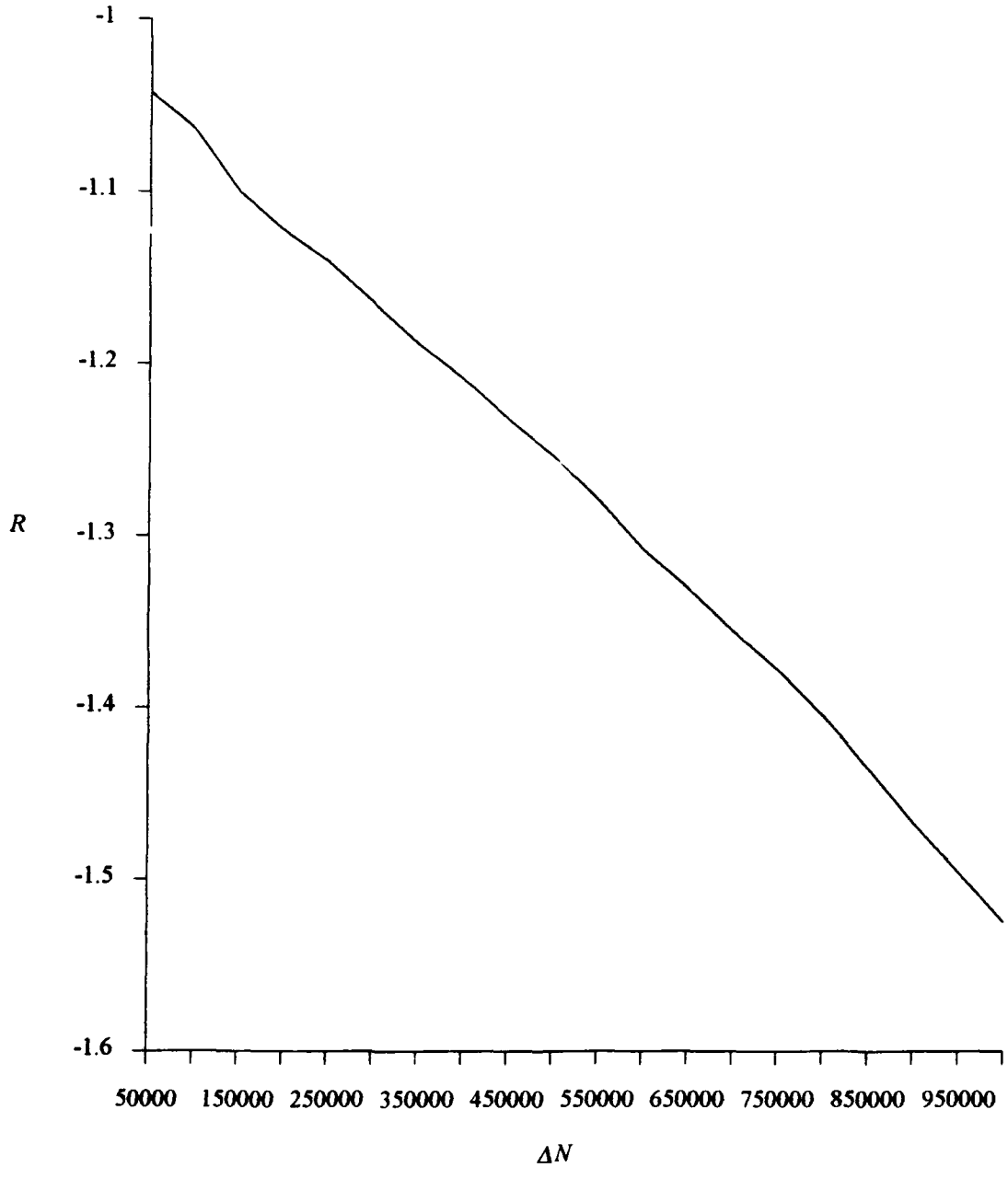


Figure 3.3. The linear relationship between the magnitude of the rate of decrease R and the increment ΔN separating series of lengths N and $N + \Delta N$ for which mean absolute differences D_{ab} were determined. The value of the correlation coefficient r^2 for this relationship is 0.99740.

expression for the rate of decrease ensures that $R > 1$, so this model predicts the expected decrease in D_{ab} with increasing values of N and $N + \Delta N$.

The power law relationship (3.1.3) between D_{ab} and N can be used to determine the proper length of a time series of data needed to approximate the attractor with a prescribed accuracy. If a power law relationship is established for a given value of ΔN and for a given level of statistical confidence, then D_{ab} between two histograms of lengths N and $N + \Delta N$ can be determined for any value of N, not just those used to create the curves. If a tolerable value of D_{ab} is established and if ΔN is fixed, then the value of N can be determined from (3.1.3) without having to generate several time series and calculating D_{ab}. This approach is used in section 3.2 to determine the appropriate length of a control series of data against which to assess the effectiveness of different sampling strategies. This value of N is of order 10^6.

3.2. Sampling Strategy and Independence of Samples

In this series of experiments, the relationship between the sampling strategy and the dependence measure $|\xi|$ of a subseries of sampled points is examined following the arguments presented in Section 2.4. In these trials, the number of points of the control series as well as the numbers I of sampled points are all powers of 2. In each case, a series of 2,097,152 (2^{21}) points was generated, from which the first 1,048,576 (2^{20}) points were discarded, leaving a control series of $N = 1{,}048{,}576$ points from which to

sample specified numbers of points according to different strategies. As before, the initial conditions for the time series were $x_0 = 0.1$, $y_0 = 0.1$, and $t_0 = 0.2$, and the parameters of the map were set at $\alpha = 0.45$ and $\beta = 0.3$.

This control series was divided into $m = 8$ equal subseries, each having 131,072 points. Thus, the total number I of points sampled from the control series is divided evenly among the 8 subseries, so each subseries contributes $I/8 \le 131{,}072$ points to the total. The independence measure $|\xi|$ is calculated using a summation version of (2.4.13):

$$|\xi| = \frac{1}{I^2}\Big|\sum_{\substack{i=1\\ j=2\\ k=3\\ l=4}}^{8}\sum_{h=1}^{128} 2\big(H_{l,h} - H_{j,h}\big)\big(H_{k,h} - H_{i,h}\big)\Big|, \quad \text{where} \quad i < j < k < l. \qquad (3.2.1)$$

A minor algebraic manipulation allows for the subtraction of alternate histograms j from l and i from k, and accounts for the factor of 2 in this expression. As in (2.3.1), each histogram consists of 128 bins. Mean absolute differences D_{ab} between the histograms composed of all the sampled data and the histogram of the control data were also calculated according to (2.3.1).

A simple but highly versatile general approach was employed to develop different sampling strategies. The 8 subseries were further subdivided into any of a number of windows, from 1 to 4,096, increasing in number by powers of 2. The gap G discussed in Section 2.4 in these experiments represents the gap between the last sampled point in any

window and the first sampled point in the following window. Points were then sampled from these windows according to one of two types of sampling strategies.

In the first type of strategy (Type I), the number of points to be sampled from each of the 8 subseries was fixed, and the number of windows was varied. This strategy is illustrated in Figure 3.4. The sampled points used in the calculation of $|\xi|$ were then evenly distributed among the windows. Points sampled from the windows included the first point in the window and then enough consecutive points to obtain the required number of points from each window. For example, if $I = 16,384$ total points were required from the $m = 8$ subseries, then each subseries would contribute 2,048 points to the sample. If these 8 subseries were each divided into 512 windows, then the first 4 points from each window would be the ones sampled.

The second type of strategy (Type II) is to fix the number of windows for each of the subseries and then to take the same increasing number of consecutive points from each window until all the required points of the subseries are sampled. This strategy is illustrated in Figure 3.5. The number $I/8$ of points per subseries ranged from 1024 to the entire 131,072 points of the subseries, increasing by a power of 2 for each trial. For example, if each subseries was divided into 64 windows, then the number of points taken from each window ranged from 16 to 2048.

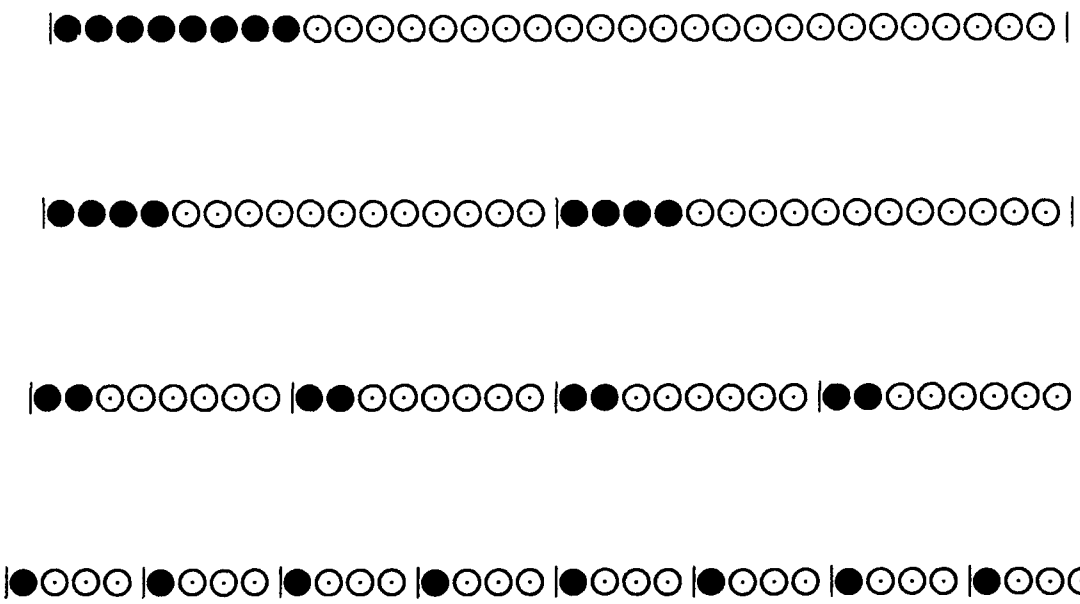


Figure 3.4. Schematic representation of Type I sampling strategy using a fixed number of points. In this example, 8 sampled points (filled circles) are divided among (from top to bottom) 1, 2, 4, and 8 windows. The unfilled circles represent the gap G of unsampled points (24, 12, 6, and 3 points, respectively). This illustration depicts only one of the 8 subseries used to create the histogram.

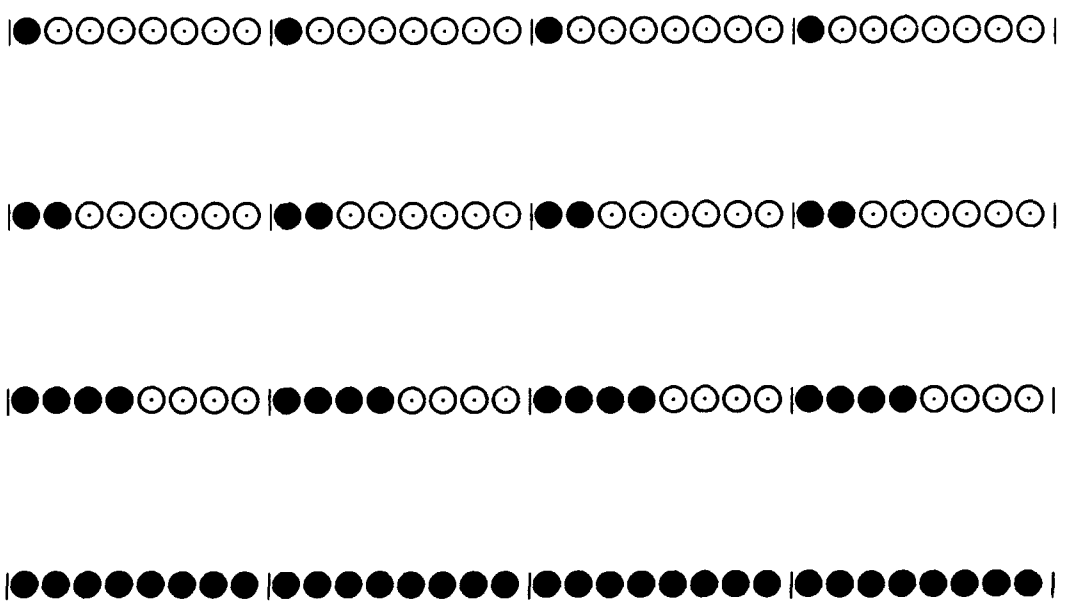


Figure 3.5. Schematic representation of Type II sampling strategy using a fixed number of windows. The number of sampled points (filled circles) increases from 1 to 8. The unfilled circles represent the gap G of unsampled points (7, 6, 4, and 0 points, respectively). This illustration depicts only one of the 8 subseries of the control histogram divided into 4 windows.

3.2.1. Type I Sampling Strategy: Fixed Number of Points

For the Type I strategy, three sets of trials were made, in which samples of $I/8 = 4096$, 16384, and 65,536 points were selected from each of the 8 subseries of the control series. The total numbers of points sampled from the control series were therefore $I = 32,768$, 131,072, and 524,288, respectively. The number of windows into which the selected points in each subseries was distributed varied from 1 to 4,096. The numbers of points selected from each window needed to satisfy the total number required for each trial are included in Table 3.2, along with the length of the gap G between the last point in a window and the first point in the next window. One of the consequences of this strategy is that for a given number of points, the gap ratio (the ratio of the number of unsampled points in a window to the number of consecutive points sampled from that window) remains constant. For example, if 4096 points are selected from a subseries, then the ratio of sampled to unsampled points is always 1:31, without regard to the number of windows into which the subseries is divided. If the subseries consists of only 1 window, then this ratio is 4,096:126,976. If the subseries is divided into 4096 windows, then this ratio is 1:31. In contrast, for 16384 points per subseries this ratio is 1:7, and for 65,536 points per subseries it is 1:1.

For each trial, the dependence measure $|\xi|$ of sampled data and the mean absolute difference D_{ab} between sampled data and the data of the control series were found. Eight histograms were constructed of the sampled data from each of the 8 subseries and the

Table 3.2. Type I sampling strategy for a specified number of points in each of 8 subseries of the control time series. The leftmost column displays the numbers of windows into which the sampled points are distributed. Numbers in parentheses indicate the number of consecutive unsampled points G between the last sampled point of a window and the first sampled point of the next window.

	Points per Subdivision //8		
Windows	4096	16384	65536
1	4096 (126976)	16384 (114688)	65536 (65536)
2	2048 (63488)	8192 (57344)	32768 (32768)
4	1024 (31744)	4096 (28672)	16384 (16384)
8	512 (15872)	2048 (14336)	8192 (8192)
16	256 (7936)	1024 (7168)	4096 (4096)
32	128 (3968)	512 (3584)	2048 (2048)
64	64 (1984)	256 (1792)	1024 (1024)
128	32 (992)	128 (896)	512 (512)
256	16 (496)	64 (448)	256 (256)
512	8 (248)	32 (224)	128 (128)
1024	4 (124)	16 (112)	64 (64)
2048	2 (62)	8 (56)	32 (32)
4096	1 (31)	4 (28)	16 (16)
Gap Ratio	1:31	1:7	1:1

dependence $|\xi|$ of the data in these 8 subseries was calculated according to (3.2.1). The sampled data from the 8 subseries for any trial was also consolidated and the histogram of this pooled data was constructed and the area normalized to that given by a histogram of 131,072 distances. The mean absolute difference D_{ab} between these combined histograms and a normalized histogram composed of all of the data of the control series was also calculated using (2.3.1) for each trial. These results were then examined for any relationship between $|\xi|$ or D_{ab} and the number of windows as well as for any relationship between $|\xi|$ and D_{ab}.

The results for $I/8 = 4096$ points are presented in Figure 3.6. No statistically significant increasing or decreasing trend in either $|\xi|$ or D_{ab} is apparent as the number of windows is varied. Four local minima are apparent in the curve for $|\xi|$. The smallest of these occur at 2, 32, and 128 windows. A single, broader minimum may exist between 16 and 256 windows, although the value of $|\xi|$ for 64 windows is relatively high. The number of points per window for these minima is 2048 for 2 windows, and 128, 64, and 32 points per window for 32, 64, and 128 windows, respectively. A significant correlation between $|\xi|$ and D_{ab} was found for this set of trials, however. The correlation coefficient r^2 for these two variables is 0.2889. Thirteen pairs of values are used to determine this correlation, and so for the associated 11 degrees of freedom, this corresponds to a fairly high confidence level of between 0.90 and 0.95. This correlation was direct; decreasing values of $|\xi|$ are directly related to decreasing values of D_{ab}.

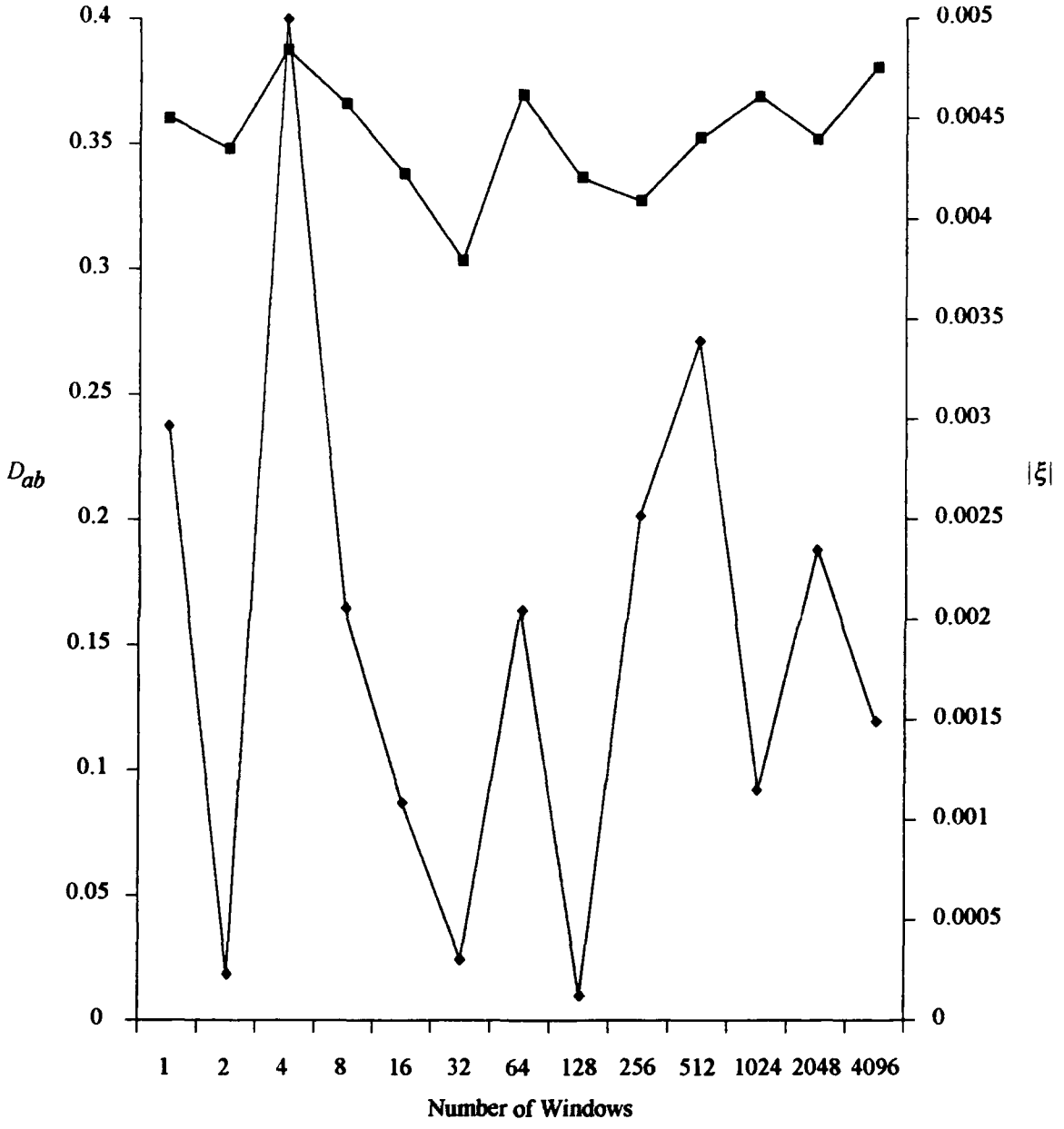


Figure 3.6. Mean absolute difference D_{ab} (squares) and the dependence measure $|\xi|$ (diamonds) as functions of the number of windows into which each subseries is divided for 4096 points.

The results for $I/8 = 16{,}384$ points are presented in Figure 3.7. Again, no statistically significant increasing or decreasing trend of either $|\xi|$ or D_{ab} with the number of windows was found. The values of $|\xi|$ that indicate the highest degrees of independence occur at 1 window, 4 windows, and 512 windows. The number of points per window for these minima are 16,384 for 1 window, 4096 for 4 windows, and 32 for 512 windows. Unlike the results for 4096 points, no significant correlation between $|\xi|$ and D_{ab} is evident for this set of trials. The correlation coefficient r^2 for these two values is 0.1164. For 24 degrees of freedom, this value indicates the confidence in the correlation is well below 0.90 and so no correlation is accepted.

The results for $I/8 = 65{,}536$ points are presented in Figure 3.8. As in the cases of 4096 points and 16,384 points no statistically significant trend of either $|\xi|$ or D_{ab} with the number of windows is apparent. The best values of $|\xi|$ for these trials are observed at 2 and 4 windows (32,768 points and 16,384 points, respectively), and at 1024 windows (64 points). As for the case of 4096 points, however, a significant correlation between $|\xi|$ and D_{ab} was found for these trials. The correlation coefficient r^2 in this case is 0.2674, corresponding to a confidence level of between 0.90 and 0.95. As before, this is a direct correlation.

The results of these trials with fixed numbers of points distributed among varying numbers of windows suggest that distributing the points among several windows serves to increase the independence of the data over that obtained from using only 1 window. It

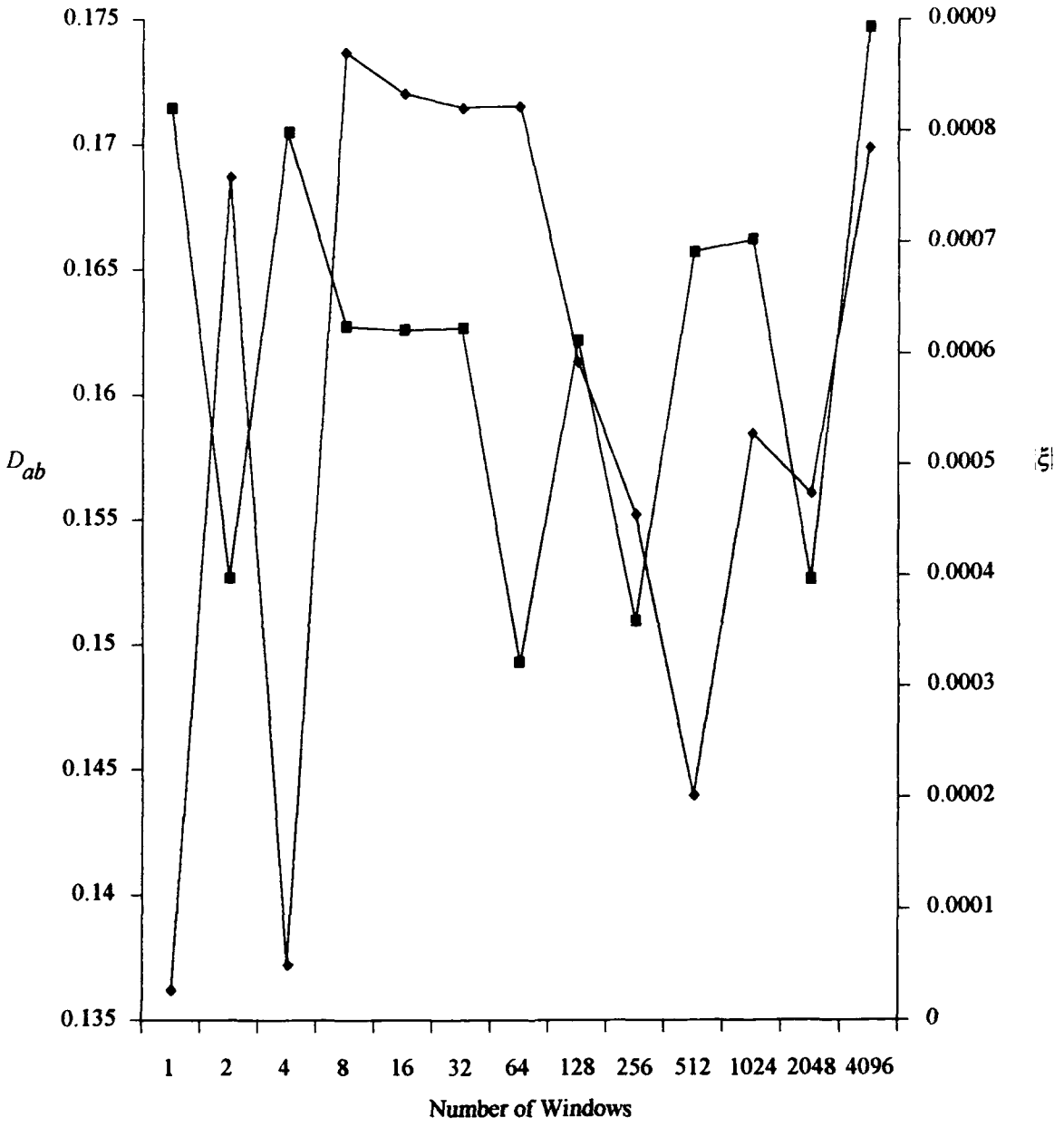


Figure 3.7. Mean absolute difference D_{ab} (squares) and the dependence measure $|\xi|$ (diamonds) as functions of the number of windows into which each subseries is divided for 16,384 points.

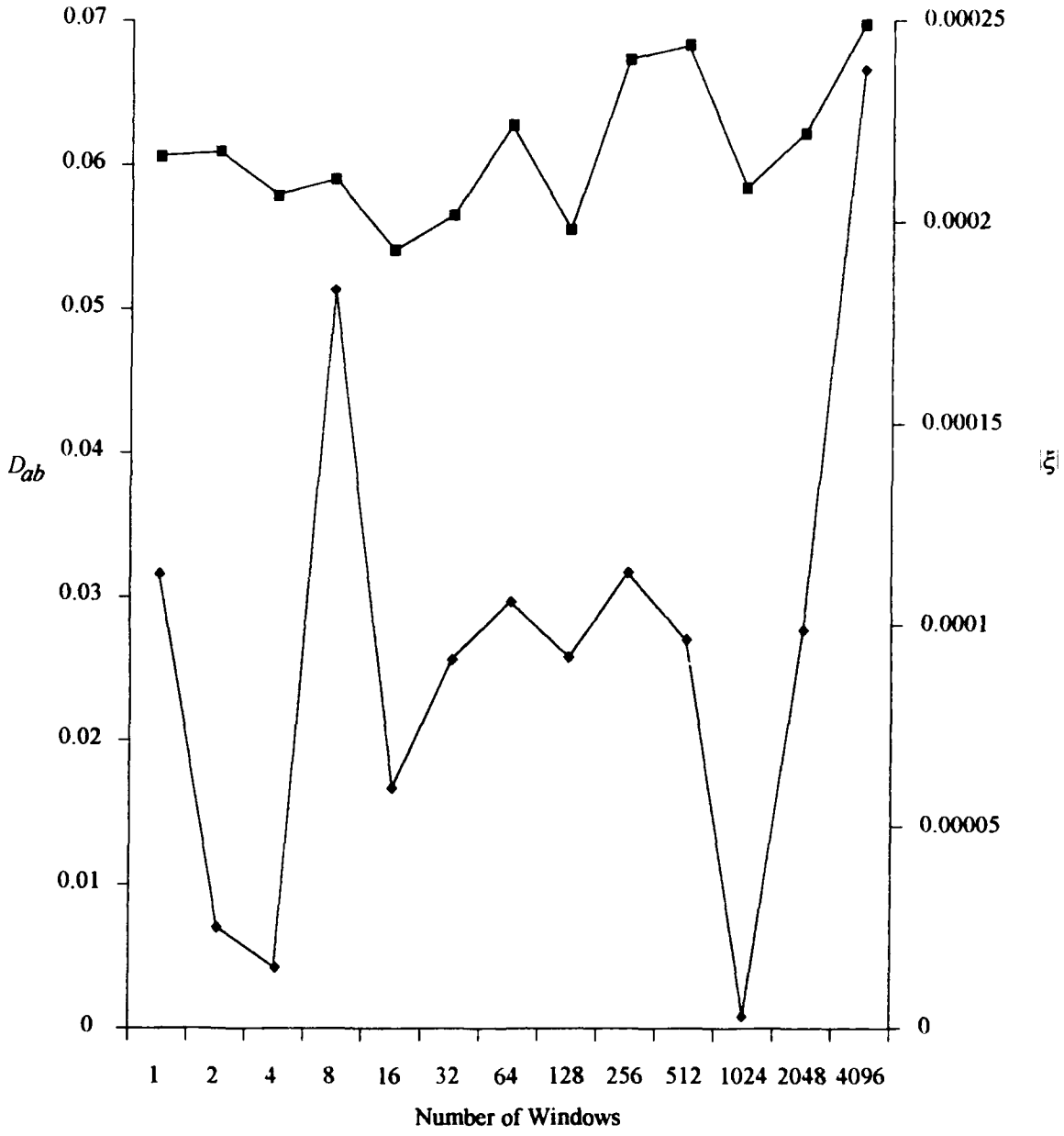


Figure 3.8. Mean absolute difference D_{ab} (squares) and the dependence measure $|\xi|$ (diamonds) as functions of the number of windows into which each subseries is divided for 65,536 points.

was not observed, however, that the greatest independence was observed when the points were distributed among the maximum possible number of windows, and so this is not necessarily an optimal sampling strategy. The best results were obtained for those subseries divided into an intermediate number of windows—those containing 32 to 64 consecutive points (64 and 128 windows)—rather than for those containing just 1 point (4096 windows). Maximizing the number of windows, or minimizing the number of points per window, does not necessarily yield the most independent data. Significant correlations between mean absolute difference D_{ab} and dependence $|\xi|$ were found in two of the three trials. This result strongly suggests that the less computationally intensive measure D_{ab} could be substituted for $|\xi|$ in evaluating sampling strategies for other time series data.

3.2.2. Type II Sampling Strategy: Fixed Number of Windows

For Type II sampling, the subseries of the control histogram are divided into specified numbers of windows and the number of sampled points varies. Five trials were performed, with each of the 8 subseries of the control histogram divided into 1, 2, 4, 64, or 1024 windows. These values for the numbers of windows into which to divide the subseries were selected to cover a three order-of-magnitude range of values, while also selecting some of the trials from the Type I experiment for which small values of the dependence measure $|\xi|$ were obtained. If cases having smaller values of $|\xi|$ are selected,

then these values could be viewed in the context of varying the number of points rather than varying the number of windows. The number of points sampled from each subseries was increased incrementally by powers of 2 from 1024 to 131,072 points. The number of points sampled from each window ranged from 1 to 128 if the subseries were divided into 1024 windows and ranged from 1024 to 131,072 if the subseries were left undivided. This strategy is illustrated in Figure 3.5. The numbers of points sampled and their corresponding gaps G are included in Table 3.3. The gap ratio given by the length of the gap of unsampled points to the length of the series of consecutive sampled points within a window decreases as the number of points sampled within a window increases; but for any given number of sampled points (column in Table 3.3), the gap ratio is the same regardless of the number of windows into which the subseries is divided.

The dependence measure $|\xi|$ of sampled data and the mean absolute difference D_{ab} between histograms of the pooled sampled data from the 8 subseries and the histogram of the data of the control series were calculated for each sampling of points using (3.2.1) and (2.3.1), respectively. In these trials, the number I of points was increased for each fixed number of windows, so the number of points could be treated as an independent variable. As for the Type I strategy, relationships between this independent variable and $|\xi|$ or D_{ab} were then investigated, as was a relationship between $|\xi|$ and D_{ab}.

The relationships between D_{ab} and the number of points sampled from each subseries for each fixed number of windows per subseries are shown in Figure 3.9. As observed in Section 3.1, the magnitude of D_{ab} decreases with increasing numbers of

Table 3.3. Type II sampling strategy in which the number of points per window increases and the number of windows per subseries is specified. Numbers in parentheses indicate the number G of unsampled points between the last sampled point of a window and the first sampled point of the next window.

	Points per Subdivision $I/8$							
Windows	1024	2048	4096	8192	16384	32768	65536	131072
1	1024 (130048)	2048 (129024)	4096 (126976)	8192 (122880)	16384 (114688)	32768 (98304)	65536 (65536)	131072 (0)
2	512 (65024)	1024 (64512)	2048 (63488)	4096 (61440)	8192 (57344)	16384 (49152)	32768 (32768)	65536 (0)
4	256 (32512)	512 (32256)	1024 (31744)	2048 (30720)	4096 (28672)	8192 (24576)	16384 (16384)	32768 (0)
64	16 (2032)	32 (2016)	64 (1984)	128 (1920)	256 (1792)	512 (1536)	1024 (1024)	2048 (0)
1024	1 (127)	2 (126)	4 (124)	8 (120)	16 (112)	32 (96)	64 (64)	128 (0)
Gap Ratio	127:1	63:1	31:1	15:1	7:1	3:1	1:1	0:1

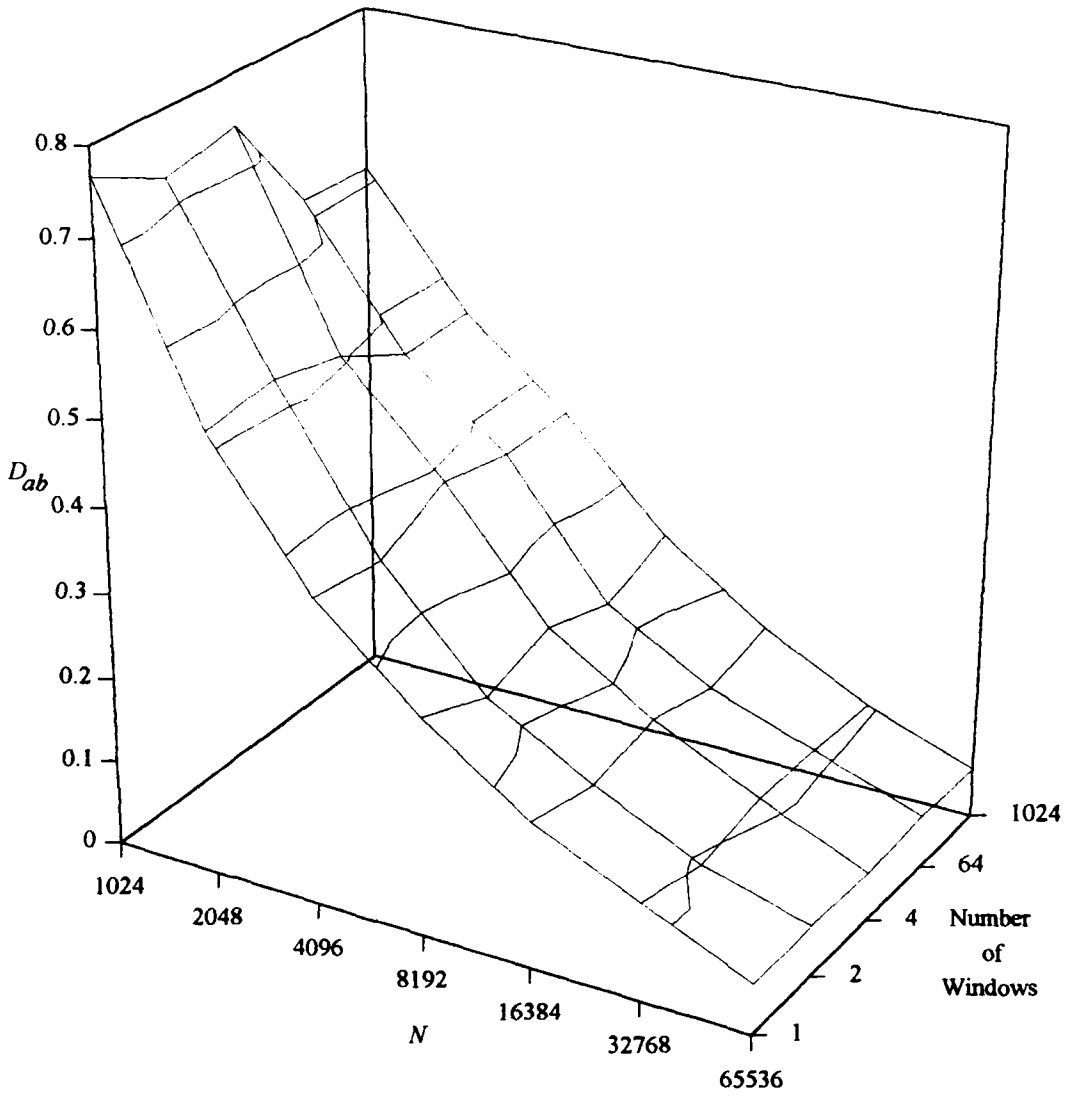


Figure 3.9. Mean absolute differences D_{ab} between histograms of sampled data and the histogram of data of the control series as functions of sample size N for five different Type II sampling strategies (1, 2, 4, 64, and 1024 windows per subseries).

sampled points. Again a log-log regression indicated a power law decrease of D_{ab} with increasing sample size similar in form to (3.1.2). Statistics for these regression analyses are presented in Table 3.4.

Unlike the discussion in Section 3.1 concerning mean absolute differences between successive histograms of increasing length, the results here, illustrated in Figure 3.9, concern mean absolute differences between histograms of data from each sample for a specified number of windows and the histogram of the data from the entire control series. The mean absolute difference between histograms of data from trials in which all points are sampled from all the windows of the 8 subseries and the histogram of the control series are therefore equal to zero. In the power law regression, these differences could

Table 3.4. Regression statistics for decrease of mean absolute difference D_{ab} with increasing sample size $l/8$ for several different numbers of windows. Statistics are for the power law decreases (3.1.2) of mean absolute difference (with $B = 0$) shown in Figure 3.9.

Number of Windows	R	$\ln A$	r^2
1	0.58496	3.83754	0.99011
2	0.60159	3.91321	0.99785
4	0.59302	3.89411	0.98724
64	0.54974	3.42916	0.98874
1024	0.55912	3.53096	0.97899

not be used because the logarithm of zero is undefined, and so the log-log regressions were performed for all values of D_{ab} and sample size excluding the final values. This regression model yielded higher correlations of mean absolute difference to sample size than the exponential model.

Higher rates R of the decrease in D_{ab} with increasing sample sizes were observed for strategies using 1, 2, or 4 windows per subseries than for 64 and 1024 windows using the power law model (3.1.2) (Table 3.4). For smaller sample sizes, the values of D_{ab} were observed to be smaller for the 64 and 1024 windows per subseries strategies (Figure 3.9) As the sample size increases, the range of values of D_{ab} for the five strategies decreases. The rates of decrease for the strategies involving smaller numbers of windows must therefore be greater because their curves converge on approximately the same values of D_{ab} as those for strategies involving many more windows. The mean absolute difference between sampled data and control series data is not so sensitive to sampling strategy for greater numbers of points sampled. For example, for 65,536 sampled points (50% of the available points of the control series), the range of values of D_{ab} is only 0.00215, or approximately 3.5% of the greatest value of D_{ab} for 65,536 point samples. However, if only 1024 points are sampled (less than one percent of the available points of the control series), then a range of values of D_{ab} of 0.15377, or approximately 20% of the greatest value of D_{ab}, is observed. For large sample sizes, finding an optimal sampling strategy may be less beneficial than for smaller sample sizes. It must be borne in mind that the results presented here are based on comparisons of sampled data with the known

standard of the control series. That which constitutes large and small for other time series data may be much more difficult to establish.

The dependence measure $|\xi|$ of sampled data for the five Type II strategies is shown in Figure 3.10. The overall trend with increasing sample size for each of the five strategies is an increase in the value of $|\xi|$, which signifies a loss of independence of the data in the 8 subseries. The highest value of $|\xi|$ occurs for 131,072 points and is the same for all five strategies. This result is not unexpected as it represents the selection of all the data from the subseries for all five strategies. The most rapid increases in $|\xi|$ occur for the 1 window and 2 window strategies for sample sizes between 4096 and 65,536 points. For sampling sizes less than 4096 points, no discernible difference exists among the values of $|\xi|$ for the five strategies, nor does any trend in the values of $|\xi|$ with increasing sample size occur.

3.3. General Conclusions and Recommendations for Further Study

Calculations involving 610 pairs of values of N and $N + \Delta N$ were performed in order to determine that the mean absolute difference D_{ab} between histograms representing the Smale-Williams attractor follows a power law decrease with increasing numbers of points. Consistently high values of the correlation coefficient r^2 (Tables 3.1 and 3.4) were observed, and so this conclusion is firmly established statistically. A correlation

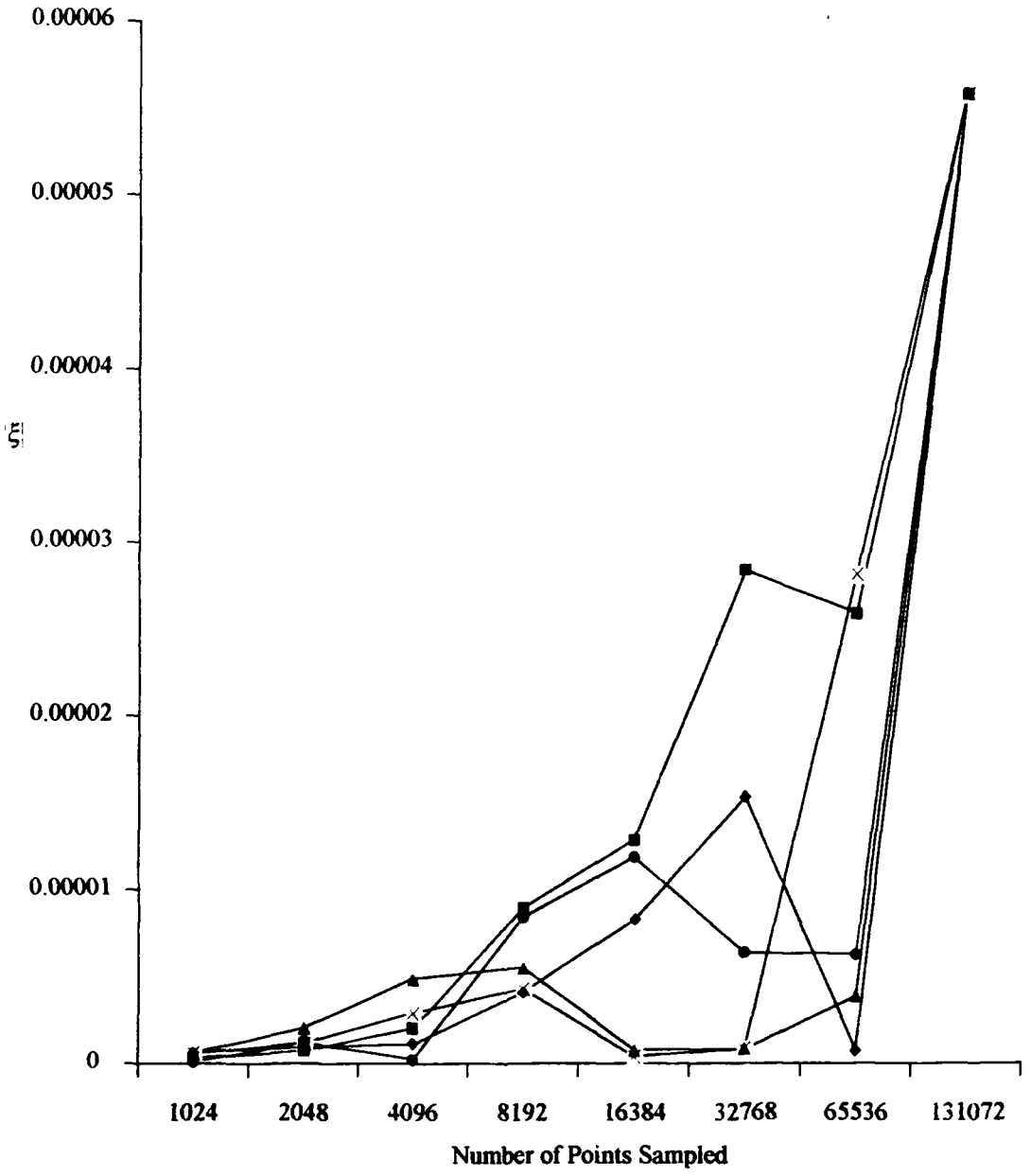


Figure 3.10. The dependence measure $|\xi|$ of sampled data for five Type II strategies as a function of the number of points sampled. The individual curves are for 1 window (squares), 2 windows (diamonds), 4 windows (circles), 64 windows (crosses), and 1024 windows (triangles).

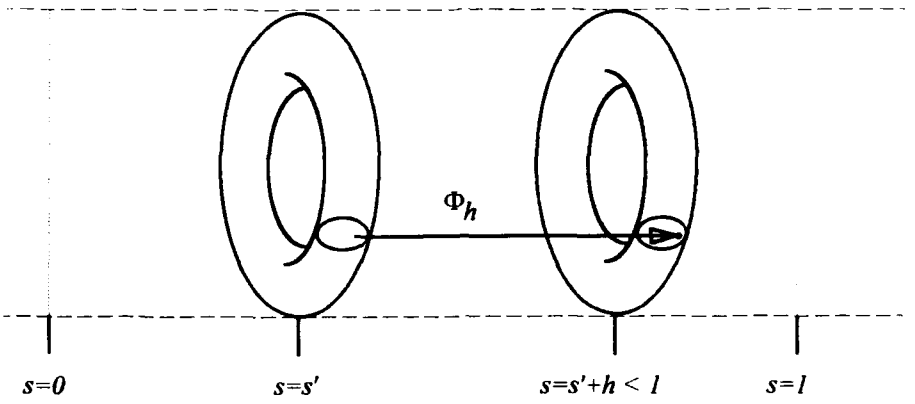
between the dependence measure $|\xi|$ and the mean absolute difference D_{ab} for Type I sampling was observed for only two of three specified numbers of points. This observation is based on samples from only one control time series. The decrease in D_{ab} with increasing numbers of windows for a small number of points using a Type II strategy is also based on samples from only a single control time series. All of the calculations of mean absolute differences and dependence measures were performed for only one set of initial conditions and one pair of values for the map parameters α and β. Additional trials for the Type I and Type II strategies should be performed using different initial conditions and map parameters so that means and standard deviations of the different values of D_{ab} and $|\xi|$ can be generated and their values compared and evaluated with more statistical rigor. This might help to confirm (or refute) some of the preliminary observations and tentative conclusions of the experiments discussed in Sections 3.2.1 and 3.2.2.

In the Type I trials for any specified number of sampled points, variations in the values of the dependence measure $|\xi|$ were very small. This result is somewhat counterintuitive because of the nature of the sampling. In the Type I strategy, a specified number of points could be sampled as a continuous series of adjacent points in a single window, as individual points isolated in separate windows, or in some combination of these patterns. The amount of independent information provided by an individual point would be expected to be greatest if the point is not influenced by the other sampled points. If the individually sampled points are separated by long gaps of unsampled points, then the value of the dependence measure $|\xi|$ for this sampling strategy should be significantly less

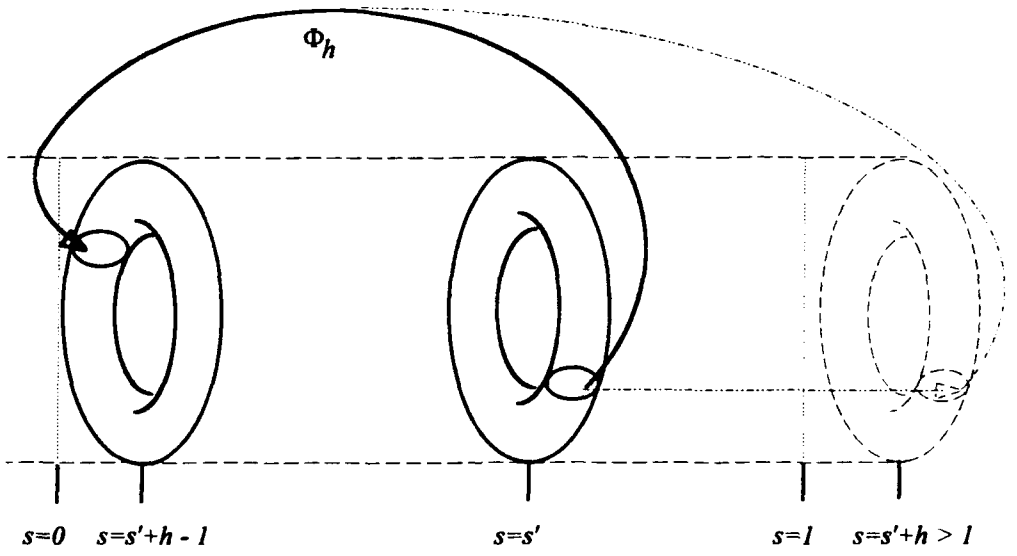
than that for a sampling of adjacent and presumably highly correlated points. This was not observed, however, and so it appears that temporally adjacent points are nearly as decorrelated as sampled points separated by several hundreds of unsampled points. This is not typical of continuous dynamical systems, particularly not of systems of ordinary differential equations such as those that describe meteorological processes.

Further investigations of the advantages of different sampling strategies should involve dynamical systems whose points decorrelate much more slowly. Differentiable equations that model atmospheric flows are solved numerically and the time steps of integration might possibly be adjusted to control decorrelation. To date, the value of the correlation dimension of the attractor of any such model can not be determined analytically and so this approach is unfeasible. A more tractable approach is to use an extension of the Smale-Williams attractor for which a continuous time step is introduced and for which an analytical dimension may be calculated.

The mapping torus of the Smale-Williams attractor is such a dynamical system. The construction of this system, illustrated in Figure 3.11, begins with the Cartesian product of a line and the Smale-Williams attractor described in Chapter 2. Taking this Cartesian product introduces a fourth coordinate s to the coordinates x, y, and t in the map described by (2.1.12). In the Cartesian product, the point (s,t,x,y) maps to the point $(s+h,t,x,y)$, where h is an arbitrary incremental value less than 1 that is set in advance. Now any point (s,t,x,y) on the attractor is identified with the point $(s+1, 3t-[3t], \alpha\cos(2\pi t)+\beta x, \alpha\sin(2\pi t)+\beta y)$ to produce the mapping torus



A



B

Figure 3.11. The mapping torus Φ_h of the Smale-Williams attractor, a Cartesian product of the Smale-Williams attractor and a line. The condition $s = s'+h < 1$ is depicted in A. A high degree of correlation between the original point and its image occurs. The condition $s = s'+h \geq 1$ is depicted in B and the Smale-Williams map is applied in this step. This mapping (B) would result in the same rapid decorrelation as occurs in the Smale-Williams attractor.

(illustrated in Figure 3.11B). In effect, the constraint $0 \le s < 1$ is applied, and so this identification requires the map Φ_h to be defined explicitly on the point (s',t,x,y) as follows:

$$\Phi_h:(s',t,x,y) \mapsto \begin{cases} (s'+h,t,x,y) \text{ if } s'+h<1 \\ (s'+h-1, 3t-[3t],\ \alpha\cos(2\pi t)+\beta x, \alpha\sin(2\pi t)+\beta y) \text{ if } s'+h \ge 1 \end{cases} \tag{3.3.1}$$

For the second condition of $s'+h \ge 1$, the Smale-Williams map described in (2.1.12) is applied. By setting h to appropriate values, the decorrelation inherent in the Smale-Williams attractor can be controlled, with smaller values of h yielding higher correlations. When $s'+h<1$, the degree of correlation between a point and its image is very high because no twist is applied, but when $s'+h \ge 1$, the same rapid decorrelation as observed for the Smale-Williams occurs. The overall degree of correlation in a time series should therefore be determined by the ratio of the number of times that the condition $s'+h<1$ is true to the number of times $s'+h \ge 1$ is true. If the decorrelation rate is moderated by choosing an appropriate value of h, then dependence measures $|\xi|$ of data from this mapping torus sampled by different strategies may show a much greater range of values than is shown by the Smale-Williams attractor.

The Hausdorff dimension d_H and box-counting dimension d_B of the Smale-Williams attractor are equal by (2.2.5), and so the Smale-Williams attractor can be said to be *reasonably regular* (Falconer 1990). The Hausdorff dimension d_H of the mapping torus of the Smale-Williams attractor is equal to the sum of the dimensions of its two

factors, the Smale-Williams attractor itself and the line . By similar arguments to those presented in Section 2.2, the correlation dimension ν of the mapping torus is equal to its Hausdorff dimension. The dimension of the line is 1, and so the values of the correlation dimension ν and the Hausdorff dimension d_H of the mapping torus of the Smale-Williams attractor can be obtained from

$$d_H = \nu = 2 + \ln 3 / \ln(1/\beta). \tag{3.3.3}$$

The analytical expression (3.3.3) for the values of dimensions and the slow rate of decorrelation together make this map suitable not only for the study of the effects of sampling on the independence of the data and the values of the mean absolute differences, but they also make it suitable for the study of the effect of sampling on estimates of the dimensions of an attractor, such as is discussed in the following chapter.

Chapter 4

SAMPLING STRATEGIES AND ESTIMATES OF THE CORRELATION DIMENSION OF THE SMALE-WILLIAMS ATTRACTOR

The mean absolute difference D_{ab} between the histogram of a sample of data and the histogram of a sufficiently long control series of data can be used to assess the accuracy of the characterization of the Smale-Williams attractor by the sampled data. The histogram of the control series to which a histogram of sampled data is compared is itself an estimate of the true histogram; however, by arguments presented in Section 3.1, this estimate is sufficiently accurate to be used as a standard of comparison. A representation of the Smale-Williams attractor by a set of sampled data can also be characterized by an estimate of its correlation dimension ν using the values of x, y, and t of the sampled points. The value of the correlation dimension can be estimated using either the approach of Grassberger and Procaccia (1983) or the approach of Wells *et al.* (1992) described in Sections 2.5 or 2.6, respectively. The accuracy of such an estimate by the method of Wells *et al.* (1992) can be assessed by directly comparing the estimated value of ν with the analytical value of ν given by (2.2.5) for a specified value of β. The conformity of the sampled data to the relationship between $C(\varepsilon)$ and ν (2.5.3) can be also be used to assess the accuracy of a characterization of the attractor by sampled data. The dependence

measure $|\xi|$ provides an additional criterion with which to assess the quality of a sample of data. Time series data can be sampled by different strategies described in Section 3.2 and the accuracy of characterizations of the Smale-Williams attractor by this sampled data can be assessed using these different criteria. Mean absolute difference, correlation dimension estimates, and the dependence measure can be determined for each set of sampled data and these values can be compared to each other for each sampling strategy to determine any correlations among the different methods of assessing the accuracy of characterizations of the attractor.

4.1. The Analytical Value of the Correlation Dimension ν for Specified Values of β and the Distance Formula for the Correlation Function

The analytical value of the correlation dimension of the Smale-Williams attractor can be determined using (2.2.5) after a value of β has been chosen. For all of the experiments concerning mean absolute differences D_{ab} and the dependence measure $|\xi|$ described in the previous chapter, the value of β was set at 0.30. This value of β yields a value of the correlation dimension ν of approximately 1.9125. Other values of β are used in preliminary experiments described in Section 4.3 that are designed to assess the utility of the multiple moment approach of Wells *et al.* (1992) in estimating the correlation dimension of the Smale-Williams attractor. In experiments described in Section 4.4 in which the value of the correlation dimension ν is estimated from data sampled according

to a Type I strategy, a value of $\beta = 0.3$ is again used. In these experiments, the value of $\nu \equiv 1.9125$ is used as a target value against which estimates of ν may be compared.

The correlation function $C(\varepsilon)$, described by (2.5.1), requires the use of a properly defined distance function $r_{i,j} = \left|\mathbf{x}_i - \mathbf{x}_j\right|$. One intuitive approach to determining the distances between points in a three-dimensional phase space would be to use a distance analogous to the Euclidean distance, such as

$$r_{i,j} = \sqrt{\left(t_i - t_j\right)^2 + \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{4.1.1}$$

However, the coordinate t in the Smale-Williams attractor is related to the angular displacement $\theta = 2\pi t$ between any two points $\mathbf{x}_i = (t_i, x_i, y_i)$ and $\mathbf{x}_j = \left(t_j, x_j, y_j\right)$ that are defined by (2.1.12). As noted in Chapter 2, the Smale-Williams attractor can be imagined as the stretching and twisting of an initial torus into a folded and contracted torus. Consequently, distances between points on the attractor are distances between points on tori, and so distances around the torus can be measured in either of two directions: displaced from each other by an angle θ or by an angle $2\pi - \theta$. Because $0 \le t \le 1$ holds by (2.1.8), a more appropriate expression that ensures that the smaller of these two angles is always used in determining the distance $r_{i,j}$ between two points is

$$r_{i,j} = \frac{\sqrt{\alpha\left(\min\left(t_i - t_j, 1 - \left(t_i - t_j\right)\right)\right)^2 + \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}}{\alpha}. \tag{4.1.2}$$

In the expressions (2.1.12) that describe the map, α is a scaling factor that determines the maximum distance between points on the attractor. Curiously, by incorporating α into the distance function as in (4.1.2), projections of sets of attractor points (illustrated in Figure 2.3) will be the same if the value of only α is changed. That is, Smale-Williams attractors belonging to the same value of β but to different values of α are dynamically equivalent.

4.2. Estimates of the Correlation Dimension Using Sampled Data

The accuracy of estimates of the correlation dimension ν of the Smale-Williams attractor can be determined in either of two ways. The first of these is to use the logarithm of the correlation function $C(\varepsilon)$ of Grassberger and Procaccia (1983) in (2.5.4) to relate $\ln C(\varepsilon)$ to the correlation dimension ν. For any value of ε, the value of $\ln F(\varepsilon)$ can be determined by

$$\ln F(\varepsilon) \cong \ln C(\varepsilon) - \nu \ln(\varepsilon). \tag{4.2.1}$$

Values of $\ln F(\varepsilon)$ are determined for all values ε for which values of $\ln C(\varepsilon)$ are determined. These values of $\ln F(\varepsilon)$ are calculated using the analytically determined value of $\nu \cong 1.9125$ in (4.2.1), and then plotted as a function of ε. The interval $[\varepsilon_1, \varepsilon_2]$ over which values of $\ln F(\varepsilon)$ are approximately constant is determined by examination of the plotted values. An example of the relationship between $\ln F(\varepsilon)$ and ε is illustrated in Figure 4.1. Once the interval $[\varepsilon_1, \varepsilon_2]$ over which the value of $\ln F(\varepsilon)$ is independent of ε

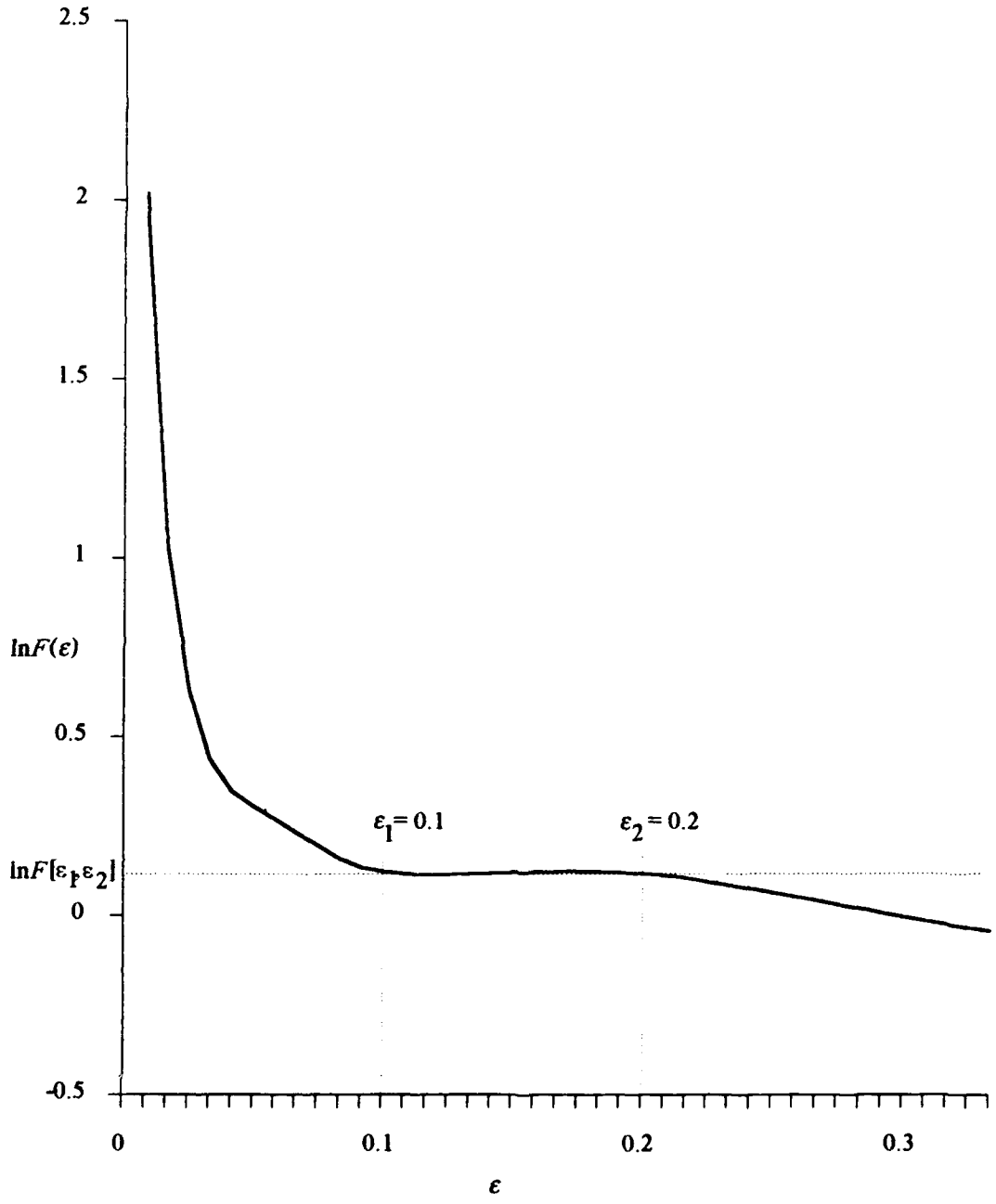


Figure 4.1. The relationship between $\ln F(\varepsilon)$ and ε for $I/8 = 256$ points in one window. The interval over which this value was determined to be constant and for which the standard deviation is calculated is $[\varepsilon_1,\varepsilon_2] = [0.1,0.2]$. The average value $\ln F[\varepsilon_1,\varepsilon_2]$ of $\ln F(\varepsilon)$ on this interval is 0.12 and the standard deviation σ of $\ln F(\varepsilon)$ for $\ln F[\varepsilon_1,\varepsilon_2]$ is 0.00386.

is determined and labeled $\ln F[\varepsilon_1, \varepsilon_2]$, the standard deviation σ of values of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ can be determined and used as an index of the conformity of the sampled data to the power law relationship (2.5.3) and therefore of the accuracy of the characterization of the attractor by the sampled data.

To assess the validity of this method when the analytical value for ν is not known, several different artificial values of ν were used in (4.2.1) for each set of sampled data to determine the interval $[\varepsilon_1, \varepsilon_2]$ over which $\ln F(\varepsilon)$ is constant and to estimate the standard deviation σ on this interval for each value of ν. The results of one such test are presented in Table 4.1 and are for the same sample of data for which values of $\ln F(\varepsilon)$ are depicted in Figure 4.1. For all samples of data, the smallest value of the standard deviation σ of $\ln F[\varepsilon_1, \varepsilon_2]$ was consistently observed for the analytical value $\nu \cong 1.9125$ and was observed on the interval $[\varepsilon_1, \varepsilon_2] = [0.1, 0.2]$.

Table 4.1. Values of the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ for several artificial values of the correlation dimension ν and the analytically determined value $\nu \cong 1.9125$. Values are for $I/8 = 256$ points in one window. Standard deviations are determined on the interval $[\varepsilon_1, \varepsilon_2] = [0.1, 0.2]$. The smallest value of σ (0.00386) occurs for the analytical value of ν.

ν	σ
1.8500	0.01333
1.9000	0.00474
1.9125	0.00386
1.9200	0.00403
1.9400	0.00648
1.9600	0.00998

The method of Wells *et al.* (1992) can be used more directly to assess the accuracy of an estimate of the correlation dimension. Values of $M(p,\varepsilon,N_T)$ can be calculated from sampled data using a predetermined value of p in (2.6.1), and these values can be used to determine the correlation dimension ν for values of ε using

$$\nu(\varepsilon,p)=\frac{pM(p,\varepsilon,N_T)}{1-M(p,\varepsilon,N_T)}. \tag{4.2.2}$$

These values of $\nu(\varepsilon,p)$ can be calculated for a range of values of ε and plotted as a function of ε. More appropriately, the average values $\overline{\nu}(\varepsilon)$ of $\nu(\varepsilon,p)$ for several values of p can then be examined over a range of values of ε and compared with the analytical value of the correlation dimension ν to assess the accuracy with which the sampled data characterizes the attractor in phase space.

4.3. Estimates of the Correlation Dimension ν for Different Values of β Using the Method of Higher Moments

To determine the specific manner for interpreting (4.2.3) to estimate the correlation dimension ν of the Smale-Williams attractor from a sample of points, three trials were performed using different values of the contraction parameter β in (2.1.12). The values used in these trials were $\beta=0.3$, $\beta=0.2$, and $\beta=1/3$, which correspond to values of the correlation dimension of $\nu\cong 1.9125$, $\nu\cong 1.6826$, and $\nu=2.0$, respectively.

No sampling strategy was applied in these trials; rather, a series of $N = 5000$ consecutive points was used for each case. These points were those that immediately followed a series of 1,048,576 points that were discarded to ensure that the transient signal present in the attractor was minimized. As before, the initial conditions of the map were $x_0 = 0.1$, $y_0 = 0.1$, and $t_0 = 0.2$, and the value of α was 0.45. The values of the moments p used in these trials were 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0. The values of the radius ε for which distances $r_{i,j}$ between points were compared ranged from 0 to 1.5, in steps of 0.0075, so that 200 values of ε were used. The average values $\overline{\nu}(\varepsilon)$ of the correlation dimension ν for all six values of p for each value of ε are presented in Figure 4.2 for the three values of β.

The results from using this approach for estimating the correlation dimension ν are somewhat ambiguous. No clear plateau is discernible in any curve for any value of β, and only for $\beta = 0.3$ ($\nu \cong 1.9125$) was one of the peaks of the curve close to the analytical value. This occurred for the peak between $\varepsilon = 0.1$ and $\varepsilon = 0.2$ and yielded a value of $\nu \cong 1.906$. The general shapes of the curves are the same for all three values of β, but peaks in the curves occurred over different ranges of ε for the three values of β examined. In the case of $\beta = 1/3$ ($\nu = 2.0$), a broader peak is observed than for the other two values of β, but the estimated value of ν is approximately 1.94. For $\beta = 0.2$ ($\nu \cong 1.6826$), the estimated value of ν is approximately 1.83. In the remainder of this thesis, the magnitude of the first peak of each curve is used to determine the correlation dimension rather than a plateau as is more commonly done.

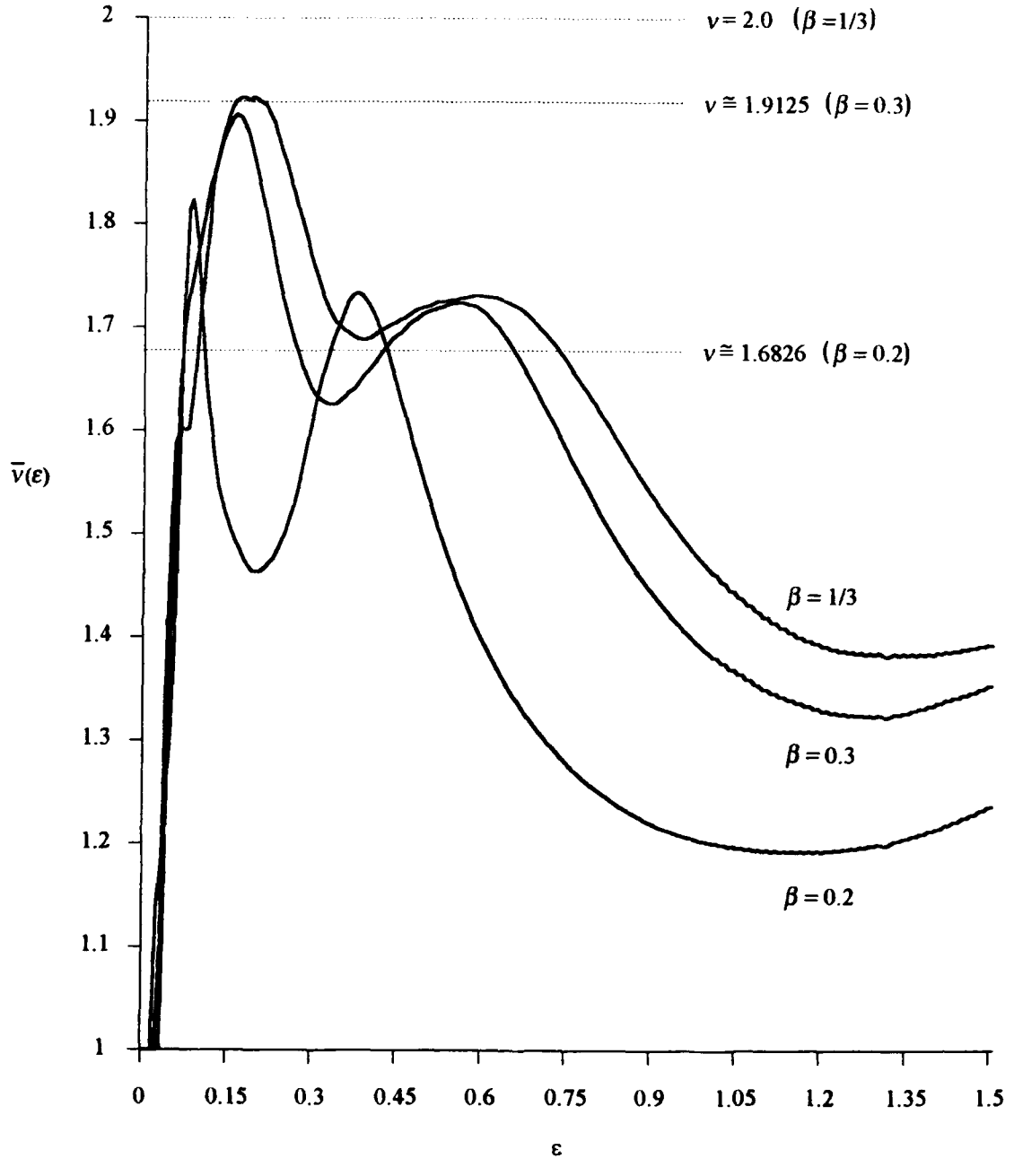


Figure 4.2. Estimated values of the average correlation dimension $\bar{\nu}(\varepsilon)$ using the values of $\nu(\varepsilon, p)$ (4.2.3) for the moments p = 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0. Values of $\bar{\nu}(\varepsilon)$ are estimated for three values of the contraction parameter β (0.2, 0.3, and 1/3). The three analytical values of ν are denoted by the dotted lines.

The only value of β for which a reasonably accurate estimate of ν was obtained is $\beta = 0.3$. For the higher value of β that was studied, the estimate of ν was lower than the analytical value, and for the lower value of β, the estimate of ν was higher than the analytical value. That a relatively accurate estimate of ν was obtained for $\beta = 0.3$ appears to be fortuitous. This estimate of ν occurred at the first peak in the curve and not at a plateau, and would have been difficult to discern if the analytical value of ν had not been independently known. This value $\beta = 0.3$ was used in subsequent experiments in which the accuracy of estimates of the correlation dimension ν from data sampled according to different Type I sampling strategies was assessed. In subsequent experiments the magnitude of the first peak is compared to the analytical value of the correlation dimension. Type II sampling strategies are not considered because far greater numbers of points must be sampled from each subseries, and the number of distances between points that must be calculated to estimate $C(\varepsilon)$ increases as the square of the number of points.

4.4. Sampling Strategy and Estimates of the Correlation Dimension

The accuracy of the estimates of the correlation dimension of the Smale-Williams attractor from sets of points sampled according to different Type I strategies can be judged by comparing calculated values of the correlation dimension $\overline{\nu}(\varepsilon)$ with the analytically determined value ν. To accomplish this, sets of 2048 and 4096 points from the control histogram of 1,048,576 points described in Section 3.1 were sampled. As

before, the initial conditions were $x_0 = 0.1$, $y_0 = 0.1$, and $t_0 = 0.2$; the values of the parameters α and β were 0.45 and 0.3, respectively; and the first 1,048,576 points of the time series were discarded to minimize the transient signal. To sample 2048 or 4096 total points from the control series, $I/8 = 256$ or $I/8 = 512$ points were chosen from each of the 8 subseries. A Type I strategy was used to sample the points, and so constant gap ratios of 1:511 and 1:255 resulted for samples of 2048 and 4096 points, respectively. For both 2048 and 4096 points, the subseries were divided into as many as 256 windows from which a number of consecutive points was sampled. In the case of 4096 points, the 8 subseries were also divided into 512 windows from each of which one point was taken. As before, the number of windows per subseries increased by powers of 2.

The dependence measure $|\xi|$ for each set of sampled data was calculated according to (3.2.1) and the mean absolute difference D_{ab} between the histogram of the pooled data from the 8 subseries and the normalized histogram of the control series was calculated according to (2.3.1). For each sample of data, values of $\nu(\varepsilon, p)$ were calculated using (4.2.3) for the 200 values of ε described in Section 4.3. For each value of ε, values of $\nu(\varepsilon, p)$ were calculated for values of p of 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0, and these six values were averaged to give $\overline{\nu}(\varepsilon)$. These average values were plotted for each set of sampled data as in Figure 4.2. An index $\delta\nu$ with which to assess the accuracy of these estimates was devised by taking the absolute difference between the maximum value of $\overline{\nu}(\varepsilon)$ and the analytical value of the correlation dimension ν, *i.e.*,

$$\delta\nu = \left|\nu - \max\left(\overline{\nu}(\varepsilon)\right)\right|. \qquad (4.4.1)$$

The use of the maximum value of $\overline{\nu}(\varepsilon)$ differs from the standard approach of using the plateau value. The results presented in Section 4.3 suggest that the first peak value yields the best estimate and should be used since no plateau values are observed.

A second method for determining the accuracy of the estimate of the correlation dimension from sampled data is to calculate the values of $\ln F(\varepsilon)$ for each of the 200 values of ε using the analytical value of the correlation dimension ν and (4.2.1). These values are then plotted against their respective values of ε, and the standard deviation σ of these values for a region over which they were judged to be constant used as an index of the accuracy of the estimate. The standard deviations σ of these values of $\ln F(\varepsilon)$ over the interval $[\varepsilon_1, \varepsilon_2] = [0.1, 0.2]$ were then used as the index of accuracy.

Relationships between sampling strategies and the values of the mean absolute difference D_{ab}, the dependence measure $|\xi|$, and the indices $\delta\nu$ and σ of the accuracy of the correlation dimension estimates were then examined. The relationship between the mean absolute difference D_{ab} and the dependence measure $|\xi|$ was also examined to determine whether a correlation between D_{ab} and $|\xi|$ exists for these values of $I/8$ as it does for the cases $I/8 = 4096$ and $I/8 = 65{,}536$ (Section 3.2.1). The relationships between the indices $\delta\nu$ and σ and the dependence measure $|\xi|$ were also examined to ascertain whether such a correlation exists. The mean absolute difference D_{ab} represents the accuracy of the representation of the histogram of the control series by the histogram

of sampled data, and so the accuracy of representations of the Smale-Williams attractor identified with a subset of the standard torus and the accuracy of estimates of its correlation dimension by sampled data can be assessed by comparing D_{ab} with both $\delta\nu$ and σ. Finally, a correlation between $\delta\nu$ and σ may help to substantiate the validity of these methods in determining the accuracy of estimates of the correlation dimension ν.

The relationship between the number of windows into which subseries were divided and the values of the mean absolute difference D_{ab} and the dependence measure $|\xi|$ for the case $I/8 = 256$ are illustrated in Figure 4.3. No direct correlation between either the values of mean absolute difference D_{ab} or of the dependence measure $|\xi|$ and the number of consecutive points sampled is observed, and no correlation between the value of the mean absolute differences D_{ab} and the value of the dependence measure $|\xi|$ is observed. The smallest value of D_{ab} is observed for 32 windows (8 consecutive points) and the smallest values of $|\xi|$ were observed for 16 and 2 windows (16 and 128 consecutive points, respectively).

Values of the accuracy indices $\delta\nu$ and σ as functions of the number of windows into which the subseries were divided are depicted in Figure 4.4. As is observed for the mean absolute difference D_{ab}, no direct correlation exists between either of these indices and the number of windows. In addition, no correlations exist between either $\delta\nu$ or σ and the dependence measure $|\xi|$. The smallest value of $\delta\nu$ is observed for 32 windows, and one of the smaller values of the standard deviation σ of $\ln F(\varepsilon)$ on $[\varepsilon_1, \varepsilon_2]$ is

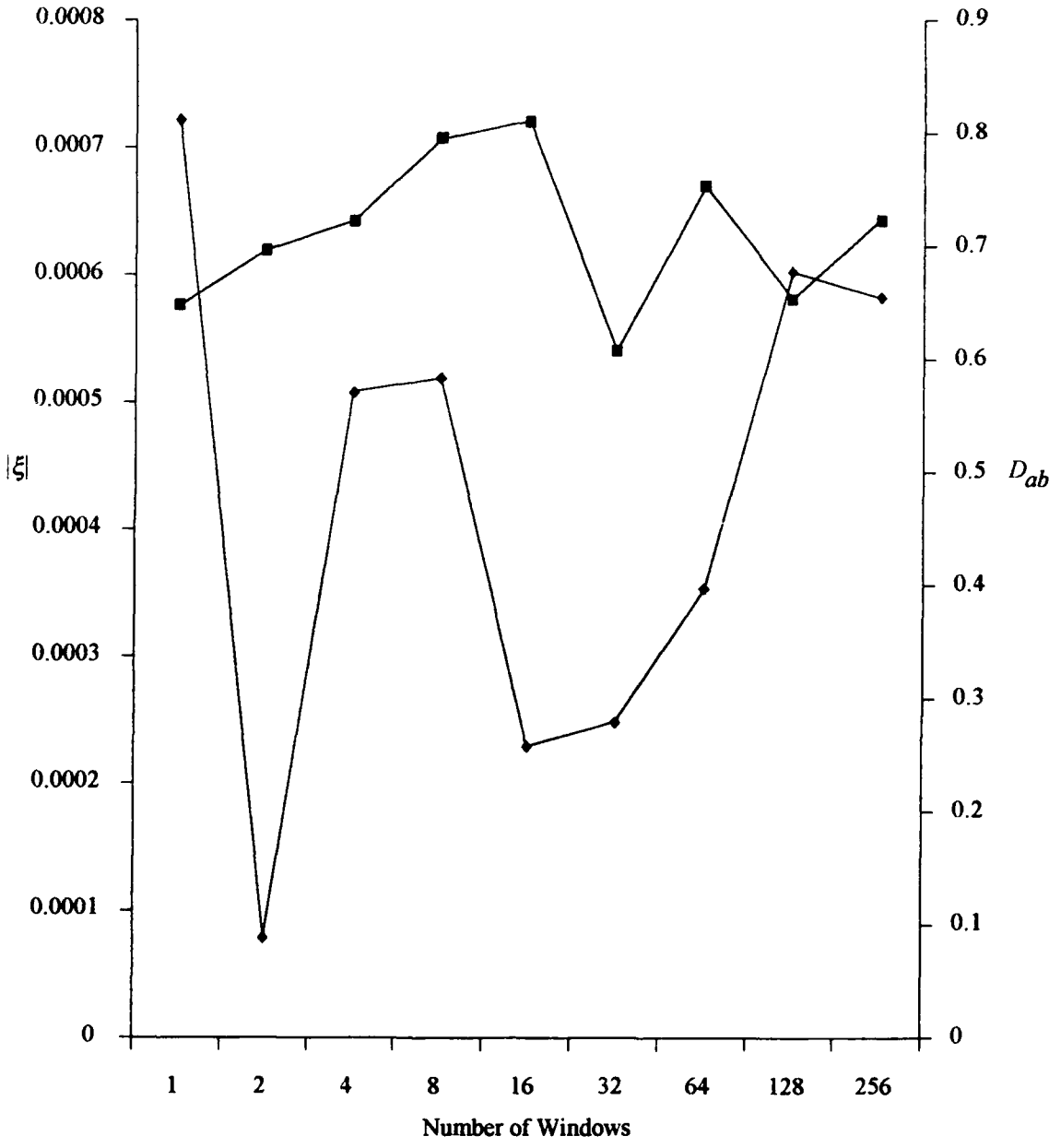


Figure 4.3. Relationship between the mean absolute difference D_{ab} (squares) or the dependence measure $|\xi|$ (diamonds) and the number of windows into which each of the 8 subseries is divided. The total number of points from each subseries is $I/8 = 256$.

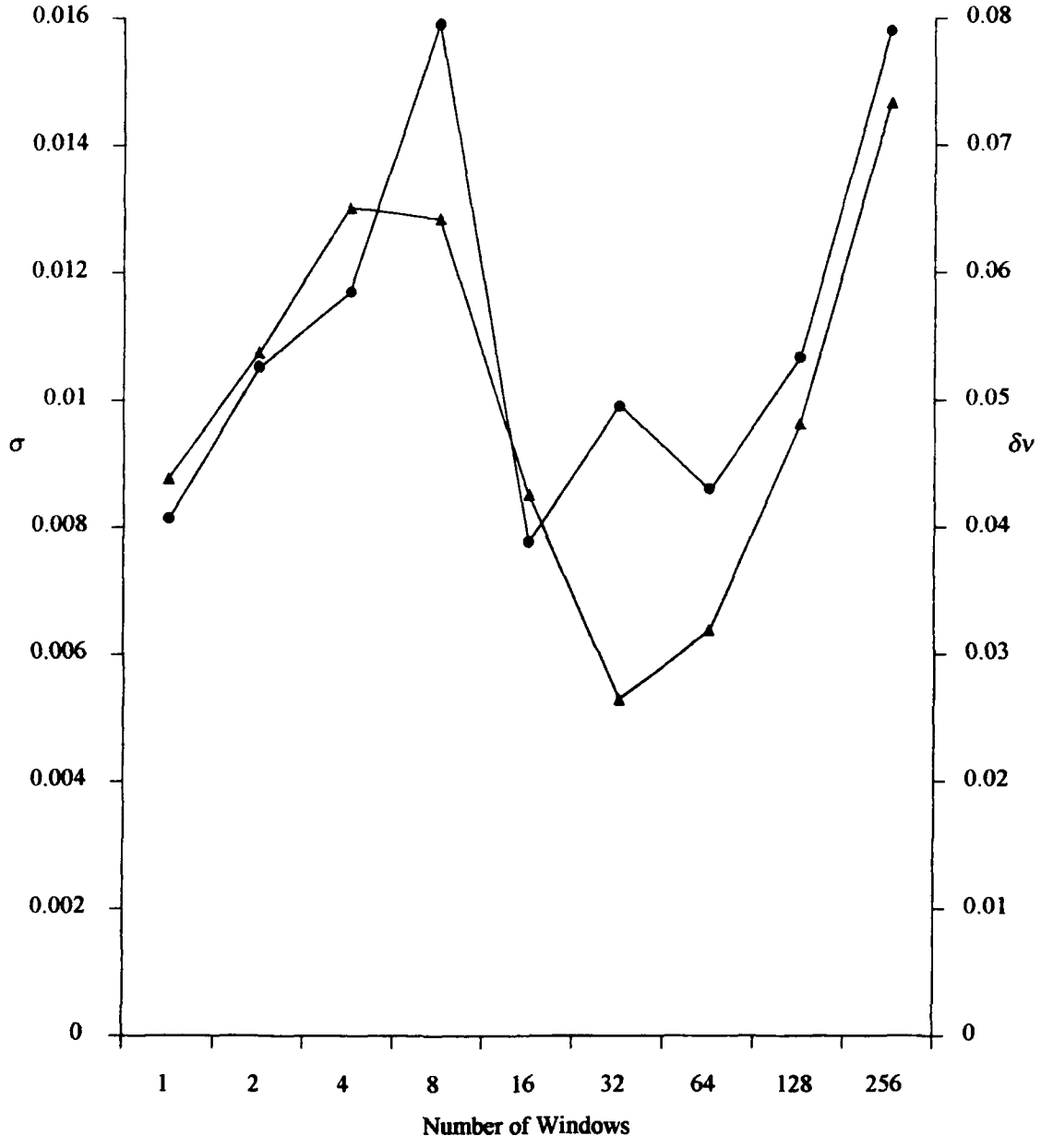


Figure 4.4. Relationship between the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ (circles) or the absolute difference $\delta\nu$ between the estimated and analytical correlation dimensions (triangles) and the number of windows into which the subseries is divided. The total number of points from each subseries is $I/8 = 256$.

observed for this many windows. Although no correlation between δv and D_{ab} or between σ and D_{ab} is observed, the concurrence of the smallest values of δv and D_{ab} along with small values of σ and $|\xi|$ for the same number of windows (32) suggests that this specific strategy is optimal for this number of sampled points. A strong correlation between the accuracy indices δv and σ is observed that is significant at a confidence level of 0.90 to 0.95 ($r^2 = 0.63043$).

The relationship between the mean absolute difference D_{ab} and the number of windows into which the subseries is divided, as well as the relationship between the dependence measure $|\xi|$ and the number of windows, are illustrated in Figure 4.5 for $I/8 = 512$ points. As for the case of $I/8 = 256$, no significant correlation between the mean absolute difference D_{ab} and the number of windows or between the dependence measure $|\xi|$ and the number of windows is observed, although a slight downward trend in $|\xi|$ is evident. Also as before, no correlation between the mean absolute difference D_{ab} and the dependence measure $|\xi|$ is observed. The smallest value of D_{ab} is again observed for 32 windows while the smallest value of $|\xi|$ is observed for 128 windows. The value of $|\xi|$ for 32 windows is among the smaller values of $|\xi|$, and so this sample retains a relatively high degree of independence.

Values of the accuracy indices δv and σ in the relationships with the number of windows are presented in Figure 4.6. Very weak correlations between the two accuracy indices δv and σ and the number of windows are observed, with the accuracy of the

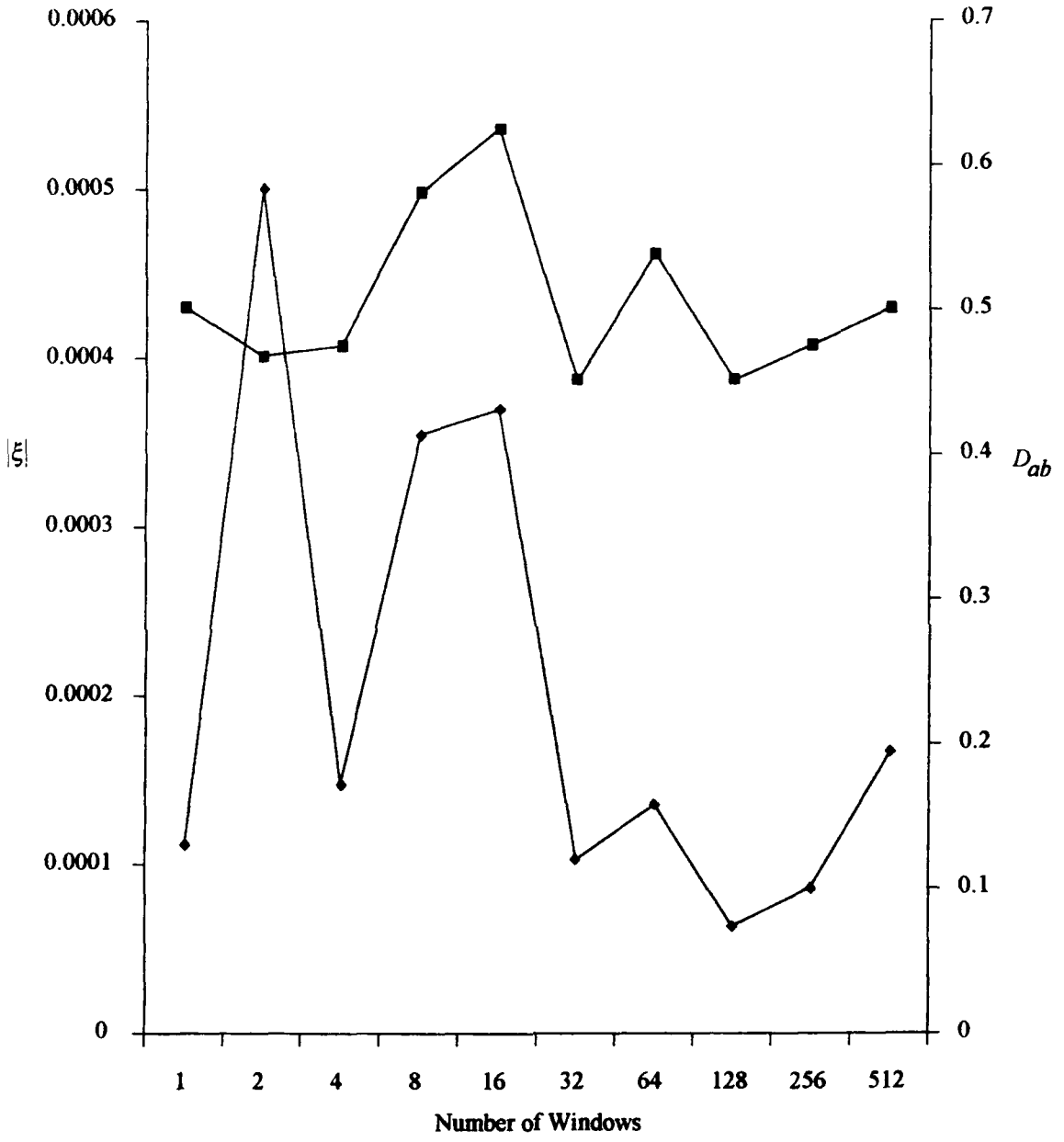


Figure 4.5. Relationship between the mean absolute difference D_{ab} (squares) or the dependence measure $|\xi|$ (diamonds) and the number of windows into which each of the 8 subseries is divided. The total number of points from each subseries is $I/8 = 512$.

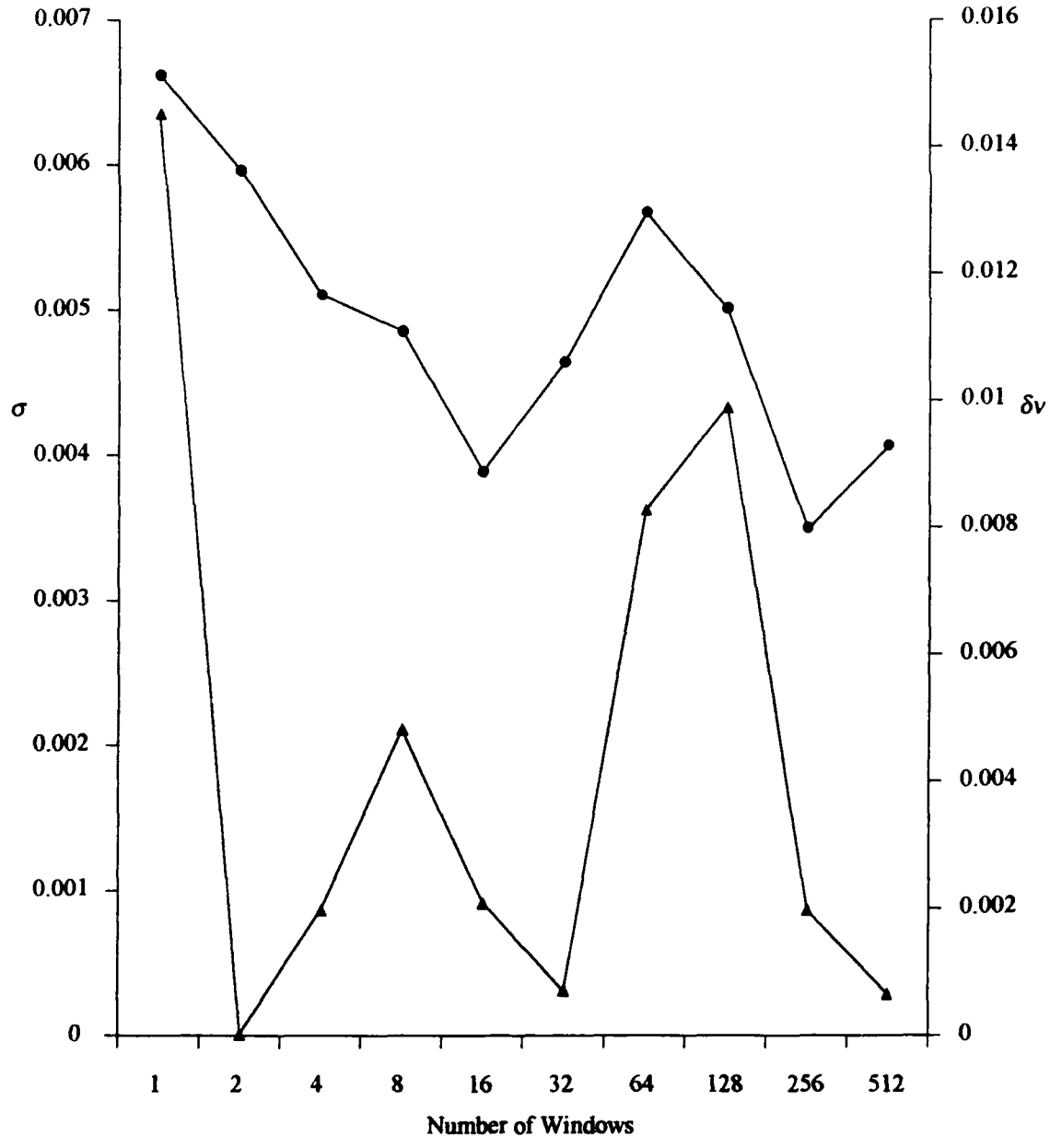


Figure 4.6. Relationship between the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ (circles) or the absolute difference $\delta\nu$ between the estimated and analytical correlation dimensions (triangles) and the number of windows into which the subseries is divided. The total number of points from each subseries is $I/8 = 512$.

estimate of the correlation dimension ν increasing as the subseries is divided into more and more windows. The correlation coefficients for these relationships are $r^2 = 0.24904$ $(\delta\nu)$ and $r^2 = 0.59309$ (σ), respectively. In addition to the overall trends in $\delta\nu$ and σ, several local minima are discernible in the two curves. One of these occurs for the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ for 16 windows, and one occurs at 32 windows for the difference $\delta\nu$ between the estimated and analytical values of the correlation dimension. Although minima in the mean absolute difference D_{ab} and $\delta\nu$ coincide, no overall correlation between D_{ab} and $\delta\nu$ is observed. In addition, no correlation between mean absolute difference D_{ab} and the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ is observed. As occurred for $I/8 = 256$ points, no correlation between the dependence measure $|\xi|$ and σ or between the dependence measure $|\xi|$ and $\delta\nu$ was observed. A significant correlation ($r^2 = 0.36288$) between the two accuracy indices $\delta\nu$ and σ is observed, as with the case $I/8 = 256$ points. This relationship is significant at the 0.90 to 0.95 confidence level. Finally, both the values of $\delta\nu$ and the values of σ are significantly less for $I/8 = 512$ points than for $I/8 = 256$ points. This reflects the greater accuracy of estimates of the correlation dimension using a greater number of points.

4.5. General Conclusions

The results presented in Section 4.4 may be considered along with those discussed in Chapter 3 to produce several general conclusions concerning the utility of different sampling strategies. Although no correlation between the dependence measure $|\xi|$ and the mean absolute difference D_{ab} is observed for $I/8 = 256$ points or $I/8 = 512$ points, correlations were observed for $I/8 = 4096$ points and for $I/8 = 65,536$ points. Thus the identification of the Smale-Williams attractor with a subset of the standard torus by a specified number of points may be better characterized by sampled data with a higher degree of independence, but also requires a minimum amount of data between 512 and 4096 points. The correlation between D_{ab} and $|\xi|$ is not firmly established and requires the sampling of a relatively large number of points. No correlation was observed between the dependence measure $|\xi|$ and either of the indices $\delta\nu$ or σ of the accuracy of estimates of the correlation dimension ν; however all estimates of ν were made for only $I/8 = 256$ points or $I/8 = 512$ points. The number of points required to obtain a correlation between the mean absolute difference D_{ab} and the dependence measure $|\xi|$ is much higher than the maximum number of points that could practically be used to estimate the correlation dimension. Unfortunately, no direct comparison can be made between the relationship between D_{ab} and $|\xi|$, and the relationship between either $\delta\nu$ or σ and $|\xi|$ because the former relationship requires a much greater number of points to establish than can be used to establish the latter.

Strong correlations between $\delta\nu$ and σ were observed for both $I/8 = 256$ points and $I/8 = 512$ points. The standard deviation σ of $\ln F(\varepsilon)$ over the interval $[\varepsilon_1, \varepsilon_2]$ provides an accurate method for determining the value of the correlation dimension ν without prior knowledge of the value of ν, as shown in Table 4.1. The method of Wells *et al.* (1992) provides an accurate estimate of the value of the correlation dimension ν only in the case of $\beta = 0.3$ and requires prior knowledge of the value of ν. The correlation between values of $\delta\nu$ and σ may therefore only occur for $\beta = 0.3$, and the use of $\delta\nu$ as an index of accuracy may be valid for only this value of β. The use of σ as an index of accuracy, however, requires that an interval over which $\ln F(\varepsilon)$ is constant be determined subjectively. This interval may not be discernible for all values of β.

No direct relationship between the accuracy of the representation of the attractor by a histogram, as measured by the mean absolute difference D_{ab}, and the accuracy of the characterization of the attractor by an estimate of the correlation dimension, as measured by $\delta\nu$ and σ, was observed. Some concurrence among the optimal values of each of these indications of accuracy were observed, however. For both $I/8 = 256$ points and $I/8 = 512$ points, minimum values of both the mean absolute difference D_{ab} and the absolute difference $\delta\nu$ were observed when the 8 subseries were divided into 32 windows. The values of the dependence measure $|\xi|$ and the standard deviation σ of $\ln F(\varepsilon)$ on the interval $[\varepsilon_1, \varepsilon_2]$ were also comparatively small for this number of windows. For $I/8 = 4096$ a local minimum of mean absolute difference D_{ab} was observed for 32 windows (128 consecutive points per window followed by a gap of 3968 unsampled

points). Thus, for the particular time series generated for this attractor, an optimal sampling strategy of 32 windows in each of 8 subseries for this time series is indicated. This conclusion does not necessarily imply that this sampling strategy is the specific optimal one for all time series generated by this attractor, but rather that if the total number of sampled points is restricted, then the use of both consecutive points and substantial gaps in the series is necessary to acquire the set of points that best characterizes the attractor.

This general approach to sampling can be used for other time series generated by this and other attractors. If the correlation dimension ν can be determined analytically, then either $\delta\nu$ or σ can be used to assess the accuracy of estimates of ν from sampled data. The correlation dimension for series of observations, for most mathematical dynamical systems such as the Henon attractor, and for meteorological models such as the Lorenz (1963) attractor, can not be known in advance, however. If the correlation dimension ν cannot be known in advance, values of $C(\varepsilon)$ must be calculated for values of ε and these values, along with values for the correlation dimension ν, substituted into (4.2.1) to determine values of $\ln F(\varepsilon)$. These values can be plotted and an interval $[\varepsilon_1, \varepsilon_2]$ over which $\ln F(\varepsilon)$ is constant can be determined. Standard deviations σ may then be determined and the value of ν that yields the smallest value of σ will be the best estimate of the correlation dimension. This approach works very well even though the graph of ν versus ε does not show a plateau at the analytic value for ν.

It is apparent that the characterization of an attractor by either a histogram or by an estimated correlation dimension for samples of data is affected by the strategy with which that data is sampled. The interpretation of this property is somewhat subjective and requires consideration of several factors, including the dependence of the data within the sample, the mean absolute difference between the histogram of the sampled data and the histogram of the control series, and the accuracy of estimates of the correlation dimension from the sampled data. If these factors are considered after several sampling strategies are applied to the time series data, then an optimal sampling strategy may be selected. The advantage of choosing an optimal sampling strategy is greatest when only a small number of points is available. The results of the experiments with Type II sampling strategies indicate that a far greater advantage is achieved if the number of sampled points is increased. It is more likely that if an observed natural variable (such as oxygen isotope ratio) or if model output is used to determine a correlation dimension, then the number of available points may be limited and the number of points that can be used to perform the calculation may be restricted. Therefore, an optimal sampling strategy should be used. The specific strategy for a particular time series can only be determined subjectively, but the results presented here indicate that some combination of consecutive points separated by substantial gaps in the data is optimum. It may be necessary to sample consecutive points to retain details of the fine structure of the attractor represented by the time series, while gaps of unsampled data in the time series may be necessary to ensure that sampled data are independent and so provide the greatest amount of new information.

REFERENCES

Baker, G. L. and J. P. Gollub, 1990: *Chaotic Dynamics: An Introduction.* Cambridge University Press, Cambridge, 192 pp.

Brandstäter, A., J. Swift, H. L. Swinney, A. Wolf, J. D. Farmer, E. Jen, and P. J. Crutchfield, 1983: Low dimensional chaos in a hydrodynamic system. *Phys. Rev. Lett., 51,* 1442-1445.

Doran, J. D., 1991: *Adequate sampling of a chaotic time series.* M. S. Thesis, Department of Meteorology, The Pennsylvania State University, University Park, PA, 172 pp.

Essex, C., T. Lookman, and M. A. H. Nerenberg, 1987: The climate attractor over short time scales. *Nature, 326,* 64-66.

Falconer, K. J., 1990: *Fractal Geometry. Mathematical Foundations and Applications.* John Wiley and Sons, New York, 288 pp.

Fraedrich, K., 1987: Estimating weather and climate predictability on attractors. *J. Atmos. Sci., 44,* 722-728.

Grassberger, P., 1986: Do climatic attractors exist? *Nature, 323,* 609-612.

Grassberger, P. and I. Procaccia, 1983: Measuring the strangeness of strange attractors. *Physica, 9D,* 189-208.

Guckenheimer, J. and G. Buzyna, 1983: Dimension measurements for geostrophic turbulence. *Phys. Rev. Lett., 51,* 1438-1441.

Henderson, H. W. and R. Wells, 1988: Obtaining attractor dimensions from meteorological time series. *Adv. Geophys., 30,* 205-237.

Hense, A., 1987: On the possible existence of a strange attractor for the southern oscillation. *Contr. Atmos. Phys., 60*, 34-47.

Higgins, R. W., 1987: From the equations of motion to spectral models. Chapter 3 in *Nonlinear Hydrodynamic Modeling: A Mathematical Introduction*, H. N. Shirer (Ed.), *Lecture Notes in Physics, 271*, Springer-Verlag, Heidelberg, 47-69.

Krishna-Mohan, T. R., J. Subba-Rao, and R. Ramaswamy, 1989: Dimension analysis of climatic data. *J. Climate, 2*, 1047-1057.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci., 20*, 130-141.

Lorenz, E. N., 1984: Irregularity: A fundamental property of the atmosphere. *Tellus, 36A*, 98-110.

Mori, H., and H. Fujisaka, 1980: Statistical dynamics of chaotic flows. *Prog. Theor. Phys., 63*, 1931-1944.

Nese, J. M., 1987: Diagnosing the structures of attractors. Chapter 16 in *Nonlinear Hydrodynamic Modeling: A Mathematical Introduction*, H. N. Shirer (Ed.), *Lecture Notes in Physics, 271*, Springer-Verlag, Heidelberg, 412-443.

Nese, J. M., J. A. Dutton, and R. Wells, 1987: Calculated attractor dimensions for low-order spectral models. *J. Atmos. Sci., 44*, 1956-1970.

Nicolis, C. and G. Nicolis, 1984: Is there a climatic attractor? *Nature, 311*: 529-532.

Pesin, Ya. B., 1992: On rigorous mathematical definitions of correlation dimension. Submitted to *J. Stat. Phys.*

Petersen, K. E., 1971: *Introductory Ergodic Theory Lecture Notes*. Department of Mathematics, University of North Carolina, Chapel Hill, 141 pp.

Smale, S., 1977: Dynamical systems and turbulence. In *Turbulence Seminar, Berkeley 1976/77*, pp. 48-70, A. Dold and B. Eckmann (Eds.). *Lecture Notes in Mathematics, 615*, Springer-Verlag, Heidelberg.

Sparrow, C., 1982: *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Applied Mathematical Series, 41*, Springer-Verlag, Heidelberg, 269 pp.

Tsonis, A. A. and J. B. Elsner, 1989: Chaos, strange attractors, and weather. *Bull. Am. Met. Soc., 70,* 14-23.

Wells, R., H. N. Shirer, C. Fosmire, and J. A. Doran, 1992: Improved algorithms for calculating the correlation dimension and for estimating the associated probable errors. Submitted to *Phys. Rev.*