available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf

Visualising and assessing the probable error from downscaling ecological time series data

P.A. Whigham

Deparment of Information Science, PO Box 56, University of Otago, Dunedin, New Zealand

ARTICLE INFO

Article history:
Received 18 December 2004
Received in revised form
26 January 2006
Accepted 8 February 2006

Keywords:
Univariate time series
Interpolation
Downscaling
Multi-scale

ABSTRACT

Collecting natural data at regular, fine scales is an onerous and often costly procedure. However, there is a basic need for fine scale data when applying inductive methods such as neural networks or genetic algorithms for the development of ecological models. This paper will address the issues involved in interpolating data for use in machine learning methods by considering how to determine if a downscaling of the data is valid. The approach is based on a multi-scale estimate of errors. The resulting function has similar properties to a time series variogram; however, the comparison at different scales is based on the variance introduced by rescaling from the original sequence. This approach has a number of properties, including the ability to detect frequencies in the data below the current sampling rate, an estimate of the probable average error introduced when a sampled variable is downscaled and a method for visualising the sequences of a time series that are most susceptible to error due to sampling. The described approach is ideal for supporting the ongoing sampling of ecological data and as a tool for assessing the impact of using interpolated data for building inductive models of ecological response.

1. Introduction

Sampling ecological time series data inevitably leads to questions regarding whether the appropriate sampling rate has been used to detect the fundamental processes of the system being measured. This question has implications in terms of the accuracy of the measured information, and the likely error that is associated with the sampling scheme being applied. Although methods such as the Fourier transform and wavelets can detect the fundamental frequencies in a time series, they are not easily applied to address the question of scale and sampling rate.

This paper introduces a simple approach that allows a quantitative and visual representation of the change in information content of a time series as the scale of the series is changed. The resulting measure may be used to detect the fundamental scales at which processes are operating in the data and, under some circumstances, is sensitive to scales smaller than the currently measured scale of the time series. This method may easily be incorporated into a sampling regime to determine whether the appropriate scale for sampling is being used and can give an estimate on the error introduced by downscaling through interpolation. Since machine learning techniques often require large amounts of data to learn a suitable model, this method can be used to estimate the validity of an interpolation procedure when downscaling is applied, and hence whether the estimates of model error are likely to be valid.

This paper is structured as follows: Section 2 gives background on related matters such as missing data and signal analysis with Section 3 describing the basic method of this paper. Following this introduction, a variety of experiments with different data characteristics are described: Section 4.1 demonstrates the approach for a random sequence, Section 4.2 a periodic sequence and Section 4.3 a chaotic time series. By altering the behaviour of these sequences information regarding properties of the time series and likely error are described. The consequences of these

E-mail address: pwhigham@infoscience.otago.ac.nz.

1574-9541/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2006.02.006

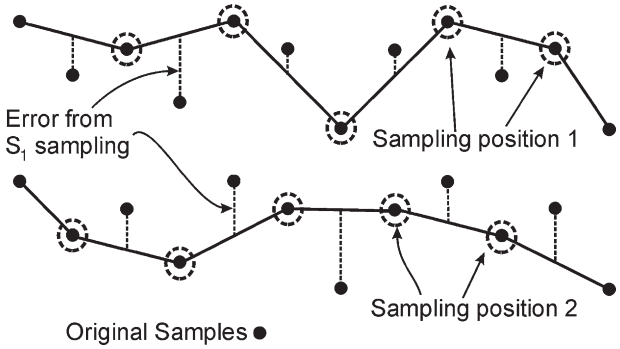

Fig. 1 – Multiple resampling and error calculation for scale S_1.

conclusions are explored in Section 5 where a variety of errors from real data are described. Finally, Section 6 demonstrates the use of the method as a tool for assessing whether data has been sampled at an appropriate scale, and therefore that interpolation between data points is valid.

2. Background

The issue of downscaling a time series dataset (say from weekly to daily samples) can be considered as an example of the prediction of missing data. This field has been extensively studied in the statistical literature, and a variety of methods have been developed. The reader is referred to the review by Scafer and Graham (2002) for an introduction to methods such as single imputation, maximum likelihood and multiple imputation. The work described here does not consider this complex problem, but addresses the simpler notion as to whether some estimate of error from downscaling can be determined, in particular for a univariate time series which has been regularly sampled. Since in principle it is not possible to know the true error produced by a particular downscaling procedure for a univariate time series, the approach described here is just one estimate of error, based on the structure of the time series at the original sampling scale.

The Nyquist–Shannon sampling theory, a fundamental theory from communication theory, states that to be able to accurately reconstruct an analog signal from a discrete sampling the sampling rate is required to be twice the highest frequency in the signal (Shannon, 1949). However, in the circumstance where reconstruction is not required, it is sometimes possible to detect the presence of higher frequency signals, which in our case can be used as information to assess whether an error will be introduced by downscaling.

Sophisticated methods for extracting frequency information from a signal have been developed in the past, including the Fourier transform and wavelet analysis (Combes et al., 1989; Resnikoff and Wells, 1998). These methods are complex and generally require some user interaction to select parameters for the analysis. Although these methods allow an accurate assessment of frequency and time-frequency information, they do not directly allow an assessment of the error associated with downscaling. The method described in this paper is a simplistic approach to frequency detection that does allow this downscaling error to be estimated.

3. Methodology

The following multi-scale approach is based on the concept introduced by Mandelbrot to address the question of the length of a coastline (Mandelbrot, 1967); however, we do not consider the fractal dimension of the time series. Rather, the time series is repeatedly resampled at all scales and then, for each scale, interpolated to produce an error measurement against the original data. The error measurement is constructed against the original sequence as the sampling scale Si is increased from 0 (the original data) to $N/2$, where N is the number of points in the sequence. Since for any scale S_i there are $i+1$ possible combinations of measurements, these are used to give the final error and standard deviation against the original sequence. For every sampled scale, the sampled points are interpolated back to the scale of the original data. The resulting function defined by this error measurement will be defined as $\lambda(\phi(S,T))$, where S is the scale, T is the original sequence of n sampling steps and $\phi(S,T)$ is the interpolated sequence, based on the sampling rate S. The standard deviation of this error is defined as $\lambda_\sigma(\phi(S,T))$. This method is also similar to the estimate of a time series variogram (Chatfield, 1989); however, the novelty of the approach is the interpolation of the resampled sequence back to the original scale and the subsequent measure of error versus the original data.

Linear interpolation, based on the two nearest measured neighbours, will be used for all interpolation experiments. This has been selected as it represents a weak, local interpolation method that is likely to give a good indication of the error due to interpolation from a smaller number of

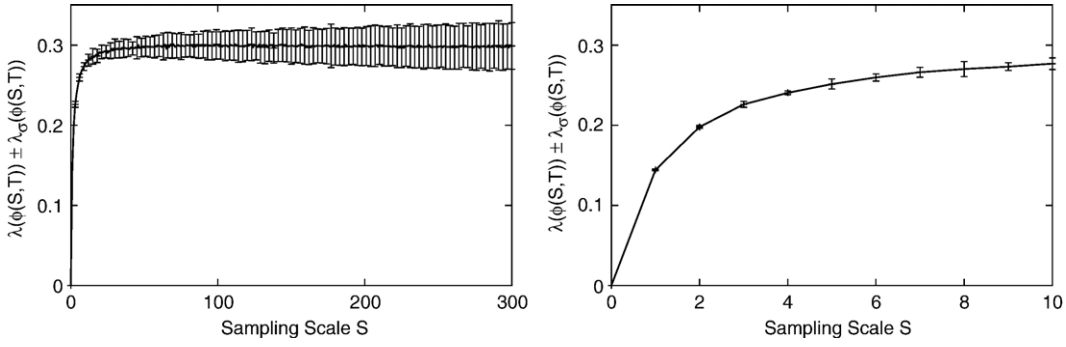

Fig. 2 – $\lambda(\phi(S,T)) \pm \lambda_\sigma(\phi(S,T))$ for a random uniform distribution.

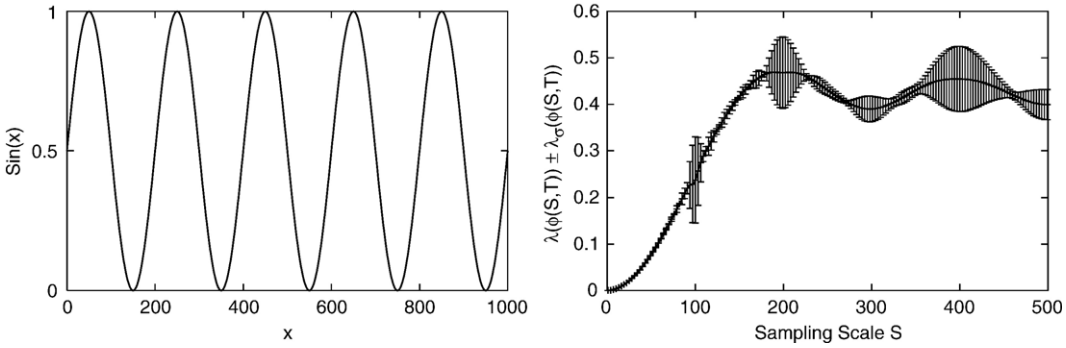

Fig. 3 – Original sin curve with five periods and $\lambda(\phi(S,T)) \pm \lambda_\sigma(\phi(S,T))$.

samples. Other methods for interpolation are likely to perform at least as well as linear interpolation, and hence this gives some expected worse case.

Root mean square error (RMSE) is used for all comparisons between the original and scaled, reinterpolated sequences. Although there are many possible ways to measure sequence similarity (Bollabas et al., 1997; Bridge, 1998) RMSE was selected as it biases towards larger errors, and hence will detect small changes in the compared sequences. Additionally, RMSE was measured against only those points that were interpolated, rather than the entire sequence. Hence, for a sequence where $N=12$, for S_1 there are $(12-2)/2=5$ predicted points used in the calculation of RMSE. Note also that the first and last points of the original sequence are always kept as samples, and therefore do not contribute to the error estimate. Fig. 1 shows the two possible resamplings of a time series that can occur for the first scale step S_1. The error created by this scale change is determined by a linear interpolation between the sampling positions to produce a new sequence with the original number of sample points. The RMSE is then determined based on the distance from each original sampling point and the interpolated points derived from the resampling at each sampling position. The standard deviation $\lambda_\sigma(\phi(S,T))$ is calculated from the variance in RMSE for each sampling position.

4. Properties of the multi-scale approach

This section will demonstrate the approach on random and periodic sequences to illustrate how the method behaves under a variety of simple sequence patterns.

4.1. *Random sequences*

To illustrate the properties of the approach a random sequence of 1000 values, with uniform distribution between 0 and 1, was generated as a univariate time series. The resulting $\lambda(\phi(S,T))$ for the random sequence is shown in Fig. 2. Since there is no information in the sequence (the sequence is random), there is a constant rate of error once $S_i \approx 10$. At this point, there is no further loss of information in the sequence, showing that a random sequence has no fundamental frequencies and rapidly converges at all scales. The subsequent interpretation of real sequences will show that, once S_i increases beyond the scale at which the processes generating the sequence occur, the error will level out. Note that, over the first 10 samples, $d\lambda/dS \approx 0.035$. This rate of change is proportional to the loss of information as the scale is increased, and hence this value shows an expected maximum loss for normalised data as scale increases. Fig. 2, focussing on S1–S10, shows the rapid approach to measuring no information in the sequence. This highlights the detection of a sequence with no pattern. Hence, $\lambda(\phi(S,T))$ can be used as a measure of the randomness in a sequence.

4.2. *Periodic sequences*

This example is based on a sin curve that goes through five complete periods. This will demonstrate how the method can be used to detect the fundamental frequencies in a sequence. Fig. 3 shows the method detecting the regular frequencies that occur at a variety of scales based

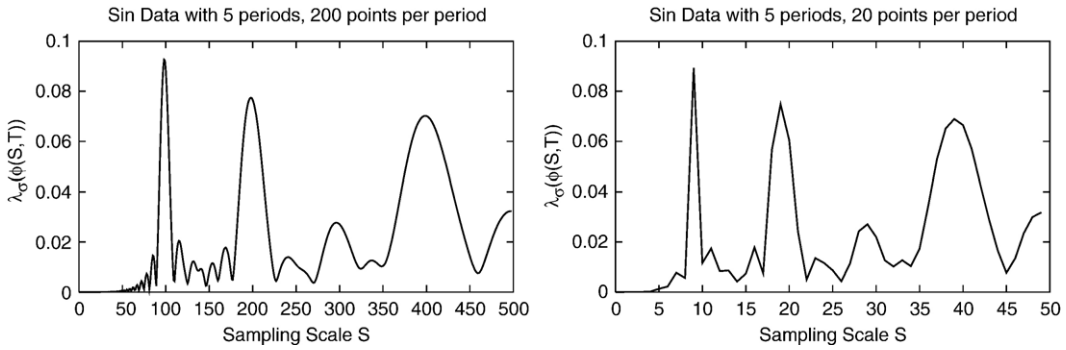

Fig. 4 – $\lambda_\sigma(\phi(S,T))$ for five-period sin curve for 200 and 20 points per period.

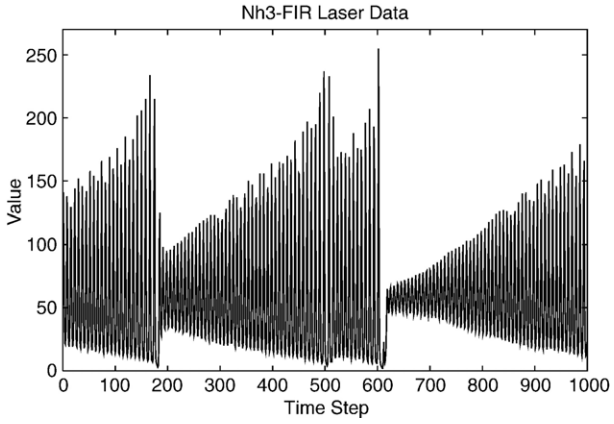

Fig. 5 – The chaotic laser data sequence.

on the sin curve shown with 5 periods. This curve has been sampled for 1000 points, corresponding to 200 points per period. The main indication that a critical scale length is being approached is that $\lambda_\sigma(\phi(S,T))$ begins to increase, as highlighted in Fig. 4. Note that the increase in λ_σ occurs around the scale at which each period of the data occurs. The influence on sampling rate is also shown in Fig. 4, where the original sin curve has been sampled at a rate of 20 points per period. Note that, although some detail has been lost (as would be expected), the major frequencies in the data are still detected at the same scales. Note that, over the first 100 scale steps, $d\lambda/dS \approx 0.002$, which is significantly lower than the random sequence.

4.3. *Chaotic sequences*

The laser dataset for NH3-FIR lasers (Hubner et al., 1994) was chosen as an example chaotic sequence. The sequence is shown in Fig. 5. The resulting $\lambda(\phi(S,T))$ and $\lambda_\sigma(\phi(S,T))$ are shown in Fig. 6. $\lambda_\sigma(\phi(S,T))$ indicates a major frequencies around 100, 300 and 420, with many low frequency scale processes also operating. Note that $d\lambda/dS \to 0$ rapidly showing characteristics of a random sequence with embedded frequency information. This indicates that $d\lambda/dS$ appears to be acting as a measure of the information loss as S_i increases.

4.4. *Properties of $\lambda(\phi(S,T))$*

The following properties exist for $\lambda(\phi(S,T))$:

- $\forall\lambda(\phi(S,T))$, $dT/dn = c \Rightarrow \lambda(\phi(S,T)) = 0$. All straight lines have no scale information;
- $\forall\lambda(\phi S,T))$, $d\lambda/dS = 0 \Rightarrow$ no scale information;
- $d\lambda/dS \propto$ rate of loss of scale information;
- $d\lambda/dS \to \infty$ implies no pattern as scale increases;
- $d\lambda/dS \to 0$ implies no loss of information as scale increases;
- As $n \to N/2$, $d\lambda/dS \to 0$;
- $d\lambda_\sigma(\phi(S,T))/dS < \varepsilon$ implies that increasing scale does not change the amount of variation in the error. Hence, if $d\lambda/dS = c$, then there is a gradual lose of accuracy with increasing scale; however, no significant processes are detected for these scales.

5. Real data sequences

This section will demonstrate the multi-scale approach to the analysis of two time series. Initially the approach will be used to show how the various periods within the data can be detected, and finally how the approach can infer the influence of periods within the data that are at a finer scale than originally sampled.

5.1. *Lake Waihola data*

The datasets used for this example were collected by Dr. Marc Schallenburg as part of a study on the characteristics of Lake Waihola, a partly sea influenced lake 30 km south of Dunedin, New Zealand. The data was collected at 5-min intervals from 4th June 1999 at 13:07 till the 8th June 1999 at 11:42, for a total number of 1136 samples. Data collected included temperature, dissolved oxygen, conductivity and water depth. The temperature and conductivity datasets are shown in Fig. 7. Note that the x-axis has been labelled with a sample number, since no assumption is made about the rate at which the samples are taken (however, there is an assumption that the samples occur regularly).

Visually, the temperature data from Fig. 7 shows a weak periodic signal, with a downward trend for the first half of the sequence, followed by a more level trend for the second half of the sample, whereas the conductivity data exhibits at least

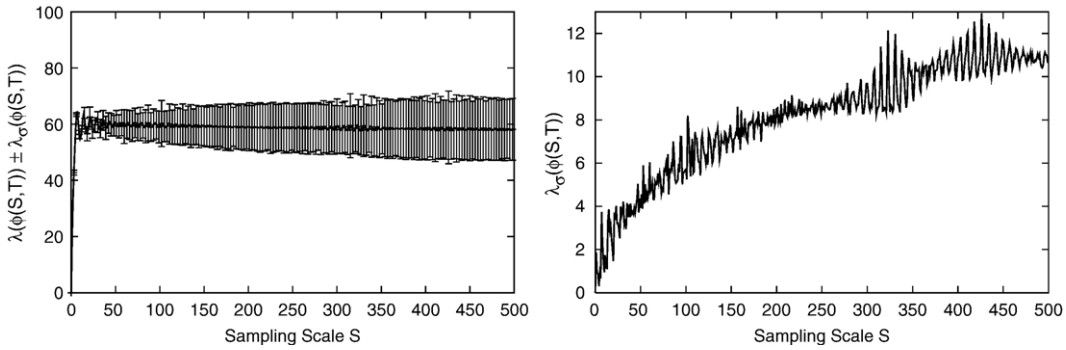

Fig. 6 – $\lambda(\phi(S,T))$ and $\lambda_\sigma(\phi(S,T))$ for the chaotic laser data.

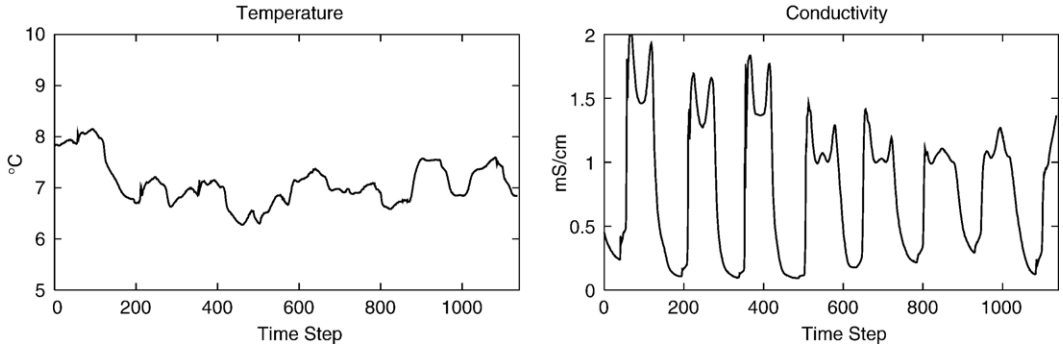

Fig. 7 – Temperature and conductivity data from Lake Waihola.

two clear periods—one at a rate of around 150 steps and a second periodic signal that occurs at a rate of approximately 75 time steps.

The resulting multi-scale analyses of these time series are shown in Figs. 8 and 9. It is clear that there is a large amount of information contained in the sequences, and that the behaviour of conductivity and temperature vary over many different scales. For example, there is a strong pulse with conductivity around the sampling scale of ≈75, corresponding to a period of 6 h. This is the pulse related to the connectivity of Lake Waihola to the sea mouth, making the lake partly tidal. This pulse is also apparent in the temperature data, although a more significant period occurs at ≈150, corresponding to a period of 12 h—the diurnal cycle. Over the first 150 sample sizes, the normalised $d\lambda/dS \approx 0.002$ for temperature, whereas for conductivity this is ≈0.004. This implies that the conductivity sequence contains more information at lower scales than temperature.

Finally, a closer look at the standard deviation for conductivity, as shown in Fig. 10, shows that there are a number of smaller periods for this dataset, and that there is a large amount of information contained in this data. This demonstrates that the multi-scale analysis of error can be used to determine fundamental periods of the data, and to show the information content of the data at all scales. Note that five scale peaks (A–E) have been identified in Fig. 10. These peaks will be considered further in Section 5.2.

5.2. *The effect of coarse sampling*

This section will demonstrate the results of applying the multi-scale approach to a coarse sampling of the previous

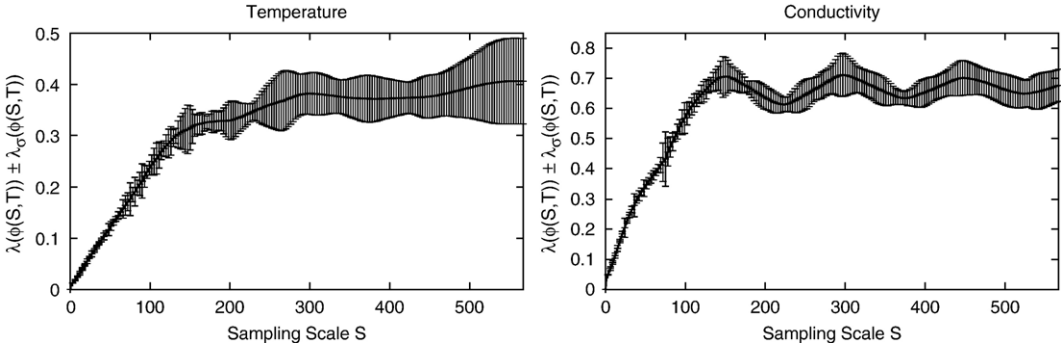

Fig. 8 – $\lambda(\phi(S,T))$ for Waihola temperature and conductivity data.

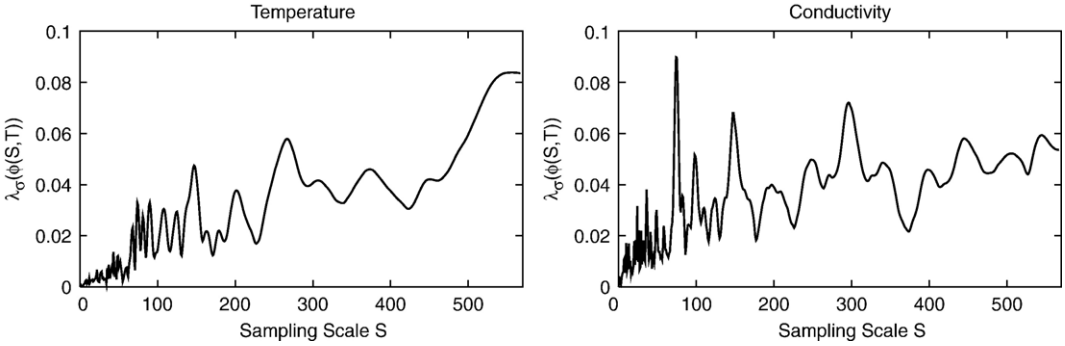

Fig. 9 – $\lambda_\sigma(\phi(S,T))$ for temperature and conductivity data.

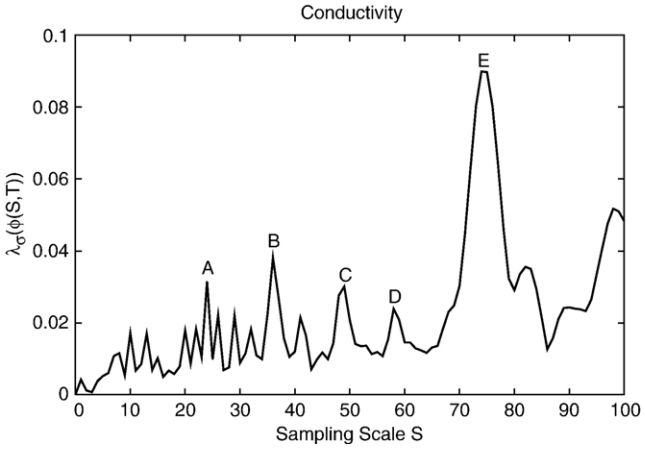

Fig. 10 – $\lambda_\sigma(\phi(S,T))$ for conductivity for S1–S100.

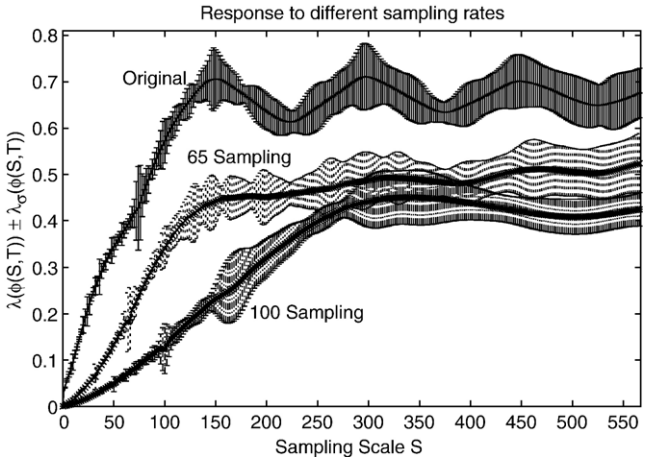

Fig. 12 – $\lambda(\phi(S,T))$ for the conductivity data at 65 and 100 sampling rates.

conductivity dataset. The results will show that the smaller scale variations in the data can be detected even though the dataset has not been sampled *at this scale*.

The original Waihola conductivity dataset was resampled at intervals of 65 and 100 steps, with a linear interpolation between neighbouring points used to reproduce the resulting full 5-min sequences. These sequences, along with the original conductivity data, are shown in Fig. 11, which shows that the majority of information has been lost due to the resampling and interpolation. However, when the multi-scale analysis is applied to each of these resampled series, some interesting details become apparent. It would be expected that the multi-scale analysis would find a peak of standard error about the scale at which the resampling occurred, but other, lower scale, information shows up with this analysis. The $\lambda(\phi(S,T))$ plots for the original, 65 and 100 resamplings are shown in Fig. 12. The resulting standard deviation plots $\lambda_\sigma(\phi(S,T))$ for each resampling are shown in Fig. 13.

Fig. 13 shows the first 100 sampling scales for both resamplings, which should be compared with the original plot from Fig. 10. The most important features of Fig. 13 are the identification of the peaks A…E and the largest peak at the sampling rate of 65. Peaks A, B, D and E have been identified, although shifted down or up in scale slightly, whereas peak C appears as a cluster of scale changes. The fact that the peaks A…D were identified, although the sampling rate was at a greater scale, shows some promise in the method identifying subscale detail in the data. Note that the position of the labels on the figures is in the same horizontal position as the original sequence (see Fig. 10), although they are shifted down to meet the curve. Fig. 13 also shows some detection of the sampling scale changes for the 100 resampling, even though all of the peaks are at a lower scale than the sampled data. Note that peaks A and B cannot be entirely resolved, and that peaks D and E have shifted. Once again, however, this type of pattern would indicate that it is likely that there is further scale information at lower sampling rates and that this sampling rate is not appropriate for the variation in the data.

6. Estimating the error from downscaling time series

The previous section has shown that information can be detected at lower scales than the sampling rate, and therefore the question arises as to an estimate of the error associated with this downscaling. Since the scaling method assumes that a linear fit is done for downscaled points, the determination of $\lambda(\phi(S,T))$ for a lower scale will be an estimate of the error associated with this downscaling. A lower bound on the error associated with the downscaling from scale i is therefore $\hat{\lambda}(\phi(S_i,T)) \pm \hat{\lambda}_\sigma(\phi(S_i,T))$. For example, if the original data for conductivity was the 65 resampling, the error associated with downscaling to the original scale is 0.18±0.08. The actual RMSE between the original and 65 resampling sequence is 0.39. Hence, this approach can be used by taking the current sampled dataset and producing a downscaling using a linear interpolation to produce a larger number of sample points. The resulting value of λ, back at the true sampling rate, gives

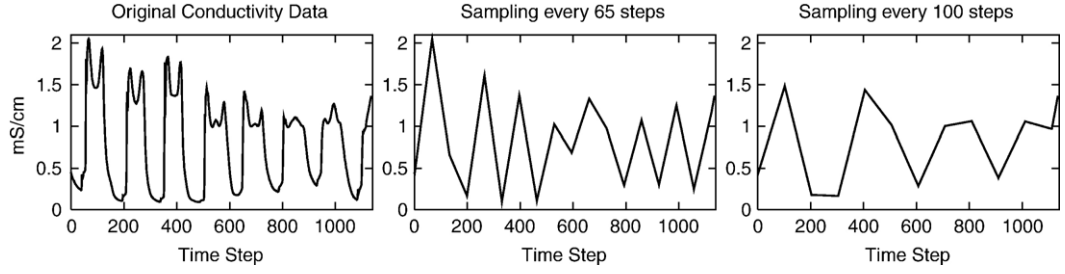

Fig. 11 – The Waihola conductivity data resampled at 65 and 100 sampling rates.

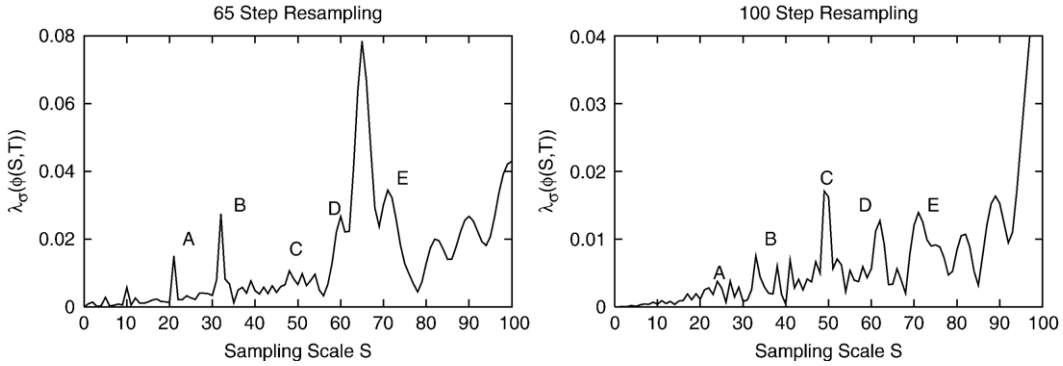

Fig. 13 – $\lambda_\sigma(\phi(S,T))$ for the 65 and 100 step resampling.

this estimate of error. In addition, peaks in λ_σ below the true sampling rate imply that the data contains information at lower scales, and therefore increasing the rate of sampling should be considered.

6.1. *Use of $\lambda(\phi(S,T))$ with resampled data—Lake Kasumigaura*

Lake Kasumigaura is situated in the southeastern part of Japan. It is a large, shallow water body where no thermal stratification occurs. Water temperatures vary widely, ranging from 4 °C in winter to over 30 °C in summer. The lake has high nutrient loadings and therefore phytoplankton abundance is high for the majority of the year. Given the reliance on light and temperature for growth, there are clear seasonal patterns in the data. The data was collected at a weekly sampling rate and has been downscaled to a daily dataset for use with a variety of machine learning algorithms (Recknagel et al., 1998, 2000). The multi-scale analysis for a common model predictor variable, chlorophyll-*a*, is shown in Fig. 14. Since the downscaling was from weekly to daily, there were 6 new points created between every sampled point. Hence, the lower bound error for the downscaling of 310 the chlorophyll-*a* data is given by $\hat{\lambda}(\phi(6,T)) \pm \hat{\lambda}_\sigma(\phi(6,T)) = 1.52 \pm 0.03$. Note that this is an approximate lower bound on the error. One interesting note is that λ_σ shows a peak around S_8, implying that the data may not have been regularly sampled each week.

A description of each variable collected for this dataset, and the corresponding downscaling errors, are shown in Table 1. One variable that shows a great deal of error is light. The plots of $\lambda(\phi(S,T))$ and $\lambda_\sigma(\phi(S,T))$ for light are shown in Fig. 15. Note the similarity in these patterns to the chaotic plots of Section 4.3. This would suggest that this data should not be used as input to a model, since it does not carry information at small scales that behave with a similar relative error to the other variables. Note that the 6-month and 12-month scales are identified

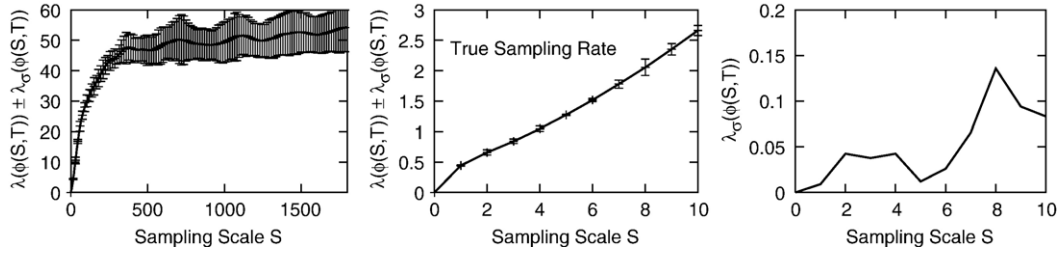

Fig. 14 – Lake Kasumigaura chlorophyll-*a* $\lambda(\phi(S,T))$.

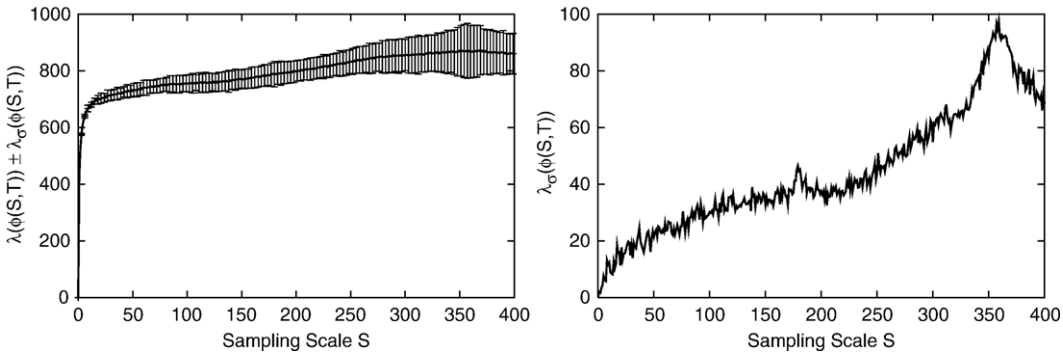

Fig. 15 – $\lambda(\phi(S,T))$ and $\lambda_\sigma(\phi(S,T))$ for light.

Table 1 – Lake Kasumigaura water quality variables and downscaling error

Variable	Ave. ± S.D.	$\hat{\lambda}(\phi(6,T)) \pm \hat{\lambda}_\sigma(\phi(6,T))$	Units
Ortho phosphate	14.14 ± 25.71	1.38 ± 0.3	mg/l
Nitrate	520.56 ± 503.4	8.85 ± 0.22	mg/l
Secchi depth	85.43 ± 44.57	0.94 ± 0.1	cm
Dissolved oxygen	11.2 ± 2.14.0	14 ± 0.02	mg/l
pH	8.74 ± 0.59	0.027 ± 0.003	–
Water temperature	16.36 ± 7.79	0.12 ± 0.009	°C
Light	1280 ± 670	640 ± 6	mJ/cm2
Rotifera	229 ± 293	20.2 ± 4.7	Ind/l
Cladocera	170 ± 222	30.85 ± 7.3	Ind/l
Copepoda	156 ± 84	6.65 ± 0.5	Ind/l
Microcystis	38563 ± 95,216	2649 ± 257	Ind/l
Oscillatoria	20160 ± 53,483	1329 ± 229	Ind/l
Anabaena	6008 ± 16,083	538 ± 26	Ind/l
Chlorophyll-α	74.43 ± 42.51	1.52 ± 0.03	μg/l

from λ_σ in Fig. 15; however, the displayed patterns suggest that the sampling of this variable has not been consistent (same time each day), or that the variable under consideration has a chaotic characteristic.

7. Conclusions

The approach described here allows a multi-scale assessment of uniformly sampled time series data. By producing a plot of $\lambda(\phi(S,T))$ and $\lambda_\sigma(\phi(S,T))$, the characteristics of a dataset can be determined and can be used to assess whether the sampling rate is appropriate for the variation in the data. The approach is simple to implement, and allows a straightforward assessment of scale properties. Additionally, $\lambda(\phi(S,T))$ may be used to estimate a lower bound on the error from downscaling a dataset and, in conjunction with $\lambda_\sigma(\phi(S,T))$, can be used to assess the appropriateness of any variable as input to a model.

Acknowledgements

The author would like to thank Dr. Frederick Recknagel and the Adelaide University for support during this work. Dr. Marc Schallenburg must also be thanked for supplying the conductivity and temperature data used in this paper. Dr. Bob McKay from the University of New South Wales must also be thanked for reading an early draft of this paper. This work was conducted during study leave from the Department of Information Science, University of Otago, New Zealand.

REFERENCES

Bollabas, B., Das, G., Gunopulos, D., 1997. Time-series similarity problems and well-separated geometric sets. Proc. of the 13th Annual Symposium on Computational Geometry, Nice, France, pp. 454–456.

Bridge, D., 1998. Defining and combining symmetric and asymmetric similarity measures. In: Smyth, B., Cunningham, P. (Eds.), EWCBR-98, Lecture Notes in Artificial Intelligence. Spinger-Verlag, Berlin Heidelberg, pp. 52–63.

Chatfield, C., 1989. The Analysis of Time Series: An Introduction. Chapman and Hall.

Combes, J.M., Grossmann, A., Tchamitchian, P. (Eds.), 1989. Wavelets. Springer-Verlag, Berlin.

Hubner, U., Weiss, C.O., Abraham, N.B., Tang, D., 1994. Lorenz-like chaos in NH3-FIR lasers. In: Weigend, A.S., Gershenfeld, N.A. (Eds.), Time Series Prediction: Forecasting the Future and Understanding the Past. Addison-Wesley, MA, pp. 73–104.

Mandelbrot, B., 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156, 636–638.

Recknagel, F., Fukushima, T., Hanazato, T., Takamura, N., Wilson, H., 1998. Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumigaura by artificial neural network. Lakes and Reservoirs: Research and Management 3, 123–133.

Recknagel, F., Bobbin, J., Whigham, P.A., Wilson, H., 2000. Multivariate time series modelling of algal blooms in freshwater lakes by machine learning. Proceedings of the 5th International Symposium WATERMATEX on Systems Analysis and Computing in Water Quality Management, Gent, Belgium, pp. 9.17–9.32.

Resnikoff, H.L., Wells, R.O., 1998. Wavelet Analysis: The Scalable Structure of Information. Springer, New York.

Scafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological Methods 7, 147–177.

Shannon, C.E., 1949. Communication in the presence of noise. Proc. Inst. Radio Eng. 37, 10–21.