Predicting Seasonal Mixed-mode Time Series

X. Rosalind Wang *,1, Vadim Gerasimov, Mikhail Prokopenko, Astrid Zeman

ICT Centre, CSIRO, Locked Bag 17, North Ryde NSW 1670

Abstract

Underlying phase space dynamics of many time series can be characterized by mixed-mode periodic and chaotic orbits. Mixed-mode dynamics create a serious pattern recognition challenge for the analysis and forecasting of such time series. We propose here a novel method, which uses a combination of attractor reconstruction and Bayesian Network modeling. Dimensionality analysis is used to find embedding dimensions of the data and possible dimensions of underlying attractors. The correct dimensions allow us to build non-linear Bayesian models, used in predicting chaotic time-series. The developed algorithm is successfully applied to Australian electricity demand data.

Key words: forecasting, mixed-mode dynamics, dimensionality analysis, chaotic, attractor, embedding dimension, Bayesian Networks, electricity demand
PACS:

1 Introduction

Mixed-mode periodic and chaotic dynamics have been identified in many physical, biological and socio-economic time series: weather [1,2], hydrodynamics [3], finance [4,5], medicine [6], etc. Analyzing and forecasting time series with mixed-mode underlying dynamics is an important, but challenging problem across many diverse disciplines. Applying classical methods, such as ARMA modeling, etc. is problematic due to multiple reasons: inherent nonlinearity of data, non-Gaussian noise, and an unknown number of contributing factors (i.e., degrees of freedom in the underlying processes), preventing accurate model selection.

* Corresponding author
Email address: Rosalind.Wang@csiro.au (X. Rosalind Wang).
[1] The authors list after the lead author is in alphabetical order.

Typically, the general selection problem for time series [7] can be reduced to a model selection. For example, selection of an autoregressive model order, if a given class of candidate estimates, which can be computed from the data can consist entirely of finite-parameter models, e.g., auto-regressions. The well-known Akaike Information Criterion (AIC) provides an asymptotically efficient solution to the model selection problem for one-step ahead prediction [8]. The AIC capitalizes on the fact that the Kullback-Leibler (KL) information represents the information loss in fitting the data, and relates the KL information to the maximum likelihood estimates. Baysian Information Criterion (BIC) [9] is another method for statistically selecting a model that best fits the data. Over recent years, different machine learning methods, such as Neural Networks [5], Support Vector Machines [10], and Self-organizing Maps [11], have also been proposed in literature for the prediction of these data — with various degrees of success.

We propose a novel method for forecasting mixed-mode time series, using a combination of dimensionality analysis and Bayesian Network modeling. The procedure used is to:

(1) Identify correlation and embedding dimensions of the time series;
(2) Group the phase-space points of the phase space using the embedding dimension;
(3) Construct a Bayesian Network model that best describes the phase points, and use it in predicting the data.

Correlation and embedding dimensions are identified by using a technique that reconstructs attractors. An attractor is a set to which a dynamical system evolves after a transient (a sufficiently long time): phase-space points that approach the attractor remain close even if slightly disturbed (geometrically, an attractor can be a point, a curve, a manifold, or even a complicated set with a fractal structure, known as a strange attractor) [12].

Ruelle and Takens showed that strange attractors are the cause for turbulent, chaotic, dynamics in fluid flow [13]. Dynamics near an attractor are characterized by stretching, which causes divergence of nearby trajectories, and folding, which constrains the dynamics to a finite region of subspace, of a dimension d. This subspace is the space in which the attractor is embedded [2]. It has been shown that for many seemingly random time series's there exists a low-dimensional attractor.

We used the Bayesian Network (BN) model to learn and predict the reconstructed time series. The Bayesian Network is a graphical model that represents the probabilistic dependencies between the variables. Therefore, it

[2] For a strange attractor, the subspace has a non-integer dimension, that is, a fractal dimension.

provides a measure of the uncertainties of the observed data as well as a modeling tool suitable for the stochastic nature of time series. The combination of attractor reconstruction and Bayesian Network modeling addresses the chaotic and stochastic nature of the time series: dimensionality analysis helps in model selection for mixed-mode dynamics, while Bayesian Network modeling deals with non-linearity and stochasticity. This combination is essentially a multi-scale pattern recognition method: reconstruction of an attractor discovers global patterns in the phase-space, while prediction in the phase-space learns localized patterns for the model.

We applied the approach to Australian electricity demand data. This time series is dependent upon many external factors, including climate variation, business cycles, etc. These factors result in complicated and random variations, representative of mixed-mode dynamics.

This paper is organized as follows: Section 2 presents the methods used to analyze dimensionality of the time series. Section 3 presents the Bayesian Network used for learning and predicting the data. Section 4 shows the results of applying the developed algorithm to the electricity data, and how these results compare to a benchmark Neural Network algorithm used as industry standard. Finally, Section 5 presents a summary.

2 Dimensionality Analysis

Let a single variable time series be $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$, where N is the total number of data points in the series. We assume that this data is embedded in a d-dimensional phase space as the following [14]:

$$\mathbf{y}_i = \left(x_t, x_{t-\tau}, \ldots, x_{t-(d-1)\tau}\right), \tag{1}$$

where τ is the time delay, d is the embedding dimension, and $i = d, d+1, \ldots, N$.

The time delay, τ, is introduced in Equation 1 to allow the possibility of skipping samples during the reconstruction. Therefore, to reconstruct the phase space, we need to determine the embedding dimension and the time delay. To find τ, Fraser and Swinney [15] suggested using the mutual information method. We used the algorithm described by Darbellay and Vajda [16] to find the mutual information in the time series data. The value of τ is determined when the mutual information first approaches a minimum.

Grassberger and Procaccia [17] showed by estimating the Kolmogorov-Sinai (KS) entropy [18–20] from a time signal that the correlation integral, $C_d(r)$

can be estimated as:

$$C_d(N,r) = \frac{1}{(N-1)N} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \Phi(r - \|\mathbf{y}_i - \mathbf{y}_j\|). \tag{2}$$

Here Φ is the Heaviside function (equal to 0 for negative argument and 1 otherwise). The vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ contain elements of the observed time series $\{x_t\}$ with the dynamical information in one-dimensional data converted or reconstructed to spatial information in the d-dimensional embedding space $\mathbf{y}$ [21] as presented in Equation 1. The norm $\|\mathbf{y}_i - \mathbf{y}_j\|$ is the distance between the vectors in the d-dimensional space, e.g., the maximum norm [22]:

$$\|\mathbf{y}_i - \mathbf{y}_j\| = \max_{\tau=0}^{d-1}(x_{i+\tau} - x_{j+\tau}) \tag{3}$$

Put simply, $C_d(r)$ computes the fraction of pairs of vectors in the d-dimensional embedding space that are separated by a distance less than or equal to r. In order to eliminate auto-correlation effects, the vectors in Equation 2 should be chosen to satisfy $|i - j| > L$, for some positive L, and at the very least $i \neq j$ [23].

The correlation dimension ν is found by:

$$\nu = \lim_{r \to 0} \lim_{N \to 0} \frac{\ln C_d(N,r)}{\ln r}. \tag{4}$$

That is, within certain ranges of r and d, the correlation integral $C_d(r)$ may be proportional to some power of r, $C_d(r) \sim r^{\nu}$ [17]. If the dynamical process is unfolded by choosing a sufficiently large $d > d_\nu$, a typical slope of the plot $\ln C_d(r)$ versus $\ln r$ becomes independent of d. Thus the common numerical practice of finding the embedding dimension d of the data set is to compute the slope from a linear region of the $C_d(N,r)$ plot. For $d \leq \lfloor \nu \rfloor$, where $\lfloor \nu \rfloor$ denotes the largest integer less than or equal to ν, the slope is equal to d. For $d > \lfloor \nu \rfloor$, the slope saturates at a constant value which is usually taken to be the estimated value of ν [24].

The correlation dimension provides useful information about the spatial structure of the process and it also provides an effective measure of its (possibly fractal) size: in particular, a random process has an "infinite" correlation dimension (its orbit is not expected to have any spatial structure). In contrast, the correlation dimension for a periodic orbit is 1, while it could be higher for some non-regular processes. A non-integer $\nu < 1$ is an indication of a strange chaotic attractor [25].

3 Time Series Prediction

Bayesian Networks are used in this paper to learn and predict the embedded phase space data. A Bayesian Network (BN) is a form of graphical model that takes a statistical approach to learning. Statistical learning uses probability distributions to model variables that represent the available data, taking into account their stochastic nature. Graphical models expose underlying relationships between probabilistic variables in a simple and clear form.

Specifically, a BN is an acyclic directed graph (ADG) [26]: if one variable of the network is dependent on another, then the reverse can not be true. This relationship between two variables is represented in BN by the direction of an arrow connecting the two. The variables of a BN are called *nodes* of a BN. The node with an arrow pointing to it is dependent on the node with the same arrow pointing away from it. The nodes connected by an arrow have a parent/child relationship, where the *child* node is dependent on its *parent* node. (See Fig. 1 for the structure of a Bayesian Network.)

In a BN, each random variable is independent of its non-descendants in the graph, given the state of its parents. This independence can be exploited to reduce the number of parameters needed to characterize the network. Thus, it is possible to efficiently compute posterior probabilities, given some evidence or observations. One set of probability parameters are encoded for each variable, in the form of the local conditional distribution, given the variable's parent. Using the independence statements encoded in the network, the joint probability distribution is uniquely determined by these local conditional distributions [27,28]. We present the general form of this joint probability distribution in the following paragraphs.

We use capital letters such as X, Y for names of random variables, and lower cases x, y for values taken by these variables. A set of variables such as $\{X_1, X_2, X_3\}$ is written as $\mathbf{X}$, likewise, a set of values such as $\{x_1, x_2, x_3\}$ is written as $\mathbf{x}$. Thus, $\mathbf{x}$ are values taken by $\mathbf{X}$.

Let $P(\mathbf{U})$ be a joint probability distribution over $\mathbf{U} = \{X_1, \ldots, X_k\}$, where X_i is a random variable expressed by a node of the network. A BN for $\mathbf{U}$ is a pair $B = \langle G, \Theta \rangle$. The first component, G, represents the graph structure of the network. G is an ADG whose vertices correspond to the random variables $X_1, \ldots, X_k$, and whose edges represent direct dependencies between the variables. The second component, Θ, represents the set of conditional probabilities that quantify the nodes of the network. It contains a set of parameters $\theta_{X_i|\Pi_{X_i}} = P_B(X_i|\Pi_{X_i})$ for each node X_i, where Π_{X_i} denotes the set of parents of X_i in G. A Bayesian network B defines a unique joint probability distribution

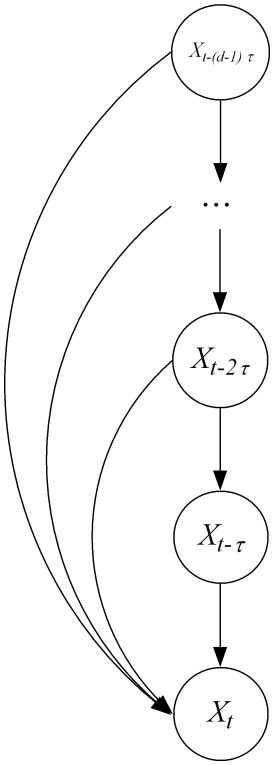


Fig. 1. The Bayesian Network used for learning and inference.

over $\mathbf{U}$, given by [27]:

$$P_B(\mathbf{U}) = \prod_{i=1}^{k} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{k} \theta_{X_i|\Pi_{X_i}}. \tag{5}$$

The learning process in a BN aims to estimate the parameter set Θ as well as to find the structure of the network, G. The objective of the learning is to find a $B = \langle G, \Theta \rangle$ that "best describes" the probability distribution over the training data [29]. In this paper, however, we will not be learning the structure of the network, using instead a simple representation of a $d-1$-th order Markov model, shown in Figure 1. The network is constructed from the underlying dependencies in a time series, that is, the data at time t is dependent on the data at time $t-\tau, \ldots, t-(d-1)\tau$. The joint distribution of the model is:

$$P(\mathbf{U}) = P(X_t|X_{t-\tau}, \ldots, X_{t-(d-1)\tau}) P(X_{i-(d-1)\tau}) \prod_{k=1}^{d-2} P(X_{t-k\tau}|X_{t-(k+1)\tau}) \tag{6}$$

where $\mathbf{U} = \{X_t, X_{t-\tau}, \ldots, X_{t-(d-1)\tau}\}$. All the nodes are modeled as one dimensional Gaussian. For example, a BN model of Fig. 1 with $d = 3$ have the dependencies as $X_{t-2} \rightarrow X_{t-1} \rightarrow X_t$ as well as $X_{t-2} \rightarrow X_t$. The joint distribution of the model will be $P(\mathbf{U}) = P(X_t|X_{t-1}, X_{t-2}) P(X_{t-1}|X_{t-2}) P(X_{t-2})$, where each $P(\cdot)$ is a Gaussian or a conditional Gaussian distribution.

Since the structure of the network is known given the value of d, only the parameter set Θ needs to be learnt. The Maximum Likelihood (ML) algorithm [30,31] is thus used to estimate Θ. In the ML estimator, the likelihood function, $p(\mathbf{x}|\theta)$, is treated as a function of θ for fixed $\mathbf{x}$, where x_j^t is the j-th data sample for the node X_t in the Bayesian Network. This *likelihood function* can be used to evaluate the choices of θ. The ML estimator chooses the value of θ that maximizes the probability of the data $\mathbf{x}$:

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(\mathbf{x}|\theta). \tag{7}$$

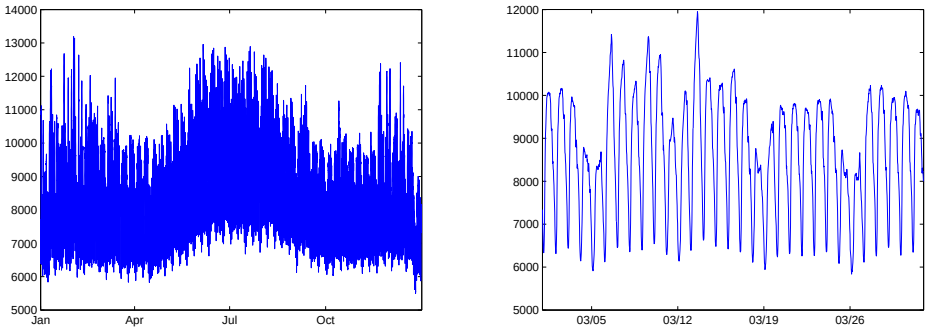
This learnt network can then be used to perform inference on new data given the parameters of the network. That is, given the observed values of some of the nodes in the network, compute the probability distribution of the other nodes using Bayes Theorem. Prediction in the Bayesian Network is achieved through inference on a child node given observations of the parents. In our case, we are to predict the value of X_t given the values observed for $X_{t-\tau}, \ldots, X_{t-(d-1)\tau}$.

4 Experimental Results and Discussion

This section is organized as follows: Section 4.1 presents the data used for testing the algorithms. Section 4.2 shows the results of the mutual information analysis of the data. Section 4.3 shows the results of dimensionality analysis. Finally, Section 4.4 presents the results of the time series prediction.

4.1 Data

We used electricity demand data from the National Electricity Market Management Company Limited (NEMMCO) in Australia [32]. NEMMCO was established by the various state governments of Australia to administer and manage the majority of the Australian electricity market. The market operates the world's longest interconnected power system. Up to AUD$7 billion of electricity is traded annually with almost 8 million end-use consumers spanning over 4000 kilometers in distance. The operation of the National Electricity Market (NEM) involves a system managing supply that meets, but does not exceed, the demand. The wholesale trading in electricity is conducted through spot prices which are determined every 30 minutes according to demand [33]. Fig. 2(a) shows the demand for electricity in 2006, and Fig. 2(b) shows the demand for the months of March in 2006. These two figures show the two aspects of mixed model dynamics described in the introduction: periodic and chaotic features of the time series.



(a) (b)

Fig. 2. The NEM electricity demand for (a) 2006, (b) March 2006.

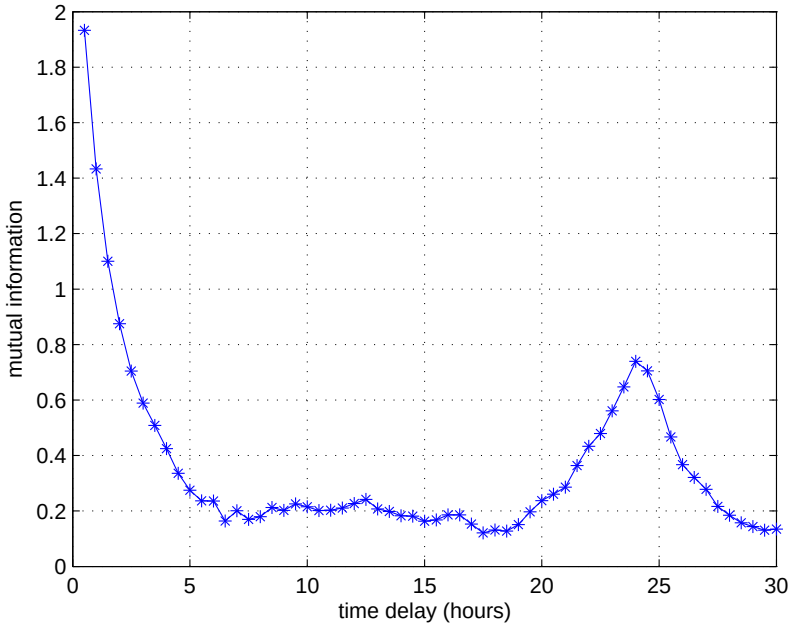


Fig. 3. Mutual information curve of the Australian electricity demand data over the year 2006.

4.2 Mutual Information

The delay time, τ, in Equation 1 is found by using mutual information as mentioned in Section 2. Figure 3 shows the mutual information curve of NEMMCO's demand data in the year 2006. For each delay τ on the horizontal axis, the curve maps the mutual information between x and $x - \tau$ for *all* data samples during the year. The mutual information curve serves to identify the delay τ, that minimizes the mutual information.

We found that the monthly and yearly mutual information curves do not differ

much. Furthermore, the mutual information curves do not differ much in shape and minimum point from year to year. Typically, we found that delay $\tau = 14$ (i.e., 7 hours) is the best delay.

It may also be observed that the mutual information between data segments reaches a minimum at time delay of 6 hours — and stays roughly at the same level for larger delays. However for a delay of 20, the mutual information increases again until it reaches a peak at $\tau = 24$. This can be explained by a 24 hour cycle in the demand for electricity.

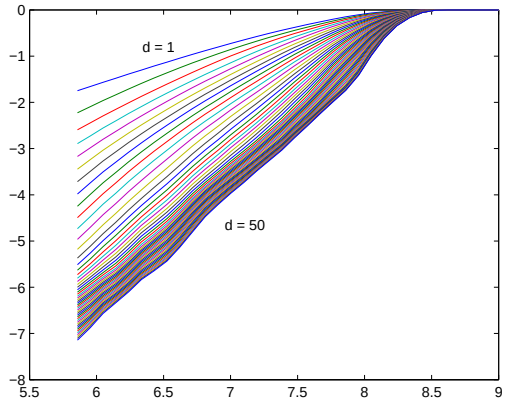
4.3 Dimensionality Analysis

To find the embedding and correlation dimensions, we divide the data into 28-day periods. With data available at 30 minute intervals, this gives us 1344 data points per period. As shown by Kugiumtzis *et al.* [34] while this number is small compared with the typical data sample sizes used in medicine and fluid dynamics, it is still a reasonable size, compared with those used in climatology. Since the number of days in a month is uneven throughout the year, the 28-day period gives a regular and meaningful division.

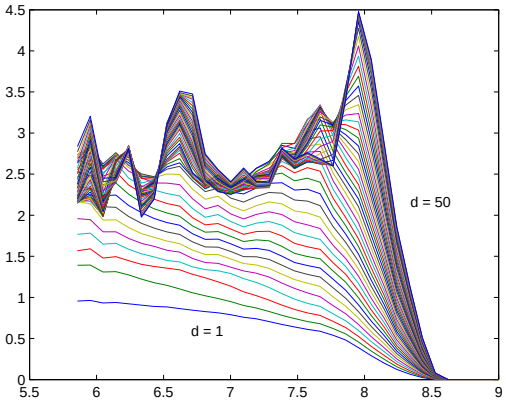
For each period of the demand data, we calculate the correlation integrals $C_d(N, r)$ for embedding dimensions d ranging from 1 to 50. A plot of $\ln C_d(r)$ versus $\ln r$ for the 6th period in the data, that is, data between 21 May 2001 and 18 June 2001, is plotted in Figure 4(a). We can observe three well-known regions: (1), the lower region distorted by fluctuations due to the small number of points; (2), a linear "scaling" region where the power law $C_d(r) \sim r^{\nu}$ holds; finally, (3), the upper region distorted due to the finite size of the data sets.

Figure 4(b) plots the gradients of the curves in Figure 4(a) versus $\ln r$, i.e., $\nu(d, r) = \frac{\partial \ln C_d(r)}{\partial \ln r}$ for different embedding dimensions d. When estimates of $\nu(d, r)$ reach a plateau in terms of $\ln r$, the corresponding value $\nu(d)$ provides an estimate of the correlation exponent for a given d. Note the near constant gradients around $\ln r = 7$ — this plateau corresponds to the second, linear, region in Figure 4(a). The exponents $\nu(d)$ can be reliably obtained from the plateau by calculating the average of the gradients from around $\ln r = 6.8$ to $\ln r = 7.5$. Figure 4(c) plots these estimates against the embedding dimension ($\nu(d)$ vs d). This plot clearly shows two regions: (1), the lower region where the values are increasing; (2), the middle flat region where the correlation exponents have converged. The point on $\nu(d)$ vs d plot, where exponents $\nu(d)$ converge, indicates the best estimate of the correlation dimension ν for the time series. This point, however, may not always be attainable, as shown in the next example.

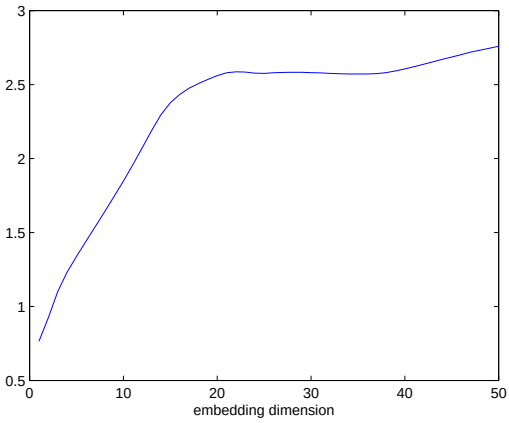
Figure 5 shows the results of dimensional analysis for the 1st period in the



(a)

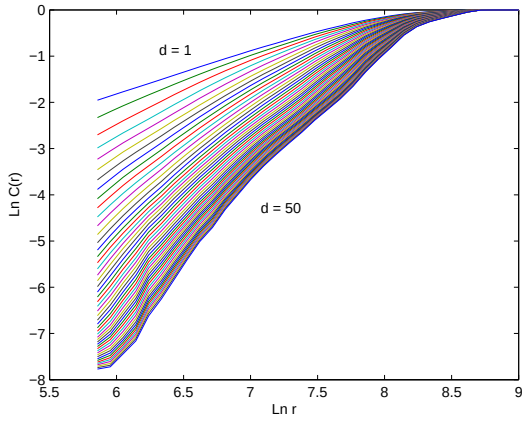


(b)

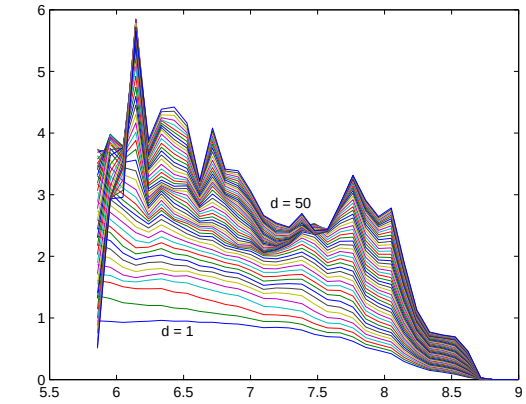


(c)

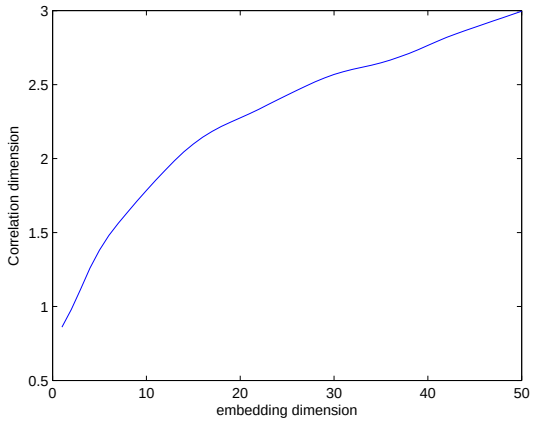
Fig. 4. Analysis of demand data of of 6th period (21 May 2001 - 18 Jun 2001): (a) correlation integral; (b) correlation integral gradients; (c) correlation dimensions vs. embedding dimension.



(a)



(b)



(c)

Fig. 5. Analysis of demand data of 1st period (01 Jan 2001 - 29 Jan 2001): (a) correlation integral; (b) correlation integral gradients; (c) correlation dimensions vs. embedding dimension.

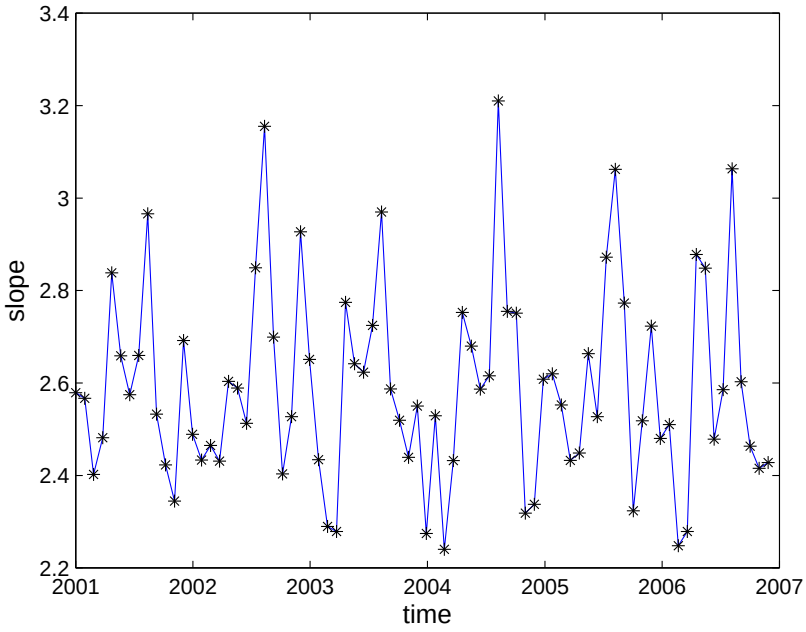
data, that is, for data between 1st January, 2001 and 29th January, 2001. As with Figure 4, we can extract estimates of the correlation exponents from the region around $\ln r = 7$ in Figure 5(b). Figure 5(c) shows the correlation exponents against the embedding dimension ($\nu(d)$ vs d) for this time period. Unlike Figure 4(c), the corresponding Figure 5(c) shows a gradual increase rather than a convergence to a certain value. In cases like this, we used a simple heuristic in determining the correct embedding and correlation dimensions for the time series: minimization of the derivative, i.e. the embedding dimension $\hat{d} = \text{argmin}\frac{\partial\nu(d)}{\partial d}$, and the corresponding correlation dimension $\hat{\nu} = \nu(\hat{d})$.

Figure 6 traces both the correlation and embedding dimensions of the demand data from January 2001 to December 2006 using the described methods. Figure 6(a) highlights chaotic character of the underlying dynamics. Moreover, correlation dimensions for many 28-day periods are non-integers, varying approximately from around $\nu = 2.2$ to around 3.2. This is indicative of a change between 2 and 3 degrees of freedom in the underlying process. In other words, the underlying dynamic process usually has 2 contributing factors, and sometimes 3 factors, while in other cases it becomes fractal. The values of embedding dimension vary as well, from low 20's to high 30's, also indicating a change in the balance between periodicity and chaos in the mixed-mode process.

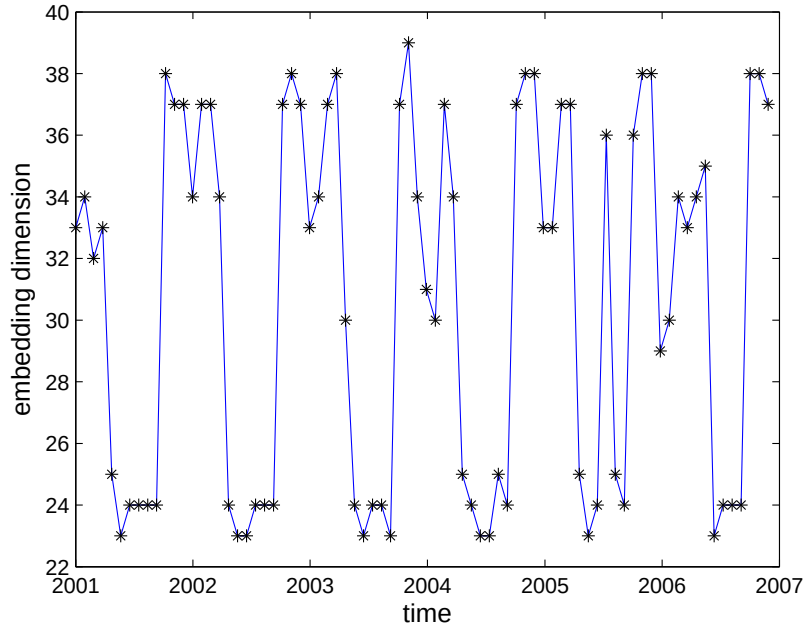
It can also be seen from both plots that the electricity demand data has a seasonal pattern over the years — with respect to the number of contributing factors. Interestingly, in each year there are two minimums with $\nu \leq 2.5$ corresponding to beginning and end of summer and two maximums with $\nu \geq 2.5$ corresponding to mid-autumn and mid-spring periods. A mid-winter period is characterized by $\nu \approx 2.5$, resulting in 'M'-shaped patterns as shown in Fig. 7. The minimums with $\nu \leq 2.5$ are explained by more stable dynamics, when weather is less chaotic during summer and winter, while the maximums with $\nu \geq 2.5$ are due to more chaotic weather variations during autumn and spring. These variations cause, obviously, less predictable demand at those periods. A more precise explanation may be given by cyclical factors: when the weather (and the demand) is more stable, it is mostly driven by daily and weekly cycles (hence, ν is closer to 2); while during change-over seasons (autumn and spring) a seasonal cycle begins to contribute, increasing ν towards 3.

4.4 Time Series Prediction

We compared our results with those obtained using a Neural Network (NN) model, chosen because NN is the model employed by NEMMCO to predict the demand [35]. NEMMCO used a two-layer feed-forward Neural Network, which is a model with one hidden layer [36].

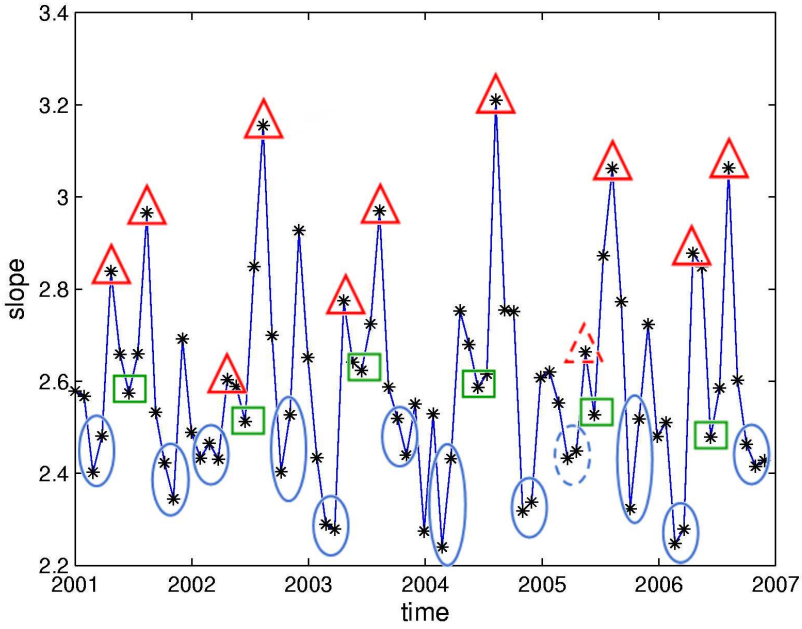


(a)

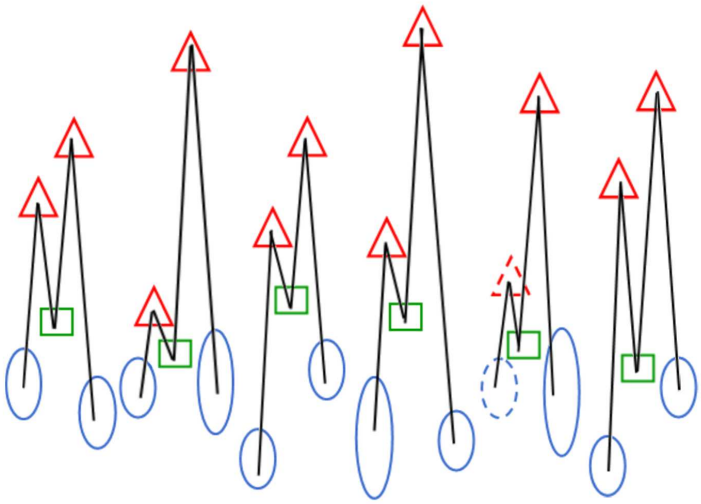


(b)

Fig. 6. Results of dimensionality analysis of the Australian electricity demand data over the years: (a) Correlation dimensions, (b) Embedding dimensions.



(a)



(b)

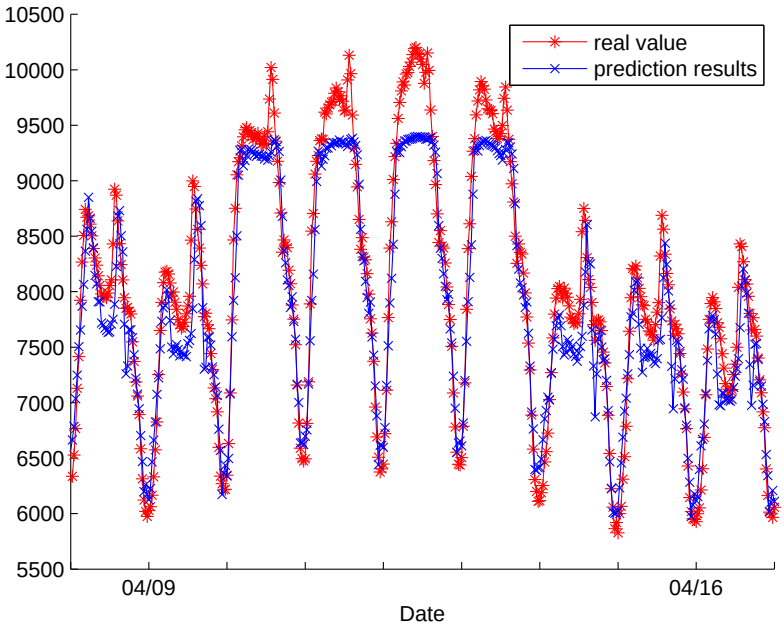
Fig. 7. The seasonal pattern in electricity demand data's correlation dimensions. The triangles show the maximums in a year, the ovals show the minimums in a year, and the squares show the local minimum between the two maximums of a year. (a) Tracing the patterns with symbols on the plot; (b) Symbols only.

To measure and compare the effectiveness of the method presented in this paper, we calculated the absolute percentage error of the prediction. This percentage error is a useful scale-independent measure of discrepancy between data sets with different ranges. The errors calculated for each data point in a data set are averaged to produce the error value for the data set.

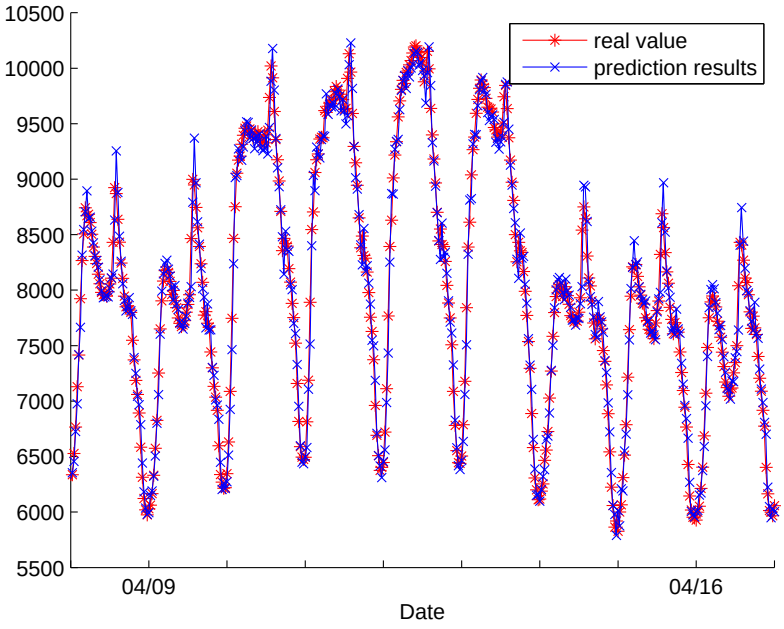
The demand data from 2006 were used for prediction. The year was divided into thirteen 28-day periods. For each period, we tested five different combinations of parameters. The first four tests involved learning the Bayesian network using data from the same period in 2005, and predicting the data from 2006. The 2005-2006 split between training and testing sets is justified by seasonal variations present in the data, see e.g. Fig. 6. We varied the values of time delay τ and embedding dimension d for each test. We used $\tau = 1$, and $\tau = 15$ as suggested by minimization of mutual information described in Section 4.2. Three different values of d were used: $d = 10$ which is the number of input neurons in the Neural Network currently used by NEMMCO for the demand prediction, a fixed value d value that is an approximate average of the values shown in Fig. 6(b), and the actual variable values d shown in these figures. The last, 5th, test learns the models using data from an immediately preceding period, with $\tau = 1$ and the actual variable values d shown in Fig. 6(b).

Table 1 shows the prediction results for the demand data. The first column of the results shows the prediction results in percentage error as determined by the Neural Network model used by NEMMCO, that is, with $d = 10$ and $\tau = 1$ (see Appendix A for full prediction results from Neural Network). The second column shows the results determined by the Bayesian Network model using the same parameters. Figure 8 (a) and (b) shows the prediction results against the actual data for one week in 2006 by NN and BN using these parameters. From these results, we can see that the Bayesian Network model outperforms the Neural Network model for the demand data. Figure 9 shows the close up view of the Bayesian Network results.

Interestingly, as can be seen from Table 1, column 4, the case with $\tau = 15$, performs significantly worse than the rest. Therefore, while the delay-time embedding procedure almost always reconstructs the original state space of a dynamic system [37], using the time delay τ for forecasting based on the Bayesian model increases the prediction errors. This is because the Bayesian model is built using the dependencies between the variables. However, the method of choosing τ picks the value when the mutual information first approaches a minimum. In other words, there is minimum dependency between the data at time t and time $t - \tau_{min}$. Further, we can see from Fig. 3 that mutual information is maximum when $\tau = 1$ and decreases almost linearly as τ increases. Therefore, when applying the Bayesian model to learn and predict time series, it is best to use time delay of $\tau = 1$.



(a)



(b)

Fig. 8. Electricity demand prediction results for a week in 2006 using: (a) Neural Network,(b) Bayesian Network. Both networks are learned using data form the same period in 2005, with $\tau = 1$ and $d = 10$

.

Table 1
Prediction Results for Demand data

period number	Prediction Results (% error)					
	NN	Bayesian Network				
	same p.	Learning using data from same period				Diff. p
	$\tau = 1$	$\tau = 1$	$\tau = 1$	$\tau = 15$	$\tau = 1$	$\tau = 1$
	$d = 10$	$d = 10$	$d = 36$	$d = 36$	var. d	var. d
1	3.070	1.002	0.930	8.431	0.954	0.932
2	3.409	1.038	0.951	8.784	0.981	1.052
3	3.209	1.028	0.973	4.657	0.962	1.070
4	3.144	1.152	1.126	4.748	1.111	1.200
5	4.580	1.266	1.195	4.202	1.257	1.239
6	4.938	1.227	1.114	5.451	1.204	1.190
7	5.176	1.206	1.122	6.017	1.183	1.164
8	2.887	1.267	1.185	5.022	1.185	1.250
9	3.142	1.230	1.176	3.866	1.220	1.340
10	4.981	1.189	1.158	3.965	1.188	1.332
11	2.282	1.086	1.044	4.102	1.044	1.108
12	2.479	1.033	0.985	5.394	0.961	1.021
13	3.827	0.997	0.894	6.195	0.875	0.880
mean	3.625	1.132	1.066	5.449	1.087	1.137

We can also see from Table 1 that using $d = 36$ and the d values found for each period give slightly better prediction results than using $d = 10$. However, using the various d values does not give much different prediction results than from just using $d = 36$. This is mainly because sometimes the process of determining the embedding dimension is not precise, as shown in Fig. 5(c), where it is hard to determine the exact value of the dimensions. Therefore, using an average of the d values is the best option in building the Bayesian model.

The last column of Table 1 shows the prediction results from a model built using the data from a different period, specifically that of data from the previous period. For example, we built a model using data from period 1 of 2006 to predict the data from period 2 of 2006. As we can see from the results, this model performs slightly worse than its counterparts in the previous column.

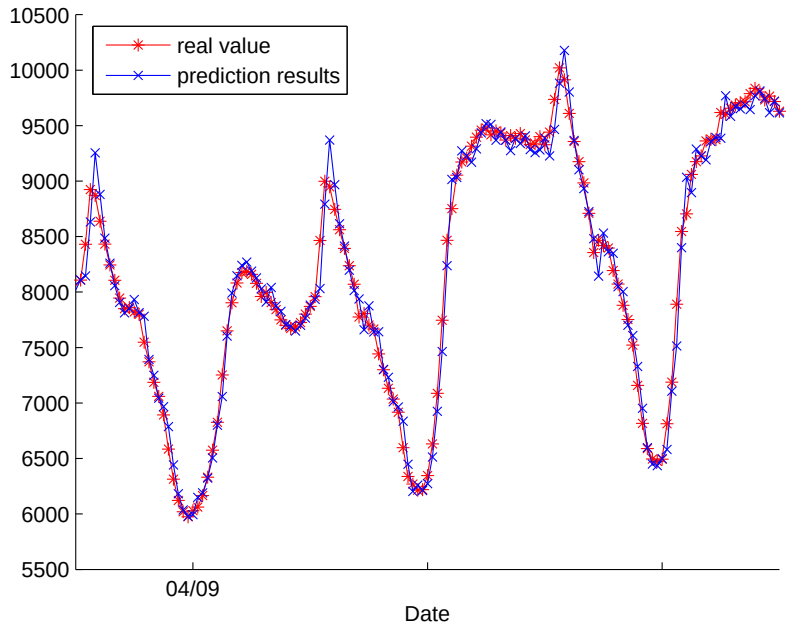


Fig. 9. Close up view of the electricity demand prediction result of Fig. 8(b).

These results show that the use of data from the same period for training is justified.

5 Conclusion

We have presented a novel approach to forecasting time series by using a combination of attractor reconstruction and Bayesian learning algorithms. The attractor reconstruction allows us to find the embedding dimension for the attractor. This embedding dimension is then used to model a Bayesian Network to learn and predict the data. We used the Australian electricity demand data to test the method.

The resulting correlation dimensions have a clear pattern over the years — highlighting in particular how changes between seasons affect chaoticity of the demand. The observed 'M' shaped pattern shows that demand has more stable dynamics in summer and winter, and has more chaotic dynamics in mid-spring and mid-autumn.

Typically a time delay τ is used in the reconstruction of the state space. It may also be used in the prediction model that is applied to the reconstructed state space. However, we found that a prediction model based on Bayesian Networks does not need $\tau > 1$. For the results where $\tau = 1$ is used for prediction, the Bayesian Network prediction is much better than those from the

Table A.1
Prediction Results for Demand data using a Neural Network.

period number	Prediction Results (%)				
	Learning using data from same period				Diff. p
	$\tau = 1$	$\tau = 1$	$\tau = 15$	$\tau = 1$	$\tau = 1$
	$d = 10$	$d = 36$	$d = 36$	var. d	var. d
1	3.070	3.723	7.328	5.350	5.404
2	3.409	1.937	8.672	5.070	1.518
3	3.209	3.635	5.167	4.700	3.286
4	3.144	4.876	5.800	4.138	2.918
5	4.580	4.367	5.962	4.562	3.016
6	4.938	3.441	5.092	2.748	4.239
7	5.176	4.158	5.916	3.713	2.913
8	2.887	5.541	6.316	3.184	4.632
9	3.142	4.082	3.606	4.590	3.354
10	4.981	4.669	4.442	5.531	4.673
11	2.282	3.718	5.013	4.172	5.844
12	2.479	4.432	4.294	3.897	3.424
13	3.827	6.714	7.861	2.768	4.839
mean	3.625	4.253	5.805	4.186	3.851

Neural Network model. By varying the embedding dimensions during the BN prediction, we observed that an embedding dimension, which is approximately the average of the estimated dimensions, is sufficient.

A Neural Networks Results

Tables A.1 show the full prediction results of the Neural Network model for the demand data.

References

[1] C. Essex, T. Lookman, M. A. H. Nerenberg, The climate attractor over short time scales, Nature 326 (1987) 63–66.

[2] M. Ghil, M. Kimoto, J. D. Neelin, Nonlinear dynamics and predictability in the atmospheric sciences, Reviews of Geophysics 29 (1991) 46–55.

[3] H. Haucke, R. Ecke, Mode-locking and chaos and Rayleigh-Benard convection, Physica D 25 (1987) 307–329.

[4] W. A. Brock, D. A. Hsieh, B. LeBaron, Nonlinear dynamics, chaos and instability: Statistical theory and economic evidence, Tech. rep., Massachusetts Institute of Technology (1991).

[5] C. L. Giles, S. Lawrence, A. C. Tsoi, Noisy time series prediction using a recurrent neural network and grammatical inference, Machine Learning 44 (1/2) (2001) 161–183.

[6] A. Babloyantz, A. Destexhe, Low-dimensional chaos in an instance of epilepsy, Proceedings of the National Academy of Sciences of the United States of America 83 (1986) 3513 – 3517.

[7] C. M. Hurvich, Selection of time series models and spectrum estimates using a bias-corrected generalization of AIC, New Direction in Time Series Analysis 46 (1990) 155–168.

[8] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control AC-19 (6) (1974) 716–723.

[9] D. Heckerman, A tutorial on learning with bayesian networks, in: M. Jordan (Ed.), Learning in Graphical Models, MIT Press, 1999.

[10] S. Mukherjee, E. Osuna, F. Girosi, Nonlinear prediction of chaotic time series using support vector machines, in: J. Principe, L. Giles, N. Morgan, E. Wilson (Eds.), IEEE Workshop on Neural Networks for Signal Processing VII, IEEE Press, 1997, p. 511.

[11] J. Vesanto, Using the SOM and local models in time-series prediction, in: Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997, pp. 209–214.

[12] Attractor, In Wikipedia, The Free Encyclopedia. Retrieved 08:24, November 15, 2007 (Nov. 2007).
URL `http://en.wikipedia.org/wiki/Strange_attractor`

[13] D. Ruelle, F. Takens, On the nature of turbulence, Communications in Mathematical Physics 20 (3) (1971) 167–192.

[14] N. H. Packard, J. P. Crutchfield, J. D. Farmer, R. S. Shaw, Geometry from a time series, Phys. Rev. Lett. 45 (9) (1980) 712–716.

[15] A. M. Fraser, H. L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (2) (1986) 1134–1140.

[16] G. Darbellay, I. Vajda, Estimation of the information by an adaptive partitioning of the observation space, IEEE Trans. on Information Theory 45 (1999) 1315–1321.

[17] P. Grassberger, I. Procaccia, Estimation of the Kolmogorov entropy from a chaotic signal, Phys. Rev. A 28 (4) (1983) 2591–2593.

[18] A. Kolmogorov, A new metric invariant of transient dynamical systems and automorphisms in lebesgue spaces, Doklady Akademii Nauk SSSR 119 (1958) 861–864, Russian.

[19] A. Kolmogorov, Entropy per unit time as a metric invariant of automorphisms, Doklady Akademii Nauk SSSR 124 (1959) 754–755, Russain.

[20] Y. Sinai, On the concept of entropy of a dynamical system, Doklady Akademii Nauk SSSR 124 (1959) 768–771, Russian.

[21] F. Takens, Detecting strange attractors in turbulence, in: Lecture Notes in Mathematics, Vol. 898, Springer-Verlag, Berlin, 1981, pp. 366–381.

[22] F. Takens, Invariants related to dimension and entropy, in: *Atas do 13* Colóquio Brasiliero do Matemática, Rio de Janeiro, 1983.

[23] J. Theiler, Spurious dimension from correlation algorithms applied to limited time-series data, Phys. Rev. A 34 (3) (1986) 2427–2432.

[24] M. Dhamala, Y. Lai, E. Kostelich, Analyses of transient chaotic time series, Phys. Rev. E 64 (5) (2001) 056207–056216.

[25] P. Grassberger, I. Procaccia, Characterization of strange attractors, Physical Review Letters 50 (1983) 346–349.

[26] N. Friedman, D. Koller, Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian Networks, Machine Learning 50 (2003) 95–126.

[27] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.

[28] F. V. Jensen, Bayesian Networks and Decision Graphs, Springer-Verlag, New York, 2001.

[29] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 1988.

[30] D. J. MacKay, Information Theory, Learning and Inference, Cambridge University Press, 2003.

[31] I. J. Myung, Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology 47 (2003) 90–100.

[32] The national electricity market management company limited.
URL http://www.nemmco.com.au/

[33] An introduction to Australia's national electricity market (Jun 2005).
URL http://www.nemmco.com.au/nemgeneral/000-0187.pdf

[34] D. Kugiumtzis, B. Lillekjendlie, N. Christophersen, Chaotic time series part I: Estimation of some invariant properties in state space, Modeling, Identification and Control 15 (1994) 205–224.

[35] Five minute electricity demand forecasting Neural Network model (Oct. 1998).
URL http://www.nemmco.com.au/

[36] R. P. Lippmann, An introduction to computing with neural nets, IEEE ASSP Magazine 1 (1987) 4–22.

[37] J. Theiler, Estimating fractal dimension, J. Opt. Soc. Am. A 7 (6) (1990) 1055–1073.