The Statistician (2002)
51, *Part* 1, *pp.* 31–40

Simulation of hidden Markov models with EXCEL

W. H. Laverty, M. J. Miket and I. W. Kelly

University of Saskatchewan, Saskatoon, Canada

[Received January 2001. Final revision November 2001]

Summary. The paper demonstrates the use of functions provided by EXCEL for simulation of two types of hidden Markov models. The graphical capabilities of EXCEL are then used to visualize the simulated models. The power of spreadsheet simulation comes through the fact that any change in the parameters defining a hidden Markov model can be seen immediately in the simulated observations and in the graphs. This can be a very valuable aid in the understanding of hidden Markov models. The paper should be accessible and useful to anyone who has some knowledge of EXCEL and basic probability concepts. Two applications are included.

Keywords: Applications of EXCEL; Hidden Markov models; Normal and Poisson probabilities; Simulation

1. Introduction

Hidden Markov models constitute a fairly recent statistical technique for coping mainly with non-standard time series data. The technique has found many interesting and important applications in sciences, engineering, finance and management; see, for example, Elliott *et al.* (1995), MacDonald and Zucchini (1997) and Rabiner and Juang (1993).

EXCEL is one of the most widely used spreadsheets in a variety of computing environments and comes with an array of options including a range of statistical functions; see Berk and Carey (1998). These, together with EXCEL's powerful graphics capabilities, are exploited in this paper, which will show how the spreadsheet can be used in aiding the understanding of hidden Markov models. The interested reader should be able to follow the steps in this paper by entering commands into EXCEL worksheets; the paper thus assumes some familiarity with EXCEL, its graphical capabilities and numerical functions.

2. Basic definitions by example

A prototypal example of a hidden Markov model would be recording the results of throwing many times one of two dice picked at random, one of which is biased and the other is unbiased. If the dice are indistinguishable to the observer, then the two 'states' (i.e. the dice) in this model are hidden; hence the name.

An example of an engineering application from signal processing in a noisy environment can be described as follows. Signals generated at discrete times t will be denoted by $\{X_t\}$. The values of X_t can be restricted to a discrete set of values called states. Signals proceed from state to state with a probability that depends only on the previous state, the Markov chain property. When the Markov chain is not observed directly, it is termed hidden. What is observed in signal

Address for correspondence: M. J. Miket, Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon, Saskatchewan, S7N 5E6, Canada.
E-mail: Miket@math.usask.ca

 0039–0526/02/51031

processing instead is the Markov chain plus noise, giving a discrete time, finite state sequence of observations $\{Y_t\}$. Given these observations, the problem is to estimate the states of the chain, i.e. the true signal.

A third example is from bioengineering and consists of counting the number of movements for consecutive 5-s intervals of a foetal lamb. The counts are made by observing a foetus through ultrasound, possibly as part of a larger experiment. Leroux (1989) contains one such data set, from which it is apparent that the counts in adjacent time intervals are dependent. This then suggests the existence of at least two states of the foetus:

(a) a normal state where the foetus spends the majority of time (this is the state of little or no movement) and
(b) a state of excitement with many movements (triggered by various stimuli).

A hidden Markov model is very well suited for this situation, where the problem is to describe details of a succession of foetal states from the observed counts.

The last example can be generalized to modelling the progression of a disease in a human body. A typical disease will consist of a number of hidden states, such as remission and relapse. The number of such states and the details of these states must be found through the analysis of actual observations on the patient. Such an analysis, using hidden Markov models, could ultimately improve the treatment of the patient.

3. Specification of hidden Markov models

A simulation of a hidden Markov model begins by choosing the number m of hidden states in the Markov chain. Throughout this paper, we take $m = 2$. The possible hidden states of the Markov chain are then $1, 2, \ldots, m$, and we write

$$\Pr(X_1 = i) = p_i$$

for the initial state probabilities. The next task, when simulating a hidden Markov model, is the specification of the transition probability matrix Γ. This is an $m \times m$ matrix, with element γ_{ij} being the probability of a transition into state j starting from state i, i.e.

$$\gamma_{ij} = \Pr(X_t = j | X_{t-1} = i),$$

where t denotes time. These two choices allow us to construct a sequence of states (known also as the Markov chain) $X_1, X_2, \ldots, X_T$ constituting the hidden part of a hidden Markov model.

When the Markov chain is in state i, it emits an observed signal Y_t, which is either a discrete or a continuous random variable with distribution conditional on the current state. In the discrete case, the probability of $Y_t = s$ (where s is an integer in some known interval $[0, N]$) is specified by the matrix Π with elements

$$\pi_{si} = \Pr(Y_t = s | X_t = i).$$

The specification of this matrix depends on the application at hand, and our knowledge of that application. For the data on foetal lamb movements, it was found that the Poisson probability distribution is an adequate description, so

$$\pi_{si} = \frac{\exp(-\lambda_i)\,\lambda_i^s}{s!},$$

where the λ_i are different for different states in the Markov chain.

The dependence structure of the entire model is illustrated in Fig. 1. The specification of m, p_i, Γ and Π allows us to simulate the entire process, and the use of EXCEL for this purpose will be described later. First, we consider the simulation of a hidden Markov model with states emitting signals that follow normal distributions.

3.1. Hidden Markov models with normal states

We first consider generating normal random variates. The starting-point is to generate a random number from a uniform distribution on the interval [0,1], using the chain Paste Function ⇒Math & Trig.⇒Rand() in EXCEL. This 'returns an evenly distributed random number greater than or equal to 0 and less than 1', giving a uniform random number in the interval [0,1].

The reason for creating a uniformly distributed random number in [0,1] is that the cumulative distribution function for a continuous random variable will have values in [0,1]; Fig. 2. Thus, to generate random numbers from a specified distribution, it is only necessary to apply the inverse function for that distribution to a uniform random number. For the normal distribution, this may be accomplished in EXCEL by using the command chain Paste Function⇒Statistical⇒NORMINV, in which we specify the probability as the uniform random number, together with the mean and standard deviation. This results in a value from the normal distribution with specified mean and standard deviation. This is a common method, suggested in Ross (2000), which also gives several other methods of simulating normal variates. By repeating these commands, we can generate a column of values from the specified normal distribution.

However, we really want a column of random numbers from a specific hidden Markov model, rather than from a single normal distribution. Thus, we need to generate the states of the model; for illustration, we specify a two-state hidden Markov model by choosing

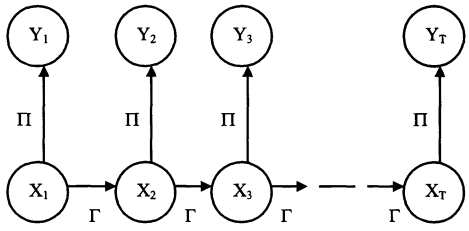

Fig. 1. Dependence structure for a hidden Markov model

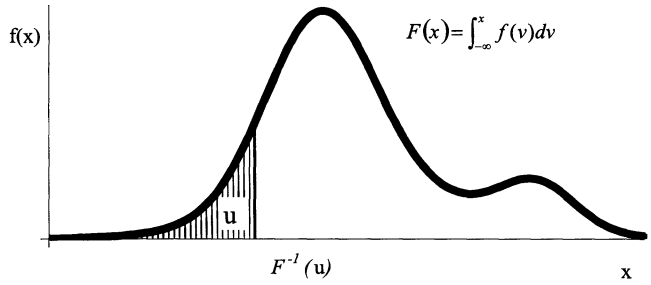

Fig. 2. Typical continuous distribution

(a) transition probability matrix

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} = \begin{pmatrix} 0.95 & 0.05 \\ 0.02 & 0.98 \end{pmatrix},$$

(b) state parameters, such that the normal distribution has $\mu = 4$ and $\sigma = 10$ in state 1, and $\mu = 4$ and $\sigma = 1$ in state 2, and

(c) initial probabilities $p_1 = 0.7$ and $p_2 = 0.3$.

In the worksheet, we now let column A contain, for convenience, integers 1–200, say, and fill column B with 200 uniform random numbers in [0,1]. Column C is now filled with 1s and 2s in the following manner, using initial probabilities in p_1 in cell E8 and p_2 in E9, and transition probabilities γ_{11} in cell E2, γ_{12} in F2, γ_{21} in E3 and γ_{22} in F3. The formula

```
=IF(B1<E8,1,2)
```

typed in cell C1 then sets C1 = 1 if the random number in cell B1 is less than p_1 (in cell E8); otherwise C1 = 2. For cell C2 (and the rest of the cells in column C on dragging down), a more complex formula is used, namely

```
=IF(OR(AND($C1=1,$B2<$E$2),AND($C1=2,$B2<E3)),1,2)
```

This combination of EXCEL's logical functions states that if the system is in state 1 at time $t = 1$ (i.e. C1 = 1) then, for time $t = 2$, the system stays in state 1 if the random number in cell B2 is less than or equal to γ_{11} in E2, and so C2 = 1; otherwise it changes to state 2, so C2 = 2. If, however, the system is in state 2 at time $t = 1$ (i.e. C1 = 2), then, for time 2, the system changes to state 1 if the random number in cell B2 is less than or equal to γ_{21} in E3, so C2 = 1; otherwise, it stays in state 2, and so C2 = 2.

Column D is then filled with 200 simulated observations by typing the formula

```
=IF($C1=1,NORMINV(RAND(),$F$8,$G$8),NORMINV(RAND(),$F$8,$G$8))
```

into cell D1, and dragging to fill the column. This formula simply says that for state 1 the mean and standard deviation are in cells F8 and G8 respectively (containing $\mu = 4$ and $\sigma = 10$ in our example), and for state 2 in cells F9 and G9 respectively (containing $\mu = 4$ and $\sigma = 1$).

The hidden states and simulated observations can then be viewed by selecting the sequence chart wizard ⇒ XY (Scatter) ⇒ data points connected with lines without markers. Fig. 3 shows an example of these states and the simulated observations, whereas Fig. 4 illustrates the EXCEL worksheet that is obtained.

The power of simulation lies in the fact that changing one or more of the parameters in the worksheet defining the hidden Markov model will recompute the entire worksheet including the graph. This allows us to develop an appreciation for various parameters in the model. Further, the F9 key can force EXCEL to recompute the entire worksheet with the existing parameters, and thus to create a new realization of the hidden Markov model.

The real problem is the opposite of simulation, namely the observations are all that is known, and from there all the parameters defining the hidden Markov model must be inferred by mathematical and statistical reasoning; see MacDonald and Zucchini (1997). However, the methods of simulating hidden Markov models described above could be used to investigate the sampling properties of various rough estimates of parameters.

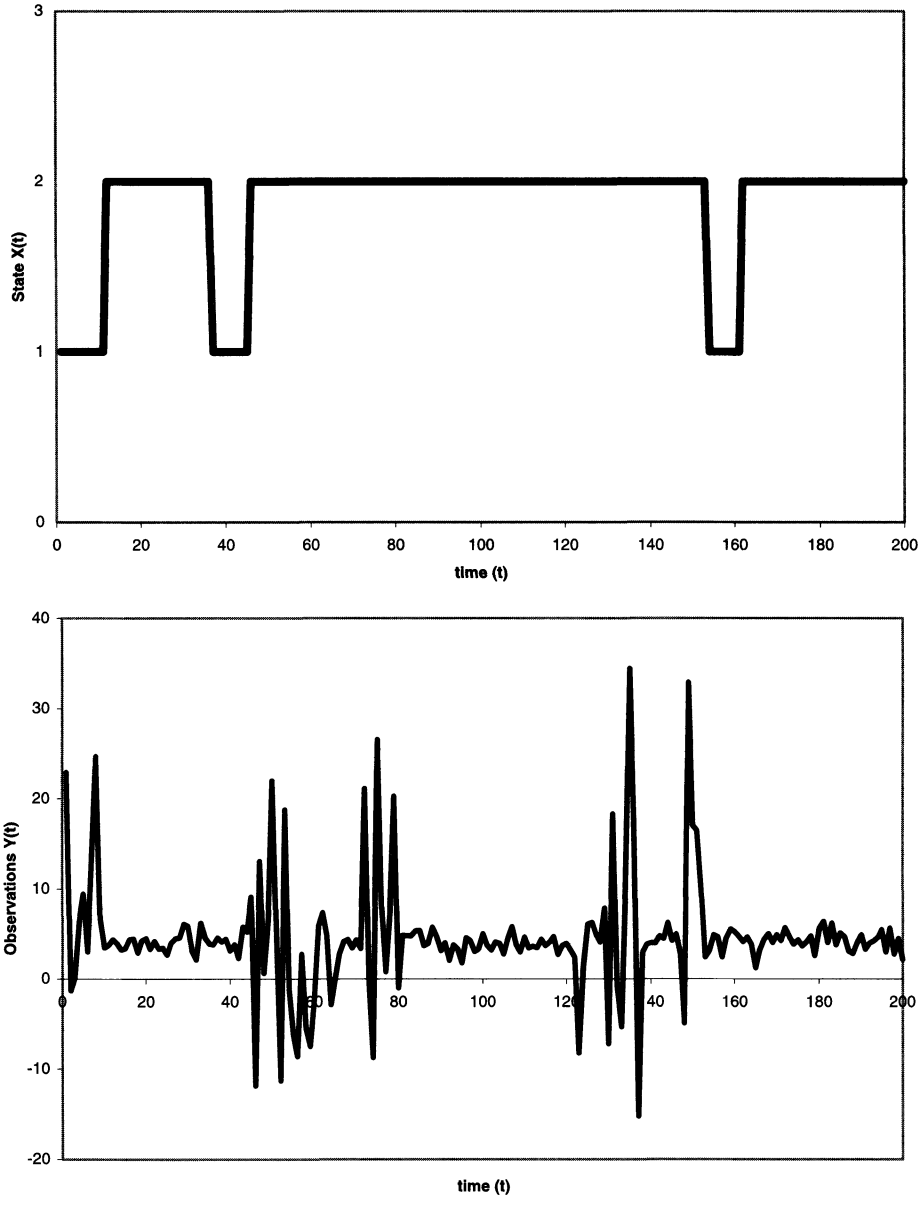

Fig. 3. Simulated observations from normal states

3.2. Hidden Markov model with Poisson states

To simulate a random variable Y from the Poisson distribution with a given mean λ, it is necessary to compute the cumulative Poisson probabilities

$$S_k = \sum_{i=0}^{i=k} \frac{\exp(-\lambda)\,\lambda^i}{i!}$$

for $k = 0, 1, 2, \ldots, N$, where N is such that S_N is equal to 1 to machine precision, and $S_{-1} = 0$.

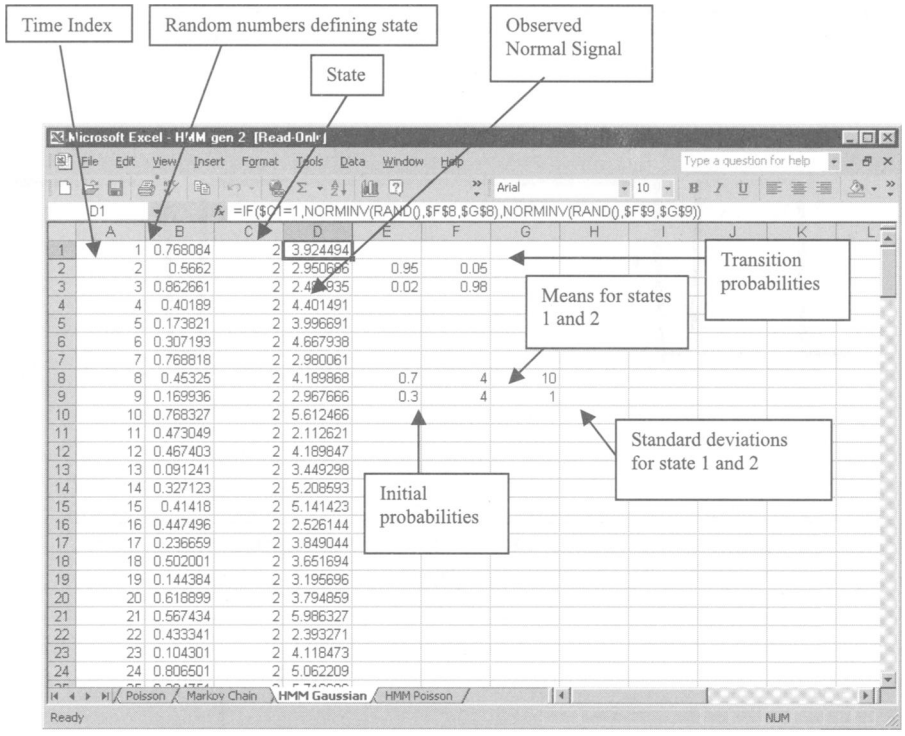

Fig. 4. Top part of a worksheet for the simulation of a hidden Markov model with normal states

Then, we select a uniform [0,1] random variable U and set $Y = s$ where s is such that

$$S_{s-1} < U \leqslant S_s;$$

see Ross (2000); nice proofs are also given in Bulgren (1982) and Gentle (1998).

To illustrate this procedure for the case $\lambda = 3$ and $N = 21$ in a new worksheet, we enter the value for λ in cell F1 and the numbers 0, 1, 2, . . . , 21 in cells A2–A23 in column A. Cells B2–B23 in column B contain cumulative Poisson probabilities $S_0, S_1, S_2, \ldots, S_{21}$, obtained by typing

```
=POISSON($A2,$F$1,TRUE)
```

in cell B2 and then dragging down to cell B23. Column C contains numbers 0, 1, 2, . . . , 22 in cells C1–C23 and is the table from which we look up numbers in column B and read off corresponding Poisson variates, as described earlier.

To illustrate the use of this table, we may then enter integers 1–200 in column G and create uniform random numbers in column H by using

```
=RAND()
```

in each cell. Column I then contains the values of the Poisson random variable, obtained by typing

```
=VLOOKUP($H1,$B$1:$C$22,2)
```

in cell I1 and then dragging to cell I200. Here, VLOOKUP is a reference function within EXCEL. A simple check of this procedure may be obtained by typing

```
=AVERAGE(I1:I200)
=STDEV(I1:I200)
```

into suitable cells; both numbers should, from the properties of the Poisson distribution, be close to λ.

Armed with this, and again moving to a new worksheet, we now outline the simulation of a hidden Markov model with $m = 2$ states. Much of this is essentially the same as for the Gaussian case considered earlier; as before, column A contains integers from 1 to 200, and column B contains random numbers between 0 and 1. We next specify the transition probability matrix

$$\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix},$$

stored in cells J2, K2, J3 and K3, the initial probabilities for the two states $p_1 = p_2 = 0.5$, stored in cells J6 and J7, and the means for state 1 in cell K6 ($\lambda_1 = 2.5$), and for state 2 in cell K7 ($\lambda_2 = 10$). Column C is then filled with 1s and 2s, denoting the states, again as for the Gaussian case.

It is now necessary to create two tables of cumulative Poisson probabilities: one for $\lambda_1 = 2.5$ and one for $\lambda_2 = 10$. This is done as before and these are stored in cells F1–G30 for λ_1 and H1–I30 for λ_2. Column D is a second set of random numbers between 0 and 1, obtained in the same way as column B, and is used to obtain the Poisson random variates (i.e. the simulated observations) in column E by means of the command

```
=If($C1=1,VLOOKUP($D1,$F$1:$G$30,2),VLOOKUP($D1,$H$1:$I$30,2))
```

entered in cell E1 and dragged down that column. Fig. 5 is an illustration of the resulting spreadsheet.

The simulated Poisson counts in column E and the corresponding states in column C can be viewed again by selecting the command chain chart wizard ⇒ XY (Scatter) ⇒ data points connected with lines without markers. Fig. 6 is an illustration.

It is again possible to experiment by changing the parameters of the model (such as the transition and initial probabilities, and state means), observing the changes graphically. Such experimentation builds an appreciation of the role for various parameters in the model. This is where the power of the spreadsheet simulation lies: seeing instantly changes in the model and its output when the input parameters are changed.

4. Other applications

This section contains brief descriptions of two interesting data sets, to which the above methods may be applied: foetal lamb movements and Dow Jones data.

The data on foetal lamb movements were first considered in Leroux (1989) and also used as an example in Leroux and Puterman (1992). It is a relatively small data set based on 240 consecutive 5-s intervals, during which counts of movements by a foetal lamb were observed through ultrasound. The data set is shown in Fig. 7. Leroux and Puterman (1992) found that this data set is best modelled by a hidden Markov model with $m = 2$ states with Poisson rates

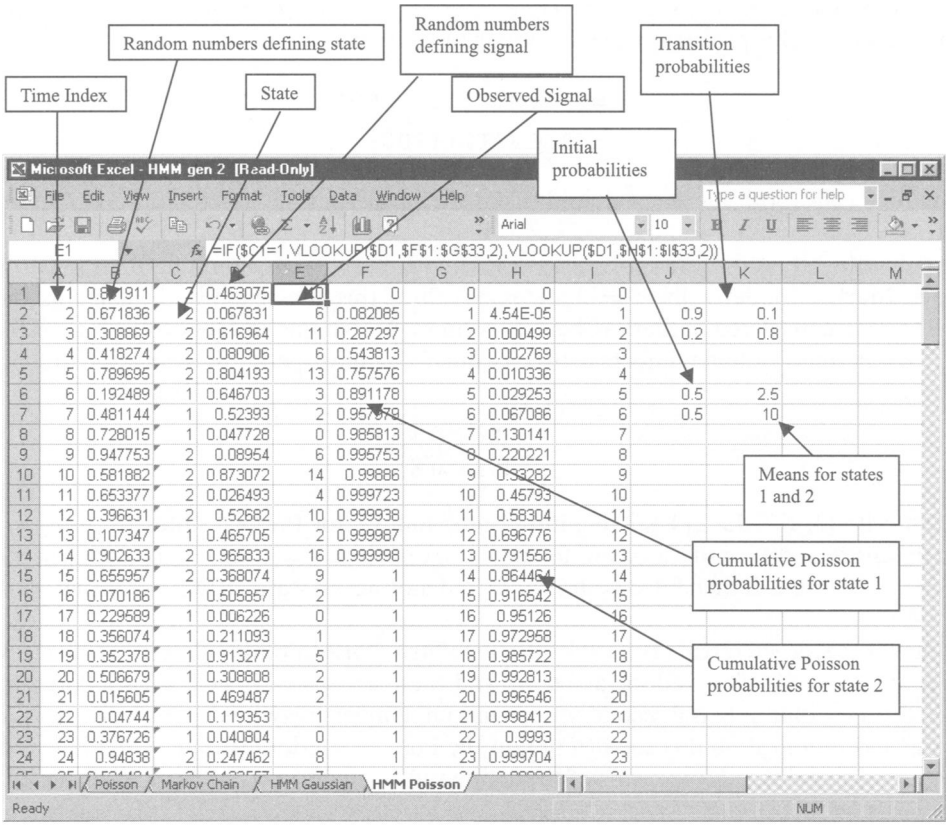

Fig. 5. Top part of a worksheet for the simulation of a hidden Markov model with Poisson states

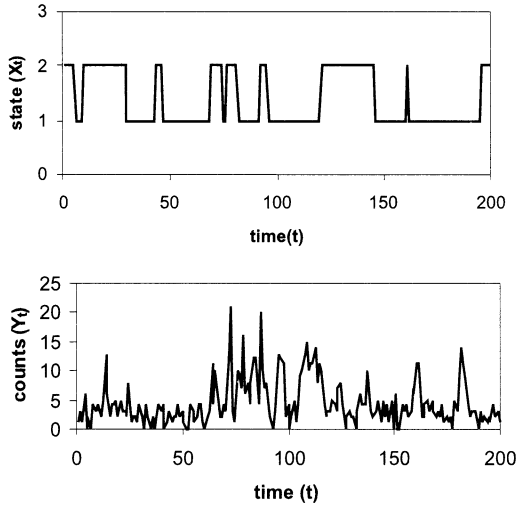

Fig. 6. Simulated counts from Poisson states

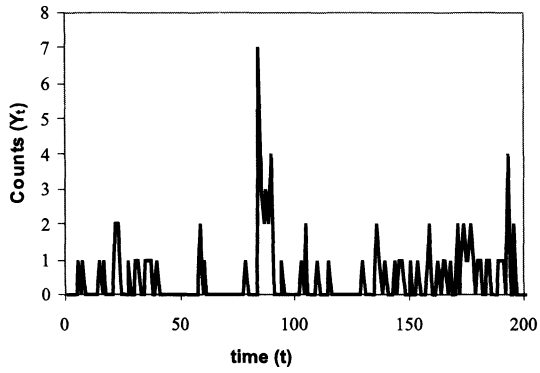

Fig. 7. Foetal lamb counts

such that one state can be interpreted as a normal state (where the foetus spends the majority of its time) and the other state as the excited state.

The Dow Jones data set is taken from Brockwell and Davis (1996) and consists of the daily value of the Dow Jones utilities index from August 28th, 1972, to December 18th, 1972. This data set is shown in Fig. 8(a) and may be differenced to form a new series Y_t, shown in Fig. 8(b). The piecewise straight lines in Fig. 8(a) and the step function in Fig. 8(b) are approximations obtained in the following way. By studying the differenced series Y_t, we might identify periods when that series is in a constant state. The step function in Fig. 8(b) shows the mean for these periods over time, and the piecewise straight lines in Fig. 8(a) are obtained by reversing the differencing. We may conclude that there are perhaps two (or possibly more) hidden states, one with a mean at about -0.2 and the other with a mean at about 0.4, corresponding to what stock analysts refer to as bear and bull markets respectively. This suggests a hidden Markov model with two states.

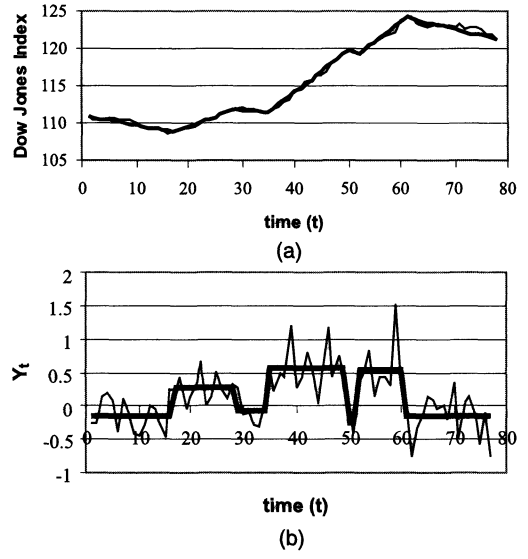

Fig. 8. Dow Jones data set

5. Discussion

The reader will, no doubt, find other examples of time series data for which a hidden Markov model might be an appropriate mathematical representation. For an interesting and recent application of hidden Markov models to image processing, see Li and Gray (2000), where the concept of hidden Markov random fields is used. This concept is also exploited in Zhang *et al.* (2000).

Choosing $m > 2$ does not present any new conceptual difficulties for simulation studies, but some formulae become longer and more involved. It would also be reasonably straightforward to make the modifications that are necessary to simulate autoregressive hidden Markov models; see Douc *et al.* (2001) for definitions and further references.

In spite of its drawbacks, outlined by McCullough and Wilson (1999) and Berk and Carey (1998), EXCEL is a powerful tool for simulation and visualization, and even computation as demonstrated in Orvis (1996). We conclude that, using EXCEL, it is straightforward to create visual simulations of hidden Markov models, thereby deepening the understanding of their behaviour.

Acknowledgement

We are grateful to the Joint Editor for many improvements in the presentation.

References

Berk, K. N. and Carey, P. (1998) *Data Analysis with Microsoft Excel.* Pacific Grove: Duxbury.
Brockwell, P. J. and Davis, R. A. (1996) *Introduction to Time Series and Forecasting*. New York: Springer.
Bulgren, W. G. (1982) *Discrete System Simulation.* Englewood Cliffs: Prentice Hall.
Douc, R., Moulines, E. and Ryden, T. (2001) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. (Available from `http://www.maths.lth.se/matstat/staff/tobias/`.)
Elliott, R. J., Aggoun, L. and Moore, J. B. (1995) *Hidden Markov Models*. New York: Springer.
Gentle, J. E. (1998) *Random Number Generation and Monte Carlo Methods.* New York: Springer.
Leroux, B. G. (1989) Maximum likelihood estimation for mixture distributions and hidden Markov models. *PhD Thesis*. University of British Columbia, Vancouver.
Leroux, B. G. and Puterman, M. L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545–558.
Li, J. and Gray, R. M. (2000) *Image Segmentation and Compression using Hidden Markov Models.* Dordrecht: Kluwer.
MacDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series.* London: Chapman and Hall.
McCullough, B. D. and Wilson, B. (1999) On the accuracy of statistical procedures in Microsoft Excel 97. *Comput. Statist. Data Anal.*, **31**, 27–37.
Orvis, W. J. (1996) *EXCEL for Scientists and Engineers.* San Francisco: Sybex.
Rabiner, L. and Juang, B. H. (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs: Prentice Hall.
Ross, S. M. (2000) *Introduction to Probability Models*. San Diego: Harcourt–Academic Press.
Zhang, Y., Smith, S. and Brady, M. (2000) Hidden Markov random field model and segmentation of brain MR images. *Technical Report TR00YZ1*. Oxford Centre for Functional Magnetic Resonance of the Brain, Oxford. (Available from `http://www.fmrib.ox.ac.uk/analysis/techrep/tr00yz1/`.)