A Survey of Recent Developments in FORECASTING Methods

Dr Chris Chatfield

Department of Mathematical Sciences, University of Bath, Bath, U.K., BA2 7AY

email: cc@maths.bath.ac.uk

Some notes prepared for the 5th Time-Series Workshop at Arrábida, Portugal on July 2-6, 2001, including key points from the overheads but excluding most examples.

Time-series Text: Chatfield (1996) *The Analysis of Time Series*, 5th edition. Chapman & Hall/CRC Press. – abbreviated ATS.
Forecasting Monograph: Chatfield (2001) *Time-Series Forecasting*. Chapman & Hall/CRC Press. – abbreviated TSF.

TOPICS TO BE COVERED

1. Preliminary questions; Problem-formulation; Types of forecasting method; Ex-post versus ex-ante forecasting. Basic ideas of time-series analysis; Time plot; Some probability models for time-series; Model selection.

2. Introduction to the many different time-series forecasting methods. Univariate and multivariate; Linear and non-linear; Can't cover everything. Give overview and avoid technical details.

3. Comparative review. Make general recommendations on choice of 'best' method; Discuss empirical evidence – forecasting competitions;

4. Calculating interval forecasts; Model uncertainty and forecast accuracy.

5. Illustrate with examples.

PRELIMINARY QUESTIONS

Why forecast?

1. Economic planning
2. Sales forecasting
3. Production planning
4. Inventory control
5. Evaluation of alternative economic strategies

What is the problem?

Important to **FORMULATE PROBLEM** properly (as in any statistical exercise). Often hardest part. *Ask Questions* to get background information and *Clarify Objectives.* Build on prior information. Use **Context**.

1) *What is purpose of forecast?*

Vital to clarify objectives. e.g production planning; planning the economy; informed management decision-making; etc. Find out exactly how forecast will be used. So talk to forecast users.

Systems Approach. Forecasts should be an integral part of planning system, not a separate exercise. A simple approach, which is widely understood, may be preferred.

Do we want:

- forecasts only?
- a useful, descriptive model?
- an interpretable model?
- forecasts for use in control?

No point in giving **RIGHT** answer to **WRONG** question – an *error of the third kind.*

A (Silly?) Example.

Statistician's forecast of sales = 34,000

Managing Director's forecast (or target?) = 40,000

Sales Director's forecast (or target?) = 37,000 – half-way between!

What should statistician's objectives be? Not much point in high accuracy unless attitudes change.

Further questions:

2) What accuracy is required? (Unrealistic expectations?)

3) What data are available?

4) How many series are there to forecast?

5) How far ahead are forecasts required?

Notation.

Observed data: $x_1, x_2, \ldots, x_N$

Problem: to forecast x_{N+h} where h = *lead time* or *forecasting horizon.*

Denote forecast by $\hat{x}(N,h)$ or $\hat{x}_N(h)$

Must specify both the time forecast is made AND the lead time - Don't use $\hat{x}_{N+h}$

Forecasting Methods – Different Approaches.

Methods can be broadly classified into:

a) Judgemental methods (subjective, qualitative)

b) Univariate time series methods (projection, extrapolation). $\hat{x}(N,h)$ depends only on $x_N, x_{N-1}, \ldots$ and/or on position in time.

c) Multivariate time series methods. $\hat{x}(N,h)$ depends also on predictor variables.

d) Combination of a), b), c).

Methods can also be classified as

Automatic or **Non**-automatic – an important distinction.

– contrast stock control with econometrics

With many items, as in stock control, need simple automatic approach. In economics, prefer non-automatic approach – probably multivariate.

Which method is best?

No simple answer. (Horses for Courses!) Wide variety of problems requiring different treatment. Big gains can be made by being selective.

Choice of method depends on:

a) Objectives

b) Type of data (eg. macro or micro) and properties of series, particularly presence/absence of trend/seasonality.

c) Number of observations available

d) Length of forecasting horizon.

e) No. of series to be forecasted and cost allowed per series. What accuracy is required?

f) Skill and experience of analyst and computer software available. Use a method you feel 'happy' with. Try more than one method?

SPECIAL WARNING :

Forecasting is extrapolation, which can be dangerous. Forecasts are conditional statements that if such-and-such behaviour continues, then..... Be prepared to modify them in light of any additional information.

Quotes from Schumacher's "Small is Beautiful," Chapter 15.

Short-term forecasts. "A refined technique rarely produces significantly different results from those of a crude technique."

Long-term forecasts "are presumptious. However long-term feasibility studies, based on clearly stated assumptions, are well worth doing."

Good idea to construct alternative *scenarios* based on different assumptions.

Some dubious predictions!!

(i) "I think there is a world market for about five computers" – Founder of IBM in 1947

(ii) "There is no reason for any individual to have a computer in their home" – President of Digital Equipment in 1977

(iii) "Stock prices have reached what looks like a permanently high plateau" – Yale Professor of Economics in September 1929 (before crash!).

Out-of-sample versus in-sample forecasts. OR

Ex-ante versus ex-post forecasts.

If $\hat{x}_N(h)$ only uses data up to time N $\Rightarrow$ *out-of-sample* (or *ex-ante*) forecast

Note: Can't use future data when computing forecasts OR in fitting the model.

In-sample 'forecasts' are NOT real forecasts.

In-sample 1-step-ahead forecast errors are the residuals

Typically find:

In-sample residual variance $\ll$ Out-of-sample forecast error variance.

For MV forecasts, may need forecasts of explanatory variables as well as of response variable. For forecasts to be genuinely ex-ante, all forecasts, including those of explanatory variables, should only use data to time N.

Ex-post forecasting.

If, instead of forecasting explanatory variables, use actual values of the future explanatory variables to compute forecasts (to assess model)

OR

Use assigned values of the future explanatory variables to compute forecasts (to assess different strategies)

There are many ways to 'cheat' – for example:

1. Divide data into training set and test set, but fit model to all the data before 'forecasting' test set.

2. Fit several models to training set and choose the one which gives 'best' forecasts of test set. Then re-use this model to give forecasts.

Some General Forecasting References

Books

Chatfield (2001); Diebold (2001); Montgomery, Johnson and Gardiner (1990);

Granger & Newbold. 1986. 2nd edn. Economic flavour.

Box & Jenkins. 1970. The famous book. NEW 3rd edn. 1994 with G. Reinsel. New chapters on intervention analysis, outliers, process control.

Note: There are some *bad* books around!

Journals

The International Institute of Forecasting (IIF) sponsored J. of Forecasting (JoF) 1981-5 and Int. J. of Forecasting (IJoF) 1985-. JoF is still published by Wiley.

See also J. Business & Econ. Stats. (JBES), Management Science, etc.

Review paper: Chatfield (1997, The Statistician).

BASIC IDEAS of TIME-SERIES ANALYSIS (TSA).

1. Introduction.

A *time series* is a collection of observations made sequentially through time.

Examples in economics, marine science, marketing,...

If observations taken at discrete times $\rightarrow$ *discrete* time series $\{X_t\}$ – though observed variable may be discrete *or* continuous.

If observations taken continuously through time $\rightarrow$ *continuous* time series $\{X(t)\}$.

May be *deterministic* or *random* (stochastic).

Discrete series may be (i) *sampled* from a continuous series (e.g. temperature), (ii) *aggregated* (e.g. sales in successive months) or (iii) really discrete as observations only taken at fixed intervals (e.g. annual dividend).

Much stat. theory is about *independent* obs. Special feature of time series is that obs. are NOT independent. Must take *order* into account.

Objectives.

a) *Description.* Plot data. Summarize features e.g. trend, seasonality, outliers. Fit model.

b) *Explanation.* Multivariate modelling.

c) *Forecasting* (or prediction). For what purpose?

d) *Control.*

Approaches.

Analysis in time domain. Based on autocorrelation function.

Analysis in frequency domain. Based on spectral analysis.

Review of Books.

Introductory time-series texts: Chatfield, 1996; Harvey, 1993; Kendall & Ord, 1990. Wei, 1990.

More advanced texts include: Brockwell and Davis (1991, 2nd edn.); Fuller (1996, 2nd edn.)

2. Simple Descriptive Techniques

See ATS, Chapter 2. Classical TSA decomposes variation into:

1. *Seasonal variation* – usually annual
2. *Other cyclic changes* – e.g. economic cycles
3. *Trend* – long-term change in mean level
4. *Irregular fluctuations* – may not be random

Good approach when variation dominated by trend and/or seasonality. But not when short-term correlation present. And decomposition into trend/seasonality not unique.

2.3 The Time Plot

Plot obs. against time.

MOST IMPORTANT STEP in any TSA or forecasting exercise.

Shows up important features such as: (i) trend; (ii) seasonality; (iii) outliers; (iv) turning points/discontinuities.

Vital to describe data and help in formulating a sensible model.

Drawing a time plot is not as easy as it sounds.

General guidelines:

a) Give clear title

b) Label axes

c) State units of measurement

d) Careful choice of scales

e) Careful choice of plotting symbol

f) Use trial-and-error to improve

- Guidelines often disregarded, especially by some PC packages. Some computer graphs are AWFUL!! So use Tippex (!) or a better package which does give control over output.
- Avoid deception

Time plot is vital for (i) describing data; (ii) formulating a model; (iii) Choosing appropriate analysis.

2.4 Transformations

May be better to analyse say $\sqrt{X_t}$ or $\log X_t$

WHY? – (a) To stabilise variance; (b) To make seasonal effect additive; (c) To make data normal. Unfortunately these requirements may conflict.

HOW? Box-Cox transformation

$$X_t^{(\lambda)} = \begin{cases} (X_t^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log X_t & \lambda = 0 \end{cases}$$

My preference is generally to analyse raw data except when growth is exponential. Then take logs to make it additive (i.e. %age increases are of interest).

2.5 Trend

Difficult to define – "Long-term" change in level per unit time.

Perception of trend may depend on length of series. Trend or low-frequency variation?

Is trend linear or non-linear?

Contrast **global** or *deterministic* linear trend – $\mu_t = a + bt$ – generally unrealistic?

with **local** linear trend – $\mu_t = a_t + b_t t$

Econometricians distinguish between *trend-stationary* – deviations from a deterministic trend are stationary – and *difference-stationary* series where stationarity can be induced by first differencing, so a 'unit root' is present: $X_t - X_{t-1} = (1 - B)X_t$, where B denotes the *backward shift operator* such that $BX_t = X_{t-1}$.

May want to measure trend, OR remove trend so as to examine local fluctuations.

Can fit a trend curve such as a *polynomial* function of time (t), or a *Gompertz* curve – $\exp(a - br^t)$, where $0 < r < 1$, or a *logistic* curve – $a/(1 + be^{-ct})$.

Can measure or remove trend with a *linear filter*, (e.g. Henderson moving average), of general form

$$y_t = \sum a_r x_{t+r}$$

If $\sum a_r = 1 \quad \rightarrow$ measure trend

If $\sum a_r = 0 \quad \rightarrow$ remove trend

Differencing is a special type of trend removal

$$y_t = x_t - x_{t-1} = \nabla x_t \quad \text{where } \nabla \text{ is the first differencing operator.}$$

Problem: No unique decomposition into trend and seasonality.

2.6 Seasonal Models

$$X_t = m_t + S_t + \varepsilon_t - \textit{additive}$$

$$X_t = m_t S_t + \varepsilon_t - \textit{multiplicative} \text{ seasonality but additive error}$$

$$X_t = m_t S_t \varepsilon_t - \text{both multiplicative}$$

$$\rightarrow \log X_t = \log m_t + \log S_t + \log \varepsilon_t - \text{additive}$$

Normalize? $\sum S_t = 0$ – additive OR $\text{Av}(S_t) = 1$ – multiplicative

Mixed seasonal: $X_t = m_t(1 + \beta_t) + \alpha_t + \varepsilon_t$

Various ways of measuring and/or removing seasonality. e.g. *seasonal differencing* for monthly data – $y_t = x_t - x_{t-12} = \nabla_{12} x_t$, where ∇_{12} is the seasonal differencing operator for data with period 12. *Henderson moving average* used in X-11 (US Census Bureau) and in X-11-ARIMA (Canada). X-12-ARIMA (Enhanced X-11; e.g. deals with holiday effects) is now available. OR use Maravall's TRAMO/SEATS – stands for 'Time-Series Regression with ARIMA noise/Signal Extraction in ARIMA Time Series'.

Important comment. The treatment of (i) Outliers (ii) Missing observations (iii) Calendar or Trading Day or Holiday effects (e.g. Easter in March or April; 4 or 5 Sundays in a month; etc) can be more crucial than other aspects of TSA. More important than choice of model or forecasting method.

2.7 The Correlogram

An important tool for assessing the behaviour and properties of a time series.

Measures correlation between observations at different distances apart.

Revision. Given N pairs of obs. on x and y, say $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, then

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[(\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]}$$

Applying to TS. Given series $x_1, x_2, \ldots, x_N$, can form $(N-1)$ pairs of obs. $(x_1, x_2), (x_2, x_3), \ldots, (x_{N-1}, x_N)$. Then calculate correlation between x_t and x_{t+1} by

$$r_1 \simeq \sum_{t=1}^{N-1} (x_t - \bar{x})(x_{t+1} - \bar{x}) / \sum_{t=1}^{N} (x_t - \bar{x})^2$$

= (auto)correlation at lag l

More generally to assess correlation between obs. k steps apart:

$$r_k = \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) / \sum_{t=1}^{N} (x_t - \bar{x})^2$$

Plot r_k against $k \rightarrow$ **correlogram**. Note : r_0 is always 1.

For random series, $r_k \simeq 0$ for all $k \neq 0$. If random, can show $r_k \sim$ approx. $N(0, 1/N)$.

So values outside $\pm 2/\sqrt{N}$ are sig. diff. from zero. For example if $N = 100$, values outside $\pm 2/\sqrt{100} = \pm 0.2$ are sig.

3. Probability Models for Time Series

See ATS, Chapter 3. Regard x_t as obs. on a **random variable** X_t. Models describe distribution of X_t and relationship with past values of X_t (and with other series?).

Stationary if distribution of X_t does not depend on t

$\rightarrow$ constant mean – μ

$\rightarrow$ constant variance – σ^2

$\rightarrow$ constant autocorrelation function (ac.f.) – $\rho(k)$ – where $\rho(k) =$ Correlation (X_t, X_{t+k})

3.4.1. Purely random process or white noise

$Z_1, Z_2, \ldots$ are uncorrelated and have same distribution. Then $\rho(k) = 1$ if $k = 0$ and zero otherwise. If Z_t are normally distributed, then they are also independent. Some writers assume independence. Need independence for non-linear models. Or use notation $\{\varepsilon_t\}$.

3.4.2 Random Walk

$X_t = X_{t-1} + Z_t$ is non-stationary

while $(X_t - X_{t-1}) = \nabla X_t = Z_t$ is stationary.

e.g. share price on day $t =$ share price on day $(t-1) +$ random perturbation.

Then share prices approx. a random walk. First differences of share prices (i.e. *change* in share prices) approx. white noise.

3.4.3 Moving Average Process

$X_t = Z_t + \beta_1 Z_{t-1}$ – MA(1)

$X_t = Z_t + \beta_1 Z_{t-1} + \ldots \beta_q Z_{t-q}$ – MA(q)

X_t and X_{t+k} are independent if $k > q$. So $\rho(k) = 0$ for $k > q$ and correlogram 'cuts off' at lag q.

Process is stationary for any values of β_i.

3.4.4 Autoregressive Process

$X_t = \alpha_1 X_{t-1} + Z_t$ – AR(1)

$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t$ – AR(p).

Like a multiple regression model but AUTO.

Find $\rho(k)$ decreases exponentially or like a damped sine wave.

For stationarity, must have α's 'reasonably small' e.g. for AR(1) $\rightarrow |\alpha| < 1$

3.4.5 Mixed ARMA Process

$X_t = \alpha_1 X_{t-1} + Z_t + \beta_1 Z_{t-1}$ – ARMA (1,1)

3.4.6 Integrated ARIMA Process

to describe non-stationary series.

e.g. Let $W_t = \nabla X_t = X_t - X_{t-1}$ and

$W_t = \alpha_1 W_{t-1} + Z_t + \beta_1 Z_{t-1}$

Then W_t is ARMA(1,1) and X_t is said to be ARIMA (1,1,1)

General notation – ARIMA (p,d,q) where p = no. of AR terms, d = no. of differences, and q = no. of MA terms.

Deterministic trend. – Trend-stationary series

$$X_t = a + bt + \varepsilon_t$$

$$\rightarrow \; X_t - X_{t-1} = b + \varepsilon_t - \varepsilon_{t-1}$$

LHS is $(1 - B)X_t$ which has a unit root. But RHS is $= b + (1 - B)\varepsilon_t$ and error term is not invertible. So don't take differences when trend stationary.

Difference-stationary series.

Suppose $(1-B)X_t$ is stationary. Then process has a unit root and does not have a deterministic trend. Then we DO need to take differences and NOT fit a deterministic trend.

Important to identify correct form of trend and correct form of differencing.

Testing for a unit root. Some financial models should contain unit roots under rational expectations and so a unit root null hypothesis is then reasonable. Can also arise in testing for co-integration – see later. But not always sensible to take a unit root as null hypothesis.

Time-Series Model-Building (Expansion of Section 4.8 of ATS; See Section 3.5 of TSF)

Some *general principles* apply in all statistical applications, but there are special problems with *(auto)correlated* time-series data. Can be difficult to separate and model effects of trend, seasonality, lagged explanatory variables, etc.

Books and journals concentrate on details of *techniques*, BUT we also need to be concerned with general *strategy*.

e.g. Problem is usually NOT "How do we fit a particular Box-Jenkins ARIMA model?", but "Is an ARIMA model appropriate here?" or "Which particular ARIMA model should we fit"?

Objectives: Why model??

1. *Description.* Describe both (i) the *systematic* variation; and (ii) the 'error' term.
2. *Prediction.*
3. To confirm or refute a priori theory
4. To facilitate comparisons
5. To give physical insight

All models are approximations. The approximation should be adequate for the task in hand, and should contain as few parameters as necessary to do this (i.e. be parsimonious). The *Principle of Parsimony* is sometimes referred to as *Occam's Razor.*

Black-Box or Structural Model? Black-Box models are also called *empirical* models. Constructed from data in fairly mechanical way. Easy to do. But may give little physical insight. Neural Net (NN) can be extreme example.
Structural models: Account for known theory and specific physical features. Need subject-matter knowledge to construct intelligent model. Hard to make general comments.
Need both types of model in different applications. Some models in-between. Autoregressive (AR) models are closer to black-box than Vector AR models.

Stages in Model-Building:

1) *Specification* of a class of models (Model formulation)
2) *Selection* of a 'best' model.
3) *Model-fitting* – the easy bit?
   – but lectures and books concentrate on estimating model parameters!
4) *Model-validation* (or model-checking) – usually a residual analysis.
   – Is model OK or does it need to be adjusted?

May be several cycles of model-fitting – an *iterative, interactive* process – as in Box-Jenkins modelling of time series (Box *et al.*, 1994). Need *inductive* and *deductive* reasoning.

Stage 1 of Model-Building: Specification

"All models are wrong but some are useful"!!
Context, costs and *objectives* are important. Ask questions; Get background information
What variables need to be included? Very important. Don't include too many but don't omit any key variables. Avoid linearly related variables – e.g. if X and Y, then not $(X + Y)$.
Use known theory; Known limiting behaviour; Known special cases;
Look at data; Look at time plot, correlogram, etc. Should suggest what assumptions are reasonable.

Strange that we say/teach so little about *model-specification*. Much more difficult than model-fitting. *Experience* and *inspiration* important. Analysts like to think modelling is *objective* but *subjective* judgement is always needed.

Stage 2. Model-Selection. Choose a 'best' model from the specified class of models. What is meant by 'best'? Best fit? Or best out-of-sample predictions? Or ...?

Various approaches.

a) Use Subjective Judgement. Look at the time plot, the ac.f. and the partial ac.f. of the observed series and of the differenced series. Use judgement to select a model.

b) Use a Model Selection Statistic Choose a model to optimize a statistic that measures fit in some way. How do you measure 'best fit'?

Can't just minimize residual sum of squares (or maximize R^2) as models with more parameters usually give better fit but not necessarily better out-of-sample predictions.

Always a danger that looking at lots of models (*data mining*) will give a model with lots of parameters, which appears to give a good fit, but which gives poor out-of-sample predictions.

Can choose a model to minimize a statistic like *Akaike's Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC) which penalize extra parameters in assessing fit.

$$\text{AIC} = \quad -2\ln(L) \quad + 2p$$

$$\text{- Fit} \qquad + \text{complexity of model}$$

where L= max. likelihood and p= no. of (independent) parameters

To a first approximation, BIC replaces $2p$ with $p + p\log(N)$, where N= no. of observations. When p becomes large compared with N, better to use bias-corrected version of AIC, denoted by AIC_C or AICC, which replaces $2p$ with $2(p+1)N/(N-p-2)$.

c) Using Tests of Hypotheses to Select a Model.

Hypothesis-testing is standard approach to model selection in experimental design (e.g. 1-way ANOVA). More controversial in time-series analysis?

Econometricians typically carry out a series of tests for presence of trend, seasonality, autocorrelated residuals, unit roots, constant variance, causal explanatory variables, etc.

But statisticians typically rely more on an initial examination of data and model-selection statistics such as AIC.

Dangerous to *over*test a single set of data?

When testing, have to:

1. Specify null hypothesis (H_0)
2. Compute test statistic to show up departures from H_0
3. Compute P-value = Prob (more extreme value than one obtained if H_0 true) (and power?)

4. Reject or fail to reject H_0.

Do not get Prob (H_0 true)!!

Example. Want to fit 'best' Autoregressive (AR) model

$$X_t = \mu + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t$$

What is p? Could introduce lagged variables one by one (which order?) and test each in turn. But doing lots of tests affects P-values and subsequent inferences.

OR choose p to minimize AIC.

Example. Testing for unit roots.

ATS Section 13.5.7. Many versions of test. For simplicity consider

$$X_t = \phi X_{t-1} + \varepsilon_t$$

$$\text{OR } (1 - \phi B) X_t = \varepsilon_t$$

If $\phi = 1$, then the equation $(1 - \phi B) = 0$ has a *unit root* (i.e. $B = 1$ satisfies equation) and process is non-stationary. Query: Why should $\phi = 1$ be H_0? Sometimes sensible economically. Tests generally have poor power and topic still controversial – see Diebold and Kilian (2000, JBES)

Prefer to fit a model which allows trend to adapt through time (e.g. as in structural state-space models) rather than impose a fixed or deterministic structure? Depends on context.

Other Reasons for Preferring Selection to Testing.

1) Get a *ranking* of models.

e.g. Clear winner? OR Several close competing models?

2) No need to decide what null hypothesis should be.

3) No need to assume *true model* exists and is in class of candidate models.

4) Valid for comparing non-nested models.

e.g. ARIMA versus Neural nets versus Econometric model

Stage 3. Model-Fitting

Plenty of packages available. Little needs to be said. Some models (e.g. GARCH; Neural Network (NN); Vector Autoregressive (VAR)) may require computationally intensive methods. Which approach to inference? Classical, Bayesian or Decision Theory. Different approaches relevant to different situations. No need to label ourselves.

Stage 4. Model-Validation

Is fitted model consistent with data? Look at *residuals* where

DATA = FIT + RESIDUAL

In time-series analysis, residual = one-step-ahead forecast error.

Plot residuals against time. Signs of change?

Large residual $\rightarrow$ outlier? or error?

– or wrong model fitted?

– or genuine extreme obs.? (use *robust* methods?)

Is systematic part of model OK? Is 'error' part OK?

If necessary go back and modify model.

Continue to check model while forecasting $\rightarrow$ *Forecast monitoring*; Look at residuals. Plot against time. If things go wrong, revise model.

Some Other Points on Modelling.

a) May not be able to find a model which describes *all* the data satisfactorily.

b) If trend and seasonal effects present, should we model them explicitly

OR remove by filtering?

c) Problems in modelling 50 monthly sales figures very different to modelling 1000 daily share prices.

d) Univariate or multivariate time-series model? Multivariate harder to fit and forecasts need not be better, but does provide extra insight and may give better forecasts.

e) In TSA, formulate and fit model to *same* data $\rightarrow$ model-selection biases. Model uncertainty.

5. FORECASTING METHODS

See ATS, Chapter 5; TSF, Chapters 4 and 5. There is a rich variety of methods. Cover the more important ones.

Judgemental Methods: Delphi method; Bold freehand extrapolation (BFE); Manager's judgement; etc. Sometimes work well – sometimes not – see for example Armstrong (1985, Chapter 15); Webby and O'Connor (1996, IJoF).

Not covered in this course as prefer method to be at least partly 'objective' and quantitative.

5.2. Some UNIVARIATE Methods

Univariate methods sometimes more relevant to everyday needs of statistician than multivariate methods – especially when lots of series to forecast.

1) **Extrapolation of trend curves** (or growth curves).

For long-term forecasting of non-seasonal data. Fit polynomial, Gompertz, or logistic curve, etc., using:

$$X_t = f(t) + \varepsilon_t$$

2) **Exponential Smoothing**

$$\hat{x}(N,1) = \alpha x_N + \alpha(1-\alpha)x_{N-1} + \alpha(1-\alpha)^2 x_{N-2} + \ldots$$

– geometric series; easily updated when rewritten as: $\hat{x}(N,1) = \alpha x_N + (1-\alpha)\hat{x}(N-1,1)$ – recurrence form, or as $\hat{x}(N,1) = \hat{x}(N,1) + \alpha e_N$ – error-correction form

where $e_N = x_N - \hat{x}(N-1,1)$

α is the **smoothing parameter** such that $0 < \alpha < 1$. Often chosen between 0.1 and 0.3. OR choose to min. $\sum_{t=2}^{N} e_t^2$.

ES is optimal for an ARIMA(0,1,1) model and for several other models (Chatfield *et al.*, 2001).

Gardner (1985, JoF) gives review.

3) **The Holt-Winters Method**

$$X_t = L_t + I_t + \varepsilon_t$$

Where L_t = current Level, I_t = current Seasonal index, and

$$L_t \simeq L_{t-1} + T_{t-1}$$

where T_t = current Trend. This is the additive seasonality case.

– Update L_t, T_t, I_t by ES – formulae in ATS

– Easy-to-use. Robust (Chen, 1997, IJoF).

There are now 3 *Smoothing Parameters* – Choose to min. $\sum e_t^2$

Choice of *starting values* for L_t, T_t, I_t also important; e.g. L_1 = average value in 1st year.

A *multiplicative* version is also available.

Can be used with *automatic* or *non-automatic approach*.

Forecasts in multiplicative case for monthly obs.

$$\hat{x}(N,h) = (L_N + hT_N)I_{N-12+h}$$

for $h = 1, 2, \ldots, 12$.

4) **The Box-Jenkins Forecasting Procedure**.

Iterative model-building procedure.

1) *Identify* appropriate ARIMA model (by looking at correlogram of various differenced series and partial autocorrelation function etc.) Not easy. Requires experience.

2) *Estimate* the parameters of the model. Easy using package such as MINITAB, AUTOBOX etc.

3) *Diagnostic checking.* Examine residuals from fitted model.

4) Consider alternative models if necessary.

Step 1. Difference data until *stationary.*

e.g. ∇ for non-seasonal data. or ∇_{12} or $\nabla\nabla_{12}$ for monthly seasonal data.

Use minimum degree of differencing so that correlogram comes down to zero fairly quickly.

Step 2. Formulate ARMA (or seasonal ARMA) model for the differenced series, by looking at correlogram, partial autocorrelation function, etc. of differenced series. – see ATS, Chapter 4.

Ex. Suppose X_t is non-stationary, but ∇X_t is stationary with 'large' r_1 but 'small' r_k otherwise $\rightarrow$ ARMA(0,1,1) has $\rho(k)$ of this form.

Step 3. Estimate the parameters of the chosen ARMA model. Nowadays this is EASY using a package such as MINITAB, GENSTAT or a specialized package such as AUTOBOX.

Still helpful to understand the procedure. Usually minimize $S = \sum e_t^2 = \sum[x_t - \hat{x}(t-1,1)]^2$

Ex. For AR(1) model, $X_t = \phi X_{t-1} + Z_t$, $\hat{x}(t-1,1) = \phi x_{t-1}$, so $S = \sum(x_t - \phi x_{t-1})^2$. This nice analytic function of ϕ can be differentiated to give $\hat{\phi} = r_1$.

Ex. For MA(1) model, $X_t = Z_t + \theta Z_{t-1}$, $dS/d\theta$ cannot be evaluated analytically. So have to find $\hat{\theta}$ in (-1,1) which min. S numerically. Can use hill-climbing.

Step 4. Check model. (Model-validation. Diagnostic checking).

Look at 1-step-ahead 'errors' (the *residuals*). If they are not random, model is not optimal. More structure waiting to be found.

Step 5. Compute forecasts if model satisfactory.

Easily done by replacing (i) future Z's with their conditional expectation (zero!), (ii) future X's with their forecasted values, (iii) past values of X and Z by their observed values. This gives the MMSE forecast: $\hat{X}(N,k) = E(X_{N+k}|X_N, X_{N-1}, \ldots)$ = conditional expectation of X_{N+k}.

Ex. AR(1). $X_t = \phi X_{t-1} + Z_t$

$\hat{x}(t,1) = E(X_{t+1}|X_t, X_{t-1}, ...) = \phi x_t$ (or rather $\hat{\phi} x_t$)

$\hat{x}(t,2) = \phi\hat{x}(t,1) = \phi^2 x_t$

Ex. MA(1). $X_t = Z_t + \theta Z_{t-1}$

$\hat{x}(t,1) = \theta Z_t$ (or rather $\hat{\theta}\hat{z}_t$)

$\hat{x}(t,2) = 0$

Ex. Independent obs. $X_t = \mu + Z_t$

$\hat{x}(t,1) = \mu = \hat{x}(t,k)$ for $k = 2, 3, \ldots,$

Ex. Random Walk or ARIMA (0,1,0). $X_t = X_{t-1} + Z_t$ or $\nabla X_t = Z_t$

$\hat{x}(t,1) = x_t = \hat{x}(t,k)$ for $k = 2, 3 \ldots,$

Ex. ARIMA (1,1,0). $\nabla X_t = \phi\nabla X_{t-1} + Z_t$, OR $X_t = X_{t-1} + \phi(X_{t-1} - X_{t-2}) + Z_t$

Then $\hat{x}(t,1) = x_t + \phi(x_t - x_{t-1})$

etc. – See ATS, Chapters 3,4, & Section 5.2.4.

Ex. SARIMA (1,0,0)(0,1,1)$_{12}$ – see ATS, page 74

Fractional ARIMA models. In ARIMA (p, d, q) model, allow d to be non-integer $\rightarrow$ *Fractional differencing.*

If $0 < d \leq 0.5$, model is stationary and an example of a *long-memory* model (e.g. Beran, 1994), where ac.f. decreases slowly. Fractional models harder to interpret? Forecast accuracy with real data is inconclusive (e.g. Crato and Ray, 96, IJoF).

5. Stepwise Autoregression.

6. State-Space and Structural Models

Harvey (1989). ATS Chap 10. – Much current interest. Andrews (1994, JBES) – Comparative empirical study. Does quite well, especially for long horizons and seasonal data.

Structural models are special case of **state-space models**. So are **Unobserved Component Models** and the West/Harrison (1997) **Dynamic Linear Models**.

Use **Kalman Filter** – a recursive method of signal processing which gives an optimal estimate of the current state of the system in the presence of noise. (Also used now for estimating parameters of ARMA model and in other applications). Forecasts can also be produced using the KF.

The Basic Structural Model (BSM) assumes additive level, trend, seasonal index and error terms.

x_t = observation at time t = $L_t + I_t + n_t$

– the **observation** equation, where

$$L_t = L_{t-1} + T_{t-1} + w_{1,t}$$

$$T_t = T_{t-1} + w_{2,t}$$

$$I_t = -\sum_{j=1}^{s-1} I_{t-j} + w_{3,t}$$

– three **transition** equations

– L_t, T_t, and I_t are unobservable state variables and s is number of observations per year.

Similar to additive HW – 3 smoothing parameters $\Longleftrightarrow \sigma_1^2/\sigma_n^2, \sigma_2^2/\sigma_n^2, \sigma_3^2/\sigma_n^2$.

Basic Structural Modelling.

KF provides local estimates of level, trend and seasonality.

– updating equations are closely related to those of HW but not the same

– more complicated to use and explain

– but based on proper probability model

– so can get prediction intervals

– can handle irregularities in data

– can be extended to incorporate explanatory variables

General Question. Is it better to difference away trend and seasonal as in BJ or use a method which models them explicitly as in HW and structural modelling? Or seasonally adjust data in some other way and fit non-seasonal model?

Bayesian Forecasting

West and Harrison (1997); Pole *et al.* (1994).

Depends on a class of models called *Dynamic Linear Models.* Really just state-space models. They derive KF as a way of updating 'priors' to get 'posteriors' using a Bayesian approach. Is it different? Recurrence relations are essentially equivalent to KF but Harrison says Bayesian forecasting is not based upon KF. Useful for short series when one really does have prior information. Can also allow *multi-process* or *mixture* models.

7. Other Univariate Methods. Many other methods such as: General Exponential Smoothing (GES); Regression on time; ARARMA.

Or **Combination** of methods – usually more accurate, but no model (e.g. see IJoF, 1989, No. 4).

5.3 MULTIVARIATE Forecasting Methods

Try to improve forecasts of y_t by including explanatory variables $x_{1t}, x_{2t}, \ldots$ in model.

Must identify all relevant variables. So ASK QUESTIONS.

Many types of model. Multivariate (MV) modelling much more difficult. Must model dependence within AND between series (auto- and cross-correlation).

One basic question: Is there a *causal relationship* between x_t's and y_t (i.e. is system *open-loop* or *closed-loop.*)

Are MV forecasts better?

Perhaps YES. But perhaps NO! – though may still improve understanding of inter-relationships.

Some problems:

MV models much more difficult to identify.

More parameters to estimate $\rightarrow$ more parameter uncertainty

More variables $\rightarrow$ more errors and outliers

Some data unsuitable for MV modelling.

May need forecasts of explanatory variables. (Are forecasts genuinely out-of-sample?).

MV models less robust

Initial Data Analysis.

Look at time plot of each variable.

Compute ac.f. for each variable.

Compute cross-correlation function for each pair of variables.

$r_{XY}(k)$ = Correlation (X_t, Y_{t+k}).

May be difficult to interpret $r_{XY}(k)$ because Variance$(r_{XY}(k))$ depends on autocorrelations.

Much easier if one series is white noise. So some identification methods aim to transform one series to white noise.

1. Multiple Regression.

Most commonly used method. Model is: $E(y|x_1, x_2, \ldots) = \mu + \beta_1 x_1 + \beta_2 x_2 + \ldots$

Problems with time-series data, especially economic data. Can often get excellent *fit* e.g. $R^2 = 0.99$; but poor *forecasts*. Good fit may be *spurious* if (i) all series correlated with time; (ii) too many x's used; (iii) x's highly correlated - often are with economic data. ('Independent' variables usually NOT independent).

Beware of x's held more or less constant in past.

Beware of feedback from y to x's (when x's not controlled). Single-equation model only suitable for open-loop data (with uni-directional causality).

M.R. can be helpful but can also be dangerous. The error structure is too simplistic.

Econometricians try to allow for this with Generalized LS and 2-stage LS. Only partially successful.

Leading indicators.

x_t is a *leading indicator* for y_t if

$$y_t = \mu + \beta x_{t-d} + e_t$$

where $d > 0$. Good for forecasting up to d steps ahead. If d small or zero, may have to forecast x to forecast y!!

Precautions

1) Remove trend and seasonality *before* regression.

2) Choose explanatory variables carefully. Limit number to say 5.

3) Include appropriate lagged values.

4) Make careful diagnostic checks.

5) Avoid if feedback thought to be present.

2. Transfer function model

Consider MR model. No reason why errors should be uncorrelated. No reason why input should only have effect at one lag. So consider more general lead-lag relationship.

$$Y_t = \nu_0 X_t + \nu_1 X_{t-1} + \ldots = \nu(B) X_t + \eta_t$$

Here $\nu(B)$ is called a transfer function while η_t may follow an ARIMA process.

This is called a *Transfer Function* or *Distributed Lag* model.

Here the *response variable*, Y_t, is *endogenous* and *input*, X_t, is *exogenous*. Good if X_t is a leading indicator. e.g. if

$$\nu(B)X_t = k_1 X_{t-d} + k_2 X_{t-d-1} + \ldots$$

Safer to fit with BJ approach than with variations of MR especially if η_t are correlated (ARIMA).

Rational Distributed Lag Model

May be able to use less parameters if write $\nu(B)$ as *ratio* of two polynomials, say $\omega(B)/\delta(B)$

$\rightarrow$ *Rational Distributed Lag Model*

Example: Consider $Y_t = \delta_1 Y_{t-1} + \omega_3 X_{t-3}$

where $\delta(B) = (1 - \delta_1 B)$ and $\omega(B) = \omega_3 B^3$.

This is easier to handle than writing Y_t in terms of lagged values of X_t only, namely

$$Y_t = \omega_3 X_{t-3} + \omega_3 \delta_1 X_{t-4} + \omega_3 \delta_1^2 X_{t-5} + \ldots$$

3. Multivariate ARMA (VARMA) models.

$$\Phi_p(B)\boldsymbol{X}_t = \Theta_q(B)\boldsymbol{\varepsilon}_t$$

where Φ and Θ are matrix polynomials in B of order p, q

Here $\boldsymbol{X}_t$ and $\boldsymbol{\varepsilon}_t$ are *vectors* of the same length.

Difficult to fit even with only 2 or 3 variables. Much current research interest (e.g. Lütkepohl 1991). But interpretation of cross-correlation function not easy.

So approx. with VAR model?

Bayesian vector autoregression shrinks coefficients towards zero.

Or use external knowledge to get *sparse* matrices (i.e. lots of zeros).

Or look for *co-integration*. e.g. X_t, Y_t, are not stationary, but $(X_t - kY_t)$ *is* stationary. An equilibrium relationship. Link with ECM models.

OR generalize to VARX models which include *exogenous* variables. The latter affect the system but are not affected by it.

Example. A bivariate VAR(1) model.

Observe $\boldsymbol{X}_t^T = (X_{1t}, X_{2t})$. Suppose

$$\left.\begin{aligned} X_{1t} &= \phi_{11} X_{1,t-1} + \phi_{12} X_{2,t-1} + \varepsilon_{1t} \\ X_{2t} &= \phi_{21} X_{1,t-1} + \phi_{22} X_{2,t-1} + \varepsilon_{2t} \end{aligned}\right\}$$

where $\{\phi_{ij}\}$ are constants and ε_{1t}, ε_{2t} are uncorrelated white noise. Can rewrite as

$$\boldsymbol{X}_t = \Phi_1 \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\varepsilon}_t^T = (\varepsilon_{1t}, \varepsilon_{2t})$, $\Phi_1 = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix}$, or as $(I - \Phi_1 B)\boldsymbol{X}_t = \boldsymbol{\varepsilon}_t$

where $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then $\Phi(B) = (I - \Phi_1 B)$ is a matrix polynomial of order one in B.

Special case: If $\phi_{12} = 0 \;\rightarrow$ no feedback from X_{2t} to $X_{1t} \;\rightarrow$ transfer function model for X_{2t} in terms of lagged values of X_{1t}. Φ_1 is a lower diagonal matrix. More generally a TFM is a special case of VARMA model when $\Phi_p(B)$ and $\Theta_q(B)$ are lower triangular.

4. Econometric Models.

A set of simultaneous equations. For 'closed-loop' data.

e.g. wages depend on prices *and* prices depend on wages

Build using economic theory and data – Want more cooperation between statisticians and econometricians.

Help to understand economy and evaluate alternative economic proposals, but forecasting performance is mixed, even after incorporating judgement to forecast exogenous variables.

5. Other multivariate methods.

Multivariate versions of *structural modelling* (Harvey, 1989, Chapters 7,8). An example is

$$\boldsymbol{X}_t = \boldsymbol{\mu}_t + \boldsymbol{n}_t$$

where

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{w}_t$$

where $\boldsymbol{n}_t, \boldsymbol{w}_t$ are vectors of innovations with appropriate correlation properties.

$\rightarrow$ *Seemingly Unrelated Time Series Equations* (SUTSE).

$\rightarrow$ *Common Trends Model* when linear combinations of some components of $\boldsymbol{\mu}_t$ are stationary

$\rightarrow$ link with *co-integration*.

For multivariate versions of *Bayesian forecasting*, see West and Harrison (1997, Chapters 9,16).

6. Intervention Analysis.

can be regarded as an extension of both univariate and multivariate forecasting methods. Let

$X_t =$ some model $+\delta_t$

where δ_t can take various forms of which the impulse function is perhaps most common, namely $\delta_t = \delta$ if $t = t_0$, but 0 otherwise, if $t \neq t_0$.

Usually t_0, the time at which the intervention takes place, is known from context. Then estimate δ from data. Box *et al.*, 1994, Chapter 11.

5.2/5.3 Some General Questions and Comments on Forecasting

1. Method or model? – An important distinction.

A *model* is a mathematical representation of reality.

A *forecasting method* is a rule or formula for computing a forecast. *It may, or may not, depend on a model.*

Ex. Exponential smoothing (ES) is a method. It is optimal for an ARIMA(0,1,1) model. It may be nearly optimal for various other models, but would not be sensible to use for some other underlying models.

2. Point or interval forecast?

Forecasts usually given as *point forecasts* – perhaps with too many significant figures implying spurious accuracy. This may be OK if forecast user cannot cope with uncertainty. But gives no idea of accuracy. So usually prefer interval forecasts

– to assess future uncertainty

– enable different strategies to be planned for a range of possible outcomes

– explore different scenarios based on different assumptions

Prediction Intervals. An interval forecast associated with a specified probability. (Better description than confidence interval). Review in Chatfield (1993, JBES) and TFS, Chapter 7.

Let $e_t(k) = X_{t+k} - \hat{X}(t,k)$. Find $Var[e_t(k)]$. Then assuming normality and unbiasedness, 95% Prediction Interval (P.I.) for X_{t+k} is given by

$$\hat{X}(t,k) \pm 1.96\sqrt{(Var[e_t(k)])}$$

Theoretical formulae for $Var[e_t(k)]$ can be found for regression, ARIMA and structural *models* (comes from Kalman filter), as well as for some *methods*, such as exponential smoothing and Holt-Winters, by assuming that the method is optimal.

OR calculate empirically from past fitted 'forecast' errors (e.g. Gardner, 1988, Man. Sci.).

OR calculate by simulation, bootstrapping or Monte Carlo methods (computationally intensive)

OR assess subjectively

P.I.s often not given because:

1) Literature is unhelpful

2) No general method of computing P.I.s

3) Impossible for some complicated models, especially if non-linearities present.

4) Resampling methods not widely understood

5) 'Approximate' methods may be invalid

6) Software packages may not produce them

P.I.s assume future is like past. But *generally too narrow*! Because:

a) May fit wrong model; (data mining $\rightarrow$ overfitting)

b) Model may change in future;

c) Outliers may be present or errors non-normal;

d) Have to estimate parameters; (though can often incorporate correction for parameter uncertainty, typically of order $1/N$.)

Out-of-sample forecast accuracy often much worse than within-sample fit, but difficult to assess effect of *Model uncertainty* – see Chatfield (1996b).

$\rightarrow$ All comparisons of different forecasting methods and models should be made on the basis of out-of-sample forecast accuracy, *not* within-sample fit.

3. Must **monitor** forecasts.

Forecast Monitoring. Check forecasts, and, if things go wrong, revise model and forecasts. Usually look at 1-step-ahead forecast errors: $e_t = x_t - \hat{x}(t-1, 1)$, and plot them against time. If, for example, find a series of positive errors, then we consistently under-forecast. Must take appropriate action.

Is mean error zero? (unbiased forecasts).

Are errors autocorrelated? (then method not optimal).

Are errors normal? (as assumed for symmetric P.I.s)

Can use CUSUM TRACKING SIGNALS – Plot cumulative sum (or cusum) of forecast errors against time. e.g. Gardner (1983, JoF).

4. Single-period forecast or **cumulative** forecast?

May want to plan production for a period. Compute by adding single-period forecasts?

5. Must be prepared to **improvise** in any particular situation. Take account of *context*.

5.4 Comparison of Forecasting Methods – TSF, Chapter 6.

Which method is best? Answer: It depends!!

What is meant by 'best'? Univariate or Multivariate? Automatic or non-automatic?

A simple automatic univariate procedure is useful

(i) as a norm

(ii) if lots of items to forecast

(iii) as a preliminary forecast to be adjusted subjectively

(iv) if analyst's skill is limited

But econometricians do not like them as they do not explain what is going on.

5.4.1 Forecasting Competitions

Compare accuracy of different methods on different series. Usually compare automatic univariate and Box-Jenkins. Newbold & Granger (74, JRSS A) suggest BJ *better* than HW.

BUT different results obtained by Makridakis & Hibon (79, JRSS A) and Makr. et al (82,JoF) in the *M-competition* – 1000 series. 24 methods. International panel. Found HW $\simeq$ BJ and seven other methods.

Which conclusions are 'true'?

M-competition results replicated by Lusk & Neves (84, JoF), Koehler (85, IJoF).

Different answers may be due to i) skill of analyst; (ii) selected sample of series; iii) different interpretation of methods.

AND comparisons must be genuinely *out-of-sample* (or *ex-ante*).

But silly to forecast automatically? Need interaction with client?

Competitions tell something but not everything. (Armstrong and Lusk, 83, JoF; Chatfield, 1988). Need more *case studies* for assessing multivariate methods and non-automatic univariate methods.

M2 competition – IJoF 1993 – 29 series from short ($n = 33$) to long ($n = 163$) plus additional information. I took part but found it unsatisfactory. No direct contact with 'client'.

M3 competition – IJoF 2000 – 3003 series! 24 methods including neural nets. Results not very clear? Too large?

Choice of an automatic method

Use if there are a large number of series or analyst's skill is limited. Forecasting competitions suggest several methods about equally accurate when applied automatically to large numbers of series. Some are rather complicated or unclear, so use a method that is simple, easily interpreted, and for which programs are available. I recommend *Holt-Winters* but several good alternatives.

5.4.2 Choosing a non-automatic method.

If no. of series is small and/or want very accurate forecasts and/or want to use external information, then use Judgemental or MV or BJ or non-automatic version of simple method.

Multivariate forecasts (MV) are sometimes worth the extra effort (and MV models usually give a better *fit*) but are not necessarily better than univariate forecasts either in theory or in practice, because:

(i) may have to forecast exogenous variables;

(ii) Economic data are generally *observational* rather than *designed* data. Often unsuitable for fitting MV models because explanatory variables are correlated and *feedback* between 'inputs' and 'outputs'.

(iii) Simple is often best. Why? Perhaps univariate methods are more robust to model misspecification or to changes in model.

VARMA models are still quite hard to fit.

Transfer function models worth considering for open-loop data.

Multiple regression – simplest multivariate procedure – beware of correlated explanatory variables.

Econometric model – to understand economy rather than forecast.

For *non-automatic univariate* method, gains to be had by being selective. Consider B-J if

i) analyst is competent

ii) extra cost can be justified

iii) variation not dominated by trend and seasonality

iv) At least 40 observations available

OR consider *structural modelling.*

Or subjective adjustment of a method such as Holt-Winters. Chatfield (78, App. Stats) demonstrates possible improvements. *Non-automatic Holt-Winters* incorporates careful choice of:

1) seasonal or non-seasonal model

2) additive or multiplicative seasonality

3) starting values for trend, seasonals

4) smoothing parameters

5) inclusion, adjustment or exclusion of outliers.

5.4.3 A general non-automatic strategy along the following lines usually works well:

a) Get background info. and define *objectives.* Point or interval forecast?

Single-period or cumulative forecasts?

b) *Plot series.* MOST IMPORTANT STEP! Look for trend, seasonality, outliers, discontinuities, etc.

c) 'Clean' the data if necessary. Adjust outliers? – Use external knowledge.

Adjust for calendar variation?

Any other adjustments? Consider possibility of transforming data.

d) Decide if *seasonal variation* is i) non-existent ii) multiplicative iii) additive iv) mixed.

e) Decide if *trend* is i) non-existent ii) global linear iii) local linear iv) non-linear.

f) Choose appropriate method. 4 types of series -

1. Discontinuities present $\rightarrow$ Univariate unwise.
2. Trend-and-seasonal $\rightarrow$ Holt-Winters
3. Short-term correlation $\rightarrow$ Box-Jenkins

4. Exponential growth $\rightarrow$ logs? or include quad. growth term

g) *Check model adequacy.* In particular study 1-step-ahead forecasts over fit period to see if improvements can be made.

h) *Compute forecasts.* Decide if they need subjective adjustment perhaps because of anticipated changes in external features.

For an alternative general approach, called *rule-based forecasting*, see Collopy and Armstrong (1992) and Adya *et al.* (2001, IJoF).

5.4.4 Summary.

a) Many different types of forecasting problem requiring different treatment.

b) No single 'best' method. Accuracy is only one criterion. Also consider cost, ease of use, etc. Automatic or non-automatic approach? *Automatic*: Holt-Winters?

Non-automatic: Judgement, or MV or non-automatic univariate?

c) Fitting 'best' model to historical data may not give best post-sample forecasts. e.g. complex methods often no better for forecasting than simple methods.

d) Combination of forecasts often better than forecasts from individual method - but no model.

e) Prediction intervals calculated assuming fitted model is 'true' are generally too narrow.

f) Be prepared to improvise.

g) Don't forget *eyeball test.* Plot forecasts. Check they look sensible. If not, more work to do!

11. NON-LINEAR MODELS

Most of theory and practice is about **linear** methods and models (e.g. ARIMA models, ES). (*What is a linear model*??!)

Many classes of *non-linear* time-series model (Tong, 1990). Increasing interest. More difficult to fit than linear models.

Threshold AR models. Simple example, with zero as threshold:

$$X_t = \begin{cases} \alpha^{(1)} X_{t-1} + Z_t & \text{if } X_{t-1} \geq 0 \\ \alpha^{(2)} X_{t-1} + Z_t & \text{if } X_{t-1} < 0 \end{cases}$$

Good example using economic data in Tiao and Tsay (1994, JoF). Little improvement in forecasts, but *more insight from modelling process.*

Bilinear models. Simple example (last term is the bilinear term):

$$X_t = \alpha X_{t-1} + Z_t + \gamma Z_{t-1} X_{t-1}$$

Why do we need NL models?

Some time series exhibit *asymmetric behaviour* in the time plot (perhaps after careful choice of scales!).

Series may go faster up than down; There may be more spikes up than down; economic series behave differently going into, rather than out of, a recession. May observe *Limit Cycles* when plotting X_t against X_{t-k}; May observe *changing variance* through time.

Examples: Sunspots; Lynx trappings; Riverflow data; Chaotic behaviour.

But difficult to tell difference between non-linearity, non-normality and changing variance.

11.3 Models for Changing Variance

Of special interest when studying financial time series, where *changes in volatility* important. See Harvey (1993, Chapter 8); Shephard (1996).

GARCH models. **G**eneralized **A**uto**R**egressive **C**onditionally **H**eteroscedastic. e.g. GARCH(1,1)

$$X_t = \sigma_t \varepsilon_t$$

where $\varepsilon_t \sim I.I.D.(0,1)$ and

$$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma X_{t-1}^2$$

$\beta = 0 \rightarrow$ ARCH

Many other models. e.g. regression models with ARCH disturbances; stochastic variance models where σ_t follows stochastic process.

Again modelling aspect more important than forecasting? But GARCH models used to forecast prices of options (derivatives) where estimation of variance is important (assessment of risk). GARCH models do not affect point forecasts and are hard to compare.

11.4 Neural Nets. What is a neural net (NN)? See ATS, Section 11.4. Inputs (predictor or lagged variables), outputs (forecasts), plus one or more hidden layer of 'nodes'. At each node, calculate linear sum of inputs and apply an 'activation function' (e.g. logistic)

Many questions in NN modelling. What architecture? How many hidden layers? How many nodes? What activation function(s)? How should NN be fitted? There may be many parameters (weights) to be estimated. Use iteration (e.g. back-propagation) to choose w_{ij} to min. $\sum(x_t - \hat{x}_t)^2$ over 1st part of series – the *training set.* Need specialist package. Do not 'overtrain' or may get spuriously good fit but poor forecasts. Prefer BIC to AIC when comparing models, to penalize the fitting of extra (spurious) parameters.

OR fit using *regularization.* Minimize $(E + \nu \Omega)$ where E is some measure of error, ν is a smoothing parameter, and Ω is a measure of 'smoothness'. – Usual bias versus variance trade-off. (c.f. AIC $= -2\log L + 2p$)

Santa Fe Competition, 1991. Six very long series (e.g. 34,000 obs.!!) Five are clearly non-linear. Only one economic series. Organisers kept holdout samples for three of them. See Weigend and Gershenfeld (1994), especially interesting introductory chapter. Little contextual information for participants. So not me!

Participants chose their own method (e.g. NN; state-space)

Some Findings

1. 'Failure to use common sense was readily apparent in many of the entries in the competition'.

2. Predictions 'based solely on visually examining and extrapolating the training data did much worse than the best techniques, but also much better than the worst'.

3. 'There was a general failure of simplistic 'black-box' approaches – In all successful entries, exploratory data analysis preceded the algorithm application'.

4. Some non-linear results much better than linear, but there are 'unprecedented opportunities for the analysis to go astray'. In particular 'the best, as well as many of the worst, forecasts of Data Set A were obtained with neural networks'.

5. The prediction methods that work well for data sets A and D, **fail** for the exchange rates time series which is close to pure randomness. Here there is a 'crucial difference between training set and test set performance' and 'out-of-sample predictions are on average worse than chance'. (So we are back to our old friend the *Random Walk*!!).

6. One successful set of predictions for Data set D used 100 hours of computer time!

Other Empirical Evidence. Many comparisons made – see Zhang et al (1998, IJoF) for review. Note that some comparisons are not fair or do not make genuine out-of-sample forecasts. Some selected examples:

Callen et al (1996, IJoF). 296 series of short ($n = 89$) accounting series. Linear methods better than NNs "even when data are financial, seasonal and non-linear".

Hal White: For many economic series 'No change' better than NNs and experts!!

Faraway and Chatfield (1998) show NNs no better than Box-Jenkins for airline data. BIC better than AIC. Plenty of scope for going badly wrong with NN modelling, so don't apply in black-box mode.

Simulations (e.g. Stern, 1996, Technometrics) show linear methods do better than NNs for *linear* data. (Of course!?).

Summary of Status of NNs

Empirical evidence unclear, especially given *publication bias* towards results in favour of a new method (e.g. Company X suppressed results showing NNs poor).

More empirical evidence is needed to establish when NNs are worth using. Good for long series

with non-linear characteristics? Poor for linear data. Need several hundred (or thousand?) observations to feel safe fitting NNs.

A black box approach. No interpretable model to help understanding. And easy to go badly wrong. The "marketing hype that NNs can be used with no experience and automatically learn whatever is required ... is nonsense" (W. Sarle).

11.5 Chaos Theory.

Best known example, leading to chaotic behaviour, is the *logistic* or quadratic map:

$x_t = kx_{t-1}(1 - x_{t-1})$ for $t = 1, 2, 3, \ldots$ and $0 < x_0 < 1$.

For small values of k get an obviously deterministic series. But for k near 4, get 'chaos'. When $k = 4$, series has flat spectrum and looks 'random' even though it is deterministic. Usual linear tests, based on second-order properties, indicate randomness. Sensitivity to initial conditions measured by *Lyapunov exponent*. *Dimension* hard to define. Random processes have infinite dimension.

Can chaotic series be forecasted? Certainly not for *long* lead times. For short lead times, might be possible for low-dimensional chaos if we knew model (Berliner, 1991, JASA; Tong, 1990). But generally don't. See Granger's (1992, IJoF) cautionary remarks. Stock market not deterministic anyway.

COMPUTER SOFTWARE FOR FORECASTING

Difficult to make general remarks. Scene changing rapidly. Sensible use vital.

GARBAGE IN $\rightarrow$ GARBAGE OUT

is still true!!

Desirable features of good software are:

i) Flexible data entry and editing facilities.

ii) Good facilities for exploring data with summary statistics and graphs.

iii) Technically sound and computationally efficient.

iv) Clear output.

v) Easy-to-learn and easy-to-use with good, clear documentation.

Rycroft (1999, IJoF) reviews 51 packages!

FORECASTING EXAMPLES

Difficult to give flavour of real-life forecasting problems. Textbook examples often artificial and designed to illustrate a particular technique. Would like to illustrate:

1) Importance of getting background information. Ask questions. Clarify objectives. Find out how data were collected. Use common sense.

2) Importance of plotting data clearly

3) Try more than one method.

4) Try to *avoid trouble*. Examples in literature usually avoid mentioning *problems, mistakes, blind alleys*, etc. – a pity! *"Many true stories are tales of woe. But these are where the lessons are to be learnt."* – Tony Greenfield. Being more positive, we learn from successes *and* failures.

5) Clear presentation of tables (as well as graphs) is important at all stages of a study, whether looking at data, or collating and presenting results.

(i) Give clear title;

(ii) Round numbers as appropriate;

(iii) Give row and column averages or totals where appropriate;

(iv) Transpose the whole table?;

(v) Re-order rows and/or columns?;

(vi) Clear spacing and layout;

(6) Must be able to *improvise* and cope with *non-standard* data. Use *context* and *common-sense*.

Example. Forecasting tyre sales. Legislation on minimum tread led to big increase in sales before law came into effect.

How do we predict in this one-off situation, both before law is passed *and* afterwards.

– Simple (univariate) methods inappropriate

– Must take tyre life distribution into account to predict replacements (c.f. forecasting population).

Example. Recent consultancy. Short series with one peculiar observation (outlier). Main problem was what to do about this observation, NOT which forecasting method to choose.

More examples will be given in the course and some can be found in Section 5.5 and Appendix D of ATS.

SOME REFERENCES

Andrews, R.L. (1994) Forecasting performance of structural time series models. *J. Business and Economic Stats.*, **12**, 129-133.

Armstrong, J.S (1985) *Long-Range Forecasting*, 2nd edn. Wiley.

Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman and Hall.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994) *Time Series Analysis, Forecasting and Control*, 3rd edn. Prentice-Hall.

Brockwell, P.J. and Davis, R.A. (1991) *Time Series: Theory and Methods*, 2nd edn. Springer.

Chatfield, C. (1988) What is the 'best' method of forecasting? *J. Applied Stats.*, **15**, 19-38.

Chatfield, C. (1993) Calculating interval forecasts. *J. Business and Econ. Stats.*, **11**, 121-144.

Chatfield, C. (1996) *The Analysis of Time Series*, 5th edn. Chapman & Hall.

Chatfield, C. (1996b) Model uncertainty and forecast accuracy. *J. Forecasting*, **15**, 495-508.

Chatfield, C. (1997) Forecasting in the 90's. *The Statistician*, **46**, 461-473.

Chatfield, C. (2001) *Time-Series Forecasting*. Chapman & Hall/CRC Press.

Chatfield, C., Koehler, A.B., Ord, J.K. and Snyder, R.D. (2001) Models for exponential smoothing: A review of recent developments. *The Statistician*, **50**, (to appear).

Collopy, F. and Armstrong, J.S. (1992) Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Sci.*, **38**, 1394-1414.

Diebold, F.X. (2001) *Elements of Forecasting*, 2nd edn. South-Western College Publishing.

Faraway, J. and Chatfield, C. (1998) Time series forecasting with neural networks: A comparative study using the airline data. *Applied Statistics*, **47**, 231-250.

Fuller, W.A. (1996) *Introduction to Statistical Time Series*, 2nd edn. Wiley.

Gardner, E.S. (1985) Exponential smoothing. *J. Forecasting*, **4**, 1-28.

Granger, C.W.J. and Newbold, P. (1986) *Forecasting Economic Time Series*, 2nd edn. Academic Press.

Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. C.U.P.

Harvey, A.C. (1993) *Time Series Models*, 2nd edn. Harvester Wheatsheaf.

Kendall, M.G. and Ord, J.K. (1990) *Time Series*, 3rd edn. Arnold.

Lütkepohl, H. (1993) *Introduction to Multiple Time Series Analysis*, 2nd edn. Springer-Verlag.

Montgomery, D.C., Johnson, L.A. and Gardiner, J.S. (1990) *Forecasting and Time Series Analysis*, 2nd edn. McGraw-Hill.

Pole, A., West, M. and Harrison, J. (1994) *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall.

Shephard, N. (1996) Statistical aspects of ARCH and stochastic volatility. In *Time Series Models* (eds. D.R. Cox *et al.*), pp 1-67. Chapman and Hall.

Tiao, G.C. and Tsay, R.S. (1994) Some advances in non-linear and adaptive modelling in time-series. *J. of Forecasting*, **13**, 109-131.

Tong, H. (1990) *Non-Linear Time Series*. Oxford Univ. Press.

Wei, W.W.S. (1990) *Time Series Analysis*. Addison-Wesley.

Weigend, A.S. and Gershenfeld, N.A. (eds.) (1994) *Time Series Prediction*. Proc. Vol. XV, Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley.

West, M and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. Springer.