Investigate times series data

Residuals

A **residual** measures the difference between an individual seasonal effect and the average seasonal effect for a particular time period.

Residual = individual seasonal effect – average seasonal effect

Residuals highlight when the largest variations from average occur in a time period.

- If a residual is positive, the raw data value is higher than average.
- If a residual is negative, the raw data value is lower than average.

The greater the absolute value of a residual (i.e. the further it is from zero), the greater the variation from a typical time period.

Example

Residuals are calculated for the a table of data about tourist industry employment (used previously). For example, for March 2011 residual = 1.125 – 1.6667 = –0.5417 (which is –542 people).

Quarter	People employed (000)	Centred moving mean (CMM)	Individual seasonal effects	Average seasonal effect	Residuals
2008 Mar	10				
Jun	15				
Sep	12	14.25	–2.250	–1.375	–0.875
Dec	16	15.5	0.5	0.4583	0.0417
2009 Mar	18	17.125	0.875	1.6667	–0.7917
Jun	17	19.25	–2.25	–0.625	–1.625
Sep	23	20.875	2.125	–1.375	3.5
Dec	22	22.125	–0.125	0.4583	–0.5833
2010 Mar	25	22	3	1.6667	1.3333
Jun	20	22	–2	–0.625	–1.375
Sep	19	23	–4	–1.375	–2.625
Dec	26	25	1	0.4583	0.5417
2011 Mar	29	27.875	1.125	1.6667	–0.5417
Jun	32	29.625	2.375	–0.625	3
Sep	30				
Dec	29				

The residuals with the greatest absolute values correspond to periods when there is the greatest variation from the average.

- Sep 09 (residual = 3.5) and Jun 11 (residual = 3) have higher numbers of people than would be expected for the September and June quarters.
- Sep 10 (residual = –2.625) has a lower number of people than would be expected for the September quarter.

The quarter with residual of lowest absolute value is Dec 08 (residual = 0.0417). This indicates that Dec 08 had a typical number of people for a December quarter.

Investigate times series data

Exercise A: Residuals with Excel

Ans. p. 27

1. The table below shows time series data for wine produced per quarter by a boutique New Zealand winery, along with individual and average seasonal effects.

Quarter	Raw data (litres of wine produced)	Individual seasonal effects	Average seasonal effects	Residuals
Mar 2000	154.3			
Jun 2000	147.5			
Sep 2000	161.4	1.325	8.958	
Dec 2000	166.8	3.5125	–0.146	
Mar 2001	174.9	8.75	–4.058	
Jun 2001	152.6	–18.5	–2.046	
Sep 2001	179.2	9.45	8.958	
Dec 2001	188.6	19.7625	–0.146	
Mar 2002	142.3	–30.775	–4.058	
Jun 2002	177.9	7.9	–2.046	
Sep 2002	187.8	16.1	8.958	
Dec 2002	155.4	–23.7125	–0.146	
Mar 2003	189.1	9.85	–4.058	
Jun 2003	190.4	4.4625	–2.046	
Sep 2003	176.4			
Dec 2003	220.3			

a. Calculate the residuals for each quarter.

b. Comment on and interpret the residuals.

2. The table shows monthly Consumer Price Index (CPI) for fruit and vegetables in New Zealand (a measure of the changing cost of purchasing a set basket of fruit and vegetables). The table also shows centred moving means of order 12 (CMM), individual seasonal effects (ISE) and average seasonal effects (ASE), to 2 d.p.

This data set is available on the ESA website RESOURCES

a. Calculate the residuals for each month.

Investigate times series data

Year/month	CPI	CMM	ISE	ASE	Residuals
2008M01	995				
2008M02	980				
2008M03	1 010				
2008M04	996				
2008M05	1 034				
2008M06	1 088				
2008M07	1 127	1 091.46	–35.54	–114.21	
2008M08	1 235	1 101.96	–133.04	–95.46	
2008M09	1 208	1 111.38	–96.63	–46.75	
2008M10	1 136	1 119.96	–16.04	–17.69	
2008M11	1 133	1 125.42	–7.58	17.42	
2008M12	1 089	1 132.5	43.50	45.78	
2009M01	1 128	1 143.96	15.96	–5.29	
2009M02	1 099	1 149.13	50.13	29.36	
2009M03	1 117	1 144.88	27.88	37.50	
2009M04	1 095	1 138.33	43.33	63.43	
2009M05	1 066	1 131.92	65.92	73.76	
2009M06	1 226	1 127.38	–98.63	–58.17	
2009M07	1 264	1 125.96	–138.04	–114.21	
2009M08	1 222	1 124.63	–97.38	–95.46	
2009M09	1 119	1 121.29	2.29	–46.75	
2009M10	1 068	1 116.54	48.54	–17.69	
2009M11	1 047	1 112.25	65.25	17.42	
2009M12	1 066	1 105.54	39.54	45.78	
2010M01	1 117	1 098.58	–18.42	–5.29	
2010M02	1 078	1 093.96	15.96	29.36	
2010M03	1 058	1 094.75	36.75	37.50	
2010M04	1 040	1 105.21	65.21	63.43	
2010M05	1 018	1 117.71	99.71	73.76	
2010M06	1 113	1 125.38	12.38	–58.17	
2010M07	1 210	1 131.63	–78.38	–114.21	
2010M08	1 165	1 139.96	–25.04	–95.46	
2010M09	1 195	1 149.08	–45.92	–46.75	
2010M10	1 243	1 157.42	–85.58	–17.69	
2010M11	1 172	1 166.58	–5.42	17.42	
2010M12	1 125	1 179.29	54.29	45.78	
2011M01	1 208	1 194.58	–13.42	–5.29	
2011M02	1 187	1 209	22.00	29.36	
2011M03	1 168	1 215.88	47.88	37.50	
2011M04	1 130	1 211.75	81.75	63.43	
2011M05	1 148	1 203.67	55.67	73.76	
2011M06	1 288	1 199.75	–88.25	–58.17	
2011M07	1 402	1 197.13	–204.88	–114.21	
2011M08	1 319	1 192.63	–126.38	–95.46	
2011M09	1 206				
2011M10	1 133				
2011M11	1 088				
2011M12	1 115				
2012M01	1 155				
2012M02	1 132				

Source: Statistics New Zealand

Investigate times series data

b. Comment on the reliability of the average seasonal effects.

c. Comment on and interpret the residuals.

Investigate times series data

Forecasting

A **trend line** can be drawn to show the linear relationship between (centred) moving means as time passes. This can be done 'by eye' so that the line passes through the 'centre' of the set of moving means.

The trend line is used when making a **forecast** (a prediction of a future time series value). The line is extrapolated (extended past the data points) to give a projected trend value. The forecast is calculated by adding the appropriate average seasonal effect to this trend value.

Forecast value = trend value + average seasonal effect

Time series data can be analysed more exactly using a spreadsheet such as Excel (unemployment data set is available as a spreadsheet on the ESA website, with instructions) or by using **iNZight** RESOURCES

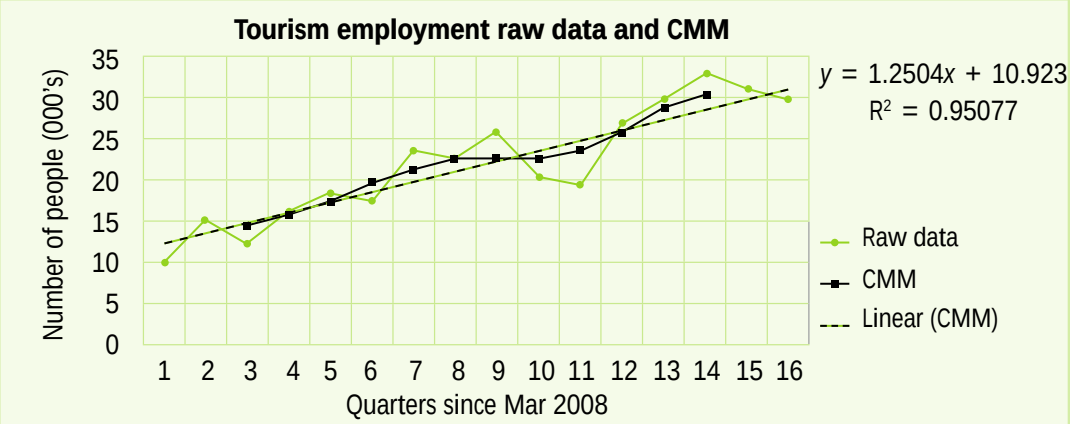
The **line of best fit** can be drawn automatically, and its equation supplied along with its R^2-value (**coefficient of determination**) which is a measure of how well the trend line fits the data. For a time series graph, R^2 indicates the proportion of the variation in the variable of interest (here the CMM value) that can be explained by the change in time. R^2 takes on values between 0 and 1. The closer R^2 is to 1, the more predictable the values of this variable will be.

Example

The graph below shows a plot of raw data and centred moving means for the previously used New Zealand tourism employment time series data. The quarters are numbered from 1 (Mar 2008) to 16 (Dec 2011). A forecast is to be made for employment in March 2012.

A trend line is fitted to the centred moving means (CMM), using Excel.

Its equation is: $y = 1.2504x + 10.923$, where x is quarter number and y is number of people (in 000s).



If the trend line is extended 'by eye' to quarter number 17 (the Mar 2012 quarter) then it can be seen that the corresponding trend value (y-value) is approximately 32. Adding on the March quarter average seasonal effect of 1.7 (1 dp) gives a forecast of 32 + 1.7 = 33.7 thousand people, i.e. 33 700 people.

Alternatively, the trend value for the Mar 2012 quarter can be calculated by substituing $x = 17$ into the equation of the Excel line of best fit for the (centred) moving means:

$y = 1.2504 \times 17 + 10.923 = 32.1798$

To make the forecast, add on the March quarter average seasonal effect of 1.6667 to the trend value.

forecast = 32.1798 + 1.6667 = 33.8465 thousand people (or 33 800 people (3 s.f.))

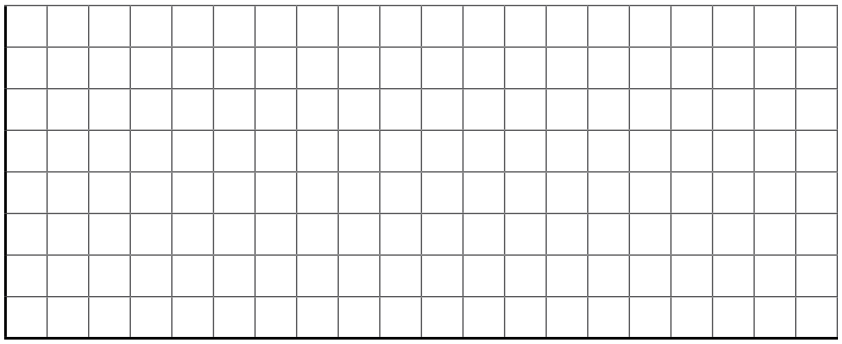
To the nearest thousand, the employment forecast for the March 2012 quarter is 34 000 people.

Exercise B: Trend-lines and forecasts using Excel

Ans. p. 27

1. Using the wine consumption data from Exercise D, question 1 and Exercise E, question 1 (litres of wine produced per quarter by a boutique New Zealand winery 2000–2003):

   a. plot a graph of the raw data vs quarter



   b. on the same graph plot the centred moving means

   c. draw (by 'eye') a line of best fit for the centred moving means

   d. extrapolate the line of best fit and use this to make a forecast for

      i. the March 2004 quarter _____

      ii. the June 2004 quarter _____

      iii. the September 2004 quarter _____

      iv. the December 2004 quarter _____

2. The time series data in the table below shows the carbon dioxide (CO_2) content of the atmosphere (in parts per million) measured by Komhyr et al. at the Mauna Loa Observatory on the Big Island of Hawaii from 2006 to 2011.
   This data set is available on the ESA website RESOURCES , and should be analysed electronically.

Period	Month/Year	CO_2 (ppm)	Moving mean	CMM	ISE
1	Jan-06	337.81			
2	Feb-06	338.26			
3	Mar-06	340.07			
4	Apr-06	340.87			
5	May-06	341.48			
6	Jun-06	341.30			
7	Jul-06	339.36			
8	Aug-06	337.84			
9	Sep-06	335.98			
10	Oct-06	336.07			
11	Nov-06	337.22			
12	Dec-06	338.33			
13	Jan-07	339.35			
14	Feb-07	340.47			

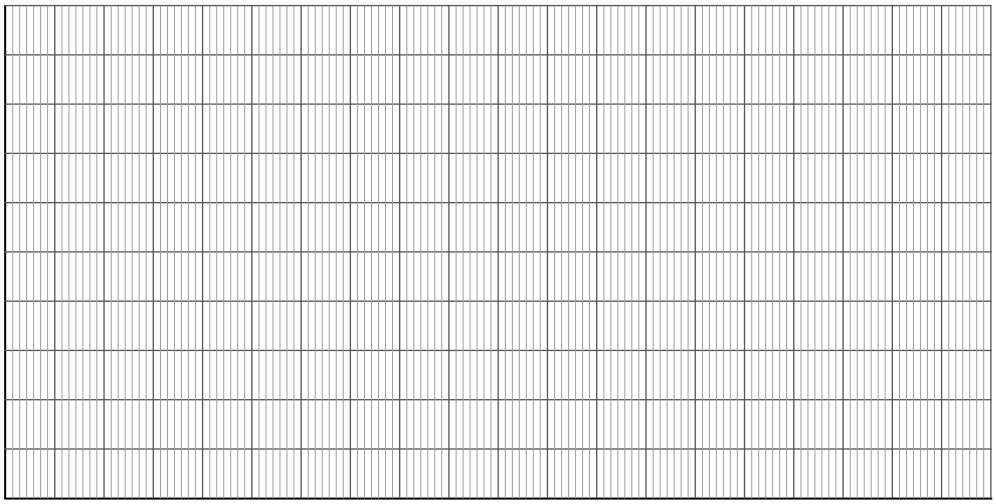
Investigate times series data

Period	Month/Year	CO_2 (ppm)	Moving mean	CMM	ISE
15	Mar-07	341.73			
16	Apr-07	342.47			
17	May-07	343.02			
18	Jun-07	342.46			
19	Jul-07	340.78			
20	Aug-07	338.61			
21	Sep-07	337.04			
22	Oct-07	337.13			
23	Nov-07	338.49			
24	Dec-07	339.85			
25	Jan-08	340.93			
26	Feb-08	341.67			
27	Mar-08	342.78			
28	Apr-08	343.71			
29	May-08	344.31			
30	Jun-08	343.41			
31	Jul-08	341.96			
32	Aug-08	339.79			
33	Sep-08	337.91			
34	Oct-08	338.10			
35	Nov-08	339.27			
36	Dec-08	340.67			
37	Jan-09	341.45			
38	Feb-09	342.69			
39	Mar-09	343.21			
40	Apr-09	345.17			
41	May-09	345.79			
42	Jun-09	345.40			
43	Jul-09	343.99			
44	Aug-09	342.00			
45	Sep-09	340.02			
46	Oct-09	340.16			
47	Nov-09	341.39			
48	Dec-09	343.01			
49	Jan-10	343.92			
50	Feb-10	344.64			
51	Mar-10	345.18			
52	Apr-10	347.10			
53	May-10	347.45			
54	Jun-10	346.80			
55	Jul-10	345.44			
56	Aug-10	343.24			
57	Sep-10	341.25			

Investigate times series data

Period	Month/Year	CO_2 (ppm)	Moving mean	CMM	ISE
58	Oct-10	341.49			
59	Nov-10	342.84			
60	Dec-10	344.37			
61	Jan-11	345.03			
62	Feb-11	345.89			
63	Mar-11	347.47			
64	Apr-11	348.02			
65	May-11	348.73			
66	Jun-11	348.11			
67	Jul-11	346.59			
68	Aug-11	344.59			
69	Sep-11	343.01			
70	Oct-11	342.88			
71	Nov-11	344.19			
72	Dec-11	345.64			

a. Smooth the raw data then draw a time series graph for the raw data and the centred moving means. Fit a trend line to the smoothed data and determine its equation.



Investigate times series data

b. Calculate estimates for the average seasonal effects. Comment on their reliability.

	Individual seasonal effects						Average seasonal effects
Year	2006	2007	2008	2009	2010	2011	
Jan							
Feb							
Mar							
Apr							
May							
Jun							
Jul							
Aug							
Sep							
Oct							
Nov							
Dec							

c. Forecast, as accurately as possible, the atmospheric CO_2 content for June 2012.

Seasonally adjusted data

Seasonally adjusted data are the 'underlying' values of the data, once seasonal effects have been removed.

Seasonally adjusted value = raw data value – average seasonal effect

Seasonally adjusted data values are often quoted in the media by reporters. For example, the numbers of people unemployed in New Zealand are usually higher in winter. Seasonally adjusted winter unemployment figures are calculated by removing the 'extra' number of unemployed people due to winter. This allows the underlying level of unemployment to be seen.

Seasonally adjusted data can be used to detect raw data values which are higher (or lower) than expected. This is done by comparing seasonally adjusted values with centred moving means (in which seasonal effects have been smoothed out by averaging).

Note that seasonally adjusted values are estimates, since they are calculated using an *estimate* of the average seasonal effect (an unknown quantity). Hence a statistician will have the same confidence in the reliability of seasonally adjusted values as he/she would have in the values of the average seasonal effects themselves.

Investigate times series data

Example

The following table is an extract from the time series data for the numbers of overseas visitors arriving per month in New Zealand, with centred moving means (CMM), average seasonal effects for each month (ASE), and seasonally adjusted values (SAV) rounded to the nearest whole number.

Period	Date Year Month (M)	Raw data number of visitors	CMM	ASE	SAV	Comparsion of SAV to CMM
1	2001M03	131 338				
2	2001M04	106 300				
3	2001M05	78 232				
4	2001M06	70 443				
5	2001M07	88 007				
6	2001M08	90 114				
7	2001M09	83 352	112 617	–37 071.6	120 424	higher
8	2001M10	94 188	114 021	–27 233.6	121 422	higher
9	2001M11	112 844	115 135	–3 136.33	115 980	higher
10	2001M12	150 571	116 139	44 308.61	106 262	lower
11	2002M01	167 950	117 105	63 866.84	104 083	lower
12	2002M02	168 080	117 958	60 469.57	107 610	lower
13	2002M03	151 318	118 663	36 031.25	115 287	lower
14	2002M04	120 003	119 483	2 097	117 906	lower
15	2002M05	91 278	120 631	–34 155.9	125 434	higher
16	2002M06	81 478	122 130	–44 927.4	126 405	higher
17	2002M07	100 170	123 729	–28 926.7	129 097	higher
18	2002M08	98 417	125 308	–32 870.1	131 287	higher
19	2002M09	91 976	126 581	–37 071.6	129 048	higher
20	2002M10	105 233	127 382	–27 233.6	132 467	higher
21	2002M11	129 342	127 725	–3 136.33	132 478	higher
22	2002M12	170 070	127 799	44 308.61	125 761	lower

From the table it can be seen that, having allowed for seasonal effects, monthly values are

- higher than expected (seasonally adjusted values higher than centred moving mean values) in 2001 months 9, 10 and 11 and in 2002, months 5–11
- lower than expected (seasonally adjusted values lower than centred moving mean values) in 2001 month 12, 2002 months 1–4 and 2002 month 12).

Note: Comparing *raw data,* 2002 Month 1 (with 167 950 visitors) had higher visitor numbers than 2002 Month 5 (with 91 278 visitors), making 2002 Month 1 look more successful than 2002 Month 5. However, 2002 Month 1's number was *lower* than expected, while 2002 Month 5's number was *higher* than expected. Once seasonal effects have been allowed for, 2002 Month 5 (with 125 434 visitors) had higher visitor numbers than 2002 Month 1 (with 104 083 visitors).

Displaying raw data, smoothed data (CMM) and seasonally adjusted data together on a graph allows comparisons to be made visually, so that underlying trends can be seen.

Investigate times series data

Example

The graph below used an extended data set for visitor numbers to New Zealand.

- For 2002 Month 1 (period 11) the raw data value is *above* the CMM value but the seasonally adjusted value is *below* the CMM value. The peak of raw data values around 2002 Month 1 (period 11) is below the height of subsequent peaks, which also shows that, once seasonal effects are allowed for, there were fewer visitors than expected for 2002 Month 1.
- For 2002 Month 5 (period 15) the raw data value is below the CMM value, but the seasonally adjusted data value is slightly above the CMM value. So 2002 Month 5 actually did better than expected.

The graph also shows that the regular seasonal lows, once seasonally adjusted, have values very close to the CMM – and hence vistor numbers are consistent with the numbers that would be expected at those times of the year.

Data source: Statistics New Zealand

Exercise C: Seasonally adjusted data using Excel

Ans. p. 28

1. The table below shows wine production data, along with centred moving means (CMM) and average seasonal effects (ASE). Calculate and interpret the seasonally adjusted values for this time series.

Quarter	Raw data (litres of wine produced)	Centred moving mean (CMM)	Average seasonal effects (ASE)	Seasonally adjusted values (SAV)	SAV compared with CMM
Mar-00	154.3				
Jun-00	147.5				
Sep-00	161.4	160.075	8.958		
Dec-00	166.8	163.2875	–0.146		
Mar-01	174.9	166.15	–4.058		
Jun-01	152.6	171.1	–2.046		
Sep-01	179.2	169.75	8.958		
Dec-01	188.6	168.8375	–0.146		
Mar-02	142.3	173.075	–4.058		
Jun-02	177.9	170	–2.046		
Sep-02	187.8	171.7	8.958		
Dec-02	155.4	179.1125	–0.146		
Mar-03	189.1	179.25	–4.058		
Jun-03	190.4	185.9375	–2.046		
Sep-03	176.4				
Dec-03	220.3				

Investigate times series data

2. The following table contains the monthly Consumer Price Index (CPI) for fruit and vegetables in New Zealand (this measures the changing cost of purchasing a set basket of fruit and vegetables). Calculate and interpret the seasonally adjusted values for this data.

This data set is available on the ESA website RESOURCES

Food price index Level 2 subgroups for New Zealand (monthly)					
Month	Fruit and vegetables index raw data	CMM	Average seasonal effects (ASE)	Seasonally adjusted values (SAV)	SAV compared with CMM
2008M01	995				
2008M02	980				
2008M03	1 010				
2008M04	996				
2008M05	1 034				
2008M06	1 088				
2008M07	1 127	1 091.46	–114.21		
2008M08	1 235	1 101.96	–95.458		
2008M09	1 208	1 111.38	–46.75		
2008M10	1 136	1 119.96	–17.694		
2008M11	1 133	1 125.42	17.4167		
2008M12	1 089	1 132.5	45.7778		
2009M01	1 128	1 143.96	–5.2917		
2009M02	1 099	1 149.13	29.3611		
2009M03	1 117	1 144.88	37.5		
2009M04	1 095	1 138.33	63.4306		
2009M05	1 066	1 131.92	73.7639		
2009M06	1 226	1 127.38	–58.167		
2009M07	1 264	1 125.96	–114.21		
2009M08	1 222	1 124.63	–95.458		
2009M09	1 119	1 121.29	–46.75		
2009M10	1 068	1 116.54	–17.694		
2009M11	1 047	1 112.25	17.4167		
2009M12	1 066	1 105.54	45.7778		
2010M01	1 117	1 098.58	–5.2917		
2010M02	1 078	1 093.96	29.3611		
2010M03	1 058	1 094.75	37.5		
2010M04	1 040	1 105.21	63.4306		
2010M05	1 018	1 117.71	73.7639		
2010M06	1 113	1 125.38	–58.167		
2010M07	1 210	1 131.63	–114.21		
2010M08	1 165	1 139.96	–95.458		
2010M09	1 195	1 149.08	–46.75		

Investigate times series data

2010M10	1 243	1 157.42	–17.694		
2010M11	1 172	1 166.58	17.4167		
2010M12	1 125	1 179.29	45.7778		
2011M01	1 208	1 194.58	–5.2917		
2011M02	1 187	1 209	29.3611		
2011M03	1 168	1 215.88	37.5		
2011M04	1 130	1 211.75	63.4306		
2011M05	1 148	1 203.67	73.7639		
2011M06	1 288	1 199.75	–58.167		
2011M07	1 402	1 197.13	–114.21		
2011M08	1 319	1 192.63	–95.458		
2011M09	1 206				
2011M10	1 133				
2011M11	1 088				
2011M12	1 115				
2012M01	1 155				
2012M02	1 132				

Interpreting the gradient of a trend line

When a line is fitted to the smoothed data (moving means), the equation of this line is of the form $y = mx + c$ where m is the gradient of the line and c is the y-intercept.

The **gradient of the trend line** can be used to interpret the time series in context:

- a positive gradient shows that the trend is an increasing one
- a negative gradient shows that the trend is a decreasing one
- the size (absolute value) of the gradient gives the rate at which the trend line is increasing or decreasing.

It is important to use units correctly, in context, e.g. if the x-axis is 'seasons' and the y-axis is 'tonnes of fish', then the gradient is given in 'tonnes of fish per season'.

Example

The trend line for the smoothed data obtained previously for the numbers of people (in thousands) employed per quarter in the tourism industry in New Zealand had equation:

$y = 1.2504x + 10.923$

The gradient of this line is 1.2504 [comparing $y = 1.2504x + 10.923$ with $y = mx + c$]

This indicates that the overall trend is an increase in the number of people employed in the tourism industry in New Zealand of approximately 1 250 people per quarter.

Ans. p. 29

Exercise D: Interpreting the gradient of an Excel trend line

1. Interpret the gradient for each of the following trend line equations:

a. $y = 10.2x + 14.5$ where x is quarters and y is number of seals

b. $y = -52x + 1\,500$ where x is months and y is number of house sales

c. $y = 1.2\text{E}{-3}x + 12.5$ where x is days and y is the number of tonnes of cherries picked.

2. Interpret the gradient and the y-intercept for the following equations. Indicate why it may be inappropriate to be concerned with the value of the y-intercept, and discuss any other limitations on x- and y-values for the model.

a. $y = 1.3x - 0.9$ where x is quarters since March 1980 and y is the number of people employed in forestry (in 000's)

b. $y = 1.4\text{E}3\ x + 1.1\text{E}2$ where x is measured in days since 1 Jan 2012 and y is the number of bacteria (in millions) in a Petri dish.

c. $y = -0.2x + 10$ where x is years since 1962 and y is the marriage rate per 1 000 people in the population.

Alternative models for the trend

Often a single straight line may not fit the data particularly well over a long period of time.

Non-linear models

If a linear trend line is not a good fit, then consider modelling the trend using non-linear functions such as:

- logarithmic
- polynomial
- power
- exponential.

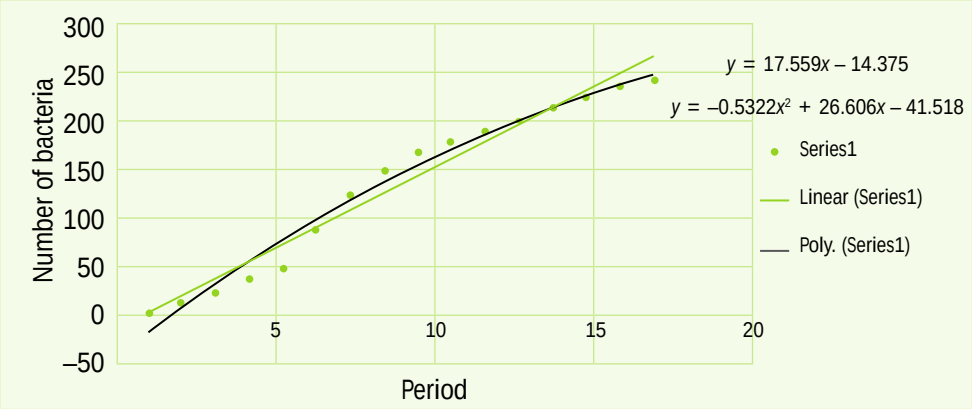
These can be fitted automatically using graphics calculators or Excel.

With alternative (non-linear) models it is most important to comment on their effect on the forecast. Non-linear graphs may behave similarly to the smoothed data for a short section of the graph, but beyond this section non-linear graphs could increase/decrease dramatically, or have other inappropriate features, so that their equations are of little use for forecasting.

Excel can automatically provide the **coefficient of determination**, R^2, for the trend line (the closer the R^2 value is to 1, the better the fit of the trend line). However, while an R^2 value close to 1 does indicate that a trend line is a very good fit to the data, it is important to consider also the *appropriateness* of the model (the closeness of its fit to the data and its behaviour beyond the data supplied), and not rely solely on the R^2-value when judging the reliability of a trend line.

Example

A trend line ($y = 17.559x - 14.375$) and a quadratic ($y = -0.5322x^2 + 26.606x - 41.518$) are fitted to the graph of the moving means (Series 1) for a time series.



From the graph it appears that the linear equation does not fit the data particularly well overall, whereas the quadratic is more consistently a better fit.

However, the quadratic will continue to decrease quite steeply beyond the 17th time period, and so will give lower forecasts for future periods. For this reason the quadratic may not be suitable for forecasting values well ahead in time.

Note: The R^2-values for the line and the quadratic are 0.9647 and 0.9796 respectively, so both the line and the curve are good fits to the moving mean values (series 1).

Piecewise trend lines

If moving means have two (or more) sections with different characteristics, a **piecewise function** may be used, in which a separate trend line is fitted to each section of data.

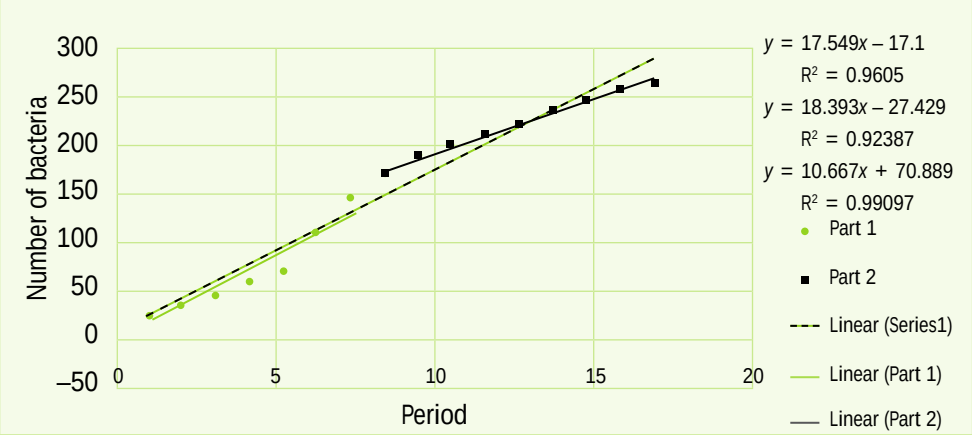
Piecewise functions arise when some external occurrence causes a step change in time series values, e.g. the time series of the numbers of fatalities prior to compulsory seat-belt use would have a different trend to that of the time series of the numbers of fatalities after using seat belts was made compulsory.

In a time series, recent data is usually more relevant (than earlier data) to what will happen in the future. So a piecewise trend line with a good fit to the most recent moving means may allow more accurate forecasts.

Investigate times series data

Example

Moving means are plotted for some smoothed time series data (Series 1).



Looking at the overall trend line ($y = 17.549x - 17.1$), it appears that this single linear trend line does not fit the data well, particularly for the later periods. A much better fit is achieved (particularly for the later periods) by splitting the moving means into two sections and fitting a linear model to each section.

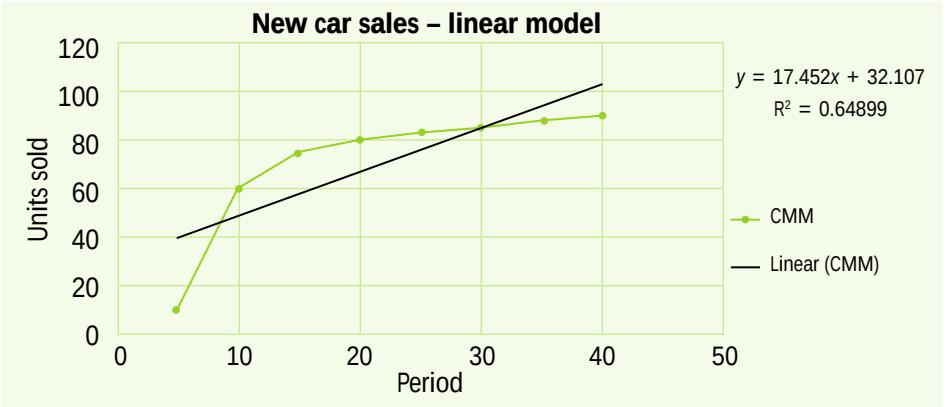
The trend line modelling the data from periods 8 to 16 ($y = 10.667x + 70.889$) can be seen to model the data closely (confirmed by an R^2-value of 0.991). Using this piecewise function instead of the overall trend line would result in lower forecasts for future periods.

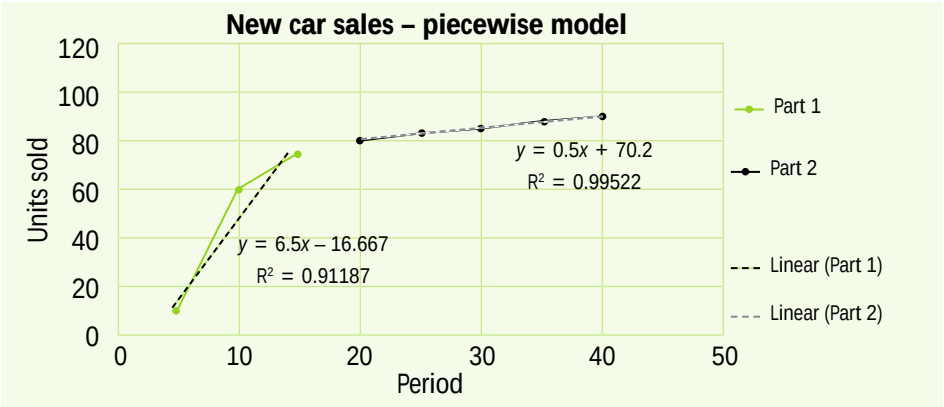
Exercise E: Alternative models for the trend using Excel

Ans. p. 29

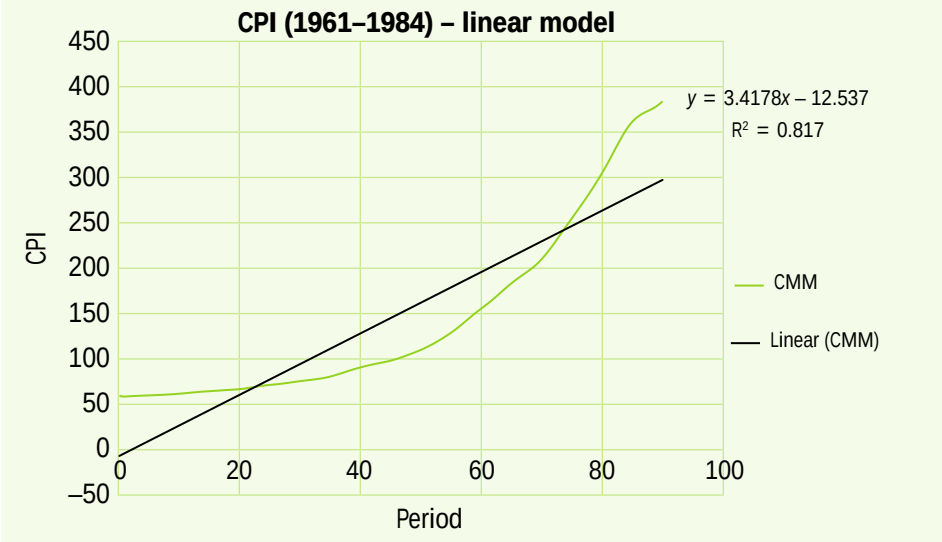
1. For each of the following time series graphs, linear models and an alternative model are shown. Select the model that would give a better fit for the data. Give reasons for your choice. Comment on the effect this alternative model would have on any forecasts.

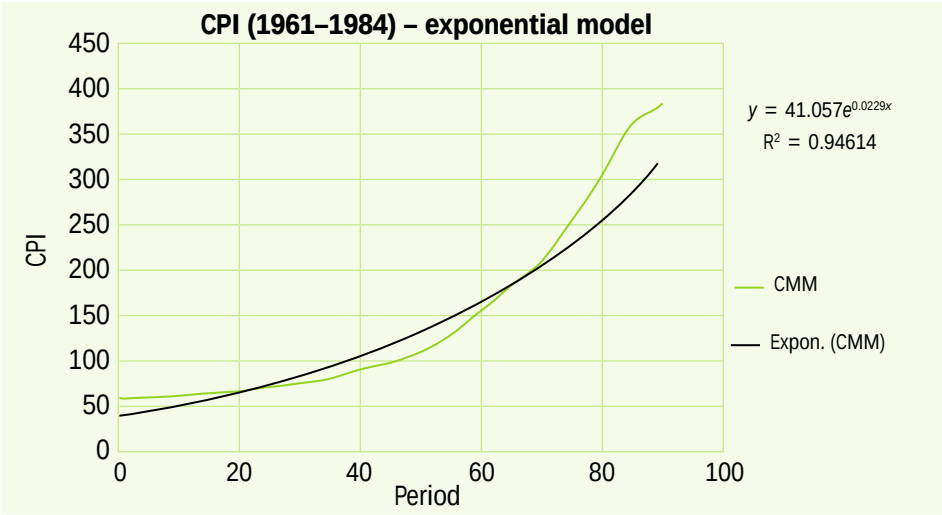
a.





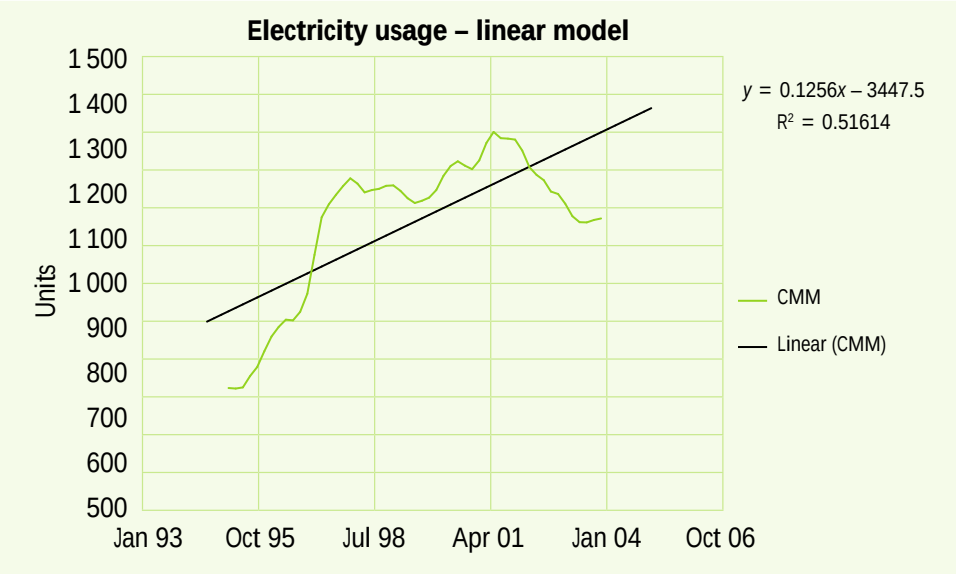
b.

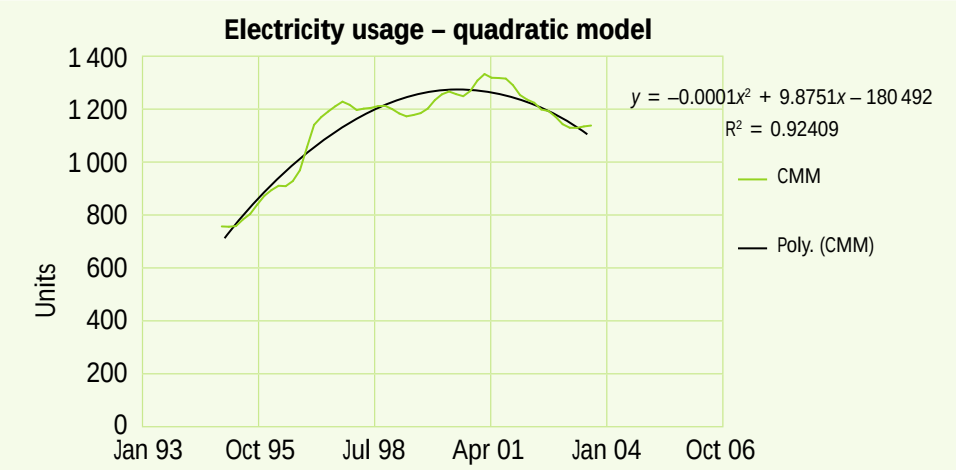




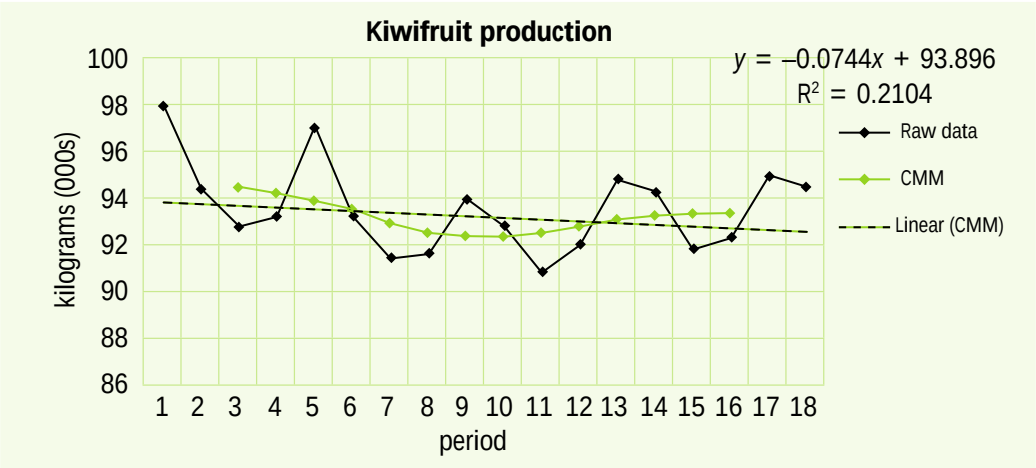
Investigate times series data

c.





2. Comment on the appropriateness of the single linear model for the time series graph of kiwifruit production. Suggest a more appropriate model for the graph. Justify your choice and explain any limitations it has.



Index series

When investigating changes in the values of a time series variable, it can be preferable to use **index numbers** to measure *relative* changes in these values instead, such as percentage increases or decreases. This allows comparisons between values to be made more easily, e.g comparing prices using the Consumer Price Index (CPI).

To calculate the index number for a data value, first select a base period and assign its raw value an index number of 100%. The raw values for other periods will be compared with the value for the base period, to create an **index series** of values.

For example, if the base period has value 28, and the next period has value 31, then the index number for the second period is $\frac{31}{28} \times 100\% = 110.7\%$.

As index numbers are usually given to 1 decimal place, it is common to avoid decimals by multiplying by a factor of 1 000 instead of 100 (or alternatively multiply the percentage value by 10), e.g. in the calculation above, the index number for 31 would be 1107.

Example

In the table below of overseas visitor numbers, the period 2001M03 (month 3, 2001) is used as the base (index number 100% or 1 000). A percentage index series is shown.

For example, the index number for 2001M04 is 106 300 ÷ 131 338 × 100 = 80.9% (or 809).

This indicates that visitor numbers were (100 – 80.9) = 19.1% lower in month 04, 2001 compared with Month 03, 2001.

Take care if you are investigating a percentage change in an index number between two periods (neither of which is the base period) as in this case you will need to investigate the ratio of the two percentages.

The percentage change between 2001M04 and 2001M05 is 59.6 ÷ 80.9 × 100 = 73.7%, which represents a 26.3% decrease.

Period	Date	Raw number of visitors	Index series as %	Index series with base 1 000
1	2001M03	131 338	100.0	1 000
2	2001M04	106 300	80.9	809
3	2001M05	78 232	59.6	596
4	2001M06	70 443	53.6	536
5	2001M07	88 007	67.0	670
6	2001M08	90 114	68.6	686
7	2001M09	83 352	63.5	635

Ans. p. 29

Exercise F: Index series

1. A table of quarterly values of the Consumer Price Index (CPI) in New Zealand (2005–2008) is shown below.

a. What quarter has been used as the base period for these CPI values?

b. Recalculate CPI values for each quarter, using the first quarter of 2005 (2005Q1) instead as the base period, in percentage and base 1 000 form.

c. Interpret the new index number for 2006 Q1.

Quarter	CPI	Index number %	Index number with base 1000
2005Q1	953		
2005Q2	961		
2005Q3	972		
2005Q4	979		
2006Q1	985		
2006Q2	1 000		
2006Q3	1 007		
2006Q4	1 005		
2007Q1	1 010		
2007Q2	1 020		
2007Q3	1 025		
2007Q4	1 037		
2008Q1	1 044		
2008Q2	1 061		
2008Q3	1 077		
2008Q4	1 072		

2. The following table shows the numbers of earthquakes registering over 7.0 (on the Richter scale) for New Zealand (1900–1920).

Year	Number of quakes > 7.0	Index number %	Index number base 1000
1900	13		
1901	14		
1902	8		
1903	10		
1904	16		
1905	26		
1906	32		
1907	27		
1908	18		
1909	32		
1910	36		
1911	24		
1912	22		
1913	23		
1914	22		
1915	18		
1916	25		
1917	21		
1918	21		
1919	14		
1920	8		

Investigate times series data

Using 1900 as the base year:

a. calculate the index number for each year in percentage and base 1 000 form

b. calculate the percentage change in the index number from:

i. 1900 to 1905

ii. 1905 to 1906

iii. 1902 to 1910

Writing the conclusion

This Achievement Standard requires you to be familiar with the process of applying the statistical enquiry cycle (PPDAC) to time series analysis, which involves:

- using existing data sets
- selecting and using appropriate display(s)
- identifying features in the data and relating this to the context
- finding an appropriate model
- using the model to make an appropriate forecast
- communicating findings in a conclusion.

At Achievement level, the conclusion to the time series investigation should show evidence of using each component of the statistical enquiry cycle. You will demonstrate your understanding of time series by:

- selecting an appropriate variable to investigate
- smoothing the data appropriately
- inserting a trend line.

At Merit level, the components of the statistical enquiry cycle should be linked to the context (the real-life setting of the time series). Statements should be supported by evidence from the data, and the trends and features of the data displays. You should:

- discuss the **variability** within the raw data – Is the variability consistent? Are there outliers? If so, is there a possible explanation?
- describe the **closeness of the fit** of the linear model to the moving means – this is based mainly on a visual inspection of the graph, with reference to the R^2 value (a good fit is confirmed by R^2 being close to 1, but keep in mind that with the smoothing of the data the R^2 value is expected to be high)
- use your analysis to make a forecast.

At Excellence level you need to evaluate and reflect on all stages of the statistical enquiry cycle. Your grade of Excellence will depend on the quality of the written analysis of the investigation, which must demonstrate statistical insight and contextual knowledge. Once you have made your forecast you will need to consider its accuracy, reliability and relevance. Points to consider include the following.

- The amount of data available – with more data, the average seasonal effects are more reliable (they will be calculated using a greater number of individual seasonal effects), and trends in the time series may be more obvious – for example, whether shifts are permanent or merely part of a long-term cycle. If there is a change in trend, try to identify a possible cause for this.
- An interpretation of the gradient of the trend line in context and consideration of the variability of the moving means about the trend line – if there is an apparent change in the variability, is it increasing or decreasing? If non-linear or piecewise functions were used to model the trend, justify their use and consider their limitations.

Investigate times series data

- The variability of the individual seasonal effects for a particular time period – significant variations in the values of the individual seasonal effects will impact on the reliability of the estimate of the average seasonal effect. In turn this affects the reliability of the forecast for that time period.
- How close the time period of the forecast is to the last time series period – proximity to recent data should mean the forecast is more reliable and useful. If the forecast is for a time period too far ahead it may be less reliable, as conditions may have changed from when the time series data was collected. You should also be aware of the gap at the end (and start) of the series of (centred) moving means – for quarterly data, this means that there is no value for either of the last two quarters with the missing values being closest to where you are likely to be making your forecast.
- The trend line gives equal weighting to all moving means; however, it may be more appropriate to put greater emphasis on the most recent data, particularly if there has been a recent change in the trend.
- The calculation and interpretation of residuals (this may make any **outliers** more obvious – try to identify a possible reason in context for outliers) and seasonally adjusted data (to identify if data is higher or lower than would be expected).
- Who could make use of the forecast – for what purpose and how the data and the forecast would allow them to do this.
- Deriving and interpreting index numbers.

The following example focuses on the conclusions once the analysis has been completed.

Example

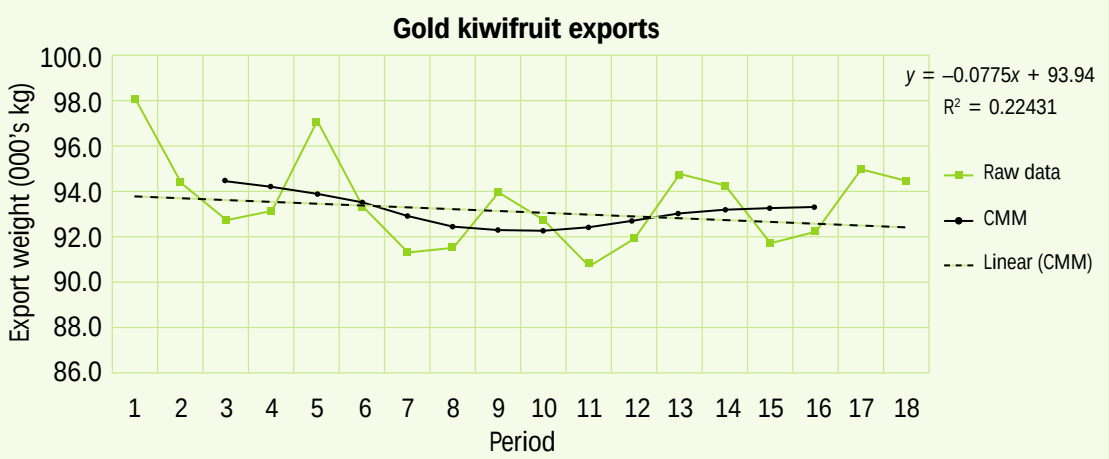
A time series investigation was carried out using the table of raw data for gold kiwifruit exports (000 kg) by quarter, as shown below. The data was analysed with moving means (CMM), individual and average seasonal effects (ISE and ASE), residuals (RSL), trend values (Trend), seasonally adjusted values (SAV) and index numbers calculated. Forecasts were made for each quarter, for the following two years.

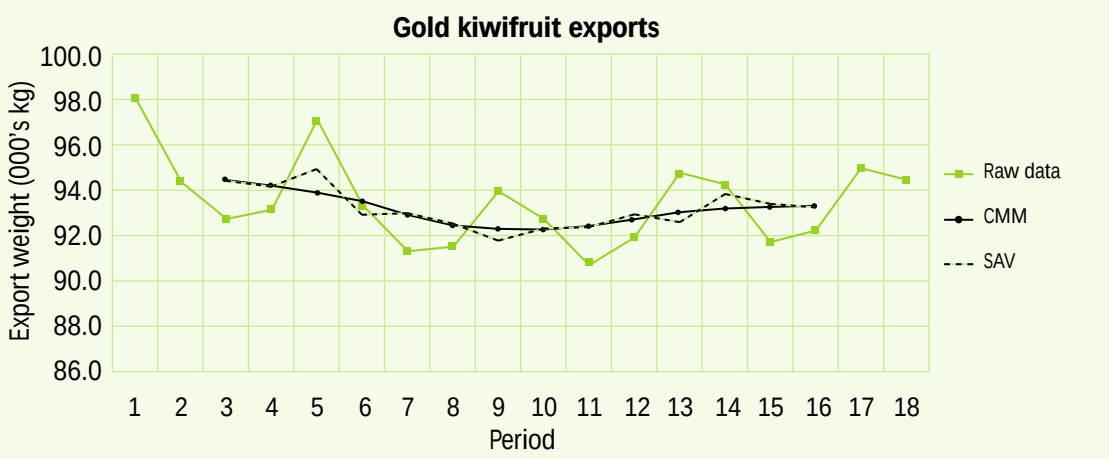
Period	Quarter	Raw	CMM	ISE	ASE	RSL	Trend	SAV	Index
1	Sep-05	98.0							100.0
2	Dec-05	94.4							96.3
3	Mar-06	92.8	94.49	–1.69	–1.63	–0.06	93.71	94.43	94.7
4	Jun-06	93.2	94.25	–1.05	–0.98	–0.07	93.63	94.18	95.1
5	Sep-06	97.1	93.95	3.15	2.16	0.99	93.55	94.94	99.1
6	Dec-06	93.4	93.58	–0.17	0.44	–0.61	93.48	92.96	95.3
7	Mar-07	91.4	92.99	–1.59	–1.63	0.04	93.40	93.03	93.3
8	Jun-07	91.6	92.53	–0.93	–0.98	0.05	93.32	92.58	93.5
9	Sep-07	94.0	92.38	1.63	2.16	–0.53	93.24	91.84	95.9
10	Oct-07	92.8	92.35	0.45	0.44	0.01	93.16	92.36	94.7
11	Mar-08	90.8	92.50	–1.70	–1.63	–0.07	93.09	92.43	92.7
12	Jun-08	92.0	92.79	–0.79	–0.98	0.19	93.01	92.98	93.9
13	Sep-08	94.8	93.10	1.70	2.16	–0.46	92.93	92.64	96.7
14	Dec-08	94.3	93.26	1.04	0.44	0.60	92.86	93.86	96.2
15	Mar-09	91.8	93.33	–1.52	–1.63	0.10	92.78	93.43	93.7
16	Jun-09	92.3	93.38	–1.08	–0.98	–0.10	92.70	93.28	94.2
17	Sep-09	95.0							96.9
18	Dec-09	94.5							96.4

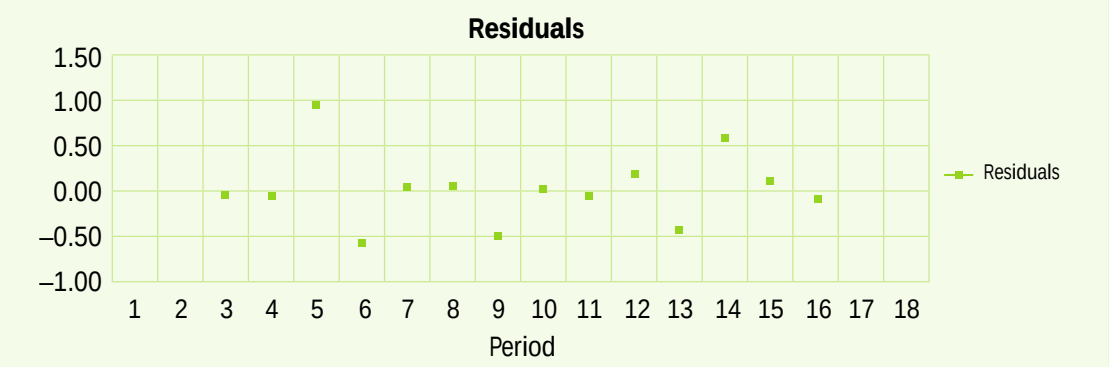
Period	19	20	21	22	23	24	25	26
Quarter	Mar-10	Jun-10	Sep-10	Dec-10	Mar-11	Jun-11	Sep-11	Dec-11
Forecasts	90.84	91.41	94.47	92.67	90.53	91.10	94.16	92.36

Investigate times series data

The following graphs were among those drawn.







Use the analysis to communicate the findings of this investigation.

Solution

The following comments could be made in the conclusion.

Trends

This time series gives the exported weight (in thousands of kilograms) of gold kiwifruit. The long-term trend is that exported weight is decreasing at an average rate of 77.5 kg per quarter (as indicated by the –0.0775 gradient of the trend line for the centred moving means). This decrease may be a result of lower demand overseas for this produce, potentially due to increased competition from alternative suppliers. However, the graph shows that more recently the decreasing trend seems to have been reversed.

There is a clear seasonal trend in this data, with exported weights highest during the September quarter, which may correspond to when the fruit is ready for use; and at their lowest in the March/June quarters, which could correspond to when the fruit is being harvested and not ready for export.

While there are no obviously unusual data values, the data points for period 1 (Sept 05) and period 5 (Sept 06) appear higher than expected. This may be due to particularly good growing seasons with higher demand, or they may be part of a pattern of changing seasonal variations – further data from previous seasons would be required to confirm if these were part of a trend.

The residuals graph shows a higher than average residual value for September 06 (period 5) – when exported quantity was higher than expected (confirming what was observed above). The residuals also suggest a lower than expected production in December 06 (period 6) and September 07 (period 9). These may have been due to a lack of supply due to previous higher volumes of exports.

Limitations of the model

The data was based on information provided by the Association of Kiwifruit growers in New Zealand. Not all growers would necessarily belong to this association and hence this data will only show trends for those who do, and not for the whole New Zealand gold kiwifruit industry. The linear model may give a variety of predictions depending on whether one line or piecewise functions are used. It would also need to be considered if the upturn in export numbers is likely to continue or whether this was part of a long-term cycle. There is significant variability in the smoothed data (CMM) about the trend line. This would impact significantly on the reliability of any forecast made using the single linear model.

Appropriateness of the model

It can be observed both visually and from the R^2 value of 0.2243 that the single linear model is not a good fit for the data. The forecast obtained is likely to be too low as the model predicts the weight of exported kiwifruit will continue to decrease over time. It would therefore be appropriate to consider other models.

While a polynomial of degree 2 visually fits the data well (not shown), with an R^2-value of 0.8936, a more appropriate model would be a piecewise function as this would allow for the two distinct portions of the graph (that prior to and after period 10). Using a piecewise function (not shown) gives an R^2-value of 0.9007 for the second portion of the graph and its positive gradient reflects the increasing nature of the trend for this portion of the graph. Using this piecewise function would result in a higher forecast for subsequent dates. For example, the forecast for December 2011 using the original model gives a quantity of 92 365 kg whereas the piecewise function gives a forecast for the same period of 95 694 kg. This higher forecast would appear to be more realistic given the most recent data.

Reliability of the calculations

Each of the estimated average seasonal effects has been calculated using the values of only three or four individual seasonal effects – which reduces confidence in their accuracy. However, the four values used for the March quarter were consistent (ranging from –1.7 to –1.52). In contrast, the three values used for the September quarter (ranging from 1.63 to 3.15) were less consistent, with the oldest value being significantly higher than the other values. Further investigation would be required to determine whether this value is an anomaly, as it has resulted in slightly higher forecasts for the September quarter than is indicated by the other values used.

Forecasts have been made for two years in advance of the most recent data point, which could be too far ahead. Within this time there could be significant events that would affect the quantity of kiwifruit exported, such as the economic downturn and the spread of the kiwifruit virus Psa-V, which would have major effects on the reliability of any forecast.

Relevance and usefulness of the model

The piecewise model appears most relevant as it reflects the increase in exported quantity over the two most recent years of data. Growers need to be aware of this to allow for planning for increased production such as staffing requirements and storage facilities. Export transporters, such as airlines, would find this information useful to make judgements about any increased capacity required.

Seasonally adjusted data

The graph comparing seasonally adjusted values (SAV) with raw and centred moving mean values (CMM) shows that in the periods September 06, March 07, June 07, etc., the seasonally adjusted values were greater than the centred moving mean values, which means export quantities of kiwifruit were greater than expected in those periods. For example in September 06 (period 5) the value for the SAV was 94 942 kg and the CMM was 93 950 kg. This means export quantities were 992 kg more than expected.

In March 06, June 06, December 06 etc., the seasonally adjusted value was less than the centred moving mean value, so export quantities were less than expected in those periods. For example, in September 07 (period 9) the SAV was 91 842 kg while the CMM was 92 380 kg. This means the export quantity was 538 kg less than expected.

Index series

From the index series, using September 05 (period 1) as the base month (index = 100% or 1 000), the index for March 2007 was 93.3 (or 933), indicating that the weight of gold kiwifruit exported in March 2007 was only 93.3% of the export weight in September 05 (i.e. 6.7% lower).

Ans. p. 30

Exercise G: Writing the conclusion using Excel

Use your own paper to write your conclusions in this exercise.

1. a. The table shows New Zealand total visitor data by month, from all countries of residence for the period 2009–2012. This data set is available on the ESA website RESOURCES.

Year, month	Total visitors (all countries)
2009M01	200 077
2009M02	194 514
2009M03	170 648
2009M04	141 757
2009M05	103 533
2009M06	89 760
2009M07	105 945
2009M08	100 007
2009M09	99 017
2009M10	107 995
2009M11	132 359
2009M12	192 464
2010M01	211 478
2010M02	203 323
2010M03	174 330
2010M04	141 268
2010M05	104 428
2010M06	91 355

Investigate times series data

2010M07	107 762
2010M08	102 211
2010M09	99 421
2010M10	106 645
2010M11	132 847
2010M12	188 288
2011M01	212 309
2011M02	202 582
2011M03	167 986
2011M04	139 770
2011M05	101 706
2011M06	88 817
2011M07	104 792
2011M08	103 339
2011M09	117 844
2011M10	124 892
2011M11	132 219
2011M12	192 155
2012M01	212 693
2012M02	191 772
2012M03	165 596
2012M04	138 406

Source: Statistics New Zealand

Using the New Zealand visitor data in the table above (available on the ESA website), carry out an analysis and make a forecast for month 6 of 2012 by:

i. smoothing the data using an appropriate moving average
ii. calculating individual and average seasonal effects
iii. plotting the raw and smoothed data on the same grid
iv. inserting the line of best fit and finding its equation
v. using the model to make an appropriate forecast.

b. Write a conclusion for your investigation. Your report may include comments about the following aspects:

- short- and long-term trends visible in your graph, with possible explanations
- variability in the data – does the trend line lie close to the smoothed data?
- interpretation of the gradient in context
- the reliability of the (average) seasonal effects
- calculation, interpretation and display of residuals
- calculation and interpretation of seasonally adjusted data
- the reliability and validity of your forecast
- who might find your forecasts useful?
- alternative models(s), such as a piecewise function, and the effects on the forecast.

2. Repeat the analysis from question 1, using the sunspot data (this data set is available on the ESA website RESOURCES), making a forecast for January 2012. (Data source: NASA / Marshall Solar Physics.)

Investigate times series data

Answers

Exercise A: Residuals with Excel (page 2)

1. a.

Quarter	Raw data (litres of wine produced)	Individual seasonal effects	Average seasonal effects	Residuals
Mar 2000	154.3			
Jun 2000	147.5			
Sep 2000	161.4	1.325	8.958	–7.633
Dec 2000	166.8	3.5125	–0.146	3.6585
Mar 2001	174.9	8.75	–4.058	12.808
Jun 2001	152.6	–18.5	–2.046	–16.454
Sep 2001	179.2	9.45	8.958	0.492
Dec 2001	188.6	19.7625	–0.146	19.9085
Mar 2002	142.3	–30.775	–4.058	–26.717
Jun 2002	177.9	7.9	–2.046	9.946
Sep 2002	187.8	16.1	8.958	7.142
Dec 2002	155.4	–23.7125	–0.146	–23.5665
Mar 2003	189.1	9.85	–4.058	13.908
Jun 2003	190.4	4.4625	–2.046	6.5085
Sep 2003	176.4			
Dec 2003	220.3			

b. Volumes produced in Mar 2001, Dec 2001 and Mar 2003 are higher than expected for those seasons (large positive residual values), while Jun 2001, Mar 2002 and Dec 2002 had significantly lower volumes than expected for those seasons (large negative residuals).

2.

Year/ month	CPI	CMM	ISE	ASE	Residuals
2008M01	995				
2008M02	980				
2008M03	1010				
2008M04	996				
2008M05	1 034				
2008M06	1 088				
2008M07	1 127	1 091.46	–35.54	–114.21	78.67
2008M08	1 235	1 101.96	–133.04	–95.46	–37.58
2008M09	1 208	1 111.38	–96.63	–46.75	–49.88
2008M10	1 136	1 119.96	–16.04	–17.69	1.65
2008M11	1 133	1 125.42	–7.58	17.42	–25.00
2008M12	1 089	1 132.5	43.50	45.78	–2.28
2009M01	1 128	1 143.96	15.96	–5.29	21.25
2009M02	1 099	1 149.13	50.13	29.36	20.76
2009M03	1 117	1 144.88	27.88	37.50	–9.63
2009M04	1 095	1 138.33	43.33	63.43	–20.10
2009M05	1 066	1 131.92	65.92	73.76	–7.85
2009M06	1 226	1 127.38	–98.63	–58.17	–40.46
2009M07	1 264	1 125.96	–138.04	–114.21	–23.83
2009M08	1 222	1 124.63	–97.38	–95.46	–1.92
2009M09	1 119	1 121.29	2.29	–46.75	49.04
2009M10	1 068	1 116.54	48.54	–17.69	66.24
2009M11	1 047	1 112.25	65.25	17.42	47.83

2.

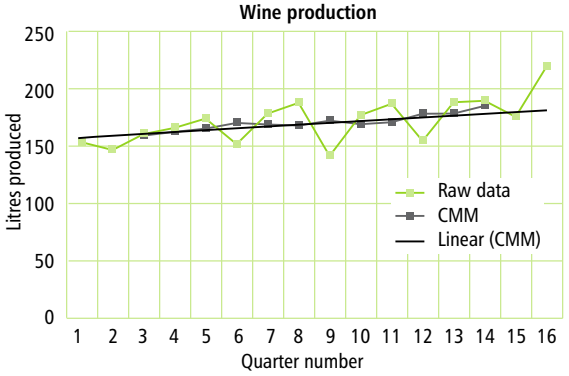
Year/ month	CPI	CMM	ISE	ASE	Residuals
2009M12	1 066	1 105.54	39.54	45.78	–6.24
2010M01	1 117	1 098.58	–18.42	–5.29	–13.13
2010M02	1 078	1 093.96	15.96	29.36	–13.40
2010M03	1 058	1 094.75	36.75	37.50	–0.75
2010M04	1 040	1105.21	65.21	63.43	1.78
2010M05	1 018	1 117.71	99.71	73.76	25.94
2010M06	1 113	1 125.38	12.38	–58.17	70.54
2010M07	1 210	1 131.63	–78.38	–114.21	35.84
2010M08	1 165	1 139.96	–25.04	–95.46	70.42
2010M09	1 195	1 149.08	–45.92	–46.75	0.83
2010M10	1 243	1 157.42	–85.58	–17.69	–67.89
2010M11	1 172	1 166.58	–5.42	17.42	–22.83
2010M12	1 125	1 179.29	54.29	45.78	8.51
2011M01	1 208	1 194.58	–13.42	–5.29	–8.13
2011M02	1 187	1 209	22.00	29.36	–7.36
2011M03	1 168	1 215.88	47.88	37.50	10.38
2011M04	1 130	1 211.75	81.75	63.43	18.32
2011M05	1 148	1 203.67	55.67	73.76	–18.10
2011M06	1 288	1 199.75	–88.25	–58.17	–30.08
2011M07	1 402	1 197.13	–204.88	–114.21	–90.67
2011M08	1 319	1 192.63	–126.38	–95.46	–30.92
2011M09	1 206				
2011M10	1 133				
2011M11	1 088				
2011M12	1 115				
2012M01	1 155				
2012M02	1 132				

b. The average seasonal effects are calculated using the values of only 3 or 4 individual seasonal effects. There is significant variability in these individual seasonal effects (e.g. month 07 ranges in values from –204.88 to – 35.54) hence the values used for average seasonal effects are not reliable.

c. The residuals show that the CPI for fruit and vegetables is higher than expected for the periods 2008 M07, 2009 M9/10/11, and 2010 M6/8. CPI for fruit and vegetables is lower than expected for the months 2008 M8/9, 2009 M6, 2010 M10 and 2011 M 6/7/8.

Exercise B: Trend lines and forecasts using Excel (page 6)

1. a. b. c. *Answers will vary for line drawn by eye.*



Investigate times series data

d. *Answers may vary for forecasts* (answers are to 3 sf)

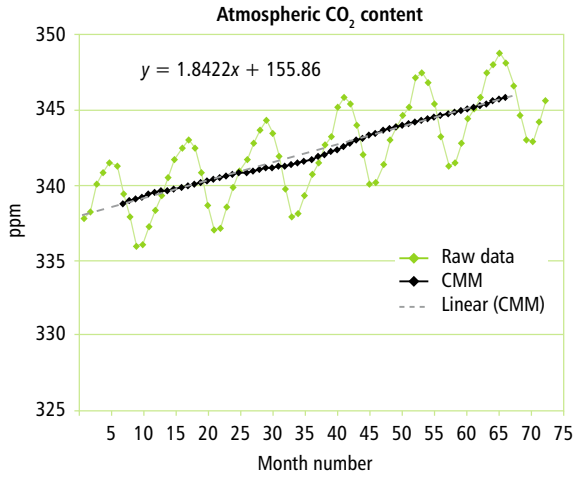
i. Mar 2004: 187 – 4.058 = 183 L

ii. Jun 2004: 189 – 2.046 = 187 L

iii. Sep 2004: 190 + 8.958 = 199 L

iv. Dec 2004: 193 – 0.146 = 193 L

2. a.



b.

Individual seasonal effects							Average seasonal effects
Year	2006	2007	2008	2009	2010	2011	
Jan		–0.23	0.16	–0.42	0.05	–0.17	–0.122
Feb		0.79	0.80	0.64	0.66	0.58	0.694
Mar		1.98	1.82	0.98	1.10	2.03	1.582
Apr		2.63	2.67	2.77	2.91	2.45	2.686
May		3.08	3.20	3.22	3.15	3.05	3.14
Jun		2.41	2.24	2.64	2.38	2.32	2.398
Jul	0.58	0.60	0.73	1.03	0.92		0.772
Aug	–1.1	–1.69	–1.5	–1.14	–1.38		–1.362
Sep	–3.12	–3.35	–3.45	–3.29	–3.52		–3.346
Oct	–3.16	–3.36	–3.33	–3.31	–3.41		–3.314
Nov	–2.14	–2.10	–2.29	–2.23	–2.15		–2.182
Dec	–1.15	–0.84	–1.03	–0.74	–0.73		–0.898

January has more variability in its individual seasonal effects (some are positive, some are negative) but as they are all close to zero the estimate of –0.122 for the average seasonal effect seems a reliable one. Generally the individual seasonal effects for each season seem to be similar enough in size to each other to make the value of the average seasonal effect appear to be reliable.

c. June 2012 = season 78, so estimate is approximately 347 plus 2.4 (June seasonal effect) = 349 ppm approximately.

Exercise C: Seasonally adjusted data using Excel (page 11)

1.

Quarter	Raw data (litres of wine produced)	Centred moving mean (CMM)	Average seasonal effects (ASE)	Seasonally adjusted values (SAV)	SAV compared with CMM
Mar-00	154.3				
Jun-00	147.5				
Sep-00	161.4	160.075	8.958	152.442	lower
Dec-00	166.8	163.2875	–0.146	166.946	higher
Mar-01	174.9	166.15	–4.058	178.958	higher
Jun-01	152.6	171.1	–2.046	154.646	lower
Sep-01	179.2	169.75	8.958	170.242	higher
Dec-01	188.6	168.8375	–0.146	188.746	higher
Mar-02	142.3	173.075	–4.058	146.358	lower
Jun-02	177.9	170	–2.046	179.946	higher
Sep-02	187.8	171.7	8.958	178.842	higher
Dec-02	155.4	179.1125	–0.146	155.546	lower
Mar-03	189.1	179.25	–4.058	193.158	higher
Jun-03	190.4	185.9375	–2.046	192.446	higher
Sep-03	176.4				
Dec-03	220.3				

Answers will vary.

The amount of wine produced in Dec-00, Mar-01, Sep-01, Dec-01, Jun-02, Sep-02, Mar-03, and Jun-03 is higher than expected once seasonal variations are adjusted for. Production for Sep-02 was higher than production in Jun-02, but once seasonally adjusted, production in Jun-02 was higher than production in Sep-02.

2.

Food price index Level 2 subgroups for New Zealand (monthly)					
Month	Fruit and vegetables index raw data	CMM	Average seasonal effects (ASE)	Seasonally adjusted values (SAV)	SAV compared with CMM
2008M01	995				
2008M02	980				
2008M03	1 010				
2008M04	996				
2008M05	1 034				
2008M06	1 088				
2008M07	1 127	1 091.46	–114.21	1 241.21	higher
2008M08	1 235	1 101.96	–95.458	1 330.46	higher
2008M09	1 208	1 111.38	–46.75	1 254.75	higher
2008M10	1 136	1 119.96	–17.694	1 153.69	higher
2008M11	1 133	1 125.42	17.4167	1 115.58	lower
2008M12	1 089	1 132.5	45.7778	1 043.22	lower
2009M01	1 128	1 143.96	–5.2917	1 133.29	lower
2009M02	1 099	1 149.13	29.3611	1 069.64	lower
2009M03	1 117	1 144.88	37.5	1 079.5	lower

Investigate times series data

2009M04	1 095	1 138.33	63.4306	1 031.57	lower
2009M05	1 066	1 131.92	73.7639	992.236	lower
2009M06	1 226	1 127.38	–58.167	1 284.17	higher
2009M07	1 264	1 125.96	–114.21	1 378.21	higher
2009M08	1 222	1 124.63	–95.458	1 317.46	higher
2009M09	1 119	1 121.29	–46.75	1 165.75	higher
2009M10	1 068	1 116.54	–17.694	1 085.69	lower
2009M11	1 047	1 112.25	17.4167	1 029.58	lower
2009M12	1 066	1 105.54	45.7778	1 020.22	lower
2010M01	1 117	1 098.58	–5.2917	1 122.29	higher
2010M02	1 078	1 093.96	29.3611	1 048.64	lower
2010M03	1 058	1 094.75	37.5	1 020.5	lower
2010M04	1 040	1 105.21	63.4306	976.569	lower
2010M05	1 018	1 117.71	73.7639	944.236	lower
2010M06	1 113	1 125.38	–58.167	1 171.17	higher
2010M07	1 210	1 131.63	–114.21	1 324.21	higher
2010M08	1 165	1 139.96	–95.458	1 260.46	higher
2010M09	1 195	1 149.08	–46.75	1 241.75	higher
2010M10	1 243	1 157.42	–17.694	1 260.69	higher
2010M11	1 172	1 166.58	17.4167	1 154.58	lower
2010M12	1 125	1 179.29	45.7778	1 079.22	lower
2011M01	1 208	1 194.58	–5.2917	1 213.29	higher
2011M02	1 187	1 209	29.3611	1 157.64	lower
2011M03	1 168	1 215.88	37.5	1 130.5	lower
2011M04	1 130	1 211.75	63.4306	1 066.57	lower
2011M05	1 148	1 203.67	73.7639	1 074.24	lower
2011M06	1 288	1 199.75	–58.167	1 346.17	higher
2011M07	1 402	1 197.13	–114.21	1 516.21	higher
2011M08	1 319	1 192.63	–95.458	1 414.46	higher
2011M09	1 206				
2011M10	1 133				
2011M11	1 088				
2011M12	1 115				
2012M01	1 155				
2012M02	1 132				

After adjusting for seasonal effects, the CPI for fruit and vegetables is higher than expected, in 2008 Months 7, 8, 9,10; in 2009 Months 6, 7, 8, 9; in 2010 Months 1, 7, 8, 9, 10 and in 2011 Months 1, 6, 7, 8.

Exercise D: Interpreting the gradient of an Excel trend line (page 14)

1. a. Seal numbers are increasing at an average rate of approximately 10.2 seals per quarter.
   b. The numbers of house sales are decreasing at a rate of approximately 52 per month.
   c. The weight of cherries picked per day is increasing at a rate of approximately 0.012 tonnes/day.
2. a. The number of people employed in forestry is increasing at a rate of 1 300 per quarter. The *y*-intercept indicates there were –900 people employed in forestry in March 1980, which is impossible, so the model does not apply for this date.
   b. The number of bacteria is increasing at a rate of 1 400 000 000 per day. According to the model, on 1 January 2012 there were 110 000 000 bacteria present. This is a possible value and could be confirmed by comparison with actual measurements.
   c. The marriage rate per 1 000 people is decreasing at a rate of 0.2 people per 1 000, per year. The *y*-intercept indicated that the marriage rate in 1962 was 10 per 1 000 people – this is a possible value and could be confirmed by comparison with recorded values. The equation indicates that after 50 years (2012) the rate will be negative, which is impossible, so the model does not apply this far into the future.

Exercise E: Alternative models for the trend using Excel (page 16)

Answers will vary – examples follow.

1. a. The model that fits the moving means better is the piecewise linear model. The second part of the piecewise model gives more weighting to recent sales figures which have moving means that can be seen to be very well modelled by a straight line (high R^2 value confirms this). As long as the sales trend continues, sales forecasts would be expected to be accurate (assuming that the initial steep growth in sales in periods 1–15 is not repeated in the future, and that there is no sales slump for economic or other reasons).
   b. The model that fits the moving means better is the exponential model, which can be seen to have a close fit to the moving means (and a high R^2 value). The exponential model is not a good fit for the most recent data however, as it lies below the CMM graph for period 80 and beyond, so forecasts would be expected to underestimate future values of the moving mean.
   c. The quadratic model is a better fit than the linear model for the given set of CMM values. However, the quadratic model can be seen to fit the CMM graph only moderately well, with actual values of the CMMs cycling above and below the model as time passes. This variation of CMM values from the trend line would make forecasts unreliable.

   Also, the graph of an inverted parabola curves downwards consistently after its peak so the quadratic model is very likely to be unrealistic as it does not allow for any upward movements in energy usage after 2000, as shown by the the last few CMM values which are trending upwards again (making quadratic forecasts dubious). It also predicts that electricity usage will fall to zero within a relatively short space of time.
2. *Answers will vary.*

   A better model for the smoothed data might be a polynomial of degree 3 (a cubic). This would result in higher forecasts for future periods (after $x = 17$) than forecasts made using the linear model. After some time the cubic model will predict production will fall again, so in the long term the cubic model will not be useful. Alternatively, a piecewise linear model may be appropriate, depending on whether the recent long-term upward trend is sustained.

Exercise F: Index series (page 20)

1. a. 2006Q2
   b.

Quarter	CPI	Index number %	Index number with base 1 000
2005Q1	953	100.0	1 000
2005Q2	961	100.8	1 008
2005Q3	972	102.0	1 020
2005Q4	979	102.7	1 027
2006Q1	985	103.4	1 034
2006Q2	1 000	104.9	1 049
2006Q3	1 007	105.7	1 057
2006Q4	1 005	105.5	1 055
2007Q1	1 010	106.0	1 060
2007Q2	1 020	107.0	1 070
2007Q3	1 025	107.6	1 076
2007Q4	1 037	108.8	1 088
2008Q1	1 044	109.5	1 095
2008Q2	1 061	111.3	1 113
2008Q3	1 077	113.0	1 130
2008Q4	1 072	112.5	1 125

Investigate times series data

c. The consumer price index is 3.4% higher in 2006 Q1 compared with the CPI in 2005 Q1.

2. a.

Year	Number of quakes >7.0	Index number %	Index number base 1 000
1900	13	100.0	1 000
1901	14	107.7	1 077
1902	8	61.5	615
1903	10	76.9	769
1904	16	123.1	1 231
1905	26	200	2 000
1906	32	246.2	246
1907	27	207.7	2 077
1908	18	138.5	1 385
1909	32	246.2	2 462
1910	36	276.9	2 769
1911	24	184.6	1 846
1912	22	169.2	1 692
1913	23	176.9	1 769
1914	22	169.2	1 692
1915	18	138.5	1 385
1916	25	192.3	1 923
1917	21	161.5	1 615
1918	21	161.5	1 615
1919	14	107.7	1 077
1920	8	61.5	615

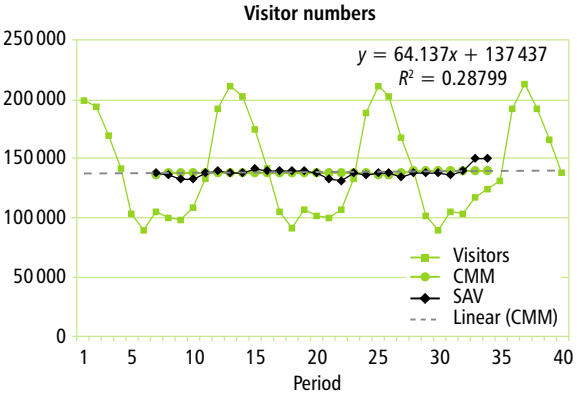
b. i. 100% increase ii. 23% increase

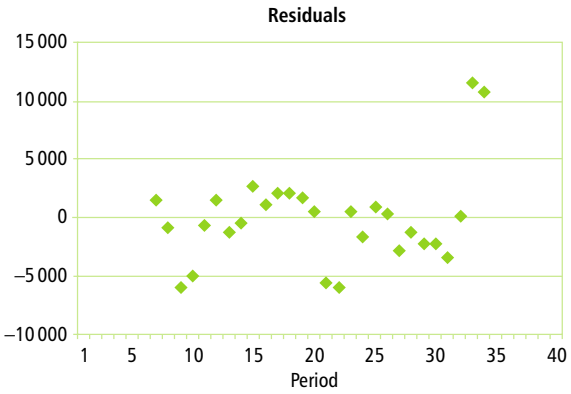
iii. 350% increase

Exercise G: Writing the conclusion using Excel (page 25)

1. *Answers will vary in detail included; an example follows.*

Investigating total visitor numbers to New Zealand over the last four years, from data sourced from Statistics NZ, the raw data indicates that in the long term, visitor numbers are gradually increasing. This could be due to the continuing emphasis on the economic importance of tourism for New Zealand, along with the increased numbers of flights to New Zealand; promotion of New Zealand as a tourist destination, and more affordable travel generally.





The data shows regular highs and lows in visitor numbers, corresponding to summer (highest visitor numbers) and winter (lowest visitor numbers). Within the winter lows there appears to be a regular small increase in numbers, this could correspond to the ski season with extra visitors arriving for snow sports.

From the raw data there do not appear to be any obviously unusual data values; however, the graph of the residuals indicates significantly higher than expected numbers of visitors in September and October 2011. This could be related to the hosting of the Rugby World Cup in those months.

The equation of the linear trend is $y = 64.137x + 137\,437$. This indicates visitor numbers overall are increasing at a rate of approximately 64 per month.

Forecasting for 2012M06 (June 2012, period 42) gives:

$(64.137 \times 42 + 137\,437) - 49\,577 = 90\,554$

i.e. expect around 90 500 visitors in Month 6, 2012.

We can be moderately confident in this forecast: the linear model fits the data well visually (although the *r*-value of 0.5367 indicates only a moderately strong correlation); the forecast is for only two periods ahead of the most recent data, so it is reasonably unlikely that conditions will change markedly before the time of the forecast. However, the estimate of the average June seasonal effect (–44 577) is the average of only two (reasonably similar) individual seasonal effects, so it is difficult to know how confident to be about this estimate of the June seasonal effect.

The seasonally adjusted data shows there were more visitors than expected for that time of year, in 2001 M9/10; 2002 M5/6/7/8/9/10/11; 2003 M8/9/10/11/12; 2004 M4/6/7, 2005 M6/7; 2006 M2/4/11, 2007 M2/3/4; 2008 M2/3/4/5/6; 2009 M4/5/12; 2010 M1/2/12; 2011 M1/2 (ignoring months where the difference was less than 1 000 visitors). For example, in 2002 M5 the value for the seasonally adjusted value was 125 434, compared with the CMM of 120 631.

From the derived index series for the raw data values of visitor numbers, using January 09 as the base month (index = 100), the index for January 10 was 105.7, indicating that visitor numbers were 5.7% higher in January 10 than in January 09.

The graph of the seasonally adjusted values dips substantially below the CMM values in 2009 months 9, 10 and in 2010 months 9,10, and is obviously well above in the final two periods 2011 months 9 and 10. As these points are close to the gaps in the centred moving mean (at the end of the series of CMMs) it would be necessary to confirm if this is a real shift in the pattern and not a random fluctuation.

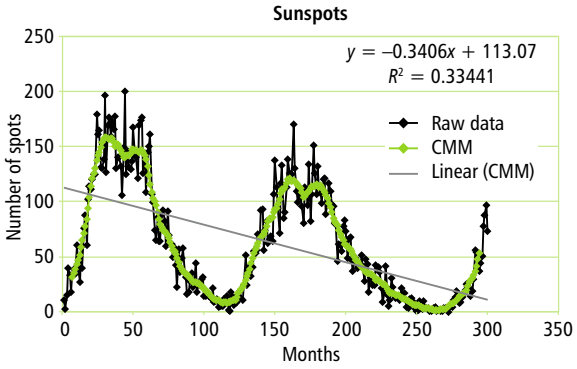
This forecast would be of use to tourism operators within the country, to allow them to cater for the potential increase in the number of visitors arriving in the off-season.

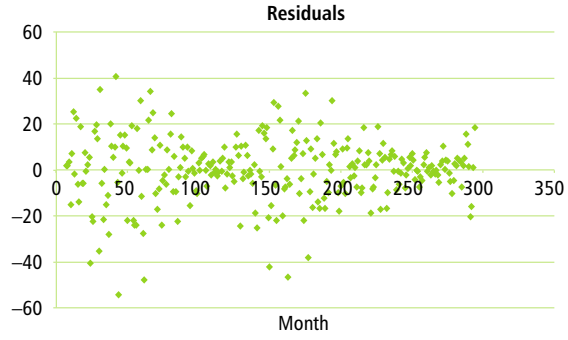
To improve the usefulness of the forecast, the country of origin and average length of stay could be used to further break down the figures, as this may have an impact on the facilities/activities required, e.g. the country of origin or length of visit may influence whether visitors travel independently or use package deals.

The information was supplied by Statistics New Zealand and hence should accurately reflect the visitor numbers to New Zealand, as all visitors should be accounted for.

Investigate times series data

2. Sunspot activity





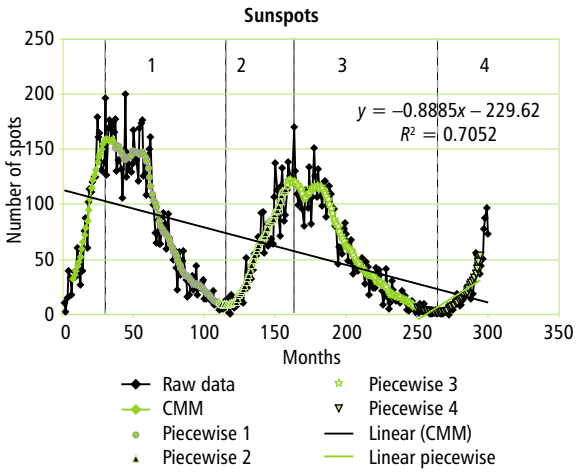
From the time series graph of the number of sunspots, there appears to be an overall decrease in sunspot numbers. There is a long-term cycle of approximately 11 years (the time between months of maximum sunspot activity). This is probably due to the underlying nature of sunspots (why the cycle occurs over 11 years is one of the mysteries of solar astronomy).

When sunspot activity is at its lowest, the number of sunspots per month does not fluctuate significantly (the size of the residuals is uniformly low, between –10 and 10); however, closer to times of maximums there is significant variation in the sunspot numbers per month (residuals vary more, mostly between –40 and 40). The graph of the residuals indicates more sunspots than expected in July 89, June 90, Dec 91 and July 01, with fewer than expected in Aug 90, Feb 92, Jun 99 Jul 00 and Sep 01.

The equation of the linear trend is given by $y = -0.3406x + 113.07$. This suggests the number of sunspots is decreasing at a rate of approximately 0.34 per month.

The forecast for January 2012 is for approximately 11.8 sunspots [–0.3406 × 301 + 113.070 + 1.227]. This forecast appears somewhat low given the number of sunspots observed in December 2011 (73 sunspots).

The linear trend line is not a good fit visually, with CMM values alternating above and below the line (the poor fit is confirmed by the low R^2-value). The model is not going to be suitable for predicting sunspot numbers in the long term as it predicts sunspot activity will continue to decline, giving predictions of negative numbers of sunspots within a relatively short space of time. Due to the cyclic nature of the trend, the use of a trigonometric model would be the most appropriate; however, due to constraints with this analysis this is impractical. Hence it may be more appropriate to use a piecewise function to model the data, modelling with the data divided into four separate sections, as shown below.
The linear model for the fourth section has equation
$y = 0.8885x - 229.62$



This gives a new forecast of 39 sunspots
(0.8885 × 301 – 229.62 + 1.227) which appears a more realistic number. Predicting the number of sunspots is the subject of much research and has proven difficult as there is a lack of understanding of the nature of sunspots.

There is significant variation in the individual seasonal effects, despite the fact that we have a large quantity of up-to-date data, hence we cannot be confident in our calculations of estimates of average seasonal effects. This only serves to emphasise the unknown nature of sunspots. There is also considerable variability in the data with significant variation in the CMM from the trend line, hence any forecast will be unreliable.

From the seasonally adjusted data, we can see there are more sunspots than expected in July 87, August 87, October 87, March 88, etc. This means, having allowed for seasonal effects, there were more sunspots than expected for the time of year.

From the derived index series for the raw data values of sunspot numbers, using 1987-01 as the base month (index = 1 000) the index for 1987-02 was 231, indicating that sunspot numbers were 76.9% lower in February 1987 than in January 1987.

Sunspot activity has implications for Earth's upper atmosphere, hence predicting the numbers of sunspots is important to Earth scientists – lack of sunspot activity in historical times has had associations with 'ice ages'.