STAT352: Applied Time Series [Assignment 1]

Question 1: 20 marks

(a) A (weakly) stationary time series is one which the mean and variance is constant regardless of time, **and** the covariance between two points of the series is only dependent upon the time difference h.

[2 marks]

(b) (i) $X_t = a + bZ_t + cZ_{t-2}$

[4 marks]

Mean:

$$\mathbb{E}[X_{t_t}] = \mathbb{E}[a + bZ_t + cZ_{t-2}] = a + b\mathbb{E}[Z_t] + c\mathbb{E}[Z_{t-2}] = a.$$

Variance:

$$\begin{aligned}
\text{Var}[X_t] = \mathbb{E}[X_t^2] - \mathbb{E}[X_t]^2 &= \mathbb{E}[(a + bZ_t + cZ_{t-2})(a + bZ_t + cZ_{t-2})] - a^2 \\
&= \mathbb{E}[a^2 + 2abZ_t + 2acZ_{t-2} + 2bcZ_tZ_{t-2} + b^2Z_t^2 + c^2Z_{t-2}^2] - a^2 \\
&= 2bc\mathbb{E}[Z_tZ_{t-2}] + b^2\mathbb{E}[Z_t^2] + c^2\mathbb{E}[Z_{t-2}^2]
\end{aligned}$$

(since the expectation of terms from the previous step with only one 'Z' are zero)

$$= \sigma^2(b^2 + c^2)$$

Autocovariance:

$$\begin{aligned}
\text{Cov}[X_t, X_{t+h}] &= \mathbb{E}[(a + bZ_t + cZ_{t-2})(a + bZ_{t+h} + cZ_{t+h-2})] - \mathbb{E}[X_t]\mathbb{E}[X_{t+h}] \\
&= \mathbb{E}\big[a^2 + abZ_{t+h} + acZ_{t+h-2} + abZ_t + b^2Z_tZ_{t+h} + bcZ_tZ_{t+h-2} \\
&\quad + acZ_{t-2} + bcZ_{t-2}Z_{t+h} + c^2Z_{t-2}Z_{t+h-2}\big] - a^2 \\
&= b^2\mathbb{E}[Z_tZ_{t+h}] + bc\mathbb{E}[Z_tZ_{t+h-2}] + bc\mathbb{E}[Z_{t-2}Z_{t+h}] + c^2\mathbb{E}[Z_{t-2}Z_{t+h-2}]
\end{aligned}$$

From the last line it is easy to see that

$$\gamma_X(h) = \gamma_X(t, t+h) = \begin{cases} \sigma^2(b^2 + c^2), & \text{if } h = 0 \text{ (the variance)} \\ bc\sigma^2, & \text{if } |h| = 2 \\ 0, & \text{otherwise.} \end{cases}$$

(ii) $X_t = Z_0 \cos(ct); \quad t \geq 0$

[5 marks]

Mean:

$$\mathbb{E}[X_t] = \mathbb{E}[Z_0 \cos(ct)] = \mathbb{E}[Z_0] \cos(ct) = 0$$

Variance:

$$\mathrm{Var}[X_t] = \mathbb{E}[X_t^2] - \mathbb{E}[X_t]^2 = \mathbb{E}[Z_0^2] \cos(ct)^2 - 0 = \sigma^2 \cos(ct)^2.$$

Covariance:

$$\begin{aligned} \mathrm{Cov}[X_t, X_{t+h}] &= \mathbb{E}[Z_0^2 \cos(ct) \cos(ct + ch)] \\ &= \sigma^2 \cos(ct)(\cos(ct) \cos(ch) - \sin(ct) \sin(ch)) \end{aligned}$$

(iii) $X_t = Z_1 \cos(bt) + Z_2 \sin(bt); \quad t \geq 0$

[6 marks]

Mean:

$$\mathbb{E}[X_t] = \mathbb{E}[Z_1] \cos(bt) + \mathbb{E}[Z_2] \sin(bt) = 0$$

Variance:

$$\begin{aligned} \mathrm{Var}[X_t] &= \mathbb{E}[X_t^2] - \mathbb{E}[X_t] = \mathbb{E}[(Z_1 \cos(bt) + Z_2 \sin(bt))(Z_1 \cos(bt) + Z_2 \sin(bt))] - 0 \\ &= \mathbb{E}[Z_1^2 \cos(bt)^2 + 2 Z_1 Z_2 \cos(bt) \sin(bt) + Z_2^2 \sin(bt)^2] \\ &= \mathbb{E}[Z_1^2] \cos(bt)^2 + \mathbb{E}[Z_1 Z_2] 2 \cos(bt) \sin(bt) + \mathbb{E}[Z_2^2] \sin(bt)^2 \\ &= \sigma^2 (\cos(bt)^2 + \sin(bt)^2) \\ &= \sigma^2 \end{aligned}$$

Autocovariance:

$$
\begin{aligned}
\mathrm{Cov}[X_t, X_{t+h}] &= \mathbb{E}[X_t X_{t+h}] - \mathbb{E}[X_t]\mathbb{E}[X_{t+h}] \\
&= \mathbb{E}[(Z_1 \cos(bt) + Z_2 \sin(bt))(Z_1 \cos(bt) + Z_2 \sin(bt))] - 0 \\
&= \mathbb{E}[Z_1^2] \cos(bt) \cos(bt + bh) + \mathbb{E}[Z_1 Z_2] \cos(bt) \sin(bt + bh) \\
&\quad + \mathbb{E}[Z_1 Z_2] \sin(bt) \cos(bt + bh) + \mathbb{E}[Z_2^2] \sin(bt) \sin(bt + bh) \\
&= \sigma^2 (\cos(bt) \cos(bt + bh) + \sin(bt) \sin(bt + bh))
\end{aligned}
$$

Now, note that

$$\cos(bt) \cos(bt + bh) = \cos(bt)(\cos(bt) \cos(bh) - \sin(bt) \sin(bh))$$

and

$$\sin(bt) \sin(bt + bh) = \sin(bt)(\sin(bt) \cos(bh) + \cos(bt) \sin(bh)),$$

so therefore the above result becomes

$$
\begin{aligned}
\sigma^2(\cos(bt) \cos(bt + bh) + \sin(bt) \sin(bt + bh)) &= \sigma^2 \Big(\cos(bt)^2 \cos(bh) - \cos(bt) \sin(bt) \sin(bh) \\
&\quad + \sin(bt)^2 \cos(bh) + \cos(bt) \sin(bt) \sin(bh) \Big) \\
&= \sigma^2 \cos(bh) \Big(\cos(bt)^2 + \sin(bt)^2 \Big) \\
&= \sigma^2 \cos(bh).
\end{aligned}
$$

(c) Which, if any, of the series from (b) are stationary? Briefly justify using your answer from (a).

[3 marks]

The series (i) is clearly stationary. Its mean, variance and covariances do not depend on t.

The series (ii) is clearly non-stationary. Its mean is constant, but its variance and covariances depend on t.

The series (iii) is stationary. Its mean and variance are constant, and its autocovariance depends only on the time difference h between two points.

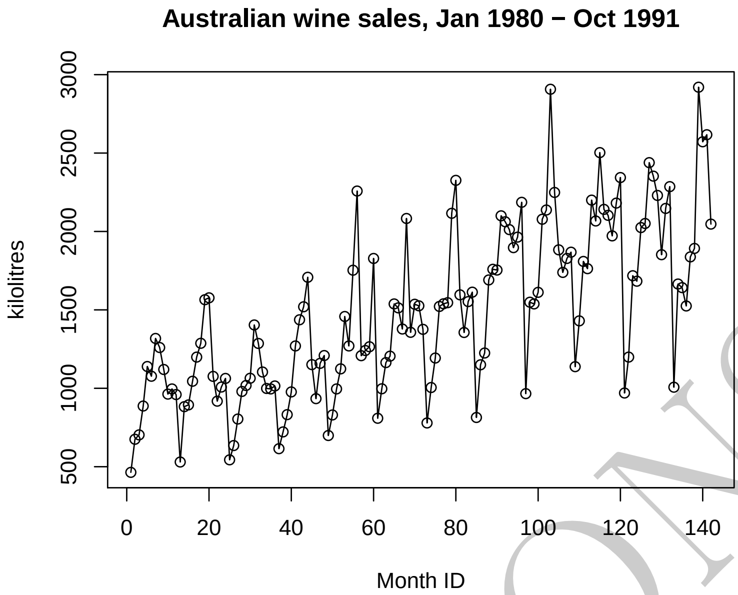

Figure 1: Aussie wine sales data.

Question 2: 10 marks

(a) Plot the time series, using appropriate axis labels/title, etc. Attach a copy to your assignment.

[2 marks]

Code is below, plot is in Figure 1.

```
> plot(wine,type="o",xlab="Month ID",ylab="kilolitres",
    main="Australian wine sales, Jan 1980 - Oct 1991")
```

(b) Suppose we wish to classically decompose these data. Would you recommend an additive or multiplicative decomposition model? Justify your answer.

[2 marks]

One of the key things that points to the need to use a multiplicative over an additive model is a perception that the amplitude of the seasonal oscillation is larger when the overall mean level of the series is higher. By inspecting Figure 1, this seems to be the case for the wine data. I would therefore recommend a multiplicative model.

(c) By commenting on your plot, suggest a reasonable period d of the seasonal component.

[2 marks]

We are looking for an approximate window of time such that the series appears to 'repeat' the same behaviour. Again examining Figure 1, we note that $d = 12$ is an appropriate choice. This is also intuitively sensible: we would expect there to be similar annual fluctuations in the sales of wine.

(d) Classically decompose your series using `classic.decomp` and your answers from (b) and (c). Use a simple linear model for your refined trend $\hat{m}$. Plot the returned object, and attach. **[2 marks]**

Code is below, plot is in Figure 2.

```
> wine.decomp <- classic.decomp(wine,12,"multiplicative",1)
> plot(wine.decomp)
```

(e) Write down your refined deseasonalised trend equation $\hat{m}$. Do you think this is an appropriate trend model to have fitted, or should we have used a higher-order polynomial? **[2 marks]**

```
> wine.decomp$m.hat

Call:
lm(formula = fm)

Coefficients:
(Intercept)   I(times^1)
    825.302        9.099
```

The fitted linear model is $\hat{m}_t = 825.302 + 9.099t$. From inspecting the deseasonalised data in Figure 2, it seems a linear model is appropriate. Little if any increase in the quality of the fit would be gained by using a higher-order polynomial.

Question 3: 10 marks

(a) Suppose we have observed a very small dataset comprised of the $N = 5$ values $x_1 = 1.182$, $x_2 = -4.132$, $x_3 = -0.669$, $x_4 = -0.157$ and $x_5 = -3.093$. Using R or by hand,

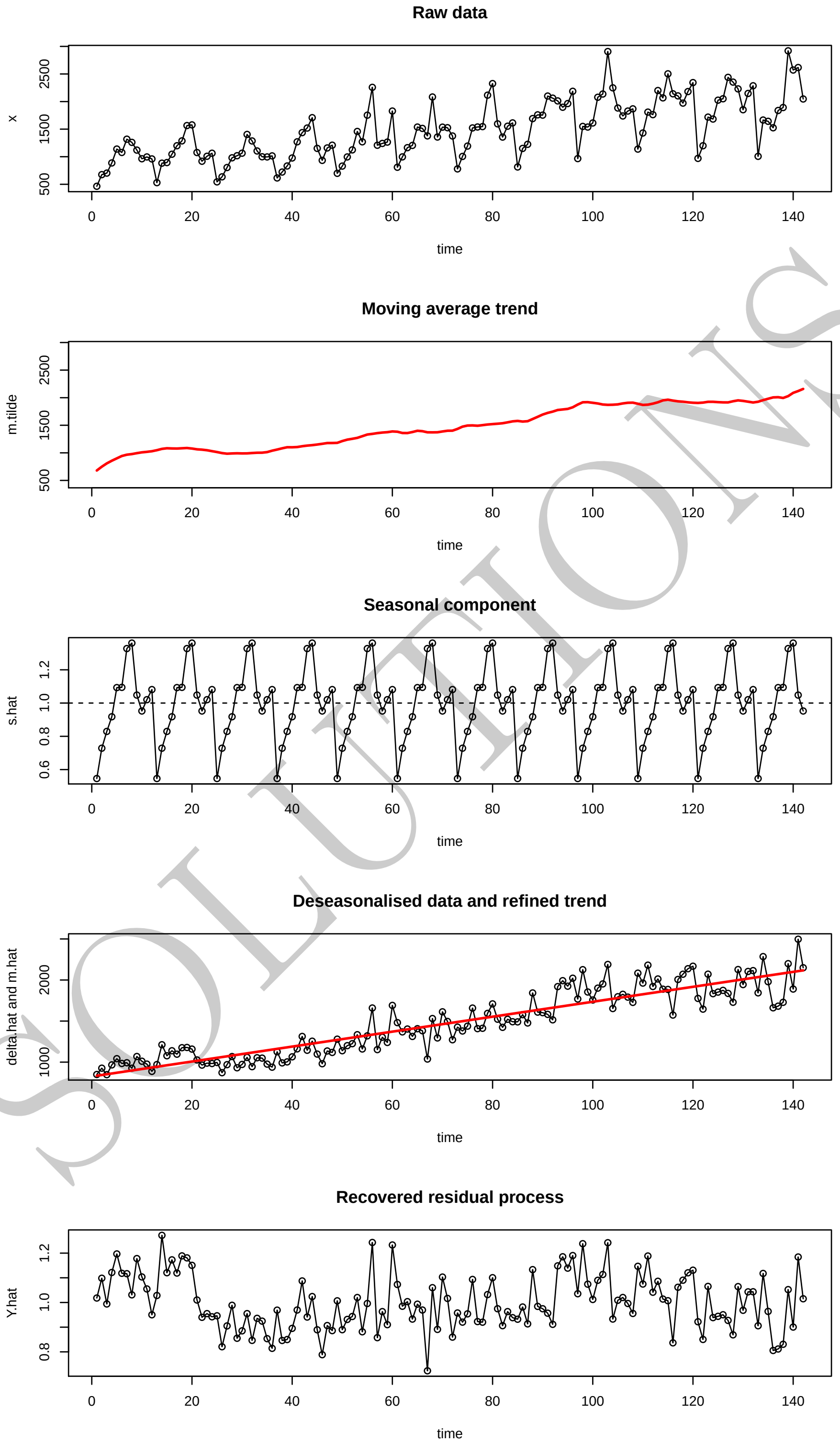

Figure 2: Classical decomposition of the Aussie wine sales data.

compute the sample autocorrelation $\hat{\rho}(h)$ for lags $h = 0$, $h = 1$ and $h = 2$.

[3 marks]

We must first compute the sample autocovariance function $\hat{\gamma}(h)$, since $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$. See Lecture 3.

Following the formulae in Lecture 3, you should be able to find that

$$\hat{\gamma}(0) = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2,$$

$$\hat{\gamma}(1) = \frac{1}{5}\sum_{i=1}^{4}(x_{i+1} - \bar{x})(x_i - \bar{x}),$$

and

$$\hat{\gamma}(1) = \frac{1}{5}\sum_{i=1}^{3}(x_{i+2} - \bar{x})(x_i - \bar{x}),$$

where $\bar{x}$ is simply the mean of the 5 observations above: $\bar{x} = (1/5) \times \sum_i x_i = -1.3738$. For the 5 observations, using the expressions above, you should have found that

$$\hat{\gamma}(0) = 3.8145550, \quad \hat{\gamma}(1) = -2.0455418, \quad \hat{\gamma}(2) = -0.5533084,$$

and following directly from those results, that

$$\hat{\rho}(0) = 1, \quad \hat{\rho}(1) = -0.5362465, \quad \hat{\rho}(2) = -0.14505194.$$

(b) Lets say you were asked to compute the autocorrelation for $h = 5$. Why isn't this possible for this data set?

[1 mark]

Computation of the autocorrelation (and autocovariance) depends on measuring differences between observations that are h units apart. Given that the maximum time is 5, and the minimum time is 1, the largest time difference in the dataset is $5 - 1 = 4$. So it is impossible to compute these figures for this dataset, because the maximum lag for which we have data available to us is $h = 4$.

(c) Using the maximum possible lag for this dataset, find the Ljung-Box test statistic.

[3 marks]

The maximum possible lag is $h = 4$. To complete this question, we are required to first find the values of $\hat{\rho}(3)$ and $\hat{\rho}(4)$. Using the same procedure as above, you should easily be able to find that

$$\hat{\rho}(3) = 0.4116755, \quad \hat{\rho}(4) = -0.2303771.$$

From Lecture 3, the Ljung-Box test statistic is

$$Q = N(N+2)\sum_{k=1}^{h} \frac{\hat{\rho}(k)^2}{N-k};$$

and for this question you should be using $N = 5$ and we already decided on $h = 4$ (the maximum possible lag) above. Having already found $\hat{\rho}(0), \hat{\rho}(1), \hat{\rho}(2), \hat{\rho}(3)$ and $\hat{\rho}(4)$, your Ljung-Box test statistic should easily found as $Q = 7.58504$

(d) The Ljung-Box test statistic follows a Chi-squared distribution. How many degrees of freedom we will be using for this test?

[1 mark]

The Ljung-Box test based on some lag h has h degrees of freedom. So we will be using 4 degrees of freedom.

(e) In the `R` language, we may find a p-value from the Chi-squared distribution with the built-in function `pchisq`. Suppose our test statistic is `Q` and our degrees of freedom is `dof`. Your p-value will be given with the command `pchisq(Q,df=dof,lower.tail=F)`. Find the p-value for your test, and draw an appropriate conclusion with respect to the data and using a significance level of 5%.

[2 marks]

```
> pchisq(Q,df=4,lower.tail=F)
[1] 0.1080173
```

Our p-value is 0.108, which is not statistically significant. Since the Ljung-Box test is a test based on independent (i.e. uncorrelated) terms in a time series, our p-value indicates insufficient evidence to reject the null hypothesis of independent terms. In other words, there is no evidence to support the notion that there exists correlation in the small dataset we have examined here.

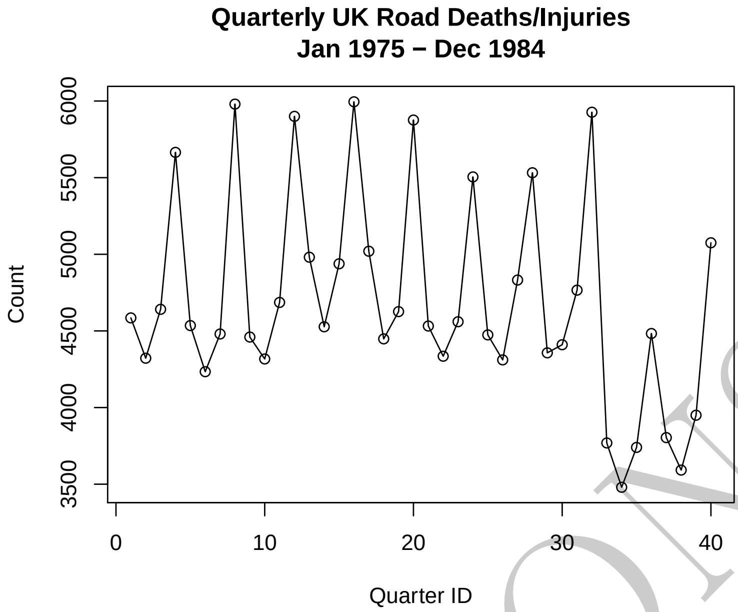

Figure 3: UK road death/injury data.

Question 4: 10 marks

(a) Plot the time series, using appropriate axis labels/title, etc. Attach a copy to your assignment. Note that seatbelts became law in the UK in the first quarter of 1983 (observation 33 onwards). What are your impressions of the data? Does the seatbelt legislation appear to have had an effect on the number of UK road deaths/injuries?

[2 marks]

Code is below, plot is in Figure 3.

```
> plot(seat,type="o",xlab="Quarter ID",ylab="Count",
    main="Quarterly UK Road Deaths/Injuries\nJan 1975 - Dec 1984")
```

There does seem to be a repeating pattern, with the overall mean level of deaths/injuries more or less constant until observation 33. It would seem that there was indeed a sharp drop in the overall level of deaths/injuries following the introduction of the seatbelt legislation.

(b) Suppose we wish to classically decompose these data. Would you recommend an additive or multiplicative decomposition model? Justify your answer.

[2 marks]

There does not appear to be any need for a multiplicative decomposition of these data. At least visually, the variation seems to remain at a constant level over the time for which we have data. I'd say additive is the way to go.

(c) By commenting on your plot, suggest a reasonable period d of the seasonal component.

[2 marks]

The fact that we noted a repeating pattern in Part (a), coupled with the fact that the data are quarterly observations, would suggest $d = 4$ as a suitable seasonal period. A closer inspection of the time plot reveals this to make sense: every four observation a similar pattern occurs. Note that the 'peak' always happens in the fourth quarter of the year, understandable since this includes the Christmas vacation. No wonder the cops always crack down on road safety over public holidays!

(d) Classically decompose your series using `classic.decomp` and your answers from (b) and (c). Use a simple linear model for your refined trend $\hat{m}$. Plot the returned object, and attach.

[2 marks]

Code is below, plot is in Figure 4.

```
> seat.decomp <- classic.decomp(seat,4,"additive",1)
> plot(seat.decomp)
```

(e) Write down your refined deseasonalised trend equation $\hat{m}$. Do you think this is an appropriate trend model to have fitted, or should we have used a higher-order polynomial?

[2 marks]

```
> seat.decomp$m.hat

Call:
lm(formula = fm)

Coefficients:
(Intercept)   I(times^1)
    5124.32       -21.13
```

The fitted linear model is $\hat{m}_t = 5124.32 - 21.13t$. From inspecting the deseasonalised data in Figure 4, it seems a linear model may not be appropriate. We are still missing that 'dip'

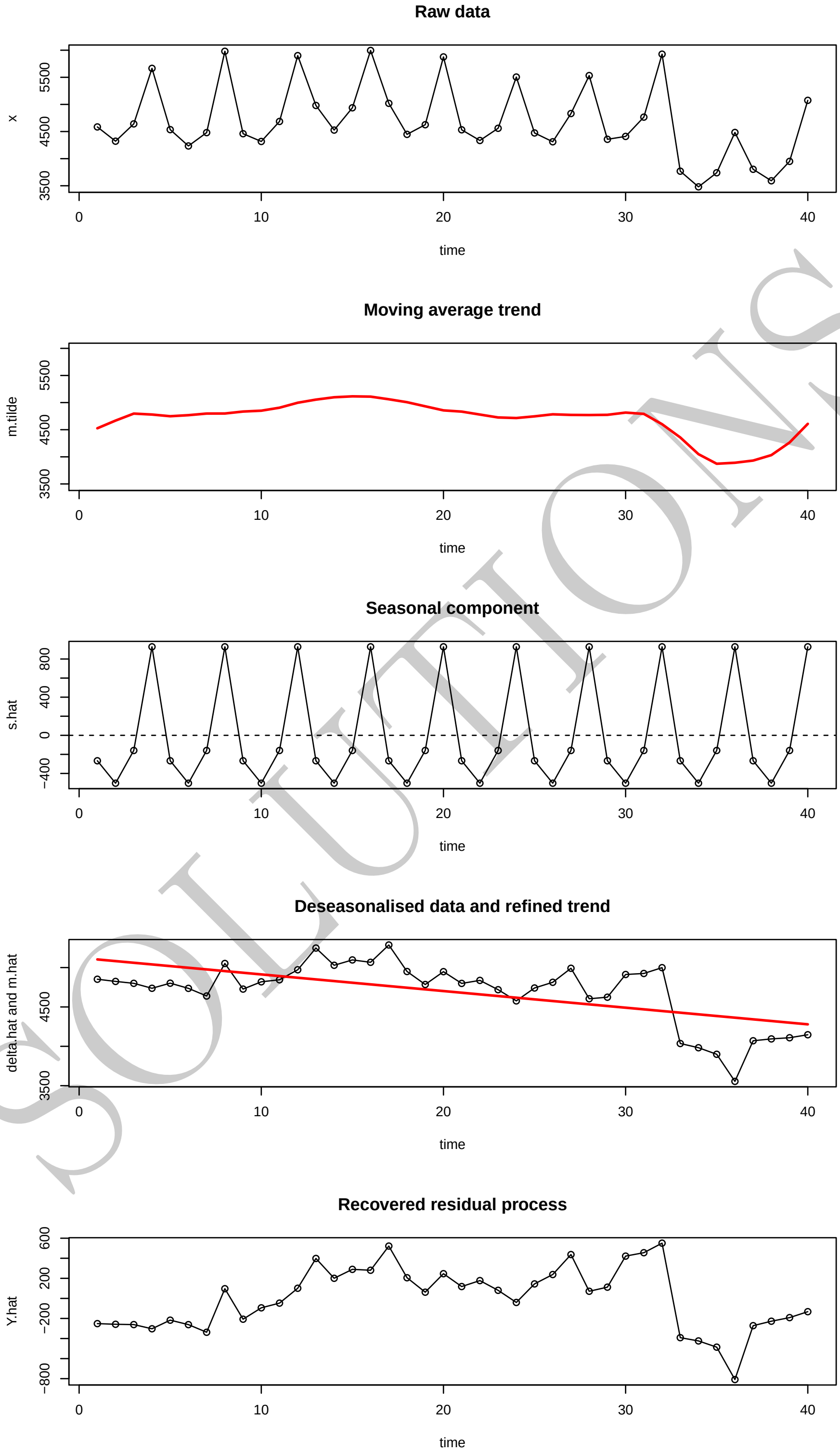

Figure 4: Classical decomposition of the UK road death/injury data.

after observation 33, and our other decomposition operations appear to have left a very slow 'bump' occurring as the recovered residual error series shows a slight but continuing increase after around observation 10, before declining again later on. It is possible a quadratic trend may be better suited to the deseasonalised data.