An automatic report for the dataset : 07-call-centre

The Automatic Statistician

Abstract

This report was produced automatically by the Gaussian process structure search algorithm. See http://arxiv.org/abs/1302.4922 and http://www-kd.iai.uni-bonn.de/cml/proceedings/papers/2.pdf for preliminary papers.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

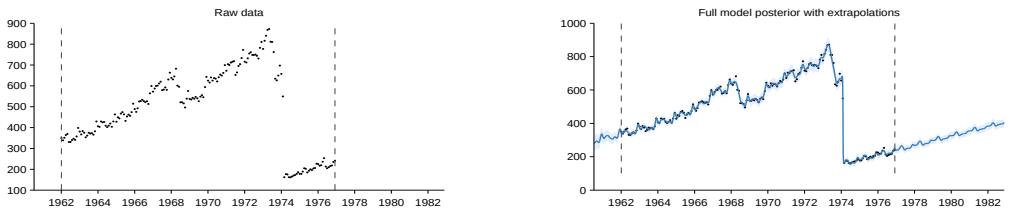

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified six additive components in the data. The first 2 additive components explain 94.6% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.2% of the variation in the data. After the first 4 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A very smooth function. This function applies until Feb 1974.
- A linearly increasing function. This function applies from Feb 1974 onwards.
- A smooth function with marginal standard deviation increasing linearly away from Sep 1962. This function applies until Feb 1974.
- An approximately periodic function with a period of 1.0 years.
- Uncorrelated noise.
- Uncorrelated noise with standard deviation increasing linearly away from Nov 1963. This function applies until Feb 1974.

Model checking statistics are summarised in table 2 in section 4. These statistics have not revealed any inconsistencies between the model and observed data.

The rest of the document is structured as follows. In section 2 the forms of the additive components are described and their posterior distributions are displayed. In section 3 the modelling assumptions

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	492.50	-
1	78.0	78.0	78.0	68.97	86.0
2	94.6	16.6	75.6	34.24	50.4
3	99.2	4.5	84.3	25.27	26.2
4	99.7	0.6	66.5	22.83	9.6
5	99.9	0.2	63.0	22.83	0.0
6	100.0	0.1	100.0	22.83	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

of each component are discussed with reference to how this affects the extrapolations made by the model. Section 4 discusses model checking statistics, with plots showing the form of any detected discrepancies between the model and observed data.

2 Detailed discussion of additive components

2.1 Component 1 : A very smooth function. This function applies until Feb 1974

This component is a very smooth function. This component applies until Feb 1974.

This component explains 78.0% of the total variance. The addition of this component reduces the cross validated MAE by 86.0% from 492.5 to 69.0.

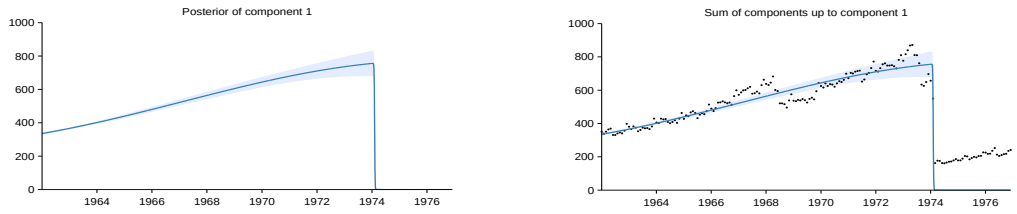

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

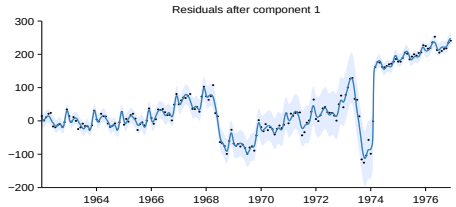

Figure 3: Pointwise posterior of residuals after adding component 1

2.2 Component 2 : A linearly increasing function. This function applies from Feb 1974 onwards

This component is linearly increasing. This component applies from Feb 1974 onwards.

This component explains 75.6% of the residual variance; this increases the total variance explained from 78.0% to 94.6%. The addition of this component reduces the cross validated MAE by 50.35% from 68.97 to 34.24.

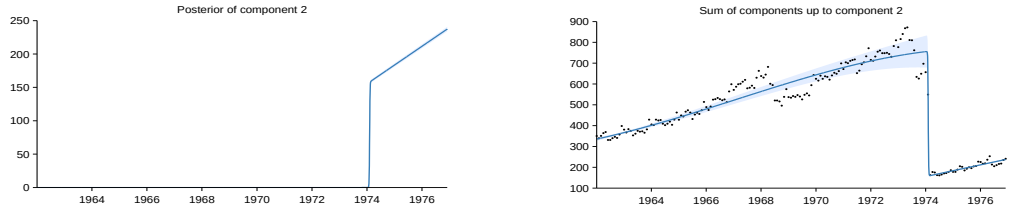

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

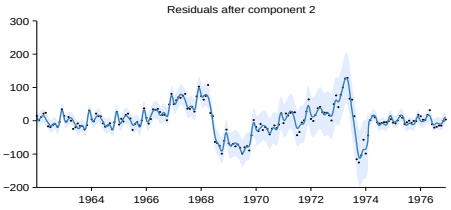

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A smooth function with marginal standard deviation increasing linearly away from Sep 1962. This function applies until Feb 1974

This component is a smooth function with a typical lengthscale of 3.6 months. The marginal standard deviation of the function increases linearly away from Sep 1962. This component applies until Feb 1974.

This component explains 84.3% of the residual variance; this increases the total variance explained from 94.6% to 99.2%. The addition of this component reduces the cross validated MAE by 26.20% from 34.24 to 25.27.

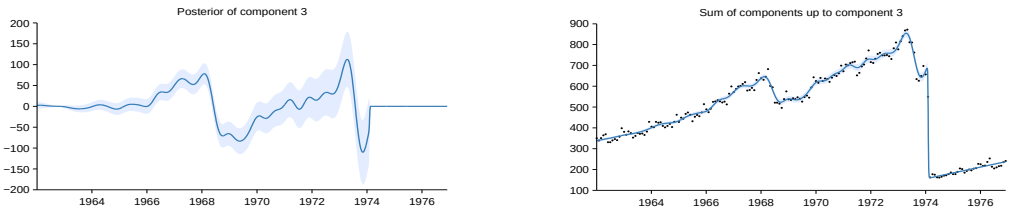

Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

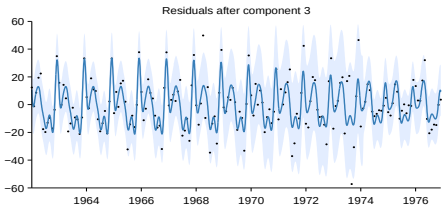

Figure 7: Pointwise posterior of residuals after adding component 3

2.4 Component 4 : An approximately periodic function with a period of 1.0 years

This component is approximately periodic with a period of 1.0 years. Across periods the shape of this function varies very smoothly. The shape of this function within each period has a typical lengthscale of 4.1 weeks.

This component explains 66.5% of the residual variance; this increases the total variance explained from 99.2% to 99.7%. The addition of this component reduces the cross validated MAE by 9.65% from 25.27 to 22.83.

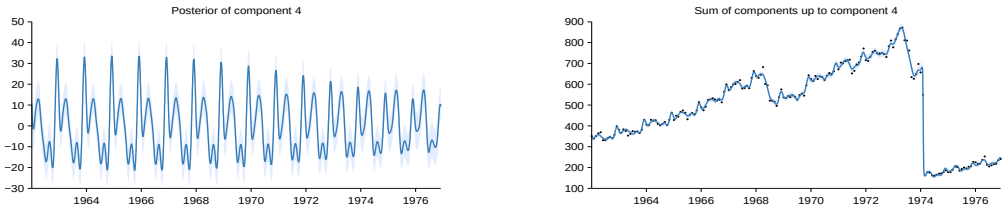

Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

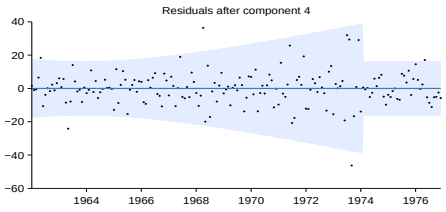

Figure 9: Pointwise posterior of residuals after adding component 4

2.5 Component 5 : Uncorrelated noise

This component models uncorrelated noise.

This component explains 63.0% of the residual variance; this increases the total variance explained from 99.7% to 99.9%. The addition of this component reduces the cross validated MAE by 0.00% from 22.83 to 22.83. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

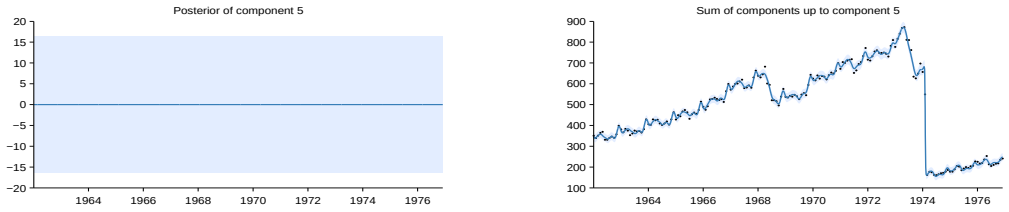

Figure 10: Pointwise posterior of component 5 (left) and the posterior of the cumulative sum of components with data (right)

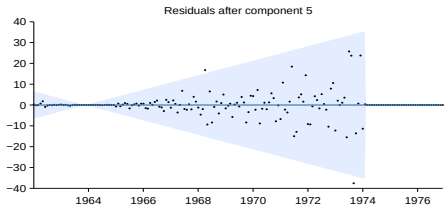

Figure 11: Pointwise posterior of residuals after adding component 5

2.6 Component 6 : Uncorrelated noise with standard deviation increasing linearly away from Nov 1963. This function applies until Feb 1974

This component models uncorrelated noise. The standard deviation of the noise increases linearly away from Nov 1963. This component applies until Feb 1974.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.9% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 22.83 to 22.83. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

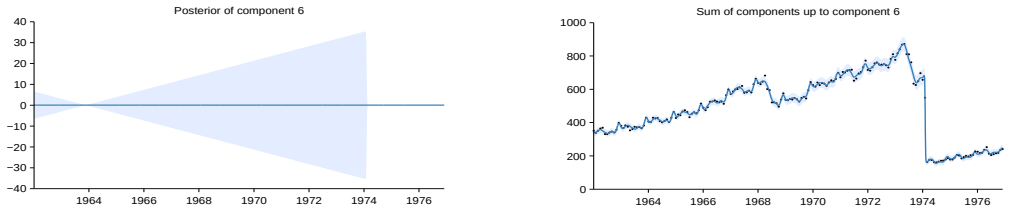

Figure 12: Pointwise posterior of component 6 (left) and the posterior of the cumulative sum of components with data (right)

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 13. The plot on the left displays the mean of the posterior together with pointwise variance. The plot on the right displays three random samples from the posterior.

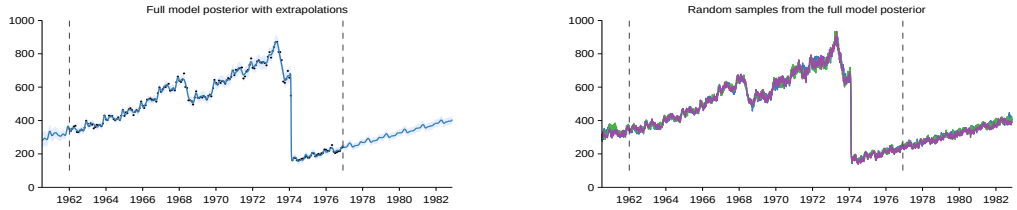

Figure 13: Full model posterior with extrapolation. Mean and pointwise variance (left) and three random samples (right)

Below are descriptions of the modelling assumptions associated with each additive component and how they affect the predictive posterior. Plots of the pointwise posterior and samples from the posterior are also presented, showing extrapolations from each component and the cuulative sum of components.

3.1 Component 1 : A very smooth function. This function applies until Feb 1974

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

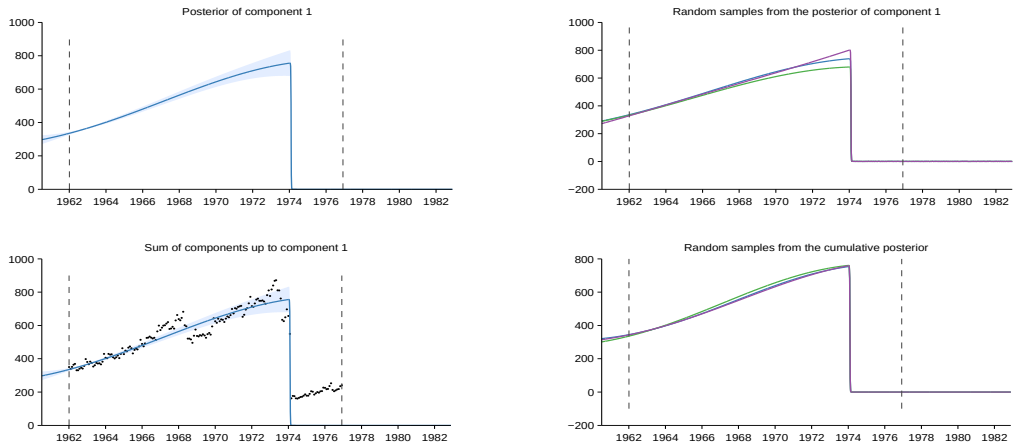

Figure 14: Posterior of component 1 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.2 Component 2 : A linearly increasing function. This function applies from Feb 1974 onwards

This component is assumed to continue to increase linearly.

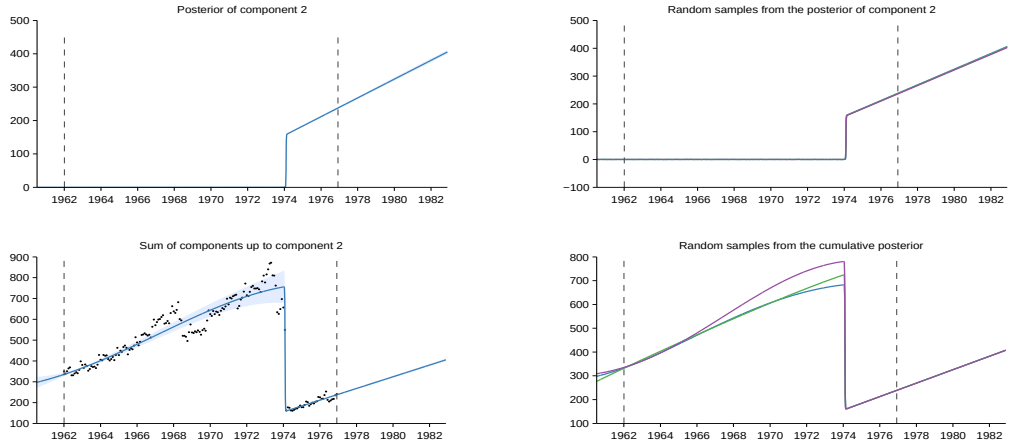

Figure 15: Posterior of component 2 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.3 Component 3 : A smooth function with marginal standard deviation increasing linearly away from Sep 1962. This function applies until Feb 1974

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

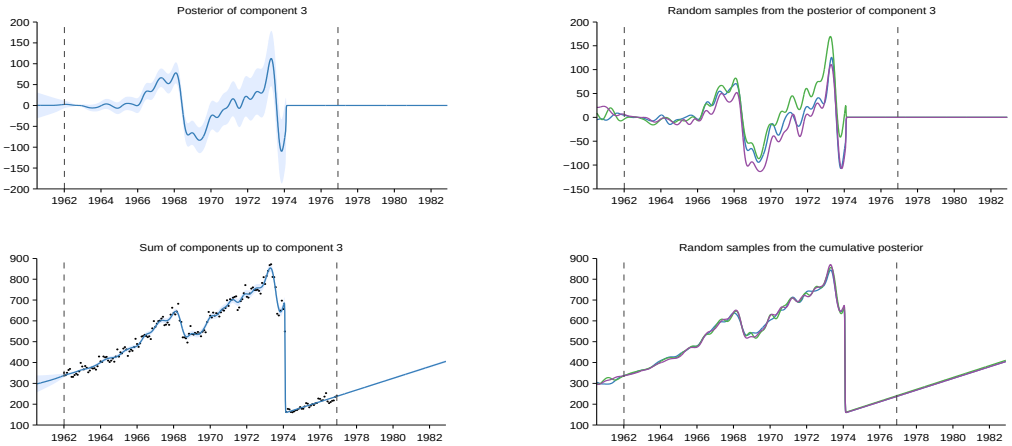

Figure 16: Posterior of component 3 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.4 Component 4 : An approximately periodic function with a period of 1.0 years

This component is assumed to continue to be approximately periodic. The shape of the function is assumed to vary very smoothly between periods but will eventually return to the prior. The prior is entirely uncertain about the phase of the periodic function. Consequently the pointwise posterior will appear to lose its periodicity, but this merely reflects the uncertainty in the shape and phase of the function. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

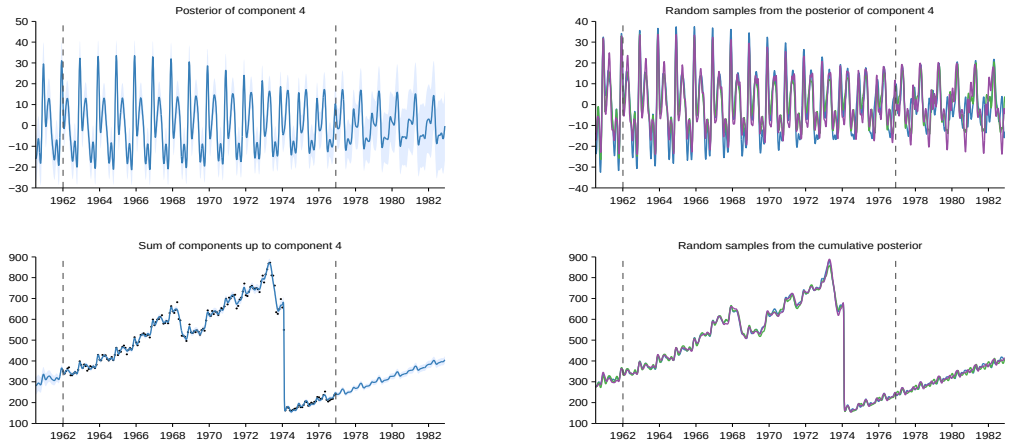

Figure 17: Posterior of component 4 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.5 Component 5 : Uncorrelated noise

This component assumes the uncorrelated noise will continue indefinitely.

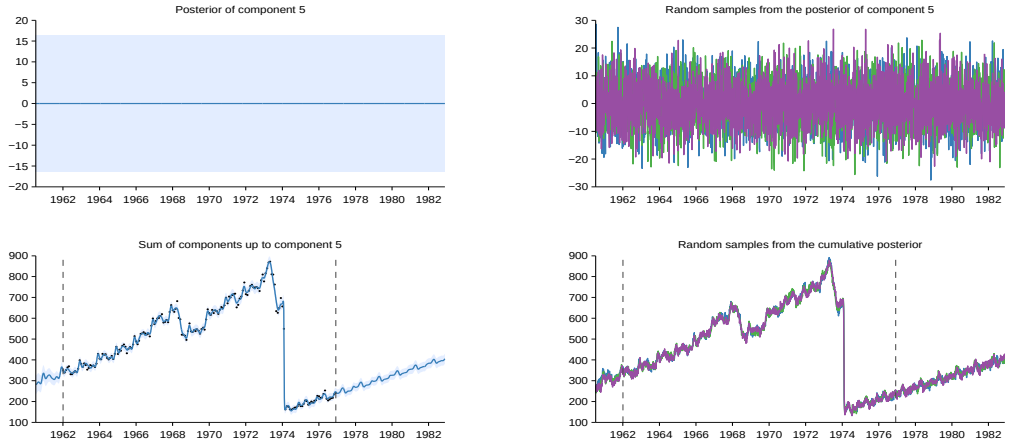

Figure 18: Posterior of component 5 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.6 Component 6 : Uncorrelated noise with standard deviation increasing linearly away from Nov 1963. This function applies until Feb 1974

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

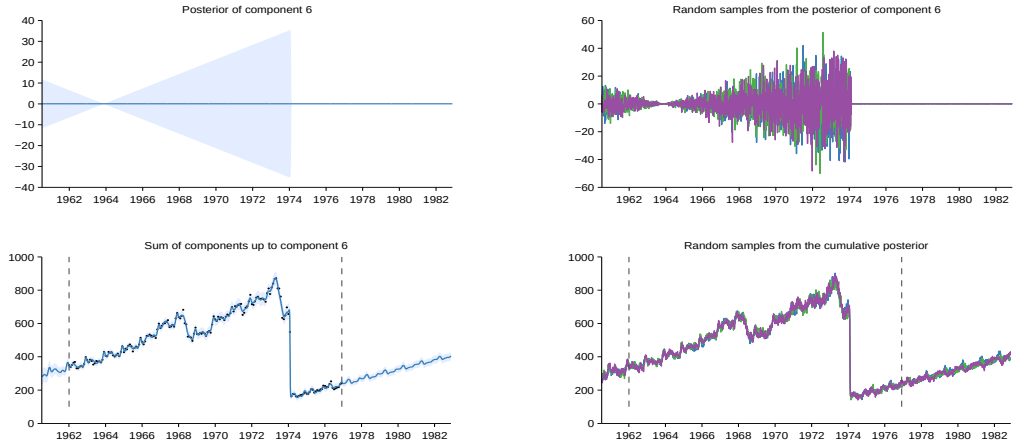

Figure 19: Posterior of component 6 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

4 Model checking

Several posterior predictive checks have been performed to assess how well the model describes the observed data. These tests take the form of comparing statistics evaluated on samples from the prior and posterior distributions for each additive component. The statistics are derived from autocorrelation function (ACF) estimates, periodograms and quantile-quantile (qq) plots.

Table 2 displays cumulative probability and p-value estimates for these quantities. Cumulative probabilities near 0/1 indicate that the test statistic was lower/higher under the posterior compared to the prior unexpectedly often i.e. they contain the same information as a p-value for a two-tailed test and they also express if the test statistic was higher or lower than expected. p-values near 0 indicate that the test statistic was larger in magnitude under the posterior compared to the prior unexpectedly often.

	ACF		Periodogram		QQ	
#	min	min loc	max	max loc	max	min
1	0.800	0.074	0.800	0.349	0.144	0.911
2	0.502	0.461	0.451	0.506	0.188	0.722
3	0.636	0.593	0.792	0.363	0.586	0.617
4	0.794	0.579	0.490	0.327	0.355	0.909
5	0.512	0.515	0.507	0.524	0.402	0.337
6	0.501	0.465	0.513	0.496	0.450	0.351

Table 2: Model checking statistics for each component. Cumulative probabilities for minimum of autocorrelation function (ACF) and its location. Cumulative probabilities for maximum of periodogram and its location. p-values for maximum and minimum deviations of QQ-plot from straight line.

No statistically significant discrepancies between the data and model have been detected but model checking plots for each component are presented below.

4.1 Model checking plots for components without statistically significant discrepancies

4.1.1 Component 1 : A very smooth function. This function applies until Feb 1974

No discrepancies between the prior and posterior of this component have been detected

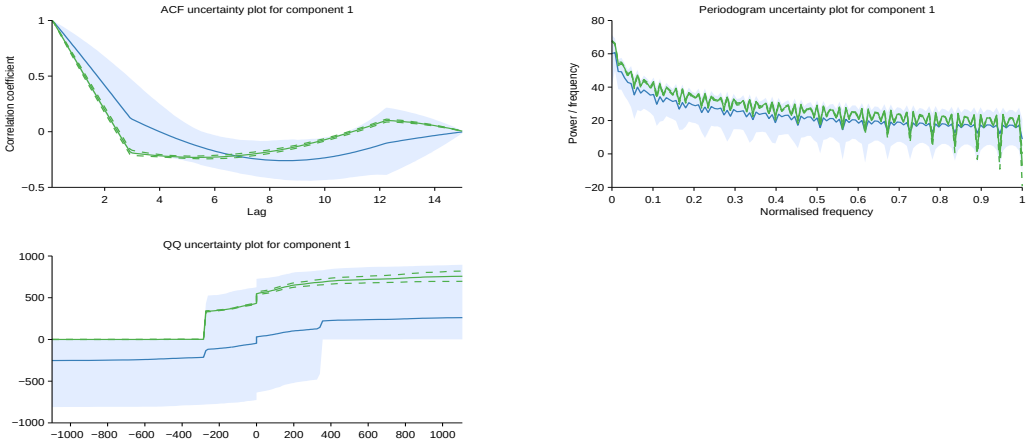

Figure 20: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.1.2 Component 2 : A linearly increasing function. This function applies from Feb 1974 onwards

No discrepancies between the prior and posterior of this component have been detected

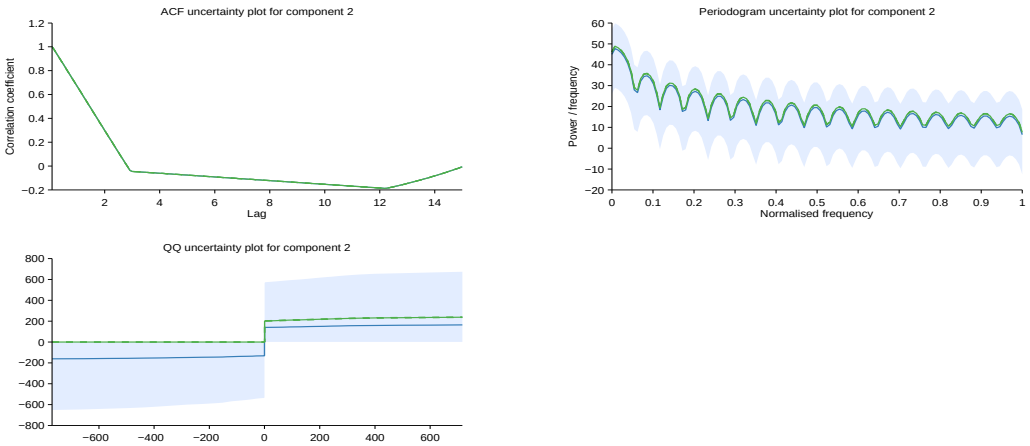

Figure 21: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 2. The green line and green dashed lines are the corresponding quantities under the posterior.

4.1.3 Component 3 : A smooth function with marginal standard deviation increasing linearly away from Sep 1962. This function applies until Feb 1974

No discrepancies between the prior and posterior of this component have been detected

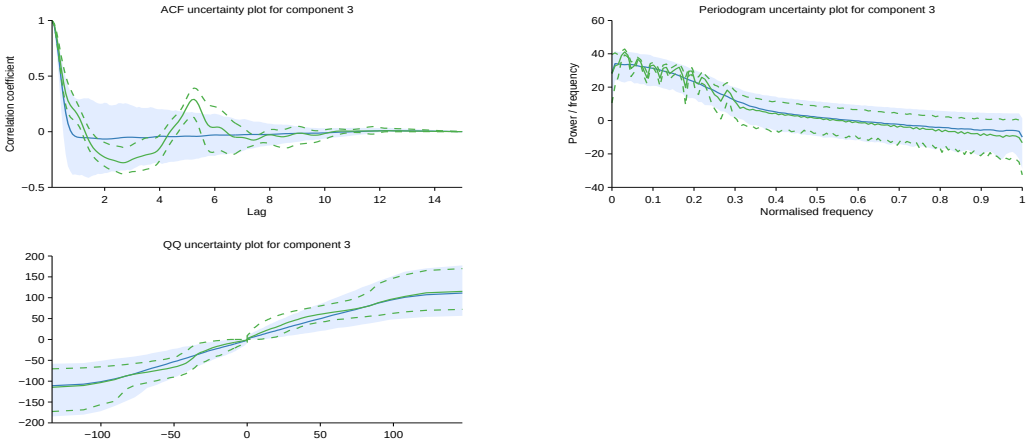

Figure 22: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 3. The green line and green dashed lines are the corresponding quantities under the posterior.

4.1.4 Component 4 : An approximately periodic function with a period of 1.0 years

No discrepancies between the prior and posterior of this component have been detected

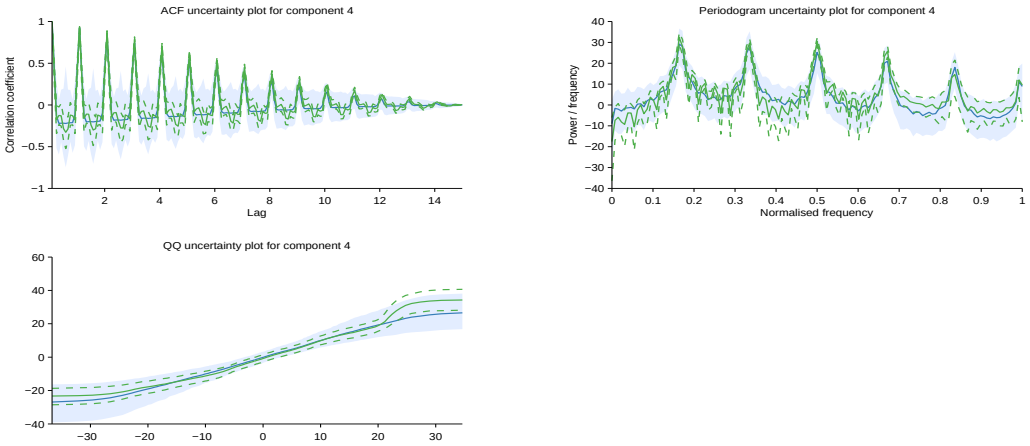

Figure 23: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 4. The green line and green dashed lines are the corresponding quantities under the posterior.

4.1.5 Component 5 : Uncorrelated noise

No discrepancies between the prior and posterior of this component have been detected

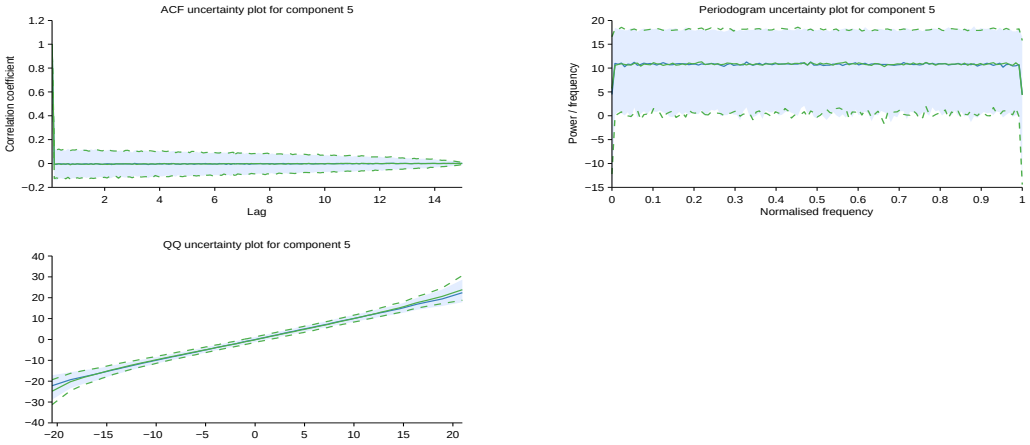

Figure 24: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 5. The green line and green dashed lines are the corresponding quantities under the posterior.

4.1.6 Component 6 : Uncorrelated noise with standard deviation increasing linearly away from Nov 1963. This function applies until Feb 1974

No discrepancies between the prior and posterior of this component have been detected

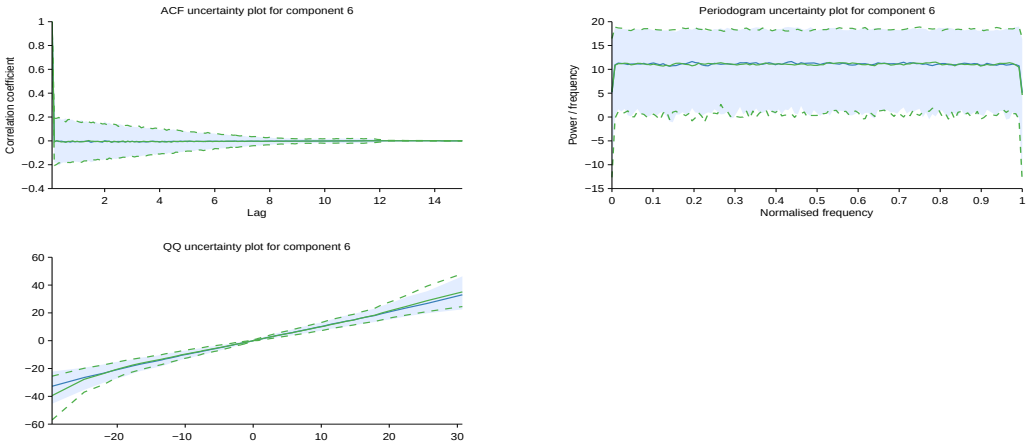

Figure 25: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 6. The green line and green dashed lines are the corresponding quantities under the posterior.