An automatic report for the dataset: 04-wheat

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

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

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.1% 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 99.3% of the variation in the data. After the first 3 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 smooth function.
- A rapidly varying smooth function with linearly increasing marginal standard deviation. This function applies until 1714 and from 1794 onwards.
- A rapidly varying smooth function. This function applies from 1714 until 1794.
- Uncorrelated noise with linearly increasing standard deviation. This function applies until 1714 and from 1794 onwards.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed statistically significant discrepancies between the data and model in component 3.

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 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.
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.

2 Detailed discussion of additive components

2.1 Component 1: A smooth function

This component is a smooth function with a typical lengthscale of 95.8 years.

This component explains 81.5% of the total variance. The addition of this component reduces the cross validated MAE by 80.6% from 107.9 to 21.0.

2.2 Component 2: A rapidly varying smooth function with linearly increasing marginal standard deviation. This function applies until 1714 and from 1794 onwards

This component is a rapidly varying but smooth function with a typical lengthscale of 1.4 years. The marginal standard deviation of the function increases linearly. This component applies until 1714 and from 1794 onwards.

This component explains 89.9% of the residual variance; this increases the total variance explained from 81.5% to 98.1%. The addition of this component reduces the cross validated MAE by 1.67% from 20.97 to 20.62.
2.3 Component 3: A rapidly varying smooth function. This function applies from 1714 until 1794

This component is a rapidly varying but smooth function with a typical lengthscale of 1.0 years. This component applies from 1714 until 1794.

This component explains 64.7% of the residual variance; this increases the total variance explained from 98.1% to 99.3%. The addition of this component reduces the cross validated MAE by 0.25% from 20.62 to 20.57.
2.4 Component 4: Uncorrelated noise with linearly increasing standard deviation. This function applies until 1714 and from 1794 onwards

This component models uncorrelated noise. The standard deviation of the noise increases linearly. This component applies until 1714 and from 1794 onwards.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.3% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 20.57 to 20.57. 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 8: Pointwise posterior of component 4 (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 9. 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 9: 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 cumulative sum of components.

3.1 Component 1: A smooth function

This component is assumed to continue smoothly but is also assumed to be stationary so its distribution will return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 95.8 years. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].
3.2 Component 2: A rapidly varying smooth function with linearly increasing marginal standard deviation. This function applies until 1714 and from 1794 onwards.

This component is assumed to continue smoothly but its distribution is assumed to quickly return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 1.4 years. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior]. The marginal standard deviation of the function is assumed to continue to increase linearly.

3.3 Component 3: A rapidly varying smooth function. This function applies from 1714 until 1794.

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.
3.4 Component 4: Uncorrelated noise with linearly increasing standard deviation. This function applies until 1714 and from 1794 onwards

This component assumes the uncorrelated noise will continue indefinitely. The standard deviation of the noise is assumed to continue to increase linearly.

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.

<table>
<thead>
<tr>
<th>#</th>
<th>ACF min</th>
<th>min loc</th>
<th>Periodogram max</th>
<th>max loc</th>
<th>QQ max</th>
<th>min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.613</td>
<td>0.946</td>
<td>0.674</td>
<td>0.175</td>
<td>0.072</td>
<td>0.879</td>
</tr>
<tr>
<td>2</td>
<td>0.300</td>
<td>0.629</td>
<td>0.904</td>
<td>0.170</td>
<td>0.140</td>
<td>0.896</td>
</tr>
<tr>
<td>3</td>
<td>0.681</td>
<td>0.406</td>
<td>0.438</td>
<td>0.451</td>
<td>0.041</td>
<td>0.461</td>
</tr>
<tr>
<td>4</td>
<td>0.517</td>
<td>0.533</td>
<td>0.585</td>
<td>0.603</td>
<td>0.363</td>
<td>0.699</td>
</tr>
</tbody>
</table>

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.

The nature of any observed discrepancies is now described and plotted and hypotheses are given for the patterns in the data that may not be captured by the model.

4.1 Moderately statistically significant discrepancies

4.1.1 Component 3: A rapidly varying smooth function. This function applies from 1714 until 1794

The following discrepancies between the prior and posterior distributions for this component have been detected.

- The qq plot has an unexpectedly large positive deviation from equality \( (x = y) \). This discrepancy has an estimated \( p \)-value of 0.041.

The positive deviation in the qq-plot can indicate heavy positive tails if it occurs at the right of the plot or light negative tails if it occurs as the left.

Figure 14: 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.
4.2 Model checking plots for components without statistically significant discrepancies

4.2.1 Component 1: A smooth function

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

Figure 15: 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.

4.2.2 Component 2: A rapidly varying smooth function with linearly increasing marginal standard deviation. This function applies until 1714 and from 1794 onwards

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

Figure 16: 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.
4.2.3 Component 4: Uncorrelated noise with linearly increasing standard deviation. This function applies until 1714 and from 1794 onwards

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

Figure 17: 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.