An automatic report for the dataset: 09-gas-production

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.

![Figure 1: Raw data (left) and model posterior with extrapolation (right)](image)

The structure search algorithm has identified seven additive components in the data. The first additive component explains 93.6% of the variation in the data as shown by the coefficient of determination ($R^2$) values in table 1. The first 4 additive components explain 99.6% of the variation in the data. After the first 5 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 linearly increasing function. This function applies from 1972 onwards.
- An approximately periodic function with a period of 1.0 years. This function applies from 1972 onwards.
- A very smooth monotonically increasing function. This function applies until 1972.
- A smooth function. This function applies from 1972 onwards.
- An approximately periodic function with a period of 1.0 years. This function applies until 1972.
- Uncorrelated noise.
- Uncorrelated noise with standard deviation increasing linearly away from 1962. This function applies from 1972 onwards.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed highly statistically significant discrepancies between the data and model in component 7. Moderate discrepancies have also been detected in components 1, 2 and 5.
<table>
<thead>
<tr>
<th>#</th>
<th>$R^2$ (%)</th>
<th>$\Delta R^2$ (%)</th>
<th>Residual $R^2$ (%)</th>
<th>Cross validated MAE</th>
<th>Reduction in MAE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>21415.27</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>93.6</td>
<td>93.6</td>
<td>93.6</td>
<td>3682.28</td>
<td>82.8</td>
</tr>
<tr>
<td>2</td>
<td>98.3</td>
<td>4.7</td>
<td>74.1</td>
<td>2383.98</td>
<td>35.3</td>
</tr>
<tr>
<td>3</td>
<td>98.5</td>
<td>0.2</td>
<td>11.2</td>
<td>1673.52</td>
<td>29.8</td>
</tr>
<tr>
<td>4</td>
<td>99.6</td>
<td>1.1</td>
<td>73.0</td>
<td>1376.65</td>
<td>17.7</td>
</tr>
<tr>
<td>5</td>
<td>99.6</td>
<td>0.0</td>
<td>4.4</td>
<td>1294.10</td>
<td>6.0</td>
</tr>
<tr>
<td>6</td>
<td>99.6</td>
<td>0.0</td>
<td>0.3</td>
<td>1294.10</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>100.0</td>
<td>0.4</td>
<td>100.0</td>
<td>1294.10</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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.

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.

## 2 Detailed discussion of additive components

### 2.1 Component 1: A linearly increasing function. This function applies from 1972 onwards

This component is linearly increasing. This component applies from 1972 onwards.

This component explains 93.6% of the total variance. The addition of this component reduces the cross validated MAE by 82.8% from 21415.3 to 3682.3.

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

![Figure 3: Pointwise posterior of residuals after adding component 1](image)
2.2 Component 2: An approximately periodic function with a period of 1.0 years. This function applies from 1972 onwards

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 6.7 weeks. This component applies from 1972 onwards.

This component explains 74.1% of the residual variance; this increases the total variance explained from 93.6% to 98.3%. The addition of this component reduces the cross validated MAE by 35.26% from 3682.28 to 2383.98.

![Image of Component 2](image1)

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

2.3 Component 3: A very smooth monotonically increasing function. This function applies until 1972

This component is a very smooth and monotonically increasing function. This component applies until 1972.

This component explains 11.2% of the residual variance; this increases the total variance explained from 98.3% to 98.5%. The addition of this component reduces the cross validated MAE by 29.81% from 2383.98 to 1673.32.

![Image of Component 3](image2)

Figure 5: Pointwise posterior of residuals after adding component 2

Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)
2.4 Component 4: A smooth function. This function applies from 1972 onwards

This component is a smooth function with a typical lengthscale of 0.9 years. This component applies from 1972 onwards.

This component explains 73.0% of the residual variance; this increases the total variance explained from 98.5% to 99.6%. The addition of this component reduces the cross validated MAE by 17.73% from 1673.32 to 1376.65.

2.5 Component 5: An approximately periodic function with a period of 1.0 years. This function applies until 1972

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 5.9 weeks. This component applies until 1972.

This component explains 4.4% of the residual variance; this increases the total variance explained from 99.6% to 99.6%. The addition of this component reduces the cross validated MAE by 6.00% from 1376.65 to 1294.10.
2.6 Component 6: Uncorrelated noise

This component models uncorrelated noise.

This component explains 0.3% of the residual variance; this increases the total variance explained from 99.6% to 99.6%. The addition of this component reduces the cross validated MAE by 0.00% from 1294.10 to 1294.10. 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.
2.7 Component 7: Uncorrelated noise with standard deviation increasing linearly away from 1962. This function applies from 1972 onwards

This component models uncorrelated noise. The standard deviation of the noise increases linearly away from 1962. This component applies from 1972 onwards.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.6% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 1294.10 to 1294.10. 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.

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 15. 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.

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 linearly increasing function. This function applies from 1972 onwards

This component is assumed to continue to increase linearly.
3.2 Component 2: An approximately periodic function with a period of 1.0 years. This function applies from 1972 onwards

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].

3.3 Component 3: A very smooth monotonically increasing function. This function applies until 1972

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.
3.4 Component 4: A smooth function. This function applies from 1972 onwards

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 0.9 years. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

3.5 Component 5: An approximately periodic function with a period of 1.0 years. This function applies until 1972

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.
3.6 Component 6: Uncorrelated noise

This component assumes the uncorrelated noise will continue indefinitely.

3.7 Component 7: Uncorrelated noise with standard deviation increasing linearly away from 1962. This function applies from 1972 onwards

This component assumes the uncorrelated noise will continue indefinitely. The standard deviation of the noise is assumed to continue to increase linearly.
Figure 22: Posterior of component 7 (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.

<table>
<thead>
<tr>
<th>#</th>
<th>ACF min</th>
<th>ACF loc</th>
<th>Periodogram max</th>
<th>Periodogram max loc</th>
<th>QQ max</th>
<th>QQ min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.519</td>
<td>0.349</td>
<td>0.020</td>
<td>0.516</td>
<td>0.084</td>
<td>0.500</td>
</tr>
<tr>
<td>2</td>
<td>0.025</td>
<td>0.492</td>
<td>0.860</td>
<td>0.439</td>
<td>0.765</td>
<td>0.984</td>
</tr>
<tr>
<td>3</td>
<td>0.147</td>
<td>0.259</td>
<td>0.459</td>
<td>0.514</td>
<td>0.171</td>
<td>0.718</td>
</tr>
<tr>
<td>4</td>
<td>0.653</td>
<td>0.691</td>
<td>0.656</td>
<td>0.315</td>
<td>0.318</td>
<td>0.659</td>
</tr>
<tr>
<td>5</td>
<td>0.012</td>
<td>0.515</td>
<td>0.915</td>
<td>0.445</td>
<td>0.667</td>
<td>0.920</td>
</tr>
<tr>
<td>6</td>
<td>0.507</td>
<td>0.488</td>
<td>0.501</td>
<td>0.506</td>
<td>0.405</td>
<td>0.608</td>
</tr>
<tr>
<td>7</td>
<td>0.495</td>
<td>0.487</td>
<td>0.520</td>
<td>0.480</td>
<td>0.493</td>
<td>0.001</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 Highly statistically significant discrepancies

4.1.1 Component 7: Uncorrelated noise with standard deviation increasing linearly away from 1962. This function applies from 1972 onwards

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

- The qq plot has an unexpectedly large negative deviation from equality \((x = y)\). This discrepancy has an estimated \(p\)-value of 0.001.

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

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 7. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2 Moderately statistically significant discrepancies

4.2.1 Component 1: A linearly increasing function. This function applies from 1972 onwards

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

- The maximum value of the periodogram is unexpectedly low. This discrepancy has an estimated \(p\)-value of 0.040.
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 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.2 Component 2: An approximately periodic function with a period of 1.0 years. This function applies from 1972 onwards

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

- The minimum value of the ACF is unexpectedly low. This discrepancy has an estimated p-value of 0.050.

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 2. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.3 Component 5: An approximately periodic function with a period of 1.0 years. This function applies until 1972

The following discrepancies between the prior and posterior distributions for this component have been detected.
- The minimum value of the ACF is unexpectedly low. This discrepancy has an estimated $p$-value of 0.024.

Figure 26: 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.

4.3 Model checking plots for components without statistically significant discrepancies

4.3.1 Component 3: A very smooth monotonically increasing function. This function applies until 1972

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

Figure 27: 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.

4.3.2 Component 4: A smooth function. This function applies from 1972 onwards

No discrepancies between the prior and posterior of this component have been detected
4.3.3 Component 6: Uncorrelated noise

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