MTHM033 - Statistical Modelling in Space and Time
Assessed Coursework 2

670033412

Strength of Overturning In The North Atlantic

The global overturning of ocean waters involves the equatorward transport of cold, deep waters and the poleward transport of warm, near-surface waters [Loz12]. Here we use data collected every 12 hours between 2004-04-02 00:00:00 and 2014-03-22 00:00:00 from a mooring positioned at 26°N measured in Sverdrup (Sv) or cubic hectometers per second (hm^3/s).

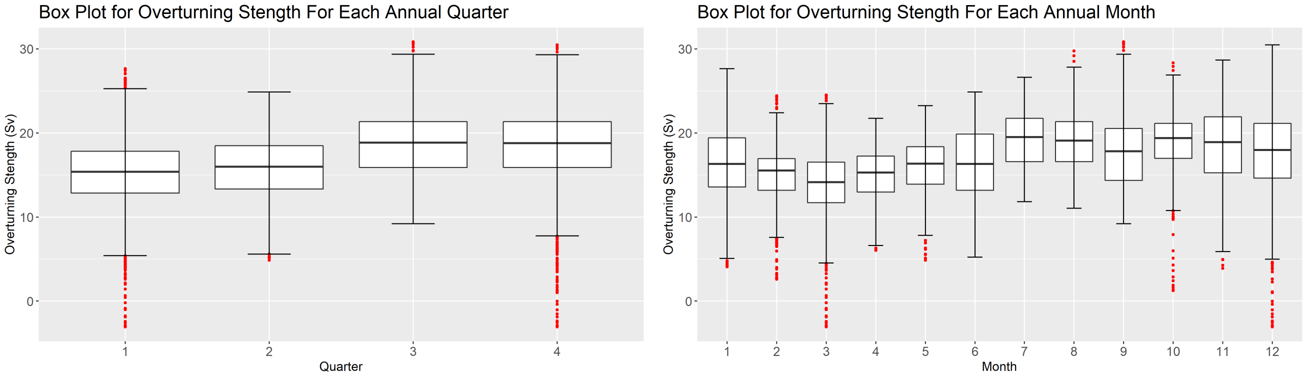
1 Data Integrity

We begin by first running a summary on the data to ensure that all the values are within the ranges we would expect.

Listing 1: Summary of Data

```
      year          month             day              hour            Quarter       Days_since_start Overturning_Strength
 Min.    :2004   Min.    : 1.000   Min.    : 1.0    Min.    : 0.000   Min.    :1.000   Min.    :   1.0   Min.    :-3.073
 Max.    :2014   Max.    :12.000   Max.    :31.0    Max.    :12.000   Max.    :4.000   Max.    :3642.0   Max.    :30.822
```

We can see that our data ranges are appropriate for our date and time columns however, without prior knowledge of the range of overturning strength it is difficult to detect outliers. We may attempt to do so using a box plot



(a) Data grouped into annual quarters.

(b) Data grouped into annual months.

Figure 1: Box plot showing the distribution of overturning strength (Sv) for both annual quarters and for each month. Outliers defined as $\frac{3}{2}$IQR from the upper or lower quartiles are highlighted in red.

We can see in Figure 1 which uses the interquartile range as a method to detect outliers we find an extremely high number of outliers especially in the first and fourth quartile. To investigate further into these outliers we can extract them from our data and plot them against time.

Looking at Figure 2 we can see that the outliers tend to be grouped together but are spread well throughout the data. We may then conclude that these outliers are instead byproducts of the complex nature and long term trends of the ocean.

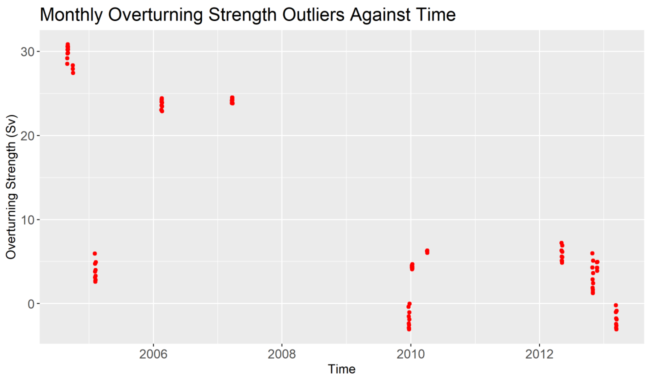


Figure 2: Monthly overturning strength outliers plotted against time.

To get an overview of our data we begin by grouping it into years and annual quarters and then take the mean of each group. This is to reduce noise in our data and also reduces the number of points which helps speeds up the model building computation. We can see this data as a time series in Figure 3.

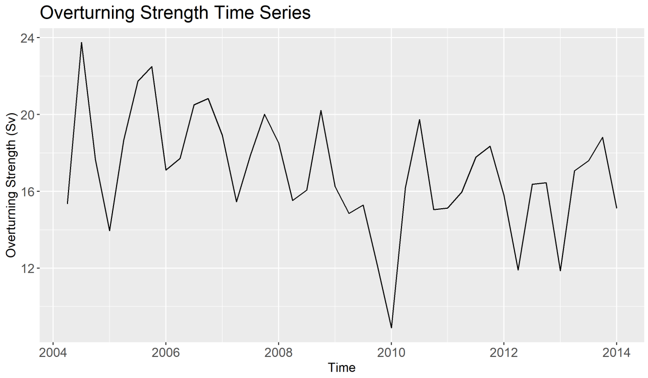


Figure 3: A time series of our original data where each point is the mean taken over the annual quarter.

Taking a look at Figure 3 we can see that there appears to be a pattern of changing between up and down within fairly consistent values roughly every half of year. However, this pattern is not followed between 2009 and 2011 when we see a drop shortly followed by another drop before a very sharp upwards tick till halfway through 2010. Whilst this significant drop may seem as an outlier it is generally accepted in the literature, and its origin is uncertain [BKMM14].

With this quick assessment of the data concluding that there are no outliers, we will then continue to build our models without removing any data from the provided data set.

2 ARMA and ARIMA Model

The first step in building an autoregressive (integrated) moving average (AR(I)MA) model is to determine which process best describes our data. Our process is considered to consist of:

Autoregressive:

$$\mathrm{AR}\,(p): x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon_t$$

Integrated: d is the degree of differencing (the number of times the data have had past values subtracted).

Moving Average:

$$\mathrm{MA}\,(q): x_t = \sum_{i=0}^{q} \beta_i \epsilon_{t-i}$$

Where x_t represents the value measured at the t^{th} time step (Δt) and ϵ_t are identical and independent realizations from a normal (Gaussian) distribution with zero mean and variance σ^2.

2.1 Autoregressive Moving Average Model (ARMA)

An ARMA model only consist off the $\mathrm{AR}\,(p)$ and $\mathrm{MA}\,(q)$ often referred to as $\mathrm{ARMA}\,(p, q)$ but is identical to $\mathrm{ARIMA}\,(p, 0, q)$, expressed mathematically as:

$$\mathrm{ARMA}\,(p,q): x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=0}^{q} \beta_i \epsilon_{t-i} \quad \text{where } \epsilon_t \sim \mathrm{N}\left(0, \sigma^2\right)$$

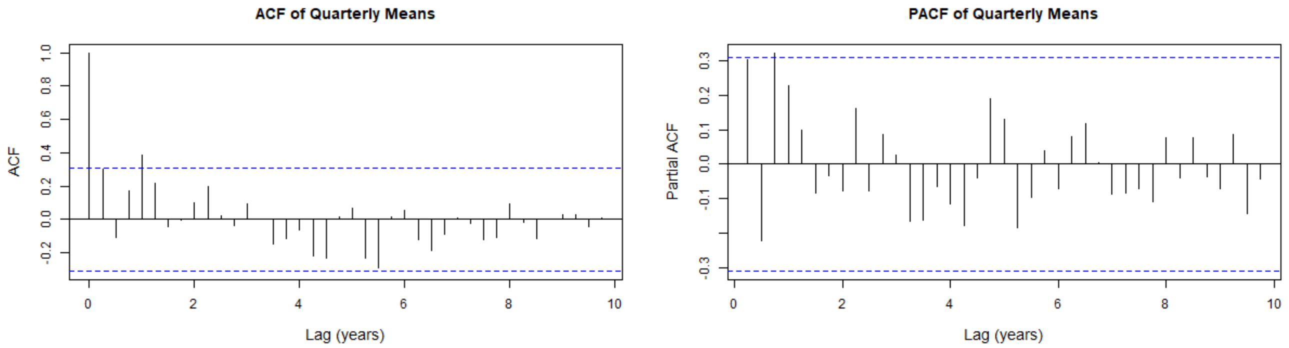
Our problem then is to fit parameters (p, q) that best describe the underlying process of data.

2.1.1 Finding optimal parameters

It is possible that our data may have one of the parameters equal 0 meaning that we just an AR or MA process. One way to check this is to calculate both the auto correlation (ACF) and partial auto correlation functions (PACF). Then using the table below we can see if our data satisfies the criteria.

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

The ACF and PACF functions of our data is shown in Figure 4. Taking a look at the ACF Figure 4a we see that we have two peaks at 0.25 and 1.00 (excluding peak at 0). Then looking at the PACF Figure 4b we have one peak at 0.75 but also one peak just below the 95% confidence interval at 0.25. Since any of these peaks could be down to noise, it is hard to exactly determine our parameters from this analysis alone.



(a) Auto correlation function

(b) Partial auto correlation function

Figure 4: Auto correlation and partial auto correlation functions of our quarterly means data plotted for different lag values in years, with 95% significance level shown in blue.

We will fit several ARMA models for varying values, $(p, q) \in [0, 3] \times [0, 7]$ and calculate both the AIC and BIC of the fit. This produces the following results:

p/q	0	1	2	3	4	5	6	7
0	412.05	399.19	399.91	405.23	403.42	404.96	408.73	413.08
1	410	400.07	405.44	405.16	407.67	409.72	414.09	418.62
2	410.03	405.24	410.92	403.7	409.39	413.58	412.62	424.28
3	401.39	404.6	408.75	409.39	415.08	412.44	417.67	421.29

Table 1: Sum of AIC and BIC for different (p, q) pairs. For $q > 7$ or $p > 3$ the parameter pair approach the end of the stationarity region.

- The minimum pair hello of (p, q) when minimizing the AIC is $p = 2$ and $q = 3$
- The minimum pair hello of (p, q) when minimizing the BIC is $p = 0$ and $q = 1$
- The minimum pair hello of (p, q) when minimizing the Sum of AIC and BIC is $p = 0$ and $q = 1$

It is worth mentioning that the models allowed for there to be a mean and the best fit was calculated by first using conditional-sum-of-squares to find starting values, then maximum likelihood.

Continuing with $p = 0$ and $q = 1$, we shall take the residuals of our model and produce the following plots:

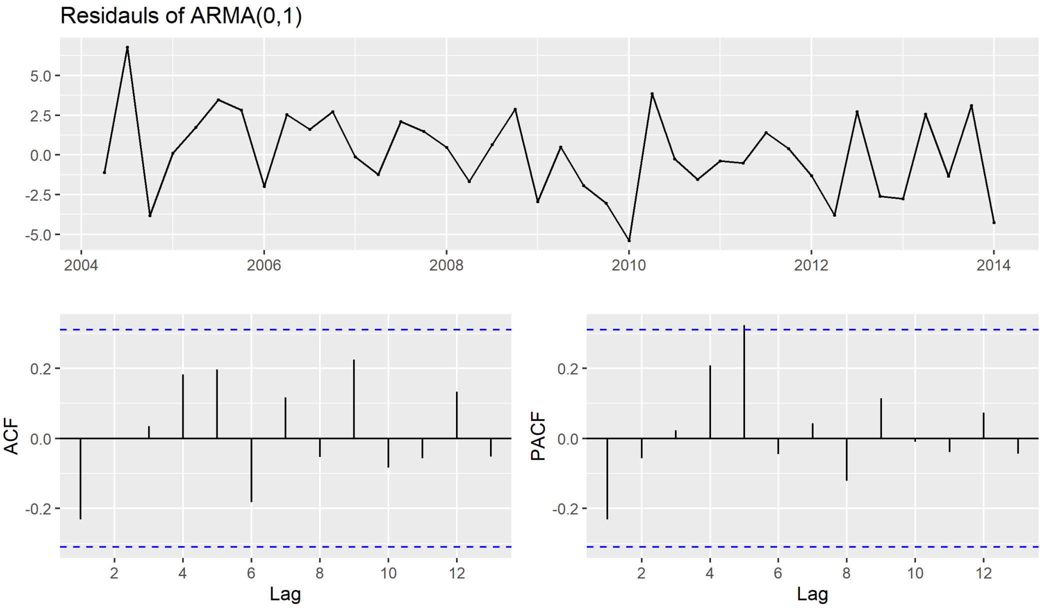


Figure 5: Residuals from our ARMA $(0, 1)$ and their ACF and PACF.

References

[BKMM14] H.L. Bryden, B.A. King, G. McCarthy, and Elaine Mcdonagh. Impact of a 30% reduction in atlantic meridional overturning during 2009-2010. *Ocean Science*, 11, 02 2014.

[Loz12] M. Susan Lozier. Overturning in the north atlantic. *Annual Review of Marine Science*, 4(1):291–315, 2012. PMID: 22457977.