Demand Forecasting

for Personnel Allocation

Case study Demonstration - UIC IDS.506

Robert Duc Bui - 660809303

Case Information

- This material is presented as a case study for UIC Liautaud School of Business, Department of Information and Decision Sciences, Course 506.27476 - Healthcare Analytics.
- The information presented here has been released for use by original owners for informational purposes only. Historical data does not indicate current business needs or data of the original provider.

Scope & Objective

- Scope:
  - Demand volume data for examination services performed by the Cardiovascular department for FHG patients at specific location in Abbeville, LA.
  - Aggregation period: from 2006 to 2013.
  - Aggregation frequency: monthly.
- Issue Statement:
  - Current demand management at the Abbeville location has resulted in booking difficulty, as staffing challenges with physicians and nurses require management to book personnel shifts well in advance of actual patient demand.
- Objective:
  - To forecast patient demand in advance of personnel plans, in service of efficient FTE utilisation.

Additional Business Context

- Current workstream: Patient -> Local Offices[LO] -> Health Centers[HC]

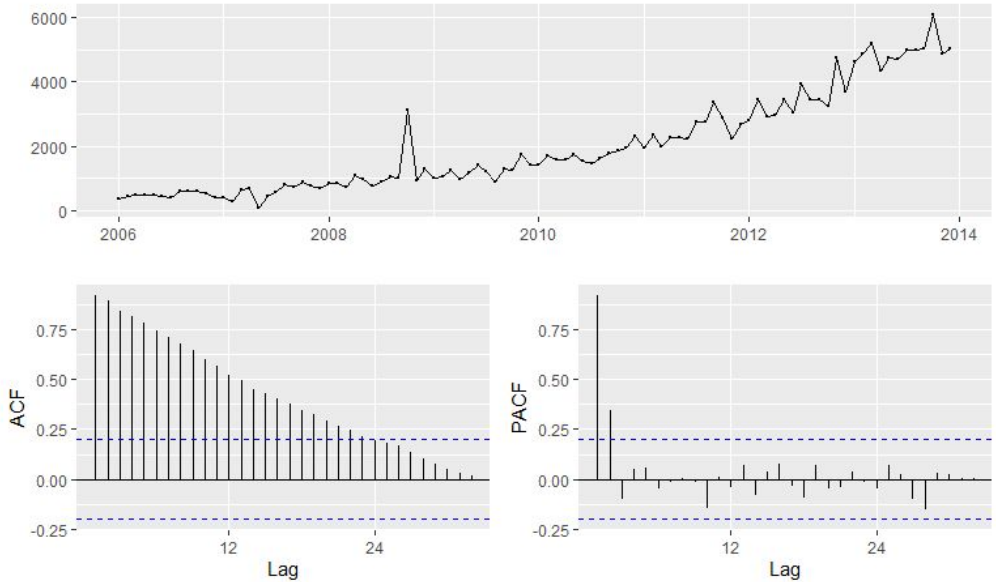
- HC constraints: 30 days to report test results
    - Regulatory penalty of 200$ per day late fees to Regional Office of Health Oversight
    - FHG HC Abbeville is severely understaffed and supplemental head counts have to be booked well in advance from other surrounding HCs.

- Current solutions include:
    - Re-routing to other HCs and out-of-network providers - not sustainable as other HCs are running into similar staffing issues. Out-of-network provision of services is a significant barrier for patients, and also creates additional transfer cost concerns for FHG.

Data Ingest

- Non-time series data with multidimensionality and many subtables, pulled from EHR and EMR records. Shortlist:
  - Booking data (by entry) at FHG Abbeville, Violet, New Orleans, Lafayette, Baton Rouge.
  - Routing SYSID : System identification codes for routed tests carried out in 2013.
  - Heart related Condition Codes
  - Condition Code Map: Codes for various health issues reported
- Preliminary cleaning:
  - Extraction of booking data from Abbeville table, grouped by date and counted for number of unique bookings
    - Joined with SYSID routing table to remove rerouted bookings.
    - Joined with Condition Code table to remove non-cardio bookings.
    - Data irregularities (non-numerics, corrupted rows) replaced with NA signifier
  - Data encoded as `ts` format for time series analysis in R.

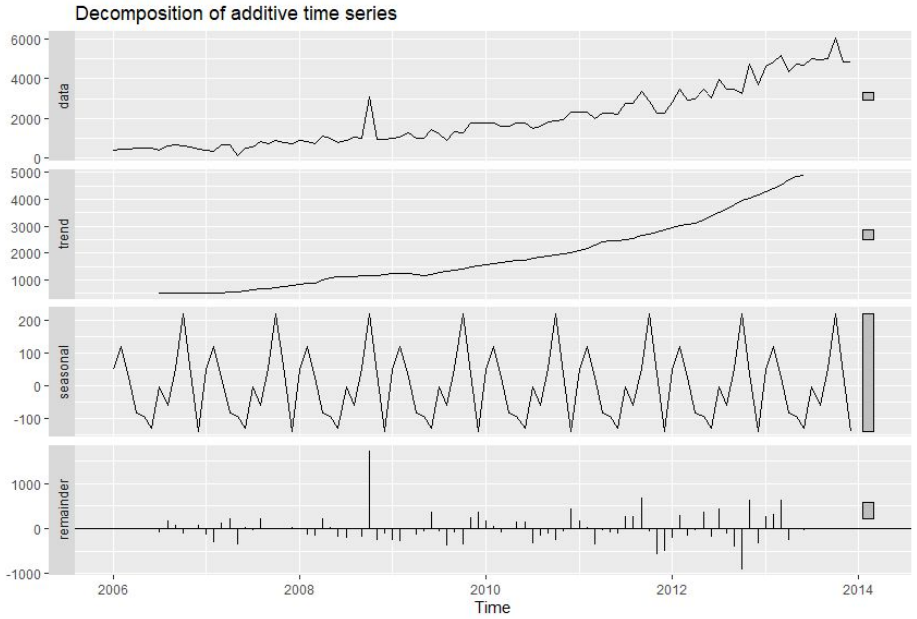
Seasonality Analysis - ACF & pACF



Observations:

- There are several non-zero autocorrelations. Time series exhibits trend or seasonality, or both.
- High degree of negative autocorrelation between adjacent values (lag=1) in PACF.
- Some degree of autocorrelation around lag=4.
- ACF plot decay - indicates trend.

Seasonality Analysis - Decomposition

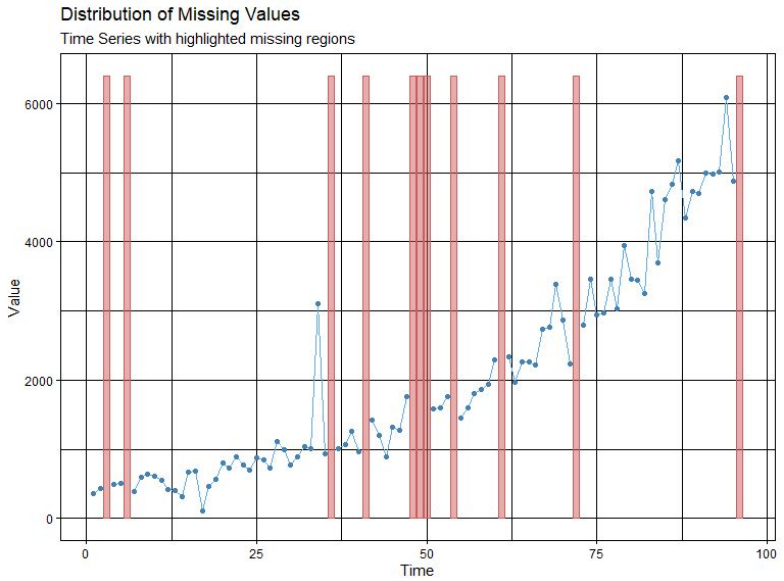


Observations:

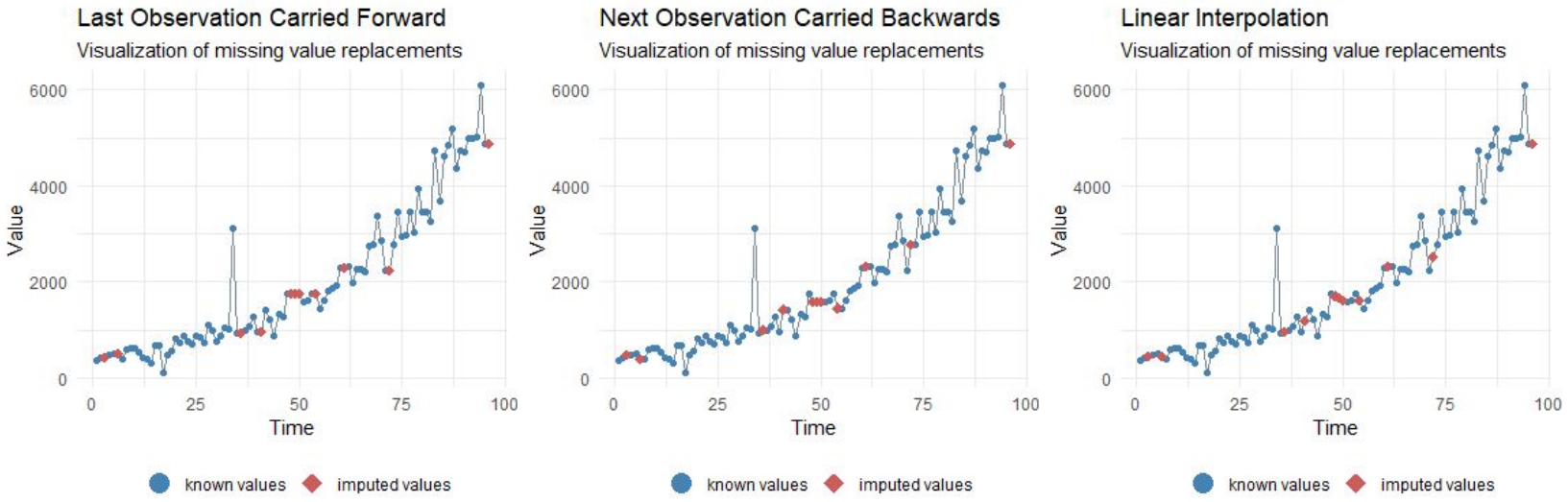
- Overall upward trend.
- When trend is removed, there is significant seasonality.
- When trend and seasonality are removed, residuals are mostly minor except for massive spike in late 2008. Could be due to transfers resulting from other facilities' closures during Hurricanes Gustav and Ike.

Missing Value Analysis & Imputation

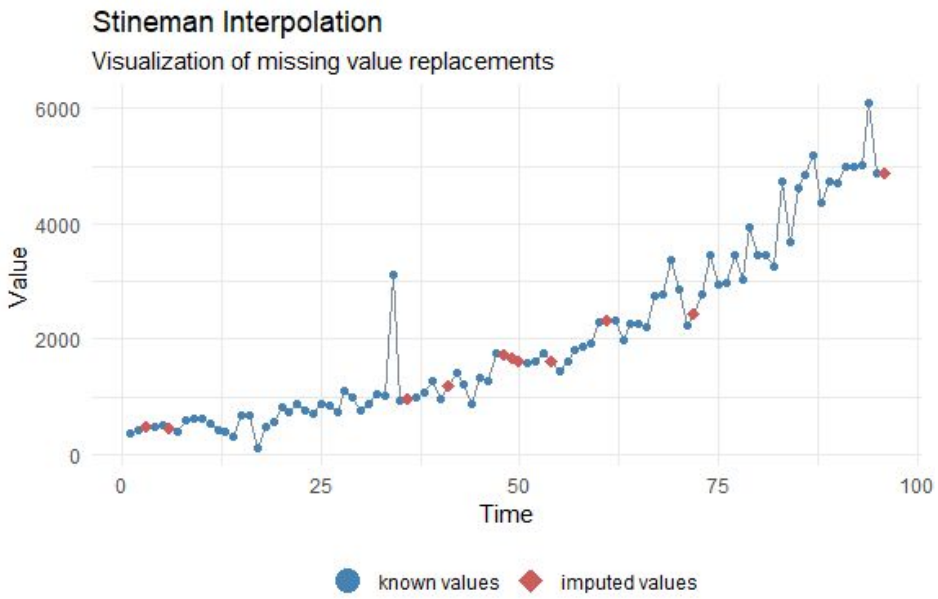
```
[1] "Length of time series:"
[1] 96
[1] "-------------------------"
[1] "Number of Missing Values:"
[1] 11
[1] "-------------------------"
[1] "Percentage of Missing Values:"
[1] "11.5%"
[1] "-------------------------"
[1] "Number of Gaps:"
[1] 9
[1] "-------------------------"
[1] "Average Gap Size:"
[1] 1.222222
[1] "-------------------------"
[1] "Stats for Bins"
[1] "  Bin 1 (24 values from 1 to 24) :      2 NAs (8.33%)"
[1] "  Bin 2 (24 values from 25 to 48) :     3 NAs (12.5%)"
[1] "  Bin 3 (24 values from 49 to 72) :     5 NAs (20.8%)"
[1] "  Bin 4 (24 values from 73 to 96) :     1 NAs (4.17%)"
[1] "-------------------------"
[1] "Longest NA gap (series of consecutive NAs)"
[1] "3 in a row"
[1] "-------------------------"
[1] "Most frequent gap size (series of consecutive NA series)"
[1] "1 NA in a row (occurring 8 times)"
[1] "-------------------------"
[1] "Gap size accounting for most NAs"
[1] "1 NA in a row (occurring 8 times, making up for overall 8 NAs)"
[1] "-------------------------"
[1] "Overview NA series"
[1] "  1 NA in a row: 8 times"
[1] "  3 NA in a row: 1 times"
```



Missing Value Imputation - Basic Methods



Missing Value Imputation - Stineman Interp.

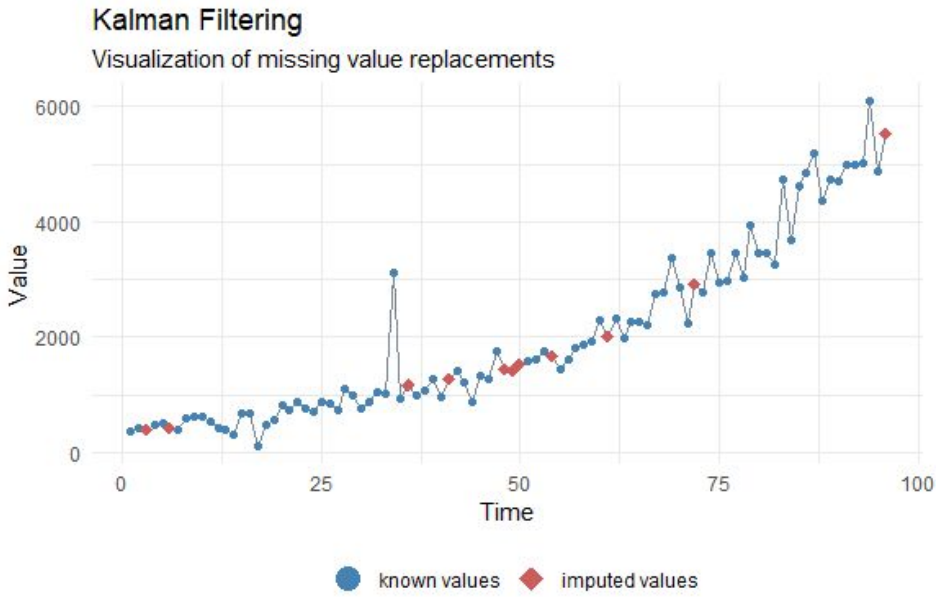


Stineman interpolation takes the concept of linear interpolation, but instead fits several splines and curves to the data.

This method works piecewise by breaking the entire series into separate segments, then fitting a curve to each segment.

This technique ensures smoother transitions between phases in the data, while preserving local extrema and allows for estimation of the derivative.

Missing Value Imputation - Kalman Filtering

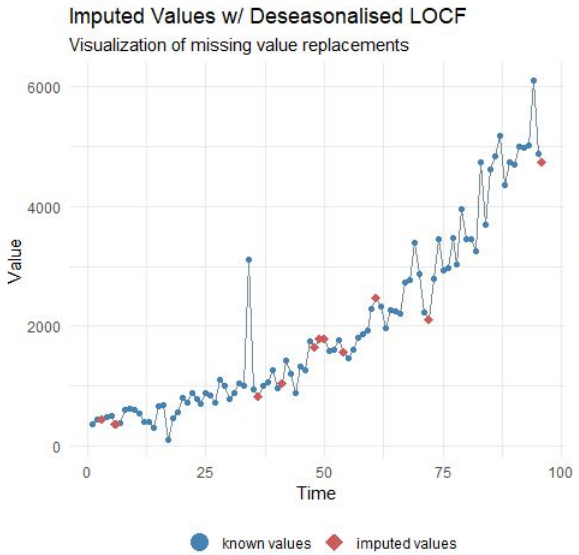


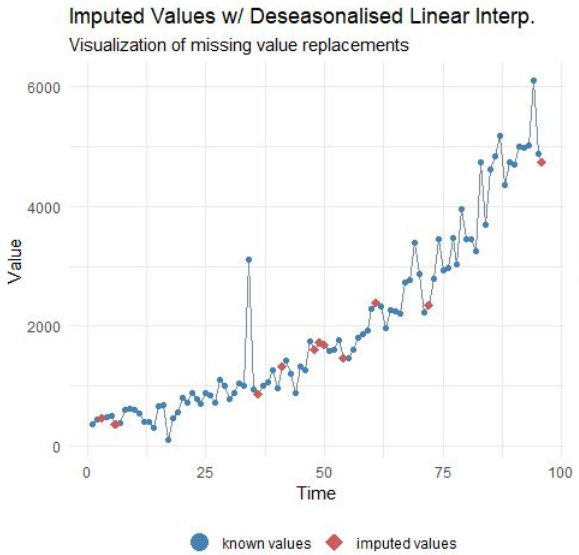
Kalman filtering for NA imputation works by using two main steps: prediction and correction.

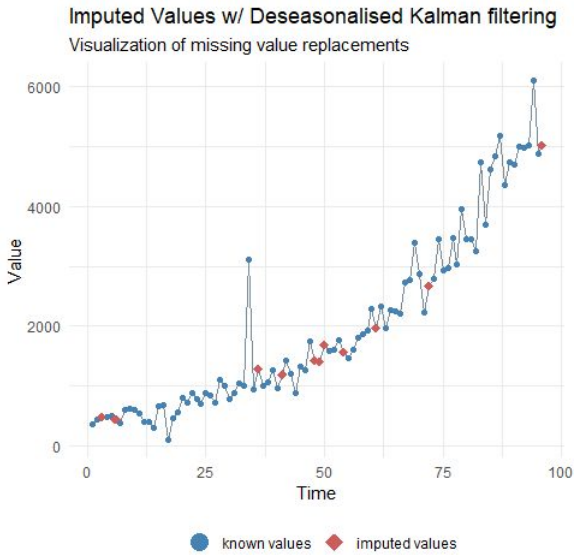
- Prediction: The filter predicts several states ahead of the system at the next time step based on data up to lag 1.
- Correction: The filter then compares the predicted state with actual measurements of the system, and adjusts the prediction to minimize the error. This is done by taking into account the uncertainties in both the prediction and the measurements, and finding the most likely true state.

Kalman filtering results have reduced effects in noise, and is generally accepted to have high accuracy.

Missing Value Imputation - Adjusted Methods

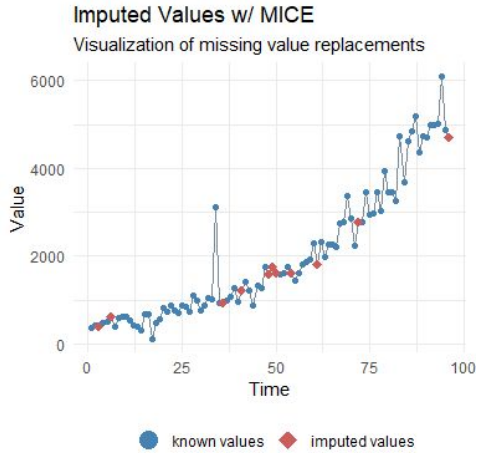


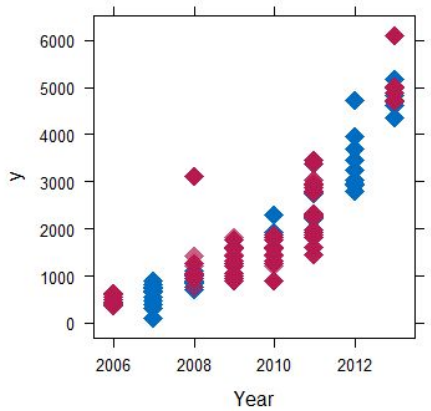


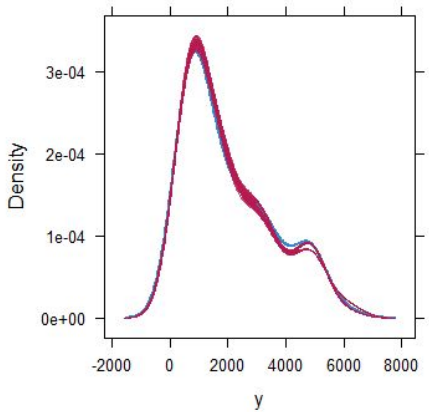


Missing Value Imputation - MICE

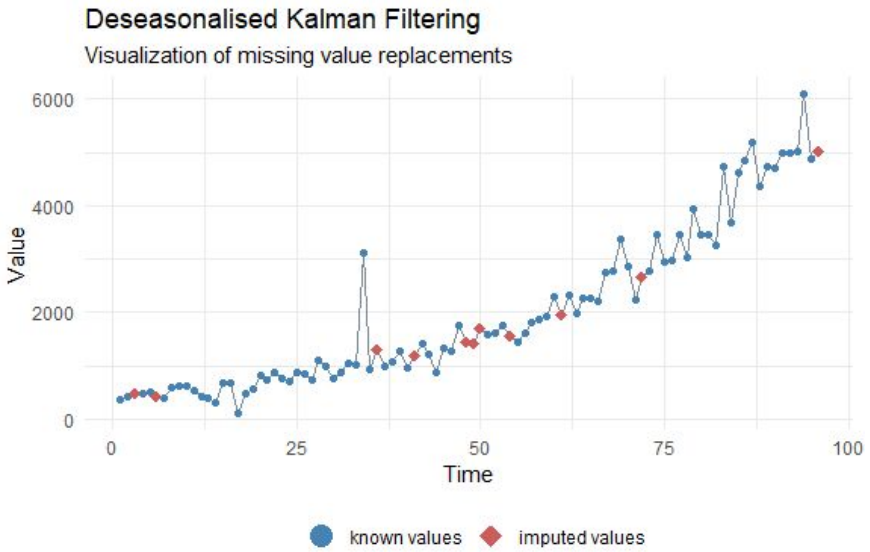
MICE is the most advanced method of general methods for imputation of NA data, and stands for Multivariate Imputation By Chained Equations algorithm. While this technique is the most advanced, it is also much more dependent on other features in the feature space, which makes it not appropriate for our use.







Missing Value Imputation - Final



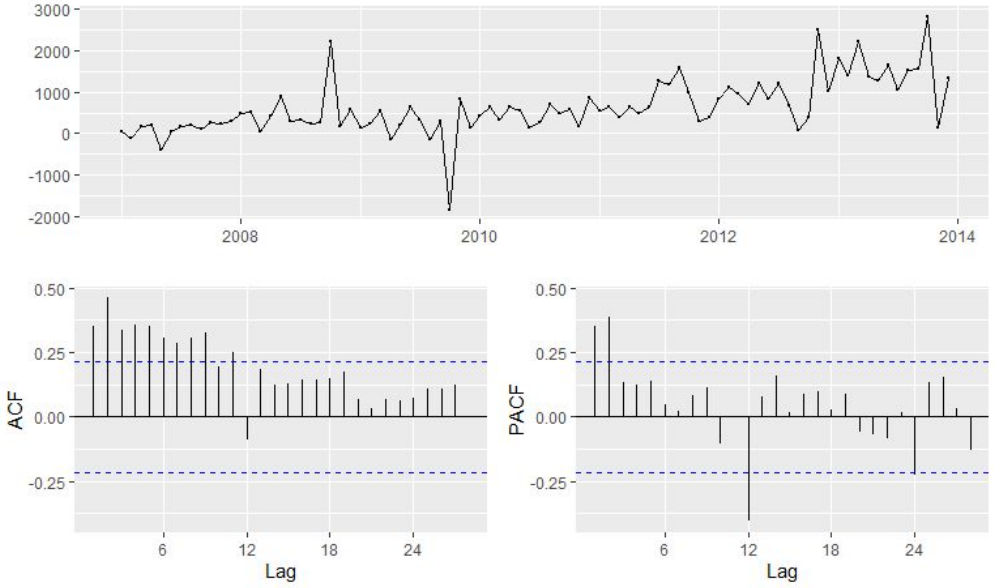
Deseasonalised Kalman filtering allows for a robust and accurate imputation process, while accounting for the established seasonality trends present in the data.

Deseasonalised Stineman is a viable alternative due to being able to preserve trend smoothness and local extrema.

Forecasting Model

- Type: Seasonal AutoRegressive Integrated Moving Average model (SARIMA)
- Tuning Strategy:
  - Observation of ACF and pACF plots, seasonal, differenced to manually determine p,d,q model terms.
  - Use of R's auto.arima to find optimal model.

Manual: Seasonal ACF & pACF

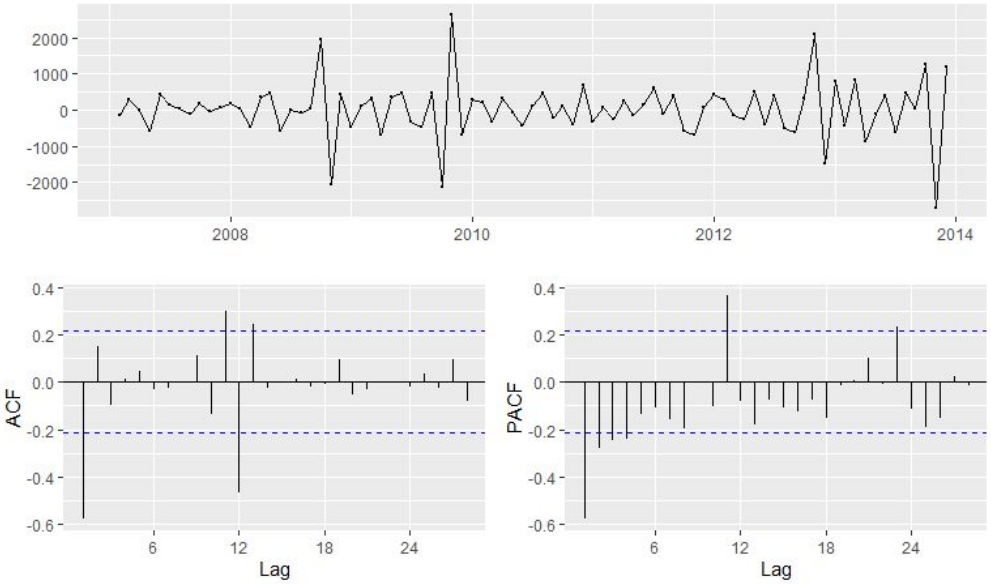


KPSS: 1.486 at p = 0.01

Observations:

- Seasonal ACF and pACF appears to be non-stationary, thus will need to be differenced again for the second-order difference.
- Thus we must start our models at SARIMA(p,1,d)(P,1,D)12.

Manual: Seasonal differenced ACF & pACF



KPSS: 0.26 at p = .1

Observations:

- Significant dip at lag 1 in the ACF suggests non-seasonal AR(1), and significant peak at lag 11/12 ACF or lag 11 pACF suggests a seasonal AR(1) component. Some other peaks are also notable.
- Thus, we can try to fit a SARIMA(1,1,q)(1,1,Q)12 or a SARIMA(11,1,q)(0,1,Q)12.
- Combined with the previous plots, we can also try SARIMA(2,0,q)(1,1,Q)12.

Auto: R's auto.arima() functionality

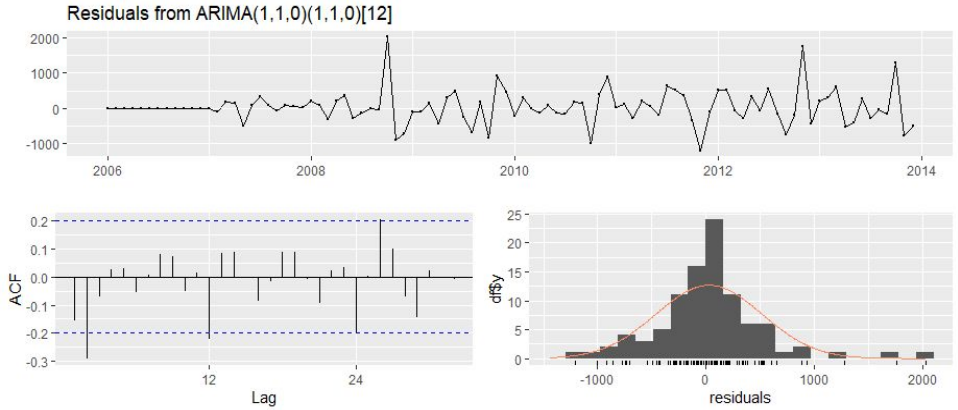
```r
model_auto <- ts %>% auto.arima(
  D = 1,
  stationary = FALSE,
  seasonal = TRUE,
  stepwise = FALSE
)
```

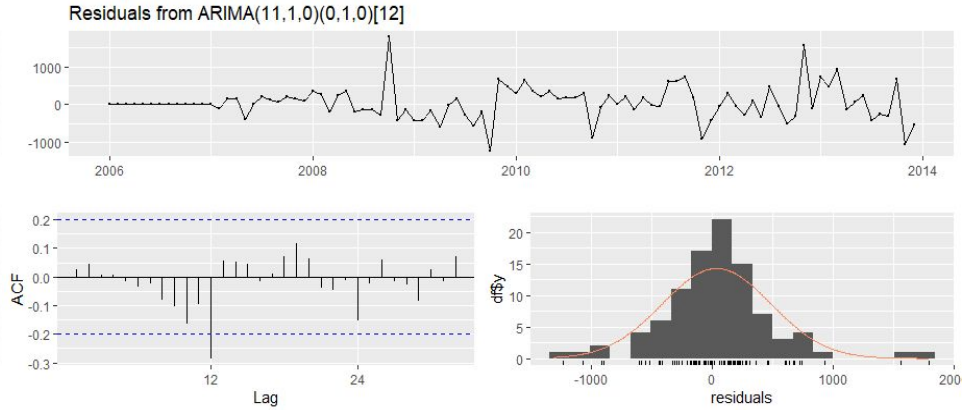
```
Series: .
ARIMA(0,1,1)(2,1,1)[12]

Coefficients:
          ma1     sar1     sar2    sma1
      -0.7395  -1.2001  -0.6780  0.6286
s.e.   0.0616   0.1601   0.0933  0.2508

sigma^2 = 181147:  log likelihood = -625.27
AIC=1260.54   AICc=1261.32   BIC=1272.63
```

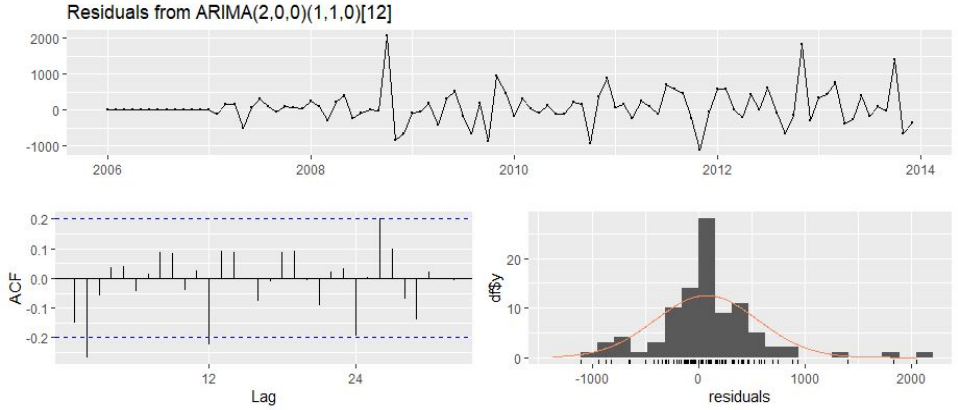
- R's forecast::auto.arima() implements a variation of the Hyndman-Khandakar algorithm, which uses either stepwise or stochastic search to determine the best value for (p,d,q) and (P,1,Q).
  - (We are setting D=1 here to limit the solution space and to let auto.arima use seasonal differencing).
- To search in a wider area of the solution space, we will be using an unrestricted stochastic Hyndman-Khandakar optimisation while accounting for seasonality.
- The output model is SARIMA$(0,1,1)(2,1,1)_{12}$

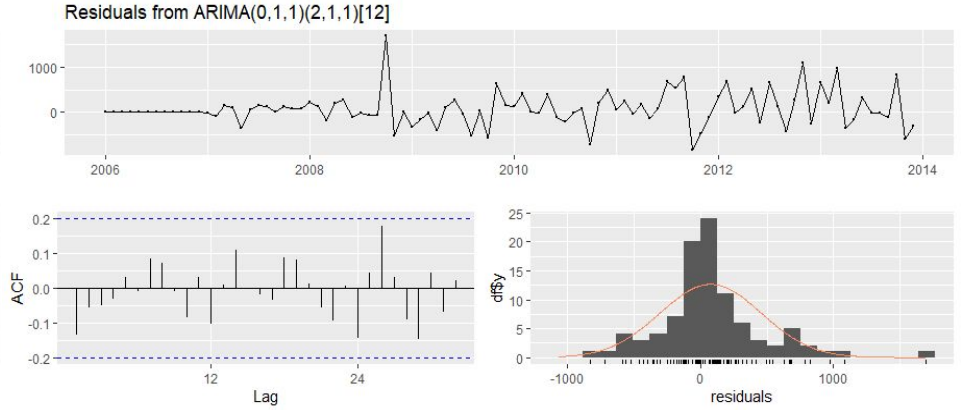
Model Assessment Statistics

Candidate Models	AICc	MAPE	Ljung-Box Test of residuals	
			Q*	p-value
ARIMA(1,1,0)(1,1,0)$_{12}$	1285.26	21.9	23.583	.1312
ARIMA(11,1,0)(0,1,0)$_{12}$	1293.57	20.76	18.703	.0165*
ARIMA(2,0,0)(1,1,0)$_{12}$	1302	21.41	22.149	.1384
ARIMA(0,1,1)(2,1,1)$_{12}$	1261.32	16.07	9.371	.8574

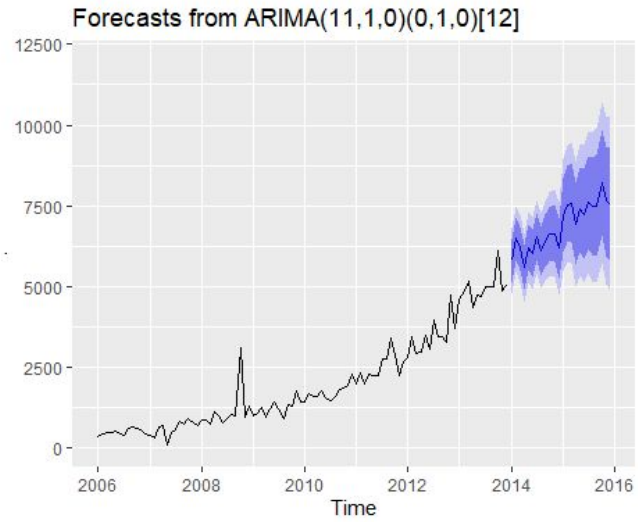

Residuals from ARIMA(1,1,0)(1,1,0)[12]
2000
1000
0
-1000
2006
2008
2010
2012
2014
ACF
0.2
0.1
0.0
-0.1
-0.2
-0.3
12
24
Lag
df$y
25
20
15
10
5
0
-1000
0
1000
2000
residuals

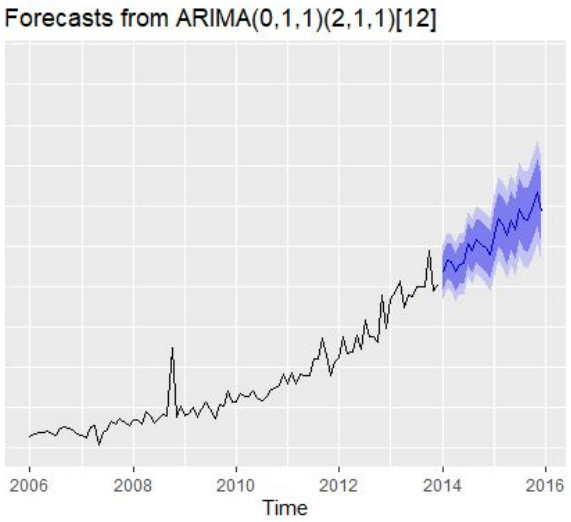

Residuals from ARIMA(11,1,0)(0,1,0)[12]
1000
0
-1000
2006
2008
2010
2012
2014
ACF
0.2
0.1
0.0
-0.1
-0.2
-0.3
12
24
Lag
df$y
20
15
10
5
0
-1000
0
1000
200
residuals


Residuals from ARIMA(2,0,0)(1,1,0)[12]
2000
1000
0
-1000
2006
2008
2010
2012
2014
ACF
0.2
0.1
0.0
-0.1
-0.2
12
24
Lag
df$y
20
10
0
-1000
0
1000
2000
residuals


Residuals from ARIMA(0,1,1)(2,1,1)[12]
1000
0
2006
2008
2010
2012
2014
ACF
0.2
0.1
0.0
-0.1
-0.2
12
24
Lag
df$y
25
20
15
10
5
0
-1000
0
1000
residuals

Forecasts (24 month)





Observations:

- While (11,1,0)(0,1,0) is the model whose residuals conform the most to a normal distribution, the confidence intervals for the forecasts themselves see higher variance.
- Thus, should more stable forecasts be of interest, the auto.arima output model serves best as the primary output, and the (11,1,0)(0,1,0) could serve as a backup.

Output Model Summary

NA Imputation Method: Deseasonalised Kalman filtering

Final output model type: AutoRegressive Integrated Moving Average (ARIMA) with seasonal component.

Model parameters:

- $(p,d,q) = (0,1,1)$
- $(P,D,Q)_{seasonality} = (2,1,1)_{12}$

In-sample Model Performance Metrics:

- AICc: 1261.32
- MAPE: 16.07%