DATA 624: Project 1

Juliann McEachern

October 22, 2019

Contents

Overview

I am leaving the project overview page here for us to compile our final report in one singular document. We will add additional information here regarding project one to include explanation of process, etc.

Dependencies

Please add all libraries used here.

The following R libraries were used to complete Project 1:

```
# General
library('easypackages')

libraries('knitr', 'kableExtra', 'default')

# Processing
libraries('readxl', 'tidyverse', 'janitor', 'lubridate')

# Graphing
libraries('ggplot2', 'grid', 'gridExtra', 'ggfortify','ggpubr')

# Timeseries
libraries('zoo', 'urca', 'tseries', 'timetk')

# Math
libraries('forecast')
```

Data

Data was stored within our group repository and imported below using the `readxl` package. Each individual question was solved within an R script and the data was sourced into our main report for discussion purposes. The R scripts are available within our appendix for replication purposes.

For grading purposes, we exported and saved all forecasts as a csv in our data folder.

```
# Data Aquisition
atm_data <- read_excel("data/ATM624Data.xlsx")
power_data <- read_excel("data/ResidentialCustomerForecastLoad-624.xlsx")
pipe1_data <- read_excel("data/Waterflow_Pipe1.xlsx")
pipe2_data <- read_excel("data/Waterflow_Pipe2.xlsx")

# Source Code
source("scripts/Part-A-JM.R")
```

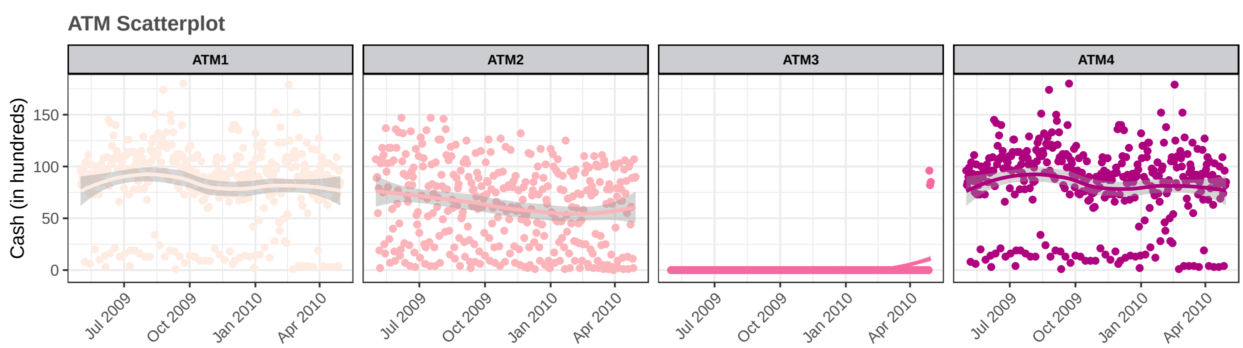
1 Part A

Instructions: In part A, I want you to forecast how much cash is taken out of 4 different ATM machines for May 2010. The data is given in a single file. The variable `Cash` is provided in hundreds of dollars, other than that it is straight forward. I am being somewhat ambiguous on purpose. I am giving you data, please provide your written report on your findings, visuals, discussion and your R code all within a Word readable document, except the forecast which you will put in an Excel readable file. I must be able to cut and paste your R code and run it in R studio. Your report must be professional - most of all - readable, EASY to follow. Let me know what you are thinking, assumptions you are making! Your forecast is a simple CSV or Excel file that MATCHES the format of the data I provide.

1.1 Exploration

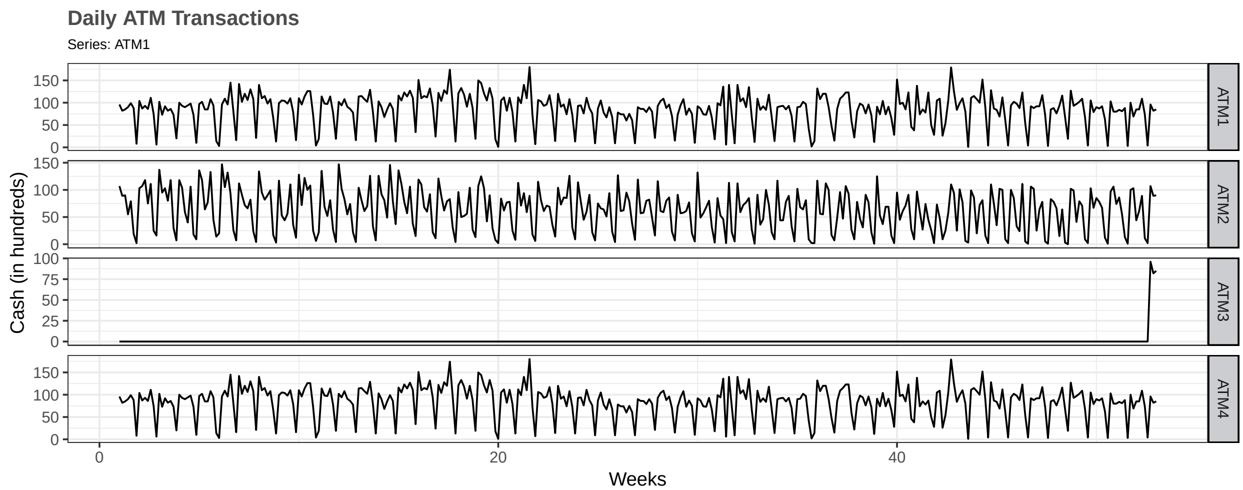
Through data exploration, we identified that the original data file contained `NA` values in our `ATM` and `Cash` columns for 14 observations in May 2010. We removed these missing values and transformed the dataset into a wide format. Our cleaned dataframe was then converted into a timeseries format using the `zoo` package for forecasting in the next section. Our initial review of the data showed that ATM2 contained one missing value on 2009-10-25 and that ATM4 contained a potential outlier of $1123 on 2010-02-09. We replaced both values with the corresponding mean value of each machine.

Next, we used a scatterplot to take an initial look at the correlation between cash withdrawals and dates for each machine. We can identified similiar patterns between ATM1 and ATM4, which show non-linear fluxuations that suggest a potential trend component in these timeseries. ATM2 follows a relatively linear path and decreases overtime. This changes in the last few observations, where withdrawals begin to increase. There are only 3 observed transactions for ATM3 that appear at the end of the captured time period.



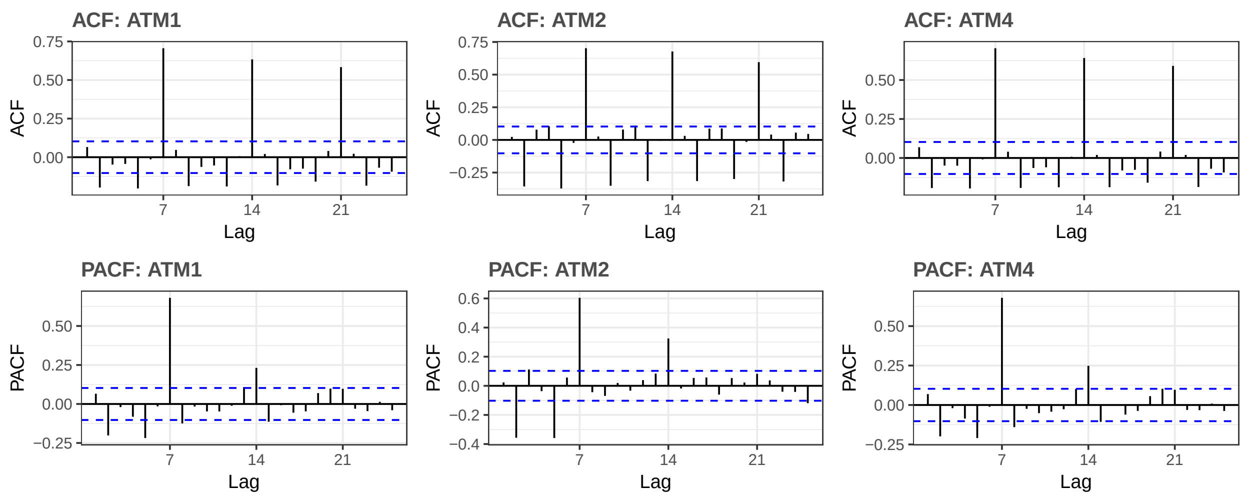
1.2 Timeseries Plots

The time series plots show high weekly variance as suspected in from our scatterplots. We can see again that time series for ATM3 only contains 3 transactions, thus we deemed this series not suitable for modeling and forecasting. As a result, our following sections will not include extrapolation on this series.



1.3 Evaluation

We constructed our initial timeseries using a weekly frequency. Our ACF plots for each ATM showcases large, decreasing lags starting at 7. This pattern continues in a multiple of seven, which confirms our assumption about seasonality within the observed data. These lags are indicative of a weekly pattern.



Our plots further suggest that the ATM data is non-stationary. We performed a unit root test using the `ur.kpss()` function to confirm this observation. The test results below show that differencing is required on all ATM2 and ATM4 series. ATM1 falls just below the cut-off critical value, but could still use differencing due to the observed seasonal pattern.

Table 1.1: KPSS unit root test

ATM	No-Diff	Diff-1
ATM1	0.4967	0.0219
ATM2	2.0006	0.016
ATM4	0.5182	0.0211

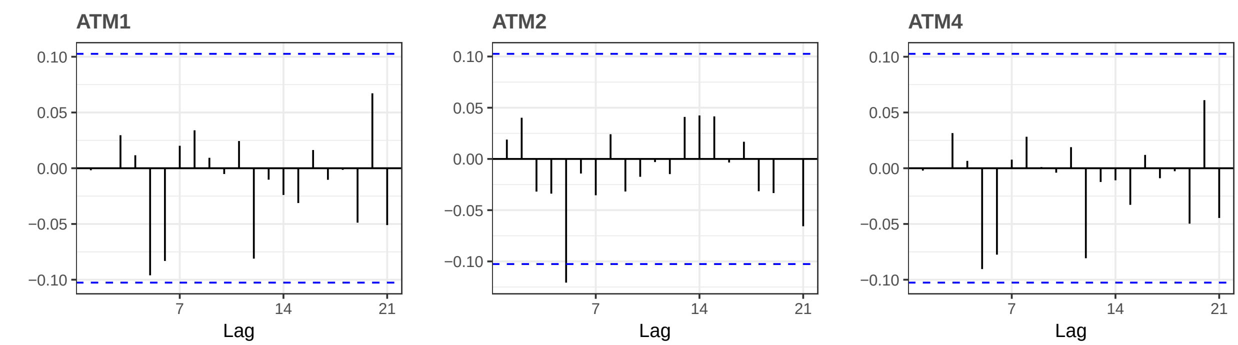
1.3.1 Modeling

We used `auto.arima()` and set `D=1` to account for seasonal differencing of our data to select the best ARIMA models. The full models and accuracy statistics for each series can be viewed in the appendix.

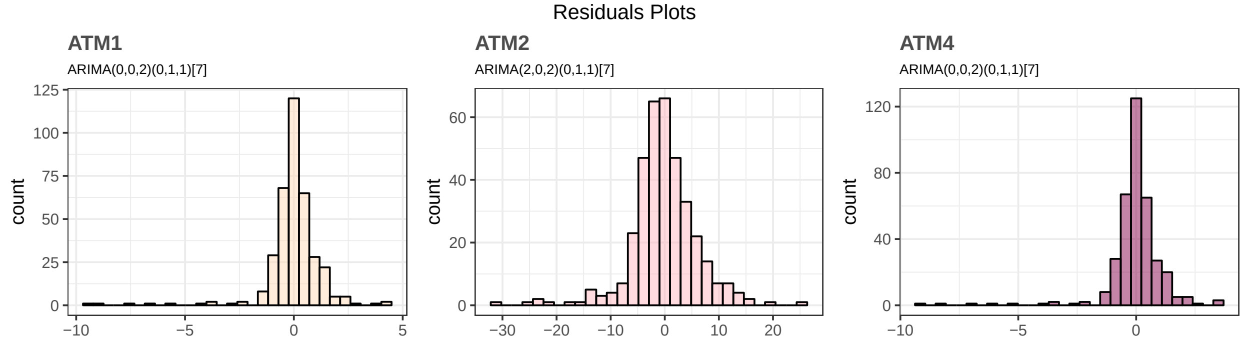
- **ATM1**: ARIMA$(0, 0, 2)(0, 1, 1)_7$
- **ATM2**: ARIMA$(2, 0, 2)(0, 1, 1)_7$
- **ATM4**: ARIMA$(0, 0, 2)(0, 1, 1)_7$

The following ACF plots show us that our differentiated data is now stationary. Further, the residual histograms follow a relatively normal distribution, which confirms that the models adequately fits the observed data.

ACF Plots



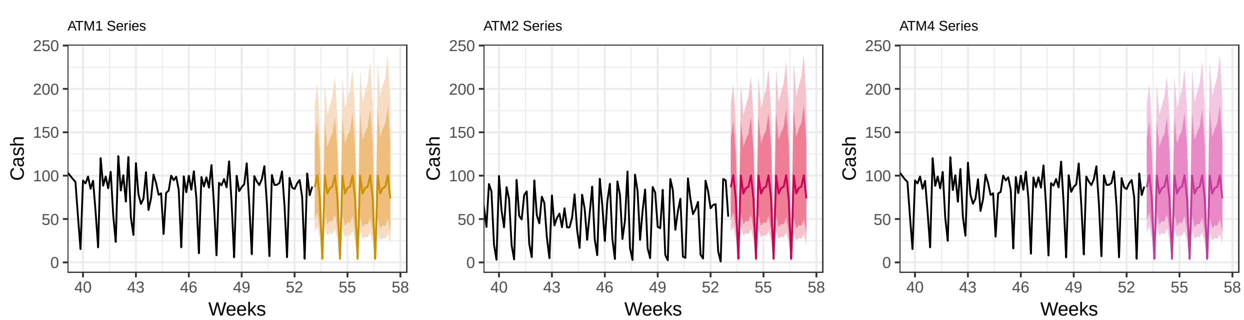
Residuals Plots



1.4 Forecast

Finally, we applied a forecast to each series for 31 days, which span across 5 weeks, in May 2010. The numeric forecasts can be viewed in a table output in the appendix section and are also located within our data output folder.

ATM Forecasts



Appendix

Part A

ARIMA Model Summary

ATM1:

```
FALSE Series: ATM1_ts
FALSE ARIMA(0,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.2584338
FALSE
FALSE Coefficients:
FALSE          ma1      ma2     sma1
FALSE       0.1085  -0.1089  -0.6425
FALSE s.e.  0.0524   0.0521   0.0431
FALSE
FALSE sigma^2 estimated as 1.726:  log likelihood=-606.1
FALSE AIC=1220.2   AICc=1220.32   BIC=1235.72
```

ATM2:

```
FALSE Series: ATM2_ts
FALSE ARIMA(2,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.661752
FALSE
FALSE Coefficients:
FALSE           ar1      ar2     ma1     ma2     sma1
FALSE       -0.4238  -0.8978  0.4766  0.7875  -0.7064
FALSE s.e.   0.0592   0.0473  0.0883  0.0608   0.0417
FALSE
FALSE sigma^2 estimated as 38.94:  log likelihood=-1162.96
FALSE AIC=2337.93   AICc=2338.17   BIC=2361.21
```

ATM4:

```
FALSE Series: ATM4_ts
FALSE ARIMA(0,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.2328582
FALSE
FALSE Coefficients:
FALSE          ma1      ma2     sma1
FALSE       0.1095  -0.1088  -0.6474
FALSE s.e.  0.0524   0.0523   0.0420
FALSE
FALSE sigma^2 estimated as 1.439:  log likelihood=-573.5
FALSE AIC=1154.99   AICc=1155.11   BIC=1170.52
```

Point Forecasts

Table 1.2: ATM Mean Point Forecast

Date	ATM1	ATM2	ATM3	ATM4
2010-05-01	86.6822230334281	65.9130078295388	NA	86.7148793513953
2010-05-02	100.569237833094	71.2678744758685	NA	100.581688969852
2010-05-03	73.710292000956	11.4694658108709	NA	73.645362194735
2010-05-04	4.22902938718943	2.46415237936744	NA	4.22143290375228
2010-05-05	100.159253208527	98.3397063045857	NA	100.159422079885
2010-05-06	79.3467329167084	89.0607216505235	NA	79.3417211032809
2010-05-07	85.7390398914344	66.0684601266365	NA	85.7781984478913
2010-05-08	87.1797624007239	65.9067170158147	NA	87.218337497544
2010-05-09	100.388112929695	71.3008782172556	NA	100.395109536387
2010-05-10	73.710292000956	11.4650527078295	NA	73.645362194735
2010-05-11	4.22902938718943	2.45577466621943	NA	4.22143290375228
2010-05-12	100.159253208527	98.3602657501897	NA	100.159422079885
2010-05-13	79.3467329167084	89.0776223553505	NA	79.3417211032809
2010-05-14	85.7390398914344	66.0458523021009	NA	85.7781984478913
2010-05-15	87.1797624007239	65.9025868757576	NA	87.218337497544
2010-05-16	100.388112929695	71.3235063276986	NA	100.395109536387
2010-05-17	73.710292000956	11.4619371370152	NA	73.645362194735
2010-05-18	4.22902938718943	2.45005983249313	NA	4.22143290375228
2010-05-19	100.159253208527	98.3744953596224	NA	100.159422079885
2010-05-20	79.3467329167084	89.0890848253535	NA	79.3417211032809
2010-05-21	85.7390398914344	66.0302973545638	NA	85.7781984478913
2010-05-22	87.1797624007239	65.8998808213307	NA	87.218337497544
2010-05-23	100.388112929695	71.3390190248587	NA	100.395109536387
2010-05-24	73.710292000956	11.4597394828691	NA	73.645362194735
2010-05-25	4.22902938718943	2.44616060460193	NA	4.22143290375228
2010-05-26	100.159253208527	98.3843423320737	NA	100.159422079885
2010-05-27	79.3467329167084	89.0968573387821	NA	79.3417211032809
2010-05-28	85.7390398914344	66.0195955091203	NA	85.7781984478913
2010-05-29	87.1797624007239	65.8981117895746	NA	87.218337497544
2010-05-30	100.388112929695	71.3496527614802	NA	100.395109536387
2010-05-31	73.710292000956	11.4581905724258	NA	73.645362194735

R Script

```r
# load data
atm_data <- read_excel("data/ATM624Data.xlsx")

# clean dataframe
atm <- atm_data %>%
  # create wide dataframe
  spread(ATM, Cash) %>%
  # remove NA column using function from janitor package
  remove_empty(which = "cols") %>%
  # filter unobserved values from May 2010
  filter(DATE < as.Date("2010-05-01")) %>%
  # ensure dates are ascending
  arrange(DATE)

atm$ATM2[is.na(atm$ATM2)] <- mean(atm$ATM2, na.rm = TRUE) ## remove NA
atm$ATM4[which.max(atm$ATM4)] <- mean(atm$ATM4, na.rm = TRUE) ## remove outlier

# create TS with weekly frequency & subset data
atm_ts <- atm %>% select(-DATE) %>% ts(start=1,  frequency = 7)
ATM1_ts <- atm_ts[,1]; ATM2_ts <- atm_ts[,2]; ATM4_ts <- atm_ts[,4]

#unit root test
## no diff
ATM1_ur <-ur.kpss(ATM1_ts)
ATM2_ur <-ur.kpss(ATM2_ts)
ATM4_ur <-ur.kpss(ATM4_ts)
## first order diff
ATM1d_ur <-ur.kpss(diff(ATM1_ts, lag=7))
ATM2d_ur <-ur.kpss(diff(ATM2_ts, lag=7))
ATM4d_ur <-ur.kpss(diff(ATM4_ts, lag=7))

# AUTO.ARIMA function; set D=1 for seasonal differencing
ATM1_AA <-auto.arima(ATM1_ts, D = 1, lambda = "auto", approximation = F, stepwise = T)
ATM2_AA <-auto.arima(ATM2_ts, D = 1, lambda = "auto", approximation = F, stepwise = T)
ATM4_AA <-auto.arima(ATM4_ts, D = 1, lambda = "auto", approximation = F, stepwise = T)

# Forecast Results
ATM1_fc <- forecast(ATM1_AA,h=31)
ATM2_fc <- forecast(ATM2_AA,h=31)
ATM4_fc <- forecast(ATM4_AA,h=31)

# Revert results back into original form
date <- as.character(seq(as.Date('2010-05-01'), length.out=31, by=1))
ATM_FC <-  cbind("Date"=date, "ATM1"=ATM1_fc$mean, "ATM2"=ATM2_fc$mean,
                 "ATM3"=c(NA,NA,NA,NA),"ATM4"=ATM4_fc$mean) %>% as.data.frame()

# Save output
write.csv(ATM_FC, file="forecasts/ATM_ARIMA_FC.csv")
```