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:

```r
# General
library('easypackages')

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

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

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

# 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.

```r
# 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")
```

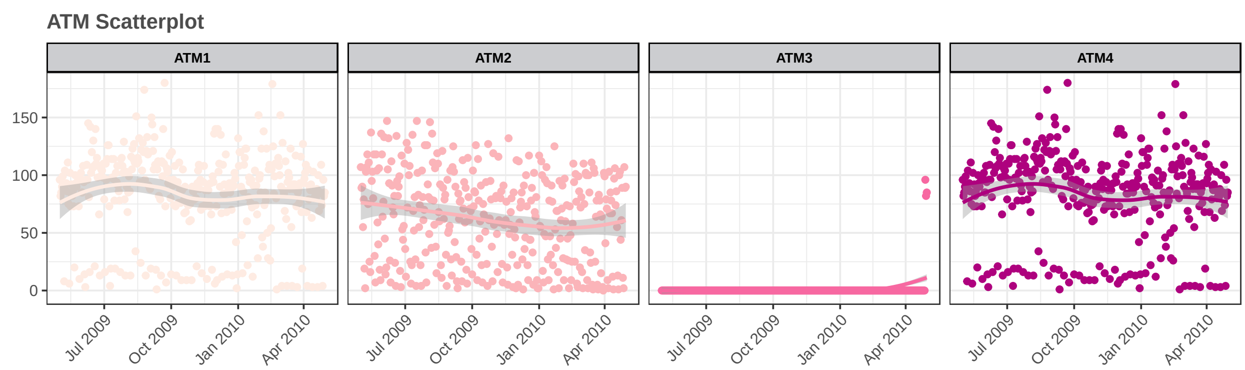
1 Part A

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

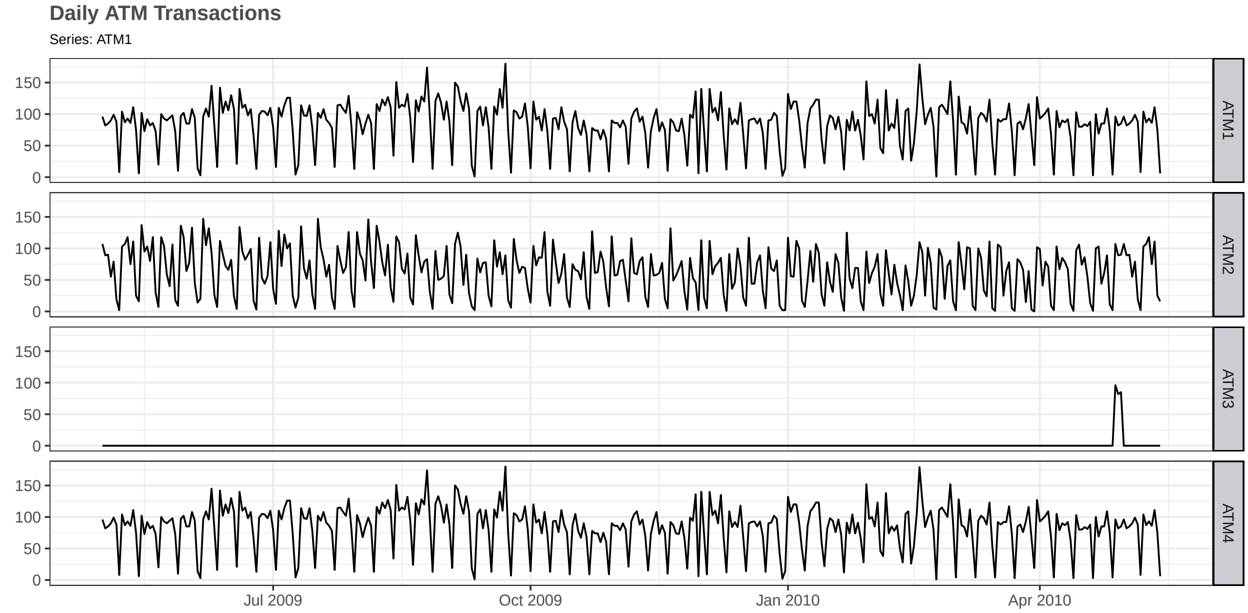
Through data exploration, we identified that the original data file contained `NA` values in our `ATM` and `Cash` columns for 14 observations in May 2010. We removed these missing values and transformed the dataset into a wide format. Our cleaned dataframe was then converted into a timeseries format 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.

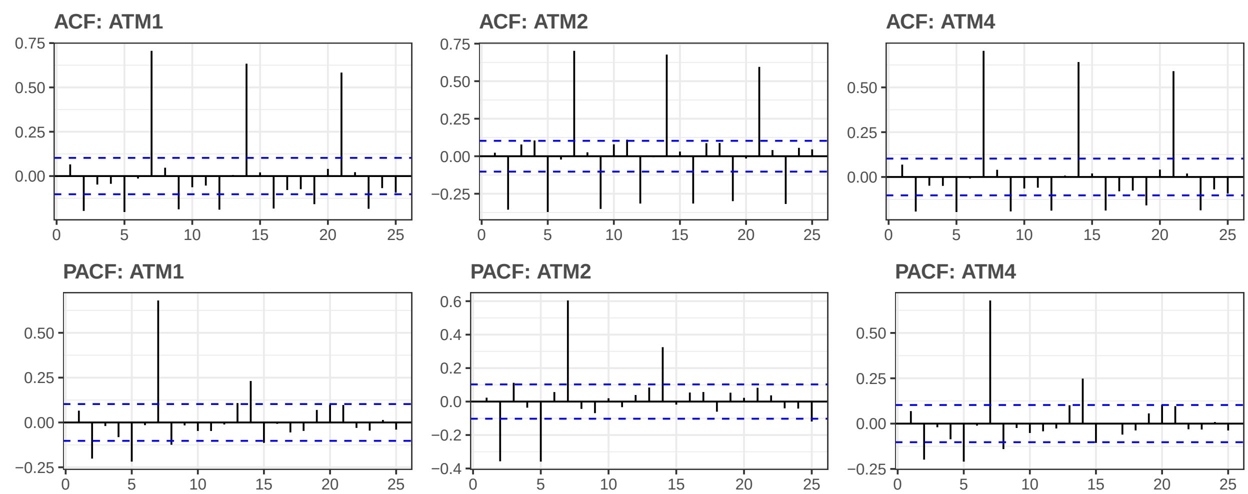


1.2 Timeseries Plots

As mentioned in our data exploration, the 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 focus on evaluating, modeling, and forecasting transactions for only the ATM1, ATM2, and ATM4 series.



1.2.1 Evaluation



Our ACF plots for each ATM showcases three large, decreasing lags at 7, 14, and 21. This confirms our assumption about seasonality within our observed data as these lags are indicative of a weekly pattern. These plot suggests our data is non-stationary, thus we performed a unit root test using the `ur.kpss()` function and determined differencing was required on all three series.

Table 1.1: KPSS unit root test

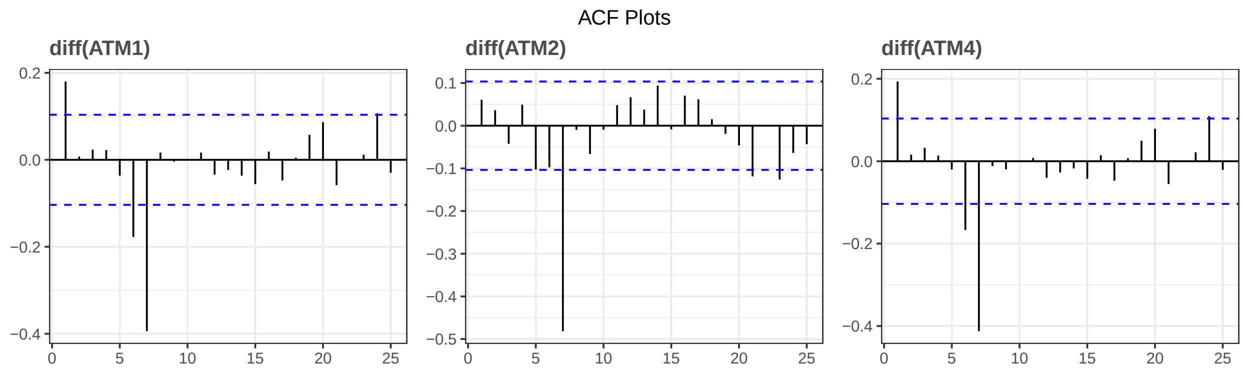
ATM	Root Test	Diff Root Test
ATM1	0.4967	0.0219
ATM2	2.0006	0.016
ATM4	0.5182	0.0211

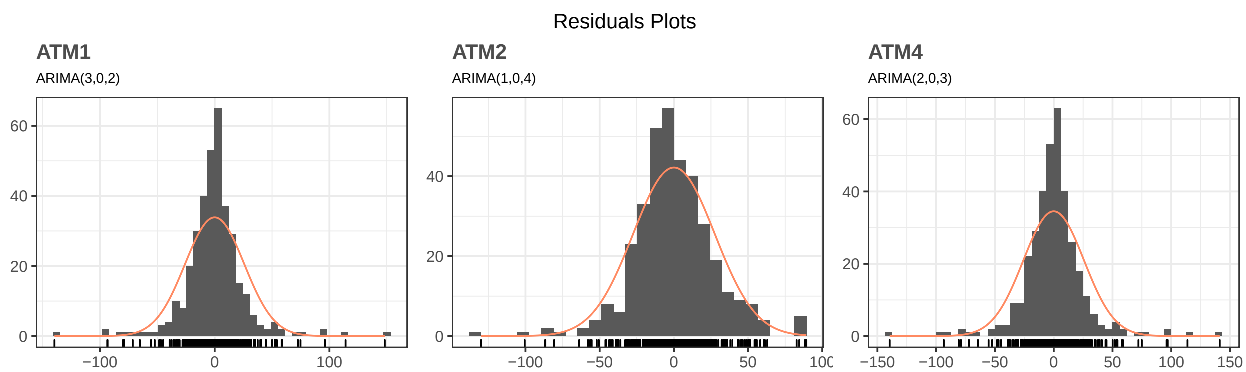
1.2.2 Modeling

We used `auto.arima()` on our differenced data to select the best ARIMA model for our series. The following models were selected for our series:

- **ATM1**: ARIMA(2,0,3) with zero mean
- **ATM2**: ARIMA(1,0,4) with zero mean
- **ATM4**: ARIMA(2,0,3) with zero mean

The following ACF plots show us that our differentiated data is now stationary and the residual histograms confirm that the model adequately fits the observed data.



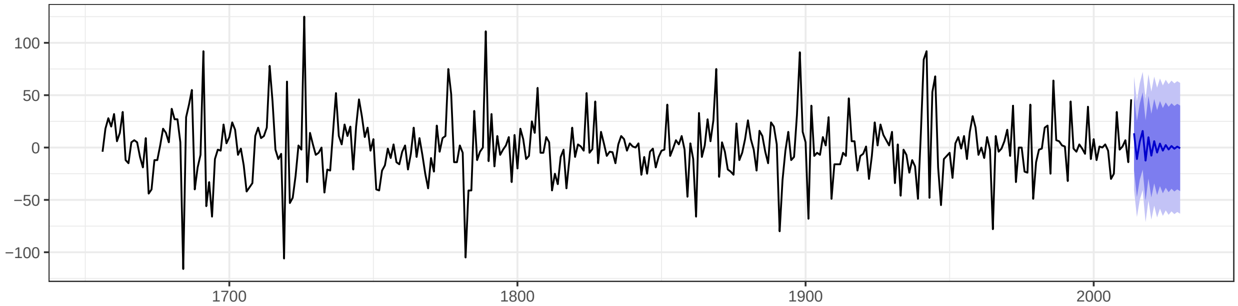


1.2.3 Forecast

Finally, we applied a forecast to each series for the remaining 17 days in May. The full forecasts can be viewed in the appendix section and are also located within our data output folder.

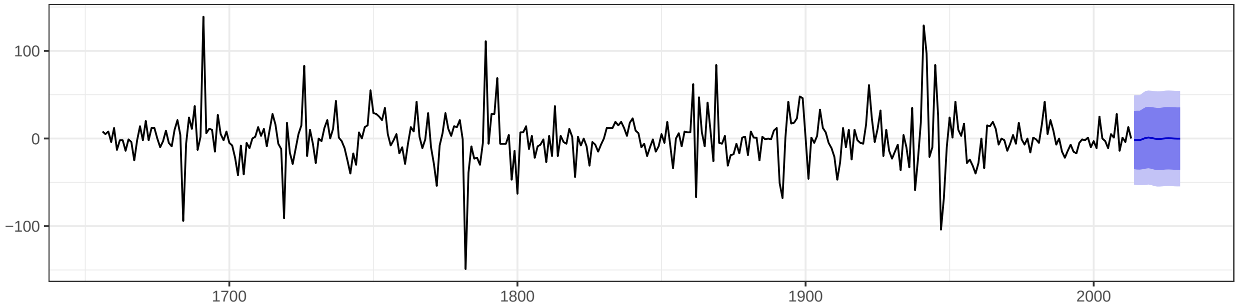
Forecasts from ARIMA(1,0,4) with zero mean

ATM2 Series



Forecasts from ARIMA(2,0,3) with zero mean

ATM4 Series



Appendix

Part A

ATM1 Forecast

Table 1.2: ATM1 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	-1.5476468	-34.83474	31.73945	-52.45586	49.36057
2015	-2.2246957	-35.76049	31.31110	-53.51327	49.06388
2016	-2.4082528	-35.98207	31.16556	-53.75497	48.93847
2017	-0.9095920	-35.28357	33.46439	-53.48006	51.66087
2018	0.7491377	-34.21898	35.71725	-52.72998	54.22825
2019	1.3990794	-33.59777	36.39593	-52.12398	54.92214
2020	0.9070426	-34.22027	36.03435	-52.81554	54.62963
2021	-0.0484768	-35.41166	35.31471	-54.13180	54.03485
2022	-0.6816097	-36.10459	34.74137	-54.85638	53.49316
2023	-0.6624510	-36.09316	34.76826	-54.84905	53.52414
2024	-0.2016687	-35.69998	35.29664	-54.49165	54.08831
2025	0.2556483	-35.28435	35.79565	-54.09809	54.60939
2026	0.4016704	-35.13908	35.94243	-53.95322	54.75656
2027	0.2321338	-35.32165	35.78592	-54.14268	54.60695
2028	-0.0427988	-35.61500	35.52940	-54.44579	54.36019
2029	-0.2054847	-35.78118	35.37021	-54.61382	54.20285
2030	-0.1800679	-35.75693	35.39680	-54.59019	54.23005

ATM2 Forecast

Table 1.3: ATM2 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	13.6179414	-21.82098	49.05686	-40.58121	67.81710
2015	-11.0244123	-47.18551	25.13668	-66.32803	44.27921
2016	5.8052816	-30.57732	42.18788	-49.83710	61.44766
2017	15.8186530	-21.06766	52.70497	-40.59410	72.23141
2018	-12.4667974	-50.90070	25.96711	-71.24638	46.31279
2019	9.8251753	-29.53934	49.18969	-50.37765	70.02800
2020	-7.7432933	-47.67490	32.18831	-68.81342	53.32683
2021	6.1025467	-34.17727	46.38237	-55.50012	67.70521
2022	-4.8094622	-45.30405	35.68513	-66.74059	57.12167
2023	3.7903727	-36.83704	44.41779	-58.34390	65.92464
2024	-2.9872208	-43.69692	37.72248	-65.24733	59.27289
2025	2.3542508	-38.40647	43.11497	-59.98389	64.69239
2026	-1.8554025	-42.64778	38.93698	-64.24196	60.53116
2027	1.4622564	-39.34977	42.27429	-60.95436	63.87887
2028	-1.1524151	-41.97665	39.67182	-63.58769	61.28286
2029	0.9082269	-39.92358	41.74003	-61.53863	63.35509
2030	-0.7157803	-41.55229	40.12073	-63.16984	61.73828

ATM4 Forecast

Table 1.4: ATM4 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	-1.7397027	-35.02165	31.54225	-52.64005	49.16065
2015	-1.8523041	-35.43051	31.72590	-53.20573	49.50112
2016	-1.8915445	-35.49987	31.71678	-53.29104	49.50795
2017	-0.6521306	-35.06829	33.76403	-53.28710	51.98284
2018	0.6409033	-34.40024	35.68205	-52.94991	54.23171
2019	1.1052058	-33.97279	36.18320	-52.54196	54.75237
2020	0.6860038	-34.51254	35.88455	-53.14553	54.51754
2021	-0.0629775	-35.49761	35.37166	-54.25558	54.12962
2022	-0.5383252	-36.03872	34.96207	-54.83149	53.75484
2023	-0.5063232	-36.01206	34.99941	-54.80766	53.79501
2024	-0.1451906	-35.71422	35.42384	-54.54333	54.25295
2025	0.2015531	-35.41007	35.81317	-54.26172	54.66483
2026	0.3061844	-35.30681	35.91918	-54.15919	54.77156
2027	0.1734676	-35.45049	35.79742	-54.30867	54.65560
2028	-0.0343836	-35.67590	35.60713	-54.54337	54.47461
2029	-0.1549780	-35.80046	35.49050	-54.67004	54.36008
2030	-0.1345348	-35.78076	35.51169	-54.65073	54.38166

R Script

```
#-----DEPENDENCIES-----#
```

```r
library(readxl); library(tidyverse); library(janitor);
library(zoo); library(urca); library(forecast)

#-----PRE-PROCESSING-----#

# 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)

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

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

# create zoo time series
atm_zoo <- atm %>%
  # remove column & generate date in timeseries using zoo
  select(-DATE) %>%
  # generate ts using zoo
  zoo(seq(from = as.Date("2009-05-01"), to = as.Date("2010-05-14"), by = 1))

# create standard time series
atm_ts <- atm %>%
  # remove column & generate date in timeseries using zoo
  select(-DATE) %>%
  # generate ts using zoo
  ts(end=c(2010,4))

#-----ATM-1-----#

#subset data
ATM1_zoo <- atm_zoo[,1]
ATM1_ts <- atm_ts[,1]

#differentiated
ATM1d <-  diff(ATM1_ts, lag=7)

#unit root test
ATM1_ur <-ur.kpss(ATM1_ts)
ATM1d_ur <-ur.kpss(ATM1d)

# ARIMA
```

```
ATM1_arima <- auto.arima(ATM1d, seasonal=F, stepwise=F, approximation=F)

# Forecast
ATM1_fc <- ATM1_arima %>% forecast(h=17)

# Save output
write.csv(ATM1_fc, file="ATM1_Forecast.csv")

#-----ATM-2-----#

#subset data
ATM2_zoo <- atm_zoo[,2]
ATM2_ts <- atm_ts[,2]

#differentiated
ATM2d <-  diff(ATM2_ts, lag=7)

#unit root test
ATM2_ur <-ur.kpss(ATM2_ts)
ATM2d_ur <-ur.kpss(ATM2d)

# ARIMA
ATM2_arima <- auto.arima(ATM2d, seasonal=F, stepwise=F, approximation=F)

# Forecast
ATM2_fc <- ATM2_arima %>% forecast(h=17)

# Save output
write.csv(ATM2_fc, file="ATM2_Forecast.csv")

#-----ATM-4-----#

#subset data
ATM4_zoo <- atm_zoo[,4]
ATM4_ts <- atm_ts[,4]

#differentiated
ATM4d <-  diff(ATM4_ts, lag=7)

#unit root test
ATM4_ur <-ur.kpss(ATM4_ts)
ATM4d_ur <-ur.kpss(ATM4d)

# ARIMA
ATM4_arima <- auto.arima(ATM4d, seasonal=F, stepwise=F, approximation=F)

# Forecast
ATM4_fc <- ATM4_arima %>% forecast(h=17)

# Save output
write.csv(ATM4_fc, file="ATM4_Forecast.csv")
```