# 590SCA_Homework2

*Da-Yae Frail, Sin Gwoo Kang, Xiao Shi, Yi Wu, Xiaoyan Yang, Xin Yuan*

*April 22, 2016*

- 1
- 2 & 3
- 4
- 5
- 6

```
set.seed(654265269)
```

## 1

The following code split the data into test set and train set.
Ideal split ratio of a sample is 80% for the train sample and 20% for the test sample.
Total sample size is equal to 456. 20% of this value equals 91.2.
After rounding, size of train sample was set to 91 while size of test sample was set to 365.

```
d = read.csv("C:/Users/Dayae Frail/Desktop/590/BikeDemandDaily.csv", header=TRUE)
n = dim(d)[1];
ind = sample.int(n, size=91);
dtrain = d[-ind,];
dtest = d[ind,];
```

## 2 & 3

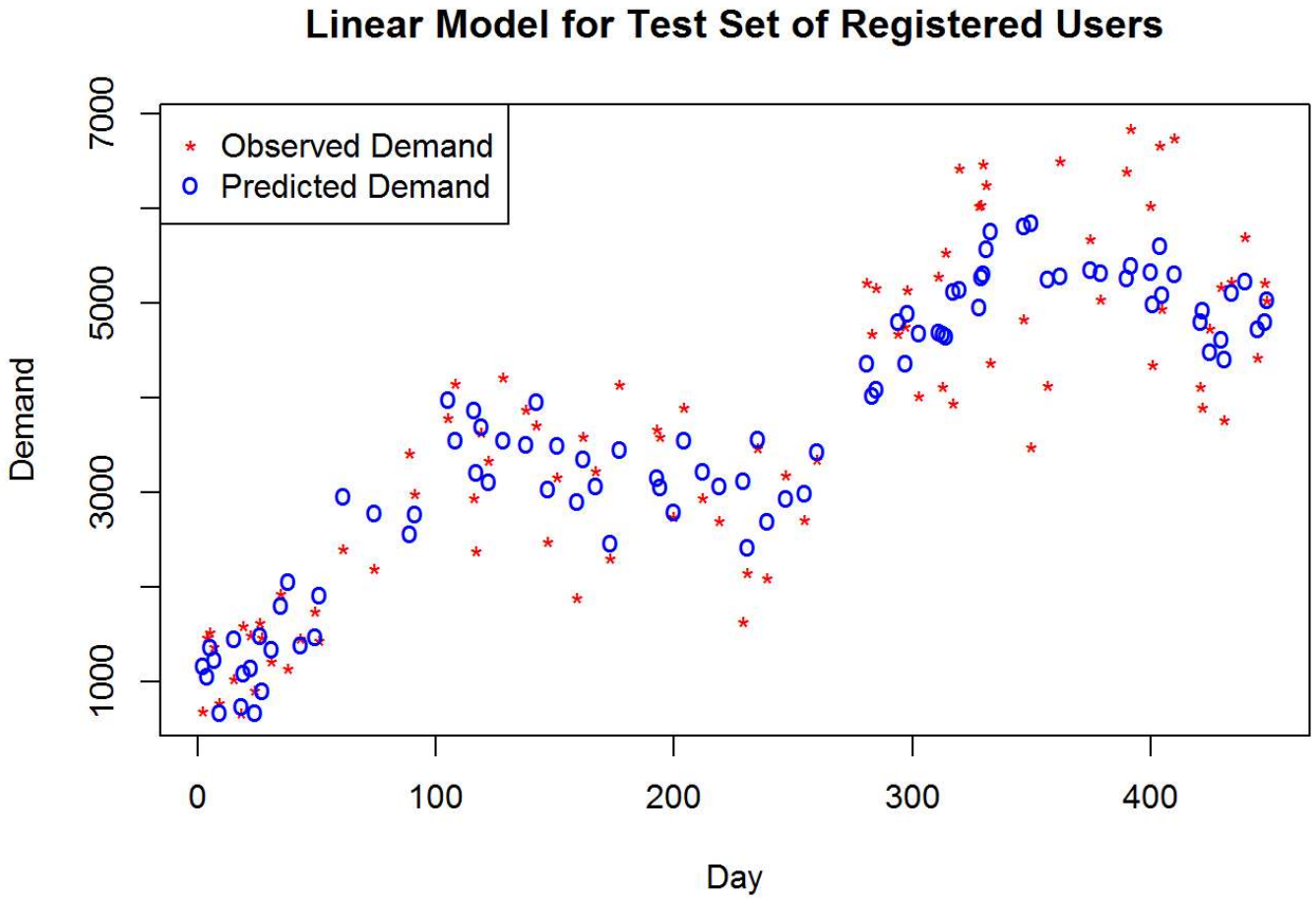
The following code is the user defined code to calculate Root Mean Square Prediction Error (RMSPE).

```
RMSPE=function (d, p){
  y1=mean((d-p)^2);
  y2=sqrt(y1)
  return(y2);
}
```

### Linear model

The following code estimates the demand forecast for the test set of registered users using linear model and calculates its RMSPE.

```
m1 = lm(Registered~Index+as.factor(season)+as.factor(holiday)+meanatemp+meanwindspeed+mea
nhumidity, data=dtrain);
p1 = predict(m1, newdata=dtest)
plot(dtest$Index, dtest$Registered, pch="*", col=2, xlab="Day", ylab="Demand",main="Linea
r Model for Test Set of Registered Users");
points(dtest$Index, p1, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```

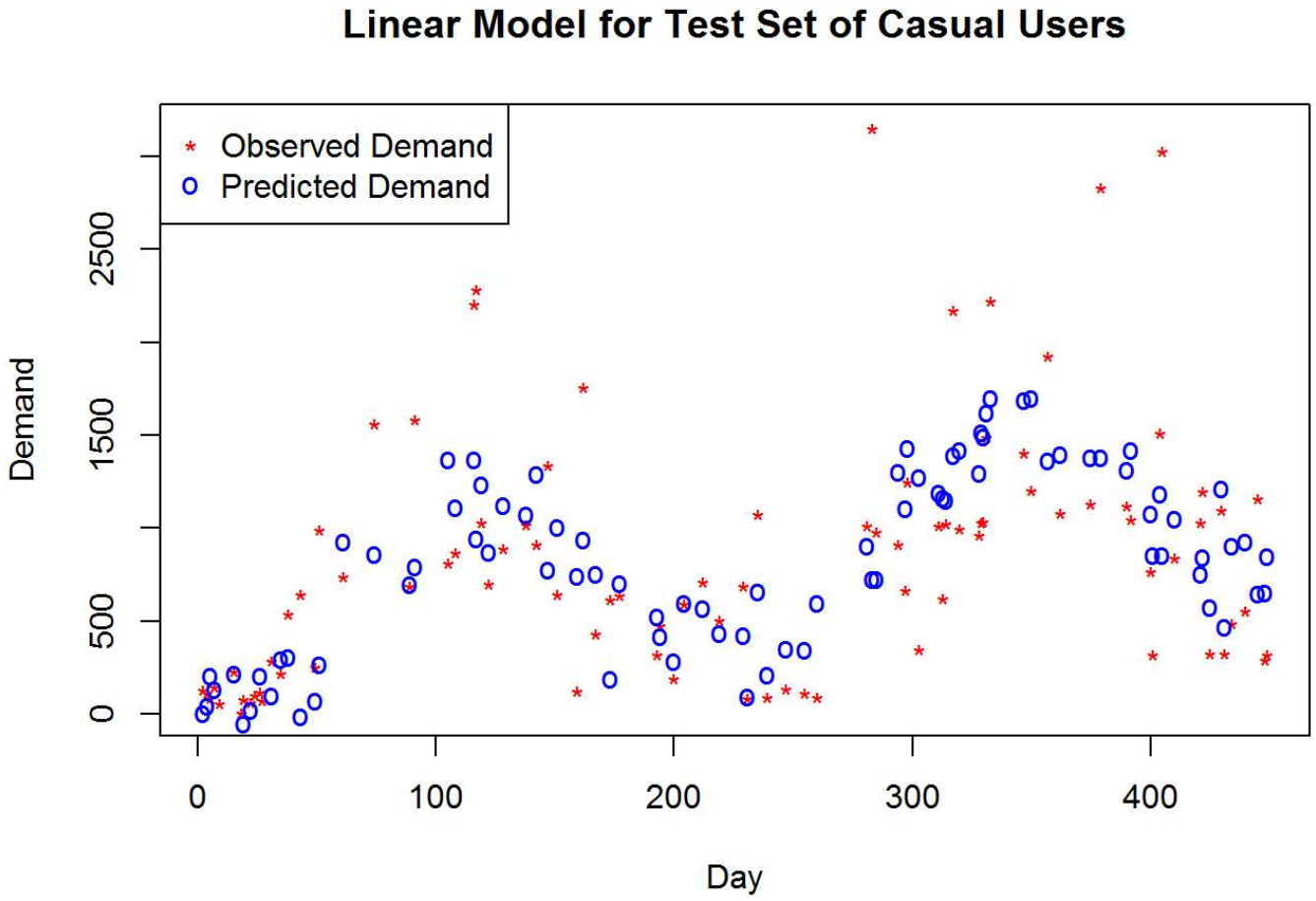



```
RMSPE(dtest$Registered, p1)
```

```
## [1] 705.1268
```

The following code estimates the demand forecast for the test set of casual users using linear model and calculates its RMSPE.

```
m2 = lm(Casual~Index+as.factor(season)+as.factor(holiday)+meanatemp+meanwindspeed+meanhum
idity, data=dtrain);
p2 = predict(m2, newdata=dtest)
plot(dtest$Index, dtest$Casual, pch="*", col=2, xlab="Day", ylab="Demand",main="Linear Mo
del for Test Set of Casual Users");
points(dtest$Index, p2, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(dtest$Casual, p2)
```

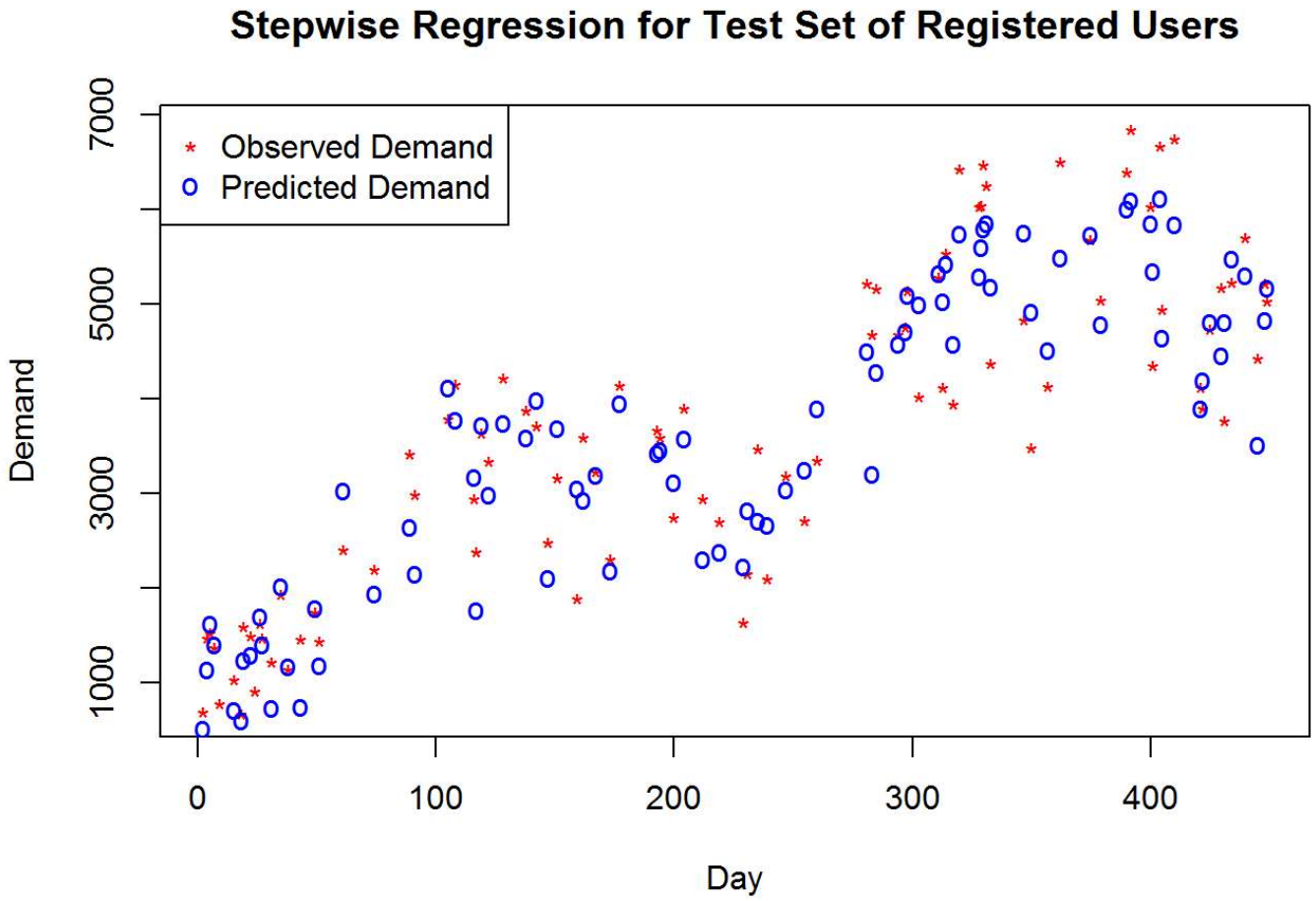
```
## [1] 545.1149
```

## Stepwise

```
m3 = glm(Registered~1, data=dtrain, family="gaussian")
m4 = glm(Registered~Index+year+as.factor(month)+as.factor(day)+as.factor(season)+as.factor(holiday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidity+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspeed, data=dtrain, family="gaussian")
s1 = step(m3, scope=list(lower=m3, upper=m4), direction="forward");
summary(s1)
```

The following code estimates the demand forecast for the test set of registered users using stepwise regression and calculates its RMSPE.

```
p3=predict(s1, newdata=dtest);
plot(dtest$Index, dtest$Registered, pch="*", col=2, xlab="Day", ylab="Demand",main="Stepwise Regression for Test Set of Registered Users");
points(dtest$Index, p3, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c(2,4))
```



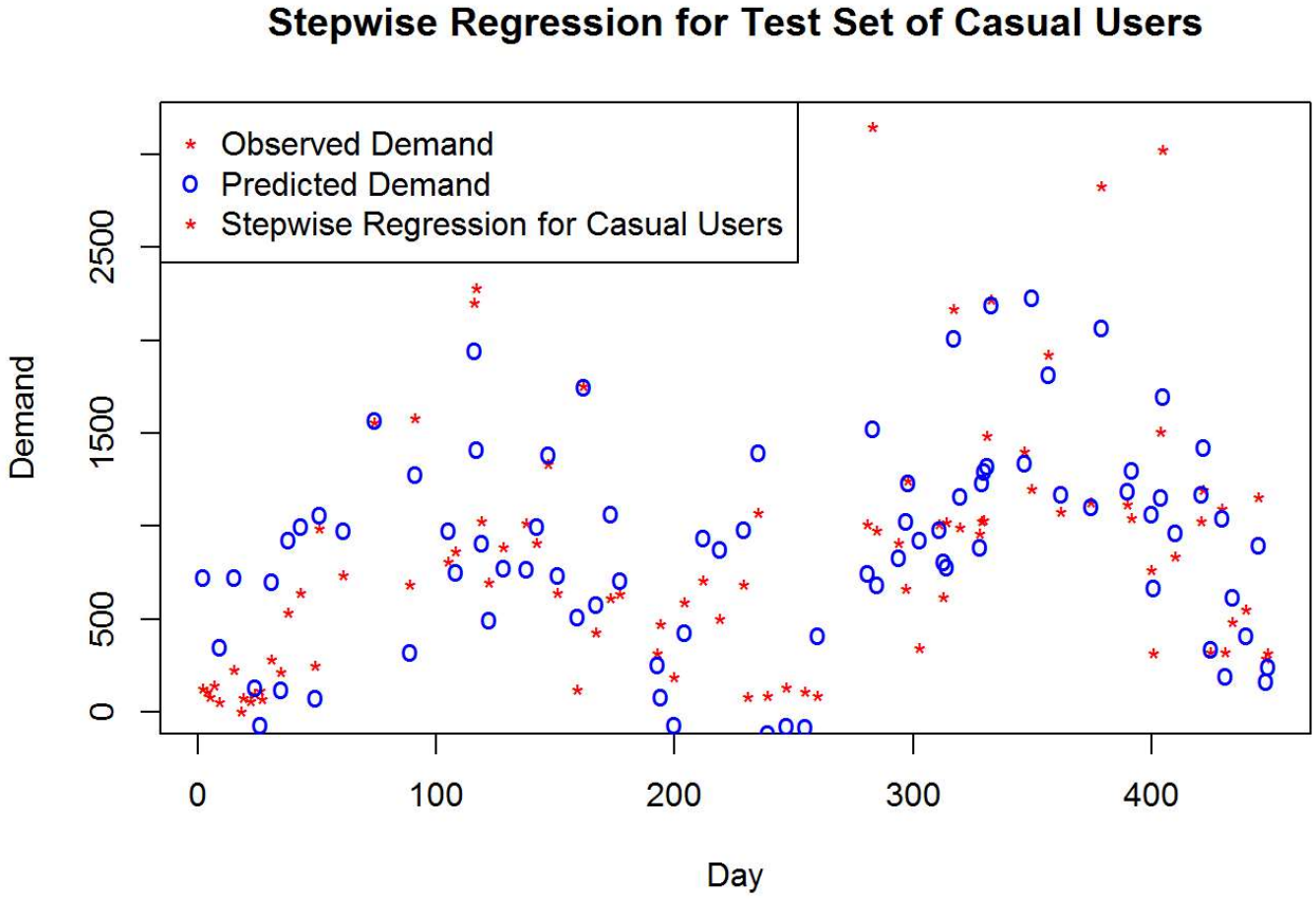

```
RMSPE(dtest$Registered, p3);
```

```
## [1] 567.1344
```

```
m5 = glm(Casual~1, data=dtrain, family="gaussian")
m6 = glm(Casual~Index+year+as.factor(month)+as.factor(day)+as.factor(season)+as.factor(ho
liday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidity+maxhumidity
+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspeed, data=dtrain,
family="gaussian")
s2 = step(m5, scope=list(lower=m5, upper=m6), direction="forward");
summary(s2)
```

The following code estimates the demand forecast for the test set of casual users using stepwise regression and calculates its RMSPE.

```
p4=predict(s2, newdata=dtest);
plot(dtest$Index, dtest$Casual, pch="*", col=2, xlab="Day", ylab="Demand",main="Stepwise
Regression for Test Set of Casual Users");
points(dtest$Index, p4, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand",main="Stepwise Regressio
n for Casual Users"), pch=c("*","o"), col=c(2,4))
```




```
RMSPE(dtest$Casual, p4);
```

```
## [1] 380.3294
```

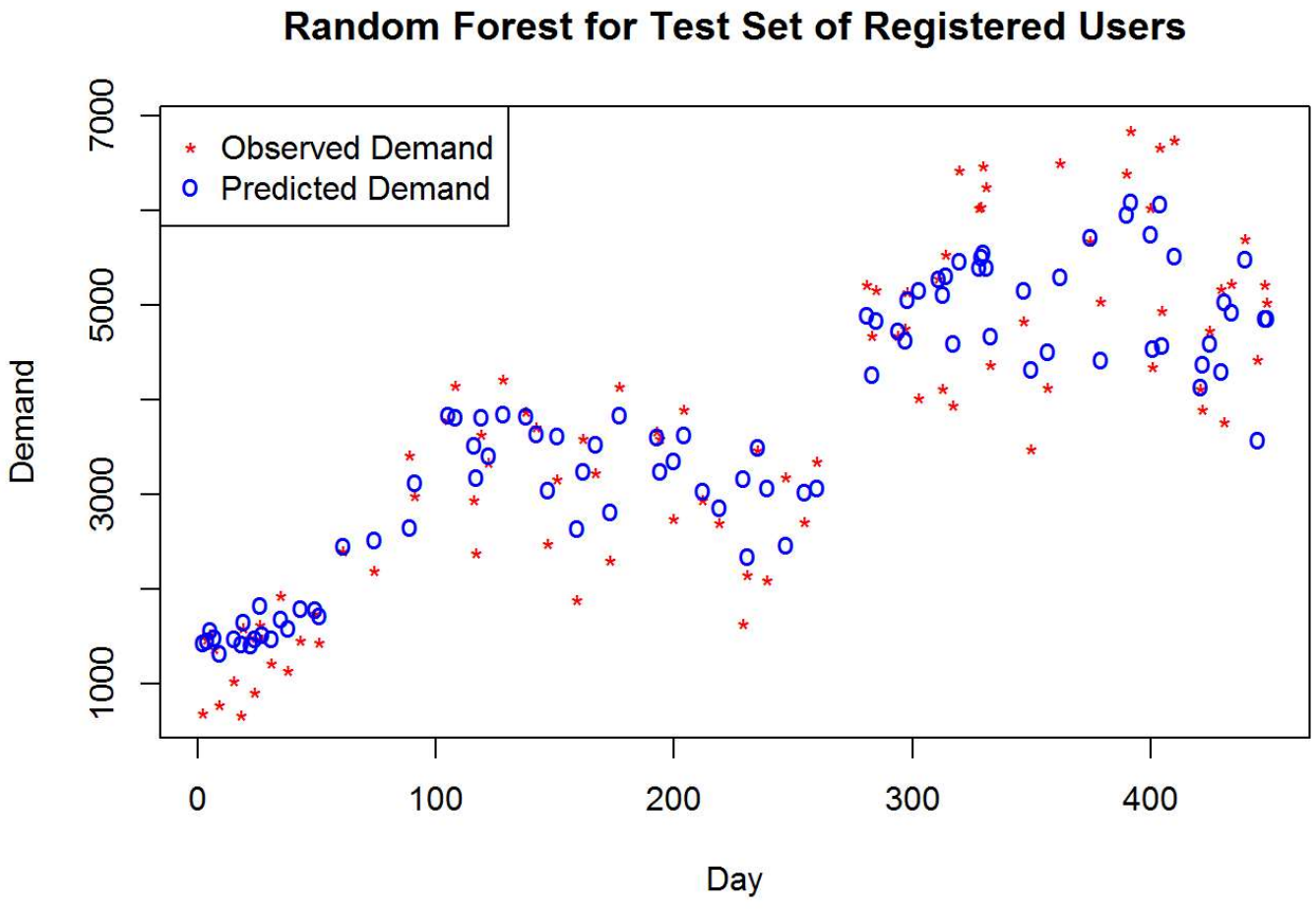
## Random forest

```
library(randomForest)
dtrain$month=as.factor(dtrain$month);
dtrain$day=as.factor(dtrain$day);
dtrain$season=as.factor(dtrain$season);
dtrain$holiday=as.factor(dtrain$holiday);
dtrain$workingday=as.factor(dtrain$workingday);
```

The following code estimates the demand forecast for the test set of registered users using random forest and calculates its RMSPE.

```
r1 = randomForest(Registered~Index+year+month+day+season+holiday+workingday+meanatemp+max
atemp+minatemp+sdatemp+meanhumidity+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxw
indspeed+minwindspeed+sdwindspeed, data=dtrain, ntree=500,do.trace=1, importance=TRUE, pr
oximity=TRUE);
dtest$month=as.factor(dtest$month);
dtest$day=as.factor(dtest$day);
dtest$season=as.factor(dtest$season);
dtest$holiday=as.factor(dtest$holiday);
dtest$workingday=as.factor(dtest$workingday);
```

```
p5 = predict(r1, newdata=dtest, type="response");
plot(dtest$Index, dtest$Registered, pch="*", col=2, xlab="Day", ylab="Demand", main="Rand
om Forest for Test Set of Registered Users");
points(dtest$Index, p5, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```



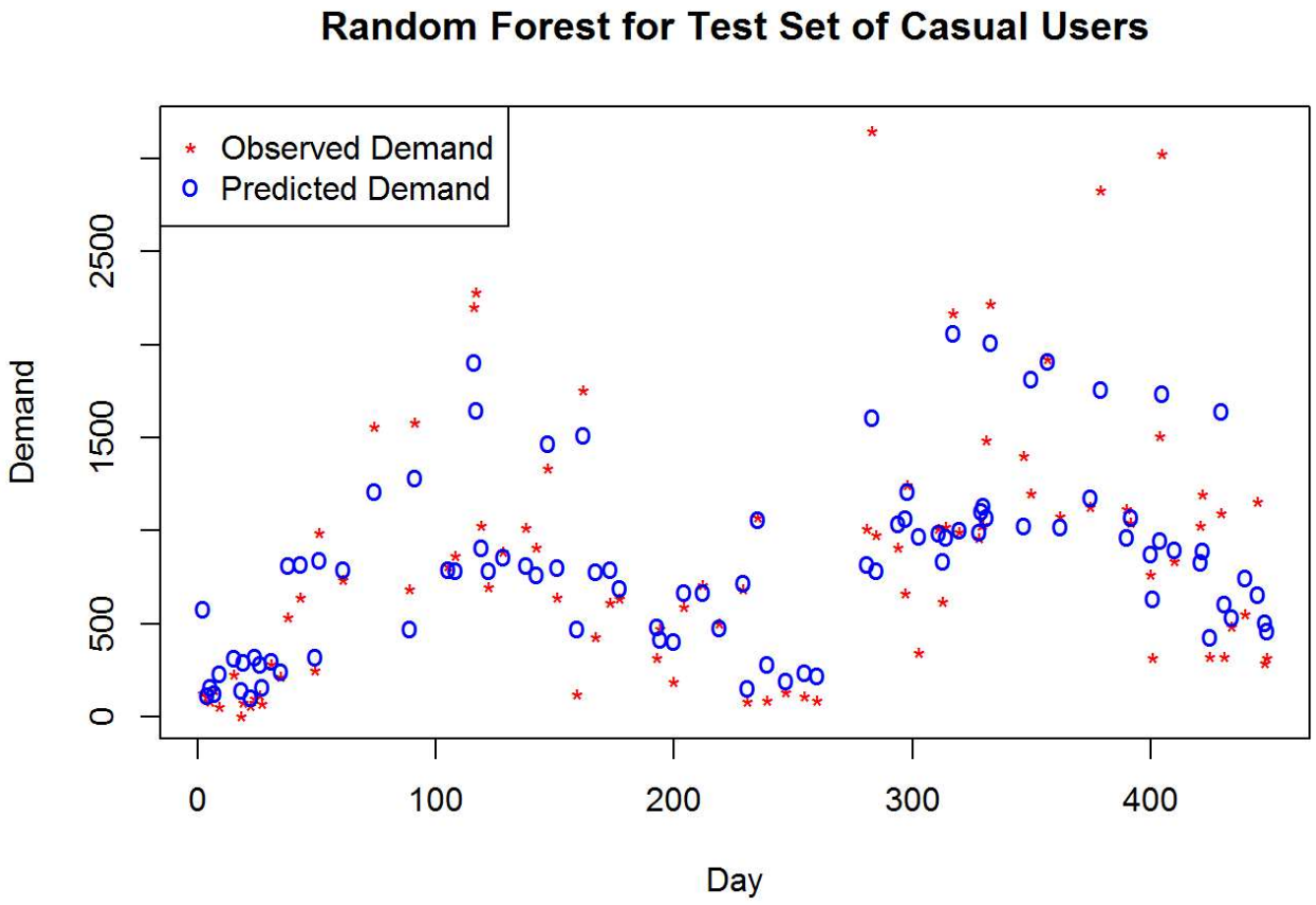

```
RMSPE(dtest$Registered, p5);
```

```
## [1] 540.6766
```

The following code estimates the demand forecast for the test set of casual users using random forest and calculates its RMSPE.

```
r2 = randomForest(Casual~Index+year+month+day+season+holiday+workingday+meanatemp+maxatem
p+minatemp+sdatemp+meanhumidity+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxwinds
peed+minwindspeed+sdwindspeed, data=dtrain, ntree=500,do.trace=1, importance=TRUE, proxim
ity=TRUE);
```

```
p6 = predict(r2, newdata=dtest, type="response");
plot(dtest$Index, dtest$Casual, pch="*", col=2, xlab="Day", ylab="Demand", main="Random F
orest for Test Set of Casual Users");
points(dtest$Index, p6, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(dtest$Casual, p6);
```

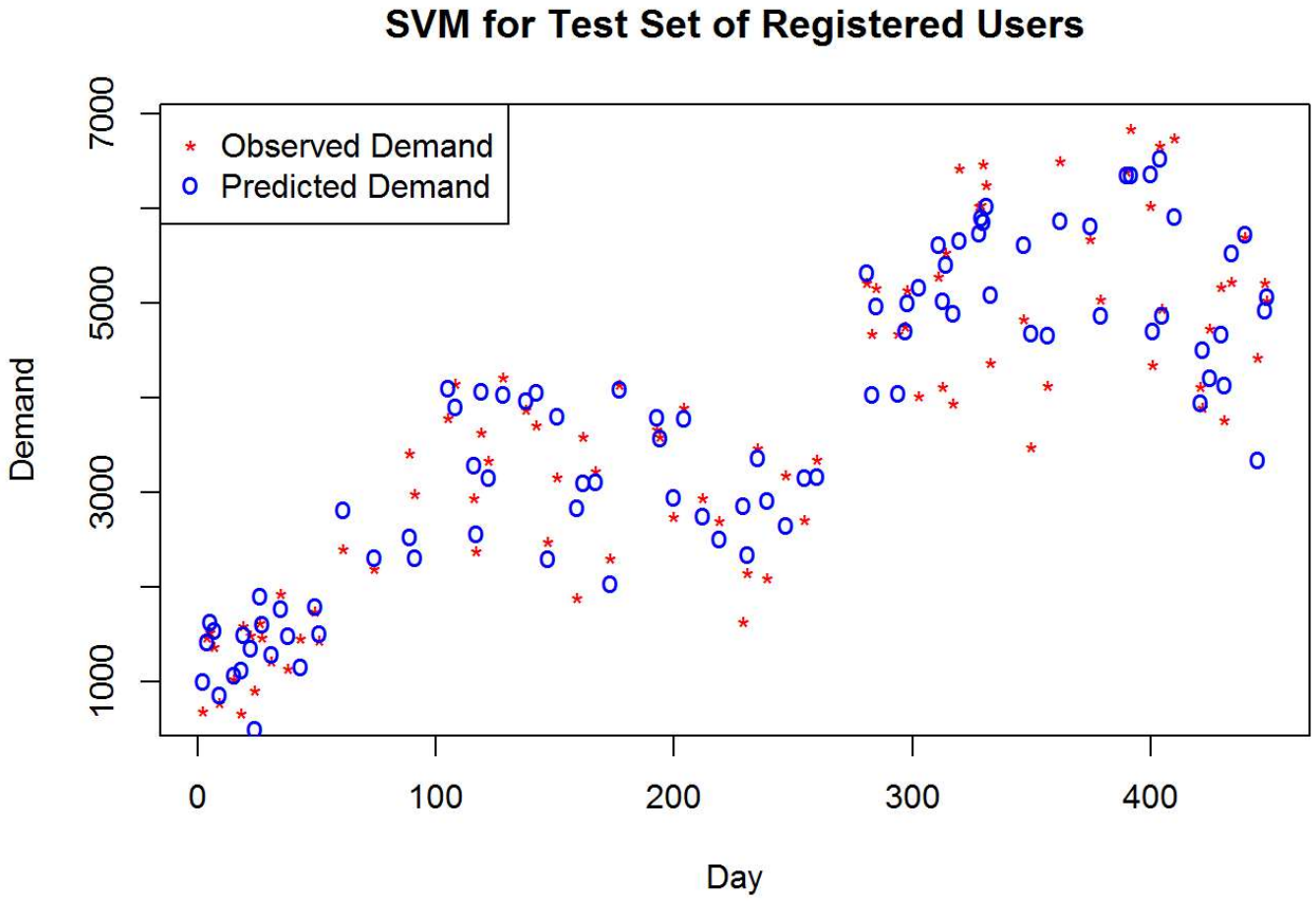
```
## [1] 330.2218
```

## SVM

```
library(e1071)
for2 = as.formula(Registered~Index+year+as.factor(month)+as.factor(day)+as.factor(season)
+as.factor(holiday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidit
y+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspee
d);
for3 = as.formula(Casual~Index+year+as.factor(month)+as.factor(day)+as.factor(season)+as.
factor(holiday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidity+ma
xhumidity+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspeed);
```

The following code estimates the SVM for the test set of registered users using SVM and calculates its RMSPE.

```
s3 = svm(for2, data=dtrain)
p7 = predict(s3, newdata=dtest, type="response")
plot(dtest$Index, dtest$Registered, pch="*", col=2, xlab="Day", ylab="Demand", main="SVM
for Test Set of Registered Users");
points(dtest$Index, p7, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```

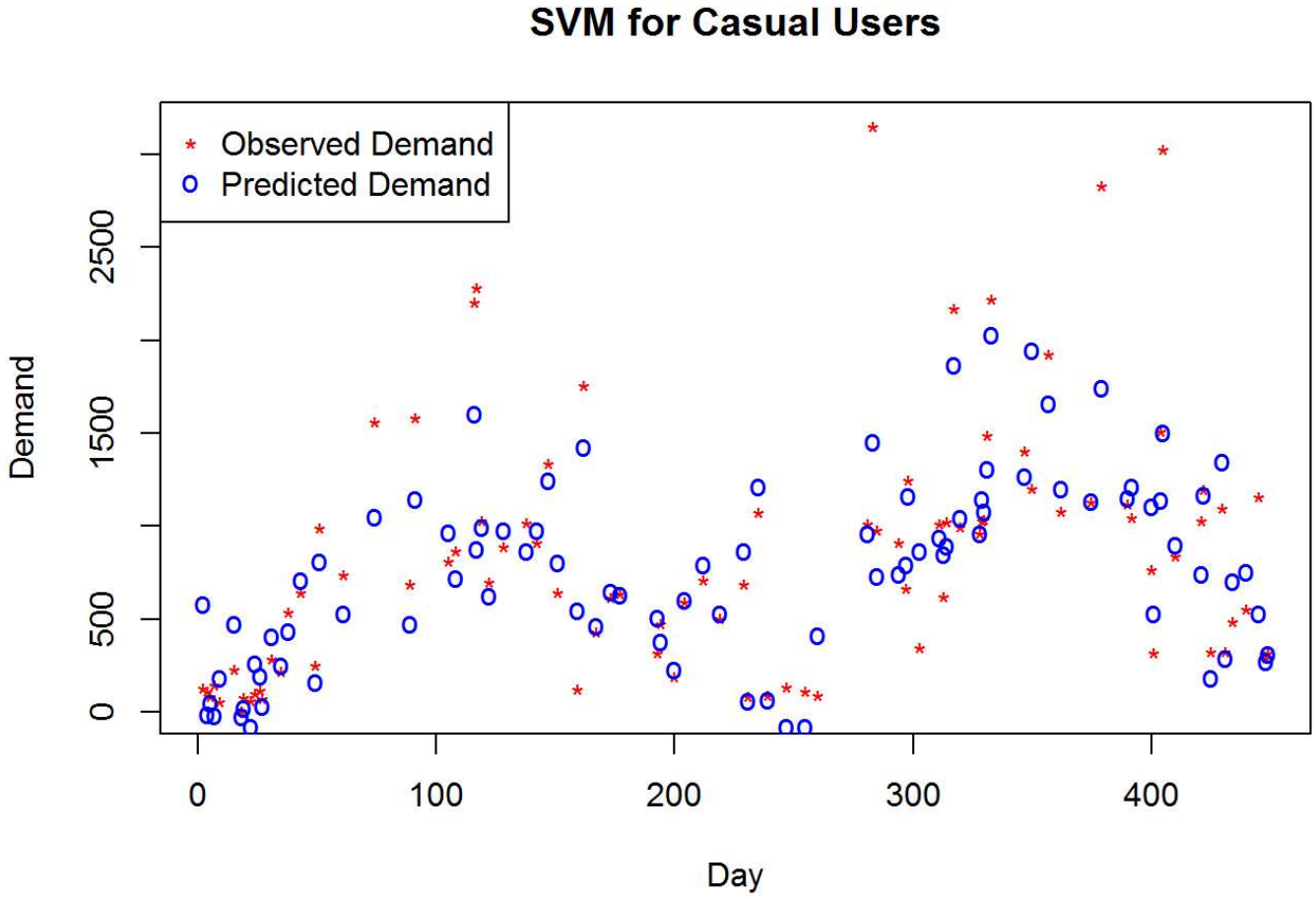



```
RMSPE(dtest$Registered, p7)
```

```
## [1] 464.9823
```

The following code estimates the SVM for the test set of casual users using SVM and calculates its RMSPE.

```
s4 = svm(for3, data=dtrain)
p8 = predict(s4, newdata=dtest, type="response")
plot(dtest$Index, dtest$Casual, pch="*", col=2, xlab="Day", ylab="Demand", main="SVM for
Casual Users");
points(dtest$Index, p8, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(dtest$Casual, p8);
```

```
## [1] 373.8003
```

In the order of linear model, stepwise regression, random forest, and SVM, respectively, the RMSPE values are as follows.
Registered RMSPE: 705.1268495, 567.1343769, 540.6766134, 464.9822894
Mean of Registered RMSPE: 569.4800323
Casual RMSPE: 545.1148654, 380.3294008, 330.2218375, 373.8003188
Mean of Casual RMSPE: 407.3666056
Since average RMSPE value of casual users is 0.2846692% less than that of registered users, the demand of casual users is easier to predict as lower RMSPE value indicates less errors.

# 4

The following code computes RMSPE values for total demand via dis-aggregated method.
The values correspond to linear model, stepwise regression, random forest, and SVM, respectively.

```
RMSPE(dtest$Total, p1+p2)
```

```
## [1] 819.9149
```

```
RMSPE(dtest$Total, p3+p4)
```

```
## [1] 778.1377
```

```
RMSPE(dtest$Total, p5+p6)
```

```
## [1] 713.6342
```

```
RMSPE(dtest$Total, p7+p8)
```

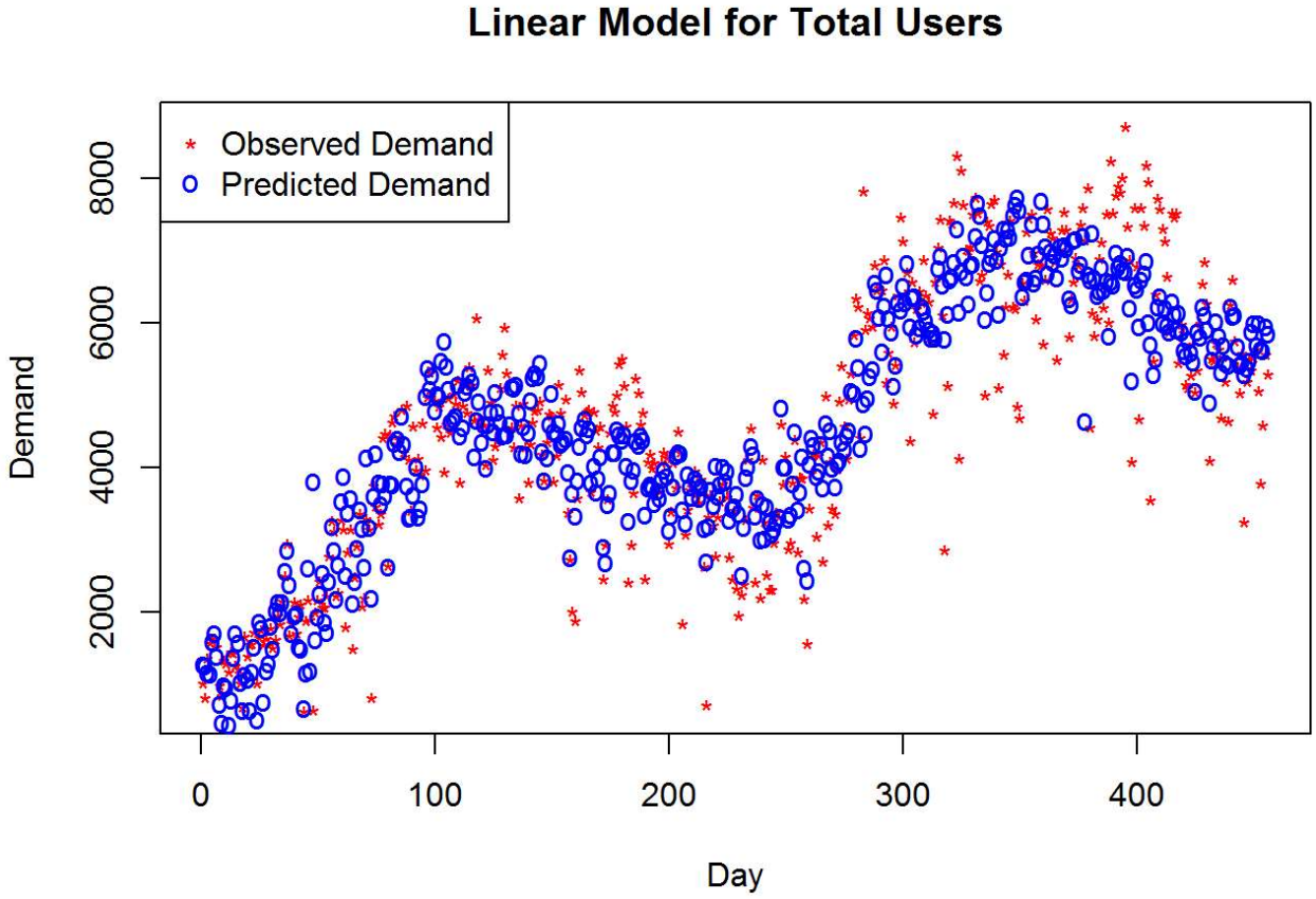
```
## [1] 667.7693
```

# 5

## Linear model

The following code uses linear model for total users and calculates its RMSPE.

```
m9 = lm(Total~Index+as.factor(season)+as.factor(holiday)+meanatemp+meanwindspeed+meanhumi
dity, data=d);
p9 = predict(m9, newdata=d)
plot(d$Index, d$Total, pch="*", col=2, xlab="Day", ylab="Demand", main="Linear Model for
Total Users");
points(d$Index, p9, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(d$Total, p9)
```

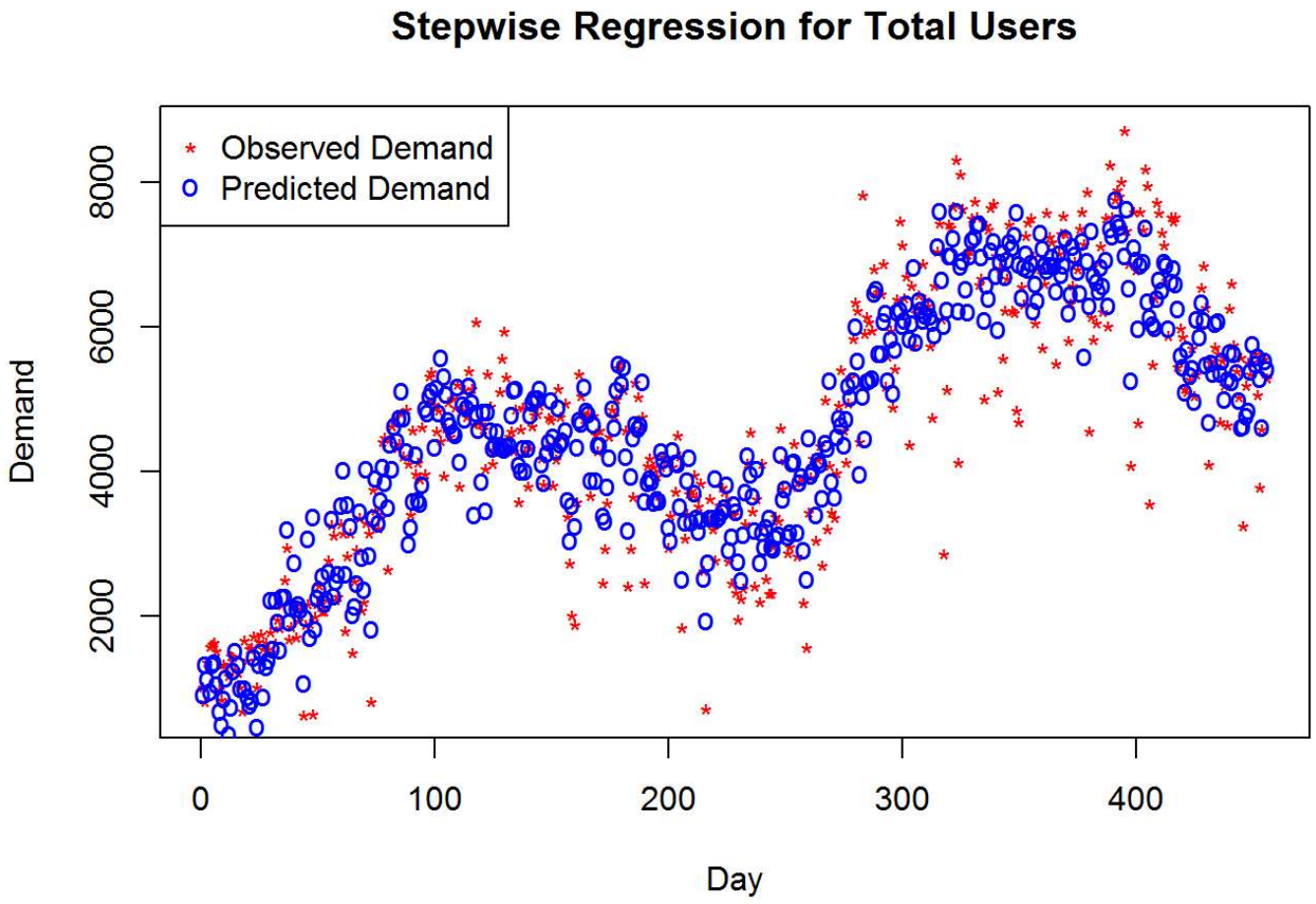
```
## [1] 757.0674
```

## Stepwise Regression

```
m10 = glm(Total~1, data=d, family="gaussian")
m11 = glm(Total~Index+year+as.factor(month)+as.factor(day)+as.factor(season)+as.factor(ho
liday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidity+maxhumidity
+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspeed, data=d, fami
ly="gaussian")
s10 = step(m10, scope=list(lower=m10, upper=m11), direction="forward");
summary(s10)
```

The following code uses stepwise regression for total users and calculates its RMSPE.

```
p10=predict(s10, newdata=d);
plot(d$Index, d$Total, pch="*", col=2, xlab="Day", ylab="Demand", main="Stepwise Regressi
on for Total Users");
points(d$Index, p10, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(d$Total, p10);
```

```
## [1] 677.6463
```

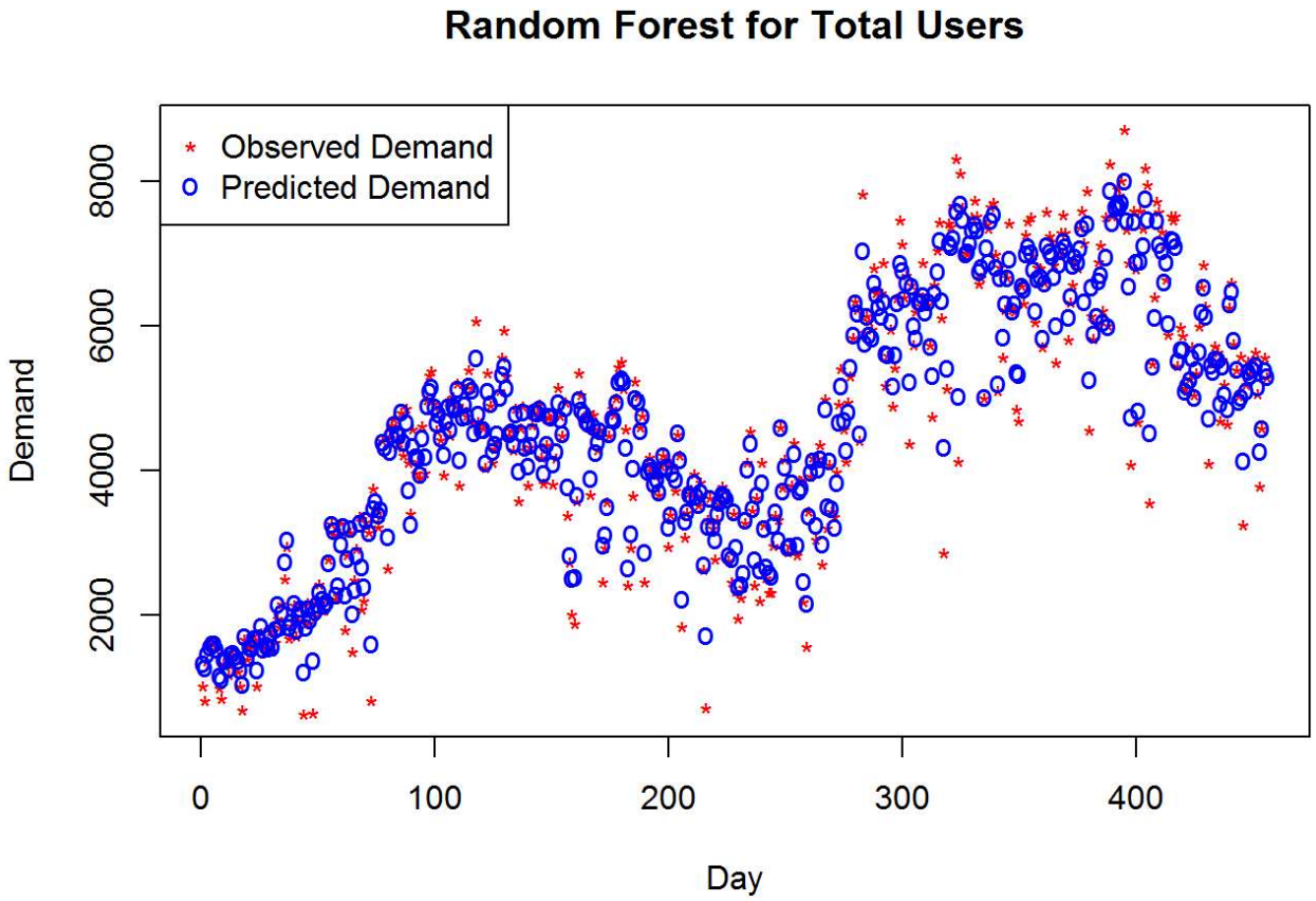
## Random forest

```
library(randomForest)
d$month=as.factor(d$month);
d$day=as.factor(d$day);
d$season=as.factor(d$season);
d$holiday=as.factor(d$holiday);
d$workingday=as.factor(d$workingday);
```

```
r3 = randomForest(Total~Index+year+month+day+season+holiday+workingday+meanatemp+maxatemp
+minatemp+sdatemp+meanhumidity+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxwindsp
eed+minwindspeed+sdwindspeed, data=d, ntree=500,do.trace=1, importance=TRUE, proximity=TR
UE);
```

The following code uses random forest for total users and calculates its RMSPE.

```
p12 = predict(r3, newdata=d, type="response");
plot(d$Index, d$Total, pch="*", col=2, xlab="Day", ylab="Demand", main="Random Forest for
Total Users");
points(d$Index, p12, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c
(2,4))
```




```
RMSPE(d$Total, p12)
```

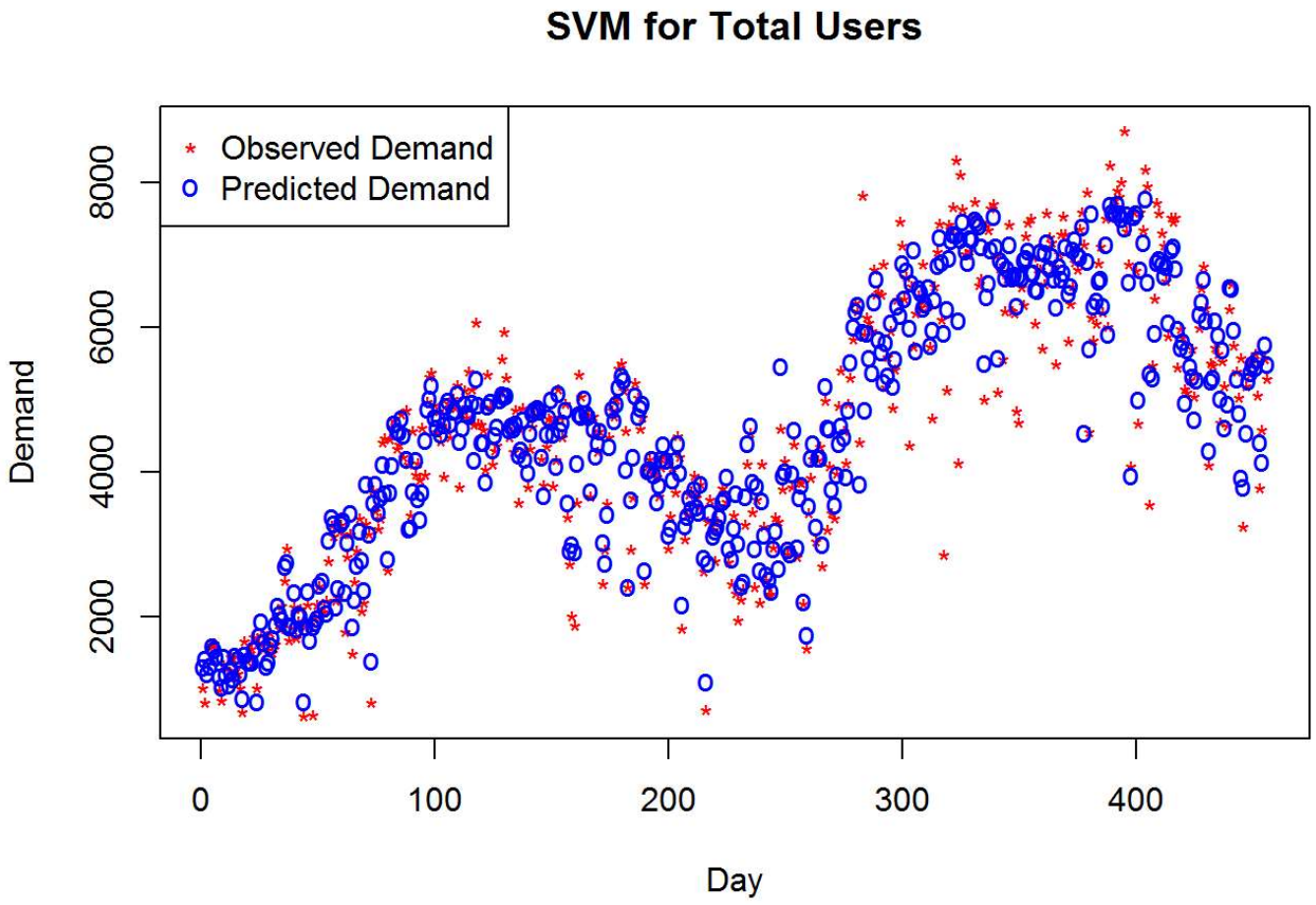
```
## [1] 269.459
```

## SVM

```
library(e1071)
for1 = as.formula(Total~Index+year+as.factor(month)+as.factor(day)+as.factor(season)+as.factor(holiday)+as.factor(workingday)+meanatemp+maxatemp+minatemp+sdatemp+meanhumidity+maxhumidity+minhumidity+sdhumidity+meanwindspeed+maxwindspeed+minwindspeed+sdwindspeed);
```

The following code uses SVM for total users and calculates its RMSPE.

```
s5 = svm(for1, data=d)
p13 = predict(s5, newdata=d, type="response")
plot(d$Index, d$Total, pch="*", col=2, xlab="Day", ylab="Demand",main="SVM for Total Users");
points(d$Index, p13, pch="o", col=4)
legend("topleft", legend=c("Observed Demand", "Predicted Demand"), pch=c("*","o"), col=c(2,4))
```




```
RMSPE(d$Total, p13)
```

```
## [1] 509.6358
```

# 6

To summarize, these were the RMSPE values for total users.

Dis-aggregated RMSPE: 819.9148992, 778.137706, 713.6341909, 667.7693348
Mean of Dis-aggregated RMSPE: 744.8640327

Aggregated RMSPE: 757.0674113, 677.6463021, 269.4589541, 509.6358398
Mean of Aggregated RMSPE: 553.4521268

With exception of the linear model, the aggregated method resulted in a lower RMSPE value than the dis-aggregated method for total demand.
Average of RMSPE value for the aggregated method was 0.2569756% less than that of the dis-aggregated value.