

Individual Household Electric Power Consumption

Data Analysis and Prediction

Abstract

This report illustrates the process and outcomes of exploratory analysis and predictive modelling for UCI's individual household electric consumption data.



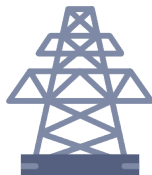


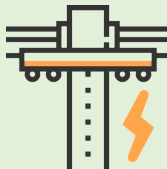

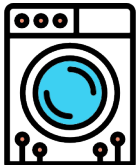

Prathyusha Sangam

Project Objectives

1. Explore and analyze UCI's individual household electric consumption dataset^[1].
2. Build a predictive model using Auto-Regressive Integrated Moving Average (ARIMA)^[2], using R's Forecast package, which can predict global active power (month averaged) for 2011.

Dataset Description

The dataset contains the below 9 attributes, collected in a single household in France, with a one-minute sampling rate, over a period of 48 months starting 16 December 2006 until 26 Nov 2010. The dataset constitutes 2,075,259 records and has missing values.

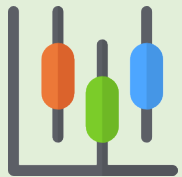
	Date DD/MM/YYYY		Time HH:MM:SS		Global Active Power in Kilowatts The real power consumption of the household
	Global Reactive Power in Kilowatts Power drawn by inductors and capacitors (not real power usage)		Voltage in Volts Electric potential difference		Global Intensity in Amperes Strength of current
	Sub-metering 1 Watt-Hour Corresponds to the kitchen, containing dishwasher, oven, etc.		Sub-metering 2 Watt-Hour Laundry room containing washing-machine drier, a refrigerator and a light		Sub-metering 3 Watt-Hour Corresponds to an electric water-heater and an air-conditioner.

Part 1: Exploratory Analysis

Data Loading and Preprocessing

Data is loaded into R global environment from the text file. Columns are transformed into their respective data types.

Insights



Missing Values

- About 1.25% of the data is missing

Attribute Ranges

- Global Active power ranges from 0.076-11.122 Kilowatts
- Global Reactive Power ranges from 0-1.39 Kilowatts
- Voltage fluctuates from 223.2 to 254.2 Volts (usually 220-240V in France)
- Global Intensity ranges from 0.2 to 48.4 Amperes
- Maximum Sub-metering in Kitchen is 88, Laundry room 80 and Air-conditioning 31 Watt-Hour

```
#Data Loading
powerData <-
read.table('household_power_consumption.txt',sep =
';',header = TRUE, stringsAsFactors = FALSE)

#Date Transformation
powerData$Date<-
as.Date(powerData$Date,format="%d/%m/%Y")

# Time Transformation
install.packages("chron")
library(chron)
time_1 <- c(powerData$Time)
powerData$Time <- chron(times = time_1)

# Numeric transformations #missing values '?' get
replaced by NA in R by coercion
for (i in 3:9)
  powerData[, i] <- as.numeric(powerData[, i])

#Initial Exploration
summary(powerData)
```

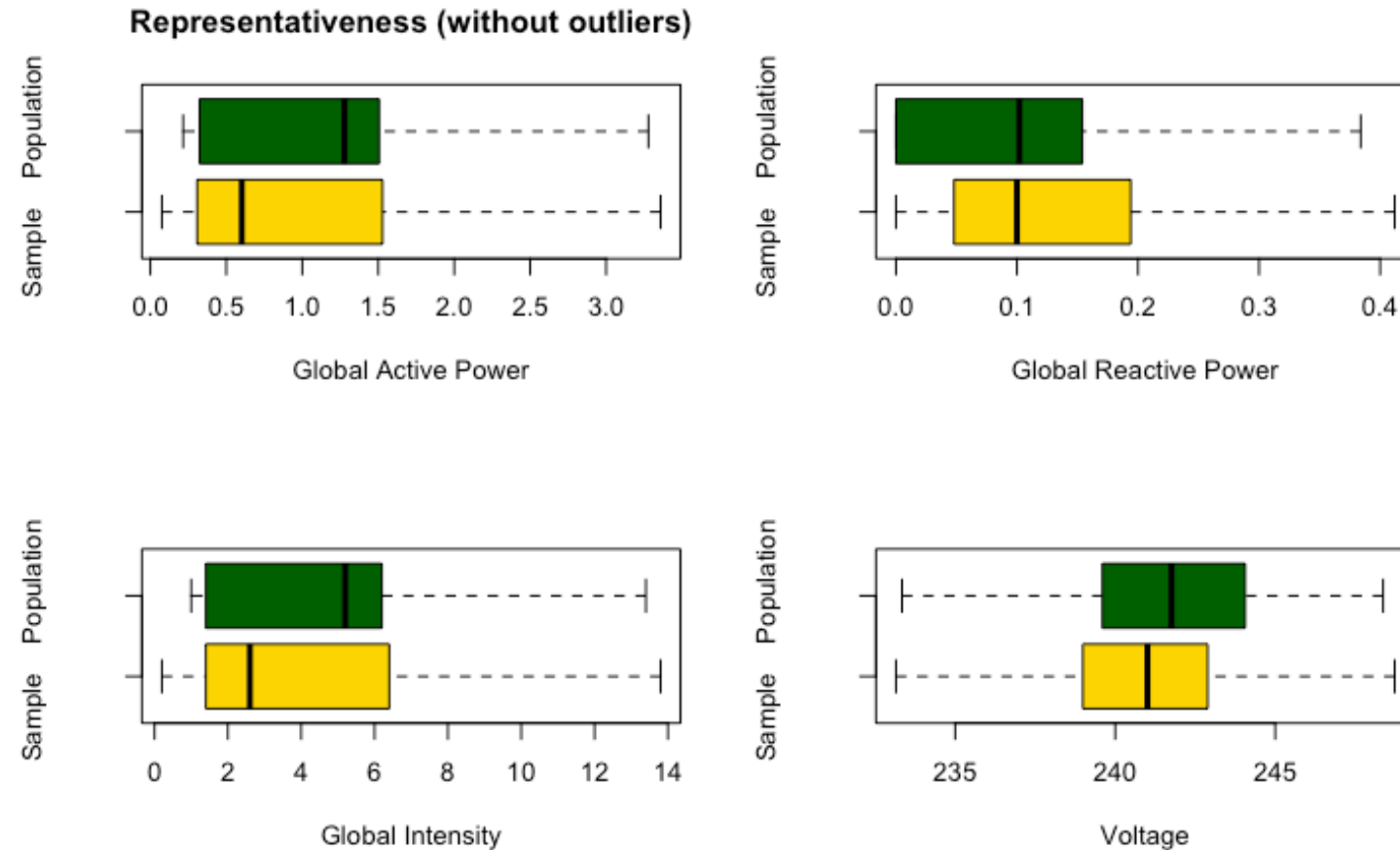
Preliminary Analysis

Global Active Power is proportional to Voltage (Ohm's Law): ***Current = Voltage/Resistance***. Sub-metering values are direct contributors to the total power consumption (Global Active Power). Hence, household power consumption can be represented as a linear model with the electrical attributes as additive components. However, for effective modelling and forecasting, a time series analysis is more appropriate, since power consumption depends on the instance of time, like, day of the week, or the season.

Sampling

To quickly gain intuition about the relationships and trends of electrical attributes and sub-metering, a small sample is drawn from the original dataset. The dates chosen are **February 8 and February 9 of 2007**. A new column fullTimeStamp is created to make visualizations easier. There are no missing values in these dates

Representativeness of the sample



Values of Global Active Power, Global Reactive Power, Voltage and Global Intensity of the subset, Feb 8 and 9, 2007, have very similar distribution to the original population. Hence, the sample can be considered as a representative of the total population.

Plotting Univariate Time Series for Electrical Attributes – Global Active Power, Sub-metering

Findings

As observed on Feb 8,9 2007

Global Active Power (Kilo Watts)

- Least power usage during the noon and midnight
- Usage spikes during the day time and evenings

Sub-metering 1 (Kilo Watt Hour):

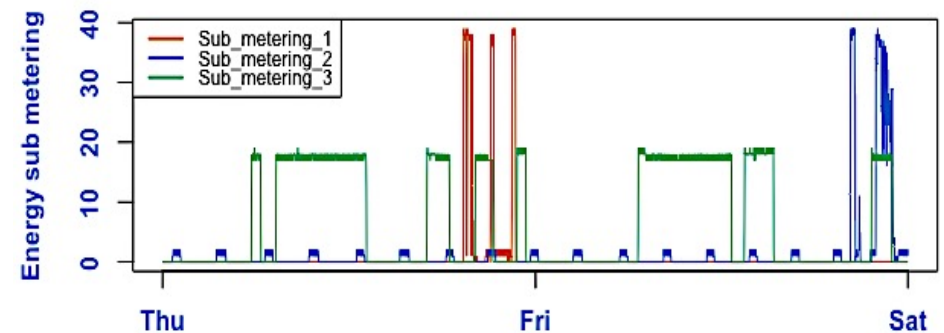
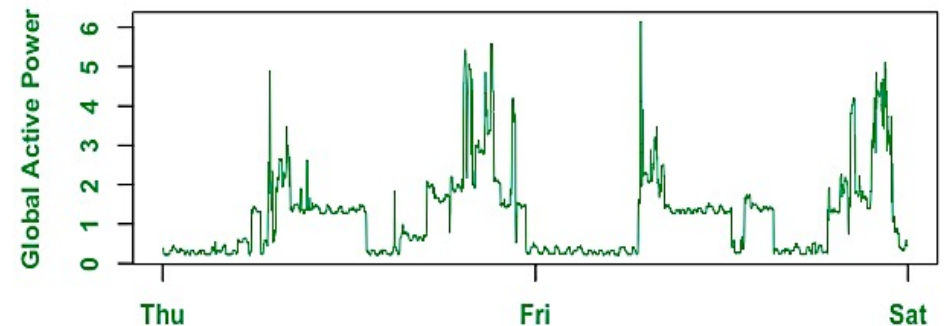
- Power usage in Kitchen is mostly un-noticeable on both days.
- Surge on Thursday night, likely due to use of dishwasher or oven

Sub-metering 2:

- Continuous cooling cycles of refrigerator (compressor) clearly seen, each lasting approximately 30 minutes, once every 3 hours.
- Spike on Friday night likely due to use of washing machine or drier in the Laundry room

Sub-metering 3:

- Air conditioning or water heating is used mostly during midday and at night
- Consumption of power is consistent once the device is turned on



```

#sampling Feb 8, 9 2007
powerData_Sample <- subset(powerData,subset = (powerData$Date>='2007-02-08' & powerData$Date<='2007-02-09'))

# Adding new column Full Time stamp
fullTimestamp <- paste(powerData_Sample$Date,powerData_Sample$Time)
powerData_Sample$fullTimestamp <- as.POSIXct(fullTimestamp)

#Initial Exploration
summary(powerData_Sample)

#representativeness: Boxplots of electrical attributes

par(mfrow = c(2,2))
boxplot(powerData$Global_active_power, powerData_Sample$Global_active_power, horizontal = TRUE,
outline = FALSE, xlab='Global Active Power', ylab='Sample      Population ', main='Representativeness
(without outliers)', col=(c("gold","darkgreen")))
boxplot(powerData$Global_reactive_power, powerData_Sample$Global_reactive_power, horizontal = TRUE,
outline = FALSE, xlab='Global Reactive Power', ylab='Sample      Population
',col=(c("gold","darkgreen")))
boxplot(powerData$Global_intensity, powerData_Sample$Global_intensity, horizontal = TRUE, outline =
FALSE, xlab='Global Intensity', ylab='Sample      Population ',col=(c("gold","darkgreen")))
boxplot(powerData$Voltage, powerData_Sample$Voltage, horizontal = TRUE, outline = FALSE,
xlab='Voltage', ylab='Sample      Population ',col=(c("gold","darkgreen")))

#plotting wrt time # electric attributes with respect to fullTimestamp

par(mfrow = c(2,1))
plot(powerData_Sample$fullTimestamp, powerData_Sample$Global_active_power, ylab = "Global Active
Power",xlab = "", type = "l",
col='#1D9867',col.axis='#1D9867',col.lab='#1D9867',font.lab='2',font.axis='2')
plot(powerData_Sample$Sub_metering_1 ~ powerData_Sample$fullTimestamp, ylab = "Energy sub metering",
xlab = "", type = "l",col='#CA5A07',col.axis='#2757BE',col.lab='#2757BE', font.lab='2',font.axis='2')
lines(powerData_Sample$Sub_metering_2 ~ powerData_Sample$fullTimestamp, col = '#2757BE')
lines(powerData_Sample$Sub_metering_3 ~ powerData_Sample$fullTimestamp, col = '#1D9867')
legend("topleft", col = c("#CA5A07", "#2757BE", "#1D9867"), legend = c("Sub_metering_1",
"Sub_metering_2", "Sub_metering_3"), lwd = 2,cex =0.75 )

```

Part 2: Predicting Global Active Power using ARIMA

The procedure for building time series forecasting model of Global Active Power, includes data preprocessing, constructing and decomposing time-series, fitting a predictive model using Auto-Regressive Integrated Moving Average: ARIMA.

Data Loading and Preprocessing

To predict power consumption, Date and Global Active Power (GAP) of the entire power consumption dataset, are loaded into the data frame trainData. For accurate time series prediction, each missing value of GAP is replaced with the previous known value, with an assumption that not much changed since last minute.

```
#Data Load
trainData <- data.frame(Date=powerData[,1],GAP=powerData[,3])
#Data Cleaning: Replace missing values of Global Active Power with previous value
library('zoo')
trainData <- transform(trainData, GAP = na.locf(GAP))
summary(trainData)
```

Aggregating Data

The original data has minute averaged values. In this step data is aggregated to monthly averages for simplification.

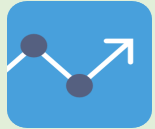
```
#aggregating for month
agg_Month <- aggregate(trainData,by=list(as.yearmon(trainData$Date,"%d%m%Y")), FUN = mean, na.rm=TRUE)
```

Constructing Time series and Decomposing components

```
# Construct, decompose time series and plot
tsmonth <- ts(agg_Month[,3],frequency = 12,start = c(2006,12))
fm <- decompose(tsmonth)
plot(tsmonth, ylab='GAP',main='Global Active Power: Month Averaged',col='#A93226', lwd=2)
abline(h=mean(tsmonth),col='#A93226')
plot(fm, col='#2757BE',lwd=2)
```

Plotting Time Series (for monthly averaged values of Global Active Power)

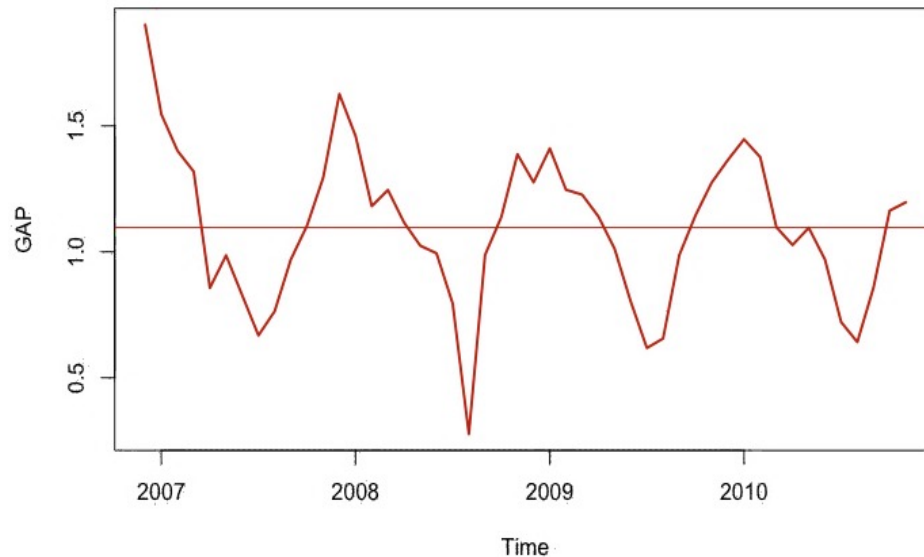
Findings



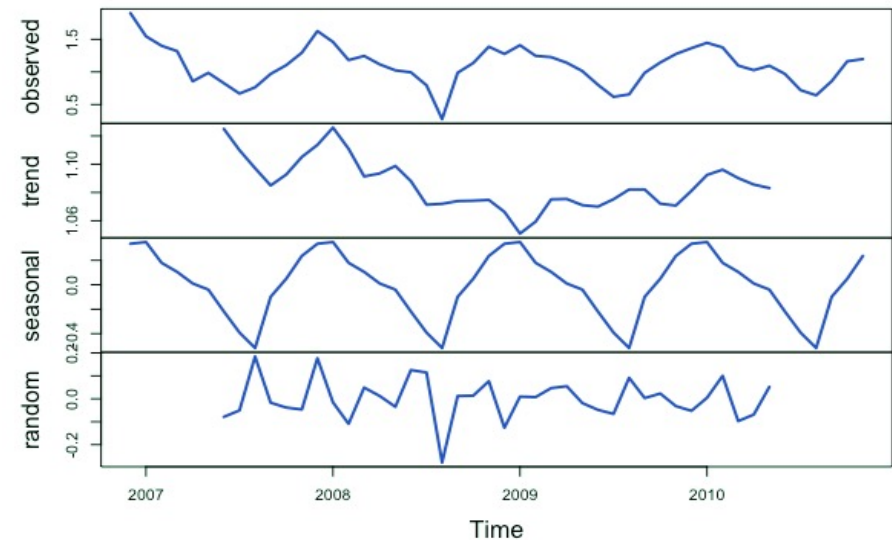
Monthly averaged values of Global Active Power from Dec 2006- Nov 2010

- There seems to be no consistent trend.
- Decrease in Global Active Power from 2007-2010 on an average is observed. Downward trend observed until 2009, gradual increase there on.
- Global Active Power = 1.09 represents the mean of the series,
- There are no obvious outliers
- Prominent seasonality. Power usage is high at start and end of the year, least usage observed mid-year.

Global Active Power: Month Averaged



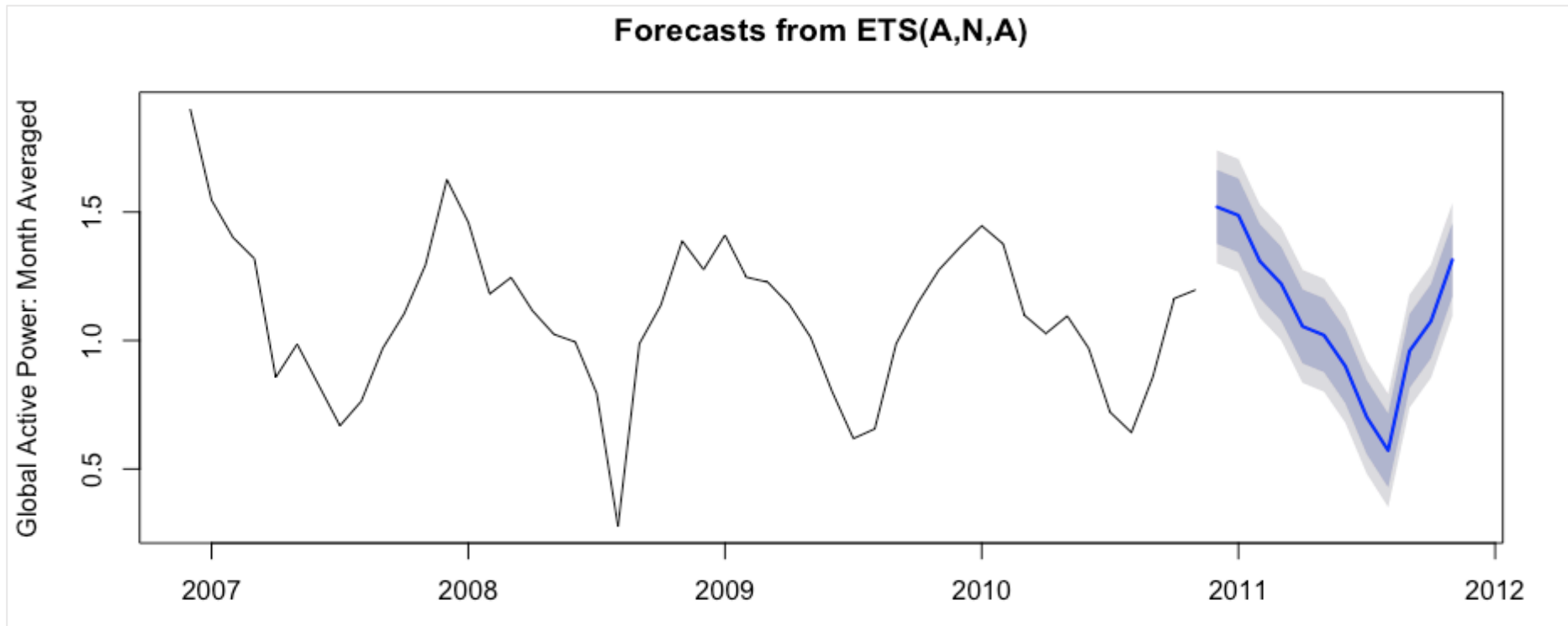
Decomposition of additive time series



Building the Model

For constructing the ARIMA model, R function called Auto ARIMA is used, which returns best model based on either AIC, AICc or BIC values, and plot prediction for specified time period.

```
#Predicting using auto.arima  
  
library('forecast')  
auto.arima(tsmmonth)  
plot(forecast(tsmmonth,h=12))
```



The blue curve represents forecast of global active power, monthly averaged values, for the year 2011. The surrounding grey area indicates error bounds at a confidence level of 95%.

Results of ARIMA

Fine-tuning the model by varying D: order of seasonal-differencing, provides better results

<code>auto.arima(tsmmonth)</code>	<code>auto.arima(tsmmonth, D=1)</code>
Series: tsmmonth ARIMA(1,0,0)(1,0,0)[12] with non-zero mean	Series: tsmmonth ARIMA(0,0,0)(0,1,1)[12]
Coefficients: ar1 sar1 mean 0.3873 0.7135 1.1109 s.e. 0.2090 0.1393 0.0896	Coefficients: sma1 -0.4395 s.e. 0.2559
sigma^2 estimated as 0.02862: log likelihood=14.38 <u>AIC=-20.76 AICc=-19.83 BIC=-13.27</u>	sigma^2 estimated as 0.02349: log likelihood=15.67 <u>AIC=-27.34 AICc=-26.98 BIC=-24.17</u>

Prescriptive Analysis: Next Course of Action

Extending the model to predict hourly, daily values of power consumption can help electricity/power board to plan an effective schedule for distribution. Monthly, quarterly and yearly predictions can be helpful in planning the power generation as per projected demand.

References

- [1]UCI repository of machine learning database [Online]. Available: <http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption>.
- [2]S. Ao, The 2013 IAENG International Conference on Artificial Intelligence and Applications, the 2013 IAENG International Conference on Bioinformatics, the 2013 IAENG International Conference on Control and Automation, the 2013 IAENG International Conference on Computer Science, the 2013 IAENG International Conference on Data Mining and Applications, the 2013 IAENG International Conference on Internet Computing and Web Services, the 2013 IAENG International Conference on Imaging Engineering, the 2013 IAENG International Conference on Software Engineering. Hong Kong: IAENG, 2013.

Image source: flaticon.com/free-icons/