Juliette Di Francesco

R FINAL PROJECT

R, through its different packages, offers a wide range of tools for data analyzes. The main objectives of this tutorial are to learn how R can be used to **plot on a Google Earth map the location of animals for which geographical coordinates were recorded** and to do statistical analyzes using a **multiple linear regression model**.

I. Geographic visualization of animal locations using the package ggmap:

The **ggmap** package is a mapping tool that can be used to access and download base maps from Google or Open Street Maps. These can then be used as background to plot the geographical location of animals for which you have longitude and latitude data. The package is quite similar to ggplot2 for graph making, as you can chose your map's zoom level and type, the way points are represented on the map, y-axis, x-axis and general titles… The function `get_map` is used to download the base map, and then the function `ggmap` enables you to add your data points.

II. Statistical analysis using a multiple linear regression model:

Multiple linear regression is a common method of multivariate statistical analysis. It is used to explain the relationship between one continuous dependent variable and two or more independent variables that can be continuous or categorical. The categorical variables with n > 2 categories have to be transformed into n-1 dummy variables.

The model has **several key assumptions**:

- The **relationship** between the continuous independent and dependent variables must to be **linear**. Outliers must be checked as multiple linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots.

- All continuous variables included in the model should have a **normal distribution**. The dependent variable should also have a normal distribution within each category of every categorical variable. This normality assumption can best be assessed by looking at the histogram, and then confirmed by using a statistical test like the Shapiro-Walk test, or by looking at the normal probability plot, where the data is plotted against a theoretical normal distribution. When the data is not normally distributed, a non-linear transformation, like a log-transformation can sometimes fix the issue.

- This model also assumes that there is **little or no multicollinearity** in the data, meaning that the independent variables shouldn't be too highly correlated.

- Residuals (measure of the variability in the dependent variable not explained by the regression model) should be independent from each other (**absence of autocorrelation**), **normally distributed**, and **homoscedastic**, meaning that the variance of errors should be the same across all levels of the independent variables.

The function `lm` from the R stats package is used to fit linear regression models. A backward stepwise model selection based on the exact AIC can be carried out automatically using the function `stepAIC` from the MASS package. This function gives you the model with the lowest AIC.

R then offers a wide range of functions to describe your final model: `summary` for a general description, `coefficients` that returns the model coefficients, `confint` that returns the chosen confidence intervals for the model parameters, `fitted` that returns the predicted values, `residuals` that returns the residuals, `extractAIC` that returns the model's AIC value, and `anova` that returns the anova table.

The assumptions of the model concerning residuals can be checked by looking at different plots. By default, four graphs are provided by R with the `plot` function: residuals against fitted values, a normal quantile-quantile plot, a scale-location plot of $\sqrt{(|\text{residuals}|)}$, and residuals against leverages.

R CODE

As an example, we will use a modified dataset from a wildlife study carried out in muskoxen. Samples and information were collected from animals during community and sports hunts around the communities of Cambridge Bay and Kugluktuk, Nunavut, in 2014. For each animal, age, sex, GPS location of the cull and season were recorded. Hair and fecal samples were analyzed in order to measure respectively hair and fecal cortisol levels. We will explain the relationship between hair cortisol level, a continuous variable and the other variables that are either continuous or categorical, using a multiple linear regression model.

```
# STEP 1: We will start by reading our data set in this manner:
rm(list=ls(all=TRUE)) # Remove all objects from the environment
setwd("C:/Users/Veronique/Downloads/R Wizardry/R project") # Set working directory
data_muskox <- read.csv("mydata_R.csv", header=T, stringsAsFactors=T, sep=";")
# Read the dataset, with all character variables as factor

# STEP 2: we will then load all the packages that will be used during the tutorial:
library(ggplot2) # Graph making
library(dplyr) # Data manipulation
library(ggmap) # Spatial visualization on a map
library(MASS)
```

```
### 1. Geographic visualization of each animal's cull location using the package
ggmap:

## STEP 1: We will first download the map from Google (the source by default). We
need to define here your map's zoom level (it can vary from 1 to 21) and type (in
our case, we will choose a satellite map)
map_Muskox <- get_map(location = c(lon = mean(data_muskox$longitude),
            lat = mean(data_muskox$latitude)), zoom = 5, maptype = "satellite")

## STEP 2: We will then plot each animal's cull geographical location using
longitude and latitude data. We choose here to represent each animal as a red
filled circle on the map (See figure 1).
ggmap(map_Muskox) +
  geom_point(data = data_muskox, aes(x = longitude, y = latitude, fill = "red"),
size = 4, shape = 21) +
  theme(legend.position = "none") + # Remove legend
  labs(list(title="Cull location of each muskox", x="Longitude", y="Latitude"))
```

```
### 2. Statistical analysis using a multiple linear regression model:

## STEP 1: We will look at the distribution of our dependent variable by plotting a
histogram using the ggplot2 package in R (see figure 2):
ggplot(data_muskox, aes(x=cortisol_hair)) +
  geom_histogram(binwidth=.5, colour="black", fill="cornflowerblue") +
  xlab(label="Cortisol Hair (pg/mg)") +
  ylab(label="Frequency")

## STEP 2: The values of hair cortisol higher than 30 pg/mg are considered as
outliers, and are due to blood contamination of the hair samples. We therefore need
to remove them for further analysis using the following line of code and the filter
function from the dyplr package:
data_muskox <- data_muskox %>% filter(cortisol_hair<=30)
```

```
## STEP 3: Once the outlier is removed, the distribution of hair cortisol values
seems normal when looking at the histogram, but we need to confirm the normality of
the distribution. This can be done in two different ways, either by using the
Shapiro-Wilk statistical test or by looking at the normal probability plot (see
figure 3):
shapiro.test(data_muskox$cortisol_hair) # we get a p-value = 0.2812 > 0.05 so the
distribution is normal
qqnorm(data_muskox$cortisol_hair, col="midnightblue")
qqline(data_muskox$cortisol_hair, col="darkred", lwd=1)


## STEP 4: We will recode the variable season of collection that has three
categories (fall, summer, and winter) into two dummy variables (called summer and
winter). Two different methods can be used to do this: you can either do it
manually (OPTION 1), or by using a function in R (OPTION 2):

# OPTION 1
data_muskox$winter <-
factor(with(data_muskox,ifelse((season_collected=="winter"),1,0)))
data_muskox$summer <-
factor(with(data_muskox,ifelse((season_collected=="summer"),1,0)))

# OPTION 2
dummies <- model.matrix(~ data_muskox$season_collected)
data_muskox$summer <- dummies[,2]
data_muskox$winter <- dummies[,3]


## STEP 5: The normality of the distribution of hair cortisol values should also be
checked within each category of every categorical variable (sex, age, location,
summer and winter). We will detail the code only for the variable sex, as the same
code will be used for the other variables.
Sub1 <- subset(data_muskox, sex=="male")
hist(sub1$cortisol_hair) # p = quick histogram to visualize the distribution of
hair cortisol values among males
shapiro.test(sub1$cortisol_hair) # p = 0.2033 > 0.05, so the distribution is normal

sub2 <- subset(data_muskox, sex=="female")
hist(sub2$cortisol_hair)
shapiro.test(sub2$cortisol_hair) # p = 0.9122 > 0.05, so the distribution is normal

# Similarly, the normality of the distribution of hair cortisol values was verified
within each category of every other categorical variable.


## STEP 6: We will now check the normality of the distributions for the independent
continuous variable (fecal cortisol level)

# The Shapiro-Walk test is run like described previously, and the distribution
turns out not to be normal (p-value = 0.01024 < 0.05). We need to transform the
variable, for example by carrying out a log-transformation using the following line
of code:
data_muskox <- mutate(data_muskox, logcortisol_feces = log(cortisol_feces))

# The normality is then checked for this newly created variable:
shapiro.test(data_muskox$logcortisol_feces) # p-value = 0.148 > 0.05, so the log-
transformed variable has a normal distribution and will therefore be used instead
for the rest of our analysis.

## STEP 7: Before carrying out the multivariate analysis by running the multiple
linear regression model, a UNIVARIATE ANALYSIS should be done.
```

```
# We will look at the linearity of the relationship between hair and fecal cortisol
levels with the following lines of code:
plot(data_muskox$logcortisol_feces, data_muskox$cortisol_hair,
     pch=21, col="black", bg="cornflowerblue",
     xlab="Feces Cortisol (ng/g wet feces)", ylab="Hair Cortisol (pg/mg)")
lines(smooth.spline(data_muskox$logcortisol_feces, data_muskox$cortisol_hair),
      lty=2, col="blue") # plots smoothing spline line
abline(lm(data_muskox$cortisol_hair ~ data_muskox$logcortisol_feces)) # plots a
régression line (See figure 4)
# The linearity of the relationship between both variables is confirmed and fecal
cortisol level can therefore be included in our multivariate analysis

cor(data_muskox$cortisol_hair, data_muskox$logcortisol_feces) # returns the
correlation coefficient: r = 0.7594599, so the two variables are positively
correlated.

# We will now compare the mean of hair cortisol between the two categories for each
categorical variable using a t-test. Its first assumption (normality of the
distribution of the continuous variable in both categories of the categorical
variable) was verified in STEP 5. The second assumption, which is the homogeneity
of variance, needs to be verified with an F-test to compare two variances before
running the t-test. We will once again detail the code only for the variable sex:
var.test(sub5$cortisol_hair, sub6$cortisol_hair) # p-value = 0.6988 > 0.05, so the
second assumption, homogeneity of variances, is verified.
t.test(sub5$cortisol_hair, sub6$cortisol_hair) # p-value = 0.3505 > 0.05, so the
mean of hair cortisol is not significantly different between males and females.

# Similarly, the homogeneity of variances was verified for all the other variables
and a t-test was run.


## STEP 8: We will now begin our multivariate analysis by running the full multiple
linear regression model that includes all the variables:
lm_full <- lm(cortisol_hair ~  logcortisol_feces + age_cat + sex + summer + winter
+ location, data = data_muskox)


## STEP 9: We will then use backward stepwise model selection to determine which
model has the lowest AIC value and will be kept as our final model (see figure 5
for the output of this command, the final model is in red).
step_lm <- stepAIC(lm_full, direction="backward")
# The model with the lowest AIC is the following:
lm(cortisol_hair ~ logcortisol_feces + summer + location)

# It's important to note that two nested models can be compared using the anova
function. We can compare our final model with the full model (see figure 6 for the
analysis of variance table) by running the following line of code:
anova(final_lm, lm_full) # We can see that p-value = 0.8858 and that both models
aren't actually significantly different.


## STEP 10: Model description.
# A general description can be viewed by using the summary function in R (see
figure 7 for the output of this command, where you can see in red the model
coefficients and their p-value, and the information concerning residuals is in
green):
final_lm <- lm(cortisol_hair ~ logcortisol_feces + summer + location)
summary(final_lm)
```

```
# Other functions can also be used:
coefficients(final_lm) # Returns the model coefficients
confint(final_lm, level=0.95) # Returns the 95% confidence intervals for all model
parameters
fitted(final_lm) # Returns predicted values
residuals(final_lm) # Returns the residuals
anova(final_lm) # Returns the anova table
extractAIC(final_lm) # Returns the AIC value


## STEP 11: Model checking

# By default, four graphs are provided by R using the function plot (see figure 8):
par(mfrow=c(2,2)) # To have all four graphs on the same image
plot(final_lm)

# The first plot (top left corner) is a scatter plot of the residuals versus fitted
values: here, the residuals look equally spread around a horizontal line without
distinct patterns. This is a good indication that we don't have non-linear
relationships.
# The second plot (top right corner) is a normal quantile-quantile plot of the
residuals. It shows if they are normally distributed. In our case, they seem to
follow a straight line and are therefore normally distributed.
# The third plot (bottom left corner) is the scale-location plot. It shows if
residuals are spread equally along the ranges of the independent variables. This
plot enables you to check the assumption of equal variance (homoscedasticity).
There should be a horizontal line with equally spread points. However, in our case,
the variance of the residuals seems to increase slightly with the fitted values.
# The fourth plot (bottom right corner) is used to find influential subjects. In
our case, there seems to be no influential individuals.

# In conclusion, all the assumptions of our model seem to be met.
```

FIGURES AND TABLES

Figure 1:

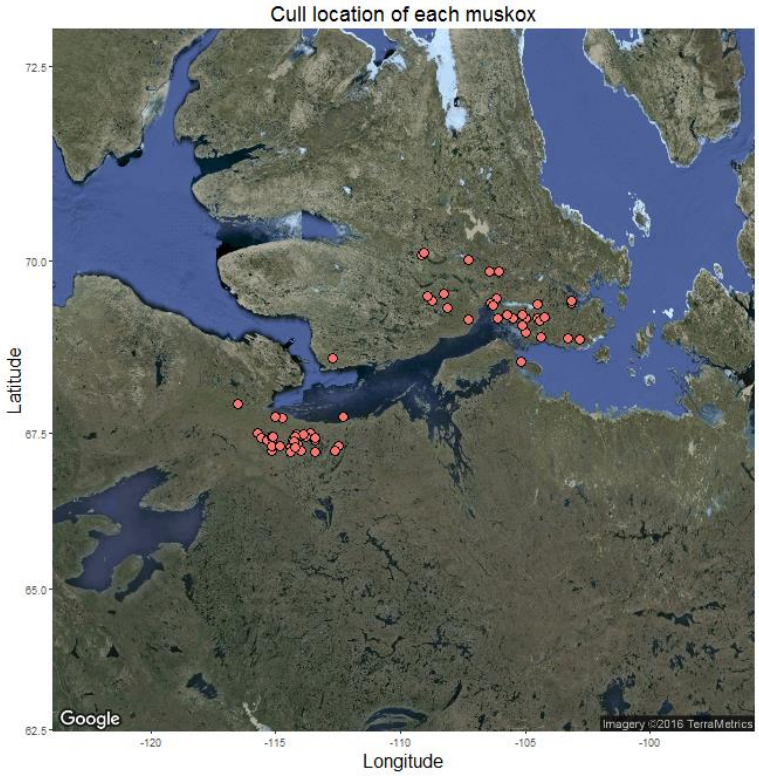

Figure 2: Histogram of hair cortisol levels

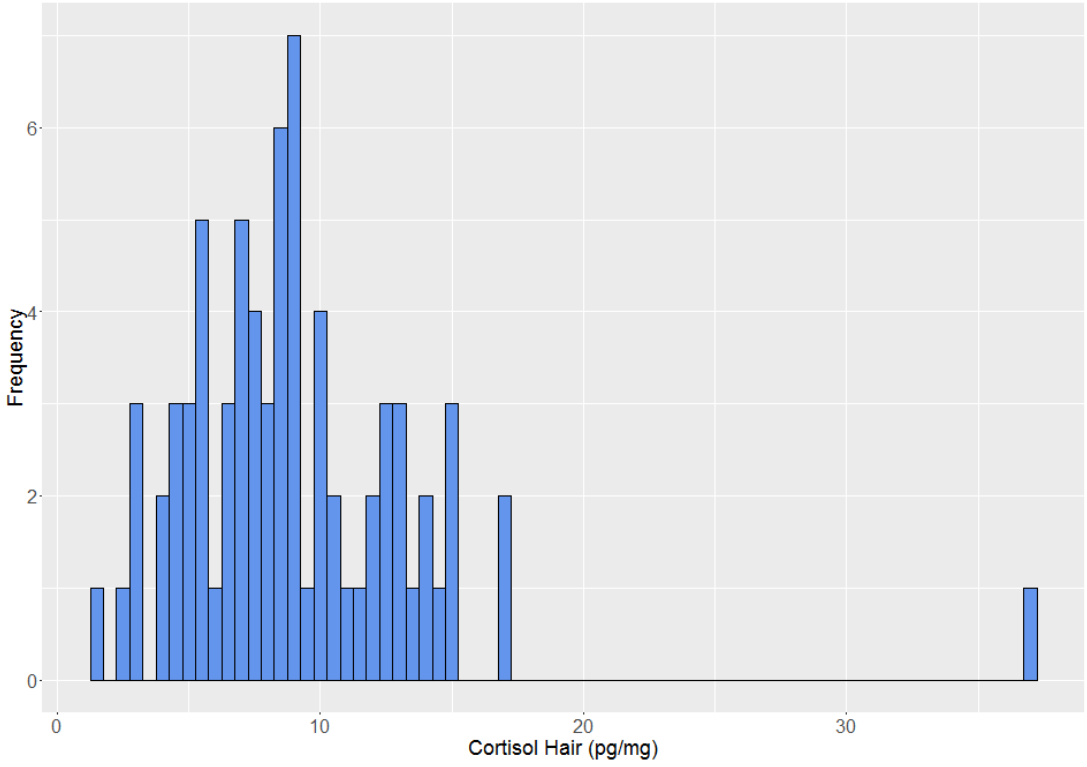

Figure 3: Normal probability plot of the hair cortisol variable

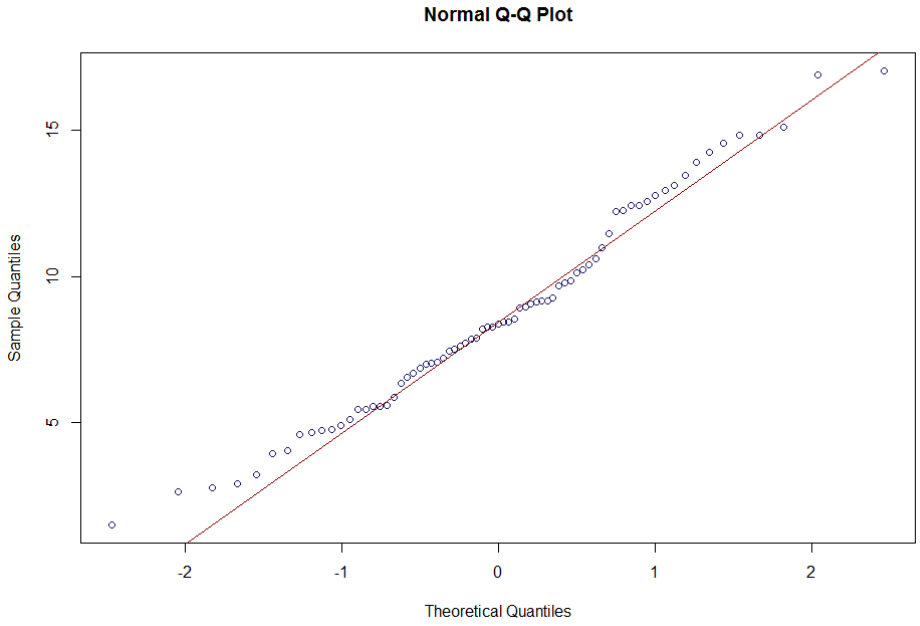

Figure 4: Plot showing the relationship between hair and feces cortisol levels.

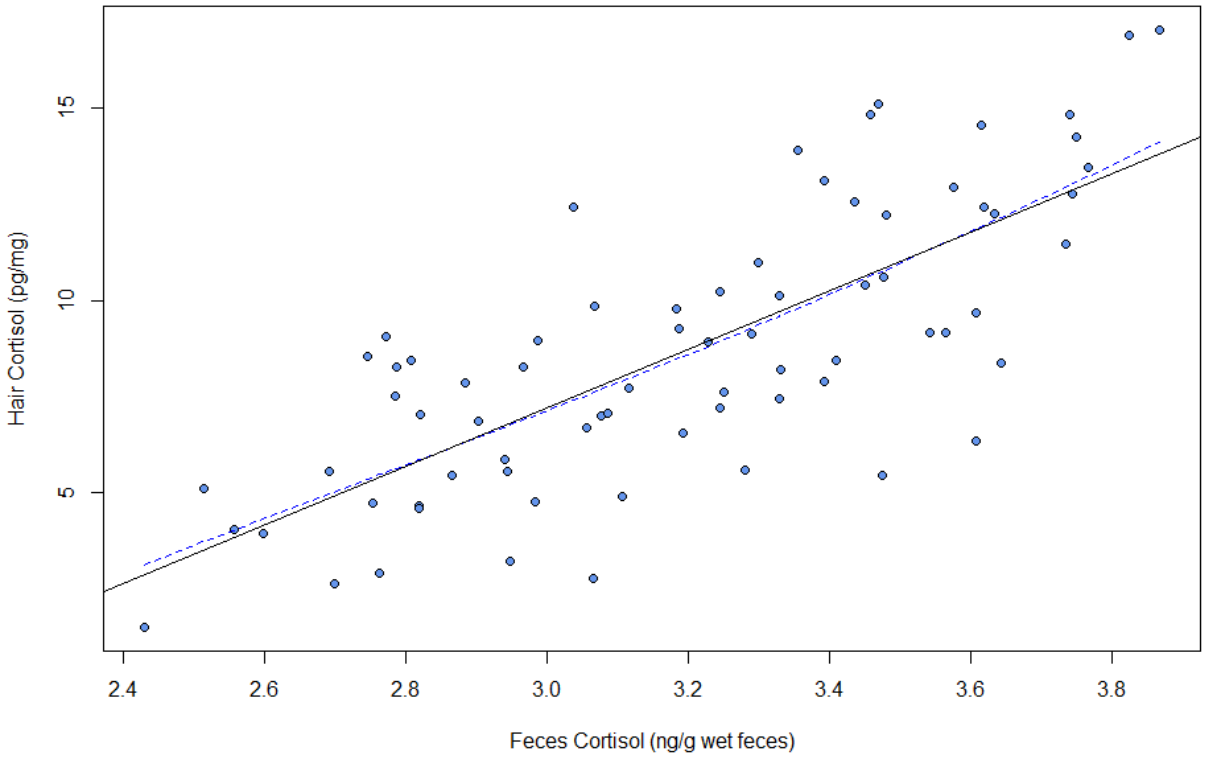

Figure 5: Results of the backward stepwise model selection

```
Start:  AIC=128.43
cortisol_hair ~ logcortisol_feces + age_cat + sex + summer +
    winter + location

                     Df Sum of Sq    RSS    AIC
- winter              1      0.01 350.04 126.43
- sex                 1      0.85 350.87 126.61
- age_cat             1      2.45 352.48 126.94
- location            1      5.54 355.56 127.58
<none>                             350.02 128.43
- summer              1     23.93 373.95 131.26
- logcortisol_feces   1    465.59 815.61 188.18

Step:  AIC=126.43
cortisol_hair ~ logcortisol_feces + age_cat + sex + summer +
    location

                     Df Sum of Sq    RSS    AIC
- sex                 1      0.86 350.90 124.61
- age_cat             1      2.57 352.61 124.97
- location            1      5.74 355.78 125.62
<none>                             350.04 126.43
- summer              1     36.85 386.88 131.74
- logcortisol_feces   1    470.90 820.94 186.66

Step:  AIC=124.61
cortisol_hair ~ logcortisol_feces + age_cat + summer + location

                     Df Sum of Sq    RSS    AIC
- age_cat             1      2.55 353.44 123.14
<none>                             350.90 124.61
- location            1      9.99 360.88 124.66
- summer              1     36.12 387.02 129.76
- logcortisol_feces   1    474.16 825.05 185.02

Step:  AIC=123.14
cortisol_hair ~ logcortisol_feces + summer + location

                     Df Sum of Sq    RSS    AIC
<none>                             353.44 123.14
- location            1     12.41 365.86 123.66
- summer              1     36.04 389.48 128.23
- logcortisol_feces   1    472.21 825.66 183.08
```

Figure 6: Comparison of the final and full models:

```
Analysis of Variance Table

Model 1: cortisol_hair ~ logcortisol_feces + summer + location
Model 2: cortisol_hair ~ logcortisol_feces + age_cat + sex + summer +
    winter + location
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     69 353.44
2     66 350.02  3    3.4175 0.2148 0.8858
```

Figure 7: General description of the final model

```
Call:
lm(formula = cortisol_hair ~ logcortisol_feces + summer + location,
    data = data_muskox)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2680 -1.3270  0.0921  1.5765  4.9038

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -13.5793     2.4957  -5.441 7.58e-07 ***
logcortisol_feces   7.2222     0.7522   9.601 2.43e-14 ***
summer1            -1.6709     0.6299  -2.653   0.0099 **
locationKugluktuk  -0.8442     0.5423  -1.557   0.1241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.263 on 69 degrees of freedom
Multiple R-squared:  0.6219,   Adjusted R-squared:  0.6054
F-statistic: 37.82 on 3 and 69 DF,  p-value: 1.431e-14
```

Figure 8: