Global warming evidences in New York city

April 19, 2015

Introduction:

Global warming is a current concern everywere in the world. Greenhouse gases, such as carbon dioxide, methane and ozone resulting from industrial activities and fossil fuel burning (diesel oil and gasoline) by automobiles cause the *greenhouse effect* and are among the major causes for this warming.

The average temperature increases resulting from global warming can usually be observed over long periods of time, such as decades or even a century. It is usually hard to find significant evidences of warming from one year to the next, for example.

The present research aims to find statistical evidence of the global warming effects over the city of New York (USA), based on the temperature records of the NOAA (National Climatic Data Center) of the climatic station located in the Central Park area (Belvedere Tower - Latitude 40.77889° / Longitude -73.96917° / Elevation 39.6 m) from December 21, 1876 to December 20, 2014.

One may believe that the average temperature increases are more noticiable in specific seasons of the year. Crossing the NOAA temperature data with the *timeanddate.com* seasons calculator, this research will try to address this possibility.

Finally, it is possible to state the research questions to be answered :

- Had the average temperature of New York city actually increased over the last 138 years, from 1876 to 2014 ?
- Is the average temperature increase more noticiable in one of the seasons of the year, or basically the same ?

It is probably redundant to deeply discuss the relevance of the global warming question in the modern world. The global warming is impacting the quality of life of almost all the humankind, creating negative climatic effects such as storms, tornadoes, expansion of subtropical deserts, unstable rain periods, heat waves, droughts, heavy snows, etc. Additionally, a huge part of the global warming effect is absorbed by the oceans, increasing significantly their average temperature and generating melted ice in the arctic regions, which consequently rises the sea levels and threatens the cities in the seashore.

Data:

The New York Central Park station is able to colect several diferent climatic variables, such as air temperature, wind, precipitation (rain and snow), sunshine index, etc. For this research, the only interesting variable is related to the air temperature.

Originally, the data provided by NOAA database gives the the minimum and maximum temperatures for each recorded day, in tenths of degrees Celcius. For the purpose of this research, the average temperature between the daily minimum and maximum was calculated.

In order to allow the investigation of the possible influence of the seasons of the year, a variable called *season* was created based on the exact dates of the climatic solstices and equinoxes which define the end of a season and the start of the next, according to *timeanddate.com* seasons calculator.

As the effect of the increase of average temperature is usually more evident after longer periods of time, the available data was split in six distinct periods of 23 years each, as follows :

- *Period 1 :* years of 1877 to 1899
- *Period 2 :* years of 1900 to 1922
- *Period 3 :* years of 1923 to 1945
- *Period 4 :* years of 1946 to 1968
- *Period 5 :* years of 1969 to 1991
- *Period 6 :* years of 1992 to 2014

The help in this identification, a variable called *period* was also created.

The details of the process to make de data set ready for this reaserch will not be detailed here, because it is not the objective of this reasearch, however it is important to inform that all the operations were performed using **R** (version 3.1.3) and **RStudio (version 0.98.1103), both versions to Mac OS X (version 10.10.3). The final data set can be downloaded at https://github.com/rodrigodmartins/Data-Analysis-and-Statistical-Inference/tree/master/Grading%20project, as a *RData* file.

Here it is possible to check some entries of the final data set :

```
if ("pander" %in% rownames(installed.packages()) == FALSE) {
    install.packages("pander")
}
library(pander)
setwd("/Volumes/Documentos importantes/Coursera/12 - Data Analysis and Statistical Inference/Project/GitHub/D
load("ny.138y.RData")
panderOptions("table.split.table", 130)
panderOptions("keep.trailing.zeros", TRUE)
pander(head(ny.full.138y, 4), "Data set used in the research")
```

Table 1: Data set used in the research

date	season	year	season.year	period	moon.phase	t.max.c	t.min.c	t.ave.c
1876-12-21	Winter	1877	Winter-1877	1	New	-7.2	-10.6	-8.9
1876-12-22	Winter	1877	Winter-1877	1	New	-0.6	-9.4	-5.0
1876-12-23	Winter	1877	Winter-1877	1	New	-1.1	-6.1	-3.6
1876-12-24	Winter	1877	Winter-1877	1	First quarter	-5.6	-10.0	-7.8

When one visualize the data set, it is possible to identify that the cases (units of observation) are the days when the temperature measurements were performed. Among the variables in the complete set, only two, **period** and **t.ave.c** (average temperature of the day, in degrees Celsius) will be used. In order to address the possible impact of the increase of average temperatures over the different seasons, the same variables will be used, however, subsetted by **season**.

The variable **period** is **categorical** and identifies the group of years each temperature measurement belongs to (as discussed in *Data* section). The varibale **t.ave.c** is numerical continuous, as it is a temperature measurement.

The study here described is **observational**, taking into account that retrospective data already sampled by NOAA was used. On the other hand, the New York Central Park climatic station was randomly selected among all others in the same area. It means that, in the context of this study, the obtained inference result can be **generalized to New York city** (New York city is the population of interest in this study), but **no causality** estatements will be possible, because the available data is not result of a experiment and it was not originally designed to have controls and different groups to compare the average temperature increases.

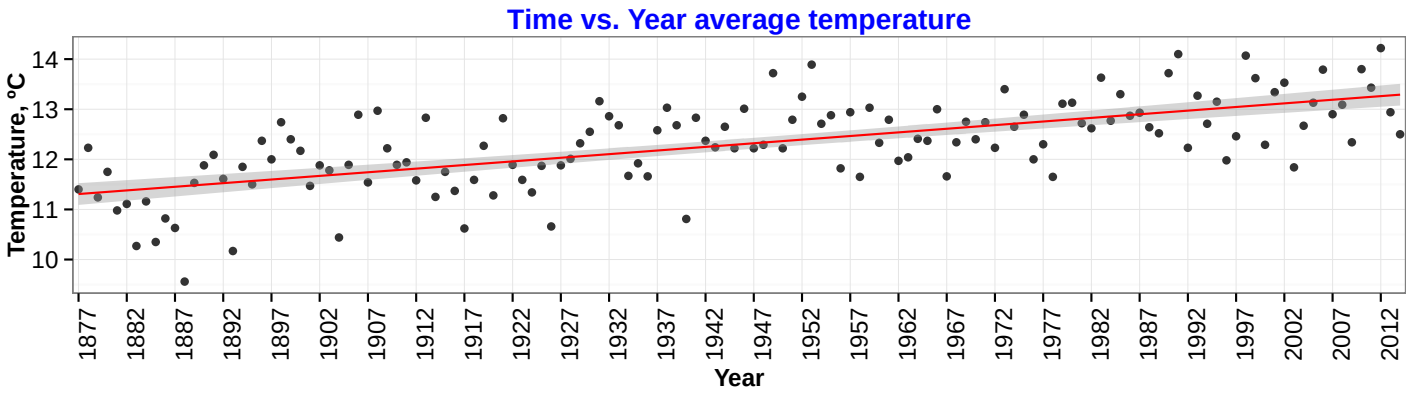
Finnaly, the main possible source of bias in the data can be attributed to the method of temperaure measurement. During the 138 years of data collection, is is expected that the technology used for data collection have changed. It means that the accuracy and precision of temperature measurements may have changed over the years, and this effect may have some influence over the result of the research.

Exploratory data analysis:

Some previous exploratory data analysis was performed, in order to look for evidences of the increase of the average temperature over time.

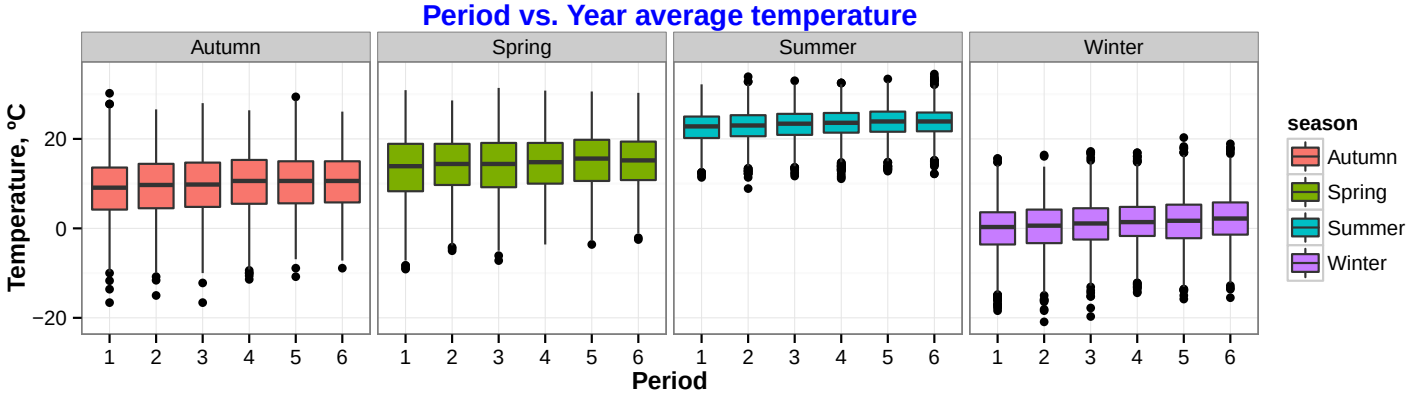
First, a scatterplot of the average annual temperatures over the years was generated. a regression line was included.

```
if ("ggplot2" %in% rownames(installed.packages()) == FALSE) {
    install.packages("ggplot2")
}
suppressPackageStartupMessages(library(ggplot2))
h1 <- ggplot(ny.138y.t.ave.year, aes(year, t.ave.c.year)) + theme_bw() + geom_point(alpha = 0.8)
h2 <- h1 + geom_smooth(method = "lm", color = "red", aes(group = 1))
h3 <- h2 + labs(x = "Year", y = "Temperature, ºC", title = "Time vs. Year average temperature")
h4 <- h3 + theme(axis.title.x = element_text(face = "bold", size = 12), axis.title.y = element_text(face = "b
    size = 12), plot.title = element_text(face = "bold", size = 14, color = "blue"))
h5 <- h4 + scale_x_discrete(breaks = seq(1877, 2014, 5))
h6 <- h5 + theme(axis.text.x = element_text(angle = 90, size = 12), axis.text.y = element_text(size = 12))
h6
```



Boxplots from the different periods over different seasons were also evaluated.

```
j1 <- ggplot(ny.full.138y, aes(period, t.ave.c, fill = season)) + theme_bw() +
    geom_boxplot() + facet_wrap(~season, nrow = 1, ncol = 4)
j2 <- j1 + labs(x = "Period", y = "Temperature, ºC", title = "Period vs. Year average temperature")
j3 <- j2 + theme(axis.title.x = element_text(face = "bold", size = 12), axis.title.y = element_text(face = "b
    size = 12), plot.title = element_text(face = "bold", size = 14, color = "blue"))
j3
```



In both graphs is possible to visualize that suggests that the average temperature of New York city have been increasing along the years (or periods). At the same time, it is still not clear if this possible increase can be more noticiable in any of the seasons. Further analysis will be needed for that.

The data summary with descriptive statistics for the data set shows :

```
panderOptions("table.split.table", 130)
panderOptions("keep.trailing.zeros", TRUE)
pander(summary(ny.full.138y[, c(1, 2, 5, 9)]), "Summary and descriptive statistics")
```

Table 2: Summary and descriptive statistics

date	season	period	t.ave.c
Min. :1876-12-21	Autumn:12383	1:8392	Min. :-20.9
1st Qu.:1911-06-26	Spring:12802	2:8392	1st Qu.: 4.5
Median :1946-01-02	Summer:12912	3:8387	Median : 12.8
Mean :1945-12-27	Winter:12269	4:8394	Mean : 12.3
3rd Qu.:1980-06-29	NA	5:8401	3rd Qu.: 20.9
Max. :2014-12-20	NA	6:8400	Max. : 34.5

Additionally, the closer values of *median* and *mean* temperature, and the almost simetric shape of the boxplots, indicate that the data is nearly normaly distributed. The big number of observations (50,366 observations, being approximatelly 12,500 in each season and 8,400 in each period) also helps to increase the power of the statistical tests to be performed.

Inference:

The average temperatures for each of the 6 periods will be compared to each other using ANOVA (analysis of variance) method.

It is relevant to state the hypothesis which is under evaluation :

- H0 : μ1877-1899 = μ1900-1922 = μ1923-1945 = μ1946-1968 = μ1969-1991 = μ1992-2014
- HA : At least one pair of means are different from each other.

The H0 hypothesis means that the mean temperature over the six periods unders evaluation are the same. On the other hand, HA, the alternative hypothesis, is that at least one pair of means are different from each other

It is important to emphasize that all conditions for a suitable ANOVA are met or are under control :

- **Independence within groups :** One may suspect that the temperature measured in one particular day may not be completely independent from the temperature measured in the next day, which can be true. On the other hand, as each of the 6 periods under evaluation include 23 years of randomly collected temperature data, there is a huge amount of observations (approximately 8,400 in each period) which increase significantly the power of the test even for minor variations. For this reason, the measurements which composes each of the periods are assumed as independent from each other.
- **Independence between groups :** Each of the 6 periods of time are independent of each other. They are not paired because they are different years and there is no relation between them.
- **Approximate normality :** As it was possible to verify in the exploratory data analysis previously performed, there is good indication that all the groups are close to normal.
- **Equal variance :** As it was possible to verify in the exploratory data analysis previously performed, all the boxplots show roughly the same size (same interquartile range), which indicates a similar variability between the groups.

As the ANOVA method is limited to indicate if one of the means between the different periods is different, additional tests are required to identify which period is different. For this reason, the methods to be used will be :

1. `aov` function : It will fit an analysis of variance (ANOVA) model for the provided data set.
2. `pairwise.t.test` function : It will compare all possible combinations of two group periods to each other, using the *Bonferroni correction* for α.
3. `TukeyHSD` function : It will create the confidence intervals on the differencs for each possible combination of two group periods, using a determined confidence level of $\alpha = 0.95$.

The first comparison will be the average year temperatures between each of the 6 group periods :

```
summary(aov(t.ave.c ~ period, ny.full.138y))
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## period          5   16420    3284   34.35 <2e-16 ***
## Residuals   50360 4814467      96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the *p-value* is almost zero, there is strong evidence that at least one pair of population average temperatures are different from each other, so H0 is rejected in favour of HA.

```
pairwise.t.test(ny.full.138y$t.ave.c, ny.full.138y$period, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ny.full.138y$t.ave.c and ny.full.138y$period
##
##   1       2       3       4       5
## 2 0.03010 -       -       -       -
## 3 3.3e-06 0.54801 -       -       -
## 4 2.3e-14 1.5e-05 0.07769 -       -
## 5 < 2e-16 8.4e-10 0.00012 1.00000 -
## 6 < 2e-16 1.2e-13 2.0e-07 0.05887 1.00000
##
## P value adjustment method: bonferroni
```

```
TukeyHSD(aov(t.ave.c ~ period, ny.full.138y), conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = t.ave.c ~ period, data = ny.full.138y)
##
## $period
##          diff          lwr       upr     p adj
## 2-1 0.4663251  0.036181466 0.8964687 0.0245547
## 3-1 0.7819924  0.351784673 1.2122001 0.0000033
## 4-1 1.2040329  0.773914893 1.6341509 0.0000000
## 5-1 1.4554379  1.025409513 1.8854663 0.0000000
## 6-1 1.6392628  1.209221657 2.0693040 0.0000000
## 3-2 0.3156673 -0.114540398 0.7458750 0.2918826
## 4-2 0.7377078  0.307589821 1.1678258 0.0000151
## 5-2 0.9891128  0.559084441 1.4191412 0.0000000
## 6-2 1.1729378  0.742896586 1.6029789 0.0000000
## 4-3 0.4220405 -0.008141598 0.8522226 0.0581049
## 5-3 0.6734455  0.243353009 1.1035380 0.0001183
## 6-3 0.8572705  0.427165155 1.2873757 0.0000002
## 5-4 0.2514050 -0.178597733 0.6814078 0.5543722
## 6-4 0.4352300  0.005214411 0.8652455 0.0453251
## 6-5 0.1838249 -0.246100996 0.6137509 0.8280277
```

Conclusion:

Insert conclusion here...

References:

- Wikipedia - Global Warming (viewed on April 18, 2015) (http://en.wikipedia.org/wiki/Global_warming)

- NOAA - National Climatic Data Center (data retrieved on April 14, 2015) (https://gis.ncdc.noaa.gov/map/viewer/#app=cdo)
- Timeanddate.com - Solstices & Equinoxes for New York (data retrieved on April 14, 2015) (http://www.timeanddate.com/calendar/seasons.html)