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 environmental warming.

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 Oceanic and Atmospheric Administration) 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.

Finally, it is possible to state the research question to be answered in this study :

- Had the average temperature of New York city actually increased over the last 138 years, from 1876 to 2014 ?

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's database gives the the minimum and maximum temperatures for each recorded day, in tenths of degrees Celsius. For the purpose of this research, the average temperature between the daily minimum and maximum was calculated and included in the original data set.

In order to allow the possible future investigation of the influence of the global warming over 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 expected to be more evident after long 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

To help in this identification, a variable called *period* was also created and included in the data set.

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). All the R packages used were the latest stable versions available at CRAN. The final data set and scripts can be downloaded at https://github.com/rodrigodmartins/Data-Analysis-and-Statistical-Inference/tree/master/Grading%20project.

Here it is possible to check some entries of the final data set. To a description of each variable, please refer to *apendix* section.

```r
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/Dat
load("ny.138y.RData")
panderOptions("table.split.table", 130)
panderOptions("keep.trailing.zeros", TRUE)
pander(head(ny.full.138y, 4))
```

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.

The variable **period** is **categorical** and identifies the group of years each temperature measurement belongs to (as discussed in *Data* section). The variable **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 of this study), but **no causality** statements 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, for example.

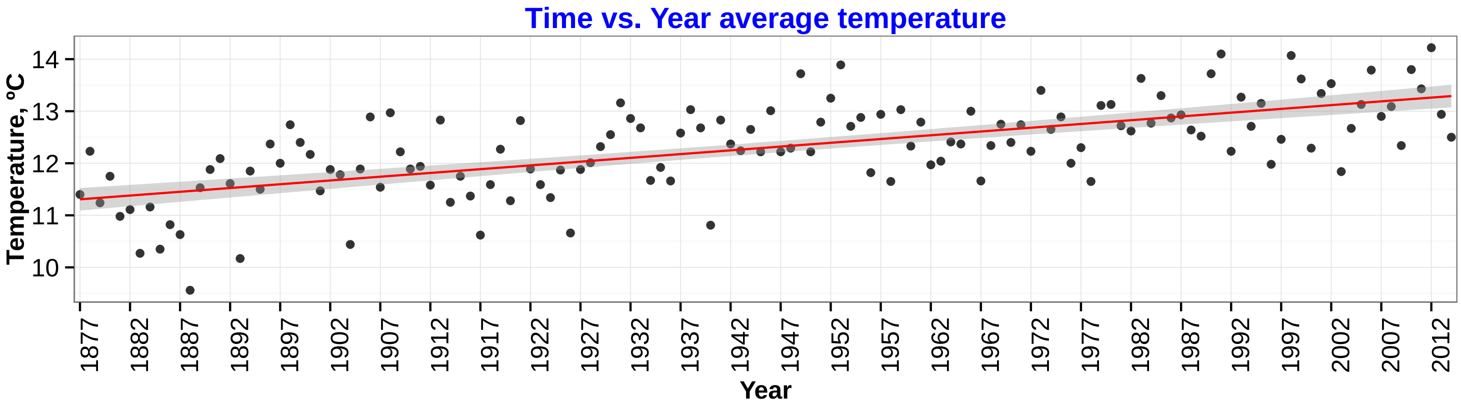
Finnaly, the main possible source of bias in the data can be attributed to the method of temperature measurement and data collection. During the 138 years, it is expected that the technology used for data collection have changed. It means that the accuracy and precision of temperature measurements may 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, with the inclusion of a regression line.

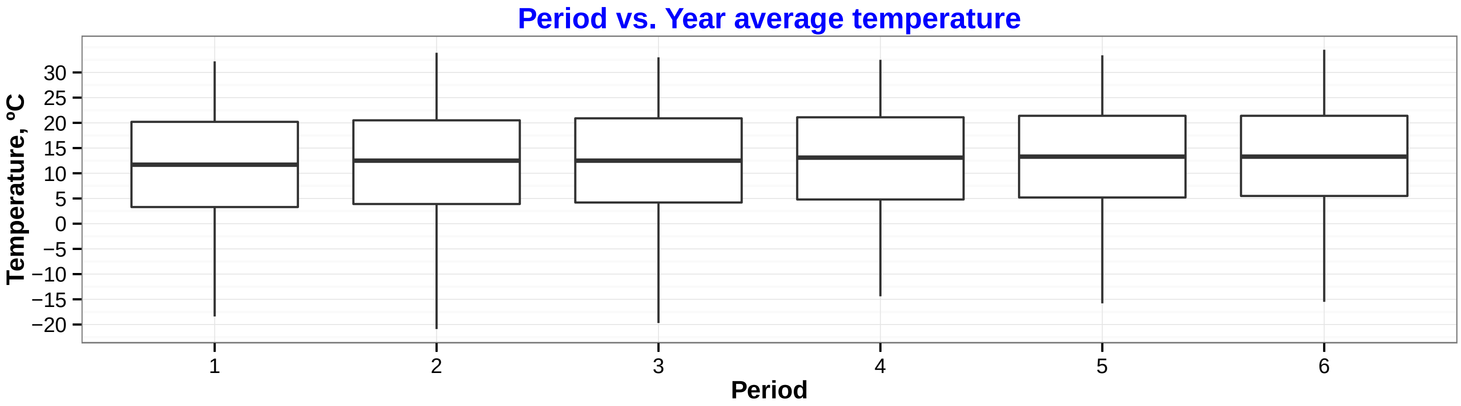
```r
if ("ggplot2" %in% rownames(installed.packages()) == FALSE) {
    install.packages("ggplot2")
}
suppressPackageStartupMessages(library(ggplot2))
h1 <- ggplot(ny.138y.t.ave.year, aes(year, t.ave.c.year)) + theme_bw() + geom_point(alpha = 0.8) +
    geom_smooth(method = "lm", color = "red", aes(group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



It is possible to visualize that the average temperature of New York city have been increasing along the years.

Boxplots were also evaluated.

```
j1 <- ggplot(ny.full.138y, aes(period, t.ave.c)) + theme_bw() + geom_boxplot() +
    labs(x = "Period", y = "Temperature, ºC", title = "Period vs. Year average temperature") +
    scale_y_continuous(breaks = seq(-20, 30, 5)) + theme(axis.title.x = element_text(face = "bold",
    size = 12), axis.title.y = element_text(face = "bold", size = 12), plot.title = element_text(face = "bold",
    size = 14, color = "blue"))
j1
```



The data set summary with descriptive statistics shows :

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

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

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 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 :

- H_0 : $\mu_{1877-1899} = \mu_{1900-1922} = \mu_{1923-1945} = \mu_{1946-1968} = \mu_{1969-1991} = \mu_{1992-2014}$
- H_A : At least one pair of means are different from each other.

The H_0 hypothesis means that the mean temperature over the six periods under evaluation are all the same. On the other hand, H_A, the alternative hypothesis, is that at least one pair of population 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 may be true in some extent. 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 in each of the periods are assumed to be 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 represent 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 among the six different periods are different, additional tests are required to identify which pair of periods 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 variables of interest in the data set.
2. `pairwise.t.test` function : It will compare all possible combinations of two periods to each other, using the *Bonferroni correction* for α.
3. `TukeyHSD` function : It will also compare the mean value of all possible combinations of two periods to each other, and additionally create the confidence intervals for the difference, using *t-statistic* at a determined confidence level of $\alpha = 0.95$. For reference, *Tukey HSD* stands for *Tukey's honest significance difference test.*

The first comparison is the average temperatures in 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 H_0 **is rejected in favour of** H_A.

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

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  ny.full.138y$t.ave.c and ny.full.138y$period
##
##   1       2       3       4       5
## 2 0.03766 -       -       -       -
## 3 5.3e-06 0.57562 -       -       -
## 4 4.0e-14 1.6e-05 0.07477 -       -
## 5 < 2e-16 1.1e-09 0.00012 1.00000 -
## 6 < 2e-16 7.1e-14 1.3e-07 0.04732 1.00000
##
## P value adjustment method: bonferroni
```

	p-value	p-value < 0.05	Significant temperature difference ?
1 vs 2	0.03010	TRUE	TRUE
1 vs 3	0.00000	TRUE	TRUE
1 vs 4	0.00000	TRUE	TRUE
1 vs 5	0.00000	TRUE	TRUE
1 vs 6	0.00000	TRUE	TRUE
2 vs 3	0.54801	FALSE	FALSE
2 vs 4	0.00002	TRUE	TRUE
2 vs 5	0.00000	TRUE	TRUE
2 vs 6	0.00000	TRUE	TRUE
3 vs 4	0.07769	FALSE	FALSE
3 vs 5	0.00012	TRUE	TRUE
3 vs 6	0.00000	TRUE	TRUE
4 vs 5	1.00000	FALSE	FALSE
4 vs 6	0.05887	FALSE	FALSE
5 vs 6	1.00000	FALSE	FALSE

Using the *pairwise t-test*, it is possible to identify which are period the combinations that actually show a significant difference in their average temperature, and which does not.

```
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
```

	average temperature diff	CI 95%	adjusted p-value	p-value < 0.05	Significant temperature difference ?
1 vs 2	0.47	0.04 to 0.90	0.02455	TRUE	TRUE
1 vs 3	0.78	0.35 to 1.21	0.00000	TRUE	TRUE
1 vs 4	1.20	0.77 to 1.63	0.00000	TRUE	TRUE
1 vs 5	1.46	1.03 to 1.89	0.00000	TRUE	TRUE

	average temperature diff	CI 95%	adjusted p-value	p-value < 0.05	Significant temperature difference ?
1 vs 6	1.64	1.21 to 2.07	0.00000	TRUE	TRUE
2 vs 3	0.32	-0.11 to 0.75	0.29188	FALSE	FALSE
2 vs 4	0.74	0.31 to 1.17	0.00002	TRUE	TRUE
2 vs 5	0.99	0.56 to 1.42	0.00000	TRUE	TRUE
2 vs 6	1.17	0.74 to 1.60	0.00000	TRUE	TRUE
3 vs 4	0.42	-0.01 to 0.85	0.05810	FALSE	FALSE
3 vs 5	0.67	0.24 to 1.10	0.00012	TRUE	TRUE
3 vs 6	0.86	0.43 to 1.29	0.00000	TRUE	TRUE
4 vs 5	0.25	-0.18 to 0.68	0.55437	FALSE	FALSE
4 vs 6	0.44	0.01 to 0.87	0.04533	TRUE	TRUE
5 vs 6	0.18	-0.25 to 0.61	0.82803	FALSE	FALSE

Finally, the *Tukey's HSD test* also performs a the *pairwise t-test* for all possible pairs of average temperatures for each period, and additionally calculates the confidence intervals (CI 95%) for the difference.

It is interesting to verify that there is a difference if the *p-values* calculated by `pairwise.t.test` and `TukeyHSD`. In general, the *p-values* from `pairwise.t.test` (with Bonferroni correction) are higher that the *p-values* from `TukeyHSD`. This bahavior is expected, because when comparing the means for the levels of a factor in an ANOVA, a simple comparison using t-tests will inflate the probability of declaring a significant difference when it is not in fact present. It happens because the intervals are calculated with a given coverage probability for each interval but the interpretation of the coverage is usually with respect to the entire family of intervals. John Tukey introduced intervals based on the range of the sample means rather than the individual differences. The intervals returned by `TukeyHSD` function are based on this Studentized range statistics.

Because of that, between both period average temperatures methods of comparison, the conclusion of this study will be based in the results obtained by the *Tukey's honest significance difference test*, only.

Conclusion:

At this point, it is important to review the stated question in the introduction section :

- Had the average temperature of New York city actually increased over the last 138 years, from 1877 to 2014 ?

After the evaluations performed, through the ANOVA hypothesis test and Tukey's honest significance test, it is possible to confirm that strong evidences could be found in favour of the increase of the average temperature of New York city over the past 138 years.

Additionally to this discovery, it was also possible to conclude that the temperature increases are in fact mild. The pairwise comparisons performed by `TukeyHSD` showed that the following periods did not show significant differences in their average temperatures :

- Period 2 vs. Period 3.
- Period 3 vs. Period 4.
- Period 4 vs. Period 5.
- Period 5 vs. Period 6.
- *(The only exception was the comparison between Period 1 vs. Period 2).*

The reason for this behavior is that the comparisons are for subsequent periods of time. It shows that in this relativelly short period of time (23 years), it is not easy to find a statistically significant difference in the mean temperatures. On the other hand, all the other comparisons of not subsequent 23 years periods, showed enough evidence of increase.

According to Tukey's honest significance test, the increase in the average temperature from period 1 (1877 to 1899) to period 6 (1992 to 2014) was of **1.64ºC**, with a CI 95% of **1.21 to 2.07ºC**.

In a future research, it would be relevant to try to answer the following question :

- Is the average temperature increase more noticiable in one of the seasons of the year, or basically all the same ?

One may believe that the average temperature increases are more noticiable in specific seasons of the year. The current data set is already prepared to be used in the evaluation of this question, regarding that the variable *season* was included since the beginning. The same variables can be used, however, subsetted by *season*.This question adds value to the conclusion of the present study, and may be helpful in the deeper explanation of the impacts of the global warming on Earth.

References:

- Wikipedia - Global Warming (viewed on April 18, 2015) (http://en.wikipedia.org/wiki/Global_warming)
- Wikipedia - Tukey's range test (viewed on April 19, 2015) (http://en.wikipedia.org/wiki/Tukey's_range_test)
- 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)

Appendix:

Below, there is the first page (first 20 entries) of the data set `ny.full.138y`.

The variables can be identified as :

- **date :** Date when the cases (units of observation) were colected.
- **season :** Season of the year (Winter, Spring, Summer or Autumn).
- **year :** Year of the observation.
- **season.year :** Season and year of the observation.
- **period :** Period of the observation, from 1 to 6 (please see the identification of each period in *data* section.
- **moon.phase :** The moon phase in the date of observation (New, First quarter, Full, Third quarter).
- **t.max.c :** Maximum temperature in the date of observation, in degrees Celsius.
- **t.min.c :** Minimum temperature in the date of observation, in degrees Celsius.
- **t.ave.c :** Average temperature in the date of observation, in degrees Celsius.

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
1876-12-25	Winter	1877	Winter-1877	1	First quarter	-5.0	-9.4	-7.2
1876-12-26	Winter	1877	Winter-1877	1	First quarter	0.6	-5.0	-2.2
1876-12-27	Winter	1877	Winter-1877	1	First quarter	0.0	-5.6	-2.8
1876-12-28	Winter	1877	Winter-1877	1	First quarter	-4.4	-7.8	-6.1
1876-12-29	Winter	1877	Winter-1877	1	First quarter	2.8	-5.0	-1.1
1876-12-30	Winter	1877	Winter-1877	1	First quarter	0.6	-5.6	-2.5
1876-12-31	Winter	1877	Winter-1877	1	Full	-1.7	-7.2	-4.5
1877-01-01	Winter	1877	Winter-1877	1	Full	-4.4	-8.9	-6.7
1877-01-02	Winter	1877	Winter-1877	1	Full	-4.4	-7.2	-5.8
1877-01-03	Winter	1877	Winter-1877	1	Full	-5.6	-11.1	-8.3
1877-01-04	Winter	1877	Winter-1877	1	Full	-7.8	-11.7	-9.8
1877-01-05	Winter	1877	Winter-1877	1	Full	-4.4	-11.1	-7.8
1877-01-06	Winter	1877	Winter-1877	1	Full	1.1	-9.4	-4.2
1877-01-07	Winter	1877	Winter-1877	1	Third quarter	7.2	1.1	4.2
1877-01-08	Winter	1877	Winter-1877	1	Third quarter	6.1	-5.6	0.2
1877-01-09	Winter	1877	Winter-1877	1	Third quarter	-5.6	-10.0	-7.8