1 – Introduction to the subject

Climate change is at the origin of many disasters that are more and more common. An example is the increase of the temperature that would leads to many issues. From the increase of sea level, the movement of population to the death of a big part of a vegetal and animal ecosystem these threats need to be taken seriously by scientist and politicians. Both play a key role; on one side the scientist needs to convince politics that there is truly a climate change issue and on the other hand politics need to change the population behaviours. The purpose of the following analysis is to adopt the scientist perspective and to check whether we observe an increase of extreme events frequency throughout the time.

First, we estimate the distribution of the average temperature of each year. Second, we compute the probability of having an average temperature higher than 86°F (equivalent to 30°C) for each year. Finally, we conclude looking at the extreme event probability evolution.

Technically, the use case purpose is to illustrate the use of Indirect Cross Validation (ICV) and to compare it to Least square cross validation (LSCV) and normal scale rule (NSR) to get the optimal bandwidth. We focus our attention on Algeria where the distribution throughout the time has good property to compare the approaches.

2 – Data and packages description

The dataset comes from the University of Dayton through Kaggle. The content concerns the daily level average temperature over major cities of the world from 1995 to 2019. We have information about location (country, state, region, city), and time (Year, Month, Day).

To do our analysis, we use ***ICV***package for Indirect cross validation estimation of the bandwidth, ***KernSmooth*** package for density estimation and local polynomial regression and ***np***package for bandwidth estimation with Least square cross validation method.

3 – Kernel density estimation

1. Descriptive analysis

Une image contenant texte, capture d’écran, diagramme

Description générée automatiquementFirst, we analyse the data with some visualizations. The distribution of the temperature in Algeria from 1995 to 2019 display two modes.

Figure 1

Une image contenant texte, capture d’écran, Caractère coloré, diagramme

Description générée automatiquementThis is due to the hot temperatures during summer and cold one during winter.

Figure 2

To be sure that we have enough variability in the distribution (unimodal, multi-modal, skewed distribution, high and low kurtosis) we focus our attention on Algeria which display good property. From figure 2 we see that we have very concentrated distribution in 1995, dispersed one in 2004, a bit of skewness in 2010 and this is what we want to compare our different approaches of bandwidth estimation.

Une image contenant capture d’écran, texte, Caractère coloré, conception

Description générée automatiquement

Figure 3

From figure 3 we can see that the distribution evolution in May and June switch progressively toward higher temperatures through the time. For other months it is not clear. Also, we see that we have cyclical evolution of temperature which is coherent with the location of Algeria on the map and the bimodal distribution in figure 1.

b – Comparison NSR,LSCV, ICV

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquement Our aim is to compare the normal scale rule to least square cross validation and indirect cross validation method. We estimate the density of the average temperature for each year from 1995 to 2019. The idea is to display advantages and drawbacks of the methods. We display only the interesting distribution.

Figure 4

Une image contenant texte, diagramme, capture d’écran, Tracé

Description générée automatiquementLooking at the distribution of the average temperature in 1996, we see that it is unimodal and approximately symmetric. These are the conditions to use NSR correctly. We see that NSR performs as good as ICV or LSCV, but this is not surprising in this case since the conditions are satisfied.

Figure 5

Une image contenant texte, diagramme, Tracé, capture d’écran

Description générée automatiquementThe distribution of the average temperature in 2008 in figure 5 is interesting since we have 3 modes. This is the type of data where most of the bandwidth estimation methods struggle. Since the condition to use NSR are not satisfied, it is not surprising that it poorly estimates the distribution. Conversely, bandwidth estimation using ICV or LSCV looks to capture the distribution and 3 modes equivalently well.

Figure 6

The distribution of the average temperature in 2016 display two modes. It is clear here that LSCV poorly estimate the distribution. There is too much variance in this estimation compared to ICV and NSR. LSCV method estimates a way too low bandwidth, and this shows the non-stability of the method.

c – Evolution of the extreme events

We have an estimate of the average temperature distribution by year. From these distributions, we computed the probability of having an average temperature higher than 86°F as follow:

FORMULA

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementIn the figure 7, we see the evolution of this probability in red. The local polynomial fit shows the increasing trend of the extreme event probability.

Figure 7

4 – Conclusion

The short application of ICV, NSR and LSCV shows that ICV outperform NSR and LSCV. The method displays much flexibility than other ones in term of conditions to be applicable. Also, the method is more stable and performs better than LSCV.

Concerning the use case, the evolution of the probability of having extreme event (a daily average temperature of 86°F ) display an increasing and warming trend. A deeper time series analysis would enable us to do forecast of what could be the extreme event probability in few years. However, this is not what we are interested in for this application part, thus we stopped our analysis to this conclusion.