Non-Parametric Models Project

CONDETTE Théo, NASR Rodrigue, NGUYEN Hoai Nam

December 2023

1 Introduction

Kernel Density Estimation (KDE) is a nonparametric technique widely used in statistical analysis to estimate the density function of continuous data. This method is particularly suitable for data sets that exhibit complex distribution patterns, making it a useful tool in a variety of scientific domains. In this report, we will discuss the implementation of KDE, focusing on the Indirect Cross Validation (ICV) approach to optimal bandwidth selection.Then we will use the strength of KDE and ICV to answer climate change questions and assess the evolution of extreme weather events in Algeria.

2 Motivation: From Histogram to Kernel

In this part, we suppose that we have n observations $(X_1, ..., X_n)$ from a random observation X, and we want to estimate the density function f of X.

2.1 Histogram Density Estimator

To construct the histogram estimator of the density f of X, we split the range of values of X to N subintervals B_j for $j = 1, ..., N$, with bandwidth h. Then, over each subinterval B_j, we estimate the density f by a constant $\hat{f}_j$ (the normalized frequence). We finally define globally $\hat{f}_h(x)$ by combining all the possible $\hat{f}_j$ for each subinterval, while ensuring that $\int \hat{f}_h(x)dx = 1$. The histogram density estimator (HDE) will be:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{N} \mathbf{1}(X_i \in B_j)\mathbf{1}(x \in B_j)$$

The optimal choice of h is found by minimizing the $AMISE(\hat{f}_h)$ with respect to h. Let's note that $\hat{f}_h(x)$ converges pointwisely to $f(x)$, since its MSE converges to 0 (locally) as n tends to ∞.

But, since in most cases f is originally unknown, it may not be easy to find the optimal value of h. Also, the estimation itself is a stair-case function, since it's defined using the sum of indicator functions. In this case, we use smoother estimators, like the Kernel Density Estimator.

2.2 Kernel Density Estimation

The Kernel Density Estimator is a smoother estimator that uses a Kernel $K(.)$, which is a function that satisfies the following conditions:

$$K(u) \geq 0, \int K(u)du = 1, \int uK(u)du = 0$$

In other words, it is a density function with a zero mean. For example, the "Epanechnikov" Kernel is defined as: $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u| \leq 1)$. For a chosen bandwidth h, the Kernel Density Estimator $\hat{f}(x)$ is defined as follows:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

Here, the bandwidth h determines the width of the kernel function and, consequently, the smoothness of the estimate. A larger bandwidth results in a smoother estimate, while a smaller bandwidth leads to a more variable estimate. Again here, the optimal global bandwidth h^* is found by minimizing the $AIMSE(\hat{f})$, which also depends on f. On the other side, the choice of Kernel is done by minimizing the $AIMSE^*(\hat{f})$, which is itself the $AIMSE(\hat{f})$ with bandwidth h^*, or equivalently solving:

$$\min_{K \geq 0; \int K(u)du=1; \int uK(u)du=0} C(K)$$

with

$$C(K) = \left(\int_{-\infty}^{+\infty} K^2(u)du\right)^{1/5} \left(\int_{-\infty}^{+\infty} u^2K^2(u)du\right)^{2/5}$$

This minimization problem has a family of solutions based on the Epanechnikov kernel presented above.

3 Theory of Cross-Validation for Kernel Density Estimation

As for any parameter-based optimization algorithm, the selection of optimal parameters to train the data always plays a vital role in the model's performance improvement. A popular and readily implemented heuristic for selecting the parameter is cross-validation. In our context of kernel estimation, cross-validation has become a classical optimization method for selecting the smoothing parameter so far. In this part, we review briefly some popular classical cross-validation methods for kernel density estimation and stay focused on Indirect Cross-Validation (ICV).

3.1 Least Squares Cross-Validation or Unbiased Cross-Validation

Least Squares Cross-Validation (LSCV) was suggested for the first time by Rudemo (1982) and Bowman (1984) and has played a fundamental role in preceding research on smoothing density estimation by the kernel function. This method is intuitive because it starts from the original point of the problem which is the minimization of the classical metrics in approximation - Mean Integrated Square Error (MISE), or simply Integrated Square Error (ISE). LSCV expands this metric and considers the following optimization program:

$$\hat{h}_{LSCV} = \arg\min_h ISE(h)$$

with

$$ISE(h) = \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx + \int f^2(x)dx$$

where f is the density function, $\hat{f}_h$ is the estimate of density with bandwidth h, $\int \hat{f}_h^2(x)dx$ is computed using the data and $\int \hat{f}_h(x)f(x)dx$ is estimated by the leave-one-out cross-validation method to avoid overfitting and make better use of the data:

$$\hat{f}_{-i,h}(X_i) = \frac{1}{(n-1)h}\sum_{j=1,j\neg i}^{n} K\left(\frac{X_i - X_j}{h}\right)$$

The bandwidth is selected as follows:

$$\hat{h}_{LSCV} = \arg\min_h CV_{LS}(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,h}(X_i)$$

The leave-one-out estimate is proved to be an unbiased estimator and the estimator of $ISE(h)$ is consequently unbiased. Stone (1984) proved that the optimal bandwidth obtained from LSCV converges in probability to the true optimal bandwidth followed by the almost sure convergence of its ISE to its minimum value.

3.2 Likelihood Cross-Validation

Likelihood Cross-Validation involves playing with different ways of looking at a set of data (bandwidths) while keeping the original data fixed. Having a group of data points with the estimated density function $\hat{f}_h$, the evaluation of how likely the estimate is for various bandwidths could be implemented by introducing a new datapoint X^* into the group of the same distribution. Its likelihood ($log\hat{f}(X^*)$) helps to pick the best bandwidth to see the data more clearly. However, there is no additional observation available in reality, the solution considers inversely omitting a randomly selected observation from the original data, says X_i, and computing $\hat{f}_{h,-i}(X_i)$ (This is a leave-one-out estimator). X_i is a random observation, and so is the estimate. The evaluated score function is then to be taken as the log-likelihood average:

$$MLCV(h) = \frac{1}{n}\sum_{i=1}^{n} log\hat{f}_{h,-i}(X_i)$$

Maximizing $MLCV(h)$ with respect to h minimizes the Kullback-Leibler information "distance" between $\hat{f}_h(x)$ and $f(x)$. So it follows:

$$\hat{h}_{ML} = \arg\max_h MLCV(h)$$

Marron (1985) demonstrated the asymptotic consistency of MLCV under certain conditions. Despite requiring more conditions compared to the two earlier approaches, this asymptotic consistency shares the same convergence rate as LSCV for smooth densities and can be faster for less smooth densities. However, MLCV faces challenges when dealing with heavy or long tails. Bandwidths derived from MLCV exhibit high variability and often result in under-smooth density estimates, introducing unwanted spurious bumpiness.

3.3 Indirect Cross-Validation (ICV)

The Indirect Cross-Validation (ICV) method, proposed by Savchuk, Hart, and Sheather (2010), slightly outperforms least squares CV in terms of mean integrated squared error.

3.3.1 The basic method

ICV initially chooses the bandwidth of an L-kernel estimator using LSCV:

$$\hat{b}_{UCV} = \arg\min_b LSCV(b) \approx \arg\min_b ISE(b) = \arg\min_b (R(\hat{f}_b) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,b}(X_i))$$

where $\hat{f}_b$ is an L-Kernel estimation. L-Kernel is a second-order kernel, a linear combination of two Gaussian kernels of the form:

$$L(x, \alpha, \sigma) = (1+\alpha)\phi(x) - \frac{\alpha}{\sigma}\phi\left(\frac{x}{\sigma}\right)$$

where ϕ is the standard normal density and α and σ are positive constants empirically determined.

Afterward, ICV multiplies the bandwidth chosen $\hat{b}_{UCV}$ by a known constant, say C, to get the indirect bandwidth $\hat{h}_{ICV}$. C depends on no unknown parameter and is defined as:

$$h_n = \left(\frac{R(\phi)\mu_{2L}^2}{R(\text{L})\mu_{2\phi}^2}\right)^{1/5} b_n = Cb_n$$

where h_n and b_n are the bandwidths that minimize the $AMISE$ of ϕ-Kernel and L-Kernel estimators respectively; $R(g) = \int g(x)^2 dx$, $\mu_{2L} = \int x^2 f_L(x) dx$;

f_L is the density function of the L-Kernel; $\mu_{2\phi} = \int x^2 f_\phi(x)dx$, f_ϕ is the density function of the Gaussian distribution.

Finally, $\hat{h}_{ICV}$ is used to estimate the density function

$$\hat{f}_{\hat{h}_{\text{ICV}}} = \frac{1}{n\hat{h}_{\text{ICV}}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\hat{h}_{\text{ICV}}}\right)$$

where K is the Gaussian kernel. In summary, we use L-Kernel to get $\hat{b}_{UCV}$ and then re-scale $\hat{b}_{UCV}$ by a constant parameter-independent multiplier C to get $\hat{h}_{ICV}$. We plug $\hat{h}_{ICV}$ in Kernel estimation with Gaussian kernel to estimate the density: $\hat{f}_{\hat{h}_{\text{ICV}}}$. We use then two kernels: L-Kernel for cross-validation to obtain $\hat{b}_{UCV}$ and ϕ-Kernel for density estimation.

3.3.2 Selection kernels

One question posed is why not use the same kernel L for both cross-validation which chooses optimal bandwidth and for density estimation which chooses the optimal kernel density estimator. We can bypass the step of rescaling $\hat{b}_{UCV}$ and simply estimate f by a L-kernel estimator with bandwidth $\hat{b}_{UCV}$, that is

$$\hat{f}_{\hat{b}_{\text{UCV}}} = \frac{1}{n\hat{b}_{\text{UCV}}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\hat{b}_{\text{UCV}}}\right)$$

where K is L-Kernel. The response is that even though the L-kernel is best for cross-validation purposes, the L-kernel is very inefficient for estimating f. L-kernel estimator based on a sequence of ICV-optimal kernels has a MISE that converges to 0 at a rate lower than $n^{-1/2}$, while ϕ-Kernel estimator has MISE that converges to 0 at a rate $n^{-4/5}$, which is much faster than L-kernel estimator. The L-kernel proves to be inefficient for density estimation; however, it exhibits efficiency in cross-validation. It lies in the principle that "the more challenging the function is to estimate, the better cross-validation tends to perform." Therefore, this inefficient kernel L is used to make the function more difficult to estimate so that better cross-validation seems to perform. Yet, LSCV is proven to outperform other methods when the density is highly structured. Consequently, L-kernel along with LSCV is the best for bandwidth selection.

3.3.3 Large sample theory

L-kernels for bandwidth selection imply in practical finding the optimal (α, σ) which simultaneously minimizes MSE. First of all, obtaining the optimal σ requires minimization of the second moment of relative error $\frac{\hat{h}_{ICV} - h_0}{h_0}$ which is the second moment of the indirect bandwidth itself $\hat{h}_{ICV}$. Under some assumptions, the second moment, or the variance of the indirect bandwidth is expressed as:

$$\text{Var}(\hat{h}_{\text{ICV}}) = S_n^2 + B_n^2 = \left[\left(\frac{1}{\sigma^{2/5} n^{1/10}}\right) \frac{R(f)^{1/2}}{R(f'')^{1/10}} C_\alpha\right]^2 + \left[\left(\frac{\alpha}{n}\right)^{2/5} \frac{R(f)'''}{R(f'')^{7/5}} D_\alpha\right]^2$$

where C_α and D_α are function of α independent from the function f itself. Minimizing the expression for σ results in the asymptotically optimal selection of σ, and substituting this value yields the corresponding asymptotically optimal mean squared error:

$$MSE_{n,opt} = n^{-1/2} C_\alpha D_\alpha \left[\frac{R(f^{'''}) R(f)^{1/2}}{R(f^{''})^{3/2}} \right]$$

The asymptotically optimal means squared error converges to 0 at the rate $n^{-1/2}$, which implies that the relative error of $\hat{h}_{ICV}$ converges to 0 at the rate $n^{-1/2}$. This convergence rate is much faster than the corresponding rates for LSCV which is $n^{-1/10}$. On the other hand, α is not confounded with f in $MSE_{n,opt}$, then a single optimal value of α, that is independent of f and minimizes the function $C_\alpha D_\alpha$ is at $\alpha_0 = 2.4233$. It is noticed that while the variance of $\hat{h}_{ICV}$ converges to 0 at a faster rate than LSCV bandwidth, the squared bias is not negligible. (α, σ) should be selected such that variance and squared bias are balanced and mean squared error converges to 0 at a faster rate.

This selection of α and σ is done by determining the minimizers of the asymptotic mean squared error of $\hat{h}_{ICV}$ for various sample sizes n and densities f throughout a polynomial regression of an appropriate degree. A single expression for the asymptotic mean squared error that is valid for either large or small values of α and a slightly enhanced version of the asymptotic bias of $\hat{h}_{ICV}$ are used so that the α depends on both n and f. The five normal mixtures defined in the article by Marron and Wand (1992) are considered, resulting in a study of a fairly representative range of density shapes. These are Gaussian density, Skewed uni-modal density, Bimodal density, Separated bimodal density, and Skewed bimodal density. ICV approach is therefore adequate to be applied in the data with one of the mentioned densities or a combination of them. The following models for α and σ are derived:

$$\alpha_{mod} = 10^{3.390 - 1.093 log10(n) + 0.025 log10(n)^3 - 0.00004 log10(n)^6}$$

$$\sigma_{mod} = 10^{-0.58 + 0.386 log10(n) - 0.012 log10(n)^2}, 100 \leq n \leq 500000$$

Given that uni-modal densities are more commonly encountered than multimodal densities in practical scenarios, the model values tend to exhibit bias toward bimodal cases. The extensive experience implemented by authors shows that the penalty for using good bimodal choices for α and σ when in fact the density is uni-modal is an increase in the upward bias of $\hat{h}_{ICV}$. The implementation of ICV, however, guards against over-smoothing by using an objective upper bound on the bandwidth. These two values of α and σ are confidently recommended in practice.

3.3.4 Local ICV

The smallest local minimizer of ICV curve is suggested by authors since ICV does not have LSCV's tendency to under-smooth. Let $\hat{f}_b$ be a kernel estimate that employs a kernel in the class L, and define, at the point x, a local ICV curve by

$$ICV(x,b) = \frac{1}{w}\int \phi\left(\frac{x-u}{w}\right)\hat{f}_b^2(u)du - \frac{2}{nw}\sum_{i=1}^{n}\phi\left(\frac{x-X_i}{w}\right)\hat{f}_{b,-i}^2(X_i), b > 0$$

The quantity w determines the degree to which the cross-validation is local, local ICV converges to global ICV as w increases. The bandwidth of a Gaussian kernel estimator at the point of x is taken to be $\hat{h}(x) = C\hat{b}(x)$ where $\hat{b} = \arg\min_b ICV(x,b)$. The choice of (α, σ) could be a reference from the global bandwidth selection since it is reasonable to select the same one as the global bandwidth for local bandwidth since locally the density should have relatively few features.

4 Practical application (R)

4.1 Introduction

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 lead to many issues. From the increase of sea level and the movement of population to the death of a big part of a plant and animal ecosystem, these threats need to be taken seriously by scientists and politicians. Both play a key role; on one side scientists have to prove to pliticians that there is truly a climate change issue and on the other hand politicians will have to change the population's behaviors. The purpose of the following analysis is to adopt the scientist's perspective and to check whether we observe an increase in extreme events frequency throughout time.

This analysis adopts a scientific perspective to assess whether there is a discernible escalation in the frequency of extreme climate events over time. The approach involves three key steps. Initially, the distribution of average annual temperatures is estimated. Subsequently, the probability of experiencing an average temperature exceeding 86°F (equivalent to 30°C) is calculated for each year. Finally, the study examines the evolution of the probability of extreme climate events. Technically, the objective of this use case is to showcase the application of Indirect Cross Validation (ICV) and compare it with Least Square Cross-Validation (LSCV) and the normal scale rule (NSR) to determine the optimal bandwidth. The focus is specifically on Algeria, as its temperature distribution over time provides a suitable basis for comparing these different approaches.

4.2 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.

4.3 Kernel density estimation

4.3.1 Descriptive analysis

First, we analyze the data with some visualizations. From Figure 1, the distribution of the temperature in Algeria from 1995 to 2019 displays two modes. This is due to the hot temperatures during summer and cold ones during winter.

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 displays good properties. From Figure 2 we see that we have a very concentrated distribution in 1995, a dispersed one in 2004, and a bit of skewness in 2010 and this is what we want to compare our different approaches to bandwidth estimation.

From Figure 3 we can see that the distribution evolution in July, August, and September switches progressively toward higher temperatures through time. For other months it is not clear. Also, we see that we have a cyclical evolution of temperature which is coherent with the location of Algeria and the bimodal distribution in Figure 1.

4.3.2 Comparison of NSR, LSCV with ICV

Our aim is to compare the NSR to LSCV and ICV methods. The idea is to display the advantages and drawbacks of the methods. To begin, we estimate the density of the average temperature for each year from 1995 to 2019. In what follows, we only display the interesting distribution.

Looking 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 well as ICV or LSCV, but this is not surprising in this case since the conditions are satisfied. The distribution of the average temperature in 2008 is interesting since we have three modes. This is the type of data where most of the bandwidth estimation methods struggle. Since the conditions 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. The distribution of the average temperature in 2016 displays two modes. It is clear here that LSCV poorly estimates 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 lack of stability of the method.

4.3.3 Evolution of the extreme events

To monitor the evolution of the extreme events in Algeria, we first need to define what is an extreme event. We define an extreme event, as an event that barely ever occurs. Thus, we define it as having a daily average temperature higher than the 99th percentile. This percentile corresponds to a daily average temperature higher than 86°F / 30°C. From our density estimation part, we have an estimation of the average temperature distribution by year. We can now compute the probability of having an average temperature higher than 86°F as follows:

$\mathbb{P}(X > p_{0.99}) = \int_{p_{0.99}}^{+\infty} \hat{f}(x)\, dx$

In Figure 5, we see the evolution of this probability in red. The local polynomial fit shows the increasing trend of the extreme event probability. This means that the probability of having a heat wave in a year is going upward with time, which is global warming.

4.4 Conclusion

The application of Indirect Cross-Validation (ICV), Normal Scale Rule (NSR), and Least Square Cross- Validation (LSCV) reveals that ICV surpasses both NSR and LSCV. This method exhibits greater flexibility in terms of conditions and looks to be more stable and effective than LSCV. Turning our attention to the use case, the analysis of the probability evolution of extreme events (defined as a daily average temperature of 86°F) indicates a discernible upward and warming trend. While a more in-depth time series analysis could allow us to forecast the potential probability of extreme events in the coming years, this specific application does not pursue that avenue.

References

[unk, 2020] (2020). Daily temperature of major cities. `https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities/data`.

[Guidoum, 2015] Guidoum, A. C. (2015). Kernel estimator and bandwidth selection for density and its derivatives. *Department of Probabilities and Statistics, University of Science and Technology, Houari Boumediene, Algeria.*

[Guidoum, 2020] Guidoum, A. C. (2020). Kernel estimator and bandwidth selection for density and its derivatives: The kedd package. *arXiv preprint arXiv:2012.06102.*

[Zambom and Ronaldo, 2013] Zambom, A. Z. and Ronaldo, D. (2013). A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42.

5 Annex

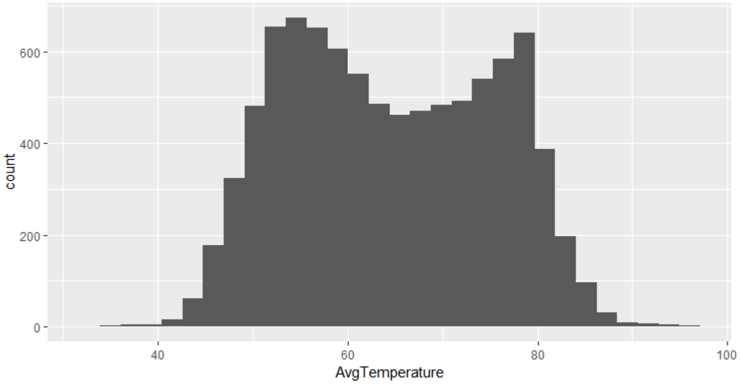

Figure 1: Distribution of the temperature in Algeria from 1995 to 2019

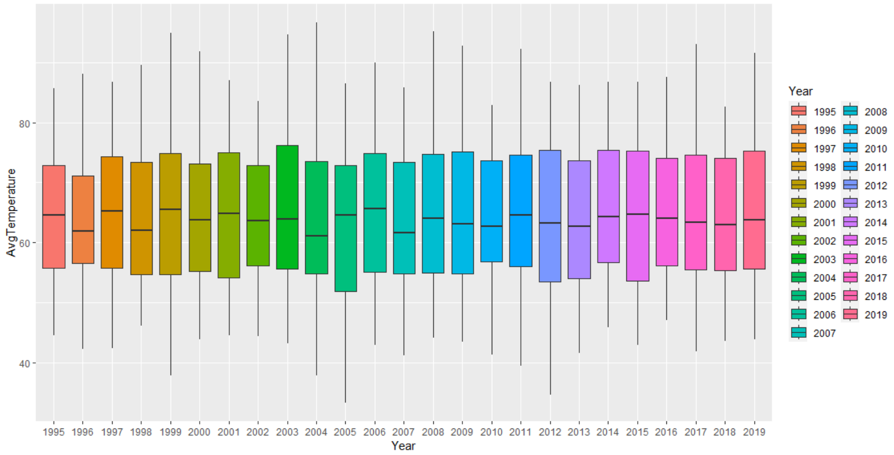

Figure 2: Evolution of the temperature distribution by year

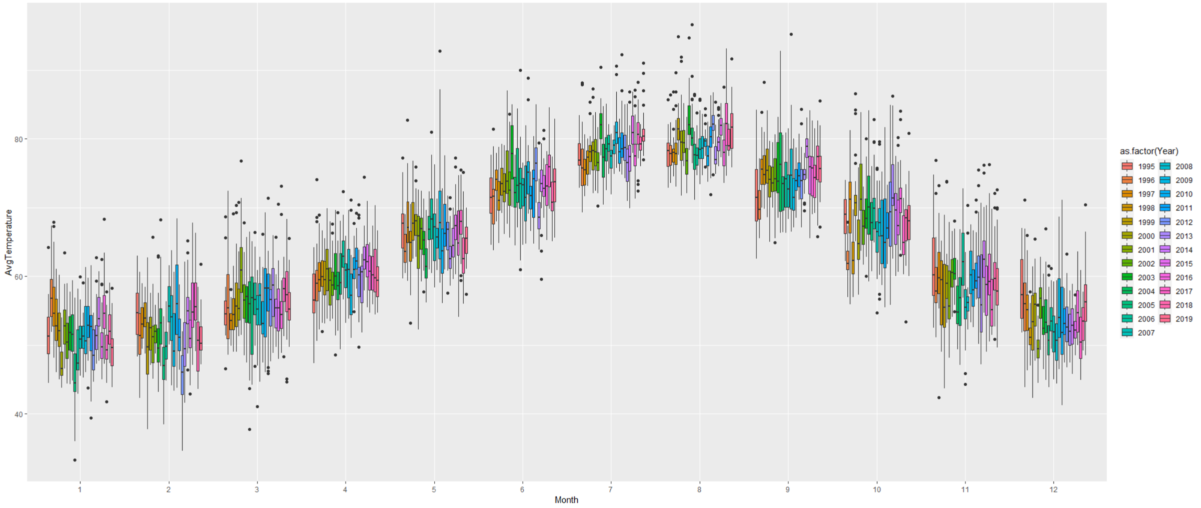

Figure 3: Evolution of the temperature distribution by month

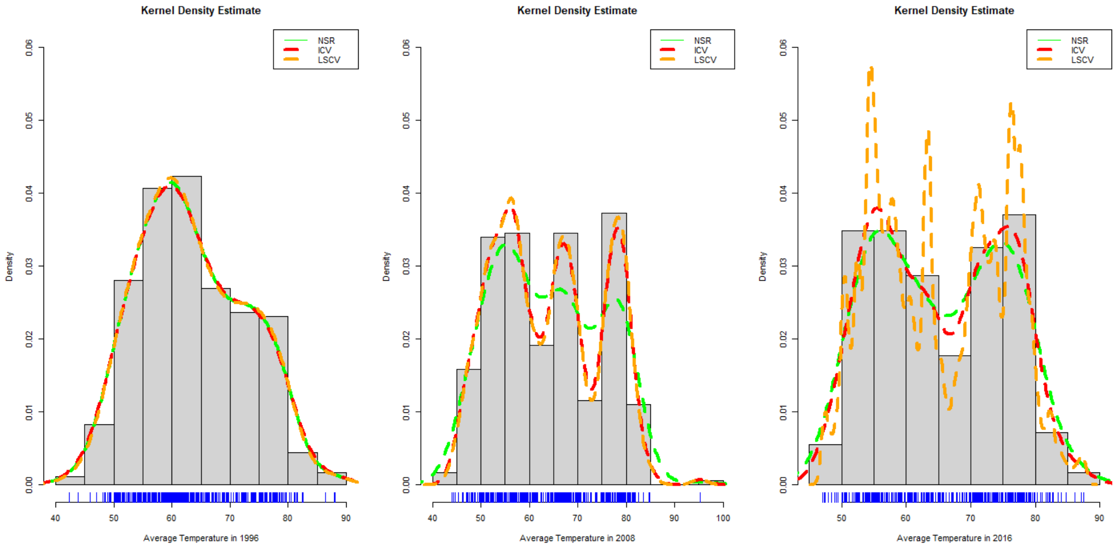

Figure 4: Some interesting kernel distribution estimation

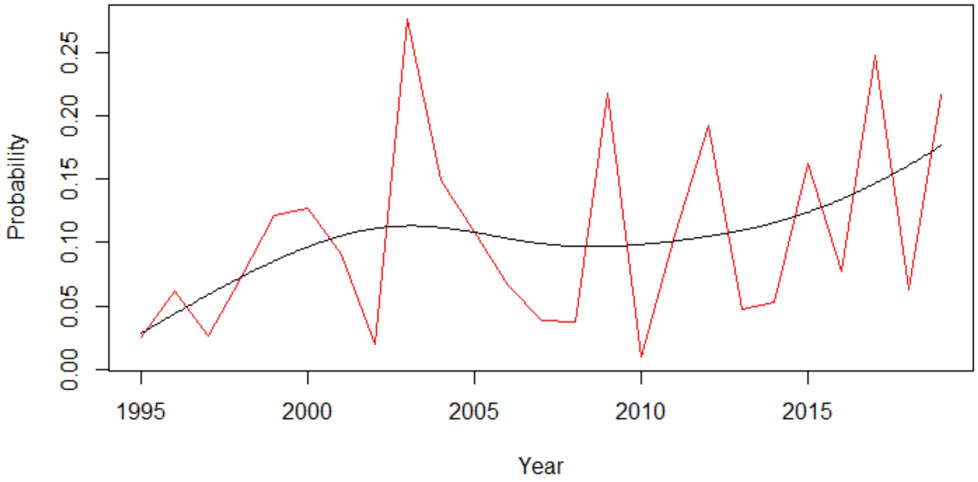

Figure 5: Evolution of extreme event probability in Algeria