



# LOW DIMENSIONAL EMBEDDING OF ENVIRONMENTAL VARIABLES

**EA MAP581**

13 mars 2018

---

Flore Martin and Lorraine Roulier



## Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Principal Component Analysis - PCA</b>	<b>4</b>
2.1	Method . . . . .	4
2.2	Results . . . . .	4
2.2.1	Two dimensions . . . . .	4
<b>3</b>	<b>Kernel Principal Component Analysis</b>	<b>6</b>
3.1	Method . . . . .	6
3.2	Results . . . . .	6
<b>4</b>	<b>Multidimensional Scaling - MDS</b>	<b>6</b>
4.1	Method . . . . .	6
4.2	Results . . . . .	7
<b>5</b>	<b>Isomap</b>	<b>8</b>
5.1	Method . . . . .	8
5.2	Results . . . . .	9
<b>6</b>	<b>Comparing the different methods</b>	<b>10</b>
6.1	Time comparison . . . . .	10
<b>7</b>	<b>Conclusion</b>	<b>11</b>
<b>8</b>	<b>Bibliography</b>	<b>11</b>

# 1 Introduction

Climate data amounts very quickly to a lot of unused data. In a day, we can collect temperature, pressure, wind data all over the world with satellites, even hourly. Our project was two sided. First, we familiarized with various dimension reduction techniques, then we attempted to show that the geographical position of a point on the planet - e.g. it's latitude and longitude - were embedded in the climate data one could gather on it.

Dimension reduction techniques can be divided in two classes, linear dimension reduction and non linear dimension reduction. However, in all methods, the main goal is to figure out a similarity function between vectors. Such a function will then enable to sort the dataset into classes of vectors with similar features, which would have been more intricate with the initial dataset. We used a set of datasets we found on the NASA website, that gathered various means on climate variables over 22 years at every given latitude and longitude. These variables are gathered in the table below

Latitude	Longitude	Temperature $^{\circ}C$	Pressure $kPa$	Relative Humidity %	Wind Speed $m/s$	Radiation $kWh/m^2/day$
----------	-----------	----------------------------	-------------------	------------------------	---------------------	----------------------------

FIGURE 1 – First row of our dataset

The latitude parameter varies from -90 to 89 and the longitude parameter varies from -180 to 179. the negative values are for the south hemisphere, the positive ones for the north. Depending on the running time of the method, we did not compute the dimension reduction with the 64800 lines, but with a subset. The subset is often a slice of longitudes containing all latitudes, as we assumed that the critical parameter to differentiate climate data was the latitude.

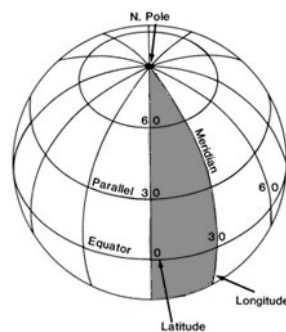


FIGURE 2 – An example of subset in grey

We classified the data according to the latitude, creating five classes listed in the table below :

	North	Temperate north	Equator	Temperate South	South
Latitudes	90 to 66	66 to 23	23 to -23	-23 to -66	-66 to -90

FIGURE 3 – First lign of our dataset

## 2 Principal Component Analysis - PCA

### 2.1 Method

Principal Component Analysis detects tendencies in the data by maximizing the variance of the dataset matrix. This yields an orthonormal matrix that can be diagonalized. The largest eigenvalues point to the eigenvectors that contain the most information about the dataset.

Let  $X \in \mathbb{R}^{d \times n}$  be our dataset, PCA maximizes the following equation :

$$\|X - MM^T X\|^2$$

subject to  $M \in \mathcal{O}^{d \times r}$  where  $r < d$ .

### 2.2 Results

#### 2.2.1 Two dimensions

We first implemented PCA and ran it with only two principal components, which yielded the following graph for the whole dataset. The associated eigenvectors were

$$y_1 = \begin{bmatrix} -0.94220902 \\ -0.31329122 \\ 0.10889051 \\ -0.04573547 \\ 0.01191187 \end{bmatrix}$$

and

$$y_2 = \begin{bmatrix} 0.02637512 \\ -0.40428947 \\ -0.9117768 \\ 0.04144017 \\ -0.05291649 \end{bmatrix}$$

This enables us to understand the meaning of these vectors.  $y_1$  is mostly related to a decreasing temperature and pressure, and  $y_2$  represents decreasing humidity and pressure.

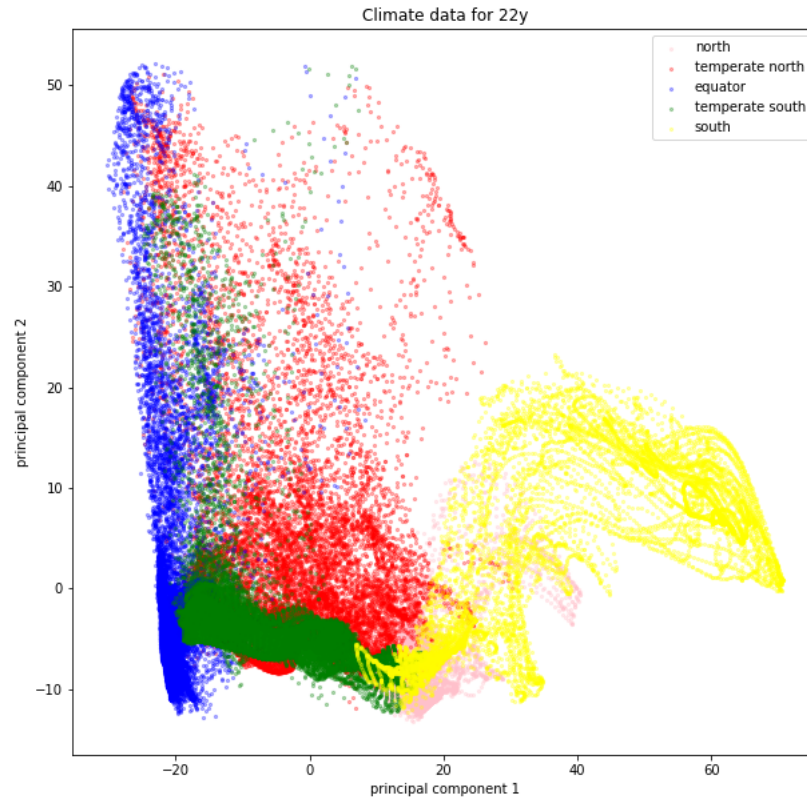


FIGURE 4 – PCA with two components

We can see on the graph that even if there is a strong dispersion, equator values are located at higher temperatures. On the contrary, north and south pole values are located at lower temperatures. Green and red classes overlap as these two categories have similar climate conditions.

Although it is an understandable figure, this is not satisfying. We plot the eigenvalues to see their relative importance in the dimension reduction.

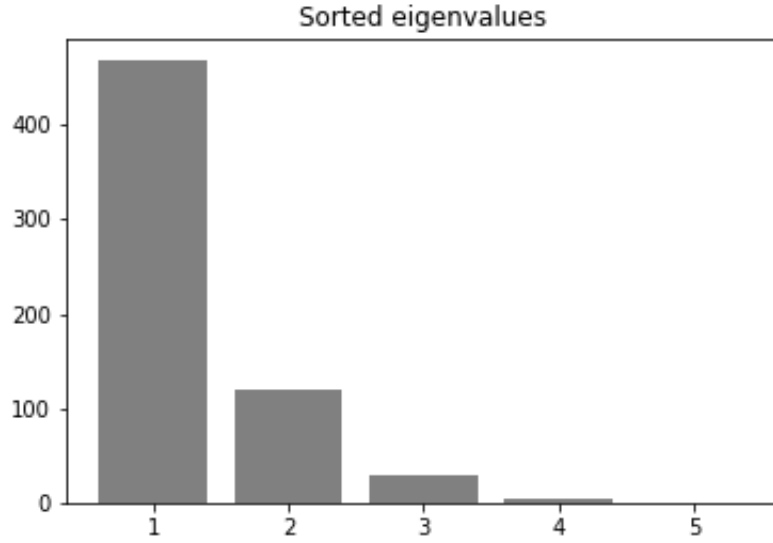


FIGURE 5 – PCA with two components

### 3 Kernel Principal Component Analysis

#### 3.1 Method

#### 3.2 Results

### 4 Multidimensional Scaling - MDS

#### 4.1 Method

Multidimensional scaling is a non-linear approach to reduce dimension of a dataset. The principle is quite different from PCA : given a matrix of distance or "dissimilarity", MDS aims at reconstructing a map preserving at best the distances between datas. In our example, the MDS algorithm aims to place each object in 2-dimensional space such that the between-object distances are preserved as well as possible. Given a distance matrix  $D$  of dimension  $n \times p$  ( $p=5$  for us) , MDS attempts to find the datapoints  $y_1, \dots, y_t$  in dimension  $d < p$  ( $d=2$  for us) that minimizes :

$$\sum_{i=1}^t \left[ \sum_{j=1}^t d_{ij}^{(X)} - d_{ij}^{(Y)} \right]$$

with  $d_{ij}^{(X)}$  and  $d_{ij}^{(Y)}$  respectively the euclidean distance between pairwise  $i$  and  $j$  in the original  $n \times p$  matrix  $D^{(X)}$  and in the computed  $n \times d$  matrix  $D^{(Y)}$ .

As the distance matrix  $D^{(X)}$  can be converted into a kernel matrix of inner products  $X^T X$  by

$$X^T X = -1/2 H D^{(X)} H$$

$H$  can be written as  $H = I - ee^T$  and  $e$  is the vector of ones. Thus, ??? becomes :

$$\sum_{i=1}^t \sum_{j=1}^t x_i^T x_j - y_i^T y_j$$

It can be shown that the solution is  $Y = \Lambda^{1/2} V^T$  with  $V$  the eigenvectors of  $X^T X$  of the top  $d$  eigenvalues present in  $\Lambda$ . The only and main parameter of this method is the norm used to calculate the distance matrix. By default we use the euclidean norm here.

## 4.2 Results

Figure 6 shows the results of MDS with our previous dataset. We use the same grandeurs (temperature, radiation, wind speed, humidity and pressure) but not the same number of lines. As MDS was extremely low and seemed to use a lot of memory, we could not run the program for more than 600 rows. Here are the result for 546 rows : we filtered our dataset with latitudes  $< 0$  (ie. south hemisphere) and longitude between 0 and 5 degrees.

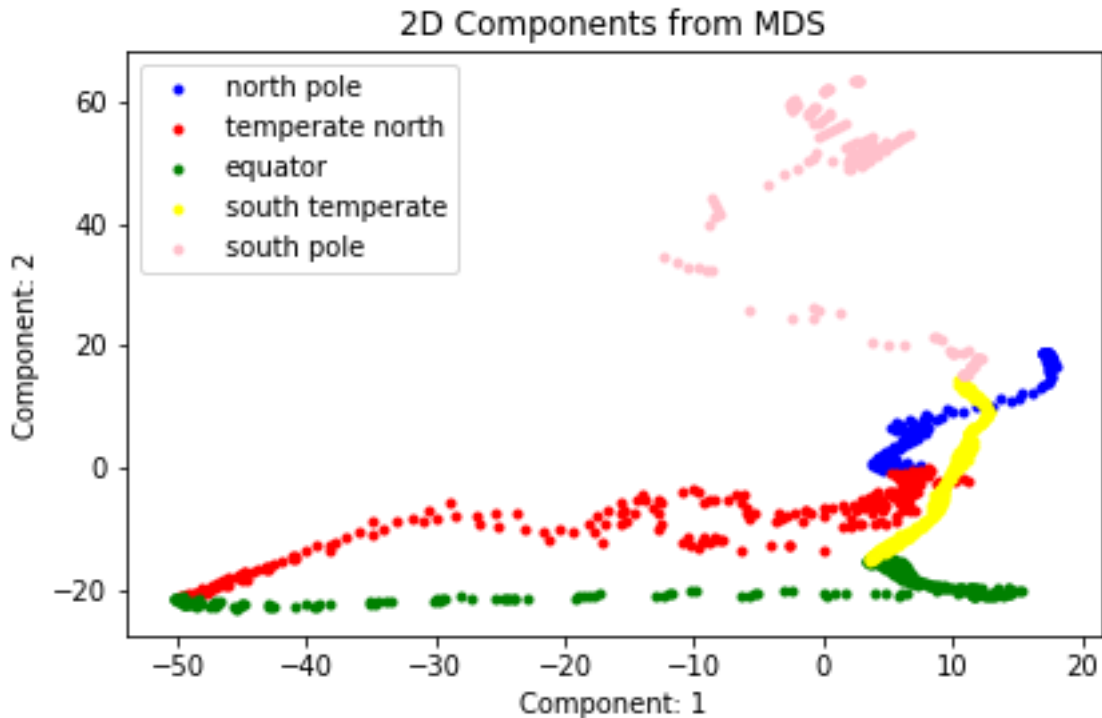


FIGURE 6 – results of MDS in 2D

One indicator to estimate of the reconstruction is accurate is the 'stress'. It is roughly the square difference between the final distance of two items in the MDS model and the true distance between them. The exact formula is :

$$s = \sqrt{\frac{\sum (d_{ij} - d(i, j))^2}{\sum d_{ij}^2}}$$

The aim of MDS is to minimize this stress. Figure 7 presents the value of stress depending on the number of lines. We can see it is relatively low ( $<0,002$ ) for more than 500 rows.

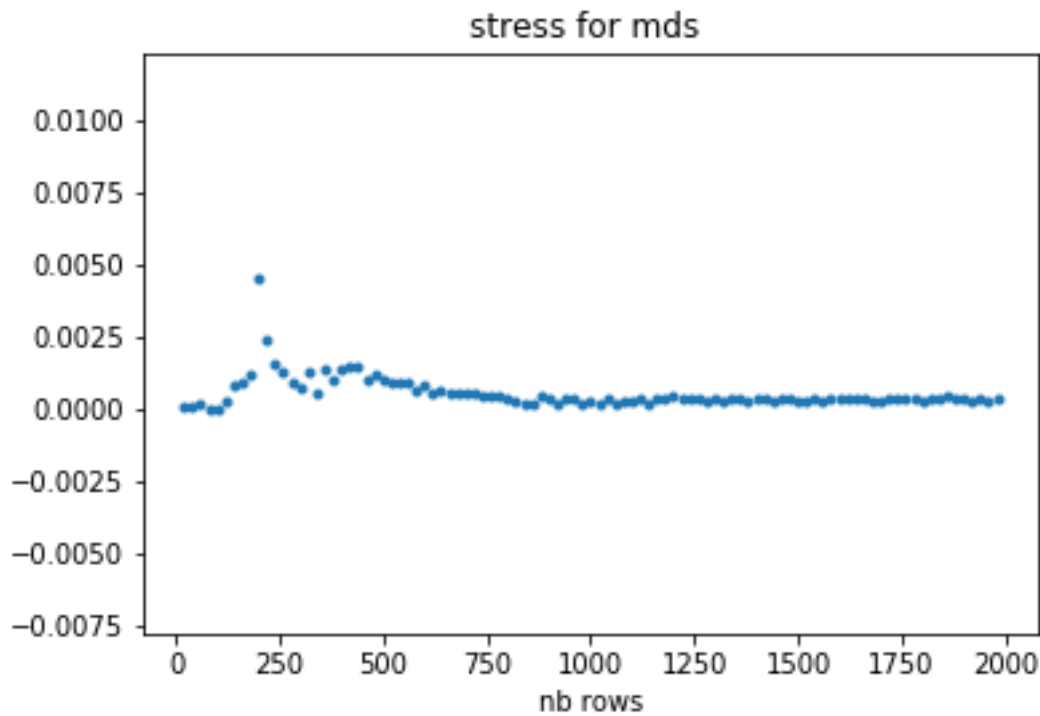


FIGURE 7 – stress as a function of number of rows

## 5 Isomap

### 5.1 Method

Isomap is a low dimensional embedding method similar to MDS. The difference is, distance is not computed with euclidean norm but with geodesic distance.



## 5.2 Results

As the main parameter of the method is the number of nearest neighbors taken into account, we present in figure 8 the result of isomap for number of neighbour from 1 to 11 neighbors.

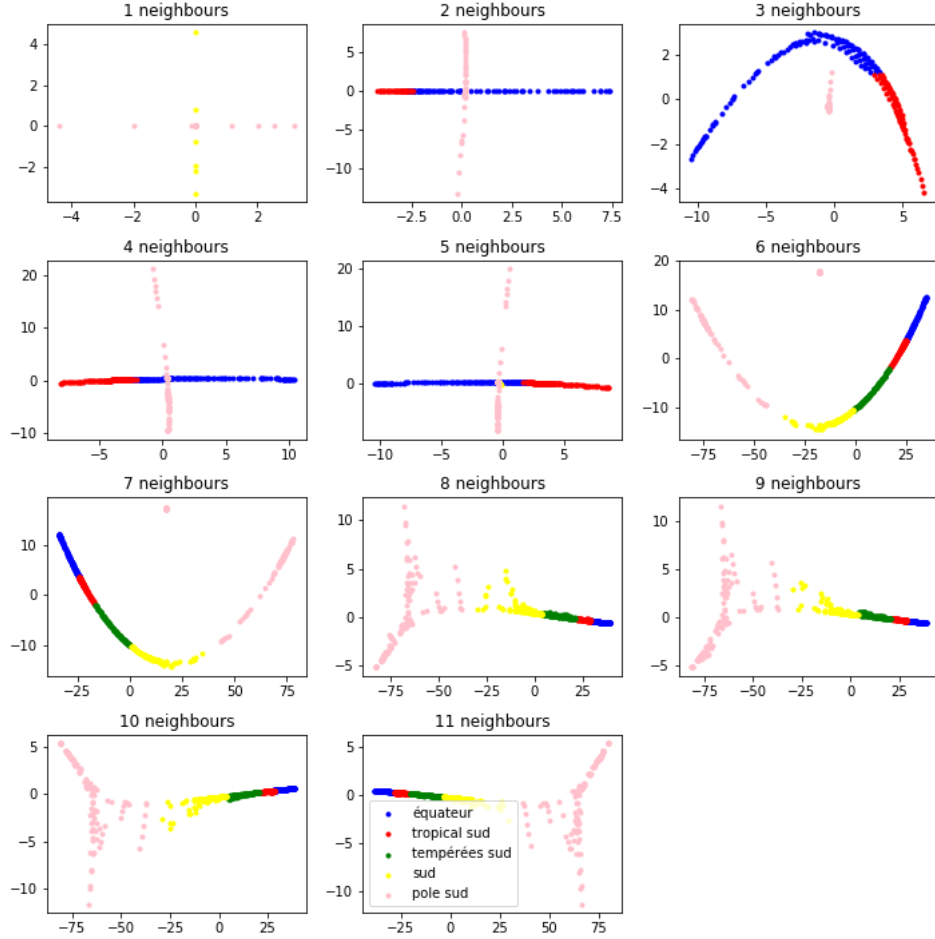


FIGURE 8 – results of isomap depending on the number of neighbors considered from 2 to 12

It is obvious that the number of neighbours considered strongly influence the 2D distribution of our dataset. To know which graph is the most accurate, we refer to the 'reconstruction error'. We computed this reconstruction error with respect to the number of neighbours considered. The results are presented in figure ??

We also computed the error reconstruction as a function of the number of rows in figure 9. This error keeps increasing with the number of rows, so isomap accuracy seems to be limited by the size of the dataset.

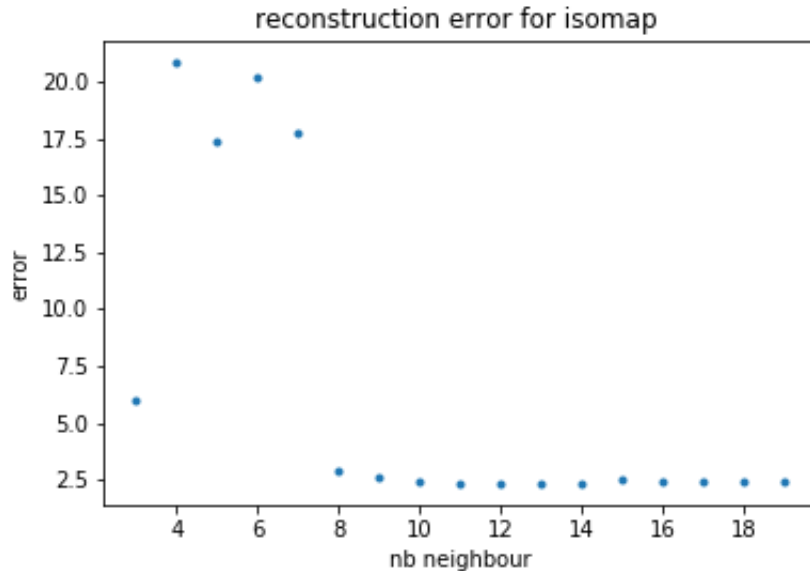


FIGURE 9 – reconstruction error as a function of number of neighbours considered

It is interesting to see that from eight neighbours, error seems to stabilize around 2,5. Indeed, we can see with graphs of figure 8 that from eight neighbours, graphs look quite all the same. Those are the most 'accurate' representations.

## 6 Comparing the different methods

### 6.1 Time comparison

In figure 10, we present computing time for each method as a function of the number of rows. From this we can conclude that the quickest method is PCA. Then comes kernel PCA, isomap and MDS. It is quite surprising to see that isomap is quicker than MDS for the same number of rows. Indeed, isomap uses the same algorithm than MDS but also needs to compute distance with neighbors first. It is thus expected to be slower. WE found an explanation on the "‘manifold guide’" of skikit.

In Scikit-Learn implementation, Isomap algorithm runs faster than Multi Dimensional Scaling on the S-Curve dataset [...]. In the third stage of algorithm, the implementation uses Partial Eigen Value decomposition instead of MDS which is the version proposed by the researchers

We did not use the "S curve dataset" as mentioned above, but we may think that due to the round shape of earth, our dataset is quite similar to the S-curve and thus Isomap runs faster than MDS.

Method	PCA	Kernel PCA	MDS	Isomap
Complexity	$O(n^2)$	$O(?)$	$O(n^3)$	

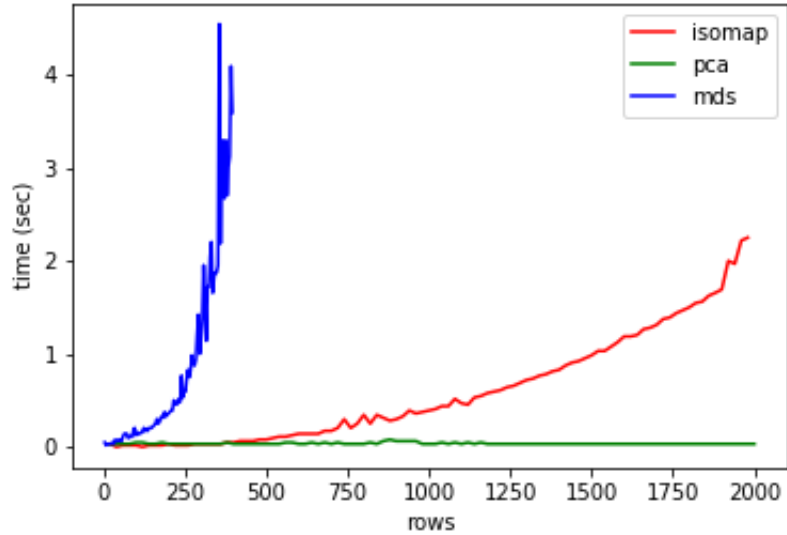


FIGURE 10 – computing time comparison

We also check how isomap computing time varies with the number of neighbor taken into account. Figure 11 shows that isomap efficiency time increases with the number of neighbors. So isomap efficiency of computing time may be limited by the number of neighbors.

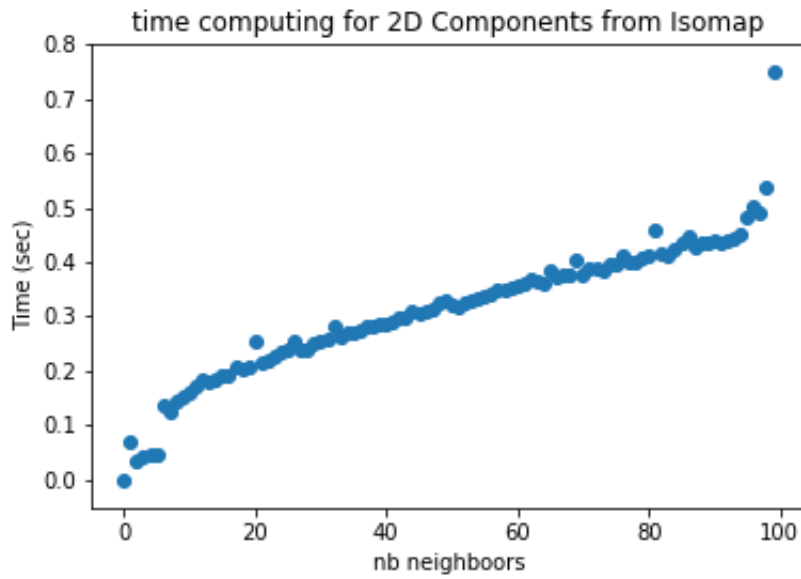


FIGURE 11 – computing time for isomap as a function of the number of neighbors considered

## 6.2 error and preserved information

## 7 Conclusion

## 8 Bibliography