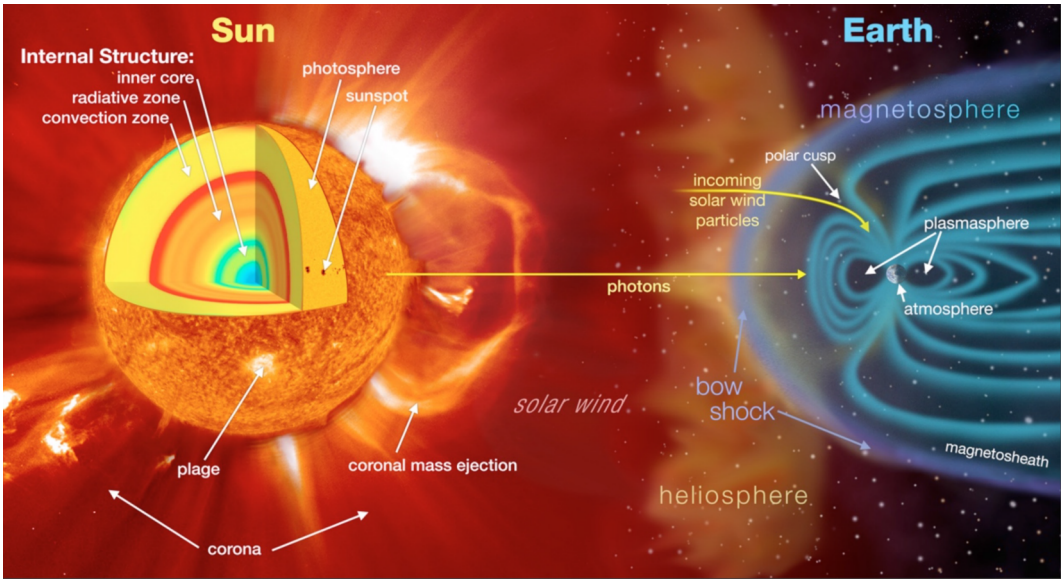
INSA Toulouse

Project

Machine Learning for Space Weather



Students :

LE Vu Nam Anh

VO Thanh Tin

Supervisor :

Antoine Brunet

18^{th} May, 2020

Contents

Introduction

Space weather is a field of space physics which studies the change of solar wind and the space surrounding the Earth including the magnetosphere, the ionosphere, the thermosphere, and the exosphere. Space weather is different from atmospheric weather, but they are closely related. Changes in solar wind, namely the distribution of electron fluxes according to magnetic activity can cause changes in atmospheric weather. By studying and analyzing space weather, we can deduce its effects on atmospheric weather and many other fields. We employ THEMIS/SST space probes to collect the electron data of solar wind. Electron data provided by SST instrument are built up through differential omnidirectional and unidirectional fluxes. In space weather, a kind of "distance" denoted **L*** represents the boundary in the space where the Earth is "the center". Scientists usually consider the boundary condition at $\mathbf{L^* > 7}$ where the difference in the distribution of electron fluxes between day side and night side is apparent. In order to construct a boundary condition at a certain value of **L***, it is essential to identify if the electron flux is isotropic or not at this **L*** value. The answer to this question will determine if unidirectional or omnidirectional fluxes resulting from THEMIS/SST measurements will be used.

Goal:

The objective of this project is the development and validation of a model of the energy spectrum of electron fluxes in radiation belts around the Earth. A neural network will be trained to model the response of the electron belt to the dynamics of the solar wind. We will compare the model obtained with the existing statistical models, as well as the network performance according to the input parameters used.

Tools and data:

The developments will be made in Python. The Scikit-Learn library will be used for learning the network. The data used for this project will be the *OMNI2* database and the measurements of the ESA instrument on *THEMIS* satellites.

1 Some basic definitions

1.1 What is solar wind and its effects? [1]

Definition

The solar wind is a stream of charged particles released from the Sun's Corona, or known as the upper atmosphere of the Sun. This plasma includes mostly electrons, protons and alpha particles.

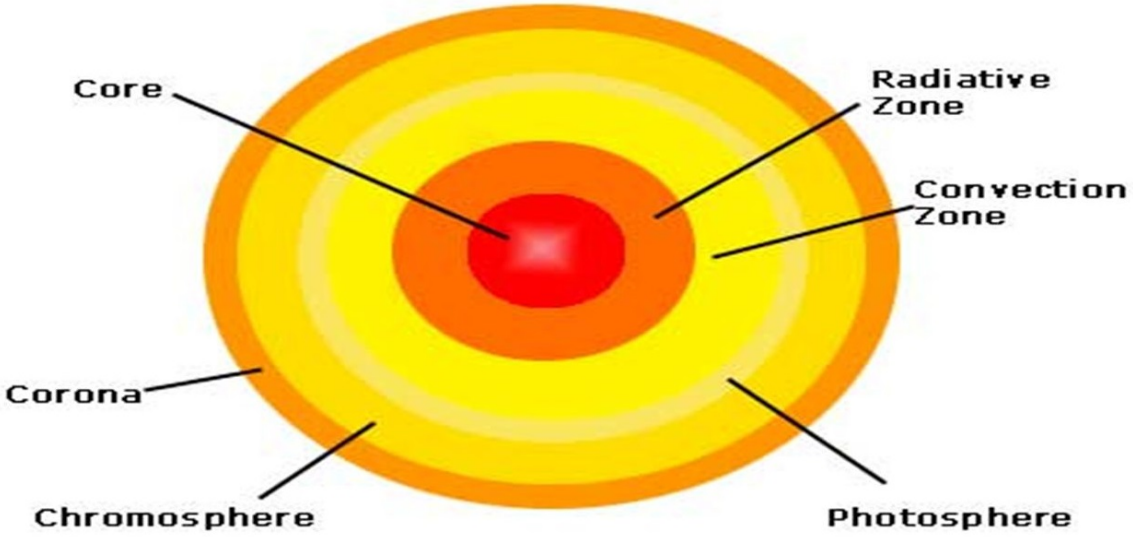


Figure 1: Anatomy of the sun

Some effects of solar wind

Over the Sun's lifetime, its surface rotation rate has importantly dropped because of the interaction of its surface layers with the escaping solar wind. It is believed that the wind is the reason for comets' tails, along with the Sun's radiation. The fluctuations in celestial radio waves observed on the Earth is also affected by the solar wind, through an effect called interplanetary scintillation.

(1) Magnetospheres

When a planet which has a well-developed magnetic field, like Earth, intersects with the solar wind, the Lorentz force deflects the particles. This region of intersection, called the magnetosphere, makes the particles to travel around the planet rather than bombarding the atmosphere or surface. The shape of the magnetosphere is like a hemisphere on the day side - facing the Sun, and out in a long wake on the opposite side.

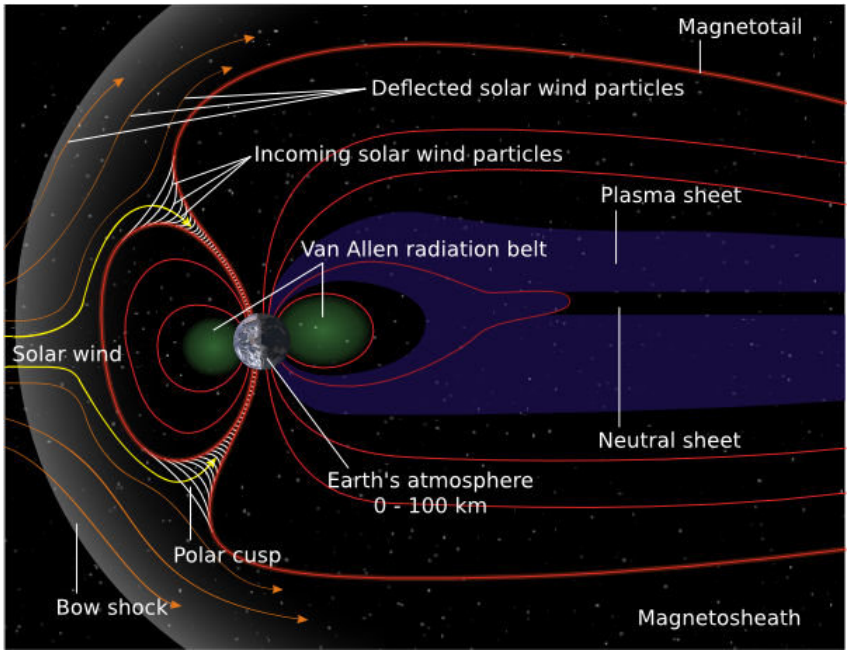


Figure 2: Schematic of Earth's magnetosphere

The reason for the overall shape of Earth's magnetosphere is the solar wind. Earth's local space environment is significantly affected by the fluctuations in solar wind velocity, density, direction, and entrained magnetic field.

For example, the levels of ionizing radiation and radio interference can vary by factors of hundreds to thousands; and the shape and location of the magnetopause and bow shock wave upstream of it can change by several Earth radii, exposing geosynchronous satellites to the direct solar wind. These phenomena are collectively called space weather.

(2) Atmospheres

The solar wind affects other incoming cosmic rays that interact with planetary atmospheres. Furthermore, the solar wind can strip the atmosphere of planets which have a weak or non-existent magnetosphere.

*The most similar and nearest planet to Earth is **Venus**, but it has 100 times denser atmosphere, with little or no geo-magnetic field. Space probes discovered a comet-like tail that extends to Earth's orbit.*

*The **Earth**'s magnetic field largely protects the planet from the solar wind, it deflects most of the charged particles; nevertheless some of the charged particles are trapped in the Van Allen radiation belt. A smaller amount of particles from the solar wind travel to the Earth's upper atmosphere and ionosphere in the auroral zones as though on an electromagnetic energy*

transmission line. We can only observe, on the Earth, the solar wind when it is strong enough to produce phenomena like the aurora and geomagnetic storms. The ionosphere is strongly heated by the bright auroras, then its plasma expands into the magnetosphere, the size of the plasma geosphere increases and atmospheric matter is injected into the solar wind.

*Even though **Mars** is larger than Mercury and four times farther from the Sun, it is believed that a third of its original atmosphere has been stripped away by the solar wind, the rest is a layer 1/100 th as dense as the Earth's. The mechanism for this atmospheric stripping might be gas caught in bubbles of magnetic field, which are ripped off by solar winds.*

(3) Moons and planetary surfaces

*The nearest planet to the Sun is **Mercury**, it bears the full brunt of the solar wind, and as its atmosphere is vestigial and transient, its surface is bathed in radiation. However, Mercury has an intrinsic magnetic field, so the normal solar wind cannot penetrate its magnetosphere and particles only reach the surface in the cusp regions.*

*The Earth's **Moon** has no atmosphere or intrinsic magnetic field, and therefore its surface is bombarded with the full solar wind.*

1.2 Earth's magnetic field [2]

Definition

The Earth's magnetic field, which is also called magnetic filed, is zone extending from the Earth interior to the space where it is affected by solar wind. The magnetic field is defined by electric currents that is the motion of the convection currents between molten iron and nickel in the Earth's outer core. These convections is explained as the affection a heat escaping from the core causes (based on the Dynamo theory). The magnitude of the Earth's magnetic field at its surface are the numbers from 25 to 65 microteslas (the tesla is a derived unit of the magnetic induction and magnetic flux density in the International System of Units). However, in some parts of this lecture, the K-index will be used instead of Tesla because it is the unit used by NASA who give the data about magnetic field and solar wind.

Significance

The Earth's magnetic is important to protect the Earth from ultraviolet radiation from the solar wind because its charged particles can eliminate and pass by Ozone layer. That is a mechanism to catch gas in bubbles of magnetic field that are broken by solar wind.

Intensity

The intensity of the field is measured in unit Gauss (G), but normally is reported in nanotesla (nT). For example, the intensity of a strong refrigerator magnet is about 10 000 000 nanotesla (100 G) in comparison to the Earth's magnetic field, it is in the interval [25000,65000].

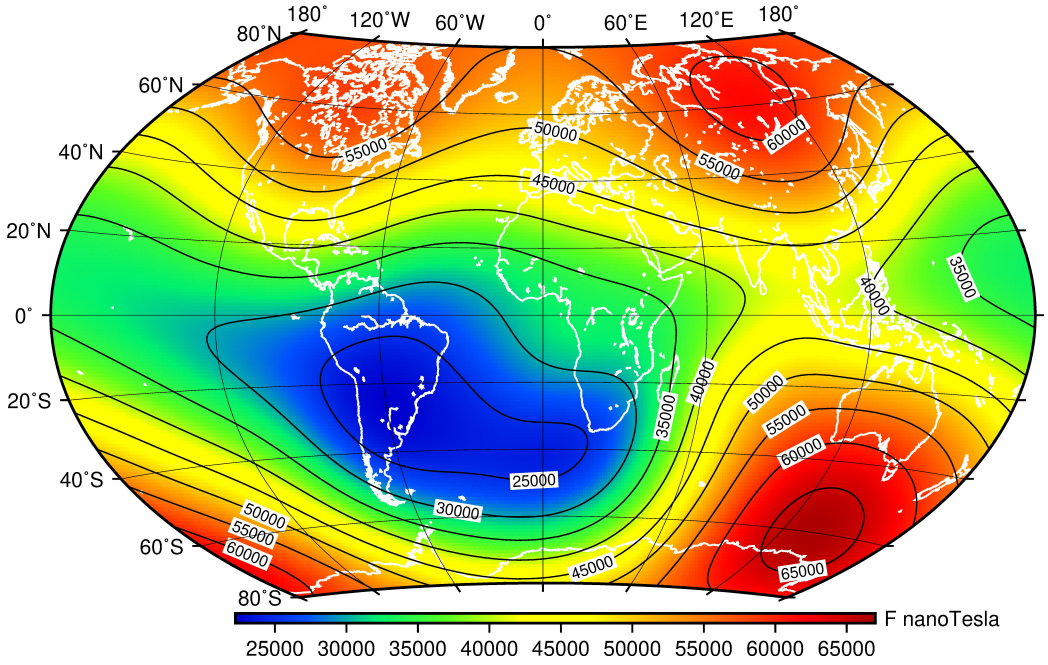


Figure 3: Intensity of the magnetic field of the Earth

In some World Magnetic Model shows, the intensity tends to decrease from the poles to the equator. The place of minimum intensity measure is South Atlantic Anomaly in the South Amercica meanwhile the maximum ones occur in the northern Canada, Siberia, and the coast of Antarctica south of Australia.

Dipolar approximation

The magnetic field of the area near the surface of the Earth can be closely defined as the field created by the magnetic dipole located at the center of the Earth and at an angle of about 11^o to the Earth's rotation. Furthermore,

the dipole is simply determined by a strong magnet whose north pole face the magnetic south pole.

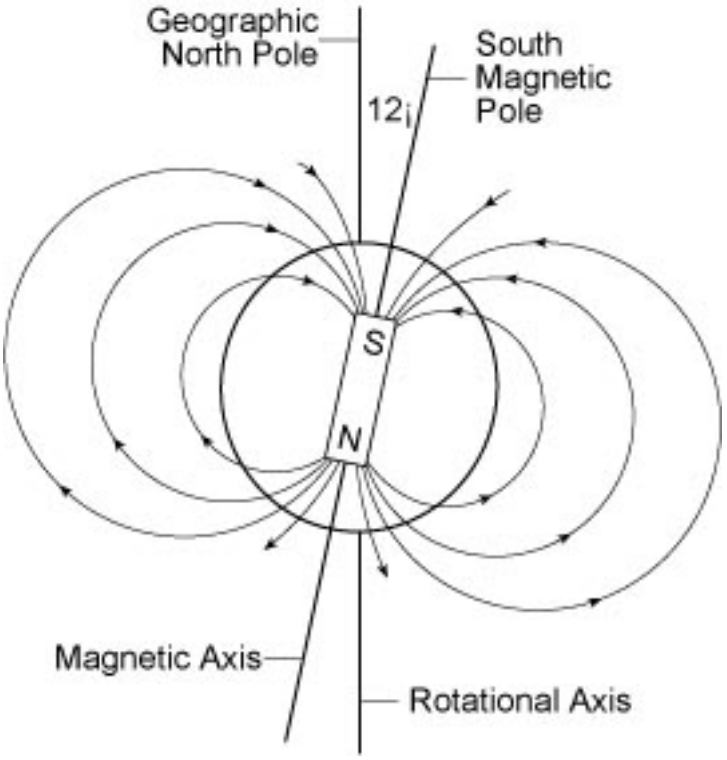


Figure 4: Magnetic dipole

L-Shell

The L-shell, L-value, or McIlwain L-parameter (after Carl E. McIlwain) is a parameter defining a particular set of planetary magnetic field lines. In the term of the Earth's magnetic field, L is usually used to describe the set of magnetic lines crossing the magnetic equator at a number of earth radius equal to the L-value. For L=3, it gives the set of the Earth's magnetic field lines crossing the Earth's magnetic equator far from the center of the earth three earth radius (about 19 113 kilometers).

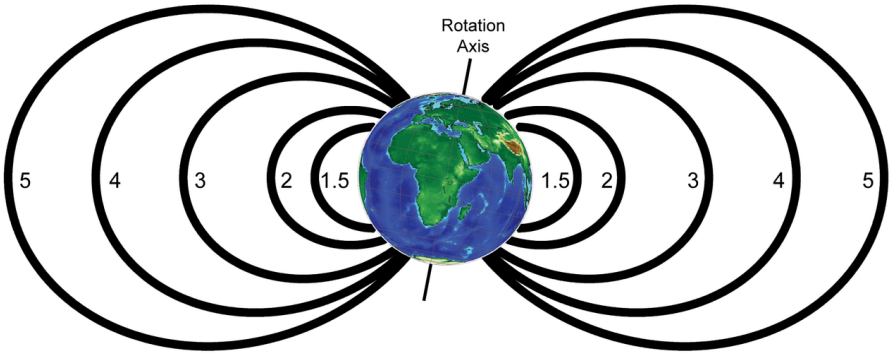


Figure 5: L-Shell

Magnetosphere

The Earth's magnetic field is mostly bipolar at its surface, however, it is impacted and distorted by the solar wind. The solar wind creates tremendous pressure, and if it can touch the earth's atmosphere, it will erode the atmosphere. However, it was prevented and deflected by the earth's magnetic field. The magnetosphere is enclosed by an area of pressure equilibrium called Magnetopause. The magnetosphere does not have a symmetrical shape, because it extends 10 times the radius of the earth on the face of the sun, but it lasts nearly 200 times on the other side.

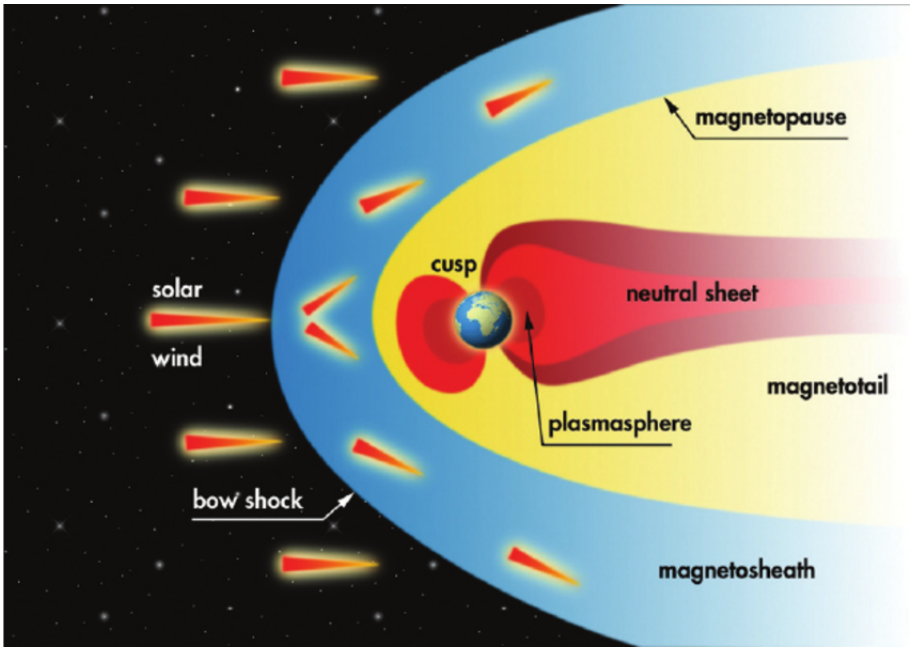


Figure 6: Operation of the magnetosphere around the Earth

Short-term Variation

The geomagnetic field is not fixed; it changes every millisecond, even smaller. This is caused by changes in the flow or due to the occurrence of a geomagnetic storm. This change according to the time scale of a year also reflects the change from within the earth, especially the iron-rich core. this short-term instability can be measured by the K-index, usually the time it takes to regain that index is about 3 hours.

1.3 Significant elements in measurement and calculation

Roederer's parameter (L*) [3]

Physically, **L*** is the radial distance to the equatorial points of symmetric L-shell on which the particles would be found if all nondipolar perturbations of the magnetic field were turned off adiabatically.

In this lecture, we will study the data at **L*=8**.

Interplanetary magnetic field [4],[5]

One of the important parts of the Sun's magnetic field is the *Interplanetary Magnetic Field* (IMF). It follows the solar wind into interplanetary space. The IMF is a vector quantity which has three directional components, two of which (**Bx** and **By**) are oriented parallel to the ecliptic. The third component, **Bz**, is created by waves and other disturbances in the solar wind and it is perpendicular to the ecliptic.

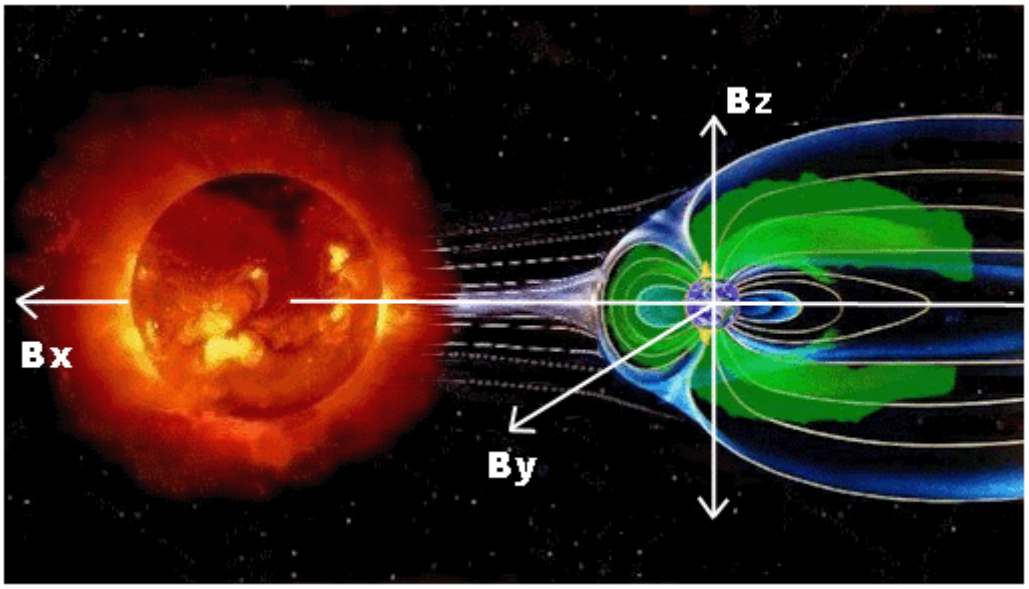


Figure 7: Interplanetary magnetic field

For measurements of solar wind, **Bz** is the most important parameter. When **Bz** goes negative, the solar wind couples strongly to the Earth's magnetosphere. We can think of **Bz** as the door that allows transferring of significant amounts of energy. The more negative **Bz** goes, the more energy that can be transferred, resulting in more geomagnetic activity.

Solar wind velocity [6]

In all directions, the solar wind streams off of the Sun at speeds of about 400 km/s. As we know, solar wind is from the Sun's hot corona, but until now, the details about how and where the coronal gases are accelerated to these high velocities have not been clarified yet. The velocity is high (800 km/s) over coronal holes and low (300 km/s) over streamers. These high and low velocity streams interact with each other and alternately pass by the Earth as the Sun rotates. These velocities variations buffet the Earth's magnetic field and can produce storms in the Earth's magnetosphere.

Density

The solar wind density depends on the position. For example, near the Earth, the density is about 3 to 6 atoms per cubic centimeter; or at the orbit of the Earth, the density is about 6 atoms per cubic centimeter. The effects of different densities are not the same such as at different positions, at day side and night side, etc. We will use frequently the solar wind density in measurements and calculation for prediction.

Planetary K-index or Kp-index (Kp)

The Planetary K-index is used to describe the extent of geomagnetic storms. It is an extremely good indicator of disturbances in the Earth's magnetic field.

In this lecture, we will study this index and try to predict it according to **Bz**, solar wind velocity and density.

2 Modelization

2.1 General Method: Cross-Validation [8]

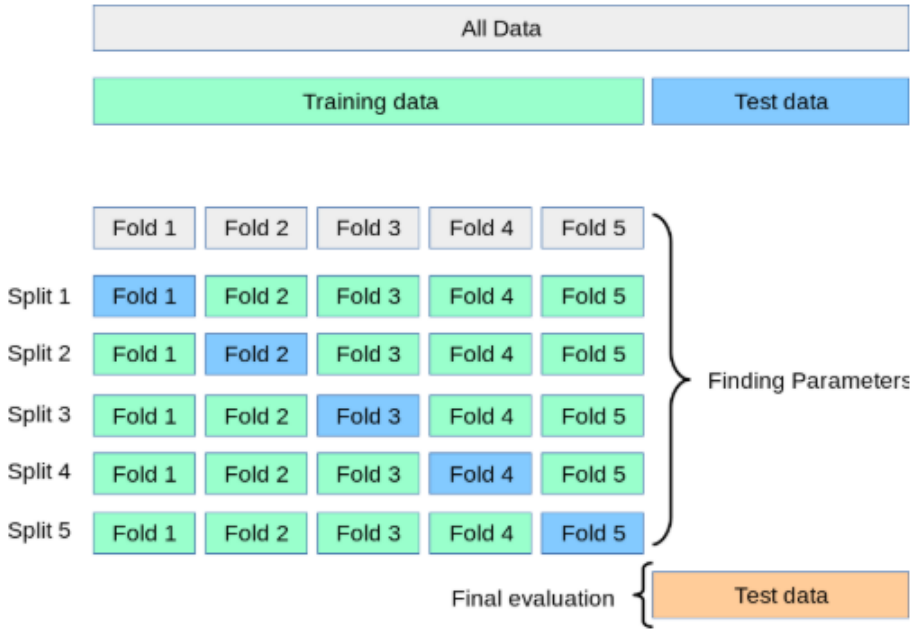


Figure 8: K-fold cross validation

Cross validation is a re-sampling procedure used to review machine learning models on a fixed data sample.

The process of dividing data into k data groups and then cross-validation of each group. Cross validation is often applied to estimate the skills of machine learning models on data that we cannot recognize trends and how numbers work. It means analyzing and predicting on a sample a limited sample so as to estimate how the model is expected to perform generally when used, and then make predictions about the data not appearing in the model.

Steps to perform:

- Divide the data into k groups
- For each unique group:
  - Take a group as data to compare with the predicted data
  - Take the remaining groups as a database to produce predicted data
  - Apply the model on a database and compare it with test data
  - Store important evaluation indicators

- Make conclusions about the skills of the model through the evaluation indicators obtained

2.2 Supporting Knowledge for analyzing

Root-mean-square deviation [7]

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a measure to define the difference of real data that measurement tools provide and those that are predicted by a mathematical model over a given time period. RMSE helps us identify the magnitudes of errors at different times which is a measure of the reliability of the model being applied. RMSE is used to know the accuracy when comparing predictive errors of different models for a part of a data set rather than the entire because it depends on the time period we consider:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_{predicted} - y_{observed})^2}{T}}$$

RMSE is a non-negative value and in fact almost impossible to be 0. Usually, a small RMSE value is better than a large RMSE, but comparing this value in different data types is meaningless. because the metric is affected by the size of the data used. In this study, RMSE is expected to receive values less than 1.5 Kp for a very large data range.

Homogeneity

Homogeneity emerges in depicting properties of a dataset or a few informational collections. They identify with the legitimacy of the suspicion that the factual properties of any piece of the overall data set are equivalent to some other. In the regression, there should then be a later phase of investigation to analyze whether the errors in the expectations from the regression carry on similarly over the data set. In this manner the inquiry gets one of the homogeneity of the distribution of the residuals.
Particularly, in this lecture, the homogeneity of the RMSE all over time will be considered.

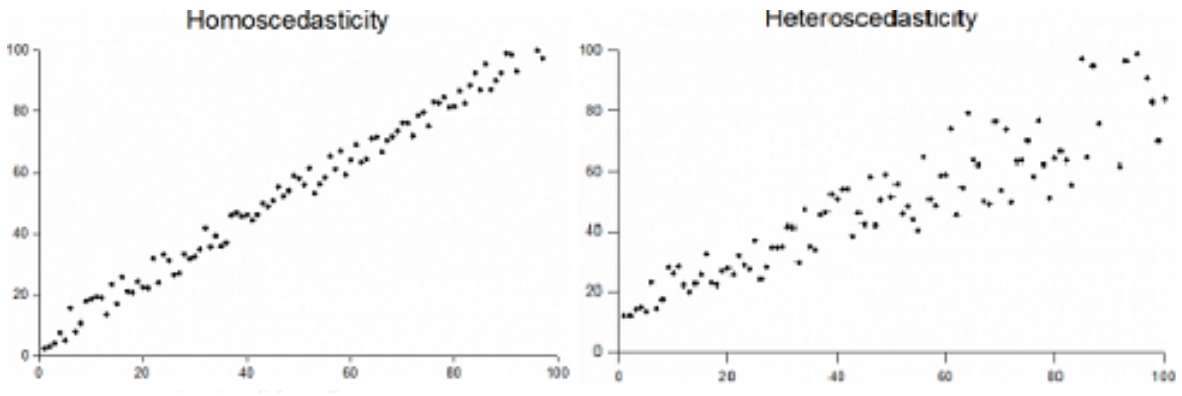


Figure 9: The difference between homogeneity and heterogeneity

3 Particular application of algorithm methods and analysis

3.1 K-nearest neighbor

The **K**-nearest neighbor algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. But first, let's take a look at *classification* and *regression*.

Classification : Overall

Classification is the way toward finding or discovering a model or function which helps in separating the data into numerous categorical classes i.e. discrete values. In classification, data is ordered under various labels according to some parameters given in input and then the labels are predicted for the data.

The derived mapping function could be shown as "IF-THEN" rules. The classification process manage the issues where the data can be separated into binary or multiple discrete labels.

For example, we want to predict the possibility of the winning of World Cup 2018 by France national football team on the basis of some parameters recorded earlier. Then there would be two labels **Yes** and **No**.

However, in this lecture, we will not focus on classification, we will only concentrate on the regression analysis, since the target values are continuous values, not discrete values.

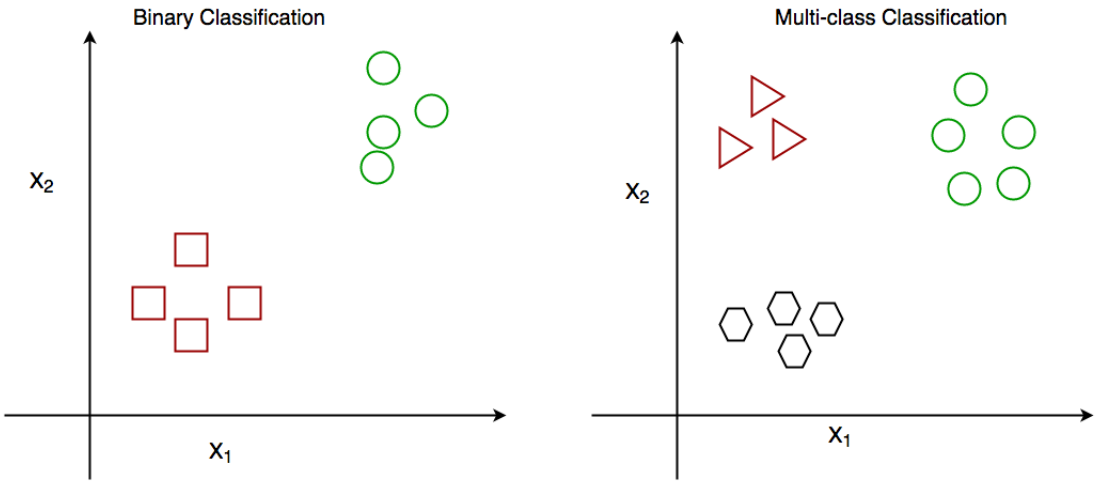


Figure 10: Examples of classification

Regression : Overall

Regression is the way toward finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can likewise distinguish the distribution movement relying upon the historical data. Since a regression predictive model predicts a quantity, thus, the ability of the model must be reported as an error in those predictions.

For example, we are researching the possibility of storm in some tropical regions with the help of some parameters recorded earlier. Then there is a probability associated with the storm.

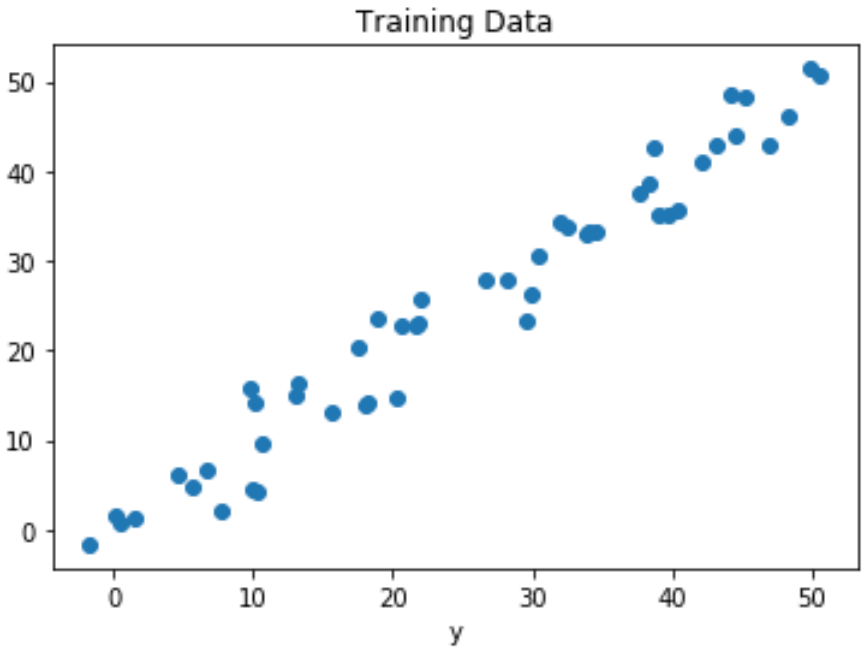


Figure 11: Example of regression

Introduction [9]

Neighbors-based regression can be utilized in situations where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors.

There are two kinds of Nearest Neighbor Regression: One is $\boldsymbol{K}$-*nearest neighbors regression* which actualizes learning based on $\boldsymbol{K}$ nearest neighbors of each query point, where $\boldsymbol{K}$ is an integer value indicated by the user; and other is *Radius nearest neighbors* which executes learning based on the neighbors inside a fixed radius $\boldsymbol{r}$ of the query point, where $\boldsymbol{r}$ is a floating-point value indicated by the user.

In this lecture, we will concentrate on the $\boldsymbol{K}$-*nearest neighbors regression.*

Method

1. Calculate the distance between the new point and each training point
2. Select the closest $\boldsymbol{K}$ data points based on the distance.
3. The final prediction for the new point is the average of these data points.

However, the problem is how the distance between points can be calculated and how the number of nearest neighbors $\boldsymbol{K}$ should be chosen.

In the first step, the distance between the new point and each training point can be calculated by various methods, but the most frequently known method for continuous variables/values is Euclidean distance:

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In the second step, in general, the conclusive outcome will change based on the $\boldsymbol{K}$ value. It depends on the error calculation of train and validation set and the goal is to minimize the error. If $\boldsymbol{K}$ is too small, $\boldsymbol{K} = 1$ for example, it will lead to a high error rate on the validation set. On the other hand, if $\boldsymbol{K}$ is too large, both train and validation set perform a high error. Therefore, it is necessary to choose a good value of $\boldsymbol{K}$.

Application to the study

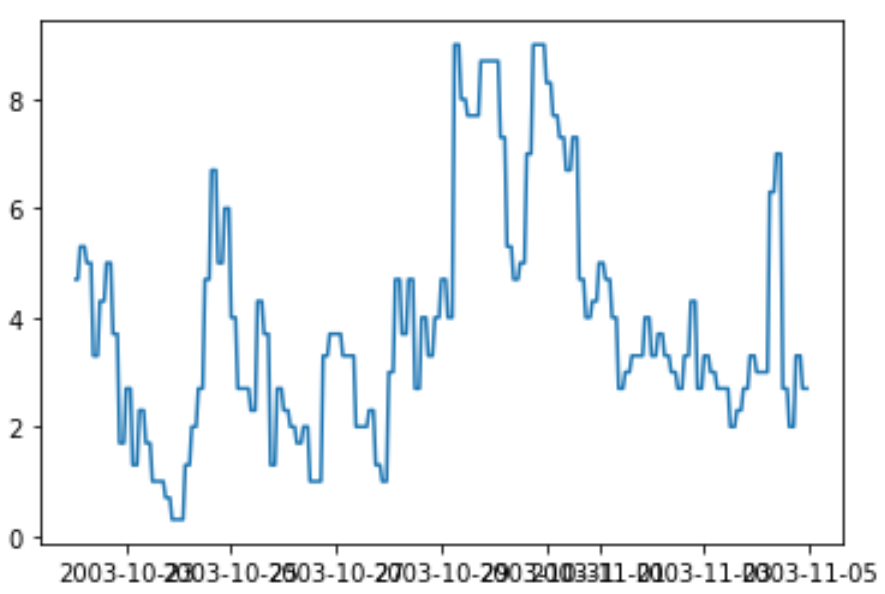


Figure 12: Test output from 22/10/2003 to 05/11/2003

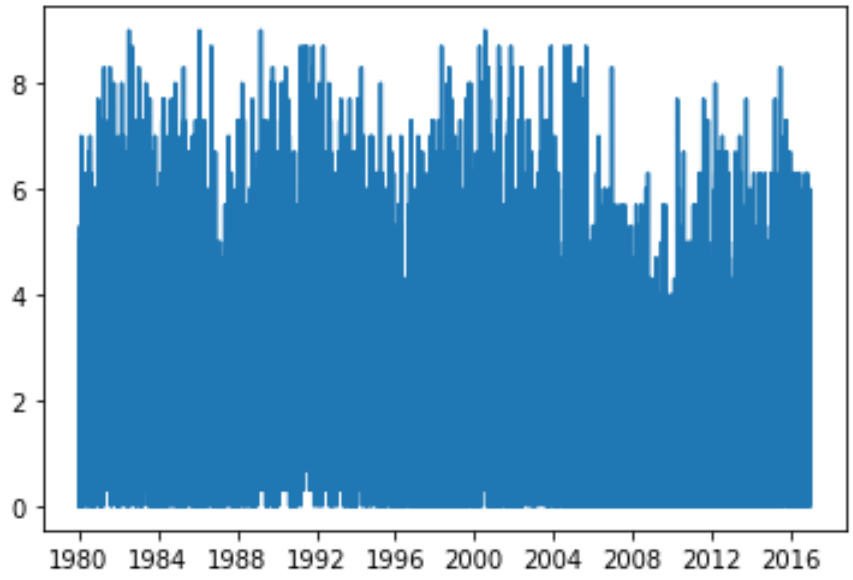


Figure 13: Data from 01/01/1980 to 01/01/2007

First, we take a large enough amount of data to be able to predict Kp values in a short span of time. It means that the size of the test data (usually just a few days) is very small compared to the train data (usually from a few years or more).

```
RMSE = 1.2300858790504017
```

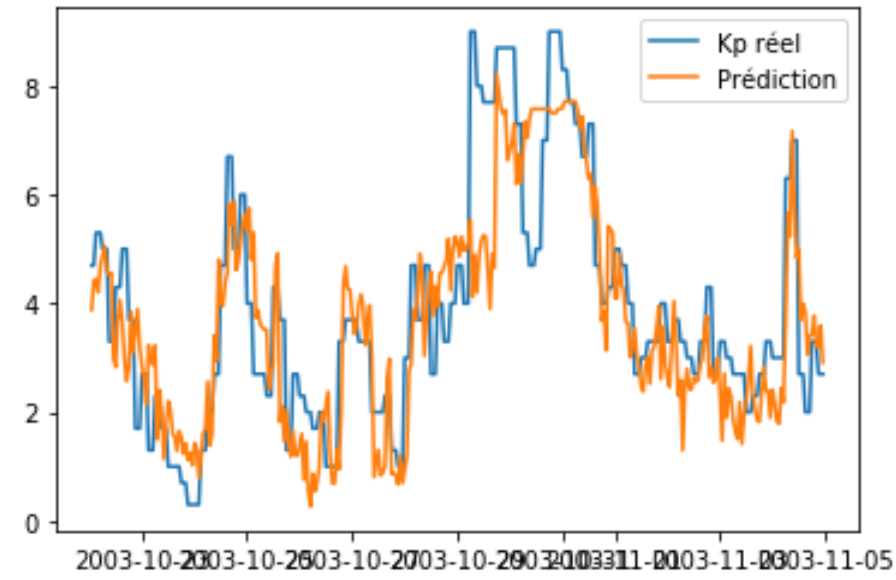


Figure 14: Observed Kp and prediction of Kp

After fitting train input and train output and predict the test output, tt can be seen that the graph of predicted Kp almost coincides with the measured data. Besides, the RMSE of this test is less than 1.5 which we expected before the simulation.

100-fold Cross-Validation: The initial data is divided into 100 subgroups, which in turn are considered as test data. With the 100-fold process, we get 100 RMSE values and graph between them and time

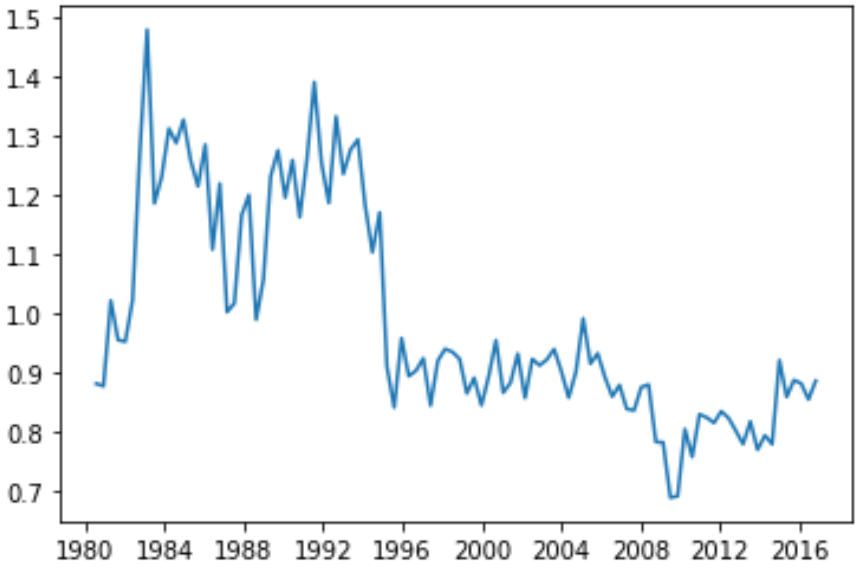


Figure 15: RMSE in 100 times of prediction

It can be seen that the RMSE before 1996 and after 1996 are very different. This can be explained by the operation of WIND and ACE satellites. Due to the progress of science, it was possible to get more accurate figures thanks to these two satellites. From 1996 to 2017, we can see the uniformity of RMSE, the numbers only fluctuate from 0.8 to 1. Thus, this method shows us that K index completely depends on three factors. The velocity, density and position of the solar wind can be estimated relatively accurately by K-nearest neighbor.

3.2 Support Vector Machines

Method [10]

We will use a method called *support-vector regression* (SVR). The model produced by SVR relies just upon a subset of the training data, because the cost function for building the model overlooks any training data close to the model prediction.

There are three different implementations of Support Vector Regression: SVR, NuSVR and LinearSVR. We note that LinearSVR provides a faster implementation than SVR but only considers the linear kernel, while NuSVR implements a slightly different formulation than SVR and LinearSVR.

SVR (or ε-SVR) solves the following primal problem:

$$\min_{\omega,b,\zeta,\zeta^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)$$

subject to

$$y_i - \omega^T\phi(x_i) - b \leq \varepsilon + \zeta_i$$

$$\omega^T\phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*$$

$$\zeta_i \ , \ zeta_i^* \geq 0 \ ; \ i = 1, ..., n$$

where $x_i \in \mathbb{R}^p \ ; \ i = 1, ..., n$ are training vectors (given) and $y \in \mathbb{R}^n$ (given).

Here, we are penalizing samples whose prediction is at least ε away from their true target. These samples penalize the objective by ζ_i or ζ_i^*, depending on whether their predictions lie above or below the ε tube.

For LinearSVR, the primal problem can be equivalently formulated as:

$$\min_{\omega,b} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n} max(0 \ , \ |y_i - \omega^T\phi(x_i) - b| - \varepsilon)$$

Application to the study

For SVR, there are 4 kernels for us to choose but Linear and Rbf (by default) are the most popular. First, we will try the Linear method (**Kernel=linear**)

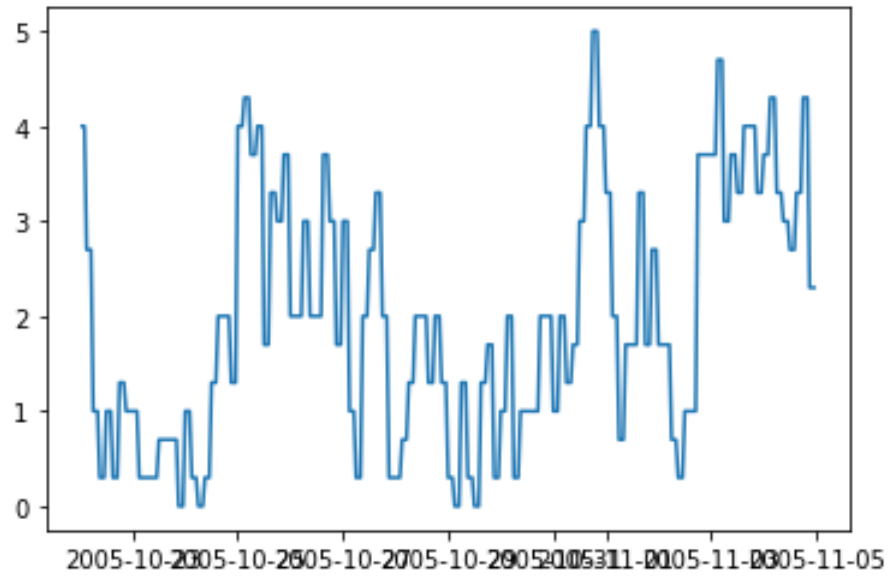


Figure 16: Test output from 22/10/2005 to 05/11/2005

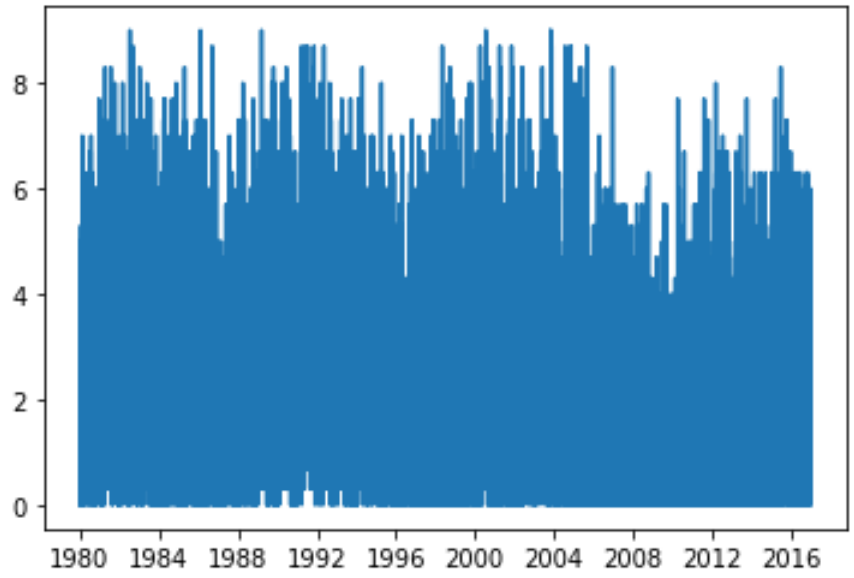


Figure 17: Data from 01/01/1980 to 01/01/2007

Similarly, we split the data into two parts "test" and "train". As we can see, the test data is tinier than the train data.

```
RMSE: 0.9897749927305153
```

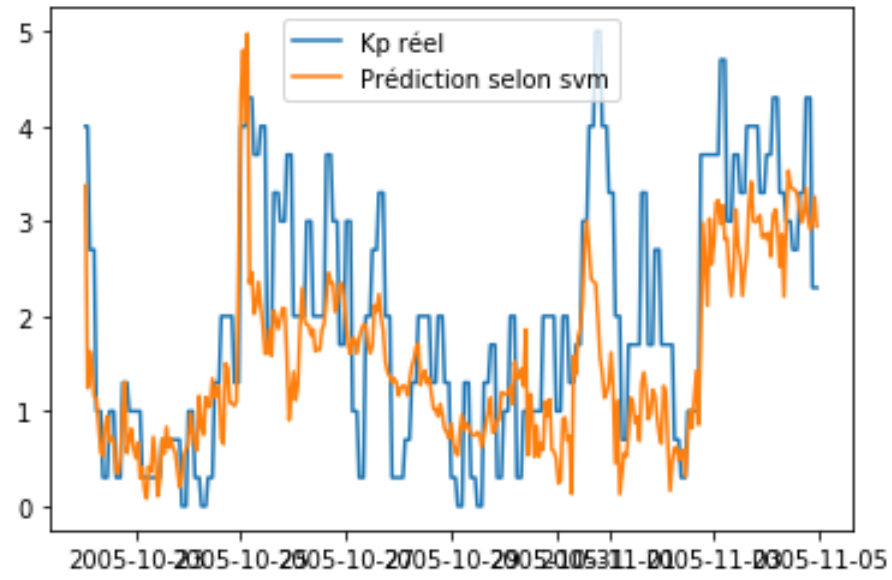


Figure 18: Observed Kp and prediction of Kp

With the test data randomly selected, we have an RMSE of less than 1.5 Kp which is exactly what we expected.

100-fold Cross-Validation: Based on the figure 19 we can see, the number of RMSE is greater than 1.5 in many more tests it is also difficult to say that the graph is homogeneous. This shows that the kernel option is not reliable Linear in this case.

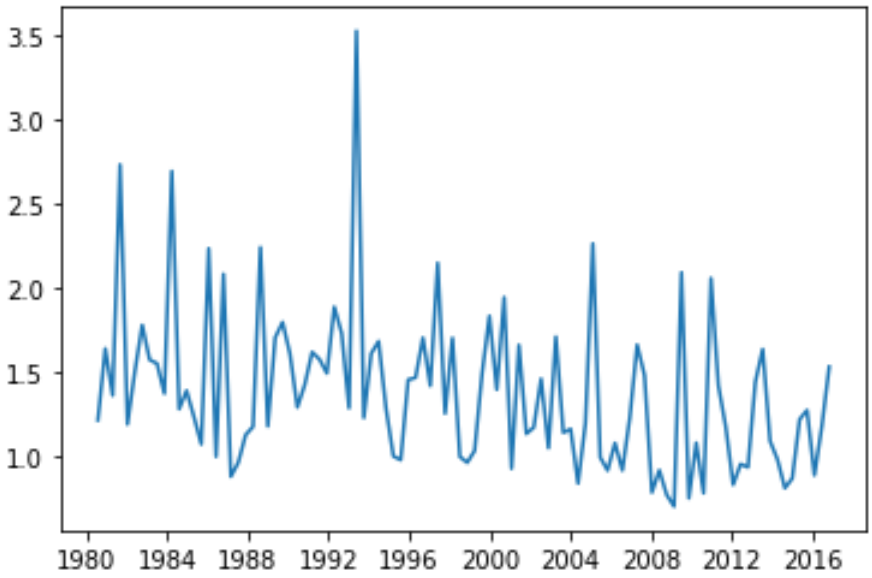


Figure 19: RMSE in 100 times of prediction

With the rbf option (**Kernel = rbf**), the fact shows us that it will take a lot of time if the data considered is from 1980 to 2017. In fact, we have run the

program with the above data for 5 hours but do not record. get results. The data was taken over the period from 2012 to 2017, which led to predictions that would not be as accurate as from 1980 to 2017. To correct this, we took test data for the shortest possible time. (1 day) and perform 1000-fold cross-validation.

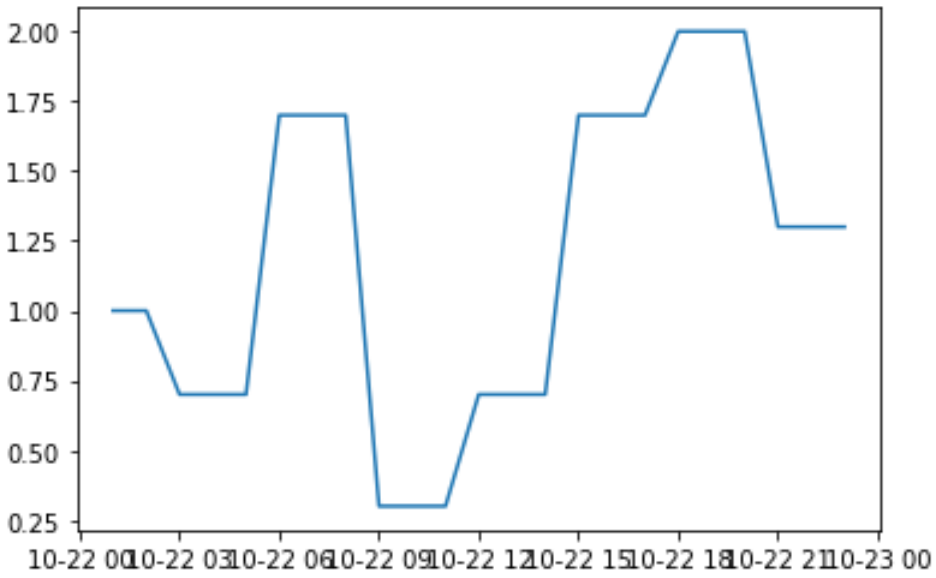


Figure 20: Test output from 22/10/2015 to 23/10/2015

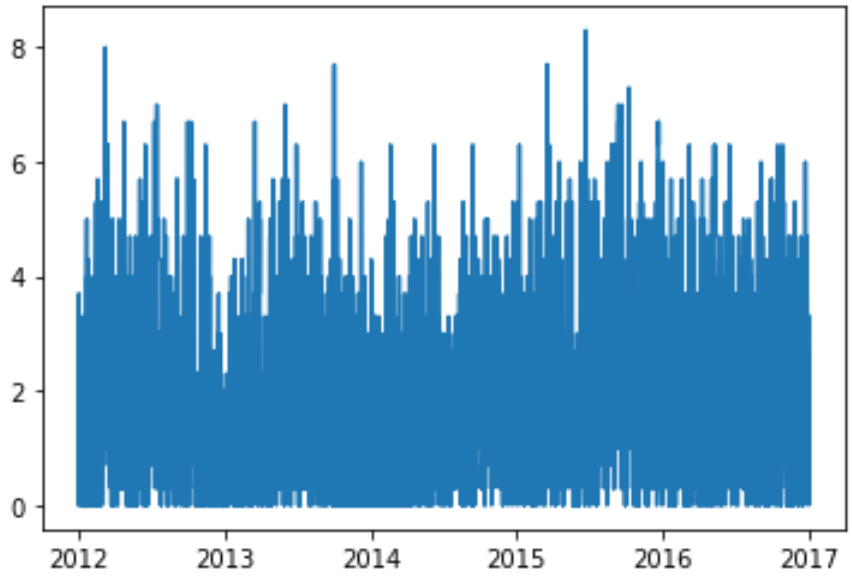


Figure 21: Data from 01/01/2012 to 01/01/2017

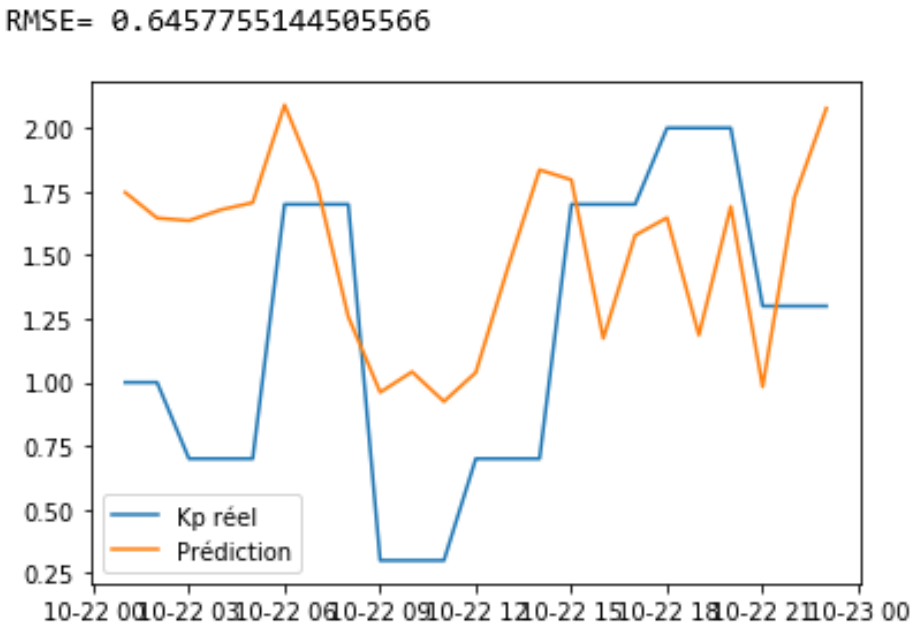


Figure 22: Observed Kp and prediction of Kp

1000-fold Cross-Validation: From figure 23, we can see that the main RMSE index ranges from 0.5 to 1.5 Kp and in addition, the homogeneity seems to appear in the graph. This again confirms that the change in Kp follows a certain pattern and is completely predictable.

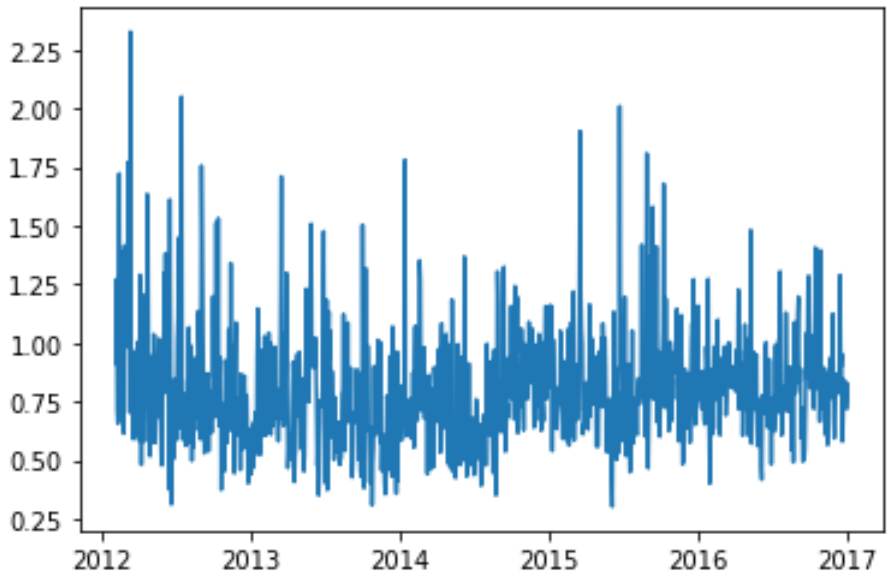


Figure 23: RMSE in 1000 times of prediction

3.3 Decision Tree Learning

Decision Tree

A decision tree is a choice help device that utilizes a tree-like model of choices and their potential results, including chance occasion results, asset expenses, and utility. It is one approach to show an algorithm that just contains conditional control statements.

Decision trees are generally utilized in activities explore, explicitly in choice examination, to help distinguish a strategy well on the way to arrive at an objective, but at the same time are a mainstream apparatus in AI.

Introduction [11]

Decision tree learning is one of the prescient displaying approaches utilized in statistics, data mining and machine learning. It utilizes a decision tree (as a predictive model) to go from perceptions about a thing (spoke to in the branches) to decisions about the thing's target value (spoke to in the leaves). Tree models where the objective variable can take a discrete arrangement of values are called characterization trees; in these tree structures, leaves speak to class names and branches speak to conjunctions of highlights that lead to those class names. Decision trees where the objective variable can take ceaseless values (regularly real numbers) are regression trees.

Decision tree learning is a method generally utilized in data mining. The objective is to make a model that predicts the estimation of an objective variable dependent on a few info factors.

Method

A tree is worked by parting the source set, establishing the root node of the tree, into subsets—which comprise the replacement youngsters. The splitting depends on a lot of splitting rules dependent on order features. This procedure is rehashed on each determined subset in a recursive way called recursive apportioning. The recursion is finished when the subset at a node has no different estimations of the objective variable, or while splitting no longer increases the value of the forecasts. This procedure of **top-down induction of deci-**

sion trees (TDIDT) is a case of a greed algorithm, and it is the most widely recognized system for taking in decision trees from data.

In data mining, decision trees can be portrayed likewise as the mix of scientific and computational procedures to help the depiction, arrangement and speculation of a given arrangement of data.

Data comes in records of the structure:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The needy variable, **Y**, is the objective variable that we are attempting to comprehend, characterize or sum up. The vector **X** is made out of the features , x_1,x_2, x_3 and so forth., that are utilized for that task.

Application to the study

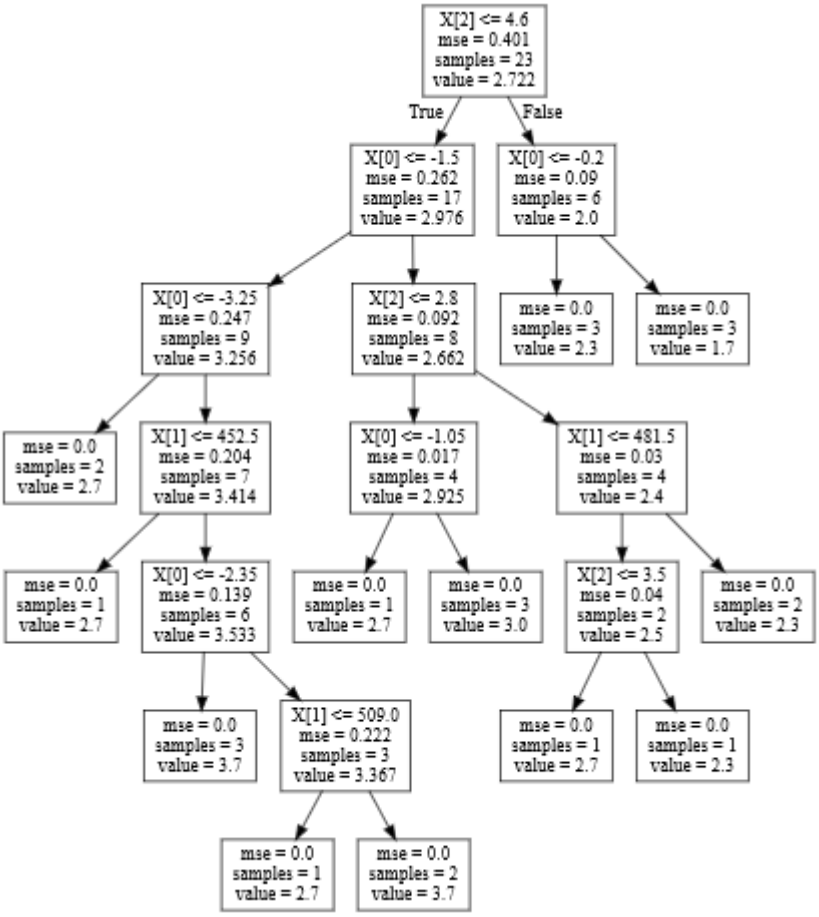


Figure 24:A tree graph

The figure 24 is an example of how Decision Tree Regression works based on the graph simulated by train data. For example, based on this graph (X[0]: BzIMF, X[1]: Velocity, X[2]: Density), if a train input has BzIMF < -3.25,

velocity < 452.5 and density < 4.6, the test output can be predicted about 2.7 Kp. [12]

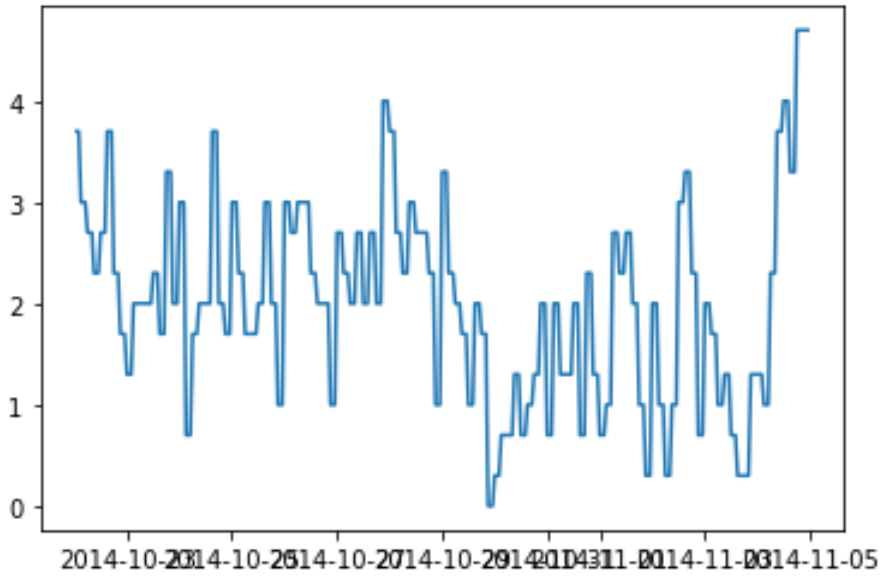


Figure 25:Test output from 22/10/2014 to 05/11/2014

Still with data from 01/01/1980 to 01/01/2017, we randomly selected data from October 22, 2014 to May 11, 2014.

```
RMSE= 1.2168622517617458
```

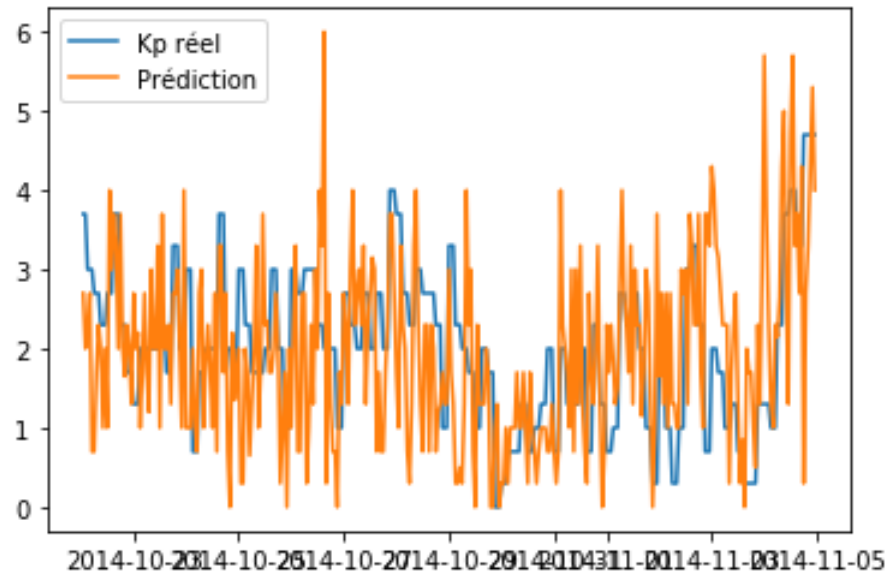


Figure 26:Observed Kp and predicted Kp

We have a RMSE value that satisfies a condition less than 1.5 Kp. However, we can see predicted Kp's very closely drawn lines, which shows that the "tree" created by the train data creates many branches.

100-fold Cross-Validation

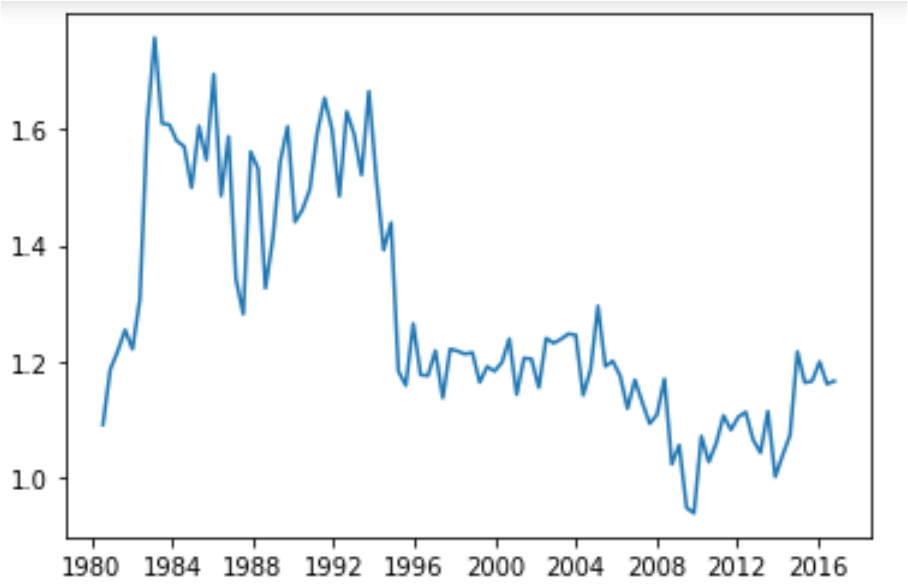


Figure 27: RMSE in 100 times of prediction

The graph nearly produced by Decision Tree is almost the same as that of K nearest neighbor.

3.4 Random Forest [13]

Introduction

Random forest, similar to its name infers, comprises of countless individual decision trees that work as a gathering. Every individual tree in the random forest lets out a class expectation and the class with the most votes turns into our model's prediction. The principal idea driving random forest is a straightforward however ground-breaking one — the astuteness of groups. In information science talk, the explanation that the random forest model works so well is:

An enormous number of generally uncorrelated models (trees) working as a board of trustees will beat any of the individual constituent models.

The key is the low correlation. The same as how speculations with low connections (like stocks and securities) meet up to shape a portfolio that is more prominent than the whole of its parts, uncorrelated models can deliver outfit forecasts that are more exact than any of the individual expectations. The explanation behind this great impact is that the trees shield each other from their individual mistakes (as long as they don't continually all blunder a similar

way). While a few trees might not be right, numerous different trees will be correct, so as a gathering the trees can move in the right heading.

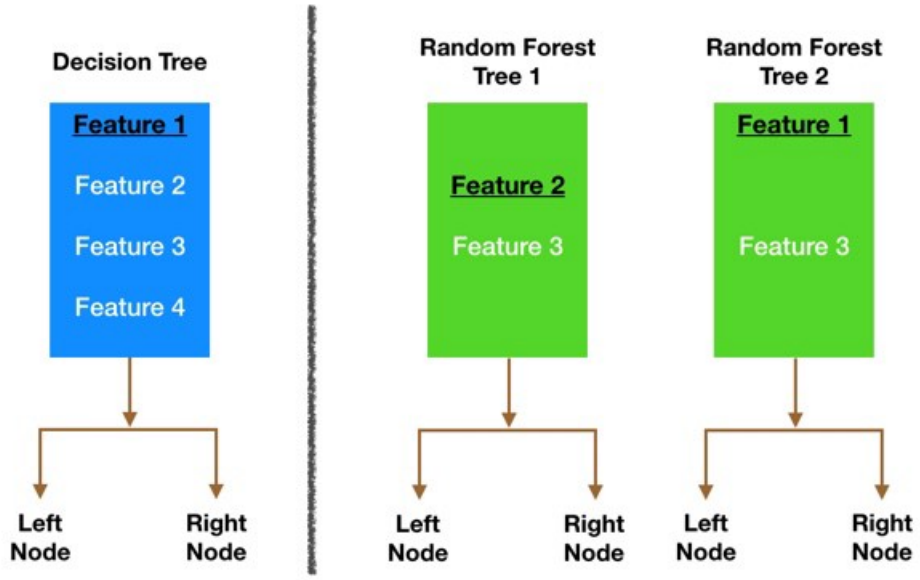


Figure 28: Visualization of decision tree and random forest

Method

- *Preliminaries: decision tree learning* (as above part)
- *Bagging*

  Training set: input $X = x_1, x_2, ..., x_n$ with output $Y = Y_1, Y_2, ..., Y_n$

  Bagging more than once (K times) chooses an irregular sample with substitution of the preparation set and fits trees:

  For k in 1..K:

  1.Choosing n training examples X_k and Y_k

  2.Create a regression tree g_k based on X_k and Y_k

  After training, the prediction on a sample X_{test} is computed by:

  $$\hat{g} = \frac{1}{B} \sum_{k=1}^{K} g_k(X_{test})$$

- *From bagging to random forests*

Application to the study

We implemented the Random Forest method on the same overall data (January 1, 1980 - January 1, 2017) and the time of the test data (October 22, 2014 -

November 5, 2014) as the Decision Tree but receive a smaller RMSE result than the Decision Tree. This is because the Random Forest is an improved version of the Decision Tree and with the success of the Decision Tree, we can believe in the success of the Random Forest.

```
RMSE=  0.9562193132512536
```

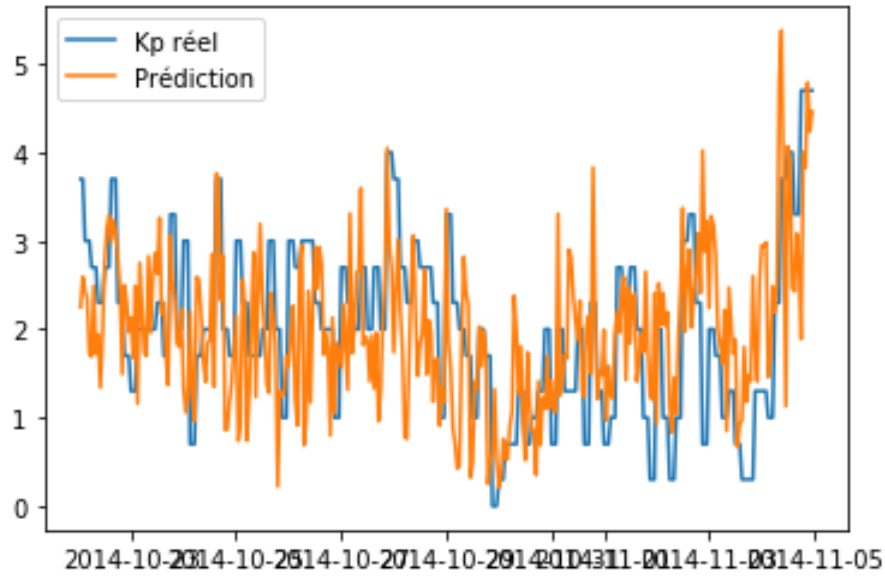


Figure 29: Observed Kp and Predicted Kp

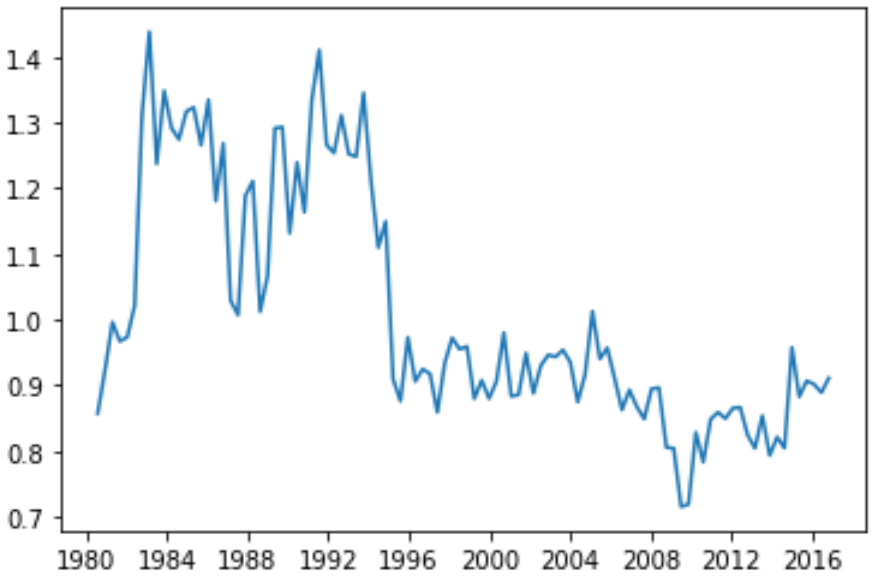


Figure 30: RMSE in 100 times of prediction

The RMSE graph generated by the Random Forest is almost similar to Decision Tree and K-nearest neighbor. This shows that the model predicting Kp is completely reliable.

4 Discussion

Through some concepts, we understand more about outer space, especially those related to solar wind and magnetic fields. From that, based on the model of K-Cross Validation model and specific algorithm methods (K-nearest neighbor, Support Vector Machine, Decision Tree and Random Forest), we can relatively assert that Kp. It can be completely modeled Kp (indicator of disturbances in the Earth's magnetic field) that is based on three important data of solar wind: Velocity, Density and Bz coordinates.

The projections of Kp index after 1996 in the graphs (figure 15, figure 26 and figure 29) are quite reliable when we use three methods of K-Nearest Neighbor, Decision Tree and Random Forest. This shows that these three methods perform quite quickly and accurately with large amounts of IMF data. In addition, the Support Vector Machine method is worth noting because it is hampered by this huge amount of data. However, if we use relative data, we can model the Kp index.

The drawback is that we cannot know for sure the exact value of Kp because these indicators are only modeled relatively. In algorithmic methods, we can only find relatively accurate Kp model while there are many times when the error is greater than 1.5 Kp. Specific formulas are very difficult to find and that is why we have to use modeling. Moreover, the amount of data aggregated over decades is huge and there are methods like Support Vector Machine that are difficult to apply.

Human understanding of space helps us to predict the phenomena impacting the earth. One of the most interesting phenomena is that the solar wind on the earth's magnetic field has long been studied by humans. Not only with physical observation and mathematical formulas, but also thanks to the development of machine learning, we can predict the effects of solar wind on the Earth's magnetic field through three main factors. velocity, position and density. In the future, thanks to advances in machinery and especially machine learning, we can fully simulate and predict the earth's magnetic field more accurately.

Bibliography

[1] https://en.wikipedia.org/wiki/Solar_wind

[2] https://en.wikipedia.org/wiki/Earth%27s_magnetic_field

[3] Hannu .E.J Koskinen, *Physics of space storms: from the Solar surface to the Earth*, p.270
https://books.google.fr/books?id=cO0nwfhXVjUC&pg=PA279&lpg=PA279&dq#v=onepage&q&f=false

[4] https://bitly.com.vn/7HfSi

[5] http://www.dartmouth.edu/~aurora/spaceweather/spaceweatherguide.html

[6] https://solarscience.msfc.nasa.gov/SolarWind.shtml

[7] https://en.wikipedia.org/wiki/Root-mean-square_deviation

[8] https://machinelearningmastery.com/k-fold-cross-validation/

[9] https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-regression

[10] https://scikit-learn.org/stable/modules/svm.html

[11] https://en.wikipedia.org/wiki/Decision_tree_learning

[12] https://scikit-learn.org/stable/modules/tree.html

[13] https://towardsdatascience.com/understanding-random-forest-58381e0602d2