Modèle d'apprentissage pour la prévision du mildiou

Detant Arthur
Steichen Antoine

Encadrants:

E.Ramat
F.Teytaud
S.Verel

M1 ISIDIS
2018-2019

Introduction

Le mildiou est un problème qui persiste depuis longtemps dans la culture, et il n'est pas présent que dans la culture de la pomme de terre, par exemple on le retrouve aussi dans les tomates. C'est une maladie qui peut ravager des cultures complètes de pommes de terre en l'espace de 2 semaines, c'est donc un gros problème alimentaire et économique.
Le projet, effectué en binôme, se compose d'un travail de recherches (lecture, étude et synthèse) et d'une partie de conception d'algorithmes (développement). Il est encadré par 3 enseignant-chercheurs : Eric Ramat, Fabien Teytaud et Sebastien Verel.
Le sujet du projet est "Modèle d'apprentissage pour la prévision du Mildiou". Les objectifs sont de construire un modèle basé sur une technique d'apprentissage supervisée et d'étendre le modèle afin de le rendre ajustable en fonction d'observations. Tout d'abord, il a fallu analyser le modèle mathématique de prévision d'apparition du mildiou (maladie des pommes de terre) puis identifier des outils d'apprentissage adéquats afin de produire le modèle.
Nous avions donc comme objectif de trouver un moyen de prédire la maladie du mildiou de la pomme de terre pour ainsi empêcher cette maladie d'atteindre les parcelles en y appliquant un traitement préventif.
Pour ce faire, le modèle appliqué est basé sur l'humidité et la température au sein de la parcelle, car comme nous l'avons vu au cours de nos recherches, le mildiou se développe le mieux entre 16 et 22 degrés et avec une exposition longue à une humidité supérieure à 90%.
L'objectif est donc de trouver un moyen informatique pour prédire la probabilité de développement du mildiou. Nous avons donc commencé par étudier le mildiou et toutes les caractéristiques autour de cette maladie, puis nous nous sommes intéressés aux techniques pouvant permettre de remédier au problème posé. Nos recherches nous ont amené aux réseaux de neurones et plus particulièrement aux Long Short Term Memory (LSTM).

Synthèse des recherches

Recherches sur le Mildiou

Les enjeux

Pour ce projet, nous avons effectué des recherches dans le but d'en savoir plus sur la maladie du mildiou dans un premier temps. Pour cela, nous avons pu étudier la thèse de Christopher Herbez : "Parallélisation massive de dynamiques spatiales: contribution à la gestion durable du mildiou de la pomme de terre". Nous avons lu toute la thèse pour bien comprendre le sujet dans la globalité mais nous avons particulièrement étudié le chapitre 5 de la thèse qui parle en particulier de la maladie du mildiou, de son développement et des modèles pour lutter contre celle-ci.
La culture de la pomme de terre est l'une des principales cultures alimentaires mondiales, cela représente donc un enjeu économique important. En effet, de part sa structure, la pomme de terre est potentiellement exposée, tout au long de son cycle de production, à un très grand nombre de bioagresseurs tels que les insectes ou les bactéries mais son principal ennemi reste le mildiou. Pour des raisons économiques et sanitaires, il est important de lutter contre le mildiou tout en respectant les quantités de pesticides pour ne pas avoir de problèmes quant aux normes dans chaque pays.

Depuis plusieurs années, la France a mis en place de nombreuses protections phytosanitaires, auxquelles est associé un haut niveau de technologie, pour garantir le contrôle et la propagation du mildiou P. infestans. Cependant, les effets nocifs de l'emploi des pesticides sur la santé des utilisateurs et sur l'environnement amènent aujourd'hui à les utiliser d'une façon plus raisonnable. Le but est donc de réduire l'utilisation des pesticides tout en contrant la propagation du mildiou.
Pour cela, il faut donc investir dans un système comme une station météo, permettant de vérifier les données utiles pour prévoir l'arrivée du mildiou.
Il est vraiment important d'anticiper la maladie, car une fois que celle-ci est visible par l'homme, il est généralement trop tard pour la contrer et cela entraîne une destruction totale des cultures de pomme de terre dans les parcelles aux alentours.

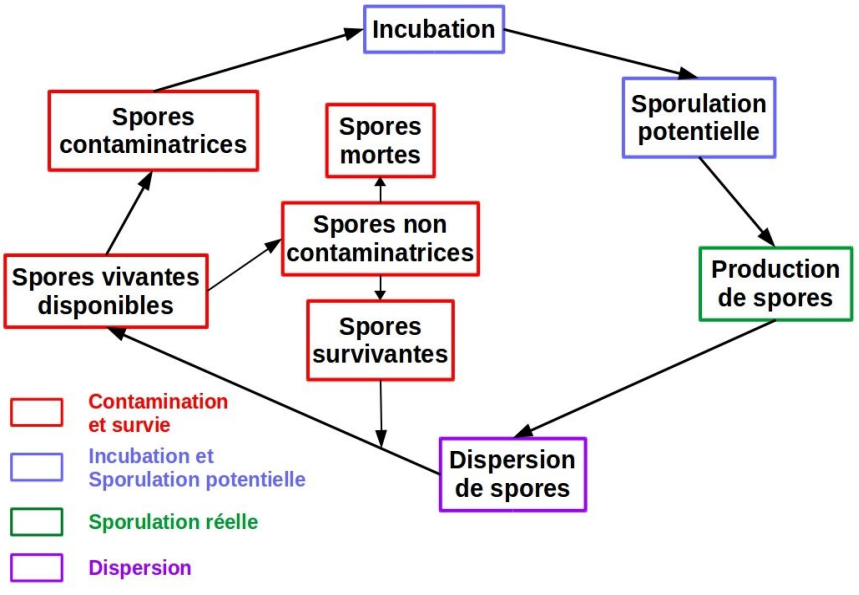
Qu’est-ce que le mildiou ?

Le mildiou de la pomme de terre, causé par Phytophthora infestans de Bary, est l’une des maladies les plus redoutables de la culture de pomme de terre. En effet, la maladie apparaît dans une parcelle sous forme de foyers isolés, à partir desquels elle peut se répandre rapidement et aboutir à une destruction totale de la végétation en moins de deux semaines. Le mildiou se propage sous forme de spores pouvant attaquer tous les organes de la plante. Cette maladie n’est généralement traitée qu’en préventif principalement.



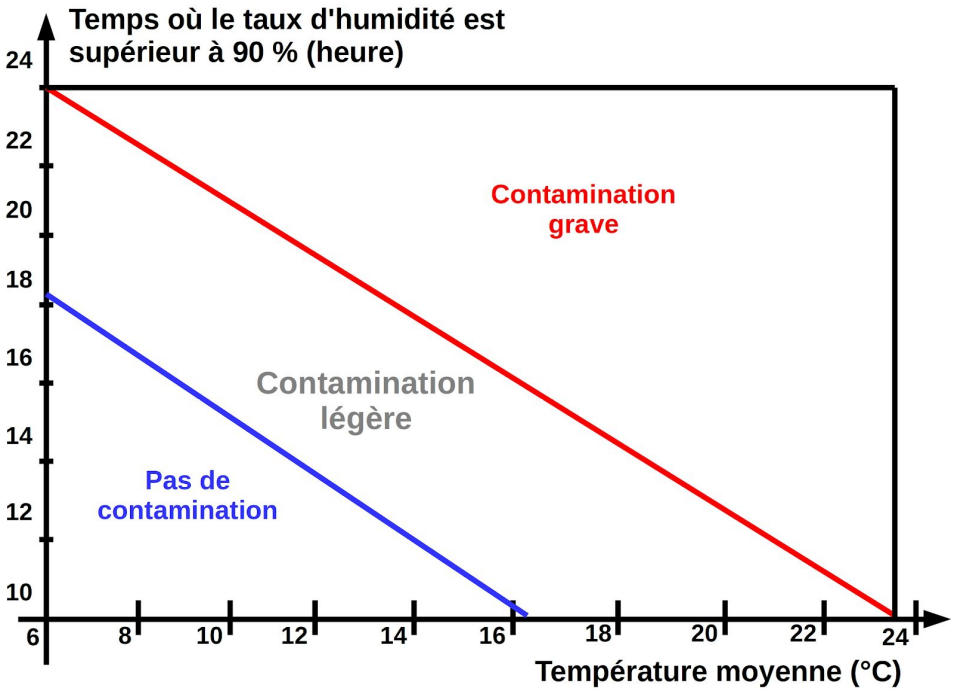
En effet, une apparition précoce provoque surtout une diminution de la photosynthèse, liée à la nécrose des feuilles, pendant la phase de tubérisation, pouvant être à l’origine de la mort des tubercules ou au ralentissement de leur croissance. Tandis qu’une apparition tardive provoque une baisse de la qualité des tubercules, notamment lorsque la contamination a lieu par le sol.

Cycle de développement du mildiou



Un cycle de la maladie correspond à la période qui s'écoule entre deux générations de spores, de l'infection à la production d'une nouvelle génération de spores.
La dispersion des spores asexuées par le vent, ou la pluie, forme le point de départ d'une épidémie de mildiou.

Voici le schéma de la gravité de la contamination :



Les conditions environnementales nécessaires à l'installation de l'agent pathogène (entre 16 et 22°C et l'humidité relative supérieure à 90 %) sont souvent atteintes dans les grands bassins de production français et notamment en Bretagne et dans le Nord de la France.

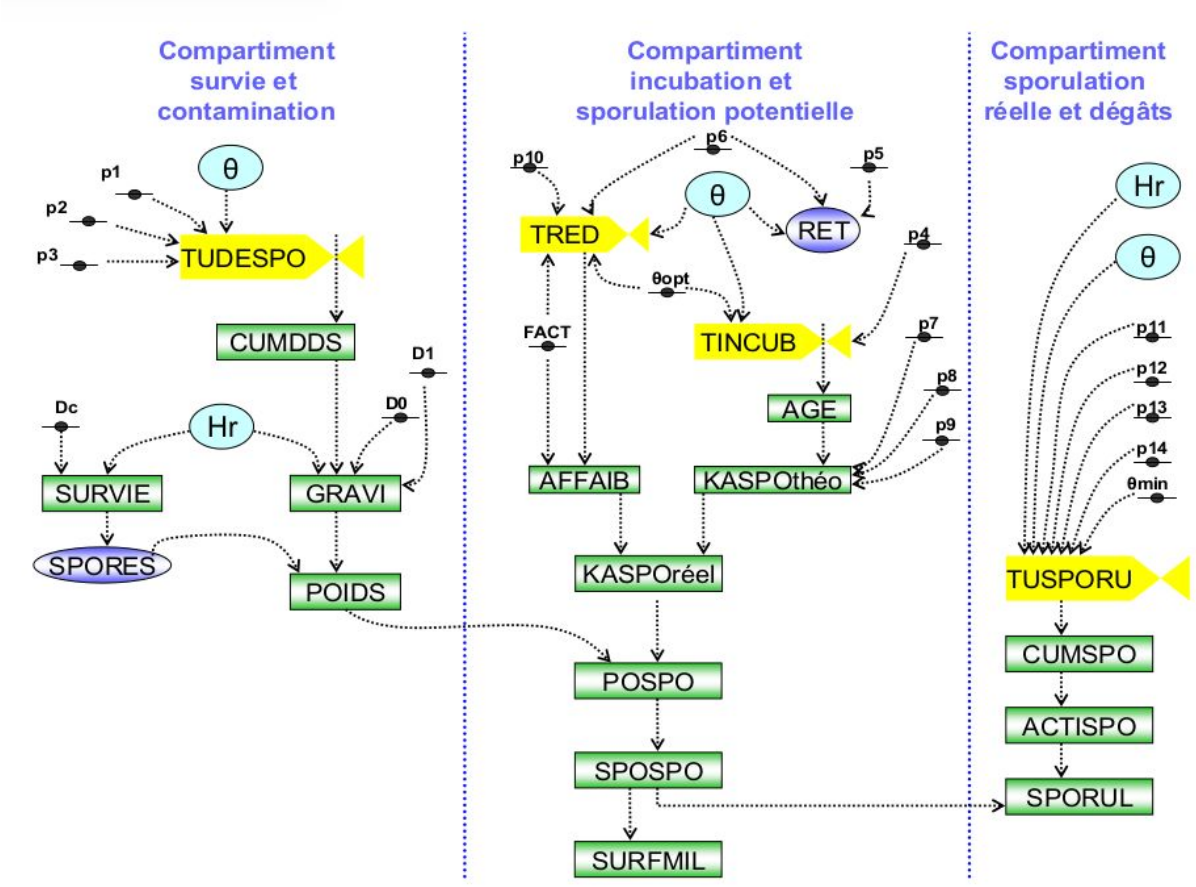
Le modèle Milsol

Dans le cadre du projet nous devions aussi nous intéresser à un modèle de développement des spores, et il se nomme le modèle Milsol.
Il simule, avec un pas de temps horaire, le niveau de risque de mildiou en se basant sur le calcul du nombre de spores vivantes présentes sur le feuillage de la culture et permet de cette manière une quantification de l'épidémie.
Ce modèle est basé sur beaucoup de calculs avec des valeurs que l'on prélève qui permettent d'engendrer de nouveaux calculs jusqu'à arriver à la sporulation réelle, représentée par la variable SPORUL.

Le mécanisme de fonctionnement de ce modèle est présenté par le schéma ci-dessous.

Nous pouvons distinguer trois catégories de variables, classées par couleur :
- Les vertes, de forme rectangulaire, correspondent aux variables d'état.
- Les bleus argentées, de forme ovale, correspondent aux variables intermédiaires.
- Les bleus ciels, de forme ovale, correspondent aux variables d'entrée.

Ce mécanisme se compose également d'un taux d'accroissement de la biomasse, représenté en jaune sur le schéma, et nécessite la présence de paramètres, représentés par des ronds noirs. Nous pouvons également distinguer des flèches en pointillés représentant les flux d'informations.

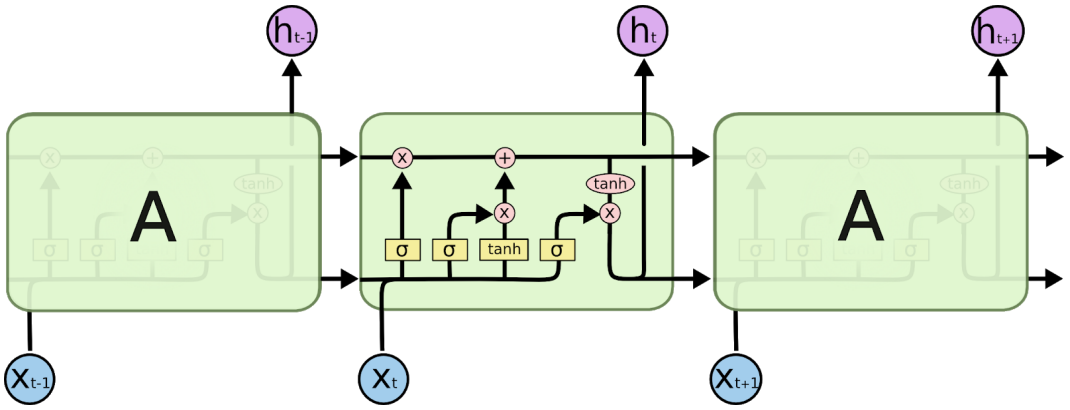


C'est avec l'aide de ce modèle que nous pouvons développer des algorithmes permettant la prédiction du mildiou.

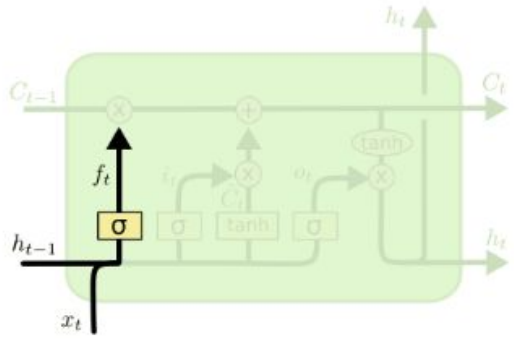
Recherches sur les LSTM

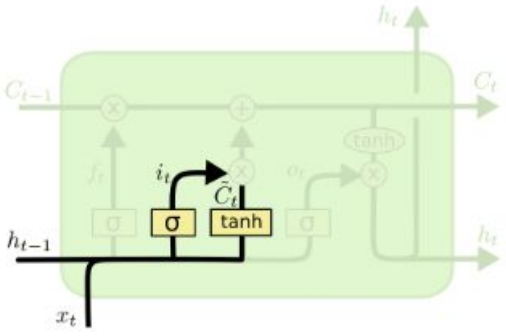
Principe des Long Short Term Memory

Contrairement aux réseaux de neurones récurrents, les LSTM sont capables de mémoriser des informations sur des séquences beaucoup plus longues.
Une cellule LSTM a 3 grosses opérations principales :
Forget Gate : capacité à oublier de l'information inutile.
Input Gate : capacité à ajouter des nouvelles informations utiles.
Output Gate : permet de définir l'état de la cellule à l'instant T.
Donc pour pouvoir faire un "forget", "input" ou "output gate", la cellule LSTM possède un nouveau vecteur qui va être la mémoire de la cellule. La cellule LSTM pourra écrire dedans, supprimer des informations et utiliser les informations à l'intérieur pour définir l'état à l'instant T.

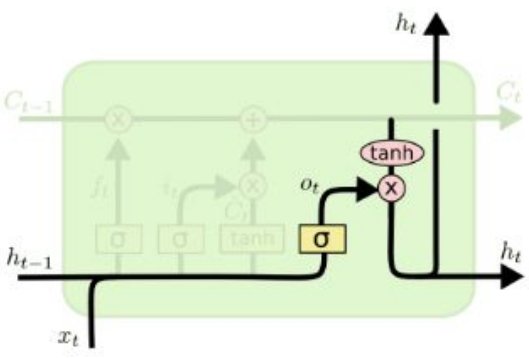


Forget Gate **Input Gate**





Output Gate



Présentation détaillée du travail

Pour le projet nous avons donc commencé par faire beaucoup de recherches concernant le mildiou, le modèle Milsol et les LSTM. La recherche à été une partie importante du projet car elle nous a permis de mieux visualiser dans quelle direction nous devions nous orienter, même si au final nous avons eu besoin de plusieurs indications sur certains détails et pour avoir de l'aide sur les technologies à utiliser.

Nous avons donc été orientés vers les réseaux de neurones pour la prédiction et plus particulièrement les Long Short Term Memory. Ce sont des réseaux de neurones très performants dans le domaine de la prédiction. Nous avons donc appris comment ils fonctionnent et comment nous pouvions les implémenter. Nous avions plusieurs choix qui étaient, en C++ avec des librairies adaptées à l'implémentation des LSTM ou en Python avec la bibliothèque Keras. Nous avons décidé de prendre le langage Python avec la bibliothèque Keras car nous avons remarqué que c'était la façon qui était la plus utilisée pour implémenter les LSTM, donc c'est avec cette méthode qu'il y avait le plus d'explication en ligne.

Keras est une API de réseaux de neurones de haut niveau écrite en Python. Elle a été développée avec pour objectif de permettre des expérimentations rapides. Keras supporte les réseaux récurrents et fonctionne de façon transparente sur CPU et GPU.

En ce qui concerne le code, il faut ouvrir le fichier CSV et traiter les données avant de commencer l'apprentissage. Nous avions fait quelques erreurs au début comme le fait de prendre les fichiers d'humidité et de température en brut, ce qui fait que nous avions énormément de données pour une même journée et de plus nous ne faisions pas ce qui était demandé. Nous prédisions l'humidité avec en entrée un fichier d'humidité. En effet il était facile d'avoir de bons résultats car l'algorithme n'avait presque rien à faire (prédiction de l'identité).

Suite à la soutenance finale, nous avons repris le travail pour essayer d'avoir de meilleurs résultats. Nous avons donc fait un programme en C++ qui nous permet de calculer les moyennes jour par jour pour l'humidité et la température. Ensuite, on récupère la sporulation ou la probabilité jour par jour dans le programme C++ fourni (implémentation du modèle Milsol). Puis nous avons assemblé les données pour avoir un fichier CSV complet avec le numéro du jour, l'humidité moyenne, la température moyenne et la probabilité ou la sporulation moyenne du jour.

Dans le code il fallait donc récupérer les colonnes qui étaient intéressantes pour l'apprentissage. Il faut donc l'humidité, la température et la sporulation, car la sporulation est en lien direct avec le temps.

Ensuite l'algorithme va devoir s'entraîner avec les valeurs qu'il a en entrée. Il va voir en fonction du jour les différentes valeurs d'humidité et de température qui correspondent à la sporulation donnée. Donc, les données d'humidité et de température sont les entrées et le nombre de sporulations est la sortie des LSTM. L'algorithme va être capable, après être passé dans les LSTM, de prédire une sporulation. On va donc donner en fin l'algorithme un fichier de test où l'on donne l'humidité et la température, pour qu'il puisse nous donner la sporulation qu'il aura estimé. Il est possible de diviser le fichier de données en deux parties pour avoir des données d'entraînement et des données de test ou alors utiliser les données des années antérieures comme données de test.

Ainsi, dans le fichier de test nous aurons la sporulation réelle et nous allons pouvoir la comparer avec celle prédite pour voir à quel point la prédiction est bonne. Bien évidemment, plus il y aura de temps d'entraînement plus la prédiction sera bonne. Si on le fait passer une seule fois dans la base de test, l'algorithme va faire de mauvaises prédictions car il n'aura pas eu le temps de comprendre. Cependant, Il faut faire attention à ne pas sur-entraîner. Pour finir, on peut afficher les résultats pour avoir un meilleur visuel sur l'état des courbes.

Malheureusement, après avoir cherché dans la documentation et tenté plusieurs implémentations, nous n'avons pas réussi à gérer l'apprentissage des LSTM avec la bibliothèque Keras.

Tests et résultats principaux

Dans un premier temps, nous allons voir le programme C++ pour le mildiou.
Le programme prend en entrée les fichiers d'humidité et de température de l'année 2018. Le fichier est en format CSV :

	A	B	C	D	E
1	1	1	2018	1:13:34	95.8
2	1	1	2018	1:28:34	95.8
3	1	1	2018	1:43:34	94.5
4	1	1	2018	1:58:34	97.6
5	1	1	2018	2:13:34	96.8
6	1	1	2018	2:28:34	98.1

Les 2 fichiers ont l'apparence ci-dessus.

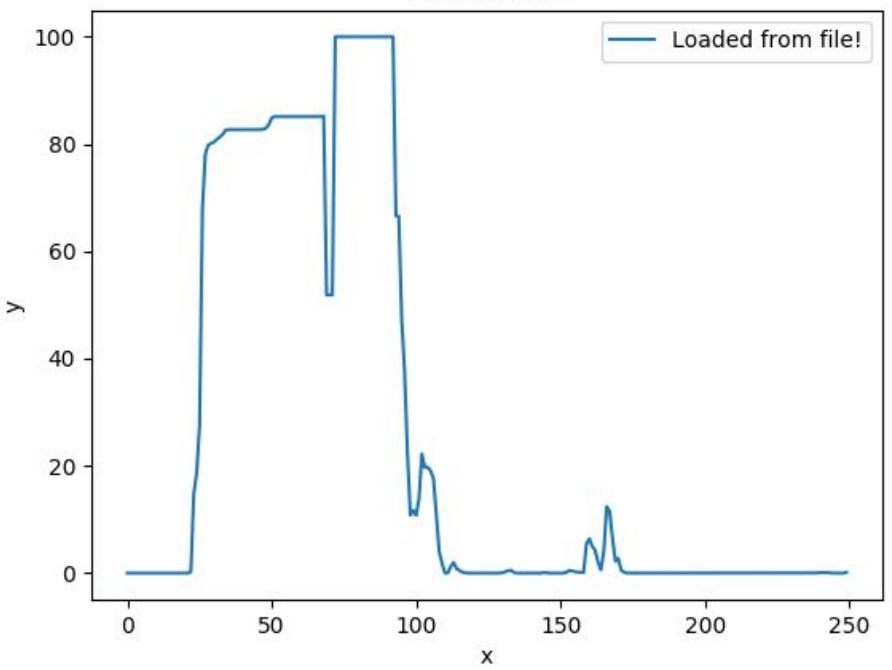
En sortie du programme, cela génère des fichiers de sporulation pour chaque jour (en fonction des fichiers donnés en entrée) :

```
3   8077.05;0
4   12115.6;0
5   16154.1;0
6   20192.6;0
7   24231.2;0
8   28269.7;0
9   32308.2;0
```

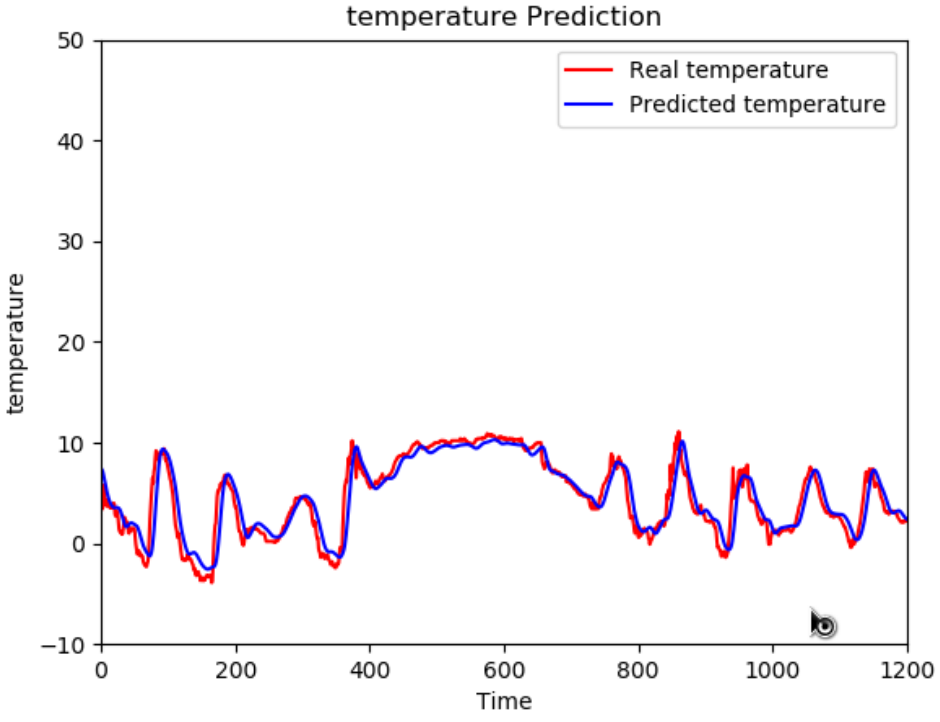
Puis il génère aussi un fichier de probabilité où l'on y voit comme son nom l'indique la probabilité en fonction du jour :

```
21;0
22;0.25
23;14.77
24;18.62
25;27.42
26;67.9
27;78.16
28;79.8
29;80.13
```

Voici donc la courbe de probabilité créée :



Suite à la compréhension de ce programme, nous avons donc implémenté notre premier réseau de LSTM, nous avions donc un fichier qui prenait seulement une entrée (humidité ou température) puis nous avons fait le nécessaire pour avoir des prédictions :



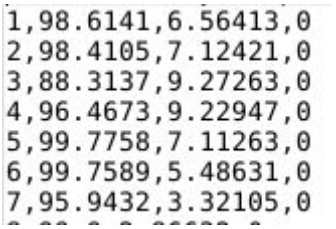
Le fichier de train était un fichier de 2018, et le fichier de test était de 2017. Cela voulait juste dire qu'il apprenait les température en fonction de la date et cela explique donc pourquoi les courbes étaient si rapprochées car tous les ans c'est plus ou moins les mêmes températures et humidités.

Nous avons pris connaissance de ce problème à la soutenance finale. Nous avons donc décidé d'essayer d'améliorer avant le rapport notre programme.
Pour cela, nous avons suivi les conseils donnés à la soutenance et nous avons commencé à coder un programme C++ pour faire des fichiers CSV qui permettront au programme d'avoir les informations nécessaires pour apprendre.
Le programme lis donc le fichier CSV de notre choix, il prend toutes les valeurs de chaque jour et il calcule la moyenne. On obtient donc en sortie nous avons un fichier qui ressemble à ceci :

```
10 99.2442
11 99.9
12 99.8189
13 99.8358
14 95.9126
15 99.8284
16 85.7621
17 74.381
```

Avec à gauche le jour et à droite l'humidité moyenne correspondante.

Nous faisons de même avec le fichier de température. Puis nous devons créer un nouveau fichier où l'on assemble le jour, l'humidité, la température, puis la sporulation moyenne:



```
1,98.6141,6.56413,0
2,98.4105,7.12421,0
3,88.3137,9.27263,0
4,96.4673,9.22947,0
5,99.7758,7.11263,0
6,99.7589,5.48631,0
7,95.9432,3.32105,0
```

Pour la suite, il fallait donc apprendre avec en entrée le fichier CSV, puis en lui donnant un fichier de test sans la sporulation, il doit pouvoir la prédire en se servant de l'humidité et de la température. Malheureusement, nous avons compris ce fonctionnement un peu tard, ce qui nous a pas permis d'implémenter le programme correspondant à ce que nous voulions faire.

Conclusion et perspectives

Pour conclure ce rapport, nous avons effectué beaucoup de recherches au cours de ce projet. Il nous a permis d'obtenir des connaissances sur un sujet nouveau. Nous avons pu constater son avancement au fil du temps grâce à nos recherches.

Cependant, nous avons fait une erreur dans l'implémentation des LSTM. La complexité des LSTM avec la bibliothèque Keras et le manque de temps ont fait que nous n'avons pas réussi à obtenir le résultat voulu. Nous n'avons pas assez fait appel aux encadrants afin de nous débloquer quand nous avions des problèmes.

L'objectif final était de pouvoir prédire le nombre de sporulations afin de déterminer le risque de mildiou avec un seuil ou une probabilité pour les jours à venir.

En ce qui concerne les perspectives du projet, nous pensons qu'il faudra obligatoirement avoir des LSTM fonctionnelles et viables afin de pouvoir vraiment nous fier aux prédictions et pourquoi pas essayer de l'intégrer dans de vraies cultures de pommes de terre.

Sources

- “Parallélisation massive de dynamiques spatiales : contribution à la gestion durable du mildiou de la pomme de terre” Thèse, Christopher Herbez
- http://www.modelia.org/html/060904_journeeProtectionCultures/SDuvauchelle_milpv.pdf
- https://medium.com/@CharlesCrouspeyre/comment-les-r%C3%A9seaux-de-neurones-%C3%A0-convolution-fonctionnent-c25041d45921
- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://perso.ens-lyon.fr/tristan.sterin/reports/sterin_BSc_internship_RNN_report_french.pdf
- https://www.youtube.com/watch?v=y7qrilE-Zlc
- https://keras.io/
- https://www.actuia.com/keras/
- http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Tensorflow_Keras_Python.pdf