

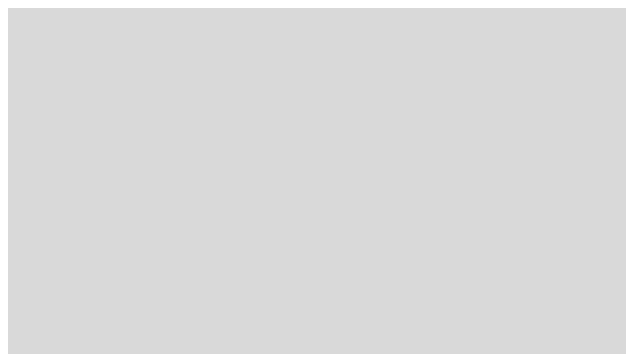


# PROJET AirIQ

Prédiction de la pollution de l'air à Lille :

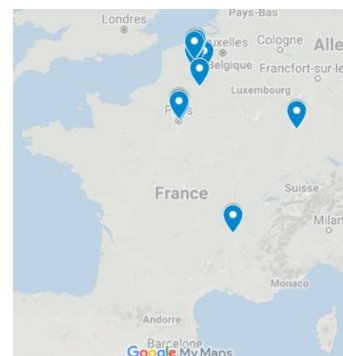
*Réalisé par :*

Alexandre VEREPT  
Mohamed BOULANOUAR  
Maxime THOOR



# Introduction

Selon un classement établi dans le rapport de AirVisual<sup>1</sup> en 2018, Lille est la neuvième ville en France où le taux de concentration en particules fines est le plus élevé (de l'ordre de  $14,3 \mu\text{g}/\text{m}^3$  contre  $17,6 \mu\text{g}/\text{m}^3$  à Saint-Denis, première du classement, en région parisienne). Aujourd'hui, le Nord de la France est la région la plus touchée par la pollution de l'air puisqu'elle ne possède pas de moins de 4 villes dans le top 10 des villes les plus polluées avec Valenciennes, Douai et Roubaix.



Ainsi, savoir prédire l'indice de qualité de l'air dans la métropole de Lille est un challenge important dans un contexte de changement climatique. Cet indicateur peut permettre aux habitants d'être davantage informés sur une des problématiques environnementales du quotidien. Plus globalement, cela a même des conséquences à court terme sur la métropole comme la gratuité des transports, la circulation alternée ou l'augmentation des pathologies respiratoires lors des pics de pollution.

## Indices de la qualité de l'air

- |   |  |
|---|--|
| <span style="color: green;">●</span> Très bon à bon<br>(1-4)        | <span style="color: orange;">●</span> Moyen à médiocre<br>(5-7)                                  |
| <span style="color: red;">●</span> Mauvais à très mauvais<br>(8-10) | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">ND</span> Non déterminé |

L'objectif de ce projet va alors consister à prédire l'indice de qualité de l'air à Lille avec 2 facteurs facilement mesurables à l'échelle locale : la température et l'humidité.

**PROBLEMATIQUE : Peut-on prédire la qualité de l'air de façon fiable avec peu de données grâce au Machine Learning ?**

<sup>1</sup> AirVisual se base sur les données gouvernementales publiques et les données de l'Agence Européenne pour l'environnement.



## **SOMMAIRE :**

**I – Etat de l’art**

**II – Modèle prédictif pour l’indice de qualité de l’air**

**III – Gestion de projet**

## I – Etat de l’art

Aujourd’hui, le calcul de l’indice de qualité de l’air est obligatoire pour les villes de plus de 100 000 habitants ; il s’agit de « l’indice ATMO ». Cependant, il en existe un pour les villes de moins de 100 000 habitants qui prend le nom de « indice de la qualité de l’air simplifié ».

L’indice de qualité de l’air varie entre 1 (très bon) et 10 (très mauvais). Il est calculé par l’observatoire agréé ATMO et est composé de 4 sous-indices, chacun étant représentatif d’un polluant de l’air : particules fines (PM10), ozone (O<sub>3</sub>), dioxyde d’azote (NO<sub>2</sub>) et dioxyde de soufre (SO<sub>2</sub>). L’indice du jour sera donc l’indice le plus élevé des 4 sous-indices.

Le calcul de cet indice est basé sur des stations de fond, c’est-à-dire des stations urbaines et périurbaines situées sur la zone géographique correspondante. Ce calcul ne prend pas en compte les stations de mesure le long du trafic (automobile et industrielle). Le tableau suivant indique les seuils réglementaires par polluant. Pour sa prédiction, l’ATMO se base sur ce tableau :

Valeurs réglementaires (concentrations en µg/m3)				
Indice	Poussières en suspension PM10	Dioxyde d’azote NO <sub>2</sub>	Ozone O <sub>3</sub>	Dioxyde de Soufre SO <sub>2</sub>
10 - très mauvais	80+	400+	240+	500+
9 - mauvais	65-79	275-399	210-239	400-499
8 - mauvais	50-64	200-274	180-209	300-399
7 - médiocre	42-49	165-199	150-179	250-299
6 - médiocre	35-41	135-164	130-149	200-249
5 - moyen	28-34	110-134	105-129	160-199
4 - bon	21-27	85-109	80-104	120-159
3 - bon	14-20	55-84	55-79	80-119
2 - très bon	07-13	30-54	30-54	40-79
1 - très bon	0-6	0-29	0-29	0-39

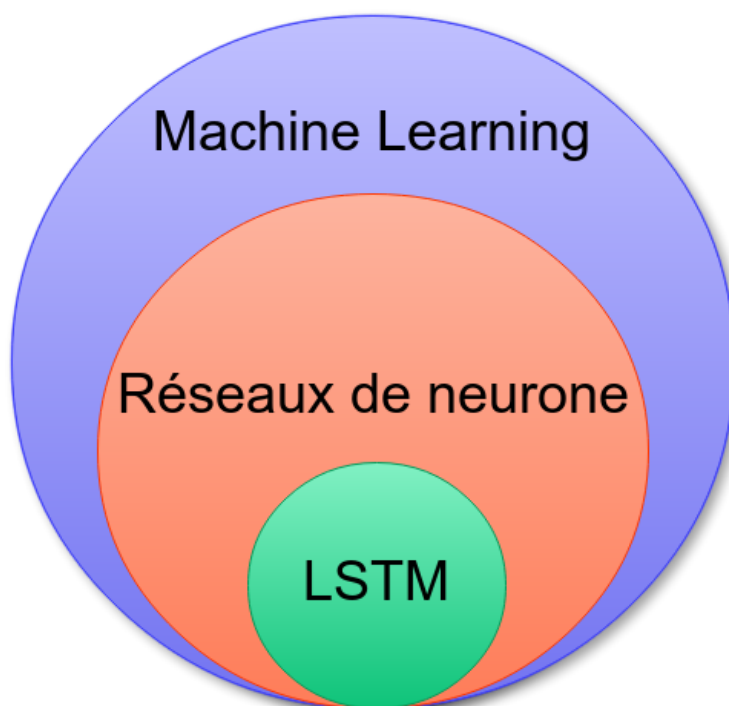
Dans son calcul, l’ATMO prévoit les indices du jour même et du lendemain. Pour obtenir un résultat assez précis, cet observatoire se base sur les modèles de prévisions de la qualité de l’air, les données météorologiques et les résultats des différentes stations. Il faut donc un certain nombre de paramètres et d’outils de précisions pour pouvoir effectuer une prédiction précise et fiable.

Ces dernières années, de nouvelles méthodes de calcul basé sur de l’autoapprentissage se sont développées, notamment le Machine Learning. Ces nouvelles techniques permettent de garantir des résultats avec plus ou moins de précision car cela dépend des données (pertinence, qualité, nombre, etc.) et de l’entraînement de l’algorithme. Il existe un grand nombre de réseau de neurones avec à chacun sa spécificité ; il suffit donc d’adapter le réseau au problème rencontré.

Dans notre cas, afin de prédire la qualité de l’air du lendemain, nous nous sommes intéressés à la catégorie des réseaux de neurones récurrents. Les réseaux de neurones récurrents permettent de traiter des données temporelles, mais également de propager l’information dans les deux sens. Ce réseau de neurones est même considéré comme plus proche du véritable fonctionnement du système nerveux. De plus, on dit que le réseau est récurrent lorsqu’il conserve l’information. En raison de ses paramètres, c’est le type de réseau qui répond à notre problématique puisque les données météo sont classées temporellement.

Nous sommes rapidement partis sur la RNN (Recurrent neural network) car elle conserve les informations, mais uniquement à court-terme. Or, nous avons besoin d'une certaine quantité de données afin que notre modèle soit précis et fiable.

Après des recherches sur le sujet nous avons découvert et conclu que la LSTM (Long short-term memory) est le réseau qui convient le mieux pour notre future prédiction (c'est une version particulière de RNN). Elle a la particularité de posséder une mémoire interne appelée cellule. Celle-ci permet de maintenir un état aussi longtemps que nécessaire.



Nous avons également trouvé quelques papiers de recherche universitaires sur la prédiction de qualité de l'air avec des réseaux récurrents : tous utilisaient des métriques telles que la direction et la force du vent, le trafic urbain mesuré, pression... en plus de la température et l'humidité. Nous n'avons pas accès à de telles mesures pour notre projet, mais il serait intéressant de s'y intéresser pour une future évolution.



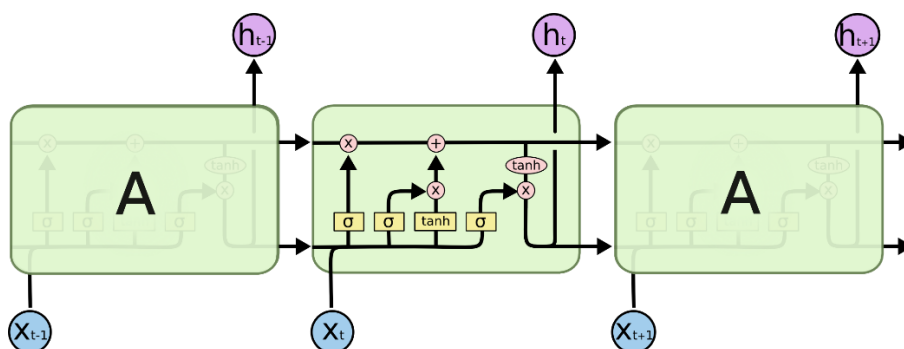
## II – Modèle prédictif pour l'indice de qualité de l'air

Le but de notre modèle est donc de réussir à prédire la pollution de l'air du lendemain à partir des données de la ruche située sur le toit de l'ISEN.

Avec ce type de données, il est évident que nous ne pouvons pas prédire l'indice de qualité de l'air avec 100% de précision, mais nous devons donner une approximation suffisamment fiable pour pouvoir être utilisée concrètement.

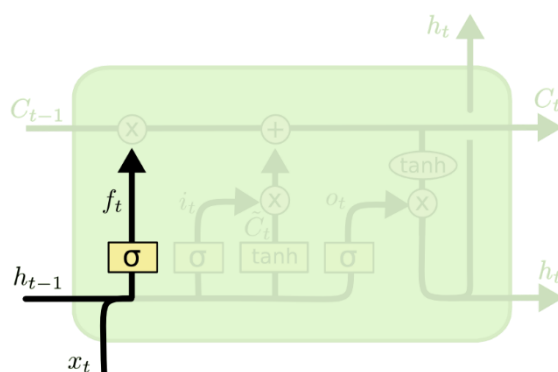
- Présentation rapide d'une LSTM

À la suite de l'état de l'art, nous avons choisi le modèle de la LSTM. C'est un cas particulier des réseaux RNN. Sur le principe elle permet d'accorder plus ou moins d'importance à des données en fonction de leur ancienneté ou de leur pertinence dans le contexte actuel.

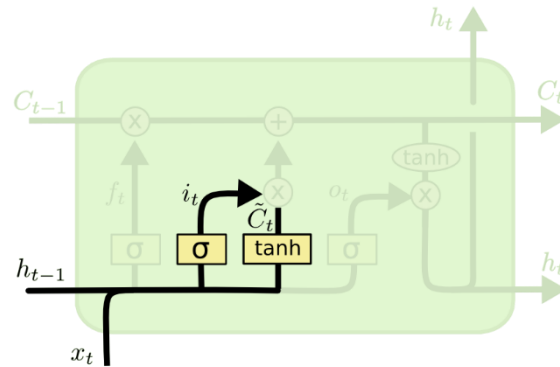


Chaque cellule est composée de plusieurs portes, avec chacune une opération mathématique différente.

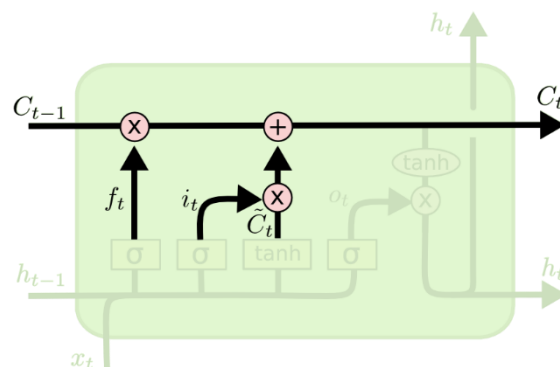
La première porte est une « forget gate », qui sélectionne les informations à garder venant de la cellule précédente à l'aide d'une fonction sigmoïde.



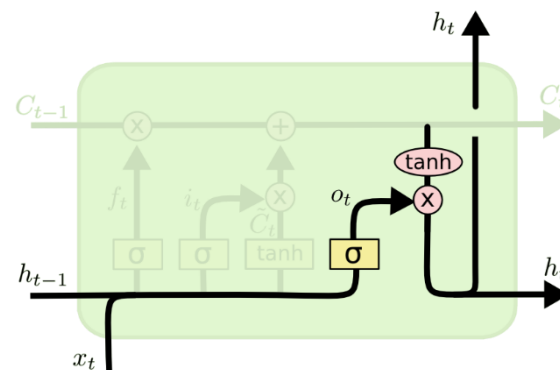
L'étape suivante a pour but de sélectionner de nouvelles informations. La sigmoïde appelée « input gate layer » décide quelles valeurs mettre à jour dans celles qui étaient déjà présentes, et la fonction tangente hyperbolique sélectionne de nouvelles valeurs.



L'opération suivante mets à jour le nouvel état de la cellule avec les informations sélectionnées lors des étapes précédentes.

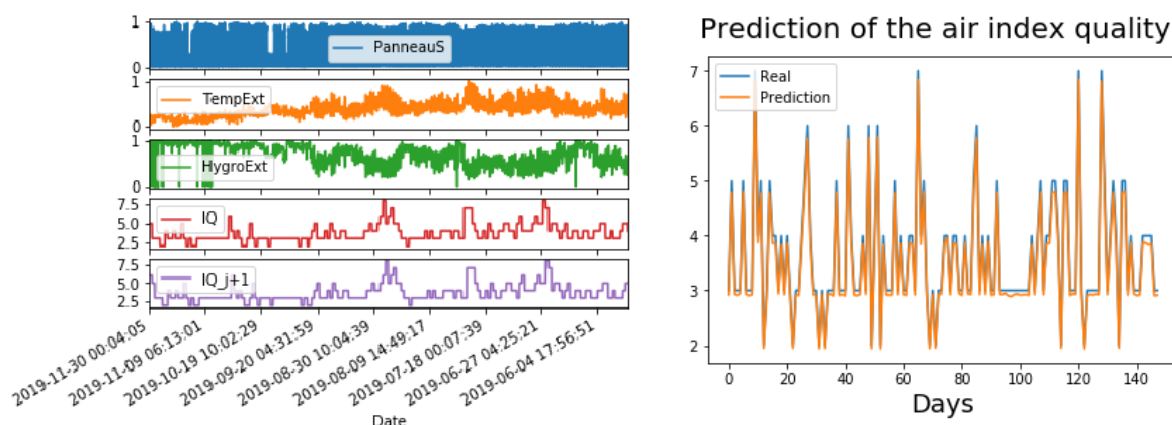


Enfin, nous utilisons une sigmoïde et une fonction tangente hyperbolique pour choisir ce qui sera sélectionné dans l'output de la cellule.



- Première application pour la ruche

En utilisant une LSTM de 96 neurones dans la hidden unit suivie d'une dense de 1 neurone, nous obtenons les résultats suivants :



Les résultats obtenus étant exploitables, nous n'avons que malgré tout une précision maximale sur les données de validation d'environ 35%, ce qui n'est pas très satisfaisant. Nous avons donc essayé de faire varier le nombre de neurones dans la couche dense sans réel changement significatif de résultat.

Nous en avons tiré plusieurs conclusions :

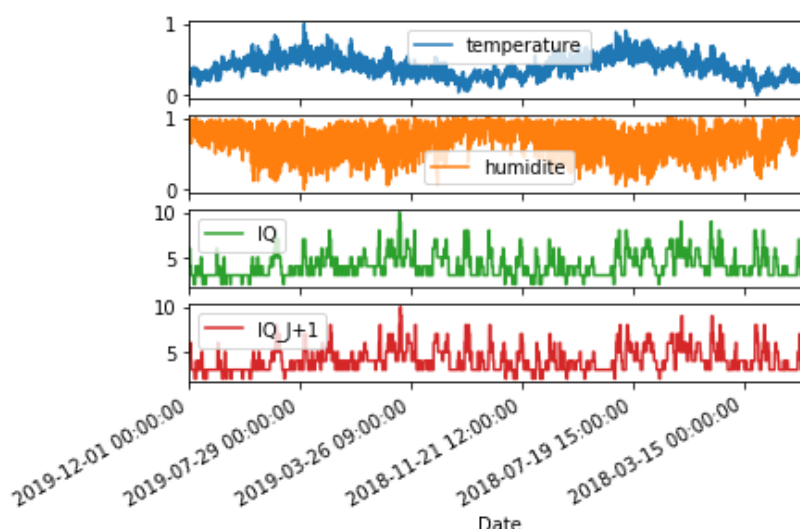
- Nous avons formé notre modèle sur les mois printemps / été et validé avec les mois d'automne, nous devons donc mélanger nos données avant le training.
- Les données normalisées produisent des résultats plus précis.
- La fonction "PanneauS" (= panneau solaire) semble effectivement inutile car la précision n'a pas bougé de manière significative avec ou sans elle.
- Sur cette période, nous n'avons pas d'indice de qualité de l'air à 10, 9, 8 et 1, il manque donc un exemple de situations possibles.
- Cependant, le problème le plus important : 131 jours de données (100 pour la formation et 30 pour la validation) sont loin d'être suffisants pour former un modèle efficace, comme on s'y attendait. En effet, les mesures à l'ISEN n'ont commencé qu'en mai, nous avons donc moins d'un an de mesures. Nous avons besoin de trouver plus de données pour le former !



- Elargissement de notre dataset

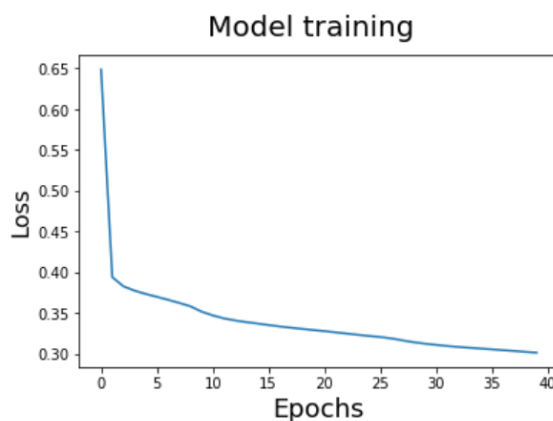
Après ces résultats mitigés mais encourageants, nous avons décidé de former notre modèle aux données publiques, disponibles gratuitement en ligne. Nous choisissons l'ensemble de données météorologiques sur le site web de Météo France. Sur ce dataset, on peut y trouver énormément de paramètres tels que la pression au niveau de la mer, la vitesse du vent ou encore les types de nuages. Cependant, afin de pouvoir par la suite le faire correspondre avec les données de la ruche, nous choisissons de prendre uniquement la température et l'humidité.

Après avoir téléchargé pour chaque mois toutes les données depuis 1996, nous avons sélectionné puis nettoyé les 2 colonnes qui nous intéressaient. Nous avons ensuite croisé cet ensemble de données avec celui contenant l'indice de qualité de tous les jours depuis 2018 et rééchantillonné pour avoir une mesure toutes les 3 heures sur les 2 dernières années.



Pour nous adapter à cette nouvelle forme des données (8 mesures correspondent à 24h), nous changeons notre réseau de neurone avec une couche LSTM de 8 neurones dans la hidden unit suivie d'une couche dense de 1.

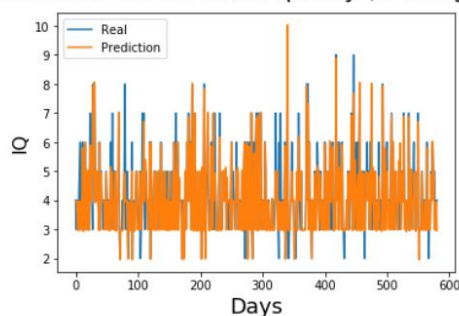
L'entraînement sur ces données a donné des résultats beaucoup plus satisfaisants. En effet, un des premiers signes a été la perte qui diminue de plus en plus, et semble tendre vers 0.3 avec un nombre d'epochs qui ne nous donne pas un résultat qui overfit comme on peut le voir sur le graphe suivant :



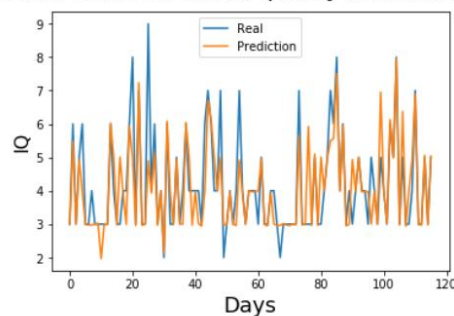
En effet avons constaté qu'au-delà d'environ 40 epochs avec notre technique d'entraînement la validation diminuait à cause de l'overfit.

Pour la précision, nous arrondissons nos valeurs et traitons notre problème comme un problème de classification. Cela nous a donc permis de mieux comparer la valeur indiquée par l'ATMO ainsi que la valeur prédite par notre modèle.

Prediction of the air index quality (training data)



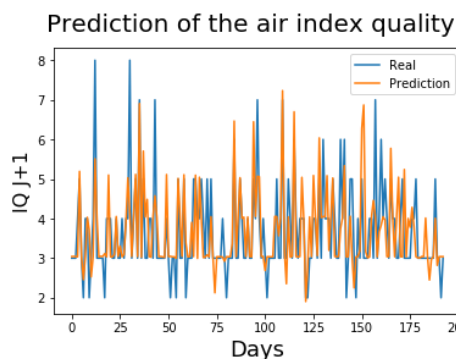
Prediction of the air index quality (validation data)



En considérant que nous avons utilisé seulement la température, l'humidité et l'indice de qualité de l'air du jour précédent, nous obtenons une précision d'environ 70% et une erreur moyenne de 0,5 en validation sur notre meilleur modèle.

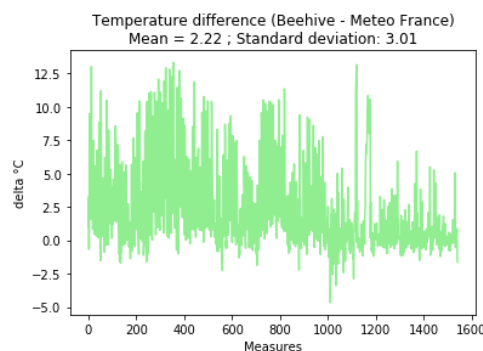
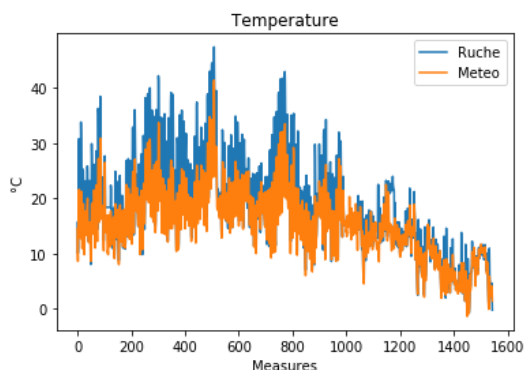
- Application de notre nouveau modèle aux mesures prises par la ruche

Nous testons maintenant le modèle que nous avons formé sur les données de Météo France pour prédire la qualité de l'air avec les données de la ruche. Voici le premier résultat que nous avons eu :



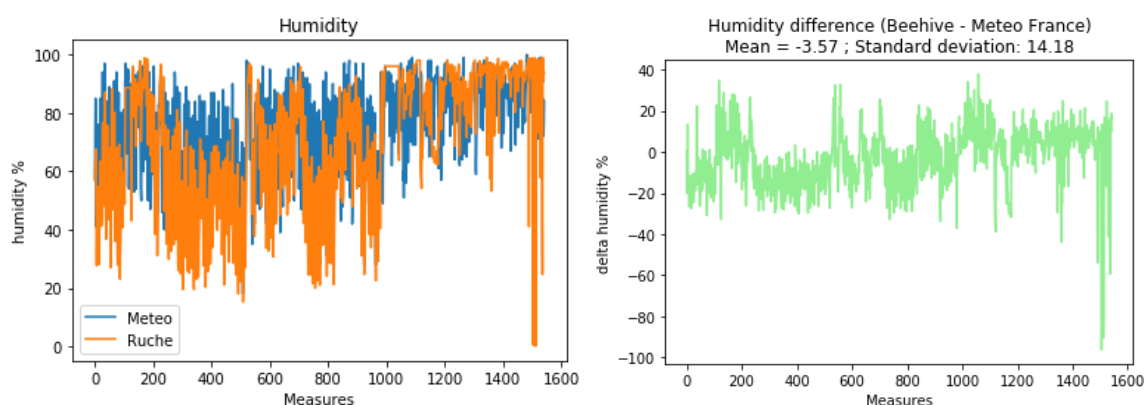
Accuracy: 50.78 %  
Mean error: 0.65

Nous passons d'environ 70% à 50% de précision. Nous voulions alors comprendre la différence entre nos données d'entraînement et nos données de production. Pour cela, nous avons donc fait des tests et des mesures, y compris celles-ci qui comparent la température de la MEL et de la ruche :



On y remarque une tendance claire de la ruche à mesurer une température plus élevée. Il est donc apparu clairement qu'il fallait ajouter un offset (calculé à 2.22 ici) à nos données d'entraînement.

Nous comparons ensuite les données d'humidité sur le même principe :



Cette fois-ci, nous n'avons pas eu de conclusion claire (à part certaines données aberrantes sur la droite). La valeur mesurée par la ruche semble osciller autour des données de la MEL.

Plus tard, la personne en charge des capteurs sur les ruches, Monsieur Capron, a suggéré que les capteurs d'humidité des ruches pouvaient être différents. Nous avons donc cherché à le vérifier :

	Moyenne dans la ruche 1 :	Moyenne dans la ruche 2 :
Température :	16.8022	16.8053
Humidité :	69.4492	69.4719

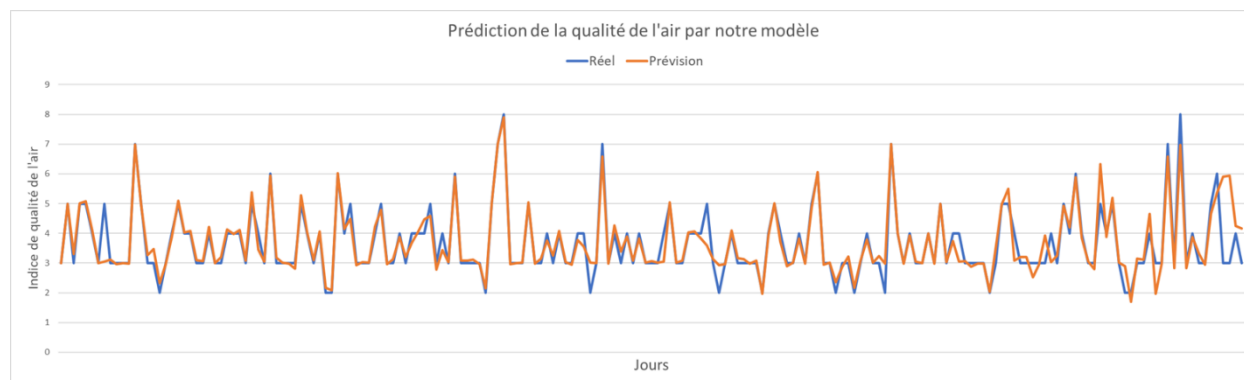
Les écarts entre les 2 ruches semblent ici clairement négligeables au regard des valeurs moyennes.

Plus tard, nous avons découvert le k-fold et la validation croisée. Nous l'avons donc utilisé de manière à réduire l'effet négatif du nombre réduit de données. En effet, cette technique consiste à découper son dataset de training en k morceaux de taille égale (nous en avons choisi 6 dans notre version finale), puis d'entraîner notre modèle en utilisant successivement chaque « fold » pour la validation de l'entraînement. Voici un schéma explicatif :



Cette méthode a 2 avantages majeurs : elle diminue les chances d'overfitting car on entraîne notre réseau avec un dataset d'entrée légèrement différent à chaque itération, mais surtout cela nous permet d'utiliser 100% de nos données de la MEL en training. Or plus on augmente la taille de notre dataset d'entraînement, plus notre modèle s'avérera adapté et précis.

Les résultats tirés de ces quelques améliorations appliquées à la ruche nous ont conduit à des résultats bien plus précis : nous sommes montés à une précision de 90% avec une perte moyenne de 0,4.



Nous avons ensuite rencontré une problématique imprévue lors de la création de la routine en temps réel : prédire l'indice de qualité de l'air tous les matins à minuit. En effet, il n'était pas très utile pour une application concrète, qui nécessite de prévoir en amont la qualité de l'air. Nous avons donc décidé de prédire un indice tous les midis.

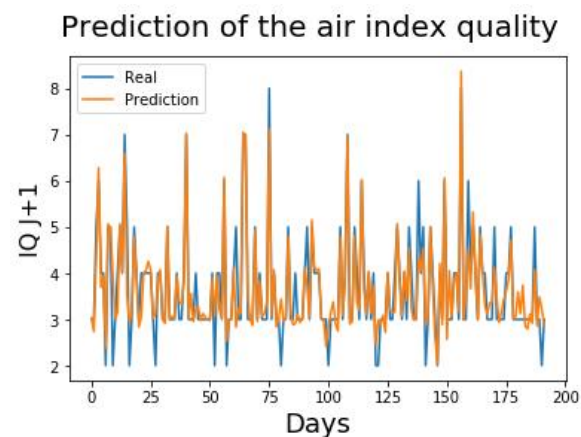
Ceci a une influence non négligeable sur le résultat de nos prédictions, puisque nous devons maintenant entraîner la LSTM de 12h à 12h et non plus sur les données d'une journée. Les résultats sont évidemment moins bons, puisque nous n'avons plus accès aux données les plus récentes (de 12h à minuit la veille du jour à prédire).

Nous avons d'abord testé de prédire l'indice sans réentraîner la LSTM sur ces nouveaux paramètres : nous avons obtenu des résultats compris entre 0.3 et 0.4 en précision et un écart moyen de 0,9 à 1.

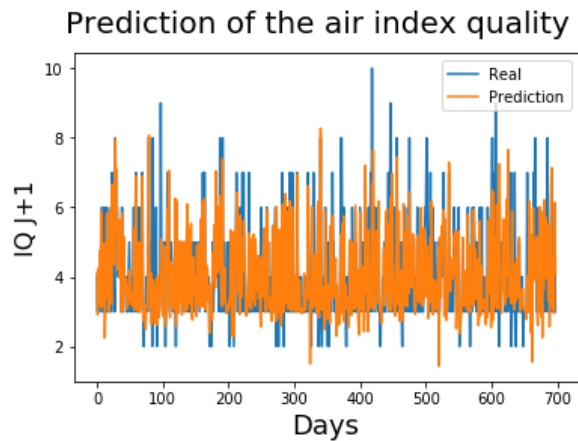
Ce n'était pas satisfaisant, et nous avons donc réarrangé notre dataset afin d'entraîner notre modèle de 12h à 12h. En entraînement sur les données de la MEL et en validation sur la ruche, les résultats étaient d'environ 0.4 à 0.6 en précision et de 0.45 à 0.6 en écart moyen.

Nous entraînons alors légèrement notre modèle sur les données de la ruche afin de coller à la forme de nos données utilisées en temps réel. Nous sommes arrivés à un bon compromis entre bon résultats en temps réel et garder une certaine précision sur les données de la MEL, et voici notre modèle final :

### Résultats sur le dataset de la ruche :



Précision : 82.3%  
Erreur moyenne : 0.32

**Résultats sur le dataset de la MEL :**

Précision : 52.44 %  
Erreur moyenne : 0.71

Nous avons sélectionné ce modèle car il est très précis sur notre dataset de la ruche, ce qui est le plus important car c'est ce que nous utilisons pour notre application en temps réel. Mais ce modèle reste plutôt correct sur le dataset de la MEL afin de pouvoir garder une prédiction correcte dans une situation qui n'est pas présente dans le dataset de la ruche.

### III – Gestion de projet

Afin d'atteindre notre objectif dans un temps imparti, nous avons dû mettre en place une gestion de projet indispensable au bon déroulement du projet. Pour ce faire, nous avons choisi d'utiliser un outil de gestion de projet en ligne, Trello.

- **Première semaine de projet (du 14 au 18 octobre) :**

Après avoir bénéficié d'une meilleure connaissance du projet, de son objectif et des deadlines à respecter au mieux lundi matin lors de la rencontre avec notre tuteur, nous avons commencé à découvrir le dataset et effectuer un état de l'art sur les modèles de prédiction de l'indice de qualité de l'air à Lille et dans certaines grandes villes du monde. Pour pouvoir avancer plus rapidement, nous nous sommes réparti le travail : Alexandre s'est occupé de récupérer les données de la MEL, Maxime s'est chargé de récupérer celle de la ruche et Mohamed a effectué une comparaison sur les différents types de modèles de Machine Learning.

Une fois ces informations récupérées, en milieu de semaine, nous avons tous suivi un tutoriel Tensorflow et Keras sur les conseils de notre tuteur pour mieux appréhender le Machine Learning sur Python. En parallèle, Mohamed a récolté toutes les informations nécessaires pour réaliser une LSTM (Long Short-Term Memory). Puis, nous avons entraîné notre premier modèle Tensorflow en réalisant d'abord une moyenne par jour de tous les paramètres que nous offrent la ruche. Cette réussite nous a permis de continuer assez rapidement.

Cependant, en fin de semaine, nous avons dû faire face à des difficultés lors de la création de notre première LSTM fonctionnelle avec un dataset propre sans utiliser la moyenne mais en regroupant toutes les données d'une journée. Ces complications ont été marquées par des erreurs dans le code sur la couche LSTM et sur le format de notre dataset.

Cette première semaine s'est conclue sur une réunion avec notre tuteur pour lui faire part de nos avancées récentes et de nos problèmes concernant le code. Lors de ce débriefing, l'objectif fixé pour la semaine 2 est d'alors finir la LSTM déjà commencée. On choisit également de laisser tomber l'idée de faire la moyenne de tous les paramètres de notre dataset.

	Assigned	Progress	OCTOBER 2019																
			14	15	16	17	18	21	22	23	24	25	28	29	30	31			
▼ 17 NOV : Modèle qui tourne (pas forcément optimal)		100%																	
Comparer modele machine learning	mohamed	100%																	
Découvrir les données	Alexandre Verept, Maxir	100%																	
Récup données API MEL	Alexandre Verept	100%																	
Récupérer données de la ruche	Maxime THOOR	100%																	
Récupérer informations LSTM	mohamed	100%																	
Mettre en forme le dataset	Alexandre Verept, Maxir	100%																	
Tutoriel Tensorflow	Alexandre Verept, Maxir	100%																	
Entraîner modèle tensorflow avec la moyenne par jour	Alexandre Verept, Maxir	100%																	



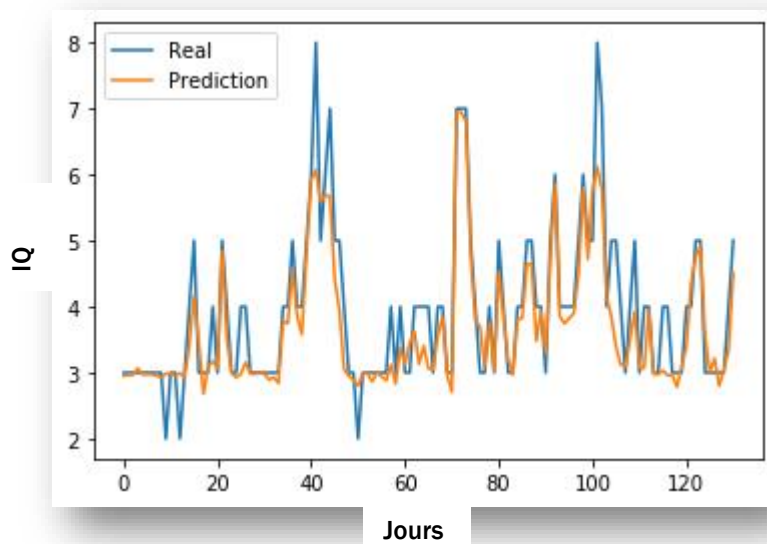
- Deuxième semaine de projet (du 12 au 15 novembre) :

Presque 1 mois après notre première semaine de projet, nous reprenons la suite de nos avancées le mardi 12 (le lundi 11 novembre étant un jour férié). Entre temps, Alexandre a eu le temps de créer des graphiques représentant les différents facteurs nécessaires à la recherche de l'indice de qualité de l'air. De plus, nous avons pris le temps ensemble de mettre à jour le GitHub, support écrit de notre projet et outil de partage du code, ainsi que le Trello.

En début de semaine, nous avons eu quelques problèmes avec Spyder à la suite d'une mise à jour de Tensorflow qui a empêché le lancement de notre code. En effet, nous étions sur la version 1.14 avec Tensorflow et le temps de trouver qu'il s'agissait de ce problème a été assez long. Suite à la résolution de ce problème, nous avons donc cherché à améliorer notre modèle. Pour cela, nous avons pris un exemple trouvé dans la documentation de Tensorflow sur la prédiction de la température à partir d'un dataset pour tenter de l'adapter à notre modèle. Ce travail a été réalisé par l'ensemble du groupe de façon à avancer sur la même dynamique pour multiplier les idées et pour améliorer la productivité.

Jeudi matin, nous avons réussi à véritablement adapter l'exemple à notre modèle en cherchant toujours à entrer un peu plus dans le détail du code. Le résultat est encourageant : il renvoie bien l'indice de qualité de l'air que l'on veut prédire en fonction de l'indice de qualité de l'air réel calculé par l'ATMO.

Graphique représentant l'indice de qualité de l'air de l'ATMO et de notre modèle



La réalisation de notre première LSTM fonctionnelle et non optimal est alors finie ; l'objectif que nous nous étions fixés en fin de première semaine est atteint.

Enfin, vendredi 15 novembre à 10h, l'ensemble du groupe a assisté à l'intervention d'un monteur professionnel pour la réalisation du support vidéo. Dans l'après-midi, nouvelle réunion avec notre tuteur pour parler des perspectives possibles pour la suite ainsi que de la réalisation de la vidéo. Dans les temps par rapport à ce que nous avons fixé en début du projet, l'objectif a été le suivant : réussir à avoir un modèle fonctionnel et optimal.

	Assigned	Progress	NOVEMBER 2019																
17 NOV : Modèle qui tourne (pas forcément optimal)		100%	11	12	13	14	15	18	19	20	21	22	25	26	27	28	29		
Créer une première LSTM fonctionnelle avec nos données	Alexandre Verept, Maxir	100%						Alexandre Verept, Maxime THOOR, mohamed											
Perfection de la première LSTM	Alexandre Verept, Maxir	100%						Alexandre Verept, Maxime THOOR, mohamed											
Modèle qui tourne (pas optimal)	Alexandre Verept, Maxir	<input checked="" type="checkbox"/>						Alexandre Verept, Maxime THOOR, mohamed											

- Les 3 dernières semaines de projet (du 2 au 20 décembre) :

Après 15 jours de « pause » dans le projet, nous sommes partis pour les 3 dernières semaines. À la suite d'une réflexion générale lors de la dernière réunion, notre tuteur nous a envoyé un mail afin d'ajouter davantage de données à notre dataset en utilisant les données publiques de Météo France. Ces données ont pour but de rendre notre modèle prédictif plus précis.

Suivant une spécialisation Big Data, Alexandre et Maxime ont alors commencé par importer et nettoyer les données Météo France. Puis, ensemble, nous avons pu entraîner les données météo en créant un nouveau dataset pour avoir des mesures toutes les 3 heures sur 2 ans. Suite à ça, nous avons à nouveau entraîné notre modèle prédictif pour la ruche et les résultats obtenus sont alors bien plus intéressants que ce que nous avons eu précédemment.

	Assigned	Progress	DECEMBER 2019																
▼ 10 DEC : Modèle fonctionnel et optimisé		88%	2	3	4	5	6	9	10	11	12	13	16	17	18	19	20		
Modèle fonctionnel et optimisé		<input checked="" type="checkbox"/>																	
Récupérer toutes les données de la ruche en temps réel	Alexandre Verept, mohamed	100%						Alexandre Verept, mohamed											
Ajout d'une page web au site pour afficher les résultats	mohamed	100%						mohamed											
Amélioration du modèle	Alexandre Verept, Maxir	100%						Alexandre Verept, Maxime THOOR											
Mettre sur un serveur pour faire la prédiction en temps réel	Alexandre Verept, mohamed	50%						Alexandre Verept, mohamed											

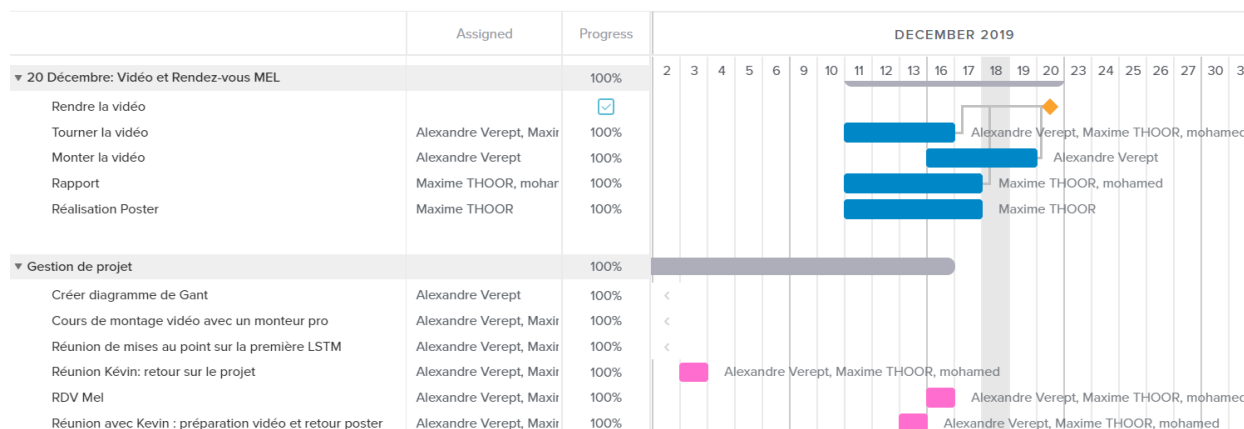
Une fois l'objectif de faire tourner un modèle optimal et fonctionnel, nous nous sommes chacun donné une tâche à réaliser. Maxime s'est occupé de voir la différence de température et d'humidité entre les données de la ruche et celle de Météo France. Quant à Alexandre et Mohamed, ils se sont mis tous les deux sur la base de données et sur l'aspect serveur. Cependant, nous n'avons guère réussi à mettre notre prédiction en temps réel sur un serveur car nous avons manqué de temps. En effet, il nous fallait tout installer et faire toutes les configurations sur un nouveau serveur. Toutefois, pendant les vacances, nous tenterons de nous renseigner sur le problème pour faire afficher notre prédiction en mettant nos données sur Google Cloud ou AWS par exemple.

Ensuite, dans l'objectif de préparer le rendez-vous avec la MEL fixé au lundi 16 décembre, Maxime a commencé à réaliser le poster pendant qu'Alexandre et Mohamed s'occupait de la gestion du site afin d'avoir une page internet où l'on pouvait afficher nos prédictions. Chacun a pu améliorer et donner son avis sur le poster afin de le rendre visuellement beau et avec un contenu de qualité.

Au début de la dernière semaine de projet, nous avons donc eu ce rendez-vous avec 2 personnes de la Métropole Européenne de Lille qui ont été très intéressés par notre projet. Ce rendez-vous, à

l'initiative de notre tuteur, a été une réelle satisfaction pour l'ensemble du groupe puisqu'elle nous a permis de vivre une nouvelle expérience professionnelle.

Le lendemain matin, nous avons commencé à tourner et à réaliser la vidéo ainsi que la rédaction du rapport. Celle-ci s'est faite de façon équitable puisque nous avons chacun traité une des 3 parties. Toutefois, nous avons également chacun relu les parties des 2 autres membres du groupe afin de compléter ou de modifier certains aspects qui nous semblaient plus cohérents (toujours en discussion avec l'ensemble du groupe).



## Conclusion

Ce semestre, nous avons tous les 3 choisis le projet de la prédiction de la pollution de l'air à Lille sans vraiment véritablement savoir ce qui nous attendait. Après la première semaine de projet, nous avons chacun vu les perspectives de ce projet et les réels bénéfices que chacun de nous pouvaient en tirer.

En respectant les deadlines, nous avons également appris davantage à maîtriser les aspects essentiels d'un projet. Ces deadlines ayant été fixé dès le début du projet, cela nous a permis de toujours nous situer quant à l'avancée de notre modèle prédictif. Il a donc fallu faire preuve de flexibilité et de persévérance face aux difficultés rencontrées.

Sur l'aspect technique, le défi de prédire l'indice de qualité de l'air a été une réelle source de motivation. En effet, dans l'ère de l'Intelligence Artificielle, apprendre à utiliser un réseau de neurones récurrent comme la LSTM a été une opportunité pour nous de mettre à profit nos compétences acquises lors des cours du premier semestre. De plus, la Big Data, domaine de spécialisation de Alexandre et Maxime, a fait sa place dans notre projet puisqu'il a fallu constamment s'appropriier nos datasets pour l'utiliser dans notre modèle de Machine Learning. Cette plurivalence a notamment pu être également mise à profit dans l'attribution des tâches et donc de la gestion de projet.

Somme toute, nous avons retrouvé lors de ces 5 semaines de projet, les compétences, les contraintes mais aussi l'excitation d'un projet. Ce dernier aspect s'est notamment révélé lors du rendez-vous avec la Métropole Européenne de Lille. En effet, cela a permis de connaître une nouvelle expérience professionnelle encore méconnue pour nous 3.

L'objectif final du projet ayant été atteint, nous nous sommes demandé de manière générale et au cours du rendez-vous s'il était possible de prédire l'indice de qualité de l'air non pas uniquement du jour d'après mais aussi à J+2 voir J+3.

## Remerciement

Dans un premier temps, nous aimerions remercier Kévin Hérissé, notre tuteur de projet, pour son temps, son implication et son aide au cours du projet. Il nous a permis d'aller toujours chercher à obtenir les meilleurs résultats possibles et il a amené à travers sa bonne humeur une bonne ambiance entre nous 4.

Nous souhaiterions remercier également la MEL avec qui nous avons pu avoir la chance d'échanger sur les perspectives d'amélioration lors de la présentation de notre projet. Leur intérêt pour notre projet a été fort appréciable.