

# Données sur le web

---

Open Data



UNIVERSITÉ DE NANTES

Charrier Romain

Ouzamou Adil

Hamon Cyril

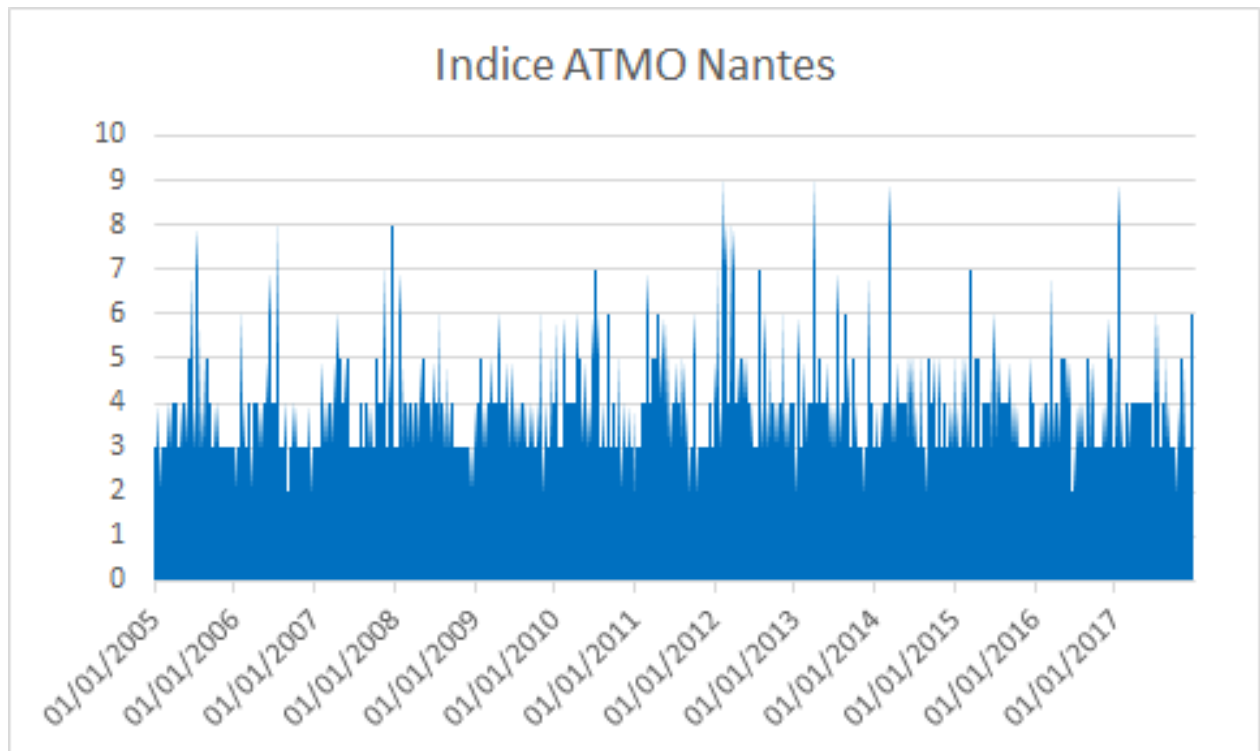
# Analyse des données

Dans un temps où le climat est un enjeu important, nous avons souhaité traiter le sujet de la pollution avec la thématique suivante: Comment Nantes et Angers essaient de diminuer la pollution dans leur ville en la rendant plus accessible à vélo ? Dans un premier temps, nous allons évaluer les différentes données fournis pour Nantes et Angers. Dans un second temps, nous allons comparer les données fournis par la région avec celle de Paris. Pour finir nous vous parlerons de l'exploitation de ses données au travers des applications.

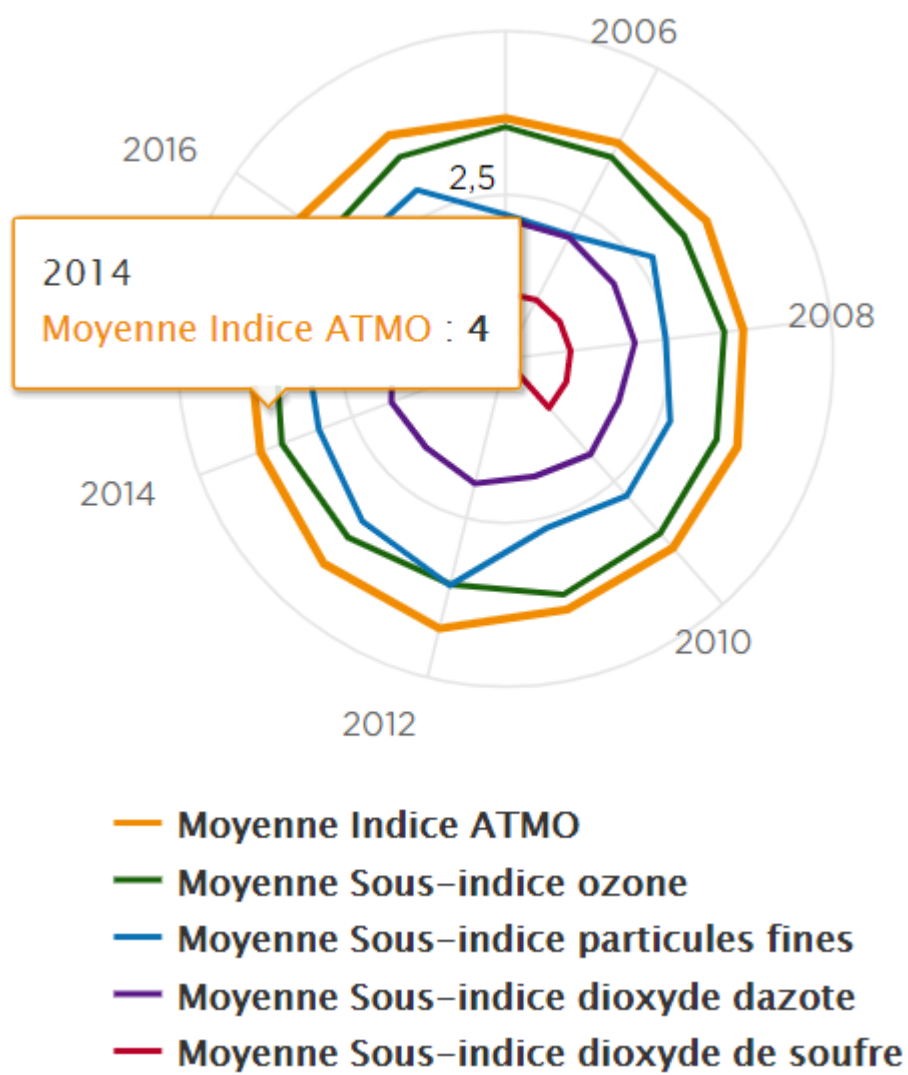
Sur la région Pays de la Loire, nous avons trouvé 2 jeux de données intéressant concernant la pollution dans la ville de Nantes et d'Angers. Elles sont basées sur l'indice ATMO. L'ATMO est l'indicateur global de la qualité de l'air. Il est calculé pour les agglomérations de plus de 50 000 habitants. Cet indice synthétique est calculé à partir de la concentration dans l'air ambiant de quatre polluants, surveillés en continu par l'ensemble des associations agréées de surveillance de la qualité de l'air (AASQA): le NO<sub>2</sub>, le SO<sub>2</sub>, les poussières d'origine industrielle et automobile, l'O<sub>3</sub> (ozone d'origine photochimique). Peuvent être également mesuré, le plomb, les composés organiques volatils.

Nous sommes allé chercher les informations de pollution sur trois sites différents qui sont <https://data.nantesmetropole.fr/pages/home/>, <https://data.angers.fr/pages/home/> et <http://www.airpl.org/Air-exterieur/indice-de-pollution>. Les données qui sont mises à disposition sur les différents sites ont plusieurs formats. On peut y retrouver des formats plus classique comme le CSV, le JSON ou Excel, mais aussi des formats plus spécifique qui se base sur des fichiers géographiques comme GeoJson, Shapefile, KML.

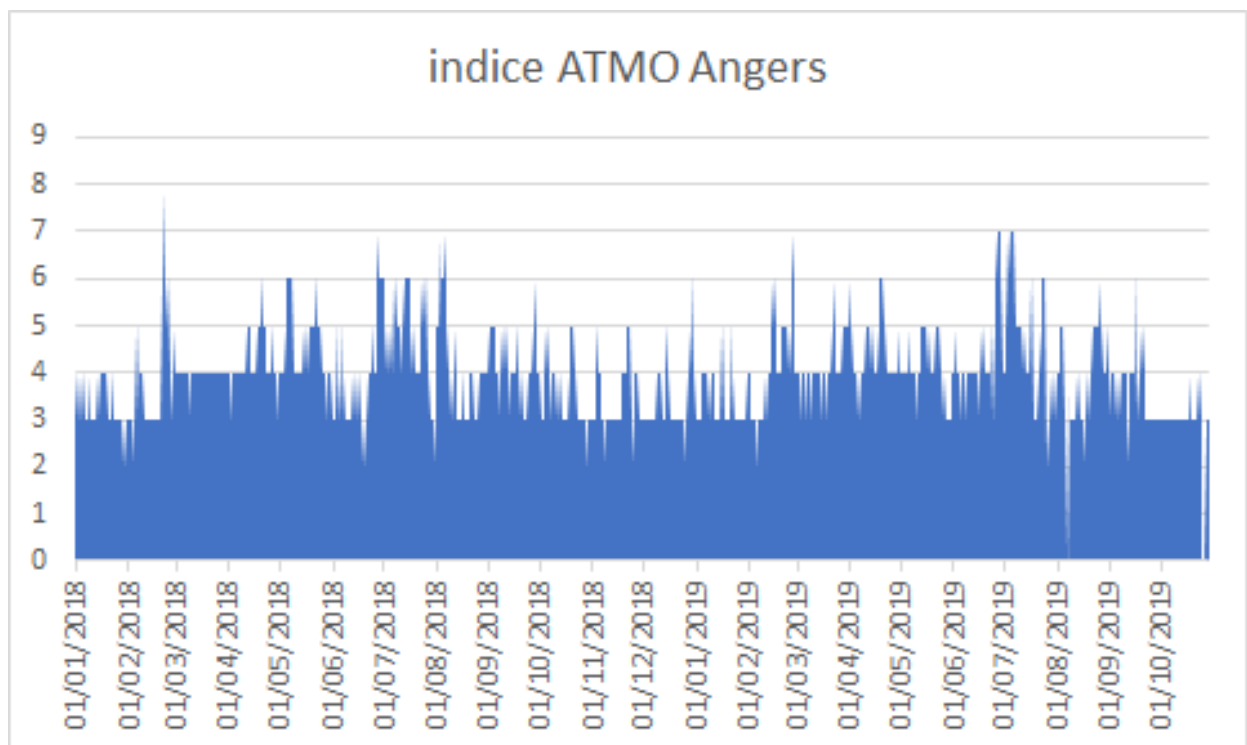
Les données concernant la pollution à Angers sont complètes, ils ont commencé à mettre à disposition leurs données à partir du 1 janvier 2018. Depuis, elles sont tenues à jour et nous avons même du prévisionnel disponible à +1 jours. À côté, Nantes met à disposition ces données depuis plus longtemps, car elles sont disponibles depuis 2005. Mais le problème, c'est qu'elles s'arrêtent en 2017 et depuis nous n'avons plus rien. Pour analyser dans le temps la pollution à Nantes, les données fournies sont très intéressantes, car nous pouvons faire des moyennes sur les années et évaluer l'évolution.



Grâce aux données fournis, nous avons la possibilité d'obtenir le schéma suivant. En abscisse, nous avons les années écoulées depuis le 1 janvier 2005 jusqu'au 31 décembre 2017. En ordonnée, nous avons l'échelle ATMO de 0 à 10 où 1 représente un air très sain et 10 une très forte pollution atmosphérique. Ici, à en juger le schéma, nous pouvons constater une très forte pollution sur l'année 2012 et 2013. De très fort pics montant jusqu'à 9 et une pollution constante assez élevée. À la différence de l'année 2005 où nous pouvons constater, que la pollution était nettement plus faible. Avoir un historique aussi fourni permet de voir l'évolution dans le temps et de comparer nos valeurs avec de l'intérêt. Savoir qu'en 2016 on polluait moins qu'en 2017, c'est intéressant mais peu utile alors que si nous pouvons comparer sur différentes années et une plage de 12 ans, nous pouvons démontrer et affirmer plus facilement les vérités. Nous avons aussi des informations plus précise concernant l'indice ATMO. Comme nous l'avons précisé, ATMO est composé d'un calcul de plusieurs éléments. Sur le site Open data Nantes, nous pouvons analyser ces différentes données au travers de graphique qui sont proposés.



Pour Angers, nous sommes capables de faire le même schéma concernant l'indice ATMO.



La différence viendra du fait que les valeurs sont sur une plage de données plus petite et une étude dans le temps serait donc plus compliquée. Ce qui peut être intéressant avec les données d'Angers, c'est d'exploiter le prévisionnel.

Dans cette première partie d'analyse, nous pouvons constater que si Nantes et Angers mettait à disposition de la même manière leurs données, nous pourrions avoir quelque chose de complet et abouti. Dans un sens, de l'historisation de données qui permettrait de comparer l'évolution dans le temps et de juger pourquoi cette pollution est de plus en plus conséquente. Ensuite, plus de données prévisionnelles, ce qui permettrait de constater une aire peu saine et pouvoir agir avant.

Nous avons choisi de mettre en corrélation les données concernant la pollution avec celle des pistes cyclables. Nous allons donc analyser maintenant ce qui existe comme données concernant les pistes cyclables, puis nous expliquerons la corrélation que nous avons faite entre ces deux jeux de données.

Dans un premier temps, nous sommes allés chercher ce qui existait du côté d'open data Nantes. La première constatation vient du nombre de données disponible concernant les vélos et les pistes cyclables à Nantes. Il y a énormément de données qui quelquefois sont identiques, pas à jour, ou complètement dépassées et inutilisables aujourd'hui. On peut en trouver sous différents formats comme précédemment, mais aussi au travers de fichier Word qui sont inexploitable. Un premier exemple de ce qu'on peut trouver. Les stations de vélos en libre-service à Nantes. On trouve deux jeux de données, l'une est mise en production par la ville de Nantes et l'autre par JCDecaux.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
13	55 #00055-MAG 11, rue de Belfort		47.21078140	False	False	OPEN	nantes	15	10	5	2019-10-28 15:31:04+01:00				
14	114 #00114 - ATC 114 - ATOUTSUD ANGLE RUE DE L'ABBE GREGOIRE/AV MARE		47.194493, -1	True	False	OPEN	nantes	25	15	9	2019-10-28 15:32:28+01:00				
15	23 #00023-BUA1118, rue de Coulmiers		47.22714546	False	False	OPEN	nantes	11	11	0	2019-10-28 15:30:58+01:00				
16	5 #00005-BROS 12, Allée Duquesne		47.21883723	False	False	OPEN	nantes	18	13	4	2019-10-28 15:32:49+01:00				
17	10 #00010- PICA 12, mail Pablo Picasso		47.216207, -1	False	False	OPEN	nantes	40	8	30	2019-10-28 15:32:44+01:00				
18	121 #00121-VICT 12, place Victor Richard		47.228614, -1	False	False	OPEN	nantes	15	5	10	2019-10-28 15:34:23+01:00				
19	82 #00082-SEBIL 120, rue Dr Jules Sébilleau		47.20675751	True	False	OPEN	nantes	16	13	3	2019-10-28 15:33:14+01:00				
20	72 #00072-BELL 138, rue Paul Bellamy		47.22862569	False	False	OPEN	nantes	15	8	7	2019-10-28 15:33:30+01:00				
21	75 #00075-LAM 14 bis, rue Lamoricière - Place Beaumanoir		47.21230895	False	False	OPEN	nantes	18	15	3	2019-10-28 15:31:56+01:00				
22	79 #00079-MAL 15, quai de Malakoff		47.213395, -1	False	False	OPEN	nantes	15	7	8	2019-10-28 15:31:59+01:00				

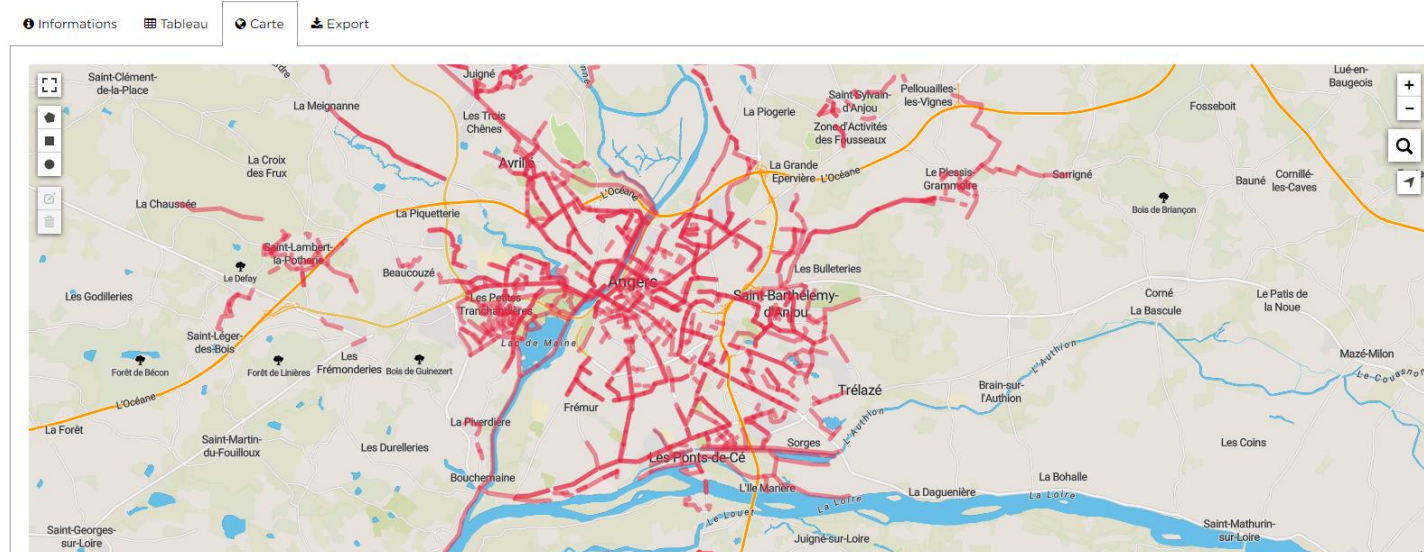
  

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
73	Station biclo Avenue Chanzy	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2254298857, -1.54799992601					
74	Station biclo Rue Dufour	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2240678792, -1.54521243477					
75	Station biclo Rue de Talensac	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			16	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2211433295, -1.55647633962					
76	Station biclo Boulevard du Petit Port	44300	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			20	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2432639146, -1.55634461051					
77	Station biclo Rue Théodore Brosseaud	44400	44143	Rezé	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.1886839996, -1.5672010003					
78	Station biclo Rue de Coulmiers	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			11	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2271454685, -1.54240474083					
79	Station biclo Rue Gaëtan Rondeau	44200	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			20	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2053961328, -1.53758186289					
80	Station biclo Rue Gaston Veil	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			27	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2114650645, -1.55682453761					
81	Station biclo Rue Paul Bellamy	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2286256903, -1.56180639272					
82	Station biclo Boulevard Allard	44100	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.211501071, -1.57717541492					
83	Station biclo Rue Guibourd de Luzinais	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			15	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.2217851832, -1.54709923329					
84	Station biclo Allée Brancas	44000	44109	Nantes	Equipement biclooPlus	ABONNEMENT JCDecaux			45	7 jours sur 7, 24 h sur 24.		bicloo.nante 01 30 79 33 4 47.213210552, -1.55739323575					

En allant se balader un peu dans les données, on peut remarquer que la majorité des stations référencées dans un fichier l'est aussi chez l'autre. La différence entre ses deux fichiers va être les colonnes d'informations supplémentaires par exemple dans le fichier de la ville de Nantes sur les conditions sur la catégorie de véhicule et l'ouverture alors que pour JCDecaux, on saura en plus si la station est toujours ouverte ou si elle a été fermée. Dans le fichier de Nantes, ils nomment comme source JCDecaux. Ils sont donc conscients que des données sont déjà mis à disposition et ils viennent en rajouter par-dessus avec un nouveau fichier. Le problème de cela c'est qu'il est plus difficile de comparer deux fichiers plutôt qu'un seul. Ici, les adresses n'ont pas la même précision, dans un cas, ils nomment les numéros et rues précises alors que dans l'autre, ils nomment les stations. On se retrouve donc à essayer de comparer deux jeux quasiment identiques mais incomparables. On est surchargé de données quasiment identique, ce qui pourrait être nettement réduit. Nous avons pris cet exemple, mais il en existe plusieurs dans ce cas.

Pour Nantes et Angers, nous trouvons pour chacun un jeu de données permettant de faire référence au infrastructure de piste cyclable.

#### Infrastructures cyclables - Angers Loire Métropole



Les données fournis peuvent être exploité sous le format d'une carte car dans les colonnes de la table on y trouve des coordonnées géographique. Cette carte permet de représenter les différentes pistes cyclables à Angers. Nous avons un jeu identique pour la ville de Nantes.

## Corrélation entre les deux jeux de données

Au cours d'une journée, 15 000 litres d'air transitent en moyenne par nos voies respiratoires. Or, nous vivons à une époque où la qualité de l'air s'est largement dégradée notamment dans les grandes villes. La principale cause de cette détérioration est le transport terrestre. Cette pollution de l'air a de nombreuses conséquences notamment sur notre santé. En effet, 5% des décès annuels mondiaux est sont dû à des complications liées à une exposition à un air pollué. Il est donc nécessaire de changer nos habitudes et de privilégier des types de transport plus sain comme par exemple le vélo. C'est ici que notre problématique rentre en jeu. Aujourd'hui, les grandes villes de France essayent de réduire un maximum la circulation en leur sein. Pour cela, ils créent de nombreux aménagements cyclables pour données un maximum d'accessibilité. Résoudre le problème de pollution en ville passe par le fait de proposer des alternatives aux voitures, car elles sont la principale cause.

## Comparaison avec Paris

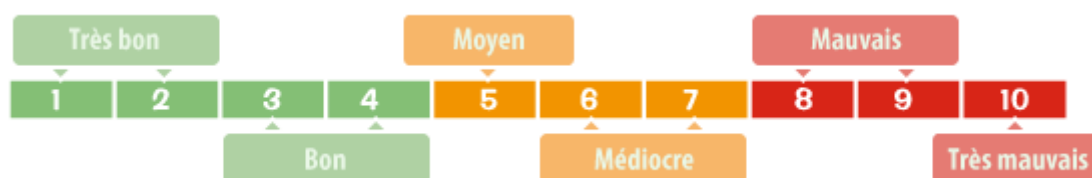
Il nous semblait intéressant de comparer les informations mises à disposition par les villes de Nantes et Angers concernant la qualité de l'air et le développement d'infrastructures de cyclisme (stations, pistes...) avec celles de l'agglomération parisienne, la première ville française ayant mis à disposition un site d'open data.

Concernant la pollution de l'air, la ville de Paris proposent de nombreux jeux de données sur le site <https://data-airparif-asso.opendata.arcgis.com/>. Les données sur la qualité de l'air sont mis à jour quotidiennement et des prévisions peuvent être déduites à partir de ces données. De plus, ces données sont récoltées depuis 2002 et sont encore mises à jour à l'heure actuelle.

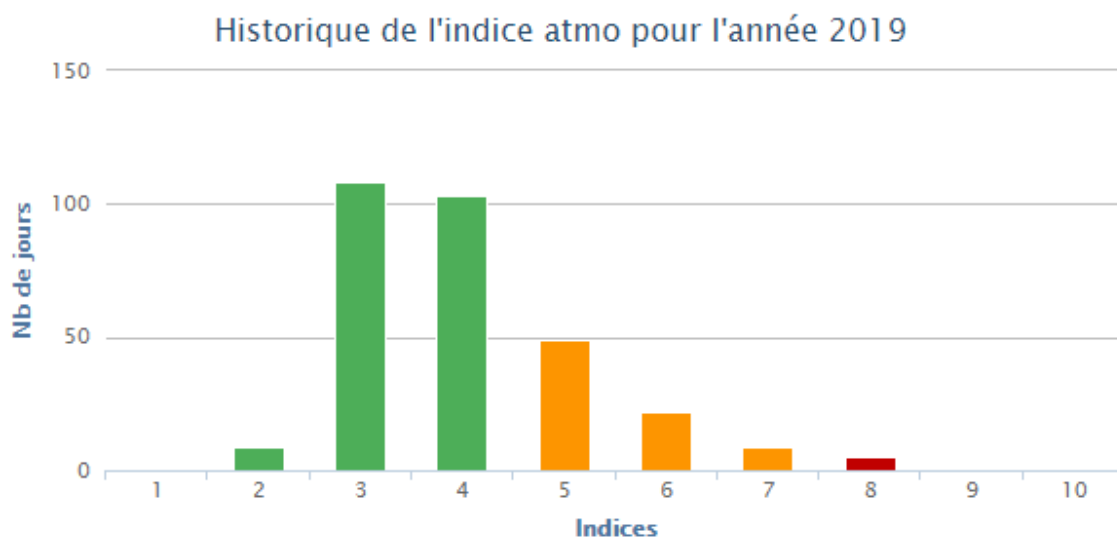
FID	valeur	source	type_zone	code_zone	lib_zone	val_no2	val_so2	val_o3	val_pm10	val_pm25	couleur	x_lamb93	y_lamb93	date_ech	qualif
91	4	Airparif	UU	00851	Unité urbaine de Paris	1	1	4	3	0	#84BF75	663550	6845563	08/07/2019 à 00:00	Bon
92	5	Airparif	UU	00851	Unité urbaine de Paris	1	1	5	3	0	#F29400	663550	6845563	07/07/2019 à 00:00	Moyen
93	7	Airparif	UU	00851	Unité urbaine de Paris	2	1	7	4	0	#F29400	663550	6845563	06/07/2019 à 00:00	Médiocre
94	5	Airparif	UU	00851	Unité urbaine de Paris	3	1	5	4	0	#F29400	663550	6845563	05/07/2019 à 00:00	Moyen
95	5	Airparif	UU	00851	Unité urbaine de Paris	2	1	5	3	0	#F29400	663550	6845563	04/07/2019 à 00:00	Moyen
96	5	Airparif	UU	00851	Unité urbaine de Paris	1	1	5	3	0	#F29400	663550	6845563	03/07/2019 à 00:00	Moyen
97	4	Airparif	UU	00851	Unité urbaine de Paris	2	1	4	3	0	#84BF75	663550	6845563	02/07/2019 à 00:00	Bon
98	4	Airparif	UU	00851	Unité urbaine de Paris	1	1	4	3	0	#84BF75	663550	6845563	01/07/2019 à 00:00	Bon
99	5	Airparif	UU	00851	Unité urbaine de Paris	2	1	5	4	0	#F29400	663550	6845563	30/06/2019 à 00:00	Moyen
100	7	Airparif	UU	00851	Unité urbaine de Paris	3	1	7	3	0	#F29400	663550	6845563	29/06/2019 à 00:00	Médiocre

En réunissant les données sur les polluants, on obtient ce jeu de données qui calcule la valeur de l'indice ATMO à un endroit donné et lui associe une qualification (Bon, Médiocre...).

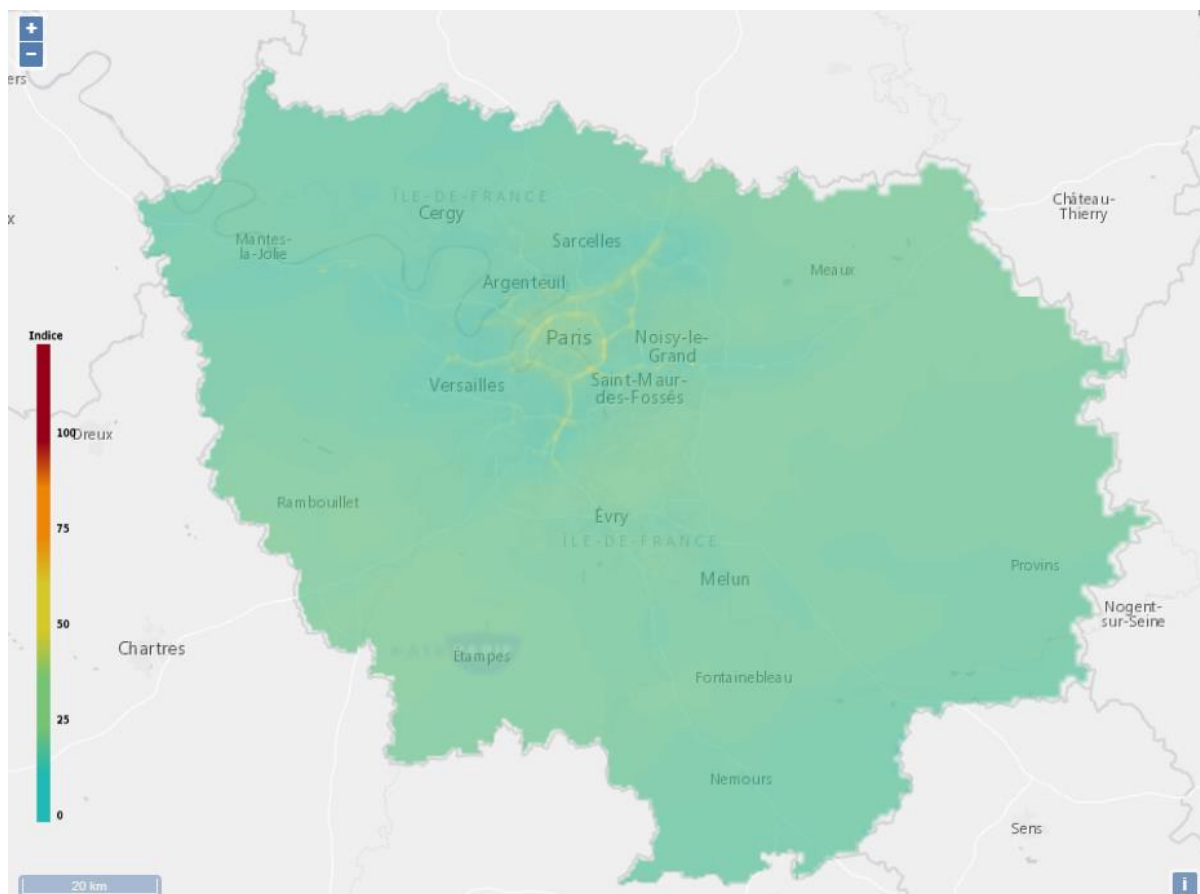
Mais ces données ne sont pas très parlantes sous forme de tableau, pour obtenir des modélisation sous forme de graphique ou de carte, il faut se rendre sur le site <https://www.airparif.asso.fr>. On peut alors trouver des graphiques qui modélisent l'indice ATMO pour chaque jour de l'année :







On peut alors en déduire les pics de pollution de l'année et comparer leur évolution par rapport aux années précédentes. De plus, en identifiant ces jours d'épisodes de pollution, on peut déterminer, à partir d'autres jeux de données donnant plus d'informations sur les polluants à partir desquels est calculé l'ATMO, la cause précises de cette pollution. Malheureusement, contrairement aux villes de Nantes et d'Angers, nous n'avons pas trouvé de graphiques comparant l'évolution de cet indice sur plusieurs années.



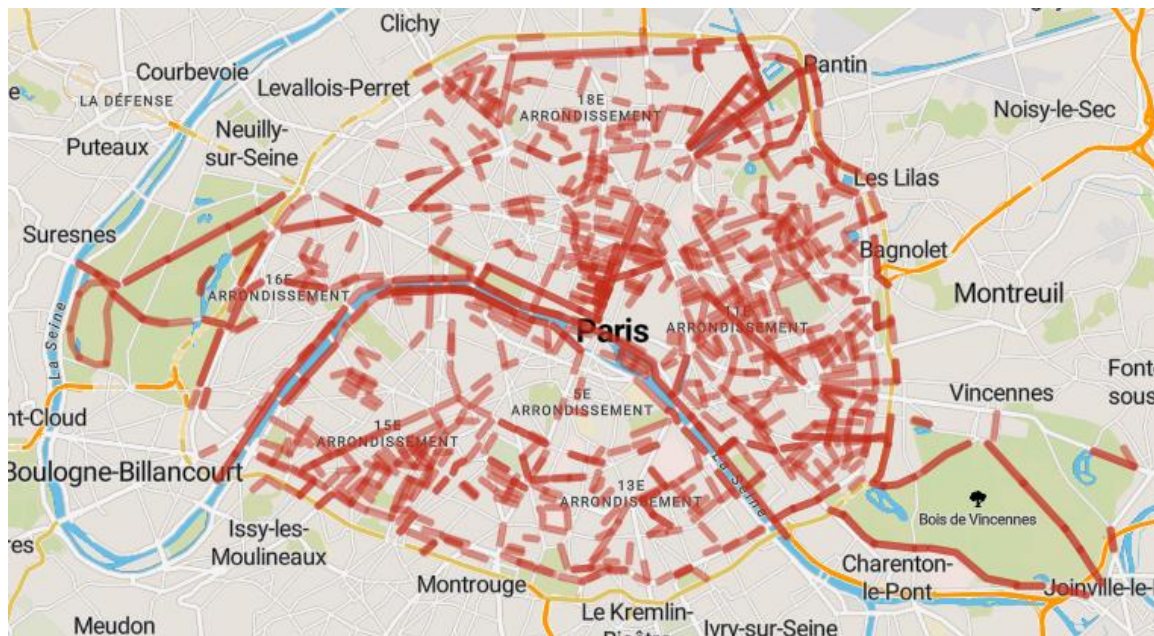
La ville de Paris possède aussi de nombreux points de relevés de la qualité de l'air, bien plus que Nantes et Angers. Cela permet d'établir avec précision une carte de la qualité de l'air.

En ce qui concerne les données sur les vélos, on trouve les différents jeux de données sur <https://opendata.paris.fr>. De la même manière que pour Nantes et Angers, on trouve de nombreux jeux de données sur la mobilité à vélo notamment sur la disponibilité des vélos en libre service ou bien l'emplacement des stations. Ces jeux de données sont disponibles au format de fichiers plats (CSV, JSON, Excel) mais aussi au format de fichiers géographiques (GeoJSON, Shapefile et KML).

Codé de la station	Nom de la station	Etat des stations	Etat du Totem	Nombres de bornes en station	Nombre de bornes disponibles	Nombre de vélo mécanique	Nombre vélo électrique	Achat possible en station (CB)	PARK + activation	Nombre vélo en PARK+
15107	Benjamin Godard - Victor Hugo	Operative	yes	35	27	2	6	no	no	0
6015	André Malraux - Saint-André des Arts	Operative	yes	32	30	10	12	yes	no	0
11037	Faubourg Du Temple - République	Operative	yes	38	32	2	4	yes	no	0
9020	Troubouze - Clauzel	Operative	yes	21	19	0	2	yes	no	0
11104	Charonne - Robert et Sonia Delaunay	Operative	yes	20	18	1	1	no	no	0
12109	Marie du 12ème	Operative	yes	30	16	3	11	no	no	0
7003	Square Boucicaut	Operative	yes	60	20	37	2	yes	no	0
5001	Harpe - Saint-Germain	Operative	yes	1	1	6	7	yes	no	0
14014	Jourdan - Stade Charlety	Operative	yes	60	39	12	8	yes	no	0
17026	Jouffroy d'Abbans - Wagram	Operative	yes	0	0	1	4	yes	no	0
17041	Guerrant - Gouvion-Saint-Cyr	Operative	yes	38	24	11	4	yes	no	0
10013	Albaret - Jemmapes	Operative	yes	59	39	11	7	yes	no	0
9104	Caumartin - Provence	Operative	yes	22	8	7	7	yes	no	0
14111	Cassini - Denfert-Rochereau	Operative	yes	25	19	0	6	yes	no	0
6003	Saint-Gulpica	Operative	yes	21	14	4	3	yes	no	0
13007	Le Brun - Gobelin	Operative	yes	48	35	6	7	yes	no	0
5110	Lacépède - Monge	Operative	yes	23	21	0	1	yes	no	0
9108	Saint-Roman - Chéche-Midi	Operative	yes	17	7	4	6	no	no	0
33006	André Krumm - République	Operative	yes	31	25	3	3	no	no	0
42916	Pierre et Marie Curie - Maurice Thorez	Operative	yes	27	14	0	13	no	no	0

Contrairement aux données de la ville de Nantes, il y a beaucoup moins d'informations dupliquées. En effet, les informations sur les vélos en libre service sont principalement réparties sur deux jeux de données ne possédant qu'une seule information en commun: ceux

sur la disponibilité des vélos en libre-service (voir ci-dessus) et ceux sur l'emplacement des stations. Sur le fichier des disponibilités on pourra retrouver, comme pour la ville de Nantes si une station est ouverte ou fermée mais aussi des informations supplémentaires comme le type de vélo mis à disposition (électrique ou mécanique) ou bien si on peut payer en CB. Cependant, on ne pourra pas avoir l'adresse d'une station directement depuis le fichier des disponibilités, il faudra se rendre dans le fichier de l'emplacement des stations où l'adresse n'est pas fournie explicitement, seulement ses coordonnées géographiques.



De la même manière que pour Nantes et Angers, on trouve un jeu de données référençant le réseau des itinéraires cyclables. Encore une fois, on trouve des informations supplémentaires dans le fichier de Paris comme par exemple le régime de vitesse (zone 30 ou voie 50) ou bien si la piste cyclable se situe au niveau du trottoir ou sur un couloir de bus.

En conclusion, la ville de Paris met à disposition des jeux de données plus fournis tout en évitant au maximum la duplication des données. En effet, toutes les données sur la qualité de l'air sont disponibles sur <https://www.airparif.asso.fr> mais ne sont pas dupliquées sur le site principal <https://opendata.paris.fr>. Cela peut aussi s'avérer un peu problématique dans la mesure où l'information n'est pas réunie au même endroit.

## Application

Il existe de nombreuses applications mesurant la qualité de l'air comme par exemple Plume une application permettant d'avoir en temps réel des données sur le niveau de pollution. L'application compare aussi les relevés avec la moyenne annuelle pour identifier les jours de pics de pollution. Plume va aussi donner des recommandations aux utilisateurs comme indiquer à quelle heure il est le plus sain de faire du vélo ou de s'asseoir en terrasse. De plus, un utilisateur peut renseigner son niveau de sensibilité à la pollution pour avoir des notifications personnalisés en cas de pics de pollution.

Ou encore l'application Itiner'AIR qui se charge de mesurer la qualité de l'air en Ile-de-France en exploitant les données fournis sur le site <https://www.airparif.asso.fr>. Itiner'AIR permet aussi d'aménager son itinéraire pour moins subir la pollution. Mais cette dernière possède de nombreux bugs, ne possède pas encore les informations de tous les polluants comme les PM2.5 (microparticules) et possède une interface de navigation peu intuitive.

### Maquette de l'application :



## Avis global de l'open data

Lors de ce rapport, nous avons pu étudier différents points concernant l'open data et nous avons soulevé certaines contraintes, mais aussi des points plus positifs.

En effet, après les recherches que nous avons effectué sur l'open data nous avons pu noter plusieurs point afin de critiquer cette démarche :

Tout d'abord l'approche Open data est une formidable mine d'information qui reste malheureusement encore largement sous exploitée par rapport au potentiel que peuvent apporter toutes ces information. En effet, on peut constater que la majorité des données que l'on peut trouver sur les sites spécialisés dépasse rarement le millier de téléchargements. On peut certainement expliquer cela grâce à plusieurs point :

Nous avons pu constater que les données pouvaient être mises en doublons avec quelquefois des différences entre les jeux de données. A cause de ça, les données sont difficilement exploitables car même si les données sont quasiment identiques, il n'y a pas de ressemblance permettant de faire un lien clair entre deux données. Dans ce cas, il faut tout regarder à la main et donc elles sont quasiment inexploitable car trop de perte de temps.

De plus, nous trouvons des données qui ne sont plus tenues à jour alors que leur objectif est d'avoir les informations en temps réel. Sont-elles toujours utiles dans ce cas ? Ne serait-il pas plus simple de les supprimer plutôt que de consommer de l'espace de stockage. Par exemple l'indice de pollution ATMO n'est plus mis à jour sur le site open data de Nantes métropole depuis maintenant plus de 1 an ce qui est un vrai problème lorsque l'on cherche, comme dans notre exemple d'application, à tenir une application à jour sur la pollution à Nantes. De plus, toujours dans le cadre de notre exemple, certains champs ne sont pas remplis pendant plusieurs mois voire années, comme par exemple le niveau d'alerte par couleur, ce qui est assez embêtant, car c'est certainement l'indicateur le plus compréhensible pour le grand public.

On remarque aussi que les données ne sont pas toujours correctement remplies et son utilisation en est quelquefois difficiles. Les caractères spéciaux, les espaces, les points, il n'y a pas de normes dans la façon de transmettre ses données. L'utilisateur doit être capable de s'adapter, mais il est très souvent difficile d'exploiter quelque chose d'intéressant.

Des points positifs sont tout de même à soulever. D'abord, la mise à disposition des données. Aujourd'hui, il est possible à n'importe qui de lire des données concernant sa ville dans un fichier Excel. Il y a une accessibilité qui rend la donnée intéressante au gens.

Enfin, pour conclure cette critique de l'open data, on peut aussi constater que certains supports d'information comme les API par exemple sont totalement inexploitable pour des personnes ne connaissant pas l'informatique ce qui limite encore plus la diffusion en masse d'information mis à disposition des citoyens. De plus, l'open data permet une transparence des entreprises, ville et tout autre entité envers ses utilisateurs et d'un point de vue citoyen, la transparence est un élément clé pour pouvoir forger une opinion construite.