UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE MATHEMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGENIEUR

Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) - UMR7357

THÈSE présentée par :

Xavier Jurado

soutenue le : **17 Décembre 2021**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Informatique

Atmospheric pollutant dispersion estimation at the scale of the neighborhood using Sensors, Numerical and Deep Learning models

THÈSE dirigée par :

Pr. WEMMERT Cédric	Professeur, ICube, Université de Strasbourg, France.
Pr. VAZQUEZ José	Professeur, ENGEES, Université de Strasbourg, France.

RAPPORTEURS :

Dr. MARTIN Fernando	Directeur de recherche, CIEMAT, Madrid, Espagne.
Pr. BENNIS ZEITOUNI Karine	Professeur, DAVID, Universté de Versailles St-Quentin, Versailles, France.

AUTRES MEMBRES DU JURY :

Pr. HOARAU Yannick	Professeur, ICUBE, Université de Strasbourg, France.
Dr. LEBBAH Mustapha	Maitre de conférence, LIPN, Université Sorbonne Paris Nord, Paris, France.

Abstract

This thesis is at the crossroad of four domains, computational fluid dynamics (CFD), data mining, deep learning and air quality. The objective of the thesis is to assess dwellers' exposures to atmospheric pollutants using the recent advances in artificial intelligence. The thesis revolves around the different time scales requested by the regulations, going from annual to real time. To do so, innovative approaches to assess mean annual concentrations were developed for modeling as well as for sensors. For modeling a statistical methodology based on wind roses frequencies associated with a flowchart to determine the numerical error from the discretization was proposed. For the sensors, data from all around France were analyzed to establish relationship between measured monthly concentrations with annual ones for particulate matter and nitrogen oxides. To determine pollution exposure in real time, a system using taking into account traffic, meteorological and 3D building layout was created built around a deep learning model was created. The system revolves around a deep learning model. This model, multiResUnet, have been chosen after comparison with other classical state-of-the-art convolutional models and optimized for the dispersion pollution issue. To train it, examples from CFD were generated efficiently following guidelines developed in this thesis. The system was then applied on a real neighborhood of $1km^2$ with real traffic data and compared with CFD. It managed to perform well on classical air quality metrics and reach a J_{3D} score of 62%.

Keywords : Deep Learning, Convolutional Neural Networks, Computational Fluid Dynamics, Data mining, Air quality, Sensors data analysis, Urban environment

Résumé

La présente thèse est à la croisée de quatre domaines, la mécanique des fluides numériques (CFD), le data mining, le deep learning et la qualité de l'air. L'objectif de la thèse est d'évaluer l'exposition des habitants aux polluants atmosphériques en utilisant les récentes avancées en intelligence artificielle. La thèse s'articule autour des différentes échelles de temps demandées par la réglementation, allant de l'annuel au temps réel. Pour ce faire, des approches innovantes pour évaluer les concentrations annuelles moyennes ont été développées pour les outils de modélisation ainsi que pour les capteurs de pollution de l'air. Pour la modélisation, une méthodologie statistique basée sur les fréquences des roses des vents associées à un organigramme permettant de déterminer l'erreur numérique due à la discrétisation a été proposée. Pour les capteurs, des données provenant de toute la France ont été analysées pour établir la relation entre les concentrations mensuelles mesurées et les concentrations annuelles pour les particules fines et les oxydes d'azote. Pour déterminer l'exposition à la pollution en temps réel, un système prenant en compte le trafic, la météorologie et la disposition des bâtiments a été créé avec en son coeur un modèle d'apprentissage profond. Le système s'articule autour d'un modèle d'apprentissage profond. Ce modèle, multiResUnet, a été choisi après comparaison avec d'autres modèles convolutifs classiques de l'état de l'art et optimisé pour la problématique de la dispersion de polluant. Pour l'entraîner, des exemples issus de la CFD ont été générés efficacement en suivant des principes développés dans cette thèse. Le système a ensuite été appliqué sur un quartier réel de $1km^2$ avec des données de trafic réels et comparé à la CFD. Il a réussi à obtenir de bonnes performances sur les mesures classiques de la qualité de l'air et à atteindre un score de similarité J_{3D} de 62%.

Mots clés : Apprentissage profond, Réseaux neuronaux convolutifs, Dynamique des fluides numérique, Extraction de données, Qualité de l'air, Analyse des données de capteurs, Environnement urbain

Contents

1 Introduction **9**

1.1 Air pollutant sources and effects 9

1.2 Air pollution regulations 12

1.2.1 Air pollution regulation in the World 12

1.2.2 Air pollution regulation in Europe 13

1.2.3 Air pollution regulation in France 14

1.2.4 Challenges 14

1.3 Dissertation content 16

1.3.1 Plan of the document 16

1.3.2 Scientific contributions 18

1.3.3 Context of the thesis 19

I Annual urban pollution assessment at the scale of the neighborhood using modelling and sensors **21**

2 Atmospheric pollution modelling **22**

2.1 Atmospheric pollutant scales 23

2.2 Atmospheric pollutant assessment for urban areas 25

2.3 Conclusion 27

3 How to assess the mean annual air pollution at the scale of a neighbourhood? **28**

3.1 Introduction 29

3.2 Material and methods 30

3.2.1 Meteorological data 30

3.2.2 Numerical model 33

3.3 Results 35

3.3.1 Wind data interpolation 35

3.3.2 Mean annual concentration assessment 39

3.4 Discussion . . . 44
3.5 Conclusion . . . 45

4 How to balance between modelling error and computational cost to assess mean annual concentration of a neighbourhood? 47
4.1 Introduction . . . 48
4.2 Material and methods . . . 49
4.2.1 Meteorological data . . . 49
4.2.2 Numerical model . . . 50
4.2.3 Annual concentration calculation . . . 52
4.2.4 Comparison cases considered in this study . . . 53
4.3 Results . . . 54
4.3.1 Annual concentration calculation when ignoring wind directions with a regular step . . . 54
4.3.2 Annual concentration calculation when considering the predominant wind directions . . . 58
4.3.3 Comparison between both methodologies . . . 61
4.3.4 Estimation of the error with respect to the complete wind rose . . . 62
4.4 Discussion . . . 66
4.5 Conclusion . . . 67

5 Assessment of mean annual NO_2 and NO_x concentration based on a partial dataset 69
5.1 Introduction . . . 70
5.2 Material and methods . . . 72
5.2.1 Study location . . . 72
5.2.2 Data availability . . . 72
5.2.3 Data range . . . 73
5.2.4 Monitoring method . . . 74
5.2.5 Empirical methods to convert concentration from NO_x to NO_2 . . . 74
5.3 Results . . . 75
5.3.1 Evaluation of annual NO_2 concentration based on NO_x data . . . 75
5.3.2 Seasonal variability of NO_2 concentration . . . 78
5.3.3 Assessment of annual NO_2 concentration . . . 80
5.4 Discussion . . . 86
5.5 Conclusion . . . 87

6 Assessment of mean annual PM_{10} and $PM_{2.5}$ concentration based on a partial dataset **89**
6.1 Introduction . . . 89
6.2 Material and method . . . 90
6.2.1 Study location . . . 90
6.2.2 Data availability . . . 91
6.2.3 Data range . . . 91
6.2.4 Statistical performance measures . . . 91
6.3 Results . . . 92
6.3.1 PM_{10} and $PM_{2.5}$ annual mean concentration trends in France . . . 92
6.3.2 Assessment of annual concentrations based on monthly data . . . 94
6.3.3 Assessment of annual concentrations based on monthly data by years 97
6.3.4 Assessment of annual concentrations based on group of months . . . 98
6.3.5 Correlation between MRE and $C_{95}RE$ . . . 100
6.3.6 Correlation between PM_{10} and $PM_{2.5}$ annual concentrations . . . 100
6.4 Discussion and perspectives . . . 103
6.5 Conclusion . . . 104
6.6 Annex . . . 106

II Deep learning models to estimate urban pollution in real time **109**

7 How can artificial intelligence be used with CFD to achieve real time pollutant dispersion? **110**
7.1 Artificial intelligence brief history . . . 111
7.2 Artificial intelligence concepts behind the words . . . 112
7.3 Deep Learning concept . . . 113
7.4 Deep learning in Fluid Dynamics . . . 117
7.5 Conclusion . . . 119

8 First approach of Deep Learning modelling to assess pollutant dispersion **121**
8.1 Introduction . . . 122
8.2 Material and methods . . . 123
8.2.1 Numerical model . . . 123
8.2.2 Data . . . 124
8.2.3 Interpolation methods . . . 124
8.2.4 Metrics . . . 128

8.3 Results . . . 130
8.3.1 Interpolation approaches comparison . . . 130
8.3.2 Deep Learning approach: U-net . . . 132
8.4 Conclusion . . . 134

9 Data generation to create examples for AI 135
9.1 Introduction . . . 136
9.2 Hypothesis on the CFD used to train the model . . . 136
9.2.1 Model assumption and equations . . . 136
9.2.2 Boundary conditions . . . 139
9.3 Creation of the data . . . 140
9.3.1 Geometries . . . 140
9.3.2 Strategies to build examples . . . 141
9.3.3 Meshing . . . 143
9.4 Data preprocessing . . . 144
9.4.1 Data collection from CFD modelling . . . 144
9.4.2 Treatment of simulations results . . . 145
9.5 Conclusion . . . 146

10 Deep learning architectures to learn air pollutant dispersion from fluids mechanics 148
10.1 Introduction . . . 149
10.2 Material and methods . . . 150
10.2.1 Physical numerical model . . . 150
10.2.2 Deep learning architectures . . . 152
10.2.3 Input and output data for the deep learning models . . . 153
10.2.4 Evaluation of the results . . . 154
10.3 Results . . . 156
10.3.1 Loss functions and filters . . . 157
10.3.2 Architectures . . . 157
10.4 Conclusion . . . 158

11 Dwelve in depth the MultiResUnet architecture 160
11.1 Introduction . . . 161
11.2 Size of the area of interest . . . 161
11.2.1 Training parameters . . . 162
11.2.2 Tested architectures . . . 164
11.2.3 Architecture without padding . . . 172

11.2.4 Comparison of the variants architectures and conclusion 174
11.3 Optimising the multiResUnet architecture 176
11.3.1 Attention layers 176
11.3.2 Tuning hyperparameter 177
11.4 3D MultiResUnet 179
11.4.1 MultiResUnet 3D architecture 179
11.4.2 3D MultiResUnet results 182
11.5 Conclusion 184

12 Case study to estimate urban pollution in real time **186**
12.1 Introduction 187
12.2 Real Time dispersion monitoring system 187
12.3 Case study 188
12.3.1 Context 188
12.3.2 Material and method 190
12.3.3 Results 195
12.4 Conclusion 195

13 Conclusion and perspectives **197**
13.1 Conclusion of the dissertation 197
13.1.1 First objective: Assessing mean annual exposure 197
13.1.2 Second objective: Assessing exposure in real time 199
13.2 Perspectives 202

14 Résumé étendu en français **203**
14.1 Introduction 203
14.2 Évaluation de la moyenne annuelle par modélisation numérique et mesures de capteurs 204
14.2.1 Comment évaluer la pollution atmosphérique moyenne annuelle à l'échelle d'un quartier ? 204
14.2.2 Optimisation de la discrétisation de la rose des vents 208
14.2.3 Évaluation de la concentration annuelle moyenne de $\mathbf{NO_2}$ sur la base de données partielles 210
14.2.4 Évaluation de la concentration annuelle moyenne de $\mathbf{PM_{10}}$ et $\mathbf{PM_{2.5}}$ sur la base de données partielles 213
14.3 Évaluation de la dispersion de polluant en temps réel 215
14.3.1 Évaluation de la capacité des modèles d'apprentissage profond pour évaluer la dispersion des polluants 216

14.3.2 Génération de données 217
14.3.3 Comparaison d'architecture d'apprentissage profond pour la dispersion de polluant 219
14.3.4 Optimisation de l'architecture multiResUnet quant à la thématique de la dispersion de polluant 222
14.3.5 Une étude de cas pour estimer la pollution urbaine en temps réel . . 223
14.3.6 Conclusion et perspectives 225

Chapter 1

Introduction

Environmental air pollution and protection is a topic that has crossed the ages. Indoor air pollution started already in early human societies, all around the world, which suffered from domestic air pollution due to domestic fire. Indeed, traces of anthracosis (blackening of the lungs) were found in corpses in Egypt, Peru or Britain (97). Outdoor air pollution issues only raised when major cities were founded. Hippocrates in his treaty "Airs, Waters, Places" written in 400 B.C. underlined the importance of pure and clean air, and water quality (97). In these cities, pollution from small manufactures and domestic uses even led to civil claims in Ancient Rome. The emperor Justinian proclaimed that it was considered a birth right to have access to clean air and water (97). In the 13th century, o tackle the issue of air pollution, King Edward I even threatened Londoners with high penalties if they continued to burn sea coal. It became an even bigger issue with the industrial revolution that eventually led to more air pollution from plants and manufactures (7). Air pollution regulations started to be really restraining during the middle of the twenty centuries especially in the aftermath of two air pollution disasters: the death of 20 persons and sickness of 7000 others in the city of Donora, Pennsylvania, and the Great smog in London in 1952 that led to the death of about 4000 persons[1].

1.1 Air pollutant sources and effects

According to the Oxford dictionary, pollution is "*The presence in or introduction into the environment of a substance which has harmful or poisonous effects.*" This definition is quite relative in what may be considered harmful or poisonous. Thus, air pollution can span from irritating particles such as pollen, to deadly chemicals such as H_2S. Nevertheless, some pollutants exist in greater quantities and have more adverse effects than others. This

[1] https://www.history.com/topics/natural-disasters-and-environment/water-and-air-pollution

balance between presence and adverse effect led The World Health Organization (WHO) to highlight four main air pollutants for human health around the world: particulate matter (PM), nitrogen dioxide (NO_2), ozone (O_3) and sulfur dioxide (SO_2). It can also be noted that Carbone Monoxyde (CO) and Lead (Pb) are also sometimes considered among the main air pollutants (91). Of course, other air pollutants exist locally and may require specific attention. For instance, heavy metals may need to be monitored near metallurgical activities.

As previously stated, these pollutants are considered to be of first importance because of their harmful effect and high presence. So what are the short and long-term effects on health and on environment of this pollutants and their sources?

Particulate Matter PM consists of carbon-based particles with added reactive metals and organic chemicals (47) within a certain size category. Particles with a diameter less or equal than a value X (in micrometers) are designed by PM_X. This leads to three commonly used sizes of particulate matter, with coarse particles PM_{10}, fine particles $PM_{2.5}$ and ultra-fine particles $PM_{0.1}$ (47).

The size is important in the adverse effect these particles may have on health since the smaller the particles are, the deeper they can penetrate the body as shown in Table 1.1.

Particle size	**Penetration degree in human respiratory system**
$> 11\mu m$	Passage into nostrils and upper respiratory tract
$7 - 11\mu m$	Passage into the nasal cavity
$4.7 - 7\mu m$	Passage into larynx
$3.3 - 4.7\mu m$	Passage into the trachea-bronchial area
$2.1 - 3.3\mu m$	Secondary bronchial area passage
$1.1 - 2.1\mu m$	Terminal bronchial area passage
$0.65 - 1.1\mu m$	Bronchial penetrability
$0.43 - 0.65\mu m$	Alveolar penetrability

Table 1.1: Penetrability according to particle size (88)

PM can cause several diseases. For instance, it was found for the short-term effects that $PM_{2.5}$ increases the number of hospitalisation by 1.04% every 10 µg/m3 in Europe and the United States (10) and by 0.47% every 10 µg/m3 in Asia (101). For the long-term effects, it has been proved that PM increases lung cancer, cardiopulmonary diseases and various diseases (47). The effects on the environment can also be detrimental: damaged forests and crops, break of the balance in nutrient and water ecosystems, acidified water bodies.(88)

PM are produced by a lot human activities. They can be emitted from domestic use of fuel such as charcoal, wood or gas for heating or cooking (62). From traffic, they can be produced by car through combustion and lubricants for thermal motors but also from the friction of tyres, clutch or brakes (14). Finally, they can also be issued from industrial processes, which vary depending on the process and exhaust of the industry, or from agricultural processes (62). But it can also be the result of natural phenomena. For example, biomass burning can emit large quantity of PM in the atmosphere (14) and natural particles are suspended by wind from sea salt, soil and dust (62).

Nitrogen Dioxide NO_2 is a highly reactive gas. Depending on its concentration, the short effect of NO2 exposure may vary from mild respiratory symptoms for low concentrations to death for very high concentration in confined space (32). Chronic exposure to NO_2 can lead to respiratory infections (bronchitis, pneumonia) especially for the elderly and young people (32; 111). NO2 can also have detrimental effect for the environment. It can provoke acid rains (79), have negative effect on vegetation, damaging leaves and slowing growth. NO2 also comes from different activities that involve combustion of fuel, from traffic through the combustion of thermal motors which produce NO that reacts with O3 and transforms it into NO_2, from industrial exhaust (151) or from the residential sector with the use of domestic fuel for heating or cooking (6).

Ozone Ozone ($O3$) is a highly reactive gas that is unstable at ambient temperature and pressure. Ground-level Ozone can cause several health problems. It can cause coughing and sore throat, difficulty of breathing and favours lung infections especially for young people. Long-term exposure can probably participate to asthma development and premature deaths, ozone can continue to damage the lungs even when symptoms are no longer present (174). Ozone is also known to be a highly phytotoxic. It can affect crop yields and will most likely pose a major threat to world food, fibre and timber production (83). Ground-level Ozone is the product of chemical reactions between oxides of nitrogen and volatile organic compounds under the heat and sunlight. Volatile Organic Compound (VOC) are generally emitted by plants exhaust but also exist naturally. It should be noted that ozone is generally not a "local pollutant" because it can be carried far from its sources (88).

Sulfur dioxide SO_2 is a relative non-toxic gas when alone but in ambient air it reacts and turns into more toxic molecules (117). It can cause acute health respiratory effects as well as on the airway as on the lungs (32). Indeed, it can cause several diseases and adverse effects such as respiratory irritation, bronchitis, mucus production, and bronchospasm, and bronchoconstriction (88). SO_2 seems to also have detrimental effect for the environment by provoking acid rains and soils (88). The main sources of sulfur dioxide are the combustion of

coals and heating oils, industrial boilers and metal smelting. Natural sources also exist such as volcanoes. (52)

1.2 Air pollution regulations

1.2.1 Air pollution regulation in the World

The World Health Organisation was founded in 1948. The first statement on air quality dates back to 1958 and since several WHO experts have updated their guidelines and estimation regularly. In 2016, the WHO reported that air pollution is responsible for 8 million deaths from related diseases around the world (164; 165). 3.8 million of these deaths can be attributed to indoor air pollution and 4.2 million to outdoor air pollution. When compared to the other causes of mortality, air pollution-related death represents around 7.6 % of deaths worldwide. To help states and regulatory agencies, the WHO provides guidelines and thresholds. Those are not legally binding for states. The threshold values for the main classical air pollutant are provided in Table 1.2.

Pollutant	**Averaging Time**	**Thresholds**	**Year of latest WHO AQGs**
PM_{10}	Annual	$20\mu/m^3$	2006
	24h	$50\mu/m^3$	2006
$PM_{2.5}$	Annual	$10\mu/m^3$	2006
	24h	$25\mu/m^3$	2006
O_3	8h daily max	$100\mu/m^3$	2006
NO_2	Annual	$40\mu/m^3$	2010
	1h daily max	$200\mu/m^3$	2010
SO_2	24h	$20\mu/m^3$	2006
	10min	$500\mu/m^3$	2006

Table 1.2: WHO guidelines on main pollutants (source: WHO).

According to the WHO, 9 out of 10 people live in areas that exceed the WHO safe health based standards. The WHO also provides guidelines for other air pollutants that are outside the scope of this work.

1.2.2 Air pollution regulation in Europe

According to the European Environment Agency air pollution is at present time the most important environmental risk to human health for European citizens and is perceived as the second-biggest environmental concern after climate change (38). According to the same report, air pollution was responsible of around 450 000 premature deaths in 2016 in European Union countries. It was estimated that 374 000 (82%) are due to $PM_{2.5}$, 71 000 (15%) due to NO_2 and 14 000 (3%) due to O_3. A more recent study from (77) estimates that around 790 000 persons die prematurely each year due to air pollution in the European Union countries. It also estimates the loss of life expectancy to be about 2.2 years. Most of the deaths 40-80 % would be related to cardiovascular events. To protect people from air pollution thresholds values have been enforced by the European Union and are presented on Table 1.3 for the main air pollutants.

Pollutant	**Averaging Time**	**Thresholds**	**Comments**
PM_{10}	Annual	$40\mu/m^3$	Limit value
	1 day	$50\mu/m^3$	Not to be exceeded on more than 35 days per year
$PM_{2.5}$	Annual	$25\mu/m^3$	Limit value
	Annual	$20\mu/m^3$	Average exposure indicator over 3 years in urban background areas
O_3	8h daily max	$120\mu/m^3$	Not be exceeded on more than 25/days/year averaged over 3 years
	1 hour	$180/240\mu/m^3$	Information / Alert threshold
NO_2	Annual	$40\mu/m^3$	Limit Value
	1 hour	$200/400\mu/m^3$	Limit Value / Alert threshold
SO_2	1 day	$125\mu/m^3$	Not be exceeded on more than 24 hours per year
	1 hour	$350/500\mu/m^3$	Limit Value / Alert threshold

Table 1.3: European Environment Agency guidelines on main pollutants (38)

These thresholds are regulatory through the 2008/50/CE directive (37) for each member state that needs to implement measures to ensure they are respected.

1.2.3 Air pollution regulation in France

In France, according to the French Agency for Ecological Transition (ADEME) report (4) 46% of French people surveyed declared to have experienced first-hand or know people who had experienced discomfort or trouble due to air pollution. Air pollution is an important issue in France despite the decrease of the number of premature deaths, with 48 000 premature deaths attributed to $PM_{2.5}$ in 2016 (41) against 40 000 in 2019 (1). Also in 2019, 7000 premature deaths were attributed to NO_2 (1). The study on air quality in France gives other interesting information such as the life expectancy loss depending on the city size. According to (41):

- For urban areas with more than 100 000 dwellers, there is a life expectancy loss between 15 months and 30 years.
- For urban areas between 2000 and 100 000 dwellers, there is a life expectancy loss around 10 months.
- For rural areas, there is a life expectancy loss around 9 months.

According to these figures, pollution in France is mostly an issue for highly densified areas with huge variation on life expectancy loss up to 30 years. To tackle the issue of pollution in densified cities, France has implemented an atmospheric protection plan, known as PPA in French (Plan de Protection de l'Atmosphère) in areas that have more than 250 000 dwellers or that exceed the European Union limit values for air pollution. PPAs are five-year plans, elaborated at local level, which defines preventive and corrective measures to implement and to respect the regulatory thresholds for air quality. Their objective is to take into account local particularities of the city and its surroundings. For instance, the PPA of Paris encompasses 1281 towns, 11.8 million inhabitants. It was reported that 3.6 million dwellers may be exposed to pollution exceeding European thresholds for NO_2 and 1.8 million for PM_{10} (2). Eleven measures have been adopted to tackle the issue, for instance aiming at reducing the traffic, limiting the heating plant exhaust and domestic exhaust (2), to name but a few. To pilot these projects, France has since the seventies started to implement an air quality monitoring network. Currently, each of 18 regions in France has its own dedicated agency, called "AASQA" (Association Agréée de Surveillance de la Qualité de l'Air), in charge of the monitoring of air quality.

1.2.4 Challenges

As previously seen, air pollution is one of the most impacting environmental threat to health worldwide and in Europe. Indeed, the main so-called classical pollutants PM, NO_2, O_3, SO_2

are responsible for several diseases from mild respiratory symptoms to acute lung cancer. Their sources are various and omnipresent in urban areas through traffic, domestic heating or industry for instance. Guidelines and regulation have been enforced to limit concentration values not to be exceeded in order to improve the air quality of dwellers. These regulations are enforced by each Member State and in France it has led to large cities of more than 250 000 dwellers to act upon it through the creation of PPA. Region's agencies have been assigned to monitor these values and to evaluate the impact of the locally implemented measures. This has led to a decrease in deaths in France going from 48 000 to 40 000 in the span of 3 years. Nevertheless, still much needs to be done, urban areas are still subjected to life expectancy loss that can go up to 30 years. How can the situation in urban areas be improved ? Through thorough knowledge of the situation. As it has been shown, annual and hourly regulations exist. Sensors can be used to evaluate the pollution of a city but sensors give only very local information. To cover a whole city, it would thus be necessary to have a high number of sensors. Yet, reliable sensors are far from cheap to purchase and maintain. In addition, for annual concentration, the sensors would need to be fixed at a specific place to determine the mean annual concentrations, thus reducing their range of action. Hence, a challenge arises : **How to assess annual concentrations with sensors without requiring a full-year period?** To tackle this issue, data mining and analyses from sensors may provide answers. Indeed, would studying seasonality with sensors allow to identify a relationship between annual concentrations and monthly ? Sensor monitoring is not the only method to determine the exposure in urban context. Another way of studying air quality in a neighborhoods is by using physical models that try to represent the complex phenomenon of pollution dispersion in a real environment. Several models exist with different scales from continents to local neighborhoods. The subject of interest of this thesis is urban areas which as seen above are the places with the biggest stakes. To study pollution from local sources several models exist. However a constant remains, the more accurate the model is and takes into account complex phenomena, the more computational expensive the model is. Computational Fluid Dynamics is one of these models that is able to have a resolution of the meter, take a lot of complex phenomena into account but require vast amount of computing resources. Thus, it is not possible to play it the whole year and in a timely manner to meet the annual requirements from the regulations. Thus, a challenge arises: **How to assess annual concentrations with a numerical model that is computing demanding?** To overcome this challenge, new methodologies need to be created and adapted for CFD models. Pollution mostly depends on meteorological conditions, such as wind speed and frequencies, and pollution emission. So, a statistical methodology using these parameters and their frequencies may help to reduce the number of simulations to a more computation-wise sustainable number. The regulation also points toward hourly threshold. As for annual

assessment, it is not possible to compute neighborhoods in real time to meet the hourly demands using traditional CFD approach. A surrogate may be to compute CFD simulations in advance in an area. However, even by doing so, the computing time would still mean that it could not cover much ground or it would be too costly. Therefore, the following challenge arises: **How to assess concentrations in real time over large areas when computational fluid dynamics modeling is too computationally expensive and slow?** Faster model must be used but at the same time, it is preferable to keep the accuracy as close as possible from CFD quality. Hence, new methods must be developed. A fast pace rising field that managed to change the paradigm in a lot of domains is machine learning and artificial intelligence. It has shown its capacity to improve both speed and accuracy in image analyses or speech recognition for instance. So, it may be possible that by using the recent advances in Deep Learning and by adapting and applying it to CFD, the computation time could be immensely reduced. By doing so, it would be possible to determine pollution dispersion in real time while approaching CFD quality.

1.3 Dissertation content

The thesis works around the two extreme time scales required by the regulation, annual and hourly. Hence, this document is divided into two main parts covering these two time scales.

1.3.1 Plan of the document

Mean annual concentration evaluation

Chapter 2.3 is an introductory chapter that presents the complexity of air pollution assessment through modelling and sensor monitoring. The different scale at play, the specificity of the urban areas and the pros and cons of the different existing classical models. Then, this part is split into two sub-parts.
The first aims at assessing annual concentration with computing expensive numerical models such as CFD. Chapter 3 aims at determining how mean annual concentration maps can be assessed from wind roses. It describes an approach based on frequency with the summation of wind direction and wind speed to have a concentration map representative of the whole year. Chapter 4 uses the concrete examples of CFD models to use the previous developed methodology. It studies the discretization of the wind rose using several real-life wind roses from all around France as well as several neighborhoods. It analyses how to optimize the discretization to spare computing time while still controlling the error made from reducing the number of used directions. The second sub-part studies how to reduce the period of time required to assess annual concentrations from sensors. Chapter 5 develops a method to take

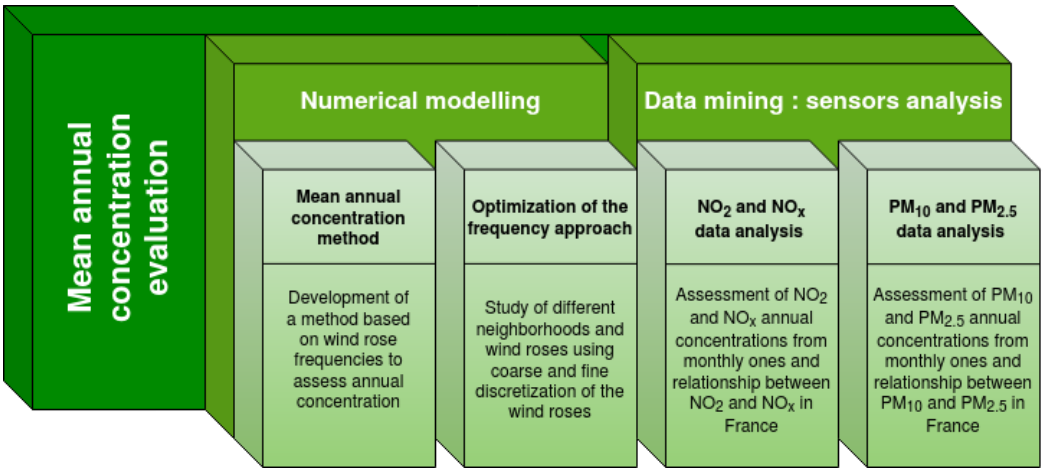


Figure 1.1: Thesis overview of the first part of the thesis

into account lacking data from NO_2 and NO_x sensors. On one hand, the chapter explores the relationship between NO_2 and NO_x in France and from previous scientific articles. On the other hand, the section investigates the seasonality of NO_2 and NO_x and the relationship between monthly and annual concentrations. Chapter 6 is quite analogous to section 3 but with PM_10 and $PM_2.5$. On one hand, the section studies the variability of particulate matter in France and its seasonality. On the other hand, it explores the relationship between PM_10 and $PM_2.5$ under the prism of the main pollution sources type of the area and seasons.

Real time pollutant dispersion monitoring

The second chapter focuses on using Deep Learning models based on CFD results to be able to determine a pollution concentration map in real time over large areas. Chapter 2.3 is an introductory chapter that presents the change of paradigm that recent advances in artificial intelligence and machine learning represent. A focus is especially made on CFD associated with Deep Learning and the improvements and advances it has managed to provide to the modeling community. This part is then divided into three sub-parts.
The first evaluates the capabilities of Deep Learning approaches to solve complex dispersion pollutant field. Section 5 apprehends the power of machine learning to interpolate or extrapolate pollutant dispersion fields. It uses several methods from simple linear interpolation to deep learning Unet model. Section 6 explores two issues. On one hand, it studies the creation of data from CFD. How to optimise their creation since CFD is computationally expensive. On the other hand, it investigates the issue of transforming the data from the CFD models into a more suitable form for Deep Learning algorithm. The second presents our proposition of deep learning architecture to automatically compute a pollution concentration map in real time. Section 7 compares several Deep Learning architectures that have proved to be effective

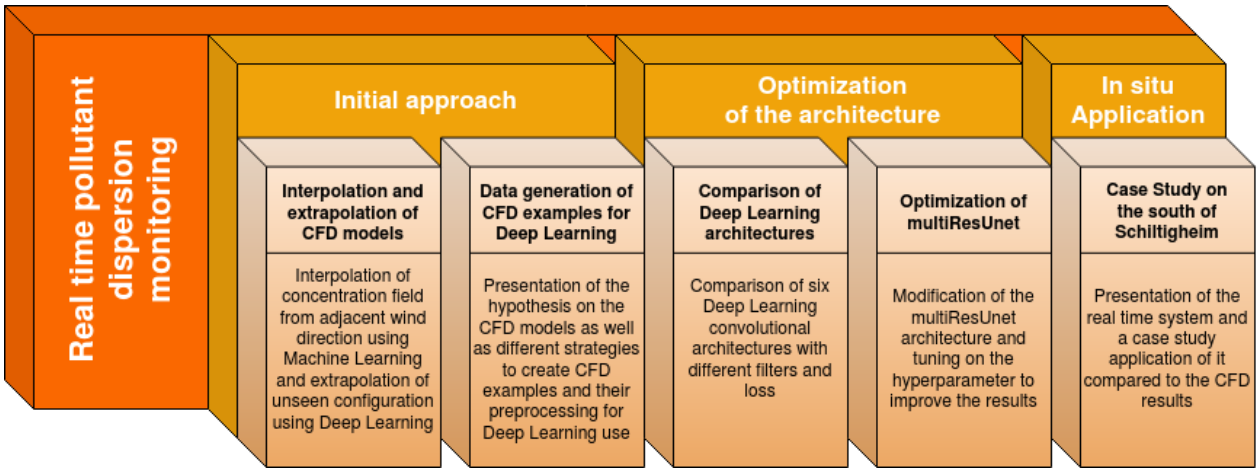


Figure 1.2: Thesis overview of the second part of the thesis

in treating spatial data. It tests several variants of the architectures with different losses, complexity of the architecture and activation functions. Section 8 tries to optimize the best architecture determined by section 7. Modifications of the architecture are carried out as well as a tuning of its hyperparameters. Finally, the third section aims at comparing the results obtained by our approach against its CFD counterparts on a real use case covering a large-scale neighborhood of $1 \times 1km^2$. Section 9 presents a functional system to determine pollution dispersion in real time. This system is then applied on a real neighborhood in the city of Schiltigheim covering about $1km^2$. The results using a CFD model of the area as the reference are compared with the Deep Learning approaches both to check on its accuracy compared to the state-of-the-art model and speed improvement.

1.3.2 Scientific contributions

During this PhD, several contributions to the research field have been achieved. Six articles as first or co-first author, one conference article as first author and two as associated author have been published or submitted.

First or co-first author (included in the dissertation)

1. Jurado, X., Reiminger, N., Vazquez, J., Wemmert, C., Dufresne, M., Blond, N., and Wertel, J. Assessment of mean annual NO2 concentration based on a partial dataset. *Atmospheric Environment 221* (Jan. 2020), 117087

2. Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Dufresne, M., Blond, N., and Wertel, J. Methodologies to assess mean annual air pollution concentration combining numerical results and wind roses. *Sustainable Cities and Society 59* (Aug.

2020), 102221

3. Jurado, X., Reiminger, N., Vazquez, J., and Wemmert, C. On the minimal wind directions required to assess mean annual air pollution concentration based on CFD results. *Sustainable Cities and Society 71* (Aug. 2021), 102920

4. Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Deep learning associated with computational fluid dynamics to predict pollution concentration fields in urban areas. In *Proceedings of the Upper Rhine-AI Conference* (October 2021)

5. Jurado, X., Reiminger, N., Maurer, L., Vazquez, J., and Wemmert, C. Assessment of mean annual pm_{10} and $pm_{2.5}$ concentration based on a partial dataset. *Submitted to Sustainable Cities and Society*

6. Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Assessment of capability of deep learning to predict air pollution dispersion. *Submitted to Computers, Environment and Urban Systems*

7. Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Deep learning methods evaluation to predict air quality based on computational fluid dynamics. *Submitted to Expert Systems with Application*

Associated author (not included in the dissertation)

1. Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Blond, N., Dufresne, M., and Wertel, J. Effects of wind speed and atmospheric stability on the air pollution reduction rate induced by noise barriers. *Journal of Wind Engineering and Industrial Aerodynamics 200* (May 2020), 104160

2. Maurer, L., Villette, C., Reiminger, N., Jurado, X., Laurent, J., Nuel, M., Mosé, R., Wanko, A., and Heintz, D. Distribution and degradation trend of micropollutants in a surface flow treatment wetland revealed by 3d numerical modelling combined with LC-MS/MS. *Water Research 190* (Feb. 2021), 116672

1.3.3 Context of the thesis

The thesis was realised thanks to a CIFRE contract (Industrial Agreements for Training through Research). It is a partnership between private and public entity. It involves the ANRt (national research technology association), a public university (University of Strasbourg), a laboratory (ICube) and an enterprise (Air&D). Air&D already sponsored a PhD thesis

prior to this one using the same package. It involved the University of Strasbourg and ICube to develop a 3D CFD solver for atmospheric pollution with Nicolas Reiminger as PhD candidates. The result of this thesis was a functional CFD models validated against experimental data and the ground basis of several research articles in international journals and conference. Air&D is a start-up that was created in 2017 that aims at improving the way air quality monitoring and diagnostic is done using state of art methods and developing new tools. It was founded by Christophe Legorgeu, Jonathan Wertel, Matthieu Dusfrene and José Vazquez. Actually, the company, excluding its founder members, has 4 employees. Nicolas Reiminger specialised in Air quality and modelling, Xavier Jurado specialised in modelling and artificial intelligence, Yohan Stephanus and Loic Saunier specialised in web development. The tools used by air&D cover several aspects of the air quality issue. Air&D developed numerical modelling methods using innovative 3D approaches for engineering issues. Air&D participated in the creation of brand new sensors to study the pollution *in situ* and validate the results from the numerical model. Air&D developed a Deep Learning model for real time pollution assessment through the grant for this thesis.

Part I

Annual urban pollution assessment at the scale of the neighborhood using modelling and sensors

Chapter 2

Atmospheric pollution modelling

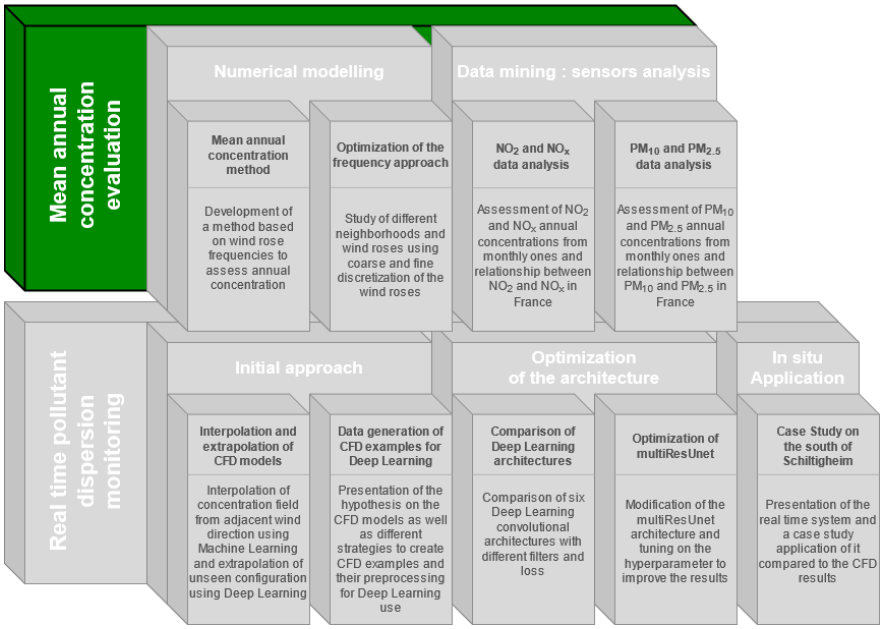


Figure 2.1: location of this chapter in the thesis

Successful atmospheric pollution modelling dates back to the early 20th century with the work of (148) and (18) according to (173). These models were Gaussian plumes model that was used to evaluate the range and concentration of the pollution from industrial plants. Since then, air modelling has gone a long way and lots of models and techniques have been developed to assess air pollution from various sources type at various temporal and spatial scales.

Nowadays, atmospheric pollution assessment can be used for several reasons. It can be used for diagnostic, to assess a situation, for example if a concentration exceeds a regulatory threshold. It can be used for the projection, to test scenarios, for instance what would happen

if the traffic would decrease of a certain amount. It can be used for risk prevention, to assess the environmental hazard, for instance if there is a leak of dangerous gaseous chemical in a plant. Thunis *et al.* (152) have studied the uses of atmospheric modelling in research projects.

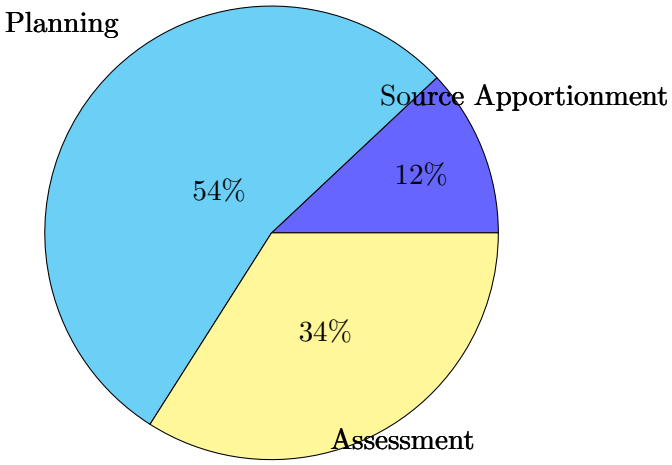


Figure 2.2: Pie chart on the uses of models in research projects according to (152) in EU

This part of the thesis mainly focuses on the assessment. However the tools can also be used for planning and testing scenarios.

2.1 Atmospheric pollutant scales

Atmospheric pollution involves a lot of phenomena different in nature, timescale and space scale. For instance, pollution emitted from traffic for NO_2 will affect mostly the vicinity close to its sources, SO_2 pollutant emitted from a plant may impact several kilometres around it and O_3 has to be modelled on a continental scale (67).

Scale	**Characteristic length**	**Atmospheric phenomena**
Micro scale	$<$ 1 km	local meteorology, turbulence
Meso scale	1 km - 1000 km	thermal ascendant movement, storm
Synoptic scale	10 000km	synoptic movement

Table 2.1: Atmospheric phenomena depending on their scale (67)

Having one model that would encompass all the different scales while keeping an accurate resolution and acceptable computing time is not feasible at the present time. Thus, different

models to assess the pollution dispersion have been created to tackle each scale. They can also be coupled to become a hybrid model. For instance, to enhance the assessment of a local model, its boundary conditions can be supplied by a bigger model. Depending on the aim of the researcher or environmental engineer, the models will be different. For atmospheric pollution there are several classes of models for different scales, the order of magnitude of each model are given on Figure 2.3 according to (96).

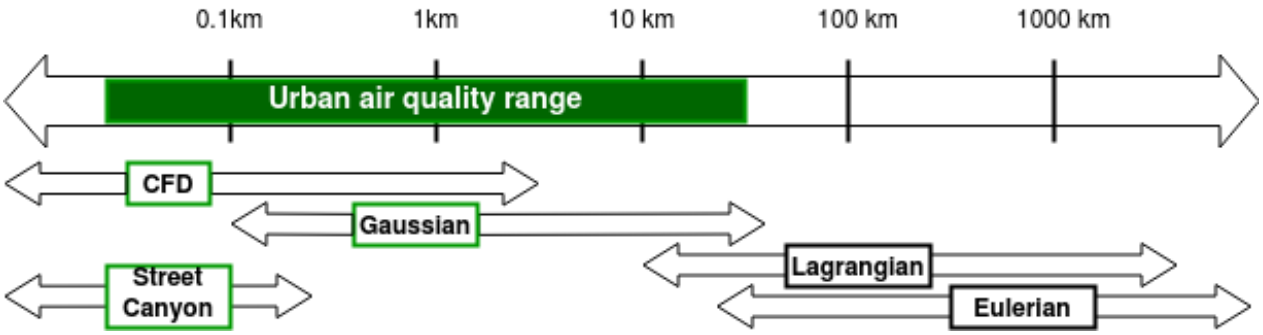
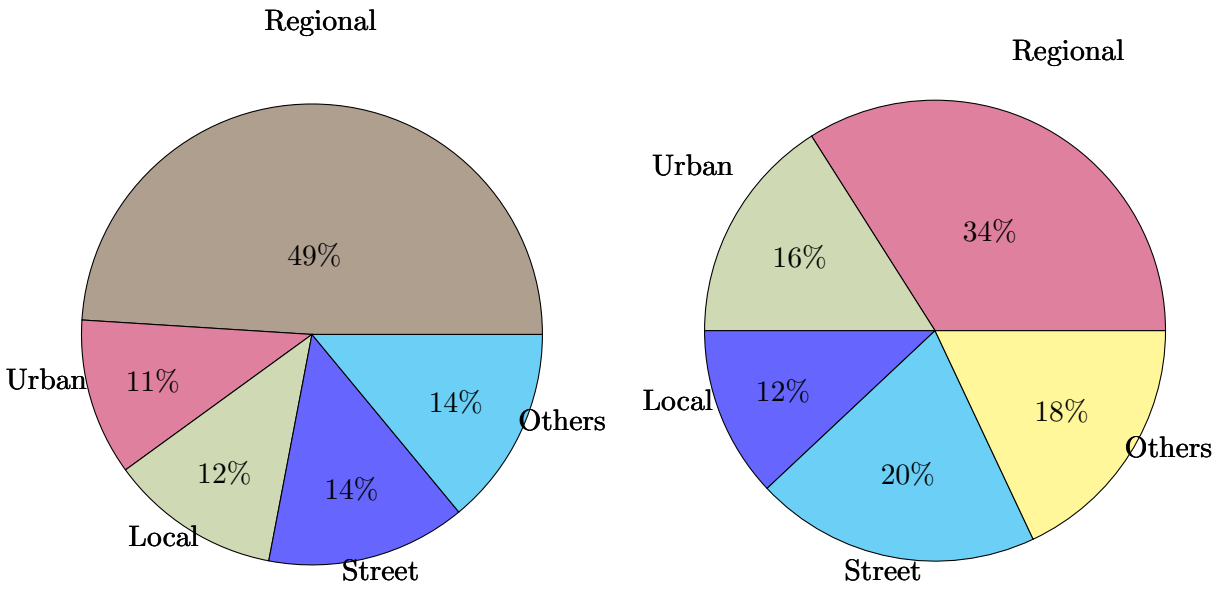


Figure 2.3: Order of magnitude on spatial length of models for atmospheric pollution

Urban air quality (encompassing every scale from the street to the entire city) is of main interest for environmental researchers representing around a third of research projects in European Union in the air quality field. It is also of main interest for public authorities representing nearly half of the Air Quality Plans as it can be seen on Figure 2.4. Indeed, as seen previously in the introduction (Chapter 1), urban areas because of its high density of dwellers and proximity to pollution sources are of primordial interest to assess the exposure of residents. Furthermore, urban dwellers are the majority of the worldwide population with around 55 % in the world that lives in urban areas and 80 % in France.



(a) Research Project

(b) Air Quality Plans

Figure 2.4: Pie chart on the scale of interest in Air Quality Planning and Research according to (152) in EU. Regional ranges from 10 to 50 km, urban from 1 to 5 km and local below 1 km.

2.2 Atmospheric pollutant assessment for urban areas

Urban air quality is a particular topic as the pollution level and sources differ greatly from the countryside. Urban background pollution is generally closely linked to the size of the city (50), the bigger the city, the higher the background pollution. Indeed, cities have numerous local pollution hotspots that tend to worsen air quality and may have poor ventilation. Cities are also within themselves subject to different phenomena with their own scale. According to (48) the smallest scale is the street canyon. Street canyons are very common in cities since they are basically every street with two high buildings on each side. They are the seat of particular air movements that depend on the height and wind speed and can lead to high concentration of pollutants emitted from the local traffic because of poor ventilation as shown in (121). The wind pattern is also affected horizontally by the upstream rows of buildings and obstacles such as trees and buildings that directly affect the wind or create differences in heat flux. Cities by their high building densely parsed and heat fluxes also impact globally the wind patterns in their vicinity. All of these phenomena make the tasks of evaluating urban wind flow and underlying pollutant dispersion a challenge.

For urban scale, as seen above on Figure 2.3, several models exist. These models can be split into two kinds of models that may or may not interact. On one hand, the semi-

empirical model such as street canyons (95), or Gaussian models (Plume and Puff) (114). The advantage of these models is their mathematical simplicity, they can provide fast and reliable results for annual average (95; 114) when they are used in the proper setting, on the right scale and knowing their limits. There are many software derived from this model, to cite some of the best known are ADMS, Aermod, CALPUFF or CALINE4 for hybrid Gaussian/street model. On the other hand, CFD models can be applied to air quality as well (122). This class of models solves fluid mechanics equation and depending on the algorithms, with more or less assumption made on the fluid and turbulence. These methods take into account more complex phenomena such as turbulence induced by buildings that cannot be done using semi-empirical law (158) and can even take into account radiation or thermal effect (119).

Studies have been done to compare Gaussian models and CFD. It was found that in flat open field they had similar result (92) but when the terrain is not flat CFD yield better results (71). And when an urban area is the area of interest, Gaussian model yield poor results when compared with CFD modelling (12). So, why is not everyone using CFD models over Gaussian if they have better accuracy on pollutant dispersion and better representation of physical phenomena? There are two principal reasons for that. First, CFD models are harder to set up and to make converge. Indeed, they can have stability issues especially when several complex phenomena are used at the same time such as radiation and turbulence. Secondly, this accuracy comes at a heavy price, the computing cost makes it unsuitable for wide areas and even for local scale ($<$ 1km) it is still costly compared to the semi-empirical models. For instance, with semi-empirical models, it is possible to an extent to play a whole year, hour by hour, with the corresponding meteorological data and then averaging the results to obtain the mean annual exposure in an area (114). With CFD models it is not possible to do so in an acceptable time. Therefore, new approaches must be developed to be able to determine mean annual concentration using CFD.

For sensors, the issue is similar. To obtain mean annual concentration it is required to have sensors of the wanted pollutant covering the whole year while remaining at the same place. This in itself would not be an issue if the sensors were cheap. But reliable sensors of medium or high quality are very costly. For instance, the sensors used by ATMO Grand Est in Strasbourg cost tenth of thousands of euros and the medium quality sensors used by Air&D cost several thousands of euros. Moreover, if the sensors have a breakdown of several months, the annual average is lost. Hence, two issues arise. First, reducing the period of time for sensors to determine annual average concentration. Is it possible from a monthly period of measurement to determine annual average and if so, what is the error ones does when doing it? Secondly, being able to determine other pollutants from the same kind from one sensor. For instance, is it possible to know the pollution concentration of $PM_{2.5}$ when

knowing PM_{10} or NO_2 when knowing NO_x?

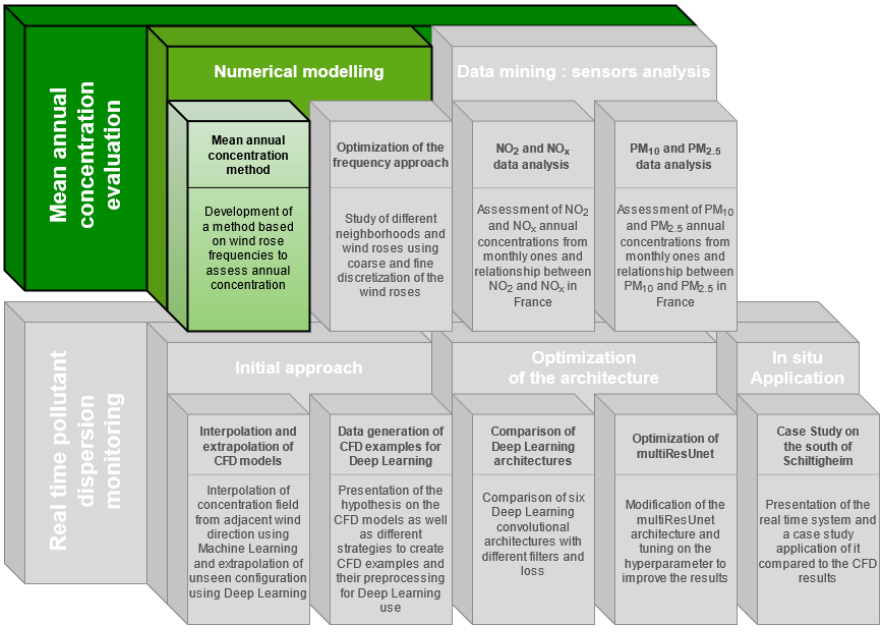
2.3 Conclusion

Air pollution modelling is a complex topic that involves different time and spatial scales. It can span from continents to the street of a city. These scales have led to the creation of several models adapted to them. There are no models that can resolve every scale with great accuracy and affordable computing time. In this thesis, the focus is on urban areas which represents in the European Union around a third of the research and half of the air quality plans. It is a topic of prime interest since urban areas atmospheric pollution sources are close to the dwellers and that it concerns the majority of the world population. At the scale of urban areas, specific phenomenon must be taken into account such as the building's impact on air flows. For urban areas, three main types of models exist street canyons, Gaussian and CFD. Street Canyons and Gaussian are semi-empirical models that can yield reliable and fast results when used in the right cases. CFD on the other hand, yield better results in urban areas than the semi-empirical model but at the cost of complexity in the set-up and convergence and computing resources. This limitation makes the CFD models harder to use to predict annual average concentration since it cannot play a whole year of meteorological data hour by hour as with semi-empirical models.

Therefore, to be able to use CFD for annual average concentration maps, new methods adapted to its limitation must be developed. A second solution for urban air quality is to use sensors. Nevertheless, sensors are costly and for the annual average, they need a whole year of data remaining at the same place. Hence, methods to reduce the period of measurements and determining other pollutants of the same kind could help reduce the cost of air quality monitoring.

Chapter 3

How to assess the mean annual air pollution at the scale of a neighbourhood?



This chapter has been published in the journal *Sustainable Cities and Societies* under the title "Methodologies to assess mean annual air pollution concentration combining numerical results and wind roses " (120).

As seen in the previous chapter, scientific community lacks tools and methods to determine annual average concentration of air pollutants using CFD. In this chapter, a novel method based on a frequency approach is presented.

3.1 Introduction

Annual concentrations can be assessed using both on-site monitoring and numerical modeling. On site monitoring requires measurements over long periods to be able to assess mean annual concentrations of pollutants, although a recent study has shown that mean annual concentration of NO_2 can be assessed using only one month of data (60), which significantly reduces the measurement time required. Monitoring nonetheless has other limitations: it does not allow assessing the future evolution of the built environment or pollutant emissions, thus, limiting its applicability to achieve the smart sustainable cities of the future as defined in (15). Numerical modelling can overcome these limitations and can help define new strategies to improve air quality in cities combining wind data, various air pollution scenarios and urban morphologies (169). Among the several models currently available, Computational Fluid Dynamics (CFD) has shown great potential for modeling pollutant dispersion from traffic-induced emissions by including numerous physical phenomena such as the effects of trees (23; 139; 159) and heat exchanges (113; 154; 161) on the scale of a neighborhood. However, this type of numerical result cannot be directly compared with the annual standards. Methodologies designed to assess mean annual concentrations based on numerical results can be found in the literature (127; 146; 159), but further work is required to improve them and assess their limits.

The aim of this study is to provide tools and methodologies to assess mean annual concentrations based on numerical results and wind rose data to improve air quality in built environment and cities. It is firstly to evaluate whether it is possible to assess continuous wind speed distributions based on wind rose data. To do so, a statistical law called Weibull distribution is compared with a new sigmoid-based function built for the purpose of this study. Secondly, it is to present and compare a discrete methodology usually used to assess mean annual concentrations based on numerical results with a continuous methodology built for the purpose of this study, and to discuss their respective advantages and limitations. The data used for the wind speed distribution assessments, the area modeled and the CFD model used for illustration purposes are presented in Section 3.2. Then, the description and the comparison of the different methodologies are presented in Section 3.3 and, finally, a discussion is provided in Section 3.4.

3.2 Material and methods

3.2.1 Meteorological data

Data location This work uses wind velocity and wind direction data from four cities in France. These cities were chosen to cover most of France to obtain representative results and include the cities of Strasbourg (Grand-Est region), Nîmes (Occitanie region), Brest (Bretagne region) and Lille (Hauts-de-France region). In particular, the data were obtained from the stations named Strasbourg-Entzheim, Nîmes-Courbessac, Brest-Guipavas and Lille-Lesquin, respectively. The location of these stations and their corresponding regions are presented in Figure 3.1.

Figure 3.1: Location of the different meteorological stations used.

Data availability and data range The data used in this work were provided by Météo-France, a public institution and France's official meteorology and climatology service. The data are mainly couples of wind velocity and wind direction over a twenty-year period from 1999 to 2018, except for the Strasbourg-Entzheim station where it is a ten-year period from 1999 to 2008. The data were obtained via a personal request addressed to Météo-France and were not available on open-access. A summary of the information of the stations is presented in Table 3.1, with the time ranges of the data and the number of data available (the coordinates are given in the World Geodetic System 1984).

All the data were monitored from wind sensors placed 10 meters from the ground and the wind frequencies are available for each wind direction with 20° steps for two distinct wind

Station				Data availability		
Location	Latitude	Longitude	Altitude	Time range	Valid	Missing
Brest	48°27'00"N	4°22'59"O	94 m	2009 - 2018	29,171	45
Lille	50°34'12"N	3°05'51"E	47 m	2009 - 2018	29,185	31
Nîmes	43°51'24"N	4°24'22"E	59 m	2009 - 2018	29,214	2
Strasbourg	48°32'58"N	7°38'25"E	150 m	1999 - 2008	29,199	25

Table 3.1: Summary of the available data

discretizations: a "basic" discretization giving wind frequencies for 4 velocity ranges (from 0 to 1.5 m/s, 1.5 to 3.5 m/s, 3.5 to 8 m/s and more than 8 m/s), illustrated in Figure 3.2 (A); and a "detailed" discretization giving wind frequencies by 1 m/s steps except between 0 and 0.5 m/s, illustrated in Figure 3.2 (B). The "basic" discretization is a common format mostly found in wind roses (possibly with different velocity ranges) while the "detailed" data are less common and more expensive.

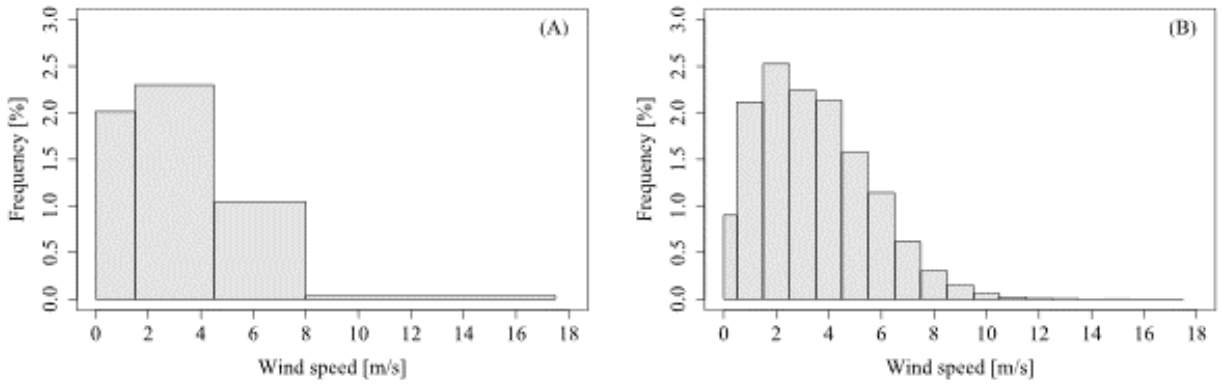


Figure 3.2: Examples of data for Strasbourg and a 200° wind direction with (A) only 4 ranges of velocities and (B) the detailed data discretized in 18 ranges.

The wind roses for each meteorological station considered in this work and based on the "basic" 4-velocity-range discretization described in Figure 3.2 (A) are provided in Figure 3.3. This shows how the monitoring locations considered in this study give distinct but complementary information, with for example many high velocities at Brest compared to Strasbourg and Nîmes, where almost no velocities were monitored over 8 m/s, and with dominant wind directions at Nîmes and Strasbourg compared to the other stations.

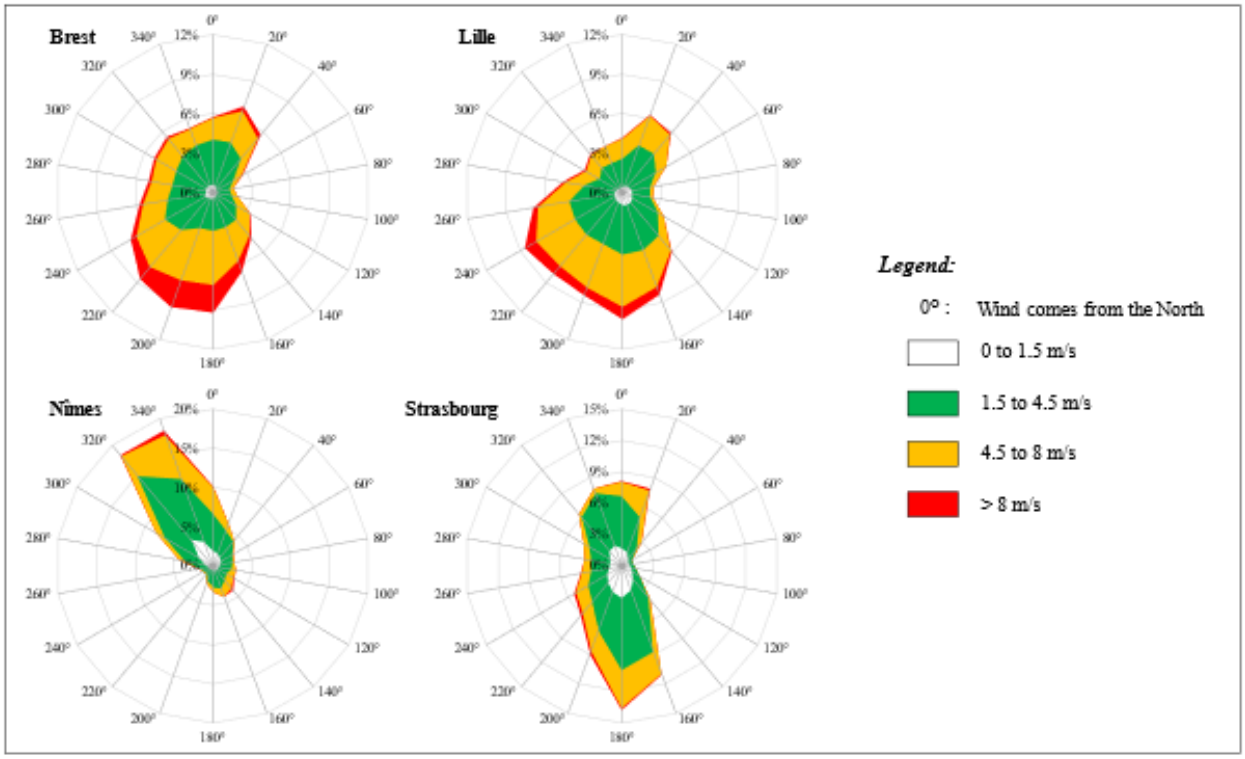


Figure 3.3: Wind roses for each location considered.

Interpolation functions A two-parametric continuous probability function, the Weibull distribution, mainly used in the wind power industry, can be used to describe wind speed distribution (70; 86). The equation of the corresponding probability density function is given in equation 3.1

$$f(v) = \frac{k}{\lambda}\left(\frac{v}{\lambda}\right)^{k-1} e^{-(v/\lambda)^k} \tag{3.1}$$

where v is the wind velocity, k is the shape parameter and λ is the scale parameter of the distribution, with k and λ being positive.

For the purpose of this study, an original 5-parametric continuous function was built to determine the "detailed" wind discretization based on the "basic" 4-velocity-range wind discretization. This function, called Sigmoid function, based on the composition of two sigmoid functions, is given in equation 3.2. The two functions will be compared in the results section.

$$f(v) = \alpha.\left(-1 + \frac{1}{1+\beta_1.e^{-\gamma_1.v}} + \frac{1}{1+\beta_2.e^{\gamma_2.v}}\right) \tag{3.2}$$

where α, β_1, β_2, γ_1 and γ_2 are positive parameters.

3.2.2 Numerical model

Simulations were performed using the unsteady and incompressible solver *pimpleFoam* from OpenFOAM 6.0. A Reynolds-Averaged Navier-Stokes (RANS) methodology was used to solve the Navier-Stokes equations with the RNG k-ε turbulence model, and the transport of particulate matter was performed using a transport equation. This solver was validated previously in (122).

The area chosen to illustrate the methodologies discussed in this paper is located in Schiltigheim, France (48°36'24", 7°44'00"), a few kilometers north of Strasbourg. This area, as well as the only road considered as an emission source in this study (D120, rue de la Paix), are illustrated in Figure 3.4 (A). PM_{10} traffic-related emissions were estimated at 1.39 mg/s using daily annual mean traffic and were applied along the street considering its length in the numerical domain (200 m), its width (9 m) and an emission height of 0.5 m to take into account initial dispersion.

The recommendations given in (42) were followed. In particular, with H being the highest building height (16 m), the distances between the buildings and the lateral boundaries are at least $5H$, the distances between the inlet and the buildings as well as for the outlet and the buildings are at least $5H$ and the domain height is around $6H$. An illustration of the resulting 3D sketch is presented in Figure 3.4 (B). A grid sensitivity test was performed and showed that hexahedral meshes of 1 m in the study area and 0.5 m near the building walls are sufficient, leading to a more comparable resolution than other CFD studies (17) and leading to a total number of around 800,000 cells. The resulting mesh is illustrated in Figure 3.5.

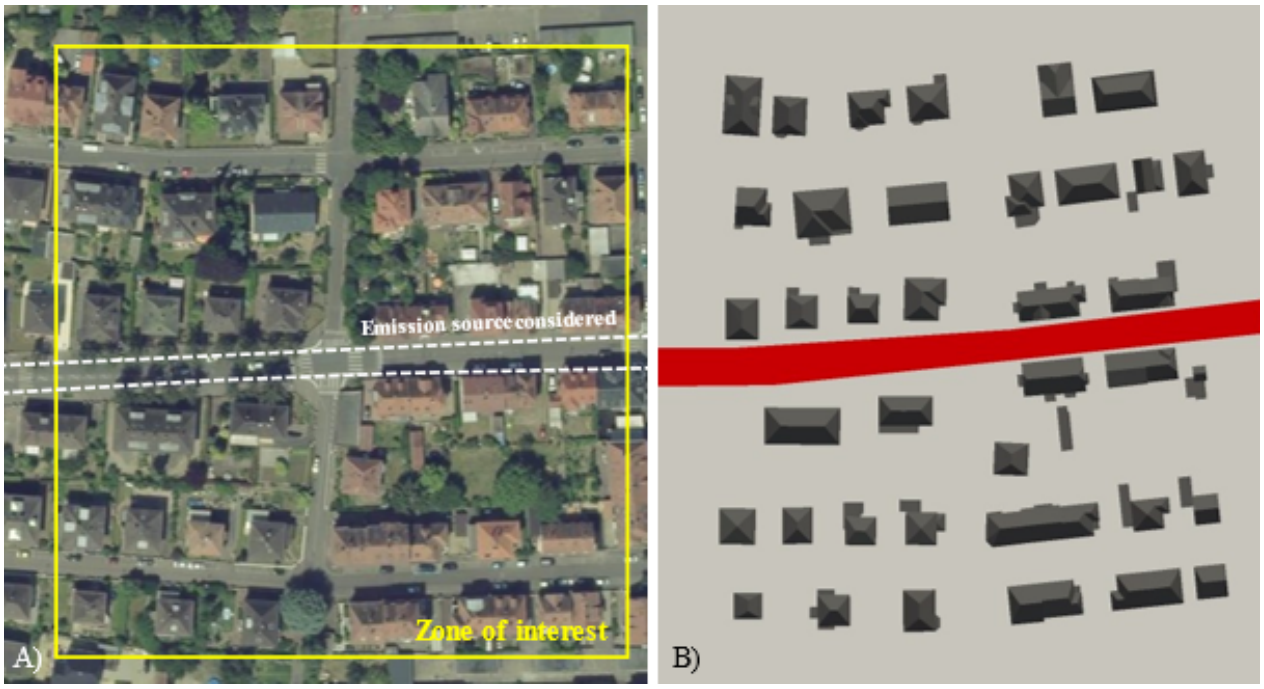


Figure 3.4: : Illustration of (A) the area of Strasbourg modeled with the road considered for the traffic-related emissions (white dashed lines), and (B) the corresponding area built in 3D for the numerical simulations with the emission source (red).

No-slip conditions ($U = 0$ m/s) were applied to the building walls and ground, and symmetry conditions to the lateral and the top boundaries. A freestream condition was applied to the outlet boundary, and neutral velocity, turbulent kinetic energy and turbulent dissipation profiles suggested in (125) were applied to the inlet boundary.

A total of 18 simulations were performed using the same wind velocity ($U_{10\ m} = 1.5$ m/s) but with different wind directions from 0° to 340° using a 20° step. Since the simulations were performed in neutral conditions and without traffic-induced turbulence, the dimensionless concentration C^* given in equation 3.3 is a function only of the wind direction (142). In other words, this means that considering the previous hypothesis, and for a given emission and building configuration (leading to constant $H.L/q$ ratio), only one simulation is needed for each wind direction simulated. The pollutant concentrations for a non-simulated wind velocity u can therefore be computed using:

$$C^* = \frac{C.U.H.L}{q} \tag{3.3}$$

where C^* is the dimensionless concentration, C is the concentration, U the wind velocity, H the characteristic building height and q/L the source strength of emission.

$$C_u = U_{ref}.\frac{C_{ref}}{u} \tag{3.4}$$

where C_u is the pollutant concentration for the wind velocity u not simulated and C_{ref} the pollutant concentration for the simulated wind velocity U_{ref}.

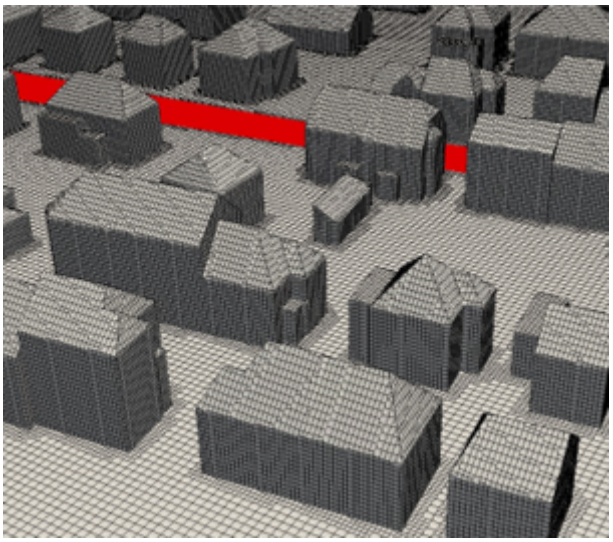


Figure 3.5: Illustration of the meshes in the computational domain with the emission source (red), with 0.5 m meshes near the buildings and 1 m in the study area.

3.3 Results

3.3.1 Wind data interpolation

Comparison between the Weibull distribution and the sigmoid function The best fitting parameters of the two functions were determined for the whole dataset using a non-linear solver and the "basic" 4-velocity-range wind data. The solver was set up to solve equation 3.5 for the four-velocity ranges [0, 1.5[, [1.5, 4.5[, [4.5, 8[and [8, $+\infty$[for both Weibull and sigmoid functions. This equation reflects that the sum of the frequencies between two wind velocities (i.e. the area under the curve) must be equal to the frequency given in the "basic" 4-velocity-range wind data. Since the sigmoid function has five parameters, a fifth equation to be solved was added only for this function and corresponds to 3.6. With this equation, it is assumed that the wind frequency tends toward 0% when the wind speed tends toward 0 m/s, as for the Weibull distribution.

$$\int_a^b f(v).dv = FVR_{[a;b[} \tag{3.5}$$

$$f(0) = 0 \tag{3.6}$$

where $f(v)$ is the Weibull or the sigmoid function and $FVR_{[a;b[}$ is the wind frequency given in the 4-velocity-range data for wind velocities ranging from a included to b excluded.

Figure 3.6 (A–D) shows a comparison between the Weibull distribution, the sigmoid function and the "detailed" 18-velocity-range data for one wind direction of each meteorological station. According to these figures, the two functions generally give the same trends, and both appear to give a good estimation of the "detailed" wind data. However, depending on the case, the Weibull function can provide improvements in comparison to the sigmoid function, as in Figure 3.6. (A), or vice versa, the sigmoid function can provide improvements in comparison to the Weibull function, as in Figure 3.6 (D).

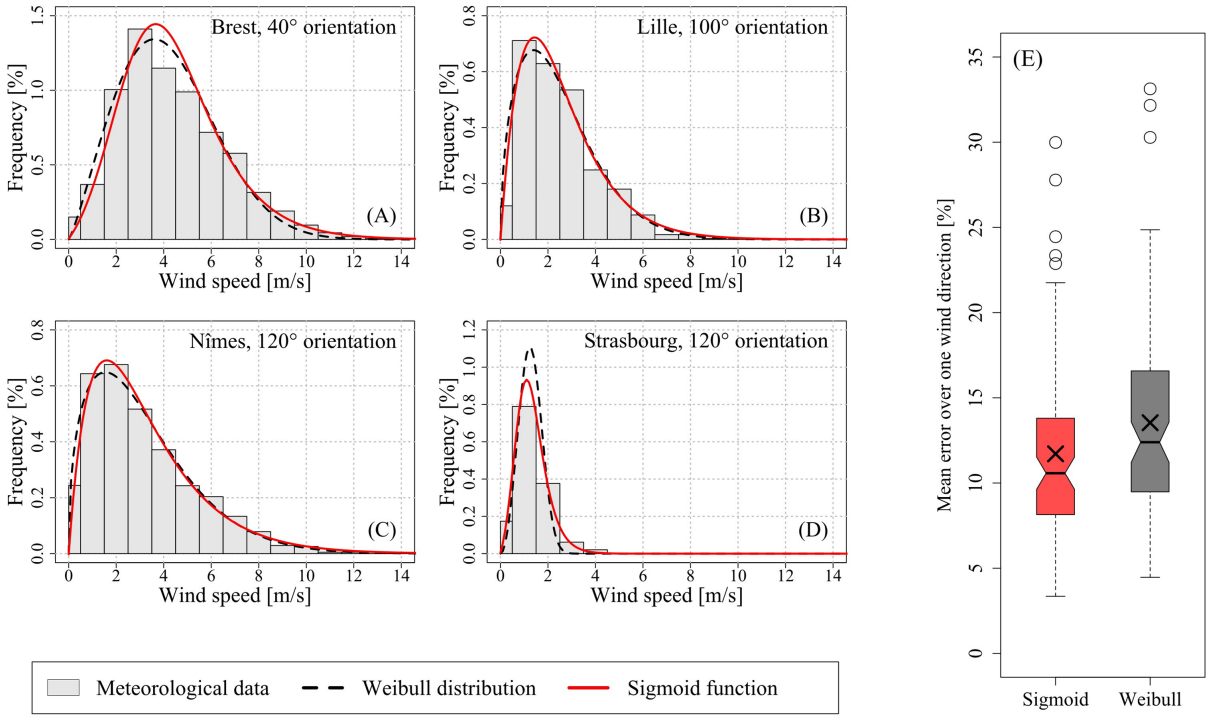


Figure 3.6: (A–D) Weibull distribution and sigmoid function results compared to the detailed meteorological wind frequency data for one wind direction at each station considered and (E) a notched box plot of the mean error over one wind direction with all stations included for both functions.

To better compare the two functions, a notched box plot of the mean error over one wind direction is given in Figure 3.6. (E). According to this figure, the sigmoid function gives generally better results compared to the Weibull distribution, with a lower maximal error (30.0% and 33.1% respectively); a lower first quartile (8.1% and 9.5% resp.); a lower third quartile (13.8% and 14.5% resp.); a lower mean (11.7% and 13.5% resp.); and a lower median (10.6% and 12.4% resp.). The differences are, however, small and may not be significant,

especially for the median because the notches slightly overlap. These differences between the Weibull distribution and the sigmoid function are also location dependent, with for example better prediction of the wind distribution in Strasbourg using the sigmoid function and an equivalent prediction in Brest. Finally, it should be noted that both functions can lead to underestimations of the lower wind velocity frequencies, as shown in Figure 3.6 (A) and (D).

According to the previous results, the Weibull distribution and the sigmoid function can accurately reproduce the "detailed" wind distribution based on a "basic" 4-velocity-range discretization with an average error of around 12% over the four stations considered in France. They can nonetheless lead to underestimations of the low wind velocity frequencies, for which the highest pollutant concentrations appear.

Optimization of the sigmoid function interpolation for low wind velocities The parametrization of the sigmoid function, called standard sigmoid function, was modified to improve the estimation of the low wind velocity frequencies in order to avoid underestimating pollutant concentrations.

Based on all the meteorological data considered in this study, it was found that the underestimation of low wind velocity frequencies occurs mostly when the frequency of the first velocity range is lower than the frequency of the second velocity range. In this specific case, the optimized sigmoid function still needs the equation 3.5 for the four-velocity ranges given in the "basic" wind data, but equation 3.6 is replaced by equation 3.7; otherwise, the previous parametrization using equations 3.5 and 3.6 is kept.

$$f(0) = FVR_{[0;\alpha[}\frac{FVR_{[0,\alpha[}}{FVR_{[\alpha,\beta[}} \tag{3.7}$$

where $FVR_{[0,\ \alpha[}$ is the wind frequency for the first range of velocities given in the 4-velocity-range data and $FVR_{[\alpha,\beta[}$ is the wind frequency for the second range of velocities (e.g., in this study $\alpha = 1.5$ and $\beta = 4.5$).

The methodology for the optimized sigmoid function is illustrated in Figure 3.7 (A–B): when the frequency of the first velocity range is higher than the second, as in Figure 3.7 (A1), the standard parametrization of the sigmoid function can be used because the low wind velocity frequencies are estimated accurately, as in Figure 3.7 (A2), when the frequency of the first velocity range is lower than the second, as in Figure 3.7 (B1), the standard parametrization leads to underestimations of low wind velocity frequencies and the optimized parametrization should be used instead, leading to a better estimation of the frequencies, as shown by the blue curve in Figure 3.7 (B2) compared to the red curve.

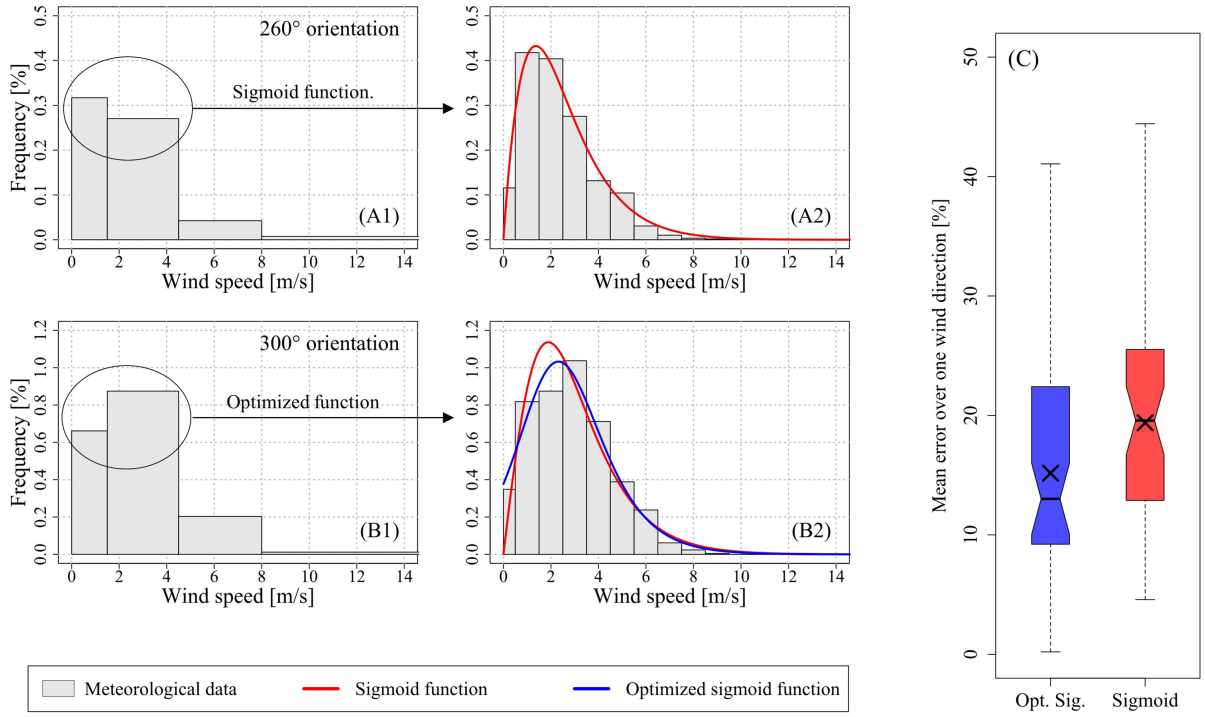


Figure 3.7: (A–B) Illustration of the optimized sigmoid function methodology and (C) comparison with the standard sigmoid function results.

The improvements with the optimized sigmoid function compared to the standard function was assessed and the results are presented in Figure 3.7 (C). For this comparison, only the wind directions where the optimized function was applied are considered (49 wind directions within the 78 previously used) and the errors compared to the "detailed" 18-velocity-range data were calculated for the low wind velocity frequencies (between 0 and 3.5 m/s). According to this figure, the optimized sigmoid function gives improvements over the standard sigmoid function with a lower maximal error (41.0% and 44.4% respectively); a lower first quartile (9.2% and 12.9% resp.); a lower third quartile (22.4% and 25.5% resp.); a lower mean error (15.2% and 19.4% resp.); and a lower median (13.0% and 19.6% resp.). The improvements using the optimized function are significative, in particular for the median since the box plot notches do not overlap; they are also location dependent. A global improvement of the wind distribution prediction ranging between 20% and 45% is observed in Strasbourg, Lille and Nîmes while no improvement is observed in Brest.

According to the previous results, using the optimized sigmoid function can improve the reproduction of the "detailed" wind distribution based on a "basic" 4-velocity-range compared to the standard sigmoid function, especially for low wind velocities.

3.3.2 Mean annual concentration assessment

Discrete methodology with intermediate velocities Initially, mean annual concentrations based on the CFD results can be calculated using a discrete methodology. This methodology considers that the mean annual concentration at a given location is composed of several small contributions of different wind velocities and wind directions. The mean concentration over one wind direction can be calculated with equation 3.8 and the mean annual concentration with equation 3.9. A similar methodology can be found in (146) and (127).

$$\overline{C}_d = \frac{\sum_{r=1}^{n} C_{d,r} . f_{d,r}}{\sum_{r=1}^{n} f_{d,r}} + C_{bg} \tag{3.8}$$

$$\overline{C} = \frac{\sum_{i=1}^{n} \overline{C}_d . f_d}{\sum_{i=1}^{n} f_d} \tag{3.9}$$

where $\overline{C}_d$ is the mean concentration over one wind direction, $C_{d,r}$ is the concentration for a given wind direction d and a given wind velocity range r, $f_{d,r}$ is the frequency for a given wind direction and a given wind velocity range, C_{bg} is the background concentration, $\overline{C}$ is the mean annual concentration and f_d the total frequency of a given wind direction.

With this methodology, it is necessary to choose a wind velocity in each velocity range for which the concentration will be calculated based on the CFD result. A simple choice is to consider an intermediate velocity, noted v_i, corresponding to the average between the minimal and the maximal value of the velocity range (e.g., for the velocity range [1.5, 4.5[, the intermediate value is 3 m/s).

A comparison of results for this methodology is given in Figure 3.8 with distinct cases considering (A) the "basic" 4-velocity-range frequencies, (B) the "detailed" 18-velocity-range frequencies, (C) the frequencies calculated with the sigmoid function, and (D) the frequencies calculated with the optimized sigmoid function. No background concentration is considered in this study to permit better comparison of the results and the CFD results used as inputs for the methodologies were strictly the same.

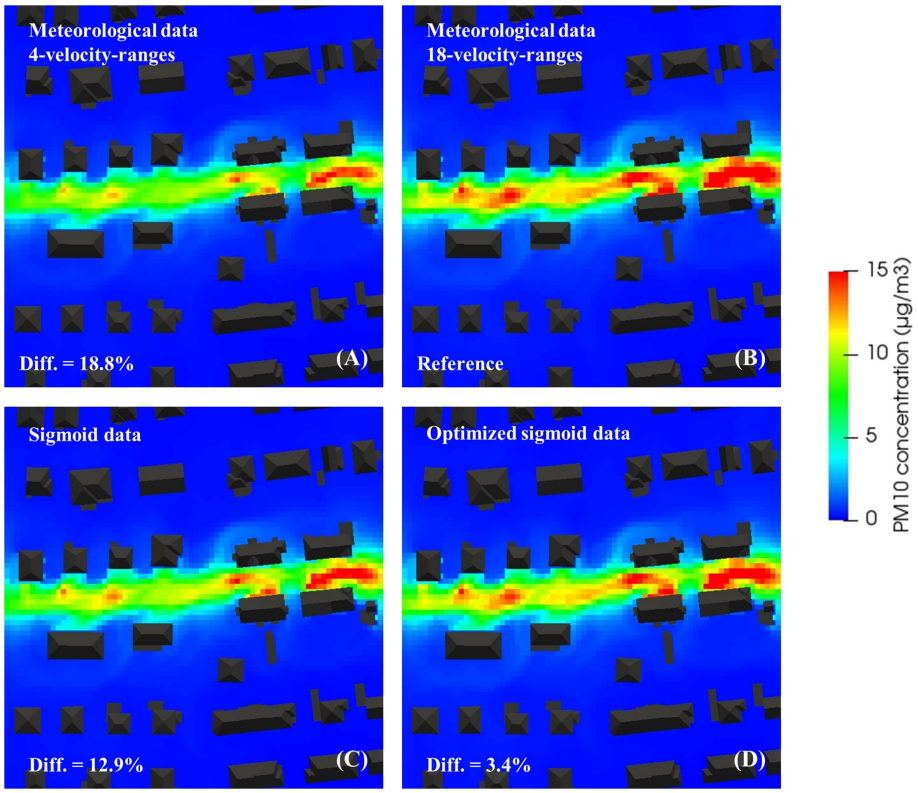


Figure 3.8: Mean annual concentrations without background concentration based on (A) the "basic" 4-velocity-range monitoring data, (B) the "detailed" 18-velocity-range monitoring data, (C) the sigmoid interpolation data and (D) the optimized sigmoid interpolation data.

Initially, it can be seen that using the "basic" 4-velocity-range data leads to an underestimation of the concentrations compared to the case using "detailed" 18-velocity-range data by around 19%. When calculating the "detailed" wind velocity distribution based on the "basic" data with the sigmoid function, the difference is reduced to 12.9%. Finally, the best results are obtained when using the optimized sigmoid function with an underestimation of 3.4%. According to these results, using the "basic" 4-velocity-range frequencies can give an estimation of the mean annual concentrations but is not sufficient to reach good accuracy compared to the mean annual concentration calculated with the "detailed" wind velocity distribution. However, using the sigmoid function and especially the optimized variant significantly improves the results, leading to almost the same results as those obtained with the "detailed" wind velocity distribution.

Discrete methodology with representative velocities The previous methodology used to compute annual concentrations, which was easy to set up, nonetheless has certain weak-

nesses that mostly concern the choice of the wind velocity for which the concentrations will be calculated, based on the CFD results. Using an intermediate velocity v_i corresponding to the average between the minimal and the maximal value of the velocity range can lead to underestimations of the mean annual concentrations. Indeed, in doing so, it is implicitly assumed that the concentration is constant with the wind velocity in a given wind velocity range. However, the concentration is not constant within a velocity range, especially when this range is large. A function describing the evolution of the concentration depending on the wind speed is therefore needed. As an example, for neutral atmosphere usually assumed in CFD, the concentration evolves hyperbolically with velocity according to equation 3.4. The representative velocity over one velocity range, considering the hyperbolic evolution of the concentration, is given in equation 3.11 as a result of equations 3.10 and 3.4.

$$\frac{1}{2}\int_{v_{min}}^{v_{max}} c\left(v\right).dv = \int_{v_{min}}^{v_r} c\left(v\right).dv \tag{3.10}$$

$$v_r = \sqrt{\frac{2}{\frac{1}{{v_{max}}^2} + \frac{1}{{v_{min}}^2}}} \tag{3.11}$$

where v_{max} and v_{min} are respectively the maximal and the minimal velocities of the velocity range, v_r is the representative velocity of the velocity range and $c\left(v\right)$ the equation describing the evolution of the concentration as a function of the wind velocity, i.e. equation 3.4.

The representative velocities v_r were calculated with equation 3.11 and compared to the intermediate velocities v_i. It is noteworthy that for a velocity range with a minimal velocity of 0 m/s, it is mathematically not possible to compute the representative velocity due to the domain definition of the function. A choice is therefore required; for the purpose of this study, the same ratio v_r/v_i as for [0.5, 1.5[was considered.

According to the results summarized in Table 3.2 for wind velocities ranging from 0 to 6.5 m/s, the intermediate velocity can be much higher than the representative velocity for low velocities. For example, for wind velocities ranging from 0.5 to 1.5 m/s, the intermediate velocity of 1 m/s is almost twice as high as the representative velocity of 0.67 m/s. For higher velocity ranges, such as [2.5, 3.5[or more, the differences can be neglected. This last statement is true for 1 m/s steps between the minimal and the maximal velocities of the velocity range but can become wrong for higher velocity steps.

Figure 3.9 shows a comparison of the mean annual concentrations when using the intermediate velocity and when using the representative velocity, based on the "detailed" 18-velocity-range wind distribution. According to the results, using the intermediate velocity leads to considerable underestimations of the mean annual concentrations compared to the use of the representative velocity. The underestimation is about 20%. When using the discrete meth-

$[v_{min}, v_{max}[$	[0,0.5[	[0.5,1.5[	[1.5,2.5[	[2.5,3.5[	[3.5,4.5[	[4.5,5.5[	[5.5,6.5[
v_i [m/s]	0.25	1.00	2.00	3.00	4.00	5.00	6.00
v_r [m/s]	0.17*	0.67	1.82	2.88	3.90	4.92	5.94
v_r/v_i	0.67*	0.67	0.91	0.96	0.97	0.98	0.99

Table 3.2: Comparison between the intermediate velocity v_i and the representative velocity v_r (* the representative velocity was calculated considering the same ratio v_r/v_i as for [0.5, 1.5[ ).

odology presented in the previous section, it is therefore suggested to use the representative velocity instead of the intermediate velocity to better take into account the hyperbolic evolution of the pollutant concentrations with the wind velocity to avoid underestimating the concentrations.

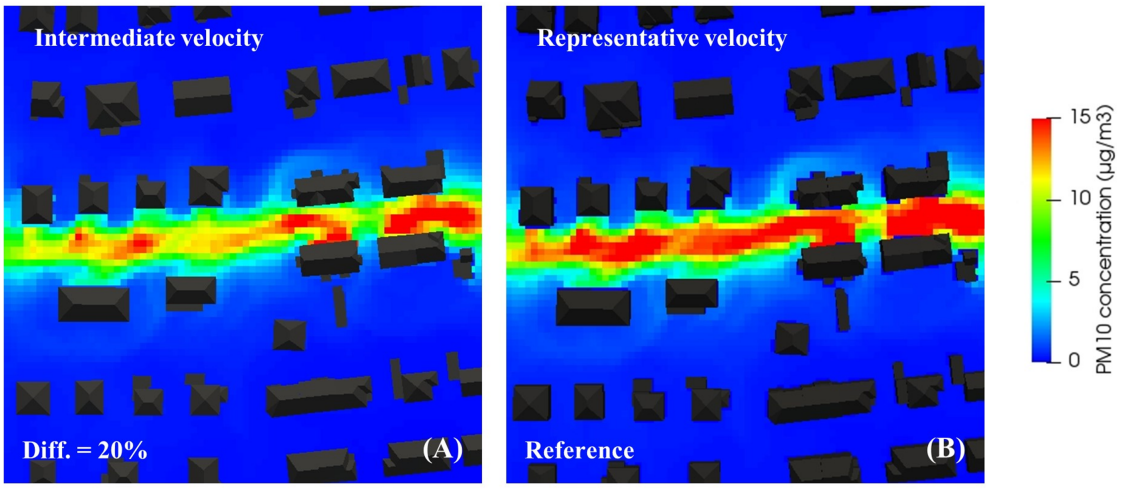


Figure 3.9: Comparison of the mean annual concentrations (A) based on the "detailed" 18-velocity-range wind distribution and using the intermediate velocity, and (B) based on the optimized sigmoid function and $v_{min} = 0.01$ m/s

Lastly, it should be noted that the representative velocities given previously were calculated with the assumption of equation 3.4 applied to equation 3.10. If the function describing the evolution of the concentration with the wind speed would change, e.g. for other types of numerical models or atmospheric conditions, equation 3.10 would need to be solved again with the new function to have a representative velocity adapted to the conditions and the numerical model considered.

Continuous methodology using the sigmoid function For the last approach, mean annual concentrations based on CFD results can be calculated using a continuous methodology. This methodology is a combination of equation 3.4, describing the evolution of pollutant concentration with wind velocity, and equation 3.2, describing the evolution of wind velocity frequency with wind velocity. The equation to compute the mean annual concentrations continuously is given in equation 3.12.

$$\overline{C_d} = \frac{\int_0^{+\infty} c\left(v\right).f\left(v\right).dv}{\int_0^{+\infty} f\left(v\right).dv} + C_{bg} \tag{3.12}$$

where $\overline{C}$ is the mean annual concentration, $c\left(v\right)$ is the function describing the evolution of the concentration with the wind velocity, $f\left(v\right)$ is the function describing the evolution of the wind velocity frequency with the wind velocity, and C_{bg} is the background concentration.

Taking equation 3.4 for $c\left(v\right)$ and equation 3.2 for $f\left(v\right)$ leads to a mathematical problem. Indeed, $c\left(v\right)$ is not defined for $v = 0$ and the limit of $c\left(v\right).f\left(v\right)$ tends toward infinity when v tends toward 0. To avoid this problem, equation 3.13 is suggested instead of equation 3.12. With this equation, it is considered that a minimal velocity (v_{min}) exists for which the pollutant concentration will no longer increase when the wind velocity decreases. This hypothesis can be justified by the additional effects, such as traffic-induced turbulence (157) and atmospheric stability (113) that may participate in pollutant dispersion for low wind velocities or become preponderant. We suggest applying a constant pollutant concentration for wind velocities ranging from 0 to v_{min} and suggest using $C_{max} = c(v_{min})$. The choice of v_{min} is particularly important when using the optimized sigmoid function.

$$\overline{C_d} = C_{max}.\frac{\int_0^{v_{min}} f\left(v\right).dv}{\int_0^{+\infty} f\left(v\right).dv} + \frac{\int_{v_{min}}^{+\infty} c\left(v\right).f\left(v\right).dv}{\int_0^{+\infty} f\left(v\right).dv} + C_{bg} \tag{3.13}$$

where $\overline{C_d}$ is the mean annual concentration, C_{max} is the maximal concentration accepted for the calculation, v_{min} is the velocity under which $c\left(v\right)$ is considered equal to C_{max}, $f\left(v\right)$ is equation 3.2, $c\left(v\right)$ is equation 3.4 and C_{bg} is the background concentration.

Figure 3.10 shows a comparison between the discrete methodology with the representative velocities and the continuous methodology using the optimized sigmoid function. It can be seen that the results of the discrete methodology given in Figure 3.10 (A) can be reached by the continuous methodology. Nonetheless, the difference of 5% reached using $v_{min} = 0.01$ m/s can increase when changing the value of v_{min}: lower values will lead to higher concentrations whereas higher values will lead to lower concentrations. The value of v_{min} must therefore be chosen carefully.

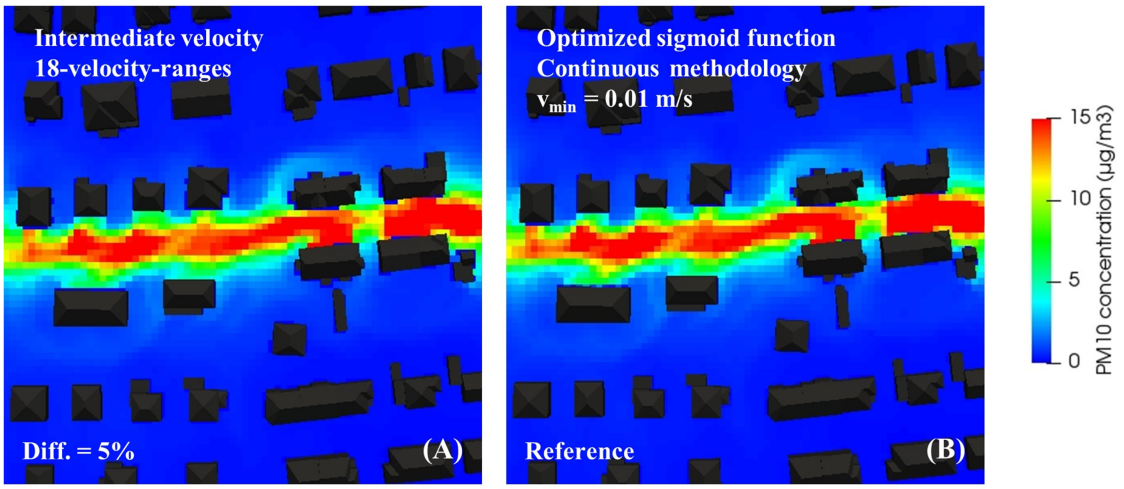


Figure 3.10: Comparison of the mean annual concentrations based on the "detailed" 18-velocity-range wind distribution using (A) the intermediate velocity and (B) the representative velocity.

3.4 Discussion

This study provides tools to assess wind velocity distributions based on "basic" data and mean annual air pollutant concentrations based on CFD results. Additional work should be done to improve the methodologies and the major issues are discussed hereafter.

The capability of the Weibull and the sigmoid functions to describe wind velocity distribution was assessed based on wind data from four meteorological stations in France. All of these stations were located in peri-urban environments close to large French cities. It is necessary to take into account that the results, and especially the interpolation-related errors, might be different for other types of stations such as urban and rural stations, and for other countries with different wind characteristics. In particular, the optimization suggested for the sigmoid function may not be suitable for different countries or type of station. Further works are therefore required in this direction.

The mean annual atmospheric pollutant concentrations can be calculated using a discrete methodology (146; 127). However, this methodology has two major problems. The first concerns the choice of wind velocity for which the pollutant concentrations will be calculated: choosing an intermediate velocity is a simple approach which can lead to considerable underestimations of pollutant concentrations, and it is better to use a representative velocity instead, as suggested in this paper. Using the representative velocity requires, however, making a choice for the first velocity range. The second problem concerns the velocity step used to build the wind velocity ranges: the result depends on the velocity step used, espe-

cially for the lower wind velocities for which a decrease in the velocity-step leads to higher mean annual concentrations. To avoid these two problems, a continuous methodology has been proposed. This methodology does not have an intrinsic limitation, but dependent on the function describing the evolution of the concentration as a function of wind velocity. If we consider a hyperbolic evolution of the concentration with wind velocity, it is necessary to choose a minimal value of velocity for which it is considered that lower velocities will not increase the concentrations due to compensatory phenomena (traffic-induced turbulence, atmospheric stability, etc.). The value of the minimal velocity is open to discussion and assessing this value is outside the scope of this paper. Further works are required, for example with infield measurement campaigns and comparisons between mean annual concentrations monitored and calculated with the continuous methodology. Lastly, two methodologies therefore exist, a discrete and a continuous with the discrete one being easier to implement in a code. However, we suggest using the continuous methodology if the user can describe the evolution of the concentration with the wind speed using a given piecewise continuous function. The discrete methodology can also be employed but, when an intermediate velocity is used, the user should be aware that the assumption of a constant pollutant concentration within velocity the range is made. To avoid this assumption, the user could consider a representative velocity instead, with as an example a linear evolution of the concentration between the limits of the velocity ranges.

Finally, it should be noted that the methodologies to assess mean annual concentrations were addressed using CFD results implying a neutral atmosphere, but can be used for any numerical results as long as a function describing the evolution of the concentration with the wind velocity is available.

3.5 Conclusion

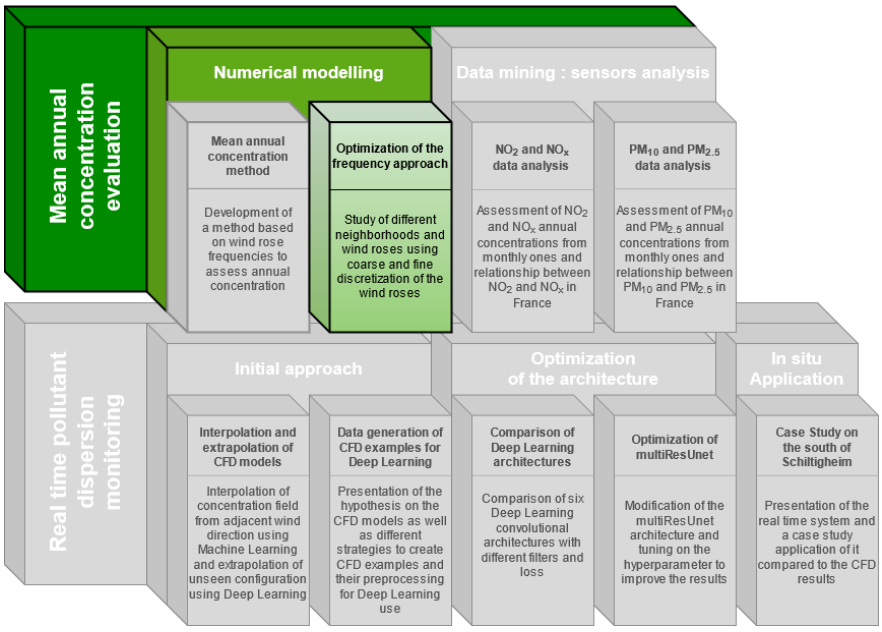
The objectives of this study were to provide methodologies to assess wind velocity distribution based on "basic" data, and to assess mean annual air pollutant concentrations based on numerical results. Three approaches for each objective were described and compared throughout this paper and the main conclusions are as follows:
(1.a) The Weibull distribution and the sigmoid function can both accurately reproduce "detailed" 18-velocity-range wind distribution based on "basic" 4-velocity-range wind data with an average error of 12%. These functions can nonetheless underestimate the frequencies of low velocities.
(1.b) The optimized sigmoid function improves the wind distribution results over the standard sigmoid function, especially for low wind velocities.
(2.a) Using "basic" 4-velocity-range wind data and the discrete methodology can provide an

estimation of the mean annual concentrations but is not sufficient to achieve high precision, leading to a difference of around 19% compared to the use of "detailed" 18-velocity-range wind data. Using the sigmoid function instead, based on the "basic" wind data improves the mean annual concentration results with a global error of less than 4%.
(2.b) When using the discrete methodology to assess mean annual concentrations, it is suggested to use a representative velocity of the function describing the evolution of pollutant concentrations with the wind velocities instead of an intermediate velocity. The intermediate velocity leads to underestimations of mean annual concentrations, especially when using CFD results with a neutral case hypothesis where the concentration evolves hyperbolically with the wind velocity.
(2.c) Mean annual concentrations can be assessed using a continuous methodology that does not have any of the limitations of discrete methodologies. It is, however, limited by the function describing the evolution of the concentrations with the wind velocities, which leads to the need to choose a minimal velocity when using the sigmoid function.

Finally, the methodologies presented in this paper can be used for outdoor air quality study purposes, which is a relevant starting point for improving both outdoor and indoor air quality and, therefore, a key-point to achieve smart sustainable cities. These results give insights to researchers and engineers on how to assess wind velocity distribution and mean annual concentrations for comparison with annual regulatory values given by the EU, the WHO or any other organization, and further works could be done to compare the results of the methodologies with monitored data.

Chapter 4

How to balance between modelling error and computational cost to assess mean annual concentration of a neighbourhood?



This chapter has been published in the journal *Sustainable cities and society* under the title "On the minimal wind directions required to assess mean annual air pollution concentration based on CFD results " (59).

In the previous Chapter 3, a methodology to determine air pollution was presented. This methodology is a statistical approach that use wind direction frequency and wind speed fre-

quency. Each wind direction is computed and depending on the assumptions that can be made, from one to several wind speed for each direction as well. Nevertheless, no specifications are made upon the discretisation of the wind rose. Thus a question arises: how many wind directions are necessary to determine the annual concentration? This section deals with this question by determining the numerical error made when using less than 18 directions and studies the best strategy to deal with it. Is it better to use the predominant wind directions or representing homogeneously all wind directions? The content is summarized in the graphical abstract given in Figure 4.1.

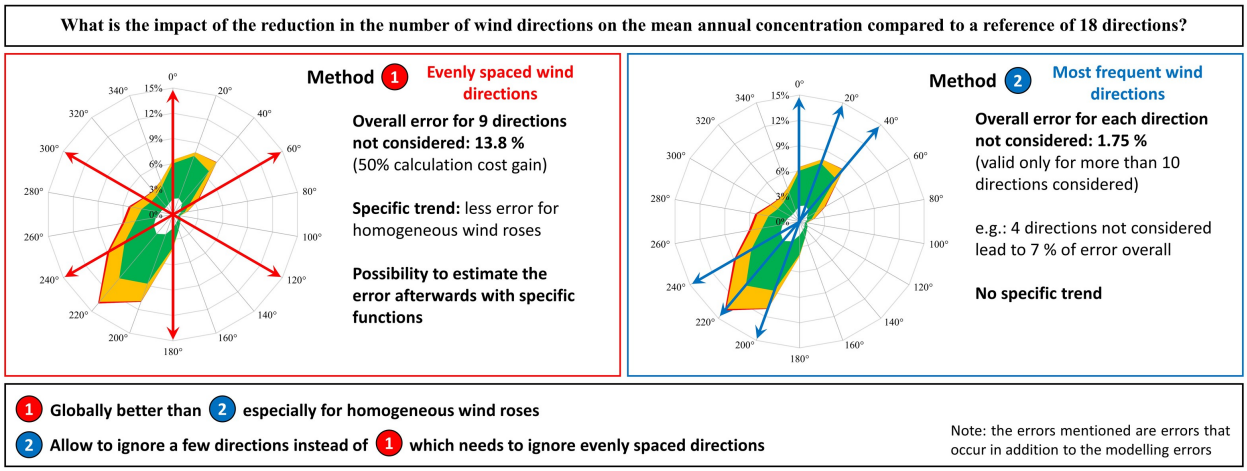


Figure 4.1: Graphical abstract of this chapter

4.1 Introduction

Among the numerous numerical models available to model air pollutant dispersion, computational fluid dynamics (CFD) has shown a great potential and a great interest from the scientific community given the many physical phenomena that can be considered. It includes notably the effects of vegetation on both airflow and pollutant deposition (23; 75; 141), the atmospheric chemistry involving nitrogen oxides (22; 137) as well as the effects of heat exchanges and solar radiation (8; 122; 154; 172). CFD models have already been used to assess annual concentrations (127; 159) and, additionally, a recent study has highlighted and discussed the different ways to assess annual concentrations based on numerical results and wind data (120)). However, these different studies always considered all the wind directions available in the wind rose which lead to a significant number of simulation to be performed. In view of the calculation time and, therefore, the calculation costs of CFD modelling, the question of reducing the number of wind directions to model in order to compute annual air pollutant concentrations is relevant.

The aim of the present work is to assess the possibility of limiting the number of wind directions needed to be modelled in order to compute annual air pollutant concentrations based on CFD results. Particularly, the novelties of this work reside on both quantitative and qualitative results using the methodology to compute mean annual concentrations presented by (120): questioning the discretization of wind roses which can change the results; allowing computing the mean annual concentration with fewer simulations to reduce the computational cost of a CFD study; challenging different ways of reducing the number of directions to compute annual concentration in an air quality CFD study; determining the order of magnitude of the additional error made by reducing the number of wind directions modelled for several building layouts and wind roses; a methodology to determine the error made once the chosen number of simulations is computed, thus enabling the user to see if the error is within a satisfying range or if it needs more directions to be modelled. To do so, different options are compared considering (1) different discretization steps in the wind directions and (2) the greatest contributions to the total wind frequencies. The meteorological data, the areas modelled, the CFD model used for the purpose of illustration and the methodology to compute the annual concentrations are presented in subsection 2. Then, the approaches to limit the number of wind directions needed are described and compared in subsection 3. Finally, a discussion is presented in subsection 4.

The novelties of this work reside on both quantitative and qualitative results using the methodology to compute mean annual concentrations from the method developed by (120): questioning the discretization of wind roses given by authorities which can change the results; allowing to compute the mean annual concentration with fewer simulations reducing the computational cost of a CFD study; challenging different ways of reducing the number of directions to compute annual concentration in an air quality CFD study; determining order of magnitude of the additional error made by reducing the number of wind directions modelled for several building layouts and wind roses; a methodology to determine the error made once the chosen number of simulations are done with its wind rose and building layouts. Thus, enabling the user to see if the error is within satisfying range or if he needs more directions for his case.

4.2 Material and methods

4.2.1 Meteorological data

The wind data used for the purpose of this work were obtained from five meteorological stations in France in Strasbourg, Brest, Nîmes, Lille and Paris respectively located in the extreme east, west, south, north and in the center of the country. These data were provided

by Météo-France, the french official climatology and meteorology service, and correspond to ten years of averaged data for the Strasbourg station (from 1999 to 2008) and twenty years of averaged data for the other stations (from 1999 to 2018). The corresponding wind roses are presented in Figure 4.2.

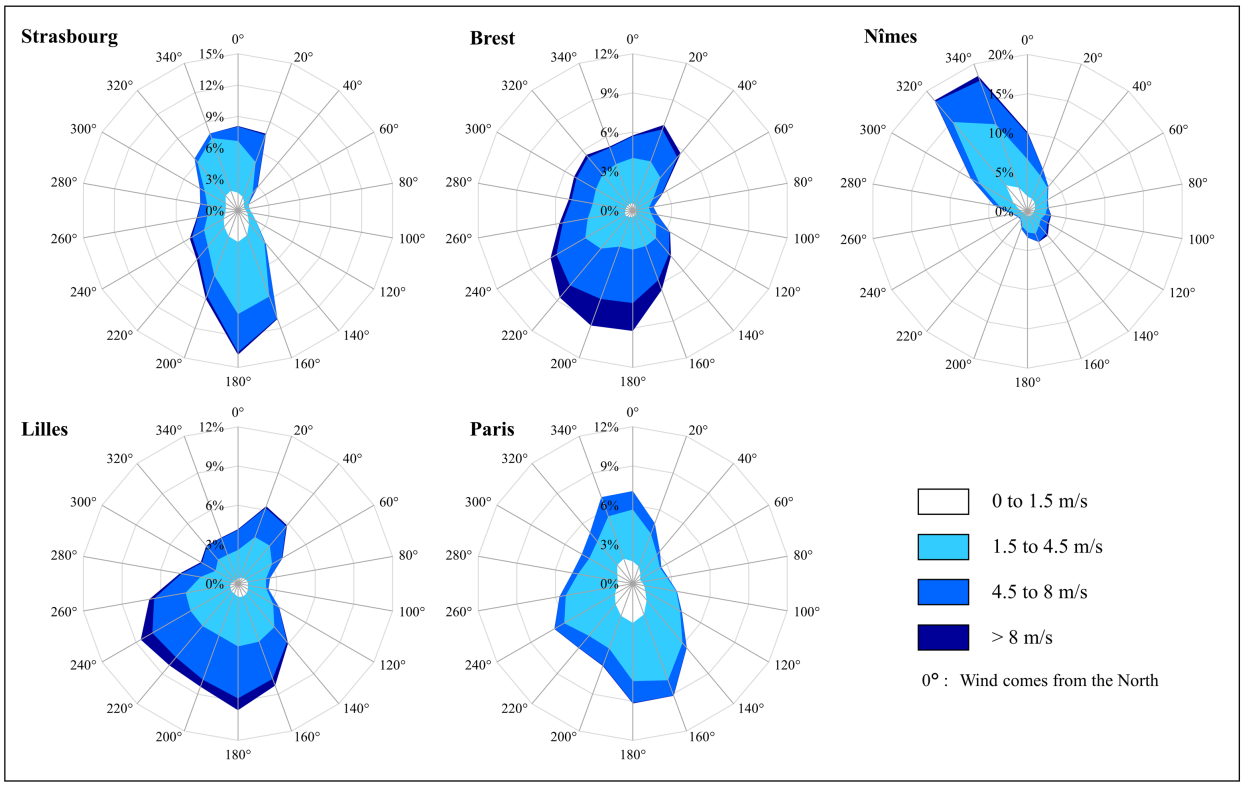


Figure 4.2: : Wind roses for the five meteorological stations considered.

These stations were chosen firstly to cover different wind types throughout France but also because of the differences observed in the wind data to improve the statistical independence of the results. Indeed, according to 4.2, the five wind roses are complementary with the station of Strasbourg having a preferential wind axis (North-South) with winds distributed in both directions while the station of Nîmes has only a preferential direction from the North-West. The other stations having finally no preferential direction but a greater variation in the share of velocities with the station of Paris having a majority of winds ranging from 1.5 to 4.5 m/s and the Brest Station having the greater frequency of winds higher than 8 m/s.

4.2.2 Numerical model

All the simulations were performed using the unsteady and incompressible pimpleFoam solver taken from the OpenFOAM 6.0 library, since unsteady simulations can improve the results for the concentration field over a steady state calculation (153). This solver was modified

Building layout	**N1**	**N2**	**N3**	**N4**	**N5**
Minimal height [m]	2	3	2	2	2
Mean height [m]	9	11	17	14	16
Maximal height [m]	11	14	26	21	22
Homogeneity	+	+	-	-	-

Table 4.1: Minimal, mean and maximal height of the buildings for the five building layouts considered.(+: homogeneous, -: heterogeneous).

to include an Eulerian passive scalar transport equation to account for pollutant dispersion, which is commonly used to model gaseous (140) or particulate matter (109) dispersion. It should be noted that, for some types of pollutants such as pollen, specific phenomenon needs to be considered (135) but is not of interest for the purpose of this work. The partial differential equations were solved using the Reynolds-Averaged Navier-Stokes (RANS) methodology and an RNG k-ε turbulence model (68; 107). This solver was previously validated in (120).

Seven different urban configurations were considered for this study: five real building layouts in Strasbourg city, noted N1 to N5, and seven road layouts (three different road layouts were applied to the N5 building layout). A top view of these different configurations is given in Figure 4.3 and additional information on building heights are given in Table 4.1.

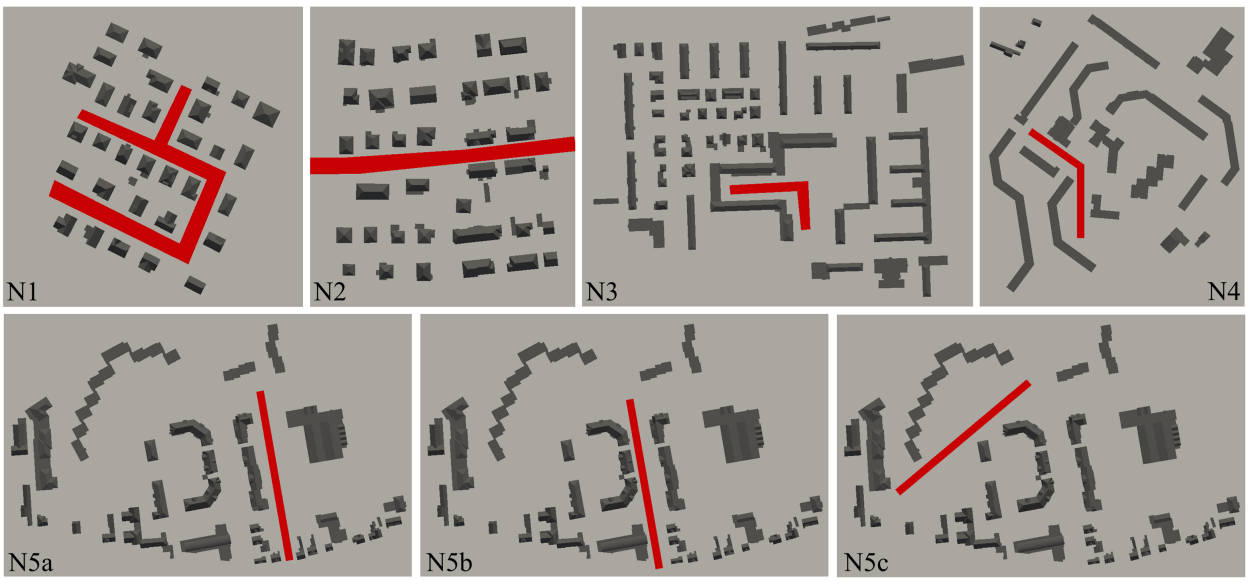


Figure 4.3: Top view of the five building layouts used in this study (N1 to N5) and the three road layouts used for the N5 case (N5a to N5c) with in red the roads considered as pollutant sources.

For each of these cases, the recommendations of the cost actions guidelines (42) were followed. For our computational domains, considering H the highest building height in each area considered, the distance between the inlet and the buildings is at least $5H$, which is also the minimal distance between the outlet boundary and the buildings, as well as between the buildings and the lateral boundaries. Lastly, the height of the computational domain was set to $6H$. After a grid sensitivity check, hexahedral meshes of 1 m in the areas of interest and 0.5 m both near the buildings wall boundaries and emission sources were used which corresponds to a comparable resolution of other studies (35; 136; 159). This resolution leads to a total number of cells ranging from 550,000 to 2.8 millions depending on the area considered and an example of the resulting meshes is presented in Figure 4.4 for the urban configurations and N2.

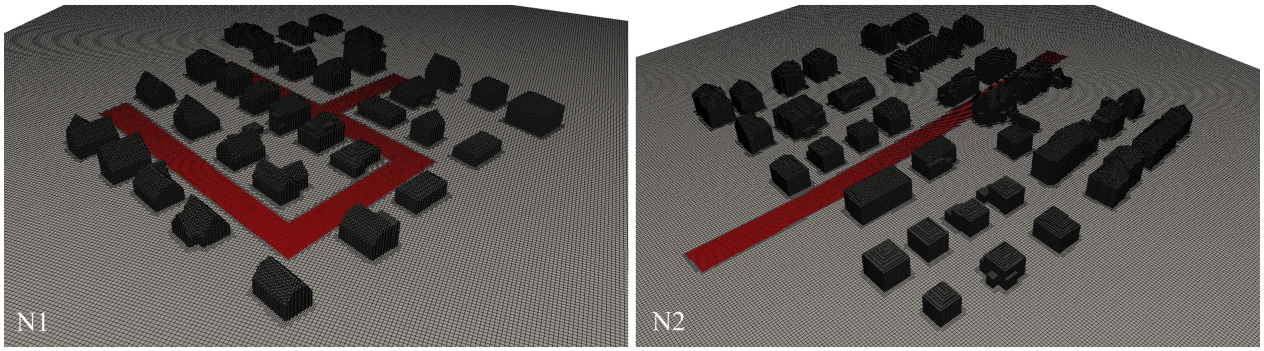


Figure 4.4: Illustration of the selected meshes for the N2 urban layouts (in red the roads considered as emission sources)

Concerning boundary conditions, symmetry conditions were applied at the top and the lateral boundaries when no-slip conditions were applied to the wall surfaces such as the building's walls or the ground. Neutral velocity, turbulent kinetic energy and turbulent dissipation rates profiles following the log-law profile suggested by (125) with a wind velocity of 1.5 m/s at 10 m high were used for the inlet boundary and a free stream condition was set at the outlet.

A total of 126 simulations were performed considering 18 wind directions (20° steps) and 7 urban configurations.

4.2.3 Annual concentration calculation

The annual air pollutant concentrations were calculated based on the 126 CFD results obtained and the continuous methodology suggested by (120). This methodology involves four equations which are given hereafter. Particularly, it corresponds to 4.1 the equation of the optimized sigmoid function used to describe the wind distribution based on the wind rose

data, 4.2 the equation of the evolution of the CFD modelled concentration with the wind velocity for neutral atmospheres, 4.3 the equation to compute the mean annual concentration for a given wind direction and 4.4 the equation to compute the mean annual concentration. Further details on how to apply this methodology and these equations can be found in the original paper of (120).

$$f(v) = \alpha . \left(-1 + \frac{1}{1 + \beta_1 . e^{-\gamma_1 . v}} + \frac{1}{1 + \beta_2 . e^{\gamma_2 . v}} \right) \tag{4.1}$$

where α, β_1, β_2, γ_1 and γ_2 are positive parameters.

$$C_u = U_{ref} . \frac{C_{ref}}{u} \tag{4.2}$$

where C_u is the pollutant concentration for the wind velocity u not simulated and C_{ref} the pollutant concentration for the simulated wind velocity U_{ref} (1.5 m/s at 10 m high).

$$\overline{C}_d = C_{max} . \frac{\int_0^{v_{min}} f(v) . dv}{\int_0^{+\infty} f(v) . dv} + \frac{\int_{v_{min}}^{+\infty} c(v) . f(v) . dv}{\int_0^{+\infty} f(v) . dv} + C_{bg} \tag{4.3}$$

$$\overline{C} = \frac{\sum_{d=1}^{n} \overline{C}_d . f_d}{\sum_{d=1}^{n} f_d} \tag{4.4}$$

where $\overline{C}_d$ is the mean annual concentration for a given wind direction, C_{max} is the maximal concentration accepted for the calculation, v_{min} is the velocity under which $c(v)$ is considered equal to C_{max}, $f(v)$ is equation 4.1, $c(v)$ is equation 4.2, C_{bg} is the background concentration, $\overline{C}$ is the mean annual concentration and f_d the total frequency of a given wind direction.

When using this continuous methodology, it is necessary to define a minimal velocity (v_{min}) for which a constant pollutant concentration (C_{max}) will be applied, considering that the pollutant concentration will not increase indefinitely with the decrease of the wind velocity but reach a threshold due to numerous new phenomena such as vehicle-induced turbulence or natural convection (120). For the purpose of this work, v_{min} was set to 1.1 m/s since it corresponds to a low wind speed were additional turbulence due to traffic start to be as important as wind speed turbulence (157). Lastly, C_{max} was calculated according to equation 4.2, with $u = v_{min}$.

For the purpose of this study, no background concentration was considered.

4.2.4 Comparison cases considered in this study

The continuous methodology described previously was applied to all wind roses and all areas considered, leading to a total of $5 \times 7 = 35$ results which are considered as the reference results.

Two approaches were studied to limit the number of wind directions needed to be modelled in order to compute annual air pollutant concentrations based on CFD results: 4.1 ignoring some wind directions with a regular step (e.g. considering only one wind direction out of two) and 4.2 considering the predominant wind directions (e.g. considering the first ten wind directions with the greatest contributions to the total wind frequency). The results of these methodologies compared to the reference results are given in the subsection Results. A comparison between the two methodologies is also provided. The errors discussed in this paper are the error between the CFD reference results considering the whole wind rose (18 directions) and the various presented methods. Thus, it is not a comparison with the error made by CFD compared with real in situ values which is another matter entirely as discussed in (127) in which they reach less than 30% error concentrations without consideration of chemical mechanisms.

4.3 Results

4.3.1 Annual concentration calculation when ignoring wind directions with a regular step

The first approach considered to decrease the number of simulations for annual concentration calculation consists in ignoring some wind directions with a regular step. In particular, we try to consider one direction out of two, three, six and nine. By doing so, the annual concentrations are calculated considering 9, 6, 3 and 2 wind directions respectively, as shown in Figure 4.5. It should be noted that depending on the number of directions considered, a more or less important number of possibilities do exist leading to various starting directions (considering one direction out of two leads to two possible starting directions: 0° and 20°, considering one direction out of three leads to three possible starting directions: 0°, 20° and 40°, etc.). The results were then compared with the reference annual concentrations obtained considering the whole wind rose, thus, 18 wind directions.

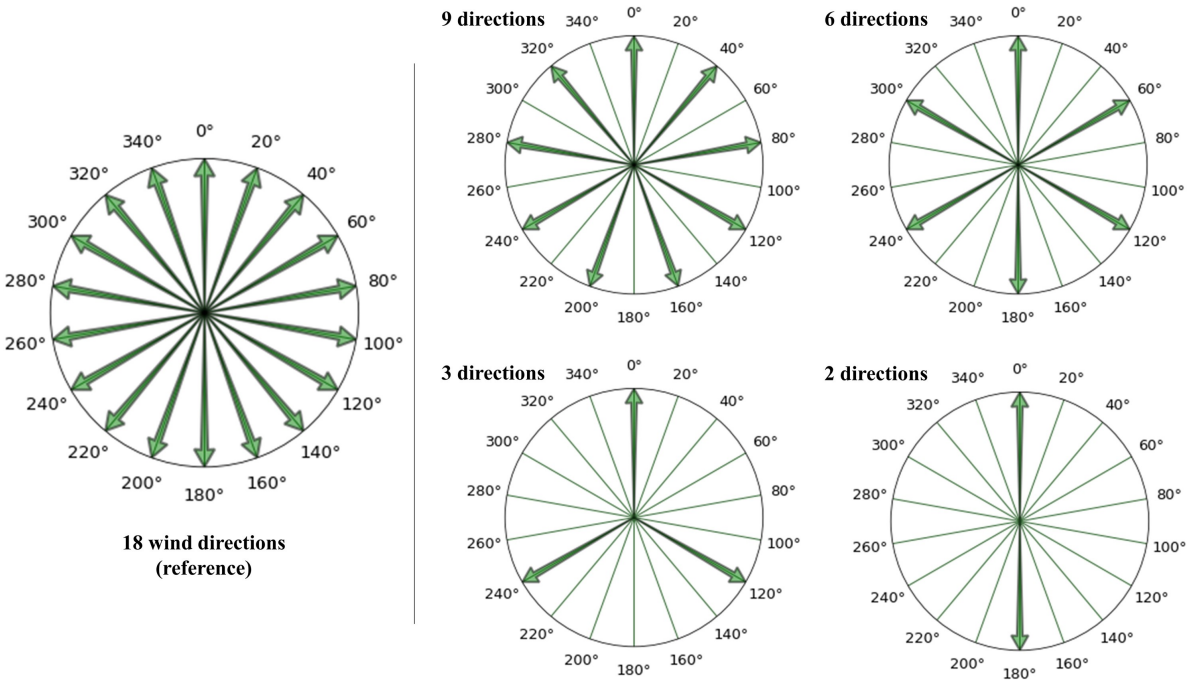


Figure 4.5: Explanation on what is referred as 18, 9, 6, 3 and 2 wind directions with regular steps for annual concentration calculation.

A first comparison is given in Figure 4.6 corresponding to the building layout N1 and the wind rose from Paris considering one direction (B1) out of two, (B2) out of three, (B3) out of six and (B4) out of nine. In spite of some local variations, it can be seen that the results obtained using 9 and 6 wind directions, respectively in Figure 4.6 (B1) and (B2) are close from the reference case which needed 18 wind directions. The results obtained with 2 and 3 wind directions, respectively in Figure 4.6 (B3) and (B4), seem more different from the reference result.

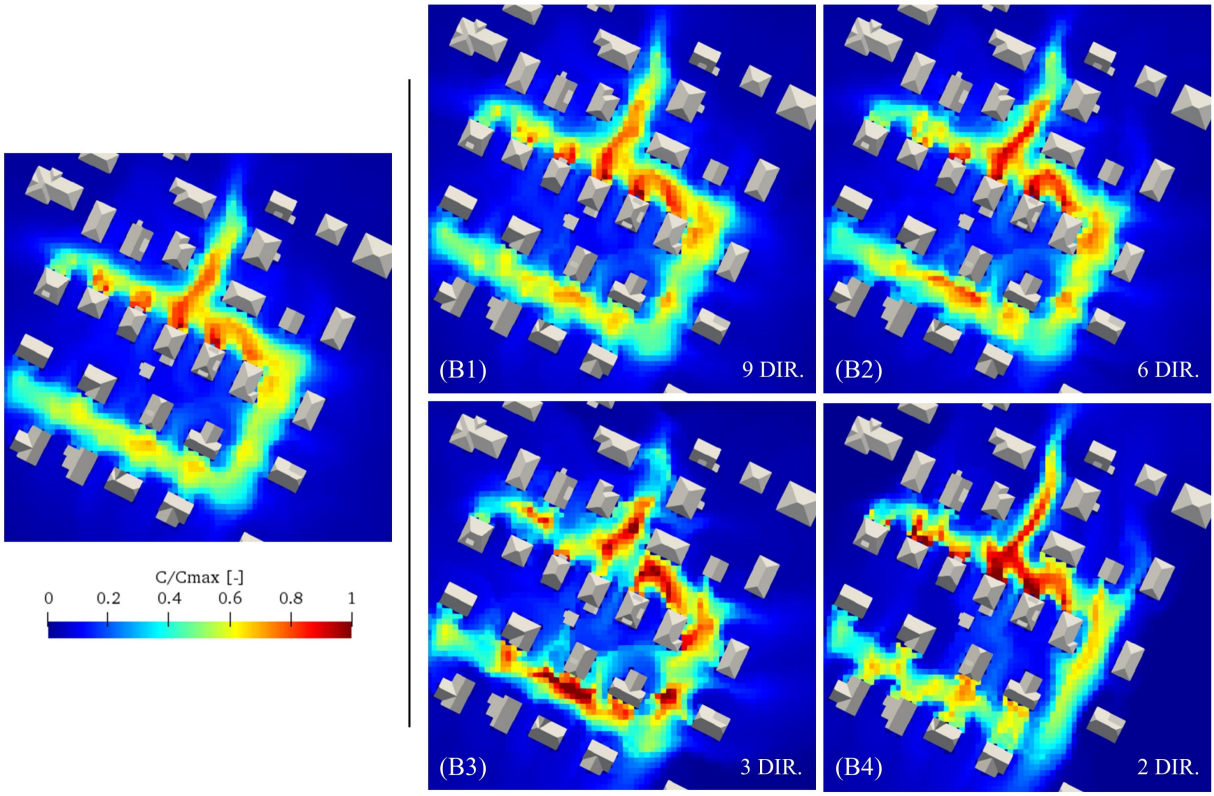


Figure 4.6: Examples of mean annual concentrations results for the building layout N1 and the wind rose of Paris using (A) the whole wind rose (reference), (B1) 9 directions, (B2) 6 directions, (B3) 3 directions and (B4) 2 directions.

In order to have a global information, global parameters were calculated. The overall results on mean error, mean relative standard deviation, calculation costs gain and the ratio between the gain and the error are given in Table 4.2 considering all the seven building layouts, the five wind roses, and the different starting directions. As previously observed, the best results compared to the reference are achieved considering one direction out of two which lead to an overall error of 13.8%. Decreasing the number of directions increases the error but considering one direction out of three lead to an error of around 21% on average. Finally, considering only two or three wind directions to compute the annual concentrations lead to high errors of more than 40%. Lastly, the best compromise between the gain in calculation costs and the induced error is obtained considering one direction out of two, with a corresponding ratio of 3.8 and a total gain of 50% with the assumption that all simulations have the same calculation cost.

Lastly, the influence of the building layout and the wind rose was assessed. The results are given in Figure 4.7 for the four cases considered (9, 6, 3 and 2 wind directions) with (A) the mean errors as a function of the wind rose and (B) as a function of the building layout.

Directions	Mean error (%)	Standard deviation	Cost gain (%)	Gain/Error
9	13.8	6.8	50	3.8
6	20.9	10.3	67	3.2
3	38.8	20.3	83	2.2
2	52.4	30.3	89	1.7

Table 4.2: Global results for annual concentration calculation using a regular wind direction step with the mean errors, the standard deviations, the calculation cost gain and the ratio between gain and error.

According to Figure 4.7 (A), the wind rose has an impact on the mean errors obtained with the four cases considered leading to an overall maximal variation of 1.7. As an example, considering one wind direction out of two (9 directions in total), an error of 10.0% is obtained with the wind rose of Brest and 17.8% with the one of Nîmes. If we consider the overall patterns of the wind roses (see Figure 4.2), the wind roses of Brest and Lille, homogeneous over wind direction and wind speed, lead to the minimal differences compared to the reference. Inversely, the wind roses of Nîmes and Strasbourg, with a preferential direction and more intermediate velocities (ranging between 1.5 and 4.5 m/s), lead to the maximal differences compared to the reference. Finally, an intermediate result is obtained with the wind rose of Paris, homogeneous but with more intermediate velocities. This observation is valid whether the case considered (9, 6, 3 or 2 wind directions). Such trends are not observed as a function of the building layout and, according to Figure 4.7 (B), an overall maximal variation of 3.9 is obtained which is higher than previously when making the comparison as a function of the wind roses. The results are hence more sensitive to the building layout than to the wind rose considered.

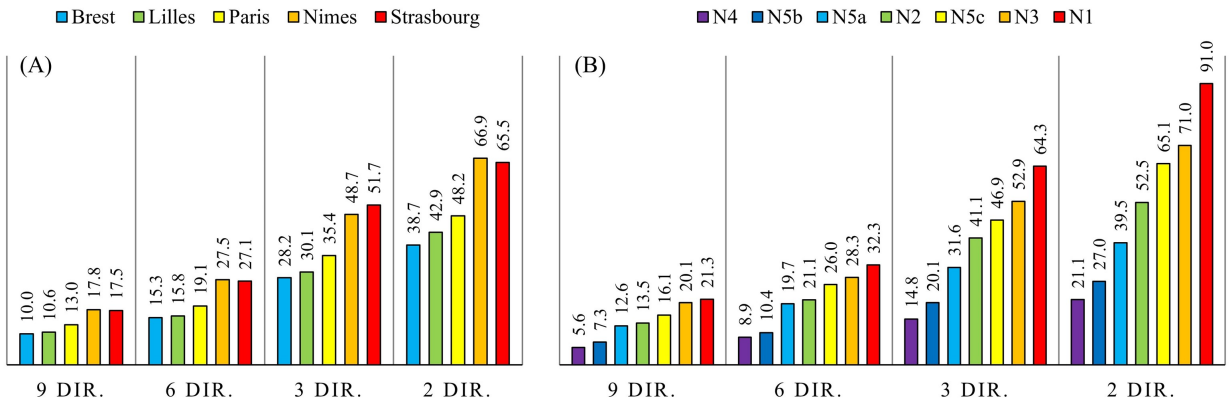


Figure 4.7: :Mean error over the mean annual concentration compared to the reference using a regular wind direction step as a function of (A) the wind rose location and (B) the building layout (DIR: Directions).

According to the previous results, calculating mean annual concentration by ignoring some wind directions with a regular step can lead to significant calculation cost reductions without leading to too much induced errors. It is particularly true when modelling one wind direction out of two, where calculation costs are reduced to 50% and an error of less than 20% can be expected (around 13.8% on average) whatever the wind rose or the building layout considered. Finally, the more a wind rose is homogeneous the smaller the error is.

4.3.2 Annual concentration calculation when considering the predominant wind directions

The second approach studied to decrease the number of simulations for annual concentration calculation is about considering the predominant wind directions. In particular, the first, the first two, the first three up to the first seventeen wind directions with the most occurrence frequencies were successively considered. The results were then compared again with the reference annual concentrations obtained considering the whole wind rose, thus, 18 wind directions.

A first comparison is given in Figure 4.8 corresponding to the building layout N1 and the wind rose from Paris and considering (B1) the first fifteen, (B2) the first nine, (B3) the first six and (B4) the first wind direction with the most occurrence frequency. In spite of some local variations, it can be seen that the results obtained using 15 and 9 wind directions, respectively in Figure 4.8 (B1) and (B2) are close from the reference case which needed 18 wind directions. The results obtained with less wind directions seems more different from the reference result, leading to higher local concentrations.

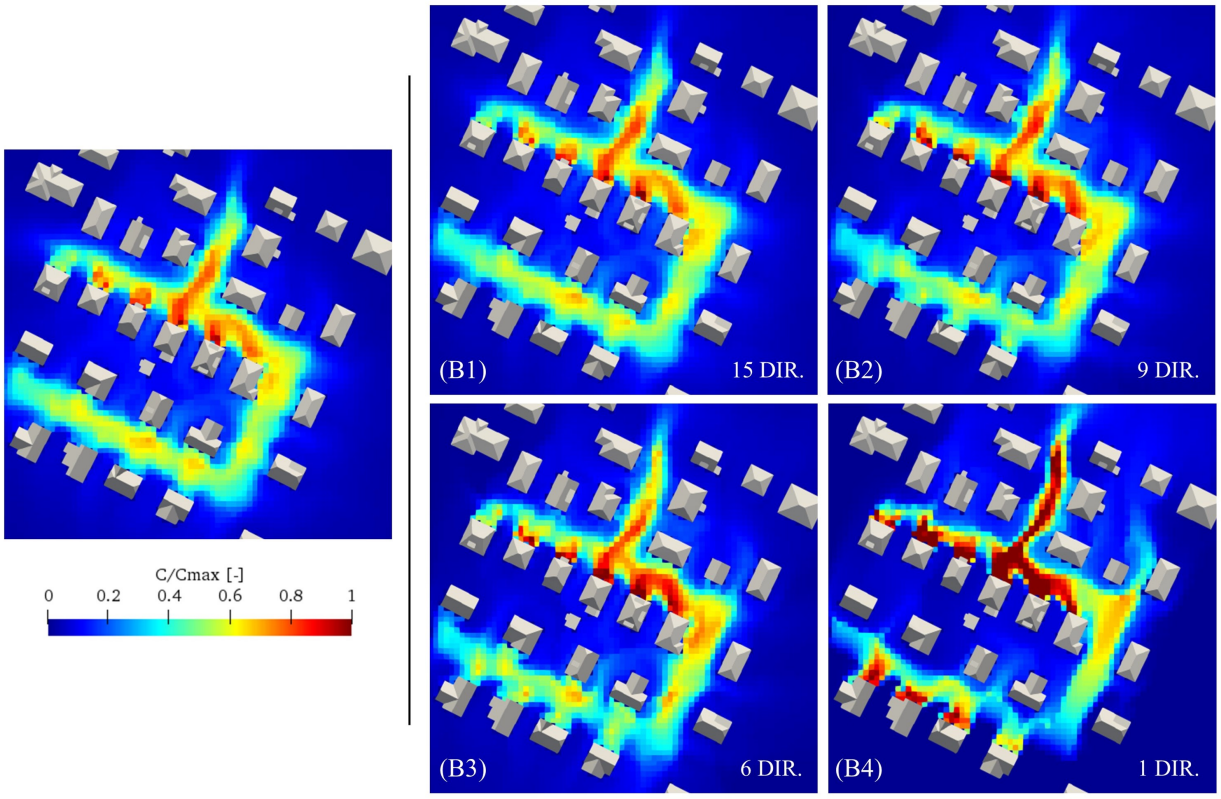


Figure 4.8: Examples of mean annual concentrations results for the building layout N1 and the wind rose of Paris using (A) whole wind rose (reference), (B1) 15 directions, (B2) 9 directions, (B3) 6 directions and (B4) 4 directions.

The evolution of the global mean error (considering all building layouts and wind roses) with their respective standard deviation is given in Figure 4.9. The gain in calculation cost and the ratio between gain and error are also plotted. According to this figure, the global evolution of the induced error while ignoring some wind directions seems to be linear between 10 and 18 wind directions considered to compute the mean annual concentration. In this case, around 1.75% of error is generated for each wind directions not considered. For fewer than 10 wind directions considered, the error starts evolving exponentially. The maximal value of the ratio between gain and error is reached for 15 wind directions with an overall value of 3.25 between 12 and 17 wind directions. This ratio starts decreasing linearly for fewer than 12 wind directions considered.

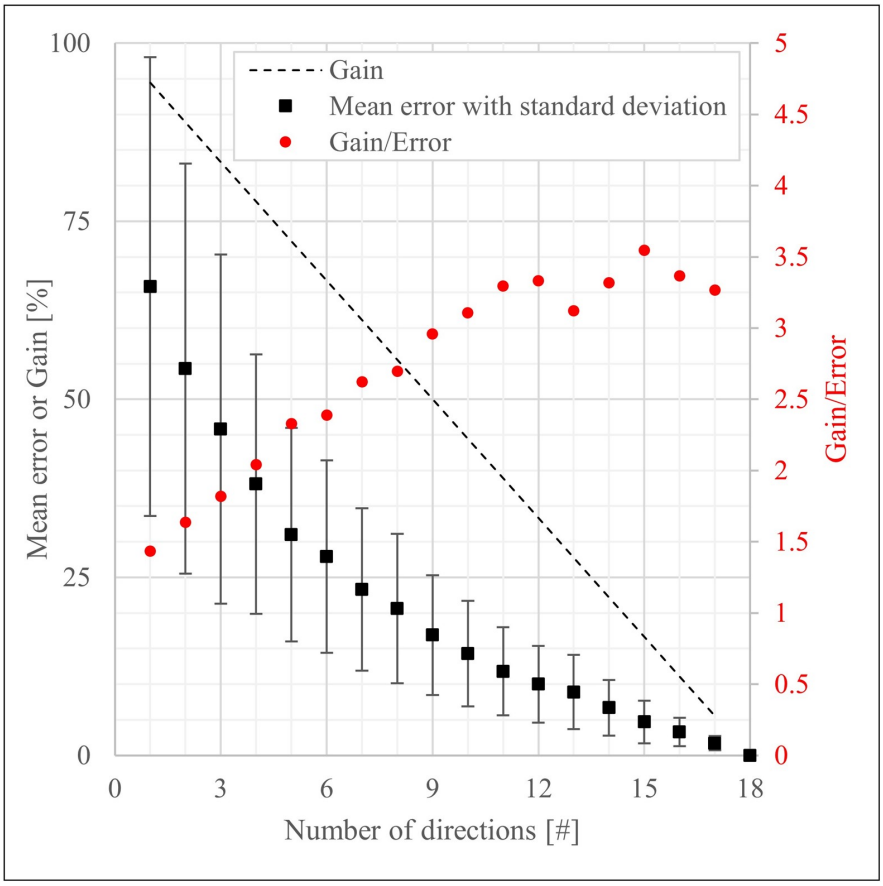


Figure 4.9: Global results for annual concentration calculation using the predominant wind directions with the mean errors, the calculation cost gain and the ratio between gain and error.

As previously, the influence of the building layout and the wind rose was assessed and the results are presented in Figure 4.10. According to Figure 4.10(A), the wind rose have an impact on the mean errors obtained leading to an overall maximal variation of 1.9. If we consider the overall patterns of the wind roses (see 4.2), there is no specific trends between the wind rose patterns and the errors using this approach. As an example, the wind rose of Paris and Nîmes are strongly different (the first one being homogeneous and the second one having a preferential direction) but neither of them gives systematically less error than the second one. According to Figure 4.10 (B), it is the same observation when comparing the results as a function of the building layout. In this case, the maximal variation is higher with an overall value of 4.2 which indicates that the error is more sensitive to the building layout than to the wind rose.

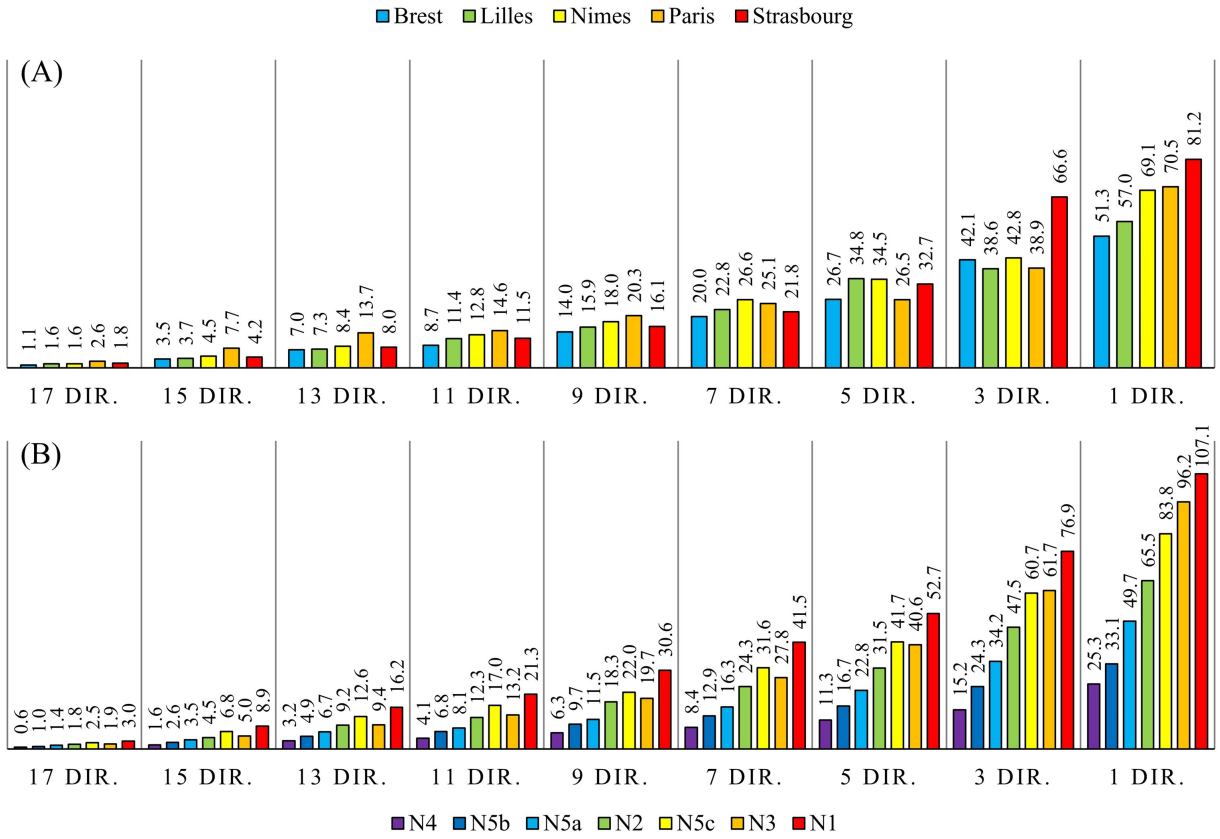


Figure 4.10: Mean error over the mean annual concentration compared to the reference using the predominant wind directions as a function of (A) the wind rose location and (B) the building layout (DIR: Directions).

4.3.3 Comparison between both methodologies

The first methodology, which uses a regular step, showed trends in the errors depending on the wind rose pattern: the more the wind rose is homogeneous and the less the error is high. Such a trend was not observed with the second methodology which uses the predominant wind directions. These two methodologies were compared as a function of the wind rose in order to find out which of the two is the best overall and for specific wind rose patterns. The results are given in Figure 4.11. No comparison was performed according to the building layout since specific trends were not observed for this parameter.

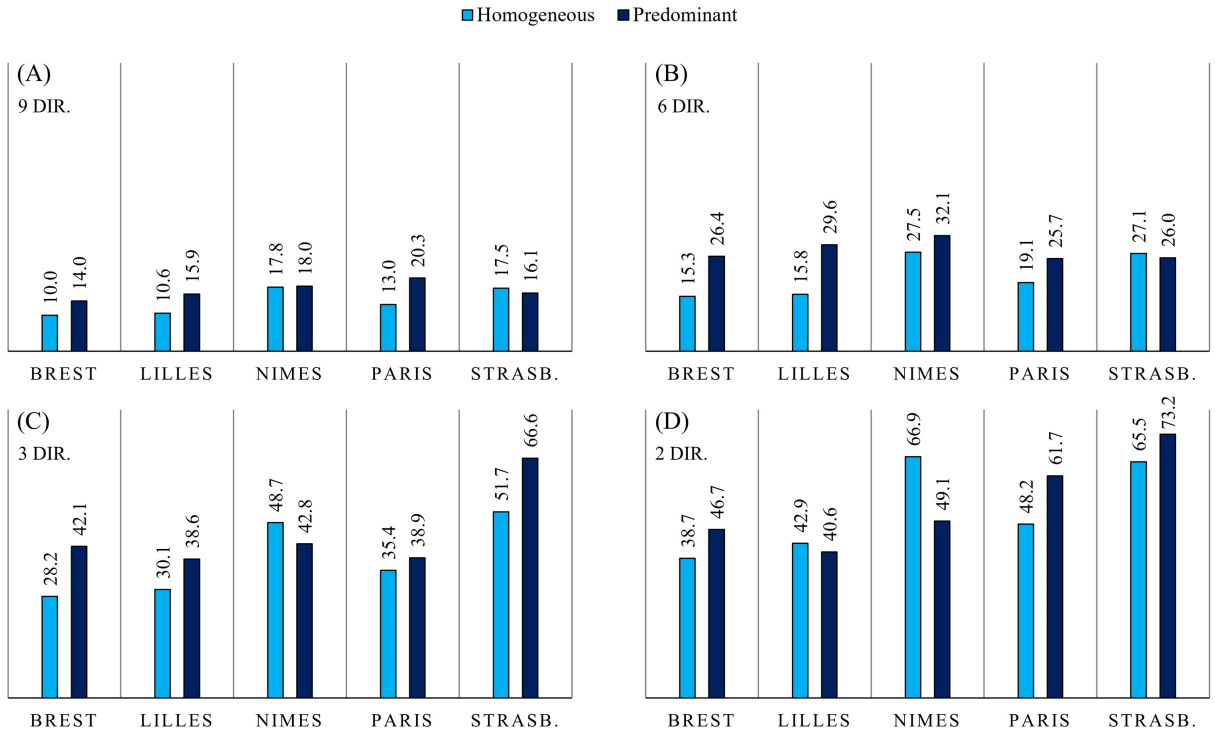


Figure 4.11: Comparison of the mean error over the mean annual concentration compared to the reference as a function of the wind rose for 9, 6, 3 and 2 wind directions using the first (homogeneous) and the second (predominant) approach.

According to Figure 4.11 it can be seen that depending on the wind rose and the number of directions considered, one approach can perform better compared to the second and vice versa. As an example, the first approach considering regular steps gives less error with the wind rose of Brest and 9 wind directions (10.0%) compared to the second approach (14.0%). Inversely, the first approach gives higher error with the wind rose of Nîmes and 3 wind directions (48.7%) compared to the second one (42.8%). The results are, however, better for three quarter of cases using the first methodology, which can be seen in Figure 4.10 (A) and (B). Additionally, when taking all cases into account a mean relative difference of 18% on the error is obtained in favor of this methodology. Nevertheless, when considering more than a half of the wind directions available in the wind rose, the first approach was not evaluated, only the second one.

4.3.4 Estimation of the error with respect to the complete wind rose

As a last point of analysis, a study has been performed to assess the possibility of estimating the error induced by considering a partial wind rose with the first approach (regular steps)

compared to a simulation of the full wind rose with 18 wind directions. Indeed, it has been shown previously that using this methodology, an overall error of 13.8% is obtained when considering 9 wind directions instead of 18. However, this error ranges from 10.0% to 17.5% as a function of the wind rose and from 5.6% and 21.3% as a function of the building layout considered in this work. Thus, even if an overall prior estimation of the error before doing the simulations is available, the specific case result can still be far from the expected ones given the large possible ranges of error. A way to assess the error more accurately once the simulations are done is therefore necessary.

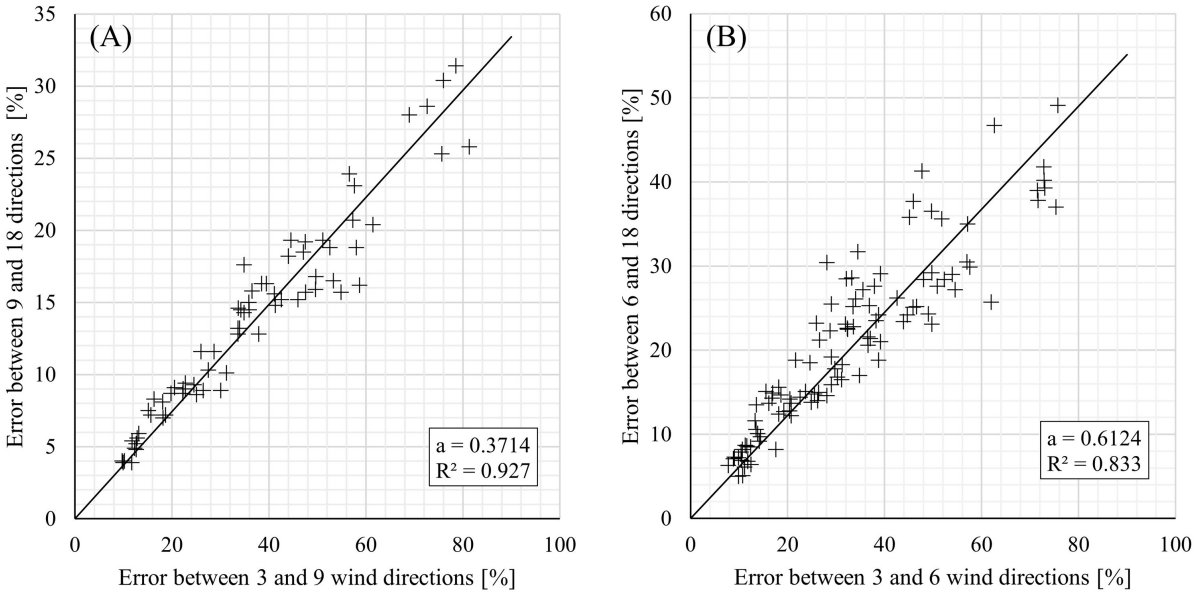


Figure 4.12: Estimation of the error using the first approach with regular steps when considering (A) 9 and (B) 6 wind directions.

In order to have a better evaluation on the error when using the approach with regular steps, the error between the results obtained with 9 and 18 wind direction has been plotted as a function of the error between the results obtained with 3 and 9 wind directions. The scatterplot is given in Figure 4.12 (A). Each point of this scatter plot corresponds to a given couple of wind rose and building layout as well as a given starting point. According to this figure, it can be seen that the scatter plot seems to be linearly correlated using a linear function with a slope of 0.3714, leading to a coefficient of determination R^2 of 0.927. The corresponding equation is given in 4.5.

$$E_{9/18} = E_{3/9} \times 0.3714 \tag{4.5}$$

Thus, using Figure 4.12(A) and results from 3 wind directions evenly spaced in a total of 9 wind directions simulated with a regular step, it is possible to assess the error compared

to considering a whole wind rose. The same work has been carried out considering 6 wind directions and the results are given in Figure 4.12 (B).

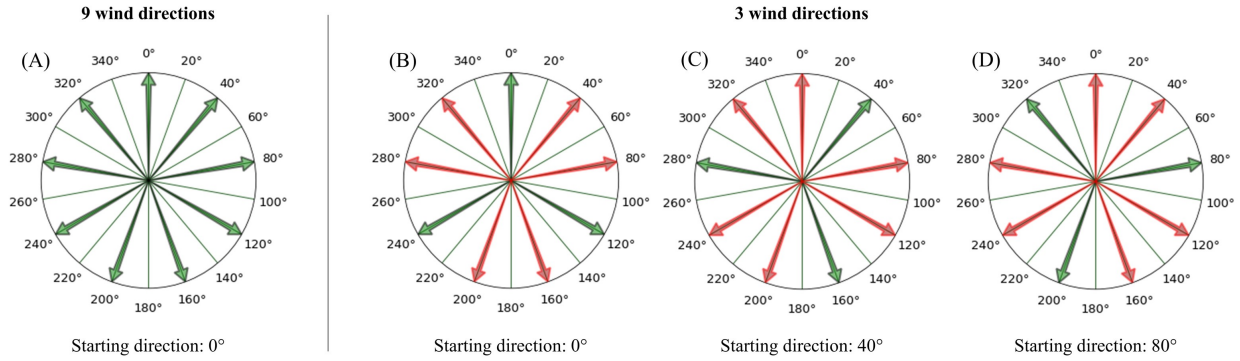


Figure 4.13: Illustration of the example to calculate the error made compared with considering the whole wind rose (green arrow: wind direction modelled / red arrow: wind direction not considered for annual concentration calculation).

As a practical example, if mean annual concentrations are calculated considering 9 wind directions starting at 0° as shown in Figure 4.13 (A), mean annual concentrations considering 3 wind directions already simulated can also be calculated with three distinct starting directions: 0°, 40° and 80° (Figure 4.13 (B), (C) and (D) respectively). If these last three annual concentrations give on average 25% of difference with the one calculated with 9 wind directions, then, according to Figure 4.12 (A), an error of around 25% $\times$ 0.3714 = 11% is made using 9 wind directions evenly spaced instead of considering the whole wind rose. The same methodology can be used when using only 6 wind directions but using Figure 4.12 (B) instead of (A). The methodology to determine the error is presented as a step-by-step flowchart on Figure 4.14.

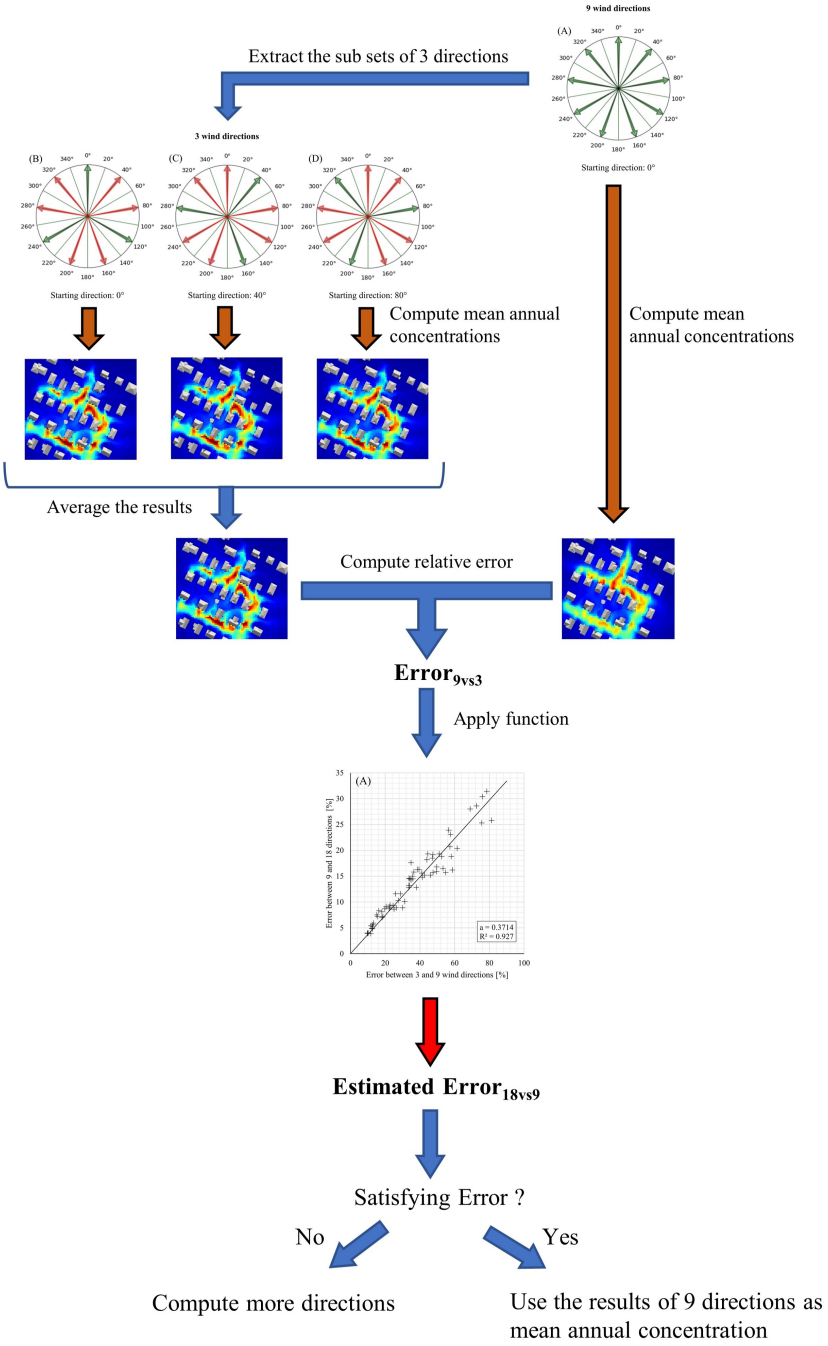


Figure 4.14: Flowchart of the methodology to determine the error for the method 1.

4.4 Discussion

This study provides information on how to decrease the number of wind directions needed for mean annual concentrations calculation based on CFD results in order to decrease the calculation costs without leading to high errors. It enables environmental engineers and scientists to assess the air pollution of a region more quickly and cost-effectively, while managing the resulting error and ensuring that it is within an acceptable range. Two approaches were considered with (1) ignoring some wind directions uniformly spaced and (2) considering the predominant wind directions. Additional work can be done to extend the use of these approaches and the major issues are discussed hereafter.

Several configurations of building layout were considered in this study which mainly included urban and peri-urban neighborhoods, with buildings overall ranging from 10 m to 25 m high with punctual structures of 2–3 m high, and both homogeneous and heterogeneous layouts. It has been shown that no specific trends are observed between the error and the building layout considered and that an overall maximal variation of 3.9 for the first approach (respectively 4.2 for the second one) is obtained. However, more densely built-up neighborhoods with higher buildings such as in city centers were not considered in the scope of this work. Since such urban configurations may lead to different results, further work can be done in this direction to extend the applicability of the methodologies studied in this paper.

Some wind roses were also considered in this study covering the four cardinal points of France (and Paris) and having different patterns: homogeneous in wind direction and velocity, homogeneous in wind direction with mostly intermediate velocities, heterogeneous with a preferential axis of wind direction and heterogeneous with a preferential direction of wind. It has been shown that homogeneous wind roses in wind direction and velocity led to the minimal errors while heterogeneous wind roses with preferential wind axis and direction led to the maximal errors when using the first approach (an overall maximal variation of 1.7 was obtained with the first approach and 1.9 with the second one). The wind roses used in this work were only located in France, nonetheless. Additional work can therefore be performed to extend the applicability of these approaches using wind roses from different countries, under different climates or with extreme wind roses highly homogeneous or heterogeneous or with mostly high and low wind velocities.

Additionally, according to the high variation in the errors as a function of the building layout, it is not possible to be sure in advance of the error made using one of the two approaches presented. Only an overall information is available prior to the choice of number of directions but given the spread of values it can be too vague. However, a methodology to assess the error afterward for the first approach has been presented, allowing the operator

to estimate the error made compared with considering the whole wind rose finely. Based on the result, the operator might choose to keep the results as they are or simulate the missing wind directions if the error is not acceptable.

The reader must nonetheless be aware that these results were achieved under the following hypothesis.

1. For the CFD simulation: RANS model, RNG k-ε turbulence model, surface emissions, passive scalar, neutral atmospheric conditions, insensitive meshing, distances between boundaries respecting COST Action 732 guidelines.

2. For the mean annual concentration: the calculation was done following the statistical approach provided in (120) and using annually averaged daily traffic emissions.

To use the raw results of our study for real-life settings, one should be aware that this work was done under this set of hypotheses, and that depending on how much the reader deviates from it (using LES models, chemical reaction, etc.) he should be careful with taking the result as they are.

The methodology developed here consists of reducing the number of directions to improve computation time while controlling subsequent error when calculating mean annual concentration. However, it can be applied to other set of hypotheses given some examples to produce adapted equations. The flowchart remains the same.

Finally, the whole work has been conducted considering wind roses with 20° steps in wind directions, thus 18 wind directions. Different discretisations can also be found such as 22.5°, 30° or 40° corresponding respectively to 15, 12 or 9 wind directions. The interest of this work was also to give an idea of the mistake that can be made by using weakly discretised wind roses. As an example, in the case of a 40° discretised wind rose, an overall error of 13.8% is thus made compared with a more discretised wind rose of 20° (18 wind directions).

4.5 Conclusion

The objectives of this study were to find out the possibilities to limit the number of wind directions needed to be simulated in order to calculate mean annual concentration based on CFD results at a lower calculation cost. Two approaches were studied and compared throughout this paper and the main conclusions are as follows:

(a) Ignoring some wind directions evenly spaced (first approach) can highly decrease the calculation costs without leading to high errors: when simulating one wind direction out of two, an overall error of 13.8% can be expected for a calculation gain of 50%.

(b) The error made when ignoring some wind directions evenly spaced is depending on both wind rose and building layout. No specific trend can be identified as a function of the

building layout. As a function of the wind rose, the trend is that the error is smaller when the wind rose is homogeneous than where there is a preferential wind axis or direction.

(c) Considering the predominant wind directions (second approach) can also decrease the calculation costs: when simulating the first twelve wind directions, an overall error of 10% can be expected for a calculation gain of 35%.

(d) The error made when considering the predominant wind directions is depending on both wind rose and building layout but no specific trend can be identified neither as a function of the building layout nor the wind rose.

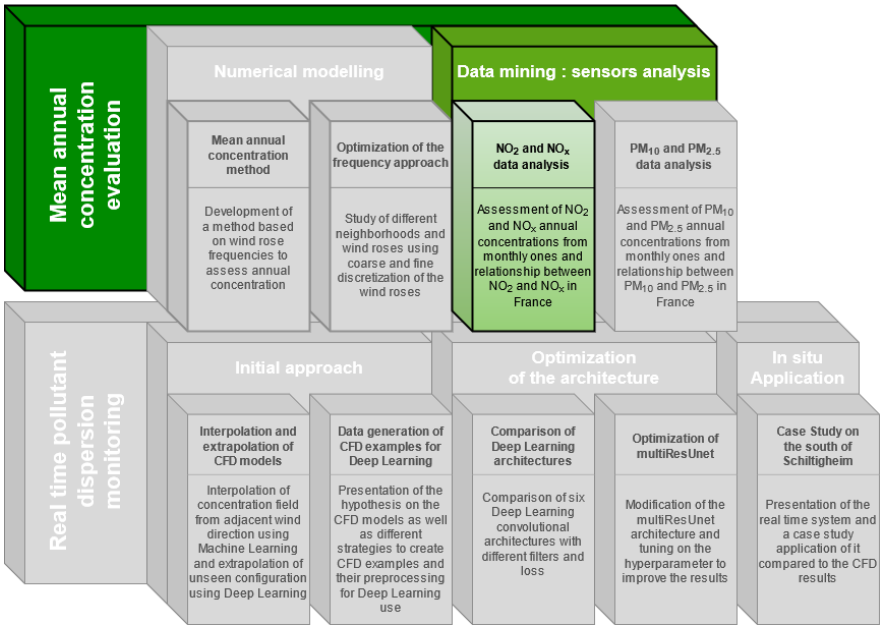
(e) The first approach considering uniformly spaced wind directions is generally better than the second, leading to lower errors for the same number of wind directions considered. The first approach should therefore be preferred, it has not been studied when more than half of the wind directions are considered with the first approach, but it could be used as well.

(f) A way to evaluate the error made considering 6 or 9 wind directions evenly spaced with respect to the full the wind rose is provided for the first approach which can be used to have a better idea of the error made or to check if additional wind directions are necessary.

Eventually, the results of this study will allow environmental engineers and scientists to assess annual outdoor air quality optimally with a wider use of numerical methods to compare with regulatory values provided by the WHO, the EU or any other organization. Indeed, the cost of modeling is the main obstacle and, with the presented methods, can be significantly reduced while managing the resulting error and ensuring that it is within an acceptable range. Further work could be done to evaluate the error and optimal strategies with others numerical models such as plume models or LES CFD models.

Chapter 5

Assessment of mean annual NO_2 and NO_x concentration based on a partial dataset



This chapter has been published in the journal *Atmospheric Environment* under the title "Assessment of mean annual NO_2 concentration based on a partial dataset " (60).
The two previous Chapters 3 and 4 deals with assessing the annual concentration using numerical modelling. The method relies on computing several wind directions and wind speed and use their frequency to ponderate their results. The discretization of the wind roses is then a balance between computation time and numerical accuracy. A flowchart is also presented to be able to assess the error that is made by doing less than 18 wind directions.

Nevertheless, numerical modelling is a powerful tool but it is often needed to be coupled with sensors to have reliable results. Indeed, sensors can have two main contributions. On one hand, they can help to determine the background pollution that is not computed with microscale models. On the other hand, they can be used to ensure that the model is valid. Yet, there is a major limit when using sensors for annual concentration, a sensor needs to measure over the whole year at the same location and measures only one pollutant. Hence, two questions arise: is it possible to determine the annual concentration over a shorter period of time and is it possible to determine other close pollutant species from the data of another pollutant?

5.1 Introduction

While many measures are implemented to improve air quality, atmospheric pollution still exceeds the thresholds of health standards. Next to particulate matter or ozone, nitrogen dioxide (NO2) has been selected as an air pollutant with the highest priority whose monitoring must be routinely carried out (163). Nitrogen oxides are known to be a source of respiratory symptoms and diseases (61), and they are also harmful to the environment as they play the role of precursor in nitric acid production, leading to acid rains (79). These air pollutants are mainly due to anthropogenic sources. Indeed (151) showed that in several cities in Europe, NO_x is mainly emitted by transport and industrial sources, with varying contributions depending on the city. For example, in dense urban areas such as Paris, 56% of NO_x comes from traffic-related emissions and 18% from the tertiary and residential sectors (6).

Nitrogen dioxide (NO_2) is, with nitric oxide (NO), one of the two components forming nitrogen oxides. In the European Union (EU) and more generally around the world, NO_2 is the most measured component. Indeed, NO_2 can have significant harmful effects on health, inducing numerous diseases like bronchitis, pneumonia, etc. (111), but it can also increase the risks of viral and bacterial infections (28).

To obtain standard values for the purposes of comparison, the European Union (EU) and the World Health Organization (WHO) have issued critical values that should not be exceeded to protect the public from the health effect of gaseous NO_2. For this purpose, two standard values have been enforced : a hourly mean of 200 µg/m^3 and an annual mean of 40 µg/m^3 not to exceed given by both the WHO (166) and the EU (37). Studies have shown that the annual standard is generally more stringent than the hourly one (26; 53). However, year-round measurements are needed to gather concentrations values that can be compared directly to this standard. This requirement is not a constraint when monitoring stations are located permanently in one area. Nonetheless, it becomes constraining when the objective is

to evaluate urban planning projects over a limited period: the heterogeneity of urban areas requires controls related to the standard at several key locations where no permanent stations have been installed and where only temporary measurements are economically viable. Moreover, these temporary measurements may only provide information on NO_x concentrations but no direct information on NO_2. Thus, one question arises in such situation: how can annual mean NO_2 concentrations be determined using only a short measurement period of NO_2 or NO_x concentrations?

The Leighton relationship provides information on the ratio between NO and NO_2 concentrations as a function of O_3, a chemical constant rate and a photolysis rate considering the photochemical steady state (76). Unfortunately, it was demonstrated that using this method with more than 10 ppb of O_3 leads to an increasing error by not taking into account VOC chemistry (137). Different methods were proposed to evaluate the photolysis rate (167), but computing an annual representative photolysis rate can still lead to a wrong evaluation of the seasonal dependencies between NO_x and NO_2. Numerical computation based on complex chemical mechanisms involving more than 300 reactions with more than 100 species gives more accurate evaluations of NO_2 ((22; 65)). Nevertheless, when NO_2 concentration measures are missing there is little chance that this information is known on other species such as VOCs. However, such information is needed in the numerical computations.

Furthermore, seasonal variability of NO_2 and NO_x concentrations differs considerably between summer and winter because NO_2 concentrations depend on photolysis conditions, and NO_x molecules play a role in several chemical mechanisms in the troposphere, involving ozone (O_3) and volatile organic compounds (VOC) (143). Robert-Semple et al. showed that there is a relative standard deviation of more than 50% when calculating the mean annual concentrations of both NO_2 and NO_x (129). Moreover, Kendrick et al. showed that there is a seasonal variability in NO_2 concentration even with constant hourly seasonal traffic (63). Thus, these results show that a few months of NO_2 monitoring are generally not representative of a mean annual concentration despite existing only slight seasonal variations of the main source, namely traffic-related emissions.

The aim of this study is first to evaluate whether one-parameter methods without any explicit chemical mechanism found in the literature are sufficiently accurate to determine NO_2 concentrations based on monitored NO_x data in France. The second aim is to present a method capable of providing the mean annual NO_2 concentration from one-month period of monitoring.

In this article, the different areas of study as well as the measurement method and the approach to turn NO_x into NO_2 used are presented in section 2. Then, the results of the study on the NO_x-based NO_2 concentration calculation in France, and the method presented for the mean annual NO_2 concentration calculation based on monthly measurement periods,

are presented in section 3.

5.2 Material and methods

5.2.1 Study location

This work uses NO_2 and NO_x concentrations monitored in a large number of regions in France, including from North to South: Hauts-de-France, Grand-Est (Strasbourg region), Ile-de-France (Paris region), Pays de la Loire, Auvergne-Rhône-Alpes and Provence-Alpes-Côte d'Azur. These areas were chosen for the availability of data and to better cover the minimum and maximum latitudes and longitudes of France. The location of these regions is presented in 5.1.

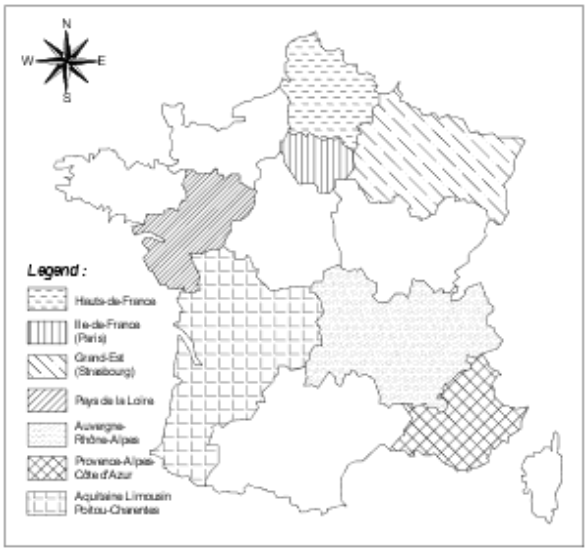


Figure 5.1: Location of the different study areas used.

5.2.2 Data availability

The data used in this work were obtained via the open access database provided by the different air quality monitoring authorities known as AASQA, the French acronym for "Approved Air Quality Monitoring Associations". In particular, the data were provided by the organisations Atmo Haut-de-France (Haut-de-France), Atmo Grand-Est (Strasbourg region), AIRPARIF (Paris region), Air Pays de la Loire (Pays de la Loire), Atmo Auvergne-Rhône-Alpes (Auvergne-Rhône-Alpes), Atmo PACA (Provence-Alpes-Côte d'Azur) and Atmo Nouvelle-Aquitaine (Aquitaine Limousin Poitou-Charentes). The data are mainly mean annual NO_2 and NO_x concentrations over a five-year period from 2013 to 2017, but other data such as

Region	**Data availability**	NO_x			NO_2			**Stations**
		A	M	H	A	M	H	
Ile-de-France (Paris)	2013 - 2017			X			X	≈ 40
Grand-Est (Strasbourg)	2018			X			X	≈ 50
Hauts-de-France	2013 - 2017	X			X			≈ 15
Pays de la Loire	2013 - 2017	X			X			≈ 50
Auvergne Rhône-Alpes	2013 - 2017	X			X	X		≈ 60
Provence-Alpes-Côte d'Azur	2013 - 2017	X			X	X		≈ 25
Aquitaine Limousin Poitou-Charentes	2013 - 2017	X			X	X		≈ 30

Table 5.1: Summary of the available data

hourly measured concentrations for the Strasbourg region in 2018 were also obtained. Additional contacts were also made with AIRPARIF to obtain more specific data for the Paris Region like hourly measured concentrations from 2013 to 2017 with their corresponding uncertainties. A summary of the available data, corresponding to about 270 different sensors, is presented in Table 5.1.

5.2.3 Data range

The annual and monthly concentrations range from 10 to 340 µg/m^3 for NO_x and from 5 to 95 µg/m^3 for NO_2, considering the complete dataset (all years, types and locations of stations included). According to these wide ranges, different types of stations were considered in this work including rural, suburban, urban and traffic stations. The dataset for the Paris region comprises 2% rural, 13% suburban, 54% urban and 31% traffic stations. The type of station was not always directly provided in the global France dataset. Thus, the percentage of each type of station was estimated based on the range of concentrations for each type of station in Paris. The corresponding results were 29%, 22%, 31% and 18% for rural, suburban, urban and traffic stations, respectively.

5.2.4 Monitoring method

The EU imposes a maximal uncertainty of 15% on AASQA for individual measurements averaged over the period considered regarding the limit values monitored by sensors. Thus, to satisfy the requirements, all AASQA use the same monitoring method in accordance with this constraint.

The reference method used for the measurement of nitrogen dioxide and oxides of nitrogen is known as chemiluminescence. Two chemiluminescence methods exist: on the one hand, chemiluminescence based on luminol reaction, and, on the other hand, chemiluminescence based on NO/O_3 reaction. The second method is the one used in France. In particular AIRPARIF uses the AC32M EN model from ENVE and the 42i model from THERMO SCIENTIFIC.

The principle of the method was well-described by (100) and is based on the reaction 5.1 between NO and O_3. This reaction produces an excited nitrogen dioxide (NO_2^*) that emits infrared radiations when returning to a stable state. The luminous radiation emitted and then measured is directly proportional to the NO concentration.

$$NO+O_3 \rightarrow NO_2^* + O_2 \tag{5.1}$$

$$NO_2^* \rightarrow NO_2 + h\nu \tag{5.2}$$

To obtain information on the NO_x concentration, it is first necessary to convert all the NO_2 into NO before the measurement. After that, the resulting NO corresponding to the initial NO and the NO derived from NO_2 are measured and the NO_x concentration is obtained. Combining both the measured NO and NO_x concentrations provides the NO_2 concentration. Thus, the uncertainties on NO_2 measurement are higher than those on NO or NO_x because the results are obtained from both NO and NO_x measurements.

Based on the work of Navas et al., this kind of technique has very low detection limits, making it a good tool for evaluating the concentration of nitrogen compounds for atmospheric purposes (100). According to a personal communication with AIRPARIF, the maximal uncertainty on the mean annual NO_2 concentration from 2015 to 2017 was lower than 10% with a mean uncertainty of 6%.

5.2.5 Empirical methods to convert concentration from NO_x to NO_2

Several one-parametric empirical methods can be found in the literature to give an estimation of NO_2 concentration based on NO_x concentration. Three methods were compared with the

entire France dataset:

1. Derwent and Middleton function, a polynomial-logarithmic function linking hourly averaged NO_x and NO_2 concentrations for NO_x concentrations in the range of 9.0 to 1145.1 ppb (34).

2. Romberg et al. function, a rational function linking annual averaged NO_x and NO_2 (130).

3. Bächlin et al., another rational function linking annual averaged NO_x and NO_x (25).

According to the above authors, the corresponding equations are 5.3, 5.4 and 5.5 respectively, with the hourly averaged NO_x and NO_2 noted $[\mathrm{NOx}]_h$ and $[\mathrm{NO2}]_h$ and annual averaged NO_x and NO_2 for the two other functions noted $[\mathrm{NOx}]_a$ and $[\mathrm{NO2}]_a$. All concentrations presented below are in µg/m^3 and A=$\log_{10}([NO_x]_h/1.91)$.

$$[NO_2]_h = \left(2.166 - \frac{[NO_x]_h}{1.91}\left(1.236 - 3.348A + 1.933A^2 - 0.326A^3\right)\right) \times 1.91 \tag{5.3}$$

$$[NO_2] = \frac{103.[NO_x]_a}{[NO_x]_a + 130} + 0.005 \times [NO_x]_a \tag{5.4}$$

$$[NO_2] = \frac{29.[NO_x]_a}{[NO_x]_a + 35} + 0.217 \times [NO_x]_a \tag{5.5}$$

For the purpose of this work, mean annual concentrations were used instead of hourly averaged concentrations for the Derwent and Middleton function.

5.3 Results

5.3.1 Evaluation of annual NO_2 concentration based on NO_x data

Best fitting function in France Figure 5.2 shows the evolution of mean annual NO_2 concentration as a function of the mean annual NO_x concentration considering the total dataset (measurements from 2013 to 2017 for the six regions considered and all types of station included). The three empirical methods cited previously are also plotted.

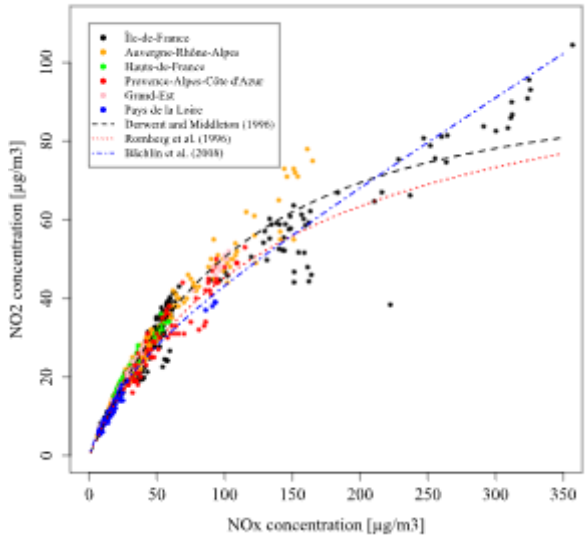


Figure 5.2: Evolution of NO_2 concentration as a function of NO_x concentration and comparison with empirical functions

To obtain a better comparison between the three functions, predicted NO_2 concentrations calculated with measured NO_x concentrations were plotted against measured NO_2 concentrations. The corresponding results are presented in 5.3. with the first bisector corresponding to ideal results. As shown in 5.3., the function from Bächlin et al. is the most appropriate for high NO_2, thus high NO_x concentrations. However, based on 5.3. (A) and 5.3. (B) the results for lower NO_2 concentrations (less than 50 µg/m^3) are better when using the function proposed by (34), and (130). Considering the difference between the predicted and measured concentrations, the function of Derwent and Middleton is the most appropriate with a deviation of less than 8%, whereas that of (130) leads to a deviation of 9.5%. Moreover, in this work, the function of (130) tends to slightly underpredict NO_2 concentrations. When choosing between two functions giving about the same deviation, the precautionary approach is to choose the function that overestimates NO_2 rather than the one which underestimates it. Hence, in France, Derwent and Middleton's function has been chosen and is advised by the authors to assess the NO_2 concentrations based on NO_x data. This is especially the case for the monitoring both in urban and rural sites. It should also be noted that these comparisons included several years of measurements and locations (various latitudes and longitudes), thus in principle giving independence to these parameters. However, for high NO_2 concentrations (higher than 70 µg/m^3) the method fits less and less well.

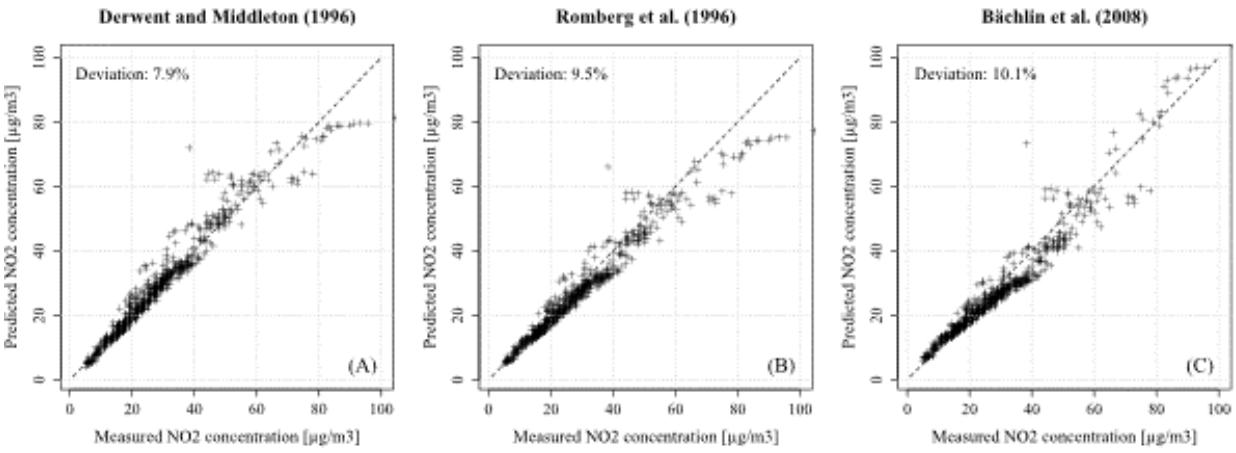


Figure 5.3: Comparison between predicted and measured NO_2 concentrations for (A) the Derwent and Middleton function, (B) the Romberg et al. function, and (C) the Bächlin et al. functions

Application to Paris region The information obtained in the Paris region was more detailed and included uncertainties as well as the type of station. 5.4 presents the mean annual NO_2 concentration for the Paris region dataset as a function of NO_x concentration with a distinction between the different types of station. Derwent and Middleton's function is also plotted.

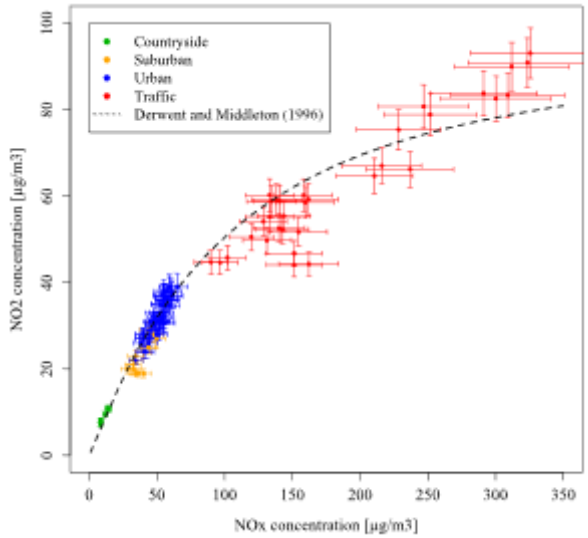


Figure 5.4: Evolution of NO_2 concentration as a function of NO_x concentration for the Paris region dataset and comparison with Derwent and Middleton's function.

These results show that in accordance with previous observations, the best range of

application for Derwent and Middleton's function is for NO_x concentrations lower than 80 µg/m^3. As can be seen in 5.4. this limit corresponds to the difference between urban and traffic stations for Parisian region. Thus, Derwent and Middleton's method applies best for rural, suburban and urban stations whereas the results are less accurate for traffic. Indeed, there are 92% of the data that are within the uncertainties range both in the countryside and in urban areas, while for traffic data it falls to 71%. The mean error on predicted NO_2 concentrations is 9% with a 95^{th} percentile of 27%.

5.3.2 Seasonal variability of NO_2 concentration

The seasonal variability of NO_2 was studied using the Paris region dataset. Hourly NO_2 concentrations were averaged for each station and each year of data, giving five mean concentrations per station and per year (one annual concentration and four seasonal concentrations). 5.5. (A) shows the differences between seasonal mean NO_x concentrations for each couple of year and station. 5.5. (B) shows the evolution of seasonal NO_2 concentrations as a function of the annual NO_2 concentration for the same year of measurement.

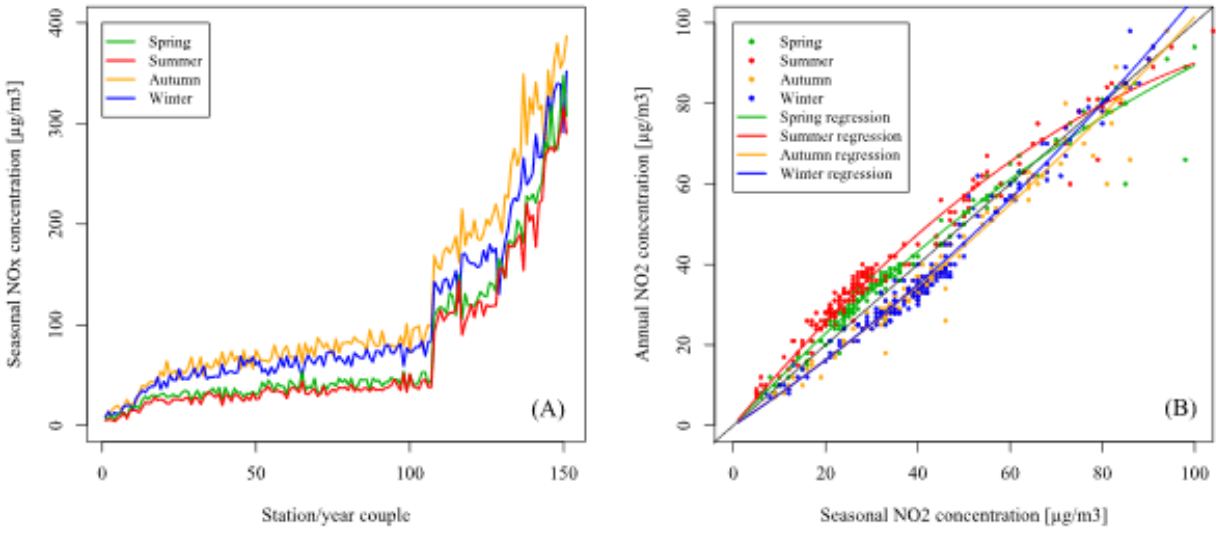


Figure 5.5: Comparison between seasonal NO_x concentrations for a given station and year of measurement in the Paris region (A) and the evolution of the annual NO_2 concentration as a function of seasonal NO_2 concentrations (B).

According to 5.5. (A), NO_x concentrations are strongly dependent on the season. Indeed, although summer and spring NO_x concentrations are similar, the concentrations are higher in winter and autumn by up to a factor of 2. These differences can be explained by several disparities between these seasons: lower boundary layer height, lower temperatures and new sources of emission due to residential heating, increased emissions by cold-started vehicles,

etc.

Since the results show that NO_x concentrations are higher in winter and autumn, for a given NO_x concentration the seasonal NO_2 concentrations should also be higher in autumn and winter than in summer and spring. However, the results for the Paris region show a different trend. The result in 5.5. (B) indicates a change of behavior when the annual NO_2 concentration increases, with the summer and spring NO_2 concentrations becoming higher than in autumn and winter. These results can be associated with those of other authors. Indeed, (63) showed that NO_2 concentrations are higher in winter and autumn than in spring and summer, with a mean annual NO_2 concentration lower than 80 µg/m^3 and for three different types of station (63). On the contrary, Mavroidis and Ilia showed that for a traffic station (i.e. giving high NO_2 concentrations), NO_2 concentrations are generally higher during the summer and spring months than in autumn and winter, with in their case a mean annual NO_2 concentration higher than 80 µg/m^3 (90). Thus, the evolution of seasonal NO_2 concentrations as a function of annual NO_2 concentration is not well represented by a linear method unable to catch these varying trends and is much better fitted by a quadratic one. With this interpolation, the spring and summer results are described by a concave quadratic function whereas the autumn and winter ones are described by a convex quadratic function. In this case, these concavities and convexities result in a NO_2 concentration of about 80 µg/m^3, where the seasonal NO_2 concentrations are equal to the annual NO_2 concentration. This concentration of 80 µg/m^3 corresponds to the value for which, in the case of a measurement station giving an annual average NO_2 concentration lower than this value, the concentrations for winter and autumn are higher than the spring and summer concentrations. Therefore, to obtain maximized measurements in order to assess an upper limit on annual NO_2 concentration over a short period of time, the measurements should be carried out in winter, in case where an annual concentration of less than 80 µg/m3 is expected, otherwise measurements should be carried out in summer.

These observations are consistent with those of other research papers, despite being counter intuitive on the first point of view. Indeed, a previous observation was that NO_x concentrations are higher during autumn and winter, in theory giving higher NO_2 concentrations. Moreover, in summer and spring, the zenithal angles are generally lower, leading to increased photochemistry with higher photolysis, including NO_2 photolysis, and the production of radicals. As shown in 5.6. (A), O_3 concentrations are globally much lower in autumn than in winter, and in winter than in spring and summer. These concentrations are about the same between spring and summer. 5.6. (B) gives supplementary information on how much ozone is available to react with NO_2, by giving the evolution of the ratio of the seasonal O_3 concentration over the seasonal NO_2 concentration as a function of the seasonal NO_2 concentration.

The first observation is that more O_3 molecules are available in spring and summer than in winter and autumn for any NO_2 concentration. This statement is always true even when the seasonal NO_2 concentration increases, leading to a systemic reduction of available O_3. For example, for a seasonal NO_2 concentration of 15 µg/m3, the ratio of seasonal O_3 concentration over seasonal NO_2 concentration is around 3 for autumn, 4 for winter and almost 5 for spring and summer. Increasing the seasonal NO_2 concentration to 30 µg/m^3 gives ratios of 1 and 1.5 for autumn and winter respectively and almost 2 for both spring and summer. The explanation of why the seasonal NO_2 concentration is higher in spring and summer than in winter and autumn for high NO_2 concentrations can be obtained from these two observations. For low NO_2 concentrations, O_3 is readily available and the reaction is not limited by the O_3 concentration but by several other factors that lead to the commonly accepted result: NO_2 concentrations are higher in winter and autumn than in spring and summer. However, when the NO_2 concentration increases, O_3 becomes less and less available until reaching a state in which it becomes the limiting reagent of the production reaction of NO_2 from NO_x. This state is reached earlier in winter and autumn than in spring and summer, leading to a higher NO_2 concentration in summer and spring than in autumn and winter.

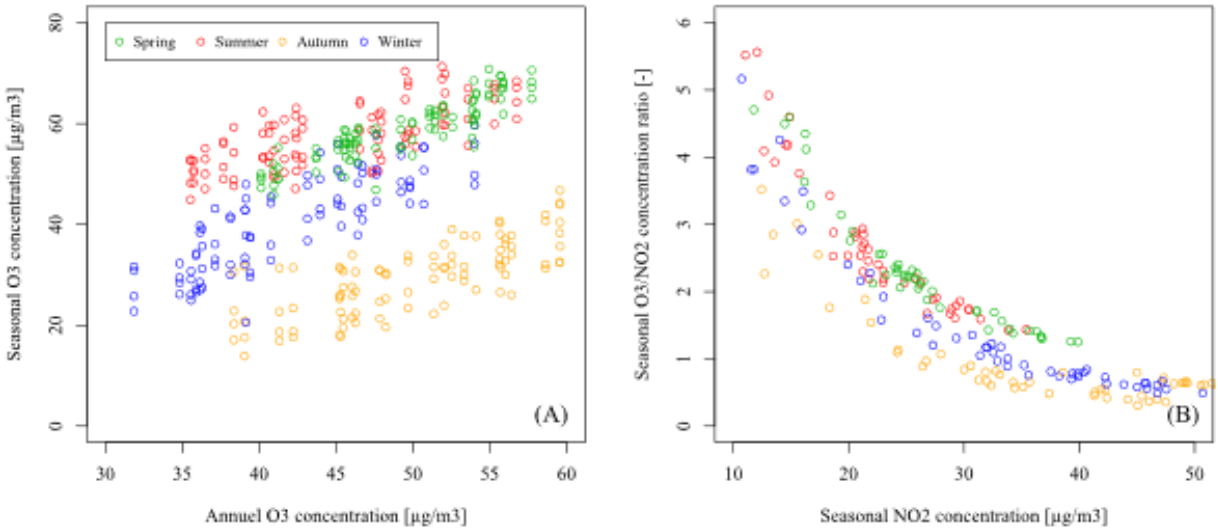


Figure 5.6: Evolution of the seasonal O_3 concentration as a function of the annual O_3 concentration (A) in the Paris region and the evolution of the ratio between seasonal O_3 and NO_2 concentrations as a function seasonal NO_2 concentrations (B).

5.3.3 Assessment of annual NO_2 concentration

Assessment of annual NO_2 concentration from monthly NO_2 concentrations As mentioned above with regards to seasonal variability, seasonal concentrations cannot be used

directly as an annual concentration. However, they seem to fit a trend and it may be possible to assess the annual mean concentration from a short period of measurement.

The NO_2 concentrations over the Paris region were first averaged for each month and then compared with annual NO_2 concentrations. The results, presented with black circles in 5.7, show that, like seasonal NO_2 concentrations, monthly averaged NO_2 concentrations as a function of annual NO_2 concentrations seem to be better fitted by a quadric function than by a linear function. These fittings are also presented with black lines in 5.7. as well as the polynomial interpolation coefficients, and the mean error between measured data and interpolation, also in black. The polynomial equation corresponds to 5.6 with $[NO_2]_a$ and $[NO_2]_m$ being the annual mean NO_2 concentration and the monthly averaged NO_2 concentration respectively in µg/m^3, and a and b the different polynomial coefficients for each month.

$$[NO_2]_a = a.[NO_2]_m{}^2 + b.[NO_2]_m \tag{5.6}$$

The polynomial methods obtained have different concavities and convexities, consistent with those obtained for seasonal variability. The maximum convexity is obtained around December and January, corresponding to the transition from autumn to winter. The maximum concavity is obtained around June and July, corresponding to the transition from spring to summer. Lastly, minimal concavity and convexity is obtained around March and September, corresponding to the transition from winter to spring and from summer to autumn, respectively. For these months, monthly averaged NO_2 concentrations are almost equal to annual NO_2 concentrations. According to these polynomial methods, the maximal mean error is around 15% and corresponds to December, and the minimal mean error is around 7% and corresponds to March. The mean error averaged over all months is below 10%.

These polynomial methods can be used to assess the annual NO_2 concentration based on only one month of measurements. However, the problem is that measurements from the first day to the last day of a month are required. If one month of data is acquired that overlaps two distinct months, say from 15^{th} January to 15^{th} February, the interpolation is no longer appropriate. An additional study was carried out to change from discrete to continuous interpolation. To achieve this, the resulting polynomial coefficients a and b were plotted as a function of the month with 1 corresponding to January and 12 to December. 5.8. shows the corresponding results.

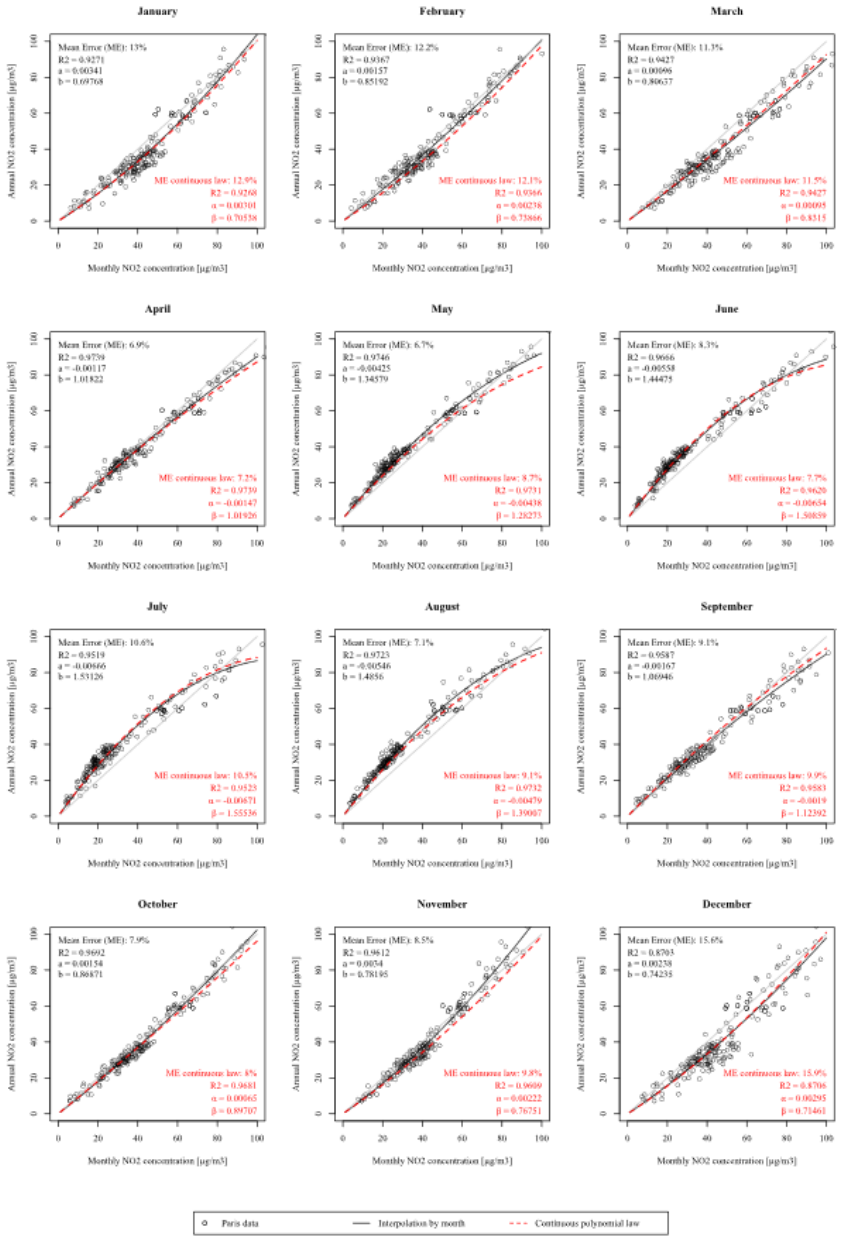


Figure 5.7: Evolution and interpolation of annual NO_2 concentration as a function of monthly NO_2 concentration.

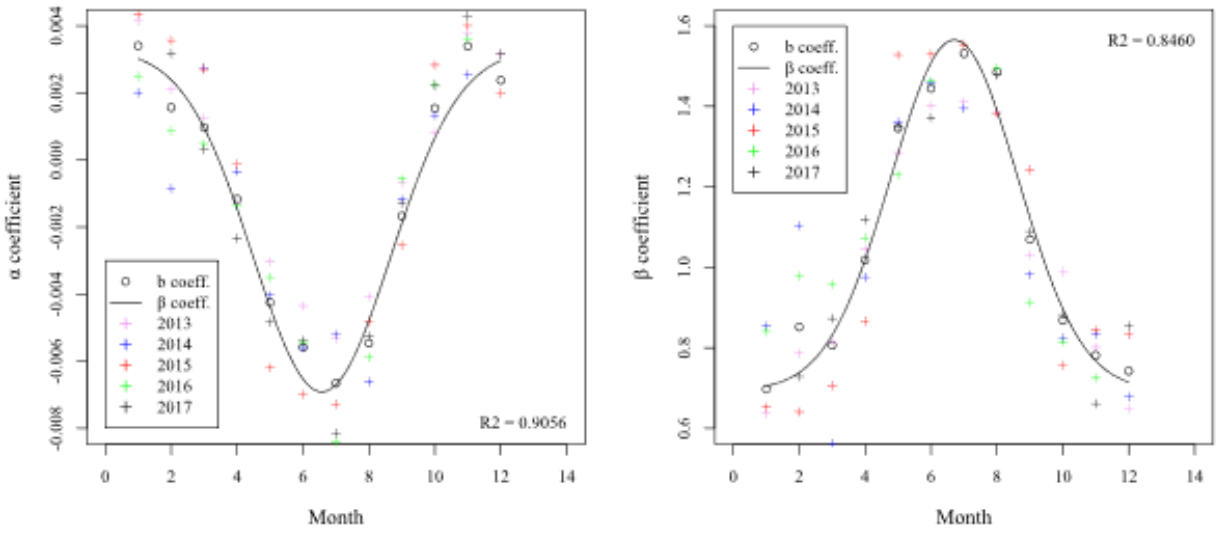


Figure 5.8: Interpolation of a and b coefficients (for each year considered and the subsequent mean) and resulting continuous α and β coefficients.

As shown in 5.8., both coefficients a and b seem to follow a cyclic trend. However, the evolution of the coefficients is inversed with a minimal value of a around June, corresponding to a maximal value of b. On the contrary, the maximal value of a is reached around January, corresponding to a minimal value of b. Considering the trends of a and b observed, a Gaussian function was used to obtain continuous values bringing two new coefficients, α and β, respectively, corresponding to the coefficients obtained from the continuous method. The corresponding equations for α and β are 5.7 and 5.8, respectively, with m being the month corresponding to the available data (e.g. $m = 1$ for the data from the first to the last day of January, $m = 3.5$ for the data from the middle of March to the middle of April, etc.).

$$\alpha = 0.0033 - 0.0102.\exp\left[\frac{-(m\ -\ 6.5749)^2}{8.6962}\right] \tag{5.7}$$

$$\beta = 0.6945 + 0.8708.\exp\left[\frac{-(m\ -\ 6.7076)^2}{7.4328}\right] \tag{5.8}$$

The new curves obtained for each month with 5.6, and the calculated α and β corresponding to a and b respectively, are presented in red dashed lines in 5.7, in addition to the corresponding values of α and β, R2 and the mean error (ME) compared to the Paris data. When comparing these new curves with the previous ones obtained with a and b, they are globally the same except for May and November, for which the curves start to deviate from each other for high monthly NO_2 concentrations. Nonetheless, the mean error for these two

months is still acceptable, with in both cases a mean error of less than 10%. The mean errors for each month are approximately equal between both cases and give an overall error of 10% and a maximal error of 16% in December.

In view to assessing the reliability of the equations, the polynomial methods were applied to several regions of France, including Aquitaine Limousin Poitou-Charentes, Auvergne-Rhône-Alpes and Provence-Alpes-Côte d'Azur from 2013 to 2017. For each month of these years, the mean annual NO_2 concentrations were calculated based on each month of data. The discrete polynomial methods were used here because the information was available for each month. The calculated annual concentrations were then compared to the measured concentrations and a mean error was obtained. The mean errors are summarised in Table 5.2. This table also gives information on the error obtained when the monthly NO_2 concentration is taken directly as an annual NO_2 concentration (called direct approach), and on the improvements between this direct approach and the approach using the suggested methods. For the three regions considered, the mean error using the discrete method is higher than for the Paris region, ranging from 12% to 20%. The errors obtained when using the direct approach range from 18% to 32%. The improvement between the two approaches depends on the regions considered and ranges from 26% to 46% with an overall improvement of 38%. According to these results, the method presented in this paper is reliable and can be used outside the Paris region in France. Overall, this simple applicable polynomial method improves the results in comparison to a direct approach by up to a factor two.

Assessment of annual NO_2 concentration from monthly NO_x concentrations. The final study was performed to give an estimation of the total error when calculating annual NO_2 concentration using monthly measured NO_x data. To manage this, data for the Paris region for the year 2017 were used. Firstly, the monthly NO_2 concentrations were calculated based on monthly NO_x concentrations measurements using the Derwent and Middleton function 5.3. Then, annual NO_2 concentrations were calculated using 5.4, 5.5 and 5.6. The resulting annual NO_2 concentrations were plotted against measured annual NO_2 concentrations and are presented in 5.9. (B). The previous results for Paris from 2013 to 2017 and for which the calculated annual NO_2 concentrations are based on monitored monthly NO_2 concentrations are also provided in 5.9. (A). According to 5.9. (A), a global error of 10% for Paris region is obtained and it can also be seen that the maximal errors occur for the highest NO_2 concentrations. The same observation can be made when comparing this result with those for Paris assessed with the monthly NO_x concentrations for 2017. The global error in this case increases but does not exceed 15%.

Region	Year	NS	$AE1$	$AE2$		$ME1$	$ME2$	
Aquitaine Limousin Poitou-Charentes	2013	31	29%	17%	↓ 41%	30%	17%	↓ 43%
	2014	29	27%	15%	↓ 46%			
	2015	29	32%	17%	↓ 46%			
	2016	35	28%	16%	↓ 44%			
	2017	29	32%	19%	↓ 42%			
Auvergne Rhône-Alpes	2013	50	29%	18%	↓ 39%	30%	18%	↓ 40%
	2014	65	29%	17%	↓ 41%			
	2015	58	30%	18%	↓ 39%			
	2016	68	30%	20%	↓ 35%			
	2017	57	30%	19%	↓ 38%			
Provence-Alpes Côte d'Azur	2013	21	19%	14%	↓ 27%	19%	13%	↓ 31%
	2014	22	19%	12%	↓ 38%			
	2015	29	19%	13%	↓ 29%			
	2016	27	20%	14%	↓ 26%			
	2017	27	18%	12%	↓ 31%			

Table 5.2: Global results of the polynomial discrete method over regions in southern France and improvements compared to the direct use of monthly concentrations as annual concentrations. NS is the number of stations with a full year of data, $AE1$ is the annual mean direct error (in %), $AE2$ is the annual mean discrete method error (in %), $ME1$ is the mean annual direct error (in %), $ME2$ is the mean annual discrete method error (%)

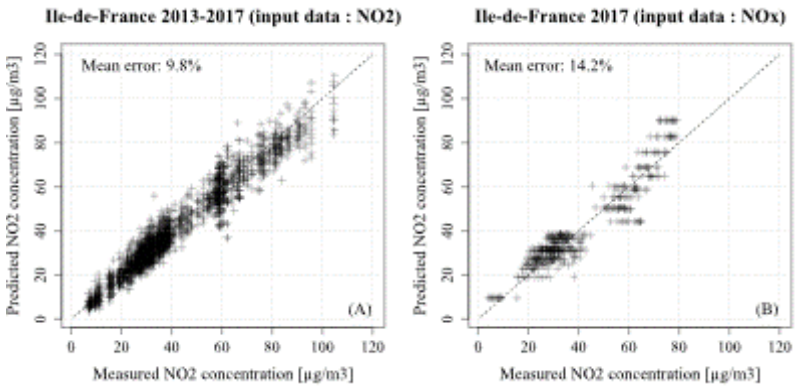


Figure 5.9: Comparison between calculated and measured annual NO_2 concentrations for the Paris region from 2013 to 2017 (A) and for the Paris region based on monthly 2017 NO_x concentrations (B).

5.4 Discussion

The seasonal variability of NO_2 concentrations was shown and leads to higher or lower seasonal NO_2 concentrations compared to annual NO_2 concentrations. An explanation for these observations was proposed and seems to be linked to the seasonal variability of ozone concentrations as well as the seasonal variability of available ozone to react with NO_2. However, this link must be quantified to better explain the phenomenon and evaluate if these observations can be fully generalized. The first hypothesis is that this phenomenon may only be generalizable to countries whose seasonal variability in ozone concentrations are like those observed in France. Thus, in countries having other types of seasons like Indonesia, with only a dry and a monsoon season or India, with winter, summer, monsoon and post-monsoon seasons, the results would be very different, and the equations presented in this paper may not be relevant. However, it may be possible to apply the methodology and adapt the coefficients of the equations to obtain good results in these countries. Nevertheless, this would require long periods of measurements.

It should also be noted that for some specific periods, monthly NO_2 concentrations are representative of annual NO_2 concentrations. Indeed, averaging monthly concentrations measured in March, April, September or October could give good estimations of the mean annual concentrations directly. For these months, it might not be necessary to use the previous methodology to assess the annual NO_2 concentration.

Lastly, the different equations obtained that could be used to asses annual NO_2 concentrations, were built for and applied to regions having around the same latitudes, from 43° to 50°. For a very different latitude, the coefficients of the equations might not be optimized,

and greater errors could occur.

5.5 Conclusion

The assessment of annual NO_2 concentrations with partial data was studied from two main approaches. The first one was to determine the annual mean NO_2 concentration with only annual mean NO_x concentration information. The second was to determine the annual mean NO_2 concentration with only a one-month period measurement. The main conclusions are as follow:

1. Three functions giving annual NO_2 concentrations based on NO_x data were compared. These functions correspond to the methods of Derwent and Middleton, Romberg et al., and Bächlin et al. The results show that the method proposed by Derwent and Middleton is the better suited to assess the annual NO_2 concentration based on NO_x concentrations for several regions of France and for several years both for rural and urban areas in particular. However, this method has some limitations for high NO_x concentrations and gives less accurate results for traffic stations with annual NO_x concentrations higher than 70 µg/m^3. The global error of this method for the regions of France considered is around 8%.

2. NO_2 concentrations are seasonally variable and depend on the concentrations of NO_x and their ratio with VOC concentrations, and on the photochemistry conditions. Hence, making it impossible to give an annual concentration directly from a seasonal concentration: for annual NO_2 concentrations lower than 80 µg/m^3, summer and spring NO_2 concentrations are lower than autumn and winter concentrations; for higher annual NO_2 concentrations, it is the summer and the spring NO_2 concentrations that become higher than the autumn and winter concentrations. Thus, to evaluate an upper limit on annual NO_2 concentration over a short period of time, measurements should be done in winter if an annual concentration of less than 80 µg/m3 is expected, otherwise they should be carried out in summer

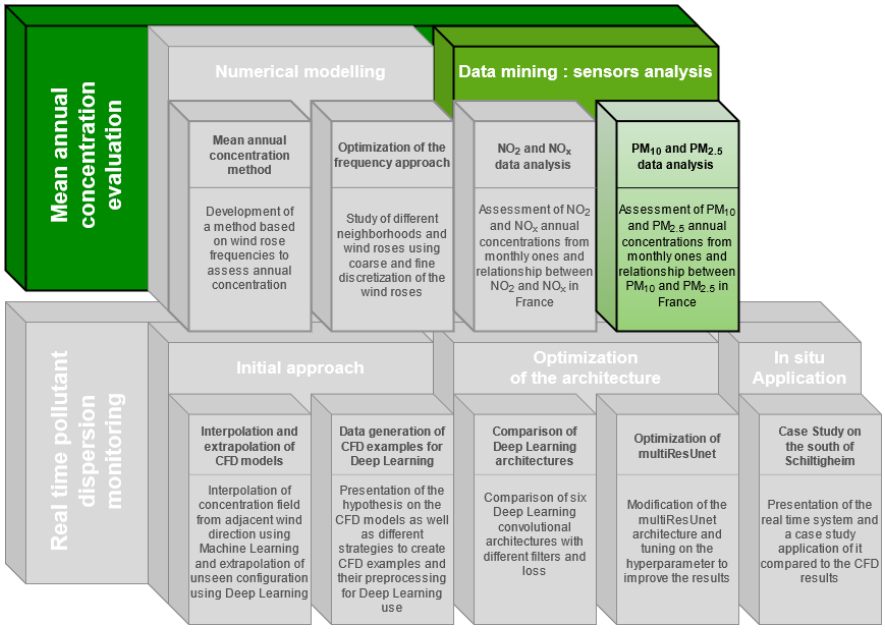
3. Monthly NO_2 concentrations follow the same variability trends as the seasonal concentrations which were quantified for each month. A discrete function was proposed to assess annual NO_2 concentrations based on monthly NO_2 concentrations, yielding a global error of 10% for the Paris region. The corresponding function was made continuous using two Gaussian methods to facilitate its use, leading also to a global error of 10% for the Paris region. The discrete methods applied to the southern regions of France yielded an overall error of 15% and provided an improvement ranging from 26% to 46% compared to the use of the direct approach.

4. Using both the Derwent and Middleton method and the quadratic equations method both presented in this work it is possible to assess annual NO_2 concentrations from monthly NO_x concentrations measurements. Those methods led to an overall error of 15% for the Paris region for the year 2017.

All the results and observations discussed in this Chapter concern NO_x and NO_2 concentrations and it was shown that interesting results can be obtained to reduce measurement periods and estimate NO_2 concentrations from NO_x data without introducing any chemical considerations. This methodology could be extended to other pollutants like particulate matter, which even if not highly chemically active, are subject to specific phenomena like deposition, resuspensions, etc.

Chapter 6

Assessment of mean annual PM_{10} and $PM_{2.5}$ concentration based on a partial dataset



6.1 Introduction

Our recent study has shown the possibility of assessing annual NO_2 concentrations based on monthly concentration and provided a methodology to assess it. Hence, allowing to reduce the necessary monitoring time and consequently the costs (60). The aim of the present study is to check if such methodology can also be implemented for particulate matters (PM_{10} and

$PM_{2.5}$), and also to find if there is any correlation between PM_{10} and $PM_{2.5}$ concentrations evolution. The data used in this study are presented in Section 6.2. The results are then presented in Section 6.3, with the evolution of PM concentration in France, the methodologies to compute annual concentrations from monthly data and to compute $PM_{2.5}$ concentrations from PM_{10} concentrations. A discussion is finally provided in Section 6.4.

6.2 Material and method

6.2.1 Study location

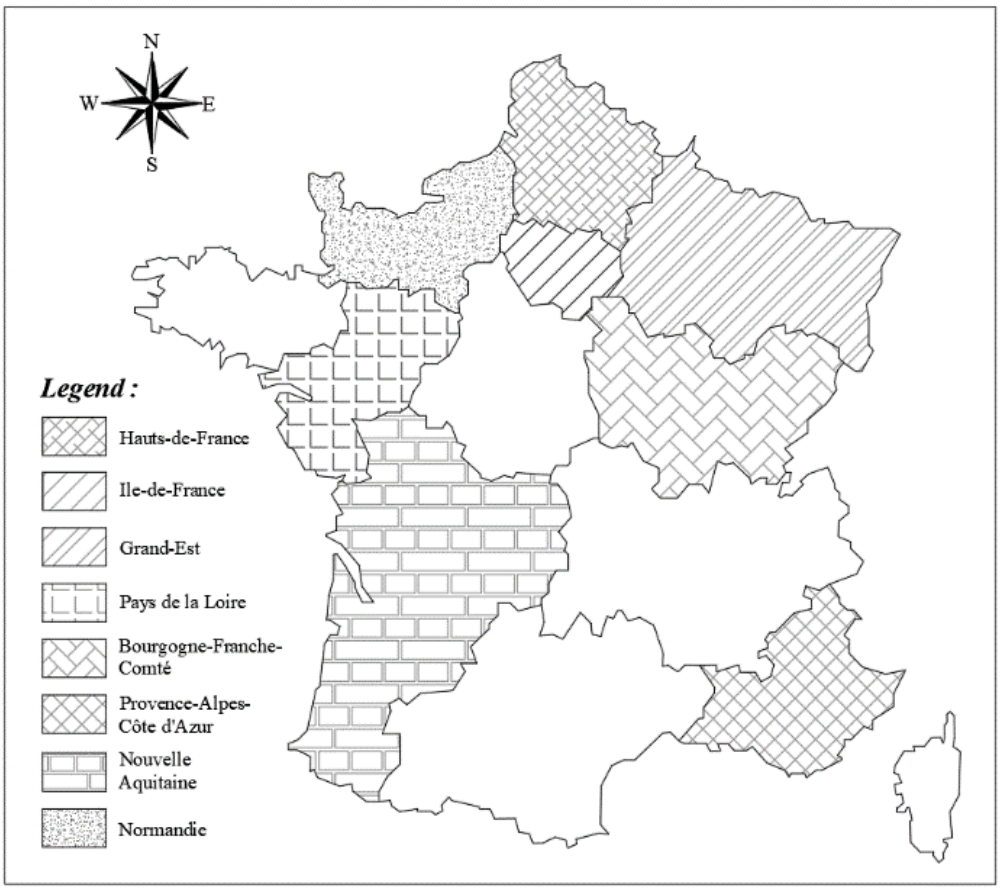


Figure 6.1: Location of the regions where data were used.

This work was performed at a national scale retrieving PM_{10} and $PM_{2.5}$ concentrations monitored in a large number of regions in France. The location of these regions is presented in Figure 6.1, and includes, from North to South: Haut-de-France, Normandie, Grand-Est, Ile-de-France, Pays de la Loire, Bourgogne-Franche-Comté, Nouvelle Aquitaine and Provence-Alpes-Côte d'Azur. A representative view of the variety of locations from France could be

covered by these regions as they are located in several latitude and longitude. But one of the most critical part to select the regions was also related to the data availability.

6.2.2 Data availability

Data were obtained through the open access database provided by the French air quality monitoring authorities known in France as ASQAA, ("Approved Air Quality Monitoring Associations"). Specific demands have also been made when the data provided in the database were not sufficient, especially in terms of temporal resolution.

The data collected are PM_{10} and $PM_{2.5}$ monthly mean concentrations, mainly observed over a nine-years period from 2011 to 2019. A summary of the available data is given in 6.1.

Region	PM_{10}			$PM_{2.5}$		
	Data availability	# data	Relative percentage	Data availability	# data	Relative percentage
Hauts-de-France	2011 - 2019	3,888	15 %	2011 - 2019	1836	20 %
Ile-de-France	2011 - 2019	3,240	13 %	2011 - 2019	1512	24 %
Grand-Est	2011 - 2019	4,536	18 %	2011 - 2019	1836	24 %
Pays de la Loire	2011 - 2019	2,060	8 %	2011 - 2019	750	10 %
Bourgogne-Franche-Comté	2011 - 2019	1,404	6 %	2011 - 2019	1080	14 %
Provence-Alpes-Côte d'Azur	2011 - 2019	3,456	14 %	2015 - 2019	490	7 %
Nouvelle Aquitaine	2012 - 2019	3,648	15 %	-	-	-
Normandie	2011 - 2019	2,808	11 %	-	-	-

Table 6.1: Summary of the available data

6.2.3 Data range

Considering the whole dataset, the monthly mean concentrations range from 24 to 76 µg/m^3 and from 1 to 47 µg/m^3 for PM_{10} and $PM_{2.5}$ respectively, while annual mean concentrations range from 30 to 55 µg/m^3 and 5 to 33 µg/m^3. These concentrations correspond to different types of stations (also called "influence") which include background (66%), industrial (10%), and traffic (24%) stations, as well as different types of area, including rural (14%), suburban (18%), and urban (68%) areas for PM_{10}. The $PM_{2.5}$ station types and influences are not known except for the one where both PM_{10} and $PM_{2.5}$ data exist. There are 25 040 monthly concentrations for the PM_{10} and 7506 for $PM_{2.5}$.

6.2.4 Statistical performance measures

In order to compare the different ways to assess annual PM concentrations from monthly PM concentrations, three statistical performance parameters were considered including 6.1

the coefficient of determination, 6.2 the relative error over a given dataset, and 6.3 the mean relative error over multiple datasets.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(C_i - \widehat{C}_i)}{\sum_{i=1}^{n}(C_i - \overline{C_i})} \tag{6.1}$$

Where R^2 is the coefficient of determination, n is the number of data, C_i is the concentration value, $\widehat{C}_i$ is the predicted corresponding value and $\overline{C_i}$ is the averaged concentration.

$$RE_i = \sum_{j=1}^{n_i} \frac{|p_{ij} - d_{ij}|}{|\overline{d_i}|} \tag{6.2}$$

Where RE_i is the Relative Error for the dataset i, n_i is the number of data in the dataset i, p_{ij} is the predicted value, d_{ij} is the actual data value, $\overline{d_{ij}}$ is the averaged actual data value.

$$MRE = \frac{1}{n}\sum_{i=1}^{n} RE_i \tag{6.3}$$

Where MRE is the Mean Relative Error, n is the number of dataset considered and RE_i is the Relative Error for the dataset i.

In the following, the 95^{th} centile relative error ($C_{95}RE$) is the centile 95 value from the RE_i.

6.3 Results

6.3.1 PM_{10} and $PM_{2.5}$ annual mean concentration trends in France

Figure 6.2 shows the evolution of the PM_{10} and $PM_{2.5}$ annual mean concentration in France between 2011 and 2019. According to these two box plots, the annual mean concentrations are globally decreasing in France since 2011 for both PM_{10} and $PM_{2.5}$. However, the most significant decrease was observed between 2011 and 2015. After 2015, the annual mean concentrations appear to be steady, even if the mean value averaged over all the stations (red line) is still decreasing. In particular, the curve's slope is about -1.3 (respectively -0.4) between 2011 and 2015 (respectively 2016 and 2019) for PM_{10} and about -1.2 (respectively -0.6) for the same periods for $PM_{2.5}$. The slopes variations are of a factor 3.25 and 2 for PM_{10} and $PM_{2.5}$ respectively.

Figure 6.2 gives additional information by distinguishing the annual mean concentrations by the influence of the station for both PM_{10} and $PM_{2.5}$. According to Figure 6.2 (A), a decrease in PM_{10} concentrations is observed whichever the influence considered (i.e. background, industrial and traffic stations), the higher decrease is observed for traffic stations nonetheless. It is the same observation for $PM_{2.5}$ concentration, according to Figure 6.2 (B),

for both background and traffic stations. No comparison can be made between PM_{10} and $PM_{2.5}$ concentrations evolution for industrial stations since there are not this kind of station for $PM_{2.5}$ available.

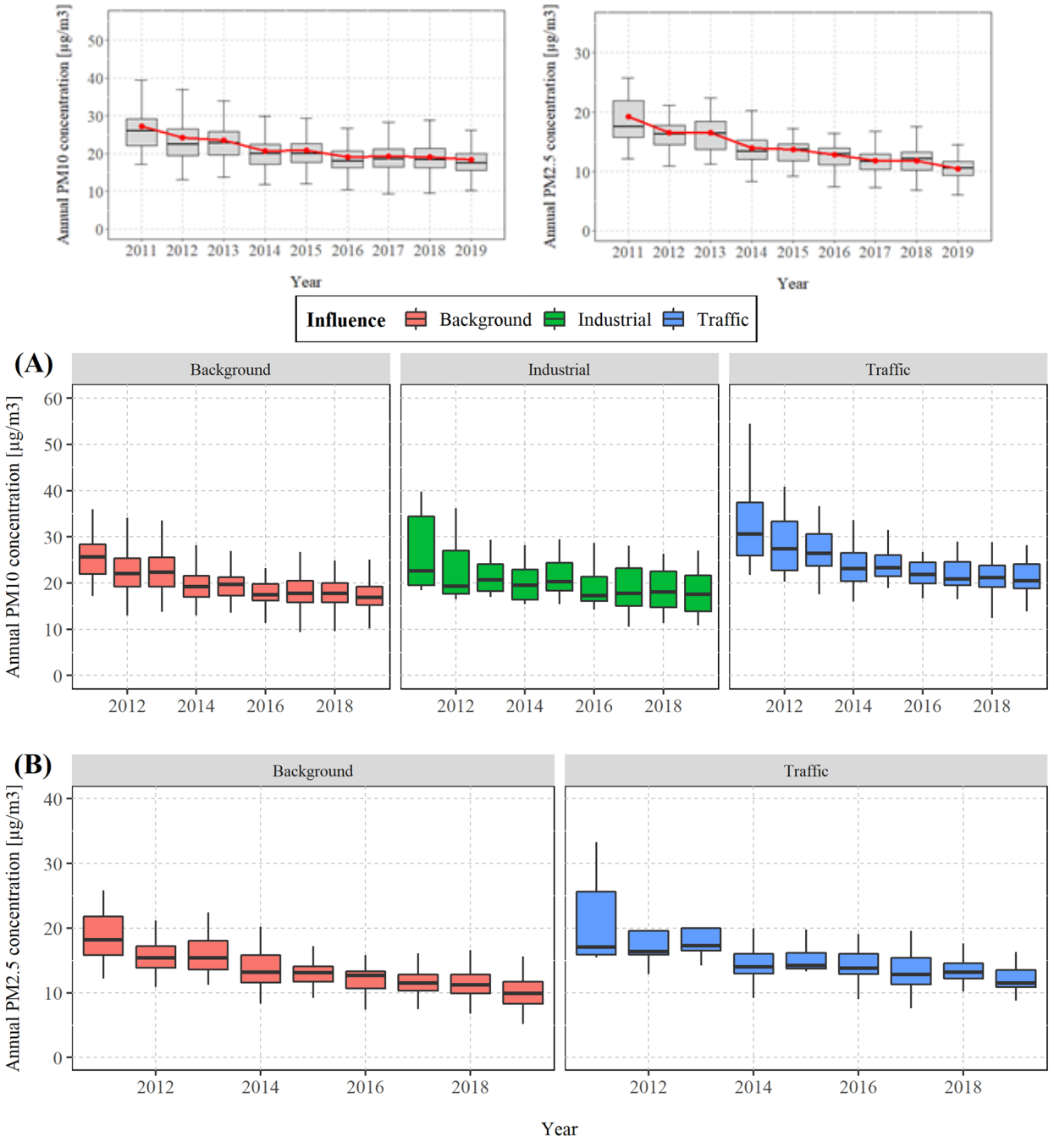


Figure 6.2: Evolution of annual (A) PM_{10} and (B) $PM_{2.5}$ concentrations between 2011 and 2019 for each type of station.

6.3.2 Assessment of annual concentrations based on monthly data

As a first approach, the annual PM_{10} concentrations were plotted as a function of the monthly PM_{10} concentrations considering the whole dataset (including all years, all regions, all types of station and all types of area). The corresponding results, along with a linear regression line, are presented in Figure 6.3 (A). Although the annual concentration appears to be linearly correlated with the monthly concentration, the scatter-plot is widely dispersed around the regression line, leading to a low coefficient of determination ($R^2 = 0.569$). Additionally, considering the obtained line to assess annual concentration from monthly concentration can lead to over or underestimation that can reach 100%.

To improve the results, the same process has been applied considering all months of given years of data. As an example, the results obtained for the year 2019 are presented in Figure 6.3 (B). According to this figure, it can be seen that the linear regression is improved considering given years of data: $R^2 = 0.783$ is obtained for 2019. The over and underestimation are also lowered but can still reach 50%.

The same methodology used for PM_{10} has been conducted on $PM_{2.5}$. As for PM_{10}, it appears that interpolating the annual $PM_{2.5}$ concentration as a function of monthly $PM_{2.5}$ concentration with a linear regression led to poor results when considering the whole data, but also each year separately.

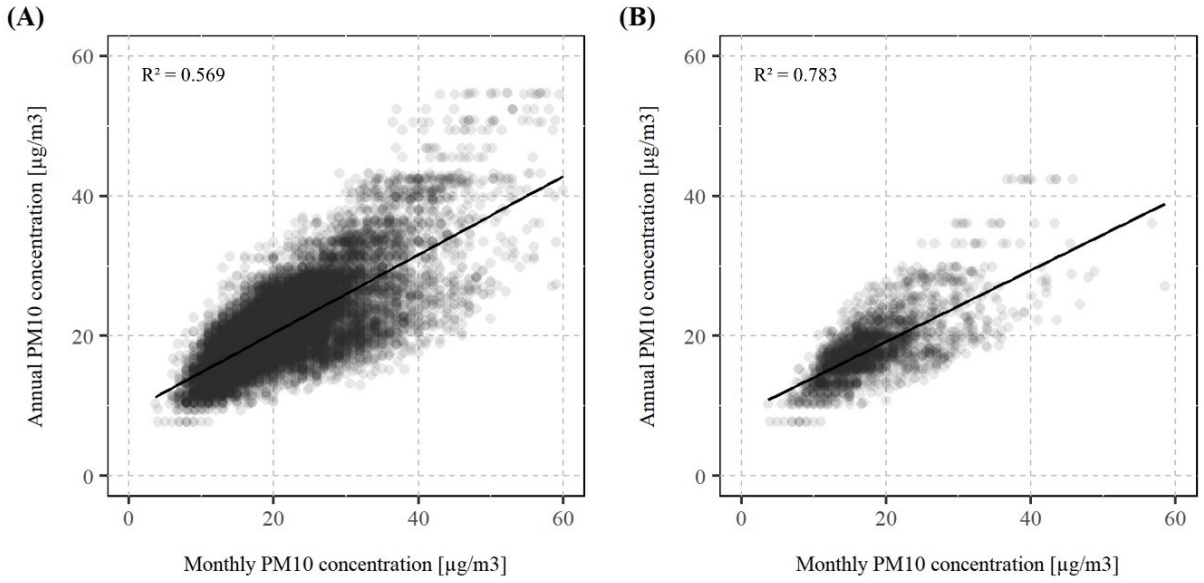


Figure 6.3: Evolution of the PM10 annual concentration as a function of the monthly one considering (A) the whole dataset and (B) only the year 2019.

As seen previously, the solution considering the whole dataset is not satisfying. To improve

the results, the same methodology as in (60) was used. Firstly, it consists to check whether there are seasonal trends on the variation of the particle matter concentration.

To do so, a clustering was performed using MetaboAnalyst package on R (version 3.6.3.). The clustering was performed by combining the heatmap and the hierarchial ascendent clustering, with distance measure using euclidean, and clustering algorithm using ward.D. The combination of heatmap and clustering was performed to evaluate whether a monthly concentration was higher or lower than the annual concentration:

$$\frac{concentration_{month} - concentration_{annual}}{concentration_{annual}} \tag{6.4}$$

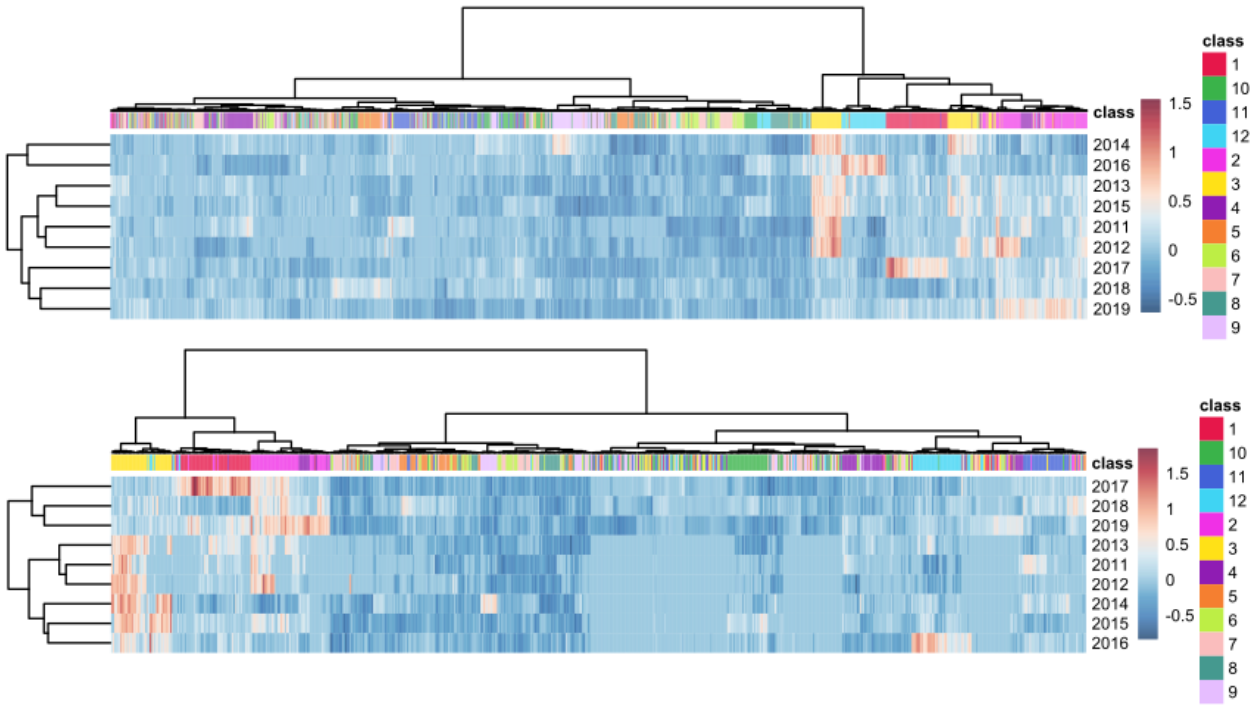


Figure 6.4: Clustering result of PM_{10} (Top) and $PM_{2.5}$ (Bottom) shown as heatmap (distance measure using euclidean, and clustering algorithm using ward.D).

Two main groups can be seen from the clustering, a group containing the winter months could be distinguish with monthly concentrations higher than the annual concentration on average and the rest of the year with monthly concentrations lower than the annual concentrations on average.

- For PM_{10} there are 594 samples representing 28.23% of the data belonging to group 1. It is composed mainly of 4 months making up around 90% of the group with February(27%), January(23%), March(25%) and December(16%).

- For $PM_{2.5}$ there are 204 samples representing 22.47% of the data belonging to group 1. It is composed mainly of 3 months making up 85% of the group with January(32%),February(30%) and March(26%).

The difference between the seasons could be explained by the more diverse source of particulate matter in winter from heating or cold starting vehicles.

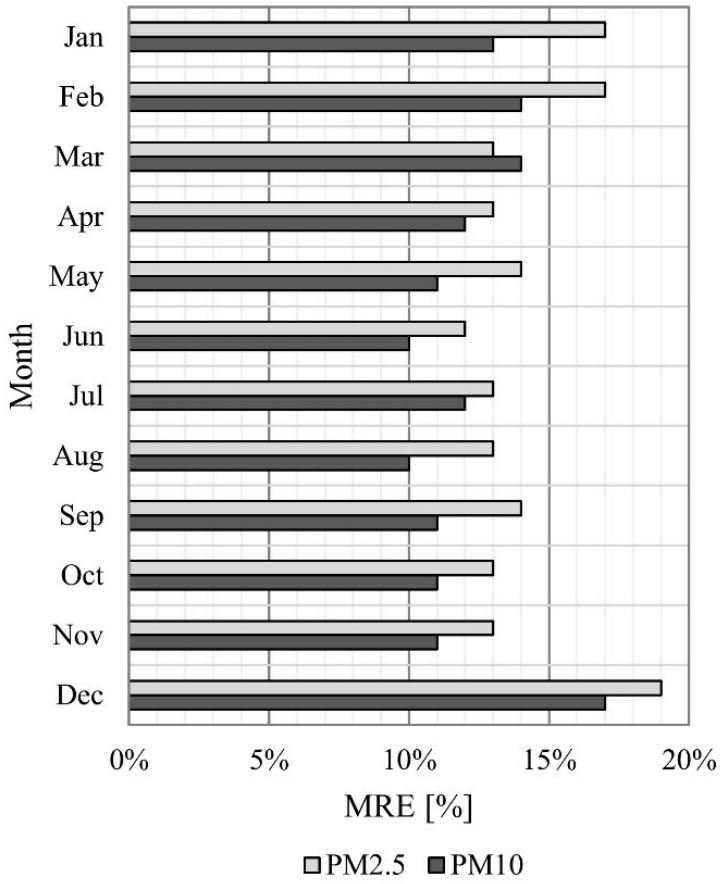


Figure 6.5: MRE from linear regression depending on the months for PM_{10} (in dark gray) and $PM_{2.5}$ (in light gray).

According to Figure 6.5, the results obtained using linear regressions considering monthly concentrations against annual ones show that, when using only one month of measurements, the worst months to evaluate the annual concentration are in winter, (MRE = 0.15 in winter vs. MRE = 0.11 the rest of the year for PM_{10} and MRE = 0.17 in winter, vs. MRE = 0.13 the rest of the year for $PM_{2.5}$). This is probably related to the fact that the mid-months represent half of the annual concentration while winter only represent a third/quarter and thus are more representative of the annual concentration. Another factor might be that

the additional particulate matter from heating may varies between years depending on the climate condition of the winter.

6.3.3 Assessment of annual concentrations based on monthly data by years

These previous predictions were made using all the years. However, as it was seen in Section 6.3.2, considering each year apart improved the results. So, is it possible to improve the results when considering each years independently? Indeed, each year could follow a different trend for instance with a warmer winter leading to less heating.

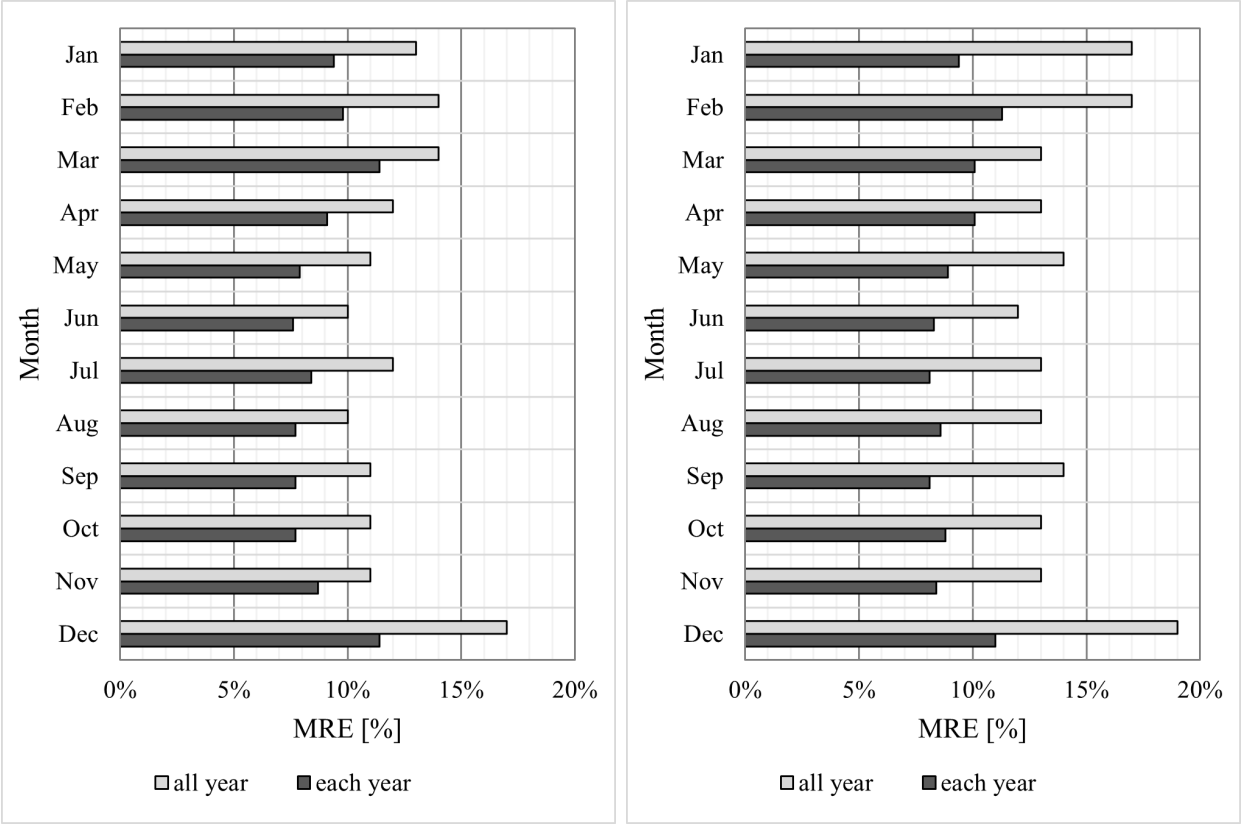


Figure 6.6: Comparison of the MRE when considering for each month all the years at once or one year at a time independently (average result) for PM_{10} on the right and $PM_{2.5}$ on the left

When comparing this approach with the previous one as it can be seen on 6.6, the MRE is lower for PM_{10} for every month with an absolute reduction of -0.033 on the MRE corresponding to a relative reduction of 27% on the MRE on average. As for PM_{10}, the prediction of $PM_{2.5}$ for annual concentration using monthly data is improved when considering each year independently. The MRE is reduced by -0.050 in absolute and 35% in relative on average.

However, there are two major limits with this method of considering each year independently. First, several sensors are necessary in the area that monitor the concentration throughout the year with enough variety to make a regression. This issue is not a big deal when the country is well covered in sensors as it is the case for France. Secondly, waiting for the end of the year to perform the correlation seems to be inconsistent with engineering issue.

6.3.4 Assessment of annual concentrations based on group of months

In France, a popular way to evaluate the annual concentration is to measure one month in summer and one month in winter to have a better representativeness of the seasonal variation during the year. Indeed, several months can be monitored throughout the year to improve the predictions on the annual mean concentration. But it lacks quantitative information on the gains in accuracy of several months measurements compared with one month measurements. To solve this issue, the mean concentration of several couple of months evenly spaced were studied. For instance, if using two months, the mean of the concentrations between January (1^{st} month) and July ($1 + 6 = 7^{th}$ month) will be computed; with 4 months, the average between January (1^{st}), April (4^{th}), July (7^{th}), and October (10^{th}) will be computed; etc.

Figure 6.7 shows a comparison of the MRE for both PM_{10} and $PM_{2.5}$ between the use of a linear regression on the average and using directly the average concentration value of the months (equivalent to a linear regression of bias 0 of and intercept of 1).

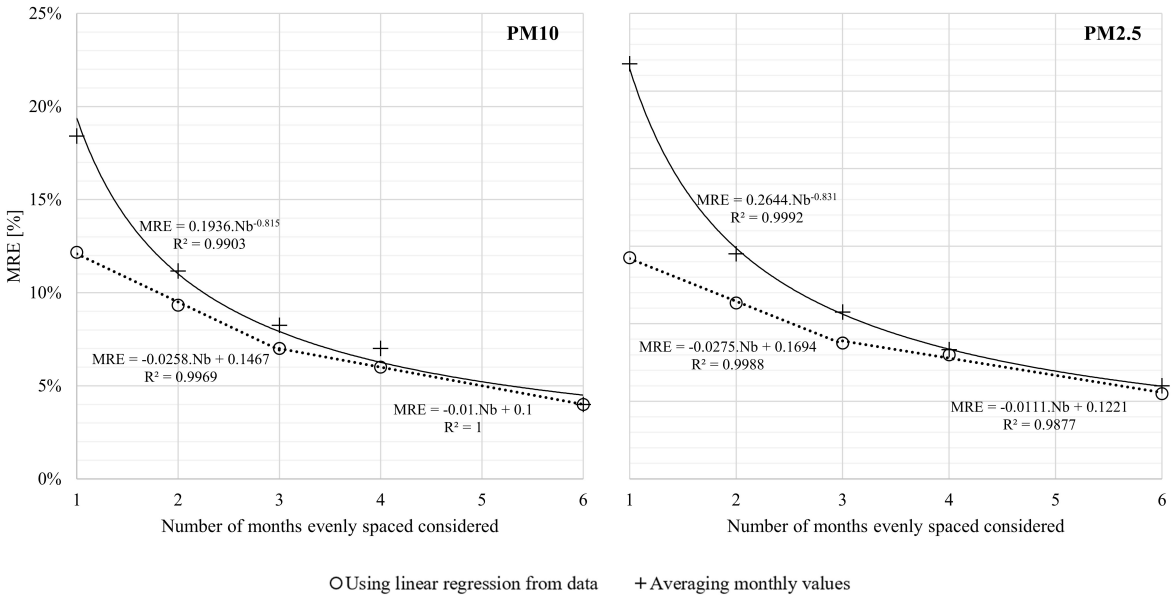


Figure 6.7: Evolution of the mean relative error (A) depending on the number of months for PM_{10} (left) and $PM_{2.5}$ (right) when using the linear regression and directly averaging.

As expected, the more month, the better the results. This results leads to two key points:

1. From one month to three months the slope is stronger than from three months to six months meaning that the gain in error is maximized up to a period of 3 months for both PM_{10} and $PM_{2.5}$.

2. The linear regression improves the results, especially when the number of months used is low. When reaching 3 months, the difference between the linear regression and averaging becomes less than 10%.

An easy way to improve the 2 months results, and also easy to be implemented, is to consider and use the observation made on the different types of months that exist: the best results are obtained when using two months that do not contain winter type of month (i.e., April to October and May to November). Therefore, most likely, if a winter month is considered and represents half of the data when averaging the concentrations, it over represents the winter season. To solve that, weighting the winter-type months (December, January, February and March) by 1/4 and the rest of the months by 3/4 improves the results as well as the stability of the predictions as it can be seen on Figure 6.8.

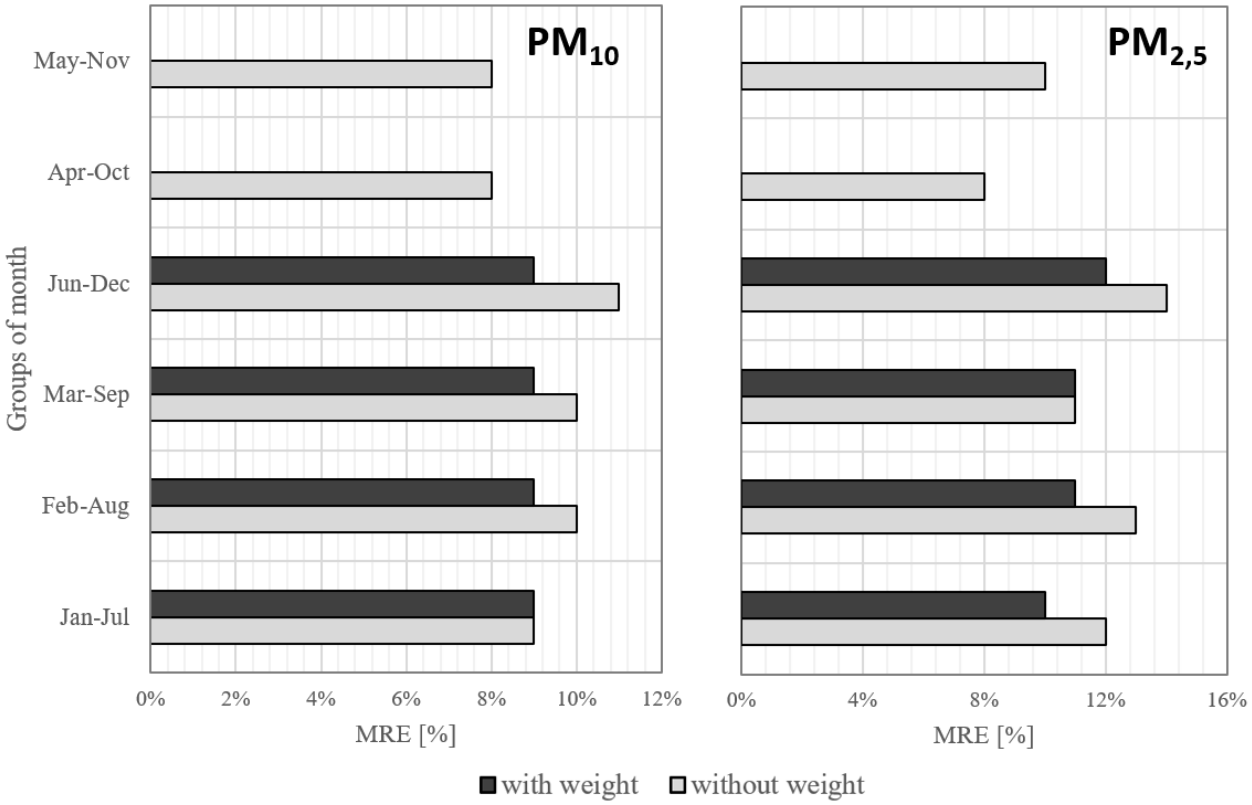


Figure 6.8: Improvement of the mean absolute error when weighting by 1/4 the winter month and rest of the year by 3/4 for PM_{10} left and $PM_{2.5}$ right.

6.3.5 Correlation between MRE and C_{95}RE

To evaluate the error made when using this laws, it can be interesting to compute the percentile 95 relative error. Indeed, for high stakes places regarding air pollution, it can be preferable to overestimate the pollution to be assured that the people in the area will not be confronted to pollution higher than what was expected. It was found that the percentile 95 relative error is linked for every relationship between monthly and annual concentration by 2.6 for PM_{10} as well as for $PM_{2.5}$. For example, if there is an MRE of 10% it means that the C_{95}RE is about 26%.

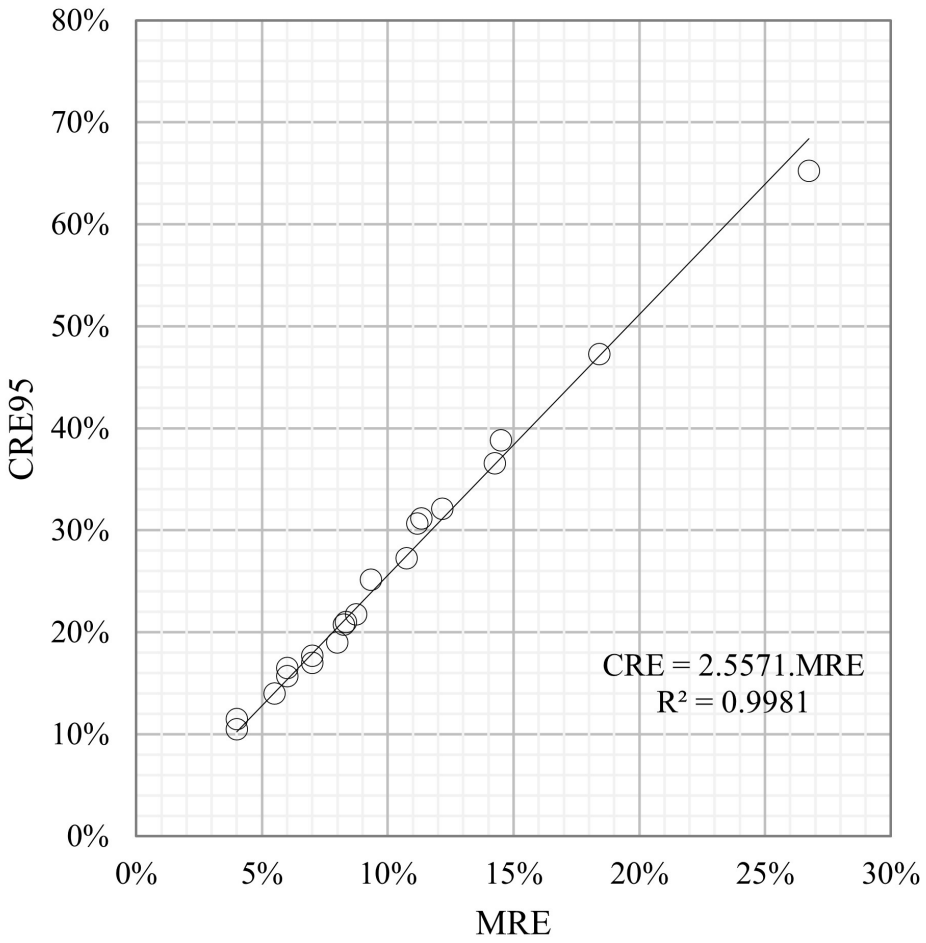


Figure 6.9: Relationship between mean error and 95 percentile error.

6.3.6 Correlation between PM_{10} and $PM_{2.5}$ annual concentrations

An issue that can often happen is that a monitoring site has only data for one of the two types of particles. France is a great example on that issue, as it can be seen with the number

of data, France is better covered for PM_{10} than $PM_{2.5}$. So, is it possible to have an idea of the concentration of $PM_{2.5}$ from PM_{10} and reciprocally PM_{10} from $PM_{2.5}$ if the data are missing?

The data from monthly concentrations when both sensors existed were merged to be compared for a total of 2941 concentrations. The ratio between PM_{10} and $PM_{2.5}$ varies depending on the emission sources, to evaluate this impact, the data were labeled according to their emission type (either traffic or background). The data range is from 0 to 76 µg/m^3 for PM_{10} and 0 to 47 µg/m^3 for $PM_{2.5}$. However, around 90% of background data is under 30 µg/m^3 PM10 and 65% for traffic data which may influence the regression. The plot is presented on the figure 6.10.

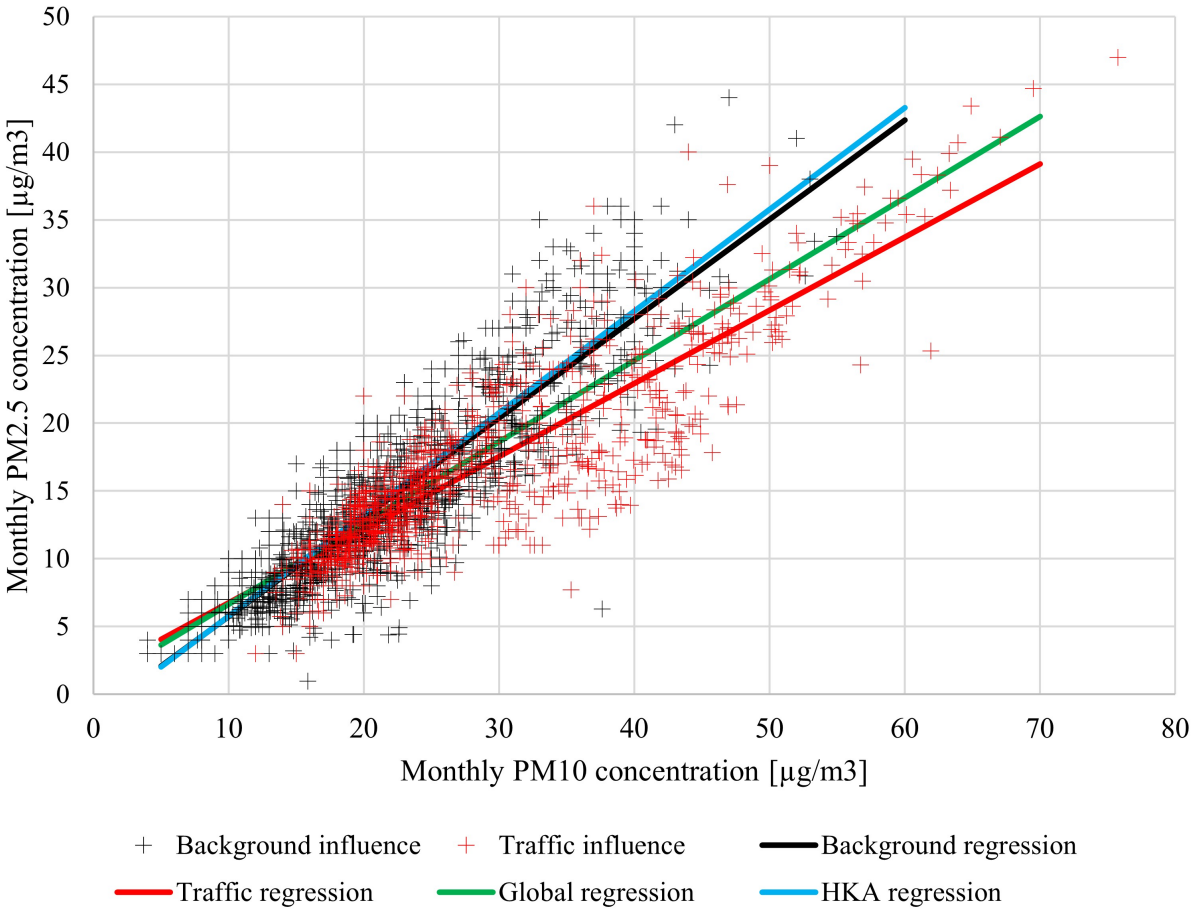


Figure 6.10: Monthly concentration of $PM_{2.5}$ against Monthly concentration of PM_{10}.

The relationship obtained from the data are given in the following table 6.2. This relationship can be used when no better data is available to evaluate either the missing PM_{10} or $PM_{2.5}$ concentrations.

Influence type	Equation	R2	MRE	$C_{95}RE$
Full dataset	$PM_{2.5} = 0.60 \times PM_{10} + 0.63$	0.74	0.17	0.50
Background	$PM_{2.5} = 0.73 \times PM_{10} - 1.58$	0.77	0.15	0.42
Traffic	$PM_{2.5} = 0.54 \times PM_{10} + 1.36$	0.75	0.17	0.44

Table 6.2: Results of the different linear regressions on the full dataset, the background and traffic influence.

The regression obtained in this study are consistent with previous work. For instance, the report from the U.S Environmental Protection Agency 2003 air quality criteria for particulate matter have a ratio $PM_{2.5}/PM_{10}$ of 0.75 for Eastern United States, 0.52 for Central United States and 0.53 for Western United States. The relationship for the background results are very close to the results obtained from Guidelines on the Estimation of $PM_{2.5}$ for Air Quality Assessment in Hong Kong (https://www.epd.gov.hk/epd/english/environmentinhk/air/guide_ref/guide_aqa_model_g5.html) in which they used 10 years (from 2002 to 2011) measures from 5 stations excluding stations that were traffic dominant and in which they reported the following relationship for the daily concentrations:

$$PM_{2.5} = 0.75 \times PM_{10} - 1.72 \tag{6.5}$$

Hence, it can probably be assumed that the monthly relationship found can also be used for daily concentrations.

Nevertheless, it can be noted that other works find significant differences between winter and summer ratio as in (131) and (49). Following our 2 types of months shown in Section 6.3.2, the regression have been calculated and are presented in table 6.3.

Season type	Equation	R2	MRE	$C_{95}RE$
Winter month	$PM_{2.5} = 0.61 \times PM_{10} + 2.37$	0.75	0.14	0.40
Rest of the year months	$PM_{2.5} = 0.54 \times PM_{10} + 1.11$	0.72	0.16	0.46

Table 6.3: Results of the different linear regressions by months types with winter, intermediate months and mid-year months respectively (Jan, Feb, March),(Apr,May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)

The results can be improved by combining the two approaches. If the knowledge on

both the influence type and season type is known, one can compute the two regression and calculate the mean.

For instance, if a value of PM_{10} is known to be from a traffic related influence in a winter month, $PM_{2.5}$= (0.54x PM_{10}+1.36 + 0.61xPM_{10}+2.37)/2 = 0.575xPM_{10}+1.875

This strategy enables to improves the results greatly with a R2 of 0.80, a MRE of 0.12 and $C_9$5RE of 0.33.

6.4 Discussion and perspectives

The results presented here provide useful information on the behavior of PM_{10} and $PM_{2.5}$all around France and should also work for comparable countries based on the climate and lifestyle (i.e. main of the West European countries). Nevertheless, the results obtained here are most likely not applicable everywhere. Indeed, for instance, the seasons impact on the particulate matter will be different in places with different climate. Thus, the results obtained in this study must be considered with geographic and lifestyle parameters. The proposed relationship most likely works in other places with 4 seasons such as the rest of Europe, but it would need to be confirmed with local data from other countries. In countries with completely different climate the reasoning could be applied if data are available to determine the correct local regressions. The relationship described here between PM_{10} and $PM_{2.5}$ most likely works in other part of the world since it is in accordance with results from previous study in China and United States. A second limit of the results proposed here are the ranges of values. Using the regressions outside the range in extrapolation may lead to greater errors. Therefore, another perspective would be to have wider ranges of values to see if the relationship still works for higher values.

The strategy elected (number of months, each year vs all year) to measure the annual concentration depend on two parameters: the needs and risk acceptance. Indeed, when using one month, the winter months gives the worst results. Nonetheless, it is often preferable to overestimate the pollution for safety reasons for the dwellers. So, it could be a strategy to measure only in winter as it does overestimate the annual concentration generally using the raw value of measurement without applying the regression law given in the annex. The results on using group of months give quantitative information about it. The MRE gives an idea of the mean error that can be made depending on the number of months elected: if the user aims to be below 10%, two months of monitoring must at least be performed. Nevertheless, it is representative of the mean result, but for safety measure, it may be required to be sure not to underestimate the annual concentration. In such case, considering the dispersion and the chances of having an underestimation, the $C_{95}RE$ may be used to lower the chances of underestimating the pollutant concentrations.

6.5 Conclusion

This work studies the assessment of annual particulate matter (PM_{10} and $PM_{2.5}$) with partial data around two aspect. First, by determining the annual concentration with monthly concentrations. Secondly the relationship between PM_{10} and $PM_{2.5}$ in France and in the world considering previous works from other authors. The main conclusions are as follows:

1. There is no general trend to assess particulate matter annual concentrations from any month.

2. Two types of behavior are highlighted regarding monthly concentrations against annual ones, winter months that overestimate annual concentrations and the rest of the year months that underestimate.

3. For each months, the annual concentration against the month concentration have been studied and regressions laws determined. The month with the best MRE are the mid years months and winters the worst (MRE = 0.14 in winter vs. MRE = 0.11 the rest of the year for PM_{10} and MRE = 0.17 in winter March excluded, vs. MRE =0.13 the rest of the year for $PM_{2.5}$).

4. The error can be decreased when restricting the sensors data to the current year to make the regression law by 27% on average for PM_{10} and 35% on average for $PM_{2.5}$. However, to make new regression law for each year, it is necessary to wait the end of the year and to have permanent sensors in the area covering a wide range of concentrations.

5. Another strategy using several months can also improve the results. The mean relative error using the linear regression laws goes from 12% and 14% with one month to 4% and 6% for six months for PM_{10} and $PM_{2.5}$ respectively. The more months, the less difference there is between the regression and directly using the average concentration of the group of months.

6. The gain in accuracy is stronger up to 3 months period than from 3 months to six months of monitoring.

7. When using two months, if a winter month is present, weighting it by 1/4 while the other month is weighted by 3/4 reduces the error and improve the stability of the predictions.

8. Quantitative measures as well as all regression laws are given with the mean absolute relative error. The choice of strategy should be done depending of the risk acceptance and cost of campaign measurement.

9. MRE and C_{95}RE are proportional by a factor of 2.6. For high stakes areas, it can be useful to multiply the concentration obtained through monitoring by the C_{95}RE to ensure not underestimating the atmospheric pollution in fine particles.

10. If no better option is available, PM_{10} and $PM_{2.5}$ can determine the other using a linear law. The results can be improved by knowing the influence either background or traffic and the month type, either winter, intermediate or mid-year or the best, knowing both and averaging the results of the two linear regressions.

The perspective of this work could be to compare the monthly/annual concentration results with other countries, applying the same methodology to other pollutant (except for NO2/NOx that were already tested in a previous work) and improving the range of concentration of particulate matter to be able to apply these solutions in more polluted areas.

6.6 Annex

nb month	group month	intercept	bias	R2	MRE reg	MRE avg
1	jan	0.59	7.66	0.6	0.13	0.18
1	feb	0.52	8.08	0.6	0.14	0.27
1	march	0.46	8.31	0.58	0.14	0.34
1	april	0.69	6.16	0.7	0.12	0.14
1	may	0.78	6.48	0.77	0.11	0.16
1	june	0.88	4.76	0.8	0.1	0.14
1	july	0.78	5.94	0.72	0.12	0.15
1	aug	0.86	6.2	0.79	0.1	0.19
1	sept	0.73	7.39	0.73	0.11	0.17
1	oct	0.8	5.08	0.78	0.11	0.13
1	nov	0.72	6.23	0.76	0.11	0.13
1	dec	0.52	10.02	0.39	0.17	0.21
2	jan-july	0.86	2.96	0.83	0.09	0.1
2	feb-aug	0.77	4.7	0.8	0.1	0.11
2	march-sept	0.68	5.15	0.77	0.1	0.16
2	april-oct	0.88	2.7	0.88	0.08	0.08
2	may-nov	0.86	4.16	0.88	0.08	0.1
2	june-dec	0.87	3.67	0.73	0.11	0.12
3	jan-may-sept	0.93	2.32	0.93	0.06	0.07
3	feb-june-oct	0.9	1.93	0.92	0.07	0.07
3	march-july-nov	0.84	2.15	0.91	0.07	0.1
3	april-aug-dec	0.98	1.3	0.88	0.08	0.09
4	jan-april-july-oct	0.96	0.95	0.94	0.05	0.06
4	feb-may-aug-nov	0.9	2.53	0.93	0.06	0.07
4	march-june-sept-dec	0.92	1.19	0.9	0.07	0.08
6	jan-march-may-july-sept-nov	0.93	1.06	0.97	0.04	0.04
6	feb-april-june-aug-oct-dec	1.01	0.21	0.97	0.04	0.04

Table 6.4: Linear regression coefficient and results between the monthly and annual concentrations for the different group of months for PM_{10}, and the MRE score, when using the regression (noted MRE reg) and when directly averaging (noted MRE avg)

nb month	group month	intercept	bias	R2	MRE reg	MRE avg
1	jan	0.4	6.78	0.39	0.17	0.31
1	feb	0.4	6.29	0.45	0.17	0.41
1	march	0.38	5.96	0.65	0.13	0.48
1	april	0.61	5.11	0.62	0.13	0.17
1	may	0.69	5.42	0.61	0.14	0.21
1	june	0.9	4.04	0.71	0.12	0.23
1	july	0.83	4.84	0.69	0.13	0.24
1	aug	0.89	4.92	0.7	0.13	0.29
1	sept	0.6	6.97	0.61	0.14	0.27
1	oct	0.67	5.09	0.65	0.13	0.18
1	nov	0.61	4.82	0.7	0.13	0.17
1	dec	0.39	7.8	0.26	0.19	0.25
2	jan-july	0.79	2.7	0.72	0.12	0.14
2	feb-aug	0.72	3.52	0.7	0.13	0.15
2	march-sept	0.56	4.95	0.76	0.11	0.21
2	april-oct	0.85	2.36	0.84	0.08	0.1
2	may-nov	0.79	3.26	0.8	0.1	0.12
2	june-dec	0.83	3.16	0.6	0.14	0.15
3	jan-may-sept	0.89	1.92	0.86	0.09	0.1
3	feb-june-oct	0.89	1.32	0.85	0.09	0.09
3	march-july-nov	0.72	2.9	0.88	0.08	0.13
3	april-aug-dec	1	0.9	0.83	0.09	0.11
4	jan-april-july-oct	0.96	0.62	0.93	0.07	0.07
4	march-june-sept-dec	0.81	2.21	0.85	0.09	0.1
4	feb-may-aug-nov	0.9	1.49	0.89	0.08	0.08
6	jan-march-may-july-sept-nov	0.88	1.32	0.96	0.05	0.06
6	feb-april-june-aug-oct-dec	1.05	-0.25	0.94	0.06	0.06

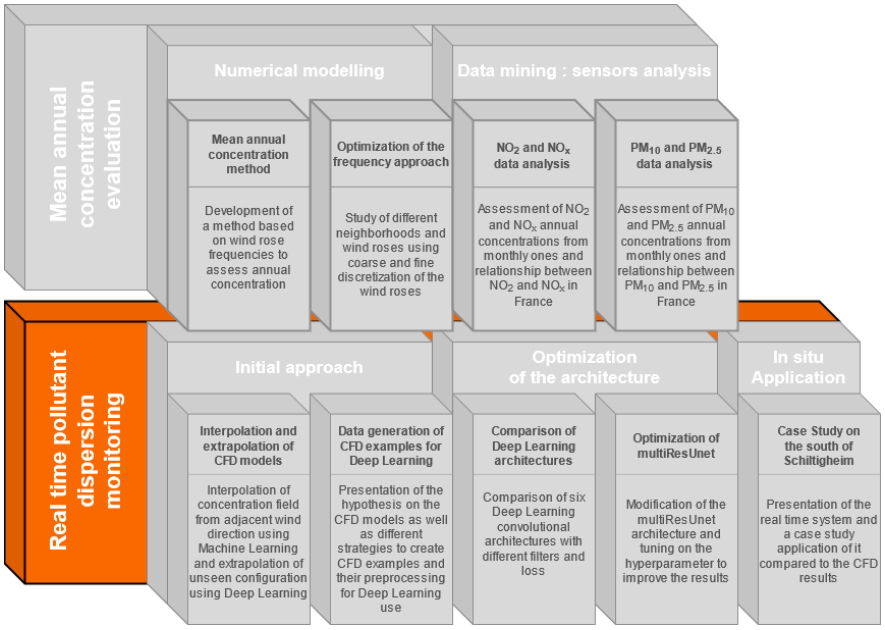
Table 6.5: Linear regression coefficient and results between the monthly and annual concentrations for the different group of months for $PM_{2.5}$ and the MRE score, when using the regression (noted MRE reg) and when directly averaging (noted MRE avg)

Part II

Deep learning models to estimate urban pollution in real time

Chapter 7

How can artificial intelligence be used with CFD to achieve real time pollutant dispersion?



CFD models for local urban pollution are among the best models in terms of accuracy since they consider complex phenomena such as turbulence induced by buildings. Nevertheless, real time solution for pollutant dispersion with CFD is nowadays not possible because it requires too much computing resources. Hence, faster model must be used. Unfortunately, these models such street canyon or gaussian plume are less accurate and do not consider buildings in a satisfying manner. Hence, new methods must be developed. A fast pace rising field that managed to change the paradigm in a lot of domains is machine learning

and artificial intelligence. Is it possible to associate the recent advances in this domain in association with CFD for real time pollutant dispersion monitoring?

7.1 Artificial intelligence brief history

Artificial intelligence is an old concept that can be traced back as far as the Greek antiquity with Homer and Hesiod around 700 B.C. and the Greek mythology. Indeed, in the myth, Hephaestus, god of fire, smith, metallurgy and volcanoes invented several machines or beings that were capable of self governance, called automaton meaning "self moving". He created for instance Talos, a giant bronze humanoid construct to defend the island of Crete or Pandora an artificial woman made out of clay and water sent to punish man by Zeus for having discover fire. Both stories finishing pretty badly for humankind.

Humans tried for long to imitate life and reasoning by creating automaton using various techniques from water machines or steam activating mechanism that would allow an object to move by itself in an "programmed" pattern. Among them, the famous Turk chess player automaton created in 1769 that could play chess against a human opponent. Even though the real reflective part was made by a human chess player, it still questioned at its time whether a machine could or could not think and apply reasoning. The French philosopher Descartes was persuaded that life was like machinery and given enough knowledge about it could be replicated. In the modern era, this idea was about to get a new light.

The beginning of modern artificial intelligence can be traced back to 1943 with (93) who described the first mathematical neuron's model that would be the groundwork for the others to follow. Then, in the early fifties with the emergence of computer with the ground founding work of Alan Turing, artificial neuron became a possibility. Alan Turing wanted machines that could learn and alter its programming. A machine that learned and evolve by itself as it grew as human kids do. In 1957, with the creation of what is considered the first artificial neuron with the perceptron (134), it was thought that many issues could be resolved using this algorithm. Nevertheless, it revealed itself as being usable in only a narrow number of issues. This caused scientists and industries to reduce their funding in the domain. Still, lots of models and advancement were made in the second half of the 20th century as it can be seen on Figure 7.1 made by Favio Vazquez (source: https://medium.com/@faviovazquez). In the past decades, the abundance of data and the improvement of processing capabilities of computer, especially thanks to the development of Graphical Processing Unit (GPU), have permitted for neural network to outperform other classical method as shown by AlexNet in 2012 (69).

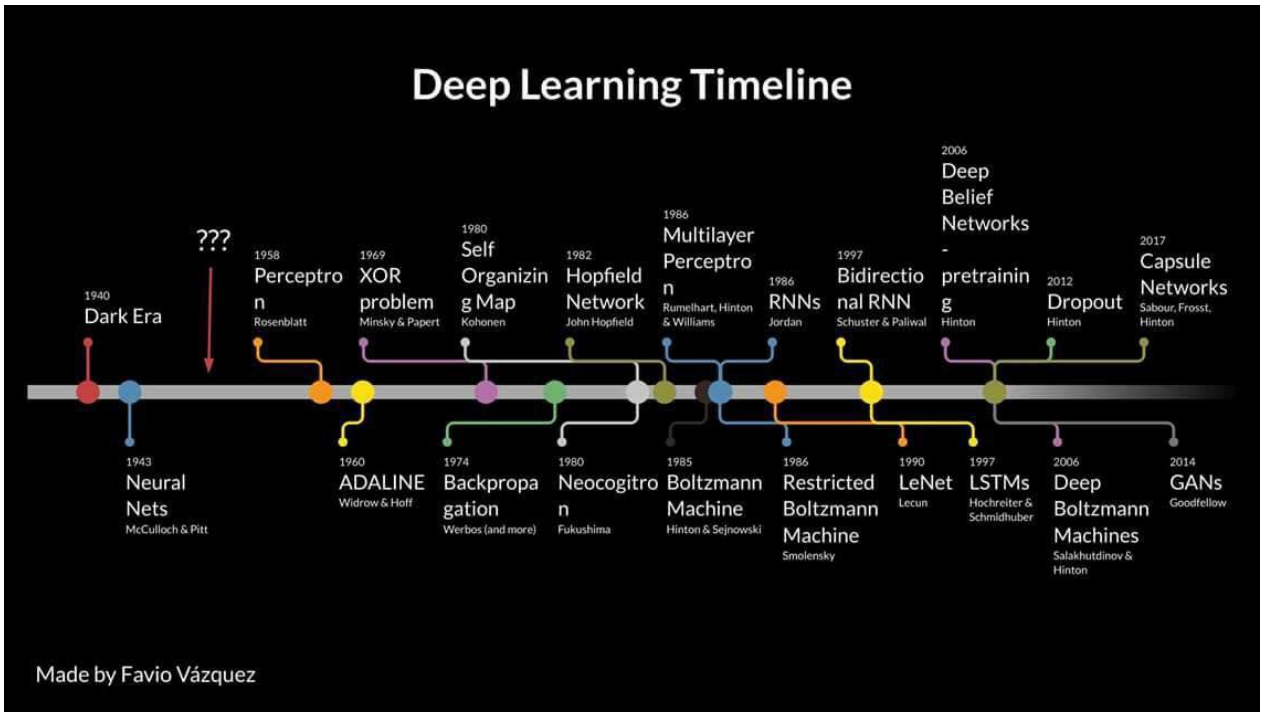


Figure 7.1: Deep Learning timeline made by Favio Vazquez

7.2 Artificial intelligence concepts behind the words

Today, *Artificial Intelligence* (AI) word is often used more as a commodity or for marketing purposes. Two kinds of artificial intelligence can be described:

- Strong AI: that are generally the concept that people see behind the words AI. It is a machine or a program that could mimic human mind, with complex thinking, decision making and planning. At the current time, this kind of AI does not really exist.
- Weak AI: most of the time, the term AI in sciences is used to describe this kind of AI. They are programs or machines very specialized in a relatively small number of tasks that resemble human intelligence such as self driving cars or object detection or language translation.

Artificial intelligence is a huge topic with different sub categories. For instance Expert systems is a computer program that uses artificial intelligence (AI) technologies to simulate the judgment and behavior of a human or an organization that has expert knowledge and experience in a particular field. The rules that the Expert system uses are given explicitly by the creator which will mimic human reasoning based on this set of rules. An example of such system are chatbot for after-sales services. Another field that has caught a lot of

attention recently is machine learning. It is defined according to Stanford University as "the science of getting computers to act without being explicitly programmed.". Machine learning is particularly interesting since it can infer rules and patterns from examples without explicit set of rules which may not be known by the researchers or would be too costly to program.

Machine Learning can be divided into three main categories:

- Supervised Learning: It consists to give to a model couples of input/output that the model must manage to reproduce. An analogy with functions would be to give x/y couples and the architecture must find the function f that link them as $f(x) = y$. Example: The work treated in this thesis, i.e., giving building layouts and pollution sources to obtain the dispersion pollution map.

- Unsupervised Learning: inputs are given to the model without explicit outputs, the model must find by itself the patterns and discovers itself the outputs. Example: Clustering the habits of users depending of their characteristics such as age, sex or income.

- Reinforced Learning: The model must interact and act on its environment, and change its behaviour depending on the modification of the environment. Example: Robotics, self driving cars, etc.

7.3 Deep Learning concept

Recently, a new category of machine learning has emerged. The field of Deep Learning. It can be dated back to 2012 when google presented an unsupervised Deep Learning architectures that was trained on millions of images to recognize cat with a success rate of 75 %. Deep learning is according to Oxford dictionary "*a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher level features from data.*". It is a neural network with a lot of layers and parameters. Each deeper layer offers a level of abstraction in which rules can be inferred. Despite the fact that the know how exist for a long time, it has only been recently used successfully because of several issues that comes with deep learning models such as the computation requirements or the exploding/vanishing gradient. The exploding/vanishing gradient to explain it simply is that the signal that the deep learning model uses to learn from the data at each step must propagate in the whole architectures. Nevertheless, the signal can either explode or disappear making the more profound layer from the signal useless since they can not learn from the signal.

Deep Learning has since been applied to many domains with several original architectures and type of models adapted to the requirements. To cover some popular architectures and

their uses :

1. The classical deep learning architecture is the multiperceptron. It is an architecture made up of many fully connected perceptrons. They have an input, several hidden layers and an output. They have been used for image classification or speech recognition for instance. A diagram of the architecture is presented below 7.2

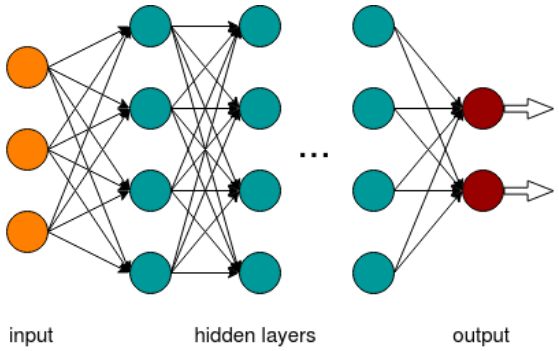


Figure 7.2: multiperceptron diagram

2. There are architectures that have been made specifically to treat spatial information. Convolutional Neural Network (CNN) uses convolutional filters to extract spatial information and then a fully connected layers network to perform regression or classification (73; 144). They have been used for instance for object detection and image classification. A diagram of the architecture is presented below 7.3

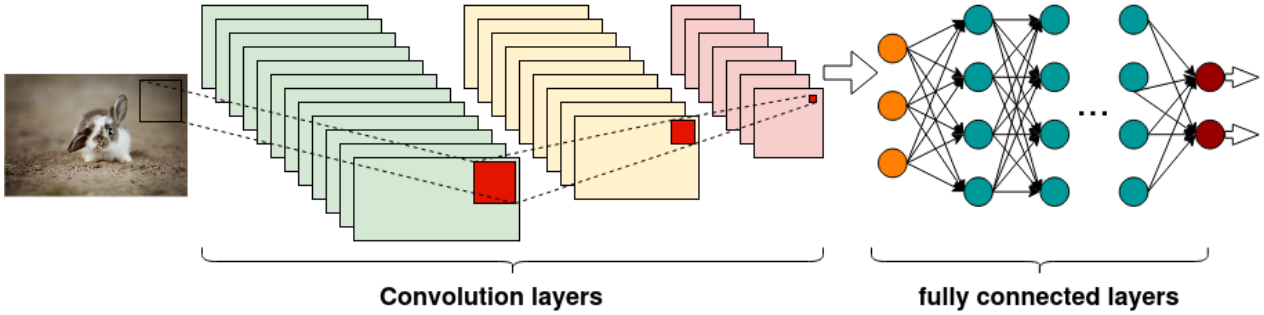


Figure 7.3: Convolutional Neural Network diagram

Autoencoder created by Geoffrey Hinton was also designed to treat spatial information. It is composed of two part, an encoding part that reduce the dimension of the input data and a decoding part which allows to extract information. Autoencoder have been used for instance for image classification. A diagram of the architecture is presented below 7.4

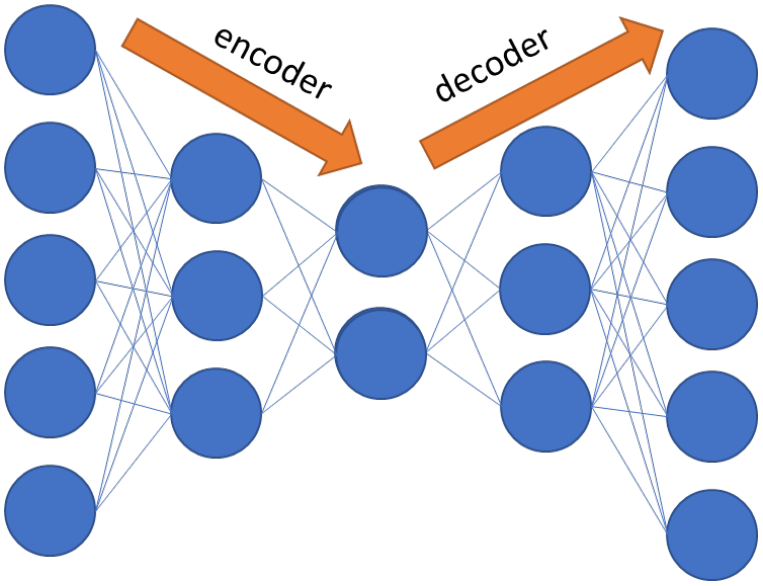


Figure 7.4: autoencoder diagram

3. Generative Adversarial Networks (GANs) are two neural networks working in pairs. Two sets of data are created, one from real data and the other one generated by the generator neural network. The discriminator neural network must then distinguish if the data is from the real dataset or the generated dataset (44). GANs have been used for instance to create fake images of faces of non existing people. A diagram of the architecture is presented below 7.5

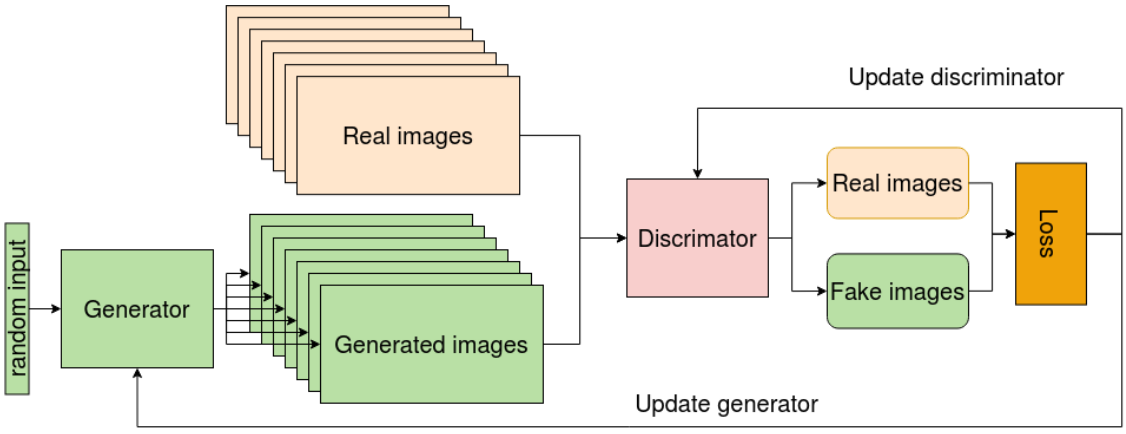


Figure 7.5: Generative adversarial networks

4. There are architectures that are made to treat sequential information. Recurrent Neural Network (RNN) gets input from both the input layer and the previous time step hidden

layer. This allows the network to have a memory of precedent occurrences. A diagram of the architecture is presented below 7.6

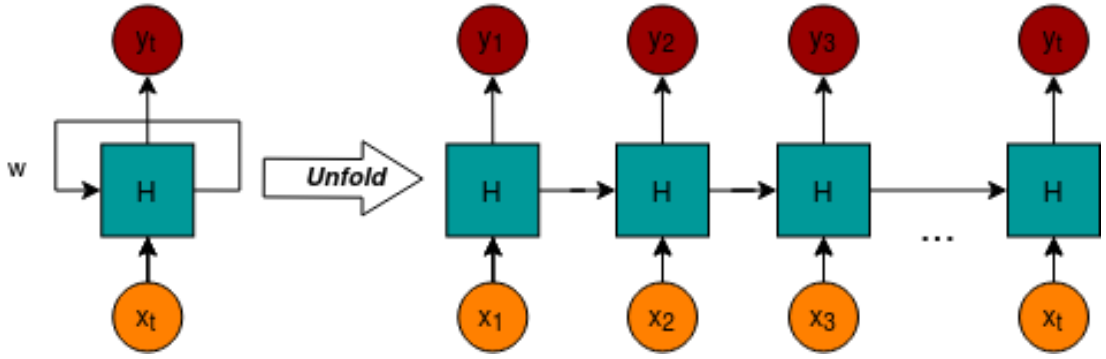


Figure 7.6: Recurrent Neural Network diagram

Long Short Term Memory (LSTM) network works like RNN, but have the capacity to forget partially previous hidden state to avoid vanishing or exploding gradient issues over long sequences. A diagram of the architecture is presented below 7.7

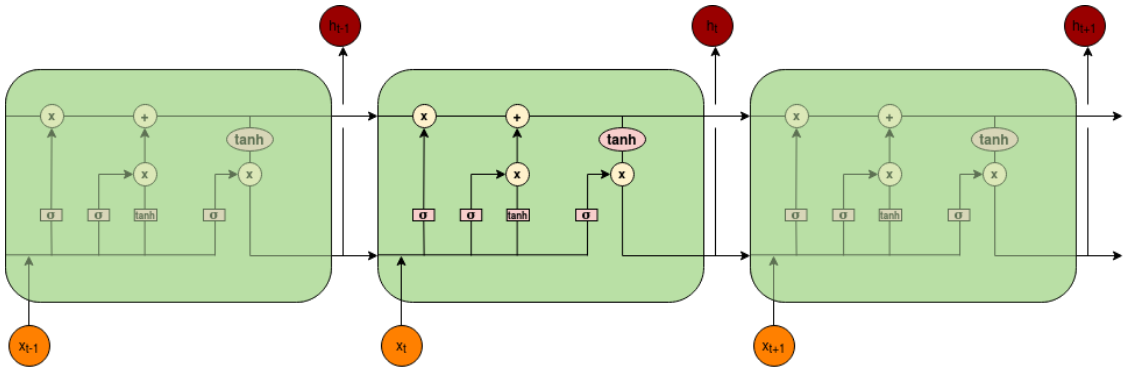


Figure 7.7: Long Short Term Memory diagram

RNN and LSTM have been used for instance for sensors analysis or language translation (85; 94).

5. Graph neural networks are network that map relationship between different elements with their characteristics and the interactions between the elements. For instance a node can be a person define by its age, sex, work, etc. and the interactions the number of messages they send to others persons. Graph neural network are used for instance by social media to map their users with same interest and predict what they may want to buy (39). A diagram of the architecture is presented below 7.8

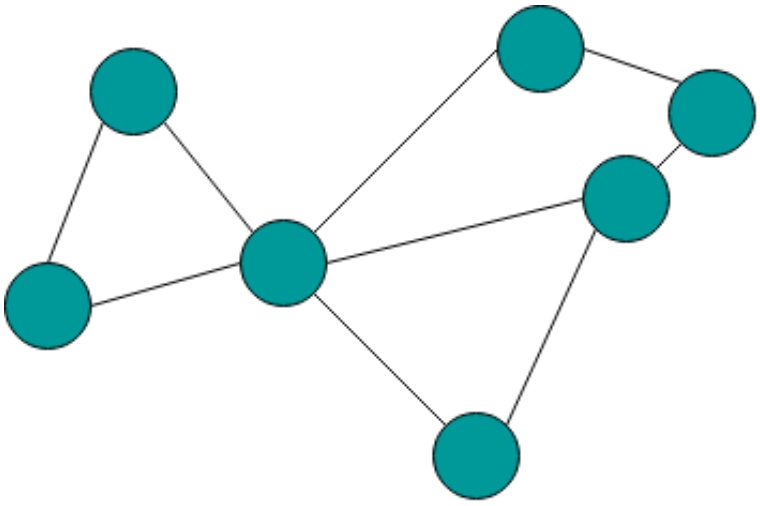


Figure 7.8: Graph neural networks. The circle are the nodes and the line linking them the interactions

The different architectures can also be merged into new hybrid ones. For instance the Unet used later in this thesis is a modified autoencoder that uses convolutional neural layers.

7.4 Deep learning in Fluid Dynamics

Deep Learning has been used recently in many real life applications and scientific researches. To cite some domains, it has been used in image analysis (51; 132; 29), speech recognition (3; 9; 103), financial predictions (156), supply chain optimization (87; 64), sensors (147; 5), and many others.

It has also been the case for physical models and CFD in every aspect of the domain. It represent a change of paradigm for the domain. For decades, since CFD creation in the seventies, the focus was done on improving using physics and mathematical models for discretization and classical resolution of equations. However, with the emergence of Deep Learning, researches and engineers have started to use the capability of these news algorithms to improve CFD in its accuracy and/or speed. Deep Learning algorithms have been applied to every aspects of CFD like:

- creating new turbulence models (155; 177),
- accelerating some aspects of CFD solver by replacing classical algorithm by deep learning models (33),
- accelerating CFD by replacing it by neural networks to converge to the final state (46),

- accelerating CFD by replacing it by neural networks to solve every steps iteration by iteration (80; 138; 150),
- reducing the complexity of a model while keeping accuracy (66; 106),
- reconstructing missing data (24),
- or predicting variables of interest such as Reynold Stress (175).

The aim of this part of the thesis is to be able to assess pollutant concentration in urban areas in real time. It is known that CFD is not capable of doing it without improvement. For instance, to predict pollution dispersion in an area of $1km^2$ it requires with our CFD model around 1 week of computation on 96 CPUs for only one wind direction. Therefore, CFD models can not be used over large areas in real time with acceptable cost. The main issue is hence speed. The Deep Learning models will be used to improve the speed of the CFD model. The approach elected in this thesis is to replace the CFD model by a neural network. The CFD model will be used to create examples of pollution dispersion to train a deep learning model to reproduce it.

Two strategies exist to replace CFD by neural networks:

- Fully data driven approach. In this approach the neural network has no physically induced equations. It will find patterns only through the training phase and the examples shown, as in (46).
- More recently physic informed neural networks have been proposed. In this approach, physical constraints such as mass conservation or momentum are added to the neural network to force it to respect physical principles such as in (115).

The future of new approaches using neural networks with CFD in the long term is probably the physical informed neural network since it adds physical constraints to the network. Nevertheless, these approaches are very recent and much work is still needed to improve them and have them ready for real life applications. In this thesis, the aim was to have an operational model by the end of the thesis, that could be used on real neighborhoods in real cities. Therefore the fully data-driven approach was used since its practicability was already demonstrated in other fields. Moreover, more arguments can be given, linked to the pollution dispersion in open areas:

- The physic informed approach requires to solve the flow field which is an harder task than to only solve the pollutant dispersion. Thus, Using a physical informed approach will add steps that will complicate the approach since our variable of concern is the pollutant dispersion field.

- Creating large examples of the size of a neighborhood is a tedious tasks, creating smaller neighborhood in numbers is much easier. But by doing so, by splitting large neighborhoods into smaller one, there is no clear boundaries in which the equations could be respected.
- To respect the equations, the architecture would have been needed to be in 3D right at the beginning of the thesis, which is a huge constraint in architecture complexity and computing power.
- This approach allows to use previous simulations of neighborhoods made by Air&D as examples for the AI.

A particularly interesting topic for the pollution dispersion and computational fluid dynamics is image analysis and classification. Indeed, image analysis basically consist to find pattern using spatial information. To this extent, a popular tool to deal with spatial information are convolutional encoder/decoder neural networks. This approach was used by (46) to predict flow in 2D and 3D settings. In this thesis, architectures that proved their performances in other fields was used, especially segmentation of images and applied it to pollutant dispersion.

When looking at images of pollutant dispersion and segmentation issues, the proximity between these two problems can be seen as shown below 7.9:

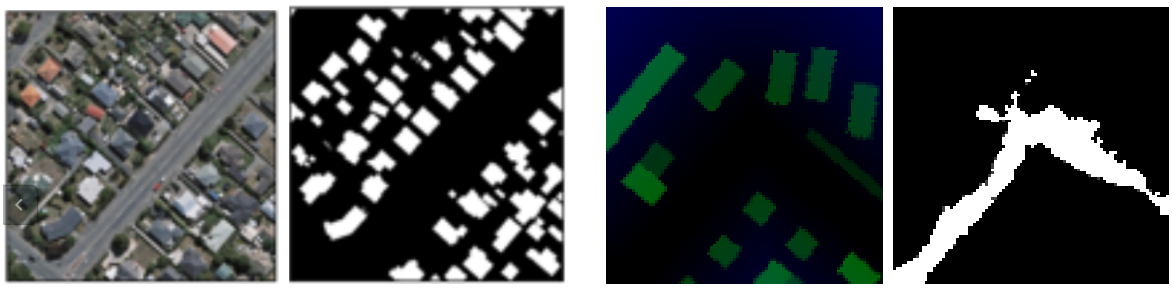


Figure 7.9: Left : Image of satellite segmentation done by (180) on an RGB satellite image and segmentation in grayscale of buildings left. Right : building layout in green and distance from road in blue and pollution dispersion in grayscale at a threshold of C/Cmax = 0.5

This strategy of using segmentation model have also been used for CFD by other researchers (150; 124) that used the classic Unet architecture to determine flow fields.

7.5 Conclusion

Artificial Intelligence has become a reality only in recent times and has known a rapid evolution in the recent years. Several types of machine learning exist, supervised, unsupervised

and reinforcement. In this thesis, a supervised approach will be developed.

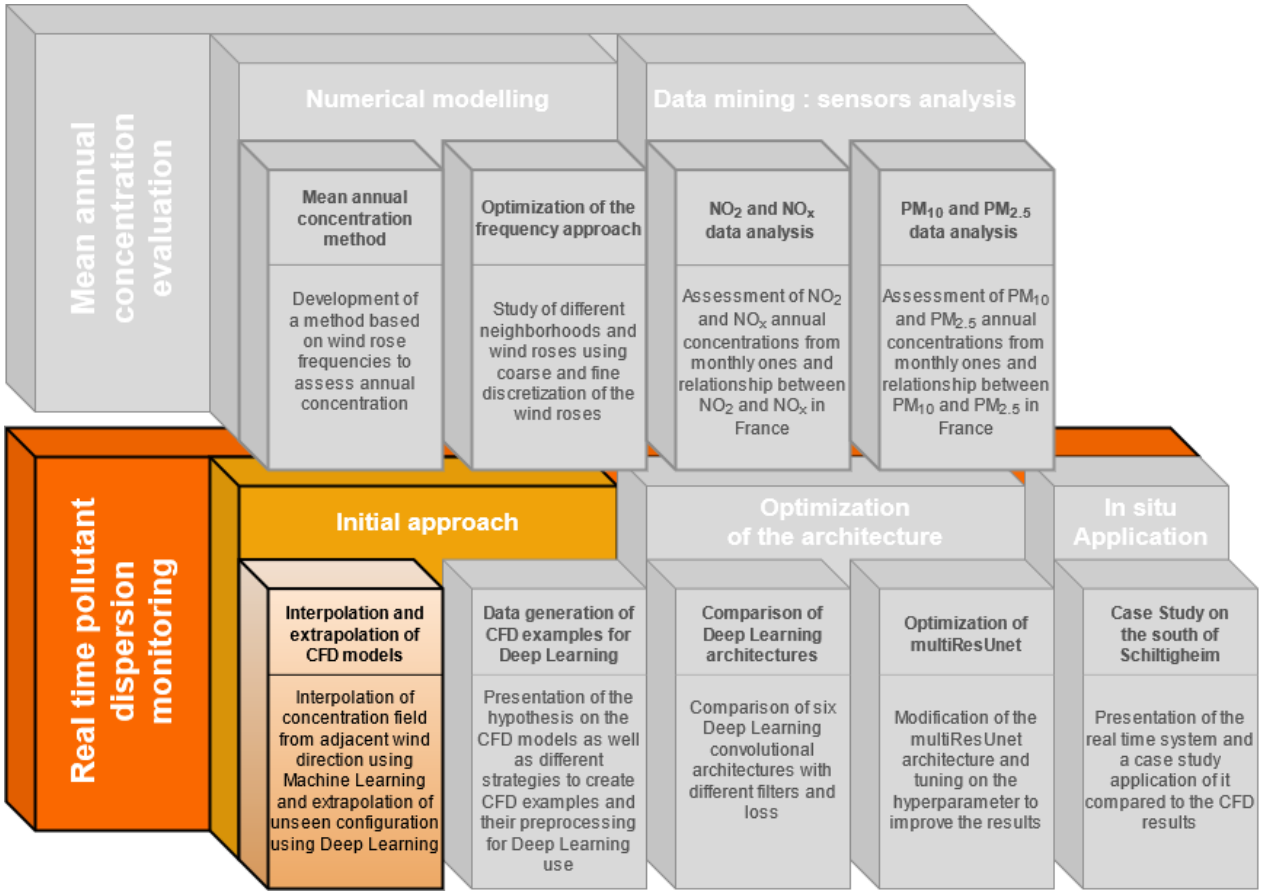
Machine learning has been applied successfully in many domains. CFD is no exception but the applications in this domain are still nascent compared to other topics much more explored, such as language translation or image analysis.

When using neural networks to deal with CFD, two main approaches are possible, either fully data driven or physical informed neural networks. Physical informed neural networks are most likely the future of deep learning associated with CFD, but the data driven approach was elected for its more mature state and various reasons linked to the dispersion of pollutant in open urban areas.

The domain of image analysis, and particularly segmentation, seems to have similarity with the pollution dispersion issue and thus, methods from these fields will be used and adapted to treat and reproduce CFD pollutant dispersion maps.

Chapter 8

First approach of Deep Learning modelling to assess pollutant dispersion



This chapter has been submitted in the journal *Computers, Environment and Urban Systems* under the title "Assessment of capability of deep learning to predict air pollution dispersion" (56).

In the previous Chapter 7, an overview of AI, particularly in CFD context, has been provided. It was shown that machine learning has been used extensively with CFD in the recent years and among the uses, to improve the speed of prediction of CFD results. Nonetheless, this theoretical concept now needs to be given flesh and its performance evaluated. To do that, a first work must be done on two main issues. On one hand, to create a way to provide CFD input and results to an AI. On the other hand, to assess the potential capability of Deep Learning to predict pollutant dispersion from CFD. It must be evaluated on its accuracy, speed and reliability.

8.1 Introduction

To achieve sustainable cities with adequate air quality, reliable and affordable methods need to be developed. Various methods exist and can have wide ranges with model spanning from the scale of continents or countries (also referenced as macroscale), to regions or cities (mesoscale), up to urban blocks (microscale) (154). Microscale modelling is particularly useful in urban context, in which many sources such as road traffic or chimney gases can locally and heavily contribute to air pollution. Among these models, Gaussian plume models are widely use since they are fast to compute. However, they can show weaknesses since they do not consider the impact of buildings on wind (71; 12) while urban morphology deeply influences air pollutant dispersion (74). Sensors can also be used to evaluate concentration from local sources and background concentration using statistical tools such as Hidden Markov Models (43). Nonetheless, they provide very local information and need a long time period of measurement to give useful values, especially for yearly standards, even if the time constraint can be improved with recent methodologies (60). Finally, Computational Fluid Dynamics (CFD) models taking into account buildings and turbulence of air flow can be used (119), but are very computational time-consuming and limited in their uses.

In recent years, artificial intelligence (AI) and specifically deep learning methods have known a tremendous development in many domains. Machine learning has been used in ever increasing fields such as medical imagery (81), electricity consumption forecasting (177), physics (110) or 3D cloud points classification (108). These advances are starting to make their way in the domain of air pollution estimation (31), urban systems (45) or CFD (116; 21).

To study air pollution of an area, wind roses and their respective speed frequencies must be considered to assess annual pollution (120). However, computing CFD for a whole wind rose is very expensive and can lead to errors. Thus, methods are needed to interpolate the

missing wind directions. Hence, the aim of this article is to develop and compare approaches to interpolate pollutant dispersion on urban blocks from existing CFD computation and even extrapolate unseen geometries.

8.2 Material and methods

8.2.1 Numerical model

To create examples for the Deep Learning model, OpenFoam v5 is used to perform numerical simulations. Air flow is considered incompressible due to the low wind speeds. The Reynolds-Averaged Navier-Stokes (RANS) approach using the renormalisation group (RNG) k-ε turbulence model has been applied to solve the fluid mechanics equations for the wind and transport equation for the pollutant dispersion. This solver was used and validated in a previous study (122).

The upper and lateral boundaries are symmetry condition ($dU/dx = cste$), and the outlet is a free stream condition. The buildings and ground are modelled by a smooth wall with a no-slip condition ($U = 0$). A logarithmic profile was used for the inlet to model a neutral atmospheric case following the guidelines of (125) and is calculated as follows:

$$U = \frac{u_*}{\kappa_{k-\epsilon}} ln \frac{z_0 + z}{z_0} \tag{8.1}$$

$$\epsilon = \frac{u_*^2}{\sqrt{C_\mu}} \tag{8.2}$$

$$k = \frac{u_*^3}{\kappa_{k-\epsilon} z} \tag{8.3}$$

where, U is the inlet speed $[m.s^{-1}]$, ε is the turbulent dissipation rate $[kg.m^{-1}.s^{-4}]$, k is the turbulent kinetic energy $[kg.m^{-1}.s^{-3}]$, u_* is the shear velocity [m/s], $\kappa_{k-\epsilon}$ is the von Kármán constant [-], z_0 is the roughness length [m] and z is the altitude [m].

To construct the meshes and environment for the simulations, the guidelines suggested by (42) were followed. For each urban block, noting H the height of highest building, a distance of $5 \times H$ was used for the lateral boundaries, the distance to the inlet/outlet from the closest building and for the top boundary to the ground. A mesh sensitivity analysis has been conducted and resulted in a mesh size of $0.5m$ for the closest cells to the building.

The numerical model results used for this study are considered as the ground truth for the machine learning algorithms that aim at reproducing the result of the CFD. The CFD model used to create the examples is the same than the one used in (119; 121) that can reach error less than 10% compared with experimental measures.

8.2.2 Data

The dataset is composed of 11 different geometries (in grey in Figure 8.1) corresponding to urban blocks of Strasbourg city (France). Each block includes a road as a source of pollutant (in red in Figure 8.1). The pollution values are collected on an area of 150x150m^2 at the centre of the block at a height of 1.50m (blue points in Figure 8.1). Each simulation is composed of approximately 700,000 cells. For 5 geometries, simulations of pollutant dispersion have been calculated every 20 degrees, for 2 others every 40 degrees, and for the 4 last ones every 60 degrees. This sums up to 136 different simulations. From this dataset, three urban blocks with 18 wind directions available were used to compute and compare the interpolation methods. For the training of the U-net model, all blocks but one with 18 directions were used for training and the last remaining to evaluate the U-net prediction.

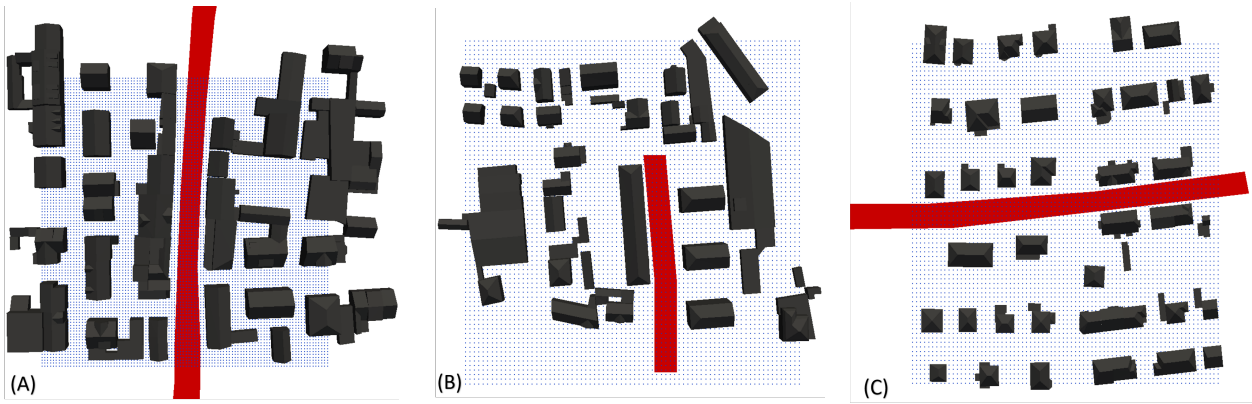


Figure 8.1: Geometry of the three urban blocks with buildings in grey, pollutant source in red and tracking points in blue.

The values of the simulations computed with CFD are acquired after convergence of the model, when residuals are below 1×10^{-5} for pressure, 1×10^{-6} for speed and 1×10^{-7} for pollutant in the form of a table "x coordinate, y coordinate, concentration". The dataset is divided into two categories, training and testing. The training set has been used to train the algorithm and interpolate. The test set has been used to evaluate the predictive capabilities of the algorithms to estimate unseen pollution dispersion direction.

8.2.3 Interpolation methods

To compute the unknown pollutant dispersion maps for the missing wind directions of a urban block from the known ones, three interpolation methods have been tested. Methods are described below.

Linear interpolation The linear approach consists in interpolating an unknown simulation from the two closest known ones. An unknown field value direction named X between two known field value directions A and B as in Figure 8.2 can be determined as following:

$$Field_X = \sum_{i=1}^{N} \frac{\frac{1}{\alpha_i^p}}{\sum_{j=1}^{N} \frac{1}{\alpha_j^p}} Field_i \tag{8.4}$$

where N is the number of closest directions taken into account, α_i is the angle between direction X and direction i, $Field_X$ is the field values of the direction X, $Field_i$ is the field values of the direction i and p is a real number that allow to give more or less impact on the closest direction.

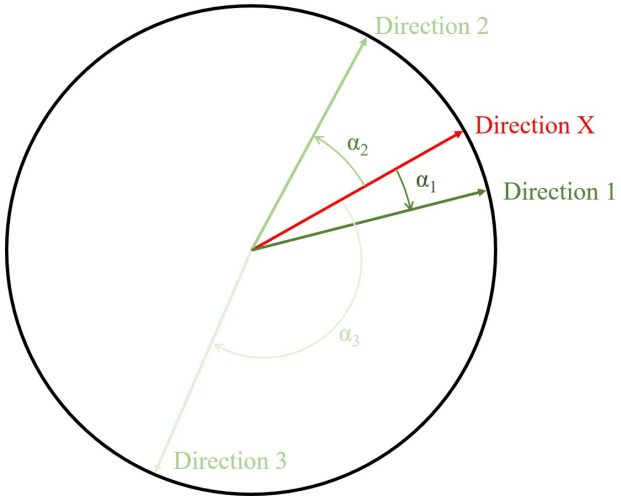


Figure 8.2: Interpolation of an unknown field at direction X by three surrounding known fields with their respective angles.

Criteria	Tested values	Best value
p	0.5, 1, 2	1 or 2
N	2, all available directions	2

Table 8.1: Tested criteria for the linear interpolation.

Random Forest Random forest (20) is a popular method of machine learning used for prediction and interpolation. It has been used in a variety of fields as broad as to measure urban poverty (102), predicting air pollution (171) or prediction of water consumption (30).

The method consists in averaging several decision trees that do not perform great on their own, but following the philosophy of Condorcet's jury theorem that averaging several mediocre opinions is better than a unique truth-worthy one. The inputs that are given to the random regressor forest are the localisation of the point, and the sinus and cosinus of the wind direction : "x coordinate, y coordinate, cosinus(direction), sinus(direction)", the expected output is the concentration value. The loss used was Mean Squared Error (mse) for its speed. Other parameters were chosen after a 10-cross-validation on the training dataset (see Table 8.2).

Criteria	**Tested values**	**Best value**
n estimators	10, 50, 100, 200, 500	200
max depth	8, 10, 14, 18, None	14
min samples leaf	2, 3, 5, 7	5
min samples split	2, 3, 5, 8	2

Table 8.2: Tested criteria for the random forest approach.

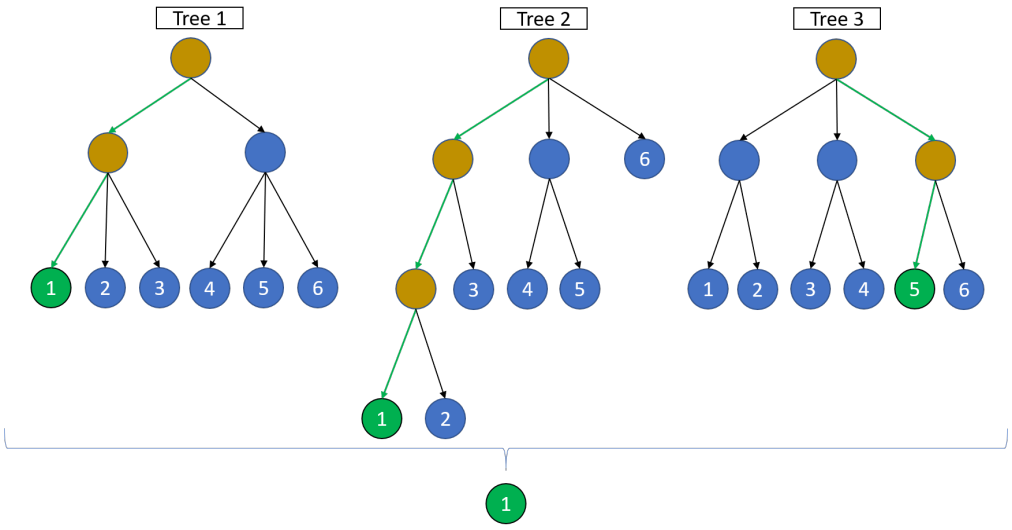
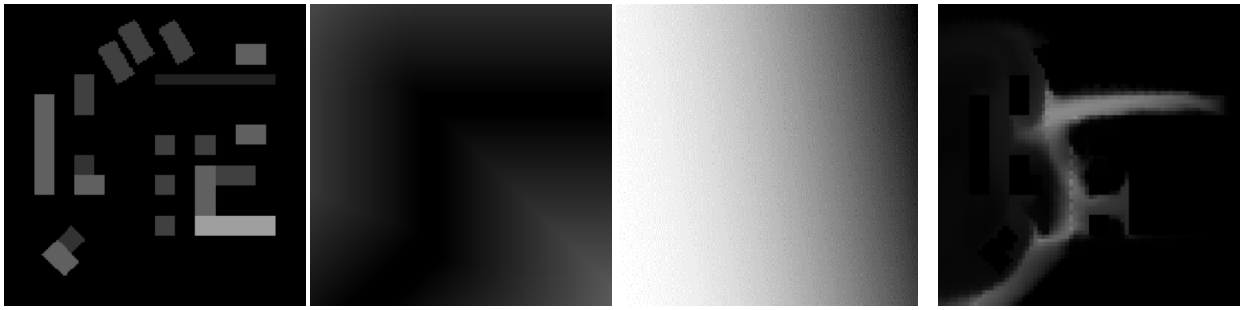


Figure 8.3: Example of a random forest made of 3 decision trees, with each tree making a prediction (green arrow and dot) and the output decided by voting from the different outcome of each tree.

U-net U-net is a deep neural network architecture created for biomedical images segmentation (132) that has also been used in many other fields such as remote sensing (176) or

image generation (84). U-net is an encoder/decoder convolutional neural network which is specialised in dealing with spatial information. It has the ability to understand the context of an image while encoding and still locate precisely spatial information thanks to its skip connections while decoding. The U-net details of the architecture used in this paper is presented in Figure 8.5. The implementation was done using python library Keras and Tensorflow. There are 4 encoding layers, 1 bottleneck and 4 decoding layers. Data have to be standardised between 0 and 1.

The input of U-net is a map of building heights (with a scale of 40m), a map representing for each pixel its distance from the pollutant source (with a scale of 250m) and a map representing the wind direction (Figure 8.4 (a-c)). The output is the pollution dispersion map, scaled for each block according to the road length and its max field pollution value (Figure 8.4 (d)). When several blocks are mixed, the average of the max field pollution value is considered. This choice has been made because some simulations have a high max value that occurs rarely and reduces the prediction capabilities.



(a) Building heights (b) Dist. to pollutant (c) Wind dir. (d) Dispersion map

Figure 8.4: Example of inputs and output of the U-net architecture. (a-c) Inputs of the network. (d) Output for a wind direction of 80°N.

Criteria	**Tested values**	**Best value**
Number of layers	4, 5	4
Minimum filters	2, 4, 8	4
Optimizer	adam, nadam, adagrad, sgd, rsmProp	adam
loss function	poisson, binary crossentropy, mse	binary crossentropy

Table 8.3: Tested criteria for U-net.

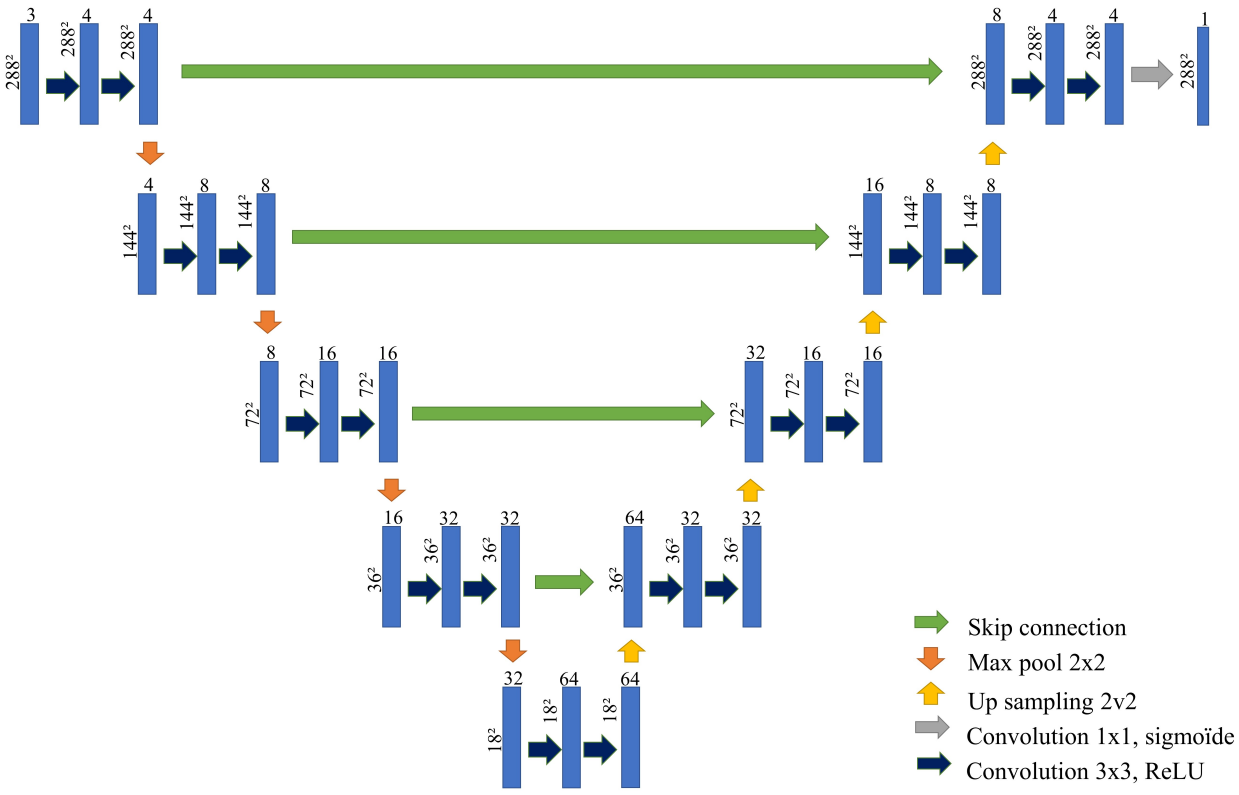


Figure 8.5: U-net architecture details.

8.2.4 Metrics

To evaluate and compare results between predictions made by the different models several metrics were used.

Metrics for air quality models To evaluate air quality models several performance measures exist, such as fractional bias (FB), geometric mean bias (MG), normalised mean squared error (NMSE), geometric variance (VG), the correlation coefficient (R) and the fraction of predictions within a factor of two of observations (FAC2) that can be found in (27) defined as:

$$FB = \frac{(\overline{C_{ref}} - \overline{C_{pred}})}{0.5(\overline{C_{pred}} + \overline{C_{ref}})}, \tag{8.5}$$

$$MG = exp[(\overline{lnC_{ref}} - \overline{lnC_{pred}})], \tag{8.6}$$

$$NMSE = \frac{\overline{(C_{ref} - C_{pred})^2}}{C_{pred}C_{ref}}, \tag{8.7}$$

$$VG = exp[\overline{(lnC_{ref} - lnC_{pred})^2}], \tag{8.8}$$

$$R = \frac{\overline{(C_{ref} - \overline{C_{ref}})(C_{pred} - \overline{C_{pred}})}}{\sigma_{C_{pred}}\sigma_{C_{ref}}}, \tag{8.9}$$

$$FAC2 = \text{fraction of data that satisfy } 0.5 < \frac{C_{pred}}{C_{ref}} < 2, \tag{8.10}$$

With C_{pred} the model prediction concentration and C_{ref} the reference concentration.

To assess air quality models, Chang *et al.* (27) proposed ranges of values for some of these parameters. However, these performance measures are more adapted to compare some measuring points with a model than a model versus another one. Indeed, the authors precise that it is harder to reach these values when the data are paired in space and/or time, which is the case here. A model starts to be good when three types of parameters are within a certain range at the same time when compared to the reference (usually sensors measures, here the CFD model):

- FAC2 > 0.5,
- NSME < 1.5 or VG < 4,
- $|\text{FB}| < 0.3$ or $0.7 <$ MG < 1.3.

Some of these measures need a threshold value such as VG and MG because of their logarithm nature. The threshold used in this study is 0.

Metrics for images Evaluating differences between two images is an ill-defined problem and results can be counter intuitive. However, for this study, three metrics were retained to compare two images, relative mean absolute error, structural similarity index and volumetric index that are presented below.

Relative mean absolute error The first ratio used is the relative mean absolute error. It consists in comparing the luminosity pixel wise and is defined as follows:

$$mae_{rel} = \frac{\sum_{i=1}^{N} |pixel_i^{true} - pixel_i^{pred}|}{\sum_{i=1}^{N} pixel_i^{true}} \tag{8.11}$$

where N is the number of pixels and pixel_i^{true} and pixel_i^{pred} are respectively the i-th pixel of the true and predicted image.

The mae_{rel} value is between 0 and Infinity. Its value is 0 for identical images, 1 when the mean error is equal to the mean of the predicted value. This ratio has to be minimised.

Structural similarity Structural similarity index (162) was originally designed to measure the visual quality between a compressed image compared to the original one. It takes into account the structure of the image and is computed between two image windows A and B as follows:

$$SSIM(A,B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c2)}{(\mu_A^2 + \mu_B^2 + c1)(\sigma_A^2 + \sigma_B^2 + c2)} \tag{8.12}$$

$$c_1 = (k_1 L)^2 \quad c_2 = (k_2 L)^2 \tag{8.13}$$

where μ_A and μ_B are the respective average of A and B, σ_A^2 and σ_B^2 are the respective variance of A and B, σ_{AB} is the covariance of A and B, L is the dynamic range of the pixel values and k_1 and k_2 are two constants respectively 0.01 and 0.03 (by default).

$SSIM$ values are between 0 and 1. Its value is 1 for two similar images 0 for dissimilar. This index has to be maximised.

Volumetric index **Note:** *I kept the notation Vol index as it is in the original submitted article that dates back to a year and a half from this thesis, but this name is later on changed to J_{3D} in the thesis.*

The Jaccard Index is a popular method used to evaluate the similarity between two binary images consisting in comparing the shared area of two images over their union area. However, the images used here are not binary but grey-scale images. Following the same idea, we propose a volumetric index, comparing the volume shared by the two images, the grey-scale level being the third dimension.

$$vol_{index} = \frac{V_{pred} \bigcap V_{true}}{V_{pred} \bigcup V_{true}} \simeq \frac{\sum_{i=1}^{N} min(pixel_i^{true}, pixel_i^{pred})}{\sum_{i=1}^{N} max(pixel_i^{true}, pixel_i^{pred})} \tag{8.14}$$

where N is the number of pixels, $pixel_i^{true}$ is the value of the i^{th} pixel of the true image and $pixel_i^{true}$ is the value of the i^{th} pixel in the predicted image.

vol_{index} values are between 0 and 1. Its value is 1 for two identical images and 0 when the images have no common pixels. This index has to be maximised.

A visual example of intersection and union volumes on two images is given in Figure 8.6 as well as the three chosen metrics.

8.3 Results

8.3.1 Interpolation approaches comparison

In our study, interpolation corresponds to generate a simulation for an unknown wind direction from known simulations calculated for different wind directions. All the simulations are

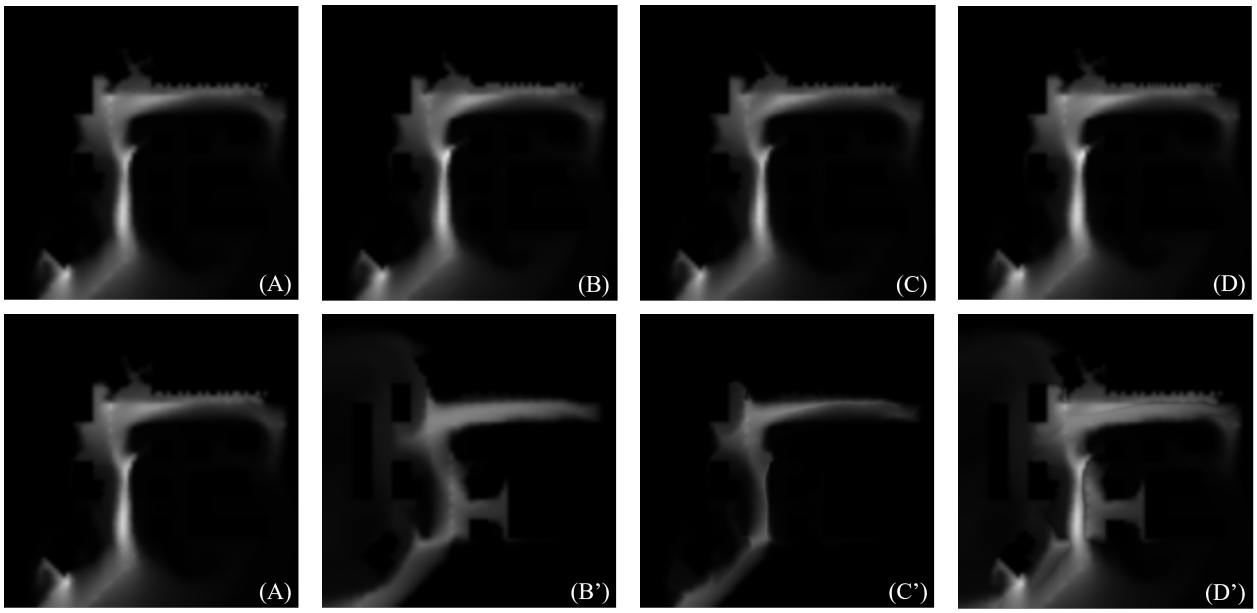


Figure 8.6: Comparison of three pollution dispersion images (A), (B) and (B') respectively with a wind direction of 20°N, 21°N and 80°N. (C) represents the volume image of the intersection of (A) and (B). (D) is the union volume image of (A) and (B).For (A) and (B') the metric gives respectively : $vol_{index} = 0.89$ and 0.34, $SSIM = 0.94$ and 0.32, $mae_{rel} = 0.12$ and 0.94, $FB = 0.02$ and 0.11, $VG = 2.6$ and 13.4, $NMSE = 0.13$ and 3.65 and $FAC2 = 0.95$ and 0.38.

performed only on one unique urban block. To evaluate and compare the three interpolation methods proposed (linear interpolation, Random Forest and U-net), we set up an experiment with three different sets:

- 9 directions out of 18 for interpolation with a step of 40 degrees between two interpolation directions;
- 6 directions with a step of 60 degrees between consecutive directions;
- 3 directions with a step of 120 degrees for the three blocks.

Results are shown on Figure 8.7.

Linear interpolation outperforms U-net and Random Forest models on the three metrics. U-net and Random Forest give approximately the same results which is coherent as they are both machine learning algorithms. As they are trained with very few examples, there results remain worse than linear interpolation.

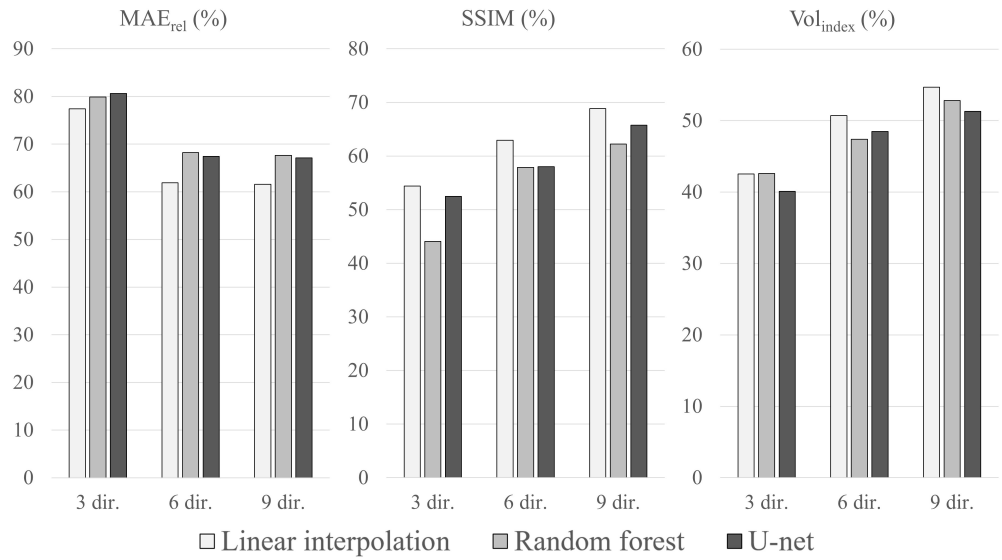


Figure 8.7: Interpolation results for the three proposed image metrics, for the different number of simulation used for interpolation and interpolation algorithms.

8.3.2 Deep Learning approach: U-net

We have shown that interpolation can be used when having already some calculated simulations for many wind directions on an urban block. But we aim to go further and study if a machine learning algorithm could be able to generate a simulation on any block without having any simulation calculated on it.

For this, we propose to train a U-net on many simulations calculated on different urban blocks (geometries) and different wind directions. U-net could then be used not to interpolate partially known geometries but extrapolate unseen geometries based on the previous geometries and wind directions it would have seen. The objective is that the trained model could determine the pollution dispersion faster than computational fluid dynamics for a trade-off of an error.

U-net was trained on all the geometries and directions presented in the dataset excluding the 4th, that was used to evaluate its performance. The pollution dispersion was predicted for the 18 directions of the 4th dataset. Results for 4 directions are presented on Figure 8.8: 80°N that performed best, 280°N that performed worst and two directions where the prediction is close to the mean value of the 18 directions (200°N and 360°N). All the numerical values on the three metrics are given in Table 8.4.

The mean result of the U-net predictions on the different metrics for the 4th block are about slightly above the ones from the linear interpolation with three directions seen on Figure 8.7. However, the geometry (buildings disposition) of this block was not used during training. Thus, without any simulation on this block, this method managed to get a result as

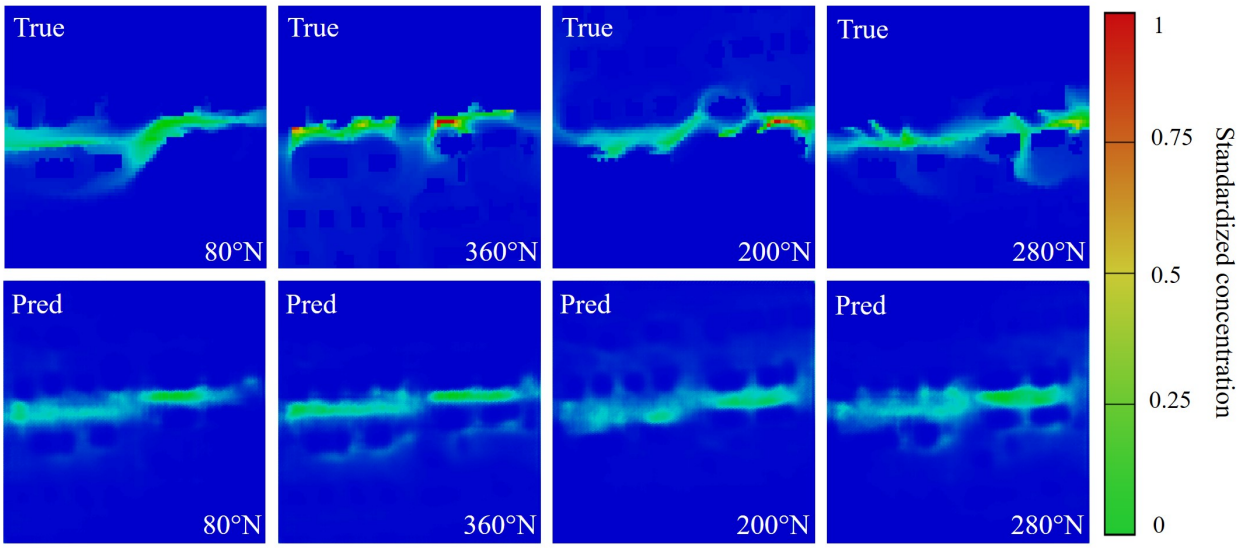


Figure 8.8: Examples of U-net predictions on urban block 4 (second row set) compared to ground truth (first row).

Direction	$mae_{rel}(\%)$	$vol_{index}(\%)$	$SSIM(\%)$	$\|FB\|(\%)$	VG	$NMSE$	$FAC2(\%)$
80°N	62.4	51.1	60.0	15	2.79	3.7	62
200°N	74.4	47.3	50.8	3	2.58	3.8	49
280°N	84.8	43.2	48.8	9.5	2.66	4.2	51
360°N	71.1	45.1	52.7	15	2.79	4.5	56
Mean on all 18 directions	**73.7**	**45.7**	**52.7**	**10**	**2.69**	**4.47**	**54**
std_{rel} on all 18 directions	8.6	7.1	12.5	64.6	3.7	22.2	10.4

Table 8.4: Evaluation of the predictions made by U-net on three metrics: mae_{rel}, vol_{index}, $SSIM$, $|FB|$, VG, $NMSE$ and $FAC2$.

good as if 3 simulations would have been made and used to interpolate. The model manages to capture well that the pollution is strong on the road, that the pollution is equal to 0 on the buildings and manage to some extent to have a dispersion following the wind direction.

The air quality metrics are all within the acceptable range but the NSME according to the values presented in section 2. The NSME poor results can be explained by the fact that there are several order of magnitude in the model results whereas VG is more adapted to measure the random scatter in this kind of case as discussed by Chang *et al.* (27).

In terms of computation time, it can be difficult to compare both methods, since CFD does its computation in 3D and here U-net performs 2D ones. However, U-net does its prediction in 0.5 second for the 18 pollutant dispersion maps on NVidia Titan V GPU whereas CFD would have required around ten hours on ten CPU processors. Deep learning seems to be a

relevant approach to estimate in real time a pollutant dispersion map on an urban block.

8.4 Conclusion

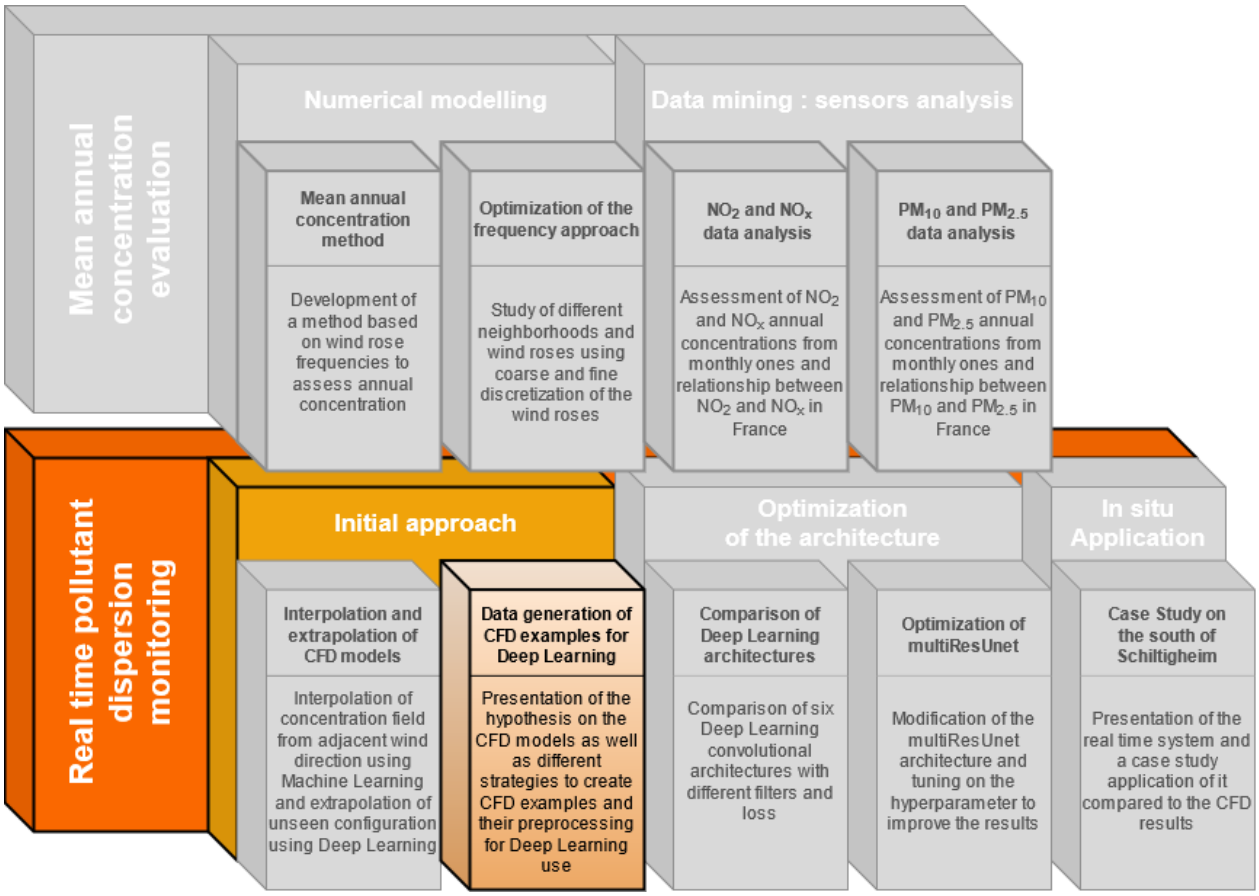
For a unique urban block for which some simulations have already been calculated, complex approaches based on machine learning do not provide any benefits compared to a classical linear interpolation. Indeed, these algorithms need to be trained with a lot of examples, to obtain relevant models. Thus, the linear approach to interpolate results over a single geometry is the easiest to implement and shows the best results. The best way to interpolate according to this study is to use the two closest directions with a power on the distance equal to two or one.

To use the full potential of deep learning neural networks, such as U-net, a global approach need to be used. This consists in using not only one geometry but several of them, so that the algorithm can learn on various cases and infer new ones. We have shown that a U-net managed to get results on an unseen geometry equivalent to the one obtained from the linear interpolation knowing 3 simulations.

The dataset for training U-net was still relatively small (about one hundred examples). We expect to be able to significantly improve the result by adding more examples (geometries and directions) and performing data augmentation. Other known architectures or loss function will be tested to try to improve the results for a given dataset. At the end, it should be possible to determine pollutant dispersion maps in real time approaching CFD quality, or to use the solution found by the deep neural network to initialise CFD simulations and reach faster convergence.

Chapter 9

Data generation to create examples for AI



9.1 Introduction

To produce more examples in less time, the first approach adopted was to accelerate the computation of a simulation for one neighborhood by reducing the number of required directions. The idea was to compute a small amount of directions and to interpolate any other directions from this small dataset of computed CFD simulations using machine learning methods. Nevertheless, as seen previously, results obtained with machine learning were not much better than a standard linear interpolation. The main issue is that to exploit to its best machine learning, a lot of examples are needed. A single neighborhoud does not provide sufficient examples. Hence the approach needs to be more general, to have more examples to train the machine learning methods.

To treat spatial information, encoder/decoder convolutional neural networks have proved to be formidable tools. The U-Net among them is a classical architecture and it was demonstrated that when training with various neighborhoods, amounting to a hundred of examples, U-Net managed to produce a dispersion map for a neighborhood that was not used during training. Even if the error was relatively high, it is still encouraging. The poor results are most likely due to the lack of data that is a must for Deep Learning methods. It becomes necessary at this point to be able to generate many examples to exploit at best Deep Learning approaches.

The issue here is that creating examples from the CFD is expensive computation wise. It is therefore necessary to exploit at best the CFD results when creating examples. In the following chapter, the CFD hypothesis and limitation will be explored. Then, rules and methods will be develop to create efficiently CFD examples for the Deep Learning architectures.

9.2 Hypothesis on the CFD used to train the model

9.2.1 Model assumption and equations

The dispersion of a pollutant is influenced by a large number of micro-scale factors. The CFD model is able to consider a certain number of them as it was seen in the previous chapter. However, each additional parameter makes the model more complex and the computation harder to perform and reproduce. We must therefore arbitrate the most preponderant parameters for the training. The other parameters can still be included in a future version of the workflow.

During a previous PhD in ICube laboratory, several models have been developed by Nicolas Reiminger. Each model is able to consider many phenomena:

- model 1: Couple model of kε model for aeraulic and advection diffusion equation for

pollutant dispersion,

- model 2: Neutral condition at inlet,

- model 3: Photochemical Equilibrum for NOx/NO_2 pollutant,

- model 4: Vegetation and Deposition model,

- model 5: Thermic.

As a first step, the second model, which consider the atmosphere stability to be neutral, will be used. This model solves three sets of equations for a given time step: the Navier-Stokes equations (flow equations), the turbulence equations and the turbulent advection diffusion equation (pollutant dispersion equation), described below.

The calculation of velocities is done by solving the Navier-Stokes equations 9.1 and 9.2 which are the classical equation to determine how a fluid behaves. The use of these equations allows to consider both the conservation of matter, the quantity of motion, the turbulence present in the atmosphere and also the effects of pressure. In addition to these different terms taken into account, the use of CFD also allows to get rid of the assumption of steady state which can be false because of turbulence and the presence of buildings.

$$\frac{\partial \bar{\bar{u}}_i}{\partial x_i} = 0 \tag{9.1}$$

$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial [\bar{u}_i \bar{u}_j]}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \bar{P}}{\partial x_i} + \nu \frac{\partial^2 \bar{u}_i}{\partial x_j \partial x_i} - \frac{\partial \bar{u_i'} \bar{u_j'}}{\partial x_j} \tag{9.2}$$

Turbulence is evaluated on the basis of the standard kϵ turbulence model, a model commonly used in the engineering field for this type of application with open free stream in atmosphere and whose equations are shown below:

$$\bar{u_i'} \bar{u_j'} = \frac{2}{3} k \delta_{ij} - \nu_t (\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i}) \tag{9.3}$$

$$\frac{\partial \bar{k}}{\partial t} + \bar{u}_j \frac{\partial k}{\partial x_j} = \frac{\partial}{\partial x_j}(\frac{\nu_t}{\sigma_k}\frac{\partial k}{\partial x_j}) + \nu_t (\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i})\frac{\partial \bar{u}_i}{\partial x_j} - \varepsilon \tag{9.4}$$

$$\frac{\partial \varepsilon}{\partial t} + \bar{u}_j \frac{\partial \varepsilon}{\partial x_j} = \frac{\partial}{\partial x_j}(\frac{\nu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}) + \frac{\varepsilon}{k}(C_{\varepsilon 1}\nu_t (\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i})\frac{\partial \bar{u}_i}{\partial x_j} - C_{\varepsilon 2}\varepsilon) \tag{9.5}$$

$$\nu_t = C_\mu \frac{k^2}{\epsilon} \tag{9.6}$$

In parallel, the spatial evolution of the concentration C of the pollutant is calculated by the turbulent advection-diffusion equation 9.7. This equation considers the velocity field in

all directions, the diffusion resulting from atmospheric turbulence and also the molecular diffusion parameter intrinsic to the pollutant under study.

$$\frac{\partial C}{\partial t} + \frac{\partial \left(u_i C\right)}{\partial x_i} - \frac{\partial}{\partial x_i}\left[\left(D_m + \frac{\nu_t}{Sc_t}\right)\frac{\partial C}{\partial x_i}\right] = E \tag{9.7}$$

The atmosphere is considered neutral for the chosen model. This condition add new equations to take into account which are provided in (126) and (125). It imposes three equations at the inlet:

$$U = \frac{u_*}{\kappa} ln\left(\frac{z}{z_0}\right) \tag{9.8}$$

$$k = \frac{{u_*}^2}{\sqrt{C_\mu}} \tag{9.9}$$

$$\varepsilon = \frac{{u_*}^3}{\kappa.z} \tag{9.10}$$

With U the speed, k the turbulent kinetic energy and ε the dissipation of the turbulent kinetic energy all three given at the entrance of the domain and z the altitude, z_0 the roughness height, u_* the friction speed and κ and Cμ two constants.

In the case of neutral atmosphere, the concentration is related hyperbolically to the inlet wind speed with the following formula:

$$c = \frac{\alpha_{wind}}{u} \tag{9.11}$$

$$\alpha_{wind} = c_{ref} v_{ref} \tag{9.12}$$

Thus the wind speed can be excluded from the necessary parameters. Indeed, if one couple concentration/wind speed is known the other ones can be easily computed using the above formulas. The concentration still depends on another factor: the emission. The concentration evolves linearly with it. Thus, when the concentration is computed with one emission it can be determined for others emission by a cross product:

$$c = \alpha_{emission} E \tag{9.13}$$

$$\alpha_{emission} = \frac{c}{E} \tag{9.14}$$

There is still a need to compute the emission from the roads. Emissions are calculated based on methods proposed by the European Environment Agency (EEA) in their "EMEP/EEA Air pollutant emission inventory guidebook 2016", Tier 3 method for engine-related NOX, PM10 and PM2.5 emissions (hot and cold emissions); 2017 metropolitan fleet data found in the "OMINEA" databases provided by the Centre Interprofessionnel Technique d'Études de la Pollution Atmosphérique (share of different vehicle types, fuels and EURO

standards in France). This method gives the emission depending on the road length, type of vehicles, speed of vehicles and number of vehicles. To sum up, the inlet wind speed and the emission can be excluded from the CFD and AI because they can be easily computed as a post process. The last crucial parameter that has a huge impact on the concentration dispersion is the wind direction. This one needs the CFD model and can not be determined another way.

Thus the concentration field under the previous assumption only depends on the wind direction, building layout and geometry of the emission source.

9.2.2 Boundary conditions

The CFD discretizes a volume in tiny volumetric element on which the Navier Stokes equation are solved depending of its neighbor cells iteratively for each physical variable. However a model can not be limitless, thus a model will have spatial limit, so it is necessary to give rule for the computation on the limit of the domain for the computation to run. These rules are called boundary conditions. To simulate the atmosphere, a box with 6 faces will be used. The 6 faces needs to have mathematical properties that represent the physic beyond the box. To simulate a neutral atmosphere the conditions presented in Table 9.1 have been used.

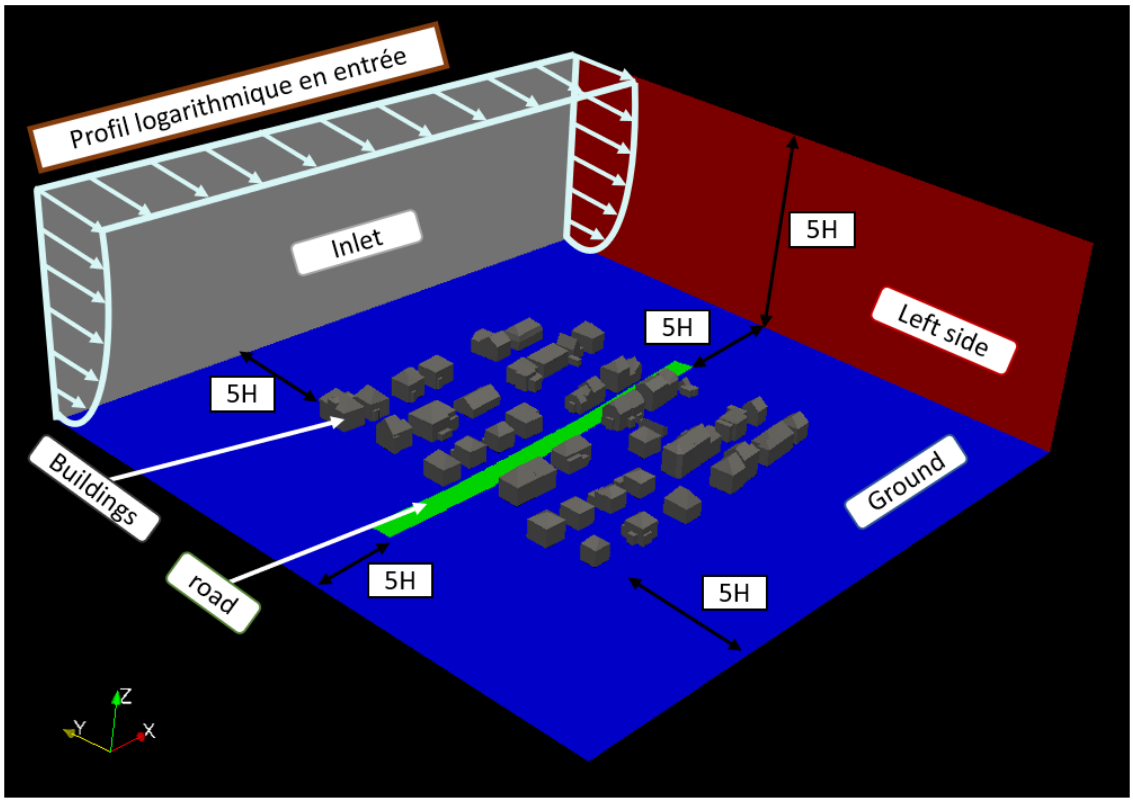


Figure 9.1: Illustration of the boundary conditions.

Boundary	U	P	k	ε	ν_t	c
Roof	symmetry	symmetry	symmetry	symmetry	symmetry	symmetry
Right side	symmetry	symmetry	symmetry	symmetry	symmetry	symmetry
Left side	symmetry	symmetry	symmetry	symmetry	symmetry	symmetry
Ground	fixed value 0	zeroGradient	wallFunction	wallFunction	AtmRoughtWallFunction	zeroGradient
Buildings	fixed value 0	zeroGradient	wallFunction	wallFunction	wallFunction	zeroGradient
Inlet	atmBoundaryLayer	zeroGradient	atmBoundaryLayer	atmBoundaryLayer	calculated	fixed value 0
Outlet	freeStream	freeStreamPressure	inletOutlet	inletOutlet	calculated	zeroGradient

Table 9.1: Boundary conditions with U the velocity field, P the pressure field,k the turbulent kinetic energy, ϵ he rate of dissipation of turbulent kinetic energy, ν_t the turbulent viscosity, c the pollutant concentration field.

For these boundary conditions to give accurate results, another requirement is needed. The box must be large enough to avoid border effects that would add numerical error while it can be easily avoided. In (42), the authors give insight on rules of thumb to respect to avoid them. For the computational domain, the distance between the highest building and the roof is at least five times the height of the highest building (noted H). When several buildings are present, the article does not specify explicitly the space required for the lateral boundaries, so here the distance between the last building and the lateral boundaries, inlet and outlet is set to 5H too.

To conclude, the model for the computation is settled with its hypothesis and assumptions. Now it is necessary to apply this model on building layouts to create examples for the Deep Learning model.

9.3 Creation of the data

9.3.1 Geometries

Now that the hypothesis for the model and boundary conditions are set, it is necessary to create concrete examples. The aim here is to make simulation of pollutant dispersion for different building layouts. To achieve that, three types of geometry will be required. A geometry representing the ground, a geometry representing the source of pollution (the road) and a geometry representing the building layouts as shown on Figure 9.2.

Some more hypothesis have been proposed in order to simplify the model to learn:

- pollution sources: volumetric homogeneous emission,

- buildings: base area is bigger than the roof area,

- ground: flat at z=0,

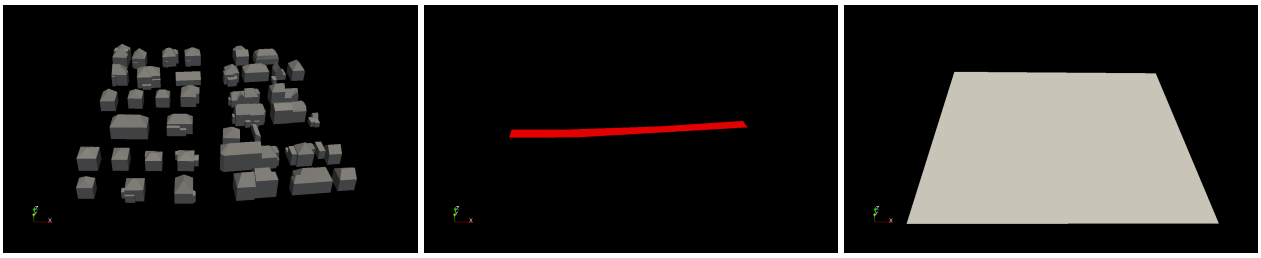


Figure 9.2: From left to right: the building layout, the pollution source (road) and the ground.

- building : starting at z=0,
- road: emitting at z=0.

9.3.2 Strategies to build examples

The first strategy that comes to mind is to associate a road with a building layout and a ground as show on Figure 9.3.

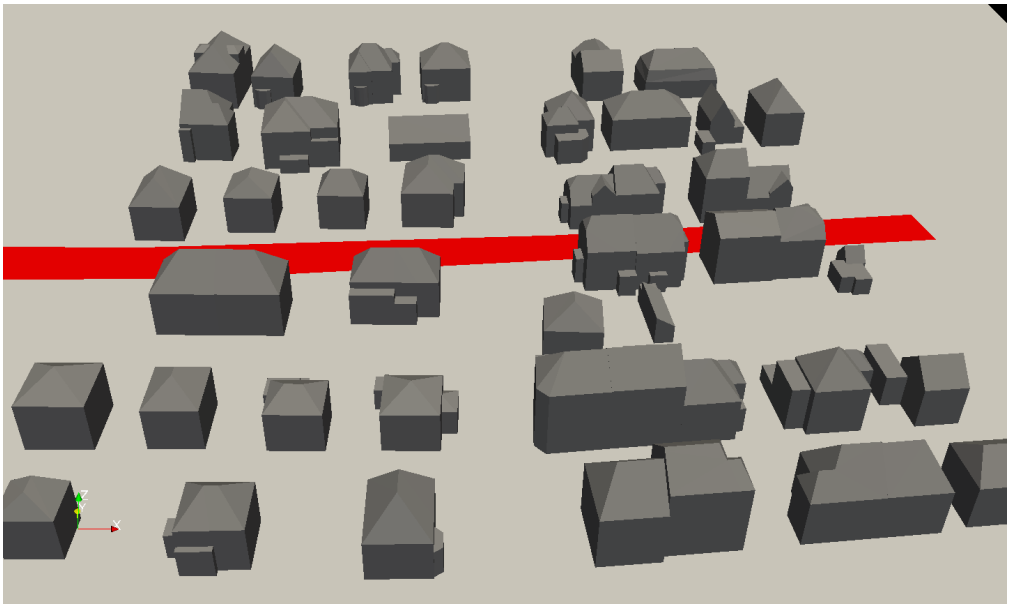


Figure 9.3: Example of a neighbourhood with one roads for one building layout.

Having one wind direction, a simulation of 300x300m^2 on 48 CPU cores will last from several hours to a day. Moreover, it is important to notice that on 300x300m^2, only 150x150m^2 can be covered with buildings if one of the building has a height of 15m because of the 5H rule of thumb explained previously. Thus, this long computation time will only provide one example for the training. This is much too long to be able to produce enough examples to have a sufficient training set.

Nevertheless, most of the time of the computation is used to resolve the Navier-Stokes equation and the aeraulic in the neighbourhood. Indeed the advection/transport equation takes less than 5% of the computation time compared to the flow field resolution. Secondly, the empty spaced required for the computation of 5H for lateral boundaries is constant given a maximum height. This means that relatively, the bigger the area of building is, the smaller the empty space will be. For instance if an area of 300x300m^2 of buildings with a maximum height of 15m, it would be required to have a total simulation of 450x450m^2. So, relatively, the empty space represents 55% of the total space of the simulation, while for a built area of 150x150m^2 with the same maximum height building, it would represent 75% of the total space of the simulation.

Thus, it would be better to make wider areas and use the advantage of having the whole velocity field calculated for several roads at the same time as it is shown on Figure 9.4. In order to do that, it is only necessary to slightly change the Openfoam model code by attributing a different variable to each different emission sources to be able to distinct them.

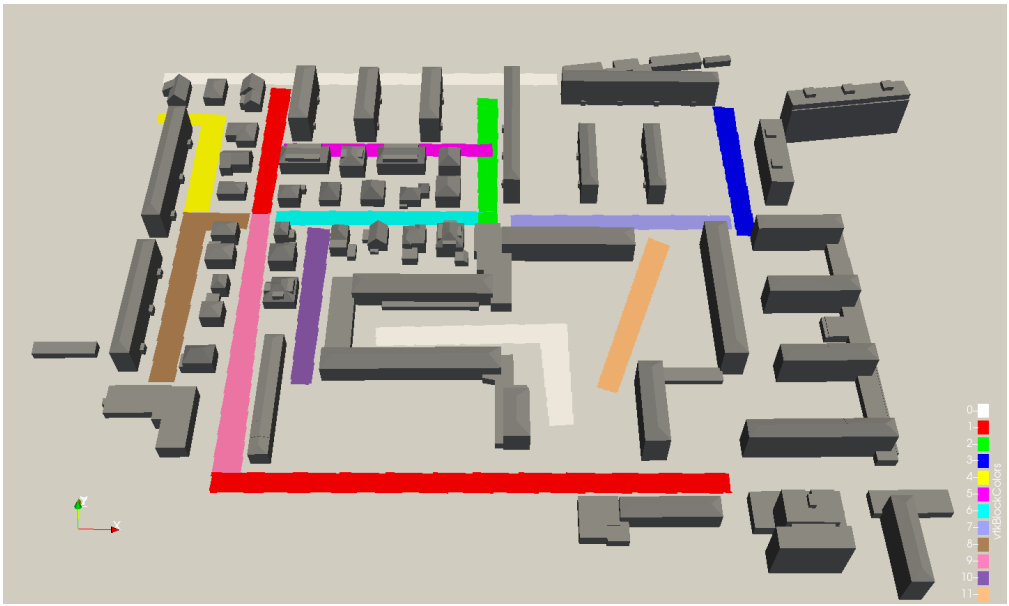


Figure 9.4: Example of a neighbourhood with several roads for one building layout.

With this strategy, one simulation can yield tenth of examples and not just one. In terms of computation cost, the gain is about 10. Moreover, it is possible to go even further. As it can be seen on Figure 9.4, roads can have different lengths. If a road is long enough, it could be used for several examples with a defined spacing, for example every 60m of road another example as shown on the figure below on Figure 9.5. With this strategy of spacing of 60m, the dataset grow by 2-3.

To conclude, this last approach with several roads, wide areas for a simulation and several examples for each road, the required time for making one example is about 20-30 times faster.

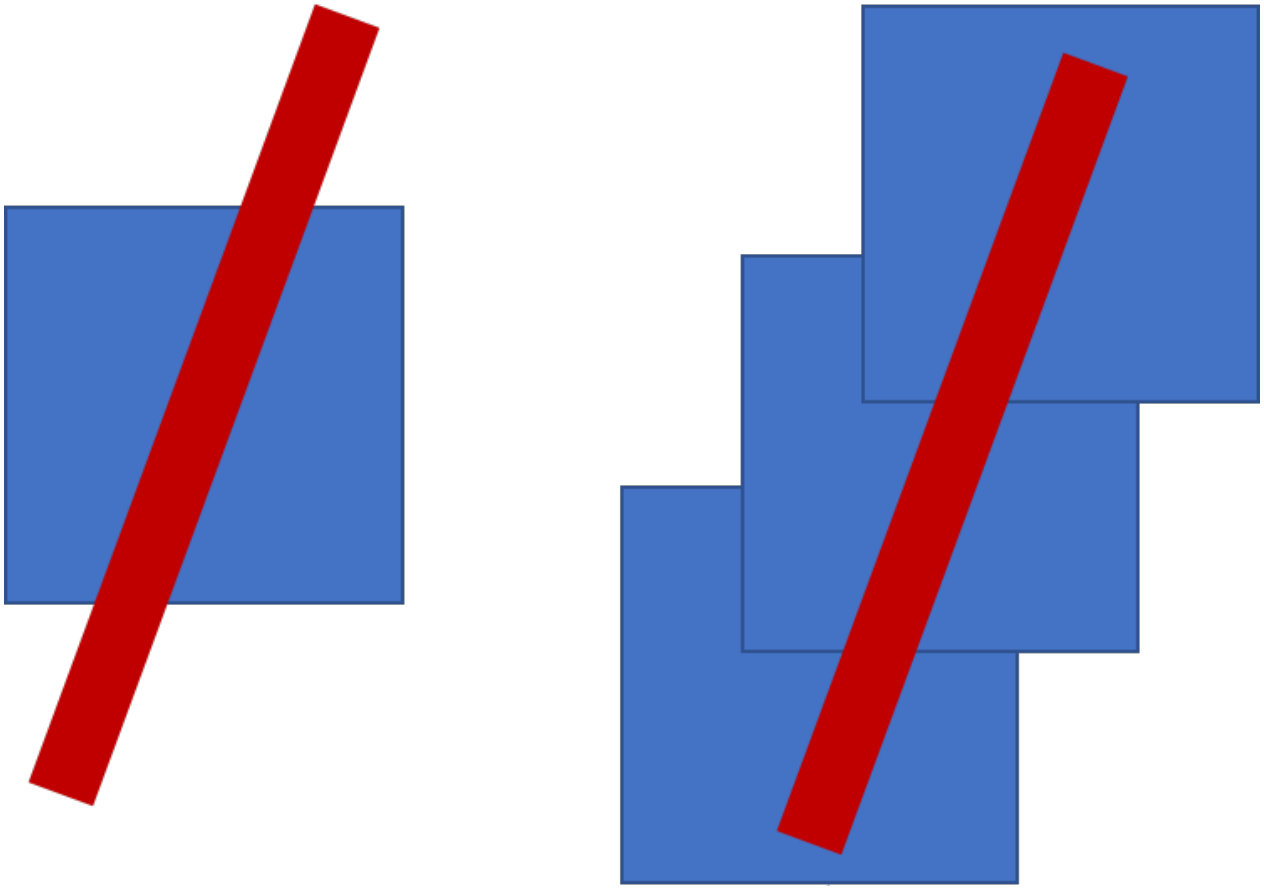


Figure 9.5: : On the left, only one example with the road, on the right, several example with one road

9.3.3 Meshing

In order to compute all transport equations, the CFD needs to discretize the full volume into smaller ones, called cells. What is considered small depends on the case and the variables of interest. The smaller the volume for the discretization, the more cells the computation will have to be solved. Thus, a good mesh is a mesh avoiding numerical errors by being small enough, an efficient mesh is one avoiding numerical errors with the biggest cells possible to reduce calculation cost. This leads to the crucial question of the choice of the size of the cells. For that, the first way is to find a study working on an analogue issue and chose the same size. Secondly, it is possible to test several meshes for the same case and study the difference between the grid. If the difference is small then it is not necessary to refine the mesh and the coarse mesh is sufficient. To evaluate this difference and the underlying errors, many methods exist (128).

In the thesis of Nicolas Reiminger, this parameter has been studied and it was found that a mesh of $0.5^3 m^3$ near the walls is sufficient to avoid numerical errors while keeping

the number of cells in the domain satisfying. However, if every cell has a size of $0.5^3 m^3$ for an area of 300x300x(15*5)m^3, that would represent 54,000,000 cells, which would require a computation time of a week or two on 128 cores for a single wind direction. To avoid that, it is only required to have a fine mesh where the gradient are large, i.e., near walls. In the free stream, it is possible to have bigger mesh, as seen on Figure 9.6. The cells can have varying sizes, depending on their proximity to walls. The range of cell sizes used here varies from 16m to 0.5m. Thanks to that, an area of 300x300x(15x5)m^3 only needs around 500-700k cells.

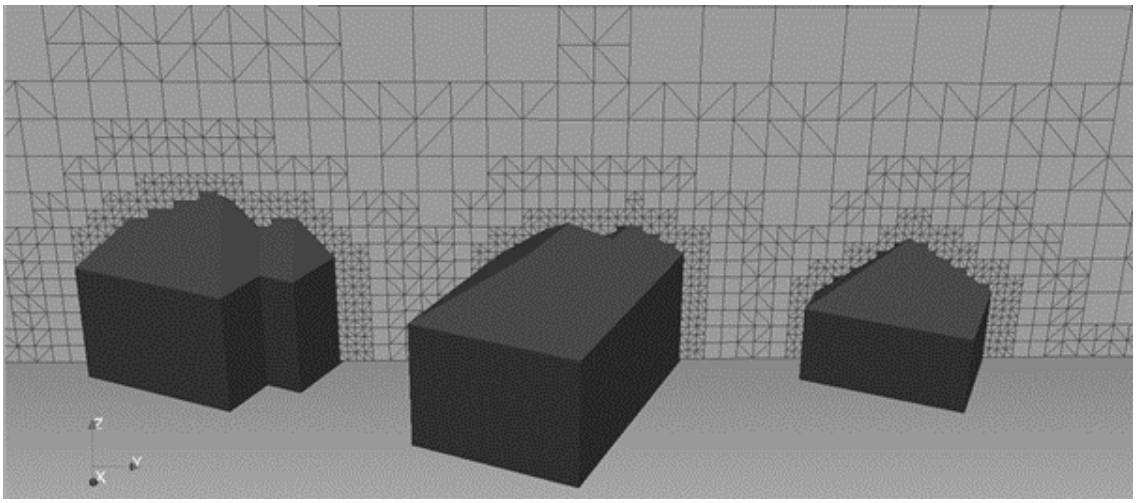


Figure 9.6: Illustration of the adaption of the size of the cells of the mesh near buildings.

The rules for the mesh can be used for every neighborhoods since they are all analogous.

9.4 Data preprocessing

9.4.1 Data collection from CFD modelling

To make an OpenFoam case, several steps are needed.

Acquiring geometries The first step to make a CFD case is to create the geometry. One can create an example by hand with a drawing software or use an available dataset of real geometries. The city of Strasbourg does provide 3D model of the building layouts of the city (https://data.strasbourg.eu/explore/dataset/odata3d_maquette/custom/). This provides a wide range of different building layouts. However, to exploit them, it is necessary to go through different conversions to get a proper format for the software. The following workflow is used:

(.skp) sketchup (.dwg) → (.dwg) autocad (.iges) → (.iges) salome (.stl) → (.stl) OpenFoam

It is possible to directly convert .skp into .stl however the blocks constituting a building are lost and only faces remained which causes issues on treatment later on. Indeed, as previously stated, for simplification, the ground is considered flat and buildings are all at the same level. However, this is not true in the real dataset, and it is necessary to realign the buildings on the ground. Concerning the roads (emission sources), they are not necessarily placed where real roads are, as the aim of a simulation case is to have a lot of roads to make more examples.

OpenFoam parametrisation and computation Once the geometry is set, it is necessary to prepare the OpenFoam simulations with the right parameters. Some files are only needed to be completed once for every simulation such as the calculation schemes and resolution. However, for each simulations there are still tens of files that must be changed for each simulation. The boundary conditions must be given for every geometry face, the meshing parameters like the size of mesh, number of refinement, inlet wind speed angle, etc. In order to do that, a script interacting with Openfoam has been written to initialise the simulation automatically, given a set of geometries (buildings) and an input parameters.

From a geometry set, several wind directions are computed. Another script is responsible for launching simulations for each wind direction while mapping the result of the previous direction to reduce computation time. This script also extracts the tracking points for each direction that will be used to create the examples for the training. The simulations are calculated until convergence or pseudo convergence.

9.4.2 Treatment of simulations results

Extracting concentration data maps In our work, we want to track air pollution at the pedestrian height, so around 1.5m. Thus, a process has been developed to extract the average concentration map at this height from the OpenFoam simulations The concentrations are then normalized by the length of the road and the inlet wind speed to have a constant linear emission µg/s/m between simulations of different road length and possibly wind speed.

Extracting spatial data maps The CFD, given a geometry and a pollutant source, is capable of determining the concentration field in an area. Thus, the machine learning model may be able to infer concentration field from these two types of spatial information. In order to do so, the map of distances from pollutant sources is determined as well as the map of altitudes of the building.

Transformation of data into images As we decided to work with convolutional architectures, all data are transformed into images to keep spatial information. To turn them into images however it is necessary to choose a scale [min,max] that will correspond for 8bit images to [0,255] with 0 being black and 255 being white. The difficult point here is to choose a scale that will on most images use the whole greyscale range:

- Building height scale: Given the type of neighbourhood that are dealt with, the building generally are not higher than 40m, thus the scale is [0,40],

- Distance from pollutant emission source scale : The areas are 300x300m^2 and the source is at the center of the image, the diagonal distance of the square area is then $300 \times \sqrt{2}/2 = 212m$,

- Concentration map: This is the hardest to choose because the concentration range can vary a lot between images. Three approaches are possible. To scale each concentration file with its own max. To fix a concentration value not to exceed a value for example 80 µg/m^3. To choose a number superior to the max pollution from all the known neighborhood. The first method issue is that for real application, on new neighborhood, it is not possible to know its max concentration value without CFD modelling. For the second method, the issue is that value that exceed the chosen pollution scale will not be differentiable. For the third, most of the time, the AI will use only a tiny portion of the greyscale range which will affect its end result.

Thus, when testing and comparing different architectures, the first method will be used. However for real time use, the second method will be used since the neighborhood that are monitored will not have CFD models.

9.5 Conclusion

To generate examples to train the Deep Learning model, it is necessary to have a reliable base. The model elected here is a CFD model for its capacity to consider complex flows and buildings in urban areas. The hypothesis on the CFD model were presented, a RANS-Kϵ RNG model coupled with a transport equation assuming a neutral atmosphere. Once the model, meshing and assumptions are well defined, several strategies to create examples were carried out. The best strategy to date is to create CFD examples on large areas with several pollutant sources, since respectively it proportionally reduces the empty areas and the computation time necessary to solve the airflow equations. As deep learning needs many examples, the creation of the data has to be automated. The last step is then to automatically transform the data from the CFD software into data for the deep learning model. For that,

the CFD data have to been scaled. For the geometric data, a scale of 40m is chosen for the height of the buildings and 210m for the distance from the pollutant source. The choice of scale for the concentration field is more tedious. When testing and comparing architectures, the strategy of scaling each concentration field to its own maximum will be used. When using the model to predict real field concentrations, a scale of 80 µg/m^3 will be used.

Chapter 10

Deep learning architectures to learn air pollutant dispersion from fluids mechanics

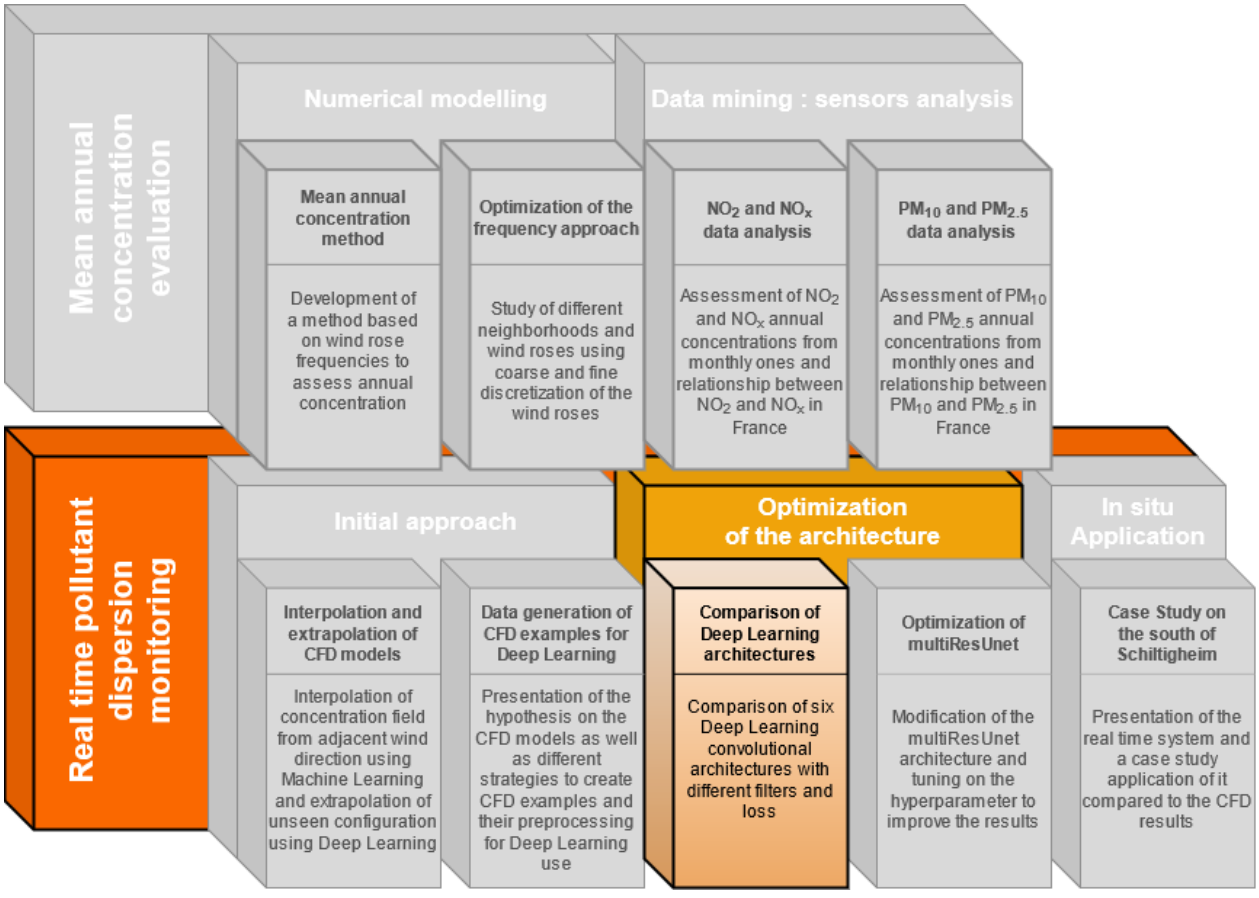


Figure 10.1: location of this chapter in the thesis

This chapter has been submitted in the journal *Expert Systems with Applications* under the title "Assessment of capability of deep learning to predict air pollution dispersion from fluid mechanics" (56).

The generation of data for the Deep Learning model needs a reliable physics model to learn from. The hypothesis and rules to create examples were covered before and using them, thousands of examples have been generated. The example for the Deep Learning model are in the end constituted of four elements a building layout, the distance from the pollutant source, the pollutant dispersion from the road and the direction of the wind taken as a rotation. Now that enough examples have been generated, it is possible to make the first deep learning models learn on several neighbourhoods and to make it predict unseen neighbourhoods. However, several deep learning architectures of the same type exist and have shown good results in different imagery field. These models are able to extract different scale of spatial features and their interactions leading to the localisation of the spatial phenomenon under study from high dimensional dataset. Thus, it is necessary to compare different architectures to determine which one performs best to solve pollutant dispersion on neighbourhoods.

10.1 Introduction

A popular approach for local pollution assessment is to simulate its dispersion with Computational Fluid Dynamics (CFD), but this requires a lot of computing resources (59). It is therefore adapted to compute mean annual average but is not ideal for large areas or use in real time. On the other hand, to cover large areas in real time, some models like plume exist. Unfortunately, they are based on hypothesis that make them unsuited for urban areas where the air pollution is the most stringent (71).

The recent advances in machine learning and deep learning may provide the answer to these limitations. Indeed, it has much progressed over the recent years especially thanks to the improvement and democratization of highly threaded parallel computing processors (19). Recently, it has proved to outperform previous state of the art methods in various fields such as speech recognition, visual object recognition, object detection and many other domains such as drug discovery or genomics (72). These new methods have not gone unnoticed in the domain of physics and numerical simulation. Their use are still nascent in these domains. For example, deep learning models were trained to perform numerical simulation to accelerate them as in (110; 46; 54). Deep learning has also been used in the domain of air quality to estimate the pollution based on pictures (31), sensors (36), to extract the main features explaining the pollution variation (112) or urban systems (45).

To build a fast and accurate system able to predict air pollution in real time based on wind, traffic and buildings geometry, we tried to use a convolutional network (CNN), that

has proven to be able to treat spatial information successfully, to learn pollutant dispersion from CFD. This will overcome the issue of speed related to standard CFD computation while proposing a model that is more appropriate to urban areas. In this paper, 6 CNN models (namely U-Net, SegNet, linkNet, MultiResUnet, PSPNet and FCN) are trained and tested, based on 5000 CFD examples. The aim of the paper is to verify the capability of such models to determine pollutant dispersion rapidly and accurately, and which of these well known CNN architecture performs better to solve this problem.

10.2 Material and methods

10.2.1 Physical numerical model

To learn pollutant dispersion in open urban areas, deep learning architectures need examples to be trained. To simulate wind and underlying pollutant dispersion, a popular technique is to use CFD as in (99; 16; 122). To perform simulations, Openfoam 5.0 was used. OpenFoam[1] is an open source software dedicated to numerical simulations, ranging from financial to radiation to fluids mechanics. Hypothesis for the simulation were the following:

- Reynolds Averaged Navier Stokes (RANS) approach was used;
- unsteady simulations were performed;
- the turbulence model for the RANS model is k-epsilon renormalization group (RNG) proposed by (168);
- a transport equation for the pollutant dispersion;
- upper and lateral boundaries are symmetry conditions;
- the outlet is a freestream condition;
- buildings have no slip conditions;
- the atmosphere is considered neutral, therefore using a logarithmic inlet profile and turbulence for k and epsilon parameter calculated as proposed in (125):

$$U = \frac{u_*}{\kappa_{k-\epsilon}} ln \frac{z_0 + z}{z_0} \tag{10.1}$$

$$\epsilon = \frac{u_*^2}{\sqrt{C_\mu}} \tag{10.2}$$

[1] https://www.openfoam.org/

$$k = \frac{u_*^3}{\kappa_{k-\epsilon} z} \tag{10.3}$$

where, U is the inlet speed $[m.s^{-1}]$, ε is the turbulent dissipation rate $[kg.m^{-1}.s^{-4}]$, k is the turbulent kinetic energy $[kg.m^{-1}.s^{-3}]$, u_* is the shear velocity [m/s], $\kappa_{k-\epsilon}$ is the von Kármán constant, z_0 is the roughness length [m] and z is the altitude [m].

Guidelines provided by (42) were respected when constructing the domain and the meshes of every simulation. For each simulation, the top of the domain is situated at a minimum distance of $5 \times H$ from highest building and the lateral, inlet and outlet boundaries at a minimum distance of $5 \times H$ from the closest building, with H the height of the tallest building in the domain. A mesh sensitivity analysis was made and a mesh with $0.5m$ for the cell closest to the building were found to be enough to be insensitive. An example of a neighborhood of the meshing is shown on Figure 10.2.

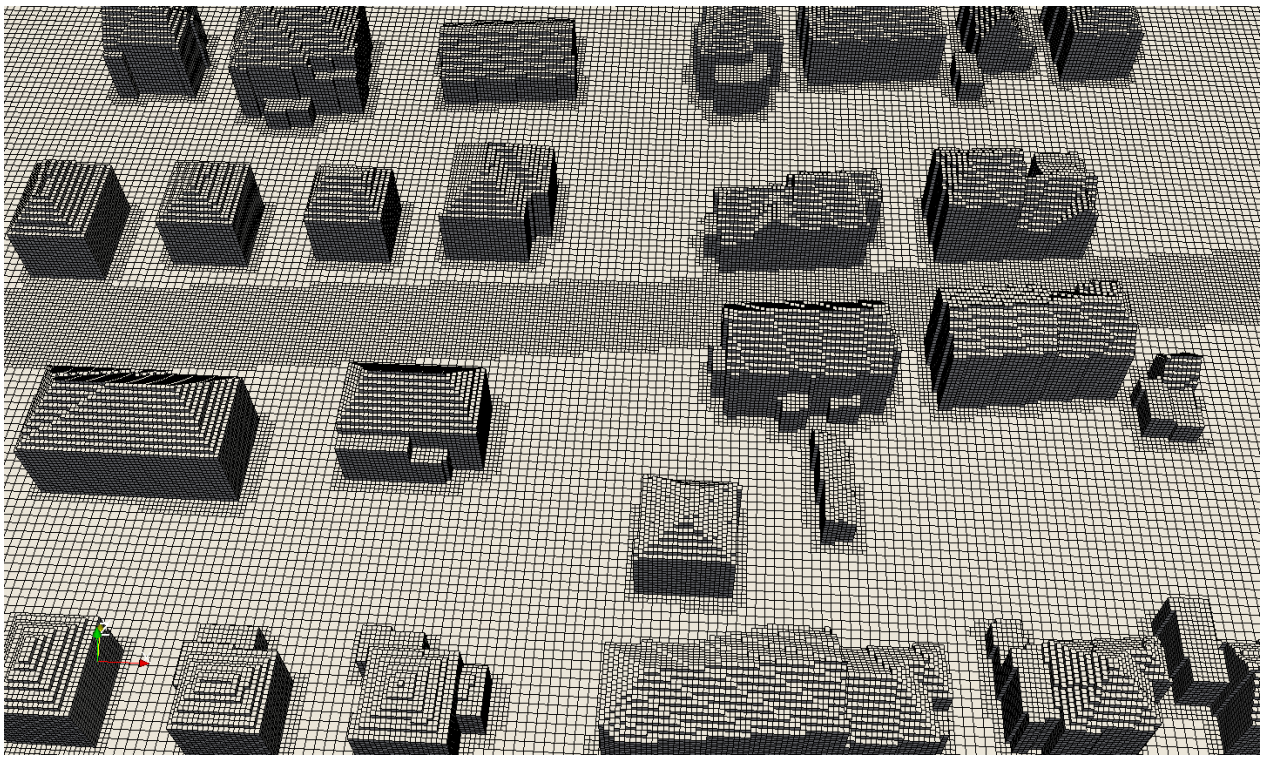


Figure 10.2: Example of the meshing on a building layout used to create the examples

More details on the model, equations and validation, please refer to (123) where the same approach has been described and properly validated.

The approach, model and meshes described above have been found to be able to reach an error which is less than 10% compared to experimental measures as show in (123) and a similar approach have been proven to have an overall error of about 30% compared to a

real *in situ* situation in urban areas (127). The numerical results will be considered as the ground truth for the deep learning algorithms.

For the sake of simplicity the wind will always come down from the y axis. Around $5,000$ examples of couples of building layouts and pollutant sources have been computed to be used for the deep learning training and validation.

10.2.2 Deep learning architectures

Deep learning architectures have shown to be very effective to tackle spatial information, for example to predict urban traffic (105), (149) or to predict citywide passenger demands (179). Furthermore, convolutional ones have shown to be very effective. Indeed, for semantic segmentation, CNNs have proven to be able to overcome issues that were not achievable before in a lot of different fields. For example, it has been used in the medical field to identify certain cell types as in (132), in face recognition as in (98), or remote sensing images analysis (160).

The strength of CNNs to treat spatial information have also started to be used to predict physical phenomena as in (46) and (54). To simulate physical phenomena, such as fluid mechanics, it is common to define a set of fundamental equations describing the phenomena and then, if needed, to implement a numerical code that will solve them step by step, until reaching convergence (or pseudo convergence) or during the transient wanted time. These steps generally require vast computing time resources.

Deep learning has already been used in fluid mechanics, especially to determine the speed vector field (46; 54). Here, we have the ambition to go further and study the ability of such architectures to build a model able to determine pollution dispersion given buildings' geometry, wind and traffic information. For that, CNN's architectures designed for image segmentation tasks will be compared. The first architectures used are encoder-decoder, with, chronologically, U-net (132), SegNet (11), linkNet (29) and multiResUnet (51). They follow the same principle of encoding the information to get the context and then decoding it to get the precise location of the wanted feature. However they have small variants on the way they handle spatial information through the layers. A multi scale representation method with PSPNet (178) will also be used. And finally, a classical full convolutional network (FCN) (82).

The models can have different number of free parameters depending on the number of layers and filters at each layer. To test different numbers of trainable parameters, the architectures will be tested with several filter per level. Each of this architectures have a level in which the number of filter is minimal as it can seen on 10.3 noted "F". This min filter will be used to describe the variation of free parameters in the models.

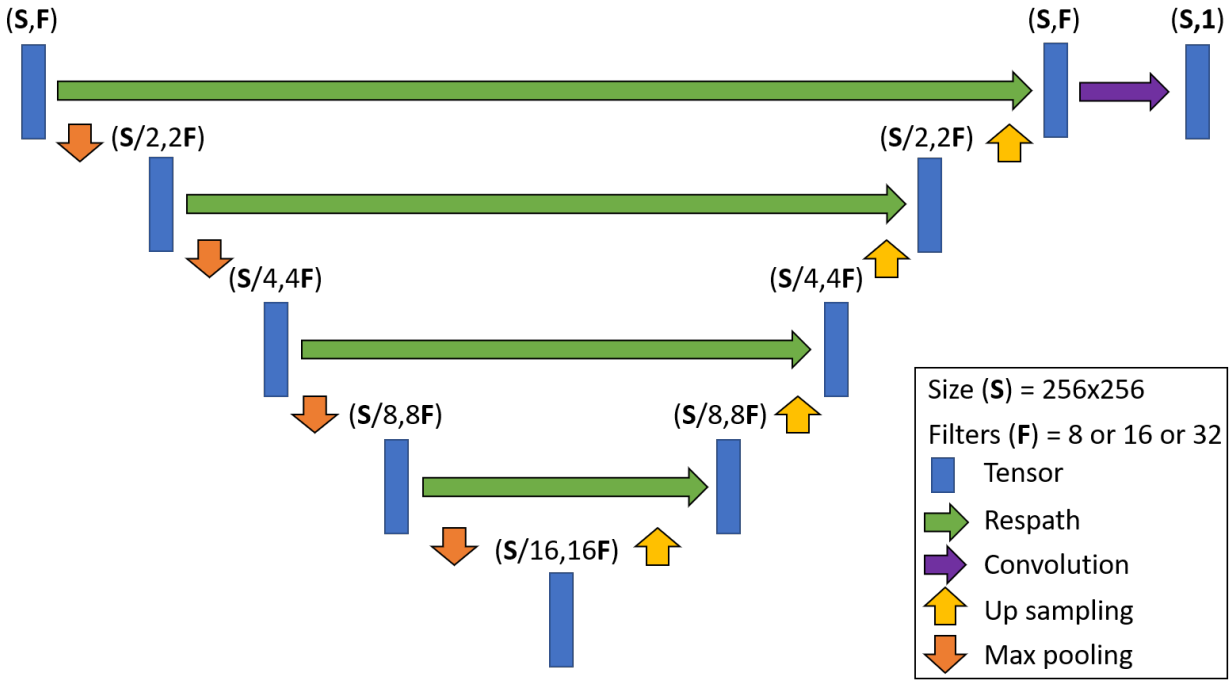


Figure 10.3: Architecture of the multiResUnet

10.2.3 Input and output data for the deep learning models

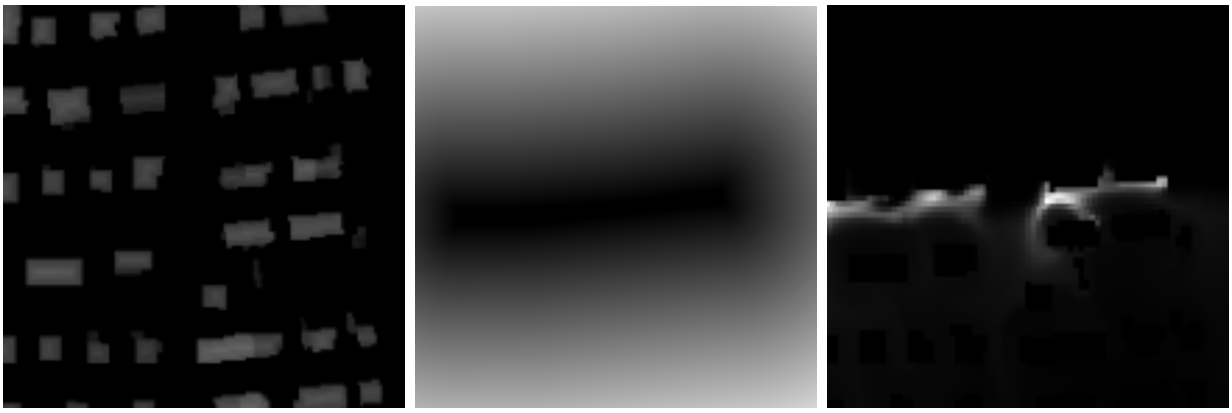
The computation from the physical model are turned into 2D maps of $150 \times 150m^2$ at a height of $1.5m$. Two maps will be used as input, the first map representing the height of the buildings and the second map the distance from the pollutant source. The last map, will be the normalized pollutant dispersion field. An example of the images used the architectures are shown below:

In this study, $4,919$ examples were produced, divided with $3,687$ for training, 410 for validation and 822 divided into 28 subsets for testing according to the methodology provided by (40). The training was performed for 25 epochs with a batch size of 6. The optimizer used is Adam. A callback patience of 5 epochs was used on the validation data loss.

10.2.3.1 Deep learning loss

For every model, three losses are tested. Two well known losses, binary crossentropy (bce) and mean squared error (mse) as defined in Equations 10.4 and 10.5.

$$bce = \frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i), \tag{10.4}$$



(a) Buildings (b) Pollution source (c) CFD result

Figure 10.4: Images given as input to the network (a) the height, shape and position of each building in the area, (b) the distance from the pollution source, and (c) the corresponding CFD simulation, considered as the right output for the CNN.

$$mse = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{10.5}$$

A custom loss, called $J_{3D}loss$, was also tested (see Eq. 10.6). It is based on the Jaccard index, originally called community coefficient, that aims at comparing the intersection with the union of two binary set. This index is often used in segmentation to compare the predicted binary mask to a ground truth segmentation mask. But here, the pollutant concentration is a continuous value, so areas can not be compared as in segmentation. However, the continuous value can be considered as a third dimension and so the intersection over the union is not computed between two surfaces but two volumes. The loss is computed between two pairs of images as following:

$$J_{3D}\ loss = 1 - \frac{V_{pred}\bigcap V_{true}}{V_{pred}\bigcup V_{true}} \simeq 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{min(y_i, \hat{y}_i)}{max(y_i, \hat{y}_i)} \tag{10.6}$$

where V_{pred} and V_{true} are the respective volume of the two images with the pixel value as the third dimension respectively for the predicted and ground truth image, N is the number of pixels, y_i^{true} is the value of the i^{th} pixel of the true image and y_i^{true} is the value of the i^{th} pixel in the predicted image.

10.2.4 Evaluation of the results

Popular metrics in the air quality field To evaluate the predictions made by the deep learning architectures, several metrics will be used. Indeed, each measures different aspects of

Models	**Min filters**	**Losses**
FCN	1 - 2 - 4 - 8	J_{3D} - *bce* - *mse*
PSPNet	8 - 16	J_{3D} - *bce* - *mse*
linkNet	8 - 16 - 32	J_{3D} - *bce* - *mse*
SegNet	8 - 16 - 32	J_{3D} - *bce* - *mse*
multiResUnet	8 - 16 - 32	J_{3D} - *bce* - *mse*
Unet	8 - 16 - 32	J_{3D} - *bce* - *mse*

Table 10.1: Summary of the different variants of each model tested in this study.

the model and helps to see strength and weaknesses better than reducing the analysis on one single metric. In the air quality field, the study of Chang *et al.* (27) provides several metrics to be used to evaluate and conclude on the quality of a model. Six metrics are provided, but some are equivalent and evaluate the same aspect of the result. Thus, we keep only four of them for the presented study. Fractional Bias (FB) measures if the prediction mean is globally the same as the ground truth mean value. Normalised Mean Squared Error ($NMSE$) measures if there are extreme differences between the prediction and the ground truth. The fraction of predictions within a factor of two of observations ($FAC2$) enables to measure that on overall, the predictions are within an accepting error margin. And finally, R index, that compares the correlation between the two datasets (ground truth and predictions). FB and $NMSE$ are to be minimised at 0, $FAC2$ and R are to be maximised at 1.

$$FB = \frac{(\overline{C_{ref}} - \overline{C_{pred}})}{0.5(\overline{C_{pred}} + \overline{C_{ref}})}, \tag{10.7}$$

$$NMSE = \frac{\overline{(C_{ref} - C_{pred})^2}}{C_{pred}C_{ref}}, \tag{10.8}$$

$$FAC2 = \text{fraction of data that satisfy } 0.5 < \frac{C_{pred}}{C_{ref}} < 2, \tag{10.9}$$

$$R = \frac{\overline{(C_{ref} - \overline{C_{ref}})(C_{pred} - \overline{C_{pred}})}}{\sigma_{C_{pred}}\sigma_{C_{ref}}}, \tag{10.10}$$

with C_{pred} the predicted concentration field and C_{ref} the reference concentration field (ground truth).

In (27), the authors propose ranges of values on the above parameters to assess if an air quality model is satisfying. They also underline that for spatial models, these values are harder to reach. The suggested values are:

– FAC2 > 0.5,

– NSME < 1.5,

– $|\text{FB}| < 0.3$.

Metrics related to images On the above metrics, three more that are commonly used to compare images will be estimated. The relative mean absolute error (MAE_{rel}), J_{3D} that is also used as a loss and described previously, and the Structural Similarity Index ($SSIM$) designed to measure the visual quality between a compressed image and the original one. MAE_{rel} is to be minimized. $SSIM$ and J_{3D} are to be maximized.

$$MAE_{rel} = \frac{|C_{ref} - C_{pred}|}{\overline{C_{pred}}} \tag{10.11}$$

$$J_{3D} \simeq \frac{min(C_{ref}, C_{pred})}{max(C_{ref}, C_{pred})} \tag{10.12}$$

with C_{pred} the model prediction concentration and C_{ref} the reference concentration (ground truth).

$$SSIM(A, B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c2)}{(\mu_A^2 + \mu_B^2 + c1)(\sigma_A^2 + \sigma_B^2 + c2)} \tag{10.13}$$

$$c_1 = (k_1 L)^2 \quad c_2 = (k_2 L)^2 \tag{10.14}$$

where μ_A and μ_B are the respective average of A and B, σ_A^2 and σ_B^2 are the respective variances of A and B, σ_{AB} is the covariance of A and B, L is the dynamic range of the pixel values and k_1 and k_2 are two constants respectively 0.01 and 0.03 (by default).

10.3 Results

To compare the architectures, the methodology provided in (40) will be used. This methodology allows to compare different models by ranking them on their performance on a metric over several datasets. This ranking can then be used to make a critical difference diagrams. To compare the models, the test dataset composed of 822 examples divided into 28 subdatasets will be used. A subdataset correspond to an emission source (road) with a building outlet.

10.3.1 Loss functions and filters

Three loss functions were tested along several number of filters for each 6 model. The difference between predictions and ground truth was evaluated according the 7 metrics presented above. Nevertheless, as this would produce $7 \times 6 = 49$ diagrams, to sum up the result, the 7 metrics of each variant were concatenated together for each model to determine the best performing variant for each model. Thus, the 6 models diagrams are presented on the critical difference diagrams in Figure 10.5. Notations on the diagram for the model are "loss"_"min filters", for example a model that uses binary crossentropy and 4 min filters will be noted "bce_4".

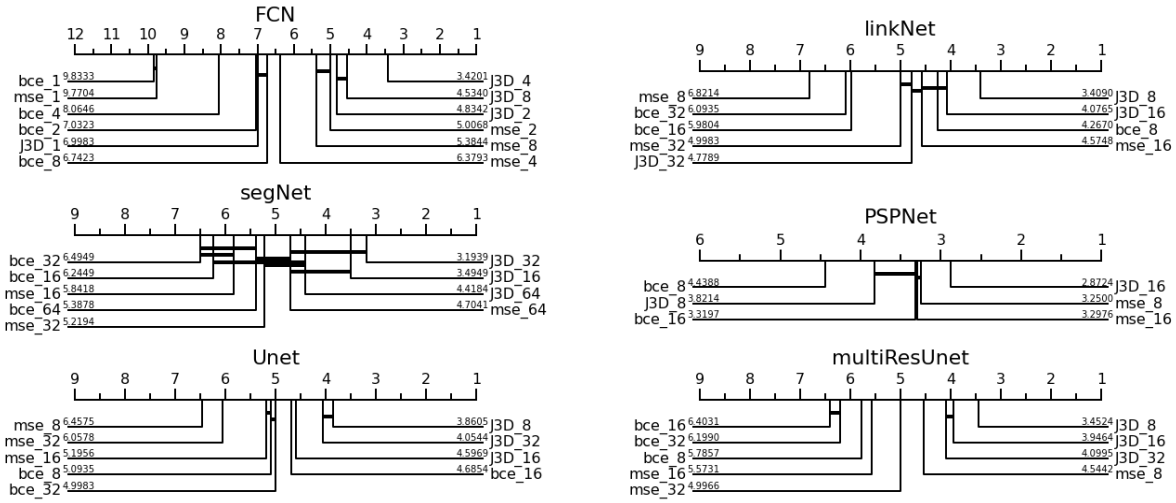


Figure 10.5: Ranking of the different variants for each model using all the metrics

As it can be seen on Figure 10.5, the J_{3D} loss always comes first for every model.

10.3.2 Architectures

Using the best variant of each model as determined in the previous subsection. The same approach of the critical difference diagram will be used to determine which model performs best. The results for all the metrics with all the best variant of each model is presented on the Figure 10.6

metric	**FAC2**	**NMSE**	**FB**	**R**	**MAE rel**	**J3D**	**ssim**
mean value	0.8	3.7	0.3	0.8	0.7	0.5	0.8
expected value	$\approx > 0.5$	$\approx < 1.5$	$\approx < 0.3$	1	0	1	1

Table 10.2: Evaluation of the results of the multiResUnet on each metric

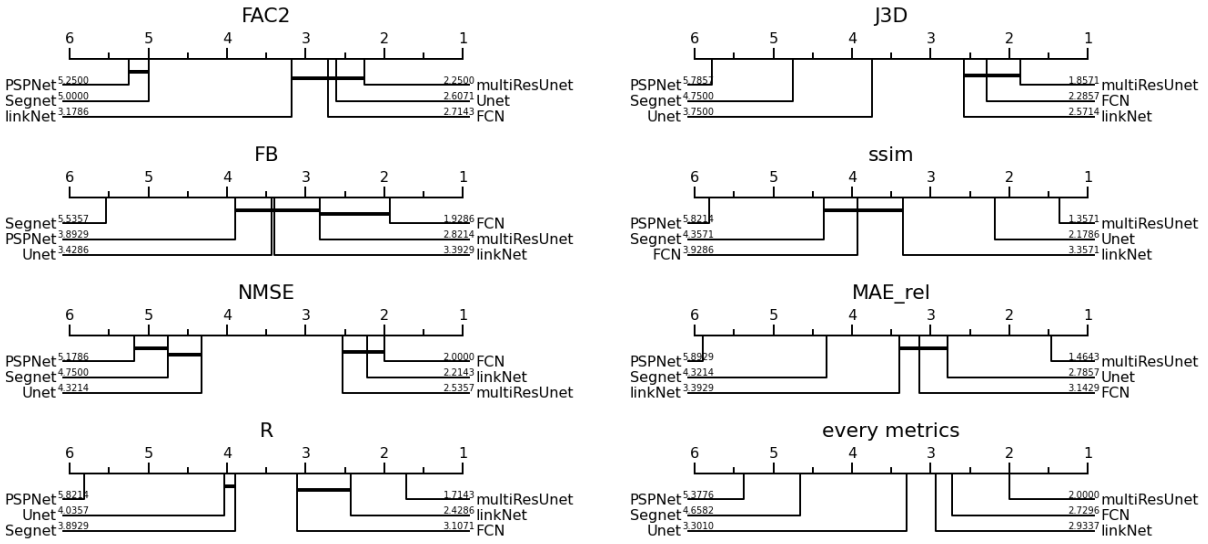


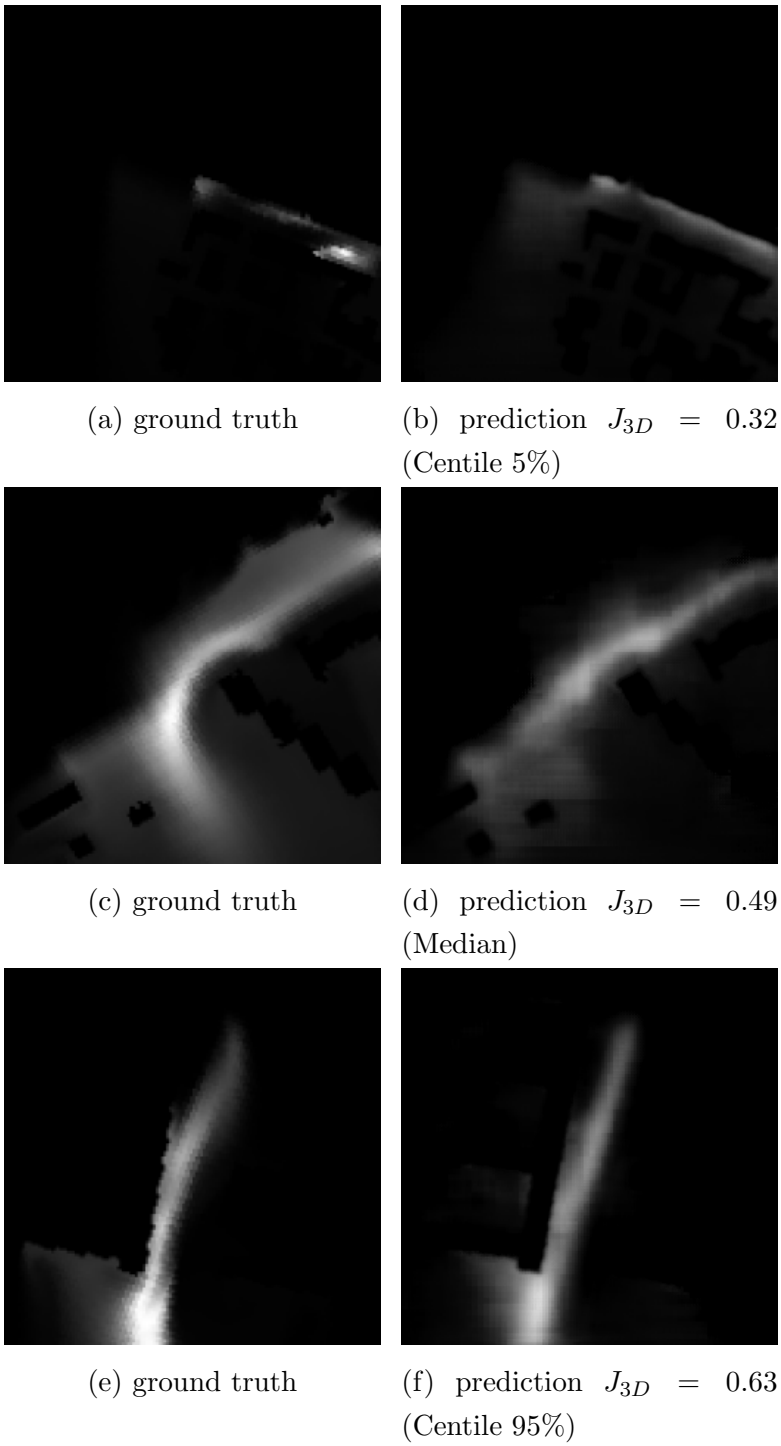
Figure 10.6: Ranking of each best variant for each model according to each metric

The architecture that manages to predict best pollutant dispersion on overall is multiResUnet which is first 5/7 times and always at least in the first statistically indistinguishable group. When all metrics are considered together, multiResUnet becomes first. The absolute results on all metrics for multiResUnet using 8 min filters and J_{3D} are given in Table 10.2. It can be seen that multiResUnet using the J_{3D} loss managed to perform within the standard performance of a good model for 2 out of 3 metrics widely used in air quality.

Examples of the multiResUnet predictions against the CFD model for the centile 5 %, the median and the centile 95 % of J_{3D} are shown on Figure 10.7.

10.4 Conclusion

Several architectures that have proved their efficiency in other fields have been applied to pollutant dispersion modelling. For each of these architectures, several variants with different amount of minimal filters were trained using three different losses. For each model, the variants were compared against several metrics and it was found that J_{3D} loss gave the best results for every model to predict airborne pollutant dispersion. The architectures were then compared one against the others and it was found that multiResUnet had the overall best results. Using metrics wildly accepted in the air quality field, 2 out of the 3 metrics are in the accepted range for a good air quality model when compared to the ground truth. The architecture was able to obtain these results in minutes compared to the computation that requires tenths of hours. These results are promising to enable real time pollutant dispersion in urban cities with CFD accuracy.



(a) ground truth

(b) prediction $J_{3D} = 0.32$ (Centile 5%)

(c) ground truth

(d) prediction $J_{3D} = 0.49$ (Median)

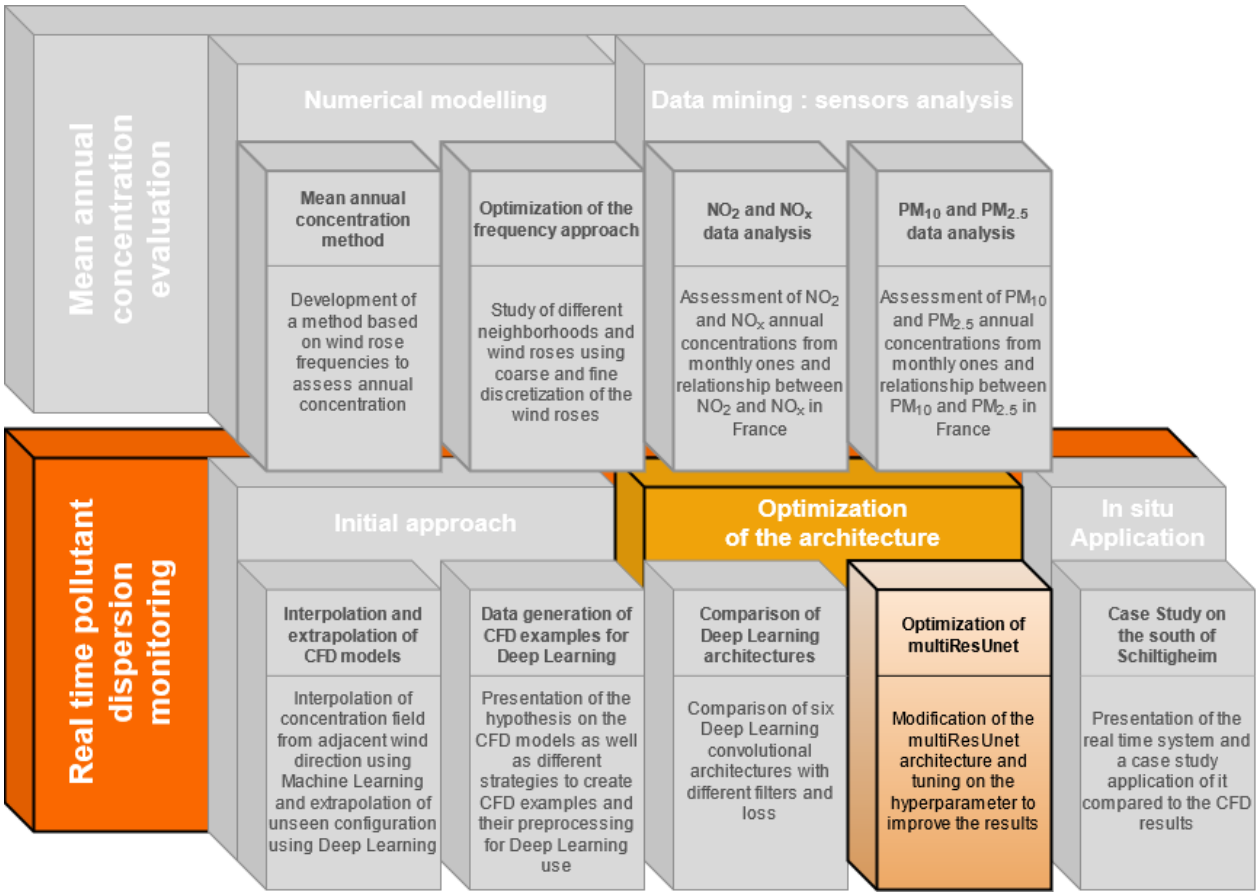
(e) ground truth

(f) prediction $J_{3D} = 0.63$ (Centile 95%)

Figure 10.7: Examples of predictions from the multiResUnet

Chapter 11

Dwelve in depth the MultiResUnet architecture



11.1 Introduction

In the previous chapter, several state of the art Deep Learning architectures in the domain of imagery and spatial data treatment have been compared. The multiResUnet gives the best results according to the conducted tests. Nevertheless, the Deep Learning architectures have not been created to deal with CFD data and hence, improvements may be done to adapt the geometry to the pollutant dispersion topic. Moreover, Deep Learning models have many hyperparameters, finding an optimised set of them could improve the results.

11.2 Size of the area of interest

The aeraulic of a neighborhood is determined in part by the buildings that make up the neighborhood forming obstacles for the wind. It is therefore necessary to consider the buildings located around the area of interest for the calculations. To do this in CFD, a bigger area is modeled than the desired area. Thus, it is important for the AI to know the surrounding of the area of interest it is working on. For that, the first hyperparameter to test, is the size of the input area regarding to the output one.

The architectures that have proven to work well on spatial data are the so-called convolutional encoder/decoder architectures. We will see below the definition of encoder/decoder and what is a convolutional filter. It is important to understand these notions to address the issue of the input image being larger than the output image because these notions play an essential role on the dimensions of the architecture and therefore of the input and output image.

What are an "encoder" and a "decoder"? *Encoder:* Encoding the information means reducing the dimensions of the problem and compressing the information contained in the input data. A simple example of encoding is to average the input information. When encoding several times in a row this allows operations to perform on different dimensions of the problem filters and to extract characteristics that the architecture will be able to exploit.

Decoder: Decoding is the reverse operation of encoding; we start from the compressed data to decompress it on several levels which allow further filter and information extraction. An example of encoding and decoding with an FCNN neural network is given in figure 11.1:

What is a convolutional filter? A convolution is an operation described on the Figure 11.2 below:

It is an encoding operation. Depending on the value that are used in the kernel, it will shade light on some features differently. For example, on Figure 11.3 a convolutional filter is

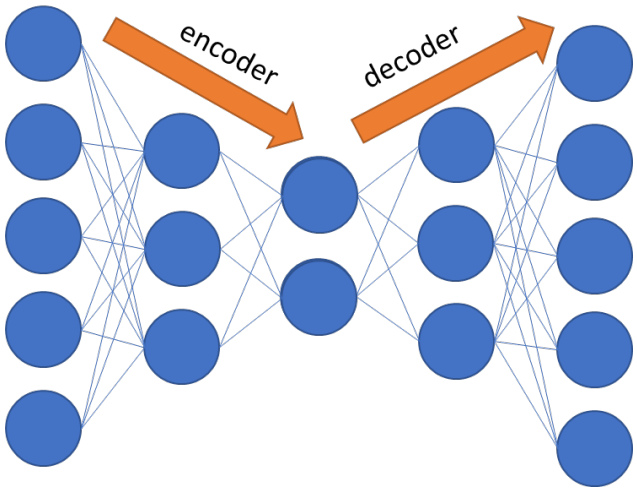


Figure 11.1: Example of encoding/decoder architecture with 5 input data, reduce to 2 data at the bottleneck to then be expanded to 5 data again at output after the decoding process.

used to isolate the water tracer from the rest of the images.

The deep learning architectures that are used to treat spatial information are the fusion of those two main concepts. A typical architecture is composed of many convolutional filters used in the encoding part and then decoded to get the wanted features as shown on Figure 11.4:

What is padding? The convolutional architectures use tensors. These tensors in our case are 2D (x,y) on which convolutional filters are applied making a third dimension, (x,y,f). Every time a convolutional filter is used it reduce the size of the tensors on its (x,y) dimensions. This can be an issue to have a stable and coherent architectures. Thus to avoid this, several strategies are possible, they are called "padding". Padding consist to artificially add information that does not exist to keep the dimensions the same. For example on Figure 11.5, it is possible to add arbitrarily 0 around the tensors to keep the dimensions constant. Another option is to make a mirror effect on the border as shown on Figure 11.6:

Finally, it is also possible not to use padding and accept the dimension reduction as it is done in the original U-Net paper (132).

11.2.1 Training parameters

Dataset: A CFD simulation were computed on about twenty neighborhoods with several different wind directions, half of which have several sources of pollutants. This led to a dataset of 4919 different examples distributed as follows:

- Train set: 3278 examples,

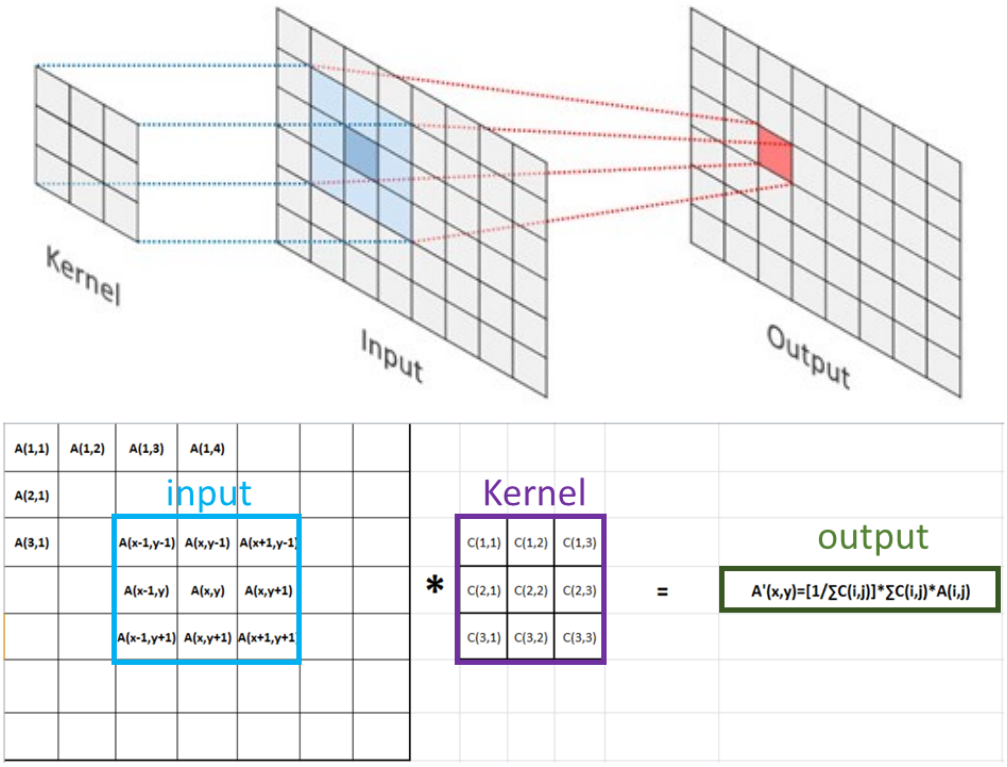


Figure 11.2: 2D convolutional filter (source : http://www.terra3d.fr/a-convolution-operator-for-point-clouds/)

- Validation set: 819 examples,
- Test: 822 examples (extracted from 4 neighborhoods never seen by the AI). This set is further divided into 28 subsets. Each subset consists in a road with its building layouts for different wind directions. These 28 subsets will then be used to compare the results following the strategy presented in (40), with critical difference diagram as in Chapter 10.

To transform the CFD data into images, it is necessary to make a bijection between the physics values with its range from [min range, max range] to [0,255] as done in the chapter 9 in the section 9.4.2. Therefore, it is necessary to define a min and max value to scale the images for the different physical parameters.

Scales:

- Scale for pollution: auto-scale (using the maximum from each image as maximum for the scale),
- Scale for buildings: [0,40] in meter,
- Scale for distance from pollutant source: [0,250] in meter.

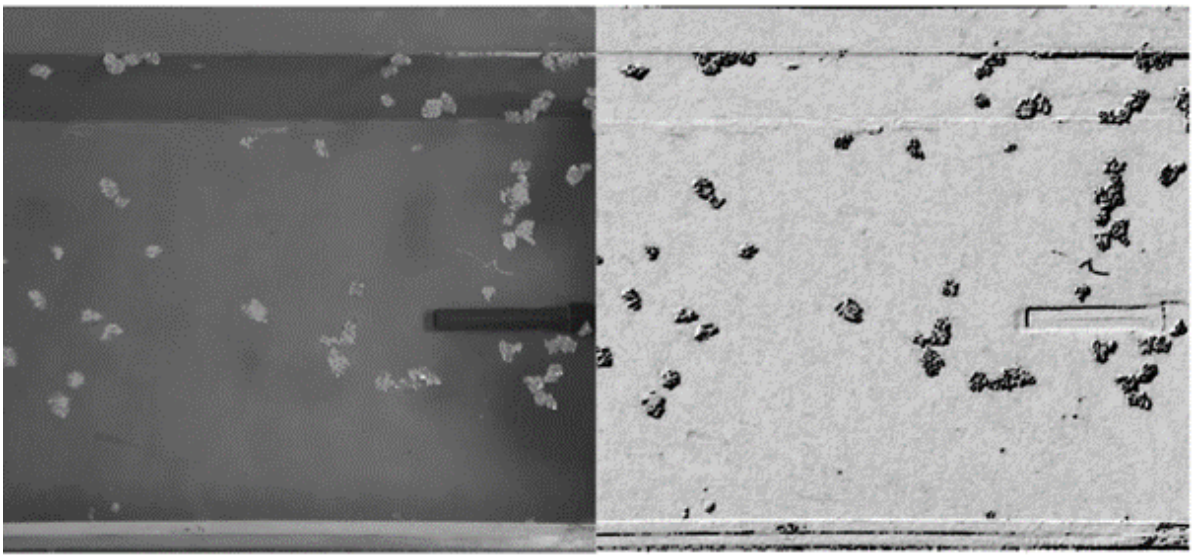


Figure 11.3: Example of an application of convolutional filter for large scale particle velocimetry.

Training parameters:

- Number of filters at bottleneck: 16,
- With padding: batch = 6, epochs = 25,
- Without padding: batch = 2, epochs = 15.

The different models presented here are trained 5 times to reduce training variability and the best result is kept.

11.2.2 Tested architectures

Architecture constrain toward dimensions The multiResUnet in its original version uses zero padding to keep its dimensions and the symmetry between the encoding and decoding parts. Thus, the input must remain at the same size as the output pixel wise.

Another important variable to consider is the resolution of each pixel in meter. To measure that, a ratio between the represented meter and the pixel noted R_{mpx} can be computed as the meter of the real dimension of the image over the number of pixels that represent it. For example, for an image of 300x300 m^2 if 288x288 px^2 are used the R_{mpx} will be equal to 300/288= 1.0417 m/px. In other words, a pixel represents 1.0417m.

So the first architecture to test is the one where input and output images are equal in pixels and with the same R_{mpx}. We choose input images of 144px for 150m. To be more concise X is used to describe the inlet and Y the output. Thus the first trial can be define by the following relations: R^X_{mpx}= R^Y_{mpx}=1.0417 and px^X=px^Y=144.

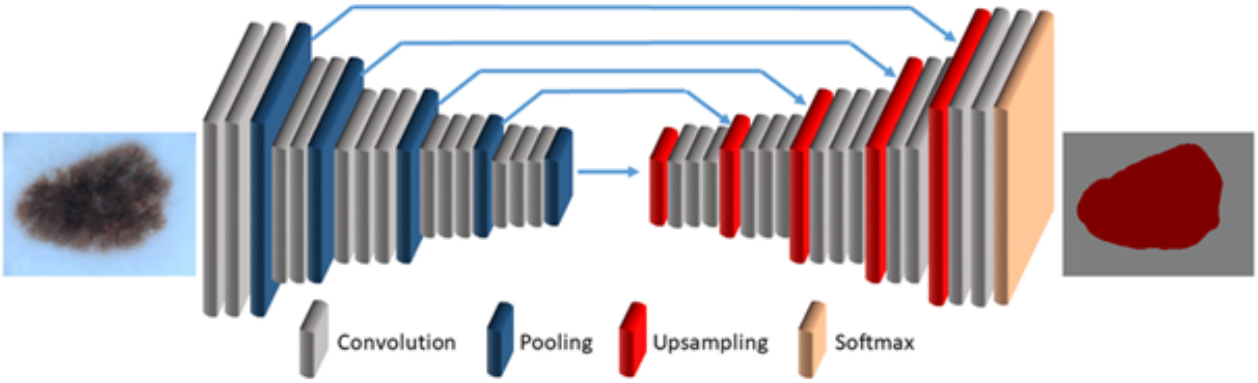


Figure 11.4: Example of a classical encoder/decoder convolutional architecture for skin lesions segmentation (170).

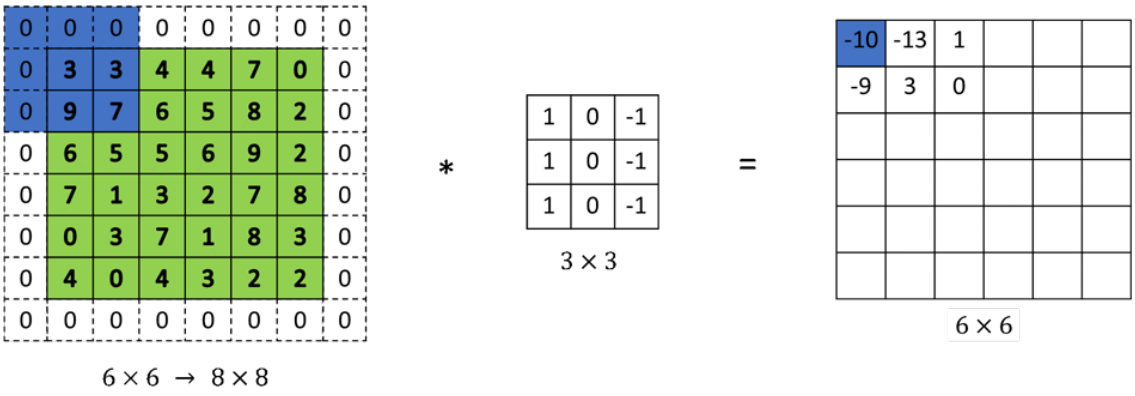


Figure 11.5: Example of "zero" padding. (source: http://datahacker.rs/what-is-padding-cnn/)

Architecture V_0: R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=px^Y=144 The architecture for the V_0 variant is presented in Figure 11.7 and an example of the images used as input and output are given in Figure 11.8.

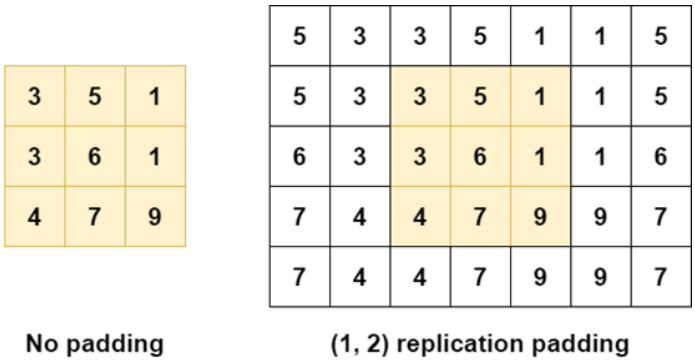


Figure 11.6: Example of "mirror" padding. (source: http://datahacker.rs/what-is-padding-cnn/)

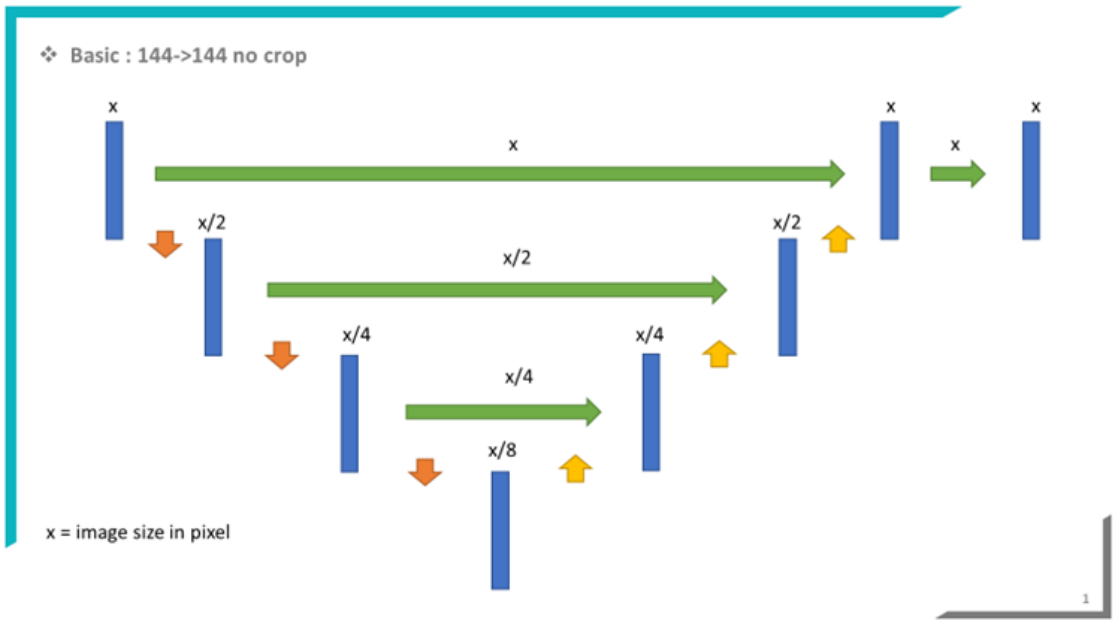


Figure 11.7: Diagram of the V_0 architecture with R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=px^Y=144

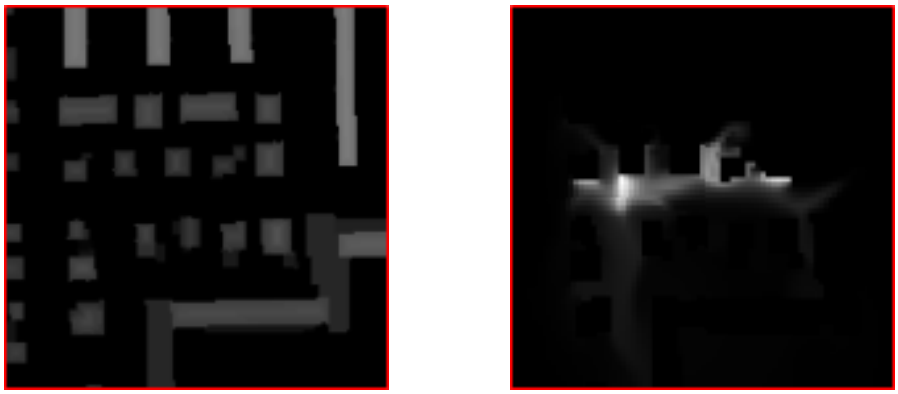


Figure 11.8: Example of images for the architecture with R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=px^Y=144, on the left an input image and on the right an output image.

Architecture V_1: R^X_{mpx}=1.0417, R^Y_{mpx}=0.5208 and px^X=px_Y=288 Is it possible to improve this result by just manipulating the input images and slightly the architecture? The first solution that comes in mind is to simply give as input an image representing more context than the output while keeping the same number of pixels.

For example, an input image representing 300x300m^2 and an output representing 150x150m^2 but both with the same number of pixels of 288. This does not require any modification of the architecture. To sum up, the architecture can be described as R^X_{mpx}=1.0417, R^Y_{mpx}=0.5208 and px^X=px_Y=288.

The architecture is given in Figure 11.9 and the input and output images for the same neighbourhood as previously are presented on Figure 11.10.

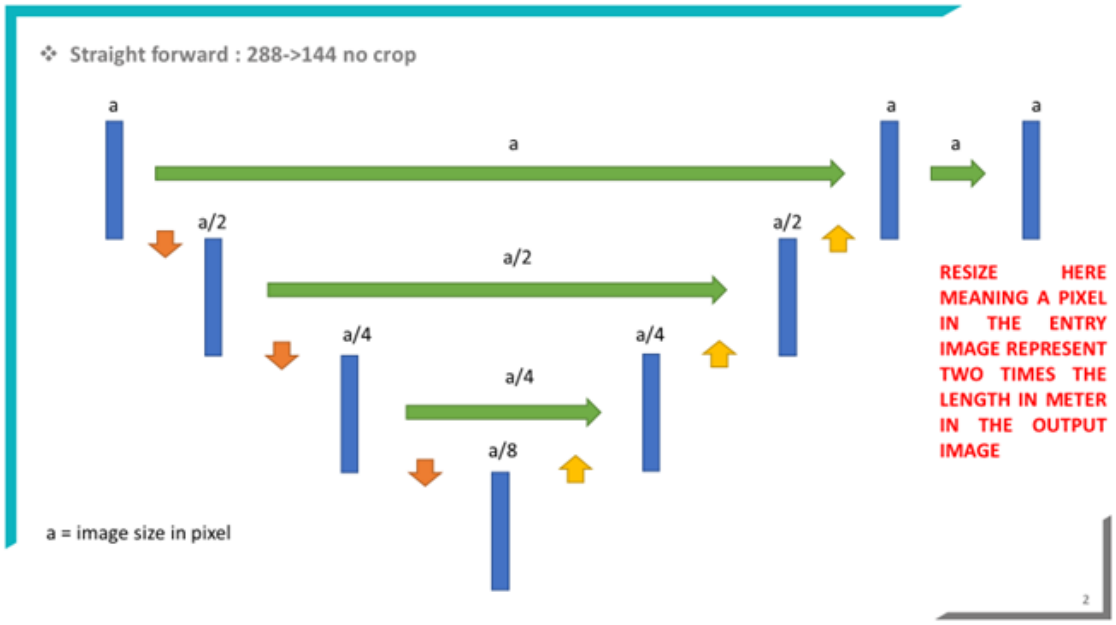


Figure 11.9: Diagram of the V_1 architecture with R^X_{mpx}=1.0417, R^Y_{mpx}=0.5208 and px^X=px_Y=288.

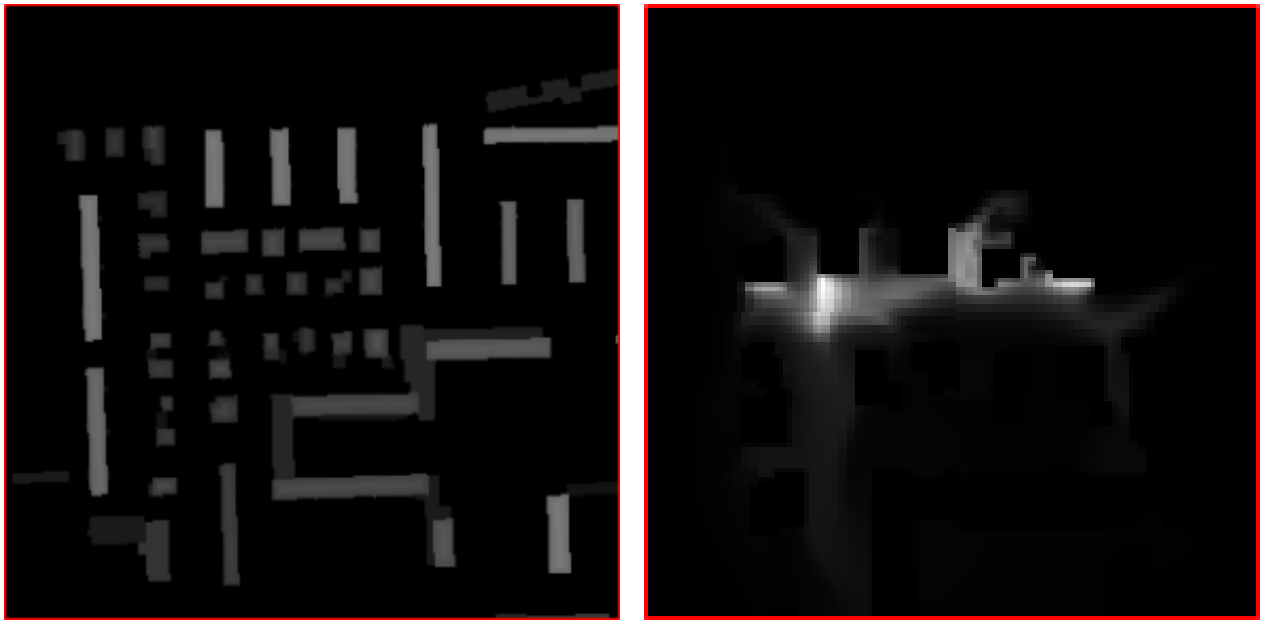


Figure 11.10: Example of images for the architecture with R^X_{mpx}=1.0417, R^Y_{mpx}=0.5208 and px^X=px_Y=288, on the left an input image and on the right an output image.

As it can be seen in Table 11.1, results are worth with this solution. The architecture worked better when both images had coherent dimensions between input and output. This could have been foreseen because of the skip connection inside the AI that bound different

layers of the encoding part with the decoding directly. Thus, this step that is crucial for the localization had the tensors representing different areas. This reduce the effectiveness of the architecture that is not able to correctly interpret the information it is given.

Architectures V_2 and V_3: R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144 In the original U-Net architecture that preludes the multiResUnet, the authors used cropping (example of cropping is shown on Figure 11.11) between the two sides of the architectures to maintain the proper dimensions because they were not using padding.

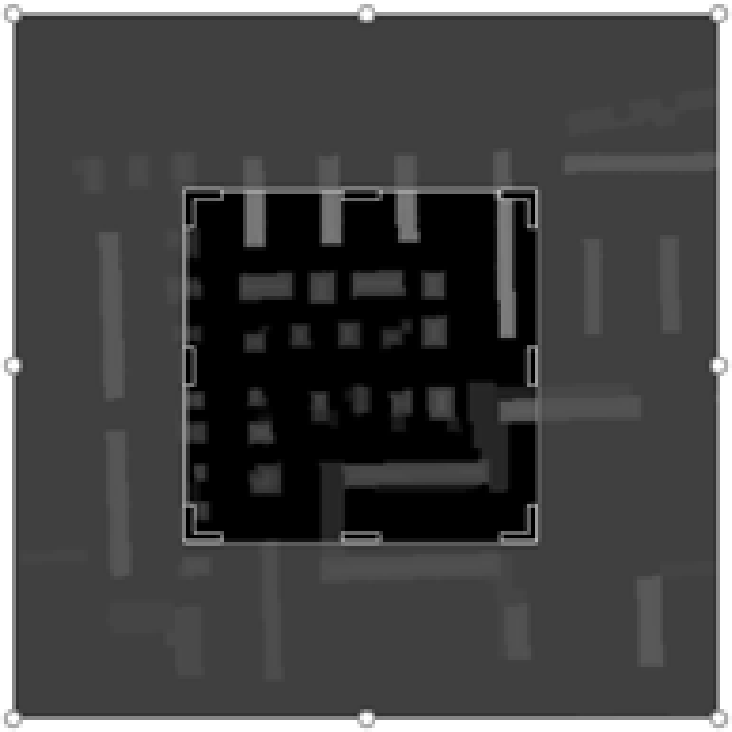


Figure 11.11: Example of cropping on an image.

However, it is possible to recycle this idea of cropping in some parts of the architecture to our benefit. The idea is to use an input image twice as big as the output. Thus, we keep a constant R_{mpx} with $px^X = 2 \times px^Y$. The coefficient of two was chosen because it is easier to make good crop since the architecture dimensions are always even numbers. Illustration of the images used in this case are presented on Figure 11.12.

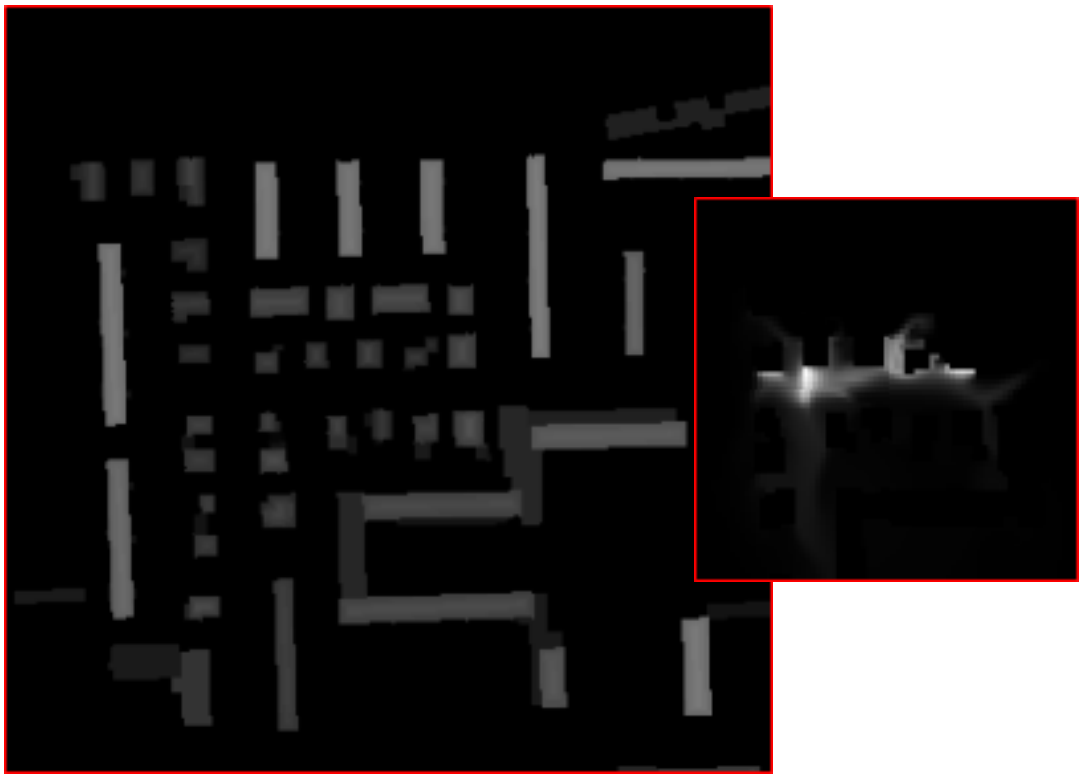


Figure 11.12: Example of images for the architecture with R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144, on the left an input image and on the right an output image.

The cropping can be performed at different places in the architecture. Either before the upsampling at the bottleneck and on the skip connections of the architectures. This allows the architecture to work on the context of the wider area while the localization part only works on the area of interest in the decoder part of the architecture. The architecture dimensions are presented on Figure 11.13. Or, at the end of the architecture which allows the architecture to work with the full context of the whole area while calculating the loss function just on the area of interest for which there are data for the training. The architecture dimensions are presented on Figure 11.14.

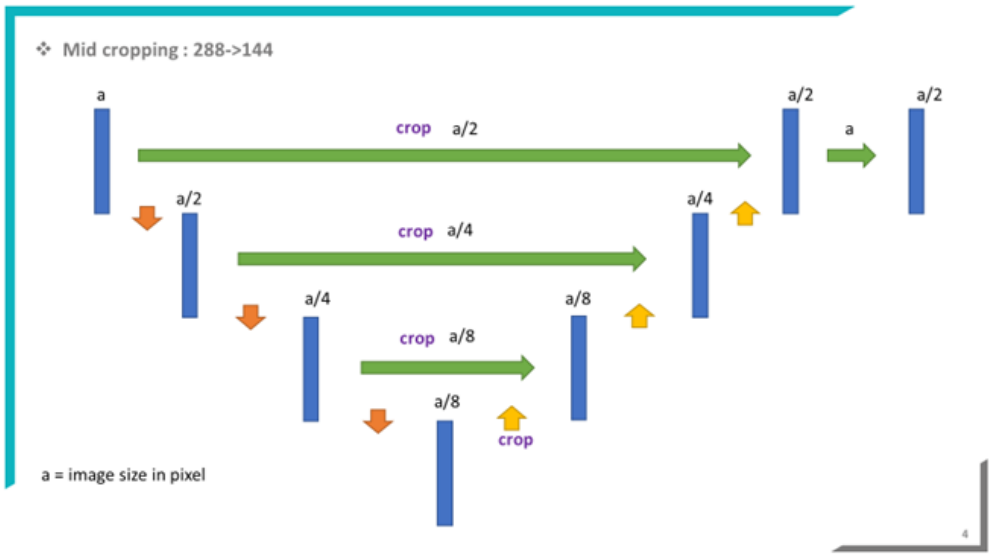


Figure 11.13: Diagram of the V_2 architecture with R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144 and a crop in the middle of architecture and skip connection.

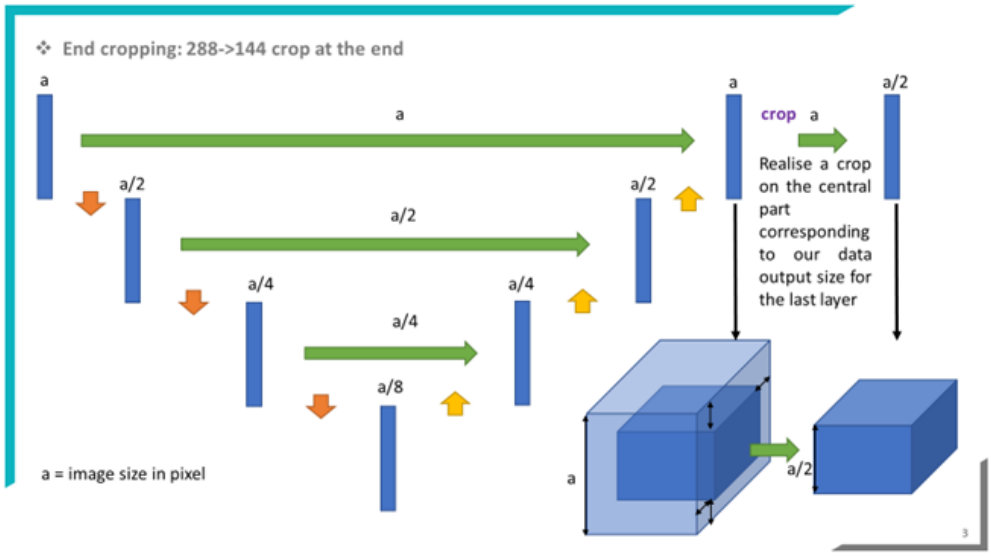


Figure 11.14: Diagram of the V_3 architecture with R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144 and a crop before the output image.

Before comparing the several variants to see if there is any meaningful results with the methodology provided in (40), the padding variant architectures will be described.

11.2.3 Architecture without padding

Study on dimension throughout the architecture without padding As previously said, the original U-net article uses the natural dimension reduction of convolutional filter to have a bigger image at the input than output. However, this strategy imposes a specific pixel size for both input and output as it will be shown below.

With padding, the dimensions of the tensors inside the architecture only changed either at the pooling stage that divided the dimensions on (x,y) by 2 compared to the previous layer or at the upsampling stage that multiplied by two the dimensions on (x,y) compared to the previous layer.

Thus, the dimension between layers could be computed easily at each layer as:

$$dim_{down} = \frac{a}{2} \tag{11.1}$$

$$dim_{up} = 2a \tag{11.2}$$

However, without pooling, each convolutional filter reduces the dimension by two, the previous equations thus become with a kernel for convolution of (3x3):

$$dim_{down} = \frac{a - 2 * nb_{conv}}{2} \tag{11.3}$$

$$dim_{up} = 2(a - 2 * nb_{conv}) \tag{11.4}$$

Now two problems emerge when no padding is use:

- the symmetry between the encoding and decoding part is broken and tensors that have not the same shape from these two parts must still correspond,
- the pooling divide by two the dimensions thus requiring that the dimensions at the end of a layer to be even.

To solve the first issue, a cropping can be used, but the cropping needs to be symmetric. For example, 64 can be cropped into 56 symmetrically since 64 = 4 + 56 + 4, however it is not possible with 63 = 5 + 56 + 4 which breaks the symmetry.

For the second issue, the image input dimensions must be chosen carefully to verify that every dimension at the end of a layer is even. The choice in the original Unet paper of 572x572 pixel for the input and 388x388 pixel for the output is not a choice at random even if it is not explained in the paper. It is the only input image size dimensions that satisfy the two-above requirement. The Unet has two convolutional filter at each layer. With the

previous equations each layer dimensions can be computed as shown in the diagram on Figure 11.15 from the Unet original paper.

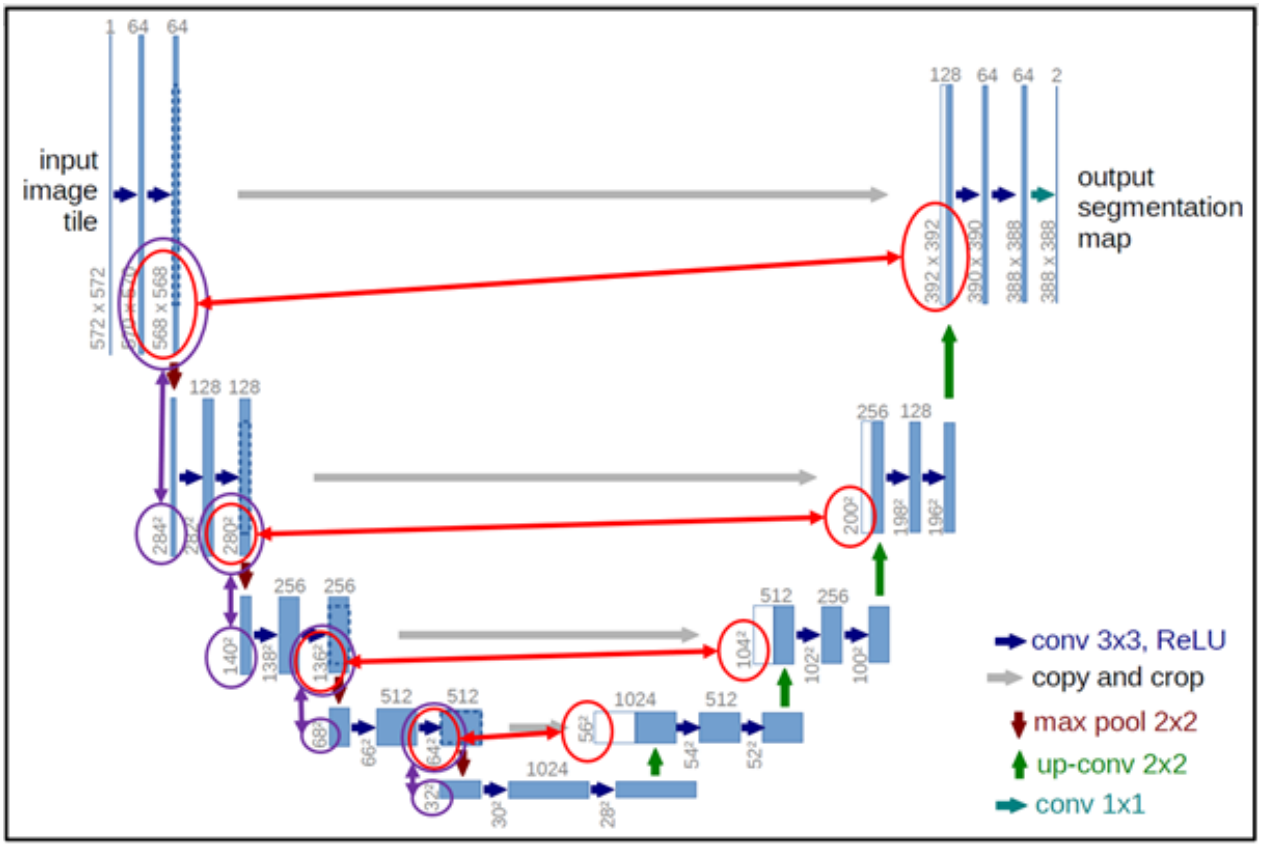


Figure 11.15: U-net architecture from original paper (132)

The areas circle in red and linked together must have matching size respecting the symmetry to crop and the purple circle must be even. Can this be also applied to the multiResUnet?

Architecture V_4: multiResUnet without padding The multiResUnet architecture contrary to the Unet does not 2 convolutional filter per layer but 3. The dimensions of each layer are then not equal to the one of the U-net and this arises a question.

Is there with 3 convolutional per layer a number that match both previous requirements, the symmetry for the crop and the even number for the pooling at each layer? Also, ideally, to use the same data as previously, another question arises, Is there a number of input dimension that satisfy px^X=2*px^Y?

Unfortunately, there is no number that satisfy the first question on the symmetry for a multiResUnet architecture of 4 layers. It is necessary to break the symmetry that may reduce the performance of the architecture. Now which is the best way to break this symmetry? It

is better to break the symmetry for high dimensions than for low ones. Indeed, for example a cropping on 285 -¿ 284 has less impact proportionally on the symmetry than a cropping on 15-¿14. The best that can be expected now is it to have a break in the symmetry at the beginning of the architecture.

For the second question, there is a number that satisfy this requirement. With an input image of 556x556 pixels the output image is 278x278 pixels. The dimensions of each layers are noted below with the first number in the parenthesis the input dimension of the layer and the second number the output dimension of the layer.

encoder: $(556, 550) \Rightarrow (275, 269) \Rightarrow (134, 128) \Rightarrow (64, 58)$

Bottleneck: (29,23)

decoder: $(46, 40) \Rightarrow (80, 74) \Rightarrow (148, 142) \Rightarrow (284, 278)$

To compare the result of these architectures with the one with padding it would require the same training and testing conditions. However, due to the large input size, the graphical card used was not able to load as big batch as before. So to compare the no padding variant with the padding variant, a new run with padding has been made using the architecture that gave the best results before, the middle cropping one, with an input image of 576(=288*2) and output of 288. This version will be noted V_2'. The version without padding will be noted $V4$.

11.2.4 Comparison of the variants architectures and conclusion

As a reminder:

- V_0 : R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=px^Y=144
- V_1 : R^X_{mpx}=1.0417 R^Y_{mpx}=0.5208 and px^X=288=px^Y=288
- V_2 : R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144 with cropping on the resPath and middle of the architecture
- V_3 : R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=288, px^Y=144 with cropping at the last layer
- V_2' : R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=572, px^Y=288 with cropping on the resPath and middle of the architecture
- V_4 : R^X_{mpx}=R^Y_{mpx}=1.0417 and px^X=556, px^Y=278 without padding

Variant	$FAC2$	$NSME$	FB	R	mae_{rel}	J_{3D}	$ssmi$
V_0	0.78	4.21	0.36	0.76	0.9	0.48	0.78
V_1	0.73	6.1	0.37	0.62	0.84	0.39	0.73
V_2	0.81	4.0	0.32	0.77	0.73	0.50	0.79
V_3	0.81	4.1	0.35	0.78	0.76	0.49	0.78
V_4	0.78	4.1	0.31	0.76	0.76	0.48	0.78
V_2'	0.78	3.6	0.33	0.77	0.78	0.49	0.79

Table 11.1: Results for all tested architecture variants

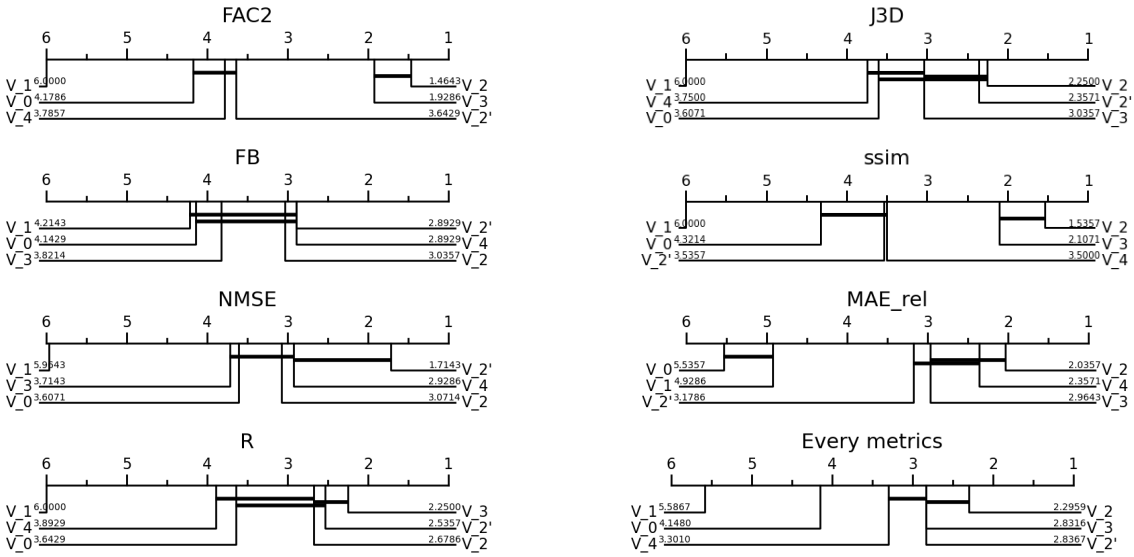


Figure 11.16: Comparison of the different architectures for each criteria

All the variants that have the input representing a bigger area than the output area gives better results than the original version except for the V_1. Since the information at the input is easy to get (height of building and distance from pollutant), this solution for the architecture will be kept as the best ones. The V_1 poor results can be easily explained by the fact that each pixel does not represent the same real meter length at the beginning of the architecture and at the end. Therefore, the skip connection that map the input to the output loses its consistency. Between the padding and no padding version, the padding version slightly outperform the no padding version. Furthermore, the padding version does not impose a constraint on the input size images that forces small batch comparatively. In overall, the V_2 seems to give the best results and this variant will be kept.

11.3 Optimising the multiResUnet architecture

11.3.1 Attention layers

Concept of Attention Gate What is meant by attention? According to the French dictionary Larousse, attention is the "activity or state by which a subject increases his efficiency with regard to certain psychological contents (perceptive, intellectual, mnemonic, etc.), most often by selecting certain parts or aspects and inhibiting or neglecting the others". For example, for a human, on a photo of a birthday party, if the matter was to count the number of people, our eyes would naturally focus on the head or body of the people while ignoring the table or the cake. Another example with the same photo, if the aim was to find the number of red T-shirt in the photo, the human eye and brain would naturally focus on the torso of people while ignoring the rest.

Attention layers was first introduced to improve the performance of Deep Learning Network by helping the layers to focus more effectively on particular words of a sentence in (13) that are relevant for the matter at hand. The concept has been successfully translated to other fields such as imagery with a Unet variant as in (104). Inspired from it, attention gate were implemented to the multiResUnet decoding part. What is an attention gate? An attention gate is described through the following chart in the original paper:

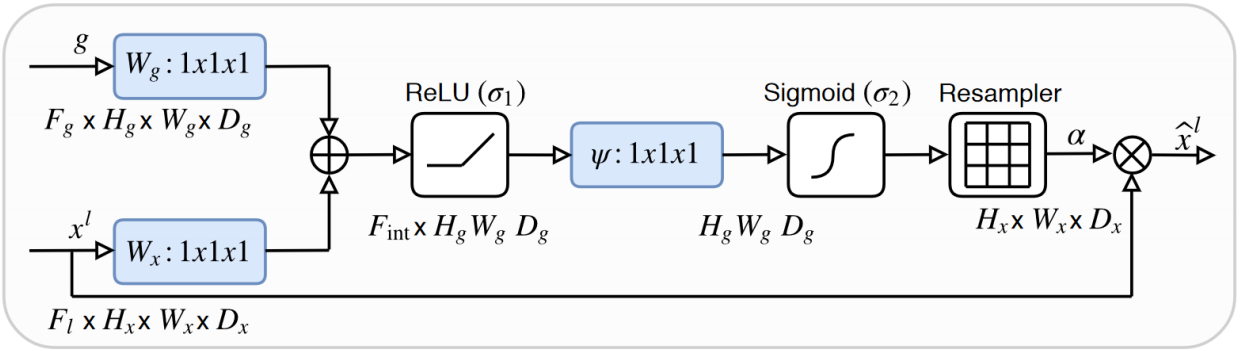


Figure 11.17: Attention gate schema from (104). Where H is the height, W the width, D the depth and F the filters.

The vector g corresponds to the previous bottom layer in the decoding part and x^l the vector from the skip connection. The vector g should have a better feature representation since it comes from a deeper part of the network. Thus, the idea is to use this better feature representation to highlight the area that are more relevant to the network to map the spatial information from the skip connection and thus help it focus on the relevant areas. This process creates a new vector with supposedly the more important features for the task at hand highlighted.

Results of attention layers Unfortunately, the results are slightly worse using this technique. Thus it is will not be kept for the later architecture.

11.3.2 Tuning hyperparameter

Concept of tuning hyperparameters Deep Learning architecture have two kinds of parameters:

- The trainable parameters are parameters that can be trained through gradient methods such as the weight and biases of the architecture.
- the hyperparameter that can not be trained during training with gradient method and are set before the training of the architecture such as the depth of the network, the number of filters, the activation functions, loss and so on. Furthermore, those parameters are inter dependent of other hyperparameters, for instance a loss can have poor result given an activation function but great result when used with another.

Hyperparameters can have a huge impact on the final architecture results. Therefore it is necessary to test several set of hyperparameters to try to find the best set. Several approaches exist to solve this issue and are listed in the diagram shown on Figure 11.18.

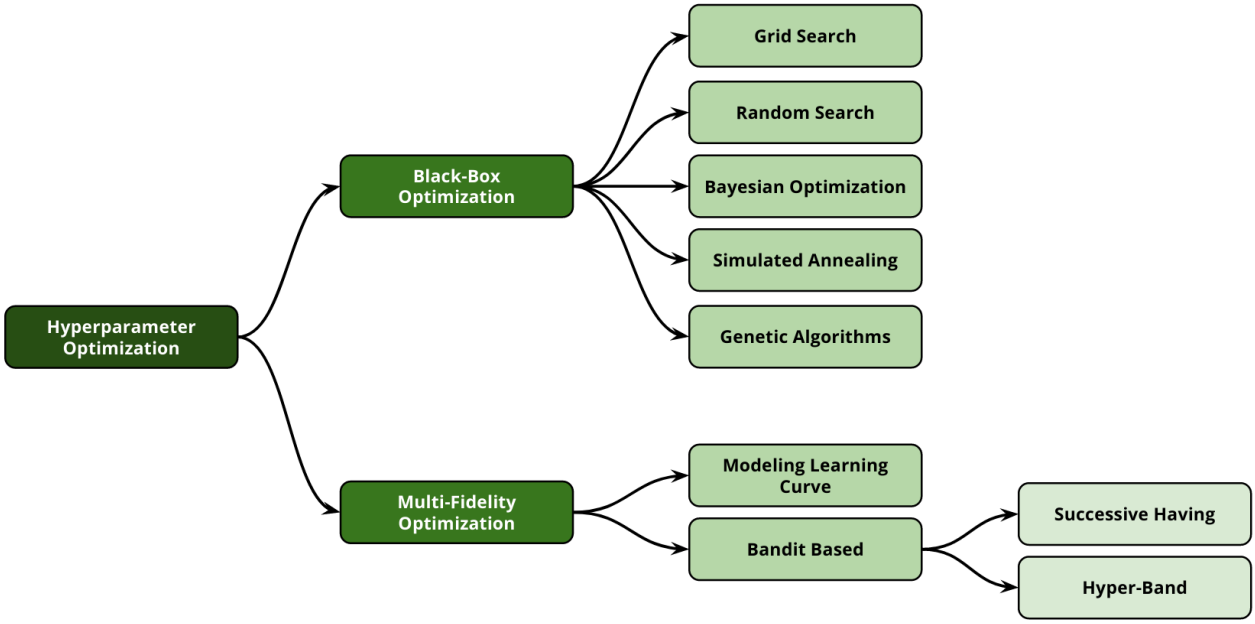


Figure 11.18: A Taxonomy for the Hyper-parameter Optimisation Techniques from (145)

A popular framework to tune hyperparameter using Keras is the library Keras tuner. This library allows to set an hypermodel with three options for the hyperparameters testing.

– the random search tuner which tries randomly several set of hyperparameters.

– the HyperBand-based algorithm originally from (78) which aims at speed up the random search method by early stopping the training.

– Bayesian Optimization tuning with Gaussian process which uses previous iterations of sets to try the next ones with the most potential using a probabilistic approach.

For the tuning the hyperband algorithm will be used.

Dropouts theory A popular regularization method to tackle overfitting is the dropout technique. This technique consists to drop randomly some neurons in the model accordingly to a threshold given by the user and training the reduced version of the network. The benefit of this approach is that it will reduce the co dependence of the neurons. Indeed, with complex model used in deep learning, it happens that several neurons carry similar information and thus overfit the data. The following diagram explains how dropout works on fully connected layers.

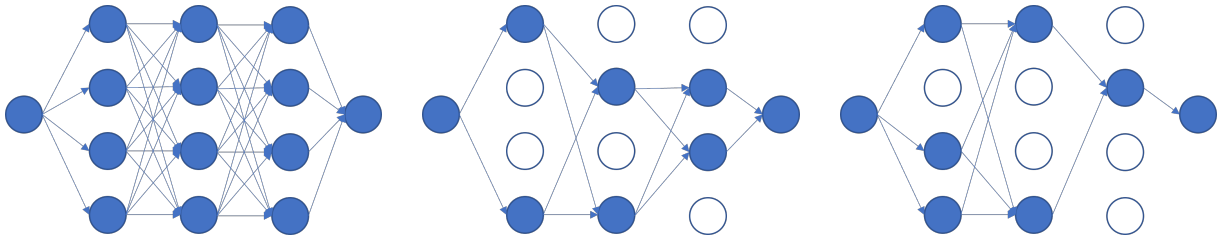


Figure 11.19: :Example of a dropout on a fully connected network. The blue filled circle means the neuron is active while white one means they are inactive. The right image is the full total network, the middle one the network at a training epoch N and the left one the network at a training epoch M

.

Hyperparameters tested and results The same dataset as in 11.2 is used. The architecture that is tuned is the $V2$ version from 11.2, which is the multiResUnet with a crop on the resPath and at the bottleneck with an input twice as big as the output. The tested hyperparameters are:

- The depth of the network, meaning the number of pooling before reaching the bottleneck. The value tested are from 2 to 6 with a step of 1. Due to this parameter that divides the input dimensions up to 2^6, the input image here is no longer (288,288) but (256,256) to not break the symmetry.

- The number of min filters. The value tested are from 4 to 32 with a step of 8.

- The percentage of dropouts. The value tested are from 0 to 0.5 with a step of 0.1.
- The activation function inside the architecture. The tested activation function are mish, elu, relu and tanh.
- The activation function at the last layer of the architecture. The activation function tested are mish, elu, relu and sigmoid,
- The optimizer choice. The optimizer tested are "adam", "adamax", "sgd", "RMSprop".

Among these set of parameters, the best set according to the hyperband algorithm is a depth of 4, a min filter of 12, a dropout percentage of 0.2, relu as the activation function inside the architecture and at the last layer and adamax optimizer. This set of hyperparameters is not far from the one already used prior to it. These changes do not improve much the architecture going from a 0.5 J3D to a 0.51.

11.4 3D MultiResUnet

11.4.1 MultiResUnet 3D architecture

The multiResUnet up until now have been used on a 2D plane at an height of 1.5m. This height was chosen because it gives good insight on the exposure of dwellers. It could be argued that the 2D multiResUnet does a bit more than just 2D since the input image on the height of building gives information on the height of the building. Nevertheless, the convolutional architectures can also deal with 3D information by using 3D convolutional filters instead of 2D. This change of paradigm could potentially allow the deep learning architecture to determine the pollutant exposure not just at a given height but at several height given its training. It may even improve the result on the 2D plane since the architecture will deal with it as the CFD, in a 3D manner.

The input instead of being an image for the height and distance to the pollutant source will becomes several images describing at each chosen altitude the height of building and distance to the pollutant. To use the architecture in a 3D way, it is possible to define plans at different height that will be used for the z axis. For example, on the Figure 11.20, there are two buildings, 30m tall and 15m tall, and 3 plans at height of 0, 7.5m and 15m.

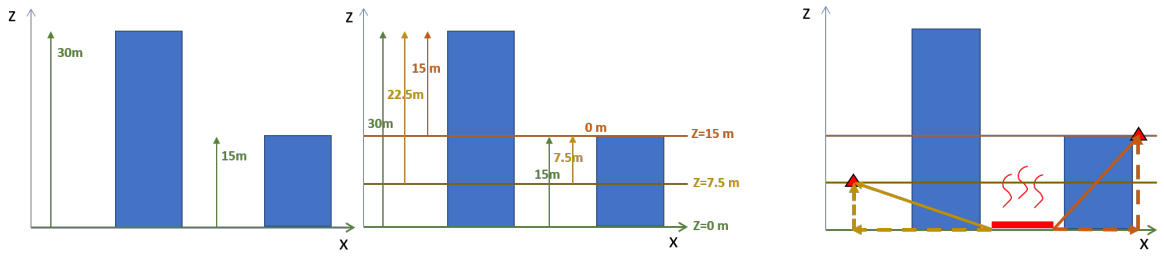
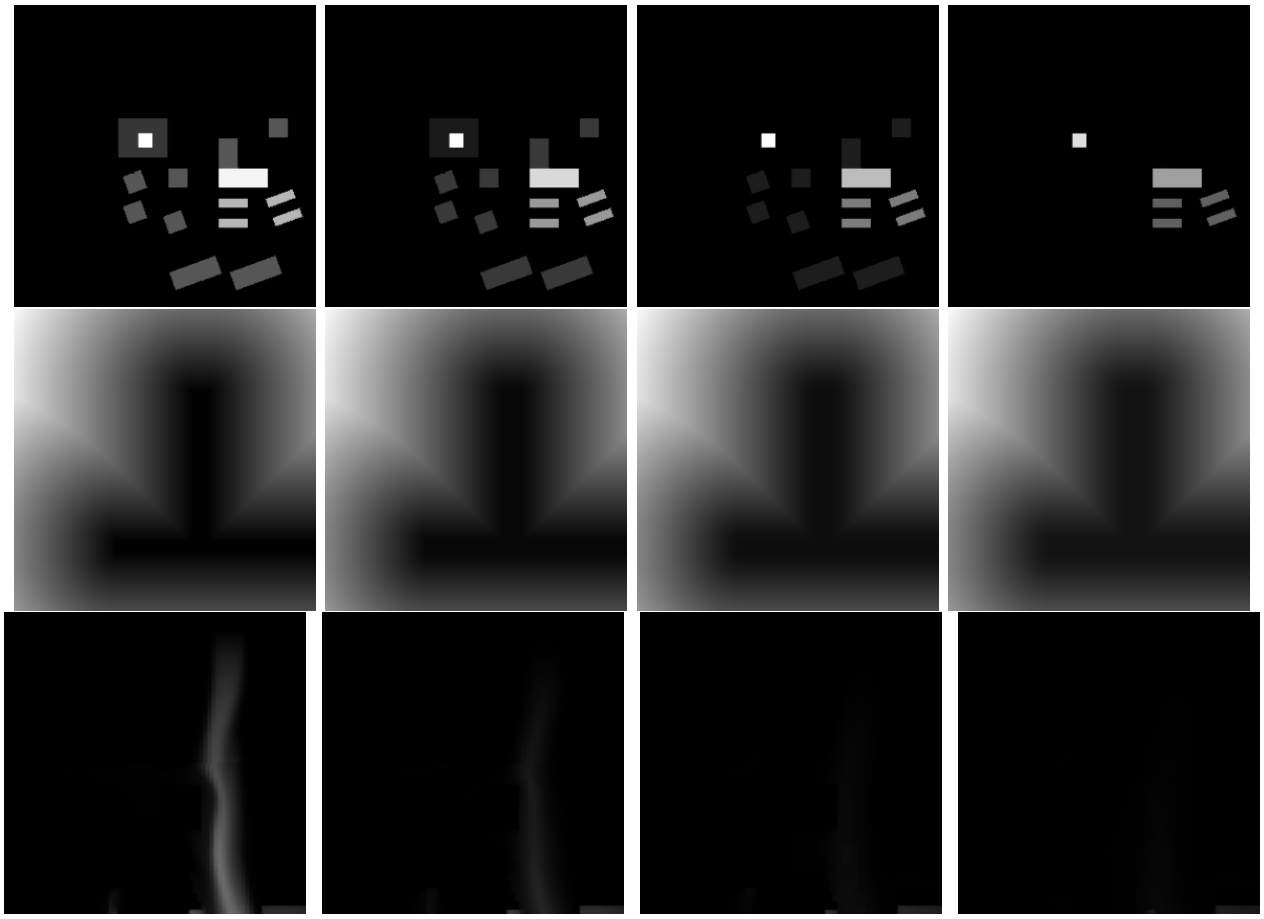


Figure 11.20: Example

For the 3D architecture, for a first try, several plan, from 1.5m to 15m every 1.5m will be used. The figure below shows an example of the input and output images for the 3D case.



(a) height of 1.5m (b) height of 6m (c) height of 10.5m (d) height of 15m

Figure 11.21: Examples of the input and output for the 3D multiResUnet, top the height of buildings, middle the distance from the pollutant source, bottom the pollution dispersion

Computing wise, the 3D architecture requires way more resources. Indeed, it will evolve linearly to the number of z plans when compared with its 2D version. The architecture

imposes a constrain on the minimal number of z plans required. Indeed, $16(=2^4)$ is the minimum dimension required at the input of architecture to do 3D convolutional computation using the multiResUnet with a depth of 4. It divides the input tensors (x,y,z) dimensions 2^4 times. Thus the 3D multiResUnet architecture will be at least 16 times bigger than the 2D version if all other parameters are kept constant. Nevertheless, due to the limitation of the graphic card at hand, it is necessary to change some of the parameter to be able to have a functional architecture. It would be also for the better to keep a R_{mpx} ratio as constant as possible between the x,y and the z axis.

Using the minimal amount of 16 z plans for 13.5m it gives a R_{mpx} of 0.84375. Since the data covers 300, 300*0.84375=252. However 252 is can not be divided by 2^4 so 256 will be retained. The input of architecture will thus be 256x256x16 and output 128x128x16 as can be seen on Figure 11.22.

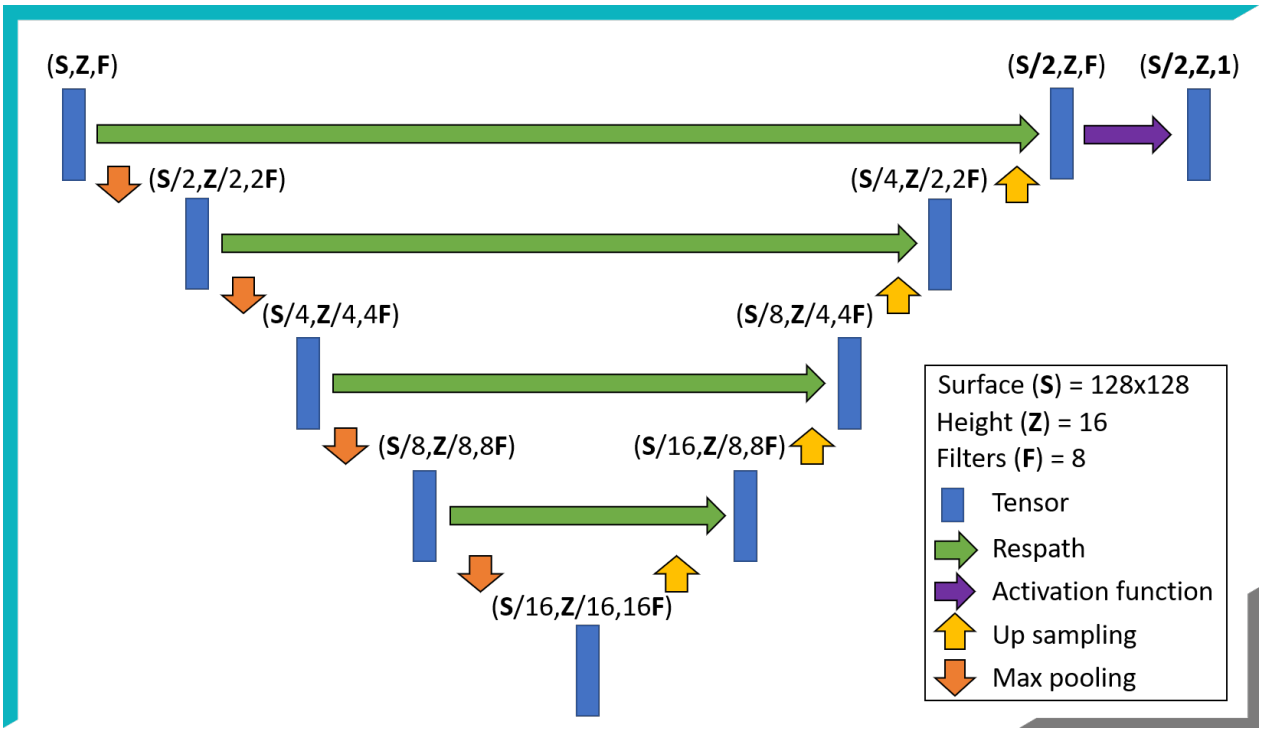


Figure 11.22: Diagram of the multiResUnet 3D architecture.

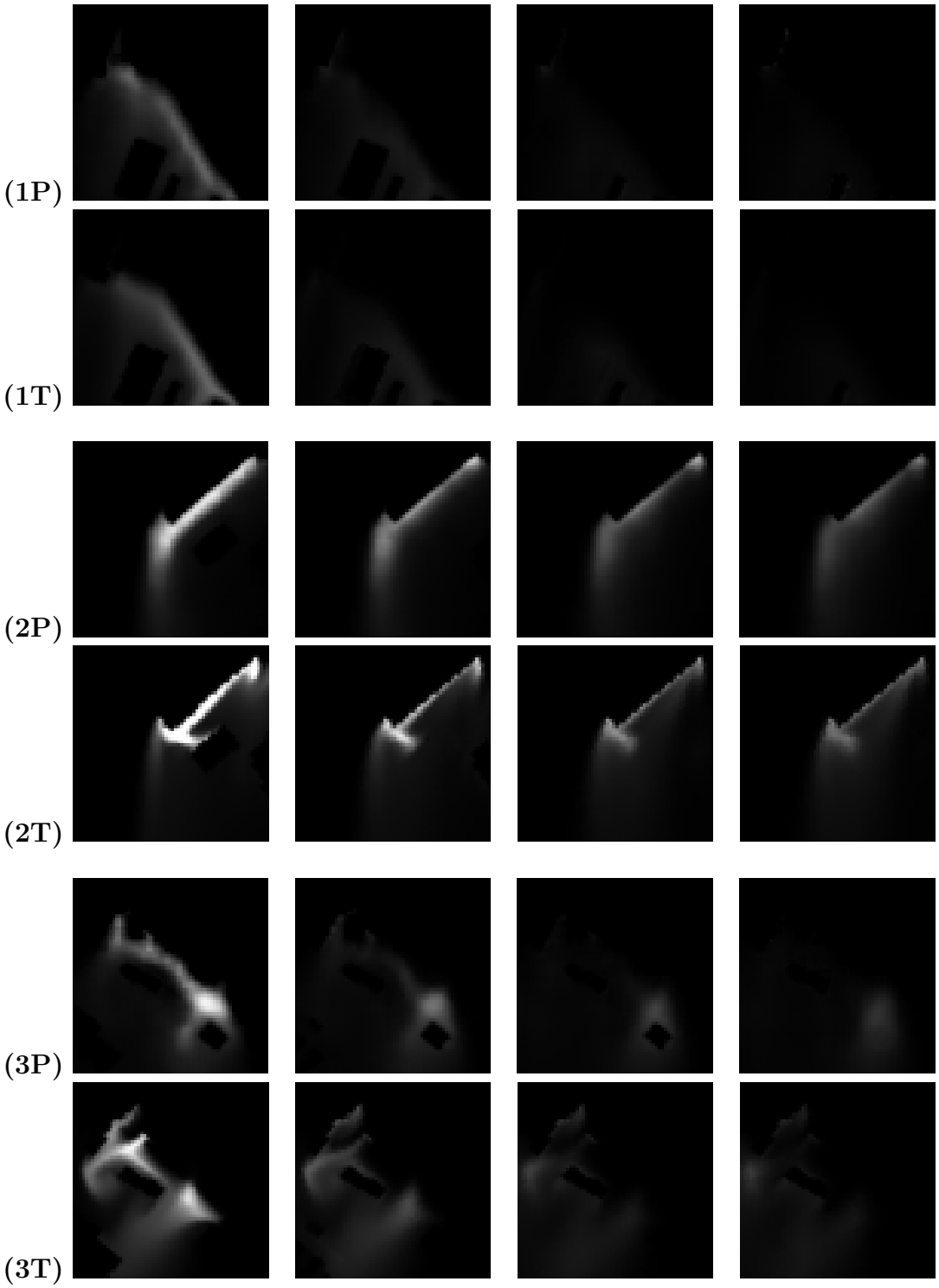
The data from the CFD have been retrieved between 1.5 and 15m with a step of 1.5m. This makes a total of 10 z plans. Nevertheless, depending of the architecture this number can vary, for example here, it is required to have 16 evenly spanned z plans. In order to do that, an interpolation is made using the zoom function from the scipy.ndimage python library.

Another important matter is that the multiResUnet 3D needs to be able to see the decreasing concentration with the altitude. The strategy used in the prior chapters of having each image scaled from the maximum level of pollution is not possible. Thus to compare

the result of the 3D architecture, a new multiResUnet 2D need to be trained with the same pollution scale. A scale of 4.07 for the pollution will be use. The scale in itself is not relevant, as long as both the architectures have the same to be comparable.

11.4.2 3D MultiResUnet results

The multiResUnet 3D and 2D have been trained and tested on the same dataset presented in 11.2. The results for the best J3D score, median score and worst score on the test dataset are presented in Figure 11.23.



(a) height of 1.5m (b) height of 4.9m (c) height of 8.3m (d) height of 11.6m

Figure 11.23: Examples of predictions from the multiResUnet (denoted by P for prediction) against CFD results (denoted by T for truth) for 3 different configuration. The top configuration achievied the best results while the bottom one achieved the worst

model	$FAC2$	$NMSE$	FB	R	NAE	MAE_{rel}	J_{3D}	$ssim$
3D	0.82	2.28	0.15	0.80	0.57	0.58	0.56	0.75
2D	0.88	3.35	0.15	0.78	0.59	0.60	0.55	0.85

Table 11.2: Results of the 2D and 3D architecture on several metrics at altitude 1.5m

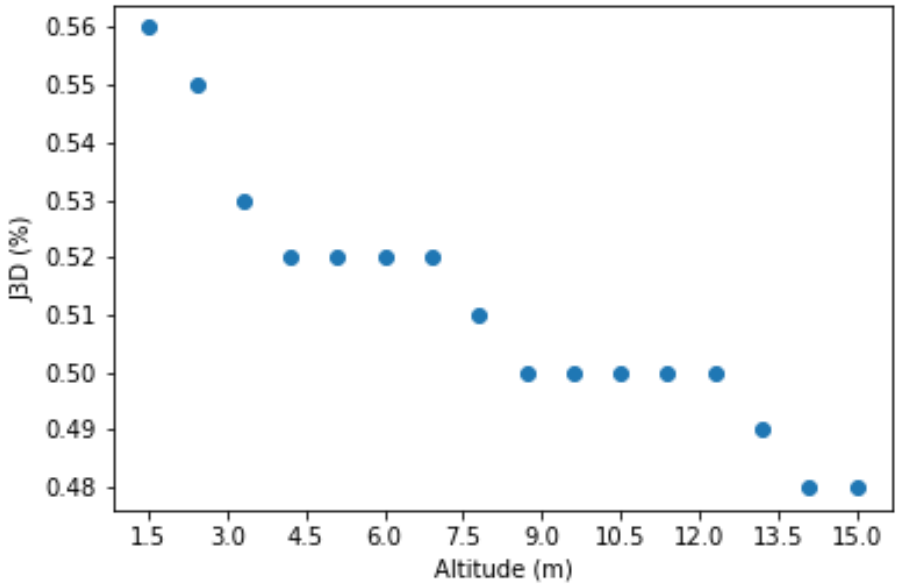


Figure 11.24: Mean results of J3D per height (altitude).

On average, the multiResUnet 3D has a J3D of 0.51. Nevertheless, it can be noted that the predictions are better on the lowest height (1.5m) and that the quality of the J3D decreases with the altitude as it can be seen on the graph 11.24. This is probably due to the fact that the concentration decreases with the altitude making these altitudes less impacting on the loss score. It may be solved by computing the loss for each altitude and then averaging it, maybe with a weight depending on the altitude.

When compared to the 2D version, they both perform similarly for the same height (1.5m) as it can be seen on the table 11.2.

11.5 Conclusion

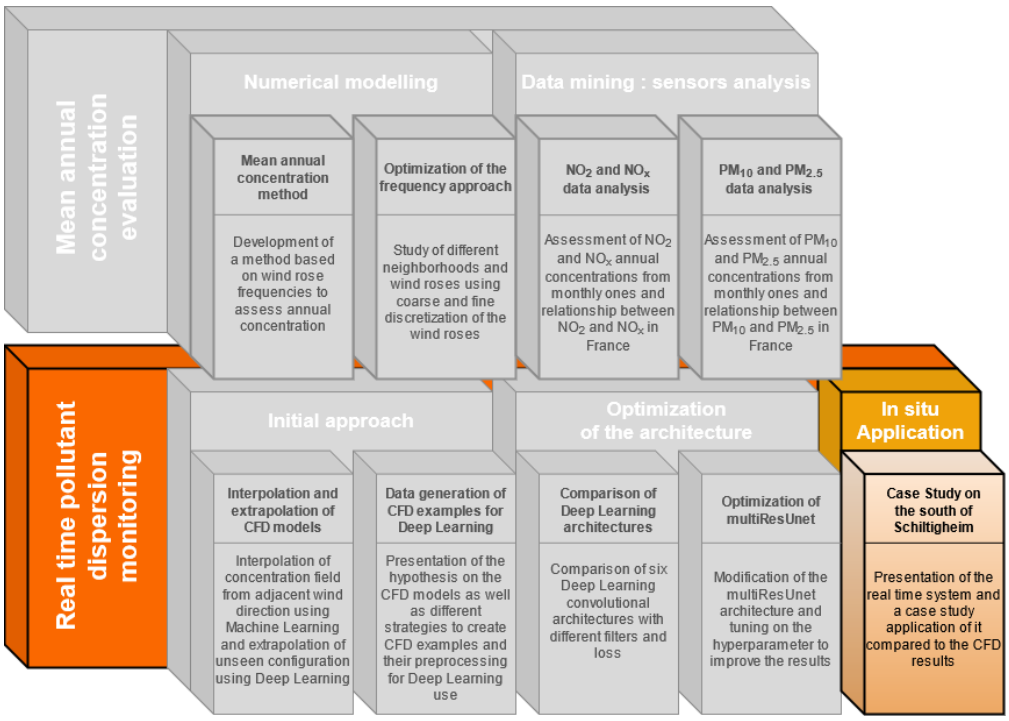
The architecture that worked the best is the multiResUnet according to the previous chapter. Several modifications of the architecture were carried out. It was found that giving an input size of an area of 300x300 m^2 improved the result over an area of 150x150 m^2 for the same

output of 150x150 $m2$. To take into account this bigger area two ways were tested, with and without padding. They both gave similar results with a slight advantage for the no padding strategy. Furthermore the no padding strategy has the benefit to not impose an input size pixel wise. Another change was to add attention gate unit to the deep neural network, nonetheless the results worsen.

The hyperparameters of the architecture were tuned using Keras tuner with the hyperband algorithm and dropout were added to the architecture. The best set of parameters are a depth of 4, a min filter of 12, a dropout percentage of 0.2, "relu" as the activation function in the architecture and at the last layer and adamax optimizer. The results did not improve much nonetheless. The last test conducted was to use the 3D version of the multiResUnet. The multiResUnet 3D needs more parameters since the tensor have one more dimension. So it needed to be scaled down to run on the graphic card at hand. However, it managed to reach similar results at the same altitude.

Chapter 12

Case study to estimate urban pollution in real time



This chapter has been presented in the international conference *Upper-Rhine Artificial Intelligence Symposium* under the title "Deep Learning associated with Computational Fluid Dynamics to predict pollution concentration fields in urban areas " (57).

12.1 Introduction

Several changes and optimization have been done to the architecture to improve its accuracy. The size of the input is twice as big as the size of the output and the hyperparameters are tuned. Now it is necessary to see how well the optimized version architecture can perform on a real test case and how it can be used in a wider system to be able to make real time pollutant dispersion predictions.

12.2 Real Time dispersion monitoring system

The Deep Learning model has proved capable of determining the pollutant from local road sources in a matter of minutes on a GTX1080Ti against several days for the CFD on 96 CPU. Thus it is possible to use the Deep Learning model to determine the pollution in real time from road traffic. The Deep Learning model is able to predict pollutant dispersion in neutral conditions on squares of $100x100m^2$. To use it on a neighborhood, each road are covered with areas that will be determined by the Deep Learning model. To improve the quality of the predictions and continuity of the results between different areas, the predictions are made on overlapping squares (every 20m) and then averaged as it can be seen on Figure 12.1.

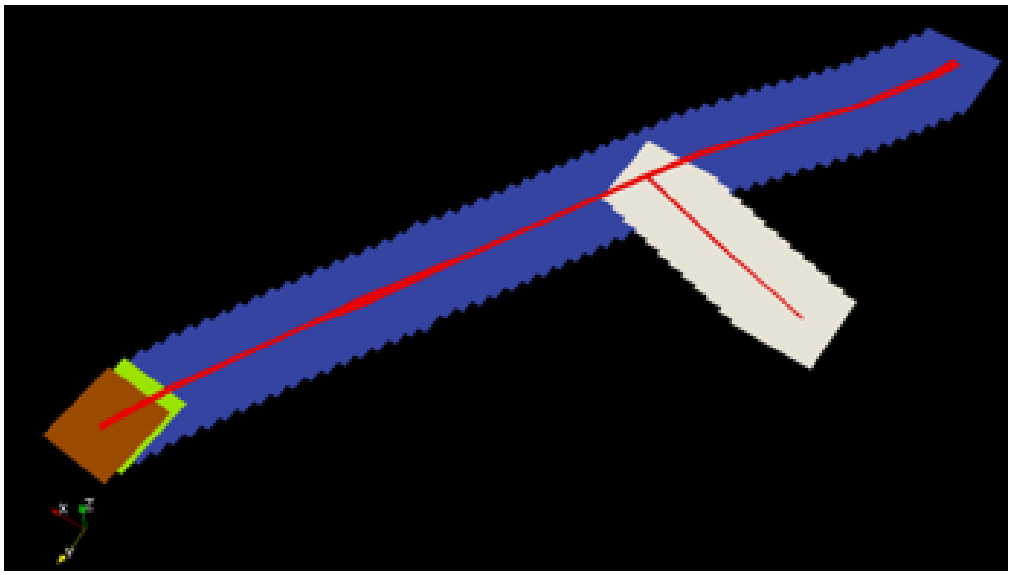


Figure 12.1: Discretization of two roads.

The second step after the computation is done is to turn this concentration field into the real estimated pollution. In order to do that, three steps are done:

– First the emission from the traffic data is computed and used to adjust the value of the dispersion field from the Deep Learning model.

– The second step requires a background sensor that is used to take into account background pollution from non local sources in the city.

– The third step is to add eventually punctual pollution sources such as factories or heating plant.

The concentration at a point x,y is thus :

$$C(x,y,t) = C^b(t) + C^l(x,y,t) = C^b(t) + \sum_{i=0}^{N_r} C_i^{DL}(x,y,t) * E_i^r(t) + \sum_{j=0}^{N_p} C_j^M(x,y,t) * E_j^p(t) \quad (12.1)$$

with C^b the background pollution,$C(x,y)^l$ the pollution from local sources, $C_i^{DL}(x,y)$ pollution from the road i at the point (x,y) determined by the Deep Learning model, N_r the number of neighboring road impacting (x,y), E_i^r the emission from the road i,N_p the number of neighboring punctual source impacting (x,y), $C_j^M(x,y)$ the pollution determined by the model at point (x,y), E_j^p the emission of the punctual source j.

The system is presented below :

12.3 Case study

12.3.1 Context

– New real estate project near pollutant sources such as heavy traffic roads, plants, or central heating system must study thoroughly air quality in the wanted area. However, these regulations are only applied at some particular timestamps and specific places.

– Sensor monitoring. But reliable sensors are expensive to acquire and maintain. For the entirety of Strasbourg city (around 80km^2), to date only 4 sensors are deployed.

– Simulation of the annual pollution dispersion on the entire city. However, models that allow large area to be simulated may not be adapted for urban areas because of buildings not taken into account.

Among the possible models of the third point, a popular approach in the scientific community is to create airborne pollutant dispersion maps in urban areas is to use Computational Fluid Dynamics (CFD) (119; 140). It allows to accurately consider a lot of different physical phenomena from building impact on the flow to solar radiation or chemical reaction. Indeed, pollutant dispersion concentration field error can reach less than 10% when compared to experimental data (121) and about 30% when compared to real life *in situ* experiments (127).

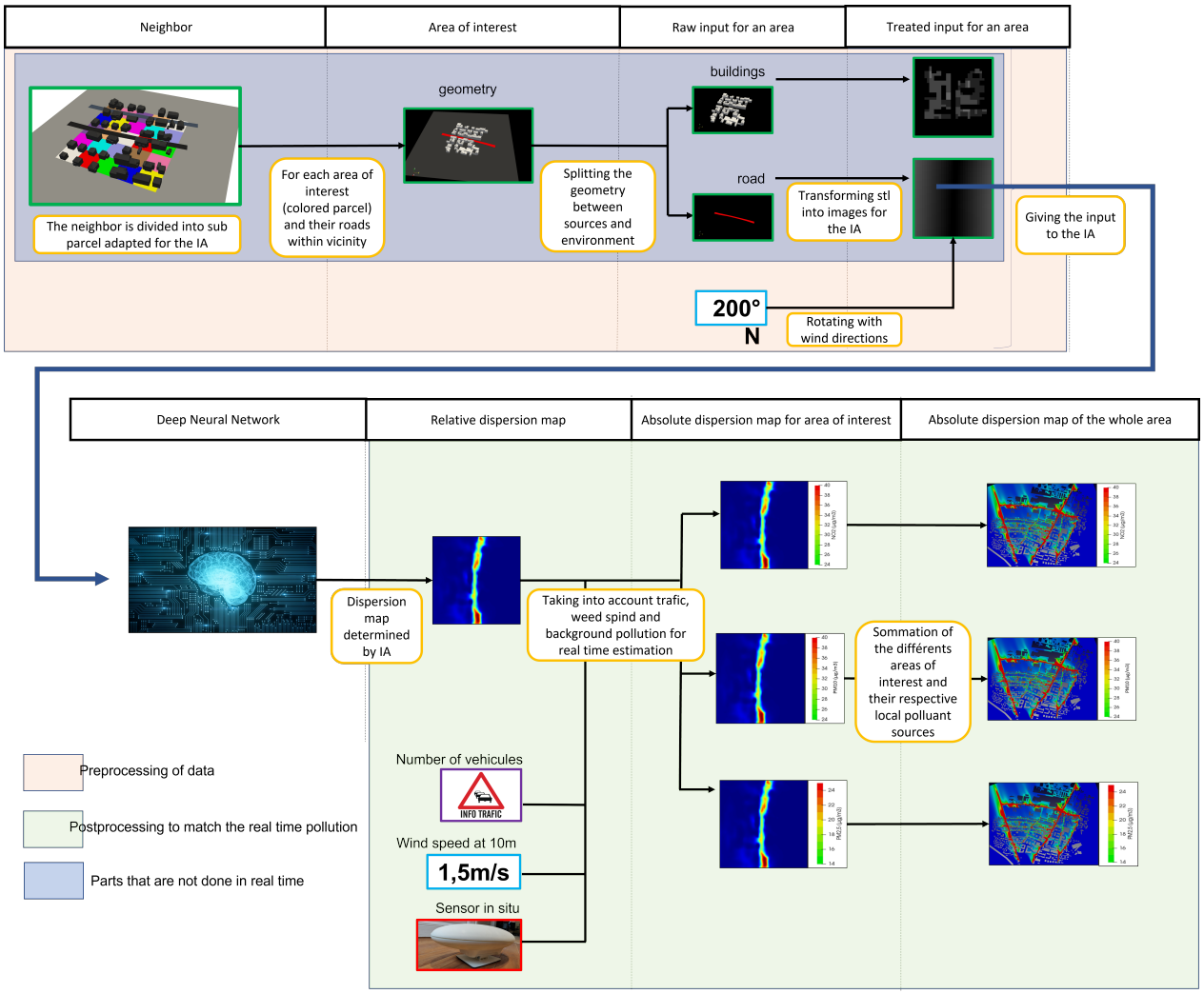


Figure 12.2: Real time dispersion monitoring system.

Nevertheless, the counterbalance of this method is that it is computationally expensive. For instance, to cover 1km^2, the method roughly needs around 30 million cells and can require a week of computation to converge on 96 CPUs. Furthermore, each time the building layout changes, it would require starting new simulations again. CFD is therefore not adapted for real time simulation, despite its great accuracy and detailed description of physical phenomena.

To accelerate the computation, an innovative solution based on Deep Learning was developed. The idea consists in training a neural network with pre-calculated CFD simulations, to create a new air quality model that can determine pollutant dispersion in a matter of minutes over a large area. Indeed, recent advances in Deep Learning for spatial information treatment with convolutional based architectures have proved to be able to solve issues, notably in semantic segmentation that was impossible before. A popular model, the multiResUnet (51), heir

of Unet (133), has proved to be particularly capable at handling spatial information. This model has been trained with about 5,000 examples of CFD results of pollutant dispersion from different urban areas. The input of the model is the 3D shape of the buildings, the wind force and direction, and the position of the roads, considered as the sources of pollution.
This Deep Learning model is then included in a wider system that uses real time meteorological, traffic and sensor data to map the concentration field in real time on an entire urban district.

12.3.2 Material and method

CFD air quality modeling To train the Deep Learning architecture examples of pollutant dispersion were obtained using Computational Fluid Dynamics (CFD). The software to compute the simulation is OpenFoam 5.0 which is an open source software for numerical simulations of different kind such as fluid mechanics or radiation. The approach elected here to solve the air flow is a Reynold Averaged Navier Stokes (RANS) with a k-epsilon renormalization group (RNG) (168) performing unsteady simulation. For the pollutant dispersion a transport equation coupled with the air flow is used.

The boundary conditions for the upper and lateral boundaries are symmetry conditions, the ground as a wall with a rugosity of $z_0 = 0.1m$, the building as a wall condition, the outlet as a freestream, the inlet as a logarithmic wind profile law as proposed by (126).

For the meshing, the guidelines from (42) are respected with the top and lateral boundaries situated at 5H from the closest building including with H the height the highest building. The mesh is insensitive with cells of 0.5m nearest to the buildings. The model, equations and validation have been detailed in previous published paper (123) where the same approach has been described and properly validated.

Deep learning network The Deep Learning network used to learn the CFD is the MultiresUNet from (51). This network is first designed to be applied for segmentation. In this work, it has been converted to solve pollutant dispersion from fluid mechanics. The input are the distance from the pollutant source and the height of the buildings in the area and the output is the pollutant dispersion field. The final result covers an area of $100 \times 100m^2$ by AI predictions as showed in Figure 12.3. The details of the MultiresUNet architecture are presented in Figure 12.4.

The loss function used is a custom loss called J_{3D} and defined as followed:

$$J_{3D} = 1 - \frac{V_{pred} \bigcap V_{true}}{V_{pred} \bigcup V_{true}} \simeq 1 - \frac{min(y_i, \hat{y_i})}{max(y_i, \hat{y_i})} \tag{12.2}$$

where V_{pred} and V_{true} is the volume represented by the grayscale value of respectively the

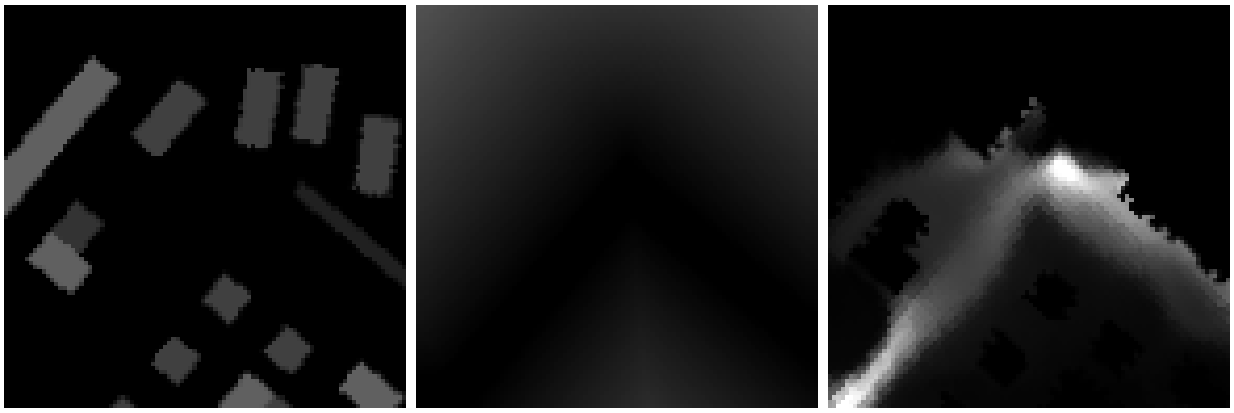


Figure 12.3: Input/output images for the Deep Learning model

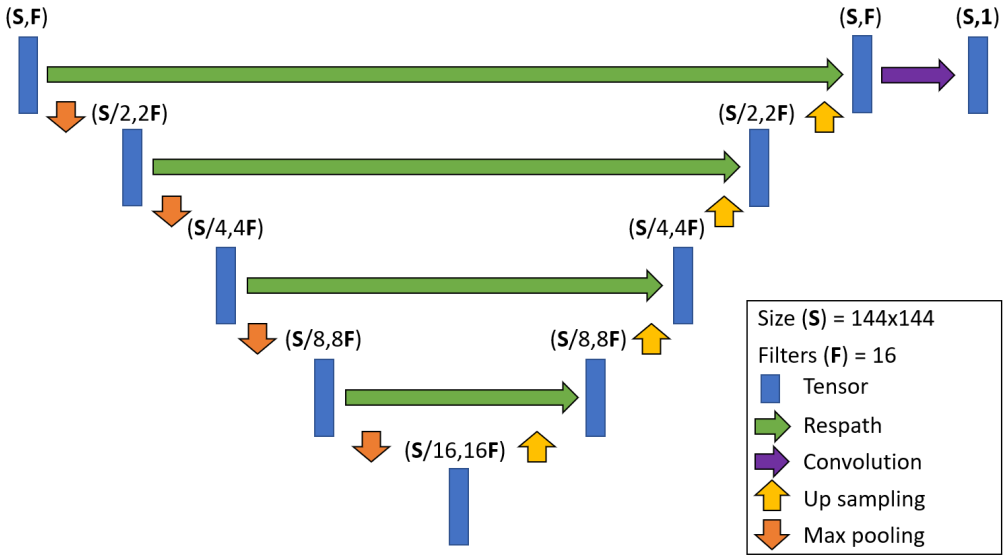


Figure 12.4: Architecture details of the MultiresUNet

ground truth and the predicted result, y_i and $\hat{y}_i$ are respectively the ground truth image and the predict deep learning result.

The dataset for the training and validation are made of around 5,000 examples of different CFD simulations with varying building layouts and pollution sources. 20% are used for the validation and 80% for the training. For the test to check on the AI capability of predicting pollutant dispersion field on unseen neighborhood, it will be compared with a real neighborhood presented in Section 12.3.2 that will be modeled in CFD. The training was made on 25 epochs with a patience of 5 epochs on the validation data.

Case study The site is located in the surrounding of Strasbourg (GPS coordinates: 48.603468, 7.743355). The building layouts of the case study is obtained thanks to the open data of the city of Strasbourg which provide digital model of the whole city (https://data.strasbourg.eu).

For the test case, a real life situation is used, the first of April of 2021 at the traffic peak which happens around 08:30 AM (to have the highest concentration related to road traffic in the area). The wind speed and directions were obtained using the API openWeatherMap with a wind speed of 2m/s and a wind direction 200°N.

Figure 12.5: Map of the Schiltigheim district with the 3 main roads used in this study

There are 27 different roads in the area. The data on traffic were obtained through the open data of the city of Strasbourg for the 4 available roads (https://data.strasbourg.eu):

- Road Bischwiller (part 1): 560 vehicles in 30 min (18.7 veh/min) with a mean velocity of 37.9km/h,
- Road Bischwiller (part 2): 784 vehicles in 30 min (26.1 veh/min) with a mean velocity of 15.5km/h,
- Street Mairie: 488 vehicles in 30 min (16.3 veh/min) with a mean velocity of 17.8km/h,
- Street General de Gaulle: 654 vehicles in 30 min (21.8 veh/min) with a mean velocity of 16.3km/h.

For other roads in the area, traffic information is lacking, thus they have been classified as secondary that will have 30% of the traffic of closest main road and tertiary that will have 5% of the closest main road. Figure 12.5 shows the map of the district of the study, with the three main roads and the secondary and tertiary roads. The choice of 30% and 5% is

arbitrary for the sake of the example since there is no study on this traffic either with sensors or models.

Emissions are calculated based on methods proposed by the European Environment Agency (EEA) in their "EMEP/EEA Air pollutant emission inventory guidebook 2016", Tier 3 method for engine-related NOX, PM10 and PM2.5 emissions (hot and cold emissions); 2017 metropolitan fleet data found in the "OMINEA" databases provided by the Centre Interprofessionnel Technique d'Études de la Pollution Atmosphérique (share of different vehicle types, fuels and EURO standards in France).

The whole neighborhood have been modeled at once with CFD spanning an area of 1 km^2 made of 28 million cells. The buildings as well as the velocity magnitude field at an height of 1.5m is shown on Fig. 12.6.

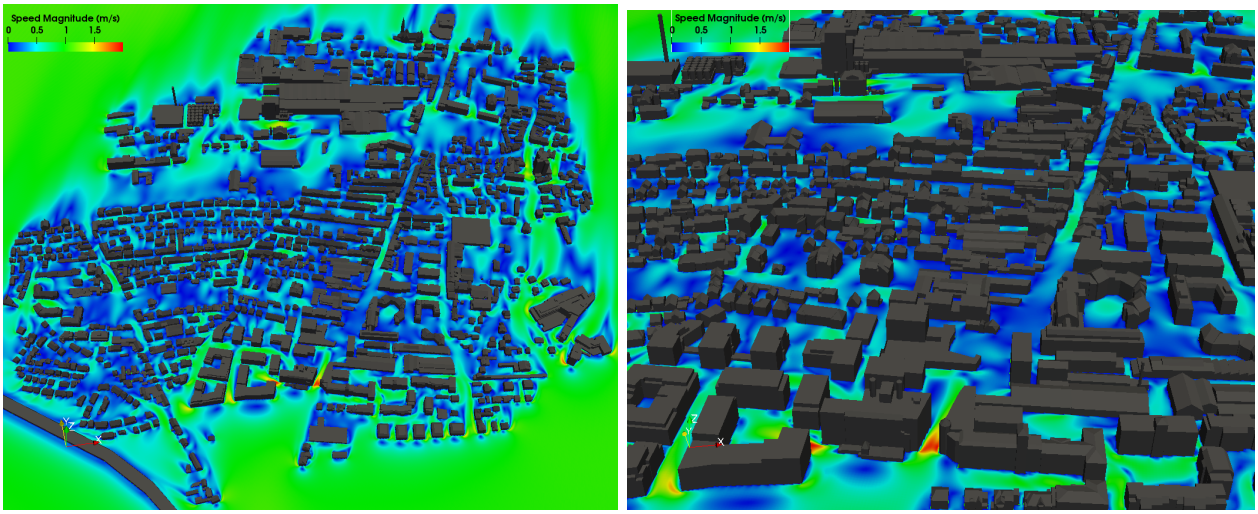


Figure 12.6: Building layouts and flow field at an height of 1.5m

Evaluation Seven metrics will be used, 4 from the air quality domain and three others from the computer vision. The air quality criteria have been chosen according to (27) in which the authors present several metrics with some overlapping since they evaluate the same aspect of the model. They also provide empirical thresholds to consider a model as making good predictions:

– Fraction of predictions within a factor of two of observation, noted $FAC2$, a good model should respect $\simeq > 0.5$,

$$FAC2 = \text{fraction of data that satisfy } 0.5 < \frac{C_{pred}}{C_{ref}} < 2 \tag{12.3}$$

– Normalised Mean Squared Error, noted NMSE, a good model should respect NMSE $\simeq< 1.5$,

$$NMSE = \frac{\overline{(C_{ref} - C_{pred})^2}}{C_{pred} C_{ref}}, \tag{12.4}$$

– Fraction Bias noted FB, $|\text{FB}| < 0.3$,

$$FB = \frac{(\overline{C_{ref}} - \overline{C_{pred}})}{0.5(\overline{C_{pred}} + \overline{C_{ref}})}, \tag{12.5}$$

– Correlation coefficient, noted R (no threshold is given for this parameter),

$$R = \frac{\overline{(C_{ref} - \overline{C_{ref}})(C_{pred} - \overline{C_{pred}})}}{\sigma_{C_{pred}} \sigma_{C_{ref}}}, \tag{12.6}$$

The three other metrics are:

– J_{3D}

$$J_{3D} \simeq \frac{min(C_{ref}, C_{pred})}{max(C_{ref}, C_{pred})} \tag{12.7}$$

– Relative mean absolute error MAE_{rel}

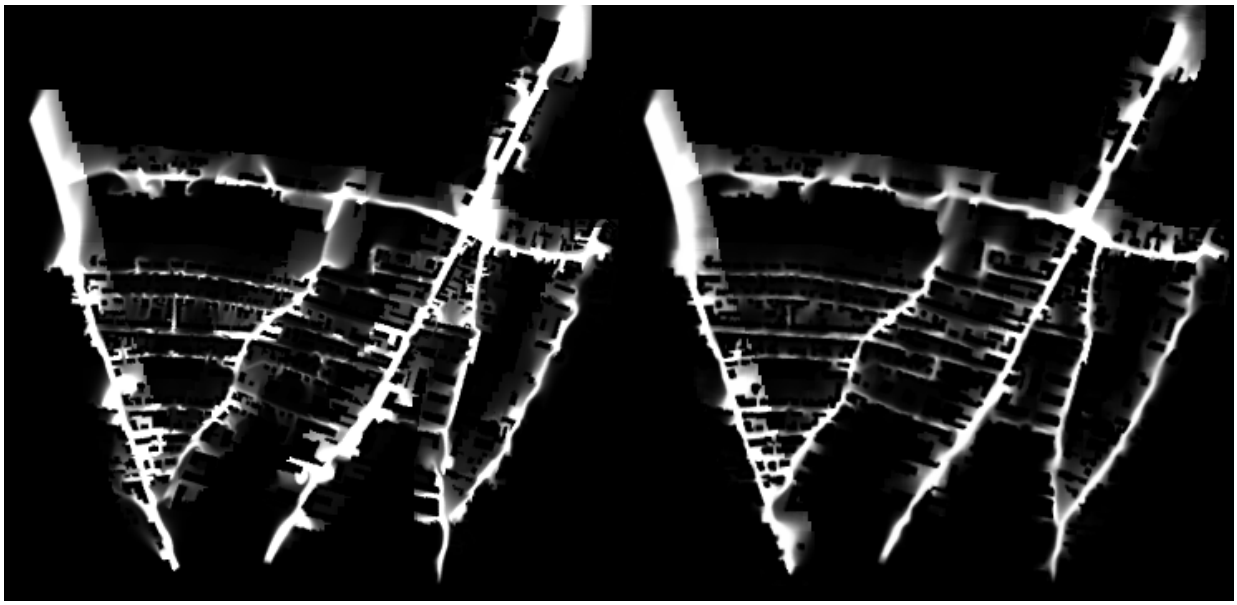
$$MAE_{rel} = \frac{|C_{ref} - C_{pred}|}{\overline{C_{pred}}} \tag{12.8}$$

– Structural similarity $SSIM$

$$SSIM(A, B) = \frac{(2\mu_A \mu_B + c_1)(2\sigma_{AB} + c2)}{(\mu_A^2 + \mu_B^2 + c1)(\sigma_A^2 + \sigma_B^2 + c2)} \tag{12.9}$$

$$c_1 = (k_1 L)^2 \quad c_2 = (k_2 L)^2 \tag{12.10}$$

with C_{pred} the model prediction concentration, C_{ref} the reference concentration (ground truth), μ_A and μ_B are the respective average of A and B, σ_A^2 and σ_B^2 are the respective variances of A and B, σ_{AB} is the covariance of A and B, L is the dynamic range of the pixel values and k_1 and k_2 are two constants respectively 0.01 and 0.03 (by default).



(a) CFD result (b) MultiresUNet result

Figure 12.7: Maps of the studied district and comparison of the two results.

12.3.3 Results

To evaluate the Deep Learning capabilities to be applied in real life situation, a comparison has been made with real world data at the traffic at 08:30AM in the south of Schiltigheim, France the first of April 2021 between results from a CFD simulation and our Deep Learning approach on the NO_x dispersion from traffic emissions. The results proposed respectively by the CFD and MultiresUNet for the whole neighborhood are shown on Fig.12.7

It can be tedious to compare the results between the CFD and the Deep Learningg network since the CFD determines the dispersion in 3D while the deep learning approach works in 2D only at a given height. Nonetheless, the CFD needed one week of computation on 96 CPU while the deep learning network needed around 3 minutes on a GTX 1080Ti GPU, representing a speed up by x3000. To evaluate the accuracy of the predictions, the metrics presented above were computed between the prediction and the CFD considered as the ground truth and are presented below on Table 12.1.

12.4 Conclusion

As demonstrated by our work, Deep Learning has proved to be able to predict results close to CFD for air pollutant dispersion. Moreover, the MultiresUNet architecture was able to compute the dispersion in a matter of minutes over a wide area against several days for the CFD. This makes the Deep Learning approach a potential model to predict in real time

Metrics	$FAC2$	$NMSE$	FB	R	MAE_{rel}	J_{3D}	$SSIM$
Score	**0.818**	1.565	**0.176**	0.851	0.431	0.620	0.768
Expected values	> 0.5	< 1.5	< 0.3	1	0	1	1

Table 12.1: Evaluation of the quality of the dispersion model given by the deep learning approach.

over large scale the pollutant dispersion from traffic related pollution.

Chapter 13

Conclusion and perspectives

13.1 Conclusion of the dissertation

Air pollution affects the environment, for example it can have detrimental effect on water quality with acid rains or crops yield. But, first and foremost, air quality stakes are on the health and well-being of dwellers. Indeed, air quality can have huge impact on life expectancy and premature death especially in urban areas.

Hence, the objective of this work was to be able to assess pollution concentrations from local pollutant sources in urban areas. This objective is twofold. On one hand, to be able to assess annual concentrations in an acceptable computing time for engineering purposes, since annual concentrations are important to determine long term exposure. On the other hand, to be able to assess local pollution in real time over wide areas. Real time pollution management is important since it can give better insight for people on their exposure and help them to act to reduce it.

13.1.1 First objective: Assessing mean annual exposure

Modelling tools are very efficient to assess pollution dispersion. Their range can span from continents to neighbourhoods. Nevertheless, assessing mean annual concentration in urban areas represents a challenge. Indeed, for modelling, the timescale makes it computationally expensive and the geographical area with buildings requires complex physical models to consider them while computing the flow. These two constrains make it impossible to use the frontal approach of modelling the whole year using computational fluid dynamics and then averaging the results to assess the annual concentration of an area. Even with sensors deployed in the studied area, the challenge remains. Indeed, to compute a mean annual concentration, it requires to have the sensors fixed at the same place all year, which is costly giving the price of purchase and maintenance, and which reduces the covered area.

To overcome both of these issues, with sensors and modelling, methods were developed during this thesis to reduce the cost of assessing annual concentration. For the modelling tools, a methodology based on a frequency approach on the wind rose was presented. This methodology is a statistical approach that allows to reduce the number of simulations required to assess annual concentration. Indeed, only a handful number of simulations is necessary to represent the concentration for each direction of the wind rose before averaging the results weighted by the direction frequency. To assess the concentration of each directions continuous laws are proposed to interpolate wind speed either the classical Weibul distribution or a new sigmoid function that improves the interpolation. This law associated with a relationship between wind speed and concentrations enables to compute the concentration of the direction. The relationship between wind speed and concentrations for instance is hyperbolic when the atmosphere is considered neutral.

Nevertheless, this theoretical methodology in itself does not solve every concrete issue. An issue that comes when one wants to apply it is the discretization of the wind roses. Indeed the wind rose is continuous and the choice of wind direction to perform is arbitrary. In France, the institute in charge, "Méteo France", discretizes up to 20 degrees (divided into 18 directions). Hence, it was considered as the reference discretization. The objective was then to assess the error made when using coarser discretization. To do that, 5 wind roses and 7 neighbourhood were considered. For each neighbourhood 18 directions were simulated. Two strategies were evaluated, the first was to homogeneously discretize the wind roses and the second was to consider the predominant wind direction frequency. Globally, the first method outperformed the second one. The error from the reference could be roughly evaluated depending on the discretization with for instance 13.8% on average and a standard deviation of 6.8 for 9 directions compared to 18. Given the standard deviation this mean value can only be used qualitatively. So it was required to come up with a better solution to evaluate the error. A solution was proposed to determine the error after the computation are done for 6 and 9 homogeneous directions. Depending on the numerical error obtained through the method, the user can decide to make new simulations to improve the results or decides that it is within a satisfying error margin for the computational cost gain.

The methodology associated with the concrete example using CFD and flowchart to determine the error gives a concrete and operational tool to assess the mean annual concentration using modelling in a computable viable time for researcher and engineer.

For the sensors the main challenge is to reduce the monitoring time as much as possible with an error within an acceptable margin. The second challenge is to be able to infer the pollution of one pollutant from another of the same kind. In order to do that, each atmospheric pollutant must be treated individually since different molecules have different

sources and are affected differently by their environment. The focus was thus made on the two main pollutants in urban areas NOx/NO2 and PM10/PM2.5 that are responsible for the most death in Europe.

Firstly, the relationship between NOx and NO2 was explored with data from all around France and compared with existing equations from previous studies. It was found that the relationship of Derwent and Middleton was the best for France with a deviation of 7.6%. The seasonality was studied and it was found that NOx and NO2 mean annual concentration could be linked with mean monthly concentration by a quadratic relationship. The concavity/convexity of the regression were linked with the seasonal transition. The worst month was December with 15% mean error and the best month March with 7%, when averaging all month, the mean error was below 10%. Therefore, it is possible to determine the annual concentration using 1 month period allowing to cover more areas with one sensor.

Secondly, the relationship between PM10 and PM2.5 was also investigated with data from all around France and compared with existing equations.

With these two propositions, it is possible to determine mean annual concentrations with shorter period of measurement for the two main urban air pollutants in Europe.

13.1.2 Second objective: Assessing exposure in real time

Deep Learning has shown to be a formidable tool in many fields to overcome issues that were impossible to solve or yielding poor results. This part of the thesis aimed at using these recent advances in Deep Learning to be able to predict air pollution at the scale of the neighborhood in real time. It was first shown that the approach needed to be global and could not be made to fit a neighborhood in particular. Machine learning method methods such as Unet or random forest did not perform better on interpolating neighborhoods compared to a classic linear interpolation. Indeed, to take advantage of machine learning and especially Deep Learning, it is necessary to have many examples which is not possible with a little amount of direction for a neighbourhood. Thus, the approach was to train a Unet architecture on several neighborhood to extrapolate an unseen one. It was shown promising result with Unet managing to perform as good as a linear interpolation of 3 directions on a geometry it had never seen in a matter of minutes. **Deep Learning was therefore proved to be promising to be able to assess pollutant dispersion in urban areas based on CFD results.**

Making a Deep Learning model that works on any urban case and extrapolates on unseen geometries requires at minimum thousands of examples to begin to have reliable results. Nonetheless, CFD requires vast amount of computing power to converge. Hence, it is important to create examples as efficiently as possible to reduce the cost. The main principles

to generate examples are:

- to use the velocity field computation for several pollution sources since it represents most of the computation time.
- to simulate wide areas to relatively reduce the part of empty space necessary to have reliable result in CFD.
- homogeneously spacing the result from a long road.

This 3 practices increase the number of examples by an order of 20 to 30. Even if it could be argue that the example are not as informative as examples that would be with entirely different building layout and pollutant sources. Then, once the CFD simulation converged, the data needs to be converted into images since the library are efficient with this format. Yet, to convert the data, scales need to be chosen to turn the data into grayscale images. Appropriate scaling is chosen for the geometric data. For the pollution data for real life application a scale of 80 μm^3 is elected and when it is to compare between architectures each concentration field is scaled by its own maximum concentration. **CFD is therefore optimised to create examples for the Deep Learning with some good practices that use the computation time as best as possible.**

Unet seemed a promising architecture, with more examples it can be tested to see how well it performs. But, it is not the only Deep Learning architectures that deal with spatial information effectively. Many architectures exist, among them some were selected to be tested. 6 architectures were tried, Unet, PSPNet, SegNet, linkNet, FCN and multiResUnet. Three loses were also compared: a custom one, J3D and two classical, mean squared error and binary cross-entropy. For each architecture, various number of minimum filters were tried. It was found that the multiResUnet with 16 minimum filters with the J3D loss outperformed the others variants. Now that the best architecture and loss were elected, the second step is to optimise it. First, to improve the results, an area twice as big as the output is given as input since the surrounding buildings influence the pollutant dispersion in the area. Several optimisation were also carried out like attention gate without success and tuning the hyperparameters that mostly confirmed the ones already empirically chosen. The last optimisation was to use the multiResUnet 3D version to improve the results. The 3D version is more computing demanding and thus the architecture had to scaled down. In spite of this, it still managed to have similar result than the 2D version at the same height. The metrics are within the good satisfactory range for 2 out of 3 air quality metrics. **The multiResUnet architecture associated with the J3D loss have thus proved to be a powerful tool to assess air quality in real time.**

Finally, the Deep Learning architecture serves as the milestone of a real pollution time system that uses data from traffic and background sensors in real time to assess the pollution

dispersion. At the present time of this thesis, the Deep Learning model is only capable of dealing with pollutant coming from roads, hence, for other local pollutant sources such as plants it would require specific modelling or sensors. To check how well the model performs it was tested on a real neighborhood of $1km^2$ with real traffic data and compared with CFD. The Deep Learning model managed to perform 62% on the J3D and the 3 air quality metrics are within satisfactory range or close to it. The 2D results over the wide areas for a given height were obtained in 3 minutes on a graphic card while the CFD computation required 1 week of computation on 96 cores.

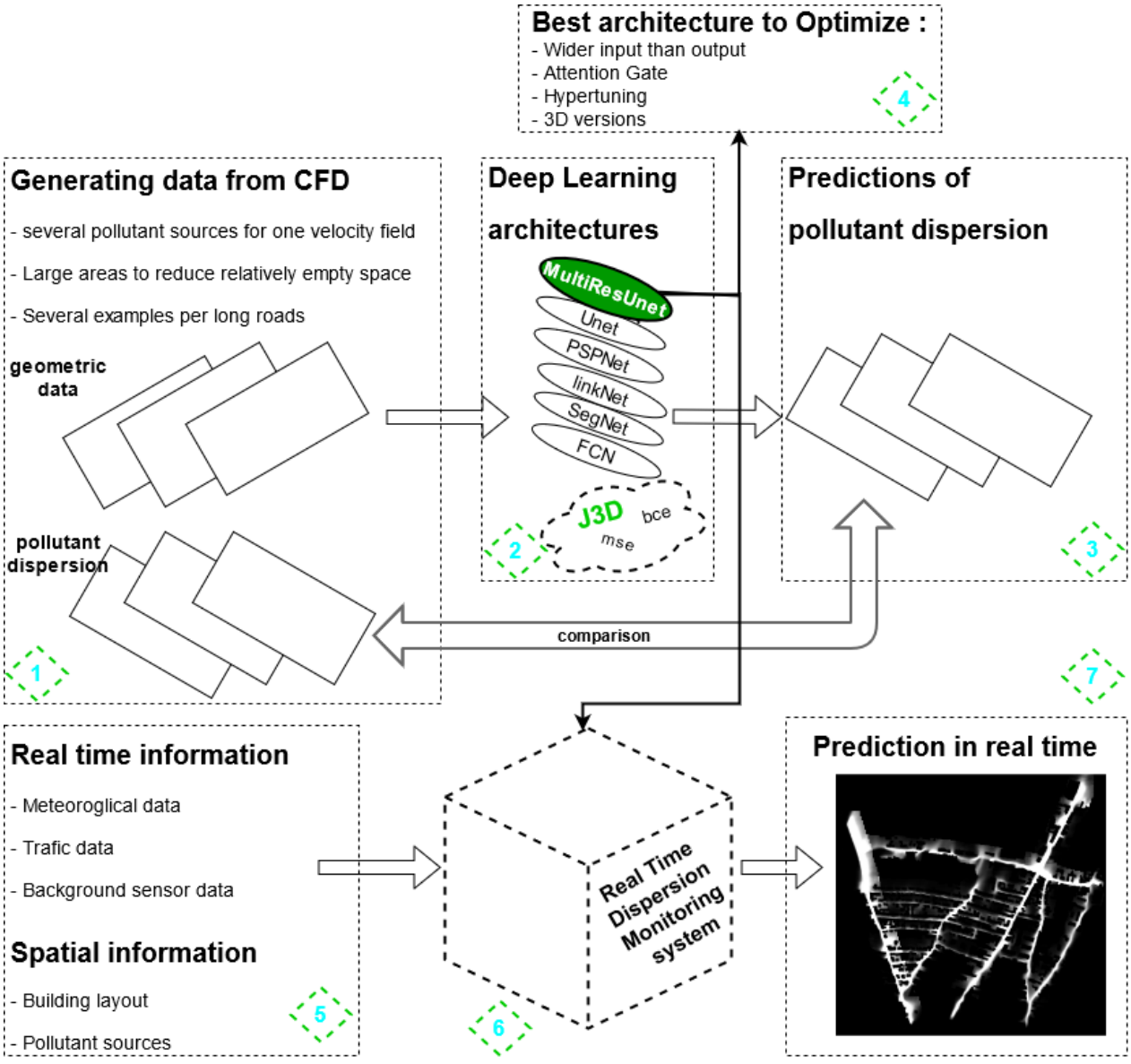


Figure 13.1: Diagram of the real time assessment part of the thesis

13.2 Perspectives

As shown before, the challenges of the thesis have been overcome and every aspect of the pollution from the long term effect of annual exposure to short-term effect of real time exposure have been covered. The thesis mixed traditional methods to assess annual pollution with CFD and enhance it with Deep Learning models, data mining from sensors and statistical analysis. Yet, much can still be done.

For the annual concentration from sensors, others pollutants annual trends such as the O3 or SO2 could be studied. The data used to establish the annual against monthly relationship only came from data sensors from France. Therefore, it could be interesting to analyse the trend in other western countries to ensure the method works in similar countries and other relationship established for countries with totally different lifestyle and climate.

For the annual concentration from modelling, it could be interesting to add the atmospheric stability into the equation. It is already present in the methodology hidden in the relationship between the concentration and wind speed. It could be added by a new summation of the different relationship between wind speed and concentration depending on the percentage of each atmospheric stability in the wanted area. Indeed, the hyperbolic relationship is only valid for neutral case. Could other relationship be established for different class of stability? How many more simulations would it required to take this into account? How impact-full would it be on the final results?

Finally, the main perspectives remain for the Deep Learning approach. First, more data should be generated to improve the results and generalisation of the model. Moreover, the examples were generated using a $k - \epsilon$ RANS model, but the results could be improved by using more complex model such as LES or more complex turbulence model such as RSM. Also, assumptions have been needed to train the Deep Learning model. For instance, it can only be applied when the road is at the same height as the building foundation. Another limitation is that it does not take into account atmospheric stability or temperature variation. These hypotheses should be leveraged to have a more robust an consistent simulation. Furthermore, Deep Learning and more broadly artificial intelligence domain know an explosion of development and, new architectures are published weekly. Hence, these new architectures should be tested and evaluated to determine if results could be improved. Recently, some architectures considering physical properties of fluid mechanics have been implemented and could be useful for the matter at hand. Finally, the model was compared against CFD models on a real neighborhood. However, to treat wide areas, Gaussian plume model are often used. It could be interesting to compare the performance between the deep learning model and this kind of model on a real city for annual concentration as well as real time concentrations compared with data obtained from *in situ* sensors.

Chapter 14

Résumé étendu en français

14.1 Introduction

Contexte de la pollution atmosphérique en milieu urbain La pollution de l'air est une thématique qui a parcouru les âges, des premiers hommes ayant subi des pollutions dues aux feux domestiques, en passant par le smog londonien des années 50 ayant tué des milliers de personnes, à aujourd'hui avec environ 8 millions de morts par an dues à la pollution de l'air.

Quatre principaux polluants ont été retenus par l'Organisation mondiale de la Santé pour leurs effets néfastes sur la santé et leur forte présence en milieu urbain. Les particules fines (PM), les dioxydes d'azote (NO_2) et le dioxyde de souffre (SO_2) étant des polluants fréquents et ayant un impact local et l'ozone (O_3) ayant un effet régional. Ces gaz affectent généralement les poumons et le système respiratoire des humains. Certains peuvent aussi avoir des effets sur l'environnement en entraînant des pluies acides ou un ralentissant de la croissance des plantes.

Pour combattre cet état de fait, des mesures ont été prises aussi bien au niveau mondial par l'Organisation mondiale de la Santé, recommandant notamment des niveaux de concentration en polluant à atteindre pour chaque polluant, qu'au niveau de l'Union européenne, en imposant dans le droit européen, et de ce fait dans le doit de ses États membres, une législation sur la pollution atmosphérique et des seuils à respecter.

En France, 40 000 décès sont attribuables aux $PM_{2,5}$ et 7000 aux NO_2. Les zones les plus touchées par la pollution de l'air sont les villes de plus de 100 000 habitants avec une perte d'espérance de vie de 15 mois à 30 ans. Comparativement, les zones rurales ont une perte de moins de 9 mois en moyenne sur l'espérance de vie. Il est donc primordial pour les autorités publiques de traiter la question de la pollution de l'air en milieu urbain.

Défi et contenu de la thèse La réglementation de la pollution atmosphérique fait intervenir différentes échelles temporelles. Il y a aussi bien des seuils sur la concentration moyenne annuelle que sur la concentration horaire. Ces deux échelles créent un double besoin. La thèse est ainsi divisée en deux parties. Chacune des parties traitent de l'un des deux besoins.

Dans un premier temps, il y a un besoin de développer des méthodologies pour déterminer la concentration moyenne annuelle efficacement. En effet, les modèles permettant de déterminer la moyenne annuelle en rejouant une année complète de données heure par heure ont des hypothèses simplificatrices très fortes. Ces hypothèses montrant leur limites en milieu urbain. Les modèles plus complexes comme ceux de dynamique des fluides ont des meilleurs résultats mais sont très coûteux en temps. Il est donc nécessaire de développer une approche statistique basée sur les fréquences de vents pour réduire le nombre de simulations nécessaires. Quant aux capteurs de pollution atmosphérique, ils doivent être fixes à un endroit donné pendant une année complète ce qui entraîne des coûts élevés. Réduire le temps des mesures pour obtenir la moyenne annuelle permettrait donc de réduire le coût associé au campagne de mesure.

Dans un second temps, connaître l'exposition des riverains et des passants en temps réel n'est pas possible avec des modèles de dynamique des fluides en un temps et pour un coût acceptable. Il faut donc pouvoir déterminer la dispersion de polluant plus rapidement et dans l'idéal en sacrifiant le moins possible en précision, avantage majeur de la dynamique des fluides. Un domaine ayant connu un fort développement et même révolutionné bons nombres de secteurs est l'apprentissage automatique et l'intelligence artificielle. Il serait donc judicieux d'utiliser les récentes avancées pour les appliquer à la thématique de la qualité de l'air et des modèles numériques afin d'en accélérer le processus et d'atteindre des performances compatibles avec le temps réel tout en conservant une précision satisfaisante.

14.2 Évaluation de la moyenne annuelle par modélisation numérique et mesures de capteurs

14.2.1 Comment évaluer la pollution atmosphérique moyenne annuelle à l'échelle d'un quartier ?

Pour évaluer la concentration moyenne annuelle une nouvelle méthodologie est nécessaire pour les modèles ne pouvant jouer une année de données heure par heure. Pour cela, une méthodologie utilisant la rose des vents annuelle avec ses fréquences de direction/vitesse, la relation entre vitesse du vent et concentration en polluant et l'émission moyenne de polluants annuelle est proposée. Cette méthodologie a été publiée et correspond à la référence suivante : Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Dufresne, M., Blond,

N., AND WERTEL, J. Methodologies to assess mean annual air pollution concentration combining numerical results and wind roses. *Sustainable Cities and Society 59* (Aug. 2020), 102221

Interpolation de la fréquence des vitesses par une fonction continue Une donnée essentielle pour évaluer la dispersion de polluant est la vitesse du vent. Cependant, pour une direction de vent celle-ci varie. Il est donc primordial de connaître la fréquence de ces variations pour pouvoir évaluer son impact sur la dispersion de polluants. Ces données sont fournies par des institutions comme météo France. Cependant, plus la discrétisation est forte, plus elles sont coûteuses. Pouvoir donc évaluer la fréquence des vitesses à partir de quelques gammes de vents est donc particulièrement intéressant. Pour interpoler ces données, deux méthodes sont comparées : la distribution de Weibul 14.1 et une interpolation utilisant des sigmoïdes 14.2.

$$f(v) = \frac{k}{\lambda}\left(\frac{v}{\lambda}\right)^{k-1} e^{-(v/\lambda)^k} \tag{14.1}$$

Où v est la vitesse du vent, k est le paramètre de forme et λ est le paramètre de distribution d'échelle, avec k et λ des nombres positifs.

$$f(v) = \alpha.\left(-1 + \frac{1}{1+\beta_1.e^{-\gamma_1.v}} + \frac{1}{1+\beta_2.e^{\gamma_2.v}}\right) \tag{14.2}$$

Où α, β_1, β_2, γ_1 and γ_2 sont des paramètres positifs.

Les deux fonctions aboutissent à des résultats similaires. Cependant, l'interpolation utilisant des sigmoïdes à un autre avantage, elle peut être optimisée pour certains profil de vitesses de vent (ayant une première gamme moins fréquente que la seconde gamme) pour améliorer l'interpolation pour les vitesses faibles entraînant les concentrations les plus importantes et ne devant donc pas être sous estimées.

$$f(0) = FVR_{[0;\alpha[}\frac{FVR_{[0,\alpha[}}{FVR_{[\alpha,\beta[}} \tag{14.3}$$

Où $FVR_{[0,\ \alpha[}$ est la fréquence du vent pour la première gamme de vitesses de la discrétisation de la rose des vents en 4 gammes et $FVR_{[\alpha,\beta[}$ est la fréquence du vent pour la deuxième gamme de vitesses (dans cette étude $\alpha = 1{,}5$ et $\beta = 4{,}5$).

Exemple d'interpolation sur la figure suivante 14.1

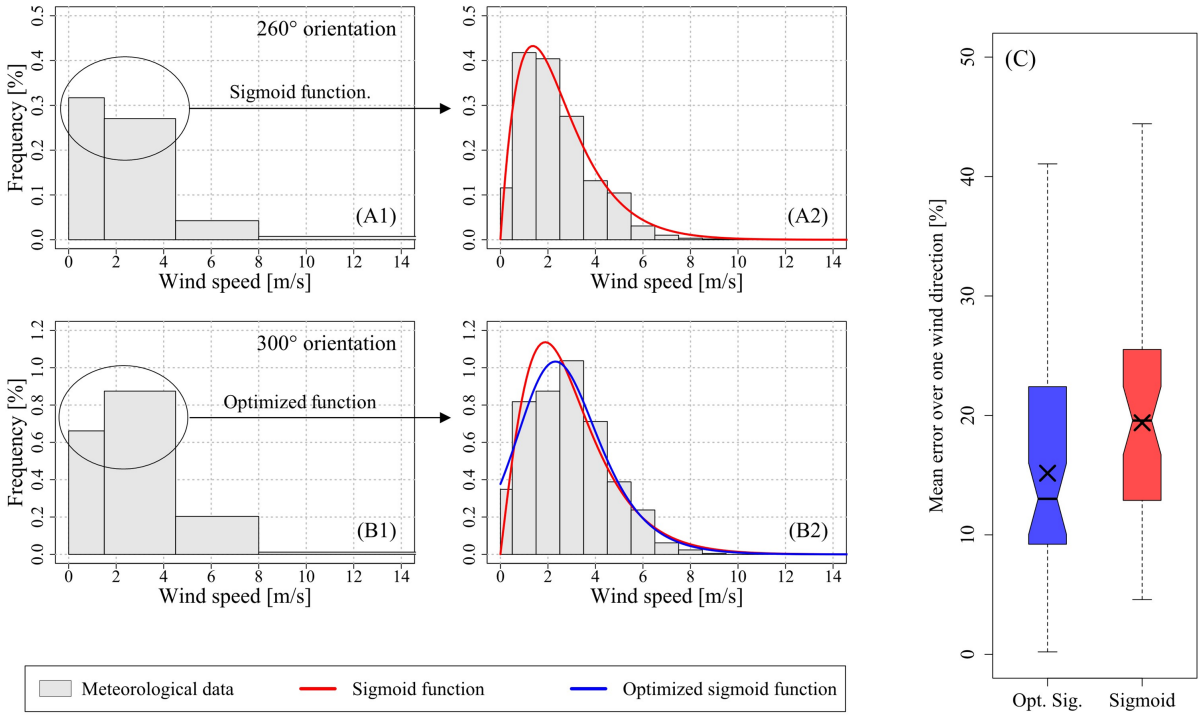


FIGURE 14.1: (A–B) Illustration de la méthodologie de la fonction sigmoïde optimisée et (C) comparaison avec les résultats de la fonction sigmoïde standard.

Évaluation de la pollution moyenne annuelle par méthode fréquentielle sur base de modèle numérique Initialement, les concentrations moyennes annuelles basées sur les résultats de la CFD peuvent être calculées en utilisant une méthodologie discrète. Cette méthodologie considère que la concentration annuelle moyenne à un endroit donné est la résultante de plusieurs contributions des différentes directions du vent et de leur fréquence de vitesses. La concentration moyenne sur une direction de vent peut être calculée avec l'équation suivante 14.4 et la concentration annuelle moyenne avec l'équation 14.5.

$$\overline{C}_d = \frac{\sum_{r=1}^{n} C_{d,r} \cdot f_{d,r}}{\sum_{r=1}^{n} f_{d,r}} + C_{bg} \tag{14.4}$$

$$\overline{C} = \frac{\sum_{i=1}^{n} \overline{C}_d \cdot f_d}{\sum_{i=1}^{n} f_d} \tag{14.5}$$

où $\overline{C}_d$ est la concentration moyenne dans une direction de vent, $C_{d,r}$ est la concentration pour une direction de vent donnée d et une plage de vitesse de vent donnée r, $f_{d,r}$ est la fréquence pour une direction de vent donnée et une plage de vitesse de vent donnée, C_{bg} est la concentration de fond, $\overline{C}$ est la concentration annuelle moyenne et f_d la fréquence totale d'une direction de vent donnée.

À noter que bon nombre de chercheurs utilisent la moyenne arithmétique pour la gamme de vitesse. Or la concentration évolue de manière hyperbolique avec la vitesse du vent. Il faut donc calculer la vitesse représentative de la gamme de la manière suivante :

$$v_r = \sqrt{\frac{2}{\frac{1}{{v_{max}}^2} + \frac{1}{{v_{min}}^2}}} \tag{14.6}$$

où v_{max} et v_{min} sont respectivement les vitesses maximale et minimale de la plage de vitesses, v_r est la vitesse représentative de la plage de vitesses et $c\,(v)$.

Pour calculer la concentration pour une direction, il est aussi possible de considérer des hypothèses sur la relation entre la vitesse du vent et la concentration comme c'est le cas pour l'atmosphère en condition neutre avec la relation suivante :

$$c(v) = v_{ref}.\frac{C_{ref}}{v} \tag{14.7}$$

Il est ainsi possible de calculer la concentration moyenne pour la direction de vent de la manière suivante :

$$\overline{C_d} = \frac{\int_0^{+\infty} c\,(v)\,.f\,(v)\,.dv}{\int_0^{+\infty} f\,(v)\,.dv} + C_{bg} \tag{14.8}$$

L'intégrale n'est pas définie en 0, la concentration tendrait vers l'infini ce qui n'est pas possible physique. En effet, la vitesse du vent n'atteint jamais 0 à proprement parlé, il y a toujours des mouvements d'air dû à la différence de température ou à la turbulence engendrée par le trafic par exemple. Une hypothèse peut donc être faite qu'il y a une vitesse minimale pour laquelle la concentration sera donc constante comme dans l'équation suivante 14.9 :

$$\overline{C_d} = C_{max}.\frac{\int_0^{v_{min}} f\,(v)\,.dv}{\int_0^{+\infty} f\,(v)\,.dv} \;+\; \frac{\int_{v_{min}}^{+\infty} c\,(v)\,.f\,(v)\,.dv}{\int_0^{+\infty} f\,(v)\,.dv} + C_{bg} \tag{14.9}$$

où $\overline{C}$ est la concentration annuelle moyenne, C_{max} est la concentration maximale pour le calcul, v_{min} est la vitesse pour laquelle $c\,(v)$ est considérée comme égale à C_{max}, $f\,(v)$ est l'équation 14.2, $c\,(v)$ est l'équation 14.7 et C_{bg} est la concentration de fond.

Les résultats entre la méthode continue utilisant une interpolation grâce à la fonction sigmoid déterminé sur une discrétisation de 4 gammes et la méthode discrète sur 18 gammes donnent des résultats similaires sont présentées sur la Figure 14.2 ci-dessous. La différence est de l'ordre de 5 %.

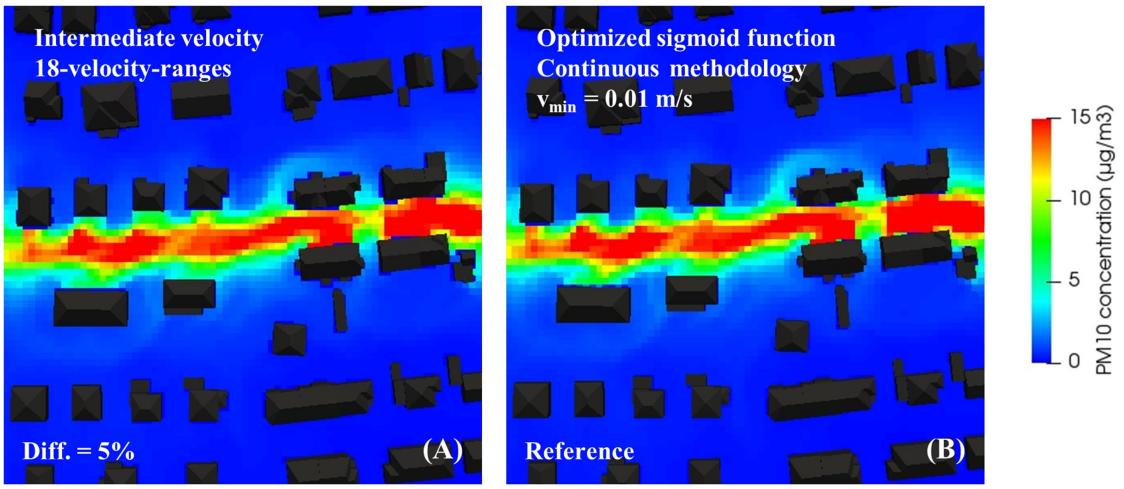


FIGURE 14.2: Comparaison des concentrations annuelles moyennes basées sur la distribution "détaillée" des vents sur 18 plages de vitesse en utilisant (A) la vitesse intermédiaire et (B) la vitesse représentative.

14.2.2 Optimisation de la discrétisation de la rose des vents

Dans le chapitre précédent, une méthodologie pour déterminer la concentration moyenne annuelle en polluant à l'aide de la rose des vents et de modèle numérique a été présentée. Cependant, aucune remarque n'est faite sur la discrétisation de la rose des vents. En effet, cette discrétisation peut varier en fonction des pays ou des chercheurs. Plus la discrétisation est fine, plus le nombre de simulations nécessaires est élevé augmentant ainsi le coût d'une étude. Il est donc important de connaître l'erreur qu'entraîne divers degrés de discrétisation afin d'optimiser le nombre de simulations. Pour se faire, plusieurs discrétisations ont été testé allant de 2 à 18 directions pour calculer la pollution moyenne annuelle sur 7 quartiers avec 5 roses des vents différentes. La discrétisation tous les 20 degrés (18 directions) est la référence dans cette étude contre laquelle les autres discrétisations plus grossière vont être comparées. Pour discrétiser la rose des vents, deux approches sont possibles, de manière homogène en répartissant équitablement les directions de vents à simuler ou simulant les directions de vents les plus fréquentes. Cette méthodologie a été publiée et correspond à la référence suivante : JURADO, X., REIMINGER, N., VAZQUEZ, J., AND WEMMERT, C. On the minimal wind directions required to assess mean annual air pollution concentration based on CFD results. *Sustainable Cities and Society 71* (Aug. 2021), 102920

Comparaison des deux approches pour discrétiser la rose des vents La première méthodologie, qui utilise un pas régulier, a montré des tendances dans les erreurs en fonction

du modèle de rose des vents : plus la rose des vents est homogène et moins l'erreur est élevée. Une telle tendance n'a pas été observée avec la deuxième méthodologie qui utilise les directions prédominantes du vent. Quelle méthodologie est alors la meilleure ? C'est-à-dire, laquelle permet d'aboutir à l'erreur la moins élevée avec le même nombre de simulations. Pour cela, leurs résultats moyens sont comparés pour différentes discrétisations. Il n'y a pas une des méthodes qui est toujours meilleure que l'autre mais la méthode homogène donnent la plupart du temps de meilleurs résultats et est donc recommandé comme montré sur la figure 14.3.

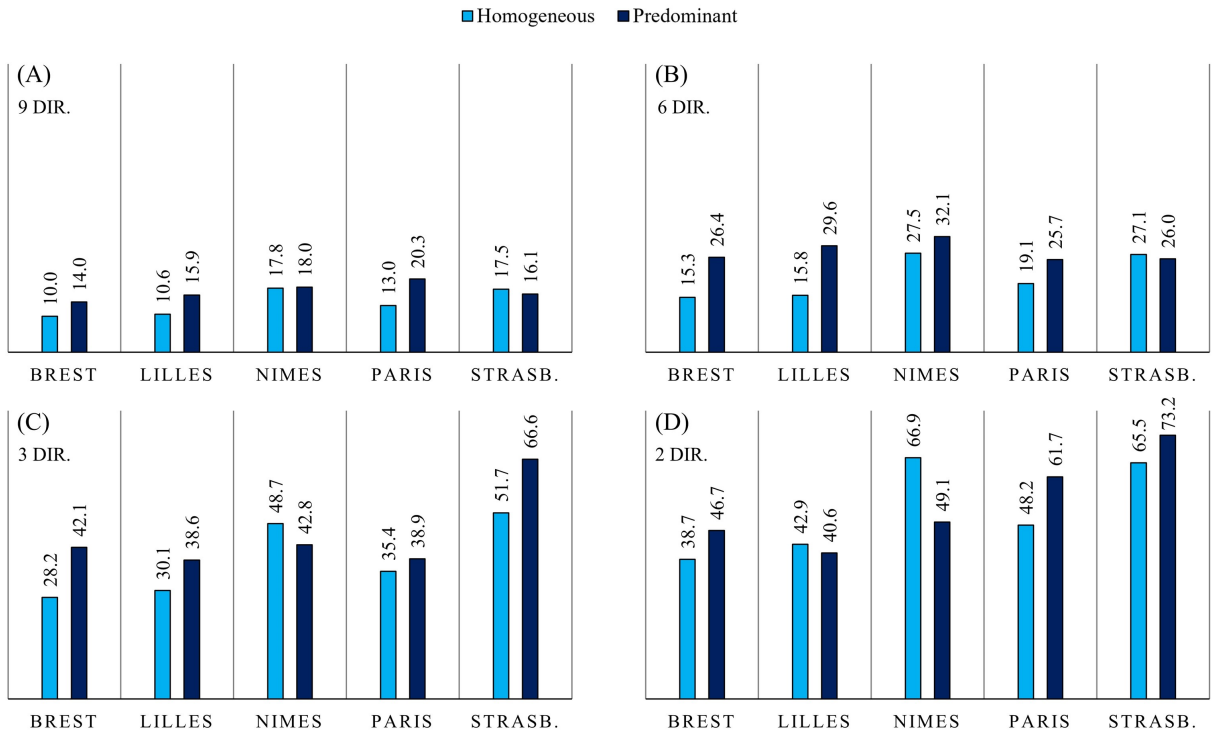


FIGURE 14.3: Comparaison de l'erreur moyenne sur la concentration annuelle moyenne par rapport à la référence en fonction de la rose des vents pour les directions de vent 9, 6, 3 et 2 en utilisant la première (homogène) et la seconde (prédominante) approche.

Cependant, il est observable sur la figure 14.3 que la variance de l'erreur est assez élevée, l'erreur moyenne pour un nombre de directions n'est donc qu'une information qualitative plus que quantitative.

Calcul de l'erreur Une valeur qualitative donne une idée, un ordre de grandeur sur l'erreur attendue. Cependant, au regard des intérêts liés à la qualité de l'air, une information quantitative plus précise est essentielle. Pour cela, une méthodologie a été développée pour la stratégie de discrétisation homogène. L'utilisateur choisit un nombre de directions qu'il

souhaite calculer, 6 ou 9. Pour la suite de l'explication, considérons que 9 directions sont choisies (discrétisation tous les 40 degrés). Il convient alors d'effectuer ces 9 simulations. Une fois que ces simulations sont terminées, il faut déterminer la carte de concentrations pour les sous-ensembles d'une discrétisation de 3 directions contenues dans l'ensemble des 9 directions réalisées. Par exemple, si on a [0,40,80,120,160,200,240,280,320] on a alors les sous-ensembles [0,120,240] [40,160,280] [80,200,320] comme indiqué sur la figure 14.4.

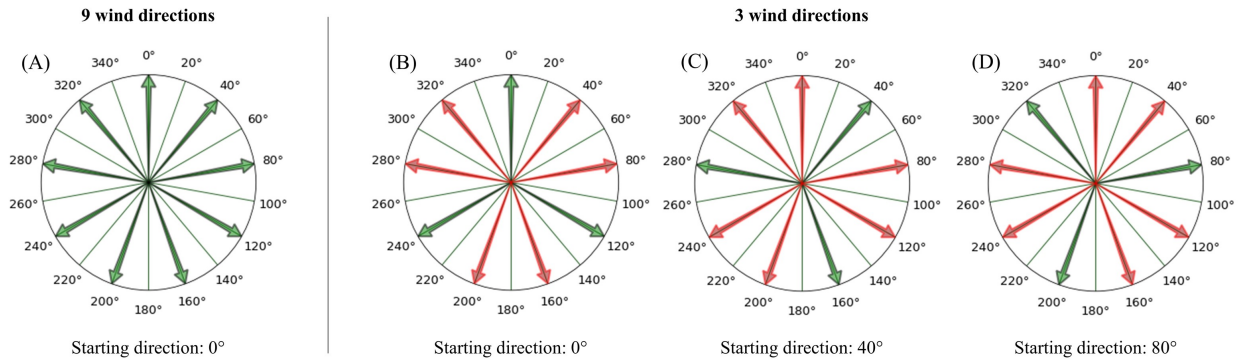


FIGURE 14.4: Illustration de l'exemple permettant de calculer l'erreur commise par rapport à la prise en compte de la totalité de la rose des vents (flèche verte : direction du vent modélisée / flèche rouge : direction du vent non prise en compte pour le calcul de la concentration annuelle).

Il suffit ensuite de moyenner les trois cartes de chaque sous-ensemble puis de comparer l'erreur entre cette carte composée des sous-ensemble de 3 directions et la carte obtenue avec les 9 directions. Cette erreur est ensuite à multiplier par un coefficient donné dans l'équation 14.10, on obtient dès lors l'erreur effectuée entre une discrétisation avec 9 directions et 18 directions.

$$E_{9/18} = E_{3/9} \times 0.3714 \tag{14.10}$$

Si l'erreur est acceptable compte tenu des enjeux et du coût, il n'y a pas besoin de directions supplémentaires, sinon il faut effectuer plus de simulations. Plus de détails sont disponibles dans la thèse quant à la manière dont ce coefficient a été déterminé, le coefficient lorsque l'on choisit 6 directions ainsi qu'un diagramme récapitulatif de la méthode.

14.2.3 Évaluation de la concentration annuelle moyenne de NO_2 sur la base de données partielles

La modélisation numérique est un outil puissant mais il est souvent nécessaire de la coupler à des capteurs pour augmenter sa fiabilité. En effet, les capteurs peuvent d'une part, aider à

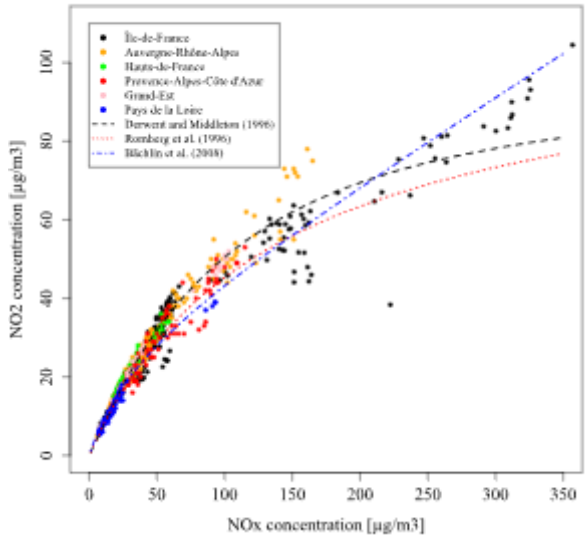


FIGURE 14.5: Évolution de la concentration de NO_2 en fonction de la concentration de NO_x et comparaison avec les fonctions empiriques

déterminer la pollution de fond qui n'est pas calculée par les modèles à micro-échelle. D'autre part, ils peuvent être utilisés pour s'assurer que les résultats du modèle sont cohérents dans la zone d'intérêt. Cependant, l'utilisation de capteurs pour la concentration annuelle présente une limite majeure : un capteur a besoin d'effectuer sa mesure sur toute l'année au même endroit et ne mesure qu'un seul polluant. Deux questions se posent dès lors : est-il possible de déterminer la concentration annuelle sur une période de temps plus courte et est est-il possible de déterminer d'autres espèces de polluant proche à partir des données d'un autre polluant ? Cette étude a été publiée et correspond à la référence suivante :JURADO, X., REIMINGER, N., VAZQUEZ, J., WEMMERT, C., DUFRESNE, M., BLOND, N., AND WERTEL, J. Assessment of mean annual NO2 concentration based on a partial dataset. *Atmospheric Environment 221* (Jan. 2020), 117087

Relation entre NO_x et NO_2 Le dioxyde d'azote fait partie de la famille des oxydes d'azote. Des auteurs ont déjà précédemment étudier la question et le lien entre les dioxydes d'azote et monoxyde d'azote. Trois équations reliant ces deux gaz ont été comparées, elles donnent des résultats proches pour des valeurs faibles mais divergent d'autant plus que la concentration augmente. Ces lois ont donc étaient comparées à des données issues de capteurs des agences régionales sur toute la France pour déterminer quelle loi semble la plus adéquat en France.

La loi qui parvient le mieux à prédire les données des capteurs en France est la loi de Derwent et Middleton présenté ci-dessous :

A=$\log_{10}([NO_x]_h/1.91)$.

$$[NO_2]_h = \left(2.166 - \frac{[NO_x]_h}{1.91}\left(1.236 - 3.348A + 1.933A^2 - 0.326A^3\right)\right) \times 1.91 \tag{14.11}$$

Cette loi parvient sur les données issues de France à atteindre 7.8% de déviation en moyenne et est donc une loi robuste et fiable pour convertir les deux gaz. À noter cependant que plus la concentration est élevée, plus la dispersion s'agrandit. Cette loi est donc à utiliser avec attention lorsqu'appliquée à des concentrations fortes comme par exemple sur une autoroute en heure de pointe.

Relation entre mesures mensuelles et mesures annuelles Le deuxième défi est donc de réduire les contraintes temporelles pour obtenir la concentration moyenne annuelle à l'aide de capteur. Pour cela la saisonnalité des dioxydes d'azotes et leur variation mois par mois a été étudiée à partir des mêmes données que celle pour comparer les différentes lois liant monoxyde et dioxyde d'azote. Il a été observée qu'une loi quadratique 14.12 permettait de relier la concentration mensuelle à la concentration annuelle pour chaque mois de l'année. En outre, les paramètres de cette loi peuvent être déterminés de manière continue tout au long de l'année par les équations 14.13,14.14 (30 jours de mesures) à la concentration annuelle suivant les lois :

$$[NO_2]_a = a.[NO_2]_m{}^{\mathbf{2}} + b.[NO_2]_m \tag{14.12}$$

$$\alpha = 0.0033 - 0.0102.\exp\left[\frac{-(m\ -\ 6.5749)^{\mathbf{2}}}{8.6962}\right] \tag{14.13}$$

$$\beta = 0.6945 + 0.8708.\exp\left[\frac{-(m\ -\ 6.7076)^{\mathbf{2}}}{7.4328}\right] \tag{14.14}$$

Ces lois ont permis de déterminer la concentration moyenne annuelle en Île de France avec une erreur de l'ordre de 10% et sur le reste de la France avec une erreur de l'ordre de 15 %. Néanmoins, tous les mois ne se valent pas, le mois d'avril est le meilleur mois pour effectuer une mesure avec une erreur qui baisse aux alentours des 7%. En outre, la valeur du mois d'avril peut être considérée comme étant représentative de la pollution moyenne annuelle car elle est reliée quasi linéairement à la moyenne annuelle avec un coefficient directeur proche de 1.

14.2.4 Évaluation de la concentration annuelle moyenne de PM_{10} et $PM_{2.5}$ sur la base de données partielles

La même étude est aussi réalisée pour les particules fines qui sont elles aussi un polluant responsable de millier de morts en France par an. Cette méthodologie a été soumise et correspond à la référence suivante : Jurado, X., Reiminger, N., Maurer, L., Vazquez, J., and Wemmert, C. Assessment of mean annual pm_{10} and $pm_{2.5}$ concentration based on a partial dataset. *Submitted to Sustainable Cities and Society*.

Évolution de la concentration en particules fines en France Les particules fines ont eu tendance à diminuer en France ces dix dernières années. La réduction a principalement eu lieu entre 2011 et 2015 avec une pente de -1.3 pour les PM_{10} et -1.2 pour les $PM_{2,5}$, puis la décroissance devient moins prononcée entre 2015 et 2019 avec une pente de -0.4 pour les PM_{10} et -0.6 pour les $PM_{2,5}$.

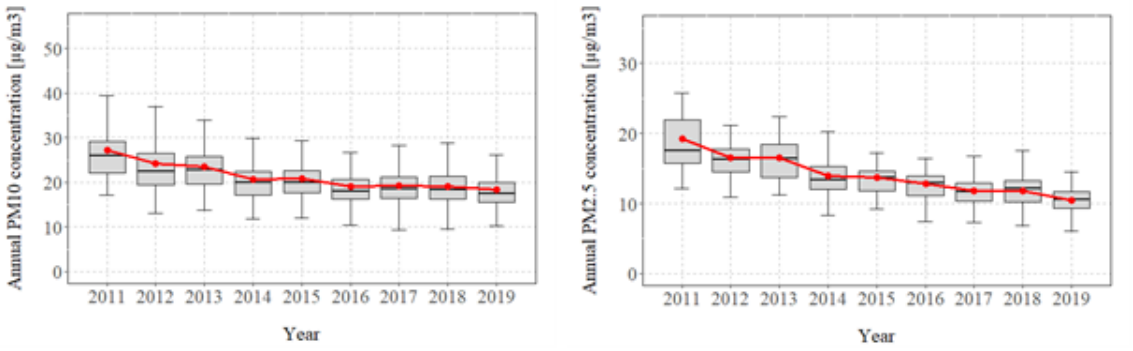


Figure 14.6: Évolution des concentrations annuelles (A) de PM_{10} et (B) de $PM_{2,5}$ entre 2011 et 2019 pour chaque type de station.

Relation entre mesures mensuelles et mesures annuelles En France, une façon populaire d'évaluer la concentration annuelle est de mesurer un mois en été et un mois en hiver pour avoir une meilleure représentativité de la variation saisonnière au cours de l'année. En effet, plusieurs mois peuvent être suivis tout au long de l'année pour améliorer les prédictions sur la concentration moyenne annuelle. Mais il manque des informations quantitatives sur le gain de précision des mesures sur plusieurs mois par rapport aux mesures sur un mois. Pour résoudre ce problème, la concentration moyenne de plusieurs couples de mois régulièrement espacés a été étudiée. Par exemple, si deux mois sont utilisés, la moyenne des concentrations sera calculée avec les mois de janvier (1^{st} mois) et juillet ($1 + 6 = 7^{th}$ mois) ; avec 4 mois, cela sera calculée avec la moyenne entre janvier (1^{st}), avril (4^{th}), juillet (7) et octobre (10^{th}) ; etc.

La figure 14.7 montre une comparaison du MRE pour les PM_{10} et $PM_{2,5}$ en utilisant soit une régression linéaire sur la moyenne soit directement la valeur de concentration moyenne des mois.

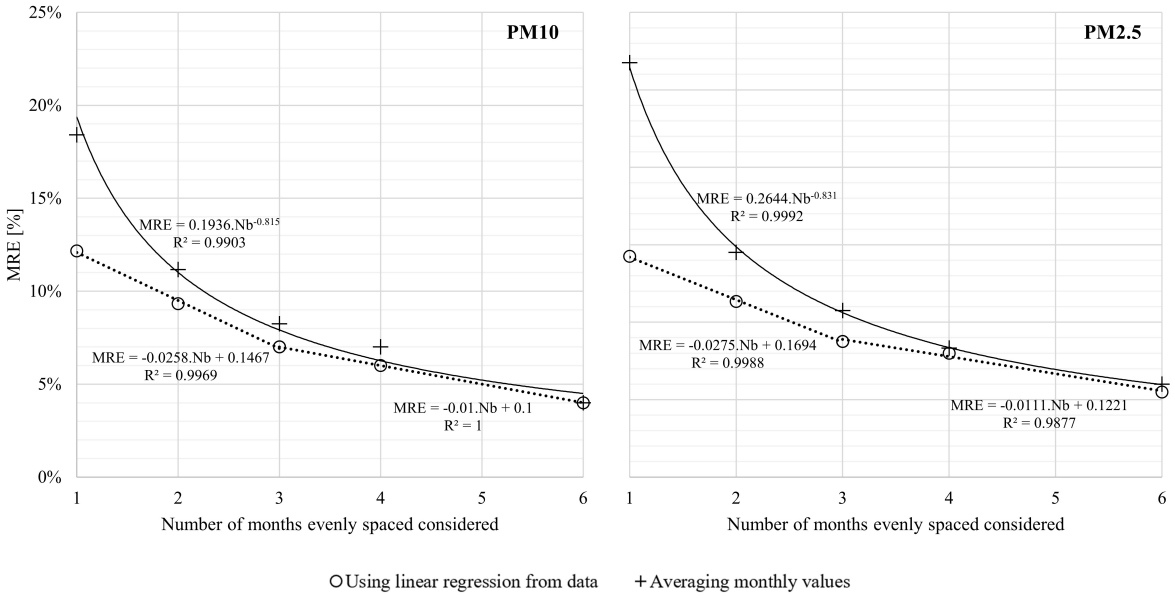


FIGURE 14.7: Évolution de l'erreur relative moyenne (A) en fonction du nombre de mois pour PM_{10} (gauche) et $PM_{2,5}$ (droite) lors de l'utilisation de la régression linéaire et de la moyenne directe.

plus le nombre de mois est élevé, meilleurs sont les résultats. Ces résultats conduisent à deux points essentiels :

- D'un mois à trois mois, la pente est plus forte que de trois mois à six mois, ce qui signifie que le gain en erreur est maximisé jusqu'à une période de 3 mois pour les PM_{10} et $PM_{2,5}$.
- La régression linéaire améliore les résultats, surtout lorsque le nombre de mois utilisés est faible. Lorsqu'on atteint 3 mois, la différence entre la régression linéaire et le calcul de la moyenne devient inférieure à 10%.

Une façon simple d'améliorer les résultats de 2 mois est d'utiliser l'observation faite sur les 3 types de mois qui existent : les meilleurs résultats sont obtenus en utilisant deux mois qui ne contiennent pas de mois de type hiver (c'est-à-dire avril à octobre et mai à novembre). Ainsi, il est fort probable que si un mois d'hiver est considéré et représente la moitié des données lors du calcul de la moyenne des concentrations, il représente trop la saison d'hiver. Pour résoudre ce problème, pondérer les mois de type hiver (décembre, janvier, février et mars) par $\frac{1}{4}$ et le reste des mois par $\frac{3}{4}$ améliore les résultats ainsi que la stabilité des prédictions.

Relation entre PM_{10} et $PM_{2,5}$ La relation entre PM_{10} et $PM_{2,5}$ a été étudiée. Des lois permettant de convertir les PM_{10} en $PM_{2,5}$ et vice versa dépendant de la saison et/ou du trafic sont proposées :

Influence type	Equation	R2	MRE	95CRE
Full dataset	$PM_{2,5} = 0,60 \times PM_{10} + 0,63$	0,74	0,17	0,50
Background	$PM_{2,5} = 0,73 \times PM_{10} - 1,58$	0,77	0,15	0,42
Traffic	$PM_{2,5} = 0,54 \times PM_{10} + 1,36$	0,75	0,17	0,44

TABLE 14.1: Résultats des différentes régressions linéaires sur l'ensemble des données, sur celle de l'influence du fond de pollution et sur celle de l'influence du trafic

Types de mois	Équation	R2	MRE	$C_9$5RE
mois d'hiver	$PM_{2.5} = 0,61 \times PM_{10} + 2,37$	0,75	0,14	0,40
Reste de l'année	$PM_{2.5} = 0.54 \times PM_{10} + 1.11$	0.72	0.16	0.46

TABLE 14.2: Résultats des différentes régressions linéaires par types de mois avec respectivement les mois d'hiver et le reste de l'année

Les deux méthodologies peuvent être utilisées en même temps. Ceci permet d'améliorer les résultats avec un R2 de 0,80, un MRE de 0,12 et un $C_9$5RE de 0,33.

14.3 Évaluation de la dispersion de polluant en temps réel

Les modèles CFD pour la pollution urbaine locale sont parmi les meilleurs modèles en termes de précision car ils prennent en compte des phénomènes complexes tels que la turbulence induite par les bâtiments. Pour l'évaluation de la moyenne annuelle, des méthodes statistiques peuvent pallier la limite engendrée par le coût des calculs en réduisant le nombre de simulations nécessaires. Néanmoins, la solution en temps réel de la dispersion des polluants avec la CFD n'est pas possible. Il faut donc utiliser des modèles plus rapides. Mais ces modèles plus rapides sont moins précis et ne prennent pas en compte les bâtiments de manière satisfaisantes. De nouvelles méthodes doivent donc être développées. L'apprentissage automatique et l'intelligence artificielle sont des domaines en plein essor qui ont réussi à changer le para-

digme dans de nombreux domaines. Est-il possible d'associer les récentes avancées dans ce domaine à la CFD pour le suivi en temps réel de la dispersion des polluants ?

14.3.1 Évaluation de la capacité des modèles d'apprentissage profond pour évaluer la dispersion des polluants

Dans un premier temps, il convient d'étudier les capacités des méthodes d'apprentissage automatique quant à leur capacité à prédire des champs de dispersion de polluants. Pour cela deux approches sont explorées. D'une part, étudier la capacité d'algorithme d'apprentissage automatique à interpoler des données CFD pour un quartier. D'autre part, la capacité d'algorithme d'apprentissage automatique à extrapoler des données CFD sur des quartiers jamais vu lors de l'entraînement. Cette méthodologie a été soumise et correspond à la référence suivante : Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Assessment of capability of deep learning to predict air pollution dispersion. *Submitted to Computers, Environment and Urban Systems*.

Interpolation de la dispersion de polluant pour un quartier Pour l'interpolation, étant donné un ensemble de simulations CFD pour un quartier, est-il possible de déterminer les directions de vents non simulées à partir de celles connues. Dans ce but, une interpolation linéaire classique sera comparée à deux méthodes d'apprentissage automatique pour voir s'il y a une valeur ajoutée à utiliser ces méthodes, la forêt aléatoire (random forest) et Unet. Les résultats des différentes approches sont présentés pour différents espacements, tous les 120 degrés (3 directions), tous les 60 degrés (6 directions) et tous les 40 degrés (9 directions). Plus de détails sont disponibles dans la thèse quant aux méthodes.

La figure 14.8 ci-dessus montre que la méthode la plus simple, l'interpolation linéaire reste la meilleure option. Ceci est facilement explicable par le fait que les méthode d'apprentissage automatique ont besoin de nombreux exemples pour être efficace. Hors, avec un seul quartier, le nombre d'exemples est très limité.

Extrapolation de la dispersion de polluant pour un quartier Une autre manière d'aborder le défi est non pas d'interpoler, mais d'entraîner un algorithme d'apprentissage profond sur différents quartiers, ce qui permet d'augmenter les nombres d'exemples et est plus approprié pour des approches d'apprentissage automatique. Une fois que l'algorithme d'apprentissage profond est entraîné, il devrait donc être capable d'extrapoler des géométries qu'il n'a jamais vues. Pour ce faire l'algorithme Unet a été entraîné sur différents quartiers et a eu pour tâche de prédire la dispersion de polluant sur un quartier qu'il n'a jamais servi lors de son entraînement. Les résultats sont présentés sur la figure 14.9 et les résultats dans le tableau 14.3 :

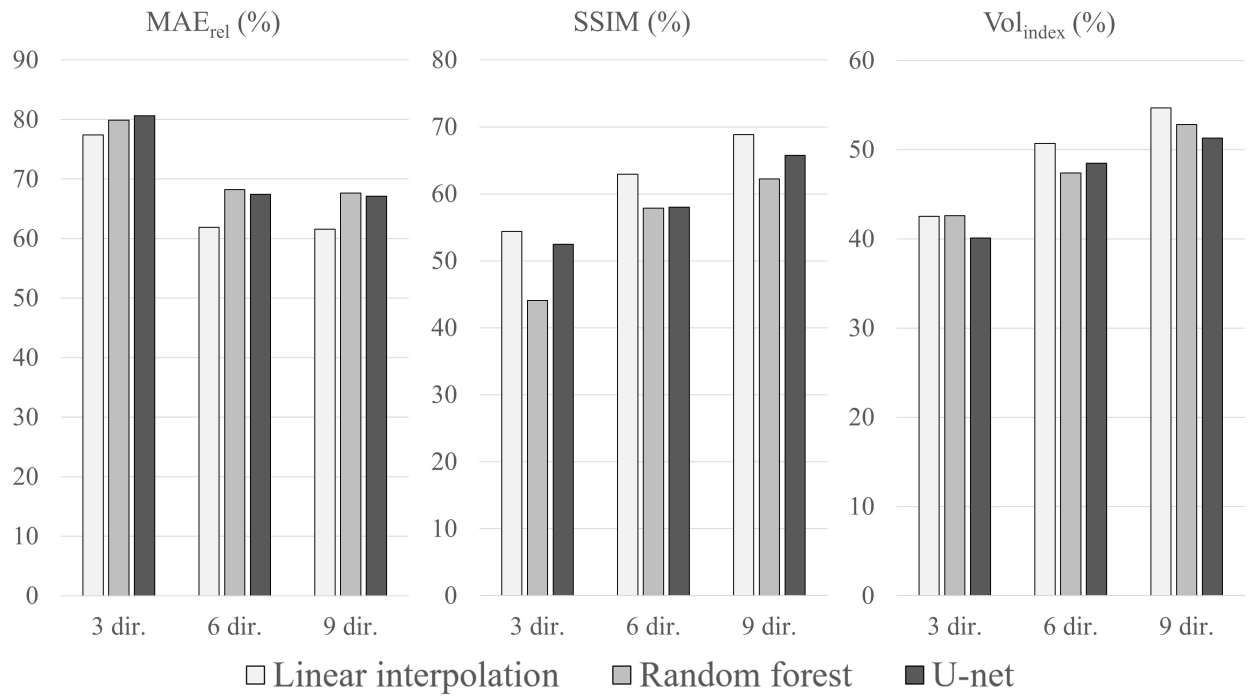


FIGURE 14.8: Résultats de l'interpolation pour les trois métriques d'image proposées, pour les différents discrétisatiosn de rose des vents utilisés pour l'interpolation pour les différentes méthode d'interpolation.

Ces résultats lorsque comparés à l'interpolation sont équivalents à ce qu'il est possible d'obtenir à partir de l'interpolation de trois simulations. Cependant, l'algorithme n'a demandé aucune simulation de ce quartier précis, les résultats sont donc encourageant quant à la capacité de méthode d'apprentissage profond à apprendre la dispersion de polluant et à l'extrapoler à des quartiers inconnus.

14.3.2 Génération de données

Comme vu précédemment, le nombre d'exemples est un paramètre primordial pour qu'un algorithme d'apprentissage profond ait de bonnes performances. Néanmoins, réaliser des études CFD est très coûteux en ressources de calcul. Il faut donc générer des exemples de manière efficiente tout en conservant la qualité des prédictions de la CFD.

Hypothèses sur la CFD et optimisation des simulations Pour cela il convient de cadrer les hypothèses faites au niveau des simulations CFD. Celle-ci suivent les lignes directrices usuelles quant au maillage, conditions limites et dimensions de la zone modélisée dans le domaine de la qualité de l'air en zone urbaine. À cela il faut ajouter que par simplicité, les routes et les fondations des bâtiments sont au même niveau que le sol qui est plat.

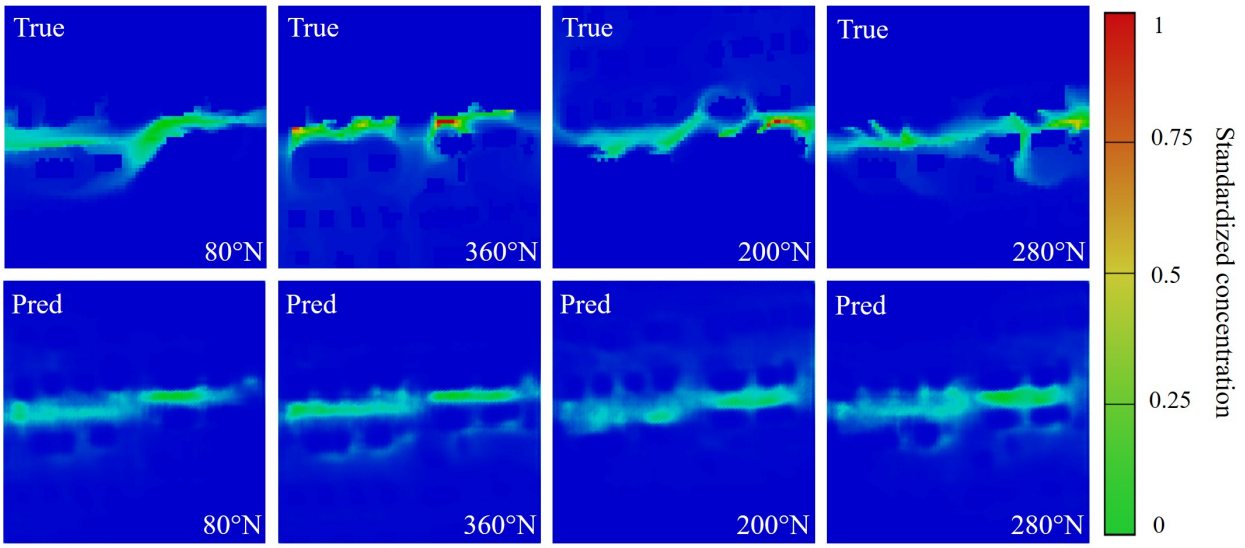


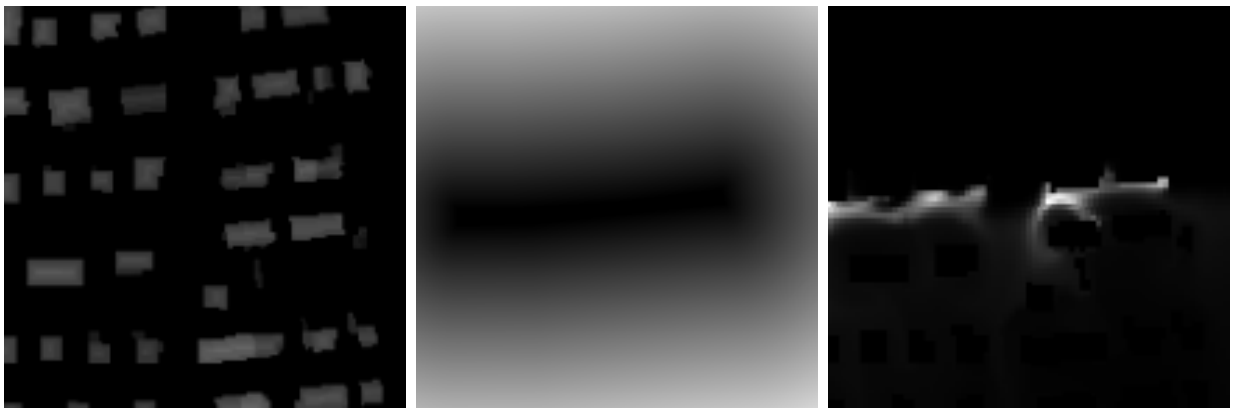
FIGURE 14.9: Exemples de prédictions U-net sur un quartier qu'il n'a jamais vu (rangée du bas) par rapport au résultat CFD (rangée du haut).

Direction	$mae_{rel}(\%)$	$vol_{index}(\%)$	$SSIM(\%)$	$\|FB\|(\%)$	VG	$NMSE$	$FAC2(\%)$
80°N	62.4	51.1	60.0	15	2.79	3.7	62
200°N	74.4	47.3	50.8	3	2.58	3.8	49
280°N	84.8	43.2	48.8	9.5	2.66	4.2	51
360°N	71.1	45.1	52.7	15	2.79	4.5	56
Mean on all 18 directions	**73.7**	**45.7**	**52.7**	**10**	**2.69**	**4.47**	**54**
std_{rel} on all 18 directions	8.6	7.1	12.5	64.6	3.7	22.2	10.4

TABLE 14.3: Évaluation des prédictions faites par le U-net pour différentes métriques : mae_{rel}, vol_{index}, $SSIM$, $|FB|$, VG, $NMSE$ and $FAC2$.

Pour réduire au maximum le temps de calcul tout en créant le plus grand nombre d'exemples quelques règles ont été mises en place. Premièrement, il convient de mettre plusieurs sources de polluant pour un quartier, ceci permet de profiter du calcul du champ de vitesses qui est l'étape la plus longue. Deuxièmement, il convient de faire de grandes zones afin de réduire proportionnellement l'espace laissé vide pour éviter les effets de bords. Troisièmement, une route longue peut être considérée comme la superposition de plusieurs petites routes. Ces trois règles permettent d'augmenter le nombre d'exemples issus d'une simulation d'un facteur x20 par rapport à faire une simulation avec un bâti et une source de polluant à petite échelle.

Conversion des données issues de la CFD en exemples pour les modèles d'apprentissage profond Les données CFD sont convertis en image car les architectures ont été construites à la base pour ce format de représentation des données et les bibliothèques associées sont rapides et optimisées. Pour se faire des traitements pour automatiser la tâche ont été codés. Les images données aux réseaux de neurones sont montrées dans la figure 14.10 ci-dessous :



(a) Buildings (b) Pollution source (c) CFD result

FIGURE 14.10: Images données en entrée du réseau (a) la hauteur, la forme et la position de chaque bâtiment dans la zone, (b) la distance de la source de pollution, et (c) la dispersion de polluant issue de la simulation CFD correspondante, considérée comme la bonne sortie pour le réseau de neurones.

14.3.3 Comparaison d'architecture d'apprentissage profond pour la dispersion de polluant

Maintenant que suffisamment d'exemples ont été générés, il est possible de faire apprendre les premiers modèles d'apprentissage profond sur plusieurs quartiers et de leur faire prédire des quartiers jamais vus. Cependant, plusieurs architectures d'apprentissage profond du même type existent et ont montré de bons résultats dans différents domaines de l'imagerie. Ainsi, il est nécessaire de comparer ces différentes architectures afin de déterminer celle qui est la plus performante pour résoudre la dispersion des polluants sur les quartiers.

Comparaison des architectures 6 architectures ont été entraînées sur le même set de données et leur prédiction sur les mêmes quartiers. Ces architectures ont ensuite été comparées sur diverses métriques liées à la qualité de l'air ou l'imagerie. Le classement des différents modèles en fonctions des métriques est présenté ci-dessous 14.11 :

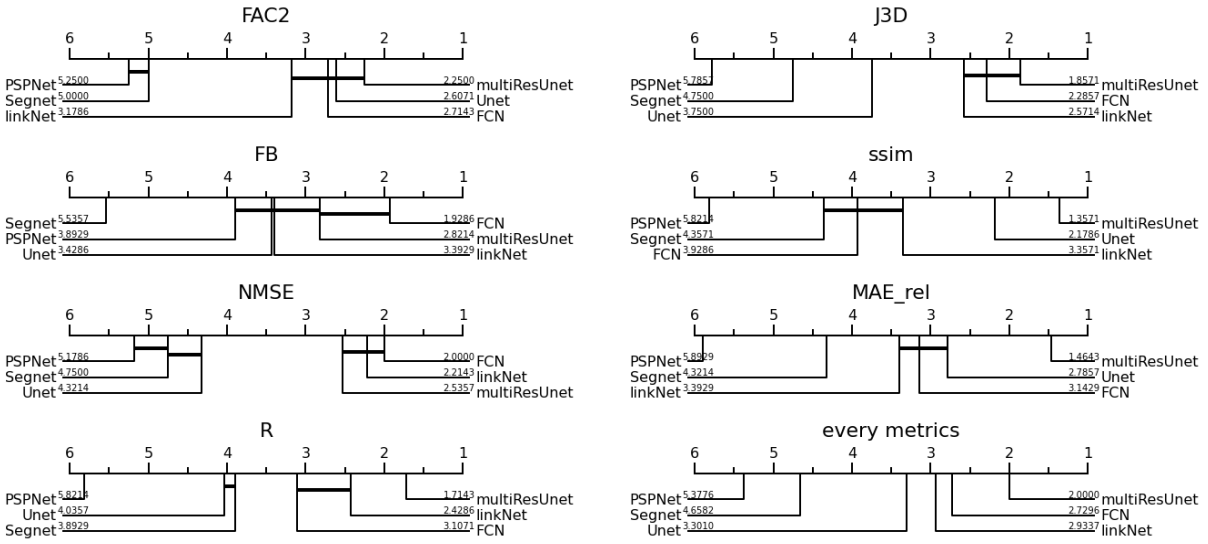


FIGURE 14.11: Classement de chaque meilleure variante pour chaque modèle en fonction de chaque métrique

L'architecture qui parvient à prédire le mieux la dispersion des polluants dans l'ensemble est le multiResUnet, qui est première 5/7 fois et toujours au moins dans le premier groupe statistiquement indiscernable. Lorsque tous les paramètres sont considérés ensemble, le multiResUnet est le premier et est statiquement discernable des autres.

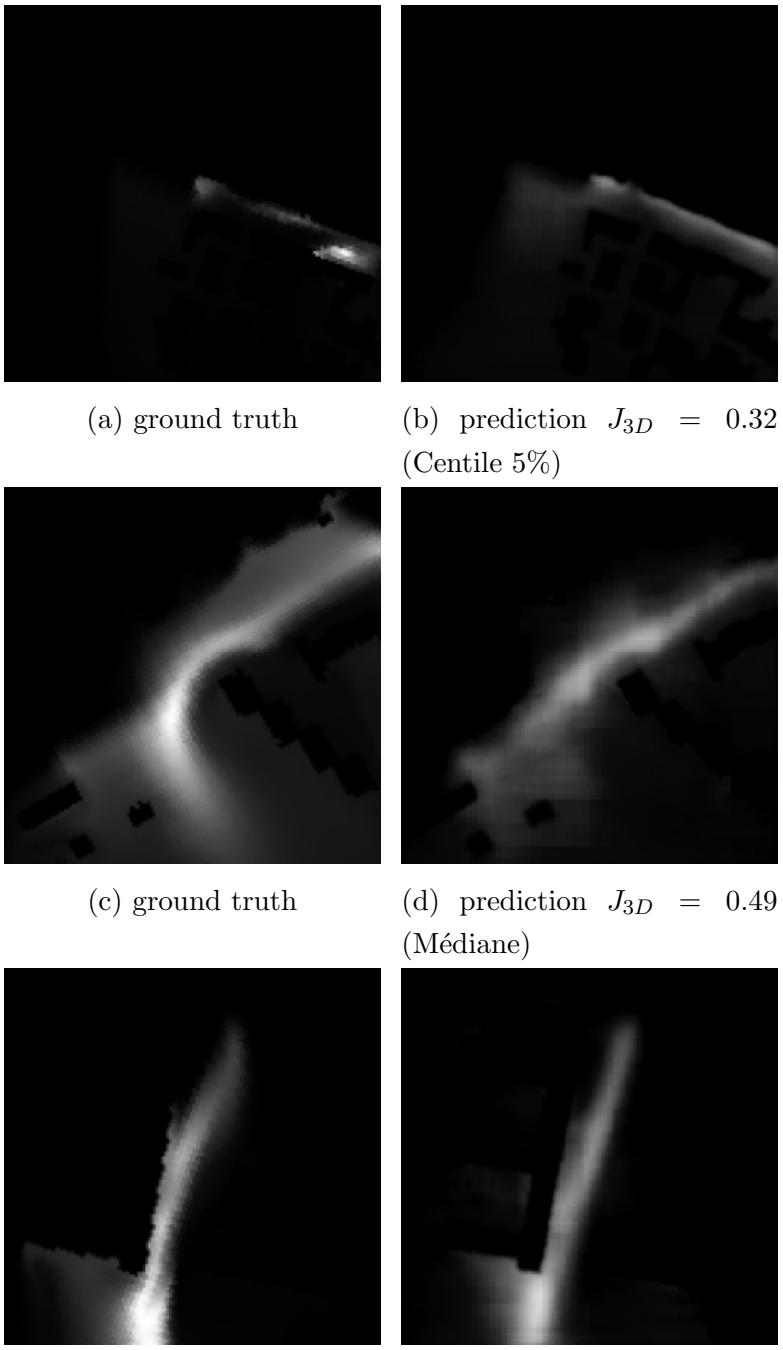
Différentes fonctions de pertes, erreur quadratique moyenne, crossentropie binaire et une fonction inventé par moi, le J_{3D} ont aussi été comparées. La fonction J_{3D} est la première au classement.

Résultat de la meilleur architecture les résultats présentés dans le tableau suivant 14.4 sont ceux de la meilleure architecture, le multiResUnet :

metric	**FAC2**	**NMSE**	**FB**	**R**	**MAE rel**	**J3D**	**ssim**
mean value	0.8	3.7	0.3	0.8	0.7	0.5	0.8
expected value	$\approx > 0.5$	$\approx < 1.5$	$\approx < 0.3$	1	0	1	1

TABLE 14.4: Évaluation des résultats du multiResUnet sur chaque métrique

Le modèle parvient à avoir deux sur trois des métriques de la qualité de l'air dans ce qui est considéré comme très satisfaisant pour un modèle de qualité de l'air. Des exemples de résultats de l'architecture sont présentés ci-contre sur la figure 14.12



(a) ground truth

(b) prediction $J_{3D} = 0.32$ (Centile 5%)

(c) ground truth

(d) prediction $J_{3D} = 0.49$ (Médiane)

(e) ground truth

(f) prediction $J_{3D} = 0.63$ (Centile 95%)

FIGURE 14.12: Exemples de prédiction du multiResUnet

14.3.4 Optimisation de l'architecture multiResUnet quant à la thématique de la dispersion de polluant

Le multiResUnet est l'architecture qui donne les meilleurs résultats selon le test effectué. Néanmoins, les architectures testées ont été à peine effacées et n'ont pas été étudiées en détail. En effet, les architectures testées n'ont pas été beaucoup modifiées par rapport à celles des articles originaux. Est-il donc possible d'améliorer encore les performances de la meilleure architecture en optimisant ses paramètres de formation ou en modifiant l'architecture elle-même pour qu'elle réponde mieux à nos besoins ?

Prise en compte du bâti environnant la zone d'intérêt L'aéraulique d'un quartier est déterminée en partie par les bâtiments qui le composent et qui constituent des obstacles pour le vent. Il est donc nécessaire de prendre en compte les bâtiments situés autour de la zone d'intérêt pour les calculs. Pour ce faire, en CFD, il est d'usage de modéliser une zone plus grande que la zone d'intérêt.

Dans sa version originale le multiResUnet a les mêmes dimensions entre son entrée et sa sortie. Plusieurs modifications ont été apportée à l'architecture pour lui permettre de donner en entrée une image couvrant une zone deux fois plus grande que son entrée.

La modification apportant les meilleurs résultats est celle utilisant du remplissage (padding) associé avec un un recadage (cropping) au milieu de l'architecture au niveau du goulot d'étrangement (bottleneck) et sur les ResPath.

Optimisation de l'architecture Pour améliorer les résultats de l'architecture, certaines techniques ont été utilisées. L'ajout de porte d'attention (attention gate) supposé aider l'architecture à se concentrer sur les parties de l'image les plus importantes quant aux problèmes à résoudre a été testé mais n'a malheureusement pas améliorer les résultats.

Une deuxième optimisation qui a été testée est l'hyper-réglage (hypertunning) des hyper paramètres. Un hyper paramètre étant un paramètre qui n'est pas entrainé par l'architecture mais qui influe tout de même sur le résultat de celle-ci. Pour se faire l'infrastructure logicielle (framework) keras tuner avec l'algorithme hyperband a été utilisé. Les hyper paramètres testés sont la profondeur du réseau, le nombre de filtres, le pourcentage de décrochement (dropout), les fonctions d'activation dans l'architecture et à la dernière couche et l'optimiseur. Parmi ces ensembles de paramètres, le meilleur ensemble selon l'algorithme hyperband est une profondeur de 4, un nombre minimal de filtre de 12, un pourcentage de décrochement de 0,2, relu comme fonction d'activation à l'intérieur de l'architecture et au niveau de la dernière couche et l'optimiseur adamax. Cet ensemble d'hyperparamètres n'est pas éloigné de celui déjà utilisé auparavant. Ceci permet donc plutôt de les confirmer.

TABLE 14.5: Résultat de l'architecture 2D et 3D sur plusieurs métriques à une hauteur de 1,5m

model	$FAC2$	$NMSE$	FB	R	NAE	MAE_rel	$J3D$	$ssmi$
3D	0.82	2.28	0.15	0.80	0.57	0.58	0.56	0.75
2D	0.88	3.35	0.15	0.78	0.59	0.60	0.55	0.85

Ouverture sur la version 3D de l'architecture Sur le plan de la ressource en calcul, l'architecture 3D nécessite beaucoup plus de ressources. En effet, elle évoluera linéairement en fonction du nombre de plans z par rapport à sa version 2D. L'architecture impose une contrainte sur le nombre minimal de plans z requis. En effet, 16(=2^4) est la dimension minimale requise à l'entrée de l'architecture pour faire du calcul convolutif 3D en utilisant le multiResUnet avec une profondeur de 4 couches. En effet, il divise les dimensions des tenseurs d'entrée (x,y,z) 2^4 fois. Ainsi, l'architecture multiResUnet 3D sera au moins 16 fois plus grande que la version 2D si tous les autres paramètres sont maintenus constants. Néanmoins, en raison de la limitation de la carte graphique dont nous disposons, il est nécessaire de modifier certains des paramètres pour pouvoir disposer d'une architecture fonctionnelle. Il serait également préférable de garder un rapport R_{mpx} aussi constant que possible entre les axes x,y et z.

En utilisant la quantité minimale de 16 plans z pour 13,5 m, on obtient un R_{mpx} de 0,84375. Comme les données couvrent 300, 300*0.84375=252. Or 252 n'est pas divisible par 2^4 donc 256 sera retenu. L'entrée de l'architecture sera donc 256x256x16 et la sortie 128x128x16.

Comparée à la version 2D, elles ont toutes deux des performances similaires pour la même hauteur (1,5 m), comme on peut le voir sur le tableau suivant 14.5 :

14.3.5 Une étude de cas pour estimer la pollution urbaine en temps réel

Plusieurs modifications et optimisations ont été apportées à l'architecture pour améliorer sa précision. À présent, il convient de voir comment l'architecture de la version optimisée peut fonctionner sur un cas de test réel et comment elle peut être utilisée dans un système plus large pour être capable de faire des prédictions de dispersion de polluants en temps réel.

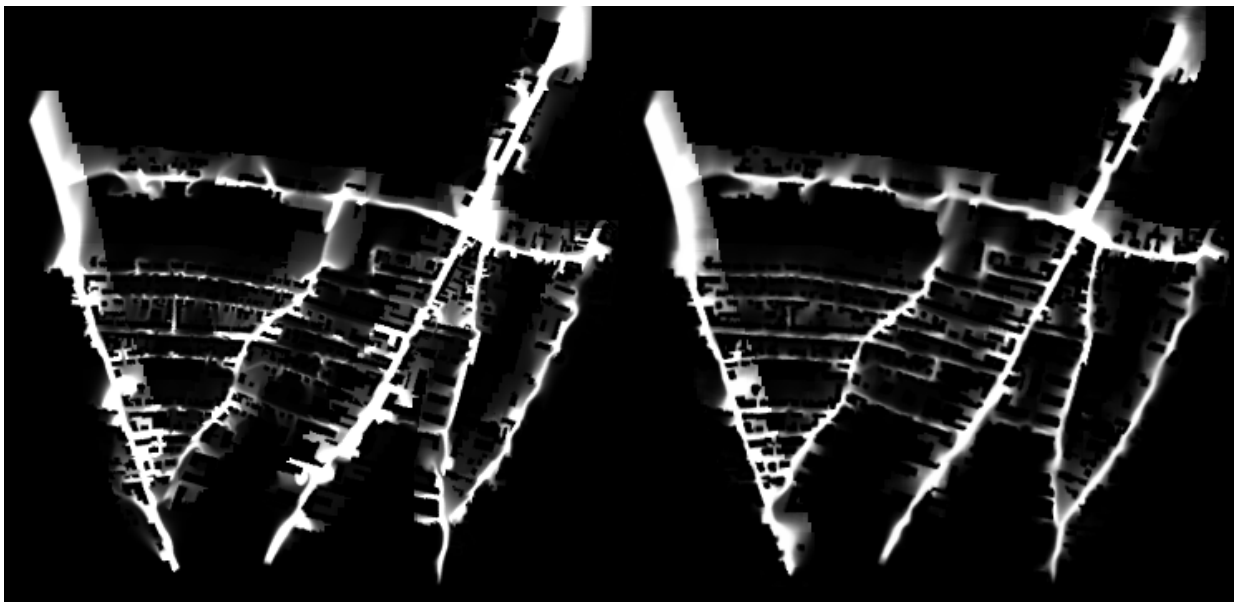
Principe du système temps réel Le système temps réel utilise des remontées d'informations sur la météo, le trafic et des capteurs de pollution de fond pour déterminer la pollution en temps réel en zone urbaine. Il est possible aussi que la pollution provienne d'autres sources locales auquel cas des modélisations annexes seront nécessaires. La concentration en un point

donné est déterminée par l'équation suivante 14.15 :

$$C(x,y,t) = C^b(t) + C^l(x,y,t) = C^b(t) + \sum_{i=0}^{N_r} C_i^{DL}(x,y,t) * E_i^r(t) + \sum_{j=0}^{N_p} C_j^M(x,y,t) * E_j^p(t) \tag{14.15}$$

avec C^b la pollution de fond,$C(x,y)^l$ la pollution issue de sources locales, $C_i^{DL}(x,y)$ la pollution provenant de la route i au point (x,y) déterminée par le modèle d'apprentissage profond, N_r le nombre de routes voisines ayant un impact au niveau du point (x, y), E_i^r l'émission de la route i,N_p le nombre de sources ponctuelles voisines impactant (x,y), $C_j^M(x,y)$ la pollution déterminée par le modèle au point (x,y), E_j^p l'émission de la source ponctuelle j.

Le cas d'étude Ce modèle a été utilisé in situ sur une zone de $1km^2$ avec des valeurs réelles provenant de capteurs de trafic et de météorologie dans les environs de Strasbourg (France) dans la ville de Schitilgheim et comparé aux résultats CFD équivalents. Le résultat est présenté ci-dessous 14.13



(a) CFD result (b) MultiresUNet result

FIGURE 14.13: Cartes du district étudié et comparaison des deux résultats

Il peut être compliqué de comparer formellement les résultats entre la CFD et le réseau d'apprentissage profond puisque la CFD détermine la dispersion en 3D alors que l'approche d'apprentissage profond travaille en 2D uniquement et à une hauteur donnée. Néanmoins, la CFD a nécessité une semaine de calcul sur 96 CPU alors que le réseau d'apprentissage

profond a nécessité environ 3 minutes sur une carte graphique GTX 1080Ti, ce qui représente un gain de vitesse de x3000.

Pour évaluer la précision des prédictions, les métriques ci-dessus ont été calculées entre la prédiction du modèle IA et la CFD considérée comme la référence et sont présentées ci-dessous sur le tableau 14.6.

Metrics	$FAC2$	$NMSE$	FB	R	MAE_{rel}	J_{3D}	$SSIM$
Score	**0.818**	1.565	**0.176**	0.851	0.431	0.620	0.768
Expected values	> 0.5	< 1.5	< 0.3	1	0	1	1

TABLE 14.6: Évaluation de la qualité du modèle de dispersion donné par l'approche d'apprentissage profond.

14.3.6 Conclusion et perspectives

La pollution de l'air affecte l'environnement comme la qualité de l'eau avec les pluies acides ou le rendement des cultures. Mais, avant tout, les enjeux de la qualité de l'air portent sur la santé et le bien-être des habitants. En effet, la qualité de l'air peut avoir un impact considérable sur l'espérance de vie et les décès prématurés, notamment dans les zones urbaines. Pour cela au cours de la thèse, de nouvelles méthodes ont été développées aussi bien pour évaluer la pollution moyenne annuelle que la pollution en temps réel.

Evaluation de l'exposition annuelle Dans un premier temps, une méthode a été développée pour déterminer la pollution moyenne annuelle dans une zone en utilisant des modèles numériques et la rose des vents. Cette méthode utilise la fréquence des directions et des vitesses pour moyenner les résultats de différentes simulations. En outre, différentes discrétisations de la rose des vents ont été testées afin de connaître l'erreur engendrée par une discrétisation plus grossière mais moins coûteuse. De cette étude est sortie une méthodologie afin de déterminer l'erreur commise par une discrétisation de 6 ou 9 directions par rapport à une discrétisation de 18 directions. Dans un second temps, des lois ont été proposés pour réduire la durée de mesures pour déterminer la moyenne annuelle des capteurs de NO_2, NO_X, PM_{10} et $PM_{2,5}$. De plus, des lois permettant de convertir les NO_2 en NO_X et les PM_{10} et $PM_{2,5}$ ont été testées et s'accordent avec celles de précédentes études.

Evaluation de l'exposition en temps réel Dans un troisième temps, afin de pouvoir déterminer la dispersion de polluant en temps réel des méthodes d'apprentissage profond ont été appliqués. Les capacités de l'apprentissage profond ont été testées et approuvées.

Néanmoins pour exploiter le plein potentiel de l'apprentissage profond il faut générer de nombreux exemples. Pour cela, des règles ont été édictées pour créer en de manière efficiente à partir de la CFD. Une fois que suffisamment d'exemples ont été créés, 6 différentes architectures populaires pour le traitement d'information spatiale ont été testé. De cette comparaison, il ressort que le multiResUnet est la meilleure architecture en association avec une fonction de perte créée dans le cadre de cette thèse, le J_{3D}. Une fois que cette architecture a été sélectionnée elle a été optimisée. Elle a pour finir, était comparée sur un cas avec des données réelles par rapport au résultat de la CFD. Le système temps réel utilisant l'apprentissage profond parvient à avoir des résultats satisfaisants sur toutes les métriques de la qualité de l'air et obtient un score final J_{3D} de 62 %. Ce résultat étant obtenu en quelques minutes contre plusieurs jours de calculs pour le résultat CFD.

Bibliography

[1] Pollution de l'air ambiant : nouvelles estimations de son impact sur la santé des français.

[2] Mar 2013.

[3] Abdel-Hamid, O., Deng, L., and Yu, D. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech 2013* (August 2013), ISCA.

[4] ADEME, o. w. Enquête "les français et l'environnement" : vague 4, Oct. 2017.

[5] Ahn, J., Shin, D., Kim, K., and Yang, J. Indoor air quality analysis using deep learning with sensor data. *Sensors 17*, 11 (Oct. 2017), 2476.

[6] AIRPARIF. Inventaire régional des émissions en Île-de-France - Année de référence 2012 - éléments synthétiques - Édition mai 2016. Tech. rep., 2016.

[7] Akatsu, M. The problem of air pollution during the industrial revolution: A reconsideration of the enactment of the smoke nuisance abatement act of 1821. In *Monograph Series of the Socio-Economic History Society, Japan*. Springer Japan, 2015, pp. 85–109.

[8] Allegrini, J., Dorer, V., and Carmeliet, J. Coupled CFD, radiation and building energy model for studying heat fluxes in an urban environment with generic building configurations. *Sustainable Cities and Society 19* (Dec. 2015), 385–394.

[9] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang,

Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 173–182.

[10] Atkinson, R. W., Kang, S., Anderson, H. R., Mills, I. C., and Walton, H. A. Epidemiological time series studies of PM2.5and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax 69*, 7 (Apr. 2014), 660–665.

[11] Badrinarayanan, V., Kendall, A., and Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 12 (Dec. 2017), 2481–2495.

[12] Bady, M. Evaluation of Gaussian Plume Model against CFD Simulations through the Estimation of CO and NO Concentrations in an Urban Area. *American Journal of Environmental Sciences 13*, 2 (Feb. 2017), 93–102.

[13] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

[14] Belis, C., Karagulian, F., Larsen, B., and Hopke, P. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in europe. *Atmospheric Environment 69* (Apr. 2013), 94–108.

[15] Bibri, S. E., and Krogstie, J. Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society 31* (May 2017), 183–212.

[16] Blocken, B. 50 years of Computational Wind Engineering: Past, present and future. *Journal of Wind Engineering and Industrial Aerodynamics 129* (June 2014), 69–102.

[17] Blocken, B., van der Hout, A., Dekker, J., and Weiler, O. CFD simulation of wind flow over natural complex terrain: Case study with validation by field measurements for Ria de Ferrol, Galicia, Spain. *Journal of Wind Engineering and Industrial Aerodynamics 147* (Dec. 2015), 43–57.

[18] Bosanquet, C. H., and Pearson, J. L. The spread of smoke and gases from chimneys. *Trans. Faraday Soc. 32* (1936), 1249–1263.

[19] Boyer, V., and Baz, D. E. Recent advances on GPU computing in operations research. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum* (May 2013), IEEE.

[20] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[21] Brenner, M. P., Eldredge, J. D., and Freund, J. B. Perspective on machine learning for advancing fluid mechanics. *Physical Review Fluids 4*, 10 (Oct. 2019).

[22] Bright, V. B., Bloss, W. J., and Cai, X. Urban street canyons: Coupling dynamics, chemistry and within-canyon chemical processing of emissions. *Atmospheric Environment 68* (Apr. 2013), 127–142.

[23] Buccolieri, R., Santiago, J.-L., Rivas, E., and Sanchez, B. Review on urban tree modelling in CFD simulations: Aerodynamic, deposition and thermal effects. *Urban Forestry & Urban Greening 31* (Apr. 2018), 212–220.

[24] Buzzicotti, M., Bonaccorso, F., Leoni, P. C. D., and Biferale, L. Reconstruction of turbulent data with deep generative models for semantic inpainting from TURB-rot database. *Physical Review Fluids 6*, 5 (May 2021).

[25] Bächlin, W., Bösinger, R., Brandt, A., and Schultz, T. Überprüfung des NO-NO2-Umwandlungsmodells für die Anwendung bei Immissionsprognosen für bodennahe Stickoxidfreisetzung. *Reinhaltung der Luft 66* (2008), 154–157.

[26] Chaloulakou, A., Mavroidis, I., and Gavriil, I. Compliance with the annual NO2 air quality standard in Athens. Required NOx levels and expected health implications. *Atmospheric Environment 42*, 3 (Jan. 2008), 454–465.

[27] Chang, J. C., and Hanna, S. R. Air quality model performance evaluation. *Meteorology and Atmospheric Physics 87*, 1-3 (Sept. 2004).

[28] Chauhan, A. J., Krishna, M. T., Frew, A. J., and Holgate, S. T. Exposure to nitrogen dioxide (NO2) and respiratory disease risk. *Reviews on Environmental Health 13*, 1-2 (June 1998), 73–90.

[29] Chaurasia, A., and Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)* (Dec. 2017), IEEE.

[30] Chen, G., Long, T., Xiong, J., and Bai, Y. Multiple random forests modelling for urban water consumption forecasting. *Water Resources Management 31*, 15 (Sept. 2017), 4715–4729.

[31] Chen, L., Ding, Y., Lyu, D., Liu, X., and Long, H. Deep Multi-Task Learning Based Urban Air Quality Index Modelling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 1 (Mar. 2019), 1–17.

[32] Chen, T.-M., Kuschner, W. G., Gokhale, J., and Shofer, S. Outdoor air pollution: Nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *The American Journal of the Medical Sciences 333*, 4 (Apr. 2007), 249–256.

[33] Chikitkin, A., and Noskov, F. Accelerating explicit method for poisson's equation using machine learning techniques. In *PROCEEDINGS OF THE X ALL-RUSSIAN CONFERENCE "Actual Problems of Applied Mathematics and Mechanics" with International Participation, Dedicated to the Memory of Academician A.F. Sidorov and 100th Anniversary of UrFU: AFSID-2020* (2020), AIP Publishing.

[34] Derwent, R., and Middleton, D. R. An empirical function for the ratio [NO2]:[NOx]. *Clean Air 26* (1996), 57–60.

[35] Di Sabatino, S., Buccolieri, R., Pulvirenti, B., and Britter, R. Simulations of pollutant dispersion within idealised urban-type geometries with CFD and integral models. *Atmospheric Environment 41*, 37 (Dec. 2007), 8316–8329.

[36] Du, S., Li, T., Yang, Y., and Horng, S. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge & Data Engineering 33*, 06 (jun 2021), 2412–2424.

[37] EU. *Directive 2008/50/EC of the european parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe.* European Union. 2008.

[38] European Environment Agency. *Air quality in Europe: 2019 report.* 2019. OCLC: 1127832561.

[39] Fan, W., Ma, Y., Li, Q., Wang, J., Cai, G., Tang, J., and Yin, D. A graph neural network framework for social recommendations. *IEEE Transactions on Knowledge and Data Engineering* (2020), 1–1.

[40] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery 33*, 4 (Mar. 2019), 917–963.

[41] France, S. P. *Impacts sanitaires de la pollution de l'air en France : nouvelles données et perspectives.* Communiqué de presse. 2016.

[42] Franke, J., Hellsten, A., Schlünzen, H., and Carissimo, B. Best practice guideline for the CFD simulation of flows in the urban environment. *COST Action 732* (2007).

[43] Gómez-Losada, Á., Santos, F. M., Gibert, K., and Pires, J. C. A data science approach for spatiotemporal modelling of low and resident air pollution in madrid (spain): Implications for epidemiological studies. *Computers, Environment and Urban Systems 75* (May 2019), 1–11.

[44] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

[45] Grekousis, G. Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems 74* (Mar. 2019), 244–256.

[46] Guo, X., Li, W., and Iorio, F. Convolutional Neural Networks for Steady Flow Approximation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (San Francisco, California, USA, 2016), ACM Press, pp. 481–490.

[47] Hamanaka, R. B., and Mutlu, G. M. Particulate matter air pollution: Effects on the cardiovascular system. *Frontiers in Endocrinology 9* (Nov. 2018).

[48] Harrison, R., Pope, F., and Shi, Z. Air pollution. In *Reference Module in Earth Systems and Environmental Sciences*. Elsevier, 2014.

[49] Harrison, R. M., Deacon, A. R., Jones, M. R., and Appleby, R. S. Sources and processes affecting concentrations of PM10 and PM2.5 particulate matter in birmingham (u.k.). *Atmospheric Environment 31*, 24 (Dec. 1997), 4103–4117.

[50] Harrison, R. M., Jones, A. M., and Lawrence, R. G. Major component composition of PM10 and PM2.5 from roadside and urban background sites. *Atmospheric Environment 38*, 27 (Sept. 2004), 4531–4538.

[51] Ibtehaz, N., and Rahman, M. S. MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Networks 121* (Jan. 2020), 74–87. arXiv: 1902.04049.

[52] Jain, R. *Environmental impact of mining and mineral processing : management, monitoring, and auditing strategies*. Butterworth-Heinemann, Oxford, UK, 2016.

[53] Jenkin, M. E. Analysis of sources and partitioning of oxidant in the UK—Part 1: the NOX-dependence of annual mean concentrations of nitrogen dioxide and ozone. *Atmospheric Environment 38*, 30 (Sept. 2004), 5117–5129.

[54] Junfeng Chen, Jonathan Viquerat, E. H. U-net architectures for fast prediction in fluidmechanics. *hal-02401465* (2019).

[55] Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Assessment of capability of deep learning to predict air pollution dispersion. *Submitted to Computers, Environment and Urban Systems*.

[56] Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Deep learning methods evaluation to predict air quality based on computational fluid dynamics. *Submitted to Expert Systems with Application*.

[57] Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., and Wemmert, C. Deep learning associated with computational fluid dynamics to predict pollution concentration fields in urban areas. In *Proceedings of the Upper Rhine-AI Conference* (October 2021).

[58] Jurado, X., Reiminger, N., Maurer, L., Vazquez, J., and Wemmert, C. Assessment of mean annual pm_{10} and $pm_{2.5}$ concentration based on a partial dataset. *Submitted to Sustainable Cities and Society*.

[59] Jurado, X., Reiminger, N., Vazquez, J., and Wemmert, C. On the minimal wind directions required to assess mean annual air pollution concentration based on CFD results. *Sustainable Cities and Society 71* (Aug. 2021), 102920.

[60] Jurado, X., Reiminger, N., Vazquez, J., Wemmert, C., Dufresne, M., Blond, N., and Wertel, J. Assessment of mean annual NO2 concentration based on a partial dataset. *Atmospheric Environment 221* (Jan. 2020), 117087.

[61] Kagawa, J. Evaluation of biological significance of nitrogen oxides exposure. *The Tokai Journal of Experimental and Clinical Medicine 10*, 4 (Aug. 1985), 348–353.

[62] Karagulian, F., Belis, C. A., Dora, C. F. C., Prüss-Ustün, A. M., Bonjour, S., Adair-Rohani, H., and Amann, M. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmospheric Environment 120* (Nov. 2015), 475–483.

[63] Kendrick, C. M., Koonce, P., and George, L. A. Diurnal and seasonal variations of NO, NO 2 and PM 2.5 mass as a function of traffic volumes alongside an urban arterial. *Atmospheric Environment 122* (Dec. 2015), 133–141.

[64] Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Bulbul, B. A., and Ekmis, M. A. An improved demand forecasting model using deep

learning approach and proposed decision integration strategy for supply chain. *Complexity 2019* (Mar. 2019), 1–15.

[65] Kim, M. J., Park, R. J., and Kim, J.-J. Urban air quality modeling with full O3–NOx–VOC chemistry: Implications for O3 and PM air quality in a street canyon. *Atmospheric Environment 47* (Feb. 2012), 330–340.

[66] Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., and Hoyer, S. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences 118*, 21 (May 2021), e2101784118.

[67] Korsakissok, I. *Changements d'échelle en modélisation de la qualité de l'air et estimation des incertitudes associées*. PhD thesis, Paris-Est, Université de Paris-Est, 2009.

[68] Koutsourakis, N., Bartzis, J. G., and Markatos, N. C. Evaluation of Reynolds stress, k-ε and RNG k-ε turbulence models in street canyon flows using various experimental datasets. *Environmental Fluid Mechanics 12*, 4 (Aug. 2012), 379–403.

[69] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Curran Associates, Inc.

[70] Kumar, M. B. H., Balasubramaniyan, S., Padmanaban, S., and Holm-Nielsen, J. B. Wind Energy Potential Assessment by Weibull Parameter Estimation Using Multiverse Optimization Method: A Case Study of Tirumala Region in India. *Energies 12*, 11 (June 2019), 2158.

[71] Kumar, P., Feiz, A.-A., Ngae, P., Singh, S. K., and Issartel, J.-P. CFD simulation of short-range plume dispersion from a point release in an urban like environment. *Atmospheric Environment 122* (Dec. 2015), 645–656.

[72] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521*, 7553 (May 2015), 436–444.

[73] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation 1*, 4 (Dec. 1989), 541–551.

[74] Lee, C. Impacts of urban form on air quality in metropolitan areas in the united states. *Computers, Environment and Urban Systems 77* (Sept. 2019), 101362.

[75] Lee, E. S., Ranasinghe, D. R., Ahangar, F. E., Amini, S., Mara, S., Choi, W., Paulson, S., and Zhu, Y. Field evaluation of vegetation and noise barriers for mitigation of near-freeway air pollution under variable wind conditions. *Atmospheric Environment 175* (Feb. 2018), 92–99.

[76] Leighton, P. *Photochemistry of air pollution*. No. 9 in Physical chemistry. New-York Acad, 1961.

[77] Lelieveld, J., Klingmüller, K., Pozzer, A., Pöschl, U., Fnais, M., Daiber, A., and Münzel, T. Cardiovascular disease burden from ambient air pollution in europe reassessed using novel hazard ratio functions. *European Heart Journal 40*, 20 (Mar. 2019), 1590–1596.

[78] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization.

[79] Likens, G. E., Wright, R. F., Galloway, J. N., and Butler, T. J. Acid Rain. *Scientific American 241*, 4 (1979), 43–51.

[80] Ling, J., Kurzawski, A., and Templeton, J. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics 807* (2016), 155–166.

[81] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis 42* (2017), 60–88.

[82] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015), IEEE.

[83] Lorenzini, G., and Saitanis, C. Ozone: A novel plant "pathogen". In *Abiotic Stresses in Plants*. Springer Netherlands, 2003, pp. 205–229.

[84] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. Pose guided person image generation. In *Advances in Neural Information Processing Systems* (2017), pp. 406–416.

[85] Mahata, S. K., Das, D., and Bandyopadhyay, S. MTIL2017: Machine translation using recurrent neural network on statistical machine translation. *Journal of Intelligent Systems 28*, 3 (July 2019), 447–453.

[86] Mahmood, F. H., Resen, A. K., and Khamees, A. B. Wind characteristic analysis based on Weibull distribution of Al-Salman site, Iraq. *Energy Reports* (Oct. 2019), S2352484719308716.

[87] Makkar, S., Devi, G. N. R., and Solanki, V. K. Applications of machine learning techniques in supply chain optimization. In *ICICCT 2019 – System Reliability, Quality Control, Safety, Maintenance and Management*. Springer Singapore, June 2019, pp. 861–869.

[88] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health 8* (Feb. 2020).

[89] Maurer, L., Villette, C., Reiminger, N., Jurado, X., Laurent, J., Nuel, M., Mosé, R., Wanko, A., and Heintz, D. Distribution and degradation trend of micropollutants in a surface flow treatment wetland revealed by 3d numerical modelling combined with LC-MS/MS. *Water Research 190* (Feb. 2021), 116672.

[90] Mavroidis, I., and Ilia, M. Trends of NOx, NO2 and O3 concentrations at three different types of air quality monitoring stations in Athens, Greece. *Atmospheric Environment 63* (Dec. 2012), 135–147.

[91] Maynard, R. *Evolution of WHO air quality guidelines past, present and future*. WHO Regional Office for Europe, Copenhagen, 2017.

[92] Mazzoldi, A., Hill, T., and Colls, J. J. CFD and Gaussian atmospheric dispersion models: A comparison for leak from carbon dioxide transportation and storage facilities. *Atmospheric Environment 42*, 34 (Nov. 2008), 8046–8054.

[93] Mcculloch, W., and Pitts, W. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5* (1943), 127–147.

[94] Medvedev, A. V., Agoureeva, G. I., and Murro, A. M. A long short-term memory neural network for the detection of epileptiform spikes and high frequency oscillations. *Scientific Reports 9*, 1 (Dec. 2019).

[95] Mensink, C., Lefebre, F., Janssen, L., and Cornelis, J. A comparison of three street canyon models with measurements at an urban station in antwerp, belgium. *Environmental Modelling & Software 21*, 4 (Apr. 2006), 514–519.

[96] Michelot, N., Carrega, P., and Rouïl, L. Panorama de la modélisation de la dispersion atmosphérique Atmospheric dispersion models: An overwiew. *POLLUTION ATMOSPHÉRIQUE* (2015), 9.

[97] Mosley, S. Environmental history of air pollution and protection. In *Environmental History*. Springer International Publishing, 2014, pp. 143–169.

[98] Müller, D., Ehlen, A., and Valeske, B. Convolutional neural networks for semantic segmentation as a tool for multiclass face analysis in thermal infrared. *Journal of Nondestructive Evaluation 40*, 1 (Jan. 2021).

[99] Murakami, S. Overview of turbulence models applied in CWE–1997. *Journal of Wind Engineering and Industrial Aerodynamics 74-76* (Apr. 1998), 1–24.

[100] Navas, M., Jiménez, A., and Galán, G. Air analysis: determination of nitrogen compounds by chemiluminescence. *Atmospheric Environment 31*, 21 (Nov. 1997), 3603–3608.

[101] Newell, K., Kartsonaki, C., Lam, K. B. H., and Kurmi, O. P. Cardiorespiratory health effects of particulate ambient air pollution exposure in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Planetary Health 1*, 9 (Dec. 2017), e368–e380.

[102] Niu, T., Chen, Y., and Yuan, Y. Measuring urban poverty using multi-source data and a random forest algorithm: A case study in guangzhou. *Sustainable Cities and Society 54* (Mar. 2020), 102014.

[103] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. Audio-visual speech recognition using deep learning. *Applied Intelligence 42*, 4 (Dec. 2014), 722–737.

[104] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. Attention u-net: Learning where to look for the pancreas, 2018.

[105] Pan, Z., Zhang, W., Liang, Y., Zhang, W., Yu, Y., Zhang, J., and Zheng, Y. Spatio-temporal meta learning for urban traffic prediction. *IEEE Transactions on Knowledge and Data Engineering* (2020), 1–1.

[106] Pant, P., and Farimani, A. Deep learning for efficient reconstruction of high-resolution turbulent dns data. *ArXiv abs/2010.11348* (2020).

[107] Papageorgakis, G. C., and Assanis, D. N. COMPARISON OF LINEAR AND NONLINEAR RNG-BASED k-epsilon MODELS FOR INCOMPRESSIBLE TURBULENT FLOWS. *Numerical Heat Transfer, Part B: Fundamentals 35*, 1 (Feb. 1999), 1–22.

[108] Park, Y., and Guldmann, J.-M. Creating 3d city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems 75* (May 2019), 76–89.

[109] Pospisil, J., and Jicha, M. Particulate matter dispersion modelling along urban traffic paths. *International Journal of Environment and Pollution 40*, 1/2/3 (2010), 26.

[110] Prieler, R., Mayrhofer, M., Gaber, C., Gerhardter, H., Schluckner, C., Landfahrer, M., Eichhorn-Gruber, M., Schwabegger, G., and Hochenauer, C. CFD-based optimization of a transient heating process in a natural gas fired furnace using neural networks and genetic algorithms. *Applied Thermal Engineering 138* (June 2018), 217–234.

[111] Purvis, M. R., and Ehrlich, R. Effect of Atmospheric Pollutants on Susceptibility to Respiratory Infection: II. Effect of Nitrogen Dioxide. *The Journal of Infectious Diseases 113*, 1 (1963), 72–76.

[112] Qi, Z., Wang, T., Song, G., Hu, W., Li, X., and Zhang, Z. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Transactions on Knowledge and Data Engineering 30*, 12 (Dec. 2018), 2285–2297.

[113] Qu, Y., Milliez, M., Musson-Genon, L., and Carissimo, B. Numerical study of the thermal effects of buildings on low-speed airflow taking into account 3D atmospheric radiation in urban canopy. *Journal of Wind Engineering and Industrial Aerodynamics 104-106* (May 2012), 474–483.

[114] Rafael, S., Rodrigues, V., Oliveira, K., Coelho, S., and Lopes, M. How to compute long-term averages for air quality assessment at urban areas? *Science of The Total Environment 795* (Nov. 2021), 148603.

[115] Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics 378* (Feb. 2019), 686–707.

[116] Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations. *arXiv:1711.10561 [cs, math, stat]* (Nov. 2017). arXiv: 1711.10561.

[117] Rall, D. P. Review of the health effects of sulfur oxides. *Environmental Health Perspectives 8* (Aug. 1974), 97–121.

[118] Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Blond, N., Dufresne, M., and Wertel, J. Effects of wind speed and atmospheric stability on the air pollution reduction rate induced by noise barriers. *Journal of Wind Engineering and Industrial Aerodynamics 200* (May 2020), 104160.

[119] Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Blond, N., Dufresne, M., and Wertel, J. Effects of wind speed and atmospheric stability on the air pollution reduction rate induced by noise barriers. *Journal of Wind Engineering and Industrial Aerodynamics 200* (May 2020), 104160.

[120] Reiminger, N., Jurado, X., Vazquez, J., Wemmert, C., Dufresne, M., Blond, N., and Wertel, J. Methodologies to assess mean annual air pollution concentration combining numerical results and wind roses. *Sustainable Cities and Society 59* (Aug. 2020), 102221.

[121] Reiminger, N., Vazquez, J., Blond, N., Dufresne, M., and Wertel, J. How pollutant concentrations evolve in step-down street canyons as a function of buildings geometric properties.

[122] Reiminger, N., Vazquez, J., Blond, N., Dufresne, M., and Wertel, J. CFD evaluation of mean pollutant concentration variations in step-down street canyons. *Journal of Wind Engineering and Industrial Aerodynamics 196* (Jan. 2020), 104032.

[123] Reiminger, N., Vazquez, J., Blond, N., Dufresne, M., and Wertel, J. CFD evaluation of mean pollutant concentration variations in step-down street canyons. *Journal of Wind Engineering and Industrial Aerodynamics 196* (Jan. 2020), 104032.

[124] Ribeiro, M. D., Rehman, A., Ahmed, S., and Dengel, A. Deepcfd: Efficient steady-state laminar flow approximation with deep convolutional neural networks, 2020.

[125] Richards, P., and Norris, S. Appropriate boundary conditions for computational wind engineering models revisited. *Journal of Wind Engineering and Industrial Aerodynamics 99*, 4 (Apr. 2011), 257–266.

[126] Richards, P. J., and Hoxey, R. P. Appropriate boundary conditions for computational wind engineering models using the k-E turbulence model. 9.

[127] Rivas, E., Santiago, J. L., Lechón, Y., Martín, F., Ariño, A., Pons, J. J., and Santamaría, J. M. CFD modelling of air quality in Pamplona City (Spain): Assessment, stations spatial representativeness and health impacts valuation. *Science of the Total Environment* (2019), 19.

[128] Roache, P. J. Perspective: A Method for Uniform Reporting of Grid Refinement Studies. *Journal of Fluids Engineering 116*, 3 (1994), 405.

[129] Roberts–Semple, D., Song, F., and Gao, Y. Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern New Jersey. *Atmospheric Pollution Research 3*, 2 (Apr. 2012), 247–257.

[130] Romberg, E., Bösinger, R., Lohmeyer, A., and Ruhnke, R. NO-NO2-Umwandlung für die Anwendung bei Immissionsprognosen für Kfz-Abgase. *Reinhaltung der Luft 56* (1996), 215–218.

[131] ROMIEU, I., and BORJA-ABURTO, V. H. Particulate air pollution and daily mortality: Can results be generalized to latin american countries? *Salud Pública de México 39*, 5 (Sept. 1997), 403–411.

[132] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[133] Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (May 2015). arXiv: 1505.04597.

[134] Rosenblatt, F. The perceptron - a perceiving and recognizing automaton. Tech. Rep. 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York, January 1957.

[135] Salesky, S. T., Giometto, M. G., Chamecki, M., and Lehning, M. The transport and deposition of heavy particles in complex terrain: insights from an Eulerian model for large eddy simulation. *Water ressources research* (2019), 21.

[136] Sanchez, B., Santiago, J. L., Martilli, A., Martin, F., Borge, R., Quaassdorff, C., and de la Paz, D. Modelling NOX concentrations through CFD-RANS in an urban hot-spot using high resolution traffic emissions and meteorology from a mesoscale model. *Atmospheric Environment 163* (Aug. 2017), 155–165.

[137] Sanchez, B., Santiago, J.-L., Martilli, A., Palacios, M., and Kirchner, F. CFD modeling of reactive pollutant dispersion in simplified urban configurations with different chemical mechanisms. *Atmospheric Chemistry and Physics 16*, 18 (Sept. 2016), 12143–12157.

[138] Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. W. Learning to simulate complex physics with graph networks, 2020.

[139] Santiago, J., Martilli, A., and Martin, F. On Dry Deposition Modelling of Atmospheric Pollutants on Vegetation at the Microscale: Application to the Impact of Street Vegetation on Air Quality. *Boundary-Layer Meteorology 162*, 3 (Mar. 2017), 451–474.

[140] Santiago, J.-L., Martilli, A., and Martin, F. On dry deposition modelling of atmospheric pollutants on vegetation at the microscale : application to the impact of street vegetation on air quality. *Boundary-Layer Meteorology 162* (2017), 451–474.

[141] Santiago, J.-L., Martilli, A., and Martin, F. On Dry Deposition Modelling of Atmospheric Pollutants on Vegetation at the Microscale: Application to the Impact of Street Vegetation on Air Quality. *Boundary-Layer Meteorology 162*, 3 (Mar. 2017), 451–474.

[142] Schatzmann, M., and Leitl, B. Issues with validation of urban flow and dispersion CFD models. *Journal of Wind Engineering and Industrial Aerodynamics 99*, 4 (Apr. 2011), 169–186.

[143] Seinfeld, J., and Pandis, S. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 3rd Edition*, wiley-blackwell ed. 2016.

[144] Sharma, N., Jain, V., and Mishra, A. An analysis of convolutional neural networks for image classification. *Procedia Computer Science 132* (2018), 377–384.

[145] Shawi, R. E., Maher, M., and Sakr, S. Automated machine learning: State-of-the-art and open challenges. *ArXiv abs/1906.02287* (2019).

[146] Solazzo, E., Vardoulakis, S., and Cai, X. A novel methodology for interpreting air quality measurements from urban streets using CFD modelling. *Atmospheric Environment 45*, 29 (Sept. 2011), 5230–5239.

[147] Sun, Q., and Ge, Z. A survey on deep learning for data-driven soft sensors. *IEEE Transactions on Industrial Informatics 17*, 9 (Sept. 2021), 5853–5866.

[148] Sutton, O. G. A theory of eddy diffusion in the atmosphere. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 135*, 826 (Feb. 1932), 143–165.

[149] Tedjopurnomo, D. A., Bao, Z., Zheng, B., Choudhury, F., and Qin, A. K. A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE Transactions on Knowledge and Data Engineering* (2020), 1–1.

[150] Thuerey, N., Weissenow, K., Prantl, L., and Hu, X. Deep learning methods for reynolds-averaged navier–stokes simulations of airfoil flows. *AIAA Journal 58*, 1 (Jan. 2020), 25–36.

[151] Thunis, P. On the validity of the incremental approach to estimate the impact of cities on air quality. *Atmospheric Environment 173* (Jan. 2018), 210–222.

[152] Thunis, P., Miranda, A., Baldasano, J., Blond, N., Douros, J., Graff, A., Janssen, S., Juda-Rezler, K., Karvosenoja, N., Maffeis, G., Martilli, A., Rasoloharimahefa, M., Real, E., Viaene, P., Volta, M., and White, L. Overview of current regional and local scale air quality modelling practices: Assessment and planning tools in the EU. *Environmental Science & Policy 65* (Nov. 2016), 13–21.

[153] Tominaga, Y., and Stathopoulos, T. Steady and unsteady RANS simulations of pollutant dispersion around isolated cubical buildings: Effect of large-scale fluctuations on the concentration field. *Journal of Wind Engineering and Industrial Aerodynamics 165* (June 2017), 23–33.

[154] Toparlar, Y., Blocken, B., Maiheu, B., and van Heijst, G. A review on the CFD analysis of urban microclimate. *Renewable and Sustainable Energy Reviews 80* (Dec. 2017), 1613–1640.

[155] Tracey, B. D., Duraisamy, K., and Alonso, J. J. A machine learning strategy to assist turbulence model development. In *53rd AIAA Aerospace Sciences Meeting* (Jan. 2015), American Institute of Aeronautics and Astronautics.

[156] Unadkat, V., Sayani, P., Kanani, P., and Doshi, P. Deep learning for financial prediction. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)* (2018), pp. 1–6.

[157] Vachon, G., Louka, P., Rosant, J.-M., Mestayer, P. G., and Sini, J.-F. Measurements of traffic-induced turbulence within a street canyon during the nantes'99 experiment. In *Urban Air Quality — Recent Advances*. Springer Netherlands, 2002, pp. 127–140.

[158] Vardoulakis, S., Fisher, B. E., Pericleous, K., and Gonzalez-Flesca, N. Modelling air quality in street canyons: a review. *Atmospheric Environment 37*, 2 (Jan. 2003), 155–182.

[159] Vranckx, S., Vos, P., Maiheu, B., and Janssen, S. Impact of trees on pollutant dispersion in street canyons: A numerical study of the annual average effects in Antwerp, Belgium. *Science of The Total Environment 532* (Nov. 2015), 474–483.

[160] Wagner, F. H., Dalagnol, R., Tarabalka, Y., Segantine, T. Y. F., Thomé, R., and Hirye, M. C. M. U-net-id, an instance segmentation model for building extraction from satellite images—case study in the joanópolis city, brazil. *Remote Sensing 12*, 10 (2020).

[161] Wang, P., Zhao, D., Wang, W., Mu, H., Cai, G., and Liao, C. Thermal Effect on Pollutant Dispersion in an Urban Street Canyon. *International Journal of Environmental Research 5*, 3 (July 2011), 813–820.

[162] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612.

[163] WHO, Ed. *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005.* World Health Organization, 2005.

[164] WHO. Mortality and burden of disease from ambient air pollution, Global Health Observatory data, 2016.

[165] WHO. Mortality from household air pollution, Global Health Observatory data, 2016.

[166] WHO. *Evolution of WHO air quality guidelines past, present and future.* Copenhagen: WHO Regional Office for Europe. 2017. OCLC: 1075973767.

[167] Wiegand, A. N., and Bo, N. D. Review of empirical methods for the calculation of the diurnal NO2 photolysis rate coefficient. *Atmospheric Environment* (2000), 10.

[168] Yakhot, V., Orszag, S. A., Thangam, S., Gatski, T. B., and Speziale, C. G. Development of turbulence models for shear flows by a double expansion technique. *Physics of Fluids A: Fluid Dynamics 4*, 7 (July 1992), 1510–1520.

[169] Yang, J., Shi, B., Shi, Y., Marvin, S., Zheng, Y., and Xia, G. Air pollution dispersal in high density urban areas: Research on the triadic relation of wind, air pollution, and urban form. *Sustainable Cities and Society 54* (Mar. 2020), 101941.

[170] Youssef, A., Bloisi, D., Muscio, M., Pennisi, A., Nardi, D., and Facchiano, A. Deep convolutional pixel-wise labeling for skin lesion image segmentation. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (2018), 1–6.

[171] Yu, Y., Kwok, K., Liu, X., and Zhang, Y. Air pollutant dispersion around high-rise buildings under different angles of wind incidence. *Journal of Wind Engineering and Industrial Aerodynamics 167* (Aug. 2017), 51–61.

[172] Yumino, S., Uchida, T., Sasaki, K., Kobayashi, H., and Mochida, A. Total assessment for various environmentally conscious techniques from three perspectives: Mitigation of global warming, mitigation of UHIs, and adaptation to urban warming. *Sustainable Cities and Society 19* (Dec. 2015), 236–249.

[173] Zannetti, A. D. P. Air pollution modeling – an overview. chapter 2 of ambient air pollution.

[174] Zhang, J. J., Wei, Y., and Fang, Z. Ozone pollution: A major health hazard worldwide. *Frontiers in Immunology 10* (Oct. 2019).

[175] Zhang, Z., dong Song, X., ran Ye, S., wei Wang, Y., guang Huang, C., ran An, Y., and song Chen, Y. Application of deep learning method to reynolds stress models of channel flow based on reduced-order modeling of DNS data. *Journal of Hydrodynamics 31*, 1 (Dec. 2018), 58–65.

[176] Zhang, Z., Liu, Q., and Wang, Y. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters 15*, 5 (2018), 749–753.

[177] Zhao, E., Zhang, Z., and Bohlooli, N. Cost and load forecasting by an integrated algorithm in intelligent electricity supply network. *Sustainable Cities and Society 60* (Sept. 2020), 102243.

[178] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), IEEE.

[179] Zhou, X., Shen, Y., Huang, L., Zang, T., and Zhu, Y. Multi-level attention networks for multi-step citywide passenger demands prediction. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.

[180] Zhu, Q., Liao, C., Hu, H., Mei, X., and Li, H. Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing 59*, 7 (2021), 6169–6181.

Université de Strasbourg

Xavier JURADO

msii ECOLE DOCTORALE

Atmospheric pollutant dispersion estimation at the scale of the neighborhood using Sensors, Numerical and Deep Learning models

Résumé

La présente thèse est à la croisée de quatre domaines, la mécanique des fluides numériques (CFD), le data mining, le deep learning et la qualité de l'air. L'objectif de la thèse est d'évaluer l'exposition des habitants aux polluants atmosphériques en utilisant les récentes avancées en intelligence artificielle. La thèse s'articule autour des différentes échelles de temps demandées par la réglementation, allant de l'annuel au temps réel. Pour ce faire, des approches innovantes pour évaluer les concentrations annuelles moyennes ont été développées pour les outils de modélisation ainsi que pour les capteurs de pollution de l'air. Pour la modélisation, une méthodologie statistique basée sur les fréquences des roses des vents associées à un organigramme permettant de déterminer l'erreur numérique due à la discrétisation a été proposée. Pour les capteurs, des données provenant de toute la France ont été analysées pour établir la relation entre les concentrations mensuelles mesurées et les concentrations annuelles pour les particules fines et les oxydes d'azote. Pour déterminer l'exposition à la pollution en temps réel, un système prenant en compte le trafic, la météorologie et la disposition des bâtiments a été créé avec en son coeur un modèle d'apprentissage profond. Le système s'articule autour d'un modèle d'apprentissage profond. Ce modèle, multiResUnet, a été choisi après comparaison avec d'autres modèles convolutifs classiques de l'état de l'art et optimisé pour la problématique de la dispersion de polluant. Pour l'entraîner, des exemples issus de la CFD ont été générés efficacement en suivant des principes développés dans cette thèse. Le système a ensuite été appliqué sur un quartier réel de 1km^2 avec des données de trafic réels et comparé à la CFD. Il a réussi à obtenir de bonnes performances sur les mesures classiques de la qualité de l'air et à atteindre un score de similarité J_{3D} de 62%.

Mots clés : Apprentissage profond, Réseaux neuronaux convolutifs, Dynamique des fluides numérique, Extraction de données, Qualité de l'air, Analyse des données de capteurs, Environnement urbain.

Résumé en anglais

This thesis is at the crossroad of four domains, computational fluid dynamics (CFD), data mining, deep learning and air quality. The objective of the thesis is to assess dwellers' exposures to atmospheric pollutants using the recent advances in artificial intelligence. The thesis revolves around the different time scales requested by the regulations, going from annual to real time. To do so, innovative approaches to assess mean annual concentrations were developed for modeling as well as for sensors. For modeling a statistical methodology based on wind roses frequencies associated with a flowchart to determine the numerical error from the discretization was proposed. For the sensors, data from all around France were analyzed to establish relationship between measured monthly concentrations with annual ones for particulate matter and nitrogen oxides. To determine pollution exposure in real time, a system using taking into account traffic, meteorological and 3D building layout was created built around a deep learning model was created. The system revolves around a deep learning model. This model, multiResUnet, have been chosen after comparison with other classical state-of-the-art convolutional models and optimized for the dispersion pollution issue. To train it, examples from CFD were generated efficiently following guidelines developed in this thesis. The system was then applied on a real neighborhood of 1km^2 with real traffic data and compared with CFD. It managed to perform well on classical air quality metrics and reach a J_{3D} score of 62%.

Keywords : Deep Learning, Convolutional Neural Networks, Computational Fluid Dynamics, Data mining, Air quality, Sensors data analysis, Urban environment.