



Distance-based indicators for evaluating environmental multi-contamination and related exposure: how far should we go?

Quentin Tenailleau^{1,2} · Caroline Lanier^{1,3} · Julie Prud'homme^{1,2} · Damien Cuny^{1,2} · Annabelle Deram^{1,3} · Florent Occelli^{1,3}

Received: 18 April 2023 / Accepted: 23 July 2024 / Published online: 5 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Assessing environmental exposure to pollution is a challenging task, and scientists often use distance-based or proximity indicators when field or modeled data are unavailable. Although buffers are commonly used to represent the impact of a pollution source on neighboring populations, they can result in high-exposure misclassification. Euclidean distance-based indicators offer a promising alternative, but practices vary significantly in the literature. In this study, we aimed to compare several distance-based indicators for multiple environmental contaminants in an industrial and urban area. At the population's grid cell resolution of 200×200 m, we compared the distance to the closest source, the average or median distance to all sources, or a restricted number of nearby sources for six types of sources (industries, railways, rail areas, roadways, road crossings, and agricultural patches) against environmental contamination data (PM_{10} , NO_2 , and multimetallic contamination in lichens). Our findings revealed that the representativeness of contamination by indicators is significantly affected by the type and number of nearby sources considered. Specifically, we found that considering the distance to the nearest source or the average distance to all sources can lead to exposure misclassifications. The optimal correlation between distance indicators and pollutant levels was observed when considering 10–14 of the closest industrial sources, located within a 4.9- to 5.5-km radius. For rail areas, the optimal number was two to three sources within a 5.4- to 7.4-km radius. For main roads, intersections, and railways, the optimal number of sources varied depending on the pollutant, generally falling within a 3- to 9.4-km radius. Environmental contamination is influenced by the diversity of nearby sources, and considering only one source increases the risk of misclassification. Our results suggest that proximity models are still appropriate for study areas where the etiology of existing health effects is unclear, providing an exploratory analysis before more sophisticated research.

Keywords Euclidian distance · Emission sources · Environmental pollution · Proximity assessment · Exposure

Introduction

Assessing environmental exposure to air, water, and soil contamination is a significant challenge in public health risk prevention. This assessment enables the identification of populations at risk and estimation of the environment's contribution to the occurrence and aggravation of chronic diseases. However, the scientific community faces difficulties in precisely measuring the population's exposure to all environmental pollutants in large areas. In situ characterization of environmental mixtures is challenging due to high costs, logistics, and spatiotemporal representativeness of the measurement. New technologies such as micro-sensors or spatial modeling (dispersion, interpolation, or land use regression models) can provide measurements or estimates of some pollutants.

Responsible Editor: Philippe Garrigues

✉ Florent Occelli
florent.occelli@univ-lille.fr

¹ ULR 4515 – LGCGE, Laboratoire de Génie Civil et géo-Environnement, Univ. Lille, Univ. Artois, IMT Lille Douai, JUNIA, 3 Rue du Pr Laguesse BP83, F-59000 Lille, France

² UFR3S - Pharmacie, 59000 Lille, France

³ UFR3S - Ingénierie et Management de la Santé (ILIS), 59120 Nord, France

Geographic Information Systems (GIS) have expanded the capacity to assist the scientific community in developing exposure proxies. Proximity or distance-based models are widely used in exposure modeling, helping to assess and prioritize multiple environmental contaminations (typologies, pollutant families, sources, and potential exposure pathways) instead of measuring a specific pollutant in a specific medium (Hoek et al. 2018). According to Tobler's *first law of geography* (Tobler 1970), these models assume that levels of pollution tend to decline with distance from the source, and people living near emission sources are usually more highly exposed than those living further away. Proximity to emission sources has been used as a proxy for exposure since the mid-1990s and appears to offer a simple, rapid, easy-to-use, and cost-effective measure, especially in a context where GIS has become a widespread tool among both the scientific and management communities (Han et al. 2017).

Most studies employing proximity methods referenced in the literature are based on the buffer methodology (Zou et al. 2009; Chakraborty et al. 2011; Pascal et al. 2013). This method involves defining a concentric surface of variable size that represents the contaminated area around sources or the populations' environment. However, several drawbacks have been identified, which could result in high-exposure misclassification (Chakraborty et al. 2011; Pascal et al. 2013; Tenailleau et al. 2015). The main limitation of buffers is the creation of an artificial boundary effect. The buffer method requires setting a distance from the source within which people are considered exposed, while those outside are not. This creates an artificial boundary that does not accurately reflect the gradient of exposure and can result in significant misclassification. The second limitation is the lack of standardization. There is no standardized directive or consensus on setting an efficient threshold for buffer distances. This often leads to arbitrary decisions, making comparisons between studies difficult and reducing the reliability of exposure assessments. Finally, there is the homogeneity assumption. Buffers assume homogeneous exposure within the defined area, ignoring variations in pollutant dispersion influenced by distance and other factors such as topography, weather conditions, and source characteristics.

To overcome this limitation, utilizing the continuous Euclidean distance between a source and the target population is an appealing alternative to the qualitative measure derived from buffers (Chakraborty et al. 2011). Current geographic and epidemiological practices define targets as individual households, while at an ecological level, it may be the centroid or boundary of official spatial units or population grids. Distance indicators used in the literature can vary greatly, such as the distance to the nearest source, the mean distance to all sources within the territory, or the mean distance from all sources within a set radius (Kihal-Talantikite et al. 2017). However, these various practices have not been

the subject of fundamental questions such regarding their association with environmental contamination.

Which distance indicator is more relevant—the closest source, the average/median distance to all sources in the study area, or the average/median distance to several nearby sources? We hypothesize that neither the distance to the closest source nor the average distance to all sources in the territory are the most suitable indicators. In this context, what is the optimal number of sources to consider? Moreover, do these distance indicators highlight differences in the contribution of various sources to environmental contamination in a study area? Specifically, we seek to understand if the spatial distribution of contamination is more closely associated with the distance indicator for a particular type of source (industrial, road, rail, agricultural).

In this study, Euclidean distance indicators (considered a vector of atmospheric exposure) were computed between population grids (the targets) and various sources of air pollutant emissions. The correlation between these indicators and the level of environmental contamination for several pollutants modeled in the same grids was then evaluated. These issues were investigated in the specific context of the Dunkirk area in the North of France, where populations are exposed to multiple emission sources and a strong industrial influence.

Material and methods

Studied location

The Dunkerque Urban Community (CUD) is an urban area consisting of 17 municipalities located on the northern coast of France (Fig. 1). It is a medium-sized urban area with a population of 198,000 inhabitants and is home to several heavy industries (such as petrochemicals, chemicals, metallurgy, and energy production) that contribute to high levels of air pollution (Lanier et al. 2019). The average annual pollution levels for PM₁₀ (particulate matter with an aerodynamic diameter $\leq 10 \mu\text{m}$) and NO₂ are 22 $\mu\text{g}/\text{m}^3$ and 16 $\mu\text{g}/\text{m}^3$, respectively, and have decreased over the past 10 years (Atmo Hauts-de-France 2019). Most of the industrial activities are concentrated in a central industrial zone along the coastline, which is surrounded by a densely populated urban area where socioeconomically disadvantaged populations have lived for generations (Occelli et al. 2016). This industrial area is served by numerous roads and railways used for freight and cargo transportation. A main highway runs through the urban area from east to west, and two others from north to south. There are several agricultural areas along the city border, used to cultivate wheat (~30%), potatoes (~10%), beets (~8%), linen (~6%), and various other crops. Approximately 30% of the agricultural land is used as

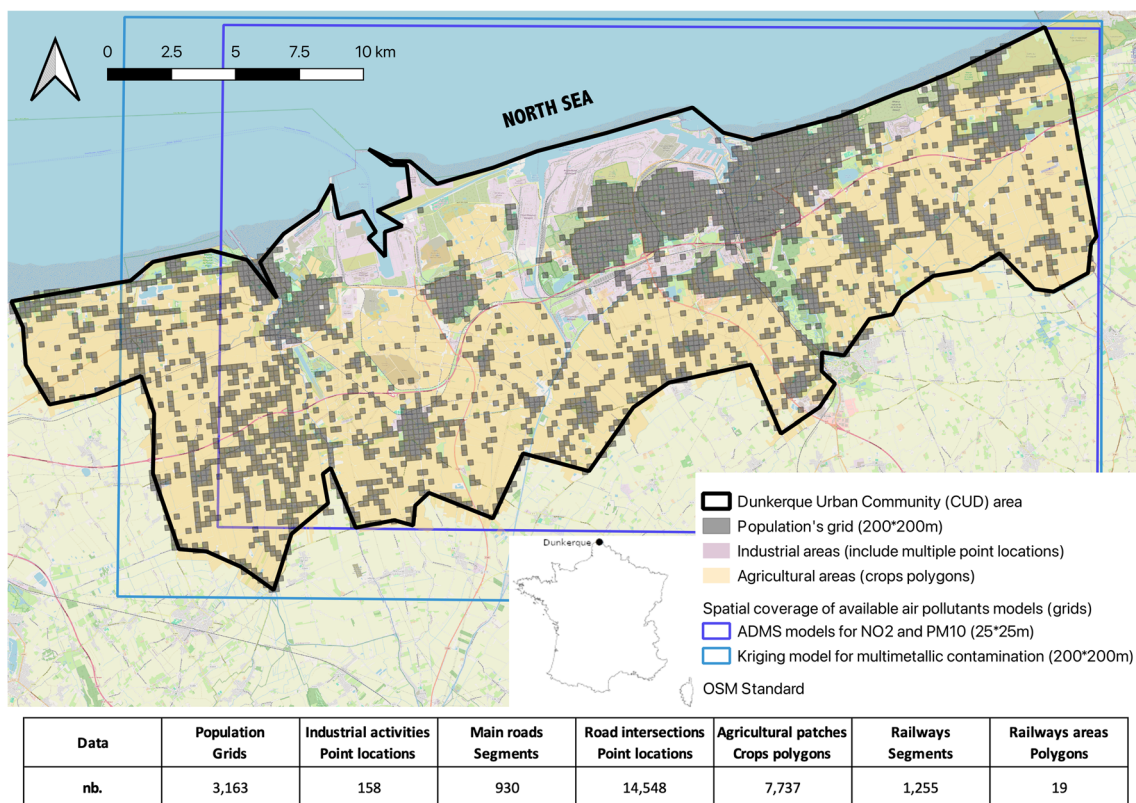


Fig. 1 The Dunkerque Urban Community area: spatial distribution of population grids and pollution sources

fallow and buffer zones between crops and industrial activities. Detailed information was not available on the environmental contamination from crop protection products.

Data origins

Pollution sources data

Industrial sources Spatial distribution of industrial sources in the CUD area was obtained from the Géorisques database (Transition and écologique et de la Cohésion des territoires 2021). This open database registers all French Installations “Classified for the Protection of the Environment” (ICPE), which are major industrial sites considered to pose a potential threat to public health or environmental quality due to the nature of their activities (e.g., petrochemistry, chemistry, metallurgy, or energy production). In the CUD area, there were 158 industrial sites. The point location was used for each activity.

Railway sources Two types of railway sources were considered, both obtained from the French National Institute for Geographic and Forest Information (IGN) BDTOPO@ database (IGN 2021a). The choice was made to separate rail

sources between (i) linear sources, which are mostly heavy-duty railways used by diesel freight trains, and (ii) surface sources which are rail stations and sorting/parking areas where trains are maintained and cargos are loaded. This represents different situations either where trains are moving at medium to high speed with a hot engine or where they are parked and running with cold and low-speed engines, possibly resulting in different emissions situations. The loading of materials, such as ores, is also a source of dust emissions. There were 1255 railway segments and 19 rail areas in the CUD area.

Road sources Road sources are of two types: road crossings and highway segments. Road crossings are considered to be indicators of intra-urban road-traffic emissions, where vehicles, mostly private cars, are traveling at medium or low speeds with frequent stops or slowdowns due to traffic congestion. Highway segments represent high traffic emissions that may be produced by both private cars and heavy vehicles traveling at high speeds in free-flowing traffic. Both sources of data were obtained from the IGN BDTOPO@ database which provides road types and spatial distribution across the country. Highway segments were selected as all roads with an “importance” value < 3. The “Importance” attribute materializes a hierarchy of the road network, based

on the importance of road sections for road traffic. Grades 1–3 correspond to the major routes between big cities. Roadway crossings were generated from the spatial distribution for all road types. The ESRI ArcGIS 10.7 software (ESRI 2021) was used to identify the intersection points where three or more road segments met. Overpasses were deleted thanks to the difference of elevation level. In the CUD area, there were 14,548 intersections and 930 main road segments.

Agricultural sources Agricultural sources were obtained from the IGN Agricultural Patch Registry (IGN 2021b) which geolocates agricultural patches (crops polygons) across the country. All the 7737 patches located inside the CUD were retained.

Population (target) data

The National Institute for Statistical and Economic Studies (INSEE) population grid (INSEE 2021) was used to provide target population data. This grid divides the county into 200×200 m cells and provides socio-occupational information on the population whose housing is located inside said cells. The centroid of each grid whose population was above zero was computed and used as a target for calculating the distance to pollutant emission sources. In the CUD area, there were 3163 target cells (Supplementary Results: SR1).

Environmental contamination data

Air pollution data were obtained from the local official association for air pollution monitoring Atmo Hauts-de-France (Atmo Hauts-de-France 2014). These data were computed by the latter following standardized protocols using official building and industrial emissions, and local traffic counts, as inputs in ADMS-urban software (CERC 2022). Yearly average PM_{10} and NO_2 concentrations were calculated for the year 2014, at 2 m above ground, following a 25×25 m grid resolution. Contamination of target population cells (200×200 m) was evaluated for each pollutant as the average value of all 25×25 m cells below. Due to the spatial coverage of available models, only 2751 population cells received NO_2 and PM_{10} contamination values (Supplementary Results: SR2 & SR3). The uncovered cells correspond to the extreme west of the study area, mostly occupied by a nature reserve. Both pollutants show a similar distribution in the urban area and a high correlation (Spearman's $\rho = 0.88$, $p < 0.01$).

Metal contamination has been evaluated using data obtained from a 60-point lichen-gathering campaign conducted in 2009 (Occelli et al. 2014). Total concentrations of 18 trace elements were evaluated and used to compute a multimetallic contamination index: the mean impregnation ratio (MIR) (Occelli et al. 2016). In this study, the MIR value at the target population cell was evaluated using empirical

Bayesian kriging computed in ESRI ArcGIS 10.7 Geostatistical Analyst tool (Krivoruchko 2012). Due to the spatial distribution of sampled values, only 2917 cells received MIR values (Supplementary Results: SR4). The uncovered cells also correspond to the extreme west of the study area. Metallic and atmospheric contamination in the metropolitan area shows some similarity and a good correlation ($SR = 0.71$ and $p < 0.01$ for both NO_2 and PM_{10}).

Distance indicators

Euclidian distance between each target (population grid centroid) and each source was computed using ESRI ArcGIS 10.7 Spatial Analyst tools. Three distance-based indicators were then computed for each target using R-statistics 3.5.1 (Posit 2022): the distance to the first/closest source (DMin); the average distance to all sources (DAvg); the median distance to all sources (DMed). Distance indicators were computed independently for all six types of sources (industries, railways, rail areas, roadways, road crossings, and agricultural patches). To study the impact of the number of considered sources on indicator quality, DAvG and DMed indicators were computed for the first to the N sources, where N is the maximum number of available sources in the area.

Indicators' evaluation and validation

To assess the relationship between distance indicators and environmental contamination, Spearman's ρ (SR) and R-statistics 3.5.1 were used to compare DMin, DAvG, or DMed with NO_2 , PM_{10} , or multi-metal contamination for each target grid centroid. As contamination levels are generally higher in the proximity of sources and decrease over distance, a negative correlation between pollution and distance to sources was expected.

To examine the correlation dynamics, graphs depicting the SRs for DMin, DAvG, and DMed were generated as a function of the number of sources considered. As DMin represents the distance to the nearest source, its correlation with pollution remained the same regardless of the number of sources included in the calculation. However, the average and median distance indicators, and therefore their correlation with pollutants, are expected to vary based on the number of nearby sources taken into account. *Knee points*, indicating a significant change in the curve dynamics, were identified.

To determine the best distance indicator, the SRs were then compared according to four conditions:

- (i) First source: the distance to the closest source (DMin).
- (ii) All sources: the average (DAvg) or median (Dmed) distance to all sources observed in the territory.

- (iii) *Plateau*: the average (DAvg) or median (Dmed) distance to the n (from 1 to N) nearby sources for which the highest SR value was observed.
- (iv) *Optimal point*: the average (DAvg) or median (Dmed) distance to the n (from 1 to N) nearby sources for which a high SR value was observed for a minimum number of nearby sources. This was graphically determined using the Pythagorean theorem ($a^2 + b^2 = c^2$), as the nearest point from the coordinate $[0; -1]$ observed on the graph. We considered a the number of sources considered on the abscissa axis, b the value corresponding to $(-1) - \text{SR}$ on the ordinate axis, and c the hypotenuse, i.e., the distance between the point $[a;b]$ and the coordinate $[0; -1]$. The *Optimal point* corresponds to the lowest value observed for c .

Results

Dynamic of correlations between distance indicators and environmental contamination

Our first results indicated that all three distance-based indicators were strongly correlated to environmental contamination and that the SR values varied significantly depending on the number of nearby sources considered (Fig. 2 and Table 1), except for the DMin indicator. Since the DMin indicator (orange line) represents the distance to the closest source, the SR did not change regardless of the number of sources considered. The DAvg (green line) and the DMed (blue line) indicators both behaved similarly: the level of correlation increased with the number of sources until it reached a first *knee point* or a succession of *knee points*, then a *plateau* beyond which correlation levels tended to slowly decrease (Fig. 2).

Results also indicated that the observed dynamic was similar for all three concerned contaminations, with similar curves (Fig. 2). SR values ranged from -0.45 to -0.84 for NO_2 , from -0.42 to -0.85 for PM_{10} , and from -0.35 to -0.85 for MIR when considering industrial, road, and rail sources (Table 1). When considering agricultural sources, they ranged from -0.13 to $+0.70$ for NO_2 , -0.05 to $+0.77$ for PM_{10} , and -0.19 to $+0.61$ for MIR.

While a single knee point was observed for distances to industries, intersections, and rail areas (point or surface sources), multiple knee points were seen for distances to main roads and railways (line sources). Here again, both DAvg and DMed behaved similarly. The DMoy and DMed curves for agricultural patches contrasted with the other sources. The positive correlation shifted towards negative as the proportion of sources increased. As the CUD is an industrial area, air emissions due to local agricultural activities seem to be in the minority compared to the others. This

accounts for only 3% of particulate emissions (Zhang et al. 2019). Thus, further results regarding agricultural sources will not be presented.

A slight difference between DAvg and DMed was observed. This varied depending on the type of source considered.

Industrial activities ($n = 158$ plants) A difference of the SR was observed for a small number of sources (less than 15 industrial plants), then a convergence from 20 to 30 nearby sources. This may be explained by two different patterns of the geographical distribution between industries and inhabited areas on the CUD territory (Fig. 1). The first pattern corresponds to population grids located very close to a high density of industrial plants (less than 2 km). Two major industrial areas surround the main urban area (30% of the population grids): on the seafront in the north housing 40% of the plants with steelworks and petrochemicals activities; in the center of the area corresponding to 20% of the plants involving diversified activities such as waste, food processing, construction, and other manufacturing activities. A third comparable area stands out in the northwest and corresponds to 10% of the population grids, and 20% of the industrial plants with nuclear power plants, steelworks, and food processing activities. The second pattern (50% of the grids) corresponds to grids located in towns and villages further away from the major industrial zones. The distance to the nearest plant is less than 3 km but could be much higher (5–10 km) for the other sites. DAvg was more strongly impacted by high distance values than DMed, so it may have seemed more representative of both profile categories.

Railways ($n = 1255$ segments) As railways are strongly linked to the economic activity in this area, their spatial distribution is close to industrial activities, and similar behaviors of SRs were observed.

Rail areas ($n = 19$ areas) SRs were close for DAvg and DMed up to 20% of the sources (three to four areas). Beyond this threshold, slight differences were observed. This could be explained by the small number of such sources and their strong clustering in the major industrial areas on the seafront, the center, and the east. This also explains the small difference with DMin. We also noted that the graph for rail areas and MIR corresponded to the only condition for which DMin performs better than DAvg and DMed.

Intersections ($n = 14,548$ points) and main roads ($n = 930$ segments) DAvg and DMed showed similar behavior up to 50% of nearby intersections, which corresponded to a distance of more than 15 km, and up to 15% of nearby main roads, which corresponded to a distance of about 6 km. This small

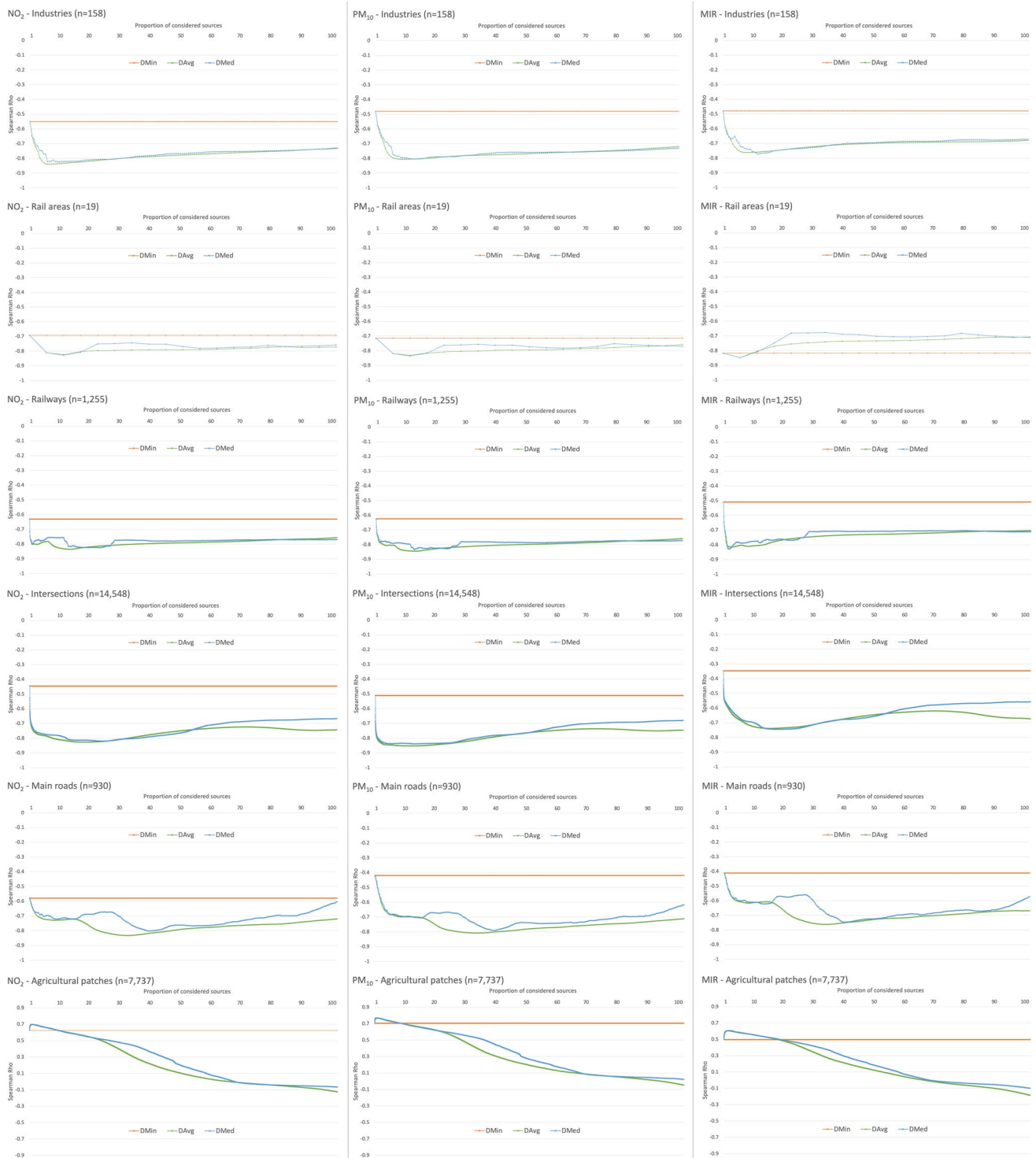


Fig. 2 Graphic matrix (six types of sources × three environmental contamination). Each graph shows Spearman’s rho evolution according to the proportion of considered sources (from the closest to 100%)

difference may be related to the more continuous distribution of roads and intersections over the territory, in contrast to the other types of sources for which there are hotspots. Beyond these thresholds, it is difficult to explain the different

behaviors of the two indicators. The differences observed between DAvG and DMed for 15–35% of main roads, corresponding to a distance of 7–8 km, could reflect the distances between the main cities of the territory.

Table 1 Range (min–max) of statistical correlations (Spearman's rho) between observed environmental contamination and distance-based indicators per type of sources

Distance indicator	Industries	Main roads	Intersections	Railways	Rail areas	Agricultural patches
Modelled NO ₂						
DMin	–0.551 to –0.551	–0.579 to –0.579	–0.446 to –0.446	–0.631 to –0.631	–0.692 to –0.692	0.625 to 0.625
DAvg	–0.840 to –0.551	–0.833 to –0.579	–0.827 to –0.446	–0.836 to –0.631	–0.824 to –0.692	–0.125 to 0.696
DMed	–0.824 to –0.551	–0.803 to –0.579	–0.820 to –0.446	–0.825 to –0.631	–0.828 to –0.692	–0.066 to 0.698
Modelled PM ₁₀						
DMin	–0.480 to –0.480	–0.419 to –0.419	–0.511 to –0.511	–0.624 to –0.624	–0.714 to –0.714	0.705 to 0.705
DAvg	–0.806 to –0.480	–0.808 to –0.419	–0.852 to –0.511	–0.845 to –0.624	–0.836 to –0.714	–0.047 to 0.769
DMed	–0.804 to –0.480	–0.791 to –0.419	–0.839 to –0.511	–0.835 to –0.624	–0.831 to –0.714	0.023 to 0.769
Multimetallic contamination (MIR)						
DMin	–0.479 to –0.479	–0.411 to –0.411	–0.347 to –0.347	–0.509 to –0.509	–0.817 to –0.817	0.495 to 0.495
DAvg	–0.762 to –0.479	–0.762 to –0.411	–0.738 to –0.347	–0.814 to –0.509	–0.848 to –0.707	–0.186 to 0.605
DMed	–0.772 to –0.479	–0.746 to –0.411	–0.745 to –0.347	–0.829 to –0.509	–0.848 to –0.677	–0.099 to 0.608

The optimal number of considered sources

Our findings indicated that the best representation of environmental contamination was neither obtained by considering all sources nor the closest (Table 2). For both D_{Avg} and D_{Med}, the correlation was always higher for the *optimal point* and for the *plateau* than when considering the closest

source or all sources. SR values observed for the *optimal point* and the *plateau* were identical or very close in nearly all situations, with usually a 0.01 difference in favor of the *plateau*. Only main road sources showed slightly higher differences between the *optimal point* and *plateau* (SR were respectively –0.72 and –0.83 for NO₂; –0.77 and –0.81 for PM₁₀; and –0.72 and –0.76 for MIR). As the number of

Table 2 Evolution of the statistical correlations (Spearman's rho) for a different proportion of sources included in the computation: the nearest source, the optimal point (determined as the nearest point

from the coordinate [0; –1]), the plateau, and all the sources. Data presented for D_{Avg} only to optimize table readability

		NO ₂			PM ₁₀			MIR		
		% sources	D _{Avg}	Rho	% sources	D _{avg}	Rho	% sources	D _{avg}	Rho
Industries	1st source	1	1338 m	–0.55	1	1338 m	–0.48	1	1338 m	–0.48
	Optimal point	10 (6%)	4852 m	–0.84	11 (7%)	5037 m	–0.80	11 (7%)	5037 m	–0.76
	Plateau	10 (6%)	4852 m	–0.84	14 (9%)	5485 m	–0.81	13 (9%)	5317 m	–0.76
	All sources	158	29,373 m	–0.73	158	29,373 m	–0.72	158	29,373 m	–0.68
Main roads	1st source	1	1278 m	–0.58	1	1278 m	–0.42	1	1278 m	–0.41
	Optimal point	52 (6%)	3173 m	–0.72	199 (21%)	6690 m	–0.77	220 (24%)	7053 m	–0.72
	Plateau	297 (32%)	8273 m	–0.83	309 (33%)	8589 m	–0.81	310 (33%)	8610 m	–0.76
	All sources	930	30,991 m	–0.72	930	30,991 m	–0.71	930	30,991 m	–0.67
Intersections	1st source	1	209 m	–0.45	1	209 m	–0.51	1	209 m	–0.35
	Optimal point	1247 (9%)	8627 m	–0.81	504 (3%)	5792 m	–0.85	1560 (11%)	9427 m	–0.73
	Plateau	2653 (18%)	12,354 m	–0.83	1947 (13%)	10,430 m	–0.85	2201 (15%)	11,075 m	–0.74
	All sources	14548	62,251 m	–0.74	14548	62,251 m	–0.75	14548	62,251 m	–0.67
Railways	1st source	1	1796 m	–0.63	1	1796 m	–0.62	1	1796 m	–0.51
	Optimal point	13 (1%)	3058 m	–0.80	108 (9%)	5961 m	–0.84	26 (2%)	3562 m	–0.81
	Plateau	242 (19%)	8380 m	–0.82	158 (13%)	6928 m	–0.85	26 (2%)	3562 m	–0.81
	All sources	1255	28,530 m	–0.76	1255	28,530 m	–0.76	1255	28,530 m	–0.70
Rail areas	1st source	1	3625 m	–0.69	1	3625 m	–0.71	1	3625 m	–0.82
	Optimal point	2 (11%)	5469 m	–0.81	2 (11%)	5469 m	–0.82	2 (11%)	5469 m	–0.85
	Plateau	3 (16%)	7458 m	–0.82	3 (16%)	7458 m	–0.84	2 (11%)	5469 m	–0.85
	All sources	19	22,713 m	–0.76	19	22,713 m	–0.76	19	22,713 m	–0.71

sources used to achieve the *optimal point* and the *plateau* may increase from single to double without a significant increase in SR value, we could consider that the *optimal point* was an efficient way to identify the number of closest sources that should be used to provide the best representation of environmental contamination.

Both *optimal point* and *plateau* were obtained for the 10–14 closest industrial activities (6 to 9% of the overall sources), located in a 4.9- to 5.5-km radius. A similar dynamic could be observed for the rail areas, with two to three mostly influential sources (11 to 16%) located on an average 5.4- to 7.4-km radius.

For the main roads, the intersections, and the railways, we could however see that the *optimal points* were obtained with a slightly lower number of sources than for the *plateau* and that this number varies depending on the selected pollutant. For the main roads, the *optimal point* was obtained for 52 (6%, 3.1 km), 199 (21%, 6.7 km), and 220 segments (24%, 7.1 km) for NO₂, PM₁₀, and MIR respectively. For the intersections, it was obtained for 1247 (9%, 8.6 km), 504 (3%, 5.8 km), and 1560 (11%, 9.4 km) elements respectively. For railways, it was obtained for 13 (1%, 3.1 km), 108 (9%, 5.9 km), and 26 elements (2%, 3.5 km).

It thus appeared that, in Dunkerque, the optimal representation of the contamination was obtained when accounting for all sources of pollutants within a 3.0 to 9.4 km radius from a targeted cell. When excluding intersections and rail areas, the radius dropped from 3.0 to 7.1 km. The maps of the DAv_g indicator calculated for the *optimal point* are given in supplementary data (Supplementary Results: SR5).

Finally, the differences in SR values for DAv_g at the *optimal point* were small and ranged from –0.72 (corresponding to the correlation between MIR or NO₂, and intersections) to –0.85 (corresponding to the correlation between PM₁₀ and intersections or between MIR and rail areas). Nevertheless, the correlations appeared to be globally higher for PM₁₀ and lower for MIR. The most influential sources seemed to be industrial activities and rail freight. Despite these observations, we could not identify the contribution of each type of source to the local contamination. The studied area is strongly influenced by industrial activities, which also implies the presence of other linked sources, such as the motorway and railway network.

Discussion

This study examines the relationship between the distance to nearby sources and the modeled levels of environmental contamination, including NO₂, PM₁₀, and multimetal concentrations. Three commonly used distance-based indicators were evaluated to determine how their behavior and correlation were influenced by the type

and number of nearby sources considered. The results indicate that the indicators are effective proxies for all three types of environmental contamination, and that their representativeness is similarly affected by the number of nearby sources. The best representation of environmental contamination is achieved by considering not the closest or all sources, but rather approximately 5–10% of nearby sources, or those within a radius of about 3–6 km.

Detailed analysis of the results

Effect of source typology on pollutants studied

Negative correlations are observed between environmental contamination and distance to industrial activities, main roads, intersections, railways, and rail areas, while SR values are mostly positive for agricultural patches. The strong industrial and urban influence on NO₂, PM₁₀, and trace element emissions in the territory generated a positive correlation between the distance to agricultural patches and the contamination considered in this study. The further away from agricultural areas (and the closer to industrial and urban areas), the higher the contamination.

Similar correlations were found between the environmental indicators and the distance indicators, regardless of the type of source (SR for *optimal points* ranged from –0.72 to –0.85). Environmental trace element concentrations were indeed influenced by industrial activities, as evidenced by major contaminated lands, but were also influenced by road and rail sources. Silva et al. (2021) recently demonstrated that the use of logarithm of distance to highways or primary roads is a good proxy for traffic-related multimetallic air pollution assessment, and Contardo et al. (2020) highlighted the influence of brake abrasion from railways on trace elements concentrations in lichens.

Very good levels of correlation are observed between PM₁₀ concentrations and all sources, clearly showing their multiple origins. This may seem more surprising for NO₂, for which road sources should be predominant. In France, more than 50% of NO₂ emissions are due to transport (CITEPA 2022). However, the CUD is an industrial-port zone, so the majority of sources focus on this zone.

The choice of distance indicator is critical

Our secondary results indicate that the performance of the indicator varies depending on the number of considered sources. On the three selected indicators, we could see that both the DAv_g and the DMed indicators exhibited similar dynamics in the SR values, whereas the DMin indicator remained linear due to its nature. This behavior was expected, and the decision to include DMin as an indicator was based on its common usage in distance-based studies

(Maheswaran and Elliott 2003; Dadvand et al. 2014; De Roos et al. 2014; Buteau et al. 2018; Heydari et al. 2021) and its intuitive appeal as a good proxy for environmental contamination. However, results indicate that DMin is not the best indicator, as it fails to account for many contributive sources.

Regarding the dynamics of correlations, it appears more effective to use DAvG or even DMed, respectively, the average and the median distance between the target and nearby sources as proxies for environmental contamination. Notable differences in SR values were observed between DMin and DAvG at the *optimal point*: approximately 0.3 for industrial sources, 0.35 for main roads (except for NO₂, which showed a difference of 0.14, likely due to significant influence of the closest road compared to other nearby roads), 0.35 for intersections, 0.25 for railways, and 0.1 for rail areas. These differences suggest that while DMin might be useful for a limited number of sources or when sources are spatially clustered (e.g., rail areas), it does not adequately represent the environmental burden when the number, diversity, and spatial distribution of sources are more complex.

The number of closest sources to consider

Moreover, our results indicate that in Dunkerque, the optimal representation of the contamination is not achieved by considering all sources, but by accounting for approximately 5–10% of nearby sources. This corresponds to all sources of emissions within a 3- to 6-km radius around targeted cells. Even though the *optimal point* for main roads is obtained for around 20% of nearby sources, the first *knee point* is also observed at approximately 6%.

The presence of multiple *knee points* when studying the distance to road sources may be linked to the nature of the study area. The significant industrial and harbor activities in Dunkerque (the 3rd most important French port and the 7th biggest port on the North Sea) create a unique local topography. The CUD is characterized by a major industrial core in the north, surrounding the harbor, and two parallel high-traffic motorways crossing the studied area from the southwest to the northeast, following the coastline. Populated areas are mostly located between those two motorways, in Dunkerque city and other smaller municipalities. While port and industrial sources constitute the majority of atmospheric emissions in the CUD (Atmo Nord-Pas de Calais 2015), the motorway's contribution should not be overlooked, especially considering the large number of commercial trucks and the daily mobility of peri-urban worker.

This underscores the need for a thorough understanding of the local topography and dynamics to accurately use distance-based indicators. The indicator used in this study combines the distance to sources in all possible directions

around the target. For specific environmental conditions, it may be appropriate to calculate indicators with greater weighting in certain directions. This would help to account for differences in landscape or prevailing winds, which could be important factors in pollutant dispersion. Complementary studies should be conducted to explore the influence of local structuration on the use of such indicators.

Strengths and limits of the study

Pollution sources data

The data used to characterize sources of pollutant emissions have certain limitations. The primary limitation is treating each source as equivalent to the others. Consequently, the actual emissions and sizes of individual industrial activities could not be considered. For road data, traffic data were not accessible. Additionally, sources located outside the study area were not considered, which may result in a boundary effect. Finally, each source-target relationship was evaluated independently rather than in combination. Such an approach would likely require weighting each type of source according to its contribution to emissions.

ADMS model for NO₂ and PM₁₀

Environmental contamination levels used to evaluate indicators performance were obtained from two different types of models representing two of the most common way to produce fine-scale environmental contamination data. A better evaluation of the indicator's performance would have been achieved by computing correlations with actual field-sampled contamination values on targeted cells instead of model outputs. This approach, however, was not feasible for the same reasons that led us to conduct this study.

Firstly, the absence of high-quality, fine-scaled, and updated database on environmental contamination in France prevented us from easily accessing such information. More recent data were not freely accessible. Secondly, conducting a sufficiently large field sampling campaign was impossible due to monetary and logistical constraints. However, modeled data were already available in the studied area, and the choice was thus made to use them for comparison. This also adds interest to the study, as modeled values are now widely considered a common and efficient basis for environmental studies.

Air pollution (PM₁₀ and NO₂) is indeed framed by law in France and routinely monitored at a large scale by local services. Our data were obtained from the fine-scale models produced annually by these services, which are commonly used for information, urban planning, or environmental health studies (Tenailleau et al. 2015, 2016; Riant et al. 2018; Havet et al. 2020). Those models are

produced following standardized protocols (Tenailleau et al. 2015; European Environment Agency. 2019). Officially validated data including local topography, meteorology, and emissions from roads, buildings, and industries have been introduced into ADMS-urban software (CERC 2022), and obtained models were validated using a field-sampling campaign conducted by the local association for air pollution monitoring. Some of the sources considered in our study (especially roadways) are used to compute pollutant diffusion in ADMS-urban. It is therefore rather reassuring that the distance indicators are strongly correlated with this data, but this probably mechanically increases the intensity of observed correlations.

In this study, the mean value of the 25×25 m models was used to estimate the contamination within the 200×200 m grids. This spatial aggregation likely reduces variability and may have an impact on the observed correlations.

Kriging the multimetallic concentrations in lichens

On the other hand, metallic pollutants are not all framed by law in France, and thus are not subject to fine-scale spatial representation. The MIR model was produced from field-sampled multimetallic contamination in lichens obtained from previous studies on the CUD area (Occelli et al. 2014, 2016; Lanier et al. 2019). The accumulation of persistent pollutants in lichens, such as trace elements, is known to reflect bulk deposition (Loppi and Paoli 2015). These data have been generalized to targeted cells using empirical Bayesian kriging (Krivoruchko 2012), and as such are independent of the source's location. Observed correlations between our indicators and MIR are usually as good as with NO_2 and PM_{10} , which leads us to consider that obtained results for the latter are strong despite the comparison to modeled values instead of sampled values. This also demonstrates the value of distance indicators for pollutants that are not routinely measured or modeled, or possibly for emerging pollutants whose sources are identified. Complementary studies would be needed to identify with more detail the limitation of this proxy, and its usefulness to represent specific pollutants such as secondary pollutants, or long-distance pollution.

The benefits of distance-based indicator for environmental health research

The concept that the proximity of sources has a direct impact on contamination levels is not a novel idea (Chakraborty et al. 2011). It underlies most of the modeling tools currently used in environmental contamination studies, such as spatial interpolation, dispersion modeling, and land use regression (Nieuwenhuijsen 2015). The strong correlation observed

between selected distance-based indicators and evaluated environmental contamination at targeted population cells is not surprising. However, it does confirm the effectiveness of distance-based indicators as a simple and effective proxy for environmental contamination at a metropolitan scale (Jerrett et al. 2005; Zou et al. 2009).

Access to environmental contamination data remains a persistent issue in environmental management and health studies worldwide (Nieuwenhuijsen 2015; aus der Beek et al. 2016; Hoek et al. 2018). Distance-based indicators can serve as a useful proxy for environmental contamination data, especially when acquiring such data is difficult due to infrastructure or budgetary constraints, which makes it impossible to produce spatial models with high levels of accuracy (Gulliver et al. 2011). This is particularly true for specific pollutants, such as pesticides, metals, or organic compounds, where monitoring and standardized modeling techniques are limited. Furthermore, distance-based approaches in environmental epidemiology can help explore the “geophysical plausibility” of a source-pathway hypothesis when dealing with unknown health outcomes (Nuckols et al. 2004). Despite its limitations, Hoek et al. (2018) argued that proximity indicators should be used in future environmental exposure assessments, particularly with continuous distance-based metrics. The use of Euclidean distance, rather than a buffer, has the advantage of not setting an arbitrary distance beyond which it is considered that the population is not exposed. These distance indicators can be generated at different levels of spatial precision, such as home or work address, the centroid of a population grid, or the administrative unit, by calculating an average value from a grid.

The population residing in the CUD has a worse health status than the French national average. In the canton of Dunkerque, the risk of premature mortality is 50% higher than that in metropolitan France. Additionally, the risk of death due to cancer or cardiovascular disease is 25% higher (Région Hauts-de-France 2022). However, we could not identify the most vulnerable areas within this territory as fine-scale health data is not available. Environmental inequalities have already been identified in this region, with the most deprived populations living in areas closest to major emission sources, making them more exposed to air pollution (Occelli et al. 2016; Brousmiche et al. 2023). However, simply considering the geographical distribution of pollution sources is not sufficient for healthy urban planning. To provide decision-makers with relevant information, it is necessary to adopt a more holistic approach that considers other sources of environmental burdens, as well as environmental amenities, territorial services, and the economic and social profiles of the population (Brousmiche et al. 2021).

Conclusion

Assessing environmental exposure to multiple air pollutants in large areas is a major challenge for preventing health risks. Distance-based indicators can be useful in situations where logistical and financial constraints are significant, and continuous distance-based metrics are a better proxy than buffers. Although the distance to the closest source and the average distance to all nearby sources are frequently used, they are not the most effective indicators for measuring environmental exposure. Our study suggests that considering the average distance to a limited number of nearby sources provides a better approximation of environmental contamination. In our case study, we found that the best representation of environmental contamination was obtained by considering around 5–10% of nearby sources, or sources within a 3–6-km radius.

However, it is important to note that distance indicators are good proxies for contamination depending on the geographical context. The presence of sources linked to industrial activities in the CUD territory shows a strong correlation with particulate, nitrogenous, and multimetallic pollution, while they are weakly dependent on agricultural sources. Similar investigations should be conducted in various territories, such as urban or agricultural areas, or more heterogeneous areas with a more random distribution of air emission sources.

We still recommend using distance-based indicators, especially Euclidean distance, in environmental health studies. When combined with other environmental or socioeconomic indicators, they are useful for investigating territorial inequalities and guiding urban planning. Moreover, they are cost-effective for assessing environmental exposure as an exploratory analysis before more sophisticated research.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11356-024-34519-9>.

Acknowledgements Source of data: INSEE, Datagouv, IGN Software: ESRI

Author contribution Quentin Tenaillon: conceptualization, methodology, software, formal analysis, writing—original draft; Caroline Lanier: conceptualization; investigation; writing—review and editing; funding acquisition; Julie Prud'homme: conceptualization, writing—review and editing; Damien Cuny: investigation, writing—review and editing; Annabelle Deram: conceptualization, investigation, writing—review and editing; Florent Occelli: conceptualization; methodology; investigation; writing—original draft; supervision.

Funding This work was supported by the Agence Régionale de Santé Hauts-de-France and Région Hauts-de-France, the Ministère de l'Enseignement Supérieur et de la Recherche (CPER Climibio), and the Fondation I-SITE Université Lille Nord-Europe.

Data availability Data could be made available on request.

Declarations

Ethical approval Not applicable. This did not involve human participants or animal.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

- Atmo Hauts-de-France (2014) Cartothèque d'Atmo Hauts-de-France. In: Atmo Hauts-de-France. <https://cartotheque-atmo-hdf.hub.arcgis.com>. Accessed 21 Nov 2022
- Atmo Nord-Pas de Calais (2015) Bilan territorial 2014 de la Communauté Urbaine de Dunkerque (Bilan territorial). Atmo Nord-Pas de Calais. <https://www.atmo-hdf.fr/joomlatools-files/docman-files/Bilans-territoriaux/BT%20Dunkerque.pdf>
- Atmo Hauts-de-France (2019) Bilan territorial 2018 de la communauté Urbaine de Dunkerque. Atmo Hauts-de-France. https://www.atmo-hdf.fr/joomlatools-files/docman-files/Bilans-territoriaux/2019/BT07-CU_dunkerque-2018.pdf
- aus der Beek T, Weber FA, Bergmann A, Hickmann S, Ebert I, Hein A, Küster A (2016) Pharmaceuticals in the environment—Global occurrences and perspectives. *Environ Toxicol Chem* 35(4):823–35. <https://doi.org/10.1002/etc.3339>
- Brousmiche D, Genin M, Occelli F, Frank L, Deram A, Cuny D, Lanier C (2021) How can we analyze environmental health resilience and vulnerability? A joint analysis with composite indices applied to the north of France. *Sci Total Environ* 763:142983. <https://doi.org/10.1016/j.scitotenv.2020.142983>
- Brousmiche D, Lanier C, Occelli F et al (2023) Spatial scale and spatial effects: shedding light on how a change of scale can affect the characterization of environmental health inequalities. *Environ Sci Policy* 144:20–30. <https://doi.org/10.1016/j.envsci.2023.03.005>
- Buteau S, Doucet M, Tétreault L-F et al (2018) A population-based birth cohort study of the association between childhood-onset asthma and exposure to industrial air pollutant emissions. *Environ Int* 121:23–30. <https://doi.org/10.1016/j.envint.2018.08.040>
- CERC (2022) ADMS-Urban. <http://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html>
- Chakraborty J, Maantay JA, Brender JD (2011) Disproportionate proximity to environmental health hazards: methods, models, and measurement. *Am J Public Health* 101:S27–S36. <https://doi.org/10.2105/AJPH.2010.300109>
- CITEPA (2022) Inventaire des émissions de polluants atmosphériques et de gaz à effet de serre en France - format SECTEN (No. 2071sec / 2022). CITEPA. https://www.citepa.org/wp-content/uploads/Citepa_Rapport-Secten-2022_Rapport-complet_v1.8.pdf
- Contardo T, Vannini A, Sharma K et al (2020) Disentangling sources of trace element air pollution in complex urban areas by lichen biomonitoring A case study in Milan (Italy). *Chemosphere* 256:127155. <https://doi.org/10.1016/j.chemosphere.2020.127155>
- Dadvand P, Ostro B, Figueras F et al (2014) Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees. *Epidemiology* 25:518–525. <https://doi.org/10.1097/EDE.0000000000000107>
- De Roos AJ, Koehoorn M, Tamburic L et al (2014) Proximity to traffic, ambient air pollution, and community noise in relation to incident

- rheumatoid arthritis. *Environ Health Perspect* 122:1075–1080. <https://doi.org/10.1289/ehp.1307413>
- ESRI (2021) ArcGIS v10.7. <https://www.esri.com/en-us/home>
- European Environment Agency (2019) EMEP/EEA air pollutant emission inventory guidebook 2019: technical guidance to prepare national emission inventories. Publications Office, LU
- Gulliver J, de Hoogh K, Fecht D et al (2011) Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. *Atmos Environ* 45:7072–7080. <https://doi.org/10.1016/j.atmosenv.2011.09.042>
- Han B, Hu L-W, Bai Z (2017) Human exposure assessment for air pollution. In: Dong G-H (ed) *Ambient air pollution and health impact in China*. Springer Singapore, Singapore, pp 27–57. https://doi.org/10.1007/978-981-10-5657-4_3
- Havet A, Hulo S, Cuny D et al (2020) Residential exposure to outdoor air pollution and adult lung function, with focus on small airway obstruction. *Environ Res* 183:109161. <https://doi.org/10.1016/j.envres.2020.109161>
- Heydari H, Najafi ML, Akbari A et al (2021) Prenatal exposure to traffic-related air pollution and glucose homeostasis: a cross-sectional study. *Environ Res* 201:111504. <https://doi.org/10.1016/j.envres.2021.111504>
- Hoek G, Ranzi A, Alimehmeti I et al (2018) A review of exposure assessment methods for epidemiological studies of health effects related to industrially contaminated sites. *Epidemiol Prev* 42:21–36
- IGN (2021a) BD TOPO®. In: French national institute for geographic and forest information. <https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo>. Accessed 11 Mar 2020
- IGN (2021b) Registre Parcellaire Graphique (RPG) : contours des parcelles et flots culturaux et leur groupe de cultures majoritaire. In: French national institute for geographic and forest information. <https://www.data.gouv.fr/fr/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-flots-culturaux-et-leur-groupe-de-cultures-majoritaire/>. Accessed 11 Mar 2020
- INSEE (2021) Données carroyées à 200 mètres. In: Institut national de la statistique et des études économiques. <https://www.insee.fr/fr/statistiques/2520034>. Accessed 11 Mar 2020
- Jerrett M, Arain A, Kanaroglou P et al (2005) A review and evaluation of intraurban air pollution exposure models. *J Exposure Sci Environ Epidemiol* 15:185
- Kihal-Talantikite W, Zmirou-Navier D, Padilla C, Deguen S (2017) Systematic literature review of reproductive outcome associated with residential proximity to polluted sites. *Int J Health Geograph* 16:1–39. <https://doi.org/10.1186/s12942-017-0091-y>
- Krivoruchko K (2012) (a) Empirical bayesian kriging and (b) Modeling contamination using empirical bayesian kriging. *ArcUser Fall* 6–10. <https://www.esri.com/news/arcuser/1012/fall2012.html>
- Lanier C, Deram A, Cuny M-A et al (2019) Spatial analysis of environmental inequalities caused by multiple air pollutants: a cumulative impact screening method, applied to the north of France. *Ecol Ind* 99:91–100. <https://doi.org/10.1016/j.ecolind.2018.12.011>
- Loppi S, Paoli L (2015) Comparison of the trace element content in transplants of the lichen *Evernia prunastri* and in bulk atmospheric deposition: a case study from a low polluted environment (C Italy). *Biologia* 70:460–466. <https://doi.org/10.1515/biolog-2015-0053>
- Maheswaran R, Elliott P (2003) Stroke mortality associated with living near main roads in England and Wales: a geographical study. *Stroke* 34:2776–2780. <https://doi.org/10.1161/01.STR.0000101750.77547.11>
- Ministère de la Transition écologique et de la Cohésion des territoires, BRGM (2021) Base des installations classées pour la protection de l’environnement. In: Géorisques. <https://www.georisques.gouv.fr>. Accessed 11 Mar 2020
- Nieuwenhuijsen MJ (ed) (2015) *Exposure assessment in environmental epidemiology*, 2nd edn. Oxford University Press, Oxford; New York
- Nuckols JR, Ward MH, Jarup L (2004) Using Geographic Information Systems for exposure assessment in environmental epidemiology studies. *Environ Health Perspect* 112:1007–1015. <https://doi.org/10.1289/ehp.6738>
- Occelli F, Cuny M-A, Devred I et al (2014) Étude de l’imprégnation de l’environnement de trois bassins de vie de la région Nord-Pas-de-Calais par les éléments traces métalliques Vers une nouvelle utilisation des données de biosurveillance lichénique. *Pollution Atmosphérique* 220:2268–3798
- Occelli F, Bavdek R, Deram A et al (2016) Using lichen biomonitoring to assess environmental justice at a neighbourhood level in an industrial area of Northern France. *Ecol Ind* 60:781–788. <https://doi.org/10.1016/j.ecolind.2015.08.026>
- Pascal M, Pascal L, Bidondo M-L et al (2013) A review of the epidemiological methods used to investigate the health impacts of air pollution around major industrial areas. *J Environ Public Health* 2013:1–17. <https://doi.org/10.1155/2013/737926>
- Posit (2022) R-statistics. <https://posit.co>
- Région Hauts-de-France (2022) Cartorhèque Hauts-de-France. <https://cartes.hautsdefrance.fr>. Accessed 29 Nov 2022
- Riant M, Meirhaeghe A, Giovannelli J et al (2018) Associations between long-term exposure to air pollution, glycosylated hemoglobin, fasting blood glucose and diabetes mellitus in northern France. *Environ Int* 120:121–129. <https://doi.org/10.1016/j.envint.2018.07.034>
- Silva E, Huang S, Lawrence J et al (2021) Trace element concentrations in ambient air as a function of distance from road. *J Air Waste Manag Assoc* 71:129–136. <https://doi.org/10.1080/10962247.2020.1866711>
- Tenaillon QM, Bernard N, Pujol S et al (2016) Do outdoor environmental noise and atmospheric NO₂ levels spatially overlap in urban areas? *Environ Pollut* 214:767–775. <https://doi.org/10.1016/j.envpol.2016.04.082>
- Tenaillon QM, Mauny F, Joly D, et al (2015) Air pollution in moderately polluted urban areas: How does the definition of “neighborhood” impact exposure assessment? *Environmental Pollution* 437–448. <https://doi.org/10.1016/j.envpol.2015.07.021>
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46:234. <https://doi.org/10.2307/143141>
- Zhang S, Bastien L, Rocq B (2019) Caractérisation des particules PM₁₀ dans la région Hauts-de-France (Rapport d’étude No. 01/2018/SZ/V0). *Atmo Hauts-de-France*. https://www.atmo-hdf.fr/sites/hdf/files/content/migrated/Rapport_et_synthese_etudes/Rapport_CARA_2018_VF.pdf
- Zou B, Wilson JG, Zhan FB, Zeng Y (2009) Air pollution exposure assessment methods utilized in epidemiological studies. *J Environ Monit* 11:475. <https://doi.org/10.1039/b813889c>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.