Universitat Oberta de Catalunya

uoc.edu

Estimació de la variació horària de la concentració d'ozó fent ús de dades meteorològiques i models LUR

Guillermo Camps Pons
Màster Universitari en Ciència de Dades
Àrea 5

Nom Consultor: Francisco Ramírez Jávega
Nom PRA: Albert Solé Ribalta

06/2021

FITXA DEL TREBALL FINAL

Títol del treball:	Estimació de la variació horària de la concentració d'ozó fent ús de dades meteorològiques i models LUR
Nom de l'autor:	*Guillermo Camps Pons*
Nom del consultor/a:	*Francisco Ramírez Jávega*
Nom del PRA:	Albert Solé Ribalta
Data de lliurament (mm/aaaa):	*06/2021*
Titulació o programa:	*Màster Universitari en Ciència de Dades*
Àrea del Treball Final:	*Àrea 5*
Idioma del treball:	*Català*
Paraules clau	*Qualitat de l'aire, LUR, ozó*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

L'increment de l'emissió de substàncies contaminants a l'atmosfera en els darrers segles ha convertit la pol·lució en una de les principals amenaces mediambientals. Davants els riscos que la contaminació de l'aire suposa pel planeta i per la salut humana, s'han de prendre mesures per a reduir la concentració de contaminants a l'atmosfera i que aquesta no superi llindars crítics. Una manera de millorar l'eficiència d'aquestes accions és a través de la creació de models de pronòstic de les concentracions de contaminants, que permeten anticipar-se al problema a partir de prediccions estadístiques, numèriques o d'algorismes d'intel·ligència artificial.

En aquest treball, es creen regressions d'ús de sòl (LUR) amb models GBM per a estimar la concentració d'ozó troposfèric (O_3) a partir de dades de concentracions de contaminants, d'estacions meteorològiques i d'ús de sòl. Es duu a procés una selecció de les millors variables i es troba que la radiació UV en superfície mitjana en les últimes 8 hores i la concentració de NO_2 són els factors que més afecten a la concentració d'ozó. Afegint més variables relacionades amb el temps, el vent, la radiació i la densitat de carrils de carreteres en la zona, s'obté un model de 7 variables que obté un R^2 de 0,911 amb un RMSE de 9,884μg/m^3. Afegint una variable relacionada amb temperatura i l'any, el model obté un R^2 de 0,933 amb un RMSE de 8,548μg/m^3. Finalment, es realitza un *downscaling* per a obtenir prediccions a resolució més elevada.

Abstract (in English, 250 words or less):

The rise in air pollutant emissions in the last centuries has turned air pollution into one of the main environmental threats. Ahead of the risks that air pollution means for the planet and for human health, steps must be taken to reduce the concentration of air pollutants in the atmosphere so that they don't exceed critical limits. One way to improve the efficiency of these actions is to create forecast models of air pollutants to be able to anticipate the issue using statistical, numeric or artificial-intelligence-based predictions.

In this work, land use regressions (LUR) are created with GBM models to estimate the concentration of tropospheric ozone (O_3) using data of the concentrations of air pollutants, data from weather stations and land use data. A selection of variables reveals that the 8-hour mean downward UV radiation at the surface and NO_2 concentration are the most important factors to predict the concentration of ozone. Adding more variables based on time, wind, radiation and lane density of nearby highways, a 7-variable model is built with a R^2 value of 0,911 and a RMSE value of 9,884μg/m^3. Adding a temperature-related variable and the year, the model obtains a R^2 value of 0,933 and a RMSE value of 8,548μg/m^3. Finally, high resolution predictions are calculated from a downscaling process.

Índex

Llista de figures

Llista de taules

1. Introducció

1.1 Context i justificació del Treball

La contaminació de l'aire s'ha convertit en una de les principals amenaces mediambientals tant pels éssers humans com pel planeta i la resta d'éssers vius, sent responsable de 4,2 milions de morts anuals [1]. Alguns dels efectes que causa la contaminació en la salut de les persones inclouen increments en la mortalitat, la incidència de càncer, la incidència i els efectes de malalties respiratòries (com l'asma), així com malestars respiratoris o mentals que afecten a les tasques del dia a dia [2]. S'estima que per l'any 2060, les morts prematures anuals degudes a la contaminació de l'aire siguin d'entre 6 i 9 milions de persones i que els efectes d'aquesta contaminació suposin un cost anual del 1% del producte interior brut (PIB) global [3].

Començant amb la revolució industrial, l'activitat humana ha anat incrementant el nivell de certes substàncies a l'atmosfera de manera directa o indirecta que, en altes concentracions, resulten novices per la salut de la gent i pel clima. Si bé en els darrers anys la contaminació de les principals ciutats del primer món (ciutats d'Europa i Nord Amèrica) ja ha començat a reduir emissions de les substàncies més perilloses, la concentració de gasos com el diòxid de sofre (SO_2) o els òxids de nitrogen (NO_x) segueix sent de gran importància en països en desenvolupament (especialment al sud i l'est d'Àsia) [4]. Aquestes substàncies, anomenades contaminants, es poden classificar en dos grups depenent del seu origen [5]. Per una banda, si el contaminant s'ha generat de manera directa per l'activitat humana o per processos naturals, per exemple, a través de processos industrials, erupcions volcàniques o emissions de gasos per part de vehicles, aquest és de tipus primari. Per una altra banda, els contaminants secundaris són aquells que es generen a partir de les reaccions dels contaminants primaris entre ells o amb altres substàncies.

En conseqüència, s'han buscat alternatives més sostenibles com l'aposta per energies renovables en lloc de l'ús de combustibles fòssils [6], el disseny de cotxes amb menys emissions elèctrics o híbrids [7] o dietes veganes per a reduir les emissions generades pel bestiar [8]. No obstant, aquestes solucions no tenen un efecte immediat sino que, en general, requereixen una transformació de la societat a gran escala, marcant objectius en futurs més o menys propers. Consegüentment, la contaminació segueix afectant a la salut de la gent a curt termini. Amb aquesta finalitat, es prenen mesures periòdicament de les concentracions d'aquests contaminants per a determinar el grau d'exposició que tenen els habitants d'una zona. Diferents organitzacions han creat estàndards que defineixen llindars de concentració per cada un dels contaminants segons el risc que suposen per a les persones o l'impacte que tenen pel planeta. A la Taula 1, es poden veure els llindars definits pels estàndards de qualitat de l'aire d'Europa [9]. D'aquesta manera, les ciutats creen normatives i protocols al voltant d'aquests estàndards on es defineixen quines mesures s'ha de prendre quan certs llindars se superen. Aquestes mesures inclouen restriccions a les fonts de contaminants, com restriccions de tràfic, que poden afectar al dia a dia de les persones. Si les mesures s'apliquen de manera reactiva, és a dir, s'activen passats els llindars, la població té molt poc marge per aplicar dites restriccions.

Taula 1. Fonts i llindars crítics de diferents contaminants.

Contaminant	**Tipus**	**Fonts [5] [10] [11]**	**Llindar [9]**	**Període de mitjana**
Matèria particulada 2,5µm ($PM_{2.5}$)	Primari i secundari	Indústria, vehicles, fum tabac	25 µg/m^3	1 any
Matèria particulada 10µm (PM_{10})	Primari i secundari	Indústria, vehicles, fum tabac	50 µg/m^3	1 dia (superar <35 dies per any)
			40 µg/m^3	1 any
Diòxid de sofre (SO_2)	Primari i secundari	Indústria, vehicles	350 µg/m^3	1 hora (superar <24 hores per any)
Diòxid de nitrogen (NO_2)	Primari i secundari	Vehicles, estufes, cuines de gas	200 µg/m^3	1 hora (superar <18 hores per any)
			40 µg/m^3	1 any
Ozó (O_3)	Secundari	Foto-oxidació de NO_X i compostos orgànics volàtils (VOCs)	120 µg/m^3	8 hores (superar <25 dies per any)
Monòxid de carboni (CO)	Primari	Vehicles, combustions en interior, fum tabac	10 mg/m^3	8 hores
Benzè (C_6H_6)	Primari	Indústria, vehicles	5 µg/m^3	1 any
Plom (Pb) (dins PM_{10})	Primari	Indústria, vehicles	0,5 µg/m^3	1 any
Arsènic (As) (dins PM_{10})	Primari	Indústria, volcans	6 ng/m^3	1 any
Cadmi (Cd) (dins PM_{10})	Primari	Indústria, incineració residus, volcans, vegetació	5 ng/m^3	1 any
Níquel (Ni) (dins PM_{10})	Primari	Indústria, vent amb pols, volcans	20 ng/m^3	1 any
Benzo(a)pirè (BaP) (dins PM_{10})	Primari	Indústria, estufes de llenya	1 ng/m^3	1 any

Consegüentment, resulta raonable intentar anticipar-se als nivells alts de contaminació fent pronòstics sobre els nivells de contaminació per tal de aplicar les mesures amb més temps i d'una manera més eficient. Les dades utilitzades per a fer tals prediccions inclouen principalment dades espaciotemporals de diferents contaminants i dades meteorològiques com el vent, que transporta les masses d'aire per l'atmosfera, i la incidència solar, que és capaç d'activar certes reaccions. Després, aquestes s'utilitzen en models de pronòstic estadístic, com regressions o models autoregressius integrats de mitjana mòbil (ARIMA), models d'intel·ligència artificial, com xarxes neuronals *long short-term memory* (LSTM), i models de pronòstic numèric, com models de transport o models comunitaris de qualitat d'aire a multiescala (CMAQ) [12]. A aquests models, s'hi afegeixen altres tècniques de mineria de dades, com anàlisis de les components principals (PCA) o màquines de vectors de suport (SVM), i la construcció de models híbrids, que poden incrementar la precisió de les prediccions [12]. Addicionalment, aquestes dades es poden enriquir amb altres dades com dades d'ús de sòl o, en anglès, *land use* (com el tipus de terreny, l'alçada d'edificis, el tràfic, etc.) que poden contenir informació dels factors que poden influir a una generació més elevada de contaminació. Per exemple, en un lloc on hi ha moltes carreteres és possible que hi hagi un volum més elevat de tràfic que generi més substàncies com el NO_2. De la mateixa manera, una zona industrial no es comporta igual que una zona residencial o una zona d'agricultura.

Aquest treball s'enfoca en la creació de models fent ús de les dades de contaminació de les diferents estacions de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica de Catalunya amb l'objectiu de realitzar estimacions de la concentració d'ozó troposfèric (O_3). Donat que l'ozó és un contaminant secundari, s'espera que tingui dependència amb els components de la reacció que el creen. Més concretament, un dels orígens de l'ozó és la foto-oxidació d'òxids de nitrogen i, per tant, s'utilitzen dades de les concentracions de diòxid de nitrogen i de radiació (fotons). Les dades de contaminants (ozó i diòxid de nitrogen) i radiació s'integren amb altres dades meteorològiques (com el vent o la temperatura) i es complementen amb dades de *land use*, dades de carreteres i dades demogràfiques per a incrementar la capacitat predictiva dels models. Més concretament, s'utilitzen models de *gradient boosting machines* (GBM), que a la literatura demostren un rendiment eficient [13]. Amb un model força bo, es pot extrapolar el comportament de l'ozó a altres punts diferents de les estacions. Per aquesta raó, el model es prova en un entorn d'alta resolució en els carrers d'una zona de Barcelona i fora de l'entorn de les estacions i podria ser de gran utilitat per a adaptar el dia a dia a les concentracions d'ozó de l'entorn de cada persona.

1.2 Objectius del Treball

Principals

- Obtenir dades de les concentracions contaminants
- Obtenir dades de caràcter meteorològic (vent, radiació en superfície, núvols, etc.) de satèl·lit o de diverses estacions meteorològiques.

- Crear un *dataset* on s'integren totes les dades anteriors amb dades de *land use*, dades de carreteres, dades poblacionals que descriuen l'entorn de les estacions de mesura de contaminants.
- Crear models per a estimar la concentració d'ozó troposfèric (O_3) i poder extrapolar el comportament a altres zones.
- Comparar els resultats dels models entre ells i amb l'estat de l'art.
- Mesurar l'eficàcia d'usar models que utilitzin dades *land use*.

Secundaris

- Aprendre a utilitzar formats de bases de dades GIS.
- Obtenir dades a través de APIs o dades restringides sota sol·licitud.
- Explorar catàlegs de dades de *land use.*
- Aprendre el funcionament de la llibreria LightGBM per a la implementació de models de *gradient boosting machines* (GBM).
- Comparar els resultats amb algun altre mètode com kriging o long short-term memory (LSTM).

1.3 Enfocament i mètode seguit

Per a poder entendre el context del treball, és necessari començar amb una recerca bibliogràfica. En primer lloc, cal recollir informació sobre la contaminació de l'aire. Això inclou entendre quins processos físics i químics generen contaminants, quins efectes generen aquests contaminants a la societat i al planeta i què en diu la legislació de la regularització dels contaminants. Per a fer aquesta recerca, es busquen articles actuals i es navega per les referències fins a trobar les fonts de la informació més importants.

En segon lloc, cal recollir informació sobre l'estat de l'art referent als models utilitzats per a estimar la concentració de contaminants a l'aire. Es tracta de saber quins models (estadístics, numèrics, d'intel·ligència artificial, etc.) funcionen millor per a tenir un punt de referència de cara a la creació de models i quina és la tendència al llarg dels anys. Aquesta recerca ens permet avaluar els resultats del treball i l'ús de certes tecnologies (com *land use*). Llavors, és important que la bibliografia consultada sigui de l'àmbit d'investigació més actual possible, alhora que es tracta de donar un sentit d'evolució al llarg dels anys.

El següent punt de recerca són les fonts de dades. L'objectiu principal és trobar tres tipus de fonts de dades: dades de concentració de contaminants, dades meteorològiques i dades d'ús de sòl. Addicionalment, es cerquen altres dades complementàries que es considerin d'importància de cara a la predicció d'ozó, com podrien ser dades de caràcter demogràfic. Prioritàriament, es busquen dades obertes però també s'utilitzen dades d'accés restringit si és necessari.

Les dades poden venir en diferents formats i cal entendre'ls per a poder manejar-les. Alguns dels formats principals a aprendre per a tractar dades amb informació geogràfica (dades GIS) que es necessiten per aquest treball poden ser imatges

GeoTIFF, arxius netCDF o arxius GRIB. Python disposa de diverses llibreries com rasterio, netCDF4 o PyGrib per a obrir i tractar, respectivament, aquests tipus de arxius. Un cop familiaritzats amb les dades, s'integren totes les dades en un mateix conjunt de dades o *dataset*, tot i descartant aquelles variables que no ens són d'interès.

Amb un únic *dataset* que engloba totes les dades, es creen models de regressió d'ús de sòl (LUR) fent servir *gradient boosting machines* (GBM) de la llibreria LightGBM de Python. S'utilitzen altres paquets de Python com pandas o scikit-learn per a complementar el tractament de les dades i l'avaluació del model. De manera iterativa, s'avaluen els resultats i s'ajusten les variables a utilitzar i els paràmetres que afecten al comportament del model, com per exemple, quin nombre de variables s'utilitza.

1.4 Planificació del Treball

Taula 2. Planificació del treball.

Número	Tasca	Inici	Final	Dies
1	Definició i planificació TFM	17/02/21	28/02/21	12
1.1	Definició del treball	17/02/21	19/02/21	3
1.2	Recerca sobre el context del treball	17/02/21	24/02/21	8
1.3	Planificació del treball	20/02/21	28/02/21	9
2	Revisió de l'estat d'art	01/03/21	21/03/21	21
2.1	Recopilació bibliogràfica	01/03/21	07/03/21	7
2.2	Lectura i síntesi dels resultats	08/03/21	21/03/21	14
2.3	Revisió dels *datasets*	01/03/21	21/03/21	21
3	Implementació	22/03/21	23/05/21	63
3.1	Creació del *dataset*	22/03/21	05/04/21	14
3.2	Creació de models	05/04/21	25/04/21	21
3.3	Ajust d'hiperparàmetres i avaluació	26/04/21	23/05/21	28
4	Redacció de la memòria	24/05/21	06/06/21	14

1.5 Breu sumari de productes obtinguts

En primer lloc, en aquest treball s'obté un conjunt de dades on es troben les dades integrades de diferents fonts, que inclouen dades d'estacions meteorològiques, dades d'ús de sòl, etc. i que conté 664878 files i 150 columnes i no conté cap valor nul. El segon producte principal que s'obté d'aquest treball és el conjunt de models que estimen la concentració d'ozó. Més concretament, s'obtenen dos models que fan d'ús de 7 i 9 variables. En tercer lloc, a partir de mapes de concentració de NO_2 i vent obtinguts a partir de models numèrics a alta resolució, s'elabora un procés de *downscaling* del qual s'obtenen animacions de mapes de la concentració d'ozó d'una zona de Barcelona per cada model. Finalment, aquesta memòria recull tot el procés per a l'obtenció dels dos productes anteriors, recerca bibliogràfica de la contaminació de l'aire i l'estat de l'art i les conclusions que es poden extreure dels resultats obtinguts. Com a productes derivats d'aquests primers, s'obtenen *notebooks* de Python (.ipynb) destinats a la elaboració dels productes anteriors.

1.6 Breu descripció dels altres capítols de la memòria

En el següent apartat, es duu a terme una recerca bibliogràfica en busca de tenir una idea general de l'estat de l'art en relació a la predicció de contaminants atmosfèrics i fent èmfasi en les prediccions de la concentració d'ozó. Conèixer l'estat de l'art i els models que utilitzen els diferents autors és important per saber quin és el punt de partida, què funciona millor i quines fites es poden assolir.

L'apartat 3 recorre tota l'elaboració dels productes descrits abans. El primer subapartat explora les diferents fonts de dades i descriu els formats i el procés de la captura d'aquestes. D'aquesta manera, es prepara pel segon subapartat, on s'integren les dades de les diferents fonts. En aquest subapartat, es creen variables i s'elabora el conjunt de dades que s'utilitza per construir els models. En el subapartat 3, es creen els models GBM. Els paràmetres dels models s'ajusten per a aconseguir un resultat òptim i es realitza una avaluació dels millors models GBM obtinguts. Per acabar, en el darrer subapartat es realitza el procés de *downscaling* per a obtenir les animacions de mapes de concentració d'ozó descrites abans.

A l'apartat 4, es descriuen les conclusions extretes d'aquest treball. Es comparen els resultats amb l'estat de l'art i s'avalua el compliment dels objectius marcats al principi del treball. Finalment, es comenten les possibles línies futures que han quedat fora de l'abast d'aquest treball. Els darrers apartats contenen el glossari dels principals termes del treball, la bibliografia consultada i un annex.

2. Estat de l'art

Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses (Ross et al., 2005) [14]. És un dels primers articles en aplicar un model de regressió d'ús de sòl (LUR) en una ciutat gran dels Estats Units. El model tracta de predir amb una regressió lineal múltiple (MLR) les concentracions de NO_2 pel comtat de San Diego, California. Es calculen *buffers* circulars de diferents radis (40m, 300m, 500m i 1000m) per a mesurar els diferents usos de sòl en l'àrea al voltant de cada punt.

S'utilitzen dades de tràfic i carreteres, zones industrials, poblacionals i distàncies a carreteres i a la costa. El predictor més rellevant resulta ser el tràfic en un *buffer* de 300m, que explica el 54% de la variació. Conjuntament amb 3 altres variables, s'aconsegueix explicar el 79% de la variació del NO_2, demostrant que LUR és un model prometedor per a la predicció de contaminants.

Mapping of background air pollution at a fine spatial scale across the European Union (Beelen et al., 2009) [15]. És un article científic on es fan prediccions per NO_2, PM_{10}, O_3, SO_2 i CO a escala europea a una resolució de 1km x 1km i per àmbits urbans, rurals i globals. Es comparen models de regressió, models de kriging ordinari i models de kriging universal.

S'utilitzen dades d'ús de sòl, de tràfic, de densitat població, de meteorologia, d'altitud, de topografia i de distància al mar. En general, els models de kriging universal obtenen els millors resultats amb R^2 de 0,61 per NO_2, 0,45 per PM_{10} i 0,70 per O_3 per totes les estacions fent ús de 3 variables. Els altres dos contaminants no tenen cap model amb resultats satisfactoris.

Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging, Land-Use Regression (LUR), and Combined Bayesian Maximum Entropy–LUR Approaches (Adam-Poupart et al., 2014) [16]. És un article científic on es fan prediccions espaciotemporals de la concentració d'ozó per a l'estiu a Quebec, Canada. Es comparen els resultats de 3 models: un model LUR, un model LUR amb *Bayesian maximum entropy* (BME) i un model kriging amb BME.

S'utilitzen dades de fins a 51 estacions de monitorització de O_3, dades de densitat de carreteres i dades meteorològiques per acabar utilitzant un total de 6 variables. El model LUR/BME obté els millors resultats amb un R^2 de 0,653. LUR obté un valor de 0,466 mentre que kriging/BME obté 0,414. Es mostra que els errors d'estimació en la interpolació de les concentracions d'ozó es poden reduir significativament amb l'ús de LUR.

Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment (Zhan et al., 2018) [17]. És un article científic on es crea un model de *random forest* (RF) per a la predicció dels màxims diaris de la mitjana de 8 hores de la concentració d'ozó troposfèric a tota China en una quadrícula de 0,1º x 0,1º.

S'utilitzen dades horàries de 1608 estacions de monitorització d'ozó del 2015, dades meteorològiques, dades d'ús de sòl (índex normalitzat de diferència de vegetació (NDVI), densitat de carreteres, etc.), densitat de població, etc. Entre les variables escollides per al model final es troben les emissions de CO, NMVOC i NH_3, l'evaporació, la duració del sol, la velocitat del vent, etc. fins a un total de 13 variables. El model RF demostra, amb un R^2 de 0,69, que pot obtenir resultats comparables o millors a altres models més complexos, com models de transport químic (CTM), per un cost computacional molt més baix.

A hybrid approach to estimating long-term and short-term exposure levels of ozone at the national scale in China using land use regression and Bayesian maximum entropy (Chen et al., 2021) [18]. És un article científic on es construeixen un model LUR per a predir els màxims anuals de la mitjana de 8 hores de la concentració d'ozó troposfèric a tota China i un model que combina LUR amb un estimador del tipus *Bayesian maximum entropy* (BME) per a predir els màxims diaris. El model LUR es dedica a explicar la variabilitat espacial de les concentracions d'ozó, construint una visió a llarg termini, mentre que el model híbrid mesura les variacions espaciotemporals, obtenint una visió a curt termini. L'objectiu és generar millors prediccions de l'exposició per a estudis epidemiològics, que es beneficien tant de dades a petita escala com a gran escala.

S'utilitzen les 5 variables predictores més significatives per a fer la regressió. Es troba que la temperatura, la llargària de les carreteres en un *buffer* de 1000m i les àrees industrials en un buffer de 3000m tenen una correlació positiva amb les concentracions d'ozó a gran escala, mentre que la velocitat del vent i l'altitud mostren una correlació negativa. En quant a resultats, el model LUR pels pics anuals obté resultats moderats, amb valors de R^2 de 0,53, 0,57 i 0,59 pels anys 2015, 2016 i 2017. El model LUR/BME obté un valor superior de R^2 de 0,80. Aquest últim es compara amb models espaciotemporals kriging, que no superen el 0,60.

A novel hybrid spatiotemporal land use regression model system at the megacity scale (Wang, J. et al., 2021) [19]. És un article científic on es construeix un model per a predir les concentracions de CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ i SO_2 a la ciutat de Tianjin a China. Segons els autors, s'ha vist a la bibliografia que els models LUR espaciotemporals poden obtenir resultats espacials pobres en alguns panels temporals. Per això, aquest article proposa un model híbrid compost per LUR i per regressions de suport vectorial (SVR). Addicionalment, regressions lineals múltiples (MLR) i un model especial espaciotemporal (ST) s'utilitzen de manera suplementària en els pitjors casos.

El model utilitza dades de tràfic, d'emissions, econòmiques, d'ús de sòl i meteorològiques i s'avalua amb *cross-validation* (cv), obtenint resultats de R^2_{cv} superiors a 0,60 (limit acceptable establert pels autors) en més del 97% dels dies de l'any. Els valors mitjans de R^2_{cv} per aquests dies es troba al voltant del 0,90 pels diferents contaminants, un rendiment elevat per a ser utilitzat pels estudis d'exposició. Addicionalment, es comenta l'importància de generar models espaciotemporals per

contaminants com l'ozó (en lloc de models espacials), ja que són contaminants secundaris i, per tant, són més sensibles a l'escala temporal i a les condicions meteorològiques.

Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States (Ren et al., 2020) [13]. És un article científic on es comparen 13 diferents models d'aprenentatge automàtic (ML) i de LUR per a estimar la mitjana de 8 hores de la concentració espacial o espaciotemporal diària d'ozó als Estats Units donat a la falta d'aquest tipus de comparacions en la literatura.

S'utilitzen predictors meteorològics, d'ús de sòl, d'emissions i de transport químic i s'ajusten els hipermaràmetres amb repetits *grid search* per a garantir extreure el màxim potencial de cada mètode. Es troba que els models no lineals d'aprenentatge automàtic obtenen millors resultats, especialment en models espaciotemporals. Dos algorismes a destacar són el *random forest* (RF) i *extreme gradient boosting* (XGBoost), els dos funcionant amb la combinació independent o seqüencial, respectivament, d'arbres de regressió. També es comenta sobre l'eficiència dels models sent els espacials molt més ràpids en general. Comparant els millors algorismes, XGBoost tarda unes 7 vegades menys que RF.

Using a land use regression model with machine learning to estimate ground level PM2.5 (Wong et al., 2021) [20]. És un article científic on es construeixen i es comparen diferents models LUR per a la predicció espaciotemporal de la concentració de $PM_{2.5}$ a Taiwan. En primer lloc, s'utilitzen per separat un model LUR i un model híbrid kriging-LUR per a determinar les variables predictores més importants. Seguidament, s'utilitzen tres algorismes diferents per a la predicció de $PM_{2.5}$ amb les variables escollides per cada un dels models LUR, resultant en 8 possibles combinacions (incloent els dos models LUR sols). Aquests algorismes són una xarxa neuronal profunda (DNN), un *random forest* (RF) i un *extreme gradient boosting* (XGBoost).

Els models utilitzen dades diàries de $PM_{2.5}$ de 73 estacions combinades amb dades meteorològiques i d'ús de sòl (índex normalitzat de diferència de vegetació (NDVI), xarxa de carreteres, terreny, etc.). Les variables 6 escollides pels models LUR són els contaminants SO_2, NO_2 i O_3, la distància a l'aeroport més proper, els boscos en un *buffer* de 5000m i els cultius en un *buffer* de 4000m. El model LUR tot sol obté un R^2 de 0,58 mentre que el model híbrid kriging-LUR obté 0,89. XGBoost i DNN, amb valors molt semblants, són els algorismes que millors resultats donen incrementant els valors anteriors a 0,73 i 0,94, respectivament, mentre que RF obté uns resultats lleugerament inferiors.

High-resolution prediction of the spatial distribution of PM2.5 concentrations in China using a long short-term memory model (Wang, Z. et al., 2021) [21]. És un article científic on es desenvolupa una xarxa neuronal recurrent del tipus *long*

short-term model (LSTM) per a la predicció d'alta resolució (1 km) de la distribució espacial de $PM_{2.5}$ a tota China.

S'utilitzen les dades dels anys 2014-2018 de 1467 estacions de monitorització, dades meteorològiques (precipitació, velocitat del vent, etc.), dades de satèl·lit d'aerosols (AOD) i dades d'ús de sòl (índex normalitzat de diferència de vegetació (NDVI) i elevació). El model LSTM, amb un R^2 de 0,83 i fent ús de 8 variables, obté millors resultats que models de *random forest* (RF) i models cubistes donat a la seva xarxa recurrent que permet capturar millor la dependència en el temps i dependències no lineals amb altres variables.

3. Disseny i implementació del treball

3.1 Descripció de les fonts de dades

En aquest apartat, es descriuen les fonts de dades utilitzades i quines dades s'obtenen. Les dades que es pretenen obtenir són molt diverses i, per tant, venen de fonts diverses i presenten formats diferents que s'han de d'obrir i tractar amb diferents llibreries dins el procés d'integració de dades. Llavors, es cerquen diferents fonts que puguin proporcionar dades obertes sobre la concentració de contaminants, mesures de variables meteorològiques, dades que descriuen l'ús del sòl i dades demogràfiques.

Per a fer unes prediccions que reflecteixin la situació d'avui en dia, es tracten dades el més actuals possibles. Per això, es busquen dades dels anys 2018 i 2019 (donat que 2020 és un any excepcional per la pandèmia del COVID-19). Amb l'objectiu de fer prediccions amb una resolució temporal elevada, es busquen dades horàries de les concentracions de contaminants i de les variables meteorològiques. En quant a la zona espacial del treball, s'utilitzen dades de les estacions de mesura de contaminants de Catalunya. Per tant, les fonts de dades a utilitzar han d'incloure la zona que ocupa Catalunya. Un exemple és la regió que va de 40,0º a 43,0º de latitud i de -0,5º a 4,5º de longitud.

3.1.1 Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica

La Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica és un sistema de detecció dels nivells d'immissió dels principals contaminants (com NO_2, O_3, SO_2, NO, CO, etc.) format per diferents estacions de mesurament distribuïdes pel territori de Catalunya. Va ser creada per la Llei 22/1983, de 21 de novembre, i actualment està adscrita administrativament al Departament de Territori i Sostenibilitat de Catalunya.

Les dades horàries de les mesures de les diferents estacions des de l'any 1991 fins al dia anterior a l'actual estan disponibles al portal de dades obertes de la Generalitat i s'actualitzen amb una freqüència diària [22]. El portal de dades permet aplicar filtres sobre les seves columnes i exportar les dades a diferents formats a través de la pàgina web o a través de l'API de Dades Obertes de Socrata (SODA).

A la Taula 3 es descriuen els diferents camps que ofereix aquesta base de dades. Cada fila correspon a una estació, un contaminant i una data i conté 24 mesures repartides en les diferents hores del dia (columnes 01h, 02h, 03h, etc.). Donat que les estacions es troben georeferenciades per la longitud, la latitud i l'altitud, es pot obtenir la distribució espaciotemporal dels contaminants d'interès.

D'aquesta manera, es creen dos conjunts de dades fent ús de la utilitat de filtratge que proporciona el portal de dades. En primer lloc, com que l'objectiu d'aquest treball és tractar amb dades de 2018 i 2019, s'aplica un filtre on el camp que descriu la data de la mesura ha de trobar-se després de 31 de desembre de 2017 i un altre filtre per a que es trobi abans de 01 de gener de 2020. Seguidament, per separat, s'aplica un tercer filtre on el contaminant ha de ser 'O3' o 'NO2'. Finalment, s'exporten les dades

en format de valors separats per comes (.csv) de manera que s'obtenen dos arxius que són anomenats O3_Catalunya_2018_2019.csv i NO2_Catalunya_2018_2019.csv, respectivament.

Taula 3. Descripció dels camps de la base de dades dels punts de mesurament de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica.

Nom del camp	Descripció	Tipus
CODI EOI	Codi identificatiu de l'estació	Text
NOM ESTACIO	Nom de l'estació	Text
DATA	Data de la mesura en format DD/MM/YYYY	Text
MAGNITUD	Codi identificatiu del contaminant	Text
CONTAMINANT	Nom del contaminant	Text
UNITATS	Unitats de la mesura del contaminant	Text
TIPUS ESTACIO	Tipus d'estació segons la localització de l'equip (Traffic, Background o Industrial)	Text
AREA URBANA	Tipus d'àrea (Urban, Peri-urban o rural)	Text
CODI INE	Codi identificatiu del municipi	Text
MUNICIPI	Nom del municipi	Text
CODI COMARCA	Codi identificatiu de la comarca	Text
NOM COMARCA	Nom de la comarca	Text
01h - 24h	24 camps corresponents a les mesures del contaminant per a cada hora. La unitat de la mesura es troba al camp UNITATS.	Numèric
ALTITUD	Altitud de l'estació en metres (m)	Numèric
LATITUD	Latitud de l'estació en graus decimals (º) cap al nord (sistema de referència WGS84 o EPSG:4326)	Numèric
LONGITUD	Longitud de l'estació en graus decimals (º) ca a l'est (sistema de referència WGS84 o EPSG:4326)	Numèric
GEOREFERENCIA	Columna de georeferència	Punt

3.1.2 ERA5

ERA5 és la quinta generació de reanàlisis de variables climàtiques atmosfèriques, terrestres i oceàniques del Centre Europeu de Prediccions Meteorològiques a Mitjà Termini (ECMWF). Els projectes de reanàlisis d'ECMWF combinen les observacions passades (que són menys completes) amb models actuals amb l'objectiu de descriure el clima global més detalladament i de manera consistent amb les lleis de la física, tot i descrivint l'evolució d'aquest al llarg dels anys.

Aquest reanàlisi proporciona diferents conjunts dades meteorològiques en freqüència horària o en mitjanes mensuals. L'objectiu del projecte és realitzar el reanàlisi cap enrere fins 1950 però les dades contingudes en els anys compresos entre 1950 i 1979 es troben, de moment, en un estat preliminar, mentre que des de 1979 fins al present es troben completes. A més a més, les dades de ERA5 es poden trobar a nivell de terra [23] o a diferents nivells de pressió. Les dades d'aquests conjunts divideixen el món en una quadrícula de 0,25º de longitud i latitud. Addicionalment, un conjunt de dades anomenat ERA5-Land proporciona dades partint de 1981 en una resolució més alta de 0,1º de longitud i latitud (uns 9km) [24].

Tots els conjunts de dades d'ERA5 estan disponibles de manera gratuïta al Climate Data Store (CDS) de Copernicus Climate Change Service (C3S) i es poden filtrar i descarregar a través de la interfície web o a través de l'API CDS en format GRIB o netCDF.

En el cas d'aquest treball, es creen dos conjunts de dades: un de ERA5-Land amb major resolució i un de ERA5 a nivell de terra (single levels) per a omplir dades buides i complementar al primer. En ambdós conjunts es seleccionen els anys 2018 i 2019 amb tots els mesos (de 1 a 12), tots els dies (de 1 a 31) i totes les hores (de 00:00 a 23:00) en una regió que va de 40,0º a 43,0º de latitud i de -0,5º a 4,5º. Per al conjunt de ERA5-Land, se seleccionen les variables de la component zonal (u) del vent a 10 metres, la component meridional (v) del vent a 10 metres, la radiació solar neta en superfície i la temperatura a 2 metres. Per al conjunt de ERA5 a nivell de terra, s'hi afegeixen la radiació ultraviolada descendent en superfície i la nuvolositat, que no estan disponibles a ERA5-Land. Per acabar, s'exporten les dades en format de netCDF (.nc) de manera que s'obtenen dos arxius que són anomenats ERA5-Land.nc i ERA5_single_level.nc, respectivament. A la Taula 4, es descriuen tant les dimensions (latitud, longitud i temps) com la resta de variables.

3.1.3 WorldPop

A partir d'investigacions revisades per experts, WorldPop desenvolupa conjunts de dades obertes geoespacials i d'alta resolució de variables demogràfiques per donar suport, principalment, a països amb ingressos baixos i mitjans, com poden ser països d'Àfrica, Àsia o d'Amèrica del Sud [25]. El Repositori Obert de Dades Poblacionals de WorldPop (WOPR) conté més de 44 mil conjunts de dades obertes que descriuen la població en graelles de diferents països segons el nombre d'habitants, l'edat i el sexe, els naixements, el flux de migració, etc.

Taula 4. Descripció dels camps dels conjunts de dades meteorològiques de ERA5.

Nom	Descripció	Dimensió	*Dataset*	Tipus
longitude	Longitud en graus decimals (º) cap a l'est (sistema de referència WGS84 o EPSG:4326)	longitude	ERA5, ERA5-Land	Numèric
latitude	Latitud en graus decimals (º) cap al nord (sistema de referència WGS84 o EPSG:4326)	latitude	ERA5, ERA5-Land	Numèric
time	Hores (h) des de 01/01/1900 00:00	time	ERA5, ERA5-Land	Numèric
u10	Component zonal (u) del vent a 10 metres en metres per segon (m/s)	time, latitude, longitude	ERA5, ERA5-Land	Numèric
v10	Component meridional (v) del vent a 10 metres en metres per segon (m/s)	time, latitude, longitude	ERA5, ERA5-Land	Numèric
t2m	Temperatura a 2 metres en graus Kelvin (K)	time, latitude, longitude	ERA5, ERA5-Land	Numèric
ssr	Radiació solar neta en superfície en Joules per metre quadrat (J/m^2)	time, latitude, longitude	ERA5, ERA5-Land	Numèric
uvb	Radiació ultraviolada descendent en superfície en Joules per metre quadrat (J/m^2)	time, latitude, longitude	ERA5	Numèric
tcc	Nuvolositat en un interval [0,1]	time, latitude, longitude	ERA5	Numèric

En el cas d'aquest treball, s'utilitzen dades sobre el nombre d'habitants d'Espanya al 2018 i al 2019 en una graella amb una resolució de 100 metres. Dins les diferents possibilitats, s'escull el conjunt de dades que s'ha creat amb un enfocament *top-down* donat que aquest és consistent amb les dades oficials espanyoles de demografia que, sent un país del primer món, es publiquen amb una freqüència elevada. Addicionalment, s'escull el *dataset* amb un model *unconstrained*, ja que el *constrained*, que ajusta la població als edificis, només està disponible per 2020.

Consegüentment, s'obtenen dos arxius GeoTIFF esp_ppp_2018.tif i esp_ppp_2019.tif corresponents als anys 2018 i 2019, respectivament. Estan formats per una imatge 2D on cada píxel descriu la població que es troba en una cel·la localitzada en una longitud i en una latitud (sistema de referència WGS84 o EPSG:4326).

3.1.4 CORINE Land Cover

CORINE Land Cover (CLC) és un projecte de l'Agència Europea de Medi Ambient (EEA) que produeix conjunts de dades per a descriure l'ús de sòl del territori europeu [26]. Pertany al programa de Copernicus Land Monitoring Service i ha publicat conjunts de dades en format GeoTIFF (.tif) pels anys 1990, 2000, 2006, 2012 i 2018 i en graelles amb una resolució de 100 metres.

Per a tenir les dades d'ús de sòl més pròximes als anys 2018 i 2019, s'obté el conjunt de dades de 2018, que conté diversos arxius. Els arxius que són d'interès per a l'objectiu del treball es troben al directori 'DATA' i s'anomenen U2018_CLC2018_V2020_20u1.tif.vat.dbf i U2018_CLC2018_V2020_20u1.tif.

El primer és un arxiu del sistema gestor de base de dades dBASE que es pot obrir amb la llibreria dbfread de Python. Conté informació sobre les metadades del segon arxiu. Entre altres coses, aquest arxiu assigna un codi per cada valor que pot tenir la imatge i que fa referència a un tipus de ús de sòl. A la Taula 5, es poden veure totes les classificacions de *land use*, que es poden descriure en un esquema de tres nivells de detall.

El segon arxiu està format per una imatge 2D on cada píxel descriu un tipus d'ús de sòl del nivell amb més detall (nivell 3 de la Taula 5), que es troba en una cel·la localitzada en un punt (en metres) dels eixos est i nord definits al sistema de referència EPSG:3050.

3.1.5 OpenStreetMap

OpenStreetMap és un projecte col·laboratiu i de lliure accés per crear mapes editables creat al 2004 i basat en l'èxit de Viquipèdia. Els usuaris poden modificar els mapes basant-se en imatges de satèl·lit o aèries, en aparells GPS o en el propi coneixement de la zona per a millorar i ampliar els mapes disponibles de manera gratuïta.

A més de l'opció d'editar mapes, OpenStreetMap permet exportar dades d'una zona sota llicència Open Data Commons Open Database License (ODbL) [27]. Una manera d'accedir a les dades és a través de l'API de Overpass, que s'implementa a Python a la llibreria overpy i que té el seu propi llenguatge de consulta.

Les dades estan estructurades segons nodes, vies o relacions. Els nodes són punts en el espai descrits per una longitud i una latitud (sistema de referència WGS84 o EPSG:4326), mentre que les vies són una llista ordenada d'aquests nodes. Les relacions són llistes ordenades de nodes, vies o altres relacions que defineixen relacions lògiques o geogràfiques amb altres elements. Tots aquests elements bàsics poden contenir etiquetes, que són parells clau-valor, i que defineixen una característica d'aquests. Si bé no hi ha una versió oficial de quines són les etiquetes a utilitzar en cada cas, la comunitat s'ha posat d'acord en molts casos per a seguir nomenclatures semblants. Molts cops aquestes es poden trobar a la wiki del projecte.

Taula 5. Classificació en nivells de l'ús de sòl de CORINE Land Cover.

Nivell 1	Nivell 2	Nivell 3	Codi
Superfícies artificials	Teixit urbà	Teixit urbà continu	111
		Teixit urbà discontinu	112
	Unitats industrials, de comerç i de transport	Unitats industrials o comercials	121
		Xarxes de carreteres	122
		Zones portuàries	123
		Aeroports	124
	Mines, abocadors i zones d'obres	Mines	131
		Abocadors	132
		Zona d'obres	133
	Zones no agrícoles amb vegetació	Zones verdes urbanes	141
		Instal·lacions d'esport i oci	142
Zones agrícoles	Terres de conreu	Terres de conreu no regades	211
		Terres de conreu permanentment regades	212
		Camps d'arròs	213
	Cultius permanents	Vinyes	221
		Arbres de fruita i plantacions de baies	222
		Olivars	223
	Pastures	Pastures	231
	Zones agrícoles heterogènies	Cultius anuals amb cultius permanents	241
		Patrons de cultiu complexos	242
		Zones agrícoles amb alguna zona de vegetació natural	243
		Zones agro-forestals	244
Boscos i zones seminaturals	Boscos	Boscos de fulla ampla	311
		Boscos de coníferes	312
		Boscos mixtos	313
	Arbustos i vegetació	Prats naturals	321
		Terra de brucs	322
		Vegetació esclerofil·le	323
		Arbust de transició a bosc	324
	Espais oberts amb poca vegetació	Platges i dunes	331
		Roques	332
		Zones amb poca vegetació	333
		Zones cremades	334
		Glaceres i neu perpètua	335
Aiguamolls	Aiguamolls interiors	Pantans	411
		Torberes	412
	Aiguamolls costaners	Aiguamolls	421
		Salines	422
		Planes de marees	423
Cossos d'aigua	Cossos d'aigua interiors	Cursos d'aigua	511
		Cossos d'aigua	512
	Cossos d'aigua marítims	Llacunes costaneres	521
		Estuaris	522
		Mar i oceà	523

Per a tractar dades de carreteres, la comunitat utilitza vies amb etiquetes amb la clau ‘highway’. Els valors que agafen les principals carreteres per aquesta etiqueta són ‘motorway’, ’trunk’, ‘primary’, ‘secondary’, ‘tertiary’, ‘unclassified’ i ‘residential’. En molts cops, les carreteres també contenen una etiqueta amb clau ‘lanes’ que fa referència al nombre de carrils que té aquesta. Un cop fet el tractament de la resta dades es realitzen consultes d’Overpass a través de la llibreria de overpy de Python per a obtenir les vies en un radi al voltant del punt d’interès que contenen l’etiqueta ‘highway’ igual a qualsevol dels 7 tipus de carreteres mencionats abans i que inclogui tots els nodes als que les vies fan referència. El resultat és un objecte ‘Result’ de la llibreria overpy que conté cada via i els nodes que la componen.

3.2 Integració i processat de les dades

En aquest apartat es descriu la integració de les dades que s’han obtingut a l’apartat anterior per a crear un únic *dataset* amb la finalitat de crear els models. El conjunt de dades final descriurà, a cada fila, una mesura en un punt espacial, que fa referència a la posició d’una estació de mesura de contaminants, i en un moment temporal, amb una freqüència temporal horària. S’utilitzen les llibreries de Python de pandas i NumPy per la creació del *dataset* i per fer els càlculs necessaris i donar el format adequat a les dades. Segons el tipus de dades i els arxius a tractar, s’utilitzen altres llibreries que es comentaran en els següents subapartats.

3.2.1 Estacions de contaminants

Els dos arxius de valors separats per comes obtinguts de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica contenen, en cada fila, les mesures d’una estació de les 24 hores d’un dia del contaminant al que fa referència cada un (NO_2 o O_3). Aquests arxius s’obren amb la llibreria pandas de Python i s’introdueixen les dades en objectes de tipus DataFrame, amb els quals es construirà el conjunt de dades final. Totes les mesures es troben en $\mu g/m^3$ per ambdós contaminats segons la columna UNITATS. Les columnes d’aquests conjunts de dades que ens interessen són les columnes de mesures (01h, 02h, etc.), els columnes de posició (LATITUD, LONGITUD i ALTITUD) i la columna de la data. No obstant això, les dades, en el format que estan i amb l’origen que tenen, presenten tres inconvenients principals.

El primer inconvenient està relacionat amb les posicions de cada estació. S’observa que les posicions de les estacions amb el mateix codi poden variar associats a un possible error o un canvi en la mesura de la posició. Per a tenir dades consistents amb unes estacions definides en un sol punt, s’arrodoneixen la latitud i la longitud a 0,0001º, corresponent a un error de fins a 6 metres en cada direcció.

El segon inconvenient està relacionat amb les estacions i els contaminants. Cada contaminant té un nombre diferent d’estacions i no són necessàriament les mateixes. Per l’ozó s’observen 49 estacions i pel diòxid de nitrogen s’observen 66, però només 40 estacions tenen observacions dels dos contaminants. Per a evitar crear valors nuls en punts on hi hagi un contaminant però no l’altre, es descarten les dades que no pertanyin a aquestes 40 estacions.

L'últim inconvenient és que el format de les mesures de cada contaminant a cada hora no es correspon amb el format que es vol per conjunt de dades final. En el conjunt de dades final, una fila correspon a un únic instant en el temps mentre que en aquests arxius correspon a 24 hores diferents. Per això, per cada fila, cal transformar els 24 valors de les columnes d'hores en 24 files diferents i 2 columnes corresponents a l'hora i el valor de la mesura.

Un cop fet això, s'ajunten les files dels dos conjunts de dades de cada contaminant i després es passen els valors que la columna CONTAMINANT (que poden ser 'O3' o 'NO2') a dues columnes corresponents a cada contaminant per cada punt espaciotemporal. Finalment, s'obté el *timestamp* sumant la columna d'hores extreta abans amb DATA, del qual es creen columnes per l'any, el mes, el dia i l'hora.

3.2.2 Dades meteorològiques

Els dos arxius de netCDF obtinguts del reanàlisi d'ERA5 contenen valors de les variables meteorològiques d'interès amb diferents resolucions. Aquests arxius s'obren amb la llibreria netCDF4 de Python, amb la qual es poden obtenir els valors de les variables a partir dels índexs de les diferents dimensions (temps, latitud i longitud). Per saber els índexs que cada una de les files de les dades dels contaminants ocupa només cal saber quins valors ocupen l'índex 0 de cada dimensió i la resolució d'aquesta.

Dit això, es prioritzen les dades de ERA5-Land, que tenen una resolució espacial més fina (0,1º) i que, per tant, poden aportar més detall a cada fila de dades. D'aquest conjunt de dades s'extreuen la radiació neta en superfície i la temperatura a 2 metres i es calculen el mòdul de la velocitat i la direcció del vent a partir de las components zonal (u) i meridional (v) del vent a 10 metres. Aquestes variables s'afegeixen com a columnes al *dataset* final. No obstant, conjunt de dades ERA5-Land conté diversos valors buits, que suposa que el *dataset* final tingui un 12,81% de files amb valors buits.

Amb l'objectiu d'omplir els valors buits creats pel conjunt de dades anterior, s'utilitza el segon conjunt de dades meteorològiques ERA5, que té una resolució menor (0,25º) però no té valors buits. D'aquesta manera, s'agafen les files que tenen algun valor nul i es tornen a extreure les variables afegides anteriorment de la mateixa manera, donat que el format és el mateix. Addicionalment, aquest conjunt de dades proporciona dues noves variables (la radiació descendent ultraviolada en superfície i la nuvolositat) que s'afegeixen al conjunt de dades final de la mateixa manera que la resta.

3.2.3 Buffers circulars

Les substàncies contaminants, com altres gasos i substàncies presents a l'atmosfera, es poden veure afectades pels processos de transport, causats pel moviment de les masses d'aire. Consegüentment, la concentració de contaminants en un punt pot no ser influenciada només pels valors que agafen certes variables, com l'ús de sòl o la densitat de carreteres, en aquest punt sinó en tota una zona al voltant del punt. En

aquesta zona anomenada zona d'influència o *buffer* s'hi poden aplicar càlculs d'agregació com la suma o la mitjana per a tenir una idea més detallada dels voltants del punt.

En general, els *buffers* es construeixen a partir de tots els elements que es troben dins una certa distància màxima o radi, construint així un cercle d'influència. No obstant això, les distàncies es poden calcular de diferents maneres segons la mètrica escollida i el sistema de referència o la projecció utilitzada per a descriure els punts. Per exemple, recórrer 0,01º cap a l'est a prop de l'equador és recórrer aproximadament 1 kilòmetre, però fer-ho a una latitud de 60º, com pot ser en un país d'Escandinàvia, redueix aquest kilòmetre a la meitat. En canvi, recórrer 0,01º cap al nord sempre serà equivalent a un kilòmetre per com el sistema de referència WGS84 està construït.

Una manera de calcular les distàncies de manera correcta en qualsevol seria calcular distàncies geodèsiques, que s'ajusten a la curvatura de l'espai (en aquest cas, la Terra). No obstant, com que els *buffers* es troben centrats en un punt d'interès, totes les distàncies es calculen respecte a aquest punt i llavors, potser resulta més eficient canviar les coordenades per a tenir una projecció on tots els punts es troben a la mateixa distància del punt central del *buffer*. Aquesta projecció s'anomena projecció azimutal equidistant. A la Figura 1, es pot veure com es transforma un mapa si es realitza aquesta projecció per la longitud i la latitud de Barcelona. També es pot veure l'equivalència de transformar punts equidistants de aquesta projecció a una projecció de latituds i longituds equiespaciadas.

D'aquesta manera, es creen *buffers* en diferents radis fent servir la projecció azimutal equidistant en cada estació per variables relacionades amb la població, amb l'ús del sòl i amb les carreteres. Això és possible ja que les dades que es tenen en aquests àmbits són imatges d'un moment temporal, o dos en el cas de la població, cosa que permet no calcular el *buffer* a cada punt espaciotemporal (cada fila del *dataset*) sinó calcular-lo un cop o dos cops a cada punt espacial (cada estació). Addicionalment, els *buffers* creats utilitzen radis mínims i màxims de tal manera que les variables amb radis grans no incloguin a les variables amb radis petits, ja que això incrementaria la informació mútua que comparteixen i, de cara als models, competirien per explicar variacions d'ozó semblants.

3.2.3.1 Població

Pel cas de les dades del nombre d'habitants, s'utilitzen els dos conjunts de dades de Worldpop corresponents a 2018 i 2019 per a crear *buffers* amb radis de 300, 500 i 1000 metres i fent ús de la projecció azimutal equidistant. Per cada estació i cada any, s'accedeix a l'arxiu de l'any corresponent i es calcula la mitjana del nombre d'habitants dels punts inclosos en el *buffer* corresponent. Els punts que no tenen dades, com pot ser el mar, es compten com a 0 en la mitjana. D'aquesta manera, s'obtenen tres variables noves pel *dataset* referents als tres radis i que comporten tres anells que van de 0 a 300 metres, de 300 a 500 metres i de 500 a 1000 metres.

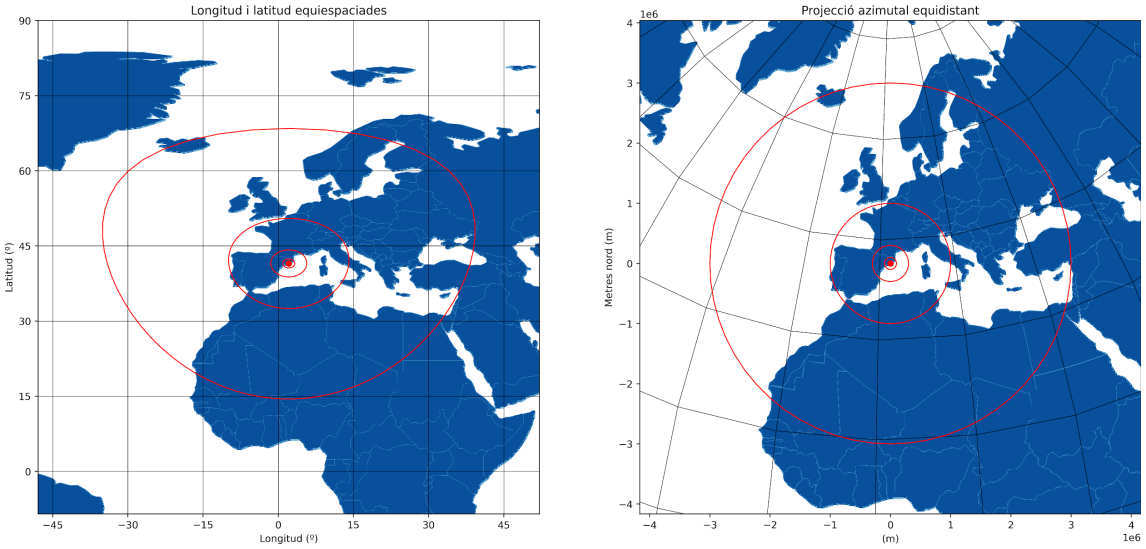

Figura 1. Comparació de la projecció azimutal equidistant (dreta) amb un mapa amb latitud i latitud equiespaciades (esquerra). Es calculen cercles (vermell) de punts a la mateixa distància a la projecció i es trasllada al mapa de longituds i latituds.

3.2.3.2 Ús de sòl

Pel cas de les dades d'ús de sòl de 2018, s'utilitzen radis de 500, 1000, 2000, 3000 i 5000 metres per a crear *buffers* per a cada estació. Per fer-ho, s'agrupen les diferents classes d'ús de sòl definides per CORINE Land Cover segons el nivell amb menys detall (nivell 1 de la Taula 5). D'aquesta manera, la imatge GeoTIFF amb les dades de *land use* pot contenir 5 valors diferents referents a superfícies artificials, zones agrícoles, boscos i zones seminaturals, aiguamolls i cossos d'aigua.

A continuació, es creen 25 noves variables referents als cinc radis i a les cinc categories d'ús de sòl fent servir la projecció azimutal equidistant per a cada estació i *buffers* en forma d'anells per a calcular el fracció o percentatge de cops que apareix la categoria dins la zona d'influència. De manera semblant al cas anterior, les possibles dades buides no es compten a cap categoria i, per tant, la suma per totes les categories en un mateix buffer pot ser menor a 1.

3.2.3.3 Carreteres

Pel cas de les carreteres, s'utilitzen radis de 500, 1000, 2000, 3000 i 5000 metres per a crear *buffers* per a cada estació. Per fer-ho, es realitza una consulta d'Overpass API per cada estació on se seleccionen totes les vies que pertanyin a algun tipus de les carreteres principals dins del radi més gran (5000 metres). A partir de les dades obtingudes de la consulta i per cada estació, es construeixen dos imatges amb una resolució de 0,001º on s'uneixen els nodes de cada carretera amb línies rectes.

Per una banda, la primera imatge permet diferenciar els diferents tipus de carreteres agafant diferents valors segons per cada tipus. A partir d'aquesta imatge es creen dos tipus de variables. En primer lloc, es crea un *buffer* per cada radi on es calcula la fracció de píxels no buits o, dit d'una altra manera, la densitat de carreteres de la zona,

independentment del seu tipus. D'aquest enfocament, s'obtenen 5 variables, una per cada radi. En segon lloc, es calcula la distància a la carretera més propera de cada tipus. Per a poder descriure la falta d'un tipus de carretera en la zona d'influència màxima (5000 metres), és calcula l'invers de la distància i es força una distància mínima de 1 metre. D'aquesta manera, l'interval d'aquestes variables es troba a l'interval [0,1]. Finalment, s'eleva aquesta distància inversa al quadrat per tractar de donar-li un sentit més físic i s'afegeix al *dataset*. D'aquest enfocament, s'obtenen 7 variables, una variable per cada tipus de carretera.

Per l'altra banda, la segona imatge compta quants carrils es troben en un mateix píxel amb l'objectiu de donar una idea del volum de tràfic en la zona. Les vies que no tinguin l'etiqueta 'lanes', és a dir, que no tenen informació sobre el nombre de carrils d'aquesta, són tractat com si tinguessin 1 carril. A partir d'aquí, es crea un *buffer* per cada radi on es recompten el nombre de carrils i es divideix pel nombre de píxels dins la zona d'influència o, dit d'una altra manera, la densitat de carrils de la zona. D'aquest enfocament, s'obtenen 5 variables, una per cada radi.

3.2.4 Buffers semicirculars

Com s'ha comentat abans, el moviment de les masses pot ser de gran influència de cara a la predicció de les concentracions de contaminants. Per això, es proposa l'ús de *buffers*, que intenten tenir en compte les zones d'on les masses d'aire poden venir. No obstant això, els *buffers* circulars no tenen en compte la direcció del vent, sinó que agafen una àrea d'on les masses d'aire poden venir. Per això, s'ha proposat l'ús de zones d'influència semicirculars amb l'objectiu de tenir variables més significatives [28]. El *buffer* circular es divideix segons l'eix perpendicular a la direcció del vent, creant una zona que està a favor del vent o, en anglès, *downwind* i una zona que està en contra del vent o *upwind*, com es pot veure a la Figura 2. Consegüentment, els *buffers* semicirculars són funció de la direcció del vent i, per tant, són diferents en cada punt espaciotemporal i, llavors, s'han de calcular per cada fila del conjunt de dades final.

Per cada *buffer* circular calculat sobre les dades de *land use* i de carreteres, es crea una variable a favor del vent i una altra en contra d'aquest, afegint un total de 70 variables noves el conjunt de dades final. En resum, el procediment a seguir és equivalent al procediment seguit en els apartats anteriors però en cada punt espaciotemporal s'alinea l'eix X (eix zonal) de la projecció azimutal equidistant amb el vent, de tal manera que les coordenades positives en el nou eix es troben a favor del vent i les coordenades negatives es troben en contra d'aquest.

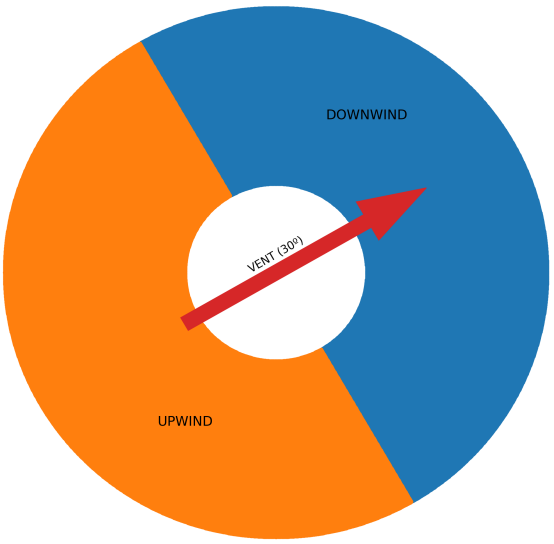

Figura 2. Representació d'una zona d'influència semicircular depenent la direcció del vent (fletxa), que en aquest cas és de 30º. La zona a favor del vent o *downwind* es mostra en color blau i la zona en contra d'aquest en color taronja.

3.2.5 Mitjanes i diferències temporals

En els apartats anteriors, s'ha comentat com els *buffers* tracten de tenir en compte com la contaminació en un punt es pot veure influenciada pels seus voltants a causa de fenòmens de transport atmosfèric. D'una manera similar, les condicions que es donen en moments anteriors poden afectar a les concentracions de contaminants en el moment actual. Saber la tendència temporal d'alguna variable o el seu valor mitjà en les últimes hores podria ser útil per a explicar la variació d'O3.

No obstant això, les variables que conté el *dataset* en el seu estat actual gairebé no contenen informació d'aquest tipus. Per aquesta raó, es calculen les mitjanes temporals del diòxid de nitrogen, de la radiació ultraviolada, de la radiació en superfície, de la temperatura i de la velocitat del vent de les últimes 3, 8 i 24 hores. Addicionalment, es calcula la diferència del NO_2 de l'instant respecte al NO_2 de 1, 2 i 3 hores abans.

D'aquesta manera, s'agreguen les últimes 18 variables al dataset final que intenten actuar com a memòria temporal pels models. El conjunt final de dades conté 664878 files i 150 columnes i no conté cap valor nul. Finalment, es guarda en un arxiu csv anomenat Dataset_final.csv.

3.3 Creació de models

Un cop creat el conjunt de dades final, ja es poden construir models per a la predicció d'ozó. El tipus de models que s'utilitzen en aquest treball són els *gradient boosting machines* (GBM) a partir d'arbres de decisió o regressió (GBDT), ja que s'ha vist que obtenen uns dels millors resultats i tarden menys que altres algorismes amb resultats similars [13].

En l'aprenentatge automàtic, el *boosting* és una manera de combinar diferents algorismes (*ensemble*) d'un grau de complexitat baix i amb una base semblant, com poden ser molts arbres de decisió. Els arbres de decisió, en aquest cas, realitzen prediccions seqüencialment de manera que cada un s'encofa en explicar millor les dades on l'arbre anterior ha obtingut més error, de manera que no es descarten les bones decisions preses [29]. Encara que el model de *boosting* es construeix seqüencialment, un cop entrenat, les prediccions es calculen paral·lelament a cada arbre, obtenint una predicció parcial per cada arbre, i després es combinen els resultats per a obtenir el resultat final.

Els *gradient boosting machines* (GBM) són una generalització de *boosting*. L'algoritme principal de *boosting* anomenat AdaBoost tracta de minimitzar la funció de pèrdua exponencial mentre que GBM pot crear models de *boosting* optimitzant una funció de pèrdua diferenciable arbitrària [30]. D'aquesta manera, s'utilitza la llibreria de Python LightGBM per a crear models de regressió de GBM fent ús d'arbres de decisió i de la funció de l'arrel de l'error quadràtic mig (RMSE). Les dades se separen en subconjunts d'entrenament, validació i testeig en proporcions de 64%, 16% i 20% (el conjunt de testeig és una quinta part del total i el de validació és una quita part de la resta). El conjunt de validació s'utilitza per a reduir el sobreentrenament del model (aturant l'entrenament abans si la validació no millora) mentre que el conjunt de testeig s'utilitza per fer l'avaluació final del model.

Els models es construeixen de dues maneres. En un primer enfocament, es calculen models per cada estació amb les dades corresponents, fins a un total de 40 models. S'espera que aquests models se sobreajustin a cada estació de tal manera que predir la concentració d'ozó fora de l'estació corresponent doni resultats molt inferiors mentre que predir-la dins l'estació doni resultats millors o iguals a qualsevol altre model. Aquest enfocament no és útil a la pràctica perquè, òbviament, no existeix un 'codi d'estació' per qualsevol punt espacial, però sí serveix per a comparar el rendiment d'altres models. En un segon enfocament, es calculen models generals per totes les dades. S'espera que aquests models siguin d'utilitat per a extrapolar el comportament de l'ozó de les estacions a altres llocs.

3.3.1 Selecció de variables

Com s'ha vist abans, el conjunt de dades per a fer el model conté un total de 150 variables, o 147 variables predictores si es descarta l'ozó, el codi de l'estació i la marca temporal (ja hi ha any, mes, dia i hora). D'aquestes 147 variables es troba que 12 són constants per totes les files. Ninguna estació conté cossos d'aigua a menys de 500 metres ni aiguamolls a menys de 2000 metres, resultant en, respectivament, 3 i 9 variables constants si es compten els diferents radis (500, 1000 i 2000 metres) i que cada cada *buffer* circular té dos *buffers* semicirculars associats. Aquestes variables constants són descartades, resultant en un total de 135 variables predictores.

Crear un model amb totes les variables predictores no és viable. Si bé els resultats serien força bons, aquests podrien sobreajustar-se al conjunt de dades, amb variables que tinguin més soroll que capacitat predictiva. A més, el nombre de variables

necessàries (135) i el temps que tarda un model d'aquestes dimensions en fer una predicció són molt grans com per a que es pugui aplicar de manera pràctica al dia a dia o, fins i tot, hora a hora.

Per aquest motiu, se seleccionen un nombre reduït de variables per a crear models amb rendiments suficientment bons que no incrementin gaire al incluir alguna variable més. No obstant això, no valen qualsevol nombre de variables: s'han de tractar d'escollir les millors variables per un cert nombre de variables. Això pot suposar fer una cerca molt extensa sobre quina combinació de variables és la millor per cada nombre de variables. Per exemple, el nombre de combinacions de 4 variables de les 135 variables possibles és 13232835. Per aquesta raó, se segueix un enfocament molt més senzill i realista: en lloc, de seleccionar la millor combinació de variables per un cert nombre de variables N, se selecciona iterativament aquella variable que és millor fins a arribar a N variables. D'aquesta manera, per 4 variables només s'han de realitzar 534 cerques (135 per una variable, 134 per dues variables, etc.) que, a canvi, es suposa que la millor combinació de 4 variables inclou a la millor combinació de 3 variables, que inclou a la millor de 2, que inclou a la millor de 1.

Un altre punt important és quin criteri s'utilitza per dir quina variable és la millor. Una opció és ajustar un model molt senzill per cada possible variable, avaluar una predicció amb una mètrica com l'arrel de l'error quadràtic mig (RMSE), el coeficient de correlació de Pearson (R) o el quadrat d'aquest (R^2) i seleccionar la variable que minimitza l'error. Algunes de les possibles mètriques es poden veure a la Taula 6. El model més senzill per a poder-ne crear molts és la regressió lineal múltiple. Si bé el temps per a crear el model i fer la predicció és mínim amb aquesta regressió degut a la seva simplicitat, les millors variables que responen millor a una dependència lineal no necessàriament són les mateixes que responen millor a un model GBM, que no és lineal. Per aquesta raó, el model senzill que s'utilitza per a avaluar les millors variables és un GBM amb menys potència (comparat amb el que s'utilitza en els apartats següents) per a reduir el temps per a crear i avaluar el model.

En conclusió, es creen models senzills de GBM per a obtenir les millors 15 variables del conjunt de dades. Aquest models utilitzen un màxim de 20 fulles per cada arbre i s'utilitzen un màxim de 100 rondes o arbres de *boosting*. L'algorisme s'atura abans si no troba millora en la validació en més de 20 rondes. S'utilitza el coeficient de correlació de Pearson al quadrat (R^2) per a determinar la millor variable per cada nombre de variables. Donat que, com s'ha comentat anteriorment, s'utilitzen dos enfocaments que creen models per cada estació o un model per tots els punts, se seleccionen per separat les millors variables per cada estació i les millors variables per tot el conjunt de dades.

Taula 6. Algunes possibles mètriques d'avaluació dels models. Els valors de y representen els valors reals i $\widehat{y}$ representen els valors predits per les mostres de N punts. Les mitjanes es representen com $\bar{y}$ i $\bar{\widehat{y}}$, respectivament.

Nom	Definició	
Error quadràtic mig	$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$	(eq.1)
Error absolut mig	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \widehat{y}_i \rvert$	(eq.2)
Correlació de Pearson	$R = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \bar{\widehat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\widehat{y}_i - \bar{\widehat{y}})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$	(eq.3)
Coeficient de determinació	$\rho^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$	(eq.4)

A la Taula 7, es poden veure les tres variables més escollides pel model d'estacions per cada nombre de variables del model. Les 15 millors variables que s'han considerat per cada estació es poden trobar a l'Annex 1, conjuntament amb els valors de R^2. S'observa que, en molts de casos, les variables de diòxid de nitrogen i de la radiació ultraviolada en una mitjana de les últimes 8 hores són seleccionades ràpidament. Addicionalment, les variables del dia i el mes i la temperatura mitjana en les últimes 24 hores apareixen freqüentment en la primera meitat de la taula. També s'observa una manca de variables relacionades amb *buffers* de ús de sòl i de carreteres. Donada la separació de les dades segons la seva estació, que ocupa només un punt espacial, els *buffers* circulars relacionats amb aquestes variables esdevenen constants i els *buffers* semicirculars són funció només de la direcció del vent i perden molta capacitat de predicció. Finalment, s'observa que els valors mitjans del coeficient de Pearson (R) arriben a 0,942 pel màxim de 15 variables i que amb 4 variables ja se supera, en general, el 0,900. A la Figura 3, es poden veure, per cada estació, els valors individuals del coeficient de Pearson al quadrat. La majoria de R^2 es troben al voltant de 0,90 per 15 variables amb algunes poques que no arriben a 0,85.

Taula 7. Les tres variables més escollides per cada nombre de variables pel model d'estacions amb el seu percentatge de selecció en les estacions. Es mostra la mitjana del coeficient de Pearson de totes les estacions per cada nombre de variables.

Nombre variables	Variables escollides	Mitjana R
1	UV_SURFACE_MEAN_8h (55,00%)	0,713
	NO2 (32,50%)	
	UV_SURFACE_MEAN_24h (12,50%)	
2	NO2 (42,50%)	0,836
	UV_SURFACE_MEAN_8h (22,50%)	
	RADIATION_SURFACE (10,00%)	
3	MES (40,00%)	0,879
	TEMPERATURE_MEAN_24h (15,00%)	
	RADIATION_SURFACE_MEAN_8h (10,00%)	
4	DIA (30,00%)	0,902
	MES (17,50%)	
	TEMPERATURE_MEAN_24h (15,00%)	
5	WIND_SPEED_MEAN_8h (17,50%)	0,915
	DIA (17,50%)	
	TEMPERATURE_MEAN_24h (12,50%)	
6	DIA (17,50%)	0,925
	TEMPERATURE_MEAN_24h (17,50%)	
	WIND_SPEED_MEAN_8h (12,50%)	
7	DIA (22,50%)	0,931
	TEMPERATURE_MEAN_24h (20,00%)	
	ANY (10,00%)	
8	ANY (17,50%)	0,935
	UV_SURFACE_MEAN_24h (12,50%)	
	NO2_MEAN_24h (10,00%)	
9	ANY (22,50%)	0,937
	NO2_MEAN_24h (12,50%)	
	RADIATION_SURFACE_MEAN_24h (10,00%)	
10	RADIATION_SURFACE_MEAN_24h (15,00%)	0,939
	UV_SURFACE_MEAN_24h (10,00%)	
	ANY (10,00%)	
11	NO2_MEAN_24h (20,00%)	0,940
	WIND_SPEED_MEAN_24h (12,50%)	
	HORA (10,00%)	
12	HORA (10,00%)	0,941
	RADIATION_SURFACE_MEAN_24h (7,50%)	
	RADIATION_SURFACE_MEAN_3h (7,50%)	
13	ANY (10,00%)	0,941
	POPULATION_CIRCULAR_0-300m (10,00%)	
	WIND_DIRECTION (7,50%)	
14	POPULATION_CIRCULAR_0-300m (12,50%)	0,942
	POPULATION_CIRCULAR_300-500m (10,00%)	
	ANY (7,50%)	
15	POPULATION_CIRCULAR_0-300m (17,50%)	0,942
	POPULATION_CIRCULAR_300-500m (12,50%)	
	POPULATION_CIRCULAR_500-1000m (10,00%)	

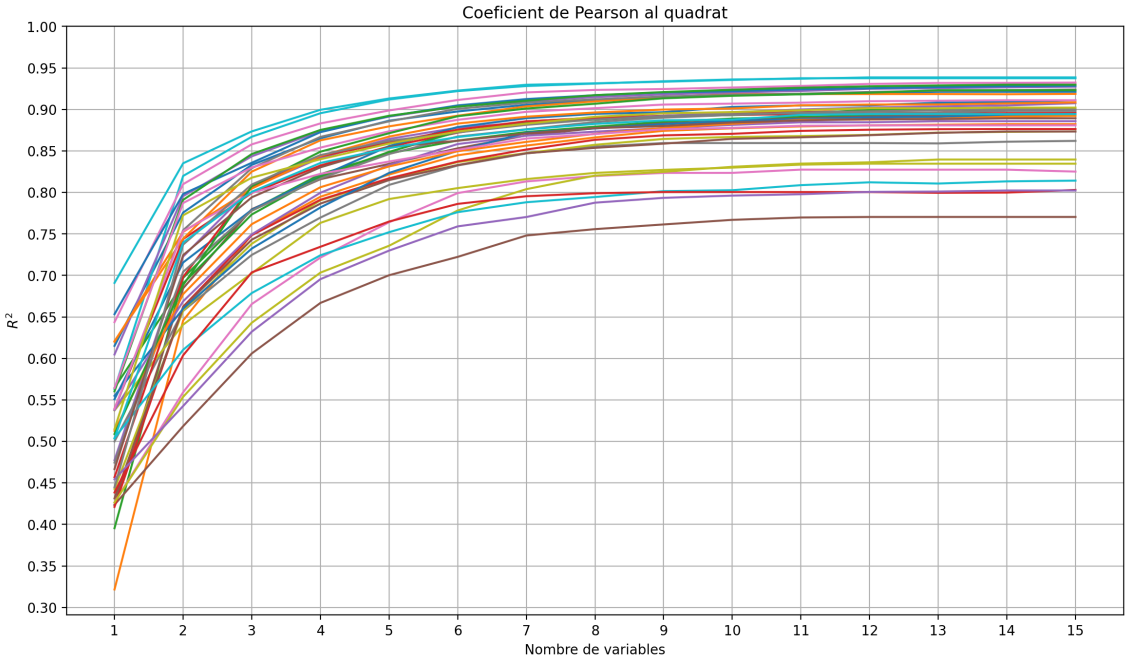

Figura 3. Valors del coeficient de Pearson al quadrat de les diferents estacions per la selecció de variables.

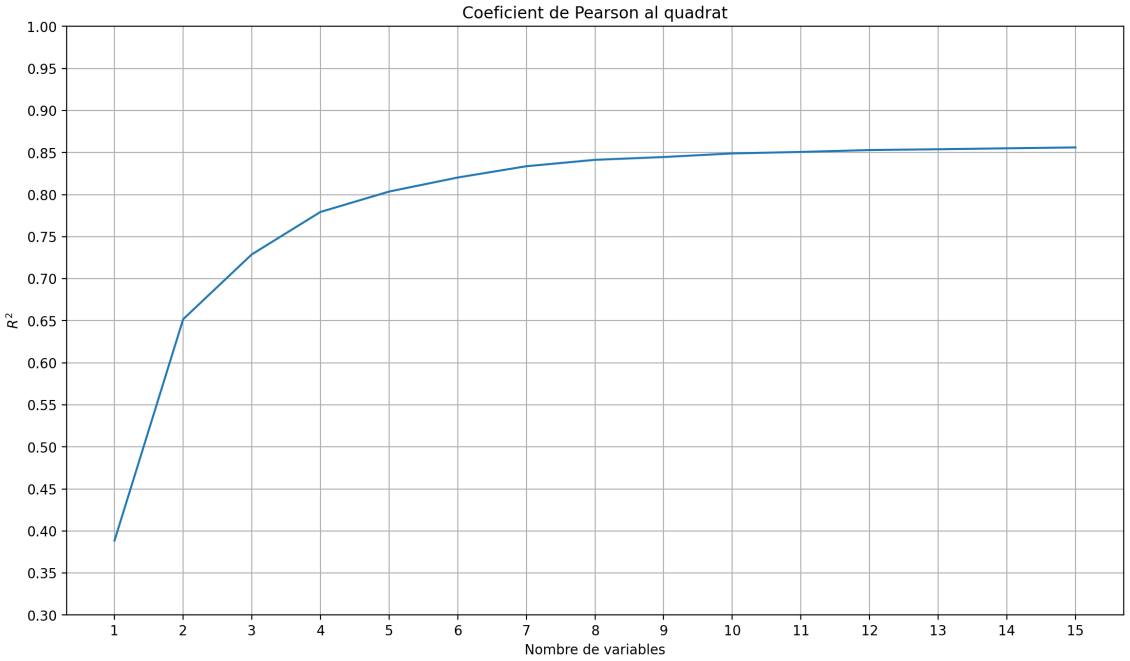

Figura 4. Valors del coeficient de Pearson al quadrat del model de totes les dades per la selecció de variables.

Taula 8. Les variables escollides per cada nombre de variables pel model de totes les dades. Es mostra el valor que agafa el coeficient de Pearson, el seu quadrat i el coeficient de determinació per cada nombre de variables.

Nombre variables	Variables escollides	R	R^2	ρ^2
1	UV_SURFACE_MEAN_8h	0,623	0,388	0,389
2	NO2	0,807	0,651	0,652
3	LANE_DENSITY_CIRCULAR_2000-3000m	0,854	0,729	0,729
4	MES	0,883	0,780	0,779
5	RADIATION_SURFACE_MEAN_3h	0,896	0,803	0,804
6	WIND_SPEED_MEAN_8h	0,906	0,821	0,820
7	DIA	0,913	0,834	0,834
8	TEMPERATURE_MEAN_8h	0,917	0,841	0,841
9	ANY	0,919	0,845	0,844
10	FOREST_CIRCULAR_1000-2000m	0,921	0,848	0,849
11	HORA	0,922	0,850	0,850
12	ARTIFICIAL_CIRCULAR_3000-5000m	0,924	0,854	0,852
13	WIND_SPEED_MEAN_24h	0,924	0,854	0,855
14	ROAD_DENSITY_CIRCULAR_1000-2000m	0,925	0,856	0,855
15	POPULATION_CIRCULAR_0-300m	0,925	0,856	0,856

A la Taula 8, es poden veure les variables escollides pel model que utilitza totes les dades del *dataset* per cada nombre de variables del model. De manera similar al cas anterior, la radiació ultraviolada mitjana de les últimes 8 hores i el diòxid de nitrogen apareixen com les dues millors variables, amb el mes i el dia apareixent a la primera meitat de la taula. S'observa que ara sí que apareixen variables de *buffers*, donat que sí hi ha una distribució espacial de les dades. Concretament, la densitat de carrils en el *buffer* circular de 2000 metres a 3000 metres és la tercera variable escollida i el *buffer* circular de boscos entre 1000 i 2000 metres és seleccionada la decena. S'observa una manca de *buffers* semicirculars, el que ens indica que no han obtingut una capacitat de predicció de l'ozó major que els *buffers* circulars. Finalment, el valor del coeficient de correlació de Pearson es troba sempre una mica per sota del cas anterior, però arriba a 0,925 amb 15 variables i és major a 0,900 per 6 variables, tan sols una variable més que en l'altre model. A la Figura 4, es representen els valors de R^2 en funció del nombre de variables, que arriben a 0,856 per 15 variables.

3.3.2 Optimització del nombre de variables

Un cop escollides les millors variables per cada nombre de variables cal ajustar quin és el nombre de variables més eficient per a fer una predicció bona de l'ozó. Com que el model amb més variables segurament obtindrà millors resultats, es tracta d'escollir un nombre de variables pel qual la millora per afegir una variable més sigui mínima.

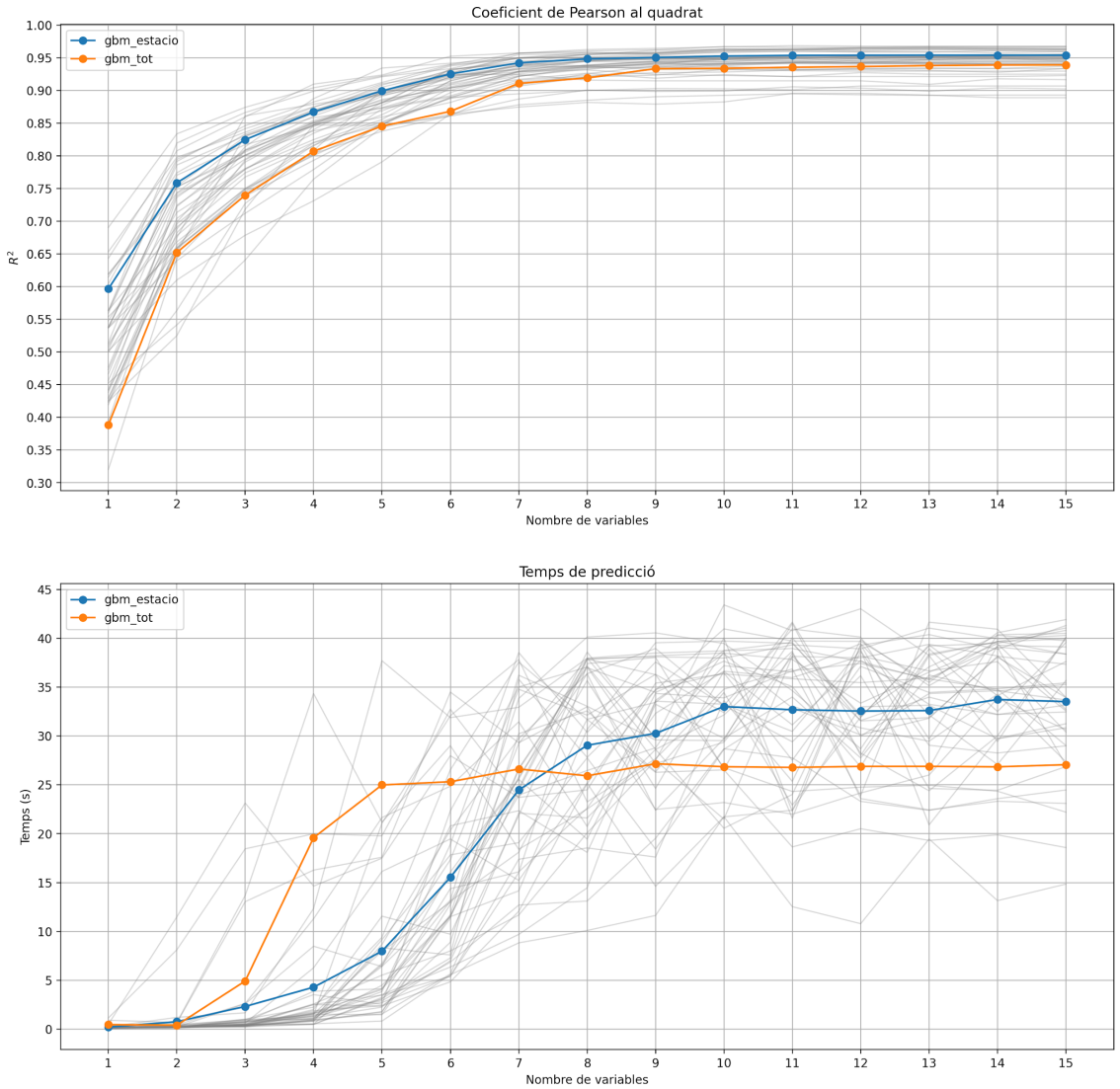

Figura 5. Coeficient de Pearson al quadrat i temps de predicció del model per estacions (blau) i model per totes les dades (taronga) segons el nombre de variables utilitzades. En gris es mostren els models de les estacions individuals. El temps dels models individuals ha estat multiplicat pel nombre d'estacions (40) per facilitar la visualització.

Es creen models GBM més complexos que els anteriors i s'avaluen també amb el coeficient de Pearson al quadrat tant pel model per estacions com pel model que utilitza totes les dades. En aquest cas, els models utilitzen un màxim de 31 fulles per cada arbre i s'utilitzen un màxim de 5000 rondes o arbres de *boosting*, aturant-se en no trobar millora en la validació en més de 100 rondes. Per a poder comparar millor els dos models, pel model d'estacions es fan prediccions parcials corresponents a cada estació i es calcula el valor de R^2 comparant tots els valors de ozó amb totes les prediccions.

A la Figura 5, es comparen les mètriques i el temps de predicció de cada model segons les variables escollides. Es pot veure que el model per estacions obté valors del coeficient de Pearson majors per qualsevol nombre de variables arribant a un R^2 de 0,954 per 15 variables mentre el segon obté un valor de 0,939. S'observa que, en ambdós casos, a partir de 9 variables la millora és gairebé nul·la. El model de totes les

dades presenta increments significatius per 7 i 9 variables en comparació amb els valors de R^2 amb més variables i, per aquesta raó, s'escolleixen com a models finals. En quant al temps, s'ha mesurat el temps total que tarda a fer la predicció de totes les dades. La mesura s'ha fet en segons i fent ús de l'eina Google Colaboratory. S'observa que per moltes variables el temps s'aproxima a dos límits constants per cada model, relacionat amb el límits definits pels hiperparàmetres del GBM (nombre de fulles i/o nombre d'arbres arriba a un màxim). El model per estacions mostra una dependència més continua segons el nombre de variables donat a que està compost de diferents models que arriben al límit en diferents models.

3.3.3 Avaluació del models finals

Taula 9. Mètriques resultants de comparar els valors reals d'ozó amb les prediccions dels models de 7 i 9 variables.

Model	**RSME (μg/m³)**	**R**	**R^2**	**ρ^2**
7 variables	9,884	0,954	0,911	0,911
9 variables	8,548	0,966	0,933	0,933

Taula 10. Selecció de variables en els nodes dels arbres i guany que generen aquestes en cada node pels models de 7 i 9 variables.

Variable	**Selecció**		**Guany**	
	7 variables	**9 variables**	**7 variables**	**9 variables**
UV_SURFACE_MEAN_8h	22004 (14,669%)	16976 (11,317%)	9,234e+08 (40,213%)	9,061e+08 (38,721%)
NO2	16702 (11,135%)	14093 (9,395%)	6,322e+08 (27,534%)	6,296e+08 (26,905%)
LANE_DENSITY_CIRCULAR_2000-3000m	19459 (12,973%)	18732 (12,488%)	1,981e+08 (8,627%)	1,951e+08 (8,336%)
MES	19019 (12,679%)	13676 (9,117%)	1,717e+08 (7,479%)	1,496e+08 (6,392%)
RADIATION_SURFACE_MEAN_3h	22538 (15,025%)	17566 (11,711%)	1,865e+08 (8,124%)	1,689e+08 (7,218%)
WIND_SPEED_MEAN_8h	24713 (16,475%)	19891 (13,261%)	1,134e+08 (4,937%)	1,091e+08 (4,664%)
DIA	25565 (17,043%)	23426 (15,617%)	7,085e+07 (3,085%)	7,846e+07 (3,353%)
TEMPERATURE_MEAN_8h		20311 (13,541%)		7,954e+07 (3,399%)
ANY		5329 (3,553%)		2,368e+07 (1,012%)

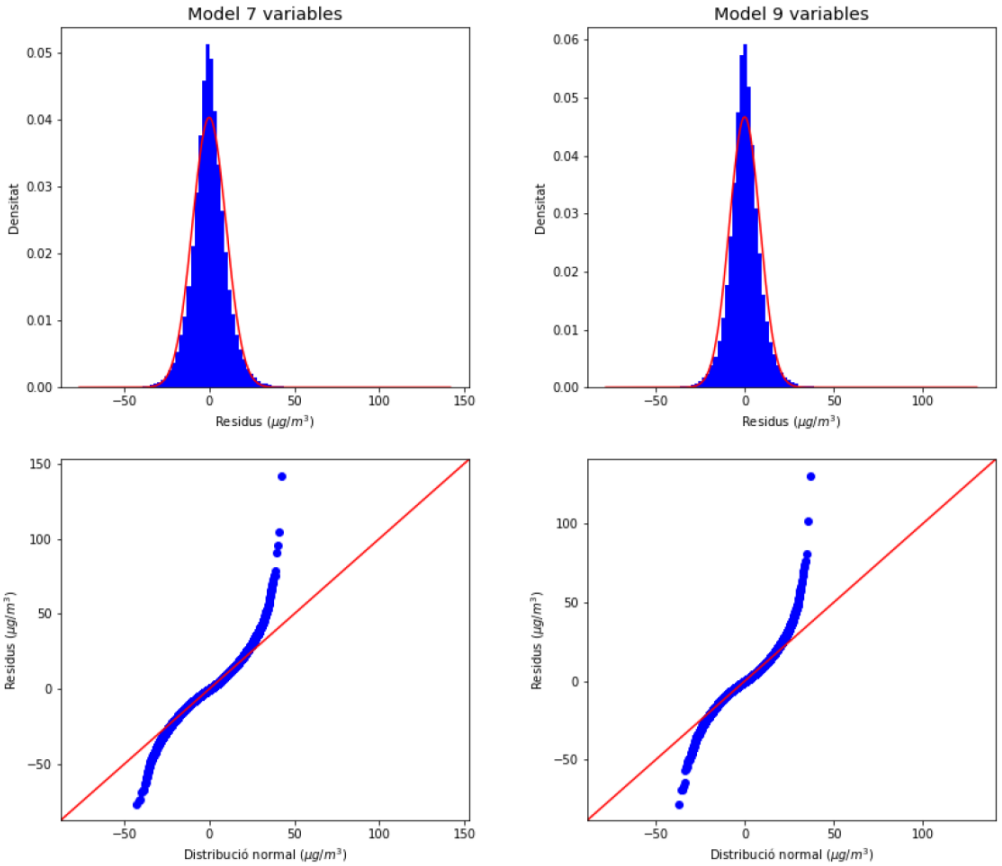

Figura 6. Avaluació de la normalitat dels residus pels models de 7 (esquerra) i 9 variables (dreta). A dalt, es compara la distribució dels residus (blau) amb una distribució normal centrada al 0 i amb la mateixa variància que els residus (vermell). A baix, un gràfic Q-Q compara els quantils de la distribució dels residus amb els quantils d'una distribució normal centrada al 0 amb la mateixa variància.

A la Taula 9, es resumeixen les diferents mètriques dels models de totes les dades per 7 i 9 variables. Com s'ha vist a l'apartat anterior, ambdós models superen un R^2 de 0,90 però les diferències entre els dos no són gaire grans. A continuació, s'estudia la importància de cada variable segons dues mètriques.

En primer lloc es mesura el nombre de cops que cada variable apareix en els nodes dels arbres que componen el model GBM, en els quals es genera una decisió que separa les dades. A la Taula 10, es poden veure el nombre de cops que s'ha seleccionat una variable i el percentatge respecte al total pels dos models sota la columna 'Selecció'. S'observa una selecció més o menys equilibrada excepte per la variable ANY pel model de 9 variables, que potser és degut a que només hi ha dos valors per a aquesta variable (2018 o 2019).

En segon lloc, s'estudia el guany que genera cada variable en separar el conjunt de dades en dos en un node. També a la Taula 10, es poden veure la suma dels guanys generats per una variable i el percentatge respecte al total pels dos models sota la columna 'Guany'. S'observa que, en general, les variables que s'havien seleccionat primer en la selecció de variables són les que més guany causen, confirmant que el procés de selecció de variables funciona com s'espera. S'hi troben algunes diferències

com que la mitjana de 3 hores de la radiació en superfície obté al voltant de un 0,8% més guany que la variable del mes pels dos models malgrat estar la segona per davant en la selecció de variables. Seguidament, es duu a terme un anàlisi de residus qualitatiu. Els residus es poden calcular com:

$$\widehat{e_i} = y_i - \widehat{y_i} \qquad \text{(eq. 5)}$$

on y és el valor real i $\widehat{y}$ és la predicció del model. Si els models de regressió construïts són correctes, els residus es poden descriure com errors aleatoris i, per tant, segueixen una distribució normal centrada al 0. Els residus del model de 7 variables tenen una mitjana de -0,0005µg/m^3 i una desviació estàndard de 9,8836µg/m^3 mentre que el model de 9 variables té una mitjana de -0,0164µg/m^3 i una desviació de 8,5479µg/m^3. A la Figura 6, es compara la distribució dels residus amb una distribució normal centrada al 0 amb la mateixa variància que els residus i s'observa que la distribució de residus s'ajusta prou a la normal però té el pic una mica més estret i alt i les cues una mica més amples. Això passa pels dos models i es pot observar millor als gràfics Q-Q a la part de baix de la Figura 6. Els gràfics Q-Q comparen els quantils de dos distribucions. En aquest cas es compara, un altre cop, la distribució dels residus amb la normal. Si les distribucions són iguals, segueixen una línia de 45º. Si no ho són i el gràfic Q-Q té una pendent més gran, significa que els residus tenen més dispersió que la normal, mentre que si té una pendent més petita, la dispersió és menor. Es pot observar que, en aquest cas, les cues tenen més pendent i un petit fragment al centre té una pendent menor a 45º (si no ho fos la línia vermell només "es creueria" un cop amb els residus), corroborant la primera observació. Un mostreig petit, amb cues encara més petites, també pot influir a la forma de la distribució mostral.

En conclusió, la distribució dels residus no acaba de ser normal, el que significa que el model s'està deixant alguna part a explicar que no es deguda a un error aleatori de la mostra. No obstant això, com s'ha vist comparant els models, incrementar el nombre de variables de 7 a 9 no ha millorat aquest fet i afegir més variables no incrementa la capacitat predictiva del model. Es conclou que aquesta diferència forma part de la complexitat de la distribució d'ozó i que aquesta no pot ser explicada amb les variables del conjunt de dades o que els models GBM, amb els paràmetres escollits, no són capaços d'explicar-la.

Per acabar, s'estudia la dispersió dels residus segons cada variable del model. A la Figura 7 i a la Figura 8, es mostra aquesta dispersió pels models de 7 i 9 variables, respectivament, representant cada punt i els quartils Q1 i Q3 en finestres de 500 punts. En primera instància, s'observa que les diferències entre els dos models en les seves variables comuns és mínima. Seguidament, s'observa que els quartils es troben simètricament distribuïts al voltant del zero, com passa amb la distribució general dels residus, és a dir, no hi ha una dependència dels residus en cap variable i, per tant, el model ha extret molta informació de cada variable. Finalment, s'observa com la dispersió, o el rang interquartil, decreix per les variables del diòxid de nitrogen i de la mitjana de les últimes 8 hores de la velocitat del vent, és a dir, si es dóna el cas que alguna d'aquestes és gran, la predicció de la concentració de l'ozó serà millor.

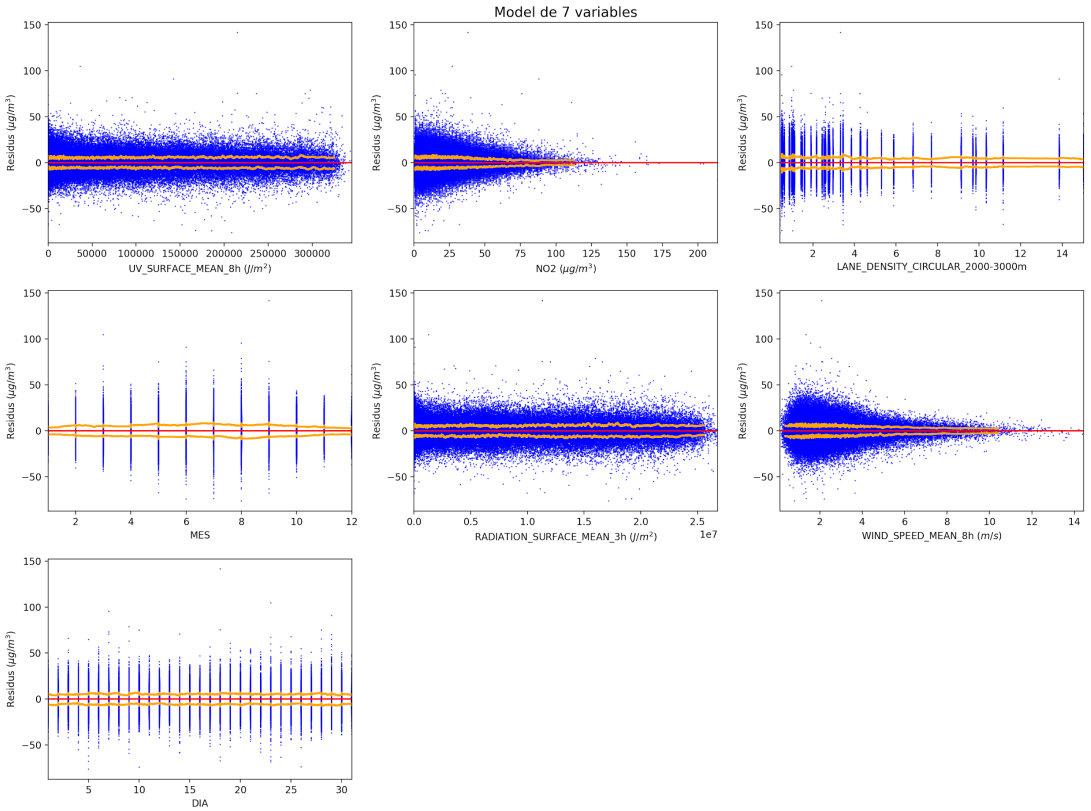

Figura 7. Residus en funció de cada variable del model de 7 variables (blau) i els quartils Q1 i Q3 en finestres de 500 punts (taronja).

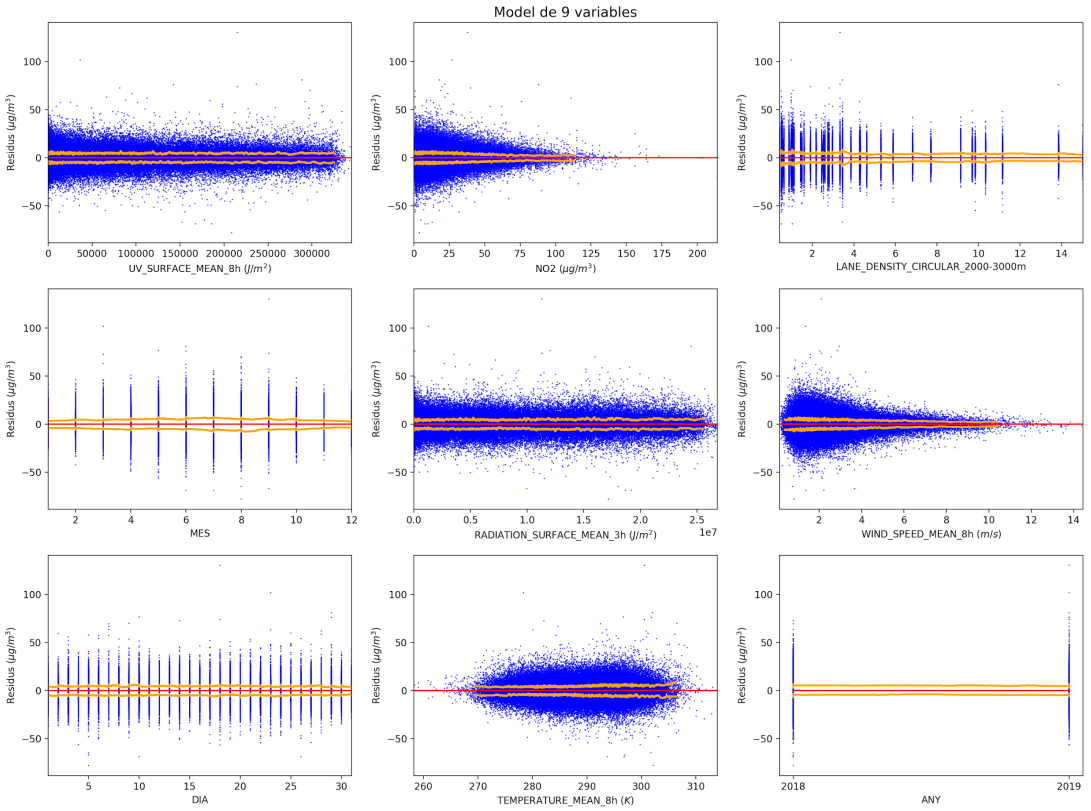

Figura 8. Residus en funció de cada variable del model de 9 variables (blau) i els quartils Q1 i Q3 en finestres de 500 punts (taronja).

3.4 Prediccions d'alta resolució

Un cop creat el model de predicció d'ozó és interessant estudiar la seva aplicació en escales encara més petites comparables amb el nostre entorn, com per exemple, discriminar els nivells d'ozó d'un carrer a un altre. El procés d'inferir informació d'alta resolució a partir de models creats amb informació amb una resolució més baixa s'anomena *downscaling*. No obstant això, és necessari tenir els valors locals d'algunes variables predictores del model, preferiblement les més importants.

Per a fer-ho, s'obtenen dades de receptors de NO_2 repartits pels diferents carrers de Barcelona i dades de models numèrics de *street canyon* dels vents, que tenen en compte les microgeometries dels carrers i les condicions ambientals del moment. Aquestes dades són d'ús restringit, han estat subministrades per Bettair Cities i són part del projecte MappingAir [31]. Es troben repartides en 4 arxius csv i descriuen la zona que va de 2,12 a 2,22 de longitud i de 41,35 a 41,43 de latitud, una zona que inclou tres estacions de contaminació de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica. Les dades tenen una freqüència temporal horària i descriuen els dies 7, 8, 9, 11, 12 i 13 de novembre de 2020.

A la Taula 11, es mostren els camps gdf_streets.csv, que descriuen les posicions dels carrers de Barcelona en el sistema de referència EPSG:3857. Les columnes de 'arxiu df_street_wind.csv es mostren a la Taula 12 i descriuen la direcció i la velocitat del vent en cada carrer descrit a l'arxiu anterior al llarg del temps. De manera similar, l'arxiu gdf_receptors.csv (Taula 13) descriu les posicions dels receptors de diòxid de nitrogen mentre que road_type.csv (Taula 14) descriu les emissions rebudes pels receptors provinents dels diferents tipus de carretera i al llarg del temps.

A partir d'aquestes dades, se segueix un procediment semblant al recorregut per a crear el *dataset* que s'ha utilitzar per definir els models però només amb les variables que els models necessiten. Per fer-ho, es descarreguen dades de temperatura, radiació en superfície i radiació UV en superfície d'ERA5 per a la zona i el temps que descriuen els receptors. A partir d'aquestes es calcula la mitja de 3 hores de la radiació en superfície i les mitjanes de 8 hores de radiació UV i de temperatura. Per a calcular la densitat de carrils en el *buffer* circular de 2000 a 3000 metres es realitza una consulta de tota la zona amb l'API d'Overpass i es calculen els *buffers* per cada receptor. El vent es calcula creant una imatge de la velocitat del vent dels carrers per cada hora en una resolució de 0,001º. Després, s'extreu la velocitat del vent per la posició de cada receptor i es calcula la mitjana de les últimes 8 hores. Finalment, el diòxid de nitrogen es calcula com a la suma de les contribucions parcials de les emissions provinents de cada carretera. El valor es normalitza per tenir un valor màxim de 250 $\mu g/m^3$, ja que els valors no es troben calibrats i no es disposa de la calibració i aquest valor és aproximadament el valor màxim que agafen les dades d'ozó utilitzades per la creació del model.

Taula 11. Camps d'interès de gdf_streets.csv que descriuen les posicions dels carrers.

Nom del camp	Descripció	Tipus
x_b	Metres cap a l'est (en coordenades EPSG:3857) del punt inicial del carrer	Numèric
x_e	Metres cap a l'est (en coordenades EPSG:3857) del punt final del carrer	Numèric
y_b	Metres cap al nord (en coordenades EPSG:3857) del punt inicial del carrer	Numèric
y_e	Metres cap al nord (en coordenades EPSG:3857) del punt final del carrer	Numèric

Taula 12. Camps de df_street_wind.csv que descriuen la velocitat del vent i la seva direcció a cada carrer.

Nom del camp	Descripció	Tipus
(index)	Clau forana que fa referència a la fila de gdf_streets.csv	Numèric
winddir	Direcció del vent en graus decimals	Numèric
windspeed	Velocitat del vent en metres per segon (m/s)	Numèric
timestamp	Marca temporal	Text

Taula 13. Camps d'interès de gdf_receptors.csv que descriuen la posició de cada receptor.

Nom del camp	Descripció	Tipus
x	Metres cap a l'est (en coordenades EPSG:3857) de la posició del receptor de NO_2	Numèric
y	Metres cap al nord (en coordenades EPSG:3857) de la posició del receptor de NO_2	Numèric

Taula 14. Camps de road_type.csv que descriuen les emissions rebudes en cada receptor en cada moment.

Nom del camp	Descripció	Tipus
(index)	Clau forana que fa referència a la fila de gdf_receptors.csv	Numèric
service, residential, tertiary, secondary, primary, trunk, motorway, unclassified	8 camps que descriuen l'emissió de NO_2 sense calibrar que arriba de cada tipus de carretera	Numèric
timestamp	Marca temporal	Text

Un cop fet això, ja es tenen les variables necessàries pel models de 7 i 9 variables. Es realitzen prediccions de l'ozó pels models de 7 i 9 variables i els resultats es representen en mapes de colors segons la longitud i la latitud, presentant arxius de vídeo animat per cada hora. Per a poder comparar els resultats amb els reals, s'afegeixen les mesures de l'ozó de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica per les tres estacions presents en la zona. A la Figura 9 i a la Figura 10, es poden veure les prediccions dels models de 7 i 9 variables, respectivament, per algunes hores del 8 de novembre de 2020, acompanyats dels valors reals de la concentració d'ozó. A excepció de la imatge de la primera hora del dia, les concentracions predites s'ajusten força bé a les concentracions mesurades a les estacions. Algunes zones amb carrers més petits, com Ciutat Vella o altres zones fora de l'Eixample, obtenen valors més constants, potser deguts a que per aquests carrers hi ha menys dades de vent. En general, s'observa com durant el dia les concentracions d'ozó incrementen per després decreixer a mida que es fa de nit, descrivint un cicle que acompanya al cicle de llum solar i al de l'activitat de les persones.

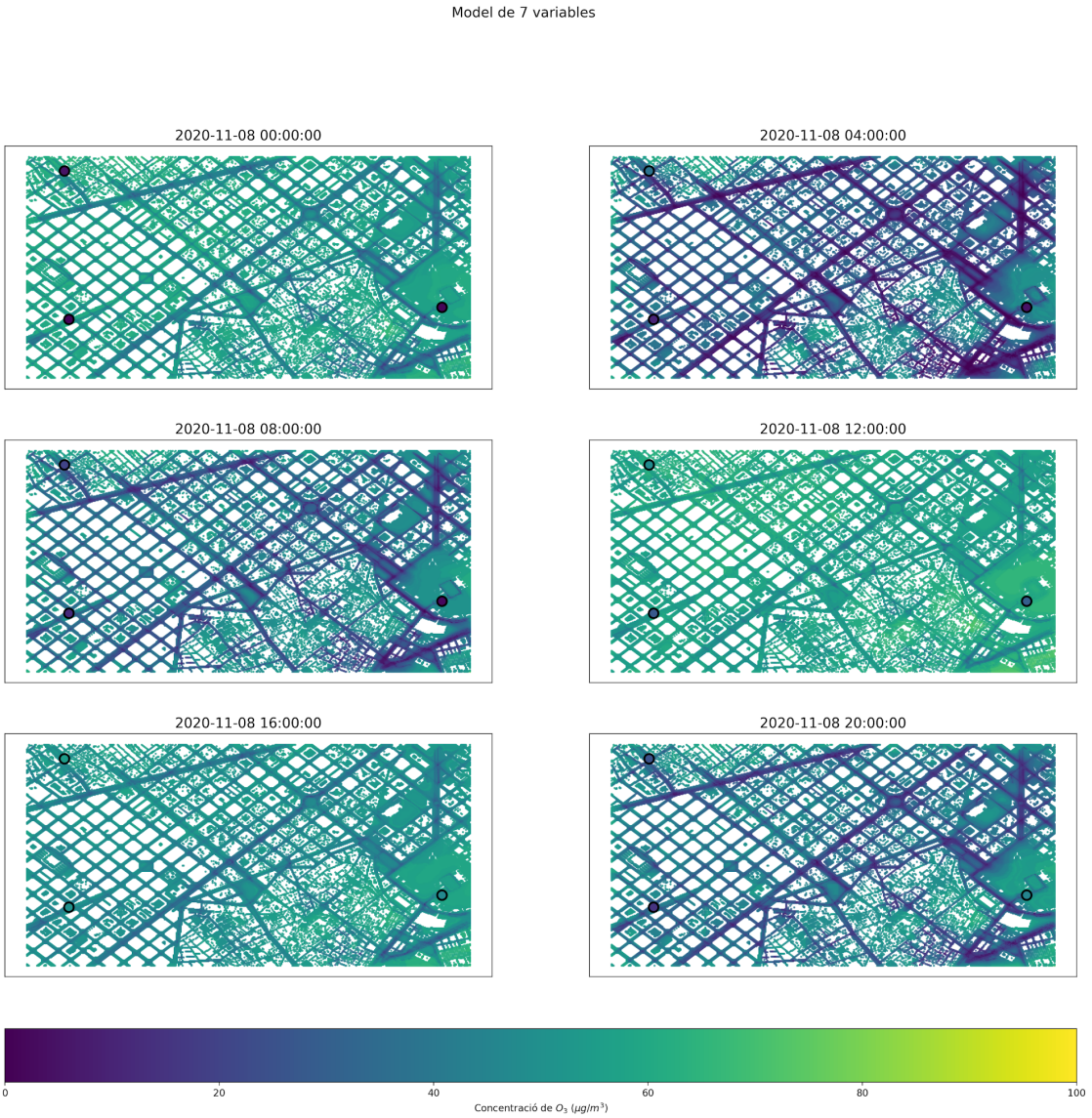

Figura 9. Predicció de la concentració d'ozó del model de 7 variables pels carrers de Barcelona per diferents hores del 8 de novembre del 2020. Els punts grans marcats amb un contorn negre fan referència al valor real de les estacions de mesura d'ozó. L'escala de color indica la concentració d'ozó en micrograms per metre al cub, on els colors més clars són concentracions més elevades.

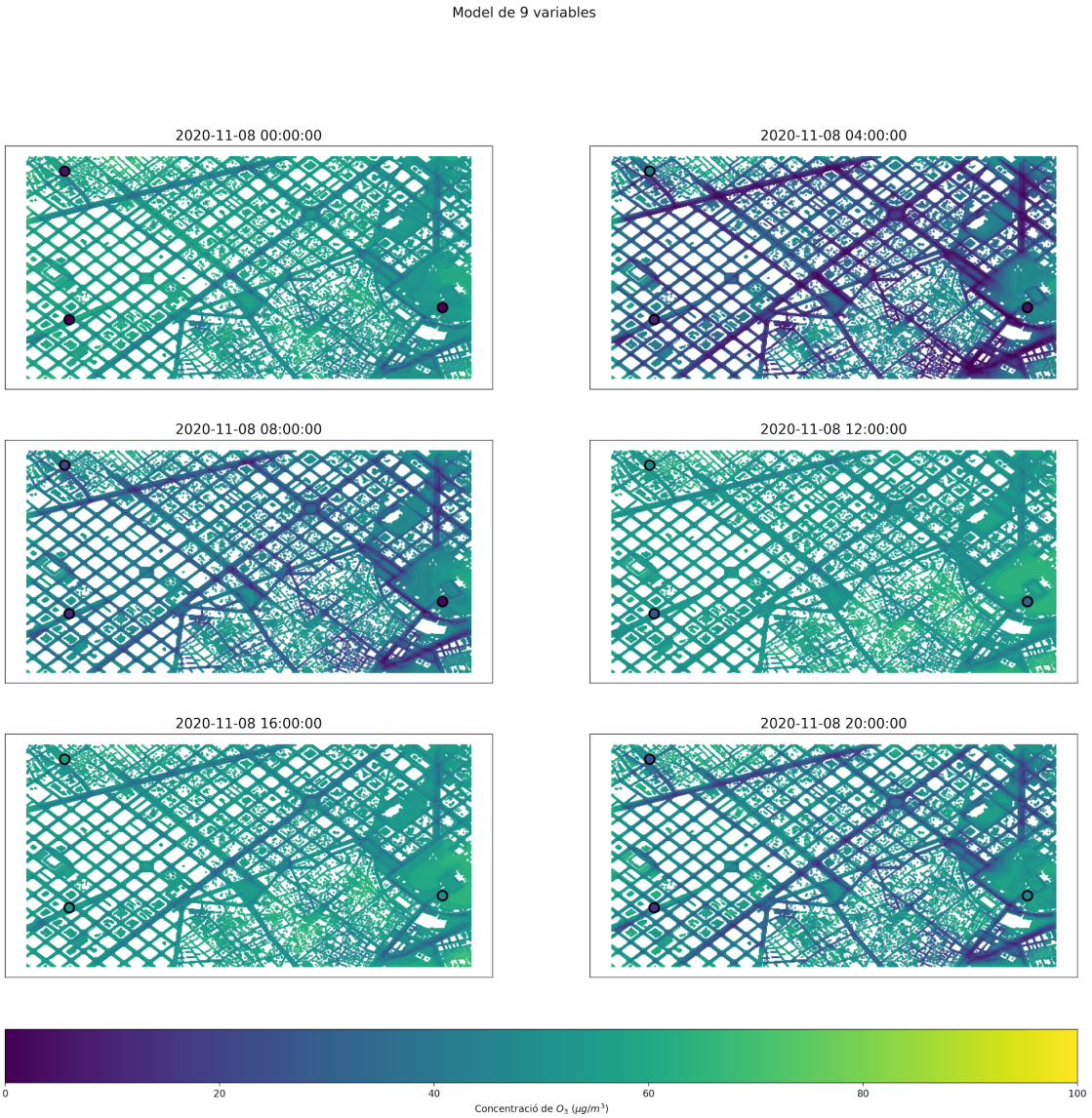

Figura 10. Predicció de la concentració d'ozó del model de 9 variables pels carrers de Barcelona per diferents hores del 8 de novembre del 2020. Els punts grans marcats amb un contorn negre fan referència al valor real de les estacions de mesura d'ozó. L'escala de color indica la concentració d'ozó en micrograms per metre al cub, on els colors més clars són concentracions més elevades.

4. Conclusions

En aquest treball s'han creat models de *gradient boosting machines* per a la predicció de la concentració d'ozó a Catalunya a partir d'un conjunt de dades creat a partir de la integració de dades de fonts d'estacions de contaminants, meteorològiques, demogràfiques i de ús de sòl i de carreteres. S'han ajustat les variables a seleccionar i el nombre d'aquestes a utilitzar fins a obtenir dos models de 7 i 9 variables que obtenen, respectivament, un valor de R^2 de 0,911 i 0,933. Aquests valors es comparen amb els valors de l'estat de l'art a la Taula 15. S'observa que, amb un enfocament senzill fent ús de GBM i sense emprar moltes més variables, s'obtenen resultats comparables als diferents models que proposen els autors pels diferents contaminants.

En general, l'anàlisi dels residus sembla mostrar que els models proposats estan explotant i extraient gran part de la informació de cada variable. S'ha observat la concentració de l'ozó comparteix força informació amb la concentració de diòxid de nitrogen i amb la radiació UV, sent els dos participants en les reaccions que creen ozó. Altres variables meteorològiques compostes per la temperatura, el vent i la radiació i variables relacionades amb el temps (any, mes i dia) les acompanyen per a construir el model. La única variable seleccionada pel model relacionada amb dades d'ús de sòl és el *buffer* circular de la densitat de carrils de 2000 a 3000 metres. Altres variables que han quedat fora per poc són els boscos en un *buffer* circular de 1000 a 2000 metres i el terreny artificial en un *buffer* circular de 3000 a 5000 metres, però la millora que oferien al models era mínima. La proposta d'utilitzar *buffers* semicirculars per part de la literatura [28] no ha mostrat millores en el cas d'aquest treball, ja que en el procés de selecció de variables no n'ha sortit cap per sobre del *buffer* circular corresponent. Igualment, s'ha observat que les variables d'ús de sòl no han aconseguit explicar gaire de la variació de l'ozó, el que pot significar que els models LUR poden obtenir resultats similars a models sense variables de *land use*. També s'hauria d'avaluar si hi ha alguna altra manera de calcular variables d'ús de sòl més significatives o si és necessari obtenir dades d'ús de sòl amb variació temporal.

Addicionalment, s'han utilitzat els models per a calcular les concentracions d'O_3 a una resolució més alta, procés que s'anomena *downscaling*, a partir de dades de NO_2 i vent a escala de carrer. En comparar l'ozó predit per alguns dies de novembre de 2020 amb les mesures de les estacions de mesura de contaminants, s'obtenen resultats força bons si es mira l'evolució al llarg del dia. Malauradament, en hores més concretes, com pot ser les 12 de la nit, les diferències amb el valor real de la concentració poden ser més grans. Els factors que poden influenciar en aquestes diferències són molts sense comptar l'error que pot incorporar el *downscaling* o tractar de predir punts diferents de l'estació (*transfer learning*). Un dels factors a destacar és la diferència de mobilitat de la gent entre 2018 o 2019, que s'han utilitzat per crear el model, i 2020 pot ser molt gran donada les restriccions de mobilitat i els confinaments per a combatre la pandèmia mundial del COVID-19. S'ha mostrat que aquesta situació afecta a la concentració de contaminants com el NO_2 [32]. Alguns altres factors poden ser la falta de calibració en les dades de NO_2 d'alta resolució o, com esmentat abans, la manca de dades de vent sobre alguns carrers més petits.

Taula 15. Comparació de l'estat de l'art amb els resultats dels models obtinguts.

Nom	Predicció	Model	Nombre variables	R^2
Ross et al. (2005) [14]	NO_2	MLR	4	0,79
Beelen et al. (2009) [15]	O_3	kriging	3	0,70
Adam-Poupart et al. (2014) [16]	O_3	BME	6	0,653
Zhan et al. (2018) [17]	Màxim diari O_3 (mitjana 8h)	RF	13	0,69
Chen et al. (2021) [18]	màxim anual O_3 (mitjana 8h)	LUR	5	0,59
	Màxim diari O_3 (mitjana 8h)	LUR + BME		0,80
Wang, J. et al. (2021) [19]	O_3	SVR + MLR/ST		0,90
Wong et al. (2021) [20]	$PM_{2.5}$	GBM	6	0,73
		kriging + GBM		0,94
Wang, Z. et al. (2021) [21]	$PM_{2.5}$	LSTM	8	0,83
Aquest treball	**O3**	**GBM**	**7**	**0,911**
			9	**0,933**

En general, els objectius plantejats en aquest treball s'han complert ja que s'han obtingut les dades dels diferents tipus proposat i s'ha après a utilitzar les eines per llegir els diferents formats de dades i per a crear models GBM. L'objectiu que no s'ha complert és comparar els models GBM obtinguts amb algun altre model, com podria ser kriging o LSTM. La complexitat d'aplicar aquests dos algorismes ha necessitat d'un temps que no del que no s'ha disposat i que no s'ha previst. Addicionalment, la creació del *dataset* ha ocupat més temps del previst donat, principalment, per la creació de *buffers* i per entendre millor les diferents transformacions de coordenades que requerien els diferents formats dades. Vist això, s'ha dut a terme un ajust i un anàlisi més extens dels models GBM. La resta de la planificació ha estat adequada pel transcurs del treball.

Finalment, les línies futures que no es cobreixen en aquest treball inclouen, a falta de la comparació amb altres models, la construcció d'altres models amb les mateixes dades. Això és necessari per poder tenir una comparació directa, que parteix de les mateixes condicions, i permet una avaluació millor que comparar el model amb l'estat de l'art. Addicionalment, els models entrenats es poden avaluar en llocs o ciutats diferents per tal d'avaluar la capacitat de *transfer learning* del model i descobrir quins són els límits d'aplicació. Per acabar, una línia interessant, si es disposa de més temps i recursos, seria explorar més l'alta resolució i/o calcular prediccions a temps real. Algun d'aquests models ja existeixen pel diòxid de nitrogen [33] i permeten que la

població pugui ajustar-se si els nivells de contaminació són perillosos. Utilitzar aproximacions més fines de carreteres o disposar de dades meteorològiques (com la radiació UV) amb més resolució, podria portar les prediccions de *downscaling* calculades en aquest treball a un nivell de precisió superior necessari per a ser un dels pocs mapes d'alta resolució d'ozó a temps real.

5. Glossari

BME: *Bayesian maximum entropy*
Dataset: Conjunt de dades
Downscaling: Procés d'inferir informació d'alta resolució a partir de models creats amb informació amb una resolució més baixa
GBM: *Gradient boosting machines*
GeoTIFF: *Georeferencing Tagged Image File Format*
GIS: Sistema d'informació geogràfica (*Geographic information system*)
LSTM: *Long short-term memory*
LUR: Regressió d'ús de sòl (*land use regression*)
MLR: Regressió lineal múltiple (*multiple linear regression)*
NO_2: Diòxid de nitrogen
O_3: Ozó
R: Coeficient de correlació de Pearson
RF: *Random forest*
ST: Model espaciotemporal (*spatiotemporal model)*
SVM: Màquines de suport vectorial (*support vector machines*)
UV: Radiació ultraviolada

6. Bibliografia

[1] Hwang, M., Kim, J., & Cheong, H. (2020). Short-Term Impacts of Ambient Air Pollution on Health-Related Quality of Life: A Korea Health Panel Survey Study. *International Journal Of Environmental Research And Public Health*, *17*(23), 9128. https://doi.org/10.3390/ijerph17239128

[2] What Constitutes an Adverse Health Effect of Air Pollution?. (2000), *161*(2), 665-673. https://doi.org/10.1164/ajrccm.161.2.ats4-00

[3] OCDE (2016). *Economic Consequences of Outdoor Air Pollution*. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264257474-en

[4] Fowler, D., Brimblecombe, P., Burrows, J., Heal, M., Grennfelt, P., & Stevenson, D. et al. (2020). A chronology of global air quality. *Philosophical Transactions Of The Royal Society A: Mathematical, Physical And Engineering Sciences*, *378*(2183), 20190314. https://doi.org/10.1098/rsta.2019.0314

[5] Ballester, F. (2005). Contaminación atmosférica, cambio climático y salud. *Revista Española de Salud Pública*, *79*(2), 159-175. Consultat des de http://ref.scielo.org/44rbfr

[6] Boudri, J., Hordijk, L., Kroeze, C., Amann, M., Cofala, J., & Bertok, I. et al. (2002). The potential contribution of renewable energy in air pollution abatement in China and India. *Energy Policy*, *30*(5), 409-424. https://doi.org/10.1016/s0301-4215(01)00107-0

[7] Duque, L., Relvas, H., Silveira, C., Ferreira, J., Monteiro, A., & Gama, C. et al. (2016). Evaluating strategies to reduce urban air pollution. *Atmospheric Environment*, *127*, 196-204. https://doi.org/10.1016/j.atmosenv.2015.12.043

[8] Poore, J., & Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. *Science*, *360*(6392), 987-992. https://doi.org/10.1126/science.aaq0216

[9] European Environment Agency. (2020). Air quality in Europe. *2020 report*. Consultat des de http://doi.org/10.2800/602793

[10] European Communities. (2000). *Ambient air pollution by AS, CD and NI compounds. Position Paper*. Consultat des de https://ec.europa.eu/environment/archives/air/pdf/pp_as_cd_ni.pdf

[11] Yi Wu, C., Cabrera-Rivera, O., Dettling, J., Asselmeier, D., McGeen, D., & Ostrander, A. et al. (2007). An Assessment of Benzo(a)pyrene Air Emissions in the Great Lakes Region, 12. Consultat des de https://www3.epa.gov/ttnchie1/conference/ei16/session6/wu.pdf

[12] Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air Pollution Forecasts: An Overview. *International Journal Of Environmental Research And Public Health*, *15*(4), 780. https://doi.org/10.3390/ijerph15040780

[13] Ren, X., Mi, Z., & Georgopoulos, P. (2020). Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environment International*, *142*, 105827. https://doi.org/10.1016/j.envint.2020.105827

[14] Ross, Z., English, P., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., & Jerrett, M. (2005). Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal Of Exposure Science & Environmental Epidemiology*, *16*(2), 106-114. https://doi.org/10.1038/sj.jea.7500442

[15] Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., & Briggs, D. (2009). Mapping of background air pollution at a fine spatial scale across the European Union. *Science Of The Total Environment*, *407*(6), 1852-1867. https://doi.org/10.1016/j.scitotenv.2008.11.048

[16] Adam-Poupart, A., Brand, A., Fournier, M., Jerrett, M., & Smargiassi, A. (2014). Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging, Land-Use Regression (LUR), and Combined Bayesian Maximum Entropy–LUR Approaches. *Environmental Health Perspectives*, *122*(9), 970-976. https://doi.org/10.1289/ehp.1306566

[17] Zhan, Y., Luo, Y., Deng, X., Grieneisen, M., Zhang, M., & Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environmental Pollution*, *233*, 464-473. https://doi.org/10.1016/j.envpol.2017.10.029

[18] Chen, L., Liang, S., Li, X., Mao, J., Gao, S., & Zhang, H. et al. (2021). A hybrid approach to estimating long-term and short-term exposure levels of ozone at the national scale in China using land use regression and Bayesian maximum entropy. *Science Of The Total Environment*, *752*, 141780. https://doi.org/10.1016/j.scitotenv.2020.141780

[19] Wang, J., & Xu, H. (2021). A novel hybrid spatiotemporal land use regression model system at the megacity scale. *Atmospheric Environment*, *244*, 117971. https://doi.org/10.1016/j.atmosenv.2020.117971

[20] Wong, P., Lee, H., Chen, Y., Zeng, Y., Chern, Y., & Chen, N. et al. (2021). Using a land use regression model with machine learning to estimate ground level PM2.5. *Environmental Pollution*, *277*, 116846. https://doi.org/10.1016/j.envpol.2021.116846

[21] Wang, Z., Zhou, Y., Zhao, R., Wang, N., Biswas, A., & Shi, Z. (2021). High-resolution prediction of the spatial distribution of PM2.5 concentrations in China

using a long short-term memory model. *Journal Of Cleaner Production*, *297*, 126493. https://doi.org/10.1016/j.jclepro.2021.126493

[22] *Qualitat de l'aire als punts de mesurament automàtics de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica*. (2020). Consultat 23 març 2021, des de https://analisi.transparenciacatalunya.cat/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu

[23] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2018). ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Consultat 23 març 2021, des de https://doi.org/10.24381/cds.adbb2d47

[24] Muñoz Sabater, J., (2019). ERA5-Land hourly data from 1981 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. Consultat 1 abril 2021, des de https://doi.org/10.24381/cds.e2161bac

[25] WorldPop. (2018). Global High Resolution Population Denominators Project. Consultat 23 març 2021, des de https://dx.doi.org/10.5258/SOTON/WP00645

[26] Corine Land Cover (CLC) 2018, Version 2020_20u1. © European Union, Copernicus Land Monitoring Service 2021, European Environment Agency (EEA). (2018). Consultat 23 març 2021, des de https://land.copernicus.eu/pan-european/corine-land-cover.

[27] *OpenStreetMap*. © OpenStreetMap contributors. (2021). Consultat 1 abril 2021, des de https://www.openstreetmap.org/

[28] Li, X., Liu, W., Chen, Z., Zeng, G., Hu, C., & León, T. et al. (2015). The application of semicircular-buffer-based land use regression models incorporating wind direction in predicting quarterly NO 2 and PM 10 concentrations. *Atmospheric Environment*, *103*, 18-24. https://doi.org/10.1016/j.atmosenv.2014.12.004

[29] Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). *Minería de datos*. Editorial UOC, Barcelona. https://www.editorialuoc.com/mineria-de-datos

[30] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed., p. 745). Springer. https://doi.org/10.1007/978-0-387-84858-7

[31] Mapping Air. Bettair. (2019). Consultat 5 juny 2021, des de https://mappingair.bettair.city/

[32] *COVID-19: Com afecta el confinament a la qualitat de l'aire de les grans ciutats*. Lobelia. (2020). Consultat 1 juny 2021, des de https://www.lobelia.earth/ca/covid-19

[33] *Calidad del aire en Barcelona*. Lobelia. Consultat 1 juny 2021, des de https://aire-barcelona.lobelia.earth/es/

7. Annexos

Annex 1. Millors variables per estació amb els coeficients de Pearson al quadrat corresponents

Vars	8015021	8019043	8019044
1	NO2 (0,615)	NO2 (0,322)	NO2 (0,395)
2	UV_SURFACE_MEAN_8h (0,776)	UV_SURFACE_MEAN_8h (0,646)	UV_SURFACE_MEAN_8h (0,697)
3	MES (0,833)	MES (0,749)	MES (0,779)
4	UV_SURFACE_MEAN_3h (0,865)	DIA (0,795)	DIA (0,822)
5	DIA (0,886)	HORA (0,822)	TEMPERATURE_MEAN_24h (0,847)
6	TEMPERATURE_MEAN_24h (0,898)	TEMPERATURE_MEAN_24h (0,845)	HORA (0,863)
7	WIND_SPEED_MEAN_24h (0,906)	WIND_SPEED_MEAN_24h (0,856)	UV_SURFACE_MEAN_24h (0,873)
8	ANY (0,911)	NO2_MEAN_24h (0,866)	WIND_SPEED_MEAN_24h (0,879)
9	HORA (0,915)	RADIATION_SURFACE_MEAN_24h (0,874)	NO2_MEAN_24h (0,885)
10	RADIATION_SURFACE_MEAN_24h (0,917)	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_3000-5000m (0,877)	ANY (0,889)
11	TEMPERATURE_MEAN_8h (0,918)	ANY (0,880)	TEMPERATURE_MEAN_3h (0,891)
12	NO2_MEAN_24h (0,920)	TEMPERATURE (0,881)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_2000-3000m (0,892)
13	WIND_DIRECTION (0,921)	FOREST_SEMICIRCULAR_UPWIND_3000-5000m (0,882)	UV_SURFACE_MEAN_3h (0,893)
14	POPULATION_CIRCULAR_0-300m (0,921)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_0-500m (0,882)	POPULATION_CIRCULAR_0-300m (0,893)
15	POPULATION_CIRCULAR_300-500m (0,921)	POPULATION_CIRCULAR_0-300m (0,882)	POPULATION_CIRCULAR_300-500m (0,893)

Vars	8019050	8019054	8019057
1	NO2 (0,539)	NO2 (0,478)	NO2 (0,423)
2	UV_SURFACE_MEAN_8h (0,744)	RADIATION_SURFACE (0,723)	RADIATION_SURFACE (0,685)
3	MES (0,799)	MES (0,806)	UV_SURFACE_MEAN_24h (0,780)
4	DIA (0,832)	TEMPERATURE_MEAN_3h (0,842)	WIND_SPEED_MEAN_24h (0,816)
5	HORA (0,855)	DIA (0,864)	UV_SURFACE_MEAN_3h (0,834)
6	WIND_SPEED_MEAN_24h (0,872)	RADIATION_SURFACE_MEAN_24h (0,878)	DIA (0,853)
7	TEMPERATURE_MEAN_24h (0,885)	TEMPERATURE_MEAN_24h (0,884)	MES (0,869)
8	RADIATION_SURFACE_MEAN_24h (0,890)	WIND_DIRECTION (0,888)	NO2_MEAN_24h (0,877)
9	NO2_MEAN_24h (0,892)	ANY (0,892)	TEMPERATURE_MEAN_24h (0,883)
10	WIND_SPEED_MEAN_8h (0,895)	WIND_SPEED_MEAN_3h (0,897)	RADIATION_SURFACE_MEAN_24h (0,886)
11	TEMPERATURE_MEAN_8h (0,897)	NO2_MEAN_24h (0,900)	HORA (0,889)
12	NO2_DIF_1h (0,898)	UV_SURFACE_MEAN_3h (0,902)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_500-1000m (0,891)
13	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,899)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,904)	ANY (0,893)
14	ANY (0,899)	WIND_SPEED_MEAN_24h (0,905)	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_2000-3000m (0,894)
15	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_2000-3000m (0,900)	HORA (0,907)	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,895)

Vars	8019058	8022006	8089005
1	UV_SURFACE_MEAN_24h (0,426)	UV_SURFACE_MEAN_24h (0,500)	UV_SURFACE_MEAN_8h (0,537)
2	NO2 (0,559)	HORA (0,657)	NO2 (0,641)
3	TEMPERATURE_MEAN_24h (0,666)	TEMPERATURE_MEAN_24h (0,725)	UV_SURFACE_MEAN_24h (0,703)
4	HORA (0,722)	MES (0,770)	WIND_SPEED_MEAN_8h (0,763)
5	DIA (0,764)	NO2 (0,809)	TEMPERATURE_MEAN_24h (0,792)
6	MES (0,799)	DIA (0,832)	DIA (0,805)
7	ANY (0,813)	ANY (0,847)	MES (0,816)
8	RADIATION_SURFACE_MEAN_24h (0,820)	WIND_SPEED_MEAN_3h (0,855)	WATER_SEMICIRCULAR_UPWIND_3000-5000m (0,824)
9	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,824)	FOREST_SEMICIRCULAR_DOWNWIND_3000-5000m (0,859)	ANY (0,827)
10	TEMPERATURE_MEAN_3h (0,824)	POPULATION_CIRCULAR_0-300m (0,859)	ARTIFICIAL_SEMICIRCULAR_UPWIND_3000-5000m (0,830)
11	ROAD_DENSITY_SEMICIRCULAR_UPWIND_1000-2000m (0,827)	POPULATION_CIRCULAR_300-500m (0,859)	NO2_DIF_2h (0,833)
12	POPULATION_CIRCULAR_0-300m (0,827)	POPULATION_CIRCULAR_500-1000m (0,859)	HORA (0,835)
13	POPULATION_CIRCULAR_300-500m (0,827)	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,859)	POPULATION_CIRCULAR_0-300m (0,835)
14	POPULATION_CIRCULAR_500-1000m (0,827)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_3000-5000m (0,861)	POPULATION_CIRCULAR_300-500m (0,835)
15	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,825)	RADIATION_SURFACE_MEAN_24h (0,862)	POPULATION_CIRCULAR_500-1000m (0,835)

Vars	8096014	8102005	8112003
1	NO2 (0,563)	UV_SURFACE_MEAN_8h (0,551)	UV_SURFACE_MEAN_8h (0,621)
2	UV_SURFACE_MEAN_8h (0,820)	NO2 (0,716)	NO2_MEAN_3h (0,744)
3	RADIATION_SURFACE_MEAN_8h (0,867)	RADIATION_SURFACE (0,779)	HORA (0,825)
4	TEMPERATURE_MEAN_24h (0,895)	MES (0,819)	MES (0,863)
5	MES (0,912)	WIND_SPEED_MEAN_8h (0,856)	DIA (0,880)
6	WIND_SPEED_MEAN_8h (0,922)	DIA (0,879)	TEMPERATURE_MEAN_24h (0,892)
7	DIA (0,928)	RADIATION_SURFACE_MEAN_24h (0,889)	NO2 (0,904)
8	NO2_MEAN_24h (0,931)	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,895)	WIND_SPEED_MEAN_24h (0,910)
9	HORA (0,934)	NO2_MEAN_24h (0,897)	ANY (0,915)
10	ANY (0,936)	TEMPERATURE_MEAN_24h (0,903)	UV_SURFACE_MEAN_24h (0,919)
11	WIND_SPEED_MEAN_24h (0,937)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_3000-5000m (0,905)	POPULATION_CIRCULAR_0-300m (0,919)
12	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_2000-3000m (0,939)	UV_SURFACE_MEAN_24h (0,905)	POPULATION_CIRCULAR_300-500m (0,919)
13	POPULATION_CIRCULAR_0-300m (0,939)	RADIATION_SURFACE_MEAN_8h (0,908)	POPULATION_CIRCULAR_500-1000m (0,919)
14	POPULATION_CIRCULAR_300-500m (0,939)	NO2_MEAN_8h (0,909)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,918)
15	POPULATION_CIRCULAR_500-1000m (0,939)	LANE_DENSITY_SEMICIRCULAR_UPWIND_3000-5000m (0,911)	RADIATION_SURFACE_MEAN_24h (0,919)

Vars	8113007	8121013	8125002
1	UV_SURFACE_MEAN_8h (0,565)	NO2 (0,457)	UV_SURFACE_MEAN_8h (0,605)
2	HORA (0,690)	UV_SURFACE_MEAN_24h (0,698)	NO2 (0,796)
3	NO2 (0,773)	HORA (0,807)	RADIATION_SURFACE_MEAN_3h (0,844)
4	MES (0,819)	TEMPERATURE_MEAN_24h (0,844)	MES (0,874)
5	WIND_SPEED_MEAN_24h (0,849)	DIA (0,860)	WIND_SPEED_MEAN_24h (0,891)
6	DIA (0,873)	MES (0,876)	TEMPERATURE_MEAN_24h (0,904)
7	TEMPERATURE_MEAN_24h (0,883)	ANY (0,886)	DIA (0,910)
8	UV_SURFACE_MEAN_24h (0,888)	WIND_SPEED_MEAN_8h (0,889)	UV_SURFACE_MEAN_24h (0,915)
9	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,892)	RADIATION_SURFACE_MEAN_24h (0,891)	WIND_SPEED (0,918)
10	WIND_SPEED_MEAN_8h (0,893)	RADIATION_SURFACE (0,894)	ANY (0,920)
11	RADIATION_SURFACE_MEAN_3h (0,895)	WIND_SPEED_MEAN_24h (0,894)	HORA (0,922)
12	RADIATION_SURFACE_MEAN_24h (0,899)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_3000-5000m (0,895)	NO2_DIF_1h (0,925)
13	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,899)	RADIATION_SURFACE_MEAN_8h (0,895)	LANE_DENSITY_SEMICIRCULAR_UPWIND_2000-3000m (0,926)
14	RADIATION_SURFACE (0,899)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_2000-3000m (0,896)	NO2_MEAN_24h (0,927)
15	UV_SURFACE (0,900)	FOREST_SEMICIRCULAR_UPWIND_3000-5000m (0,896)	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_3000-5000m (0,927)

Vars	8137001	8169009	8184006
1	UV_SURFACE_MEAN_24h (0,424)	NO2 (0,644)	UV_SURFACE_MEAN_8h (0,538)
2	HORA (0,518)	RADIATION_SURFACE (0,810)	NO2 (0,754)
3	TEMPERATURE_MEAN_24h (0,606)	MES (0,858)	RADIATION_SURFACE_MEAN_8h (0,829)
4	DIA (0,667)	DIA (0,883)	MES (0,867)
5	WIND_SPEED_MEAN_8h (0,700)	HORA (0,899)	DIA (0,886)
6	MES (0,722)	TEMPERATURE_MEAN_24h (0,911)	WIND_SPEED_MEAN_24h (0,901)
7	NO2 (0,748)	WIND_SPEED_MEAN_8h (0,921)	TEMPERATURE_MEAN_24h (0,908)
8	ANY (0,756)	WATER_SEMICIRCULAR_UPWIND_2000-3000m (0,923)	WIND_DIRECTION (0,912)
9	RADIATION_SURFACE_MEAN_24h (0,761)	NO2_DIF_2h (0,925)	NO2_MEAN_24h (0,916)
10	WIND_SPEED_MEAN_24h (0,767)	ANY (0,926)	UV_SURFACE_MEAN_24h (0,918)
11	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,770)	NO2_MEAN_24h (0,928)	ANY (0,919)
12	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,770)	UV_SURFACE_MEAN_24h (0,931)	WIND_SPEED_MEAN_8h (0,921)
13	FOREST_SEMICIRCULAR_DOWNWIND_500-1000m (0,770)	WIND_SPEED_MEAN_24h (0,932)	HORA (0,922)
14	AGRICULTURAL_SEMICIRCULAR_UPWIND_500-1000m (0,770)	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,932)	TEMPERATURE_MEAN_8h (0,923)
15	FOREST_SEMICIRCULAR_UPWIND_500-1000m (0,770)	WATER_SEMICIRCULAR_UPWIND_3000-5000m (0,932)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_500-1000m (0,924)

Vars	8187012	8194008	8202001
1	UV_SURFACE_MEAN_8h (0,443)	NO2 (0,691)	UV_SURFACE_MEAN_8h (0,653)
2	NO2 (0,658)	UV_SURFACE_MEAN_8h (0,835)	NO2 (0,798)
3	TEMPERATURE_MEAN_24h (0,738)	UV_SURFACE_MEAN_3h (0,874)	HORA (0,836)
4	RADIATION_SURFACE_MEAN_3h (0,791)	TEMPERATURE_MEAN_24h (0,900)	WIND_SPEED_MEAN_24h (0,873)
5	UV_SURFACE_MEAN_24h (0,816)	MES (0,913)	TEMPERATURE_MEAN_24h (0,892)
6	WIND_SPEED_MEAN_8h (0,837)	WIND_SPEED_MEAN_24h (0,923)	MES (0,905)
7	DIA (0,847)	DIA (0,930)	DIA (0,912)
8	MES (0,857)	ANY (0,931)	NO2_MEAN_8h (0,917)
9	ANY (0,864)	HORA (0,933)	ANY (0,920)
10	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_2000-3000m (0,867)	RADIATION_SURFACE_MEAN_24h (0,936)	RADIATION_SURFACE_MEAN_24h (0,922)
11	RADIATION_SURFACE_MEAN_24h (0,868)	NO2_DIF_2h (0,938)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,924)
12	WIND_SPEED_MEAN_24h (0,869)	POPULATION_CIRCULAR_0-300m (0,938)	WIND_SPEED_MEAN_8h (0,926)
13	RADIATION_SURFACE (0,872)	POPULATION_CIRCULAR_300-500m (0,938)	NO2_MEAN_24h (0,927)
14	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_2000-3000m (0,873)	POPULATION_CIRCULAR_500-1000m (0,938)	NO2_DIF_1h (0,928)
15	NO2_MEAN_24h (0,874)	TEMPERATURE_MEAN_3h (0,938)	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,929)

Vars	8205002	8263001	8279011
1	UV_SURFACE_MEAN_8h (0,620)	UV_SURFACE_MEAN_8h (0,560)	UV_SURFACE_MEAN_8h (0,421)
2	NO2 (0,752)	NO2 (0,791)	NO2 (0,662)
3	TEMPERATURE_MEAN_24h (0,805)	TEMPERATURE_MEAN_24h (0,846)	MES (0,749)
4	RADIATION_SURFACE_MEAN_3h (0,844)	RADIATION_SURFACE (0,875)	TEMPERATURE_MEAN_24h (0,791)
5	WIND_SPEED_MEAN_8h (0,867)	WIND_SPEED_MEAN_24h (0,892)	WIND_SPEED_MEAN_8h (0,817)
6	MES (0,883)	UV_SURFACE_MEAN_24h (0,903)	RADIATION_SURFACE_MEAN_8h (0,837)
7	UV_SURFACE_MEAN_24h (0,891)	NO2_MEAN_24h (0,910)	DIA (0,852)
8	WIND_DIRECTION (0,896)	ANY (0,917)	ANY (0,863)
9	ANY (0,900)	DIA (0,921)	UV_SURFACE_MEAN_24h (0,869)
10	DIA (0,901)	MES (0,924)	RADIATION_SURFACE_MEAN_24h (0,871)
11	NO2_MEAN_24h (0,905)	WIND_SPEED_MEAN_8h (0,926)	WIND_SPEED_MEAN_24h (0,874)
12	TEMPERATURE_MEAN_8h (0,906)	HORA (0,928)	FOREST_SEMICIRCULAR_DOWNWIND_1000-2000m (0,876)
13	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_500-1000m (0,906)	RADIATION_SURFACE_MEAN_24h (0,929)	ARTIFICIAL_SEMICIRCULAR_UPWIND_3000-5000m (0,876)
14	WIND_SPEED_MEAN_24h (0,907)	LANE_DENSITY_SEMICIRCULAR_UPWIND_2000-3000m (0,930)	ARTIFICIAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,876)
15	HORA (0,909)	POPULATION_CIRCULAR_0-300m (0,930)	FOREST_SEMICIRCULAR_UPWIND_1000-2000m (0,876)

Vars	8283004	8301004	8305006
1	UV_SURFACE_MEAN_8h (0,538)	UV_SURFACE_MEAN_8h (0,431)	UV_SURFACE_MEAN_8h (0,564)
2	NO2 (0,669)	NO2 (0,662)	NO2 (0,787)
3	MES (0,750)	MES (0,743)	MES (0,830)
4	HORA (0,800)	DIA (0,787)	DIA (0,854)
5	TEMPERATURE_MEAN_24h (0,833)	WIND_SPEED_MEAN_8h (0,815)	HORA (0,874)
6	WIND_SPEED_MEAN_24h (0,858)	ANY (0,833)	WIND_SPEED_MEAN_8h (0,887)
7	DIA (0,868)	TEMPERATURE_MEAN_8h (0,847)	TEMPERATURE_MEAN_24h (0,897)
8	UV_SURFACE_MEAN_24h (0,873)	NO2_MEAN_24h (0,854)	RADIATION_SURFACE_MEAN_24h (0,902)
9	ANY (0,877)	TEMPERATURE_MEAN_24h (0,859)	ANY (0,906)
10	WIND_SPEED_MEAN_8h (0,881)	UV_SURFACE_MEAN_24h (0,865)	LANE_DENSITY_SEMICIRCULAR_UPWIND_2000-3000m (0,907)
11	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_1000-2000m (0,885)	UV_SURFACE (0,866)	NO2_MEAN_24h (0,908)
12	FOREST_SEMICIRCULAR_UPWIND_0-500m (0,885)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_500-1000m (0,869)	UV_SURFACE_MEAN_24h (0,910)
13	FOREST_SEMICIRCULAR_DOWNWIND_2000-3000m (0,886)	WIND_SPEED_MEAN_24h (0,872)	WIND_DIRECTION (0,911)
14	POPULATION_CIRCULAR_0-300m (0,886)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_2000-3000m (0,873)	WIND_SPEED (0,911)
15	POPULATION_CIRCULAR_300-500m (0,886)	POPULATION_CIRCULAR_0-300m (0,873)	RADIATION_SURFACE_MEAN_3h (0,911)

Vars	8307012	17013001	17184001
1	UV_SURFACE_MEAN_8h (0,475)	UV_SURFACE_MEAN_24h (0,426)	UV_SURFACE_MEAN_8h (0,504)
2	NO2 (0,704)	TEMPERATURE_MEAN_24h (0,554)	WIND_SPEED_MEAN_8h (0,610)
3	MES (0,778)	DIA (0,643)	MES (0,679)
4	DIA (0,817)	RADIATION_SURFACE (0,703)	RADIATION_SURFACE_MEAN_3h (0,724)
5	RADIATION_SURFACE (0,846)	ANY (0,736)	UV_SURFACE_MEAN_24h (0,752)
6	ANY (0,867)	NO2 (0,778)	DIA (0,776)
7	TEMPERATURE_MEAN_24h (0,876)	MES (0,804)	NO2_MEAN_24h (0,788)
8	WIND_SPEED_MEAN_8h (0,885)	HORA (0,820)	TEMPERATURE_MEAN_24h (0,794)
9	NO2_MEAN_24h (0,890)	FOREST_SEMICIRCULAR_UPWIND_2000-3000m (0,825)	ARTIFICIAL_SEMICIRCULAR_UPWIND_3000-5000m (0,802)
10	AGRICULTURAL_SEMICIRCULAR_UPWIND_3000-5000m (0,893)	WIND_SPEED_MEAN_8h (0,831)	HORA (0,803)
11	UV_SURFACE_MEAN_24h (0,896)	ARTIFICIAL_SEMICIRCULAR_UPWIND_500-1000m (0,835)	WIND_SPEED_MEAN_24h (0,809)
12	HORA (0,898)	RADIATION_SURFACE_MEAN_24h (0,836)	FOREST_SEMICIRCULAR_DOWNWIND_1000-2000m (0,812)
13	WIND_SPEED_MEAN_24h (0,899)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_2000-3000m (0,840)	WIND_DIRECTION (0,811)
14	AGRICULTURAL_SEMICIRCULAR_UPWIND_1000-2000m (0,900)	POPULATION_CIRCULAR_0-300m (0,840)	ANY (0,813)
15	WIND_SPEED_MEAN_3h (0,900)	POPULATION_CIRCULAR_300-500m (0,840)	NO2_DIF_3h (0,814)

Vars	25051001	25119002	25120001
1	UV_SURFACE_MEAN_8h (0,555)	UV_SURFACE_MEAN_8h (0,513)	UV_SURFACE_MEAN_8h (0,509)
2	RADIATION_SURFACE (0,660)	WIND_SPEED_MEAN_8h (0,678)	NO2 (0,689)
3	WIND_SPEED_MEAN_24h (0,732)	RADIATION_SURFACE_MEAN_24h (0,762)	RADIATION_SURFACE_MEAN_8h (0,808)
4	NO2 (0,782)	NO2 (0,806)	WIND_SPEED_MEAN_24h (0,849)
5	MES (0,824)	TEMPERATURE_MEAN_24h (0,831)	MES (0,871)
6	TEMPERATURE_MEAN_24h (0,850)	UV_SURFACE_MEAN_3h (0,850)	DIA (0,892)
7	UV_SURFACE_MEAN_24h (0,872)	DIA (0,861)	WIND_DIRECTION (0,901)
8	DIA (0,877)	MES (0,871)	TEMPERATURE_MEAN_24h (0,907)
9	RADIATION_SURFACE_MEAN_24h (0,882)	WIND_DIRECTION (0,878)	ANY (0,913)
10	FOREST_SEMICIRCULAR_UPWIND_500-1000m (0,886)	WIND_SPEED_MEAN_24h (0,883)	UV_SURFACE_MEAN_24h (0,916)
11	HORA (0,888)	NO2_MEAN_24h (0,888)	NO2_MEAN_24h (0,918)
12	RADIATION_SURFACE_MEAN_3h (0,890)	RADIATION_SURFACE_MEAN_3h (0,888)	RADIATION_SURFACE_MEAN_3h (0,920)
13	ANY (0,891)	UV_SURFACE_MEAN_24h (0,889)	RADIATION_SURFACE_MEAN_24h (0,923)
14	WIND_SPEED_MEAN_8h (0,893)	ANY (0,891)	POPULATION_CIRCULAR_0-300m (0,923)
15	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_500-1000m (0,894)	TEMPERATURE_MEAN_8h (0,892)	POPULATION_CIRCULAR_300-500m (0,923)

Vars	25196001	43005002	43014001
1	UV_SURFACE_MEAN_24h (0,438)	UV_SURFACE_MEAN_8h (0,454)	NO2 (0,467)
2	TEMPERATURE_MEAN_24h (0,604)	UV_SURFACE_MEAN_24h (0,543)	UV_SURFACE_MEAN_8h (0,725)
3	DIA (0,704)	NO2 (0,632)	MES (0,794)
4	MES (0,735)	TEMPERATURE_MEAN_24h (0,696)	DIA (0,830)
5	ANY (0,765)	DIA (0,730)	WIND_SPEED_MEAN_8h (0,855)
6	HORA (0,786)	WIND_SPEED_MEAN_8h (0,759)	RADIATION_SURFACE (0,862)
7	RADIATION_SURFACE_MEAN_24h (0,795)	MES (0,770)	ANY (0,871)
8	LANE_DENSITY_SEMICIRCULAR_UPWIND_1000-2000m (0,799)	ANY (0,788)	UV_SURFACE_MEAN_24h (0,877)
9	ROAD_DENSITY_SEMICIRCULAR_UPWIND_1000-2000m (0,800)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_500-1000m (0,793)	TEMPERATURE_MEAN_24h (0,880)
10	POPULATION_CIRCULAR_0-300m (0,800)	UV_SURFACE_MEAN_3h (0,796)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_3000-5000m (0,884)
11	POPULATION_CIRCULAR_300-500m (0,800)	NO2_MEAN_24h (0,798)	WIND_SPEED_MEAN_24h (0,887)
12	POPULATION_CIRCULAR_500-1000m (0,800)	HORA (0,801)	LANE_DENSITY_SEMICIRCULAR_DOWNWIND_3000-5000m (0,888)
13	TEMPERATURE_MEAN_8h (0,800)	AGRICULTURAL_SEMICIRCULAR_UPWIND_500-1000m (0,801)	WATER_SEMICIRCULAR_UPWIND_1000-2000m (0,888)
14	NO2 (0,800)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_1000-2000m (0,802)	AGRICULTURAL_SEMICIRCULAR_DOWNWIND_2000-3000m (0,889)
15	WATER_SEMICIRCULAR_UPWIND_3000-5000m (0,803)	POPULATION_CIRCULAR_0-300m (0,802)	POPULATION_CIRCULAR_0-300m (0,889)

Vars	43047001	43123005
1	UV_SURFACE_MEAN_8h (0,539)	UV_SURFACE_MEAN_8h (0,445)
2	NO2 (0,753)	NO2 (0,737)
3	MES (0,800)	MES (0,810)
4	DIA (0,822)	DIA (0,845)
5	WIND_SPEED_MEAN_8h (0,837)	WIND_SPEED_MEAN_24h (0,862)
6	RADIATION_SURFACE (0,851)	TEMPERATURE_MEAN_24h (0,872)
7	TEMPERATURE_MEAN_24h (0,865)	NO2_MEAN_24h (0,881)
8	ANY (0,870)	RADIATION_SURFACE_MEAN_24h (0,888)
9	LANE_DENSITY_SEMICIRCULAR_UPWIND_1000-2000m (0,876)	UV_SURFACE_MEAN_3h (0,892)
10	RADIATION_SURFACE_MEAN_24h (0,878)	FOREST_SEMICIRCULAR_DOWNWIND_3000-5000m (0,894)
11	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_500-1000 m (0,880)	HORA (0,896)
12	ROAD_DENSITY_SEMICIRCULAR_UPWIND_500-1000m (0,881)	ROAD_DENSITY_SEMICIRCULAR_UPWIND_0-500m (0,896)
13	POPULATION_CIRCULAR_0-300m (0,881)	ANY (0,896)
14	POPULATION_CIRCULAR_300-500m (0,881)	ROAD_DENSITY_SEMICIRCULAR_DOWNWIND_0-500m (0,898)
15	POPULATION_CIRCULAR_500-1000m (0,881)	POPULATION_CIRCULAR_0-300m (0,898)

Vars	43148028	43171001
1	NO2 (0,515)	NO2 (0,504)
2	UV_SURFACE_MEAN_8h (0,772)	UV_SURFACE_MEAN_8h (0,740)
3	RADIATION_SURFACE_MEAN_8h (0,818)	MES (0,801)
4	TEMPERATURE_MEAN_24h (0,840)	DIA (0,836)
5	UV_SURFACE_MEAN_24h (0,859)	HORA (0,853)
6	WIND_SPEED_MEAN_24h (0,873)	WIND_SPEED_MEAN_8h (0,867)
7	DIA (0,883)	TEMPERATURE_MEAN_24h (0,876)
8	MES (0,891)	UV_SURFACE_MEAN_24h (0,883)
9	ROAD_DENSITY_SEMICIRCULAR_UPWIND_500-1000m (0,895)	FOREST_SEMICIRCULAR_DOWNWIND_3000-5000m (0,887)
10	LANE_DENSITY_SEMICIRCULAR_UPWIND_1000-2000m (0,897)	WIND_SPEED_MEAN_24h (0,888)
11	NO2_MEAN_24h (0,899)	ANY (0,893)
12	RADIATION_SURFACE_MEAN_24h (0,902)	WIND_SPEED (0,894)
13	ANY (0,901)	POPULATION_CIRCULAR_0-300m (0,894)
14	WIND_SPEED_MEAN_8h (0,902)	POPULATION_CIRCULAR_300-500m (0,894)
15	POPULATION_CIRCULAR_0-300m (0,902)	POPULATION_CIRCULAR_500-1000m (0,894)