

POLITECNICO DI MILANO  
Scuola di Ingegneria Industriale e dell'Informazione  
Corso di Laurea Magistrale in Computer Science and Engineering



**SVILUPPO DI UN FRAMEWORK DI HEALTH  
GEOMATICS PER LA CORRELAZIONE DI  
VARIABILI AMBIENTALI ED EVENTI  
CLINICI: IL CASO ICTUS IN LOMBARDIA**

Relatore: Prof. Enrico Gianluca Caiani

Tesi di laurea di:  
Davide Cattaneo Matr. 876724

Anno accademico 2018/2019

## **Ringraziamenti**

Esprimo un sentito ringraziamento al Prof. Enrico Caiani per avermi affiancato nello sviluppo di questo elaborato, offrendomi la possibilità di lavorare ad una tesi dal tema molto attuale ed inerente al mio percorso di specializzazione.

Dedico poi un ringraziamento particolare a Guido Francesco Villa e Maurizio Migliori di AREU per la loro disponibilità nel fornirmi tutto il materiale e le spiegazioni necessarie ad approfondire un tema complesso come quello dell'ictus e dell'infrastruttura regionale che regola gli interventi di soccorso.

In ultimo, ma non per importanza, voglio ringraziare il Dott. Mauro Mussin di ARPA ed il Prof. Elio Agostoni dell'Ospedale Niguarda per avermi fornito aiuto nel comprendere al meglio lo stato dell'arte in termini di interpolazioni ed eziologia dell'ictus rispettivamente.

# Indice dei contenuti

<b>SVILUPPO DI UN FRAMEWORK DI HEALTH GEOMATICS PER LA CORRELAZIONE DI VARIABILI AMBIENTALI ED EVENTI CLINICI: IL CASO ICTUS IN LOMBARDIA.....</b>	<b>I</b>
<b>RINGRAZIAMENTI .....</b>	<b>II</b>
<b>INDICE DEI CONTENUTI .....</b>	<b>III</b>
<b>INDICE DELLE FIGURE.....</b>	<b>VI</b>
<b>INDICE DELLE TABELLE.....</b>	<b>XII</b>
<b>SOMMARIO.....</b>	<b>XIII</b>
<b>ABSTRACT .....</b>	<b>XVII</b>
<b>CAPITOLO 1 INTRODUZIONE .....</b>	<b>1</b>
1.1 L'ICTUS .....	1
1.1.1 <i>Ictus e impatto sul Sistema Sanitario Nazionale .....</i>	1
1.1.2 <i>I fattori di rischio .....</i>	3
1.1.3 <i>Il protocollo di soccorso .....</i>	3
1.1.4 <i>Il trattamento e le Stroke Unit in Lombardia .....</i>	4
1.2 ICTUS, FATTORI METEOROLOGICI ED AGENTI INQUINANTI .....	6
1.2.1 <i>Lo stato dell'arte.....</i>	7
1.3 INTRODUZIONE AI SISTEMI GIS E ALLA HEALTH GEOMATICS .....	9
1.3.1 <i>Sistemi di coordinate.....</i>	10
1.4 INTERPOLAZIONE DEI DATI.....	12
1.4.1 <i>Lo stato dell'arte nelle interpolazioni ambientali.....</i>	12
1.4.2 <i>Differenze circostanziali e definizione di un modello differente .....</i>	13
1.5 OBIETTIVO DEL LAVORO.....	14

<b>CAPITOLO 2 MATERIALI E METODI .....</b>	<b>15</b>
2.1 IL DATABASE DEGLI ICTUS.....	15
2.1.1 <i>L'architettura del sistema informativo di AREU</i> .....	15
2.1.2 <i>Metodologia di acquisizione dei dati</i> .....	16
2.1.3 <i>Descrizione dei campi del database</i> .....	16
2.2 DATI MISURATI: FATTORI METEOROLOGICI E AGENTI INQUINANTI .....	22
2.3 DATASET DELLE STAZIONI METEOROLOGICHE.....	25
2.3.1 <i>Distribuzione delle stazioni</i> .....	25
2.3.2 <i>Intervallo di campionamento</i> .....	27
2.4 DATASET DELLE STAZIONI PER LA QUALITÀ DELL'ARIA.....	28
2.4.1 <i>Distribuzione delle stazioni</i> .....	29
2.4.2 <i>Intervallo di campionamento</i> .....	35
2.5 DATI DEMOGRAFICI .....	36
2.5.1 <i>Popolazione residente in ogni provincia</i> .....	36
2.5.2 <i>Popolazione residente in ogni capoluogo</i> .....	37
2.5.3 <i>Incidenza demografica</i> .....	38
2.6 PRE-ELABORAZIONE DEI DATI.....	39
2.6.1 <i>GIS e sistemi di coordinate</i> .....	39
2.6.2 <i>Kriging</i> .....	40
2.6.3 <i>Semivariogramma</i> .....	40
2.6.4 <i>Fitting del semivariogramma</i> .....	42
2.6.5 <i>Generazione del modello di interpolazione</i> .....	45
2.6.6 <i>Analisi e validazione del modello</i> .....	48
2.6.7 <i>Validazione</i> .....	67
2.6.8 <i>Manipolazione dei dati</i> .....	70
2.7 ELABORAZIONE DEI DATI.....	73
2.7.1 <i>Interpolazione dei valori</i> .....	73
2.7.2 <i>Lag period ed intervalli di analisi</i> .....	76
2.7.3 <i>Analisi temporale</i> .....	76
2.7.4 <i>Analisi di correlazione</i> .....	77
<b>CAPITOLO 3 RISULTATI.....</b>	<b>78</b>
3.1 ANALISI DEMOGRAFICA.....	78
3.2 ANALISI TEMPORALE .....	85
3.3 ANALISI DI CORRELAZIONE .....	87
<b>CAPITOLO 4 DISCUSSIONE E CONCLUSIONI .....</b>	<b>100</b>

4.1 VALUTAZIONE DEL PROCESSO DI INTERPOLAZIONE E DEL CAMPIONE DI RIFERIMENTO.....	100
4.2 VALUTAZIONE DEL PROCESSO DI ANALISI DEMOGRAFICA E TEMPORALE .....	102
4.3 VALUTAZIONE DEL PROCESSO DI ANALISI DI CORRELAZIONE .....	104
4.4 SVILUPPI FUTURI E LIMITI DEL METODO D'ANALISI.....	109
<b>BIBLIOGRAFIA.....</b>	<b>111</b>
<b>SITOGRAFIA.....</b>	<b>113</b>
<b>APPENDICE A.....</b>	<b>114</b>
A.1 VALIDAZIONE DEI MODELLI DI INTERPOLAZIONE .....	114
A.2 CROSS VALIDAZIONE DEL MODELLO DI INTERPOLAZIONE.....	118
A.3 ALGORITMI.....	122
A.3.1 <i>Pre-processing dei dati ARPA</i> .....	122
A.3.2 <i>Pre-processing dei dati AREU</i> .....	124
A.3.3 <i>Pre-processing dei dati giornalieri sugli inquinanti</i> .....	125
A.3.4 <i>Pre-processing dei dati AREU in base all'anno</i> .....	126
A.3.5 <i>Elaborazione dei dati e interpolazione valori</i> .....	127
A.3.6 <i>Selezione dei record validi e ricostruzione tabella iniziale</i> .....	133
A.3.7 <i>Calcolo della distribuzione degli eventi</i> .....	136
A.3.8 <i>Rappresentazione della posizione delle centraline ARPA</i> .....	136
A.3.9 <i>Generazione del variogramma e suo fitting</i> .....	137
A.3.10 <i>Generazione di una mappa interpolata</i> .....	141
A.3.11 <i>Generazione delle medie</i> .....	143
A.3.12 <i>Rappresentazione dei fenomeni espressi in bin</i> .....	144

# Indice delle figure

Figura 1 Rappresentazione dei 10 livelli di disabilità espressi tramite la scala EDSS ( <a href="https://notiziemediche.it/info/la-valutazione-della-progressione-malattia/">https://notiziemediche.it/info/la-valutazione-della-progressione-malattia/</a> )	2
Figura 2 Rappresentazione di una proiezione pseudo-cilindrica ( <a href="https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/">https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/</a> )	10
Figura 3 Rappresentazione di una proiezione ellissoidica ( <a href="https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/">https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/</a> )	10
Figura 4 Rappresentazione del sistema cartografico UTM ( <a href="https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/">https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/</a> )	11
Figura 5 Rappresentazione del suolo italiano e dei fusi che lo attraversano nei sistemi di coordinate UTM e Gauss Boaga ( <a href="https://3dmetrica.it/i-codici-epsg/">https://3dmetrica.it/i-codici-epsg/</a> )	11
Figura 6 Centraline rilevanti la temperatura in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	26
Figura 7 Centraline rilevanti l'umidità relativa in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	27
Figura 8 Centraline rilevanti ozono in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	29
Figura 9 Centraline rilevanti biossido di azoto in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	30
Figura 10 Centraline rilevanti ossidi di azoto in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	31

Figura 11 Centraline rilevanti monossido di carbonio in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	32
Figura 12 Centraline rilevanti PM <sub>10</sub> in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	33
Figura 13 Centraline rilevanti PM <sub>2,5</sub> in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	34
Figura 14 Centraline rilevanti benzene in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso	35
Figura 15 Rappresentazione dei confini provinciali e dei capoluoghi della Lombardia	38
Figura 16 Rappresentazione grafica del calcolo della semivarianza ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	41
Figura 17 Rappresentazione grafica di un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	42
Figura 18 Rappresentazione dei parametri che caratterizzano un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	43
Figura 19 Rappresentazione di un modello di fitting lineare per un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	43
Figura 20 Rappresentazione di un modello di fitting sferico per un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	44
Figura 21 Rappresentazione di un modello di fitting esponenziale per un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	44
Figura 22 Rappresentazione di un modello di fitting gaussiano per un semivariogramma ( <a href="http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm">http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm</a> )	44

Figura 23 Rappresentazione grafica della relazione che lega varianza e covarianza ( <a href="http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/seminvariogram-and-covariance-functions.htm">http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/seminvariogram-and-covariance-functions.htm</a> )	45
Figura 24 Rappresentazione grafica delle curve di varianza e covarianza ( <a href="http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/seminvariogram-and-covariance-functions.htm">http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/seminvariogram-and-covariance-functions.htm</a> )	46
Figura 25 Rappresentazione parametrica di un semivariogramma [Lichtenstern A. 2013]	48
Figura 26 Semivariogramma della temperatura	50
Figura 27 Fitting esponenziale per il semivariogramma della temperatura	50
Figura 28 Fitting sferico per il semivariogramma della temperatura	50
Figura 29 Rappresentazione di una mappa di interpolazione per la temperatura in Lombardia	51
Figura 30 Semivariogramma dell'umidità relativa	52
Figura 31 Fitting esponenziale per il semivariogramma dell'umidità relativa	52
Figura 32 Fitting sferico per il semivariogramma dell'umidità relativa	52
Figura 33 Rappresentazione di una mappa di interpolazione per l'umidità relativa in Lombardia	53
Figura 34 Semivariogramma dell'ozono	54
Figura 35 Fitting esponenziale per il semivariogramma dell'ozono	54
Figura 36 Fitting sferico per il semivariogramma dell'ozono	54
Figura 37 Rappresentazione di una mappa di interpolazione per l'ozono in Lombardia	55
Figura 38 Semivariogramma del biossido di azoto	56
Figura 39 Fitting esponenziale per il semivariogramma del biossido di azoto	56
Figura 40 Fitting sferico per il semivariogramma del biossido di azoto	56
Figura 41 Rappresentazione di una mappa di interpolazione per il biossido di azoto in Lombardia	57
Figura 42 Semivariogramma degli ossidi di azoto	58
Figura 43 Fitting esponenziale per il semivariogramma degli ossidi di azoto	58
Figura 44 Fitting sferico per il semivariogramma degli ossidi di azoto	58
Figura 45 Rappresentazione di una mappa di interpolazione per gli ossidi di azoto in Lombardia	59
Figura 46 Semivariogramma del monossido di carbonio	60
Figura 47 Fitting esponenziale per il semivariogramma del monossido di carbonio	60
Figura 48 Fitting sferico per il semivariogramma del monossido di carbonio	60
Figura 49 Rappresentazione di una mappa di interpolazione per il monossido di carbonio in Lombardia	61

Figura 50 Semivariogramma del benzene	62
Figura 51 Fitting esponenziale per il semivariogramma del benzene	62
Figura 52 Fitting sferico per il semivariogramma del benzene	62
Figura 53 Rappresentazione di una mappa di interpolazione per il benzene in Lombardia	63
Figura 54 Semivariogramma del PM10	64
Figura 55 Fitting esponenziale per il semivariogramma del PM10	64
Figura 56 Fitting sferico per il semivariogramma del PM10	64
Figura 57 Rappresentazione di una mappa di interpolazione per il PM10 in Lombardia	65
Figura 58 Semivariogramma del PM2.5	66
Figura 59 Fitting esponenziale per il semivariogramma del PM2.5	66
Figura 60 Fitting sferico per il semivariogramma del PM2.5	66
Figura 61 Rappresentazione di una mappa di interpolazione per il PM2.5 in Lombardia	67
Figura 62 Rappresentazione grafica di un grid file della temperatura sul territorio della Lombardia elaborato da ARPA	68
Figura 63 Rappresentazione grafica di una mappa di interpolazione della temperatura sul territorio della Lombardia	68
Figura 64 Rappresentazione grafica di una mappa di interpolazione della temperatura per il comune di Milano	69
Figura 65 Incidenza percentuale negli anni 2015-2017 in relazione alle province di residenza	82
Figura 66 Incidenza percentuale riportata in base al sesso dei pazienti ed esplicitata per ognuno dei tre anni di interesse	83
Figura 67 Rappresentazione della distribuzione del numero di ictus in base al sesso e per ciascuna delle fasce d'età	84
Figura 68 Distribuzione aggregata degli ictus in base all'orario di intervento dei soccorsi	86
Figura 69 Distribuzione degli ictus in base all'orario di intervento dei soccorsi negli anni 2015-2017	86
Figura 70 Distribuzione aggregata degli ictus in base al giorno di intervento dei soccorsi (0 = lunedì, 6 = domenica)	86
Figura 71 Distribuzione degli ictus in base al giorno di intervento dei soccorsi negli anni 2015-2017	86
Figura 72 Distribuzione aggregata degli ictus in base al mese di intervento dei soccorsi (1 = gennaio, 12 = dicembre)	87

Figura 73 Distribuzione degli ictus in base al mese di intervento dei soccorsi negli anni 2015-2017	87
Figura 74 Dislocazione delle centraline meteorologiche nella città di Milano	88
Figura 75 Dislocazione delle centraline per la qualità dell'aria nella città di Milano	88
Figura 76 Distribuzione dei casi di ictus accertati dal 2015 al 2017 sul territorio della città di Milano	88
Figura 77 Distribuzione degli ictus nella città di Milano nel corso del triennio 2015-2017	89
Figura 78 Grafico esprimente la relazione ictus - temperatura	90
Figura 79 Grafico esprimente la relazione ictus - umidità relativa	90
Figura 80 Grafico esprimente la relazione ictus - ozono	90
Figura 81 Grafico esprimente la relazione ictus - biossido di azoto	90
Figura 82 Grafico esprimente la relazione ictus - ossidi di azoto	90
Figura 83 Grafico esprimente la relazione ictus - monossido di carbonio	90
Figura 84 Grafico esprimente la relazione ictus - benzene	91
Figura 85 Grafico esprimente la relazione ictus - PM10	91
Figura 86 Grafico esprimente la relazione ictus - PM2.5	91
Figura 87 Distribuzione degli ictus e della temperatura in funzione dell'aggregazione in bin	94
Figura 88 Distribuzione degli ictus e dell'umidità relativa in funzione dell'aggregazione in bin	94
Figura 89 Distribuzione degli ictus e dell'ozono in funzione dell'aggregazione in bin	94
Figura 90 Distribuzione degli ictus e del biossido di azoto in funzione dell'aggregazione in bin	95
Figura 91 Distribuzione degli ictus e degli ossidi di azoto in funzione dell'aggregazione in bin	95
Figura 92 Distribuzione degli ictus e del monossido di carbonio in funzione dell'aggregazione in bin	95
Figura 93 Distribuzione degli ictus e del benzene in funzione dell'aggregazione in bin	96
Figura 94 Distribuzione degli ictus e del PM10 in funzione dell'aggregazione in bin	96
Figura 95 Distribuzione degli ictus e del PM2.5 in funzione dell'aggregazione in bin	96
Figura 96 Rappresentazione grafica di una mappa di interpolazione dell'umidità relativa per il comune di Milano	115
Figura 97 Rappresentazione grafica di una mappa di interpolazione dell'ozono per il comune di Milano	115
Figura 98 Rappresentazione grafica di una mappa di interpolazione del biossido di azoto per il comune di Milano	116

Figura 99 Rappresentazione grafica di una mappa di interpolazione degli ossidi di azoto per il comune di Milano	116
Figura 100 Rappresentazione grafica di una mappa di interpolazione del monossido di carbonio per il comune di Milano	117
Figura 101 Rappresentazione grafica di una mappa di interpolazione del benzene per il comune di Milano	117
Figura 102 Rappresentazione grafica di una mappa di interpolazione del PM10 per il comune di Milano	118
Figura 103 Rappresentazione grafica di una mappa di interpolazione del PM2.5 per il comune di Milano	118
Figura 104 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano Lambrate	119
Figura 105 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Brera	119
Figura 106 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Juvara	120
Figura 107 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Marche	120
Figura 108 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano p.zza Zavattari	121
Figura 109 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Feltre	121

# **Indice delle tabelle**

Tabella 1 Elenco delle UCV in Lombardia	5
Tabella 2 Pubblicazioni sul tema ictus note in letteratura	7
Tabella 3 Campi del database AREU	16
Tabella 4 Formato dei dati salvati nel database ARPA	23
Tabella 5 Campi associati ad ogni centralina gestita da ARPA	24
Tabella 6 Misure rilevate dalle stazioni meteorologiche ARPA	25
Tabella 7 Misure rilevate dalle stazioni per la qualità dell'aria ARPA	28
Tabella 8 Campi contenuti nella base dati ISTAT	37
Tabella 9 Campi relativi ai capoluoghi di provincia	37
Tabella 10 Validazione del modello di interpolazione per la temperatura	69
Tabella 11 Validazione del modello di interpolazione per tutte le variabili indagate	70
Tabella 12 Popolazione residente in base alla provincia	78
Tabella 13 Vittime di ictus in base alla provincia di residenza	80
Tabella 14 Incidenza percentuale in base alla provincia	81
Tabella 15 Incidenza dell'ictus per fasce d'età	84
Tabella 16 Risultati dell'analisi di correlazione subitanea con le variabili in esame	91
Tabella 17 Risultati dell'analisi di correlazione con la media giornaliera delle variabili in esame	92
Tabella 18 Risultati della correlazione tra la distribuzione degli ictus e la distribuzione delle variabili	94
Tabella 19 Risultati dell'analisi di correlazione con le medie mensili delle variabili in esame	98

## Sommario

L'ictus ischemico in Lombardia rappresenta la seconda causa di chiamata al numero unico di emergenza (112) e colpisce ogni anno circa 18.000 persone su una popolazione totale di 10 milioni di abitanti.

La patologia si manifesta in seguito alla formazione di un trombo, dovuto ad arteriosclerosi o ad embolia cardiaca, che occlude vasi arteriosi all'interno del cervello, impedendo di fatto l'ossigenazione dei tessuti. Questa condizione è trattabile tramite apposite terapie fibrinolitiche o mediante asportazione meccanica del trombo effettuata in ospedali attrezzati con Stroke Unit e dunque abilitati all'intervento. In caso di azione rapida, cioè entro 4-6 ore, si riesce a prevenire il decesso. Nei casi non fatali ci si trova comunque a dover affrontare disabilità permanenti che portano a danni economici ed umani incalcolabili. Solo la disostruzione effettuata entro al più un'ora dall'insorgenza dei sintomi è in grado di scongiurare handikap a lungo termine e di garantire un recupero prossimo al 100%.

La morte per ictus sopraggiunge in seguito all'anossia cerebrale che danneggia il lobo frontale del cervello, responsabile delle attività di controllo del cuore.

La malattia è notoriamente associata ad un'incidenza superiore nella popolazione over 65 anni e in un periodo storico in cui l'età media della popolazione è in costante aumento, i casi di ictus registrati, stando ai dati, sembrano seguire il medesimo andamento crescente.

Assume dunque una particolare importanza lo studio dettagliato di questo fenomeno in relazione alle possibili cause scatenanti, così da circoscrivere le aree a rischio, stilare piani d'intervento efficaci e possibilmente predire i casi.

In letteratura, particolare interesse è stato dato all'individuazione di un legame tra ictus, fattori meteorologici ed agenti inquinanti. Le prime ricerche in tal senso risalgono agli inizi degli anni 2000 ed in particolare sono attribuibili a [Hong Y. et al. 2002], che osservando i dati raccolti nella città di Seoul nel periodo compreso tra il 1991 e il 1997 giungeva alla conclusione che TSP (particolato totale sospeso), SO<sub>2</sub>, NO<sub>2</sub>, CO e O<sub>3</sub> fossero significativamente riconducibili ad un incremento del rischio collegato all'insorgenza di ictus per via dell'azione infiammatoria che questi inquinanti esercitano sull'organismo.

Altri lavori di ricerca si sono succeduti ed altre città del mondo sono state utilizzate come base per le analisi. La maggior parte di queste si trova grossomodo concorde con quanto scoperto da [Hong Y. et al. 2002].

Lo scopo di questo lavoro di tesi, svolto in collaborazione con AREU (Agenzia Regionale Emergenze urgenze), è quello di proporre un approccio innovativo nell'ambito degli studi clinici. Si propone infatti un framework di Health Geomatics, ossia la scienza che studia i fenomeni medici associati al territorio, che riconduca ai record dei pazienti informazioni sulla demografia e sulle variabili atmosferiche campionate nell'ambiente di riferimento.

Si usano metodologie tipiche della geomatica per cercare di ottenere dati più affidabili dal punto di vista qualitativo, sfruttando le risorse a disposizione per offrire una visione più ampia e dettagliata dei fenomeni sotto indagine.

Questa metodologia viene applicata al caso degli ictus registrati in Lombardia negli anni 2015-2017 ed in particolare indaga il fenomeno restringendolo all'area della città di Milano, che con una popolazione di oltre un milione di abitanti risulta particolarmente interessata dal fenomeno.

I dati territoriali utilizzati sono forniti da ARPA (Agenzia Regionale per la Protezione dell'Ambiente) tramite la piattaforma open data della Regione Lombardia mentre i dati sugli ictus sono forniti, in forma anonimizzata, direttamente da AREU. Le informazioni demografiche sono reperite tramite il portale web pubblico messo a disposizione da ISTAT (Istituto nazionale di STATistica).

I fenomeni indagati in questo elaborato sono: temperatura, umidità relativa, ozono, biossido di azoto, ossidi di azoto, monossido di carbonio, benzene, PM10 e PM2.5. La scelta è dettata dal fatto che quanto presentato corrisponda alla massima selezione di fenomeni possibile.

I record all'interno del database AREU sono geolocalizzati e presentano le diverse posizioni degli eventi di ictus. I dati sono stati pre-processati in modo da attribuire a ciascuno di questi record un valore per ognuna delle variabili ambientali sotto esame. E' stato attribuito un valore a tutte le variabili studiate per ognuno dei 5 giorni di lag period antecedenti all'evento; tale scelta è da ricondurre agli studi in letteratura, ed in particolare a [Guo P. et al. 2017], che individua in 5 il numero massimo di giorni antecedenti all'ictus entro cui si individuano correlazioni significative con i parametri studiati.

Nella fase di analisi si sono valutate innanzitutto le percentuali di incidenza degli eventi di ictus in relazione alla provincia di residenza, al sesso e all'età del paziente. In seguito si sono studiati i trend caratteristici del fenomeno, scomponendolo secondo tre diverse granularità: oraria, giornaliera e mensile, in modo da mettere in luce eventuali ciclicità.

In queste prime due fasi dell'analisi, il dataset fornito da AREU è stato utilizzato nella sua completezza e i dati sono stati elaborati su base regionale in quanto le operazioni non risultano inficiate dalla qualità delle interpolazioni per le variabili atmosferiche.

L'ultimo spunto d'analisi ha invece visto protagonista l'analisi di correlazione tra gli eventi di ictus ed i dati ambientali interpolati con metodo Kriging a partire dalle centraline ARPA.

Per quest'ultima fase dell'analisi si è deciso di effettuare un focus specifico sulla città di Milano, che grazie all'elevata densità di centraline presenti sul territorio consente di ottenere interpolazioni di grande precisione e allo stesso tempo conta una popolazione campione adatta allo scopo.

L'analisi di correlazione è stata eseguita con tre granularità differenti: dapprima si è correlato il numero di ictus giornaliero con la media calcolata sulla base dei valori interpolati e associati a ciascuno dei record (per ognuno dei lag period). Questa analisi preliminare ha permesso di verificare se sussistesse un'associazione tra l'insorgenza di ictus ed i valori istantanei delle variabili.

In secondo luogo si è proceduto nel relazionare il numero di ictus alle medie giornaliere della città per verificare se fosse invece l'esposizione giornaliera a svolgere un ruolo determinante nella manifestazione dei sintomi. Anche in questo caso si è tenuto conto dei 5 giorni di lag period antecedenti alla chiamata al 112.

In ultima istanza si è invece proceduto ad associare il conteggio dei casi aggregati su base mensile alle medie, sempre mensili, dei diversi fenomeni monitorati. Quest'ultima fase ha lo scopo di verificare se l'insorgenza della patologia non sia dovuta ad un'interazione con le variabili ambientali che ecceda i 5 giorni noti in letteratura e sia dunque da ascrivere ad un'esposizione più duratura nel tempo.

Tutte le analisi di correlazione, visto il dominio discreto delle variabili in esame, sono state verificate usando la correlazione a ranghi di Spearman.

I risultati ottenuti mostrano un'incidenza in costante aumento durante il triennio 2015-2017 in pressoché tutte le province lombarde e le donne risultano essere più interessate dal fenomeno rispetto agli uomini. Entrando nello specifico si evidenziano due comportamenti distinti, che vedono la popolazione maschile più colpita nelle fasce d'età al di sotto dei 77 anni mentre superati i 78 anni la situazione si inverte.

Le fasce orarie più attive per quanto riguarda gli interventi erogati dal 112 sono quelle della prima mattinata, periodo in cui verosimilmente ci si accorge dei sintomi dopo la notte.

Il trend giornaliero vede una predominanza dell'incidenza nei giorni lavorativi mentre nel fine settimana si assiste ad un calo dell'8% nel numero dei casi.

Annualmente si assiste ad una distribuzione che vede i picchi massimi nel periodo freddo dell'anno, da ottobre a marzo, mentre si assiste ad un minimo in corrispondenza di agosto,

mese nel quale si registra tuttavia un forte aumento nelle partenze dei vacanzieri, fenomeno che contribuisce a ridurre la popolazione campione.

Per quanto riguarda l'analisi di correlazione ci si trova di fronte a correlazioni deboli in tutti i casi valutati su base giornaliera, con al più una leggera indicazione di causalità data dagli inquinanti quali NO<sub>2</sub>, NO<sub>x</sub>, benzene, PM<sub>10</sub> e PM<sub>2.5</sub>, che sebbene risultino essere i medesimi inquinanti riportati in letteratura, non forniscono numeri tali da poter considerare indicativi i risultati ottenuti.

L'analisi effettuata su base mensile mostra invece correlazioni moderate con le variabili sopra citate, segno che effettivamente la correlazione è da ricercarsi più in un 'esposizione a lungo termine piuttosto che non nei 5 giorni precedenti all'ictus.

In conclusione, si è sviluppato un framework di Health Geomatics per la ricerca di correlazione tra i dati clinici e le variabili ambientali, presentando nel dettaglio l'applicazione al caso degli ictus in Lombardia nel triennio 2015-2017.

I risultati mostrano come la componente puramente ambientale non sia il solo fattore scatenante della patologia, che come indicano i dati sull'incidenza giornaliera, potrebbe trovare terreno fertile nell'esposizione a stress lavorativo o di altra natura accumulato in settimana. Si è mostrato come, tra le granularità analizzate, quella relativa all'esposizione a lungo termine risulti tuttavia la strada verso cui dirigere futuri ed ulteriori approfondimenti, legando possibilmente le analisi territoriali ai dati biometrici e alle valutazioni sullo stress personale.

## **Abstract**

The ischemic stroke in Lombardy represents the second cause of emergency call to the national emergency number (112) and affects every about 18,000 people every year, over a total population of 10 million inhabitants.

The disease manifests itself following the formation of a thrombus, due to arteriosclerosis or cardiac embolism, which occludes arterial vessels inside the brain, preventing the oxygenation of tissues. This condition is treatable through specific fibrinolytic therapies or by mechanical asportation of the thrombus performed in hospitals having a Stroke Unit and thus qualified for surgery. In the case of rapid action, ie within 4-6 hours, death can be prevented. In non-fatal cases, however, the patients can face permanent disabilities that lead to incalculable economic and human damage. Only the intervention carried out within one hour from the onset of symptoms is able to ward off long-term handicap and guarantee an almost certain recovery.

Stroke death occurs as a result of cerebral anoxia, which damages the frontal lobe of the brain, being responsible for controlling the heartbeat.

The disease is notoriously associated with a higher incidence in the population over 65 years and in a historical period in which the average age of the population is constantly increasing, the recorded cases of stroke, according to the data, seem to follow the same increasing trend. Therefore, the detailed study of this phenomenon assumes particular importance in relation to the possible triggering causes, so to circumscribe the areas at risk, draw up effective action plans and possibly predict the cases.

In the literature, particular interest has been given to the identification of a link between stroke, meteorological factors and pollutants. The first researches in this sense date back to the beginning of the 2000s and in particular are attributable to [Hong Y. et al. 2002], which, by observing the data collected in the city of Seoul in the period between 1991 and 1997, concluded that TSP (total suspended particulate), SO<sub>2</sub>, NO<sub>2</sub>, CO and O<sub>3</sub> were significantly attributable to an increase in the risk linked to onset of stroke due to the inflammatory action these pollutants exert on the organism.

Other research work followed and other cities in the world were used as a basis for analysis. Most of these are roughly in agreement with what was discovered by [Hong Y. et al. 2002].

The aim of this thesis work, carried out in collaboration with AREU (Agenzia Regionale Emergenze Urgenze), is to propose an innovative approach in clinical studies. A Health Geomatics (namely the science that studies the medical phenomena associated with the territory) framework is proposed to associate information on demography and atmospheric variables sampled in the reference environment directly to the patients' records.

A methodology typical of geomatics is used to try to obtain more reliable data from the qualitative point of view, using the resources available to offer a broader and more detailed view of the phenomena under investigation.

This methodology is applied to the case of stroke registered in Lombardy in the years 2015-2017 and, in particular, it investigates the phenomenon by restricting it to the area of the city of Milan, which, counting on a population of over one million inhabitants, is particularly affected by the phenomenon.

The territorial data used for the analysis are provided by ARPA (Agenzia Regionale per la Protezione dell'Ambiente) through the open data platform held by Regione Lombardia, while the data about strokes are provided, in an anonymised form, directly by AREU. The demographic information is obtained through the public web portal made available by ISTAT (Istituto Nazionale di STATistica).

The phenomena investigated in this paper are: temperature, relative humidity, ozone, nitrogen dioxide, nitrogen oxides, carbon monoxide, benzene, PM<sub>10</sub> and PM<sub>2.5</sub>. The choice is dictated by the fact that what is presented corresponds to the maximum selection of available variables.

The records within the AREU database are geolocated and present the different locations of stroke events. The data were pre-processed in such a way as to give each of these records a value for each of the environmental variables under examination. A value has been assigned to all the variables studied for each of the 5 days of lag period prior to the event; this choice is due to studies in the literature, and in particular to [Guo P. et al. 2017], which identifies in the number 5 the maximum amount of days prior to the stroke within which significant correlations are identified with the studied parameters.

In the analysis phase, the percentages of incidence of stroke events were assessed in relation to the province of residence, sex and age of the patient. Subsequently, the characteristic trends of the phenomenon were studied on the basis of three different granularities: hourly, daily and monthly, in order to highlight any cyclicity.

In these first two phases of the analysis, the dataset provided by AREU was used in its entirety and the data were processed on a regional basis as the operations are not affected by the quality of interpolations for atmospheric variables.

On the other hand, the last analysis was focused on the correlation analysis between the stroke events and the environmental data interpolated with the Kriging method starting from the ARPA stations.

For this last phase of the analysis it was decided to carry out a specific focus on the city of Milan, which thanks to the high density of environmental station present on the territory allows to obtain interpolations of great precision and at the same time counts a sample population suitable for the purpose.

The correlation analysis was performed with three different granularities: first the daily stroke number was correlated with the average of the variables calculated on the basis of the interpolated values associated with each of the records (for each of the lag periods). This preliminary analysis allowed to verify if there was an association between the onset of stroke and the instantaneous values of the monitored parameters.

Secondly, the number of strokes was reported to the daily averages of the city to see if daily exposure would play a decisive role in the manifestation of symptoms. Also in this case the 5 days of lag period prior to the call to 112 were taken into account.

As a last resort, the aggregate cases count on a monthly basis was associated with the monthly average of the different monitored phenomena. This last phase aims to verify the existence of a dependency relation to be ascribed to a longer lasting exposure over time.

All correlation analyzes, given the discrete domain of the variables under examination, were verified using Spearman's rank correlation.

The obtained results show an increasing incidence during the three year period 2015-2017 in almost all the provinces in Lombardy and also prove that women are more interested in the phenomenon than men do. Specifically, it worth to highlight two distinct behaviors, which see the male population most affected in the age groups below 77 years, while over 78 years the situation is reversed.

The most active time slots for the interventions provided by the 112 are those of the early morning, a period in which one most likely notice the symptoms after the night.

The daily trend sees a predominance of the incidence in the working days while at the weekend there is a decrease of 8% in the number of cases.

Yearly distribution of the events sees the maximum peak in the cold period of the year, from October to March, while there is a minimum in August, a month in which there is a strong increase in the departures for vacationers, a phenomenon that helps to reduce the sample population.

Analysis results report a weak correlations in all the cases assessed on a daily basis, with at the very least a slight indication of causality given by pollutants such as NO<sub>2</sub>, NO<sub>x</sub>, benzene, PM<sub>10</sub> and PM<sub>2.5</sub>, which although they appear to be the same pollutants reported in the

literature, do not provide numbers high enough to be able to consider the results as indicative.

On the other hand, the analysis carried out on a monthly basis shows moderate correlations with the variables previously mentioned, a sign that the correlation is to be found more in a long-term exposure rather than in the 5 days preceding the stroke.

In conclusion, a Health Geomatics framework was developed for the correlation research between clinical data and environmental variables, presenting in detail the application to the case of stroke in Lombardy in the 2015-2017 three-year period.

The results show that the purely environmental component is not the only triggering factor of the disease, which, as the data on the daily incidence indicate, could find fertile ground in exposure to work stress or other stress accumulated during the week. It has been shown that, among the analyzed granularities, that relating to long-term exposure, is the one to direct future and further investigations, possibly linking territorial analysis to biometric data and assessments on personal stress.

# CAPITOLO 1

## INTRODUZIONE

### 1.1 L' ICTUS

#### 1.1.1 Ictus e impatto sul Sistema Sanitario Nazionale

L'ictus ischemico è una patologia vascolare causata dall'ostruzione di uno o più vasi sanguigni arteriosi localizzati per lo più nell'area legata al lobo frontale del cervello e in quella carotide.

L'ictus ischemico si differenzia dall'ictus emorragico per la natura ostruttiva del fenomeno, causato dalla formazione di trombi che vanno a bloccare la circolazione sanguigna e portano all'anossia cerebrale (<https://en.wikipedia.org/wiki/Stroke>).

L'origine di questi trombi è attribuibile a due fenomeni:

- **Arteriosclerosi:** l'origine arteriosclerotica è dovuta alla formazione di placche all'interno dei principali vasi sanguigni. Queste placche sono originate da un'infiammazione dell'endotelio che porta alla calcificazione della ferita e all'aggregazione di materiale plasmatico quale lipidi, piastrine e globuli rossi. Il distaccamento di questi coaguli porta alla formazione di trombi, che in seguito generano l'ostruzione arteriosa responsabile dell'ictus.

- **Embolia:** l'origine embolica è dovuta alla formazione di coaguli generatisi all'interno delle cavità atriali. Questa genesi è attribuibile principalmente a malfunzionamenti del muscolo cardiaco, all'età del paziente e a difetti congeniti. L'embolia è spesso legata a problemi di aritmia cardiaca, frequente nei pazienti over 65.

In Italia l'ictus rappresenta la terza causa di morte dopo le malattie cardiache e le neoplasie. Sono circa 18.000 i pazienti colpiti ogni anno in Lombardia e le chiamate raccolte dal numero di soccorso per questo genere di eventi raggiunge il 33.9% del totale [<http://www.ospedalivarese.net/files/corsiformazione/1503/3433/DELODOVICI.pdf>].

Se non trattato l'ictus porta alla morte del paziente nel 20% delle diagnosi; il decesso è dovuto nella maggior parte dei casi ad arresto cardiaco in quanto l'anossia cerebrale intacca il tronco encefalico responsabile del controllo delle attività del cuore.

Il decorso dell'ictus è tempo correlato ed è tanto migliore quanto più rapidi sono i soccorsi. Nel migliore dei casi, ossia quello in cui si è in grado di intervenire entro un'ora dell'insorgenza, si raggiungono percentuali di perfetta guarigione prossime al 100%. Non sono tuttavia rari i casi di insorgenza di handikap o disabilità motorie dovute al danneggiamento del lobo cerebrale frontale.

In Italia si stimano un milione di individui colpiti con vari livelli di disabilità acquisita. Questa disabilità è misurata utilizzando la scala EDSS (Extended Disability Status Scale), che definisce un range da 0 a 10 con cui quantificare l'invalidità:



Figura 1 Rappresentazione dei 10 livelli di disabilità espressi tramite la scala EDSS (<https://notiziemediche.it/info/la-valutazione-della-progressione-malattia/>)

Il decorso è considerato positivo solamente se non viene superato il livello 2 della EDSS. Il danno economico causato dagli ictus è difficilmente quantificabile in quanto non limitato ad una cifra da destinare alle pensioni di invalidità permanente ma bensì esteso all'impatto che la condizione ha sulla vita delle vittime e su quella dei loro parenti.

---

## **1.1.2 I fattori di rischio**

I fattori di rischio noti per essere cause di ictus riguardano sia cause modificabili e dipendenti dallo stile di vita sia cause non direttamente modificabili.

Fra i rischi modificabili si annoverano fumo, alcool, ipertensione, diabete, sovrappeso e sedentarietà.

I rischi non modificabili contemplano invece età, sesso e predisposizione familiare.

I dati mostrano come tra i fattori indipendenti l'età giochi un ruolo fondamentale nell'insorgenza di ictus ischemico. Il rischio, infatti, aumenta di 10 volte ogni 10 anni una volta superati i 65 anni [<https://www.humanitas.it/malattie/ictus-cerebrale>].

Un ultimo fattore di rischio è collegato ad eventi di ictus transitorio, che normalmente si risolvono autonomamente con decorso positivo. In questi casi si è osservato che l'incidenza di ictus non transitori nei pazienti aumenta fino a 30 volte, a parità di condizioni, rispetto a quanti non hanno mai manifestato episodi transitori [[https://en.wikipedia.org/wiki/Transient\\_ischemic\\_attack](https://en.wikipedia.org/wiki/Transient_ischemic_attack)].

## **1.1.3 Il protocollo di soccorso**

Il protocollo di soccorso viene avviato non appena viene ricevuta una chiamata di soccorso al numero unico 112, con conseguente avvio di una missione legata alla richiesta di intervento.

Per individuare un sospetto caso di ictus, l'operatore fa uso della CPSS (Cincinnati Prehospital Stroke Scale), ossia una metrica dalla comprovata efficacia [Kothari RU et al. 1999] per valutare la condizione del paziente sulla base delle risposte ad alcune semplici domande:

- Parla male?
- Ha la bocca storta?
- Riesce a tenere le braccia sollevate?
- Da quanto tempo ha questo problema?

Tali domande mirano ad individuare rispettivamente:

- Anomalie del linguaggio
- Paresi facciale
- Deficit motorio degli arti superiori

---

Sulla base delle risposte ai precedenti quesiti, gli operatori attribuiscono un codice di gravità all'intervento e dispongono l'invio dei mezzi più adeguati. Tra questi mezzi figurano:

- Mezzo di soccorso di base (MSB): un'ambulanza con a bordo personale tecnico soccorritore.
- Mezzo di soccorso intermedio (MSI): un'ambulanza che preveda a bordo un infermiere attrezzato e autorizzato ad operare procedure di intervento avanzato.
- Mezzo di soccorso avanzato (MSA): un mezzo su gomma (ambulanza o automedica) e a pala (elicottero) con a bordo un infermiere ed un medico di pronto soccorso.

Giunti sul posto, gli operatori del 118 verificano le condizioni del paziente ed eventualmente attribuiscono alla missione il codice ictus, ossia se la CPSS restituisce nuovamente esito positivo, il paziente ha più di 18 anni e i sintomi risultano presenti da non oltre 4 ore. Il tal caso gli operatori comunicano con l'hub centrale al fine di farsi indirizzare verso la struttura ospedaliera più adeguata al trattamento. In caso di rischio particolarmente elevato, con annessa necessità di stabilizzare il paziente, si procede invece al trasporto presso il pronto soccorso più vicino.

E' compito dell'hub del 118 inviare l'allerta circa un paziente in arrivo all'ospedale di destinazione, che procederà ad allertare il neurologo di riferimento. Il neurologo svolge il ruolo di attending doctor.

Dopo l'arrivo in pronto soccorso, il paziente viene immediatamente inviato in radiologia e sottoposto ad una TAC per verificare lo stato cerebrale. Successivamente si procede con la terapia fibrinolitica o con l'asportazione meccanica del trombo.

#### **1.1.4 Il trattamento e le Stroke Unit in Lombardia**

Il trattamento di un caso di ictus prevede la dissoluzione o la rimozione meccanica di un trombo. Il trombo è costituito da una massa ematica solida che si origina all'interno del sistema cardiocircolatorio. Generalmente si tratta di un coagulo di sangue costituito da globuli rossi, bianchi, piastrine e fibrina.

- Trattamento fibrinolitico: l'obiettivo è quello di disgregare la fibrina che tiene unito il coagulo in modo da ripristinare la normale circolazione. Il trattamento può sortire esito positivo solo se erogato entro 4.5 ore dall'insorgenza dei sintomi.

- 
- Asportazione meccanica: tramite intervento chirurgico si individua il trombo e si procede alla sua asportazione mediante l'uso di un catetere endovenoso. Il trattamento meccanico è sempre eseguito congiuntamente alla terapia fibrinolitica e può sortire esito positivo solo se erogato entro 6 ore dall'insorgenza dei sintomi.

Sul territorio italiano i presidi ospedalieri dotati di Stroke Unit (o UCV, Unità Cardio Vascolare) sono classificati su due livelli: Livello II e Livello III [[http://www.quadernidellasalute.it/imgs/C\\_17\\_pubblicazioni\\_1698\\_allegato.pdf](http://www.quadernidellasalute.it/imgs/C_17_pubblicazioni_1698_allegato.pdf)].

Gli ospedali in cui è presente una Stroke Unit di secondo livello sono quelli adibiti al solo trattamento fibrinolitico mentre si delega alle Stroke Unit di terzo livello l'asportazione meccanica dei trombi.

In Lombardia si contano 42 strutture attrezzate con UCV e il loro elenco è riportato nella tabella sottostante.

Tabella 1 Elenco delle UCV in Lombardia

<b>PRESIDIO OSPEDALIERO</b>	<b>CITTA'</b>
A. O. Spedali Civili di Brescia	Brescia
A.O. Mellino Mellini	Chiari (BS)
Azienda Ospedaliera Carlo Poma	Mantova
Azienda Ospedaliera Niguarda Cà Granda	Milano
Azienda Ospedaliera Ospedali Riuniti di Bergamo	Bergamo
Azienda Ospedaliera S. Anna	Como
Azienda Ospedaliera S. Antonio Abate	Gallarate
Azienda Ospedaliera Treviglio	Caravaggio (BG)
Fond. Macchi	Varese
IRCCS Fond. Istituto Neurologico Naz. C. Mondino	Pavia
IRCCS Policlinico S. Matteo	Pavia
Ist. Auxologico Italiano IRCCS Milano Istituto Scientifico San Luca	Milano
Istituto Clinico Beato Matteo	Vigevano
Istituto Clinico Città Studi	Milano
Istituto Humanitas	Rozzano
Osp. Sondrio Azienda Ospedaliera Valtellina	Valchiavenna
Ospedale A. Manzoni	Lecco

---

Ospedale Bolognini	Seriate (BG)
Ospedale di Busto Arsizio	Busto Arsizio
Ospedale di Crema	Crema
Ospedale di Cremona	Cremona
Ospedale di Desio	Desio (MB)
Ospedale di Esine	Esine (BS)
Ospedale di Legnano	Legnano
Ospedale di Lodi	Lodi
Ospedale di Magenta	Magenta (MI)
Ospedale di Merate	Merate
Ospedale di Saronno	Saronno
Ospedale di Vimercate	Vimercate (MB)
Ospedale di Vizzolo Predabissi	Vizzolo Predabissi
Ospedale di Voghera	Voghera
Ospedale Guido Salvini	Garbagnate Milanese
Ospedale Ponte San Pietro	Bergamo
Ospedale S. Donato	San Donato Milanese
Ospedale Sacco	Milano
Ospedale San Carlo	Milano
Ospedale San Gerardo	Monza
Ospedale San Giuseppe	Milano
Ospedale San Raffaele	Milano
Ospedale Valduce	Como
Poliambulanza Brescia	Brescia
Policlinico San Marco	Zingonia (BG)

## 1.2 Ictus, fattori meteorologici ed agenti inquinanti

Il tentativo di correlare gli ictus agli eventi meteorologici e agli agenti inquinanti non è totalmente nuovo in letteratura. Esistono infatti diverse pubblicazioni che trattano l'argomento seppur le conclusioni siano talvolta discordanti.

---

Una conclusione definitiva sembra non essere ancora stata tratta e seppur nella maggior parte delle pubblicazioni venga stimato un aumento del rischio relativo, i risultati variano anche considerevolmente in base al luogo e al clima in cui sono condotte le indagini.

### 1.2.1 Lo stato dell'arte

Si è proceduto ad un'attenta analisi della letteratura in merito al fenomeno ictus relazionato in particolar modo alle condizioni meteorologiche e agli agenti inquinanti.

Nella tabella seguente sono riportate le pubblicazioni oggetto di analisi:

Tabella 2 Pubblicazioni sul tema ictus note in letteratura

TITOLO	AUTORI	ANNO
Air Pollution A New Risk Factor in Ischemic Stroke Mortality	Hong Y. et al.	2002
Evidence for an Association Between Air Pollution and daily Stroke Admissions in Kaohsiung, Taiwan	Tsai et al.	2003
Air Pollution and Hospital Admissions for Ischemic and Hemorrhagic Stroke Among Medicare Beneficiaries	Gregory A. et al.	2005
Ambient Air Pollution and Risk for Ischemic Stroke: A Short-Term Exposure Assessment in South China	Guo P. et al.	2017
Ambient Temperature and Stroke Risk Evidence Supporting a Short-Term Effect at a Population Level from Acute Environmental Exposures	Lavados Pablo M. et al.	2017
Air Pollution and Stroke	Ken K. et al.	2018

Di seguito sono riportati, per maggior chiarezza, i contenuti delle pubblicazioni sopra elencate:

---

**[Hong Y. et al.]:** sono stati analizzati gli ictus registrati nel periodo 1991-1997 a Seoul, Korea, che hanno portato alla morte del paziente, riconosciuta in base al codice ICD (International Classification of Diseases), in relazione a ciascuno degli inquinanti presentati (particolato sospeso totale, SO<sub>2</sub>, NO<sub>2</sub>, O<sub>3</sub> e CO) verificando se sussistessero trend temporali e giornalieri. I fattori meteorologici presi in considerazione sono la temperatura, l'umidità relativa e la pressione atmosferica. I risultati riportano incrementi di rischio relativo rapportati ad incrementi inter-quartile di particolato sospeso totale (TSP) e SO<sub>2</sub> nello stesso giorno e ad incrementi inter-quartile di NO<sub>2</sub> e CO nelle 24 ore precedenti e O<sub>3</sub> nei tre giorni precedenti.

**[Tsai et al.]:** sono stati analizzati i dati sulle ammissioni in ospedale dovute a stroke nel periodo 1997-2000 a Kaohsiung, Taiwan. Nell'analisi a singolo inquinante si è osservato che nei giorni caldi, ossia con temperatura superiore ai 20°C, si riscontravano associazioni positive tra l'insorgenza di ictus ed i valori di PM<sub>10</sub>, NO<sub>2</sub>, SO<sub>2</sub>, CO e O<sub>3</sub>. Nei giorni freddi, ossia con temperature inferiori ai 20°C solo i valori di CO mostrano correlazione. Il modello a due inquinanti registra invece un legame tra le ammissioni in ospedale ed i valori di PM<sub>10</sub> e NO<sub>2</sub> ed un aumento del rischio correlato ad aumenti inter-quartile della coppia di inquinanti. Gli effetti di SO<sub>2</sub>, CO e O<sub>3</sub> si sono rivelati non significativi.

**[Gregory A. et al.]:** è stata valutata l'associazione tra i livelli giornalieri di PM<sub>10</sub>, CO, NO<sub>2</sub> e SO<sub>2</sub> e le ammissioni ospedaliere dovute ad ictus che coinvolgono pazienti di età superiore ai 65 anni in nove città statunitensi in periodi di lunghezza variabile compresi tra il 1986 e il 1999. Si sono valutati gli effetti del PM<sub>10</sub> in ognuna delle città in un arco temporale (lag period) fino a 3 giorni prima dell'insorgere della patologia. Si è valutato che un aumento inter-quartile nei livelli di PM<sub>10</sub> è associabile ad un aumento del rischio di ospedalizzazione nello stesso giorno. Risultati simili sono stati ottenuti anche per quanto riguarda CO, NO<sub>2</sub> e SO<sub>2</sub>. Si suggerisce che serviranno ulteriori studi per verificare i risultati.

**[Guo et al.]:** si è valutata la correlazione tra ictus e agenti inquinanti (PM<sub>2.5</sub>, O<sub>3</sub>, SO<sub>2</sub> e NO<sub>2</sub>) in Cina, nella provincia di Guangzhou, nel biennio 2013-2015. Nel giorno stesso e per un lag period di un giorno si registrano correlazioni tra le ospedalizzazioni e i valori di tutti gli inquinanti presi in esame. Nel modello a due inquinanti risulta prevalente la correlazione tra SO<sub>2</sub> e NO<sub>2</sub>, che porta ad un aumento del rischio di ictus. Risultati simili sono stati constatati per le 24 ore precedenti nel caso di un modello a due inquinanti che comprenda NO<sub>2</sub> e O<sub>3</sub>.

**[Lavados Pablo M. et al.]:** si sono valutate diverse pubblicazioni in materia di ictus correlati ai fattori ambientali e in particolar modo si è basata la review sull'influenza della temperatura. Aumenti di temperatura anche molto contenuti sono messi in relazione ad un

---

---

aumento del rischio e si osserva una maggior rilevanza del fenomeno se correlato alla mortalità indotta nella popolazione di età superiore ai 65 anni. Non si osserva nessuna correlazione tra temperature elevate e l'incidenza di ictus mentre le basse temperature si correlano con un rischio di incidenza superiore, soprattutto se si considera un lag period di due settimane.

**[Ken K. et al.]:** nella review si è valutata l'incidenza di ictus in relazione agli agenti inquinanti più noti quali SO<sub>2</sub>, CO, NO<sub>2</sub>, PM<sub>2.5</sub> e PM<sub>10</sub> e si è osservato che sia nel lungo sia nel breve periodo risultano confermati gli effetti che il PM<sub>2.5</sub> ha notoriamente sull'insorgere di malattie cardiovascolari. Si osserva inoltre una provata incidenza dovuta all'esposizione al restante insieme di inquinanti. Si analizzano infine i possibili effetti biologici che il particolato induce nel corpo umano cercando di spiegare dal punto di vista biologico come l'esposizione possa tramutarsi in un concreto emergere dell'ictus ischemico.

## 1.3 Introduzione ai sistemi GIS e alla health geomatics

L'idea che il luogo in cui un individuo vive possa influenzare la sua salute è sempre stata attuale in medicina. Sin dall'antichità, infatti, si era osservato che alcune malattie avevano la tendenza a manifestarsi in determinati ambienti piuttosto che in altri, identificabili tramite caratteristiche geografiche ben precise (vicinanza alle fonti d'acqua, aree rurali o montane). Uno dei casi più noti di analisi geografica relazionata alla salute umana è legato all'epidemia di colera che colpì Londra nel 1854. Al tempo fu il Dr. John Snow, usando una mappa della città, a riuscire a circoscrivere l'origine dell'epidemia ad un pozzo inquinato semplicemente tracciando in maniera puntuale ognuno dei casi riportati [[https://it.wikipedia.org/wiki/Epidemia\\_di\\_colera\\_a\\_Broad\\_Street\\_del\\_1854](https://it.wikipedia.org/wiki/Epidemia_di_colera_a_Broad_Street_del_1854)].

Oggi giorno l'invecchiamento della popolazione e il conseguente aumento di malattie legate all'anzianità, come l'ictus ci pongono dinanzi a nuove sfide e questo richiede l'impiego di strumenti e tecniche che possano garantire un valore aggiunto rispetto alle metodologie di studio tradizionali.

La health geomatics in particolare, ossia la scienza che studia il legame tra salute e territorio, consente di svolgere un'analisi esaustiva su quelli che sono i fenomeni ambientali in una determinata area geografica ricorrendo alle procedure tipiche della geomatica, relazionando poi queste informazioni spaziali ai dati sulla salute della popolazione.

---

Far ricorso ad una disciplina che sfrutti le tecnologie informatiche nel processo di analisi dei dati ambientali e territoriali è di notevole aiuto in quanto consente di automatizzare processi di modellazione ed interpretazioni di informazioni geolocalizzate, ossia riconducibili a coordinate geografiche ben precise. Lo scopo è quello di fornire una visione d'insieme dei processi in atto sul territorio.

### 1.3.1 Sistemi di coordinate

Lo scopo dei sistemi di coordinate geografiche è quello di mappare al meglio un punto sulla superficie terrestre e il loro impiego si rende necessario poiché la Terra presenta una superficie geoidale non regolare e ciò rende impossibile una sua descrizione matematica precisa.

I sistemi di coordinate costituiscono per tanto delle approssimazioni e come tali non risultano valide ovunque sul pianeta. La necessità di mappare al meglio territori diversi a latitudini e longitudini differenti ha portato alla necessità di sviluppare diversi di questi sistemi, che meglio si adattano alle diverse aree del globo.

I metodi di rappresentazione più diffusi a livello globale sono quello UTM (Universal Transverse of Mercator), figlio di una proiezione pseudo-cilindrica, e quello basato su rappresentazione ellissoïdica espressa tramite la coppia latitudine-longitudine [<https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/>].

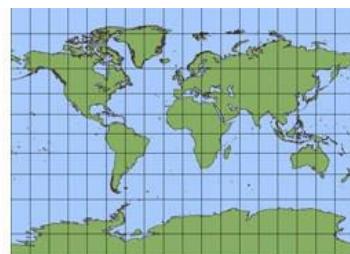


Figura 2 Rappresentazione di una proiezione pseudo-cilindrica  
(<https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/>)

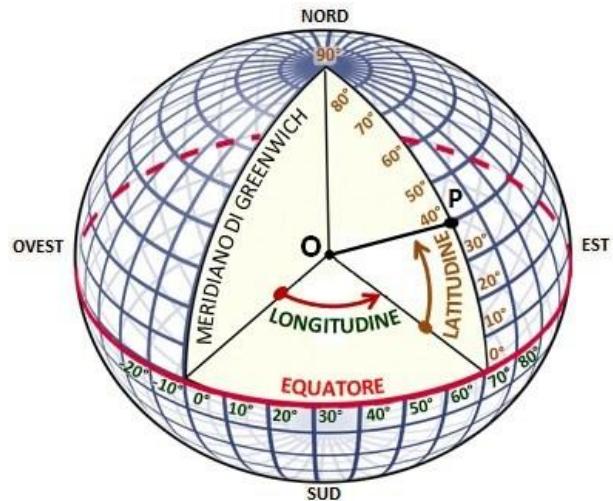


Figura 3 Rappresentazione di una proiezione ellissoïdica  
(<https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/>)

In Italia si utilizzano principalmente due sistemi di coordinate. Il primo è il sistema EPSG:3003, noto anche come proiezione di Gauss-Boaga e specifico per la cartografia italiana. Si tratta di una proiezione di tipo cilindrico che riporta le coordinate di latitudine e longitudine in un piano cartesiano. La proiezione assume il passaggio del fuso zero per la località di Roma Monte Mario. Il secondo usato in Italia è invece UTM, utilizzante una proiezione della Terra che la vede suddivisa in 60 fusi di  $6^\circ$  individuati da un numero progressivo crescente a partire dall'antimeridiano di Greenwich in direzione est. La lettera che segue il numero del fuso indica l'emisfero in considerazione: N per l'emisfero Nord e S per l'emisfero Sud.

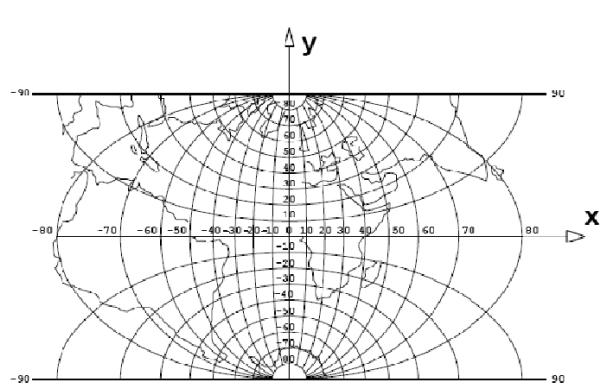


Figura 4 Rappresentazione del sistema cartografico UTM  
[\(https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/\)](https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/)

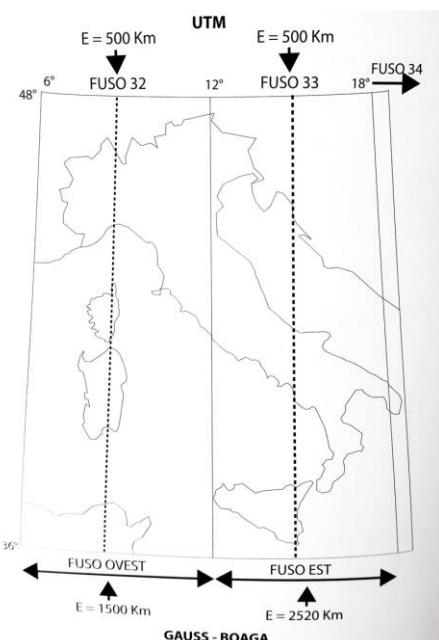


Figura 5 Rappresentazione del suolo italiano e dei fusi che lo attraversano nei sistemi di coordinate UTM e Gauss Boaga (<https://3dmetrica.it/i-codici-epsg/>)

La proiezione utilizzata è quella conforme di Gauss, derivante dalla proiezione di Mercatore, e consente di rappresentare la superficie terrestre su un piano cartesiano. Questa proprietà torna molto utile in quanto consente di calcolare la distanza fra due punti ricorrendo alla semplice distanza euclidea.

---

## **1.4 Interpolazione dei dati**

In Lombardia i dati relativi ai fenomeni atmosferici e agli agenti inquinanti sono rilevati dalle centraline ARPA (Agenzia Regionale per la Protezione Ambientale) sparse sul territorio regionale.

Le stazioni, tuttavia, sono in numero finito e questo rende necessario procedere all'interpolazione spaziale dei dati con l'obiettivo di stimare i valori in prossimità dei punti di interesse.

### **1.4.1 Lo stato dell'arte nelle interpolazioni ambientali**

Lo stato dell'arte per quanto concerne lo studio, l'analisi e le interpolazioni dei dati ambientali in Italia è rappresentato dal sistema ARPA (Agenzia Regionale per la Protezione dell'Ambiente).

ARPA ([http://www.arpalombardia.it/Pages/ARPA\\_Home\\_Page.aspx](http://www.arpalombardia.it/Pages/ARPA_Home_Page.aspx)) si occupa di raccogliere dati riguardanti sia i fenomeni meteorologici sia gli agenti inquinanti in tutta la penisola; questa raccolta avviene mediante l'impiego di stazioni fisse e mobili sparse su tutto il territorio. Questi dati concorrono alla creazione di mappe e previsioni consultabili direttamente sul portale web di ARPA.

Poiché i dati sono raccolti dalle centraline in maniera puntuale, è di fondamentale importanza interpolare correttamente questi valori al fine di creare mappe che coprano l'intero territorio di interesse con valori quanto più accurati possibile.

Il modello adottato da ARPA consiste in una interpolazione spaziale tridimensionale che prende il nome di Optimal Interpolation [Uboldi F. et al. 2007]. I parametri forniti in ingresso al modello sono le coordinate delle centraline e le loro altitudini.

L'algoritmo utilizzato è una versione sviluppata in proprio da ARPA e consiste in un'implementazione ad hoc di Optimal Interpolation. I calcoli per le interpolazioni vengono eseguiti su un server situato nella sede di Milano, che ogni sei ore produce un aggregato dei dati rilevati ed aggiorna le previsioni. Giornalmente si provvede alla rianalisi dei dati e si correggono i modelli creati.

L'output dell'algoritmo consiste in una serie di grid file, ossia file in codifica ASCII rappresentanti una griglia, che coprono l'intero territorio regionale e sono prodotti su base oraria. Le maglie della griglia coprono un'area di 10Km x 10Km per quanto riguarda i dati meteorologici mentre la risoluzione della stessa sale a 4Km x 4Km nel caso dei dati sulla qualità dell'aria.

---

## **1.4.2 Differenze circostanziali e definizione di un modello differente**

Nonostante ARPA disponga già di un modello di interpolazione efficace, gli algoritmi da loro utilizzati fanno uso di librerie personalizzate e soggette a licenza d'uso, nonché ottimizzate appositamente per l'hardware su cui il software viene eseguito. Per questo motivo non è stato possibile far uso di un algoritmo che rappresenti lo stato dell'arte nelle analisi geospaziali.

Sfortunatamente si è reso impossibile anche reperire le mappe già interpolate in quanto disponibili solo per gli anni 2017 e 2018 e solamente per la temperatura. Il ricalcolo delle stesse per gli anni di interesse (2015-2017), dalle prove effettuate, avrebbe richiesto oltre 42 giorni di attività sul server ARPA per la sola temperatura. L'infattibilità dell'operazione e la volontà di non occupare le risorse di calcolo di un sistema pubblico hanno portato alla scelta di agire diversamente.

Si è dunque deciso di far ricorso ad un altro modello di interpolazione noto con il nome di Kriging [Lichtenstern A. 2013, Sunila R. 2015]. L'algoritmo di Kriging viene spesso utilizzato nei software di analisi geospaziale in quanto rappresenta un modo efficace per ottenere interpolazioni formalmente equivalenti [Lorenc, 1986; Daley, 1991] al metodo di Optimal Interpolation.

Gli algoritmi di Kriging possono essere eseguiti in spazi dimensionali a due o tre variabili e in tempi computazionali accettabili in entrambi i casi. L'interpolazione effettuata con questo metodo è tra le più impiegate in ambito geoinformatico e rappresenta il riferimento per le analisi relative alla temperatura [Holdaway M. R. 1996], alla distribuzione delle piogge [S. Ly et al. 2011] e alla diffusione degli agenti inquinanti [Tyagi A. et al. 2013].

Essendo un metodo molto diffuso è anche ben documentato e consente lo sfruttamento di librerie software già esistenti; per queste ragioni e data la comprovata efficacia della tecnica, si è deciso di usarla come strumento di interpolazione all'interno di questa analisi.

L'ostacolo principale all'implementazione di un Kriging tridimensionale è dato però dalla natura dei dati da analizzare, che, nel caso trattato in questo lavoro di tesi, contemplano infatti la sola posizione geografica definita dalle coordinate X e Y del luogo in cui si verifica l'ictus, ma mancano della componente altimetrica necessaria ad un'interpolazione tridimensionale.

Per ovviare a tale mancanza si è tentato di ricondurre ogni evento georeferenziato all'interno del database alla propria altitudine facendo ricorso all'API Open Elevation [<https://open-elevation.com/>], un'alternativa open source concorrente al servizio Elevation [<https://developers.google.com/maps/documentation/javascript/examples/elevation-simple>] offerto a pagamento da Google. Il servizio sfrutta la mappa SRTM250 (Shuttle Radar

---

Topology Mission con precisione di 250m) generata dalla NASA, che associa ad ogni coppia di coordinate un valore altimetrico.

La mappa consiste in un file formato .tiff che associa le coordinate nel formato (Longitudine, Latitudine) delle quali si vuole richiedere l'altitudine alle coordinate nel formato (X,Y) dell'immagine. Il colore di ogni pixel nell'immagine rappresenta il valore dell'altitudine nelle date coordinate.

Sfortunatamente la precisione a 250m risulta non sufficiente ad eseguire un'interpolazione corretta, soprattutto in aree ad altro gradiente altimetrico. Inoltre la complessità computazionale dell'algoritmo di Kriging passa da  $O(n^2)$  a  $O(n^3)$ , il che si traduce in un dilatarsi dei tempi d'esecuzione quantificabile in un 11x medio.

## 1.5 Obiettivo del lavoro

L'ipotesi alla base di questo lavoro è che l'integrazione dei dati provenienti da diverse origini, quali open data demografici e variabili ambientali, unitamente con informazioni geolocalizzate relative ad una particolare categoria di eventi clinici, possa fornire un nuovo contenuto informativo in grado di studiare la diffusione geospaziale di un fenomeno clinico, e le sue possibili relazioni con le variabili in gioco.

A partire da tale ipotesi, l'obiettivo specifico di questo lavoro consiste nello sviluppo di un framework di health geomatics che operi tale integrazione, utile allo studio del possibile impatto delle variabili ambientali (meteo e inquinamento) su un evento clinico geolocalizzato. A tale proposito, si applicherà tale framework al database (fornito da AREU) degli ictus in Regione Lombardia per il triennio 2015-2017, al fine di studiare la distribuzione sul territorio e l'incidenza di tale patologia, così come le possibili interazioni con le variabili meteo e di inquinamento disponibili tramite portale dell'ARPA.

# CAPITOLO 2

## MATERIALI E METODI

### 2.1 Il database degli ictus

I dati utilizzati per lo svolgimento di questo progetto di tesi sono stati forniti da AREU (Agenzia Regionale Emergenze Urgenze), ossia l'agenzia che si occupa della gestione delle chiamate di emergenza a livello regionale in Lombardia (<https://www.areu.lombardia.it/>). AREU è infatti in grado di raccogliere dati all'interno del proprio sistema informativo a partire dalla ricezione di una chiamata al 112. Queste informazioni comprendono i dati del paziente, gli orari di dispatch dei mezzi, le coordinate del luogo dell'evento e una serie di altre informazioni meglio descritte in seguito.

I dati a disposizione sono stati anonimizzati e rappresentano un sottoinsieme di quanto presente nel sistema informativo AREU.

#### 2.1.1 L'architettura del sistema informativo di AREU

La base dati detenuta da AREU trova posto all'interno di un sistema informativo ben più ampio.

L'architettura dello stesso è basata su quattro entità fondamentali:

- **Evento:** è un'entità atomica alla quale vengono ricondotte una o più chiamate al 112
- **Missione:** per ogni evento viene avviata una missione, alla quale farà riferimento ogni mezzo coinvolto nell'intervento

- 
- **Tratta:** rappresenta un frammento del tragitto di soccorso che separa la centrale dal paziente e il paziente dall'ospedale e monitora i tempi di interarrivo.
  - **Paziente:** contiene una descrizione anagrafica del paziente e del suo stato di salute pregresso

Alcuni dei campi presenti all'interno del database risultano essere duplicati. La causa è da ricercarsi nel normale processo evolutivo a cui le sorgenti dati vanno in contro nel tempo ed è da associare alla necessità di raccogliere maggiori informazioni sia sui pazienti sia sull'efficienza interna dei processi aziendali, al fine di migliorare la prestazione erogata. Alcuni campi invece, quali ad esempio i tempi di interarrivo, sono stati materializzati al fine di ottimizzare le operazioni di interrogazione a scopo di business intelligence.

### **2.1.2 Metodologia di acquisizione dei dati**

I dati inseriti all'interno del sistema informativo AREU provengono da una molteplicità di fonti. Derivano infatti da un processo di raccolta realizzato durante le due principali fasi dell'intervento:

- Chiamata al 112
- Intervento sul posto dei sanitari

Lo scopo della chiamata al 112 è quello di raccogliere i sintomi del paziente, accettare l'indirizzo ed erogare un codice di intervento.

Spetta poi ai sanitari che si recano sul luogo verificare l'anagrafica, raccogliere la posizione GPS, accettare patologie pregresse ed eventualmente correggere il codice di intervento. I dati sugli intertempi sono calcolati in modo automatico e vengono impiegati per lo più a livello statistico.

### **2.1.3 Descrizione dei campi del database**

Tabella 3 Campi del database AREU

VOCE	TIPO	DESCRIZIONE
AAT	String	Articolazioni Aziendali Territoriali
ANNO	Integer	Anno
MESE	Integer	Mese

---

ID_EMERGENCY	ID	ID dell'evento
ID_EMERG_HOUR_IN_DAY	Integer	Ora dell'evento
CD_SEVERITY_LEVEL	Code	Gravità
CD_EMERG_CLASS	Code	Classificazione
CD_EMERG_CLASS_DTL	Code	Classificazione dettagliata
CD_REASON	Code	Ragione
CD_REASON_DTL	Code	Ragione dettagliata
CD_CALLER	Code	Chiamante
CD_PLACE	Code	Luogo
CD_PROVINCE_ISTAT	Code	Codice ISTAT della provincia
CD_TOWN_ISTAT	Code	Codice ISTAT della città
DS_TOWN	Description	Nome della città
CD_PROVINCE	Code	Codice della provincia
CD_STREET	Code	Codice della strada
DS_STREET	Description	Nome della strada
DS_STREET_TYP	Description	Tipo di strada
CD_TOWN_ISTAT_STREET	Code	Codice ISTAT della strada
CD_ZONE	Code	Codice della zona
CD_INSTITUT_TRANSPO_TO	Code	Codice della zona in cui il paziente è trasportato
CD_INSTITUT_TRANSPO_FROM	Code	Codice della zona da cui il paziente è trasportato
CD_METEO	Code	Condizioni meteo
ID_EMERG_MONTH	Date-Time	AnnoMese dell'evento

---

DT_EMERG_DAY	Date-Time	Giorno dell'emergenza
DT_EMERG_OPEN	Date-Time	Orario di apertura dell'evento
DT_EMERG_OPEN_HH24MI	Date-Time	Ora di apertura dell'evento
DT_EMERG_CLOSE	Date-Time	Orario di chiusura dell'evento
DT_MISSION_OPEN_1	Date-Time	Orario di apertura della missione 1
DT_CAR_START_1	Date-Time	Orario di partenza del primo veicolo
DT_CAR_H_1	Date-Time	Orario in cui il primo veicolo raggiunge l'ospedale
DT_CAR_I_1	Date-Time	Orario in cui il primo veicolo raggiunge il luogo dell'emergenza
VL_PEOPLE	Integer	Numero di persone coinvolte
VL_PATIENT	Integer	Numero di pazienti
VL_PATIENT_ECG	Integer	Flag per ECG fatta al paziente
VL_EMERG_INTERVAL	Integer	Durata dell'evento
VL_MISSION_OPEN_1_INTERVAL	Time	Tempo di allerta del primo veicolo

---

VL_CAR_START_1_INTERVAL	Time	Tempo di partenza del primo veicolo
VL_CAR_H_1_INTERVAL	Time	Tempo di arrivo del primo veicolo in ospedale
VL_CAR_H_1_INTERVAL_HH24_MI	Time	Tempo richiesto per l'arrivo in ospedale
VL_CAR_H_1_INT_MINUTI	Time	Tempo richiesto per raggiungere l'ospedale in minuti
VL_CAR_I_1_INTERVAL	Time	Tempo richiesto per raggiungere il paziente con il primo veicolo
VL_CAR_I_1_INT_HH24_MI	Time	Tempo richiesto per raggiungere il paziente
VL_CAR_I_1_INT_MINUTI	Time	Tempo richiesto per raggiungere il paziente in minuti
VL_AMOUNT_MISSION_MSB	Integer	Numero di MSB coinvolti
VL_AMOUNT_MISSION_MSI	Integer	Numero di MSI coinvolti
VL_AMOUNT_MISSION_MSA	Integer	Numero di MSA coinvolti
DT_CAR_I_1_MSB	Date-Time	Orario in cui il MSB raggiunge il paziente
DT_CAR_I_1_MSI	Date-Time	Orario in cui il MSI raggiunge il paziente

---

DT_CAR_I_1_MSA	Date-Time	Orario in cui il MSA raggiunge il paziente
VL_CAR_I_1_INTERVAL_MSB	Integer	Tempo richiesto dal MSB per raggiungere il paziente
VL_CAR_I_1_INTERVAL_MSI	Integer	Tempo richiesto dal MSI per raggiungere il paziente
VL_CAR_I_1_INTERVAL_MSA	Integer	Tempo richiesto dal MSA per raggiungere il paziente
VL_GEO_X	Coordinate	Coordinata X dell'evento espressa in EPSG:3003
VL_GEO_Y	Coordinate	Coordinata Y dell'evento espressa in EPSG:3003
AC_SESSO	Categorical	Sesso
VL_ETA	Integer	Età
DS_PAT_1	Description	Patologia1 nota di cui il paziente ha sofferto
DS_PAT_2	Description	Patologia2 nota di cui il paziente ha sofferto
ID_CODICE	Code	Codice di gravità dell'evento
ID_CODICE_E	Code	Codice di gravità filtrato dagli operatori

---

OSPEDALE_DENOMINAZIONE	String	Ospedale di destinazione
OSPEDALE_LOCALITA	String	Città dell'ospedale di destinazione
OSPEDALE_REPARTO	String	Reparto dell'ospedale
OSPEDALE_CODICE	Code	Codice dell'ospedale
OSPEDALE_DESCR_SINTETICA	String	Descrizione sintetica dell'ospedale
OSPEDALE_DESCR_COMPLETA	String	Descrizione completa dell'ospedale
OSPEDALE_PROVINCIA	String	Provincia dell'ospedale
DS_REPARTO_R	Descrizione	Descrizione del reparto
DT_APERTURA	Date-Time	Orario di apertura della missione
ID_CODICE_R	Code	Codice di ritorno
ID_CODICE_E2	Code	Codice di emergenza
ID_PAT1_LIV1	Code	Codice della patologia1 livello1 nota al paziente
ID_PAT1_LIV2	Code	Codice della patologia2 livello2 nota al paziente
ID_PAT2_LIV1	Code	Codice della patologia2 livello1 nota al paziente

---

ID_PAT2_LIV2	Code	Codice della patologia2 livello2 nota al paziente
DT_COMPARSA_EV	Date-Time	Orario di manifestazione dei sintomi
DT_REGISTRAZ_EV	Date-Time	Orario di registrazione dell'evento
VL_CAR_H_1_COMPARSA_EV_MIN	Time	Tempo trascorso dall'insorgenza dei sintomi e l'arrivo del primo mezzo di soccorso
codice ictus	Code	Attribuzione codice ictus

## 2.2 Dati misurati: fattori meteorologici e agenti inquinanti

I dati delle misurazioni sono stati prelevati dalla repository open data della regione Lombardia all'indirizzo web <https://dati.lombardia.it/>. Sfruttando lo strumento di ricerca integrato è possibile individuare i dataset relativi ai “Dati sensori meteo” per ciascuno degli anni interessati dagli eventi di ictus, che in questo caso coprono un arco temporale che si estende dal 2015 fino al 2018.

Volendo svolgere un'analisi esaustiva, si è deciso di non includere nel processo di ricerca i dati relativi agli eventi rilevati nell'arco del corrente anno. La fonte dati AREU fornita, infatti, contiene le occorrenze di ictus rilevate fino al giorno 29/05/2018, data in cui è stata estratta la vista dal database.

Non essendo in grado di coprire interamente l'anno in corso, si è preferito restringere di 5 mesi l'orizzonte temporale a cui i dati fanno riferimento al fine di poter invece osservare con completezza gli anni 2015, 2016 e 2017, che essendo coperti interamente da gennaio a

---

dicembre, consentono di svolgere un’analisi esaustiva circa trend, periodicità e incidenza degli eventi oggetto di indagine.

I dati raccolti tramite il sito open data contemplano anche i campionamenti relativi all’anno 2014 in quanto le analisi sugli agenti inquinanti e i fattori meteo possono interessare, come suggerito in letteratura, un lag period ben superiore all’unità giornaliera. Al fine di poter avere dati anche per i primi giorni dell’anno 2015, è stato utilizzato anche il dataset relativo al 2014, i cui record sono stati conservati unicamente e relativamente agli ultimi 5 giorni dell’anno, essendo 5 giorni il lag period massimo suggerito dall’analisi dello stato dell’arte in materia entro cui siano mai state osservate correlazioni significative.

In alcuni casi i dati sono risultati mancanti per una o più centraline in un dato istante temporale. In tali circostanze non si è proceduto all’induzione sul valore del dato ma si è semplicemente ignorata l’assenza, facendo affidamento sull’esistenza dei valori di tutte le restanti stazioni. La metodologia è stata suggerita dal modus operandi seguito da ARPA, che agisce nel medesimo modo e sfrutta le centraline attive per soppiare alla mancanza puntuale di misurazioni.

Tra le variabili rilevate dalle centraline si è deciso di scartare quelle relative ai fenomeni le cui misurazioni non risultavano correlabili spazialmente tra di loro, impedendo di fatto la creazione di un modello di interpolazione.

I dati sono infatti salvati nel seguente formato:

Tabella 4 Formato dei dati salvati nel database ARPA

NOME CAMPO	TIPO	SIGNIFICATO
IdSensore	String	Id sensore della stazione
Data	Date	Data
Valore	Integer	Valore misurato dal sensore
Stato	Categorical	Misura valida o no
idOperatore	Categorical	1 = valore medio 3 = valore massimo 4 = valore cumulativo

Nel caso in cui un dato risulti non disponibile, il suo valore è equivalente a -999 e lo Stato viene posto a “NA” per indicarne la non validità.

Per poter associare le rilevazioni alle centraline e alla posizione delle stesse, è necessario scaricare il file all’indirizzo web <https://www.dati.lombardia.it/Ambiente/Stazioni->

---

[Meteorologiche/nf78-nj6b](#), per quanto riguarda le stazioni meteorologiche, oppure <https://www.dati.lombardia.it/Ambiente/Stazioni-qualit-dell-aria/ib47-atvt> per quanto riguarda le stazioni sulla qualità dell'aria. Questo file contiene i seguenti campi associati ad ogni centralina presente in Lombardia:

Tabella 5 Campi associati ad ogni centralina gestita da ARPA

NOME CAMPO	TIPO	SIGNIFICATO
IdSensore	String	Id sensore della stazione
Tipologia	String	Tipo di misura rilevata
Unità DiMisura	String	Unità di misura dei valori rilevati
IdStazione	String	Id della stazione
NomeStazione	String	Nome della stazione
Quota	Integer	Altitudine della stazione
Provincia	Categorical	Provincia
DataStart	Date	Data di inizio del campionamento
DataStop	Date	Data di fine del campionamento
Storico	Categorical	N = ancora misurato S = la stazione non è più attiva
UTM_Nord	Coordinate	Coordinata UTM 32N Nord
UTM_EST	Coordinate	Coordinata UTM 32N Est
Lng	Coordinate	Longitudine
Lat	Coordinate	Latitudine
Location	Position	Posizione GPS

Utilizzando questo file è possibile ricavare la descrizione delle centraline ed è possibile ricondurre ogni misurazione effettuata dai sensori alla stazione di appartenenza e soprattutto alla posizione della stessa grazie ad una semplice operazione di join sul campo IdSensore, presente in entrambe le tabelle. In questo modo si è in grado di ricostruire l'intero database

---

ARPA e per ogni istante temporale si è in grado di risalire alle misurazioni di interesse, associandole ad una posizione geografica ben precisa. Ciò sarà utile in seguito nella generazione di un modello di interpolazione che possa associare ad ogni evento di ictus i campi necessari nella posizione dell'intervento.

E' bene notare che una centralina può disporre di più di un sensore al suo interno.

## 2.3 Dataset delle stazioni meteorologiche

Per quanto riguarda i dati meteorologici rilevati dai sensori ARPA ci si è trovati dinanzi alle seguenti misurazioni:

Tabella 6 Misure rilevate dalle stazioni meteorologiche ARPA

TIPOLOGIA	UNITA' DI MISURA
Radiazione Globale	W/m <sup>2</sup>
Altezza Neve	cm
Temperatura	°C
Precipitazione	mm
Livello Idrometrico	cm
Velocità Vento	m/s
Direzione Vento	°
Umidità Relativa	%

Le variabili atmosferiche su cui si è deciso di basare il processo di analisi sono temperatura e umidità relativa, entrambe interpolabili ed entrambe oggetto di interesse in letteratura.

### 2.3.1 Distribuzione delle stazioni

In questo capitolo viene analizzata la ripartizione delle centraline meteorologiche sul territorio lombardo al fine di quantificarne la distribuzione e trarre in prima approssimazione delle conclusioni circa la significatività delle misurazioni nelle diverse aree della regione.

Di seguito sono riportate le misure prese in esame.

---

## **Temperatura:**

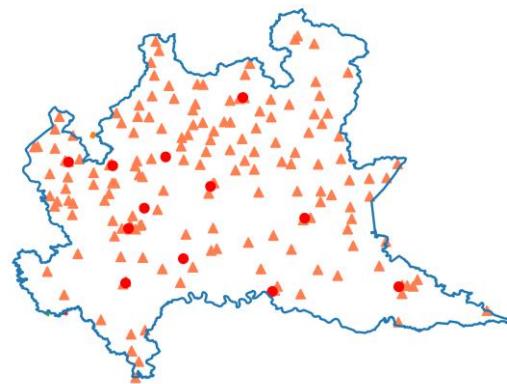


Figura 6 Centraline rilevanti la temperatura in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 163 centraline in grado di rilevare la temperatura.

Si può osservare una capillare distribuzione delle stesse in tutta la regione mentre in colore rosso si riportano i capoluoghi delle 12 province.

---

## **Umidità relativa:**



Figura 7 Centraline rilevanti l'umidità relativa in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 122 centraline in grado di rilevare l'umidità relativa.

Si può osservare una discreta capillarità nella distribuzione delle stesse in tutta la regione mentre in colore rosso si riportano i capoluoghi delle 12 province.

### **2.3.2 Intervallo di campionamento**

Per quanto riguarda le misurazioni di carattere meteorologico usate alla base dell'analisi, si osservano le seguenti frequenze di campionamento dei dati:

- **Temperatura:** record raccolti ogni 10 minuti
- **Umidità Relativa:** record raccolti ogni 10 minuti

Com'è possibile notare, i dati vengono raccolti in maniera continua e risultano disponibili con un'elevata granularità durante tutto l'arco della giornata.

---

## 2.4 Dataset delle stazioni per la qualità dell'aria

Per quanto riguarda i dati sugli agenti inquinanti rilevati dai sensori ARPA ci si è trovati dinanzi alle seguenti misurazioni:

Tabella 7 Misure rilevate dalle stazioni per la qualità dell'aria ARPA

TIPOLOGIA	UNITA' DI MISURA
Ammoniaca	$\mu\text{g}/\text{m}^3$
Arsenico	$\text{ng}/\text{m}^3$
Benzene	$\mu\text{g}/\text{m}^3$
Benzo(a)pirene	$\text{ng}/\text{m}^3$
Biossido di Azoto	$\mu\text{g}/\text{m}^3$
Biossido di Zolfo	$\mu\text{g}/\text{m}^3$
Black Carbon	$\mu\text{g}/\text{m}^3$
Cadmio	$\text{ng}/\text{m}^3$
Monossido di Carbonio	$\text{mg}/\text{m}^3$
Nikel	$\text{ng}/\text{m}^3$
Ossidi di Azoto	$\mu\text{g}/\text{m}^3$
Ozono	$\mu\text{g}/\text{m}^3$
Particelle sospese PM2.5	$\mu\text{g}/\text{m}^3$
Particolato Totale Sospeso	$\mu\text{g}/\text{m}^3$
Piombo	$\text{ng}/\text{m}^3$
PM10 (SM2005)	$\mu\text{g}/\text{m}^3$

Le variabili su cui si è deciso di basare il processo di analisi sono benzene, biossido di azoto, monossido di carbonio, ossidi di azoto, ozono, PM2.5 e PM10. Tutti questi fenomeni risultano interpolabili ed oggetto di interesse in letteratura.

---

## 2.4.1 Distribuzione delle stazioni

In questo capitolo viene analizzata la ripartizione delle centraline per la qualità dell'aria sul territorio lombardo al fine di quantificarne la distribuzione e trarre in prima approssimazione delle conclusioni circa la significatività delle misurazioni nelle diverse aree della regione.

Di seguito sono riportate le misure prese in esame.

### **Ozono:**

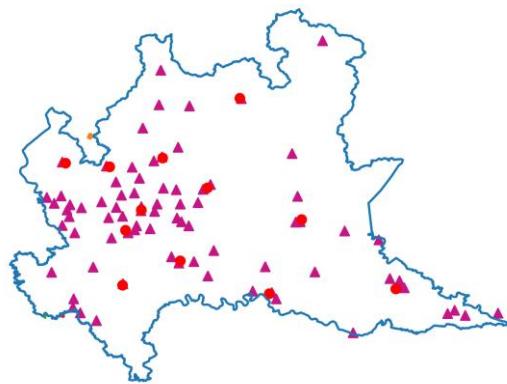


Figura 8 Centraline rilevanti ozono in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 79 centraline in grado di rilevare il livello di ozono.

Si può osservare una distribuzione circoscritta alle aree densamente popolate. In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

## **Biossido di Azoto:**

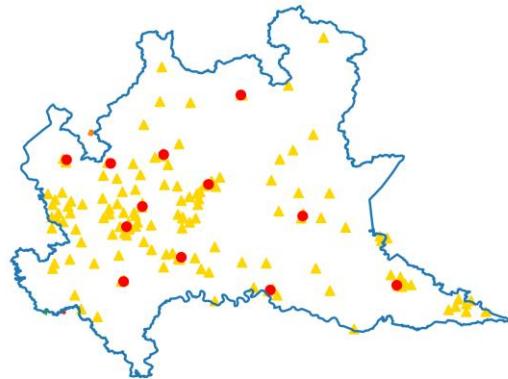


Figura 9 Centraline rilevanti biossido di azoto in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 135 centraline in grado di rilevare il biossido di azoto ( $\text{NO}_2$ ).

Si può osservare una discreta capillarità nella distribuzione delle stesse in tutta la regione, con particolare concentrazione individuabile nelle aree a maggior densità abitativa.

In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

## **Ossidi di Azoto:**

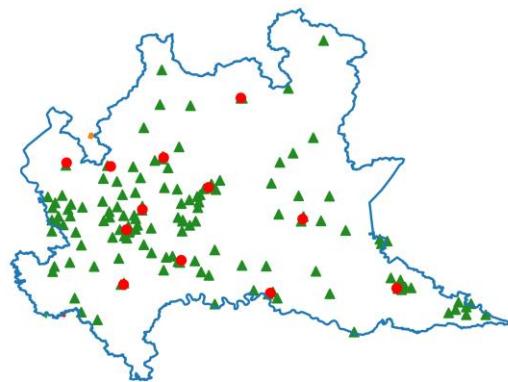


Figura 10 Centraline rilevanti ossidi di azoto in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 134 centraline in grado di rilevare gli ossidi di azoto ( $\text{NO}_x$ ).

Si può osservare una discreta capillarità nella distribuzione delle stesse in tutta la regione, con particolare concentrazione individuabile nelle aree a maggior densità abitativa.

In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

## **Monossido di Carbonio:**

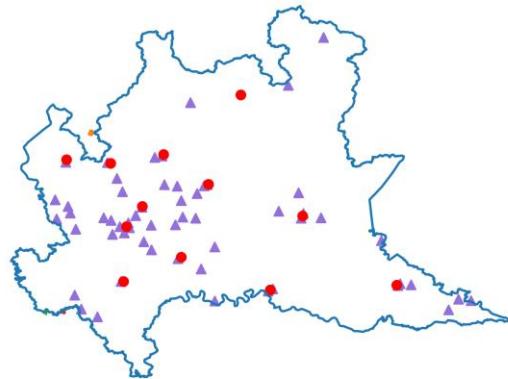


Figura 11 Centraline rilevanti monossido di carbonio in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 57 centraline in grado di rilevare il livello di monossido di carbonio (CO).

Si può osservare una distribuzione circoscritta alle sole aree densamente popolate. In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

**PM<sub>10</sub>:**

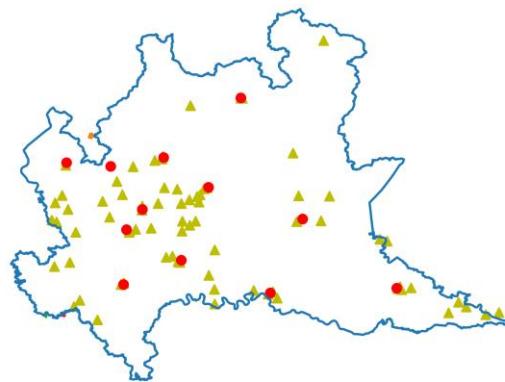


Figura 12 Centraline rilevanti PM<sub>10</sub> in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 74 centraline in grado di rilevare il livello di PM10.

Si può osservare una distribuzione circoscritta alle sole aree densamente popolate. In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

## **PM<sub>2.5</sub>:**

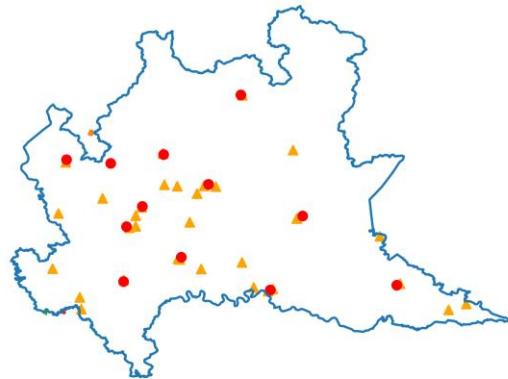


Figura 13 Centraline rilevanti PM<sub>2.5</sub> in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 32 centraline in grado di rilevare il livello di PM<sub>2.5</sub>.

Si può osservare una distribuzione decisamente non capillare e incentrata in alcuni punti di interesse rilevante, quali i grandi centri abitativi della regione. In colore rosso sono rappresentati i 12 capoluoghi di provincia.

---

## **Benzene:**

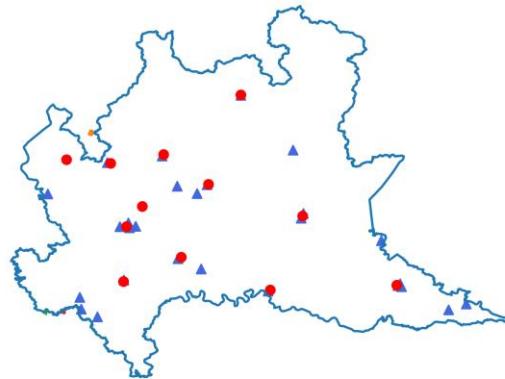


Figura 14 Centraline rilevanti benzene in Lombardia identificate con il simbolo di triangolo. I capoluoghi di provincia sono rappresentati con un pallino rosso

La figura rappresenta la distribuzione sul territorio della Lombardia delle 27 centraline in grado di rilevare il livello di Benzene.

Si può osservare una distribuzione assolutamente non capillare e incentrata in alcuni punti di interesse rilevante quali i grandi capoluoghi regionali, rappresentati in rosso sulla mappa.

### **2.4.2 Intervallo di campionamento**

Per quanto riguarda le misurazioni degli agenti inquinanti usate alla base dell'analisi, si osservano le seguenti frequenze di campionamento dei dati, suddivisibili in due categorie:

- Dati continui:
  - Monossido di carbonio: record raccolti ogni 10 minuti
  - Ozono: record raccolti ogni 10 minuti
  - Ossidi di azoto: record raccolti ogni 10 minuti
  - Biossido di azoto: record raccolti ogni 10 minuti

- 
- Dati giornalieri:
    - Benzene: record basati su media giornaliera
    - PM10: record basati su media giornaliera
    - PM2.5: record basati su media giornaliera

I dati continui vengono normalmente campionati ad intervalli di 10 minuti, come avviene per i dati meteorologici.

Per i dati giornalieri è invece fornito il dato in forma aggregata e corrispondente alla media giornaliera dell'inquinante. In quest'ultimo caso la misura viene resa disponibile alle ore 00:00 del giorno successivo a quello delle rilevazioni. Non è chiara la scelta che ha portato a questa distinzione di trattamento ma si ipotizza una variabilità limitata dei fenomeni in questione, che dunque ne giustificano l'impiego di una media giornaliera.

## 2.5 Dati demografici

Al fine di comprendere meglio la relazione tra gli eventi di ictus e l'incidenza in relazione al luogo di residenza si è deciso di ricorrere ad un'analisi demografica che prendesse in esame la popolazione residente in Lombardia, in particolar modo nelle province e nei capoluoghi di regione.

Questi dati sono generati da ISTAT (Istituto nazionale di STATistica) e fanno riferimento alla popolazione residente campionata al 1º gennaio di ogni anno.

### 2.5.1 Popolazione residente in ogni provincia

La base dati che si è utilizzata nell'ambito dell'analisi è liberamente scaricabile dal sito web [http://dati.istat.it/Index.aspx?DataSetCode=DCIS\\_POPRES1](http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1). Per ottenere i dati relativi agli anni 2015, 2016 e 2017 è sufficiente impostare il filtro alla voce “Selezione periodo”. Per il download della base dati è necessario richiamare la funzione “Esporta” in cima alla pagina. Dalla prima pagina a cui il link riconduce è possibile ricavare i dati sulla popolazione residente nelle 12 province lombarde mentre selezionando “Regioni e comuni”, “Lombardia” ed agendo sempre sul periodo di interesse, è possibile scaricare una tabella contenente i dati sulla popolazione residente nei vari comuni della regione. E' facile a quel punto estrarre i capoluoghi di regione ed osservarne il bilancio demografico.

I campi di maggior interesse utilizzati nel corso dell'analisi demografica sono:

---

Tabella 8 Campi contenuti nella base dati ISTAT

NOME CAMPO	VALORE	DESCRIZIONE
Provincia	String	Provincia
Maschi	Integer	Numero di maschi
Femmine	Integer	Numero di femmine
Totale	Integer	Popolazione totale

Come verrà approfondito in seguito, le metriche oggetto di indagine risultano dalla comparazione dei casi riportati nel database AREU, identificabili per provincia e sesso, e i valori ISTAT.

## 2.5.2 Popolazione residente in ogni capoluogo

La base dati relativa alla popolazione residente nei capoluoghi di provincia è scaricabile dal sito ISTAT <http://demo.istat.it/pop2015/index3.html> nella sezione dedicata alla regione Lombardia.

Modificando il sottodominio popANNO all'interno dell'URL della fonte e inserendo l'anno di interesse è altresì possibile ricavare i dati per tutti i periodi di interesse.

La fonte dati è organizzata in modo da registrare per ogni comune all'interno del suolo provinciale il numero di residenti suddivisi in base all'età anagrafica, al sesso e allo stato civile. Per l'estrazione dei dati del capoluogo è sufficiente filtrare i record in base alla lista dei comuni.

Di seguito si riportano i campi di maggior interesse presenti nel database:

Tabella 9 Campi relativi ai capoluoghi di provincia

NOME CAMPO	VALORE	DESCRIZIONE
Codice comune	Code	Codice ISTAT della provincia
Denominazione	String	Nome del comune
Età	Integer	Età degli abitanti
Totale Maschi	Integer	Numero totale di maschi per età
Totale Femmine	Integer	Numero totale di femmine per età

---

I dati sopra descritti vengono usati nel corso dell'analisi demografica al fine di valutare l'incidenza del fenomeno ictus in relazione al fattore di rischio noto come fibrillazione atriale, frequentemente associato ad un'età superiore ai 65 anni.

### 2.5.3 Incidenza demografica

Come constatato in precedenza, in un processo di analisi esplorativa dei dati appare di sicuro interesse andare ad osservare come gli eventi siano distribuiti sul territorio al fine di provare ad individuare eventuali pattern o ricorrenze nella loro diffusione.

A tale scopo di è deciso di suddividere i casi di ictus avvenuti sul territorio lombardo in base alla provincia presso cui è stato registrato l'evento. Come si può osservare nell'immagine sottostante, in Lombardia sono presenti 12 province, che nel complesso spartiscono una popolazione regionale di oltre 10 milioni di abitanti.

Di seguito si può apprezzare la mappa riportante i limiti provinciali e i rispettivi capoluoghi:

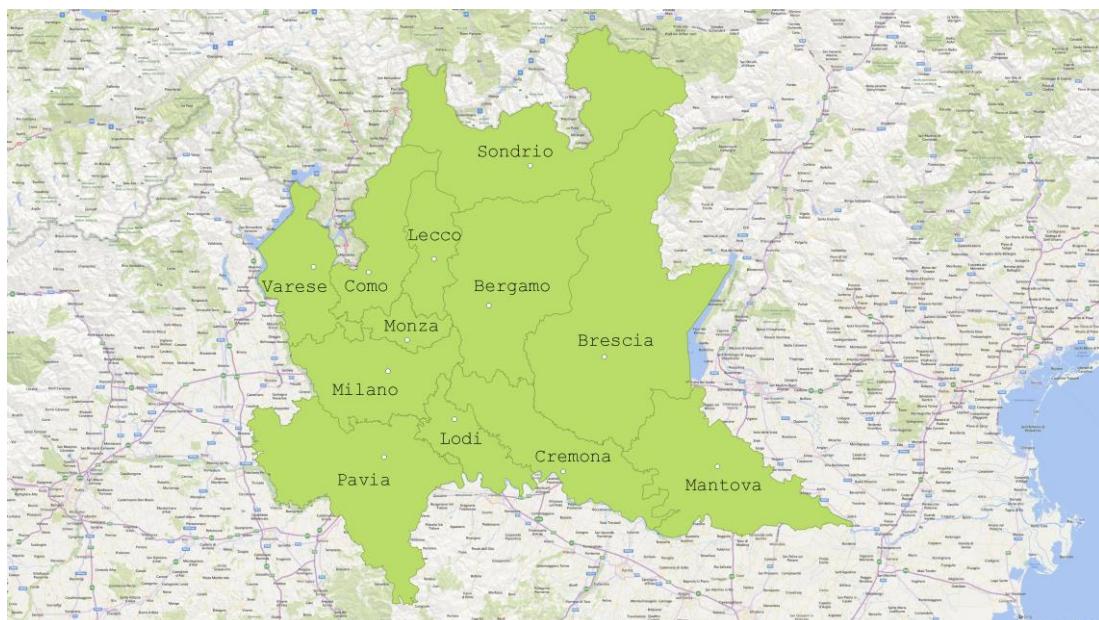


Figura 15 Rappresentazione dei confini provinciali e dei capoluoghi della Lombardia

Per valutare se esistano province più o meno colpite dal fenomeno si è operato sul database fornito da AREU al fine di assegnare ogni evento al territorio corretto. Il processo è stato facilitato dal fatto che un campo che servisse allo scopo fosse già presente all'interno della base dati.

---

Per prima cosa si è proceduto ad estrarre dai dati ISTAT il valore relativo alla popolazione maschile e femminile residente nelle diverse province lombarde al 1º gennaio di ogni anno dal 2015 al 2017.

In seguito, per ogni anno dal 2015 al 2017 è stata calcolata la popolazione colpita da ictus ischemico associata a tutte e 12 le province.

Si è dapprima discriminato in base al sesso del colpito, dopodiché si sono analizzati i dati anche in forma aggregata. Per alcuni dei record non era disponibile la classificazione in base al sesso dell'individuo, che veniva riportato con la dicitura "N" pari a "non classificato".

Andando infine ad eseguire il rapporto tra individui colpiti da ictus e popolazione residente in una data provincia si è giunti al calcolo dell'incidenza del fenomeno, riportata in percentuale.

## 2.6 Pre-elaborazione dei dati

I capitoli seguenti trattano il tema della pre-elaborazione dei dati al fine di renderli consoni ad un'elaborazione automatica.

Verranno descritti i processi di selezione delle coordinate, interpolazione, filtraggio, selezione e aggregazione dei record.

### 2.6.1 GIS e sistemi di coordinate

Con l'acronimo GIS (Geographic Information System) si intende un sistema informatico geografico, ossia una serie di strumenti informatici utilizzabili per effettuare analisi su un territorio.

Si è fatto uso di un software GIS chiamato qGIS nella sua versione 2.18.18 al fine di geolocalizzare su una mappa gli eventi di ictus verificatisi sul territorio della Lombardia e oggetto di indagine. Il software sfrutta i sistemi di coordinate geografiche per creare un modello della superficie terrestre e in base al tipo di coordinate utilizzate, identificabili tramite il loro codice EPSG (European Petroleum Survey Group), è in grado di localizzare con precisione un determinato record georeferenziato.

Al fine di ottenere una consistenza tra i dati e di poter avere uniformità tra gli stessi, le analisi spaziali svolte in questo elaborato sono sempre condotte facendo uso delle coordinate EPSG:32632, riconducibili al sistema UTM (Universal Transverse of Mercator) specifico per l'area 32N del globo terrestre. L'Italia si estende su tre di questi fusi: 32N, 33N e 34N (<http://host154-194-static.207-37-b.business.telecomitalia.it/epsg/NotaSistemiEPSG.pdf>).

Il territorio della Lombardia tuttavia rimane centrato all'interno del fuso 32. Per questa ragione si è fatto uso delle coordinate UTM 32N.

---

## 2.6.2 Kriging

Come osservato in precedenza, le stazioni di rilevazione dell'ARPA sono in numero finito e dislocate in posizioni fisse del territorio. L'interpolazione dei dati si prefigge lo scopo di poter assegnare ad ogni evento di ictus una misurazione attendibile di quelle che fossero le condizioni nel luogo in cui si è verificato.

I fenomeni di cui si desidera ottenere un'interpolazione sono quelli ritenuti di interesse per quanto riguarda meteorologia e qualità dell'aria.

La procedura di interpolazione sfrutta un modello di Kriging bidimensionale e si compone di quattro fasi [Lichtenstern A. 2013]:

- Generazione del semivariogramma
- Fitting del modello sul semivariogramma
- Generazione del modello di interpolazione
- Interpolazione dei dati di input

## 2.6.3 Semivariogramma

Lo scopo del semivariogramma è quello di mostrare le regole di dipendenza che legano fenomeni misurati in maniera puntuale e frutto del campionamento di una superficie, mettendo in evidenza la loro autocorrelazione spaziale.

La dipendenza trovata è una dipendenza di tipo statistico e si basa sull'identificazione di una funzione di covarianza da utilizzare per determinare l'autocorrelazione spaziale fra due punti, ossia la relazione che lega i valori di campionamenti differenti in relazione alla loro distanza, al fine di poter estendere la conoscenza del fenomeno osservato anche a coordinate prive di campionamento diretto.

Il semivariogramma altro non è che la versione riportata su grafico della semivarianza calcolata per ogni coppia di punti presso cui un valore è campionato nello stesso istante e si basa sulla formula:

$$\text{Semivariogram}(\text{distance}_h) = 0.5 * \text{average}((\text{value}_i - \text{value}_j)^2)$$

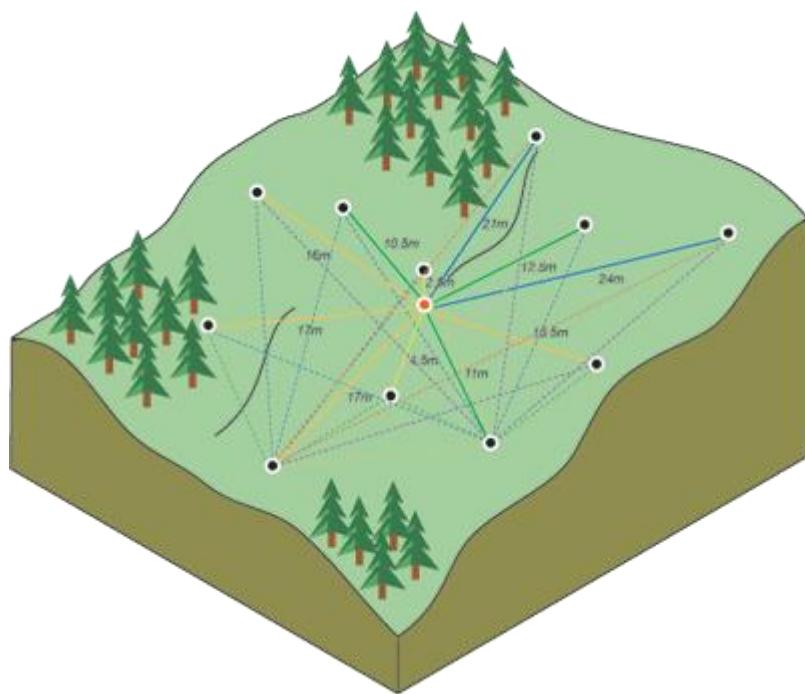


Figura 16 Rappresentazione grafica del calcolo della semivarianza  
<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>

Spesso ogni coppia di punti risulta avere tuttavia una distanza univoca. La generazione di un semivariogramma completo per ognuna delle coppie di punti potrebbe quindi risultare esosa dal punto di vista computazionale ed ingestibile nel caso in cui i punti fossero numerosi. Per questa ragione si usa ricorrere all'impiego di intervalli spaziali che raggruppino valori prossimi delle distanze. A seconda della finezza che si desidera ottenere, della superficie su cui si estendono le misurazioni e del tempo in cui si vuole generare il semivariogramma, si possono scegliere intervalli di 50, 100, 200, 500 o 1000 metri.

La scelta di un intervallo piuttosto che un altro non influisce sul risultato finale, che si presenta come un insieme di punti esprimenti l'evolversi della semivarianza in relazione alla distanza.

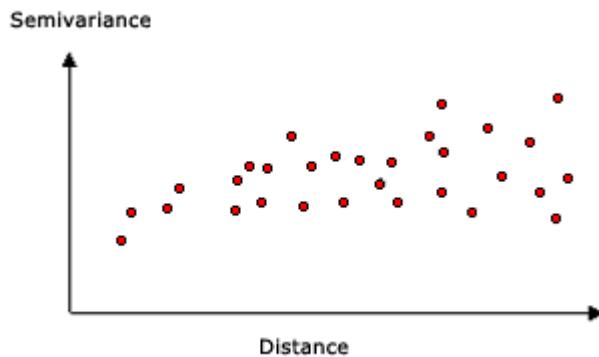


Figura 17 Rappresentazione grafica di un semivariogramma  
[\(<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>\)](http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm)

## 2.6.4 Fitting del semivariogramma

Dopo aver generato il semivariogramma è necessario fare il fitting di un modello che possa descrivere l'andamento del fenomeno. E' un processo cruciale in quanto lega la descrizione spaziale di un fenomeno alla sua predizione. Inoltre, il fitting con un modello continuo fa sì che l'interpolazione successiva sia ricavabile per ogni possibile punto nello spazio.

Una volta noto il modello dell'andamento si sarà in grado di predirlo facendo ricorso alla curva appresa.

Prima di procedere al fitting del modello è però necessario conosce i tre parametri che caratterizzano un semivariogramma:

- **Range:** la distanza entro cui la semivarianza si stabilizza definitivamente. Di norma si assiste ad un assestamento asintotico, per cui si preferisce usare come range il 95% del valore a cui la semivarianza tende
- **Sill:** è il valore assunto dalla semivarianza al raggiungimento del Range
- **Nugget:** è il gap che separa la semivarianza dal valore 0 in coincidenza con l'origine. È un parametro che assume un valore maggiore o uguale a zero.

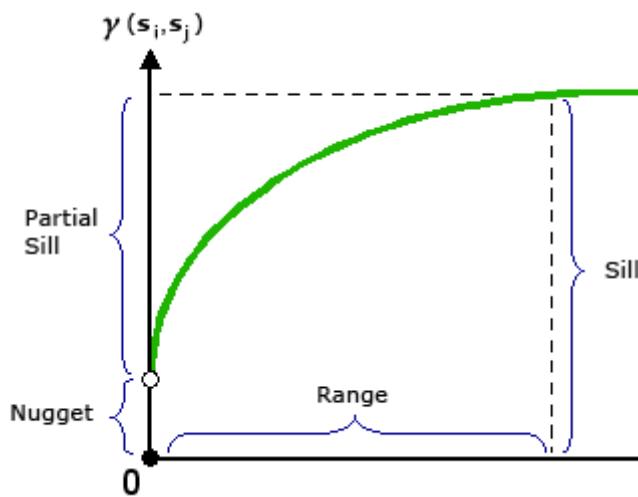


Figura 18 Rappresentazione dei parametri che caratterizzano un semivariogramma  
<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>)

Questi parametri sono cruciali nell'identificazione del modello migliore in quanto aiutano a modellare la funzione di fitting per adattarla al semivariogramma.

Le funzioni normalmente utilizzate sono le seguenti, dove per tutti i modelli ‘r’ e ‘c’ rappresentano rispettivamente Range e Sill:

#### LINEAR

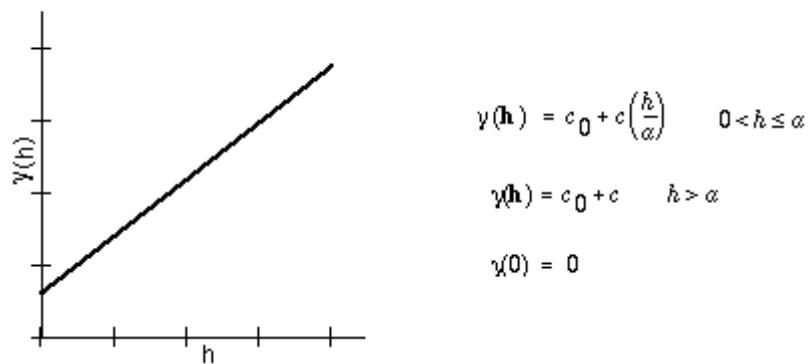


Figura 19 Rappresentazione di un modello di fitting lineare per un semivariogramma  
<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>)

### SPHERICAL

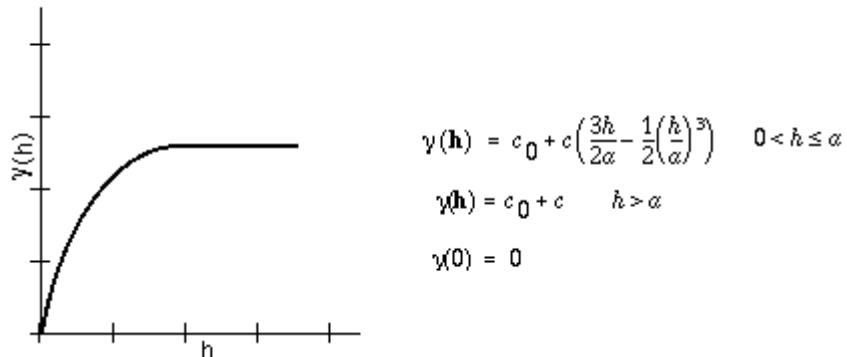


Figura 20 Rappresentazione di un modello di fitting sferico per un semivariogramma  
(<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>)

### EXPONENTIAL

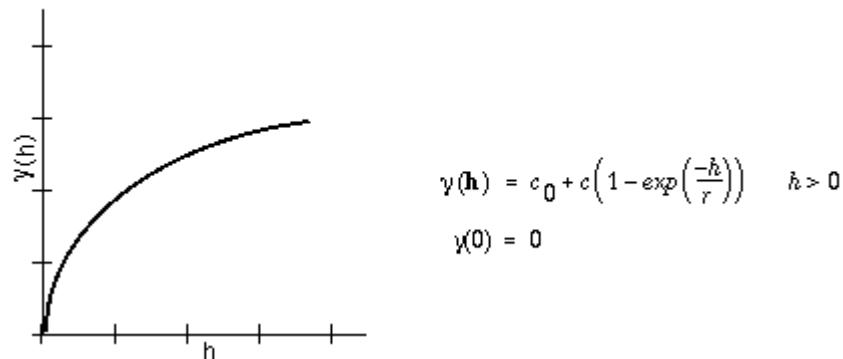


Figura 21 Rappresentazione di un modello di fitting esponenziale per un semivariogramma  
(<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>)

### GAUSSIAN

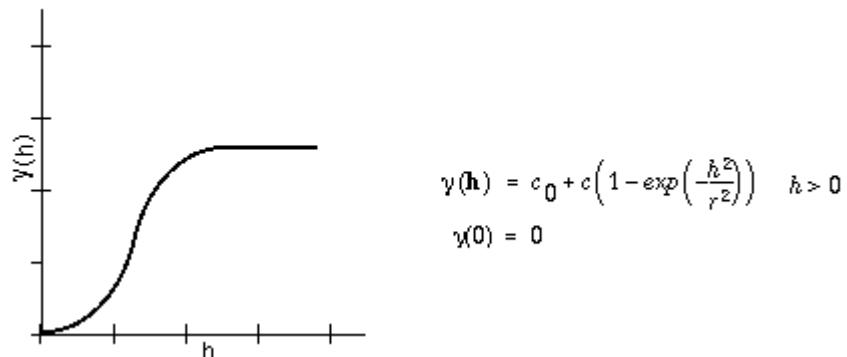


Figura 22 Rappresentazione di un modello di fitting gaussiano per un semivariogramma  
(<http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kriging-works.htm>)

## 2.6.5 Generazione del modello di interpolazione

Dopo aver individuato la funzione che meglio modella l'autocorrelazione spaziale rappresentata dal semivariogramma, si procede con la generazione del modello vero e proprio. Prima di procedere con la descrizione dell'algoritmo, è bene ricordare la relazione che lega varianza e covarianza, come mostrato nella figura sotto riportata (<http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/semivariogram-and-covariance-functions.htm>).

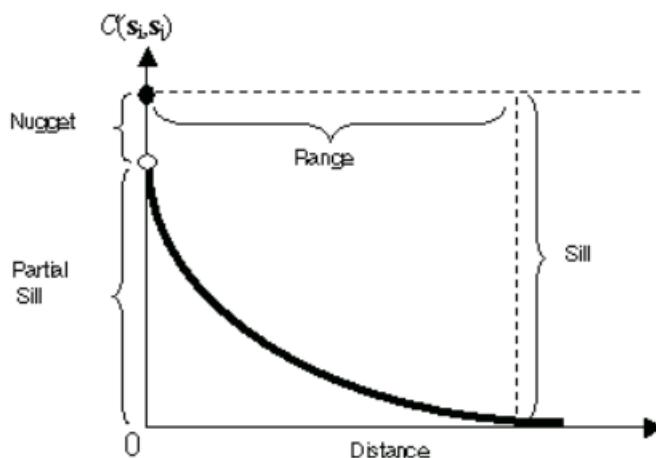


Figura 23 Rappresentazione grafica della relazione che lega varianza e covarianza (<http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/semivariogram-and-covariance-functions.htm>)

Come si osserva nell'immagine, è possibile ricondurre la funzione di covarianza  $C$  ai medesimi parametri Nugget, Sill e Range già osservati nel semivariogramma.

Indicando con  $\gamma$  la funzione di semivarianza, si ricostruisce la seguente relazione:

$$\gamma(S_i, S_j) = \text{Sill} - C(S_i, S_j)$$

Dove  $S_i$  e  $S_j$  rappresentano due punti spaziali, distinti fra loro, all'interno del medesimo piano di interesse.

In figura si riporta il grafico delle due curve:

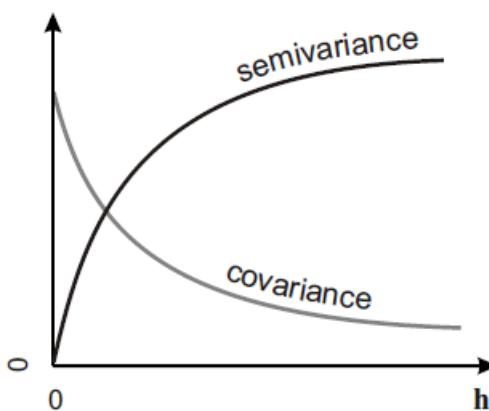


Figura 24 Rappresentazione grafica delle curve di varianza e covarianza  
<http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/semivariogram-and-covariance-functions.htm>

Si consideri ora un processo casuale  $Z$  di cui sono note una serie di realizzazioni  $Z(S_i)$ , con  $i$  che rappresenta il numero di centraline ARPA per cui è disponibile un determinato dato. Generare un modello significa trovare uno stimatore in grado di definire il processo  $Z$  per ogni punto della superficie spaziale in esame.

Le tecniche di interpolazione con metodo Kriging che producono questo stimatore sono principalmente due: Ordinary Kriging e Universal Kriging (Tomislav H. 2007).

Il primo modello sia basa sull'ipotesi che la media del fenomeno analizzato sia stazionaria e non conosciuta mentre il secondo si basa sull'assunzione che nel territorio in analisi sussista un trend generalizzato che invalidi l'ipotesi di stazionarietà.

Universal Kriging rappresenta dunque una generalizzazione di Ordinary Kriging e presuppone l'esistenza di una media costante alla quale si sommi una componente di deriva variabile che caratterizzi il trend. E' inoltre previsto che la media risulti essere una combinazione lineare di tutti i campionamenti sul territorio.

Come è possibile intuire, questa seconda tecnica di Kriging pone dei requisiti piuttosto stringenti circa la natura dei fenomeni, spesso influenzati da una morfologia variabile del territorio, e risulta applicabile solo in scenari contraddistinti da un effettivo trend spaziale. Queste ipotesi, inoltre, fanno sì che il metodo di Universal Kriging mal si adatti all'applicazione in vasti territori e mostri invece i suoi pregi nell'interpolazione di fenomeni più circoscritti.

Il suolo lombardo è un mix tra montagna, collina e pianura, con ambienti e microclimi alpini, lacustri e rurali molto differenti tra di loro. Questa diversità climatica e morfologica rende inapplicabile una tecnica basata su un trend generalizzato all'intera regione ed è per questo

---

motivo che si è deciso di utilizzare un modello di Ordinary Kriging nell'interpolazione dei dati sul territorio.

Come accennato in precedenza, Ordinary Kriging si basa sull'assunzione che la media  $\mu$  del processo Z sia ignota e stazionaria, ossia costante. L'algoritmo di interpolazione fa ricorso ad una media pesata tra le misure campionate in modo da ricondurre il valore interpolato nel punto  $s_0$  ad una combinazione lineare dei dati a disposizione nelle centraline site nei punti i:

$$\hat{z}_{OK}(s_0) = \sum_{i=1}^n w_i(s_0) \cdot z(s_i) = \lambda_0^T \cdot \mathbf{z}$$

Il vettore dei moltiplicatori  $\lambda_0$  prende il nome di vettore dei coefficienti e costituisce l'incognita del problema. Creare un modello di Kriging significa calcolare tale vettore dei coefficienti, la cui definizione è assegnata alla formula:

$$\lambda_0 = \mathbf{C}^{-1} \cdot \mathbf{c}_0$$

$\mathbf{C}$  rappresenta la matrice di covarianza derivata dalle osservazioni puntuali in  $S_i$  mentre  $\mathbf{c}_0$  esprime il vettore della covarianza rispetto ai campionamenti nel nuovo punto  $s_0$ . Espandendo la notazione si ottiene:

$$\begin{bmatrix} C(s_1, s_1) & \cdots & C(s_1, s_n) & 1 \\ \vdots & & \vdots & \vdots \\ C(s_n, s_1) & \cdots & C(s_n, s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(s_0, s_1) \\ \vdots \\ C(s_0, s_n) \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(s_0) \\ \vdots \\ w_n(s_0) \\ \varphi \end{bmatrix}$$

Si noti che sono state aggiunte una colonna e una riga alla matrice  $\mathbf{C}$  al fine di rendere la somma dei coefficienti uguale a 1.  $\varphi$  rappresenta invece il moltiplicatore di Lagrange.

La covarianza tra due punti della superficie è ricavabile dal semivariogramma secondo la relazione sopra descritta. Il calcolo dei coefficienti è il risultato di un semplice prodotto matriciale.

Essendo il metodo di Kriging una predizione spaziale di tipo statistico, offre la possibilità di associare alla misura interpolata un livello di incertezza sotto forma di varianza, che viene calcolata come:

$$\begin{aligned} \hat{\sigma}_{OK}^2(s_0) &= (C_0 + C_1) - \mathbf{c}_0^T \cdot \lambda_0 \\ &= C_0 + C_1 - \sum_{i=1}^n w_i(s_0) \cdot C(s_0, s_i) + \varphi \end{aligned}$$

Come si può osservare dall'equazione, la varianza risulta definita come la somma pesata delle covarianze nel nuovo punto nei confronti dei campionamenti, a cui si somma il moltiplicatore di Lagrange.  $C_0 + C_1$  rappresenta il Sill del semivariogramma, ossia la covarianza spaziale a distanza pari a zero.

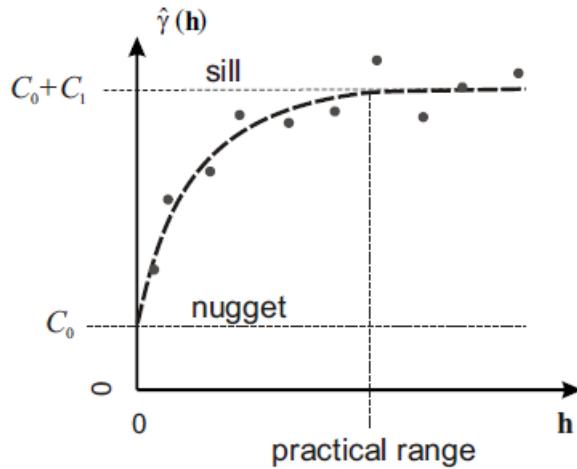


Figura 25 Rappresentazione parametrica di un semivariogramma [Lichtenstern A. 2013]

Applicando un modello di Ordinary Kriging si produce uno stimatore unbiased in quanto l'errore di predizione è in media pari a zero. Questa proprietà è assicurata dal fatto che la somma dei coefficienti sia unitaria ed è verificabile tramite la seguente equazione:

$$\mathbb{E}[Z_{\omega}^*(x_0) - Z(x_0)] = \mathbb{E} \left[ \sum_{i=1}^n \omega_i Z(x_i) - Z(x_0) \underbrace{\sum_{i=1}^n \omega_i}_{=1} \right] = \sum_{i=1}^n \omega_i \underbrace{\mathbb{E}[Z(x_i) - Z(x_0)]}_{=0} = 0$$

L'equazione mostra come la media della differenza tra il valore interpolato  $Z^*\omega(x_0)$  ed il valore reale  $Z(x_0)$ , per un generico punto  $x_0$ , sia pari alla somma pesata degli incrementi  $\mathbb{E}[Z(x_i) - Z(x_0)]$ , che tuttavia è pari a zero a causa dell'ipotesi di stazionarietà alla base del modello di Ordinary Kriging.

Ne consegue quindi che in media l'errore di predizione sia zero e lo stimatore risulti di conseguenza unbiased [Lichtenstern A. 2013].

## 2.6.6 Analisi e validazione del modello

Come osservato in precedenza, la creazione di un modello di kriging tridimensionale risulta particolarmente esosa in termini di risorse computazionali e per giunta il suo impiego non fornisce risultati apprezzabilmente superiori al modello bidimensionale.

Essendo questa una tesi non focalizzata sul metodo di interpolazione ma sulla rilevanza dei risultati, si è deciso di optare per il modello a due dimensioni al fine di rendere sostenibile l'operazione di l'interpolazione estesa ad un lag period di 5 giorni per ognuno dei fattori indagati.

---

Di vitale importanza per la generazione di un modello corretto è l'operazione di fitting sul semivariogramma, in modo che sia possibile ricondurre il trend di autocorrelazione spaziale ad una funzione continua ed utilizzabile dunque per il calcolo dei coefficienti di Kriging.

Questo paragrafo si prefigge lo scopo di entrare nel dettaglio in merito alle scelte adottate nella selezione dei modelli di fitting, analizzando dunque la struttura dei semivariogrammi e misurando quanto le funzioni utilizzate per le interpolazioni descrivano effettivamente i vari fenomeni.

Di seguito si riportano i semivariogrammi generati nel corso dell'analisi. Si sottolinea che al fine di offrire una maggior comprensione dei fenomeni in esame si è deciso di rappresentare la semideviazione anziché la semivarianza, reputando questa prima grandezza come più indicativa e di immediata percezione in quanto ha la stessa unità di misura della variabile rilevata.

Passati in rassegna i grafici di correlazione spaziale si indaga poi quali siano i modelli che offrano il miglior adattamento per ciascuno dei fenomeni indagati.

Come appreso in letteratura (Cao W. et al. 2009, Nur Falah A. et al. 2016), tutti i fenomeni oggetto d'indagine presentano modelli ideali di fitting a carattere esponenziale o sferico, tipici dei fenomeni meteorologici e di diffusione degli agenti inquinanti. In talune situazioni si è voluto andare oltre verificando se non sussistessero modelli di diffusione lineare. L'indagine non ha sortito esito positivo e ha confermato che gli adattamenti sono effettivamente esponenziali o sferici.

Per dare una maggiore comprensione delle scelte effettuate in fase di selezione del modello, si riportano i confronti tra le due funzioni per ognuna delle misure di interesse e si motivano le ragioni che hanno portato all'impiego dell'uno o dell'altro adattamento.

---

## **Stazioni meteorologiche:**

### **Temperatura:**

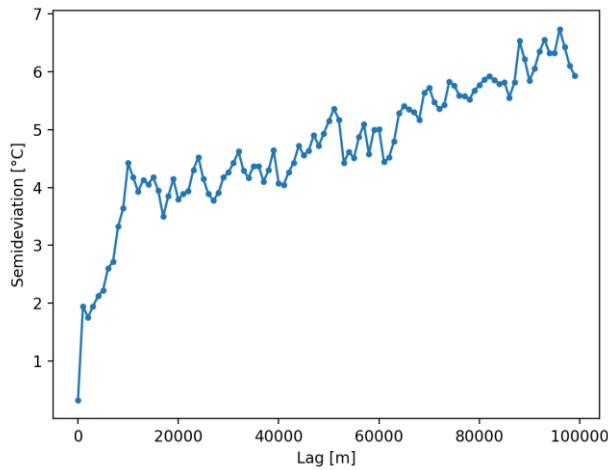


Figura 26 Semivariogramma della temperatura

Nel caso in esame si può osservare una correlazione più stretta tra i valori a minor distanza, com'è lecito aspettarsi. La semideviazione è contenuta a valori prossimi ai 5°C mentre si riduce a circa 2°C nei primi 10Km dalle stazioni di rilevamento.

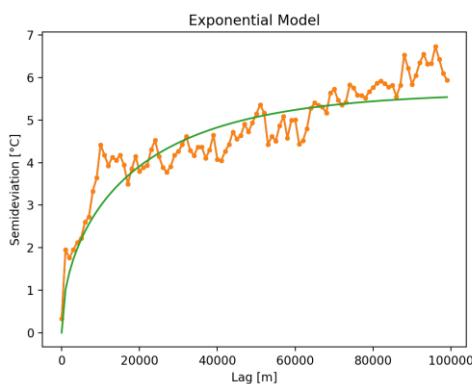


Figura 27 Fitting esponenziale per il semivariogramma della temperatura

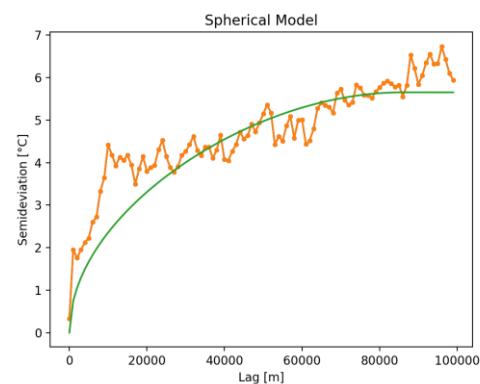


Figura 28 Fitting sferico per il semivariogramma della temperatura

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma della temperatura.

---

Come è possibile osservare il modello esponenziale è in grado di approssimare in maniera migliore il fenomeno in questione offrendo una rappresentazione più significativa soprattutto per quanto concerne distanze inferiori ai 30Km, dopodichè si comporta in maniera assolutamente paragonabile al modello sferico.

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.77 per il modello esponenziale ed uno di 0.8 per il modello sferico. Restringendo invece il semivariogramma ai primi 30Km si calcolano valori di  $R^2$  pari a 0.69 e 0.61 relativamente ai fitting esponenziale e sferico.

Data l'importanza di avere una corretta interpolazione dei dati anche in località densamente coperte da centraline meteorologiche, la maggior capacità approssimativa della funzione esponenziale sulle brevi distanze è risultata una caratteristica fondamentale nella scelta del modello vista l'elevata quantità di centraline in grado di coprire uniformemente il territorio.

Per quanto riguarda le interpolazioni si è per tanto optato per un fitting esponenziale. Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

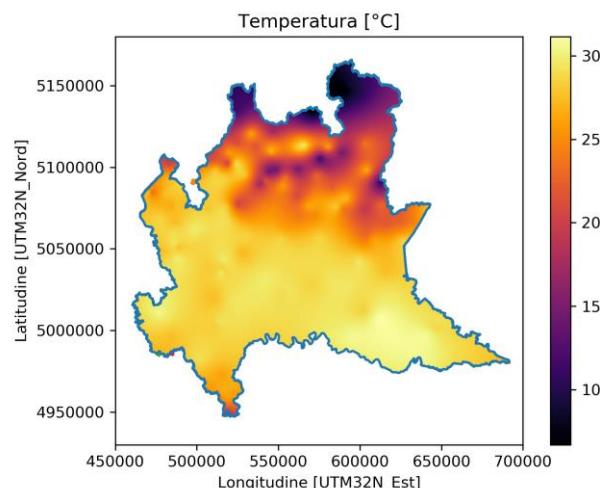


Figura 29 Rappresentazione di una mappa di interpolazione per la temperatura in Lombardia

---

## **Umidità relativa:**

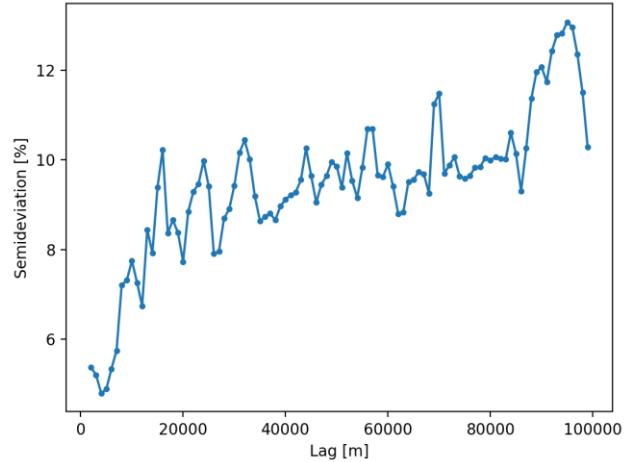


Figura 30 Semivariogramma dell'umidità relativa

L'umidità relativa viene espressa in percentuale e come è possibile vedere nel grafico soprastante si mantiene pressoché costante con variazioni comprese tra l'8% e il 10% una volta superati i 15Km dalle stazioni. Per centraline vicine tra loro si osserva una deviazione standard con crescita esponenziale compresa tra il 6% e l'8%, segno che le misurazioni rilevano tuttavia un fenomeno piuttosto omogeneo.

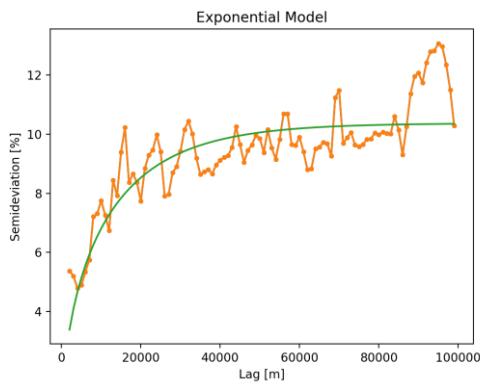


Figura 31 Fitting esponenziale per il semivariogramma dell'umidità relativa

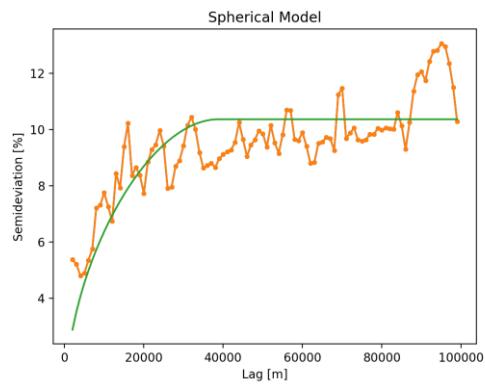


Figura 32 Fitting sferico per il semivariogramma dell'umidità relativa

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma dell'umidità relativa.

Come è possibile osservare il modello esponenziale è in grado di approssimare in maniera migliore il fenomeno in questione offrendo una rappresentazione più significativa sia a brevi

---

distanze sia a medie distanze. Il modello sferico sembra generare una sovrastima del fenomeno e per giunta converge troppo rapidamente al valore di Sill pregiudicando la precisione delle interpolazioni comprese tra i 35Km e i 70Km.

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.56 per il modello esponenziale ed uno di 0.49 per il modello sferico. Restringendo invece il semivariogramma ai primi 30Km si calcolano valori di  $R^2$  pari a 0.70 e 0.67 relativamente ai fitting esponenziale e sferico.

Ottenendo benefici sia impiegando il modello nella sua globalità sia impiegandolo in maniera ristretta alle brevi distanze si può apprezzare la miglior copertura offerta dalla funzione esponenziale, che è stata dunque impiegata nella creazione del modello di previsione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

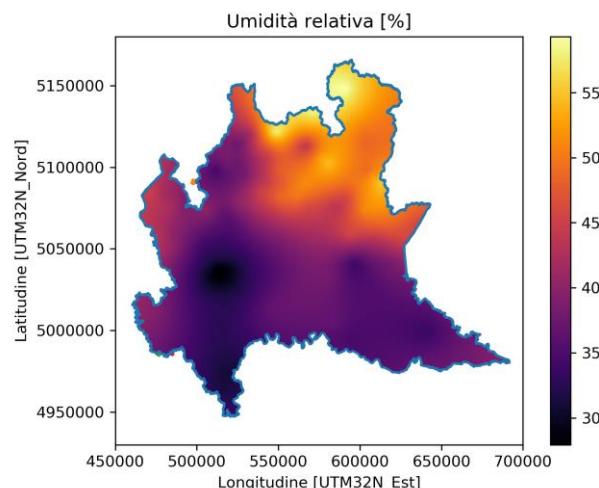


Figura 33 Rappresentazione di una mappa di interpolazione per l'umidità relativa in Lombardia

---

## **Stazioni per la qualità dell'aria:**

### **Ozono:**

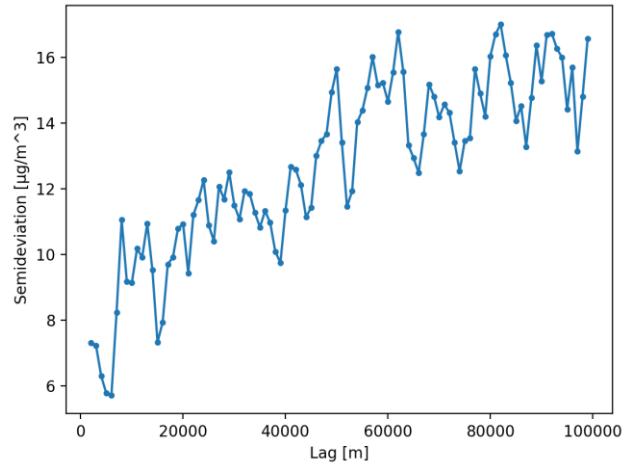


Figura 34 Semivariogramma dell'ozono

La semideviazione delle rilevazioni d'ozono mostra come le centraline vicine misurino dati simili mentre ad elevate distanze la concentrazione dell'inquinante risulta variare molto più rapidamente, in particolare a partire dai 10Km.

L'andamento è di tipo crescente e presenta una tendenza a stabilizzarsi, segno che una decorrelazione interessa il fenomeno a partire dai 60Km dalle sorgenti di rilevazione.

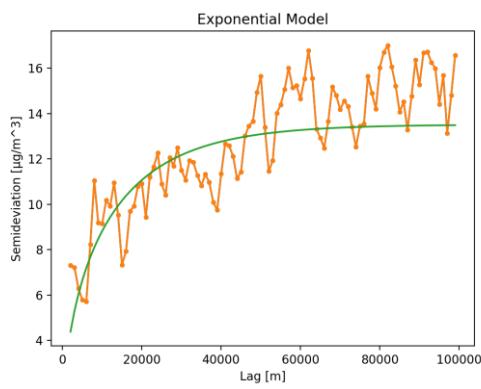


Figura 35 Fitting esponenziale per il semivariogramma dell'ozono

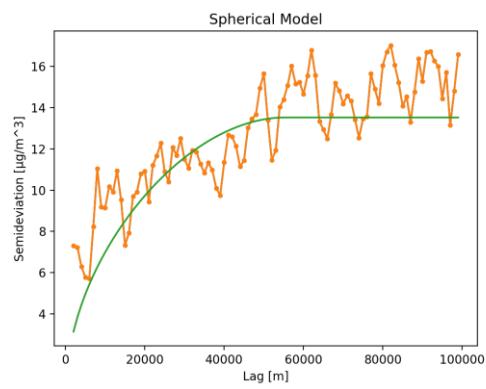


Figura 36 Fitting sferico per il semivariogramma dell'ozono

---

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma dell'ozono.

Come è possibile osservare dai grafici, la sovrapposizione offerta dalle due funzioni risulta molto simile.

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.64 per il modello esponenziale ed uno di 0.65 per il modello sferico. Restringendo invece il semivariogramma ai primi 30Km si calcolano valori di  $R^2$  pari a 0.65 e 0.66 relativamente ai fitting esponenziale e sferico.

Ottenendo benefici sia impiegando il modello nella sua globalità sia impiegandolo in maniera ristretta alle brevi distanze si può apprezzare la miglior copertura offerta dalla funzione sferica, che è stata dunque impiegata nella creazione del modello di previsione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

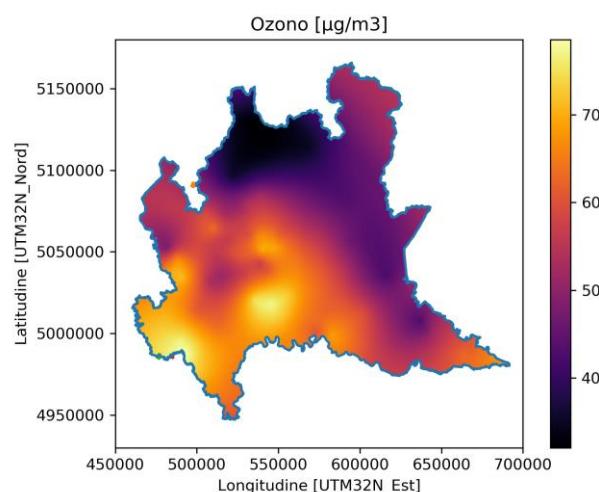


Figura 37 Rappresentazione di una mappa di interpolazione per l'ozono in Lombardia

---

## Biossido di azoto:

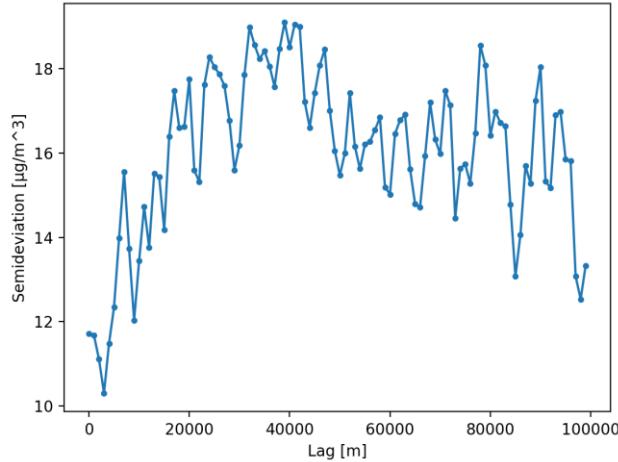


Figura 38 Semivariogramma del biossido di azoto

Come nel caso precedente, il biossido di azoto viene rilevato in maniera similare per stazioni prossime tra loro mentre si evidenzia un comportamento che tende a stabilizzarsi superati i 40Km tra le centraline.

Osservando bene il grafico si può osservare come nei primi 5Km la situazione sia pressoché stazionaria e come la correlazione permanga di interesse fin oltre i 10Km.

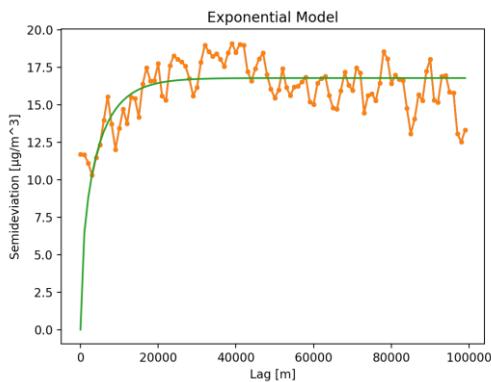


Figura 39 Fitting esponenziale per il semivariogramma del biossido di azoto

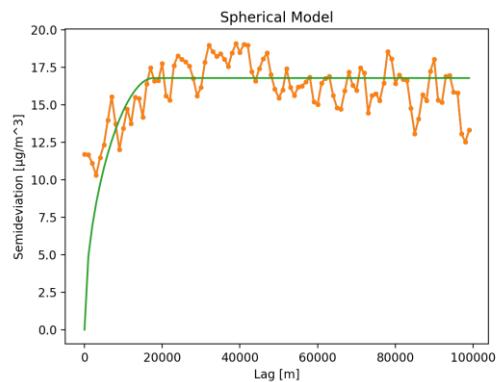


Figura 40 Fitting sferico per il semivariogramma del biossido di azoto

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma del biossido di azoto.

---

Ad una prima osservazione si può osservare che entrambi i modelli risultano parecchio simili e che sussiste un trend decrescente superati i 40Km, che porta da uno scostamento dal modello.

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.38 per il modello esponenziale ed uno di 0.40 per il modello sferico. Restringendo invece il semivariogramma ai primi 30Km si calcolano valori di  $R^2$  pari a 0.65 e 0.69 relativamente ai fitting esponenziale e sferico.

Ottenendo benefici sia impiegando il modello nella sua globalità sia impiegandolo in maniera ristretta alle brevi distanze si può apprezzare la miglior copertura offerta dalla funzione sferica, che è stata dunque impiegata nella creazione del modello di previsione. E' bene osservare che nonostante  $R^2$  calcolato sull'intero semivariogramma risulti molto basso, l'elevata distribuzione delle centraline sul territorio fa sì che interpolazioni a distanze superiori ai 40Km, punto in cui sembra innescarsi un trend decrescente, siano nella pratica mai effettuate.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

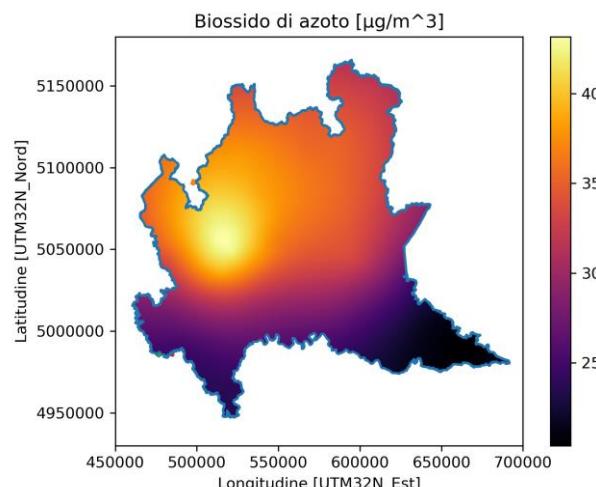


Figura 41 Rappresentazione di una mappa di interpolazione per il birossido di azoto in Lombardia

---

## Ossidi di azoto:

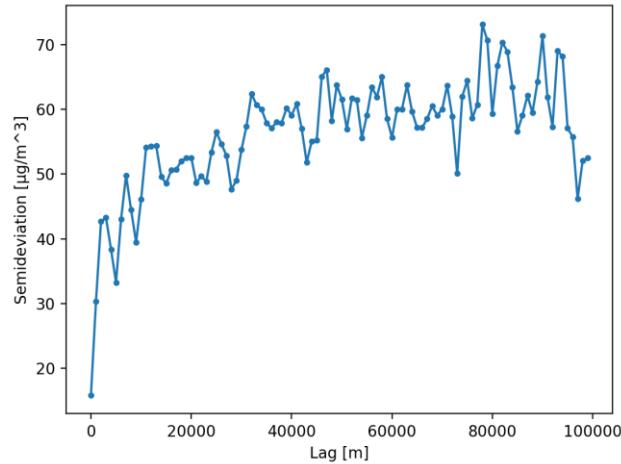


Figura 42 Semivariogramma segli ossidi di azoto

Gli ossidi di azoto presentano una semideviazione a crescita esponenziale che tende a stabilizzarsi superati i 40Km tra due centraline.

Il grafico sembra indicare una decorrelazione molto rapida tra le misurazioni.

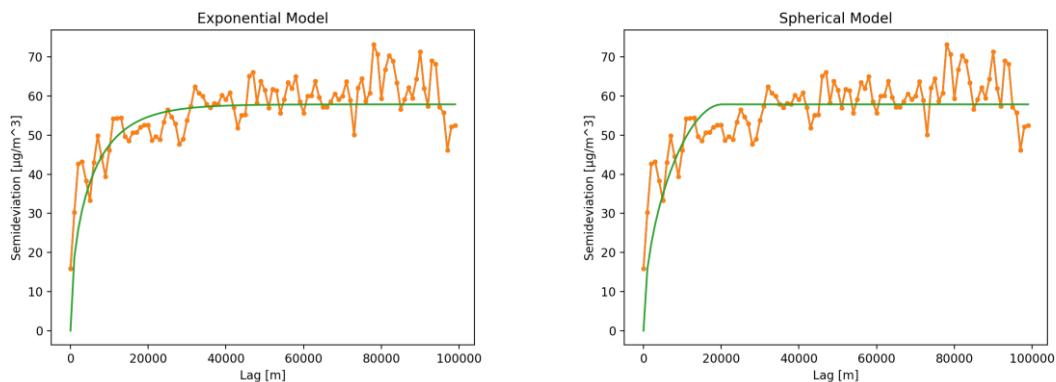


Figura 43 Fitting esponenziale per il semivariogramma degli ossidi di azoto

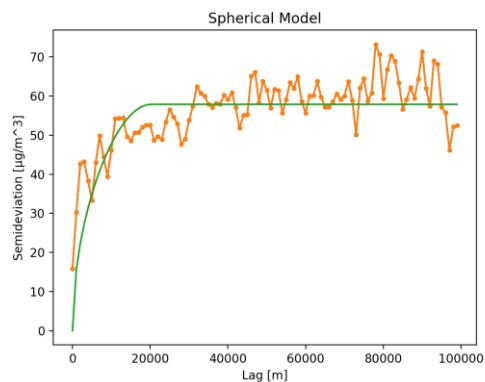


Figura 44 Fitting sferico per il semivariogramma degli ossidi di azoto

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma degli ossidi di azoto.

Osservando i due grafici si può notare un comportamento molto simile di ambedue le curve. Entrambi i fitting riescono ad approssimare bene i valori sulle brevi distanze e risultano consistenti anche considerando lag in un intorno dei 30Km.

---

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.55 per il modello esponenziale ed uno di 0.47 per il modello sferico. Restringendo invece il semivariogramma ai primi 30Km si calcolano valori di  $R^2$  pari a 0.71 e 0.70 relativamente ai fitting esponenziale e sferico.

Ottenendo benefici sia impiegando il modello nella sua globalità sia impiegandolo in maniera ristretta alle brevi distanze si può apprezzare la miglior copertura offerta dalla funzione esponenziale, che è stata dunque impiegata nella creazione del modello di previsione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

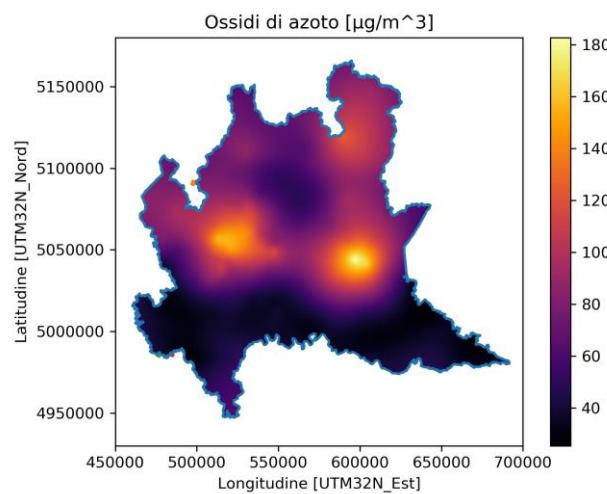


Figura 45 Rappresentazione di una mappa di interpolazione per gli ossidi di azoto in Lombardia

---

## Monossido di carbonio:

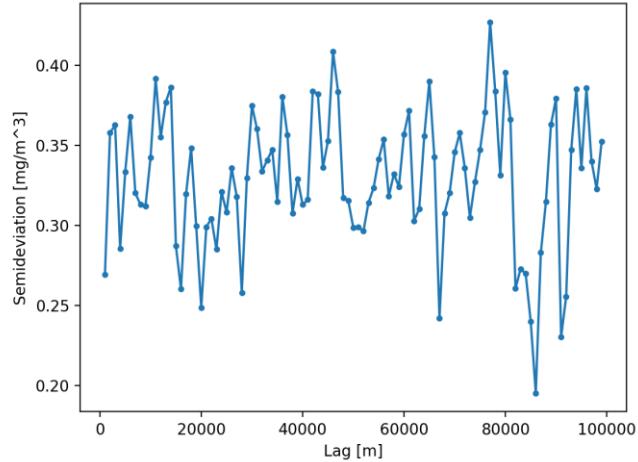


Figura 46 Semivariogramma del monossido di carbonio

Il monossido di carbonio ha un comportamento differente da quelli precedentemente analizzati. Ha un fronte di crescita per stazioni prossime fra loro mentre tende a stabilizzarsi molto presto e ad avere un'oscillazione importante, segno che ad un trend di fondo si somma una maggior rilevanza puntuale piuttosto che spaziale.

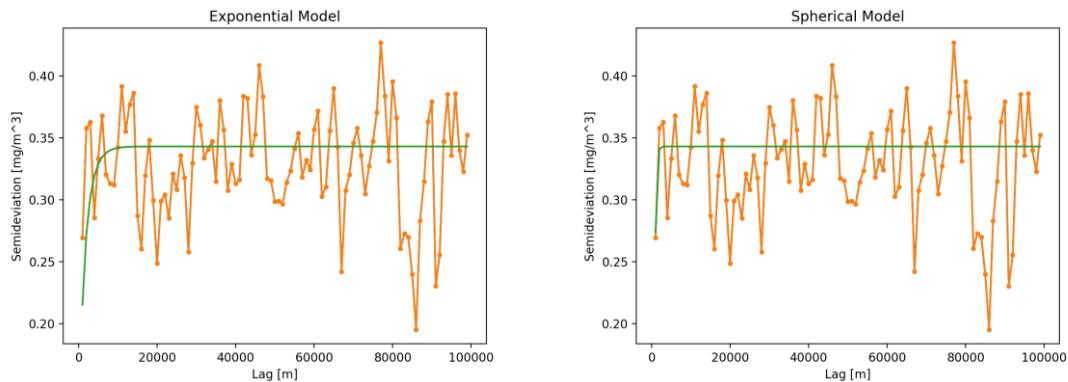


Figura 47 Fitting esponenziale per il semivariogramma del monossido di carbonio

Figura 48 Fitting sferico per il semivariogramma del monossido di carbonio

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma monossido di carbonio.

Osservando i due grafici si può notare un trend a cui è possibile ricondurre campionamenti molto prossimi tra di loro mentre in fenomeno risulta globalmente ricco di rumore nelle misurazioni e non delinea un andamento preciso.

---

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.005 per il modello esponenziale ed uno di 0.019 per il modello sferico. Restringendo invece il semivariogramma ai primi 10Km, ossia la copertura di una città di grandi dimensioni, si calcolano valori di  $R^2$  pari a 0.08 e 0.32 relativamente ai fitting esponenziale e sferico.

Entrambi i modelli non offrono una copertura esaustiva del fenomeno ma ottenendo benefici soprattutto sull'interpolazione nelle brevi distanze si è scelto di impiegare la funzione sferica nella creazione del modello di previsione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

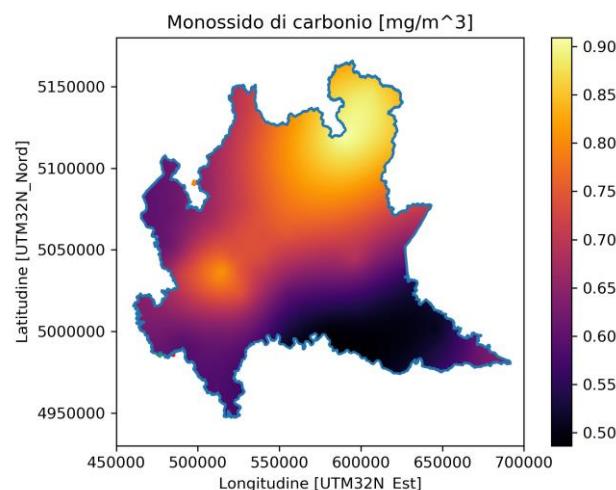


Figura 49 Rappresentazione di una mappa di interpolazione per il monossido di carbonio in Lombardia

---

## Benzene:

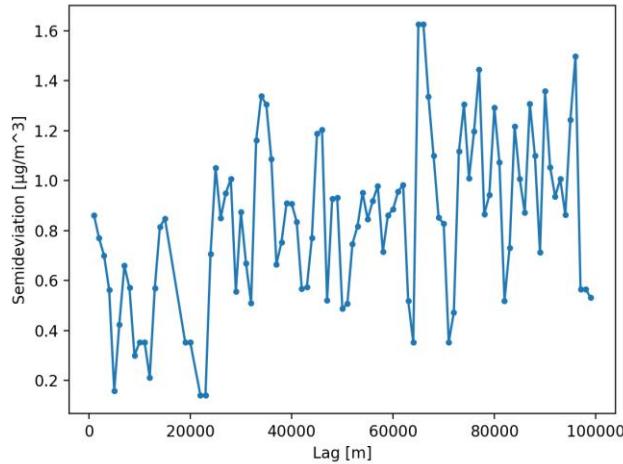


Figura 50 Semivariogramma del benzene

Il benzene presenta una semideviazione molto ricca di rumore. Nonostante ciò si osserva che il trend è di crescita ed è interessante notare come le centraline distanti fino a 20Km siano in grado di catturare valori molto simili tra loro.

Il rumore si può verosimilmente ricondurre al numero inferiore di centraline dispiegate sul territorio, che non consentono una rappresentazione estremamente precisa del fenomeno.

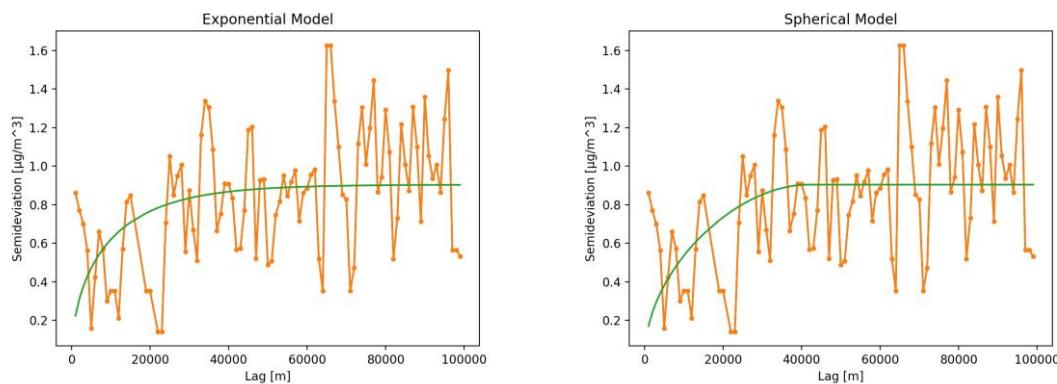


Figura 51 Fitting esponenziale per il semivariogramma del benzene

Figura 52 Fitting sferico per il semivariogramma del benzene

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma del benzene.

Calcolando  $R^2$  per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.148 per il modello esponenziale ed uno di 0.153 per il modello

---

sferico. Restringendo invece il semivariogramma ai primi 10Km si calcolano valori di  $R^2$  pari a 0.58 e 0.55 relativamente ai fitting esponenziale e sferico.

Entrambi i modelli non offrono una copertura esaustiva del fenomeno ma ottenendo benefici soprattutto sull'interpolazione nelle brevi distanze si è scelto di impiegare la funzione sferica nella creazione del modello di previsione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

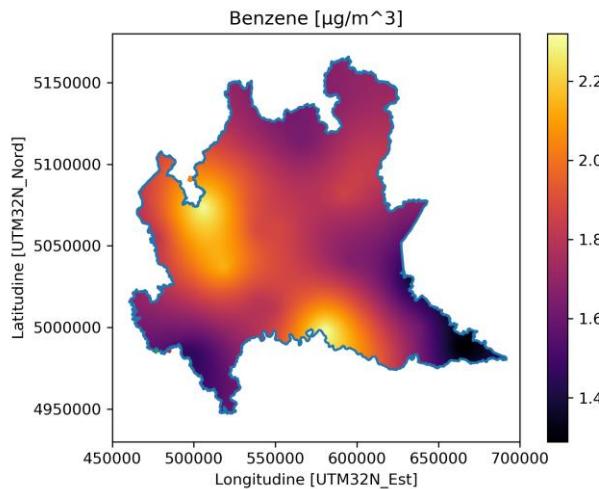


Figura 53 Rappresentazione di una mappa di interpolazione per il benzene in Lombardia

---

## **PM<sub>10</sub>:**

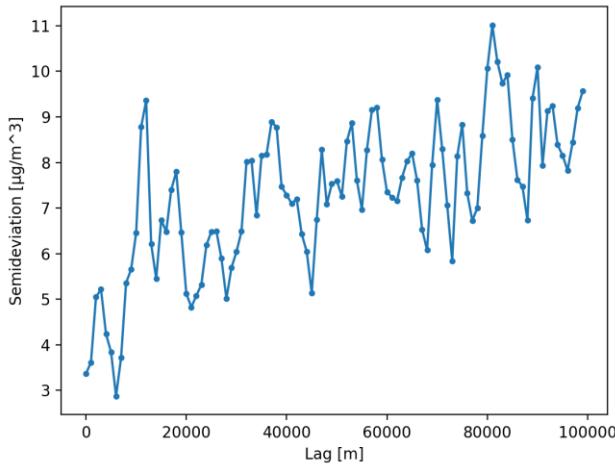


Figura 54 Semivariogramma del PM<sub>10</sub>

Nel caso del PM<sub>10</sub> si osservano misurazioni molto simili per centraline distanti pochi chilometri mentre in seguito l'andamento ha trend crescente. La deviazione standard è contenuta, segno di rilevazioni consistenti sul territorio.

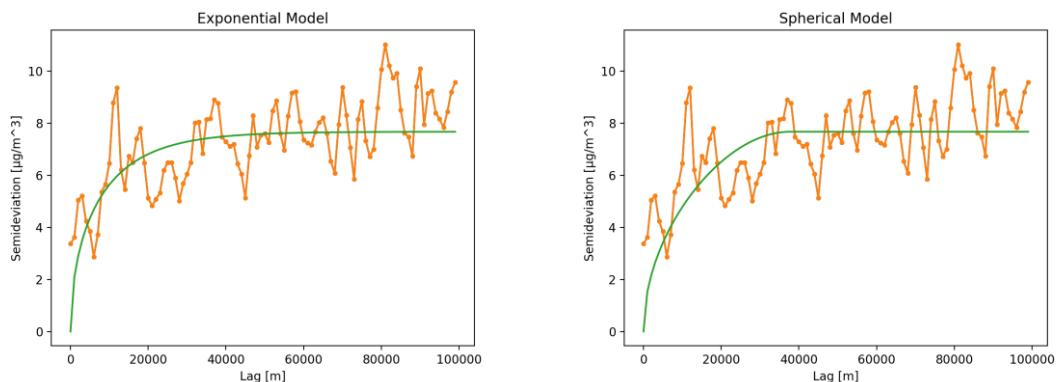


Figura 55 Fitting esponenziale per il semivariogramma del PM<sub>10</sub>

Figura 56 Fitting sferico per il semivariogramma del PM<sub>10</sub>

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma del PM<sub>10</sub>.

Calcolando R<sup>2</sup> per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.35 per il modello esponenziale ed uno di 0.33 per il modello sferico. Restringendo invece il semivariogramma ai primi 10Km si calcolano valori di R<sup>2</sup> pari a 0.29 e 0.33 relativamente ai fitting esponenziale e sferico.

---

In base ai dati registrati si può apprezzare la miglior copertura offerta dalla funzione sferica, che è stata dunque impiegata nella creazione del modello di previsione. E' bene osservare che nonostante  $R^2$  risulti piuttosto basso, nei contesti urbani in cui le centraline sono più densamente concentrate, il modello è ugualmente in grado di offrire risultati di valore. Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

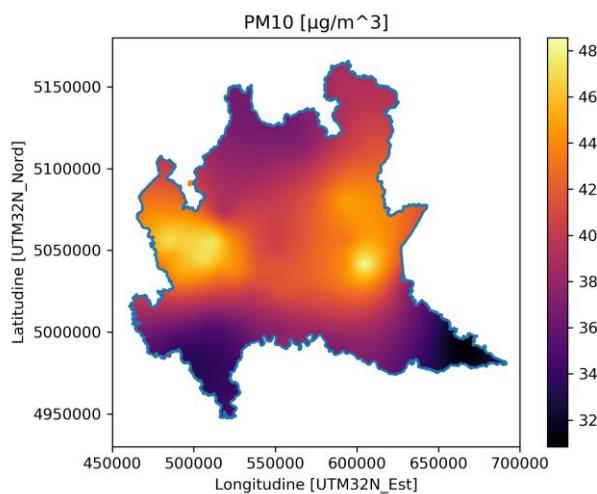


Figura 57 Rappresentazione di una mappa di interpolazione per il PM10 in Lombardia

---

## **PM<sub>2.5</sub>:**

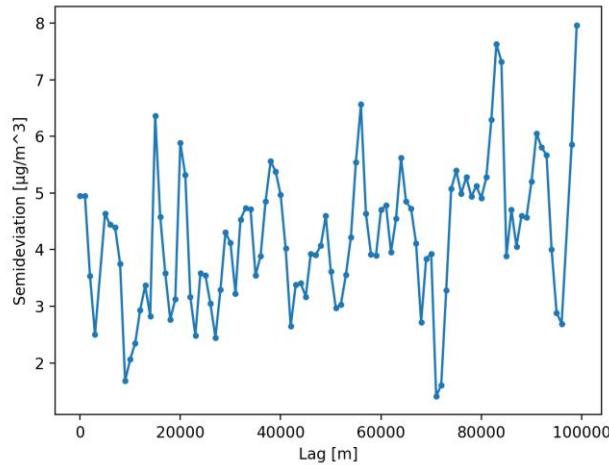


Figura 58 Semivariogramma del PM<sub>2.5</sub>

Il caso del PM<sub>2.5</sub> lascia intravedere un trend crescente seppur il grafico risulti molto rumoroso a causa del basso numero di stazioni presenti sul territorio. La deviazione standard è tuttavia contenuta ed è indice di consistenza tra le misurazioni.

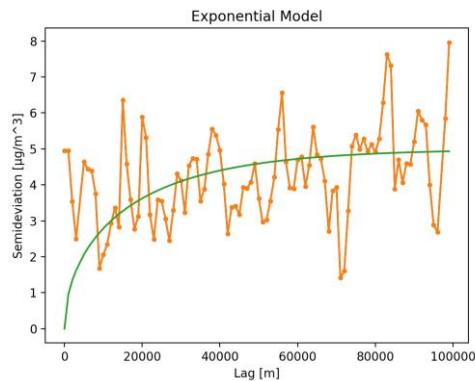


Figura 59 Fitting esponenziale per il semivariogramma del PM<sub>2.5</sub>

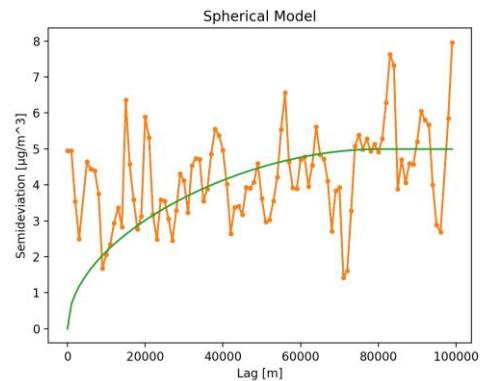


Figura 60 Fitting sferico per il semivariogramma del PM<sub>2.5</sub>

Le figure sovrastanti riportano il modello esponenziale e il modello sferico applicati al semivariogramma del PM<sub>2.5</sub>.

Calcolando R<sup>2</sup> per determinare la copertura del modello circa la reale distribuzione si osservano un valore di 0.08 per il modello esponenziale ed uno di 0.1 per il modello sferico.

Restringendo invece il semivariogramma ai primi 10Km si calcolano valori di  $R^2$  pari a 0.34 e 0.22 relativamente ai fitting esponenziale e sferico.

E' bene osservare che  $R^2$  risulti piuttosto basso. Il modello deve infatti coprire un fenomeno piuttosto variabile e campionato in maniera non consistente sul territorio.

Si è tuttavia deciso di utilizzare il fenomeno in quanto registrato in maniera ottimale soprattutto nei grandi centri abitati. Per questo motivo si è deciso di privilegiare la copertura nei primi 10Km, utilizzando una funzione esponenziale per la sua descrizione.

Di seguito si riporta un esempio di risultato ottenuto con tale modello applicato all'intera regione Lombardia:

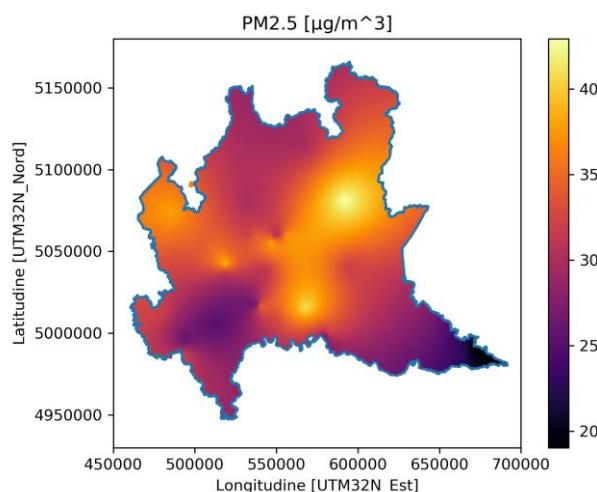


Figura 61 Rappresentazione di una mappa di interpolazione per il PM<sub>2.5</sub> in Lombardia

## 2.6.7 Validazione

Questo capitolo si prefigge di chiudere la sezione dedicata alle interpolazioni andando a comparare l'efficacia della modellazione utilizzata in relazione a quanto offerto dallo stato dell'arte in letteratura.

Come descritto in precedenza l'optimum nel campo delle interpolazioni è raggiunto dal sistema ARPA, basato su una modifica proprietaria dell'algoritmo di Optimal Interpolation. ARPA esegue le interpolazioni diverse volte durante il giorno e in media ogni 6 ore. La riesecuzione degli algoritmi consente di migliorare le previsioni passate. L'output generato consiste in una serie di grid file salvati in formato .txt e rappresentanti una griglia coprente l'intera Lombardia. Questa griglia ha una risoluzione di 10Km x 10Km nel caso delle interpolazioni di temperatura e 4Km x 4Km nel caso delle interpolazioni del PM<sub>10</sub>.

Su gentile concessione di ARPA si sono potuti avere alcuni file di esempio per quanto riguarda la temperatura al fine di poter mostrare le differenze che intercorrono con il modello utilizzato in questo progetto di tesi.

---

Di seguito si riportano la mappa generata tramite Optimal Interpolation e quella realizzata mediante Kriging bidimensionale:

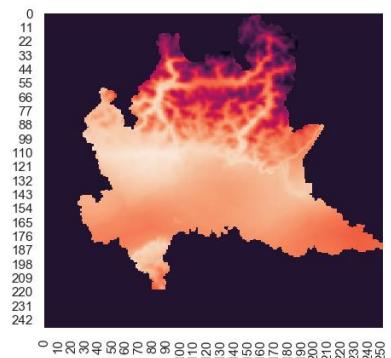


Figura 62 Rappresentazione grafica di un grid file della temperatura sul territorio della Lombardia elaborato da ARPA

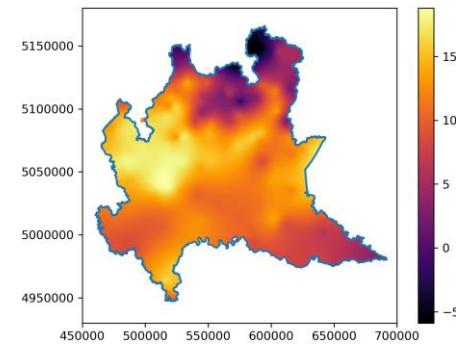


Figura 63 Rappresentazione grafica di una mappa di interpolazione della temperatura sul territorio della Lombardia

Come si può osservare il modello di Kriging è in grado di catturare la quasi totalità delle informazioni rappresentate nel riferimento.

La morfologia lombarda presenta infatti zone per lo più pianeggianti che consentono al modello bidimensionale di risultare pienamente significativo. Al contrario, come constatato in precedenza, le aree ad alto gradiente altimetrico risultano non catturabili a causa della natura dei dati a disposizione, per cui non è prevista la variabile legata all'altitudine.

Questo deficit è evidente soprattutto nelle aree settentrionali della regione, che appaiono le uniche a non essere conformi al modello guida.

Allo scopo di approfondire la validità del modello da un punto di vista quantitativo si è invece proceduto alla valutazione dell'attendibilità dei dati interpolati sull'area della città di Milano. L'elevata presenza di centraline consente infatti di svolgere analisi dettagliate in merito alla capacità predittiva, ai differenziali e alla deviazione standard legate al Kriging bidimensionale.

Per svolgere questo tipo di indagine si è provveduto a generare un modello di interpolazione a livello regionale e si sono confrontati i valori interpolati in coincidenza con le centraline presenti sul suolo milanese, valutando le differenze in termini di precisione rispetto ai dati registrati dalle centraline stesse.

Di seguito si riportano la mappa interpolata della temperatura a Milano e la tabella indicante temperature lette, temperature interpolate, differenziale di predizione e deviazione standard. In rosso sono evidenziate le posizioni delle 6 centraline meteorologiche.

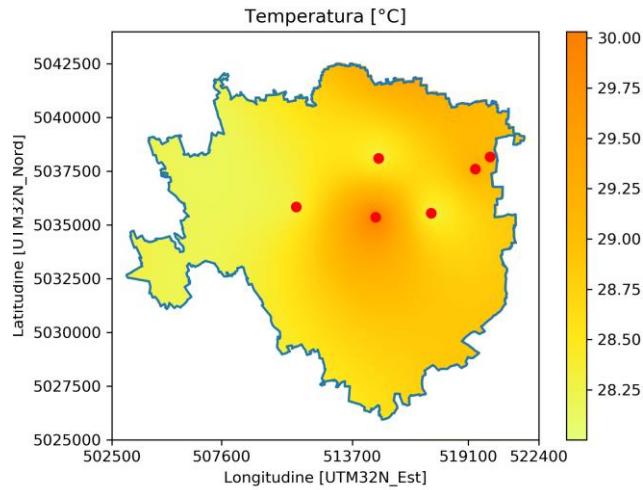


Figura 64 Rappresentazione grafica di una mappa di interpolazione della temperatura per il comune di Milano

Tabella 10 Validazione del modello di interpolazione per la temperatura

NOME CENTRALINA	VALORE CENTRALINA	VALORE INTERPOLATO	DIFFERENZIALE	DEVIAZIONE STANDARD
Milano Lambrate	29.0	29.0	-5.97e-13	8.66e-07
Milano v.Brera	30.6	30.6	-5.29e-13	8.06e-07
Milano v.Juvara	27.9	27.9	-7.78e-13	6.85e-07
Milano v.Marche	28.0	28.0	-8.42e-13	1.08e-06
Milano p.zza Zavattari	28.2	28.2	-6.36e-13	7.01e-07
Milano v.Feltre	29.4	29.4	2.03e-13	6.012e-07

Come si può osservare dalla tabella sopra, i valori vengono interpolati con estrema precisione e risultano pressoché identici ai dati registrati dalle centraline stesse. Il modello riesce dunque a catturare molto bene le relazioni di correlazione spaziale e a fornire una rappresentazione quasi esatta del fenomeno.

Per approfondire ulteriormente la relazione che lega le stazioni al modello si è scelto di ricorrere alla cross validazione dei dati generando modelli privi di una delle centraline e verificando la capacità predittiva nei confronti della stessa. A rotazione sono state escluse, una alla volta, tutte e 6 le centraline.

---

Il processo ha mostrato come il modello risulti decisamente consistente nel predire i valori corretti. I differenziali si rivelano sempre molto contenuti e la deviazione standard è indice di consistenza dei valori.

Ne consegue che il modello dimostra una buona robustezza e si dimostra un affidabile strumento per l'interpolazione dei valori. I risultati ottenuti sono riportati in appendice.

In questo capitolo si è mostrato come solo esempio il modello utilizzato per l'interpolazione della temperatura al fine di non appesantire l'elaborato con immagini ridondanti rispetto all'obiettivo finale. Si rimanda all'appendice per la visione delle mappe complete di tutti i fenomeni mentre di seguito si riporta una tabella riassuntiva riportante il differenziale medio e la deviazione standard media per ciascuna delle variabili analizzate:

Tabella 11 Validazione del modello di interpolazione per tutte le variabili indagate

NOME FENOMENO	DIFFERENZIALE MEDIO	DEVIAZIONE STANDARD MEDIA
Temperatura	-5.30e-13	7.92e-07
Umidità relativa	-7.46e-14	5.48e-07
Ozono	4.26e-14	5.18e-07
Biossido di azoto	-1.23e-13	8.71e-07
Ossidi di azoto	1.90e-13	2.24e-06
Monossido di carbonio	2.11e-15	1.51e-08
Benzene	-1.85e-15	1.51e-08
PM10	-4.74e-15	1.93e-07
PM2.5	-7.99e-14	3.29e-07

da questi risultati, si può osservare una ottima corrispondenza tra il valore stimato dal modello utilizzato ed il valore misurato, per ogni parametro considerato.

## 2.6.8 Manipolazione dei dati

Il processo di pre-elaborazione dei dati si prefigge di svolgere due compiti essenziali:

- Screamatura dei record
- Strutturazione delle sorgenti dati

Il processo di screamatura rappresenta il primo passo da compiere precedentemente all'utilizzo delle informazioni contenute in una base dati e ricopre un ruolo fondamentale in quanto consente di filtrare le informazioni non significative o inconsistenti che riducono il

---

livello di data quality assegnato alla sorgente. In questa fase si individuano i campi di interesse all'interno del database e si rimuovono i record che ne presentano una rappresentazione errata o inesistente.

Nel corso del progetto in questione si è provveduto innanzitutto ad una scrematura dei record presenti all'interno del database fornito da AREU in modo da mantenere solamente le tuple a cui fossero riconducibili delle coordinate geografiche che consentissero di localizzare l'evento in maniera precisa. Questa scelta è stata dettata dalla necessità di dover individuare correttamente i vari eventi al fine di poter svolgere le analisi successive.

Per quanto riguarda invece i database dei fenomeni meteorologici e degli agenti inquinanti forniti da ARPA, ci si è limitati al filtraggio dei record che presentassero un'invalidità dei dati rilevati dai sensori e dunque non risultassero significativi. Come osservato nei capitoli 2.2.3 e 2.3.3 questa operazione è stata resa possibile dalla presenza del campo "Stato" che esprime la validità o meno di un campionamento. La scelta in questo caso ha ricalcato quello che è il comportamento tenuto da ARPA stessa nell'impiego dei dati, che prevede lo scarto dei sensori mal funzionanti piuttosto che la loro imputazione.

La seconda metodologia di pre-processing seguita in questa prima fase ha riguardato invece la strutturazione unificata delle sorgenti in maniera tale da rendere omogenea la rappresentazione delle informazioni grazie all'unificazione dei formati, dei tipi di dato e soprattutto del loro significato semantico.

Il database AREU è stato dunque arricchito di tre nuovi campi:

- UTM32N\_Est
- UTM32N\_Nord
- DATE\_TIME

Intuibilmente i primi due campi rappresentano le coordinate geografiche associate agli eventi di ictus ed espresse nel formato EPSG:32632, che come descritto nel capitolo 1.3.2 rappresenta lo standard per la rappresentazione dei dati territoriali nel fuso 32N, a cui la Lombardia è riconducibile. L'aggiunta dei nuovi valori di longitudine e latitudine ha richiesto un'operazione di conversione delle coordinate già associate all'evento (ed individuabili tramite i campi VL\_GEO\_X e VL\_GEO\_Y) in quanto rappresentate nel formato EPSG:3003. Per ottenere una conversione corretta, si è fatto uso della libreria python pyproj, come riportato nello script 2.

L'operazione di standardizzazione del database AREU ha portato inoltre alla creazione del campo DATE\_TIME, ossia un timestamp nel formato "%d-%m-%Y %H:%M:%S" formato dall'aggregazione dei campi DT\_EMERG\_DAY e DT\_EMERG\_OPEN\_HH24MI.

---

Al fine di uniformare l'intervallo di campionamento con quello del database ARPA, che si ricorda raccogliere nuovi dati con granularità pari a 10 minuti, si è deciso di arrotondare l'orario riportato dal timestamp al multiplo di 10 minuti più prossimo. Così facendo, il nuovo campo ha consentito di assegnare ad ogni evento di ictus un orario che potesse essere messo in relazione diretta con i dati meteorologici e ambientali tramite una semplice intersezione degli orari.

Inoltre, per poter garantire un maggior agilità nel trattamento di un'elevata mole di dati, si è deciso di suddividere per anno il database in modo da ridurre l'impatto sulla memoria occupata nella fase di analisi. Sono dunque stati generati i file:

- Tabella\_completa\_2015
- Tabella\_completa\_2016
- Tabella\_completa\_2017

L'operazione di strutturazione dei dati rilevati dalle centraline ARPA ha invece seguito un processo differente. Innanzitutto si è provveduto a collezionare i dati degli ultimi 5 giorni dell'anno precedente a quello riportato dal timestamp dei campionamenti in modo da poter supportare l'individuazione delle misurazioni in un lag period di cinque giorni anche per le prime giornate di ogni anno. Per questo motivo ad ogni annualità sono stati ricondotti i record successivi al 26 dicembre dell'anno precedente.

In seguito si è provveduto a standardizzare il formato del timestamp secondo quanto descritto in precedenza per la sorgente dati AREU e seguendo dunque il formato “%d-%m-%Y %H:%M:%S” e sostituendo al nome del campo “Data” la stringa DATE\_TIME, in modo da garantire una continuità semantica con quanto descritto in precedenza.

Infine si è voluto separare i dati record raccolti su base giornaliera da quelli campionati ogni 10 minuti. L'operazione è stata motivata da una volontà di semplificare le operazioni di join successivamente svolte in fase di analisi in quanto un'unica fonte dati più corposa avrebbe richiesto un quantitativo di memoria superiore a quella impiegata per processare i dati di due database differenti in maniera seriale. L'individuazione dei fenomeni campionati su base giornaliera è stata effettuata applicando un filtro sul nome dei campi da considerare in questa tipologia.

Queste operazioni sono state svolte sui dati meteorologici e su quelli ambientali in maniera separata. Sono stati dunque generati i seguenti file:

- 
- DATI\_ARIA\_2015
  - DATI\_ARIA\_2016
  - DATI\_ARIA\_2017
  - DATI\_METEO\_2015
  - DATI\_METEO\_2016
  - DATI\_METEO\_2017
  - INQUINANTI\_GIORNALIERI\_2015
  - INQUINANTI\_GIORNALIERI\_2016
  - INQUINANTI\_GIORNALIERI\_2017

## 2.7 Elaborazione dei dati

Il presente capitolo esplora le metodologie di elaborazione dei dati a disposizione e le procedure tramite cui si è giunti a risultati proposti nella sezione Risultati dell'elaborato.

### 2.7.1 Interpolazione dei valori

Come accennato in precedenza, al fine di assegnare ad ogni evento di ictus valori specifici per quanto riguarda i fenomeni atmosferici e inquinanti analizzati, si è dovuto far ricorso ad un sistema di interpolazioni spaziali effettuate mediante un modello di Kriging bidimensionale.

La procedura di interpolazione prevede che ad ogni evento sia ricondotto, mediante un'operazione di join fra tabelle, un set di centraline attive all'istante in cui esso si è verificato. Questo insieme di stazioni costituisce la base per il modello di interpolazione e viene usato come campionamento di base per definire la funzione di mappatura.

Dopo aver individuato il modello di fitting del semivariogramma per un determinato fenomeno, viene richiamata la creazione del modello tramite l'apposita libreria software. Una volta creato il modello, esso può essere usato per generare interpolazioni lungo la superficie geografica d'interesse.

E' bene ricordare che il modello creato è valido solo e unicamente per il fenomeno preso in esame nell'ora in cui esso è stato misurato. Ogni qual volta si desideri generare interpolazioni da attribuire ad un nuovo record nel database degli ictus, è necessario definire nuovamente il modello in base al fenomeno osservato, alla data e all'ora delle rilevazioni delle centraline.

Il processo di interpolazione per l'intero database consta quindi di tre fasi:

- Attribuzione dei valori delle centraline all'evento di ictus nella data/ora in cui esso si è verificato

- 
- Generazione del modello di interpolazione per record e per ogni fenomeno di interesse
  - Utilizzo del modello per l'individuazione del valore interpolato

Osservando l'elenco di operazioni è evidente come il primo passaggio risulti essere quello computazionalmente più intenso a causa dell'elevato quantitativo di memoria richiesta all'elaboratore per ospitare i dataset ed effettuarne il join.

Per cercare di alleggerire questa fase dell'analisi e rendere il processo di interpolazione modulare, si è deciso di effettuare una riduzione della tabella contenente tutti gli eventi di ictus. Per fare ciò si è introdotto nella stessa il campo IDX, indicante l'indice intero associato ai record, ed in seguito si è provveduto all'estrazione dei soli campi di interesse utili per l'interpolazione. Questi campi sono:

- DATE\_TIME
- UTM32N\_Est
- UTM32N\_Nord
- IDX

Ridurre il numero di campi consente di occupare uno spazio inferiore in memoria e di ridurre quindi il rischio di overflow, con conseguente operazione di swap sulla memoria di massa e risultante degrado prestazionale.

Le tabelle ridotte sono generate per ogni annualità e per ciascuno dei lag period in esame. Sempre per una questione di ottimizzazione, le tabelle vengono materializzate, salvate su disco e caricate all'occorrenza durante il processo di interpolazione.

Il campo IDX svolge il ruolo di indice sia nelle tabelle ridotte sia nella tabella principale e consente dunque, con una facile operazione di join, di riassociare i risultati ottenuti al corrispondente record della tabella originaria.

Una seconda ottimizzazione consta nella scelta di creare una tabella ridotta per ogni possibile lag period. Tale decisione è stata presa a fronte della possibilità di esecuzione di task concorrenti, nonché della non necessità di eseguire i calcoli in maniera temporalmente dipendente.

Il processo di interpolazione viene eseguito operando sui dati anno dopo anno ed è svolto in tre fasi:

- **Interpolazione dei dati meteorologici:** in questa fase viene utilizzata la tabella ridotta precedentemente descritta e vengono generate le interpolazioni di temperatura e umidità relativa. Il risultato finale consta in una tabella nel formato

---

(DATE\_TIME, UTM32N\_Est, UTM32N\_Nord, IDX, TEMPERATURE, HUMIDITY). Viene creata una tabella di questo tipo per ognuno dei lag period in esame e viene salvata nel file il cui nome segue il formato weather\_ANNO\_LAGPERIOD.

- **Interpolazione dati sugli inquinanti continui:** in questa seconda fase vengono svolte le interpolazioni sugli inquinanti il cui campionamento avviene ogni 10 minuti e dunque presentano un fattore di continuità temporale. Si usa sempre la tabella ridotta come base di partenza mentre le misure interpolate vengono aggiunte come nuovi campi. La tabella finale si mostra nel formato (DATE\_TIME, UTM32N\_Est, UTM32N\_Nord, IDX, CO, O<sub>3</sub>, NO<sub>x</sub>, NO<sub>2</sub>). Come nel caso precedente, il risultato è materializzato nel file il cui nome segue il formato pollutants\_ANNO\_LAGPERIOD.
- **Interpolazione dati sugli inquinanti giornalieri:** quest'ultima fase prevede l'interpolazione degli inquinanti raccolti come media su base giornaliera. Il procedimento è analogo ai due casi sopra riportati e l'output è rappresentato da una tabella nel formato (DATE\_TIME, UTM32N\_Est, UTM32N\_Nord, IDX, PM10, PM25, BENZENE) che viene successivamente salvata in un file il cui nome segue il formato pollutants\_daily\_ANNO\_LAGPERIOD.

Al termine di questa serie di interpolazioni si otterrà un elenco di file con tutti i fenomeni di interessi suddivisi per tipologia.

Il penultimo passo del processo di interpolazione consta nell'utilizzo del campo IDX presente sia nella tabella originale sia nelle tabelle ridotte, per ricongiungere i risultati interpolati ai record completi sugli ictus. Al termine di questa fase vengono creati i file nel formato FINAL\_ANNO, contenenti tutti i record suddivisi per annualità e associati ai valori interpolati in tutti i lag period di interesse.

L'ultimo step dell'intera opera di interpolazione consiste nel filtraggio dei file precedentemente creati. L'operazione si rende necessaria in quanto può capitare che alcuni dati sui fenomeni meteorologici o sugli agenti inquinanti presentino lacune nella campionatura e dunque risultino mancanti in talune date o orari. Queste assenze portano all'impossibilità di generare un'interpolazione valida. Si è dunque deciso di porre un filtro a questi dati mancanti eliminando il record interessato, così da non intaccare l'analisi aggregata portata avanti nelle fasi successive.

Oltre alla semplice rimozione dei record incompleti, si è poi provveduto ad una razionalizzazione dei risultati, ossia all'eliminazione di quelle tuple che presentassero valori fuori scala per i fenomeni osservati. Si è dunque deciso di limitare la temperatura a valori

---

ragionevoli compresi tra i  $-20^{\circ}\text{C}$  e i  $+50^{\circ}\text{C}$ , così come i valori di umidità relativa sono stati contenuti in una scala percentuale tra 0 e 100. L'obbligo di correzione sorge dalla natura dei modelli e in relazione alla distribuzione delle centraline. Nelle aree scarsamente coperte da sensori viene infatti lasciata libertà al modello di crescere o decrescere senza correzioni e questo porta in rari casi ad interpolazioni inconsistenti ed in contrasto con una ragionevole misura.

L'output contenente le tabelle finali è rappresentato dai file nel formato GOAL\_ANNO.

## 2.7.2 Lag period ed intervalli di analisi

Come suggerito dall'analisi in letteratura, l'insorgenza di ictus sembra potersi ricondurre all'andamento che i fenomeni indagati hanno avuto nei giorni precedenti al manifestarsi della patologia.

In questo progetto di analisi si è dunque deciso di seguire quanto suggerito dagli studi in materia e dunque utilizzare un lag period esteso a 5 giorni.

Per approfondire meglio le dinamiche legate alla variazione dei fenomeni in esame, si è deciso di analizzare in maniera ancor più dettagliata il lag period, campionando le misurazioni ogni 12 ore nell'arco dei sopra citati 5 giorni.

Così facendo si hanno a disposizione il doppio dei dati e si ha la possibilità di valutare al meglio le variazioni non solo giornaliere ma anche giorno-notte. La scelta è giustificata dal fatto che in letteratura si trova riscontro della possibile influenza di ampi differenziali piuttosto che di alti valori dei campionamenti in sé.

Ad ogni record vengono dunque associati 11 valori relativi alla misurazione all'istante dell'evento e al lag period campionato ogni 12 ore. Queste assegnazioni vengono riproposte per ciascuno dei fenomeni in esame, siano essi meteorologici o legati alla qualità dell'aria.

## 2.7.3 Analisi temporale

Al fine di individuare potenziali ciclicità che possano dare un'idea migliore sulla distribuzione degli ictus, si è deciso di indagare in merito alla distribuzione temporale degli eventi.

Si è fatto uso del database fornito da AREU e si è effettuato un pre-processing dei dati per associare ogni record all'ora intera più prossima, al giorno della settimana e al mese dell'evento, già presente all'interno della base dati.

Fatto ciò si è proceduto ad un'operazione di group by su base oraria, giornaliera e mensile, contando poi le tuple associate ad ogni chiave per identificare il numero di eventi ad essa

---

riconducibili. L'operazione è stata ripetuta sia sui dati cumulativi legati al triennio 2015-2017 sia per ciascuna delle singole annualità al fine di identificare possibili trend o associazioni.

## 2.7.4 Analisi di correlazione

Individuate le percentuali di incidenza sulla popolazione si è deciso di investigare in merito alla sussistenza di eventuali relazioni di correlazione tra gli eventi di ictus e i fenomeni presi in esame.

Per fare ciò si è deciso di ridurre il campo di ricerca, restringendolo alla sola città di Milano. La scelta è stata dettata dalla volontà di avere dati quanto più precisi possibile ed un campione numeroso e significativo di eventi osservati. Milano, con la sua capillare copertura di stazioni ARPA e la sua popolazione di 1.3 milioni di abitanti offre le condizioni ideali per questo tipo di investigazione.

Si sono cercate dapprima tracce di correlazione lineare che legassero l'occorrenza di ictus ad ognuno dei fenomeni meteorologici e degli agenti inquinanti trattati in precedenza. Si è infine proceduto a realizzare una matrice di correlazione che legasse ognuno dei parametri al fine di rilevare possibili dipendenze tra di essi.

L'analisi è stata svolta prendendo in esame il valore dei fenomeni sia nel momento della manifestazione dell'ictus sia nei precedenti lag period. Nella fase di correlazione sono stati indagati i legami in tre varianti temporali:

- Legame istantaneo: si è cercata la correlazione tra il numero di ictus giornalieri e la media dei fenomeni associati all'istante della chiamata al 112. Tale istante è stato utilizzato anche nella definizione dei lag period.
- Legame giornaliero: si è cercata la correlazione tra il numero di ictus in un dato giorno e la media giornaliera dei fenomeni. La stessa procedura è applicata ai lag period.
- Legame mensile: si è cercata la correlazione tra il numero di ictus su base mensile e le medie mensile dei diversi fenomeni.

In ultima istanza si è proceduto a verificare eventuali correlazioni tra gli ictus e i parametri sotto indagine associando ogni evento ad un range del fenomeno, suddividendolo così in cluster. Si è poi fatto uso della correlazione a ranghi di Spearman per calcolare l'indice di correlazione e la sua significatività.

---

# CAPITOLO 3

## RISULTATI

### 3.1 Analisi demografica

Il punto di partenza dell'analisi demografica consiste nella raccolta dei dati relativi alla popolazione residente in ognuna delle 12 province della Lombardia.

Di seguito si riporta una tabella riassuntiva che conta il numero di abitanti in base al sesso dell'individuo e alla provincia di residenza:

Tabella 12 Popolazione residente in base alla provincia

ANNO	PROVINCIA	MASCHI RESIDENTI	FEMMINE RESIDENTI	TOTALE RESIDENTI
2015	BG	548992	559861	1108853
	BS	622658	642419	1265077
	CO	293530	306375	599905
	CR	177276	184334	361610
	LC	167195	173056	340251
	LO	113033	116543	229576
	MB	422740	441817	864557
	MI	1545121	1651704	3196825
	MN	202886	212033	414919
	PV	266468	282254	548722
	SO	89012	93074	182086
	VA	432704	457530	890234

	<b>TOTALE</b>	4881615	5120000	10002615
2016	BG	548643	559655	1108298
	BS	621957	642148	1264105
	CO	293494	306160	599654
	CR	176835	183609	360444
	LC	166897	172357	339254
	LO	112908	116505	229413
	MB	423516	442560	866076
	MI	1552258	1656251	3208509
	MN	202074	210794	412868
	PV	266380	281546	547926
	SO	88854	92858	181712
	VA	432727	457363	890090
<b>TOTALE</b>		<b>4888543</b>	<b>5121806</b>	<b>10008349</b>
2017	BG	549853	560080	1109933
	BS	621253	641425	1262678
	CO	293861	306329	600190
	CR	176295	183093	359388
	LC	166975	172263	339238
	LO	112928	116410	229338
	MB	425127	443732	868859
	MI	1557612	1660589	3218201
	MN	202255	210355	412610
	PV	266487	280764	547251
	SO	88922	92515	181437
	VA	432795	457248	890043
<b>TOTALE</b>		<b>4894363</b>	<b>5124803</b>	<b>10019166</b>

Utilizzando la base dati fornita da AREU si è poi proceduto all'estrazione del numero di individui colpiti, così da evidenziare la distribuzione degli eventi di icuts sul territorio regionale.

Di seguito si riporta il numero di pazienti in base alla provincia di residenza e al sesso della persona. I dati coprono un intervallo temporale di tre anni ed espongono le informazioni raccolte tra il 2015 e il 2017.

Tabella 13 Vittime di ictus in base alla provincia di residenza

<b>ANNO</b>	<b>PROVINCIA</b>	<b>MASCHI COLPITI</b>	<b>FEMMINE COLPITE</b>	<b>NON CLASSIFICATI</b>	<b>TOTALE COLPITI</b>
2015	BG	231	248	30	509
	BS	218	260	18	496
	CO	308	339	49	696
	CR	88	102	16	206
	LC	165	248	53	466
	LO	50	55	21	126
	MB	309	382	5	696
	MI	1341	1516	35	2892
	MN	115	141	10	266
	PV	132	196	46	374
	SO	44	61	7	112
	VA	413	544	46	1003
<b>TOTALE</b>		<b>3414</b>	<b>4092</b>	<b>336</b>	<b>7842</b>
2016	BG	284	274	32	590
	BS	294	377	21	692
	CO	317	350	57	724
	CR	117	135	14	266
	LC	188	224	39	451
	LO	72	86	26	184
	MB	386	414	8	808
	MI	1547	1884	39	3470
	MN	149	200	11	360
	PV	212	252	67	531
	SO	56	50	11	117
	VA	414	536	51	1001
<b>TOTALE</b>		<b>4036</b>	<b>4782</b>	<b>376</b>	<b>9194</b>
2017	BG	326	343	19	688
	BS	362	397	17	776
	CO	297	385	31	713
	CR	149	211	11	371
	LC	181	183	25	389
	LO	79	81	12	172
	MB	413	476	3	892

	MI	1616	1884	28	3528
	MN	179	210	12	401
	PV	265	340	27	632
	SO	57	57	5	119
	VA	408	527	36	971
	<b>TOTALE</b>	<b>4332</b>	<b>4782</b>	<b>376</b>	<b>9652</b>

Infine si è provveduto a calcolare il rapporto tra la popolazione colpita e quella residente in ciascuna delle provincie prese in esame.

Tabella 14 Incidenza percentuale in base alla provincia

ANNO	PROVINCIA	INCIDENZA MASCHI %	INCIDENZA FEMMINE %	INCIDENZA TOTALE %
2015	BG	0,042	0,044	0,046
	BS	0,035	0,040	0,039
	CO	0,105	0,111	0,116
	CR	0,050	0,055	0,057
	LC	0,099	0,143	0,137
	LO	0,044	0,047	0,055
	MB	0,073	0,086	0,081
	MI	0,087	0,092	0,090
	MN	0,057	0,066	0,064
	PV	0,050	0,069	0,068
	SO	0,049	0,066	0,062
	VA	0,095	0,119	0,113
2016	BG	0,052	0,049	0,053
	BS	0,047	0,059	0,055
	CO	0,108	0,114	0,121
	CR	0,066	0,074	0,074
	LC	0,113	0,130	0,133
	LO	0,064	0,074	0,080
	MB	0,091	0,094	0,093
	MI	0,100	0,114	0,108
	MN	0,074	0,095	0,087
	PV	0,080	0,090	0,097

	SO	0,063	0,054	0,064
	VA	0,096	0,117	0,112
2017	BG	0,059	0,061	0,062
	BS	0,058	0,062	0,061
	CO	0,101	0,126	0,119
	CR	0,085	0,115	0,103
	LC	0,108	0,106	0,115
	LO	0,070	0,070	0,075
	MB	0,097	0,107	0,103
	MI	0,104	0,113	0,110
	MN	0,089	0,100	0,097
	PV	0,099	0,121	0,115
	SO	0,064	0,062	0,066
	VA	0,094	0,115	0,109

Al fine di presentare al meglio i risultati ottenuti si è scelto di mostrare l'incidenza percentuale mediante istogramma.

Di seguito è riportato il valore di incidenza per ciascuna delle province lombarde in relazioni agli anni 2015-2017:

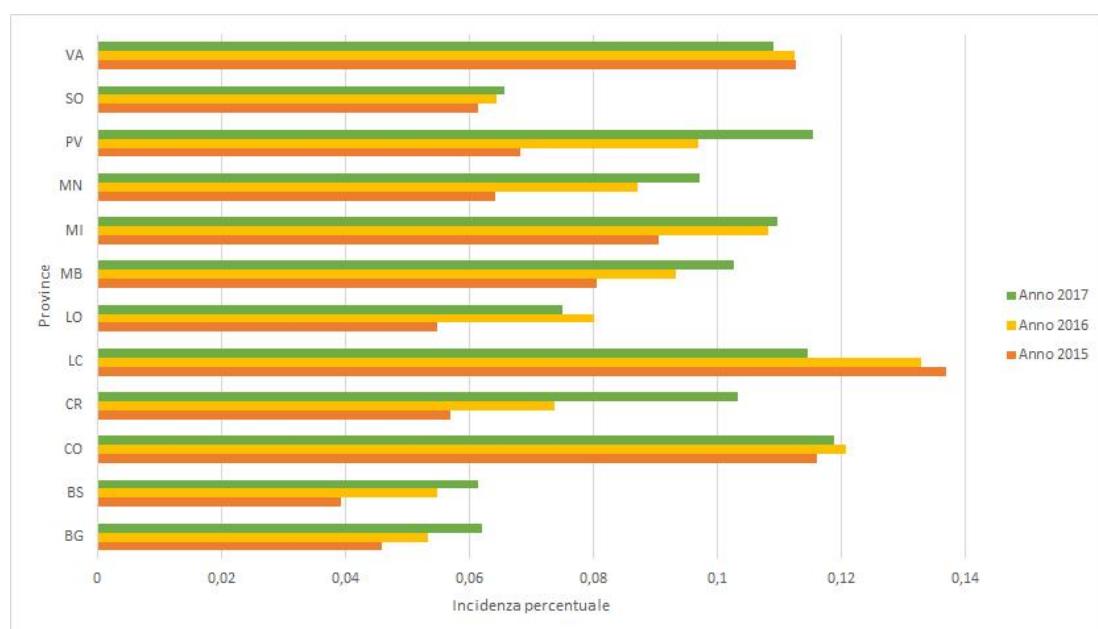


Figura 65 Incidenza percentuale negli anni 2015-2017 in relazione alle province di residenza

Da tale grafico è possibile osservare una diversità di incidenza nelle varie province lombarde, con picchi associati alle province di Lecco, Como e Varese. Inoltre, è possibile osservare nel triennio considerato un aumento generalizzato dei casi in gran parte del territorio regionale. L'ultimo passo dell'analisi demografica ha portato all'individuazione dell'incidenza aggregata in relazione al sesso dei pazienti e alle fasce d'età cui essi appartengono.

Di seguito si apprezzano i grafici esprimenti i risultati ottenuti:

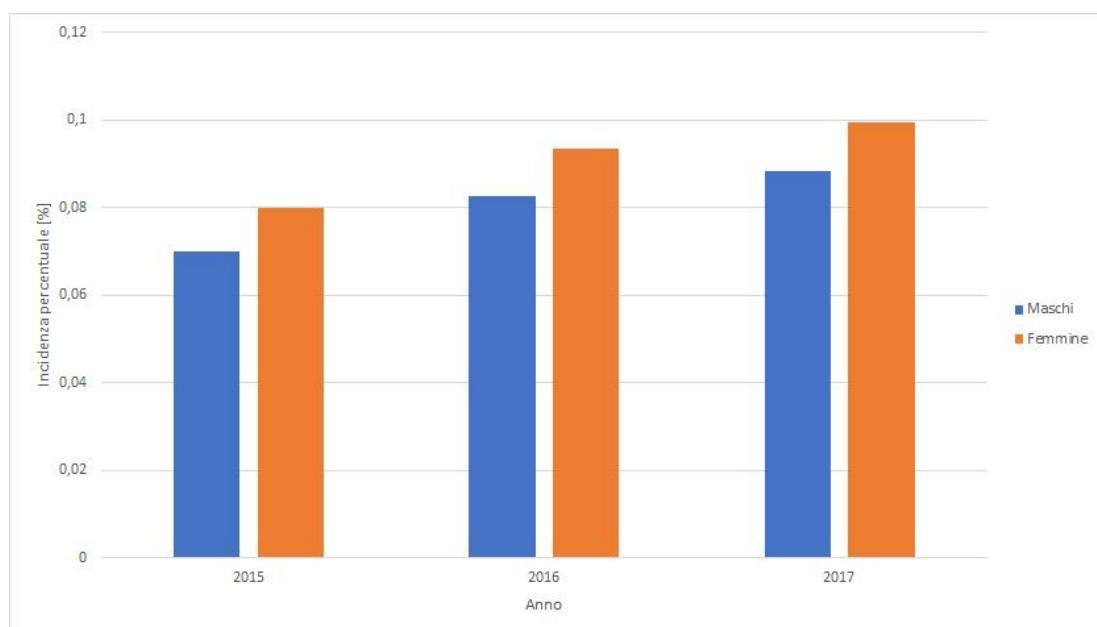


Figura 66 Incidenza percentuale riportata in base al sesso dei pazienti ed esplicitata per ognuno dei tre anni di interesse

E' possibile osservare una maggiore numerosità nei casi che hanno interessato la popolazione femminile, sia in termini di numero totale degli eventi (13565 negli uomini e 16171 nelle donne) sia in termini di incidenza annuale.

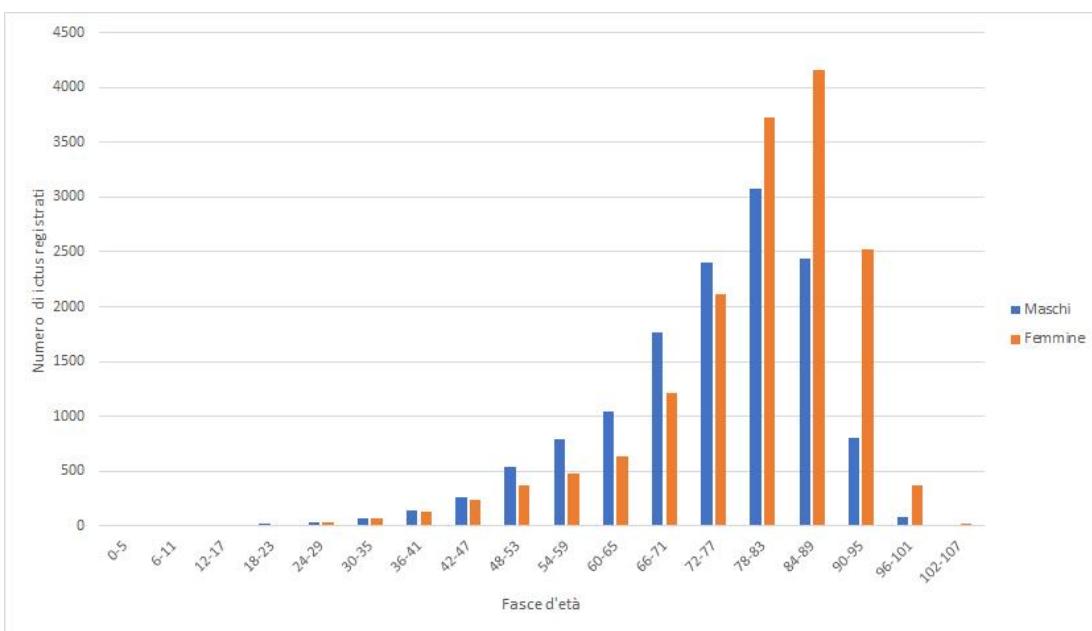


Figura 67 Rappresentazione della distribuzione del numero di ictus in base al sesso e per ciascuna delle fasce d'età

Considerando la distribuzione anagrafica dei soggetti colpiti si può osservare una maggior incidenza associata agli individui di sesso maschile aventi età inferiore a 77 anni. Superata tale soglia si assiste invece ad un'inversione del fenomeno, che vede le persone di sesso femminile significativamente più interessate dal fenomeno.

Allo scopo di fornire una visione d'insieme del fenomeno appena descritto, si riporta in forma tabulare l'incidenza percentuale legata alle fasce d'età sopra definite:

Tabella 15 Incidenza dell'ictus per fasce d'età

FASCIA D'ETÀ	INCIDENZA PERCENTUALE MASCHILE [%]	INCIDENZA PERCENTUALE FEMMINILE [%]	DIFFERENZIALE DI INCIDENZA (MASCHILE – FEMMINILE) [%]
0-5	0,001	0,002	0,000
6-11	0,001	0,001	0,000
12-17	0,001	0,002	-0,001
18-23	0,005	0,003	0,002
24-29	0,006	0,008	-0,001
30-35	0,014	0,013	0,001

---

36-41	0,030	0,027	0,003
42-47	0,053	0,046	0,007
48-53	0,111	0,073	0,039
54-59	0,163	0,093	0,070
60-65	0,214	0,124	0,090
66-71	0,362	0,237	0,125
72-77	0,491	0,414	0,077
78-83	0,630	0,728	-0,097
84-89	0,499	0,812	-0,313
90-95	0,165	0,492	-0,328
96-101	0,017	0,073	-0,056
102-107	0,001	0,004	-0,003

Come si può notare il differenziale di incidenza si mantiene pressoché costante per entrambi i sessi al di sotto dei 47 anni mentre assume valori positivi nella macro-fascia dai 48 ai 77 anni, segno che gli uomini risultano più colpiti delle donne.

Si osserva altresì l'inversione di tendenza a partire dai 78, con il differenziale che cambia di segno e assume valori negativi segnalando una dominanza nell'incidenza femminile nella macro-fascia dai 78 ai 101 anni.

Si può inoltre osservare come per gli uomini l'incidenza maggiore si registri nella fascia d'età 78-83 anni, e per quella delle donne nella fascia 84-89.

## 3.2 Analisi temporale

Nello svolgere l'analisi temporale dei fenomeni di ictus nei tre anni a disposizione si è voluto indagare circa la possibile sussistenza di ciclicità, periodicità o stagionalità nell'insorgenza dei fenomeni.

Il primo passo dell'analisi consiste nella rappresentazione dei dati forniti da AREU in relazione all'orario di intervento dei soccorsi.

Di seguito si riportano in maniera comparativa i due grafici relativi alla rappresentazione aggregata dei dati e alla rappresentazione annuale degli stessi per quanto riguarda il triennio 2015-2017:

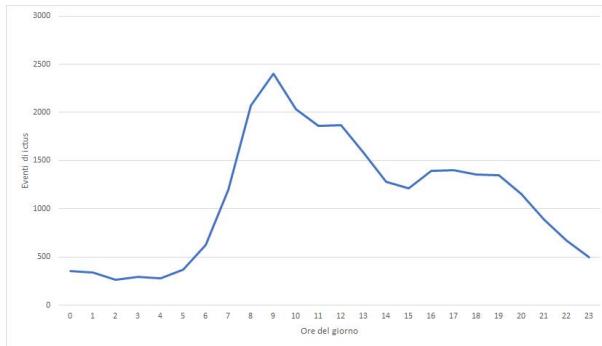


Figura 68 Distribuzione aggregata degli ictus in base all'orario di intervento dei soccorsi

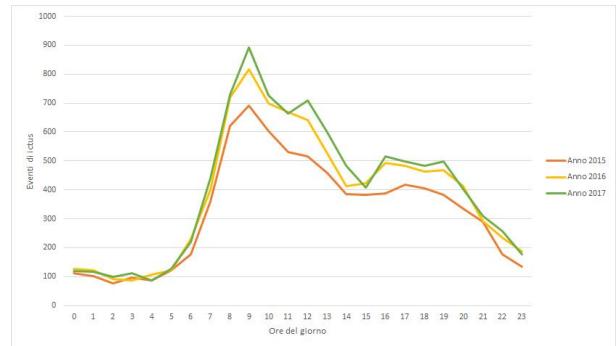


Figura 69 Distribuzione degli ictus in base all'orario di intervento dei soccorsi negli anni 2015-2017

Facendo seguito all'analisi oraria si è proseguito con la rappresentazione dei dati espressi giornalmente.

Si può osservare un picco di eventi, con relative chiamate al 112, nelle prime ore del mattino, probabilmente legato alla scoperta dell'ictus notturno alla ripresa delle attività quotidiane. Si rilevano inoltre una valle in corrispondenza delle prime ore pomeridiane ed un ulteriore aumento nella fascia oraria dalle 17 alle 19, con una successiva e progressiva diminuzione dei casi nelle ore serali. Tale pattern si è rivelato costante nei tre anni considerati.

Come nel caso precedente si è provveduto a fornire una rappresentazione aggregata del fenomeno, nonché una rappresentazione cumulativa esprimente la distribuzione giornaliera per gli anni compresi nel triennio 2015-2017.

Di seguito sono riportati i due grafici:

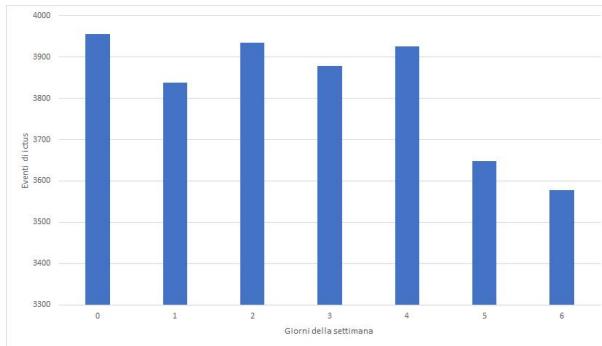


Figura 70 Distribuzione aggregata degli ictus in base al giorno di intervento dei soccorsi (0 = lunedì, 6 = domenica)

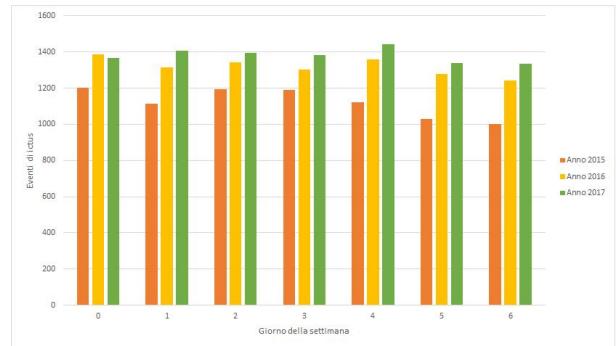


Figura 71 Distribuzione degli ictus in base al giorno di intervento dei soccorsi negli anni 2015-2017

Da tali grafici si può osservare il progressivo e globale aumento del numero di ictus in tutta la Lombardia nei tre anni considerati, con un numero minore di casi riscontrati nei giorni 5 e 6, corrispondenti relativamente ai sabati e alle domeniche.

L'ultimo spunto d'analisi è dato dalla ricerca di trend manifesti durante l'intero anno di riferimento.

I dati contenuti all'interno della base dati AREU sono stati raggruppati in base al mese di insorgenza della patologia e mostrati sia in forma aggregata sia in forma estesa per ognuna delle annualità oggetto d'indagine.

Di seguito si riportano i grafici in forma cumulativa ed in forma individuale per il triennio 2015-2017:

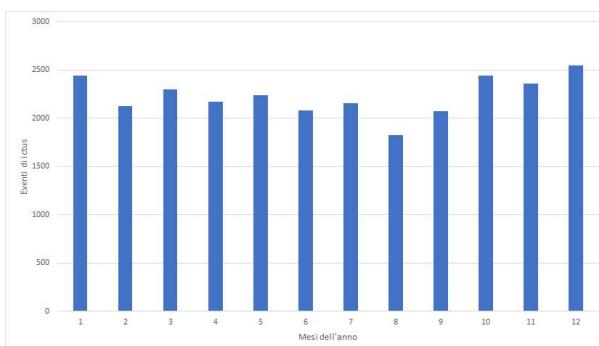


Figura 72 Distribuzione aggregata degli ictus in base al mese di intervento dei soccorsi (1 = gennaio, 12 = dicembre)

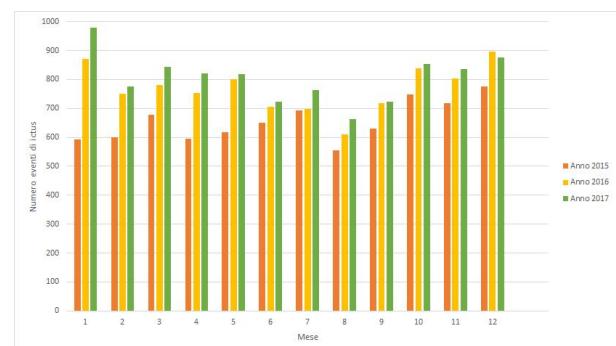


Figura 73 Distribuzione degli ictus in base al mese di intervento dei soccorsi negli anni 2015-2017

Si può osservare una diminuzione riscontrata nel mese di agosto, probabilmente legata ad una minor presenza della popolazione sul territorio di residenza, mentre si osserva un maggior numero di eventi associati ai mesi più freddi (ottobre-gennaio).

### 3.3 Analisi di correlazione

Il processo di analisi di correlazione mira a verificare l'esistenza di possibili relazioni sussistenti tra il numero di ictus e le variabili ambientali registrate sul territorio.

Al fine di conseguire risultati quanto più precisi si è scelto di ridurre il focus di interesse alla sola città di Milano in quanto offre un'ottima distribuzione delle centraline e garantisce al contempo una distribuzione degli ictus significativa, con 5193 casi registrati nei tre anni di interesse.

Di seguito si riportano le mappe riportanti le centraline dislocate sul territorio cittadino e gli eventi di ictus rilevati nel triennio 2015-2017 sul suolo meneghino:

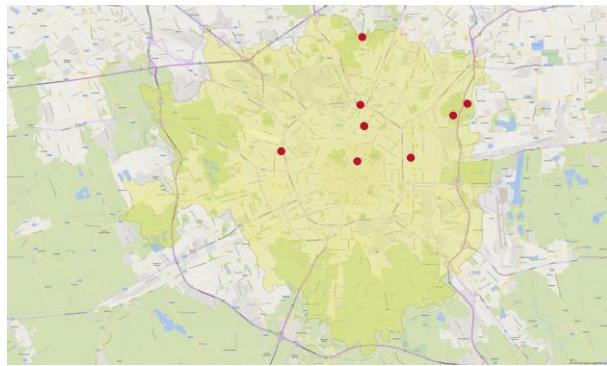


Figura 74 Dislocazione delle centraline meteorologiche nella città di Milano

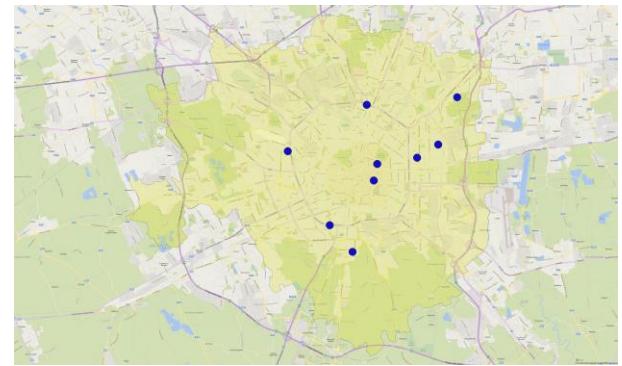


Figura 75 Dislocazione delle centraline per la qualità dell'aria nella città di Milano

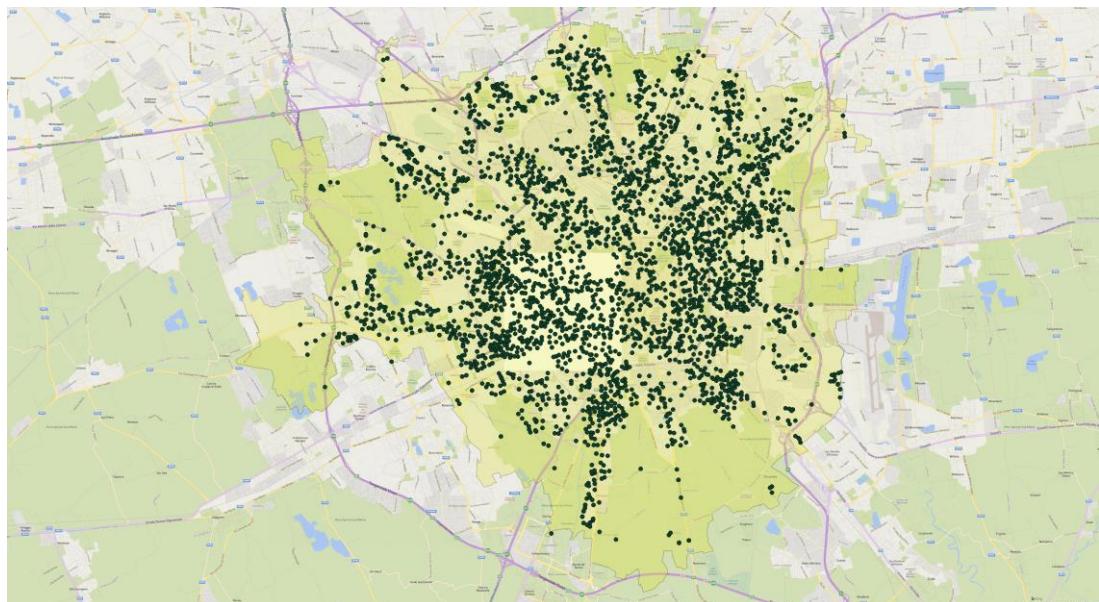


Figura 76 Distribuzione dei casi di ictus accertati dal 2015 al 2017 sul territorio della città di Milano

Dopo aver provveduto alla rappresentazione spaziale degli eventi si è proceduto con l'analisi di correlazione. Si sono ricavate le medie giornaliere istantanee, ossia le medie associate ai fenomeni misurati negli istanti della chiamata al 112, le medie giornaliere, calcolate dunque sulle 24 ore della giornata, e le medie mensili, associando poi questi dati alla distribuzione degli ictus nei rispettivi lassi temporali.

Di seguito si riporta la distribuzione degli ictus nel triennio d'indagine:

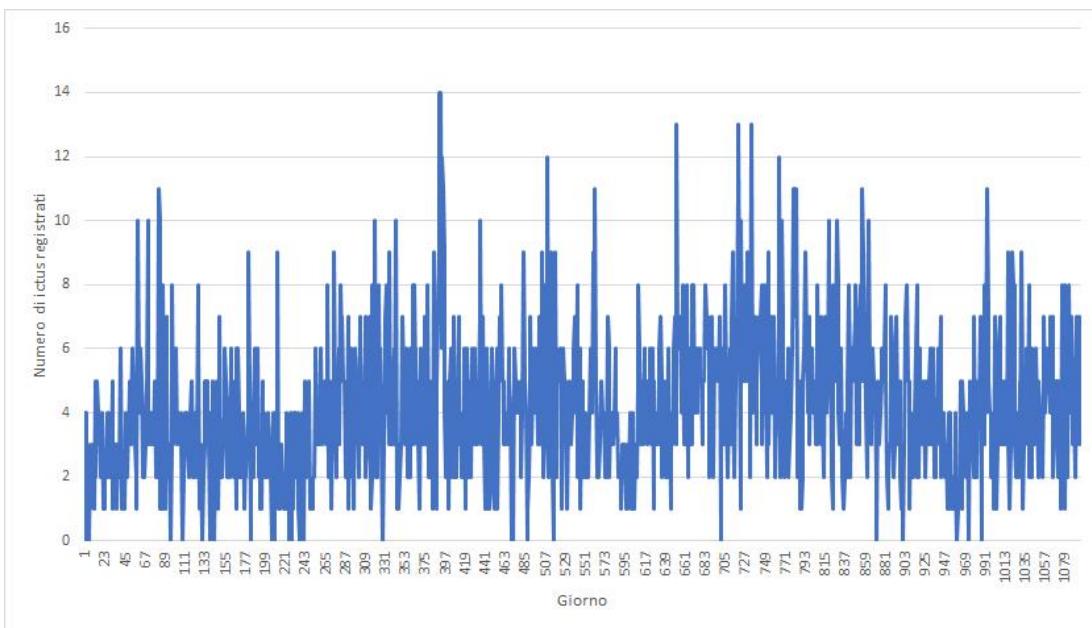


Figura 77 Distribuzione degli ictus nella città di Milano nel corso del triennio 2015-2017

Come detto, una prima analisi è stata svolta relazionando il numero di ictus registrati durante la giornata alla media dei fenomeni associati all'istante dell'insorgenza, estendendo il calcolo ai lag period precedenti. Questa media è calcolata solo sui valori istantanei delle misurazioni e non tiene conto dello storico giornaliero delle misure.

Per calcolare la correlazione tra ictus e fattori ambientali si è proceduto alla rappresentazione grafica della relazione e dato il dominio discreto della variabile “numero di ictus”, si è fatto ricorso alla correlazione a ranghi di Spearman al fine di ricavare la relazione tra le misure.

Di seguito si riportano i gli scatter plot relativi a ciascuna delle variabili in esame. I dati riportati sono relativi al lag period oH, ossia all'istante temporale in cui si è manifestato l'ictus. I successivi lag period presentano grafici paragonabili e si è dunque omessa la loro rappresentazione per non appesantire l'elaborato. La rappresentazione mira a porre in evidenza le differenze tra le diverse annualità ma si ricorda che il calcolo della correlazione è svolto sui dati aggregati al fine di avere un campione più numeroso ed una significatività più elevata.

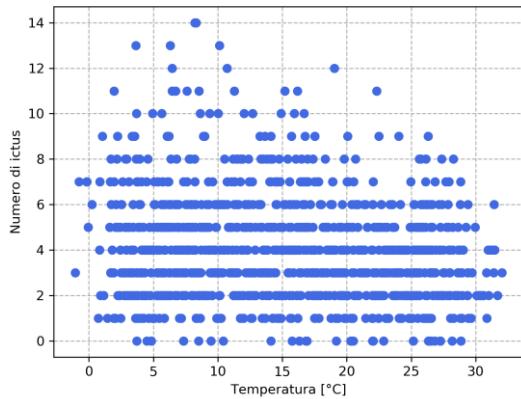


Figura 78 Grafico esprime la relazione ictus - temperatura

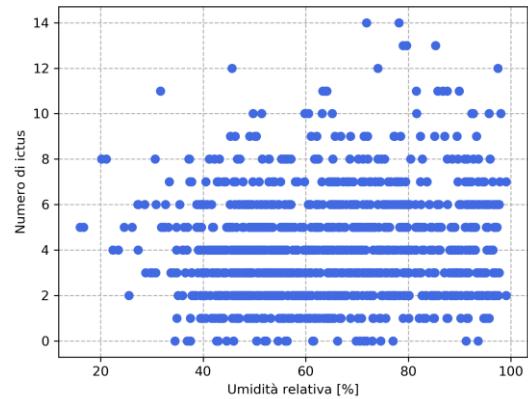


Figura 79 Grafico esprime la relazione ictus - umidità relativa

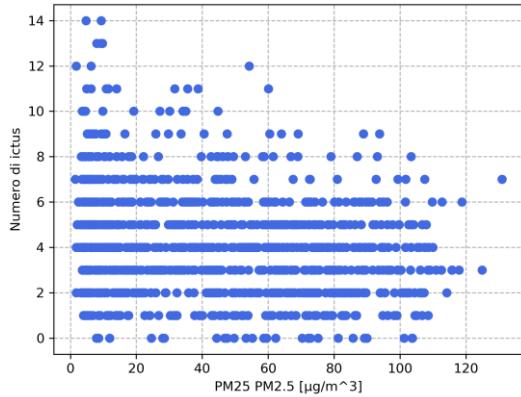


Figura 80 Grafico esprime la relazione ictus - ozono

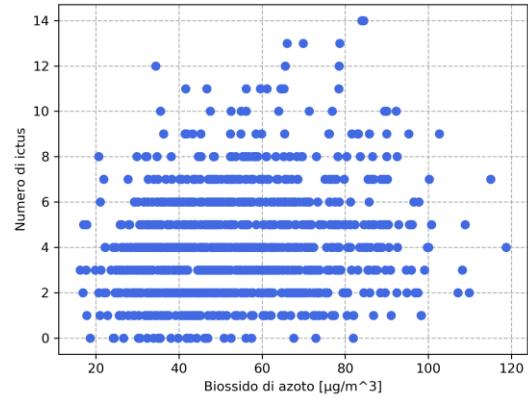


Figura 81 Grafico esprime la relazione ictus - biossido di azoto

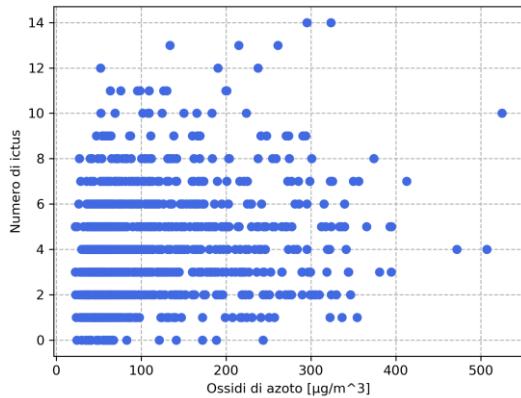


Figura 82 Grafico esprime la relazione ictus - ossidi di azoto

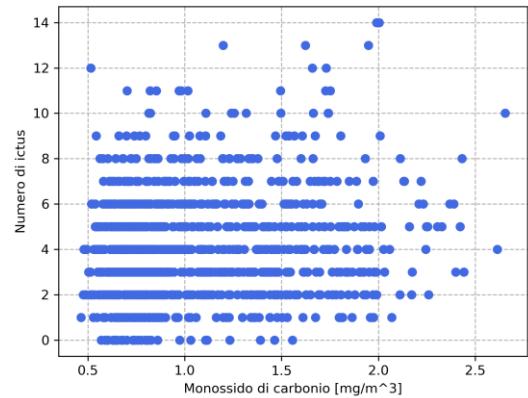


Figura 83 Grafico esprime la relazione ictus - monossido di carbonio

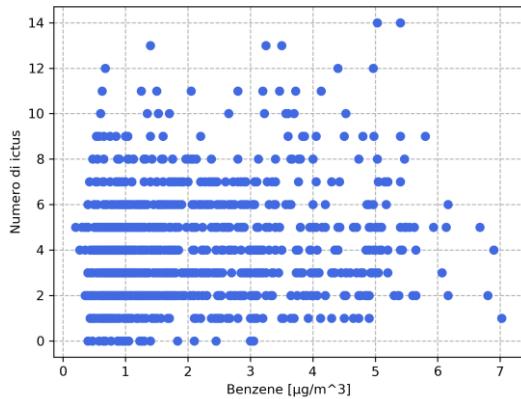


Figura 84 Grafico esprime la relazione ictus - benzene

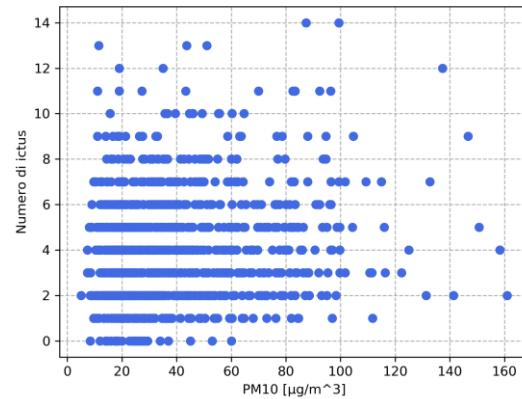


Figura 85 Grafico esprime la relazione ictus - PM10

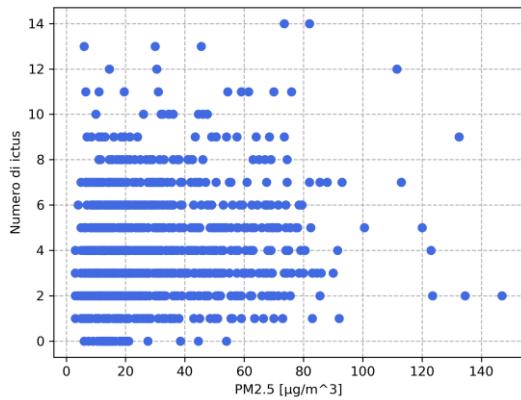


Figura 86 Grafico esprime la relazione ictus - PM2.5

Di seguito si riportano in forma tabulare i valori relativi alle correlazioni. Vengono sottolineate ed evidenziate le correlazioni significative, ossia con p-value < 0.05.

Tabella 16 Risultati dell’analisi di correlazione subitanea con le variabili in esame

FENOMENO	LAG PERIOD <b>0H</b>	LAG PERIOD <b>24H</b>	LAG PERIOD <b>48H</b>	LAG PERIOD <b>72H</b>	LAG PERIOD <b>96H</b>	LAG PERIOD <b>120H</b>
Temperatura	<b>-0.075</b>	<b>-0.091</b>	<b>-0.100</b>	<b>-0.106</b>	<b>-0.101</b>	<b>-0.104</b>
Umidità relativa	0.028	0.058	<b>0.066</b>	<b>0.067</b>	0.036	0.054
Ozono	<b>-0.071</b>	<b>-0.091</b>	<b>-0.073</b>	<b>-0.065</b>	-0.061	<b>-0.067</b>

Biossido di azoto	<b><u>0.122</u></b>	<b><u>0.093</u></b>	0.059	<b><u>0.080</u></b>	<b><u>0.068</u></b>	<b><u>0.068</u></b>
Ossidi di azoto	<b><u>0.141</u></b>	<b><u>0.110</u></b>	<b><u>0.103</u></b>	<b><u>0.098</u></b>	<b><u>0.106</u></b>	<b><u>0.093</u></b>
Monossido di carbonio	<b><u>0.103</u></b>	<b><u>0.101</u></b>	<b><u>0.076</u></b>	0.058	0.053	0.060
Benzene	<b><u>0.106</u></b>	<b><u>0.143</u></b>	<b><u>0.125</u></b>	<b><u>0.099</u></b>	<b><u>0.094</u></b>	<b><u>0.094</u></b>
PM <sub>10</sub>	<b><u>0.069</u></b>	<b><u>0.077</u></b>	<b><u>0.065</u></b>	0.037	0.042	0.020
PM <sub>2.5</sub>	<b><u>0.073</u></b>	<b><u>0.084</u></b>	<b><u>0.072</u></b>	0.041	0.047	0.027

Da questa tabella si notano alcune variabili correlate negativamente con il numero di ictus giornalieri (temperatura e ozono), mentre le altre sono correlate positivamente.

Si osservano inoltre differenti dipendenze temporali, con alcune variabili maggiormente correlate al valore attuale (biossido di azoto, ossidi di azoto e monossido di carbonio), altre invece maggiormente correlate ad un effetto cumulativo nelle 24 h (ozono, benzene, PM).

Dopo aver analizzato la correlazione con le misurazioni istantanee si è deciso di approfondire la relazione di correlazione con le medie giornaliere della città anziché i valori puntuali, sempre facendo uso della tecnica sviluppata da Spearman dato il dominio discreto della variabile “numero di ictus”.

Di seguito si riportano i risultati ottenuti per quanto riguarda il triennio 2015-2017. Come nel caso precedente i dati sono proposti in forma aggregata al fine di avere una numerosità del campione che consenta di avere risultati significativi. I grafici, risultando in tutto paragonabili a quanto già mostrato in precedenza, sono stati omessi per brevità. Vengono sottolineate ed evidenziate le correlazioni significative, ossia con p-value < 0.05.

Tabella 17 Risultati dell’analisi di correlazione con la media giornaliera delle variabili in esame

FENOMENO	LAG PERIOD oH	LAG PERIOD 24H	LAG PERIOD 48H	LAG PERIOD 72H	LAG PERIOD 96H	LAG PERIOD 120H
Temperatura	<b><u>-0.163</u></b>	<b><u>-0.172</u></b>	<b><u>-0.175</u></b>	<b><u>-0.172</u></b>	<b><u>-0.171</u></b>	<b><u>-0.176</u></b>
Umidità relativa	<b><u>0.092</u></b>	<b><u>0.138</u></b>	<b><u>0.168</u></b>	<b><u>0.145</u></b>	<b><u>0.109</u></b>	<b><u>0.121</u></b>
Ozono	<b><u>-0.201</u></b>	<b><u>-0.188</u></b>	<b><u>-0.196</u></b>	<b><u>-0.193</u></b>	<b><u>-0.181</u></b>	<b><u>-0.176</u></b>
Biossido di azoto	<b><u>0.182</u></b>	<b><u>0.137</u></b>	<b><u>0.111</u></b>	<b><u>0.108</u></b>	<b><u>0.090</u></b>	<b><u>0.098</u></b>

---

Ossidi di azoto	<b><u>0.204</u></b>	<b><u>0.163</u></b>	<b><u>0.146</u></b>	<b><u>0.153</u></b>	<b><u>0.135</u></b>	<b><u>0.150</u></b>
Monossido di carbonio	<b><u>0.137</u></b>	<b><u>0.104</u></b>	<b><u>0.104</u></b>	<b><u>0.093</u></b>	<b><u>0.068</u></b>	<b><u>0.079</u></b>
Benzene	<b><u>0.158</u></b>	<b><u>0.156</u></b>	<b><u>0.150</u></b>	<b><u>0.134</u></b>	<b><u>0.137</u></b>	<b><u>0.148</u></b>
PM <sub>10</sub>	<b><u>0.127</u></b>	<b><u>0.133</u></b>	<b><u>0.104</u></b>	<b><u>0.086</u></b>	<b><u>0.077</u></b>	<b><u>0.082</u></b>
PM <sub>2.5</sub>	<b><u>0.123</u></b>	<b><u>0.137</u></b>	<b><u>0.115</u></b>	<b><u>0.098</u></b>	<b><u>0.093</u></b>	<b><u>0.100</u></b>

In questo caso le correlazioni mostrano valori maggiori che nella tabella precedente. Si notano alcune variabili correlate negativamente con il numero di ictus giornalieri (temperatura e ozono), mentre le altre sono correlate positivamente.

Si notano inoltre differenti dipendenze temporali, con alcune variabili maggiormente correlate al valore attuale (ozono, biossido di azoto, ossidi di azoto e monossido di carbonio), altre invece maggiormente correlate ad un effetto cumulativo nelle 24h (benzene e PM) o 48h (umidità relativa).

Come ultimo spunto d'indagine circa le medie giornaliere si è deciso di approfondire il legame tra ictus e variabili ambientali andando a suddividere la media giornaliera in bin e riferendo a ciascuno di essi il numero di ictus associato. Nei grafici che seguono si mostra la distribuzione degli eventi in relazione ai 9 fenomeni indagati e si riporta la distribuzione degli stessi al fine di offrire una prospettiva circa la normale distribuzione delle variabili e della loro correlazione con i bin, calcolata sempre con il metodo a ranghi di Spearman.

Nei vari grafici viene inoltre indicato il numero di giorni in cui si è misurato uno specifico valore del parametro sotto indagine, così da poter valutare se i picchi di eventi, associabili a determinati valori, non siano legati al fatto che quei valori del parametro risultano più comunemente misurati.

Quanto riportato afferisce alla media giornaliera, reputata più significativa di quella istantanea, per l'istante OH. Si lascia al paragrafo in appendice la rappresentazione dei restanti lag period. Vengono sottolineate ed evidenziate le correlazioni significative, ossia con p-value < 0.05.

Tabella 18 Risultati della correlazione tra la distribuzione degli ictus e la distribuzione delle variabili

FENOMENO	AMPIEZZA BIN	NUMERO DI BIN	R ICTUS
	1 [°C]	38	-0.17
	5 [%]	20	<b>0.63</b>
	5 [µg/m³]	28	<b>-0.47</b>

Figura 87 Distribuzione degli ictus e della temperatura in funzione dell'aggregazione in bin

Figura 88 Distribuzione degli ictus e dell'umidità relativa in funzione dell'aggregazione in bin

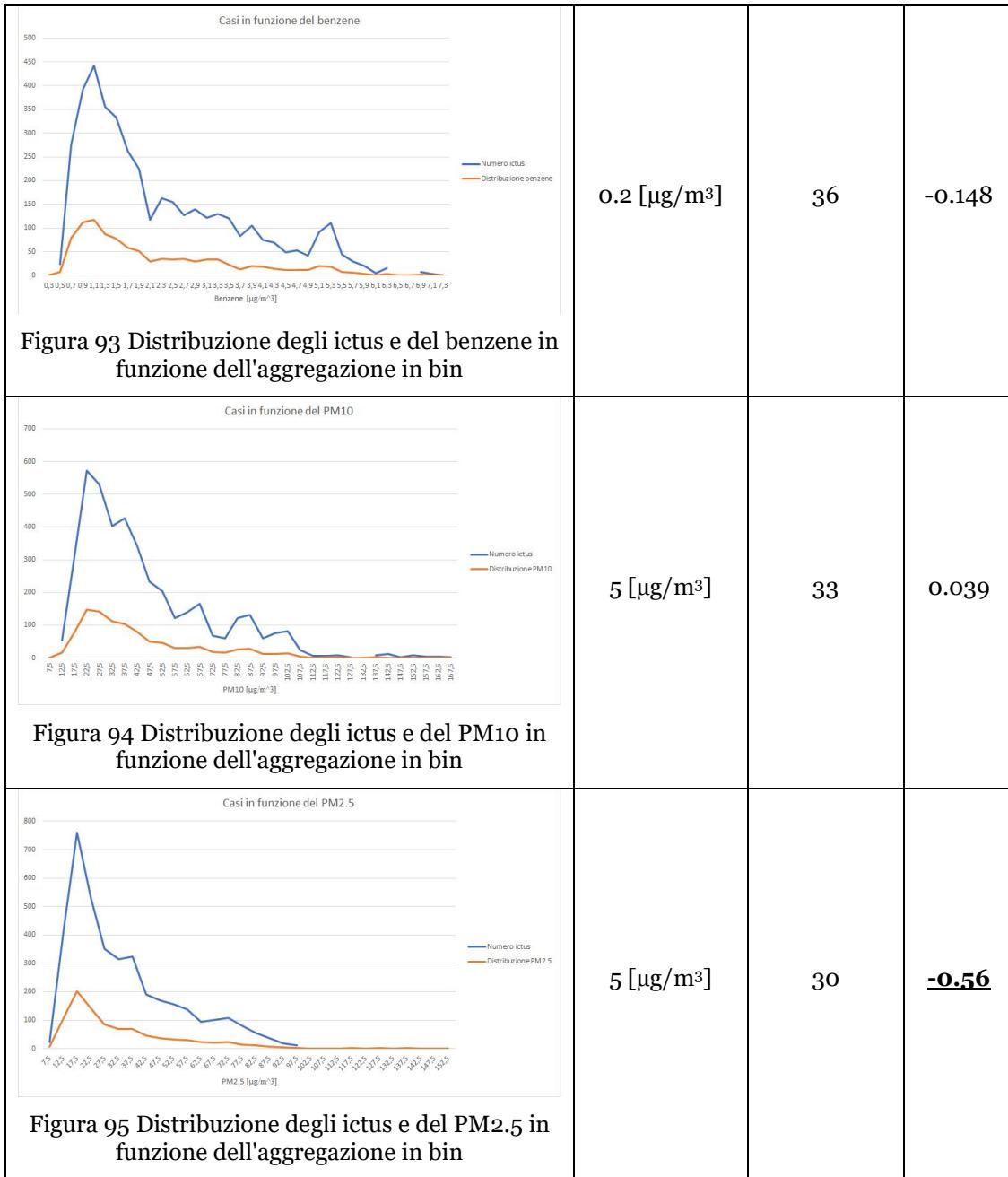
Figura 89 Distribuzione degli ictus e dell'ozono in funzione dell'aggregazione in bin

<p><b>Casi in funzione del biossido di azoto</b></p> <p>Y-axis: Numero ictus (blue line), Distribuzione NO2 (orange line)</p> <p>X-axis: Biossido di azoto [<math>\mu\text{g}/\text{m}^3</math>] (bins: 15, 30, 45, 60, 75, 90, 105, 120, 135)</p>	5 [ $\mu\text{g}/\text{m}^3$ ]	26	0.005
<p><b>Casi in funzione degli ossidi di azoto</b></p> <p>Y-axis: Numero ictus (blue line), Distribuzione NOx (orange line)</p> <p>X-axis: Ossidi di azoto [<math>\mu\text{g}/\text{m}^3</math>] (bins: 30, 50, 70, 90, 110, 130, 150, 170, 190, 210, 230, 250, 270, 290, 310, 330, 350, 370, 390, 410, 430, 450, 470, 490, 510, 530, 550)</p>	20 [ $\mu\text{g}/\text{m}^3$ ]	26	-0.227
<p><b>Casi in funzione del monossido di carbonio</b></p> <p>Y-axis: Numero ictus (blue line), Distribuzione CO (orange line)</p> <p>X-axis: Monossido di carbonio [<math>\text{mg}/\text{m}^3</math>] (bins: 0,150, 0,350, 0,450, 0,550, 0,650, 0,750, 0,850, 0,950, 0,051, 0,151, 0,251, 0,351, 0,451, 0,551, 0,651, 0,751, 0,851, 0,952, 0,052, 0,152, 0,252, 0,352, 0,452, 0,552, 0,652, 0,752, 0,852, 0,953, 0,053)</p>	0.1 [ $\text{mg}/\text{m}^3$ ]	30	-0.002

**Figura 90 Distribuzione degli ictus e del biossido di azoto in funzione dell'aggregazione in bin**

**Figura 91 Distribuzione degli ictus e degli ossidi di azoto in funzione dell'aggregazione in bin**

**Figura 92 Distribuzione degli ictus e del monossido di carbonio in funzione dell'aggregazione in bin**



Come si può notare non emerge alcun tipo di correlazione tra i bin ed il numero di ictus, mentre appare forte la correlazione tra la distribuzione degli eventi e la distribuzione dei fenomeni, con valori prossimi o superiori a 0.98 in tutti i casi.

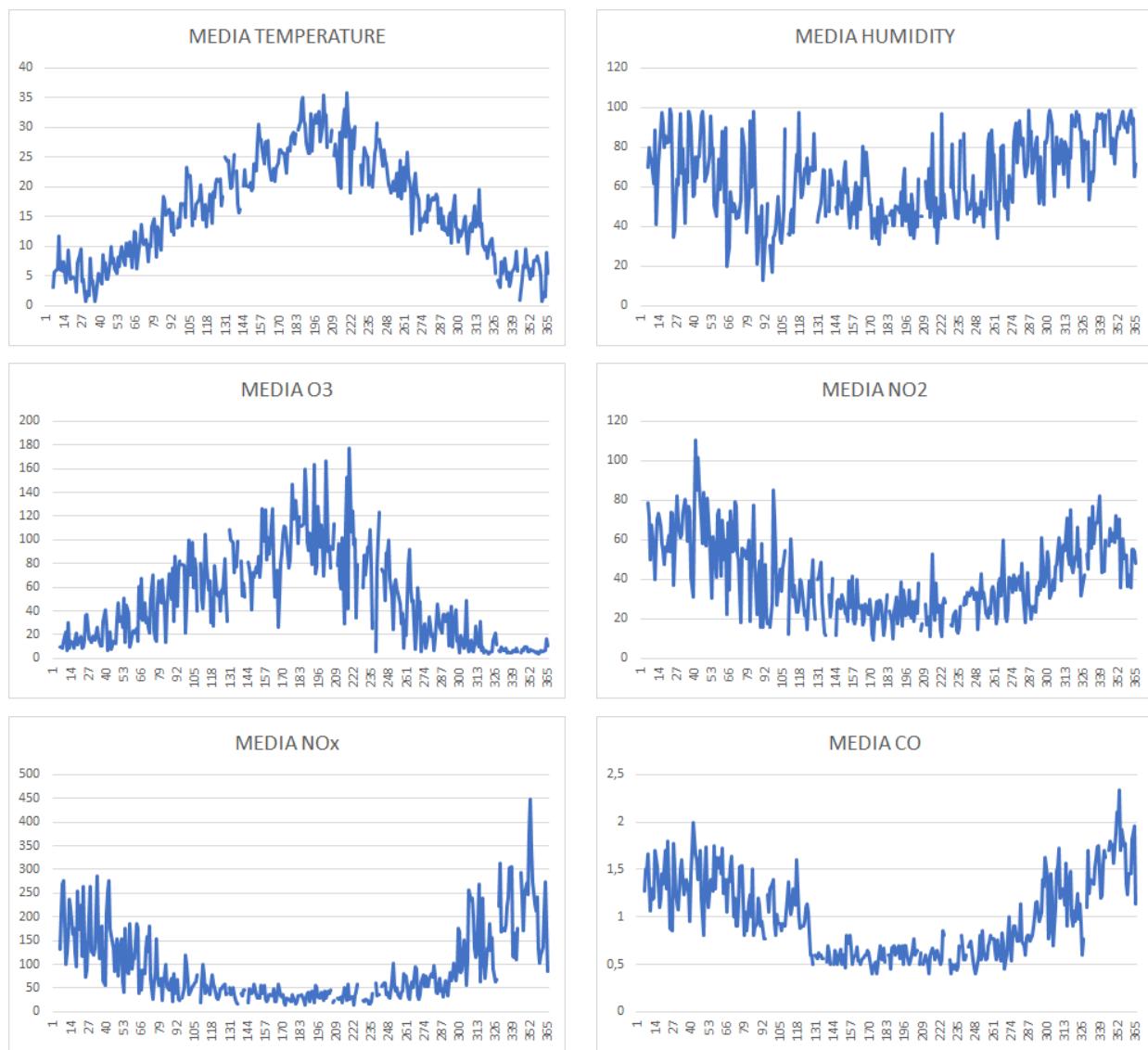
Si osserva come alcune correlazioni risultino significative e moderate, seppur negative. Ciò indica che al crescere del valore dei bin non corrisponde un relativo aumento nel conteggio

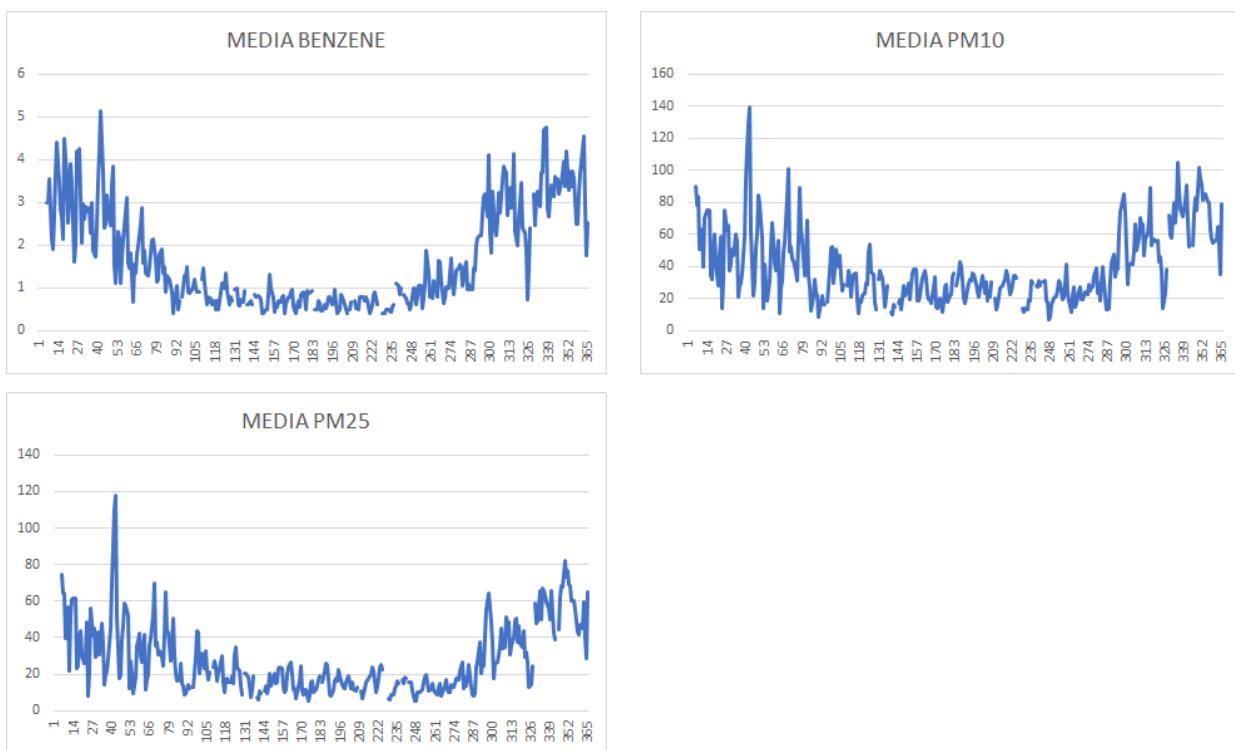
dei casi di ictus, mettendo in luce che non sono gli alti valori delle variabili a determinare direttamente effetti negativi sulla salute.

L'unica eccezione alla negatività delle correlazioni è data dall'umidità relativa. Quanto si osserva è spiegabile in base al fatto che l'umidità contribuisca al ristagno di tutti i fenomeni atmosferici, studiati e non, che in maniera cumulata contribuiscono a generare una correlazione positiva non necessariamente ascrivibile al singolo fattore umidità.

Quanto descritto dimostra che a livello giornaliero un numero di ictus maggiore è semplicemente da associare ad una media dei fenomeni più frequente durante l'anno.

Di seguito si mostrano esempi della distribuzione calcolata sulla media dei valori interpolati all'istante oH. I dati sono espressi sulla singola annualità data la ciclicità dei fenomeni:





In ultimo luogo si è proceduto ad un’analisi di correlazione lineare per indagare l’esistenza di possibili relazioni che leghino gli eventi di ictus aggregati su base mensile alle medie mensili dei fenomeni meteorologici e degli agenti inquinanti indagati, al fine di determinare se l’insorgenza della patologia sia attribuibile a periodi di esposizione più lunghi rispetto al lag period di 5 giorni noto in letteratura.

Di seguito si riportano le correlazioni individuate in base ai fenomeni di riferimento:

Tabella 19 Risultati dell’analisi di correlazione con le medie mensili delle variabili in esame

FENOMENO	R	SIGNIFICATIVITÀ
Temperatura	-0.49	0.0025
Umidità relativa	0.41	0.0139
Ozono	-0.504	0.0017
Biossido di azoto	0.40	0.0145
Ossidi di azoto	0.48	0.0033
Monossido di carbonio	0.22	0.1981
Benzene	0.48	0.0034
PM10	0.43	0.0092
PM2.5	0.41	0.0133

---

Tale analisi ha messo in evidenza valori di correlazione in modulo più elevati che nei brevi periodi prima considerati, segno che l'influenza delle variabili ambientali sull'insorgenza di ictus sia maggiormente connessa a fenomeni legati ad esposizioni prolungate alla variabile considerata e non a variazioni repentine. Inoltre, si conferma una correlazione negativa con temperatura ed ozono, e positiva con tutte le altre variabili, escluso il monossido di carbonio ove la significatività non viene raggiunta.

---

# **CAPITOLO 4**

## **DISCUSSIONE E CONCLUSIONI**

### **4.1 Valutazione del processo di interpolazione e del campione di riferimento**

Il primo spunto di discussione nasce naturalmente dalle considerazioni fatte per quanto riguarda la necessità di interpolare i dati registrati sul territorio allo scopo di creare un modello di interpolazione che consenta di attribuire ai record clinici informazioni puntuale in merito alle variabili oggetto di indagine.

Nel caso trattato si è fatto uso dei dati ambientali raccolti da ARPA ed in particolare si è indagato in merito a nove fattori: temperatura, umidità relativa, ozono, biossido di azoto, ossidi di azoto, monossido di carbonio, benzene, PM10 e PM2.5.

Sfruttando questi dati è stato possibile ricondurre ad un evento clinico geolocalizzato tutte le informazioni ad esso relative per quanto riguarda le variabili appena citate. Ciò è permesso grazie all'impiego di un algoritmo di interpolazione noto come Kriging, che a partire dalla correlazione spaziale tra i dati atmosferici genera un modello in grado di descrivere territorialmente ed in maniera continua i diversi fenomeni ambientali, per ogni possibile coppia di coordinate (X, Y) nella regione.

Il processo di validazione del modello di Kriging ha permesso di mettere in luce la bontà dell'approssimazione matematica creata, che, come mostrato, è in grado di raggiungere precisioni estremamente elevate, dimostrandosi un ottimo strumento di rappresentazione della realtà.

---

Come osservato, le stazioni di monitoraggio dell'ARPA dislocate sul suolo regionale dimostrano avere un campionamento adeguato solamente per alcuni dei fenomeni indagati. Temperatura, umidità relativa e ossidi di azoto su tutti palesano l'eccellente capillarità delle centraline mentre se si cerca di raccogliere dati su benzene, polveri sottili e monossido di carbonio la rete di monitoraggio sembra mostrare il fianco alla penuria di punti di rilevamento in quanto tali misurazioni risultano disponibili limitatamente ai capoluoghi di provincia.

Per questo motivo si è deciso di restringere il processo di analisi alla sola città di Milano, che essendo coperta in maniera brillante dai sensori ambientali si rivela un ottimo terreno di studio, data anche l'ingente mole di casi di ictus intercettati da AREU nella città meneghina. Questa limitazione dovuta alla scarsità di centraline ha portato necessariamente ad un abbattimento dei record utilizzabili all'interno del database dei casi di ictus, ridotti da un totale di 30759 per l'intera Lombardia a 4180 (14% circa) per la città di Milano. Ciò, tuttavia, consente di innalzare la qualità dei risultati ottenuti rendendoli più affidabili poiché il modello assume un'elevata precisione nei luoghi in cui i rilevamenti si fanno più densi (in quanto i dati divengono relativi ad un territorio più ristretto e con caratteristiche simili), come mostrato anche nel capitolo inerente alla validazione dei modelli di Kriging.

Comparando poi la dimensione del campione sotto indagine con i numeri presentati in letteratura, si nota una certa affinità con quanto analizzato da [Hong Y. et al. 2002], il cui studio si riconduce alla città di Seoul e all'osservazione di 7137 casi di ictus nell'arco di 7 anni. Cifre simili per il campione analizzato sono riscontrabili anche negli studi di [Gregory A. W. et al. 2005], in particolare per quanto riguarda i centri urbani minori come Birmingham (6642 casi tra il 1986 e il 1993), Minneapolis (6903 casi tra il 1987 e il 1997), New Heaven (7546 casi tra il 1988 e il 1999), Salt Lake City (2559 casi tra il 1986 e il 1999) e Seattle (7190 casi tra il 1986 e il 1995).

Sempre in termini di dimensione del campione di riferimento, le pubblicazioni di [Tsai S. et al. 2003] e [Guo P. et al. 2017] godono invece di una popolazione analizzata di gran lunga superiore, pari relativamente a 24.000 casi (in 4 anni dal 1997 al 2000 a Kaohsiung, Taiwan) e ad oltre 35.000 casi nel solo 2015 (Guangzhou, Cina). Una sproporzione del genere nel numero dei record a disposizione rende un confronto solo parzialmente sovrapponibile.

In tutte le pubblicazioni analizzate in letteratura non emerge alcun tentativo di geolocalizzare gli eventi di ictus e di fatto non vengono citate interpolazioni puntuali dei valori per ogni singolo caso. Si parla infatti in tutti i casi di medie cittadine, verosimilmente riconducibili al medesimo genere di analisi effettuata in questo elaborato sulla media giornaliera dei valori letti dalle centrali. Non si hanno riferimenti in merito al numero e alla distribuzione dei

---

sensori utilizzati in letteratura e ciò impossibile un confronto diretto circa la loro distribuzione sul territorio e dunque sulla rappresentatività di cui godono i dati presentati.

Il framework qui proposto consente invece di approfondire nel dettaglio le condizioni ambientali presenti nel luogo e nell'ora in cui si è registrato il caso di ictus, permettendo un'analisi temporale e spaziale estremamente dettagliata in merito alle circostanze che hanno condotto all'insorgenza dei sintomi. La metodologia proposta garantisce inoltre la possibilità di esplorare grandezze temporali estremamente variegate tra di loro, spaziando dal riferimento orario, a quello giornaliero ed in infine a quello mensile. Ciò consente di valutare gli effetti di un evento clinico in relazione al suo evolversi nel tempo, una novità in letteratura.

Per quanto riguarda le variabili adottate per lo svolgimento delle analisi e la lunghezza dei lag period, si è fatto particolare riferimento agli scritti di [Hong Y. et al. 2002], [Tsai S. et al. 2003], [Gregory A. W. et al. 2005] e [Guo P. et al. 2017], prendendo come riferimento i fenomeni meteorologici e gli agenti inquinanti già indiziati di correlazione al fine di verificare l'effettiva sussistenza delle relazioni di causalità.

Nelle pubblicazioni lette, il riferimento ai lag period analizzati varia tra i due e i cinque giorni. Per completezza di è scelto l'estremo superiore tra i due valori.

## **4.2 Valutazione del processo di analisi demografica e temporale**

A differenza di quanto mai fatto in letteratura, questa tesi ha preso in esame anche le componenti demografica e temporale nello studio dell'ictus in relazione al territorio.

Ciò che emerge dalle analisi è che il numero di ictus è in costante crescita negli ultimi anni, in linea con quanto confermano i trend nazionali ed internazionali, e si riscontra un'incidenza alle volte più che doppia nelle città maggiormente industrializzate.

Gli uomini sembrano essere i più colpiti nella fascia d'età compresa tra i 48 e i 77 anni mentre a partire dai 78 anni in poi le donne fanno registrare il maggior numero di casi.

Per quanto riguarda i maschi, si osserva un picco di incidenza in corrispondenza della fascia d'età tra i 78 e gli 83 anni mentre la maggior parte dei casi ricade nel range 66 – 89 anni. La distribuzione relativa alle femmine mostra invece un picco associato al bin 84-89 anni, concentrando altresì la maggior parte dei casi nella macro-fascia 77 – 101 anni.

Da questi dati si può notare uno scostamento di 5-6 anni che separa le due età a maggior vulnerabilità per entrambi i sessi. In letteratura è nota la maggior incidenza associabile agli individui over 65, che come mostrano i grafici risultano effettivamente i più interessati dal

---

fenomeno. Guardando i grafici dell’incidenza percentuale calcolata sul triennio 2015-2017 si nota poi come l’incidenza sulla popolazione femminile sia costantemente superiore a quella maschile.

Questi dati trovano pareri discordanti nella letteratura scientifica, che sembra dividersi circa la predominanza dell’incidenza maschile piuttosto che femminile. A tal proposito si ipotizza l’esistenza di una certa variabilità da attribuire alle differenti aree geografiche e ai diversi stili di vita delle popolazioni coinvolte.

I dati ottenuti mostrano come l’incidenza dell’ictus in Lombardia oscilli tra lo 0.06% e lo 0.11% a seconda delle province, con una propensione agli ictus individuata principalmente nelle città di Como, Lecco, Milano e Varese, noti poli industriali della regione.

Quanto misurato si pone ben al di sotto della media nazionale [[http://www.salute.gov.it/portale/salute/p1\\_5.jsp?lingua=italiano&id=28&area=Malattie\\_cardiovasscolari](http://www.salute.gov.it/portale/salute/p1_5.jsp?lingua=italiano&id=28&area=Malattie_cardiovasscolari)], che si attesta a circa 195.000 casi l’anno, per un’incidenza totale che si aggira attorno allo 0.3%.

Le medie europee del fenomeno variano invece da un minimo di 30 casi su 100.000 abitanti (pari ad un’incidenza dello 0.03%) ad un massimo di 170 casi su 100.000 abitanti (pari allo 0.17%) [<https://www.osservatorioictusitalia.it/pubbllicazioni/impatto-delictus-in-europa/>]. Confrontando queste misure con quanto ottenuto in fase di analisi emerge una evidente similarità ad unire incidenza lombarda ed europea, che si dimostrano assolutamente in linea. Passando infine al valutare i dati sull’incidenza mondiale, si osservano valori globali compresi tra lo 0.14% e lo 0.37% [[http://www.iso-spread.it/capitoli/capitolo\\_04.pdf](http://www.iso-spread.it/capitoli/capitolo_04.pdf)], segno che la Lombardia non risulta poi così interessata dal fenomeno, a differenza dell’Italia presa nella sua interezza.

Per quanto riguarda invece l’analisi temporale, si è deciso di analizzare i fenomeni con tre granularità differenti: oraria, giornaliera e mensile.

L’analisi oraria mostra un evidente picco di interventi nelle prime ore della giornata, attribuibili verosimilmente alla scoperta dei sintomi di un ictus occorso durante la notte. Il trend vede una flessione in prossimità del primo pomeriggio, per poi riprendere leggermente corpo prima di scemare a fine giornata. Il risultato è in linea con quanto ci si possa aspettare da un vasto range di patologie, i cui sintomi emergono solo al risveglio dell’individuo.

L’informazione più interessante è forse quella relativa al trend settimanale, che mostra un’incidenza superiore (8%) durante l’arco dei 5 giorni lavorativi piuttosto che nel weekend, segno che sussiste un fenomeno di base al quale si somma una componente più rilevante da ricercarsi all’interno dei luoghi e delle attività svolte nell’arco della settimana lavorativa. In merito a tale evidenza non si riscontrano tuttavia spiegazioni deducibili dai dati in possesso.

Di interesse è anche la distribuzione annua del fenomeno, che vede una concentrazione di ictus superiore nei mesi più freddi e notoriamente più inquinati dell’anno, sempre in netta

---

antitesi con quanto sostenuto da [Lavados P. M. 2017]. Il trend sembra seguire quello vissuto dai principali fenomeni inquinanti quali ossidi di azoto e particolato, attivi principalmente durante l'inverno a causa dell'inquinamento causato dal riscaldamento domestico e dal traffico urbano.

## **4.3 Valutazione del processo di analisi di correlazione**

Lo scopo di questo progetto di tesi è quello di sviluppare un framework di Health Geomatics in grado di integrare dati clinici geolocalizzati con dati ambientali, così da investigare la loro possibile relazione.

L'approccio prevede di utilizzare i dati ambientali forniti dalle centraline ARPA presenti sul territorio della Lombardia al fine di generare un modello di interpolazione che consenta di attribuire ad ogni evento di ictus, che si ricorda essere geolocalizzato sul territorio regionale, un ben preciso riferimento numerico per ciascuna delle variabili ambientali di interesse.

Come suggerito in letteratura [Hong Y. et al. 2002, Guo P. et al. 2017] si è scelto di indagare i fenomeni non solo in relazione all'ora zero in cui essi si verificano ma anche in funzione di un lag period precedente. Le due pubblicazioni sopra citate suggeriscono una finestra temporale, di 4 e 5 giorni rispettivamente, entro cui sono state notate relazioni di correlazione tra ictus e fenomeni ambientali. Al fine di catturare quante più informazioni possibili, si è dunque scelto di operare usando un lag period di 5 giorni, così da poter confrontare in seguito i risultati ottenuti.

Ciò che emerge dalle analisi, effettuate per ogni lag temporale antecedente all'evento e per l'orario dell'evento stesso, è una correlazione debole sia con gli inquinanti sia con i parametri meteorologici.

La metodologia d'analisi seguita in questo progetto di tesi ha portato inoltre ad approfondire periodi di esposizione più lunghi rispetto a quanto sia mai stato fatto in letteratura. Si è infatti correlata la media di ictus di ogni mese con le corrispettive medie mensili per ognuna delle variabili in considerazione. In questo caso, come in quello delle medie giornaliere, si è però associato il numero di ictus aggregati su base mensile alla media dei valori rilevati dalle centraline, senza ricondurre le misurazioni specifiche ad ogni singola posizione. Questo è stato fatto per supportare l'ipotesi di dinamicità, che vede gli individui spostarsi all'interno della città e dunque rimanere esposti a valori di particolato variabili in base alla zona. Una media puntuale e statica non sarebbe stata in grado di catturare il fenomeno nella sua interezza.

---

Allo scopo di fornire un dettaglio maggiore ed una miglior comprensione dei risultati, si riportano ora le conclusioni ed i confronti con quanto noto in letteratura per ogni singola variabile ambientale:

**Temperatura:** i risultati delle analisi effettuate sulla temperatura mostrano una debole correlazione negativa con i dati sugli ictus. Il risultato trova pieno accordo con quanto rilevato da [Hong Y. et al. 2002], che riporta allo stesso modo l'esistenza di una correlazione negativa. Per quanto riguarda gli studi condotti da [Tsai S. et al. 2003], sebbene gli autori decidano di includere la temperatura nello studio e la usino per filtrare i record ( $> 0 < 20^{\circ}\text{C}$ ), nulla viene concluso in merito al suo impatto sull'insorgenza di ictus ischemico. Verosimilmente la variabile è stata impiegata come mezzo arbitrario per la selezione dei record più significativi data la sua correlazione inversa con il caso clinico in esame.

Anche gli studi condotti da [Gregory A. et al. 2005], sebbene includano la variabile nel processo d'analisi, non giungono a nessuna conclusione in merito alla sua influenza, segno che verosimilmente non sia stato trovato nulla di significativo. Le medesime osservazioni valgono per quanto studiato da [Guo P. et al. 2017].

Si registra invece un totale disaccordo con quanto proposto da [Lavados P. M. et al. 2018], che evidenzia un significativo aumento del rischio associato alle alte temperature. La dissonanza accomuna anche le restanti pubblicazioni citate e fa propendere per una metodologia meritevole di riconferma dati i risultati ampiamente fuori dal coro.

**Umidità relativa:** i risultati circa l'influenza dell'umidità relativa forniscono valori di correlazione molto bassi, quando significativi. Si osserva una lieve predominanza dei valori misurati per i lag period corrispondenti a 2 o 3 giorni ma la correlazione resta molto debole. Gli studi avanzati da [Hong Y. et al. 2002, Tsai S. et al. 2003] utilizzano la variabile unicamente per aggiustare le misurazioni degli altri inquinanti mentre [Gregory A. et al. 2005] utilizza questo fenomeno per modulare il valore della temperatura apparente.

In nessuno dei tre casi si fa uso della misura, che verosimilmente non trova correlazione con il fenomeno ictus. Prova ne è data dal fatto che non pervengano altre pubblicazioni a farne menzione.

**Ozono:** i risultati delle analisi effettuate sull'ozono mostrano una correlazione negativa che si sposa ai valori di correlazione della temperatura, anch'essa in egual modo negativa. Le due variabili risultano strettamente legate tra di loro e mostrano un andamento analogo durante tutti i lag period. Giustificazione è data dal fatto che a giornate più calde si accompagnano solitamente cieli tersi e valori di ozono più elevati.

Conferma di questa scoperta è data da [Hong Y. et al. 2002], che ribadisce la negatività della correlazione.

---

Nelle analisi svolte da [Tsai S. et al. 2003] si registra invece un'associazione positiva e statisticamente significativa tra ozono ed ictus, in totale disaccordo con quanto emerso. Gli stessi autori, però, indicano una non significatività della variabile, se posta in relazione a NO<sub>2</sub> o PM<sub>10</sub>.

Anche la pubblicazione ad opera di [Guo P. et al. 2017] pone in evidenza una relazione significativa tra O<sub>3</sub> ed ictus, segnalando un incremento del rischio relativo, che secondo le analisi degli autori si assesta a 1.0173. Gli stessi segnalano però la non significatività dell'inquinante in relazione al biossido di azoto.

Si scopre quindi che l'ozono non risulta essere un inquinante predominante ed anzi si assoggetta a variabili ben più indicative come gli ossidi di azoto o le polveri sottili. I risultati ottenuti in questo elaborato trovano accordo in una delle pubblicazioni, quella riguardante i test svolti a Seoul, probabilmente la città più climaticamente vicina a Milano. Le altre due pubblicazioni citate fanno invece riferimento a Kaohsiung (Taiwan) e Guangzhou (Cina), che è lecito supporre più esposte alle alte temperature e dunque ad alti valori d'ozono, vista la collocazione tropicale delle due città.

**Biossido di azoto:** i valori di correlazione legati al biossido di azoto risultano tra i più alti rilevati nel corso del processo di analisi. Seppur debole, la correlazione appare significativa. Nello studio condotto da [Hong Y. et al 2002] si identifica l'inquinante come moderatamente correlato alla patologia per quanto riguarda i valori registrati nello stesso giorno e nel precedente. I dati analizzati in questo elaborato ricalcano grossomodo queste considerazioni, seppur mostrino valori di correlazione inferiori.

Anche [Tsai S. at al. 2003] è in grado di individuare un significativo aumento del rischio correlato al biossido di azoto, segnalando inoltre la sua significatività anche in concomitanza con il PM<sub>10</sub>.

In aggiunta, gli studi condotti da [Gregory A. et al. 2005, Guo P. et al. 2017] confermano nuovamente tracce di correlazione significativa tra ictus e i valori di NO<sub>2</sub> giornalieri, segnalando infine che i valori del biossido permangano significativamente correlati anche in concomitanza con altri inquinanti.

**Ossidi di azoto:** i risultati ottenuti mostrano come gli ossidi di azoto risultino l'inquinante che maggiormente si correla agli ictus. Il valore di R ottenuto è il più elevato tra quelli di tutte le variabili analizzate e la correlazione debole ma significativa.

Tra le pubblicazioni studiate in letteratura, nessuna sembra prendere in considerazione gli ossidi di azoto come inquinante, probabilmente poiché non rilevati dalle centraline a disposizione. Ciononostante, la misura è paragonabile in toto a quanto visto per il solo NO<sub>2</sub> ed il comportamento rilevato in tutte le analisi svolte si mantiene coerente con tale variabile.

---

Si può dunque concludere che NO<sub>2</sub> e NOx presentino un comportamento simile ed è lecito aspettarsi che le pubblicazioni in esame, se avessero disposto dei valori relativi a tale inquinante, sarebbero giunte alle medesime conclusioni tratte per quanto riguarda il biossido di azoto.

**Monossido di carbonio:** i risultati ottenuti dalla correlazione del monossido di carbonio con gli ictus registrati sul territorio milanese mostrano una relazione debole e valida unicamente per il giorno stesso di insorgenza della patologia.

[Hong Y. et al 2002] identifica invece nel CO una correlazione moderata con gli ictus, estendendo la relazione anche ad un lag period pari a un giorno.

[Tsai S. et al. 2003] conferma l'esistenza di una significativa relazione di correlazione, relegata però ai soli giorni con temperatura inferiore a 20°C. La rilevanza dell'inquinante viene verificata anche in presenza di altre variabili ambientali come O<sub>3</sub> ed SO<sub>2</sub>.

Lo studio condotto da [Gregory A. et al.] non giunge ad una conclusione significativa, limitandosi a constatare l'esistenza di un legame tra PM<sub>10</sub> e CO.

I dati in tal senso sembrano non riuscire a confermare completamente quanto noto in letteratura e non si è in grado di giungere ad una conclusione esaustiva. Osservando il risultato di correlazione ottenuto sulle medie mensili, tuttavia, si sarebbe portati a trarre una conclusione circa la non primarietà degli effetti derivanti dall'esposizione al CO.

**Benzene:** le analisi svolte in relazione al benzene evidenziano valori di correlazione deboli, che però si mantengono costanti per i tre giorni precedenti al manifestarsi dell'ictus.

In letteratura non sono noti casi di analisi svolta su questo particolare inquinante e a giudicare da quanto osservato in questo elaborato di tesi, la variabile sembra avere un effetto poco indicativo durante tutto l'arco temporale dei 5 giorni.

Stupisce invece il risultato legato alla correlazione con il numero di casi calcolati su base mensile, che vede il benzene aggiudicarsi il valore di R più elevato insieme agli ossidi di azoto.

Si deduce quindi che l'inquinante non debba essere considerato un rischio nel breve periodo, mentre un'esposizione prolungata è significativamente associata ad una correlazione moderata con il fenomeno ictus.

**PM<sub>10</sub>:** i risultati ottenuti dalla correlazione del PM<sub>10</sub> con gli ictus mostrano una relazione piuttosto debole tra i due fenomeni. Osservando i valori di R estesi al lag period completo di 5 giorni si può notare come i valori misurati il giorno prima dell'evento vantino una correlazione maggiore rispetto ai giorni restanti.

In letteratura si individua nello studio condotto da [Hong Y. et al. 2002] una correlazione significativa ed elevata tra TSP (particolato totale sospeso) ed ictus nello stesso giorno del

---

ricovero. In questo caso l'associazione afferisce sia ad ictus ischemico sia ad ictus emorragico. Pur non trattandosi direttamente di PM10, la relazione assume comunque un certo significato data la natura comune dei due inquinanti.

La correlazione tra PM10 vero e proprio ed ictus è trattata nello scritto di [Tsai S. et al. 2003], che osserva un significativo aumento del rischio associato al particolato nei giorni caldi ( $>20^{\circ}\text{C}$ ), specialmente in associazione con il biossido di azoto. Entrambe queste associazioni sono rilevate nello stesso giorno in cui insorgono i sintomi dell'ischemia.

A conclusioni simili giunge anche [Gregory A. et al. 2005], che conferma l'esistenza di una relazione di correlazione significativa tra PM10 ed ictus per un lag period pari a 0, ossia nello stesso giorno.

Si assiste dunque ad una certa discrepanza nelle conclusioni ottenute, soprattutto in merito all'entità della correlazione, che nello studio qui presentato appare molto bassa. Una possibile spiegazione può essere data dalla diversità di ambiente o più verosimilmente dalla qualità dell'aria respirata nelle diverse città, che tende a non far emergere legami particolarmente significativi nel giro dei 5 giorni usati come finestra di valutazione.

Osservando infatti il valore di correlazione derivante dall'aggregazione mensile dei fenomeni, si nota una significativa e moderata correlazione proprio tra PM10 ed ictus, segno che anche in questo caso è l'esposizione prolungata all'inquinante ad avere le implicazioni più rilevanti.

**PM2.5:** la correlazione tra ictus e PM2.5 porta alla luce una correlazione debole ma significativa tra l'inquinante e la patologia oggetto d'indagine.

I livelli di correlazione più elevati si apprezzano per il giorno precedente al ricovero, così come accade per il PM10. Le due misure sono infatti strettamente correlate tra di loro in quanto figlie dei medesimi processi di combustione.

L'inquinante non è stato oggetto di molte analisi in letteratura ma è comunque possibile individuare pubblicazioni come quella di [Hong Y. et al. 2002], che identifica nel TSP (particelle totali sospese), ossia un insieme di materia particolata di diverso diametro, una fonte significativa di correlazione con gli ictus.

Parlando invece di PM2.5 vero e proprio si riporta lo studio condotto da [Guo P. et al. 2017], che identifica nell'inquinante una significativa correlazione con gli ictus registrati nello stesso giorno. Non si osserva tuttavia altrettanta rilevanza quando la variabile è considerata in concomitanza con NO<sub>2</sub>, O<sub>3</sub> o SO<sub>2</sub>, che sembrano assumere un ruolo predominante come agenti inquinanti.

Il caso del PM2.5 segna un netto contrasto con quanto osservato in questa tesi. I valori di correlazione identificati sono infatti bassi e non possono fornire informazioni deboli circa la natura patogena del particolato. Tuttavia, come nel caso del PM10 o in quello del benzene, quando si osservano i dati aggregati su base mensile si nota una correlazione

---

significativa e moderata. Si riporta dunque un possibile effetto a lungo termine collegato all'inquinante.

Curiosamente in nessuno degli elaborati analizzati si riescono ad individuare numeri specifici circa le correlazioni individuate dagli autori, che invece si limitano a riportare valori inerenti all'aumento di rischio relativo. Solo [Hong Y. et al. 2002] sembra dare una minima indicazione da questo punto di vista, parlando di valori compresi tra 0.5 e 0.9, senza però fornire ulteriori spiegazioni. Queste lacune fanno sì che un confronto diretto e quantitativo risulti una strada non percorribile, limitando di fatto le conclusioni in tal senso.

In conclusione, si sono analizzate 9 variabili ambientali, ponendole in relazione con gli eventi di ictus occorsi sul territorio della città di Milano negli anni 2015-2017.

Dal processo di analisi sono emerse correlazioni deboli riconducibili alle variabili indagate, con effetti valutati nella finestra temporale di 5 giorni normalmente proposta in letteratura.

Ciò che emerge da questo studio, tuttavia, è che la correlazione risulta tanto più forte quanto più è duratura l'esposizione agli inquinanti, segno che nel breve periodo il corpo umano sembra adattarsi alle condizioni ambientali e le variazioni subite dei parametri considerati non sono in grado di giustificare l'insorgenza di ictus. Al contrario, nel lungo periodo i dati suggeriscono che l'esposizione continuata a certi valori dei parametri considerati possa portare all'insorgenza di episodi infiammatori che possono generare arteriosclerosi e dunque portare all'ictus nei soggetti più vulnerabili, come esposto nella pubblicazione di [Ken K. et al. 2018], che presentando una review sull'argomento "Air Pollution and Stroke" mostra quelli che sono i risultati scientificamente noti dell'esposizione a lungo termine ai fattori inquinanti sopra citati.

## 4.4 Sviluppi futuri e limiti del metodo d'analisi

Sviluppi futuri di quanto trattato all'interno di questo progetto di tesi prevedono l'interfacciamento della realtà universitaria con il mondo medico al fine di poter analizzare dati circa lo stato di salute del paziente, il suo stile di vita e la sua attività lavorativa. Emerge anche la necessità di poter associare ad ogni record un outcome post trattamento che consenta di identificare quali siano i metodi e i parametri che meglio possono modellare il fenomeno ictus allo scopo finale di poterne predire l'insorgenza in determinate categorie di soggetti.

L'ipotesi al momento più indicativa è quella relativa alla località del fenomeno, circoscritto e caratteristico di aree verosimilmente omogenee sotto tutti i punti di vista: qualità della vita,

---

inquinamento e stress giornaliero. In tal senso emergono alcuni limiti del framework presentato e inerenti per lo più alla natura delle informazioni trattati.

Il posizionamento delle centraline, il loro numero e la tipologia di dati raccolti rappresentano al momento un ostacolo per lo svolgimento di un'analisi esaustiva su vasta scala che sia in grado di individuare queste aree omogenee del territorio.

Un avanzamento in tal senso sarebbe da ricercare nella collaborazione con le autorità regionali al fine di predisporre un piano di campionamento dei fenomeni di interesse attuale e futuro. Al momento, infatti, un altro limite del metodo in oggetto è rappresentato dalla penuria di variabili rilevate dalle centraline ARPA. L'ideale sarebbe predisporre un piano di ampliamento pianificato in modo da interessare territori affini (ad esempio Milano e provincia) anziché aree molto distanti tra di loro come lo sono i capoluoghi. In tal senso potrebbe valere la pena di indagare sull'utilità di sensori IoT concepiti all'uopo.

Un ulteriore aiuto nascerebbe dalla possibilità di monitorare un gruppo campione di individui misurandone i principali indici di stress, particolarmente elevato durante la settimana lavorativa e che potrebbe spiegare la maggior incidenza della patologia durante quei giorni, o di salute fisica (massa grassa, frequenza cardiaca e pressione sarebbero ottimi spunti da cui partire) tramite dispositivi economici e facilmente reperibili quali smart band o fitness tracker.

Al momento mancano infatti informazioni a corredo dei dati nudi e crudi, senza le quali è difficile poter trarre delle conclusioni che siano esaustive e replicabili in altre parti del globo.

Sarebbe altresì utile poter avere i dati altimetrici relativi alle posizioni in cui si verificano i casi di ictus e l'adozione dell'algoritmo di Optimal Interpolation sviluppato da ARPA rappresenterebbe il perfezionamento della tecnica di assegnazione delle variabili ambientali.

A parere di chi scrive si rende dunque necessaria la definizione di uno schema dati completo e specifico che funga da base di partenza per tracciare in toto il fenomeno ictus.

In conclusione, si è sviluppato un framework operativo che sfrutta le recenti nozioni di Health Geomatics al fine di individuare le relazioni che legano le variabili ambientali misurate sul territorio al fenomeno degli ictus in Lombardia e con un focus particolare sulla città di Milano, che garantisce al momento una precisione superiore dei risultati.

Tale studio pone le basi per una ricerca più esaustiva, che sperabilmente segua gli sviluppi futuri proposti.

---

## Bibliografia

- Cao W., Hu J., Yu X. 2009, *A study on temperature interpolation methods based on GIS*, 2009 17th International Conference on Geoinformatics DOI:10.1109/GEOINFORMATICS.2009.5293422
- Gregory A. W., Schwartz J., Mittleman A. 2005, *Air Pollution and Hospital Admissions for Ischemic and Hemorrhagic Stroke Among Medicare Beneficiaries*, Stroke. 2005;36:2549-2553
- Guo P., Wang Y., Feng W., Wu J., Fu C., Deng H., Huang J., Wang L., Zheng M., Liu H. 2017, *Ambient Air Pollution and Risk for Ischemic Stroke: A Short-Term Exposure Assessment in South China*, International Journal of Environmental Research and Public Health DOI:10.3390/ijerph14091091
- Holdaway M. R. 1996, *Spatial modeling and interpolation of monthly temperature using kriging*, CLIMATE RESEAR Vol. 6: 215-225,1996
- Hong Y., Lee J., Kim H., Kwon H. 2002, *Air Pollution A New Risk Factor in Ischemic Stroke Mortality*, Stroke. DOI: 10.1161/01.STR.0000026865.52610.5B
- Jerret M. et al. 2005, *A review and evaluation of intraurban air pollution exposure models*, Journal of Exposure Analysis and Environmental Epidemiology volume 15, pages 185–204 (2005) DOI: 10.1038/sj.jea.7500388
- Kamel Boulos MN, Rousdari AV, Carson ER 2001, *Health Geomatics: An Enabling Suite of Technologies in Health and Healthcare*, Journal of Biomedical Informatics 34, 495-2019 (2001) PMID: 11723701
- Ken K., Millaer M., Shah A. 2018, *Air Pollution and Stroke*, Journal of Stroke 2018;20(1):2-11 DOI: 10.5853/jos.2017.02894
- Kothari RU, Pancioli A., Liu T. Brott T., Broderick J. 1999, *Cincinnati Prehospital Stroke Scale: reproducibility and validity*, Ann Emerg Med. 1999 Apr;33(4):373-8
- Lavados P. M. Olavarria V., Hoffmeister L. 2017, *Ambient Temperature and Stroke Risk Evidence Supporting a Short-Term Effect at a Population Level From Acute Environmental Exposures*, Stroke. DOI: 10.1161/STROKEAHA.117.017838
- Lichtenstern A. 2013, *Kriging methods in spatial statistics*, Bachelor's Thesis at Technische Universitat Munchen
- Nur Falah A., Subartini B., Ruchjana B. 2016, *Application of universal kriging for prediction pollutant using GStat R*, IOP Conf. Series: Journal of Physics: Conf. Series 893 (2017) 012022 DOI :10.1088/1742-6596/893/1/012022

---

S. Ly, Charles C., Degré A. 2011, *Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium*, Hydrol. Earth Syst. Sci., 15, 2259–2274 DOI: 10.5194/hess-15-2259-2011

Sunila R. 2015, *Geostatistics: Kriging*, Konetekniikka 1, Otakaari 4, 150 10-12 [[https://mycourses.aalto.fi/pluginfile.php/141841/mod\\_resource/content/1/Geostatistics.pdf](https://mycourses.aalto.fi/pluginfile.php/141841/mod_resource/content/1/Geostatistics.pdf)]

Tomislav H. 2007, *A Practical Guide to Geostatistical Mapping of Environmental Variables*, Institute for Environment and Sustainability EUR 22904 EN

Tsai S., Goggins W., Chiu H., Yang C. 2003, *Evidence for an Association Between Air Pollution and Daily Stroke Admissions in Kaohsiung, Taiwan*, Stroke. DOI: 10.1161/01.STR.0000095564.33543.64

Tyagi A., Singh P. 2013, *Applying Kriging Approach on Pollution Data Using GIS Software*, International Journal of Environmental Engineering and Management. ISSN 2231-1319, Volume 4, Number 3 (2013), pp. 185-190

Uboldi F., Lussana C., Salvati M. 2007, *Three-dimensional spatial interpolation of surface meteorological observations from high resolution local networks*, ARPA Lombardia DOI: 10.1002/met.76

Zhou Y., Levy J. 2007, *Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis*, BMC Public Health DOI: 10.1186/1471-2458-7-89

# Sitografia

EDSS

<https://notiziemediche.it/info/la-valutazione-della-progressione-malattia/>

Geomatica

<https://it.wikipedia.org/wiki/Geomatica>

Geographic information system

[https://it.wikipedia.org/wiki/Geographic\\_information\\_system](https://it.wikipedia.org/wiki/Geographic_information_system)

Coordinate geografiche

[https://it.wikipedia.org/wiki/Coordinate\\_geografiche](https://it.wikipedia.org/wiki/Coordinate_geografiche)

EPSG Italia

<http://host154-194-static.207-37-b.business.telecomitalia.it/epsg/NotaSistemiEPSG.pdf>

Sistemi di riferimento e coordinate

<https://geomappando.com/2016/01/26/sistemi-di-riferimento-coordinate-1parte/>

CODICI EPSG

<https://3dmetrica.it/i-codici-epsg/>

EPSG:32632

<http://spatialreference.org/ref/epsg/wgs-84-utm-zone-32n/>

Simple Kriging in python

<http://connor-johnson.com/2014/03/20/simple-kriging-in-python/>

How Kriging works

<http://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm>

Semivariogram and covariance function

<http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/seminvariogram-and-covariance-functions.htm>

---

## APPENDICE A

### A.1 Validazione dei modelli di interpolazione

Di seguito si riportano le tabelle descriventi in maniera quantitativa la precisione ottenuta in ciascuno dei modelli utilizzati per interpolare i valori delle variabili ambientali descritte in questo progetto di tesi.

Questa descrizione è effettuata mediante il calcolo del differenziale di predizione rispetto al riferimento fornito dai valori letti dalle centraline e mediante l'individuazione della deviazione standard sul medesimo campione.

Si esclude il caso della temperatura in quanto già trattato nell'elaborato. Si ricorda che in rosso sono segnalate le centraline dedito alla lettura dei fenomeni presentati.

## ***Umidità relativa***

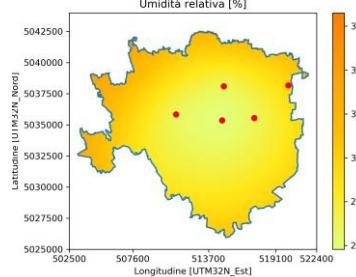


Figura 96 Rappresentazione grafica di una mappa di interpolazione dell'umidità relativa per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	-6.75e-14	6.08e-07
Milano v.Brera	25.2	25.2	-2.84e-14	4.35e-07
Milano v.Juvara	26.1	26.1	-4.97e-14	4.51e-07
Milano v.Marche	22.5	22.5	-1.07e-13	6.52e-07
Milano v.Zavattari	23.5	23.5	-1.21e-13	5.93e-07

## ***Ozono***

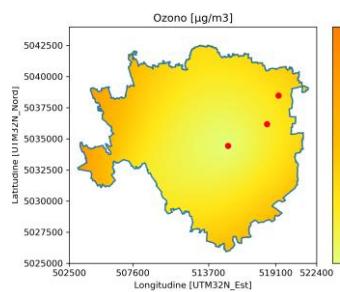


Figura 97 Rappresentazione grafica di una mappa di interpolazione dell'ozono per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Parco Lambro	53.4	53.4	1.99e-13	7.85e-07
Milano Pascal Città Studi	53.8	53.8	-9.95e-14	3.68e-07
Milano Verziere	46.2	46.2	2.84e-14	4e-07

## Biossido di azoto

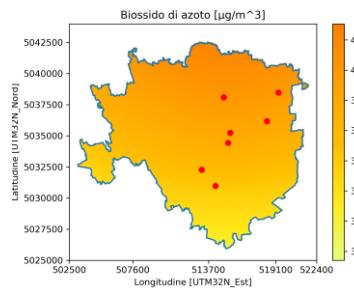


Figura 98 Rappresentazione grafica di una mappa di interpolazione del biossido di azoto per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano P.zza Abbiategrasso	25.8	25.8	-1.49e-13	9.1e-07
Milano Parco Lambro	46.6	46.6	-9.24e-14	8.85e-07
Milano Pascal Città Studi	40.7	40.7	-1.49e-13	7.87e-07
Milano Verziere	48.7	48.7	-1.71e-13	1.2e-07
Milano via Senato	58.5	58.5	-1.35e-13	9.95e-07
Milano viale Liguria	44.4	44.4	-1.49e-13	8.45e-07
Milano viale Marche	72.6	72.6	-1.42e-14	4.73e-07

## Ossidi di azoto

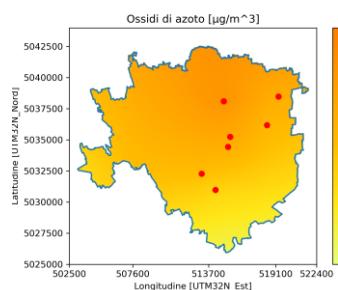


Figura 99 Rappresentazione grafica di una mappa di interpolazione degli ossidi di azoto per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano P.zza Abbiategrasso	40.9	40.9	4.97e-13	1.65e-06
Milano Parco Lambro	71.0	71.0	-9.95e-14	2.91e-06
Milano Pascal Città Studi	55.6	55.6	7.82e-14	1.63e-05
Milano Verziere	96.4	96.4	5.54e-13	2e-06
Milano via Senato	130.8	130.8	1.99e-13	1.73e-06
Milano viale Liguria	110.1	110.1	4.26e-14	2.28e-06
Milano viale Marche	191.0	191.0	5.68e-14	3.5e-06

## Monossido di carbonio

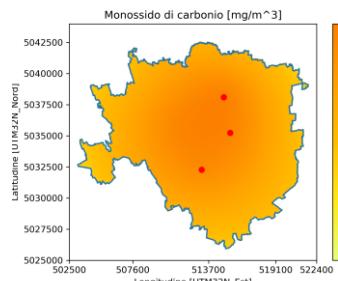


Figura 100 Rappresentazione grafica di una mappa di interpolazione del monossido di carbonio per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano via Senato	0.9	0.9	3e-15	1.96e-08
Milano viale Liguria	1.1	1.1	1.78e-15	1.53e-08
Milano viale Marche	1.6	1.6	1.55e-15	1.23e-08

## Benzene

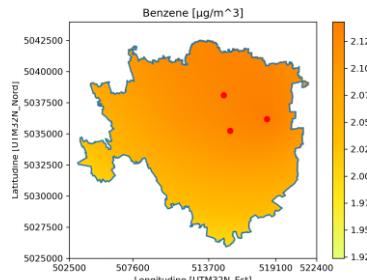


Figura 101 Rappresentazione grafica di una mappa di interpolazione del benzene per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Pascal Città Studi	3.2	3.2	-2.66e-15	8.82e-09
Milano via Senato	2.2	2.2	-1.78e-15	1.11e-08
Milano viale Marche	2.0	2.0	-1.11e-15	2.54e-08

## PM10

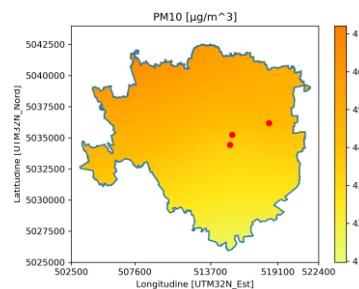


Figura 102 Rappresentazione grafica di una mappa di interpolazione del PM10 per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Pascal Città Studi	48.0	48.0	7.82e-14	4.88e-08
Milano Verziere	36.0	36.0	-9.95e-14	3.91e-07
Milano via Senato	44.0	44.0	7.11e-15	1.4e-07

## PM2.5

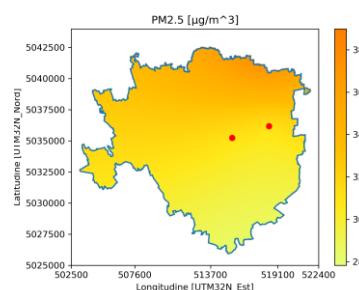


Figura 103 Rappresentazione grafica di una mappa di interpolazione del PM2.5 per il comune di Milano

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Pascal Città Studi	31.0	31.0	-2.21e-13	4.56e-07
Milano via Senato	31.0	31.0	6.04e-04	2.02e-07

## A.2 Cross validazione del modello di interpolazione

Di seguito si riportano le mappe interpolate della temperatura a Milano relative a ciascuna delle stazioni escluse nel processo di cross validazione. Si riporta unicamente il caso della temperatura in quanto rappresentativo di tutti gli altri fenomeni, che conducono a conclusioni identiche. A fianco ad ogni mappa si riporta la tabella contenente temperature lette, temperature interpolate, differenziale di predizione e deviazione standard. In rosso sono evidenziate le posizioni delle centraline.

## Milano Lambrate

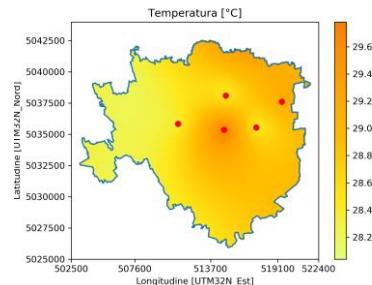


Figura 104 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano Lambrate

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	0.13	1.71
Milano v.Brera	30.6	30.6	-3.80e-13	3.42e-07
Milano v.Juvara	27.9	27.9	-4.33e-13	2.48e-07
Milano v.Marche	28.0	28.0	-3.80e-13	2.72e-07
Milano p.zza Zavattari	28.2	28.2	-2.45e-13	6.96e-07
Milano v.Feltre	29.4	29.4	-1.06e-14	8.22e-07

## Milano v.Brera

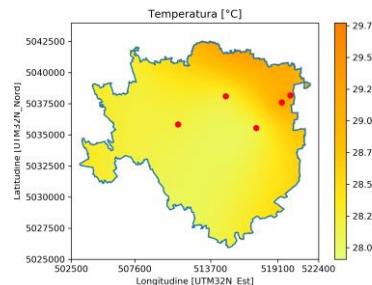


Figura 105 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Brera

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	2.24e-13	4.45e-07
Milano v.Brera	30.6	28.06	-2.54	1.83
Milano v.Juvara	27.9	27.9	9.24e-14	8.32e-07
Milano v.Marche	28.0	28.0	9.98e-13	1.28e-06
Milano p.zza Zavattari	28.2	28.2	3.41e-13	9.41e-07
Milano v.Feltre	29.4	29.4	5.93e-13	1.16e-06

## Milano v.Juvara

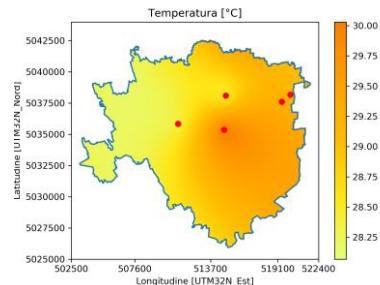


Figura 106 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Juvara

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	-5.26e-13	9.37e-07
Milano v.Brera	30.6	30.6	-3.77e-13	5.88e-07
Milano v.Juvara	27.9	29.58	1.68	1.82
Milano v.Marche	28.0	28.0	-2.81e-13	3.61e-07
Milano p.zza Zavattari	28.2	28.2	-3.77e-13	4.38e-07
Milano v.Feltre	29.4	29.4	-8.16e-13	9.45e-07

## Milano v.Marche

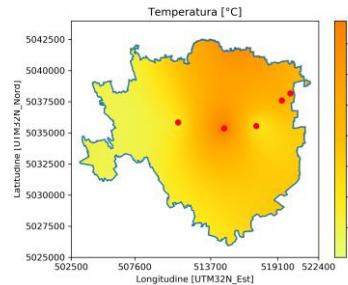


Figura 107 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano v.Marche

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	-3.91e-17	4.60e-07
Milano v.Brera	30.6	30.6	-4.72e-13	7.66e-07
Milano v.Juvara	27.9	27.9	-8.88e-13	1.02e-06
Milano v.Marche	28.0	29.42	1.42	1.87
Milano p.zza Zavattari	28.2	28.2	-1.07e-13	5.78e-07
Milano v.Feltre	29.4	29.4	-7.82e-14	5.71e-07

## *Milano p.zza Zavattari*

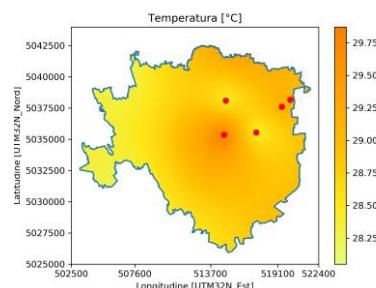


Figura 108 Mappa di interpolazione della temperatura comunale elaborata senza la stazione di Milano p.zza Zavattari

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	9.88e-13	1.08e-06
Milano v.Brera	30.6	30.6	8.35e-13	1.06e-06
Milano v.Juvara	27.9	27.9	6.54e-13	1.08e-06
Milano v.Marche	28.0	28.0	6.04e-13	8.40e-07
Milano p.zza Zavattari	28.2	28.91	0.71	2.04
Milano v.Feltre	29.4	29.4	1.23e-12	1.35e-06

*Milano v.Feltre*

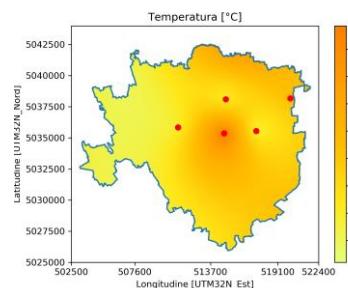


Figura 109 Mappa di  
interpolazione della temperatura  
comunale elaborata senza la  
stazione di Milano v. Feltre

Nome centralina	Valore centralina	Valore interpolato	Differenziale	Deviazione standard
Milano Lambrate	29.0	29.0	-2.98e-13	6.41e-07
Milano v.Brera	30.6	30.6	-6.25e-13	9.28e-07
Milano v.Juvara	27.9	27.9	-7.53e-13	1.06e-06
Milano v.Marche	28.0	28.0	-1.06e-12	1.07e-06
Milano p.zza Zavattari	28.2	28.2	-9.17e-13	1.25e-06
Milano v.Feltre	29.4	29.76	-0.64	1.62

---

## A.3 Algoritmi

Questa sezione raccoglie gli script in linguaggio Python (versione 3.6.3) utilizzati per effettuare pre-processing e processing dei dati. Si mostra una selezione di quanto ritenuto più significativo.

I sottocapitoli definiscono nel lo scopo di ognuno degli script, riportando in seguito il codice completo.

### A.3.1 Pre-processing dei dati ARPA

Si procede alla rimozione dei record non validi e alla standardizzazione del formato data-ora

```
1. import pandas as pd
2. from multiprocessing import Process
3. import time
4. import multiprocessing as mp
5. from datetime import timedelta
6.
7. def run(year, type):
8.     print("\nLoading", type, "data for year:", year)
9.     # loading the current year and the previous year. The previous year is
10.    # used to extract the last 6 days used to gather measures for lag periods
11.    table = pd.read_csv(str(year) + "_" + type + '.csv', sep=',', decimal='.')
12.    table_prev = pd.read_csv(str(year - 1) + "_" + type + '.csv', sep=',', decimal='.')
13.    # removing null records
14.    table = table[~table['Stato'].isnull()]
15.    table_prev = table_prev[~table_prev['Stato'].isnull()]
16.
17.    # parsing date-time and rounding to the closest 10 minutes interval
18.    print("Changing dates format...")
19.    # data about air have a different format for 2018 =
20.    if type == "ARIA" and year == 2018:
21.        table["DATE_TIME"] = pd.to_datetime(table.Data, format='%d/%m/%Y %I
22.        :%M:%S %p')
23.    else:
24.        table["DATE_TIME"] = pd.to_datetime(table.Data, format='%d/%m/%Y %H
25.        :%M:%S')
26.
27.    # dropping useless columns
28.    table.drop("Data", axis=1, inplace=True)
29.    table.drop("Stato", axis=1, inplace=True)
30.    table.drop("idOperatore", axis=1, inplace=True)
31.
32.    table_prev.drop("Data", axis=1, inplace=True)
33.    table_prev.drop("Stato", axis=1, inplace=True)
34.    table_prev.drop("idOperatore", axis=1, inplace=True)
35.
```

---

```
36. # selecting just the last 6 days for the previous year
37. print("Selecting data for the last 6 days of the previous year...")
38. table_prev = table_prev.loc[table_prev['DATE_TIME'] >= str(year -
1) + '-12-26 00:00:00']
39.
40. print("Concatenating frames...")
41. frames = [table_prev, table]
42. table = pd.concat(frames)
43.
44. print(table.head(20), "\n\n", table.tail(20))
45.
46. print("Saving...")
47. table.to_csv("DATI_" + type + "_" + str(year) + ".csv", sep=',', decimal='.', header=True, index=False)
48.
49.
50. if __name__ == '__main__':
51.
52.     n_years = 4
53.     processes = []
54.     year = 2015
55.
56.     start = time.time()
57.
58.     # the process is memory intensive so the data are processed sequentially
59.     # first weather data
60.     for i in range(n_years):
61.         p = Process(target=run, args=(year + i, 'METEO'))
62.         p.start()
63.         p.join()
64.
65.         end = time.time()
66.         print("The process took ", str(timedelta(seconds=(end-start))))
67.
68.     # then pollution data
69.     for i in range(n_years):
70.         p = Process(target=run, args=(year + i, 'ARIA'))
71.         p.start()
72.         p.join()
73.
74.         end = time.time()
75.         print("The process took ", str(timedelta(seconds=(end-start))))
76.
77.     end = time.time()
78.     print("Creation completed. The process took ", str(timedelta(seconds=(end-start))))
```

---

### A.3.2 Pre-processing dei dati AREU

Si procede alla conversione delle coordinate nel sistema EPSG:36632 ed alla standardizzazione del formato data-ora

```
1. import pandas as pd
2. from pyproj import Proj, transform
3. from osgeo import gdal
4.
5. def changeCoordinates(row):
6.     inProj = Proj('+init=EPSG:3003')
7.     outProj_UTM32N = Proj('+init=EPSG:32632')
8.     outProj_LONG_LAT = Proj('+init=EPSG:4326')
9.
10.    x1,y1 = row["VL_GEO_X"],row["VL_GEO_Y"]
11.    x2,y2 = transform(inProj,outProj_UTM32N,x1,y1)
12.    x3,y3 = transform(inProj,outProj_LONG_LAT,x1,y1)
13.    row["UTM32N_Est"] = x2
14.    row["UTM32N_Nord"] = y2
15.    row["Longitude"] = x3
16.    row["Latitude"] = y3
17.
18.    return row
19.
20. print("Loading data...")
21. tabella = pd.read_csv('Tabella ictus completa.csv',sep=';', decimal=',')
22.
23. # dropping duplicates
24. tabella = tabella.drop_duplicates()
25.
26. # removing null coordinates
27. tabella = tabella[~tabella['VL_GEO_X'].isnull()]
28. tabella = tabella[~tabella['VL_GEO_Y'].isnull()]
29.
30. # converting coordinates to UTM32N
31. print("Converting coordinates...")
32. tabella = tabella.apply(changeCoordinates, axis=1)
33.
34. # gathering altitude
35. print("Gathering altitudes...")
36. coord = tabella[['Longitude','Latitude']]
37. coord = [tuple(x) for x in coord.values]
38.
39. tabella.drop("Longitude", axis=1, inplace=True)
40. tabella.drop("Latitude", axis=1, inplace=True)
41.
42.
43. # parsing date-time and rounding to the closest 10 minutes interval
44. print("Adjusting dates...")
45. tabella["DATE_TIME"] = pd.to_datetime(tabella.DT_EMERG_DAY, format='%d/%m/%Y') + pd.to_timedelta(tabella.DT_EMRG_OPEN_HH24MI+":00")
46. tabella["DATE_TIME"] = tabella["DATE_TIME"].dt.round('10min')
47.
48. print(tabella)
49.
50. print("Saving...")
```

---

```
51. tabella.to_csv("tabella.csv", sep=',', decimal='.', header=True, index=False)
```

### A.3.3 Pre-processing dei dati giornalieri sugli inquinanti

Si separano i dati campionati in forma continua da quelli elaborati su base giornaliera. Lo scopo è quello di snellire i seguenti passi computazionali.

```
1. import pandas as pd
2. from datetime import timedelta
3.
4. for year in range (2015, 2018 + 1):
5.     # carico i dati dell'anno X
6.     print("\nLoading data for year", year)
7.     data = pd.read_csv("DATI_ARIA_" + str(year) + ".csv", sep=',', decimal='.')
8.     stations = pd.read_csv('Stazioni_qualit_dell_aria.csv', sep=',', decimal='.')
9.
10.    if year != 2018:
11.        # carico il dataset dell'anno precedente per poter includere il 31 dicembre
12.        data_prev = pd.read_csv("DATI_ARIA_" + str(year+1) + ".csv", sep=',', decimal='.')
13.
14.        data_prev["DATE_TIME"] = pd.to_datetime(data_prev["DATE_TIME"])
15.        data_prev["DATE_TIME"] = data_prev["DATE_TIME"] + timedelta(hours=-24)
16.
17.        data_prev = data_prev.loc[data_prev["DATE_TIME"].dt.year == year]
18.
19.        print("Merging data for previous year...")
20.        # joining the data to add stations' locations to every record
21.        merged = pd.merge(data_prev, stations, how='inner', on=['IdSensore', 'IdSensore'])
22.        merged = merged[merged["NomeTipoSensore"].isin(["PM10 (SM2005)", "Particelle sospese PM2.5", "Benzene"])]
23.
24.        data_prev = merged
25.        # keeping only the last day of the year
26.        data_prev = data_prev.loc[data_prev["DATE_TIME"] == str(year)+"-12-31"]
27.
28.        print("Merging data...")
29.        # joining the data to add stations' locations to every record
30.        merged = pd.merge(data, stations, how='inner', on=['IdSensore', 'IdSensore'])
31.        merged = merged[merged["NomeTipoSensore"].isin(["PM10 (SM2005)", "Particelle sospese PM2.5", "Benzene"])]
32.        merged["DATE_TIME"] = pd.to_datetime(merged["DATE_TIME"])
33.
34.        # scalo di un giorno perchè i dati si riferiscono alla media del giorno precedente
35.        merged["DATE_TIME"] = merged["DATE_TIME"] + timedelta(hours=-24)
36.
37.    if year != 2018:
```

---

```

38.         frames = [merged, data_prev]
39.         result = pd.concat(frames)
40.     else:
41.         result = merged
42.
43.     print(result["DATE_TIME"].head(20))
44.     print(result["DATE_TIME"].tail(20))
45.
46.     result.to_csv("INQUINANTI_GIORNALIERI_" + str(year) + ".csv", sep=',',
decimal='.', header=True, index=False)

```

### A.3.4 Pre-processing dei dati AREU in base all'anno

Si suddividono i record del database AREU in base all'anno di rilevamento, così da snellire i seguenti step computazionali.

```

1. import pandas as pd
2. import datetime
3. from datetime import timedelta
4.
5. # disables a useless warning
6. pd.options.mode.chained_assignment = None
7.
8. # reading the data about strokes
9. print("Loading data...")
10. data = pd.read_csv('tabella.csv',sep=',', decimal='.')
11. data["DATE_TIME"] = pd.to_datetime(data["DATE_TIME"])
12.
13. col_list = ["DATE_TIME", "UTM32N_Est", "UTM32N_Nord", "ALTITUDE"]
14. selection = data[col_list]
15.
16. print(selection)
17.
18. N_DAYS = 5
19. N_HOURS = 12
20.
21. for year in range (2015, 2018 + 1):
22.     # salvo: tabella intera dell'anno X, tabella ridotta dell'anno X e le tabelle ridotte dei lag period
23.
24.     print("Working on data for year:", year)
25.     total = data.loc[data.DATE_TIME.dt.year == year]
26.     reduced = selection.loc[selection.DATE_TIME.dt.year == year]
27.
28.     total["IDX"] = total.index
29.     reduced["IDX"] = reduced.index
30.
31.     print("Saving original and reduced table...")
32.     total.to_csv("tabella_completa_" + str(year) + ".csv", sep=',', decimal='.',
= '.', header=True, index=False)
33.     reduced.to_csv("tabella_ridotta_" + str(year) + ".csv", sep=',', decimal='.',
= '.', header=True, index=False)
34.
35.     double = reduced
36.
37.     print("Generating lag periods...")

```

---

```

38.     for i in range(int(24 * N_DAYS / N_HOURS)):
39.         # generating the datetime every N_HOURS hours, up to N_DAYS before
        the stroke event
40.         hour = N_HOURS * (i+1)
41.         lag = pd.read_csv("tabella_ridotta_" + str(year) + ".csv", sep=',',
        decimal='.')
42.
43.         lag[ "DATE_TIME"] = pd.to_datetime(lag[ "DATE_TIME"])
44.         lag[ "DATE_TIME"] = pd.to_datetime(lag[ "DATE_TIME"] -
        pd.Timedelta(hour, unit='h'))
45.
46.         lag.to_csv("tabella_ridotta_" + str(year) + "_" + str(hour) + "H.cs
        v", sep=',', decimal='.', header=True, index=False)

```

### A.3.5 Elaborazione dei dati e interpolazione valori

Si utilizzano i dati geolocalizzati per assegnare un valore a ciascuna della variabili di interesse, sfruttando modelli di interpolazione creati on demand.

```

1. from pyproj import Proj, transform
2. import pandas as pd
3. import datetime
4. from datetime import timedelta
5. import time
6. import pykrige.kriging_tools as kt
7. from pykrige.rk import OrdinaryKriging
8. from pykrige.rk import UniversalKriging
9. import numpy as np
10. import warnings
11. warnings.filterwarnings("ignore", category=DeprecationWarning)
12.
13. def interpolateWeather(data, measure):
14.     start = time.time()
15.     print("\nLoading weather stations...")
16.     stations = pd.read_csv('Stazioni_Meteorologiche.csv', sep=',', decimal='.
        ')
17.
18.     # filtering on stations that are currently active and measures the requ
        ired measure
19.     print("Filtering ", measure.lower(), " stations...")
20.     #stations = stations[stations['DataStop'].isnull()]
21.     stations = stations.loc[stations['Tipologia'] == str(measure)]
22.
23.     # keeping only the useful fields
24.     to_keep = ["IdSensore", "UTM_Nord", "UTM_Est"]
25.     stations = stations[to_keep]
26.
27.     # joining the data to add stations' locations to every record
28.     merged = pd.merge(data, stations, how='inner', on=['IdSensore', 'IdSens
        ore'])
29.
30.     # grouping by date and coordinates of the stroke event. In this way I r
        etrieve a group containing all the sensor reads, for all the stations, for e
        very stroke (ca 189 values, i.e.

```

---

```

31.      # the number of stations in Lombardy, for every stroke. Then I use the
32.      # coordinates of those stations and the read values to generate an interpolation
33.      # for the value at the location
34.      # of the stroke. Every interpolation uses all data available data in Lo
35.      # mbardy to increase precision.
36.      group = merged.groupby(['IDX', 'DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord'
37.      ])
38.
39.      print("# of groups:", len(group))
40.
41.      if measure == "Temperatura":
42.          model = 'exponential'
43.          print("Interpolation model:", model)
44.      elif measure == "Umidità Relativa":
45.          model='exponential'
46.          print("Interpolation model:", model)
47.
48.      id = []
49.      value = []
50.
51.      print("Interpolating values...")
52.
53.      for key, item in group:
54.          df = group.get_group(key)
55.
56.          # if the location coincides with one of the weather stations, a sin
57.          # gular matrix is generated. Being not invertible, the algorithm cannot inver
58.          # t it to generate the weights
59.          # if such an error occurs, I add the fake value -
60.          # 999.0 to the list of temperatures and then I cut the entire record to keep
61.          # only the valid ones
62.
63.          try:
64.              OK = OrdinaryKriging(df["UTM_Est"], df["UTM_Nord"], df["Valore"]
65.              ], variogram_model=model, verbose=False, enable_plotting=False)
66.              z = OK.execute('grid', key[2], key[3])
67.
68.              id.append(key[0])
69.              value.append(round(z[0][0][0], 1))
70.          except Exception:
71.              id.append(key[0])
72.              value.append(-999.0)
73.
74.          if measure == "Temperatura":
75.              out = pd.DataFrame({'IDX': id, 'TEMPERATURE':value})
76.          elif measure == "Umidità Relativa":
77.              out = pd.DataFrame({'IDX': id, 'HUMIDITY':value})
78.
79.          print("# of stations used for the interpolation:", len(df["UTM_Est"]))
80.
81.          print("# of records analyzed:", len(value))
82.
83.          end = time.time()
84.          print("The process took ", end - start, " seconds", "\n")
85.          return out
86.
87. def interpolatePollutants(data, measure, flag):
88.     start = time.time()
89.     if flag == 1:
90.         print("\nLoading air quality stations...")
91.         stations = pd.read_csv('Stazioni_qualit_dell_aria.csv',sep=',', de
92.         cimal='.')

```

---

```

81.
82.      # filtering on stations that are currently active and measures the
     required measure
83.      print("Filtering ", measure.lower(), " stations...")
84.      #stations = stations[stations['DataStop'].isnull()]
85.      stations = stations.loc[stations['NomeTipoSensore'] == str(measure)
     )]
86.
87.      # keeping only the useful fields
88.      to_keep = ["IdSensore", "Utm_Nord", "UTM_Est"]
89.      stations = stations[to_keep]
90.
91.      # joining the data to add stations' locations to every record
92.      merged = pd.merge(data, stations, how='inner', on=['IdSensore', 'Id
     Sensore'])
93.      else:
94.          print("\nLoading air quality stations...")
95.          print("Filtering ", measure.lower(), " stations...")
96.          data = data.loc[data['NomeTipoSensore'] == str(measure)]
97.          to_keep = ["DATE_TIME", "UTM32N_Est", "UTM32N_Nord", "ALTITUDE", "I
     DX", "IdSensore", "Utm_Nord", "UTM_Est", "Valore"]
98.          data = data[to_keep]
99.          merged = data
100.
101.      # grouping by date and coordinates of the stroke event. In this
     way I retrieve a group containing all the sensor reads, for all the stations
     , for every stroke (ca 189 values, i.e.
102.      # the number of stations in Lombardy, for every stroke. Then I u
     se the coordinates of those stations and the read values to generate an int
     erpolation for the value at the location
103.      # of the stroke. Every interpolation uses all data available dat
     a in Lombardy to increase precision.
104.      group = merged.groupby(['IDX', 'DATE_TIME', 'UTM32N_Est', 'UTM32
     N_Nord'])
105.
106.      print("# of groups:", len(group))
107.
108.      if measure == "Monossido di Carbonio":
109.          model = 'spherical'
110.          print("Interpolation model:", model)
111.      elif measure == "Ozono":
112.          model='exponential'
113.          print("Interpolation model:", model)
114.      elif measure == "Ossidi di Azoto":
115.          model = 'spherical'
116.          print("Interpolation model:", model)
117.      elif measure == "Biossido di Azoto":
118.          model = 'exponential'
119.          print("Interpolation model:", model)
120.      elif measure == "PM10 (SM2005)":
121.          model='spherical'
122.          print("Interpolation model:", model)
123.      elif measure == "Particelle sospese PM2.5":
124.          model='spherical'
125.          print("Interpolation model:", model)
126.      elif measure == "Benzene":
127.          model='spherical'
128.          print("Interpolation model:", model)
129.
130.      id = []
131.      value = []

```

---

```

132.
133.         print("Interpolating values...")
134.
135.         for key, item in group:
136.             df = group.get_group(key)
137.
138.                 # if the location coincides with one of the weather stations
139.                 , a singular matrix is generated. Being not invertible, the algorithm canno
140.                 t invert it to generate the weights
141.                 # if such an error occurs, I add the fake value -
142.                 999.0 to the list of temperatures and then I cut the entire record to keep
143.                 only the valid ones
144.             try:
145.                 OK = OrdinaryKriging(df["UTM_Est"], df["Utm_Nord"], df[
146.                     "Valore"], variogram_model=model, verbose=False, enable_plotting=False)
147.                 z = OK.execute('grid', key[2], key[3])
148.
149.                 id.append(key[0])
150.                 value.append(round(z[0][0][0], 1))
151.             except Exception:
152.                 id.append(key[0])
153.                 value.append(-999.0)
154.
155.             if measure == "Monossido di Carbonio":
156.                 out = pd.DataFrame({'IDX': id, 'CO':value})
157.             elif measure == "Ozono":
158.                 out = pd.DataFrame({'IDX': id, 'O3':value})
159.             elif measure == "Ossidi di Azoto":
160.                 out = pd.DataFrame({'IDX': id, 'NOx':value})
161.             elif measure == "Biossido di Azoto":
162.                 out = pd.DataFrame({'IDX': id, 'NO2':value})
163.             elif measure == "PM10 (SM2005)":
164.                 out = pd.DataFrame({'IDX': id, 'PM10':value})
165.             elif measure == "Particelle sospese PM2.5":
166.                 out = pd.DataFrame({'IDX': id, 'PM25':value})
167.             elif measure == "Benzene":
168.                 out = pd.DataFrame({'IDX': id, 'BENZENE':value})
169.
170.             print("# of stations used for the interpolation:", len(df["UTM_E
171. st"]))
172.             print("# of records analyzed:", len(value))
173.
174.             end = time.time()
175.             print("The process took ", end - start, " seconds", "\n")
176.             return out
177.
178.             N_DAYS = 5
179.             N_HOURS = 12
180.
181.             file_to_open = '_120H'
182.
183.             start = time.time()
184.
185.             for year in range (2015, 2018 + 1):
186.                 # carico i dati dell'anno X
187.                 print("Loading data for year", year, file_to_open)
188.                 data = pd.read_csv("DATI_METEO_" + str(year) + ".csv", sep=',', d
189.                 ecimal='.')
190.
191.                 print("Loading table for year", year)

```

---

```

185.         table = pd.read_csv("tabella_ridotta_" + str(year) + file_to_ope
n + '.csv',sep=',', decimal='.')
186.         print("# of records: ", len(table))
187.
188.         # converto le date in date time prima del join
189.         data["DATE_TIME"] = pd.to_datetime(data["DATE_TIME"])
190.         table["DATE_TIME"] = pd.to_datetime(table["DATE_TIME"])
191.
192.         # joining the sensors data with the strokes data on the basis of
the date
193.         print("Extracting common elements...")
194.         common = pd.merge(table, data, how='inner', on=['DATE_TIME', 'DA
TE_TIME'])
195.
196.         common_start = interpolateWeather(common, "Temperatura")
197.         table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
X'])
198.
199.         common_start = interpolateWeather(common, "Umidità Relativa")
200.         table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
X'])
201.
202.         table.to_csv("weather_" + str(year) + file_to_open + ".csv", sep
= ',', decimal='.', header=True, index=False)
203.         print("=====\n")
204.
205.         for year in range (2015, 2018 + 1):
206.             # carico i dati dell'anno X
207.             print("Loading data for year", year, file_to_open)
208.             data = pd.read_csv("DATI_ARIA_" + str(year) + '.csv',sep=',', de
cimal='.')
209.
210.             print("Loading table for year", year)
211.             table = pd.read_csv("tabella_ridotta_" + str(year) + file_to_ope
n + '.csv',sep=',', decimal='.')
212.
213.             print("# of records: ", len(table))
214.
215.             # converto le date in date time prima del join
216.             data["DATE_TIME"] = pd.to_datetime(data["DATE_TIME"])
217.             table["DATE_TIME"] = pd.to_datetime(table["DATE_TIME"])
218.
219.             # i dati meteo sono campionati ogni ora quindi arrotondo gli ora
ri dei dati sugli ictus
220.             table["DATE_TIME"] = table["DATE_TIME"].dt.round('1h')
221.
222.             # joining the sensors data with the strokes data on the basis of
the date
223.             print("Extracting common elements...")
224.             common = pd.merge(table, data, how='inner', on=['DATE_TIME', 'DA
TE_TIME'])
225.
226.             common_start = interpolatePollutants(common, "Monossido di Carbo
nio", 1)
227.             table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
X'])
228.
229.             common_start = interpolatePollutants(common, "Ozono", 1)
230.             table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
X'])

```

---

---

```

232.
233.         common_start = interpolatePollutants(common, "Ossidi di Azoto",
234.             1)
235.         table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
236.             X'])
237.         common_start = interpolatePollutants(common, "Biossido di Azoto"
238.             , 1)
239.         table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
240.             X'])
241.         table.to_csv("pollutants_" + str(year) + file_to_open + ".csv",
242.             sep=',', decimal='.', header=True, index=False)
243.         print("=====\\n")
244.         # inquinanti su base giornaliera
245.         for year in range(2015, 2018 + 1):
246.             # carico i dati dell'anno X
247.             print("Loading data for year", year, file_to_open)
248.             data = pd.read_csv("INQUINANTI_GIORNALIERI_" + str(year) + '.csv'
249.                 , sep=',', decimal='.')
250.             print("Loading table for year", year)
251.             table = pd.read_csv("tabella_ridotta_" + str(year) + file_to_ope
252.                 n + '.csv', sep=',', decimal='.')
253.             print("# of records: ", len(table))
254.             # converto le date in date time prima del join
255.             data["DATE_TIME"] = pd.to_datetime(data["DATE_TIME"])
256.             table["DATE_TIME"] = pd.to_datetime(table["DATE_TIME"])
257.             table["DATE_TIME"] = table["DATE_TIME"].dt.strftime('%Y-%m-
258.                 %d')
259.             table["DATE_TIME"] = pd.to_datetime(table["DATE_TIME"])
260.             # joining the sensors data with the strokes data on the basis of
261.             the date
262.             print("Extracting common elements...")
263.             common = pd.merge(table, data, how='inner', on=['DATE_TIME', 'DA
264.                 TE_TIME'])
265.             common_start = interpolatePollutants(common, "PM10 (SM2005)", 0)
266.             table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
267.                 X'])
268.             common_start = interpolatePollutants(common, "Particelle sospese
269.                 PM2.5", 0)
270.             table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
271.                 X'])
272.             common_start = interpolatePollutants(common, "Benzene", 0)
273.             table = pd.merge(table, common_start, how='left', on=['IDX', 'ID
274.                 X'])
275.             table.to_csv("pollutants_daily_" + str(year) + file_to_open + ".c
276.                 sv", sep=',', decimal='.', header=True, index=False)
277.             print("=====\\n")
278.             end = time.time()

```

---

```
276.     print("The whole process took", str(timedelta(seconds=(end-
    start))), "seconds")
```

### A.3.6 Selezione dei record validi e ricostruzione tabella iniziale

Si ricostruisce la tabella completa dei record utilizzando i valori che rispondono ai vincoli di integrità.

```
1. import pandas as pd
2. import time
3.
4. N_DAYS = 5
5. N_HOURS = 12
6.
7. file_to_open = ''
8.
9. start = time.time()
10.
11. for year in range (2015, 2018 + 1):
12.     # carico i dati dell'anno X
13.     print("Loading data for year", year, file_to_open)
14.     weather = pd.read_csv("weather_" + str(year) + ".csv",sep=',', decimal='.')
15.     pollutants = pd.read_csv("pollutants_" + str(year) + ".csv",sep=',', decimal='.')
16.     pollutants_daily = pd.read_csv("pollutants_daily_" + str(year) + ".csv", sep=',', decimal='.')
17.
18.     pollutants = pollutants.drop(['DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord', 'ALTITUDE'], axis=1)
19.     pollutants_daily = pollutants_daily.drop(['DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord', 'ALTITUDE'], axis=1)
20.
21.     # joining the sensors data with the strokes data on the basis of the date
22.     print("Joining elements...")
23.     common = pd.merge(weather, pollutants, how='inner', on=['IDX', 'IDX'])

24.     common = pd.merge(common, pollutants_daily, how='inner', on=['IDX', 'IDX'])
25.
26.     # merging lag periods
27.     for i in range(int(24 * N_DAYS / N_HOURS)):
28.         # generating the datetime every N_HOURS hours, up to N_DAYS before the stroke event
29.         hour = N_HOURS * (i+1)
30.         print("Loading gap period for hour", hour)
31.         weather = pd.read_csv("weather_" + str(year) + "_" + str(hour) + "H.csv",sep=',', decimal='.')
32.         pollutants = pd.read_csv("pollutants_" + str(year) + "_" + str(hour) + "H.csv",sep=',', decimal='.')
33.         pollutants_daily = pd.read_csv("pollutants_daily_" + str(year) + "_" + str(hour) + "H.csv",sep=',', decimal='.'
```

---

```

34.
35.     weather = weather.drop(['DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord', 'ALTITUDE'], axis=1)
36.     pollutants = pollutants.drop(['DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord', 'ALTITUDE'], axis=1)
37.     pollutants_daily = pollutants_daily.drop(['DATE_TIME', 'UTM32N_Est', 'UTM32N_Nord', 'ALTITUDE'], axis=1)
38.
39.     weather = weather.rename(index=str, columns={"TEMPERATURE": "TEMPERATURE_" + str(hour) + "H", "HUMIDITY": "HUMIDITY_" + str(hour) + "H"})
40.     pollutants = pollutants.rename(index=str, columns={"CO": "CO_" + str(hour) + "H", "O3": "O3_" + str(hour) + "H", "NOx": "NOx_" + str(hour) + "H", "NO2": "NO2_" + str(hour) + "H"})
41.     pollutants_daily = pollutants_daily.rename(index=str, columns={"PM10": "PM10_" + str(hour) + "H", "PM25": "PM25_" + str(hour) + "H", "BENZENE": "BENZENE_" + str(hour) + "H"})
42.
43.     # joining the sensors data with the strokes data on the basis of the date
44.     print("Joining elements...")
45.     common = pd.merge(common, weather, how='inner', on=['IDX', 'IDX'])
46.     common = pd.merge(common, pollutants, how='inner', on=['IDX', 'IDX'])
47.     common = pd.merge(common, pollutants_daily, how='inner', on=['IDX', 'IDX'])
48.
49.     common.to_csv("FINAL_" + str(year) + ".csv", sep=',', decimal='.', header=True, index=False)
50.     print("=====\\n")

```

```

1. import pandas as pd
2. import time
3.
4. N_DAYS = 5
5. N_HOURS = 12
6.
7. for year in range(2015, 2018 + 1):
8.     print("Loading data for year", year)
9.     df = pd.read_csv("FINAL_" + str(year) + '.csv', sep=',', decimal='.')
10.
11.    # filters are used to keep reasonable values
12.    # in some cases, when coordinates are far from the closest station, the exponential model can lead to extremely high/low values. Those measures are cut
13.    df = df[df['TEMPERATURE'].between(-20, 40, inclusive=True)]
14.
15.    # humidity % ranges between 0 and 100
16.    df = df[df['HUMIDITY'].between(0, 100, inclusive=True)]
17.
18.    # values > 0
19.    df = df[df['CO'] > 0]
20.    df = df[df['O3'] > 0]
21.    df = df[df['NOx'] > 0]
22.    df = df[df['NO2'] > 0]
23.    df = df[df['PM10'] > 0]
24.    df = df[df['PM25'] > 0]
25.    df = df[df['BENZENE'] > 0]

```

---

```

26.
27.     # other measures do not suffer for such problems
28.
29.     # filter on existance
30.     df.dropna(inplace=True)
31.
32.     # filter for lags
33.     for i in range(int(24 * N_DAYS / N_HOURS)):
34.         # generating the datetime every N_HOURS hours, up to N_DAYS before
            the stroke event
35.         hour = N_HOURS * (i+1)
36.
37.
38.         # filters are used to keep reasonable values
39.         # in some cases, when coordinates are fare from the closest station
            , the exponential model can lead to extremely high/low values. Those measur
            es are cut
40.         df = df[df['TEMPERATURE' + '_' + str(hour) + 'H'].between(-
            20, 40, inclusive=True)]
41.
42.         # humidity % ranges between 0 and 100
43.         df = df[df['HUMIDITY' + '_' + str(hour) + 'H'].between(0, 100, incl
            usive=True)]
44.
45.         # values > 0
46.         df = df[df['CO' + '_' + str(hour) + 'H'] > 0]
47.         df = df[df['O3' + '_' + str(hour) + 'H'] > 0]
48.         df = df[df['NOx' + '_' + str(hour) + 'H'] > 0]
49.         df = df[df['NO2' + '_' + str(hour) + 'H'] > 0]
50.         df = df[df['PM10' + '_' + str(hour) + 'H'] > 0]
51.         df = df[df['PM25' + '_' + str(hour) + 'H'] > 0]
52.         df = df[df['BENZENE' + '_' + str(hour) + 'H'] > 0]
53.
54.         # other measures do not suffer for such problems
55.
56.         # filter on existance
57.         df.dropna(inplace=True)
58.
59.         df.to_csv("GOAL_" + str(year) + ".csv", sep=',', decimal='.', header=Tr
            ue, index=False)
60.         print("=====\\n")

```

```

1. import pandas as pd
2. from datetime import timedelta
3. import sys
4.
5. for year in range(2015, 2018):
6.     print("\nLoading data for year", year)
7.     tabella = pd.read_csv('tabella_completa_' + str(year) + '.csv', sep=',',
            decimal='.')
8.     dati = pd.read_csv('GOAL_' + str(year) + '.csv', sep=',', decimal='.')
9.
10.    data = pd.merge(tabella, dati, how='inner', left_on=['IDX', 'DATE_TIME',
            'UTM32N_Nord', 'UTM32N_Est'], right_on=['IDX', 'DATE_TIME', 'UTM32N_Nord',
            'UTM32N_Est'])
11.    print("Saving...")
12.    data.to_csv('CONCLUSIONE_' + str(year) + '.csv', sep=';', decimal=',',
            header=True, index=False)

```

---

### A.3.7 Calcolo della distribuzione degli eventi

Si calcolano la distribuzione degli ictus e le medie istantanee associate alle variabili di interesse.

```
1. import pandas as pd
2. import time
3. import sys
4.
5. data1 = pd.read_csv('CONCLUSIONE_2015.csv',sep=';', decimal=',')
6. data2 = pd.read_csv('CONCLUSIONE_2016.csv',sep=';', decimal=',')
7. data3 = pd.read_csv('CONCLUSIONE_2017.csv',sep=';', decimal=',')
8.
9. year = 2015
10.
11. frames = [data1, data2, data3]
12. data = pd.concat(frames)
13. data = data.loc[data.DS_TOWN == "MILANO"]
14. #data = data.loc[data.ANNO == year]
15.
16. data.to_csv("CONCLUSIONE.csv", sep=';', decimal=',', header=True, index=False)
17.
18. # calcolo ictus e media per ogni giorno
19. data["DATE_TIME"] = pd.to_datetime(data["DATE_TIME"])
20. data["DAY"] = data.DATE_TIME.dt.day
21.
22. numero = data.groupby(['ANNO', 'MESE', 'DAY'])["DATE_TIME"].size()
23. media = data.groupby(['ANNO', 'MESE', 'DAY'])[str(sys.argv[1])].mean()
24.
25. data = pd.DataFrame({'EVENTI': numero, 'MEDIA '+str(sys.argv[1]): media})
26. data.to_csv("OUT_GIORNO.csv", sep=';', decimal=',', header=True)
27.
28. print("Correlazione con", str(sys.argv[1]), "anni 2015-"
29.       2017:", data["EVENTI"].corr(data['MEDIA '+str(sys.argv[1])]))
30. print(round(data["EVENTI"].corr(data['MEDIA '+str(sys.argv[1])])), 3))
```

### A.3.8 Rappresentazione della posizione delle centraline ARPA

Si rappresenta la distribuzione delle centraline ARPA presenti sul territorio della Lombardia.

```
1. import matplotlib.pyplot as plt
2. import pandas as pd
3. import shapefile as shp
4. from pyproj import Proj, transform
5.
6. df = pd.read_csv('result.csv',sep=';', decimal='.')
7. sf = shp.Reader("Regione_polygon.shp")
8.
9. X = df["UTM_Est"]
10. Y = df["UTM_Nord"]
```

---

```

11. Z = df["Valore"]
12.
13. citta = [("Milano",45.4773,9.1815), ("Monza",45.5834,9.2759), ("Bergamo",45
   .6989,9.67), ("Brescia",45.5257,10.2283), ("Como",45.8109,9.0885), ("Cremon
   a",45.1371,10.029),
14.   ("Lecco",45.8566,9.4039), ("Lodi",45.3145,9.5039), ("Mantova",45.15
   3,10.7748), ("Pavia",45.1854,9.1625), ("Sondrio",46.1699,9.8702), ("Varese"
   ,45.83,8.823)]
15.
16. plt.figure()
17.
18. for shape in sf.shapeRecords():
19.     x = [i[0] for i in shape.shape.points[:]]
20.     y = [i[1] for i in shape.shape.points[:]]
21.
22.     plt.plot(x,y)
23.
24. for i in range(len(X)):
25.     # plt.scatter(X,Y,edgecolors='none', marker='^')
26.     plt.scatter(X[i], Y[i], color='royalblue', marker='^')
27.
28. for city in citta:
29.     inProj = Proj('+init=EPSG:4326')
30.     outProj_UTM32N = Proj('+init=EPSG:32632')
31.
32.     x1,y1 = city[2],city[1]
33.     x2,y2 = transform(inProj,outProj_UTM32N,x1,y1)
34.     plt.scatter(x2, y2, color='red', marker='o')
35.
36. plt.title('Temperature stations')
37. plt.ylabel('UTM32N_NORTH')
38. plt.xlabel('UTM32N_EAST')
39. plt.axis('off')
40. plt.savefig("Stazioni ", dpi=300)
41. plt.show()
42.
43. print("Numero stazioni:", len(X))

```

### A.3.9 Generazione del variogramma e suo fitting

Si riporta lo script dedito alla generazione dei variogrammi ed al fitting dei modelli.

```

1. import pandas as pd
2. import numpy as np
3. from scipy.spatial.distance import pdist, squareform
4. import matplotlib
5. import matplotlib.pyplot as plt
6. import matplotlib.cm as cm
7. import matplotlib.colors as colors
8. import matplotlib.patches as mpatches
9.
10. def SVh( P, h, bw ):
11.     """
12.     Experimental semivariogram for a single lag
13.     """
14.     pd = squareform( pdist( P[:,2] ) )

```

---

```

15.     N = pd.shape[0]
16.     Z = list()
17.     for i in range(N):
18.         for j in range(i+1,N):
19.             if( pd[i,j] >= h-bw )and( pd[i,j] <= h+bw ):
20.                 Z.append( ( P[i,2] - P[j,2] )**2.0 )
21.     return np.sum( Z ) / ( 2.0 * len( Z ) )
22.
23. def SV( P, hs, bw ):
24.     """
25.     Experimental variogram for a collection of lags
26.     """
27.     sv = list()
28.     for h in hs:
29.         sv.append( SVh( P, h, bw ) )
30.     sv = [ [ hs[i], sv[i] ] for i in range( len( hs ) ) if sv[i] > 0 ]
31.     return np.array( sv ).T
32.
33. def C( P, h, bw ):
34.     """
35.     Calculate the sill
36.     """
37.     c0 = np.var( P[:,2] )
38.     if h == 0:
39.         return c0
40.     return c0 - SVh( P, h, bw )
41.
42. def spherical( h, a, C0 ):
43.     """
44.     Spherical model of the semivariogram
45.     """
46.     # if h is a single digit
47.     if type(h) == np.float64:
48.         # calculate the spherical function
49.         if h <= a:
50.             return C0*( 1.5*h/a - 0.5*(h/a)**3.0 )
51.         else:
52.             return C0
53.     # if h is an iterable
54.     else:
55.         # calcualte the spherical function for all elements
56.         a = np.ones( h.size ) * a
57.         C0 = np.ones( h.size ) * C0
58.         return list(map( spherical, h, a, C0 ))
59.
60. def exponential( h, a, c ):
61.     """
62.     Exponential model of the semivariogram
63.     """
64.     a, c = float( a ), float( c )
65.     return c*( 1.0 - np.exp( -h/a ) )
66.
67. def gaussian( h, a, c ):
68.     """
69.     Gaussian model of the semivariogram
70.     """
71.     a, c = float( a ), float( c )
72.     return c*( 1.0 - np.exp( -h**2.0/a**2.0 ) )
73.
74. def opt( fct, x, y, C0, parameterRange=None, meshSize=1000 ):
75.     if parameterRange == None:

```

---

```

76.         parameterRange = [ x[1], x[-1] ]
77.         mse = np.zeros( meshSize )
78.         a = np.linspace( parameterRange[0], parameterRange[1], meshSize )
79.         for i in range( meshSize ):
80.             mse[i] = np.mean( ( y - fct( x, a[i], C0 ) )**2.0 )
81.         return a[ mse.argmin() ]
82.
83. def cvmodel( P, model, hs, bw ):
84.     """
85.     Input: (P)      ndarray, data
86.            (model) modeling function
87.                  - spherical
88.                  - exponential
89.                  - gaussian
90.            (hs)    distances
91.            (bw)    bandwidth
92.     Output: (covfct) function modeling the covariance
93.     """
94.     # calculate the semivariogram
95.     sv = SV( P, hs, bw )
96.     # calculate the nugget
97.     nugget = sv[1][0] - 0
98.     # calculate the sill
99.     C0 = C( P, hs[0], bw )# + nugget
100.    # calculate the optimal parameters
101.    param = opt( model, sv[0], sv[1], C0 )
102.    # return a covariance function
103.    covfct = lambda h, a=param: model( h, a, C0 )
104.    return covfct
105.
106. def krig( P, model, hs, bw, u, N ):
107.     """
108.     Input (P)      ndarray, data
109.            (model) modeling function
110.                  - spherical
111.                  - exponential
112.                  - gaussian
113.            (hs)    kriging distances
114.            (bw)    kriging bandwidth
115.            (u)     unsampled point
116.            (N)     number of neighboring
117.                      points to consider
118.      ...
119.
120.      # covariance function
121.      covfct = cvmodel( P, model, hs, bw )
122.      # mean of the variable
123.      mu = np.mean( P[:,2] )
124.
125.      # distance between u and each data point in P
126.      d = np.sqrt( ( P[:,0]-u[0] )**2.0 + ( P[:,1]-u[1] )**2.0 )
127.      # add these distances to P
128.      P = np.vstack(( P.T, d )).T
129.      # sort P by these distances
130.      # take the first N of them
131.      P = P[d.argsort()[:N]]
132.
133.      # apply the covariance model to the distances
134.      k = covfct( P[:,3] )
135.      # cast as a matrix
136.      k = np.matrix( k ).T

```

---

```

137.
138.      # form a matrix of distances between existing data points
139.      K = squareform( pdist( P[:,2:] ) )
140.      # apply the covariance model to these distances
141.      K = covfct( K.ravel() )
142.      # re-cast as a NumPy array -- thanks M.L.
143.      K = np.array( K )
144.      # reshape into an array
145.      K = K.reshape(N,N)
146.      # cast as a matrix
147.      K = np.matrix( K )
148.
149.      # calculate the kriging weights
150.      weights = np.linalg.inv( K ) * k
151.      weights = np.array( weights )
152.
153.      # calculate the residuals
154.      residuals = P[:,2] - mu
155.
156.      # calculate the estimation
157.      estimation = np.dot( weights.T, residuals ) + mu
158.
159.      return float( estimation )
160.
161.      # -----
162.      #
163.
164.      DISTANCE_KM = 100
165.      LAG_M = 1000
166.
167.      # reading the data
168.      data = pd.read_csv('result.csv',sep=';', decimal='.')
169.      col_list = ["UTM_Est", "UTM_Nord", "Valore"]
170.
171.      data = np.array(data[col_list])
172.
173.      # bandwidth, plus or minus 500 meters
174.      bw = LAG_M
175.
176.      # lags in 1000 meter increments from zero to 14Km
177.      hs = np.arange(0,DISTANCE_KM * 1000, LAG_M)
178.      sv = SV( data, hs, bw )
179.
180.      import pandas as pd
181.      df = pd.DataFrame(columns=["X", "Y", "M"])
182.      df["X"] = pd.Series(sv[0])
183.      df["Y"] = pd.Series(sv[1])
184.      df.to_csv("exp.csv", sep=';', decimal=',', index=False)
185.
186.      plt.plot( sv[0], np.sqrt(sv[1]), '.-' )
187.      plt.xlabel('Lag [m]')
188.      plt.ylabel('Semideviation [°C]')
189.      plt.savefig('sample_semivariogram.png',fmt='png',dpi=200)
190.
191.      # model fitting
192.      sp = cvmodel(data, model=exponential, hs=np.arange(0,DISTANCE_KM * 1
193.          000, LAG_M), bw=LAG_M)
194.      plt.plot( sv[0], np.sqrt(sv[1]), '.-' )

```

---

```

195.     plt.plot( sv[0], np.sqrt(sp(sv[0])))
196.
197.     df["M"] = pd.Series(sp(sv[0]))
198.     df.to_csv("exp.csv", sep=';', decimal=',', index=False)
199.
200.     plt.title('Exponential Model')
201.     plt.ylabel('Semideviation [°C]')
202.     plt.xlabel('Lag [m]')
203.     plt.savefig('semivariogram_model_exponential.png',fmt='png',dpi=200)

```

### A.3.10 Generazione di una mappa interpolata

Si riporta lo script per la generazione di una mappa interpolata (caso relativo alla città di Milano).

```

1. import numpy as np
2. import pykrige.kriging_tools as kt
3. from pykrige.rk import OrdinaryKriging
4. from pykrige.uk import UniversalKriging
5. import pandas as pd
6.
7. import matplotlib
8. import matplotlib.pyplot as plt
9. import matplotlib.cm as cm
10. import matplotlib.colors as colors
11. import matplotlib.patches as mpatches
12.
13. import warnings
14. warnings.filterwarnings("ignore", category=UserWarning)
15.
16. import shapefile as shp
17. sf = shp.Reader("L090102_CittàMilano.shp")
18.
19. # reading the data
20. data = pd.read_csv('result.csv',sep=';', decimal='.')
21. col_list = ["UTM_Est", "UTM_Nord", "Valore"]
22.
23. rr = data
24. #leave one out
25. #non = "Milano v.Feltre"
26. #data = data.loc[data.NomeStazione != non]
27.
28. #selezione colonne
29. data = np.array(data[col_list])
30.
31. X0, X1 = data[:,0].min(), data[:,0].max()
32. Y0, Y1 = data[:,1].min(), data[:,1].max()
33.
34. ## milano
35. X0 = 502500
36. X1 = 522400
37. Y0 = 5025000
38. Y1 = 5044000
39.
40. NUM = 400.0

```

```

41.
42. gridx = np.arange(X0, X1, (X1-X0)/NUM)
43. gridy = np.arange(Y0, Y1, (Y1-Y0)/NUM)
44.
45. OK = OrdinaryKriging(data[:, 0], data[:, 1], data[:, 2], variogram_model='e
   xponential', verbose=False, enable_plotting=False)
46. z, ss = OK.execute('grid', gridx, gridy)
47.
48. print("AVG sigma:", np.average(ss), "\nMAX sigma:", np.max(ss), "\nMIN sigm
   a:", np.min(ss), "\nMEDIAN sigma:", np.median(ss), "\n75° percentile sigma:
   ",
   np.percentile(ss, 75), "\n90° percentile sigma:", np.percentile(ss,
   90), "\n95° percentile sigma:", np.percentile(ss, 95), "\n99° percentile s
   igma:", np.percentile(ss, 99))
49.
50.
51. extent = (X0,X1,Y0,Y1)
52. im = plt.imshow(z, cmap='Wistia', aspect='auto', extent=extent, origin="low
   er")
53. plt.colorbar(im)
54.
55. from scipy.ndimage import imread
56. im = imread('Maschera_Milano.png', mode='RGBA')
57. plt.imshow(im, extent=extent)
58.
59. for shape in sf.shapeRecords():
60.     x = [i[0] for i in shape.shape.points[:]]
61.     y = [i[1] for i in shape.shape.points[:]]
62.
63.     plt.plot(x,y)
64.
65. data = rr
66. delta = []
67. dev = []
68. import math
69.
70. print("\n\n---\nDelta temperature rispetto alle misurazioni delle centraline ---\n\n")
71. for staz in ["Milano Lambrate", "Milano v.Brera", "Milano v.Juvara", "Milan
   o v.Manche", "Milano P.zza Zavattari", "Milano v.Feltre"]:
72.     if len(data.loc[data.NomeStazione == staz]["Valore"]) == 1:
73.         val = float(data.loc[data.NomeStazione == staz]["Valore"])
74.         res, sigma = OK.execute('grid', float(data.loc[data.NomeStazione ==
   staz]["UTM_Est"]), float(data.loc[data.NomeStazione == staz]["UTM_Nord"]))
75.
76.         res = res[0][0]
77.         dev_std = math.sqrt(abs(sigma[0][0]))
78.         #if staz != non:
79.         plt.scatter(float(data.loc[data.NomeStazione == staz]["UTM_Est"]),
   float(data.loc[data.NomeStazione == staz]["UTM_Nord"])), color='red', marker
   ='o')
80.         dev.append(dev_std)
81.         delta.append(res-val)
82.         print("Centralina di "+staz+"\nVal. letto:", val, "\tVal. interpola
   to:", res, "\tDelta:", res-val, "\tDev_std:", dev_std)
83. print("\n\nAVG delta:", np.mean(delta), "\nAVG STD_DEV:", np.mean(dev))
84.
85.
86. #plt.axis('off')
87. plt.title('Temperatura [°C]')
88. plt.xlabel('Longitudine [UTM32N_Est]')

```

---

```
89. plt.xticks([502500,507600, 513700, 519100,522400])
90. plt.ylabel('Latitudine [UTM32N_Nord]')
91. plt.savefig("Mappa_temperatura.png", dpi=300)
92. plt.show()
```

### A.3.11 Generazione delle medie

Si riporta lo script generante le medie mensili relative ai diversi fenomeni indagati

```
1. import pandas as pd
2. from datetime import timedelta
3.
4. for year in range (2015, 2018 + 1):
5.     # carico i dati dell'anno X
6.     print("\nLoading data for year", year)
7.     data = pd.read_csv("DATI_ARIA_" + str(year) + ".csv",sep=',', decimal='.')
8.     stations = pd.read_csv('Stazioni_qualit_dell_aria.csv',sep=',', decimal='.')
9.
10.    print("Merging data...")
11.    # joining the data to add stations' locations to every record
12.    merged = pd.merge(data, stations, how='inner', on=['IdSensore', 'IdSensore'])
13.    merged = merged[merged["NomeTipoSensore"].isin(["Ossidi di Azoto"])]
14.    merged[ "DATE_TIME"] = pd.to_datetime(merged[ "DATE_TIME"])
15.
16.    # scalo di un giorno perchè i dati si riferiscono alla media del giorno precedente
17.    merged[ "DATE_TIME"] = merged[ "DATE_TIME"] + timedelta(hours=-24)
18.
19.    result = merged
20.
21.    result = result.loc[result.Comune == "Milano"]
22.    result[ "MONTH"] = result[ "DATE_TIME"].dt.month
23.    result = result.loc[result[ "DATE_TIME"].dt.year == year]
24.
25.    to_keep = ["Valore", "DATE_TIME", "MONTH", "Comune", "IdSensore"]
26.    result = result[to_keep]
27.
28.    result = result.groupby(['MONTH']).mean()
29.    result = result["Valore"]
30.
31.    print(result)
32.
33.    result.to_csv("MEDIA_" + str(year) + ".csv", sep=';', decimal=',', header=True, index=True)
```

---

### A.3.12 Rappresentazione dei fenomeni espressi in bin

Si riporta lo script dedito alla rappresentazione in bin dei fenomeni indagati

```
1. import pandas as pd
2. data = pd.read_csv('CORR_PM10.csv',sep=';', decimal=',')
3.
4. step = 5
5. base = 0
6. top = 150
7. current = base
8.
9. dict = {}
10. current = base
11.
12. while True:
13.     if current == top:
14.         break
15.
16.     out = data.query(str(current) + ' < V0 < ' + str(current + step))
17.     tot = out.COUNT.sum()
18.     current += step
19.     dict[current + step/2] = tot
20.     #print("Range:", current-1, current, tot, out)
21.
22. print(dict)
23. out = pd.DataFrame(list(dict.items()), columns=['BIN', 'COUNT'])
24. out.to_csv("OUT__.csv", sep=';', decimal=',', header=True, index=False)
```

I restanti frammenti di codice non sono stati inseriti per evitare un appesantimento eccessivo dell'elaborato. La loro significatività è stata ritenuta secondaria.