

ANALISI DELLA QUALITÀ DELL'ARIA IN LONDON

Progetto Finale - Data Mining & Analytics 2021
E.Carta, D.Caputo, L.Cefaloni

Introduzione

La città di **Londra** come anche moltissime delle città Europee stanno cercando pian piano di cambiare il loro aspetto dal punto di vista ecologico, dando importanza al riciclo, alla diminuzione dell'utilizzo dei combustibili fossili e a vari altri aspetti che puntano a migliorare la qualità di vita dei suoi abitanti.

In questi ultimi decenni a Londra sono state adottate moltissime politiche green proprio per migliorare la qualità dell'aria come per esempio il **Carbon Descent Plan** del 2009, che punta alla diminuzione degli inquinanti principalmente dovuti ai combustibili fossili.

L'analisi che andremo a condurre cercherà di capire se queste azioni adottate dalla capitale Inglese stanno portando miglioramenti.

Dataset

Nel dataset che andremo a usare, per compiere questa analisi, sono presenti le misurazioni prese ora per ora di 6 anni, che vanno dal **2015** al **2020**, nel quartiere di Londra, per la precisione nel quartiere di **N. Kensington**.

Abbiamo deciso di prendere solo questo quartiere poiché è uno di quelli più centrali.

Le misurazioni riguardano **8 inquinanti e 3 caratteristiche metereologiche**.

Andiamo a vederle nel particolare:

Inquinanti:

- **co**, monossido di carbonio: proviene principalmente dal fumo di tabacco e da fonti di combustione;
- **nox**, ossidi di azoto: la fonte principale deriva dal traffico veicolare e gli impianti di riscaldamento;
- **no₂**, diossido di azoto;
- **no**, monossido di azoto;
- **o₃**, ozono: si forma in atmosfera per effetto di reazioni favorite dalla radiazione solare in presenza dei cosiddetti inquinanti precursori;
- **so₂**, anidride solforosa: combustione di combustione di carburanti fossili (traffico, riscaldamenti). Con le reazioni acido-base formano le piogge acide;
- **pm 10**: lunghi tempi di permanenza in atmosfera. Origine sia naturale che antropica;
- **pm 2.5**: polveri sottili, sono delle particelle inquinanti presenti nell'aria che respiriamo, fonti naturali, traffico veicolare.

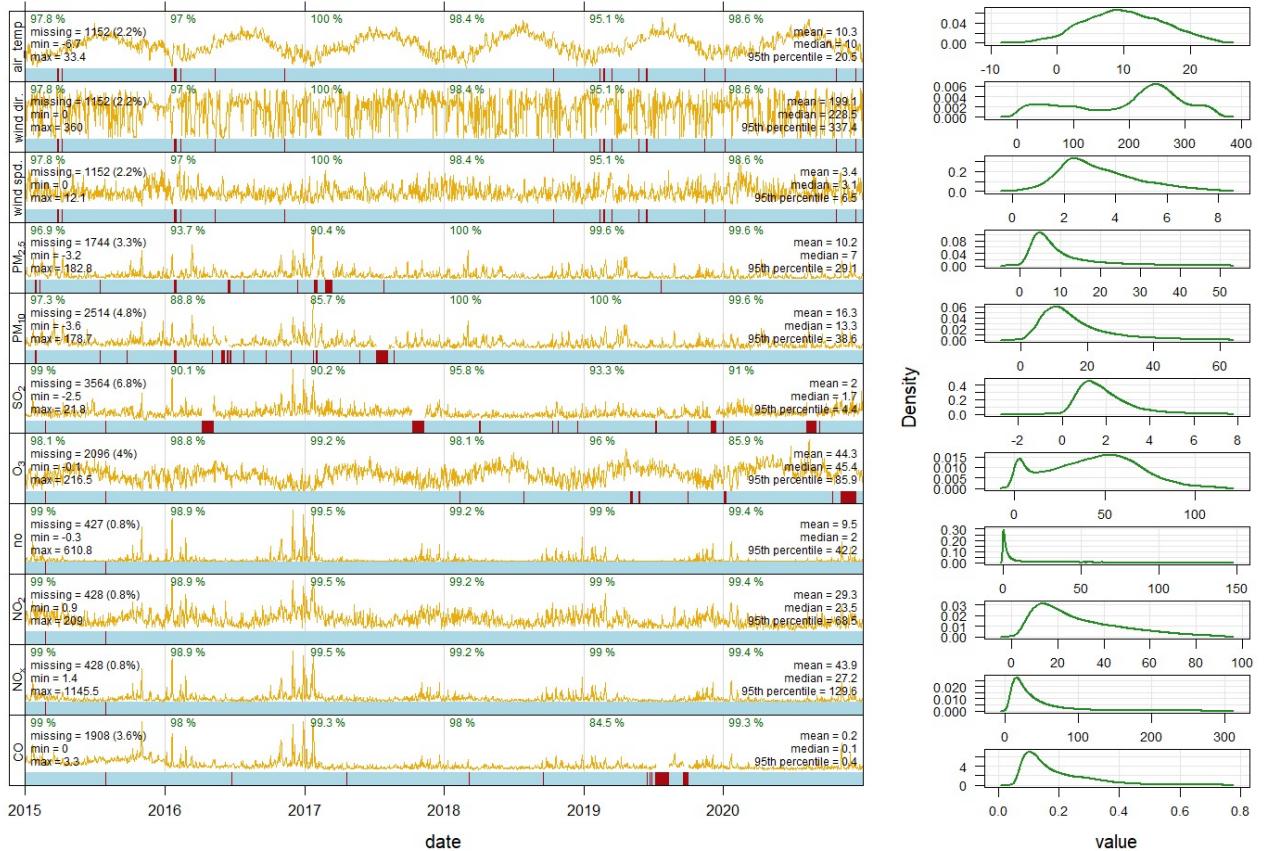
Caratteristiche merceologiche:

- ws, intensità dell'aria
- wd, direzione dell'aria
- air temp, temperatura dell'aria

1 Analisi Preliminare

Grafico Generale

Come prima cosa andiamo a vedere graficamente tutti i dati che abbiamo in possesso con annesse le loro distribuzioni.

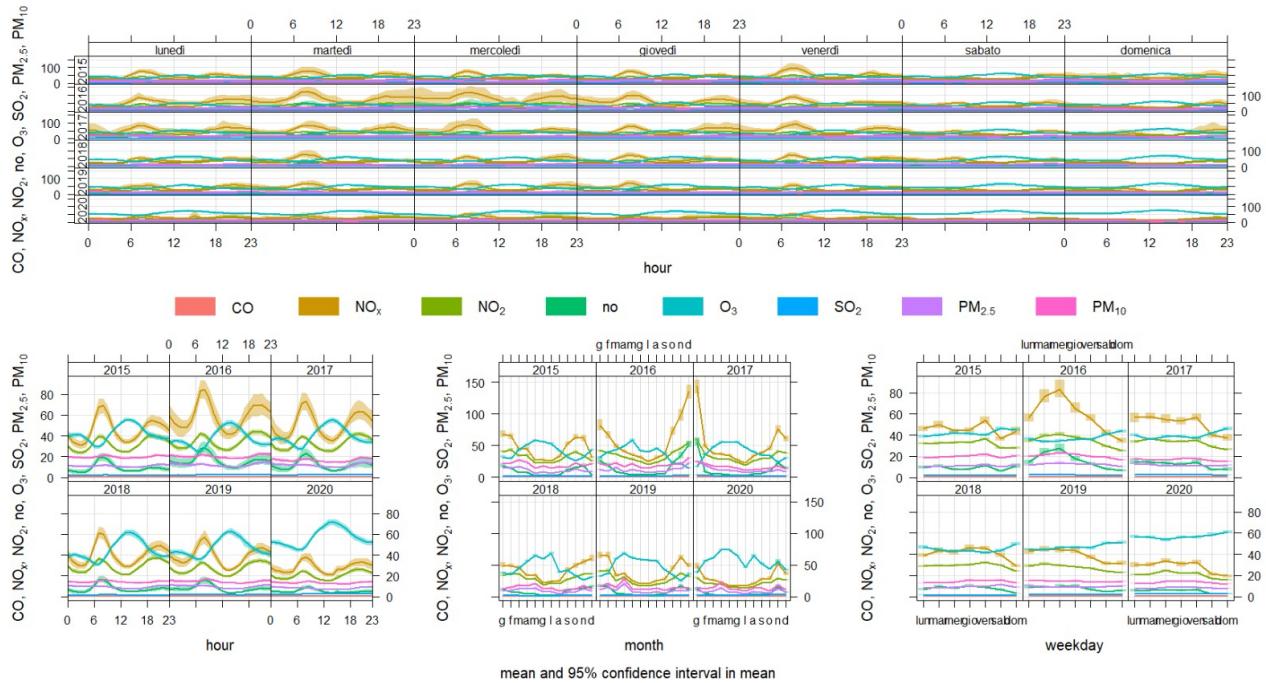


Possiamo notare che molti dei nostri dati hanno una cadenza stagionale, a seconda di quella in cui ricascano avremo più o meno emissioni di quell'inquinante.

Ogni stagione può essere descritta con un anno. Nella maggior parte degli inquinanti i valori maggiori si presentano durante i mesi invernali, tranne che per **O3** che presenta valori molto alti durante nei mesi più caldi. Infine nei grafici a destra viene mostrata la densità per ogni variabile del nostro dataset.

1.1 Time Variation

Descriviamo adesso l'andamento annuale di ogni inquinante.

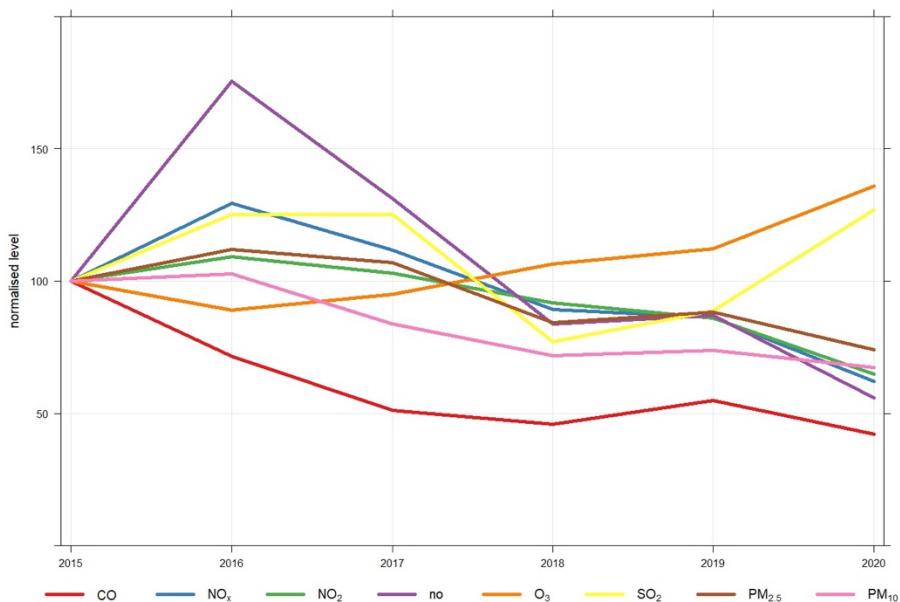


Come possiamo notare da questo grafico le emissioni di inquinanti anno dopo anno vanno a calare. In particolare gli **nox** vediamo che le loro misurazioni durante i primi anni (2015,2016,2017) rimangono molto alte a confronto degli altri inquinanti, avendo un grande picco nel 2016. Anche se i livelli delle sue emissioni sono molto elevati, durante i sabati e le domeniche di tutti e cinque gli anni i loro valori diminuiscono.

Dal 2018 questo andamento cambia, infatti passa dall'inquinante prevalente, venendo superato dall'**Ozono**.

Infatti quest'ultimo presenta un aumento costante delle sue emissioni, questo comporta una preoccupante prospettiva per gli anni seguenti.

Questo aumento potrebbe essere causato dalla diminuzione repentina dei combustibili fossili e nella richiesta costante di energia da parte delle aziende.

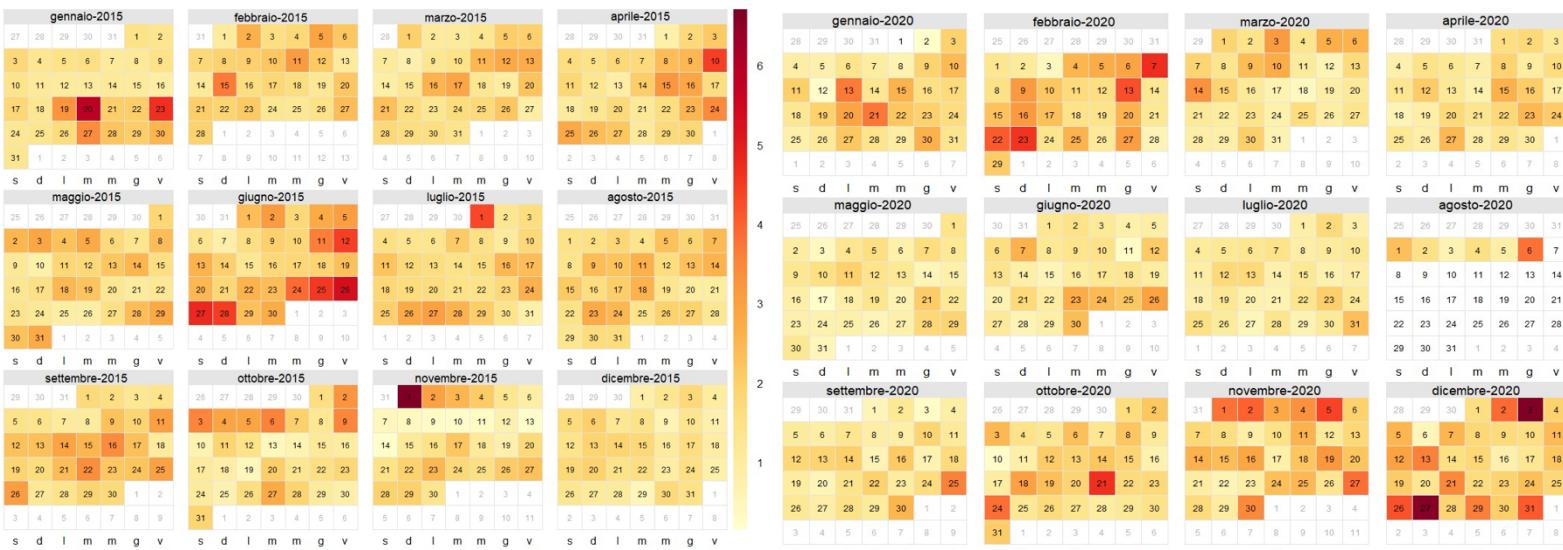


Dal seguente grafico vediamo l'andamento durante gli anni dei vari inquinati.

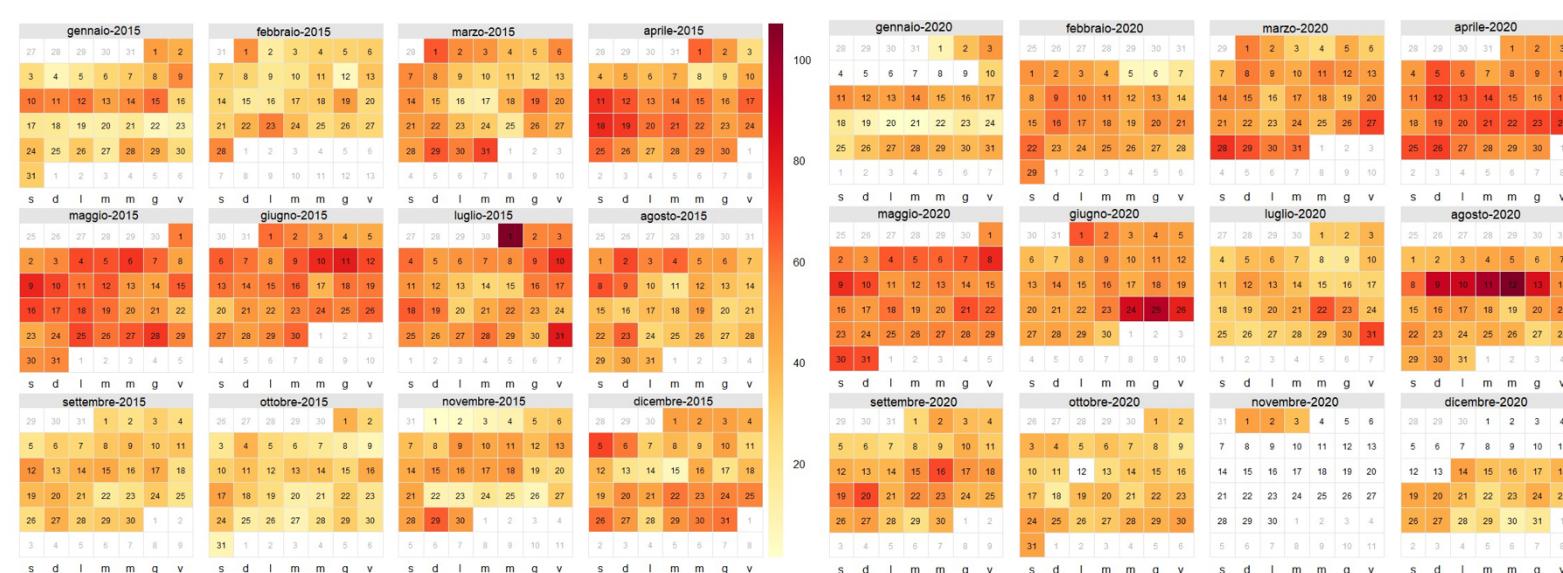
1.2 Calendar Plot

In questi grafici andremo ad osservare le emissioni di due inquinanti: **so2** e **o3**, negli anni 2015 e 2020, mostrando in quali giorni le misurazioni sono più alte.

Calendario so2:

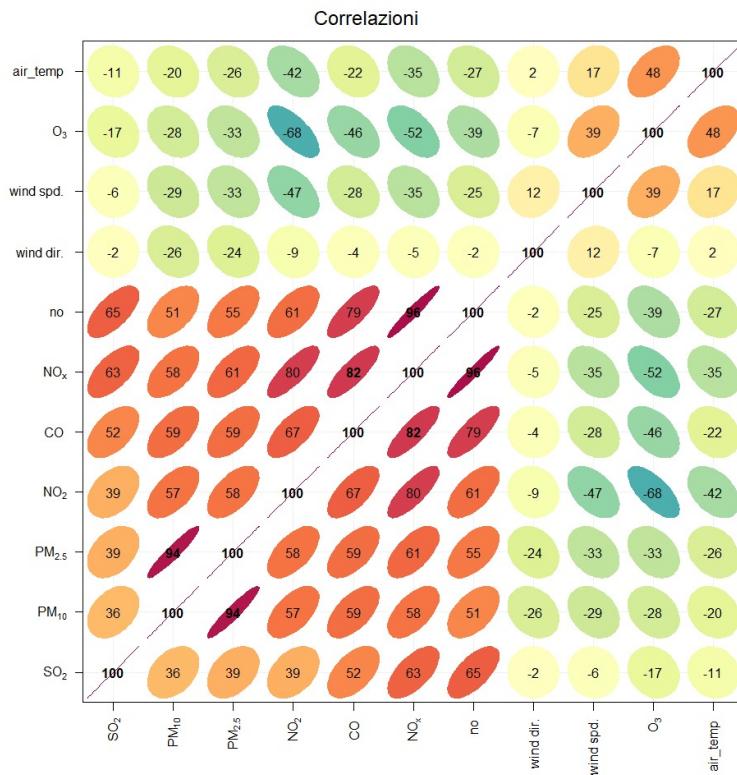


Calendario o3:



1.3 Correlazione

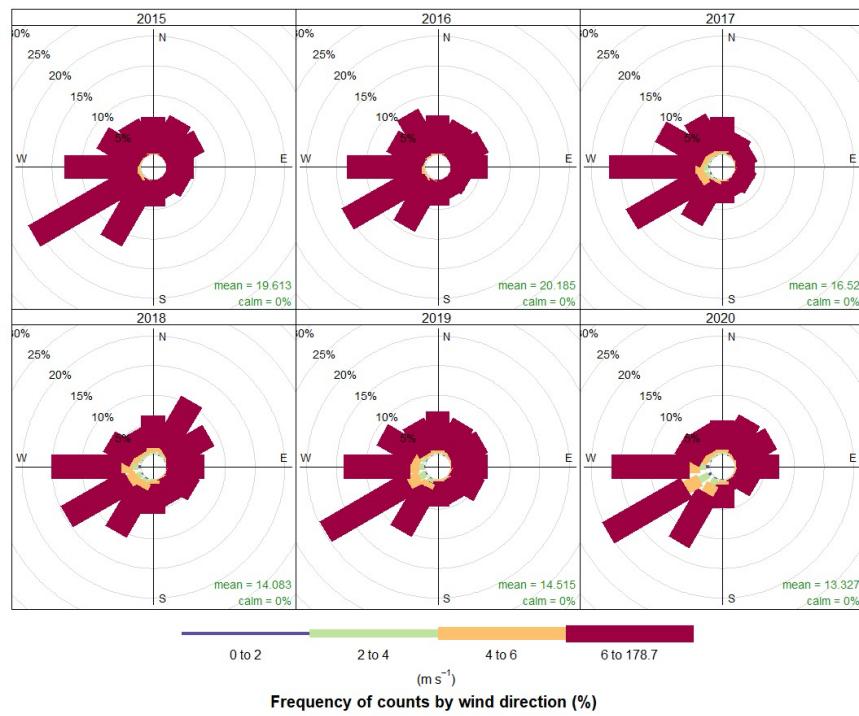
In questo grafico andiamo a studiare le correlazioni tra i vari inquinanti e le caratteristiche atmosferiche.



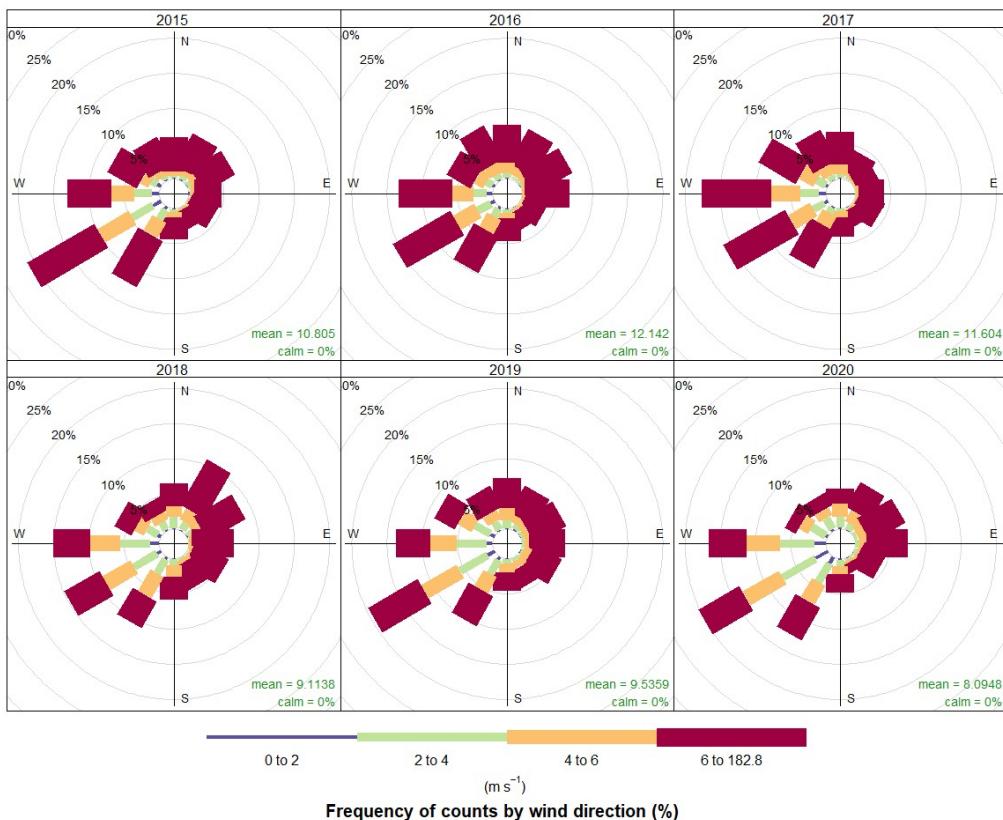
1.4 WindRose e PolarPlot

Andando a eseguire questa analisi puntiamo a comprendere le concentrazioni di inquinanti in base alla direzione del vento, o più specificamente la percentuale di tempo in cui la concentrazione si trova in un particolare intervallo. Questo tipo di approccio può essere molto istruttivo per le specie di inquinanti atmosferici. infine ci permette con il polar plot di comprendere dove sia la porzione di quartiere con il più alto tasso di inquinante.

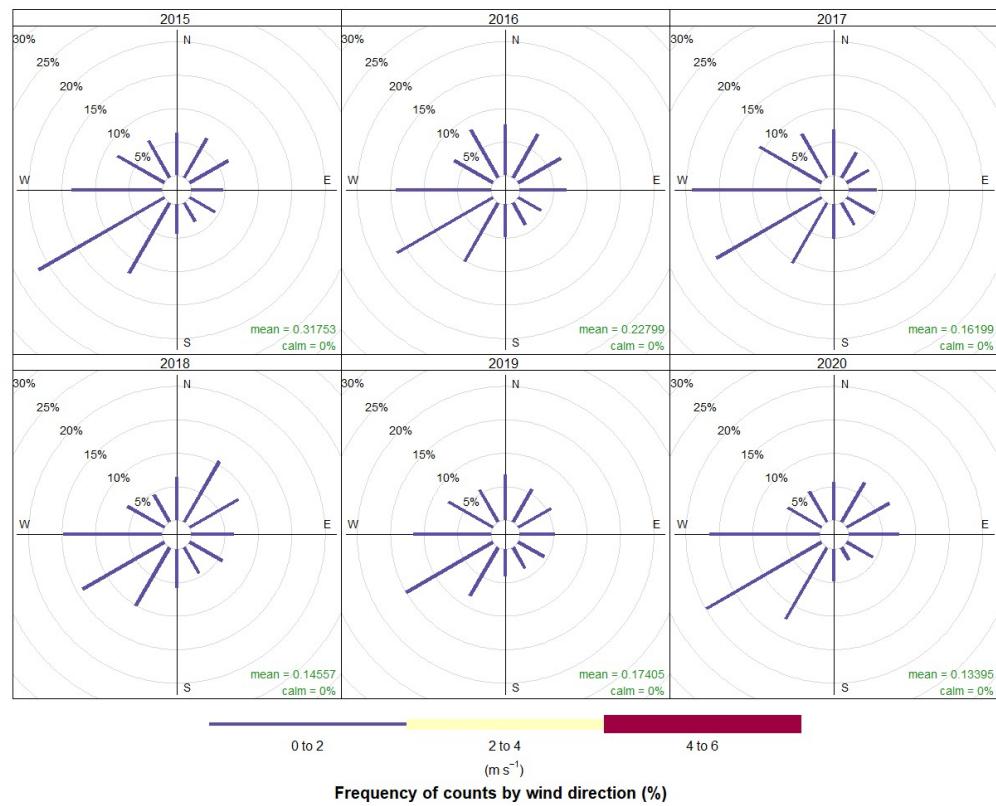
WindPlot Pm10:



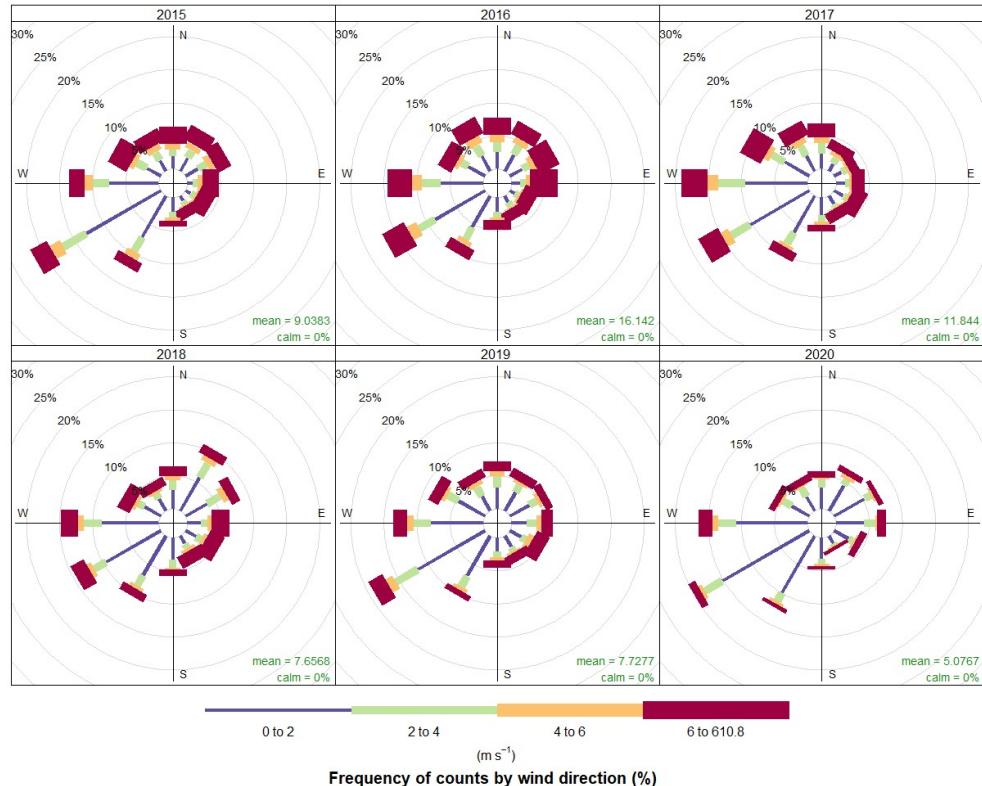
WindPlot Pm 2.5:



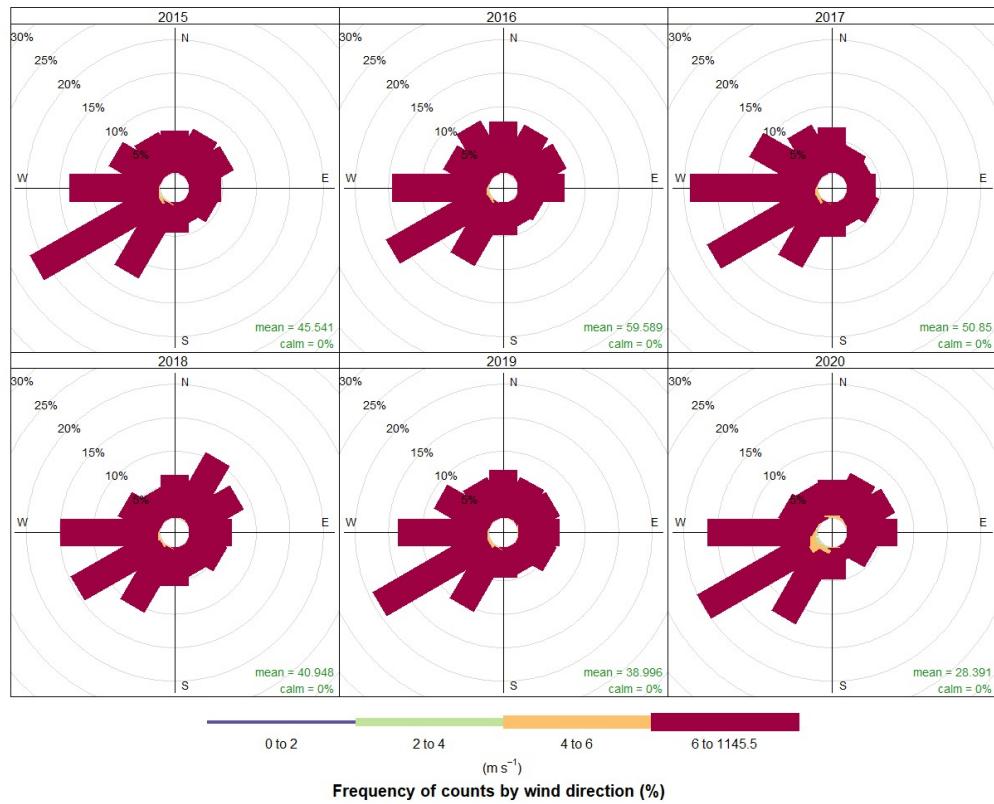
WindPlot CO:



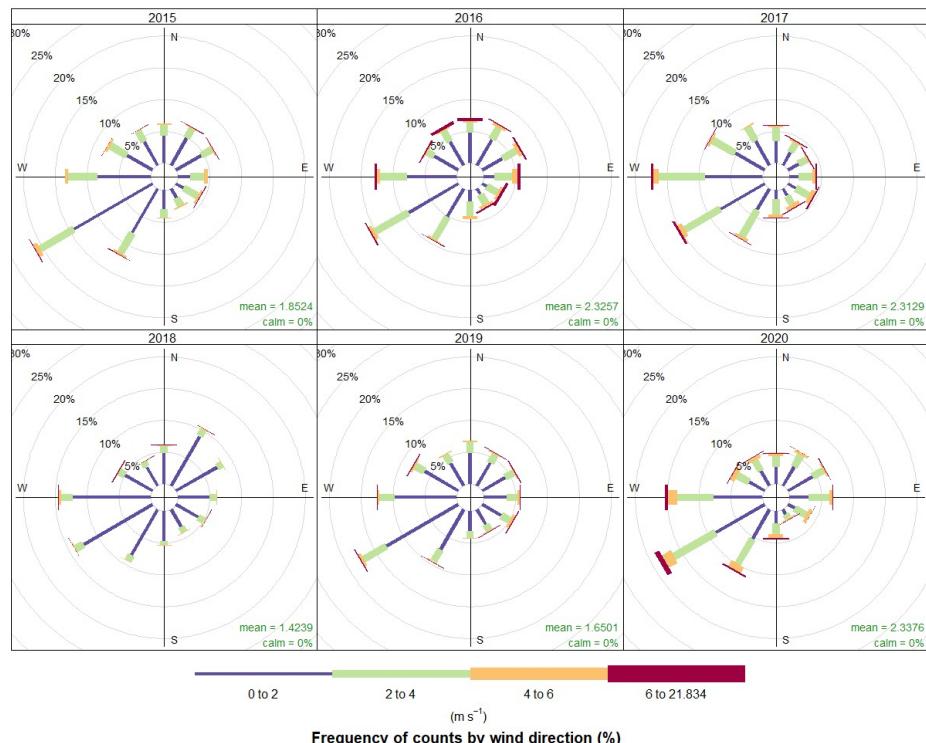
WindPlot NO:



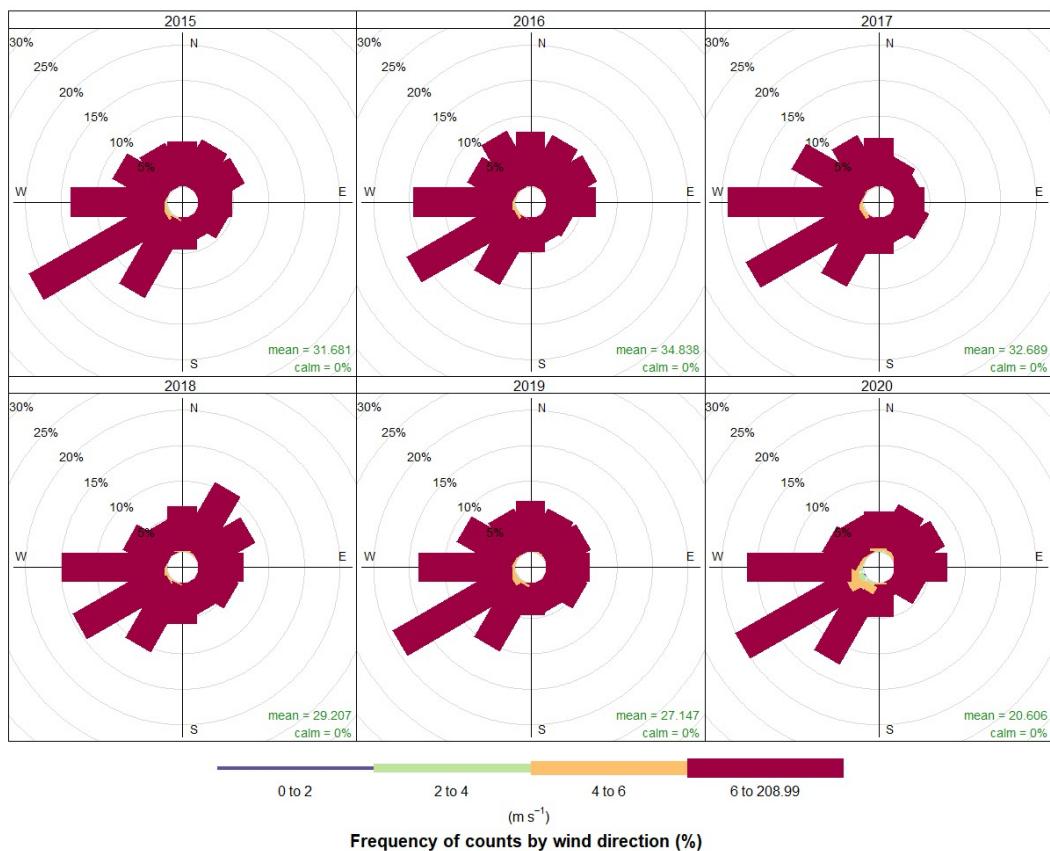
WindPlot NOx:



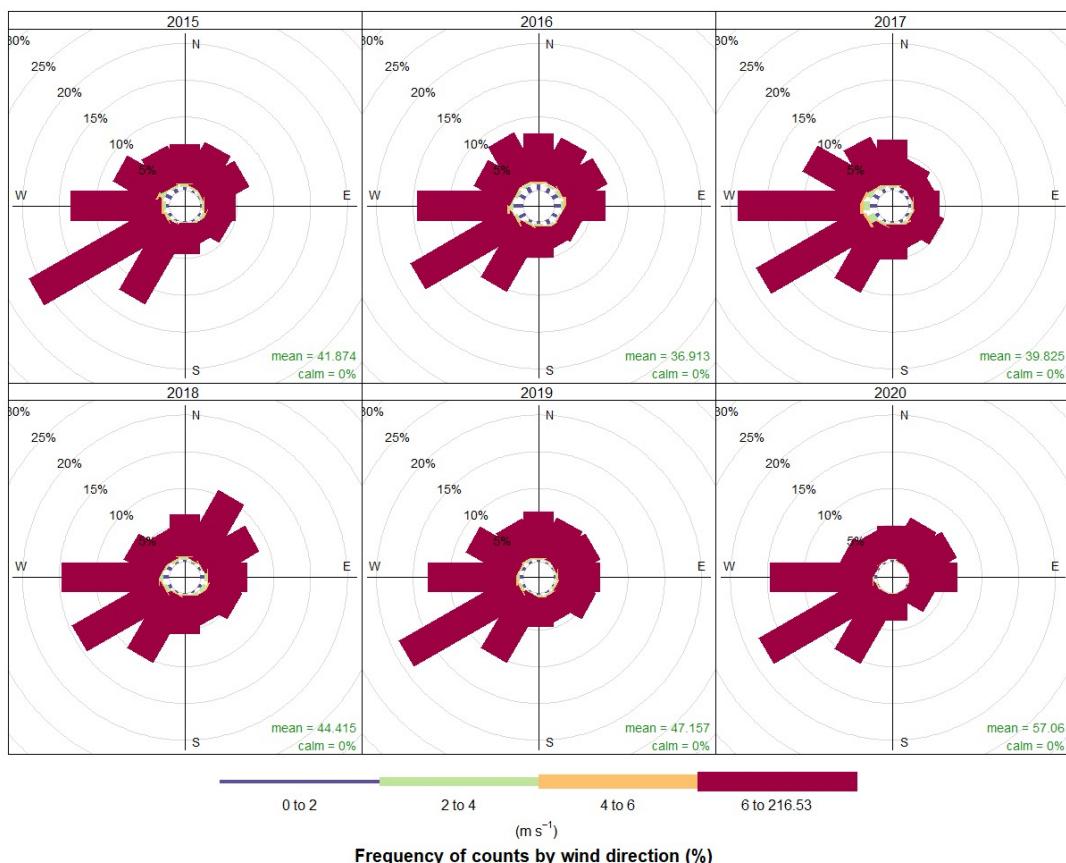
WindPlot SO2:



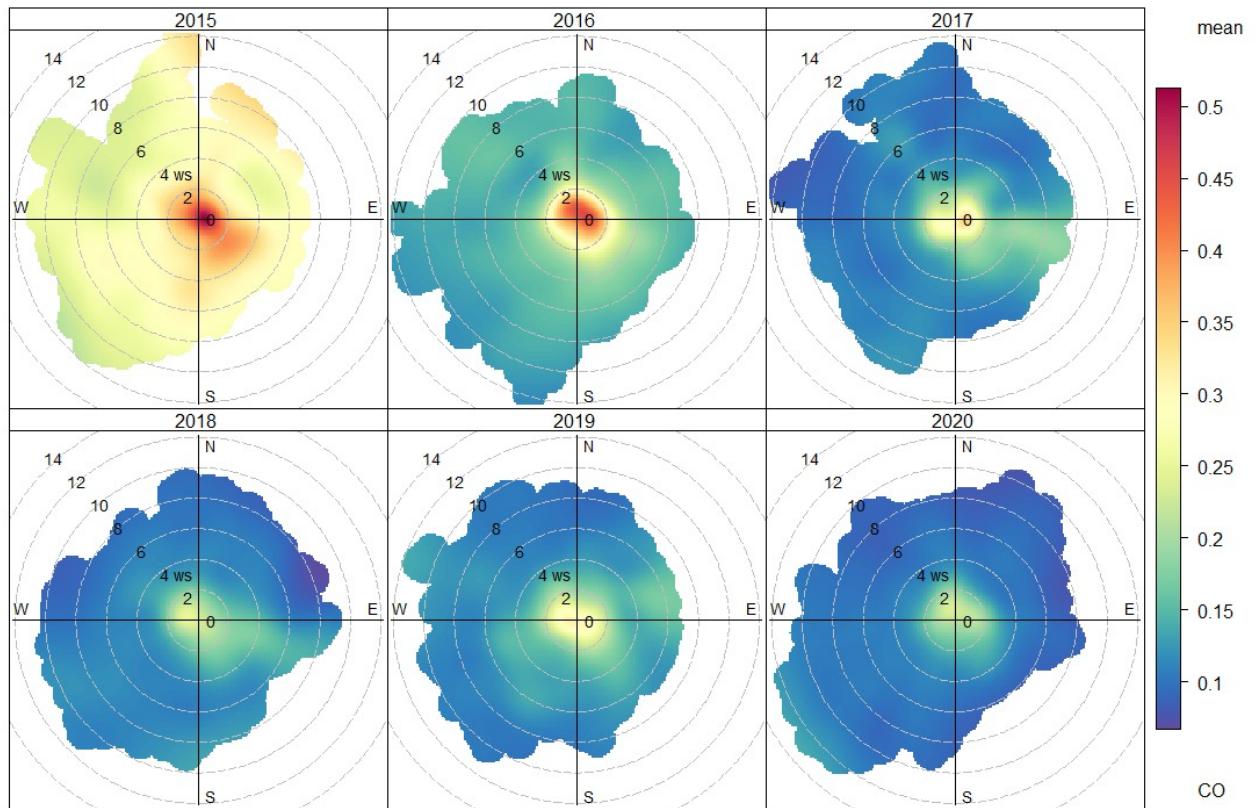
WindPlot NO2:



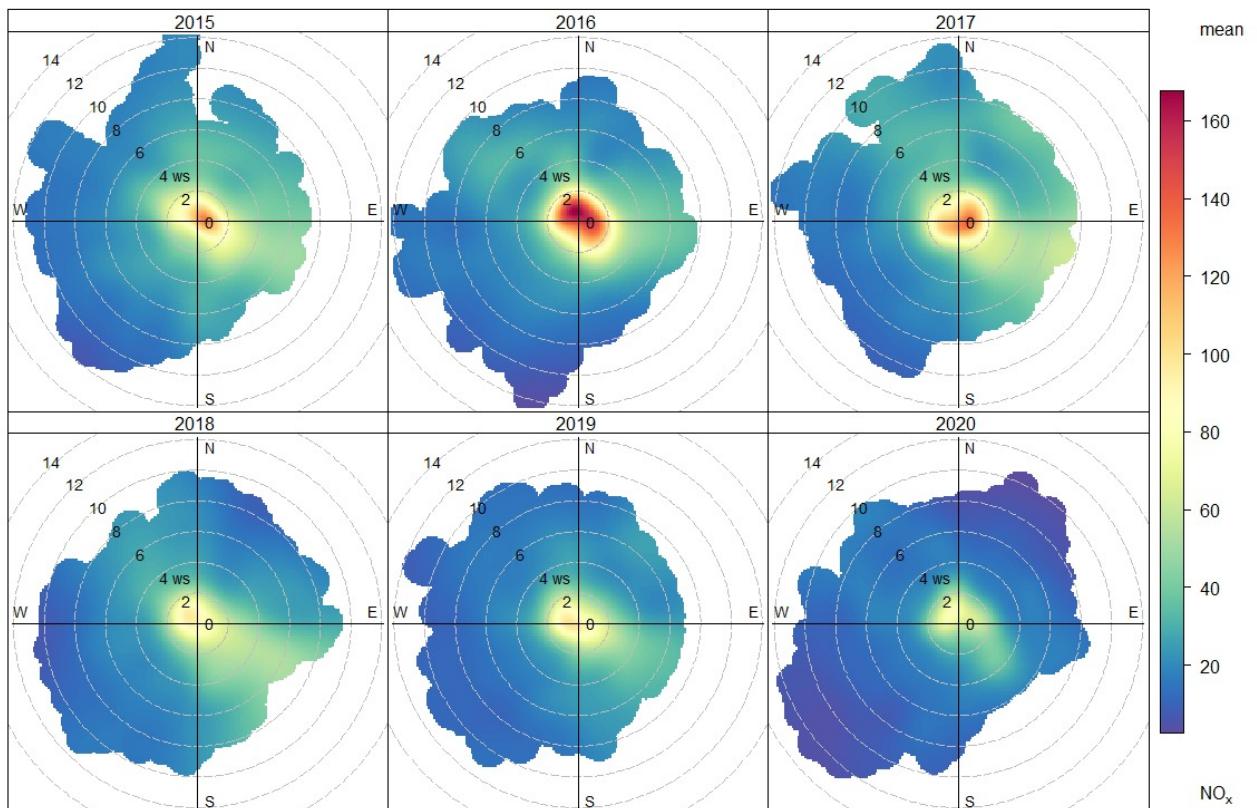
WindPlot O3:



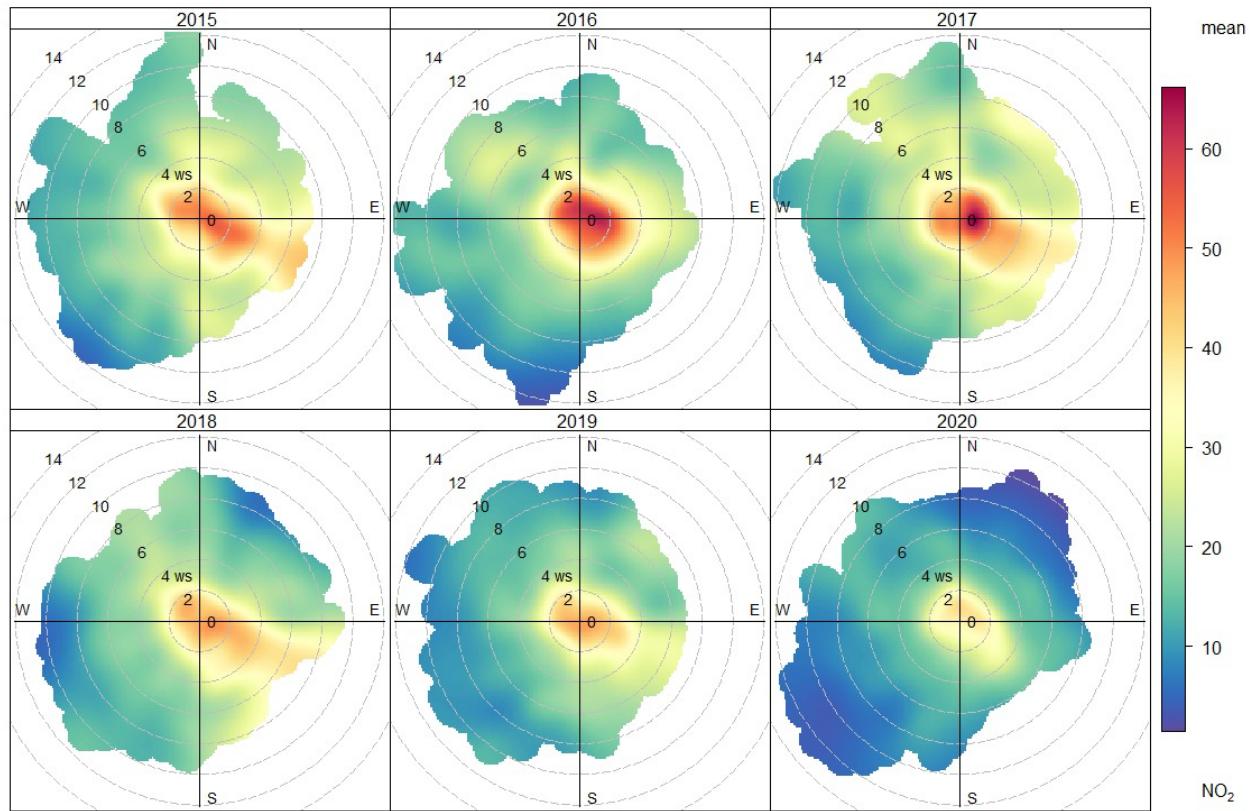
PolarPlot CO:



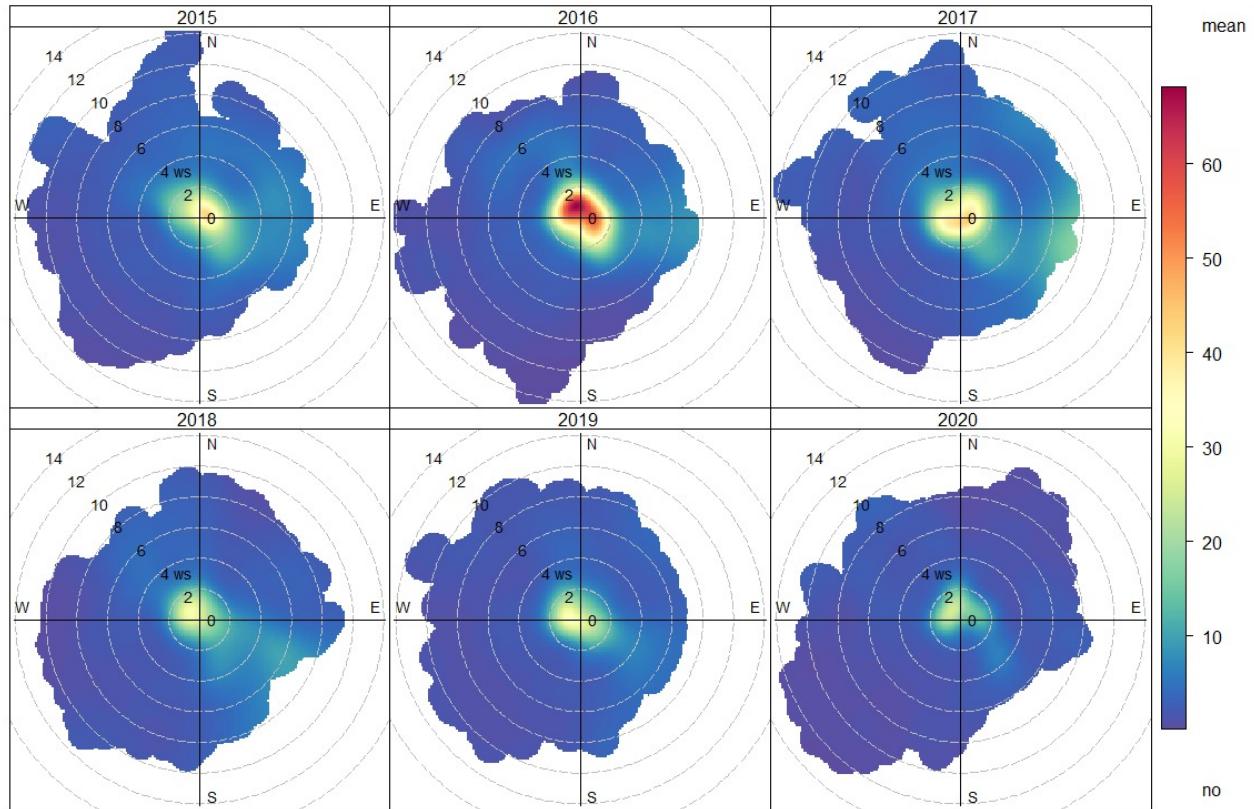
PolarPlot NOx:



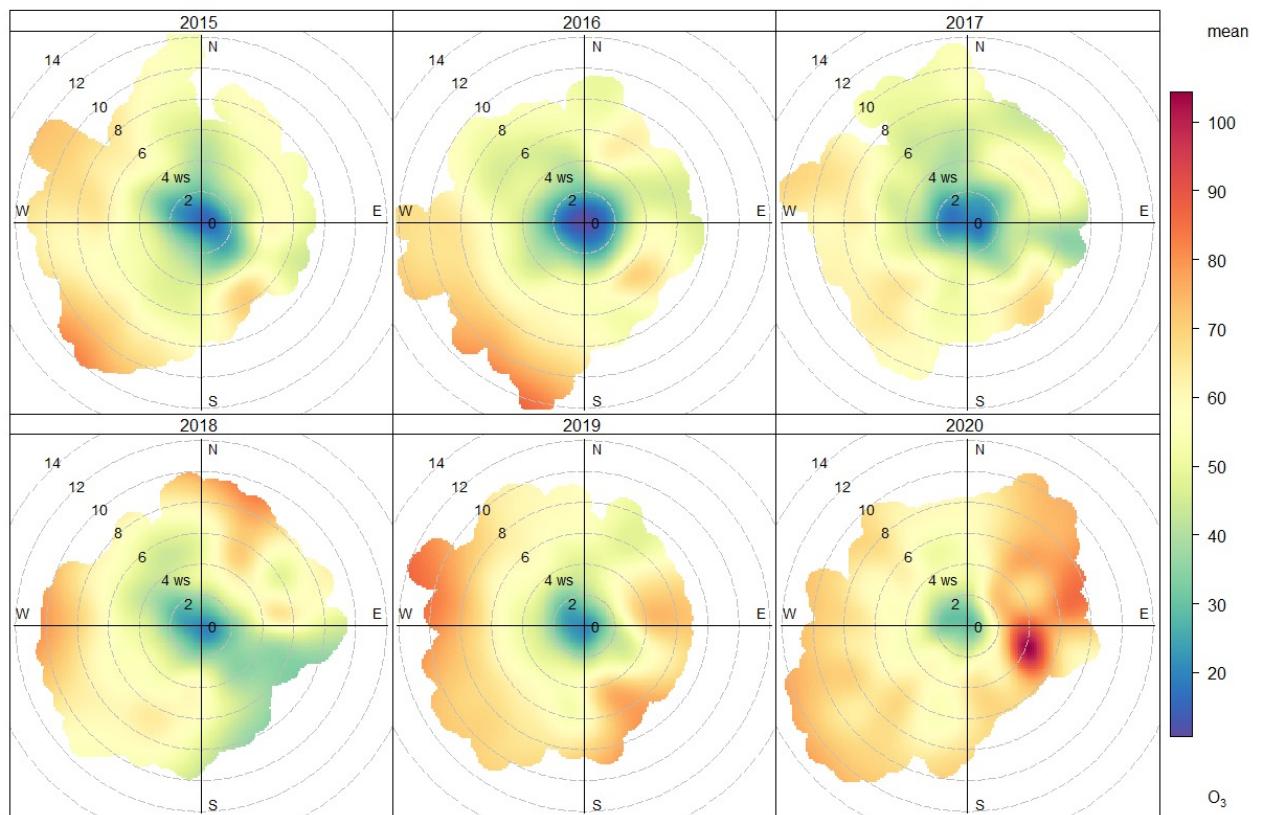
PolarPlot NO2:



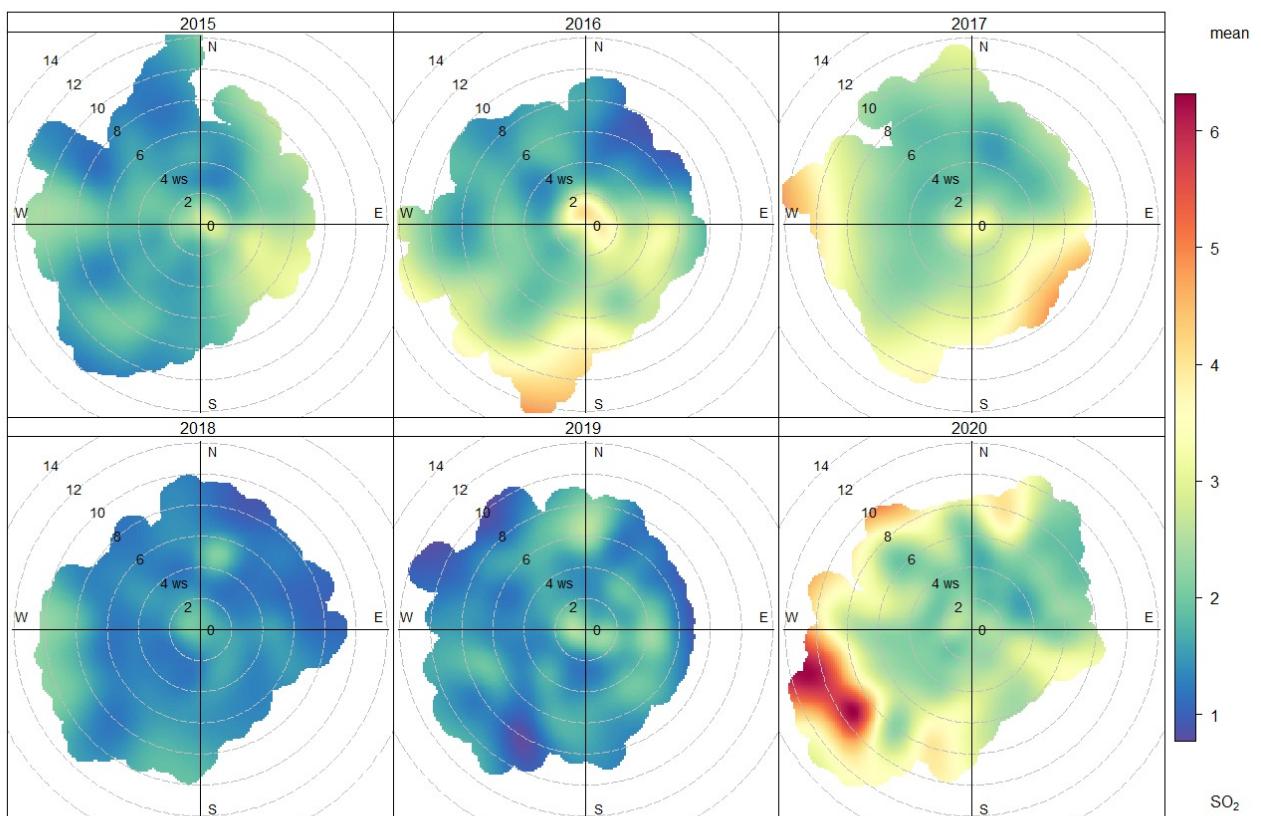
PolarPlot NO:



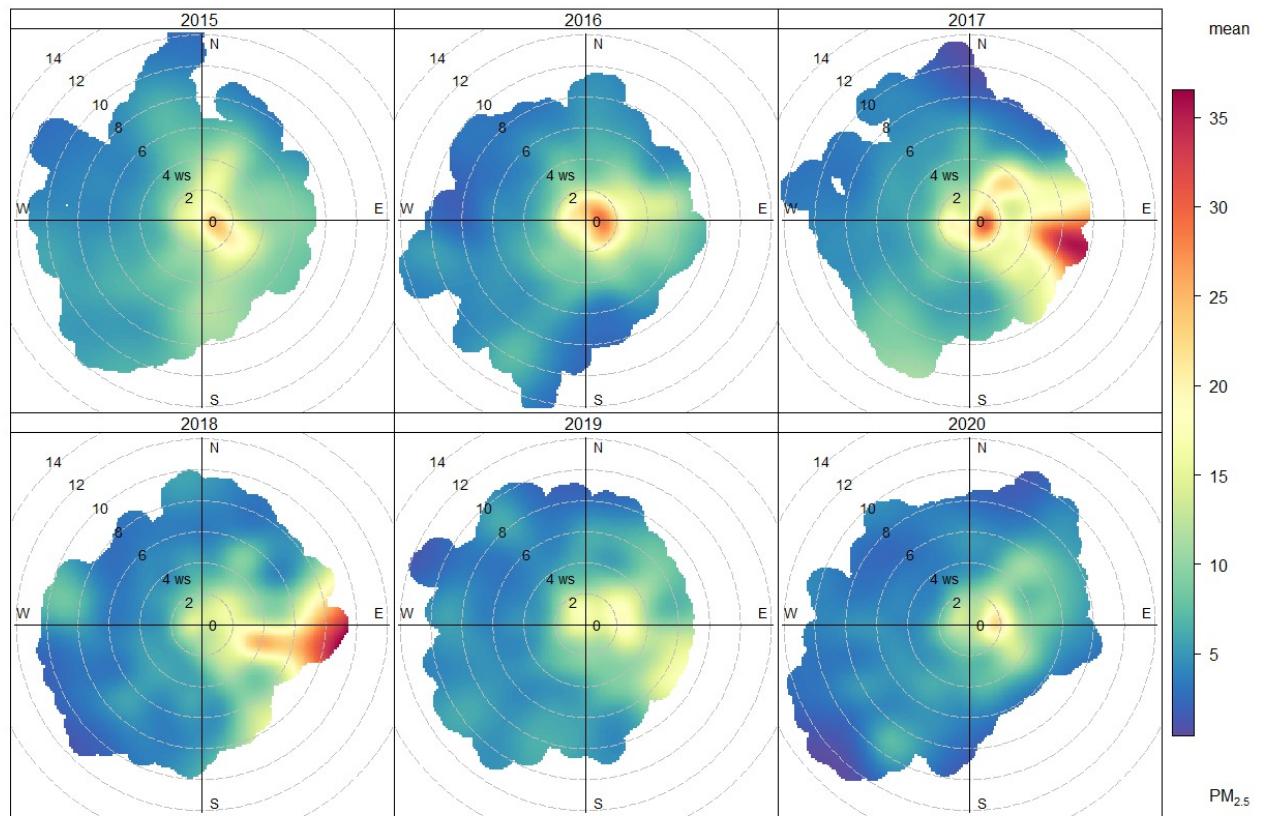
PolarPlot O3:



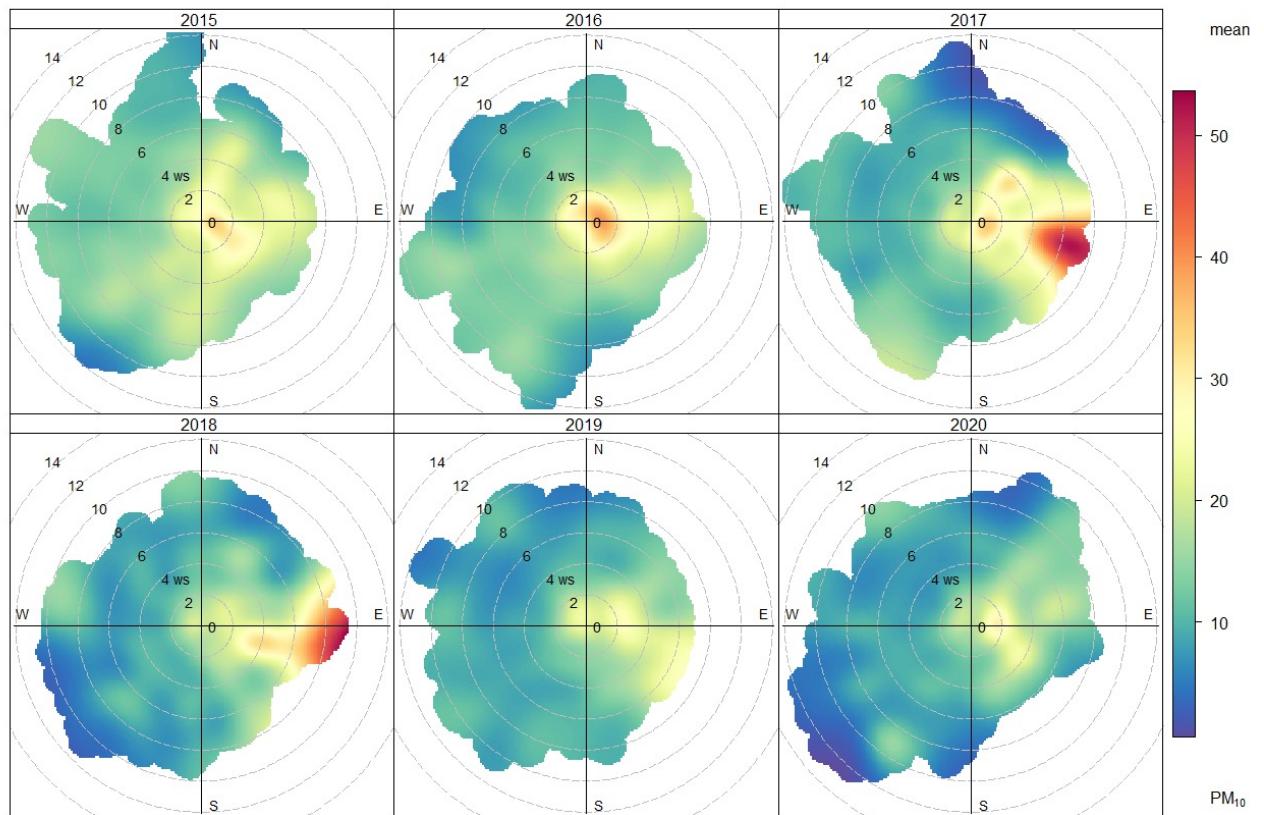
PolarPlot SO₂:



PolarPlot PM 2.5:



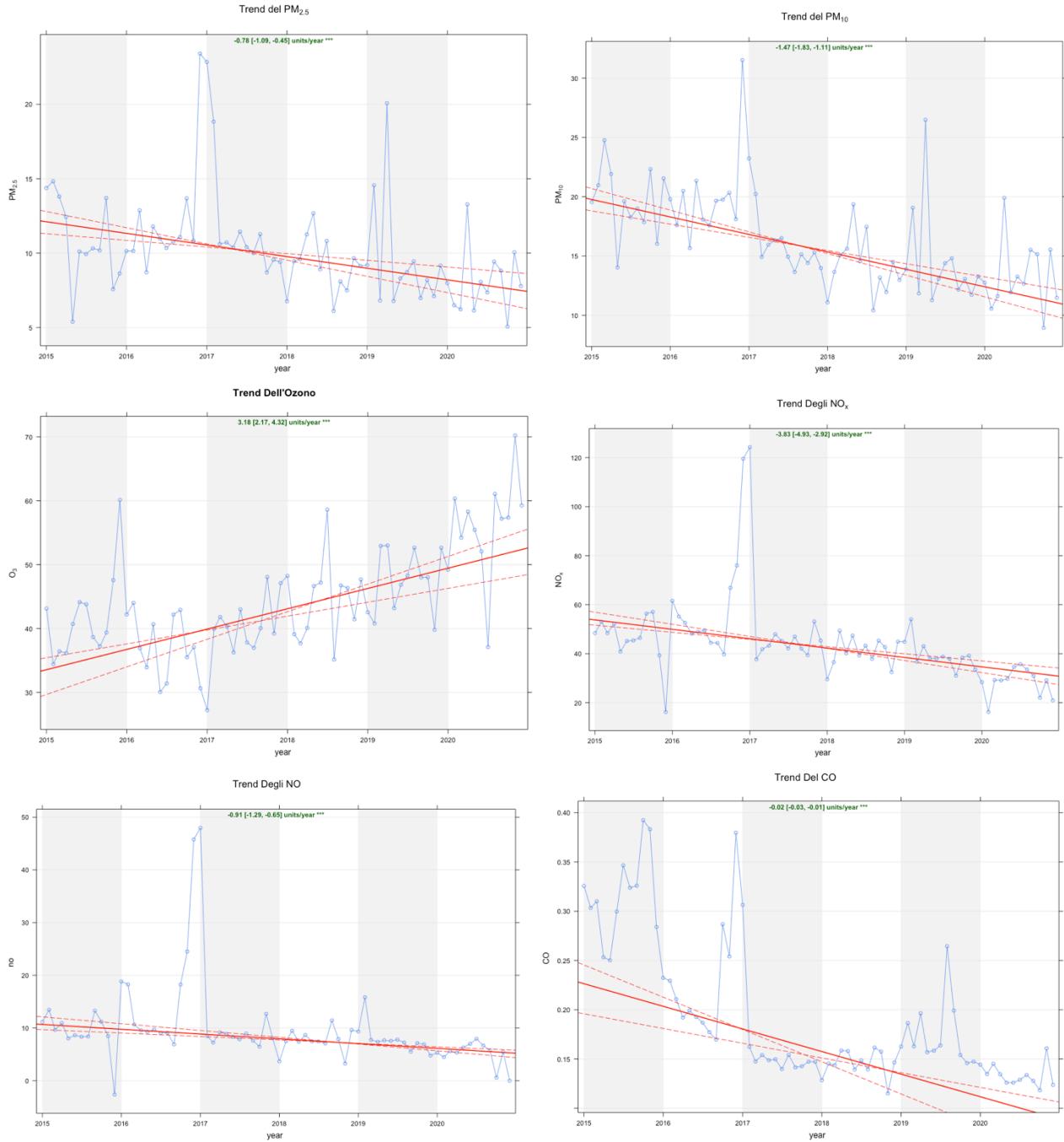
PolarPlot PM 10:

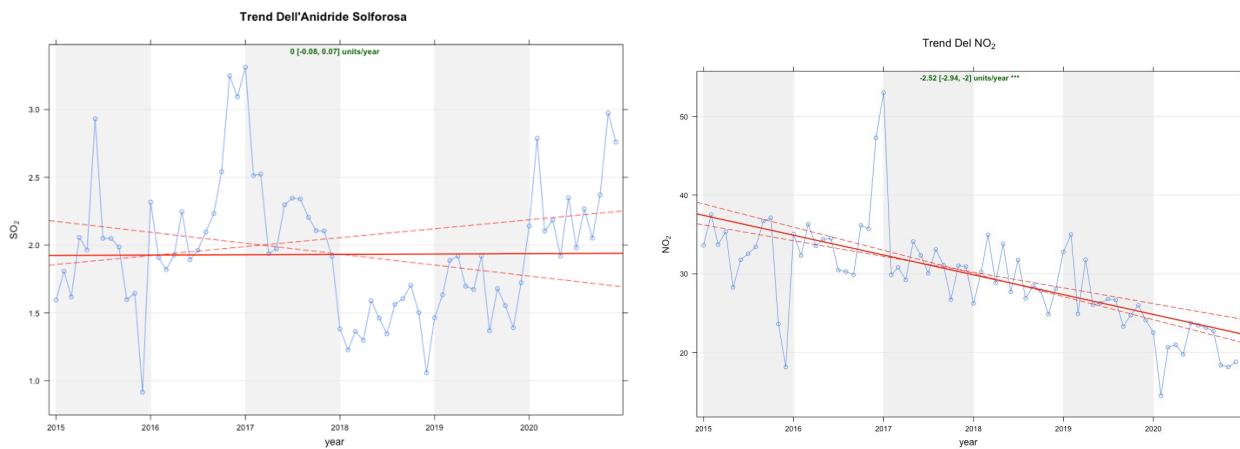


1.5 Analisi del Trend

Il calcolo delle tendenze per gli inquinanti atmosferici è uno dei compiti più importanti e comuni che possono essere intrapresi. Le tendenze sono calcolate per tutti i tipi di motivi. A volte è utile avere un'idea generale di come potrebbero essere cambiate le concentrazioni.

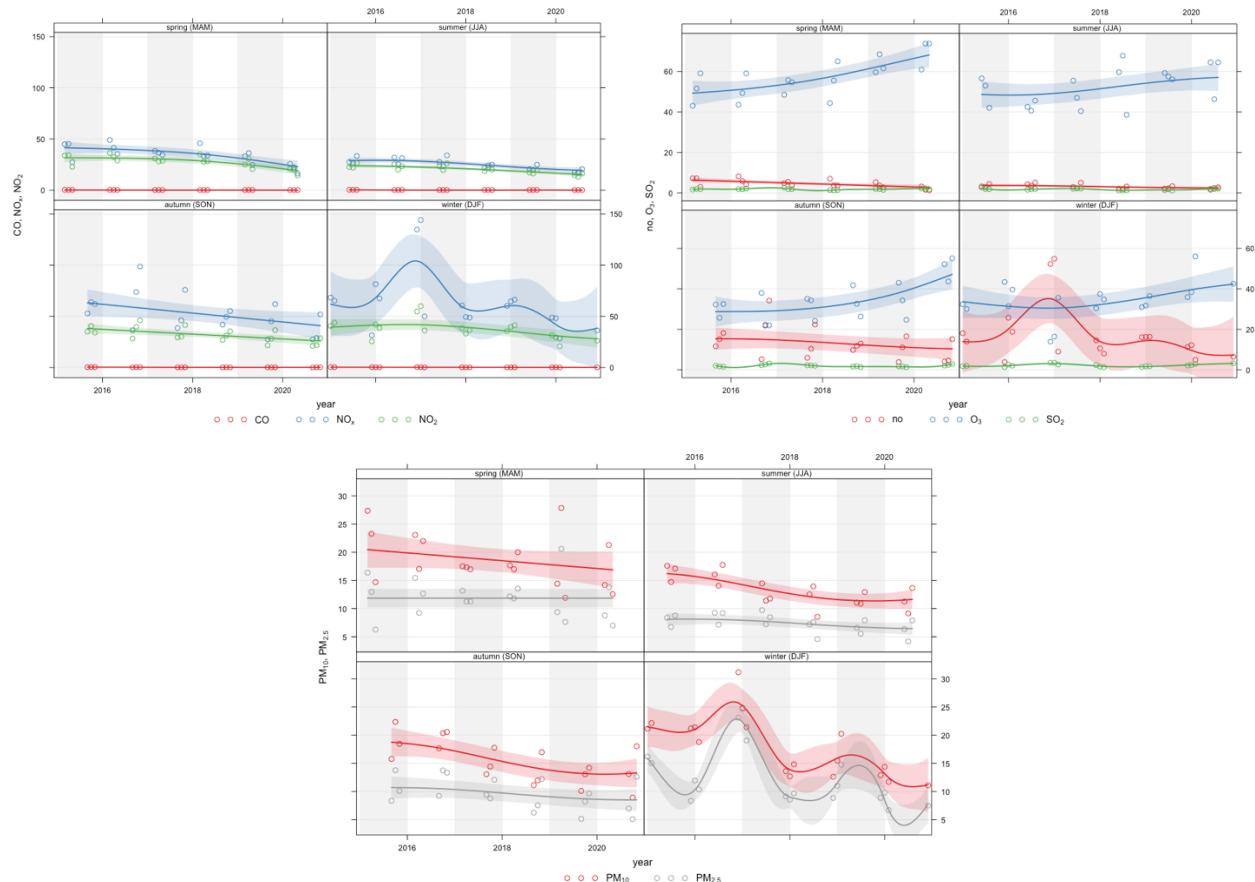
I vantaggi degli approcci non parametrici e delle simulazioni bootstrap. Si noti inoltre che tutti i parametri di regressione vengono stimati tramite il ricampionamento bootstrap.





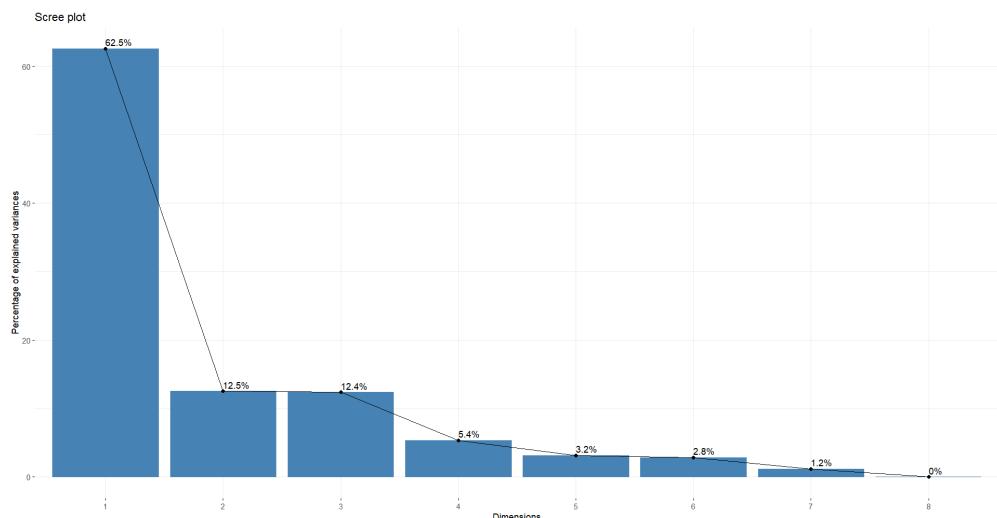
Lo studio di questi trend ci permette di capire l'andamento probabile nei prossimi anni. Tutte le misure effettuate per ridurre i combustibili fossili stanno funzionando facendo sì che la maggior parte degli inquinanti punti a scendere, ma **O₃** e **SO₂** hanno un trend, per il primo pienamente crescente e nel secondo il neutro, che probabilmente dei prossimi anni andrà a crescere.

Andiamo adesso a vedere dove le sottanze aumentano di più o al contrario diminuiscono in base alle stagioni.

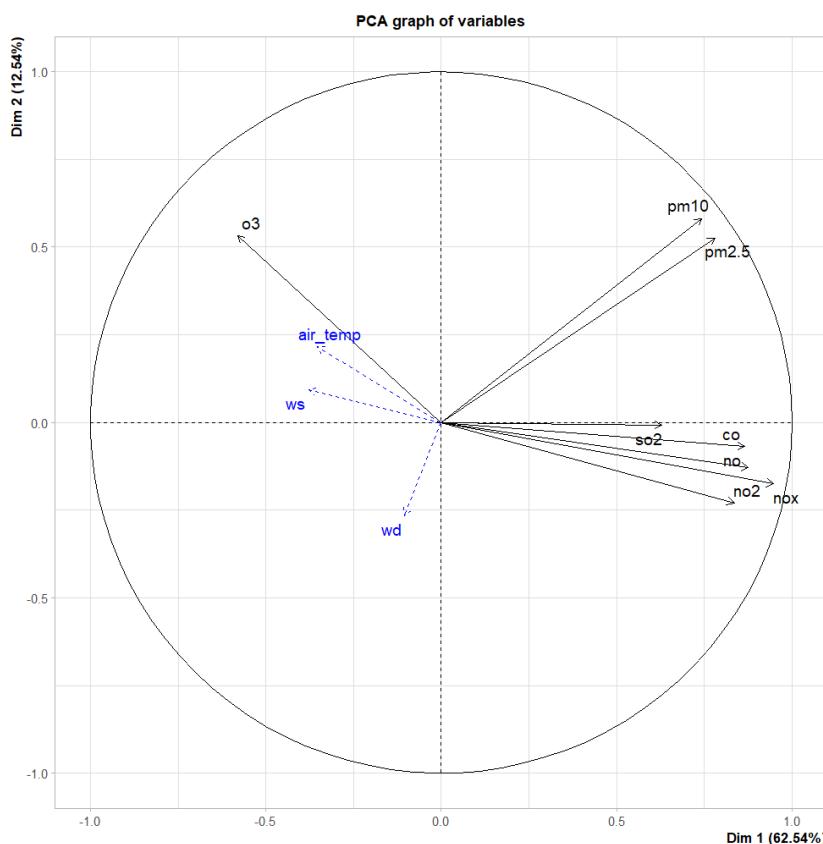


2 PCA

Effettuiamo ora l'analisi sulle componenti principali.

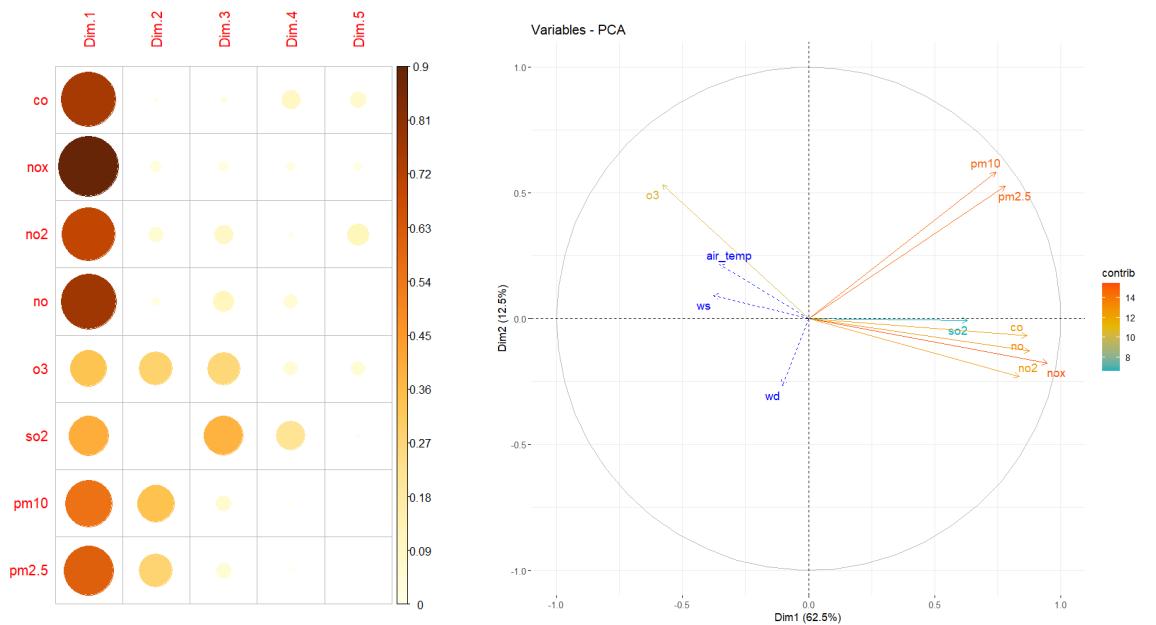


Dallo ScreePlot ne emerge che la prima dimensione spiega la maggior parte dei dati, la seconda e la terza dimensione rimangono quasi sulle stesse percentuali.

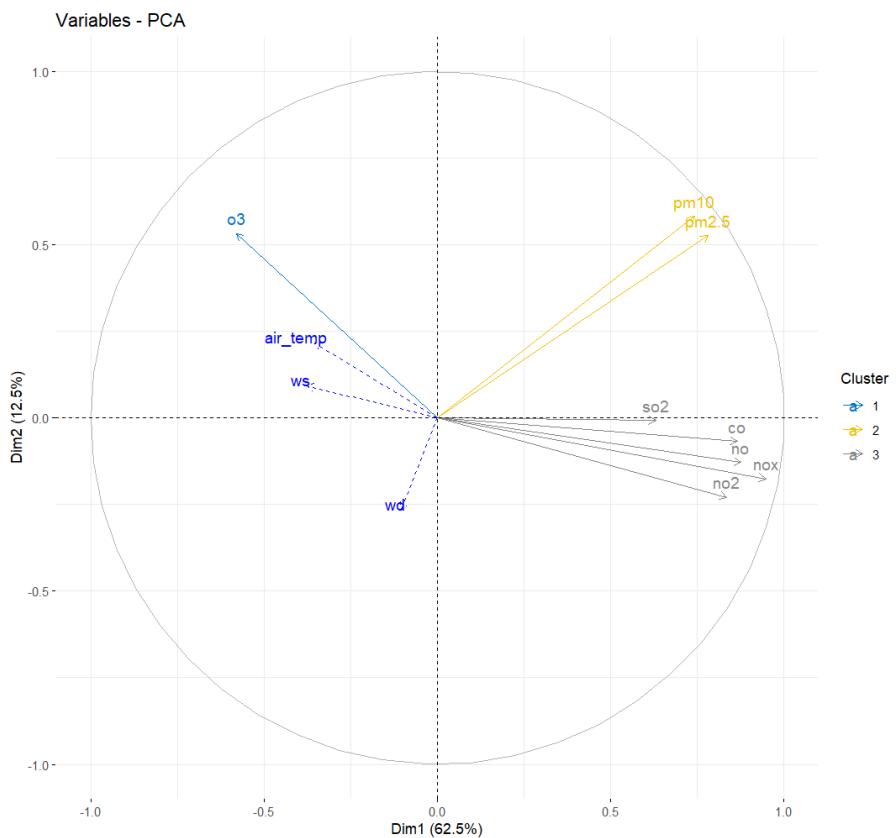


Come si evince da questo grafico la prima dimensione ha verso destra valori che hanno basse **temperature**, bassi livelli di **O₃** e bassa **forza del vento**. Verso sinistra giacciono valori con alte **temperature** alti livelli di **O₃** e alta **intensità di vento**.

Nella seconda dimensione vengono messi agli opposti elementi con alti valori di **Pm10**, **Pm2.5**, **O₃**, **temperatura** dell'aria ed elementi che hanno gli elementi precedenti bassi.



Attraverso questo grafico di sinistra andiamo a comprendere quali sono gli inquinanti che contribuiscono maggiormente nelle varie dimensioni, nel grafico di destra vediamo il contributo di ogni variabile per la prima e la seconda dimensione.



Con questo grafico andiamo a clusterizzare le variabili tramite le k-means, formando così dei gruppi che hanno tutte caratteristiche simili.

3 PCR

La regressione delle componenti principali (PCR) è una tecnica di regressione basata sull'analisi delle componenti principali (PCA).

L'idea alla base della PCR è calcolare i componenti principali e quindi utilizzare alcuni di questi componenti come predittori in un modello di regressione lineare adattato utilizzando la tipica procedura dei minimi quadrati.

Come puoi facilmente notare, l'idea centrale della PCR è strettamente correlata a quella sottostante PCA e il "trucco" è molto simile. In alcuni casi è sufficiente un piccolo numero di componenti principali per spiegare la stragrande maggioranza della variabilità dei dati. Ad esempio, supponiamo che tu abbia un set di dati di 50 variabili che desideri utilizzare per prevedere una singola variabile. Utilizzando la PCR potresti scoprire che 4 o 5 componenti principali sono sufficienti per spiegare il 90% della varianza dei tuoi dati. In questo caso, potrebbe essere meglio eseguire la PCR con questi 5 componenti invece di eseguire un modello lineare su tutte le 50 variabili. Questo è un esempio approssimativo, ma spero che abbia aiutato a raggiungere il punto.

Un presupposto fondamentale della PCR è che le direzioni in cui i predittori mostrano la maggior variazione siano le direzioni esatte associate alla variabile di risposta. Da un lato, questa ipotesi non è garantita per il 100% delle volte, tuttavia, anche se l'ipotesi non è completamente vera, può essere una buona approssimazione e produrre risultati interessanti.

Alcuni dei vantaggi più notevoli dell'esecuzione della PCR sono i seguenti:

- Riduzione della dimensionalità
- Prevenzione della multicollinearità tra predittori
- Mitigazione del sovraccarico

Per lo svolgimento di questa analisi abbiamo posto come variabile dipendente ogni nostro inquinante e come variabili indipendenti tutti gli altri inquinanti.

Y= NOx:

VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
comps	59.72	24.3	24.3	17.65	16.13	15.4	0.5329
CV	59.72	24.3	24.3	17.65	16.13	15.4	0.5328

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
X	59.96	74.25	87.60	93.13	96.44	99.20	100
nox	83.45	83.46	91.27	92.71	93.36	99.99	100

Y= NO:

VALIDATION: RMSEP

Cross-validated using 10 random segments.

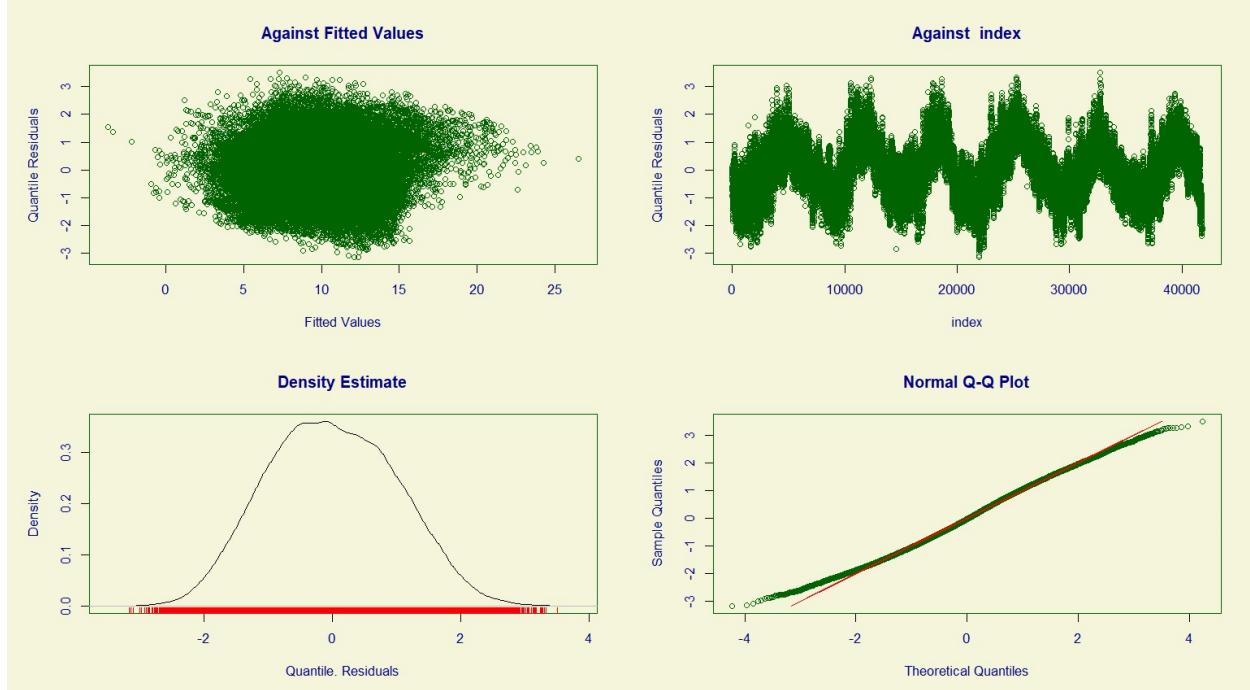
	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	29.61	17.21	17.21	13.46	12.48	12.42	1.705
adjCV	29.61	17.21	17.21	13.46	12.48	12.42	1.704

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
X	62.12	76.44	88.78	94.01	97.55	99.21	100
no	66.25	66.26	79.37	82.25	82.42	99.67	100

4 Regressione

L'obiettivo di questa regressione è cercare di capire quanto sia influenzata la temperatura da parte di tutti gli agenti inquinanti.



	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.946884	0.113227	52.522	<2e-16	***
co	10.058627	0.301893	33.319	<2e-16	***
nox	2.285784	3.052683	0.749	0.454	
no2	-2.340692	3.052691	-0.767	0.443	
no	-3.550171	4.680714	-0.758	0.448	
o3	0.078867	0.001375	57.342	<2e-16	***
so2	0.360338	0.023515	15.324	<2e-16	***
pm10	0.100007	0.006453	15.498	<2e-16	***
pm2.5	-0.165794	0.007551	-21.956	<2e-16	***

L'obiettivo di questa regressione invece è quella di capire se con l'aumento dell'intensità del vento i valori Pm10 e Pm2.5 diminuiscono.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.8314538	0.0228437	123.949	< 2e-16	***
co	-0.0241868	0.0648595	-0.373	0.709	
o3	0.0196944	0.0003127	62.985	< 2e-16	***
pm10	0.0149039	0.0018211	8.184	2.82e-16	***
pm2.5	-0.0520181	0.0021466	-24.233	< 2e-16	***

Da quello che si nota quando il vento aumenta d'intensità il particolato più fine tende a diminuire, ciò è causato dal fatto che essendo più vicino al terreno il vento tenda a spostarlo in modo maggiore che i Pm10 che giacciono in altitudini più elevate.

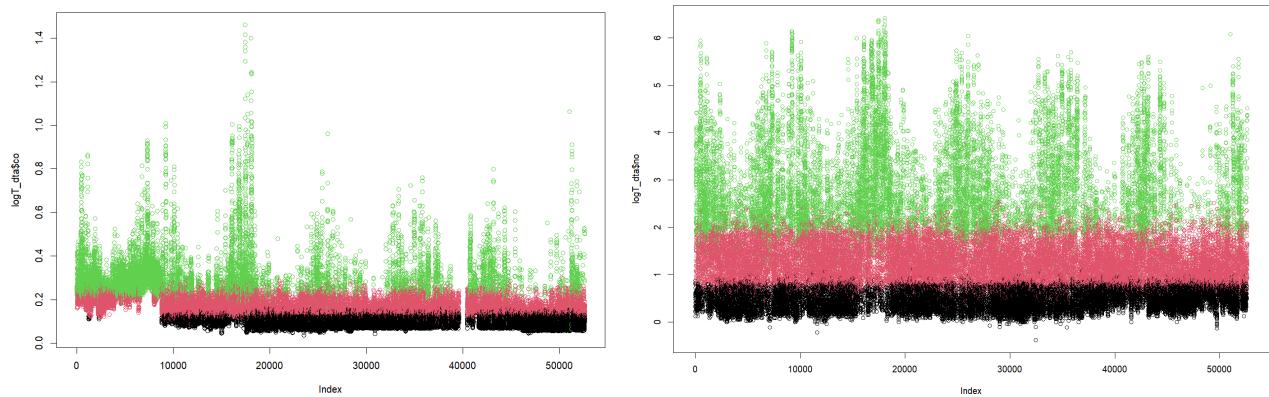
5 HMM

5.1 Univariati

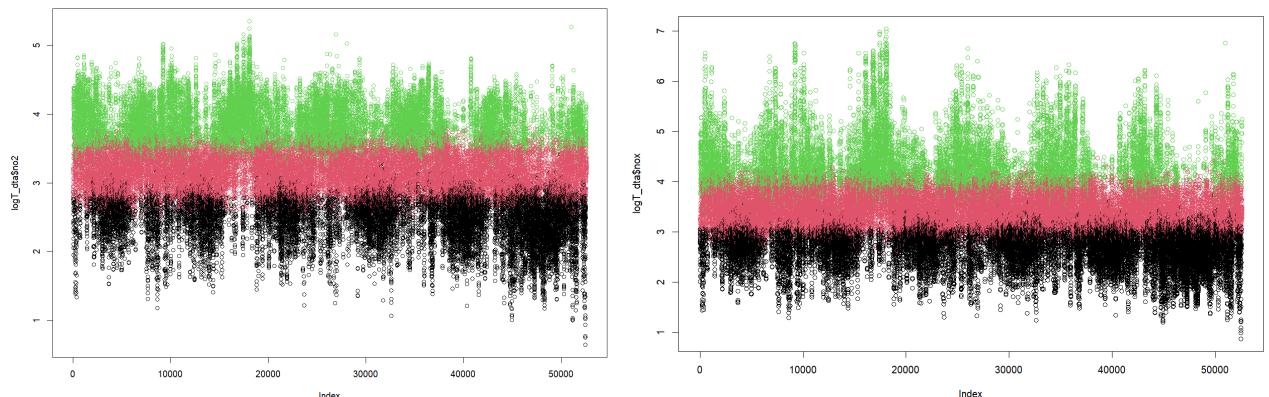
Per questa analisi andremo ad utilizzare gli hidden markovs models uni-variatii per andare a creare delle "classi" di elementi nel nostro dataset.

Il primo approccio che abbiamo scelto di eseguire è quello che divide gli elementi di ogni colonna in 3 gruppi. L'idea è quella di andare a descrivere i valori: alti, medi e bassi.

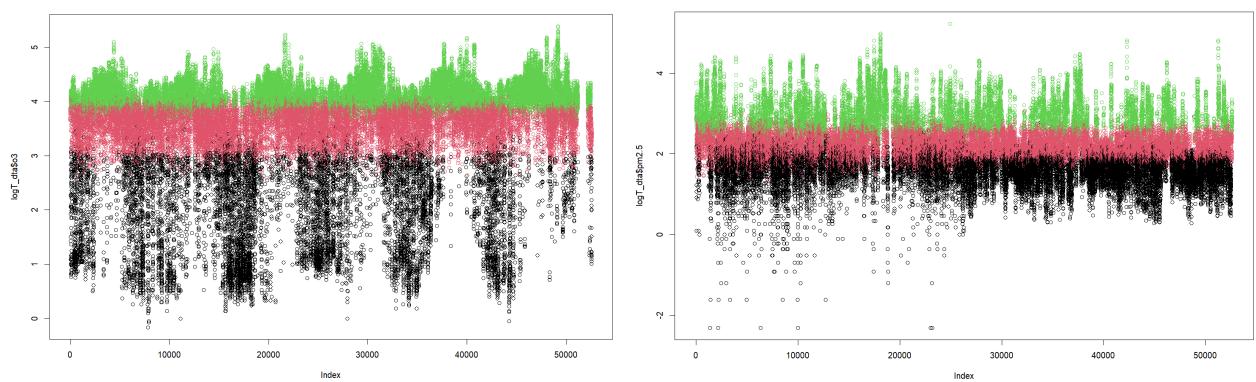
CO e NO:



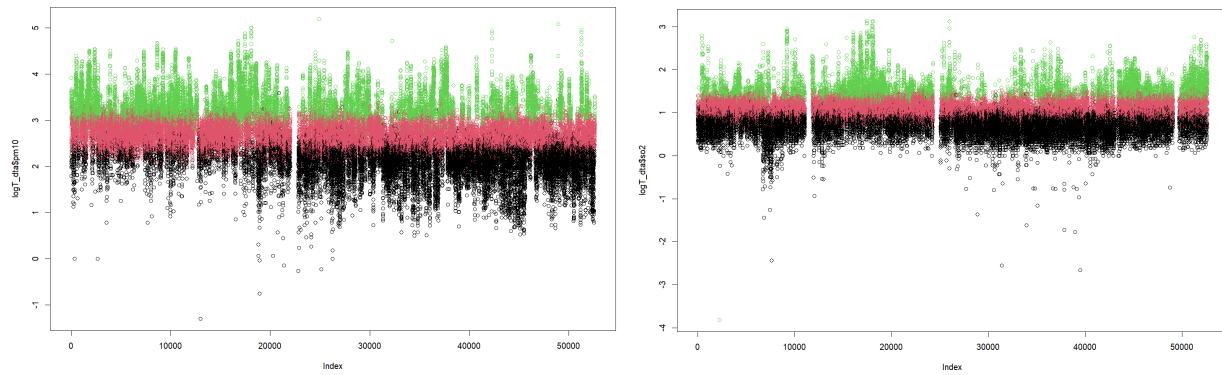
NO₂ e NO_x:



O₃ e Pm2.5:

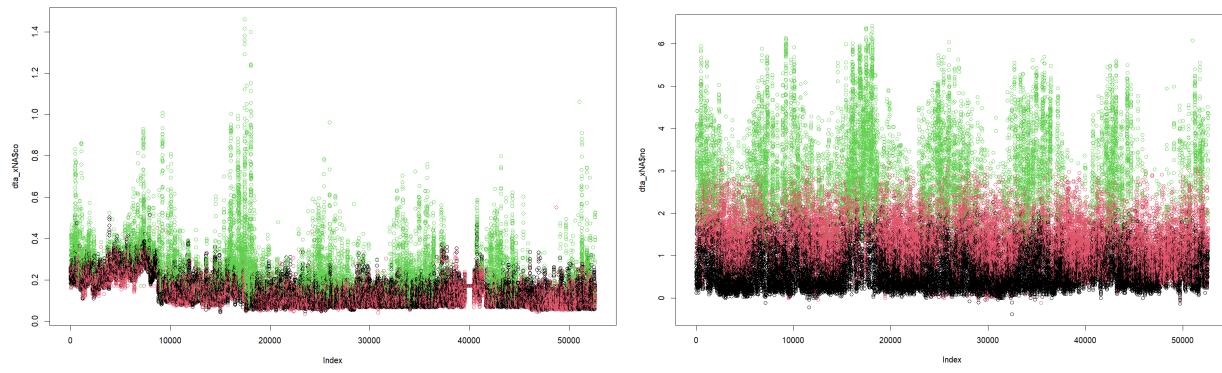


Pm10 e SO2:

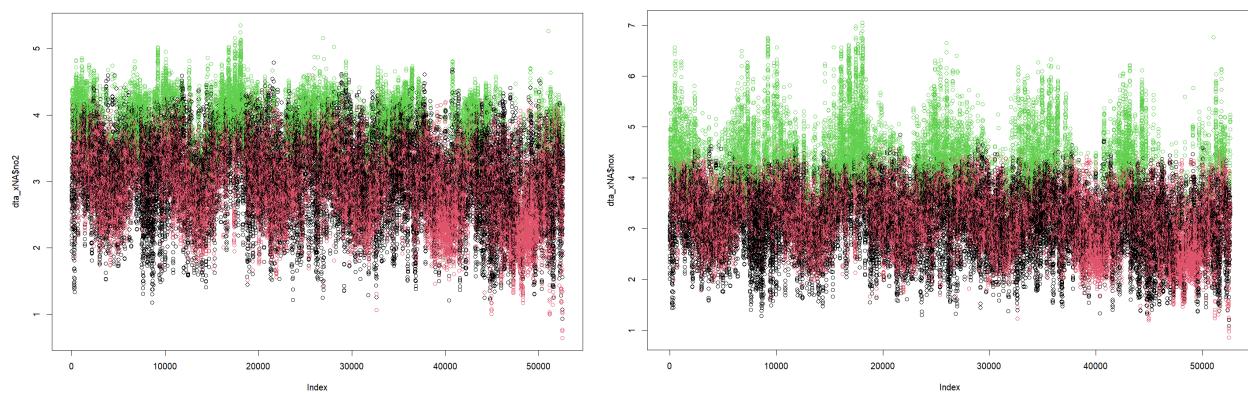


5.2 Multivariati

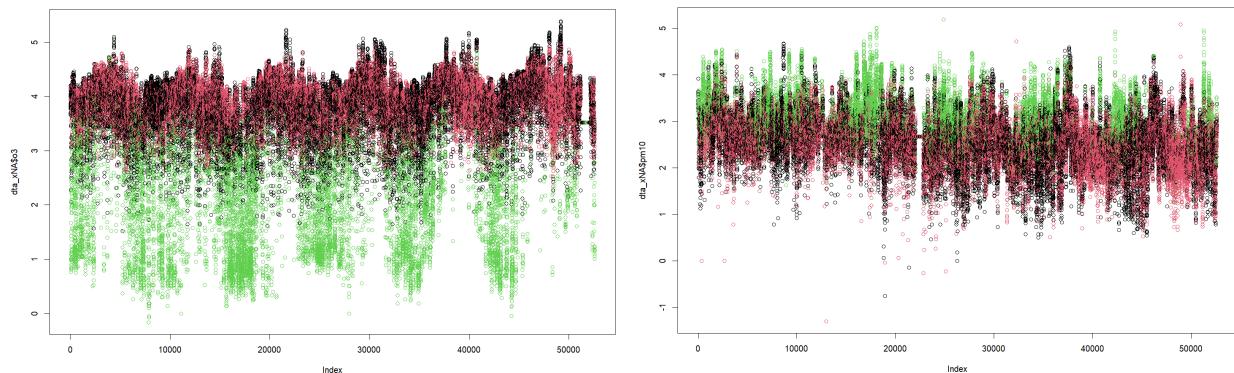
CO e NO:



NO2 e NOx:



O3 e Pm10:



SO2:

