

---

## CITY STATISTICS

---

### Analisi caratteristiche delle FUA dei paesi OECD

Pasquale Somma  
Marco Di Maio  
*Università Degli Studi di Salerno*

# Contents

<b>1</b>	<b>Presentazione Dataset</b>	<b>3</b>
<b>2</b>	<b>Domande di Ricerca</b>	<b>3</b>
<b>3</b>	<b>Preprocessing del Dataset</b>	<b>4</b>
3.1	Data imputation . . . . .	5
<b>4</b>	<b>Analisi esplorativa</b>	<b>7</b>
4.1	GDP . . . . .	7
4.2	GDP pro capite . . . . .	11
4.3	Labour force . . . . .	15
4.4	Mean population exposure to PM2.5 air pollution . . . . .	17
4.5	Population density . . . . .	22
4.6	Population, all ages, administrative data . . . . .	25
4.7	Unemployment rate . . . . .	29
<b>5</b>	<b>Risoluzione domande di ricerca</b>	<b>33</b>
5.1	Domanda di ricerca 1 . . . . .	33
5.2	Domanda di ricerca 2 . . . . .	37
5.3	Domanda di ricerca 3 . . . . .	41
5.4	Domanda di ricerca 4 . . . . .	44
5.4.1	Cluster gerarchici . . . . .	45
5.4.2	Cluster non gerarchici . . . . .	50
5.4.3	Confronto metodi . . . . .	51
5.5	Domanda di ricerca 5 . . . . .	52
5.5.1	Selezione del Campione . . . . .	52
5.5.2	Adattamento alla Distribuzione Normale . . . . .	53
5.5.3	Stima Puntuale dei Parametri . . . . .	53
5.5.4	Stima Intervallare . . . . .	54
5.5.5	Confronto tra Due Popolazioni . . . . .	54
5.5.6	Verifica delle Ipotesi . . . . .	56
5.5.7	Conclusioni . . . . .	57

# 1 Presentazione Dataset

Il dataset **City Statistics** è stato realizzato dall'OECD, in collaborazione con l'UE, per sviluppare una definizione armonizzata di aree urbane funzionali (AUF) che comprendono una città e la sua zona pendolare, riflettendo l'estensione economica basata sugli spostamenti giornalieri delle persone. Questo supera le limitazioni degli approcci amministrativi, fornendo una definizione funzionale ed economica delle città. Il database fornisce indicatori socio-economici e ambientali per oltre 1.256 AUF in 40 paesi. Le AUF sono definite utilizzando una griglia di popolazione e flussi pendolari, garantendo un collegamento minimo al livello governativo. Il database include indicatori modellati su dati aggregati o geospatiali, offrendo una visione completa delle dinamiche urbane, spaziando tra demografia, migrazione, economia, lavoro, inquinamento, digitalizzazione, trasporti pubblici, disastri naturali e organizzazione territoriale.

## 2 Domande di Ricerca

Nell'ambito dell'analisi statistica del nostro dataset, emergono domande di ricerca fondamentali che rappresentano la colonna vertebrale della nostra indagine. Queste domande non solo guidano l'analisi, ma ci aiutano anche a scavare in profondità nei dati, per scoprire pattern, relazioni e tendenze. Ciascuna di queste domande è stata attentamente formulata per sfruttare al massimo il potenziale informativo offerto dai nostri dataset, che spaziano da temi economici a demografici, dalla forza lavoro all'inquinamento atmosferico.

1. **Indagare la possibile correlazione tra il PIL pro capite delle città e la media di esposizione alla PM2.5.** verificando se è possibile prevedere il **PIL pro capite di una città, basandoci sul suo livello di inquinamento.** Questa domanda ci permette di esplorare se esiste un legame significativo tra il PIL pro capite e l'inquinamento di un'area geografica.
2. **Esaminare in che modo combinazioni di variabili come forza lavoro, tasso di disoccupazione e popolazione totale possono congiuntamente influenzare il PIL pro capite di una città.** Questo modello più complesso ci permette di considerare l'interazione tra diverse variabili.
3. **Osservare come il PIL si è evoluto nel tempo nelle varie città.** Questo approccio ci permette di comprendere le tendenze a lungo termine e le fluttuazioni economiche nel contesto delle città analizzate.
4. **Cercare di raggruppare le nazioni in cluster basandoci su variabili come PIL pro capite, tasso di disoccupazione, densità di popolazione e inquinamento.** Questo metodo ci aiuta a identificare schemi e somiglianze tra nazioni diverse, fornendo nuove intuizioni sulla loro classificazione socioeconomica e ambientale.
5. **Come è cambiata l'esposizione media della popolazione alle PM2.5 dal 2009 al 2019, e quali conclusioni possiamo trarre in termini di qualità dell'aria e politiche ambientali nei diversi paesi?** Questa domanda ci permette di indagare le tendenze temporali nell'esposizione della popolazione alle particelle fini PM2.5, che

sono un importante indicatore di inquinamento atmosferico e di rischi per la salute pubblica, Analizzando i cambiamenti nell'esposizione alle PM2.5 su un arco di tempo di dieci anni.

Ognuna di queste domande apre una finestra su aspetti differenti del nostro dataset, permettendoci di tessere insieme una narrazione complessa e dettagliata delle dinamiche urbane ed economiche. L'obiettivo è quello di trarre conclusioni significative e basate su dati solidi, che possano illuminare i fenomeni studiati e guidare decisioni informate.

### 3 Preprocessing del Dataset

Il processo di selezione e trasformazione del dataset originale è stato effettuato con attenzione e precisione, focalizzandosi su specifiche variabili rilevanti per le domande di ricerca (sezione 2). Inizialmente, dal dataset originale di 114 variabili, che copriva un arco temporale dal 2001 al 2022, sono state selezionate 7 variabili principali:

- **GDP (Million USD, constant prices, constant PPP, base year 2015)**: fornisce una misura complessiva dell'attività economica di una città. È fondamentale per studiare l'impatto economico su altre variabili come la demografia o il mercato del lavoro.
- **GDP pro capite (USD, constant prices, constant PPP, base year 2015)**: indica il reddito medio per persona, utile per confrontare il benessere economico tra diverse città. È cruciale per l'analisi di correlazione e regressione, in particolare quando si esaminano le relazioni tra il livello di vita e altri fattori come l'occupazione o l'inquinamento.
- **Labour force (15-64 years old)**: indica il numero di persone disponibili per il lavoro; importante per studiare la relazione tra mercato del lavoro ed economia.
- **Mean population exposure to PM2.5 air pollution**: è importante per studiare l'impatto dell'ambiente sulla salute pubblica e potenzialmente sull'economia.
- **Population density (inhabitants per km<sup>2</sup>)**: è importante per capire come la concentrazione della popolazione influenzi vari aspetti come l'economia, l'inquinamento e l'accesso ai servizi.
- **Population, all ages, administrative data**: essenziale per comprendere la base di riferimento di molte altre variabili, come il PIL pro capite o l'accesso ai servizi. È anche utile nelle analisi di regressione e correlazione.
- **Unemployment rate (unemployment 15-64 over labour force 15-64)**: fornisce un'indicazione del mercato del lavoro non utilizzato. Alti tassi di disoccupazione possono avere implicazioni economiche significative.

Nella configurazione iniziale del dataset, ogni riga corrispondeva a un'osservazione distinta relativa a una specifica area geografica per ciascun anno. Tale struttura presentava limitazioni nell'analisi delle tendenze temporali dei dati. Per superare questa restrizione, il processo di trasformazione attuato mediante lo script "script/format.R" ha comportato una

ristrutturazione significativa del formato dei dati. In questa nuova configurazione, ogni riga del dataset è stata riorganizzata per rappresentare l'intero arco temporale di interesse per ogni area geografica, disponendo i dati di ogni anno in colonne distinte. Parallelamente, si è proceduto con un'accurata selezione delle colonne da conservare. Lo script ha eliminato tutte quelle giudicate non essenziali (semplici sigle o duplicazioni) per gli obiettivi della ricerca, mantenendo solamente le colonne cruciali. Queste includevano "country", "city" (sostituito al termine originale "Geography"), "variable", e le colonne corrispondenti agli anni rilevanti per lo studio. Questa operazione di semplificazione ha ulteriormente affinato il dataset, rendendolo più diretto e funzionale agli scopi analitici prefissati.

A causa dell'elevato numero di valori mancanti (NA) in alcuni anni, in particolare agli estremi dell'intervallo temporale considerato, lo script "script/preprocessing.R" è stato impiegato per rimuovere questi anni dal dataset. Questo ha ridotto l'intervallo temporale ai dati tra il 2009 e il 2019, permettendo di concentrarsi su un periodo più ristretto ma probabilmente più significativo per analisi e studi.

### 3.1 Data imputation

Tuttavia, rimangono ancora alcuni valori NA nel dataset. Considerando che il dataset contiene serie temporali di varie variabili relative a diverse città e paesi e che queste serie temporali sono caratterizzate da misurazioni sequenziali nel tempo (rappresentate dai valori annuali delle variabili dal 2009 al 2019), abbiamo optato per l'utilizzo dell'interpolazione lineare per vari motivi:

1. Natura Progressiva e Continua dei Dati Temporali:
  - Le serie temporali nel nostro dataset presumibilmente mostrano cambiamenti graduali e progressivi da un anno all'altro.
  - L'interpolazione lineare è particolarmente adatta per dati che cambiano in modo relativamente uniforme e prevedibile nel tempo.
2. Semplicità ed Efficacia:
  - L'interpolazione lineare è un metodo semplice ma efficace per stimare i valori mancanti in serie temporali.
  - Questo metodo assume che il cambiamento tra due punti temporali sia lineare, il che è una supposizione ragionevole per molte variabili su brevi periodi di tempo, come da un anno all'altro.
3. Preservazione delle Tendenze Temporali:
  - L'interpolazione lineare aiuta a mantenere l'integrità delle tendenze temporali all'interno dei dati, collegando direttamente i punti dati adiacenti con una linea retta.
  - Questo aiuta a garantire che l'imputazione rispetti la continuità temporale e la sequenza dei dati, mantenendo le tendenze generali e i modelli osservati nei dati completi.
4. Minimizzazione delle Distorsioni:

- A differenza di altri metodi, come l'imputazione con la media o la mediana, l'interpolazione lineare evita di appiattire le tendenze temporali o introdurre potenziali distorsioni.
- Questo metodo evita di sovrascrivere le variazioni specifiche di ciascuna serie temporale con una misura centrale che potrebbe non riflettere accuratamente le variazioni annuali.

##### 5. Adattabilità a Serie Temporali con Dati Mancanti Sporadici:

- Nel nostro caso, i dati mancanti non erano ampiamente distribuiti ma piuttosto concentrati in specifici anni.
- L'interpolazione lineare è particolarmente utile in situazioni dove i dati mancanti sono sporadici o isolati, poiché sfrutta i dati esistenti più prossimi per creare una stima accurata.

In sintesi, l'interpolazione lineare è stata scelta per il suo equilibrio tra semplicità, efficienza e capacità di mantenere le tendenze e le caratteristiche intrinseche delle serie temporali. Questa scelta è particolarmente appropriata quando si lavora con variabili economiche e ambientali, dove le variazioni annuali sono spesso graduali e sequenziali. Tuttavia, è importante notare che l'interpolazione lineare funziona meglio quando i dati mancanti non sono eccessivi e quando i dati esistenti offrono un contesto affidabile per l'imputazione.

Per implementare questo approccio, abbiamo sviluppato uno script denominato "script/imputazione\_interpolazione\_lineare.R", che gestisce la pulizia dei dati, la conversione del formato dei dati, l'interpolazione dei valori mancanti e il salvataggio del dataset aggiornato in un nuovo file CSV chiamato "Clean\_dataset\_interpolated".

Tuttavia, dopo l'aggiornamento del dataset attraverso lo script, sono rimaste 31 righe su cui non è stato possibile effettuare l'interpolazione lineare. Queste righe avevano 3 o 4 valori NA, che si concentravano all'inizio o alla fine della serie temporale, e quindi una semplice interpolazione lineare non avrebbe fornito una giusta imputazione. Per affrontare questa situazione, abbiamo sviluppato uno script aggiuntivo denominato "script/perc\_variazione\_per\_imputazione.R". Questo script è stato progettato per analizzare il dataset e calcolare le variazioni percentuali medie annuali delle variabili per ciascuna città, prendendo in considerazione solo le righe con valori mancanti (NA). Esso comprende una serie di passaggi, tra cui il filtraggio dei dati, il calcolo delle variazioni percentuali e la creazione di un nuovo file CSV chiamato "percentuale\_media\_per\_citta\_variabili\_con\_NA" per conservare i risultati delle analisi, grazie a questi valori abbiamo imputato i restanti valori NA utilizzando la percentuale media di variazione.

Dopo l'elaborazione e l'imputazione dei dati, abbiamo ottenuto un dataset pulito denominato "Clean\_dataset.csv", che conta 5106 righe, eliminando il precedente. Questo dataset rappresenta la base per le nostre successive analisi statistiche. Inoltre, abbiamo suddiviso nuovamente il dataset principale in sette sotto-dataset tramite lo script "script/split\_dataset.R", ognuno corrispondente a una delle sette variabili di interesse. Questa suddivisione ci consentirà di condurre analisi statistiche più dettagliate e fornire una panoramica completa delle dinamiche urbane in studio.

## 4 Analisi esplorativa

Una volta definito il nostro dataset, è arrivato il momento di visualizzare i nostri dati. Per avere una panoramica generale e visualizzare la loro distribuzione, il primo grafico che abbiamo realizzato, per tutte le 7 variabili, è stato un grafico a bolle su una cartografia del mondo. Il grafico è stato realizzato tramite lo script "script/plot\_leaf.R". Le coordinate delle varie aree geografiche sono state reperite tramite lo script "script/coordinate.R", che utilizza una funzione per accedere ai servizi di Google Maps. Dalle immagini mostrate nelle sottosezioni successive, possiamo notare come i dati siano distribuiti principalmente in Nord e Centro America, in Europa Centrale ed Europa dell'Est, Giappone, Australia e Nuova Zelanda. Notiamo l'assenza totale di Africa, buona parte dell'Asia e del Sud America.

Successivamente, per comprendere la distribuzione dei dati del nostro dataset e identificare eventuali outlier abbiamo deciso di utilizzare lo script "script/boxplot\_tutti.R" che ci ha permesso di generare i boxplot al fine di effettuare una rappresentazione della distribuzione mostrando chiaramente la mediana, il primo quartile, il terzo quartile e i possibili outlier, consentendo di avere una visione rapida della distribuzione dei dati. Ciò ci ha permesso di individuare punti che si discostano significativamente dalla distribuzione principale, aiutandoci a rilevare potenziali errori o valori anomali. Infine, ci ha fornito informazioni sulla variabilità dei dati, inclusi i range interquartili (IQR), che possono essere utili per valutare la dispersione dei dati. Abbiamo prima deciso di analizzare visivamente i vari boxplot generati e in seguito per un'analisi più dettagliata abbiamo utilizzato lo script denominato "script/search\_outliers.R", che ci ha permesso di identificare con precisione i vari outlier per ogni Paese.

### 4.1 GDP

Nei grafici nella figura 1, possiamo notare come il paese, che presenta il PIL maggiore e costante nel tempo sia il Giappone, principalmente con la grande metropoli di Tokyo. Tuttavia, il maggior incremento del PIL è avvenuto negli Stati Uniti d'America, nella città di New York. Questa crescita è attribuibile al suo status di centro finanziario e commerciale globale, beneficiando della ripresa e della crescita seguite alla crisi finanziaria del 2008. Al contrario, il calo del PIL di Atene riflette la grave crisi economica della Grecia durante questo periodo, caratterizzata da recessione, misure di austerità e salvataggi finanziari.

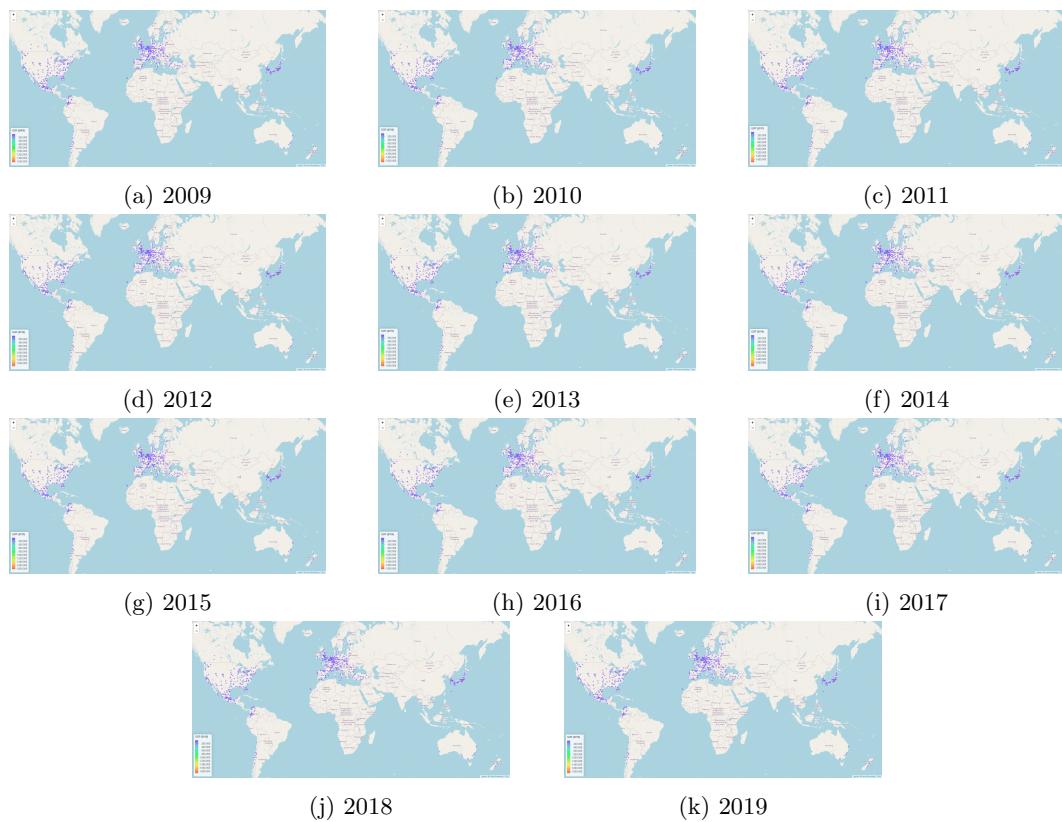


Figure 1: GDP rilevato tra il 2009-2019

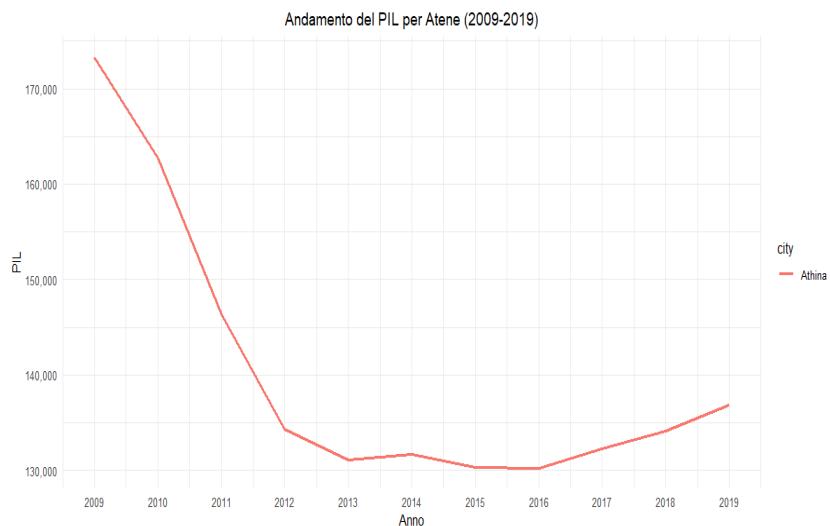


Figure 2: Andamento del PIL per atene

Nella figura 2, ottenuta tramite lo script "script/gdp\_grafico\_linee.R", abbiamo voluto evidenziare con un grafico a linee la città di Atene che ha avuto un decremento notevole del PIL.

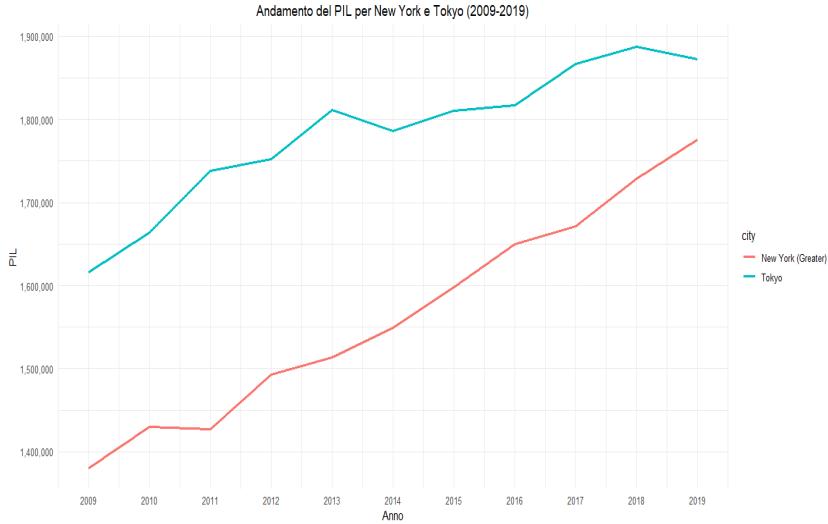


Figure 3: Confronto tra la città di New York e Tokyo

Nella figura 3, ottenuta tramite lo script ”script/gdp\_grafico\_linee.R” abbiamo voluto mettere in risalto con un grafico a linee la città di Tokyo con il PIL maggiore e costante nel tempo e la città di New York che presenta il maggior incremento del PIL.

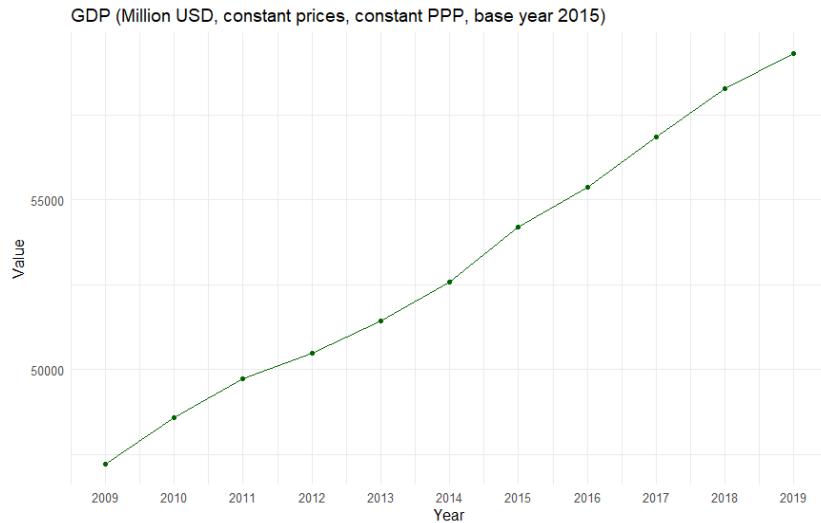


Figure 4: Evoluzione media del PIL annuale dei Paesi OECD

Nella figura 4 possiamo verificare l’andamento del PIL medio dei Paesi OECD, ottenuto con lo script ”script/plot\_global\_mean”, che negli anni ha continuato ad aumentare costantemente.

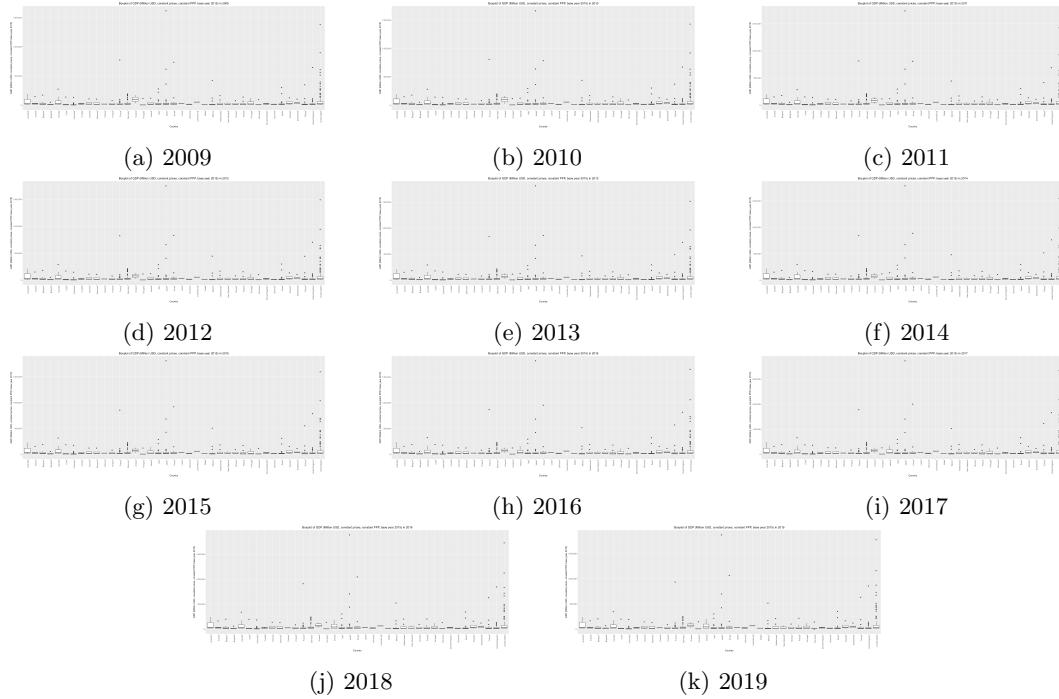


Figure 5: Boxplot GDP rilevato tra il 2009-2019

Per quanto riguarda l'analisi dei boxplot relativi alla variabile Prodotto Interno Lordo (PIL), è emerso un pattern che ha richiamato la nostra attenzione. È stato osservato che tali boxplot presentano una notevole presenza di outlier. Questo fenomeno, come ci si poteva aspettare, può essere attribuito all'ineguaglianza economica presente nei dati. In molte situazioni, si verifica una notevole diseguaglianza economica all'interno dei paesi, con alcune regioni o gruppi di popolazione che contribuiscono in modo significativo al PIL, mentre altri hanno un impatto molto minore.

La presenza di queste diseguaglianze economiche all'interno dei paesi può generare la comparsa di outlier nei dati relativi al PIL. Questo si verifica perché alcuni paesi registrano un PIL notevolmente superiore alla media, a causa delle regioni o dei settori economici particolarmente prosperi, mentre altri paesi registrano un PIL sensibilmente al di sotto della media, a causa di regioni o settori economici meno sviluppati.

In sostanza, l'analisi dei boxplot relativi al PIL ha permesso di mettere in luce questo fenomeno di diseguaglianza economica, che contribuisce alla presenza di outlier nei dati. Tale informazione è essenziale per comprendere meglio la variazione e la distribuzione dei dati relativi al PIL nei diversi paesi e negli anni considerati.

## 4.2 GDP pro capite

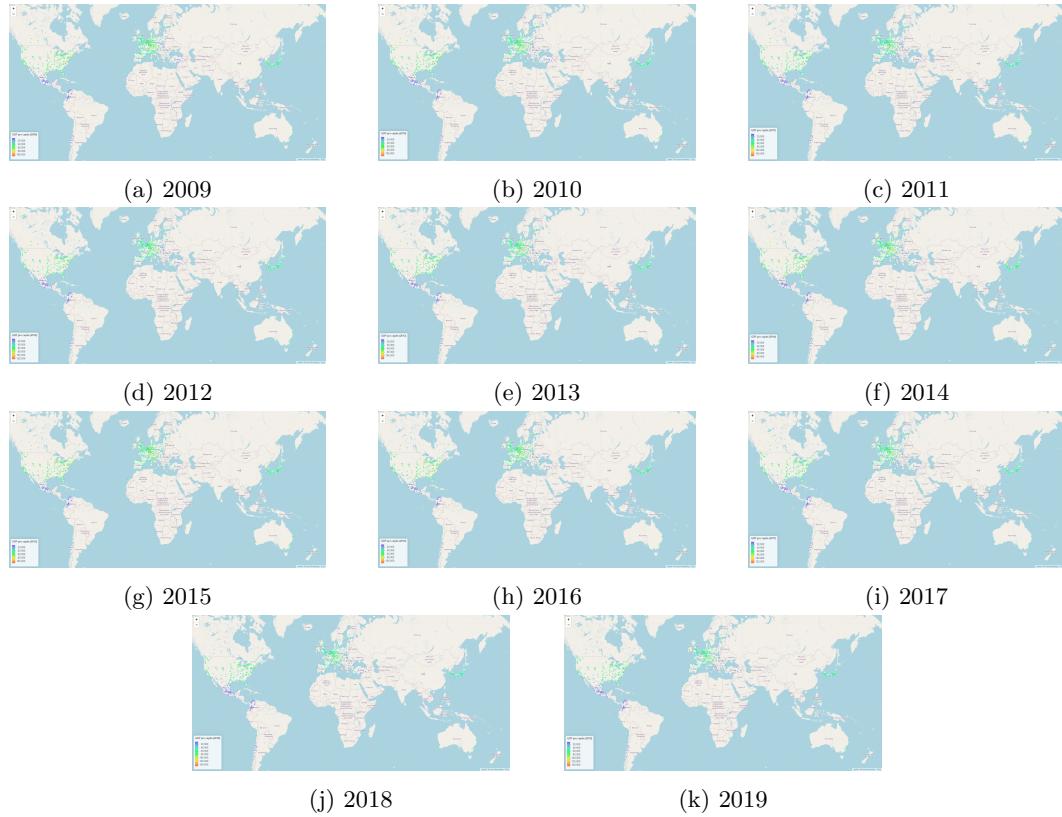


Figure 6: GDP per capite rilevato tra il 2009-2019

Possiamo osservare come ci siano cambiamenti significativi nel primato del PIL pro capite quasi ogni anno, con diverse città che si alternano in questo ruolo nel corso del tempo. La città di Cork, in Irlanda, presenta il maggior aumento di PIL pro capite. Questo aumento potrebbe riflettere la crescita economica dell'Irlanda, spesso denominata "Celtic Tiger", dovuta a investimenti stranieri e una forte economia del settore tecnologico. D'altra parte, il Messico mostra segnali di difficoltà economica, in particolare con l'area urbana di Carmen, che potrebbe aver risentito delle conseguenze legate al calo dei prezzi del petrolio o alla riduzione della produzione petrolifera, settori chiave per l'economia locale.

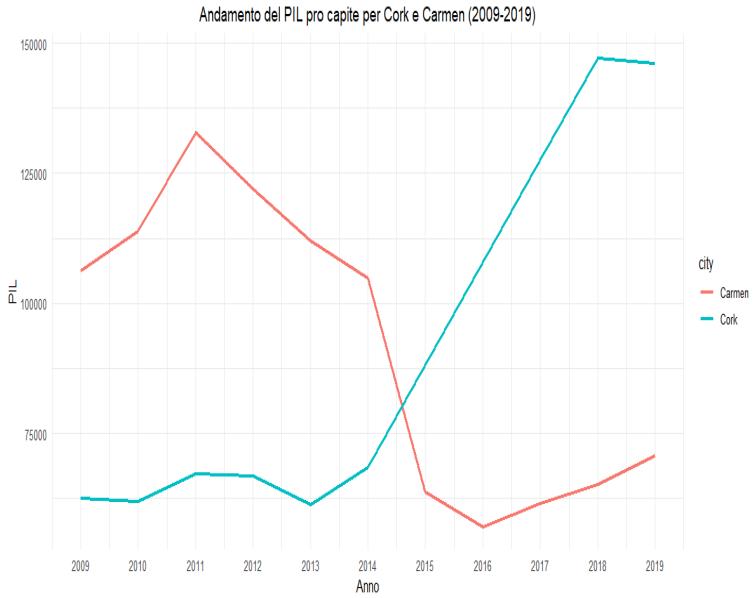


Figure 7: Confronto tra i paesi col PIL pro capite migliore e peggiore

Nella figura 7, ottenuta tramite lo script "script/gdp\_per\_capita\_grafico\_linee.R" abbiamo voluto evidenziare con un grafico a linee la città di Carmen, che è quella che è risultata con un decremento notevole del PIL pro capite, paragonandola alla città di Cork che ha presentato un incremento notevole del PIL. In particolare, la città di Cork ha presentato un incremento di 83551 USD.

Nella figura 8 possiamo verificare l'andamento del PIL pro capite medio dei Paesi OECD, che negli anni ha continuato ad aumentare costantemente.

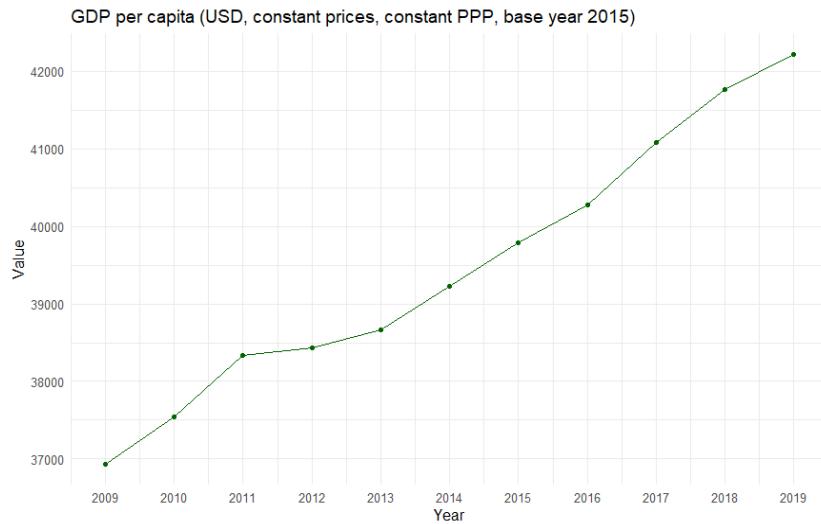


Figure 8: Evoluzione media del PIL pro capite annuale dei Paesi OECD

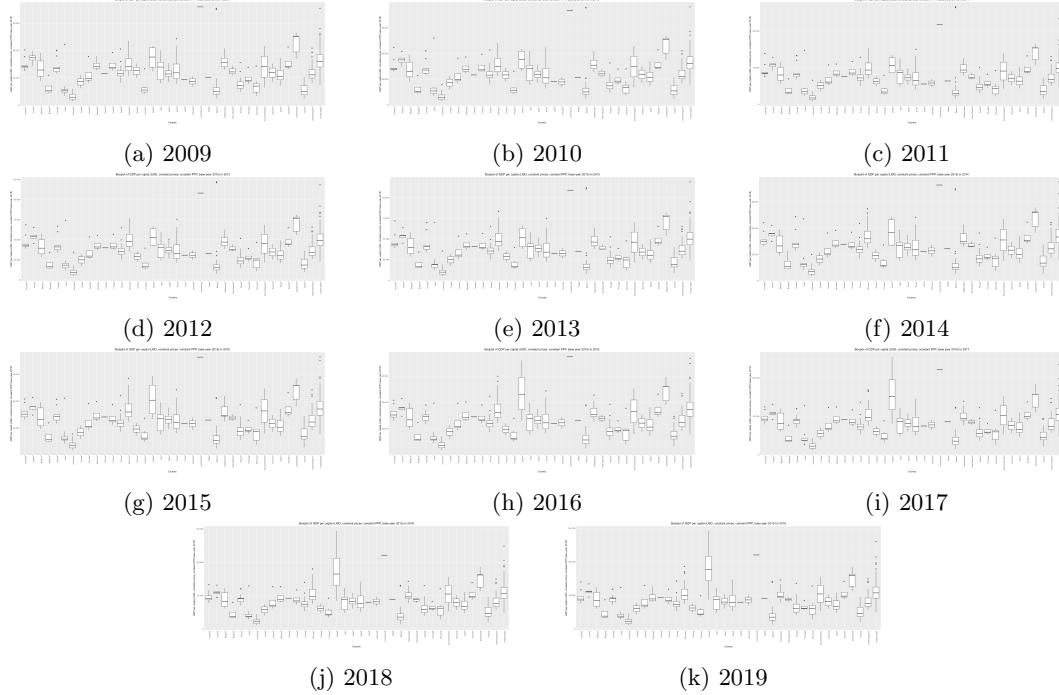


Figure 9: Boxplot GDP pro capite rilevato tra il 2009-2019

Analizzando i boxplot per la variabile inherente al PIL pro capite abbiamo notato diverse informazioni utili.

1. Maggiore Variabilità (IQR): L'Irlanda mostrava la maggiore variabilità nel GDP pro capite nel 2019, con un IQR di 36,236 USD. Questo suggerisce una significativa differenza nei livelli di reddito all'interno del paese.
2. Valori di GDP pro capite più Altì (Media):
  - Lussemburgo: Ha mantenuto una media di GDP pro capite più elevata, con circa 109,592 USD.
  - Svizzera: Segue con una media di circa 71,596 USD.
  - Irlanda: Con una media di circa 64,900 USD.
3. Valori di GDP pro capite più Bassi (Media):
  - Colombia: Ha avuto la media più bassa di GDP pro capite, con circa 11,103 USD.
  - Messico: Segue con una media di circa 19,133 USD.
  - Bulgaria: Con una media di circa 21,321 USD.
4. Maggiore Variabilità nel GDP pro capite:
  - Messico: Ha mostrato la maggiore variazione totale (differenza tra il valore più alto e più basso) nel GDP pro capite, con una variazione di 126,388 USD.
  - Irlanda: Con una variazione di 110,012 USD.

- Stati Uniti: Con una variazione di 107,800 USD.

Analizzando gli outlier nei boxplot relativi al tasso di disoccupazione tramite lo script "search\_outliers" abbiamo prestato particolare attenzione al Messico e agli Stati Uniti.

Città	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
<b>Campeche</b>	105050	113083	131788	120964	110787	103770	63102	56259	60670	64322	69694
<b>Carmen</b>	106199	113911	132882	122084	111916	104923	63859	56984	61530	65259	70765
<b>Monterrey</b>	-	30949	30575	30264	30036	30928	-	-	-	-	-
<b>Villahermosa</b>	-	31339	36243	35460	32882	33273	-	-	-	-	-

Table 1: Valore degli outlier del Messico

Città	2009	2010	2011	2012	2013	2014	2015	2016
<b>Boston</b>	78644	81405	81718	82903	82610	83295	87060	88688
<b>Durham</b>	105957	114351	116598	118366	119109	112917	109558	104230
<b>New Haven</b>	76815	-	-	-	-	-	-	-
<b>San Francisco</b>	84160	85853	87578	91936	95218	99355	105272	-
<b>Peoria</b>	-	-	-	84119	-	-	-	-
<b>Seattle</b>	-	-	-	79247	81167	83046	85700	-

Table 2: Valore degli outlier degli Stati Uniti d'America

### 4.3 Labour force

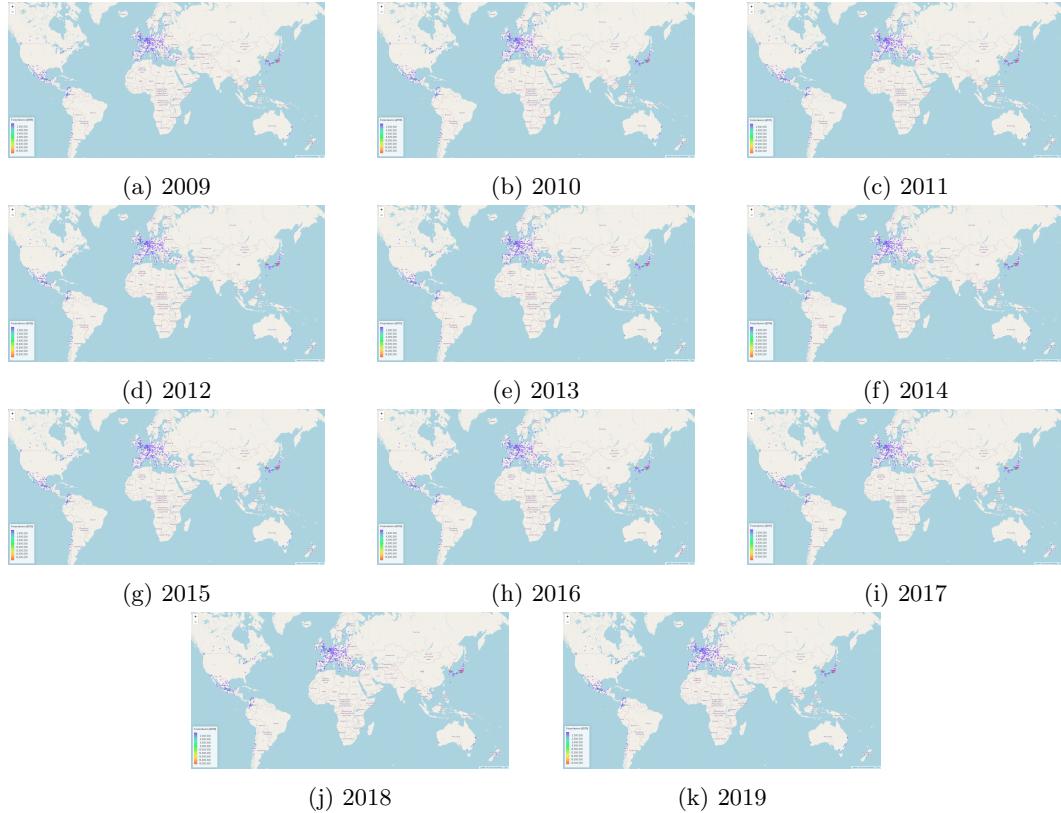
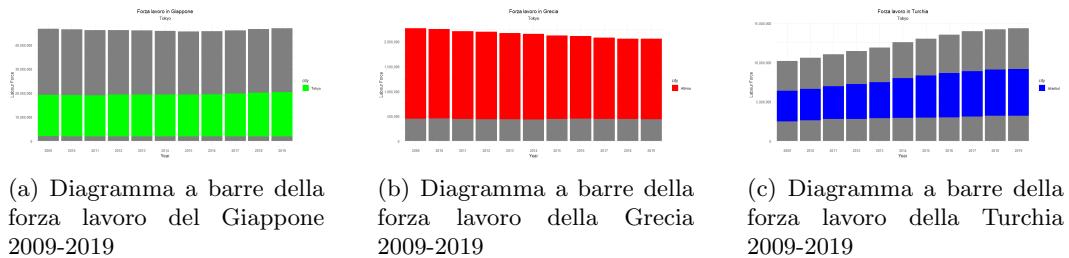


Figure 10: Labour force rilevata tra il 2009-2019

Per la città di Tokyo, in Giappone, si osserva il mantenimento di un'elevata forza lavoro durante tutti gli anni considerati dal 2009 al 2019. Questa stabilità può essere attribuita alla sua posizione come capitale economica del Giappone e uno dei principali centri finanziari del mondo, riflettendo l'importanza delle grandi metropoli nel panorama economico globale.

La città di Istanbul, in Turchia, ha registrato il maggiore incremento della forza lavoro nello stesso periodo. Questo notevole aumento può essere attribuito alla sua posizione strategica come ponte tra Europa e Asia, alla rapida urbanizzazione e alla crescita economica che hanno attirato sia investimenti interni che esteri, stimolando così l'espansione del mercato del lavoro.

Infine, la città di Atene, in Grecia, è quella dove si è notato il calo più marcato della forza lavoro. Ciò riflette la crisi economica della Grecia, con una prolungata recessione e misure di austerità che hanno impattato il mercato del lavoro.



La figura 11a, ottenuta tramite lo script "script/labour\_force\_barplot.R", sottolinea con una diagramma a barre il mantenimento di un'elevata forza lavoro per il Giappone, in particolare per la città di Tokyo.

La figura 11b, ottenuta tramite lo script "script/labour\_force\_barplot.R", sottolinea con una diagramma a barre la notevole riduzione della forza lavoro per la Grecia, in particolare per la città di Atene.

La figura 11c, ottenuta tramite lo script "script/labour\_force\_barplot.R", sottolinea con una diagramma a barre il notevole incremento della forza lavoro per la Turchia, in particolare per la città di Instanbul.

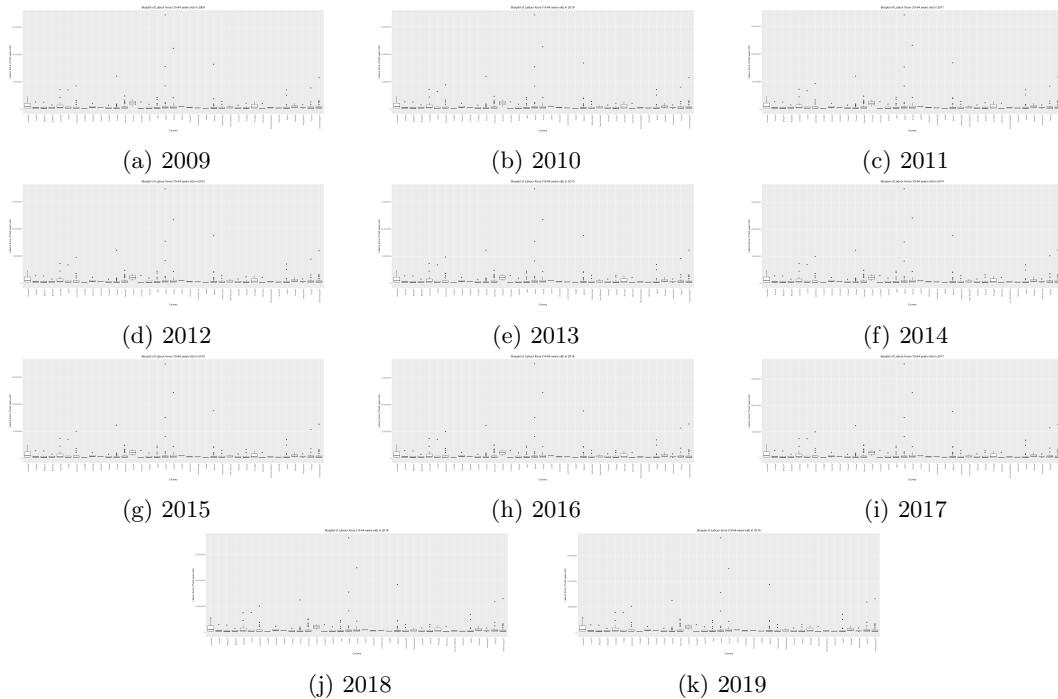


Figure 12: Boxplot labour force rilevata tra il 2009-2019

Per quanto riguarda l'analisi dei boxplot relativi alla forza lavoro abbiamo prestato particolare attenzione ai country Giappone, Corea, Messico e Francia che hanno presentato un grande numero di outlier e che hanno catturato la nostra attenzione per i loro valori estremamente elevati. Questi sotto sono per l'appunto i risultati ottenuti tramite lo script "script/search\_outliers.R".

Nel contesto dell'analisi dei boxplot relativi alla variabile "Labour force (15-64 years old)", abbiamo concentrato la nostra attenzione su alcune nazioni specifiche, tra cui il Giappone, la Corea, il Messico e la Francia. Questi paesi hanno dimostrato di essere di particolare rilevanza nel nostro studio a causa del notevole numero di outlier identificati e dei valori estremamente elevati che hanno suscitato grande interesse nella nostra analisi. Qui di seguito, presentiamo i risultati ottenuti attraverso l'esecuzione dello script "script/search\_outliers.R".

Città	2009	2010	2011	2012	2013
Bordeaux	535279	533032	532014	535406	544522
Lille	620414	624603	628670	629993	627470
Lyon	932880	948971	945769	963047	967601
Marseille	763586	764977	775810	780934	793708
Paris	6034636	5983589	5989576	6036326	6100996
Toulouse	566160	577008	580128	577512	583944

Table 3: Valore degli outlier della Francia

Città	2009	2010	2011	2012	2013	2014
Fukuoka	1242680	1250200	1246440	1235160	1225760	1218240
Nagoya	4252490	4223640	4200560	4137090	4119780	4096700
Osaka	7817220	7710360	7644600	7718580	7718580	7628160
Sapporo	1036320	1036320	1024080	1007760	991440	975120
Tokyo	17238610	17186560	17087580	17354740	17382770	17438190

Table 4: Valore degli outlier del Giappone

Città	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Gimhae	2097137	2125146	2160832	2160286	2184810	2188282	2179394	2188008	2142979	2139807
Seoul	11370483	11578766	11670342	11745891	12060234	12160982	12261463	12424317	12442676	12464322

Table 5: Valore degli outlier della Corea

#### 4.4 Mean population exposure to PM2.5 air pollution

L’evoluzione del primato dell’esposizione alla PM2.5 dal 2009 al 2019 ha visto inizialmente la Turchia, poi la Polonia e, infine, per un bel po’ di tempo il Cile. In questo caso l’Australia mostra un aumento dell’esposizione alla PM2.5, che potrebbe essere dovuto all’incremento del traffico veicolare, al riscaldamento domestico e a incendi boschivi, che hanno colpito frequentemente il paese negli ultimi anni. Mentre l’Italia, con la città di Torino, grazie agli sforzi di miglioramento della qualità dell’aria attraverso regolamenti ambientali più rigorosi e iniziative per ridurre l’inquinamento da veicoli e industrie, mostra la maggiore diminuzione di esposizione alle PM2.5.

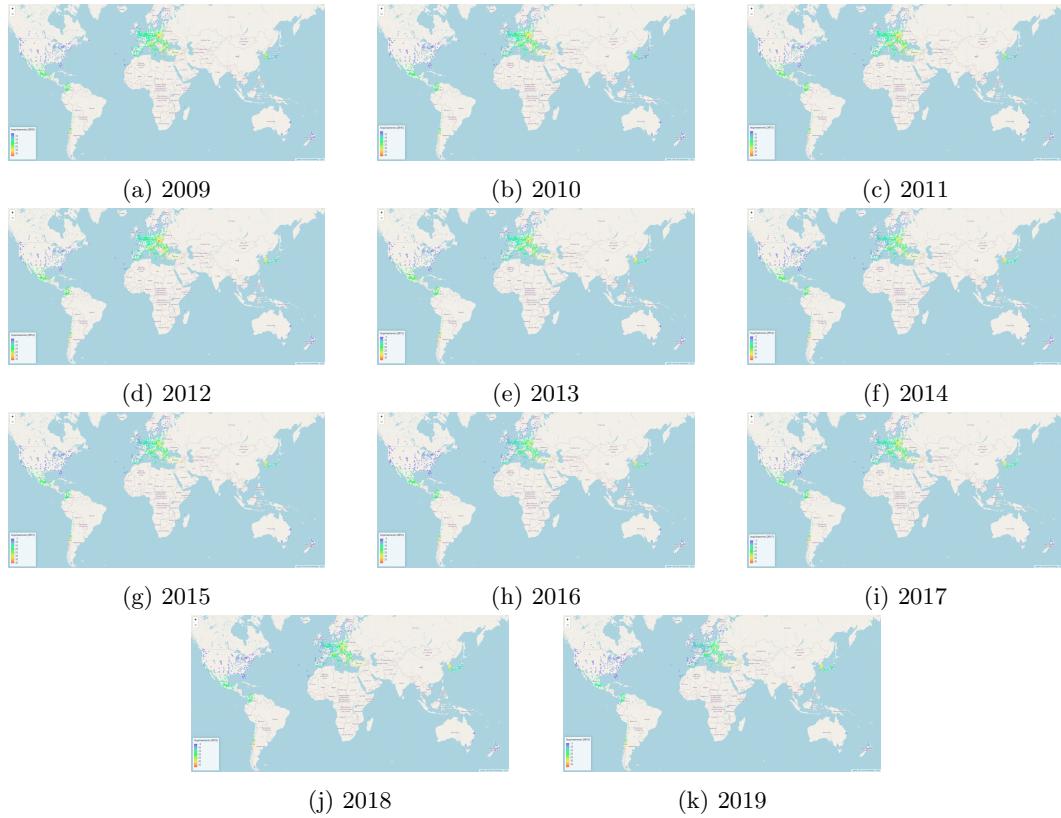
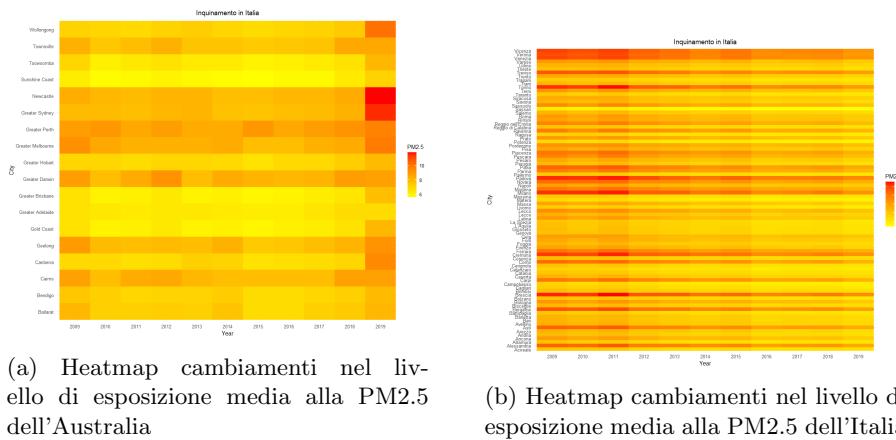


Figure 13: Mean population exposure to PM2.5 air pollution rilevata tra il 2009-2019



La figura 14a e la figura 14b, ottenute tramite lo script ”script/ mappa\_calore\_mean\_population.R”, mostrano con una mappa di calore l’aumento dell’esposizione alla PM 2.5 dell’Australia e la notevole riduzione generale dell’esposizione alla PM 2.5 dell’Italia, anche se rimane abbastanza alta in alcune zone.

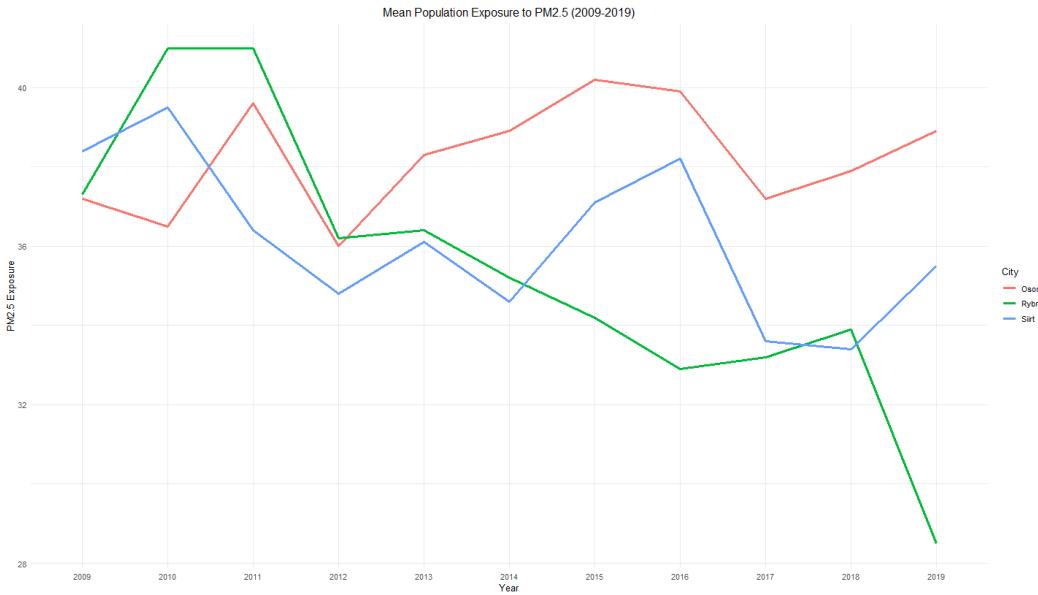


Figure 15: Le peggiori città tra il 2009-2019

La figura 15, ottenuta tramite lo script ”script/worse\_city\_mean\_population.R”, evidenzia tramite un grafico a linee il primato dell’esposizione alla PM 2.5 che ha visto alternarsi le città di Siirt, Rybnik e Osorno, rispettivamente della Turchia, Polonia e Cile.

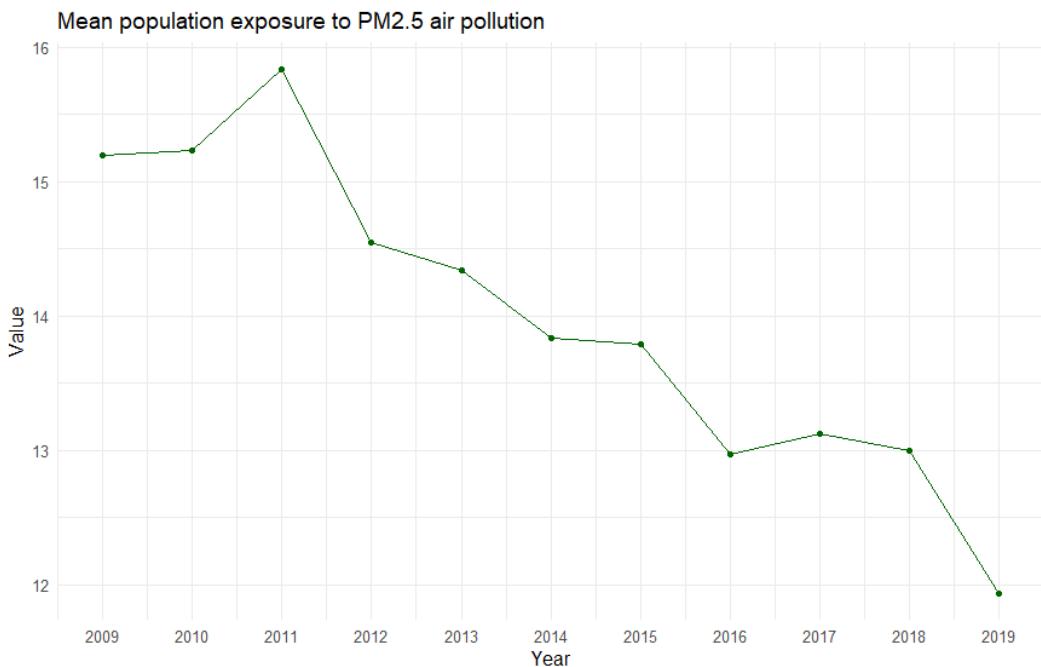


Figure 16: Evoluzione media dell’esposizione media alla PM2.5 dei Paesi OECD

Nella figura 16 possiamo verificare l’andamento dell’esposizione media alla PM2.5 dei Paesi OECD, che negli anni ha continuato a diminuire, tranne nel 2011 e nel 2017.

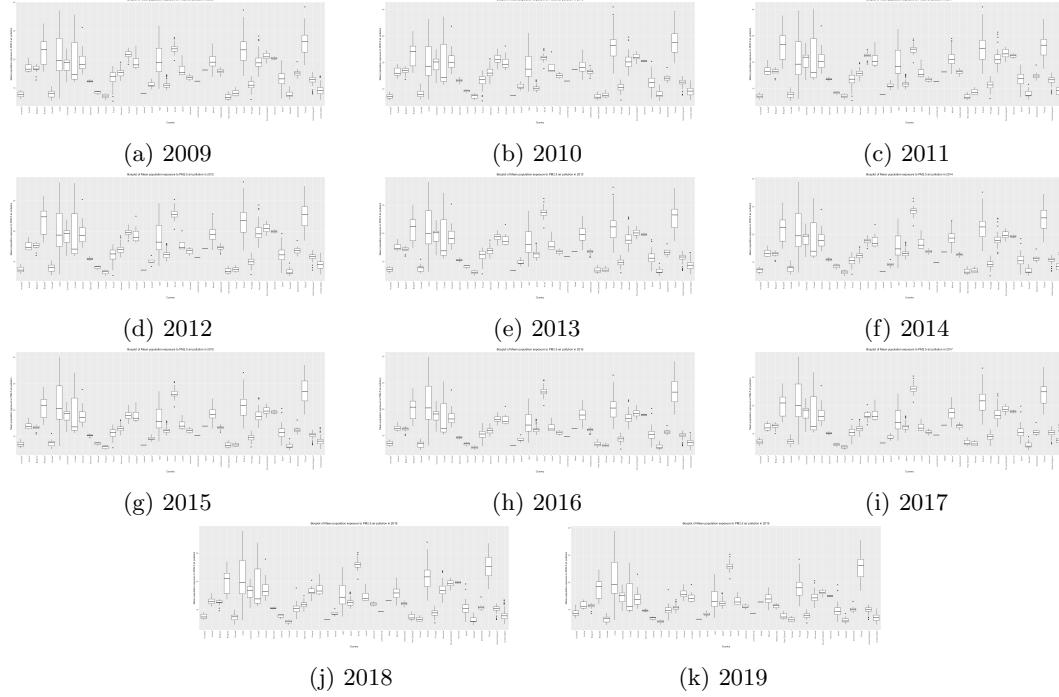


Figure 17: Boxplot Mean exposure to PM2.5 rilevata tra il 2009-2019

Analizzando i boxplot per la variabile inerente all'inquinamento possiamo notare:

1. Turchia ha la mediana più alta di esposizione alla PM2.5, con un valore di 26.05. Questo significa che la Turchia ha mostrato livelli relativamente elevati di inquinamento da PM2.5 durante il periodo considerato. La media di esposizione è di 26.02, e la deviazione standard è di 4.77, indicando una certa variabilità ma non estrema.
2. Finlandia presenta la mediana più bassa di esposizione alla PM2.5, con un valore di 6.20. Ciò suggerisce che la Finlandia ha avuto livelli generalmente bassi di inquinamento da PM2.5. La media è di 6.22, con una deviazione standard relativamente bassa di 1.03, indicando una certa consistenza nei livelli di esposizione.
3. Croazia mostra la maggiore variabilità nell'esposizione alla PM2.5, con una deviazione standard di 8.56. Questo indica che i livelli di inquinamento in Croazia sono stati piuttosto variabili durante gli anni, con una media di esposizione di 18.92 e una mediana di 14.10.
4. Islanda, al contrario, ha la minore variabilità, con una deviazione standard di solo 0.58. Questo implica che l'Islanda ha avuto livelli di inquinamento da PM2.5 piuttosto stabili e consistenti nel tempo, con una media di 6.99 e una mediana di 6.80.

Analizzando gli outlier nei boxplot relativi all'inquinamento tramite lo script "script/search\_outliers.R" abbiamo prestato particolare attenzione al Giappone e al Regno Unito.

Città	2009	2010	2011	2012	2013	2014
<b>Fukuoka</b>	14.3	13.2	15.4	15.9	16.3	17.1
<b>Kagoshima</b>	14.5	13.4	14.8	15.6	16.4	17.2
<b>Kumamoto</b>	13.9	12.8	14.7	-	-	16.5
<b>Kurume</b>	13.3	-	14.3	-	-	15.9
<b>Nagasaki</b>	14.6	13.3	15.2	15.8	16.5	17.1
<b>Omuta</b>	14	12.8	14.9	15.4	-	16.6

Table 6: Valori outlier del Giappone

Città	2009	2010	2011	2012	2013	2014	2015
<b>Aberdeen</b>	7.9	7.7	8	7.2	7	6.8	6.8
<b>Dundee City</b>	8	8	8	7.2	6.8	7.0	-
<b>Edinburgh</b>	8.7	9	8.8	8	7.5	7.7	-
<b>Falkirk</b>	8.5	9	8.6	8	7.3	7.8	-
<b>Glasgow</b>	9.3	-	9.4	8.8	8.1	8.5	-

Table 7: Valori outlier del Regno Unito

## 4.5 Population density

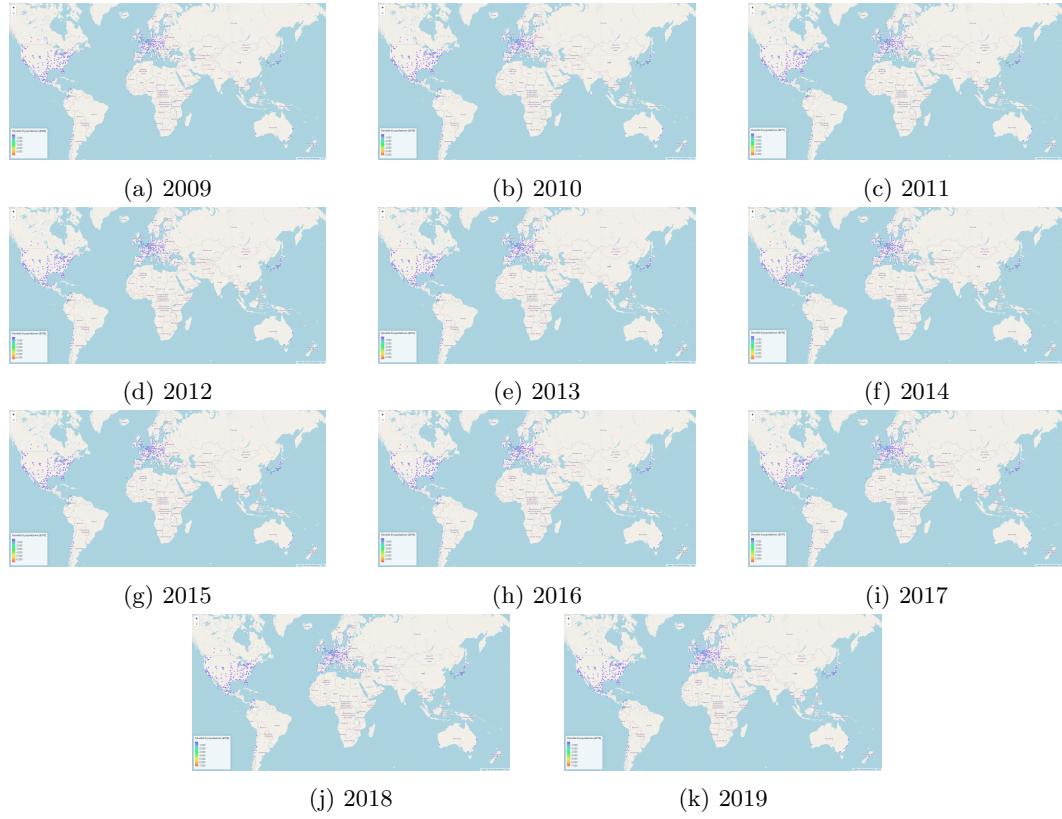


Figure 18: Population density rilevata tra il 2009-2019

Istanbul, in Turchia, ha mantenuto il primato per la densità della popolazione più alta per tutti gli anni dal 2009 al 2019, con una crescita costante. Ed è anche la città col maggiore incremento della densità di popolazione, probabilmente dovuto alla rapida urbanizzazione, alla crescita economica e all'attrattiva della città come centro culturale e commerciale. Queste variazioni potrebbero riflettere l'urbanizzazione e la migrazione interna verso le grandi città per opportunità di lavoro, con conseguente calo nelle aree rurali. Infine, la Grecia, con la città di Atene, presenta la maggiore dimunizione di densità di popolazione, confermando ciò che è stato accennato precedentemente, ovvero un'emigrazione significativa a causa della recessione e della disoccupazione.

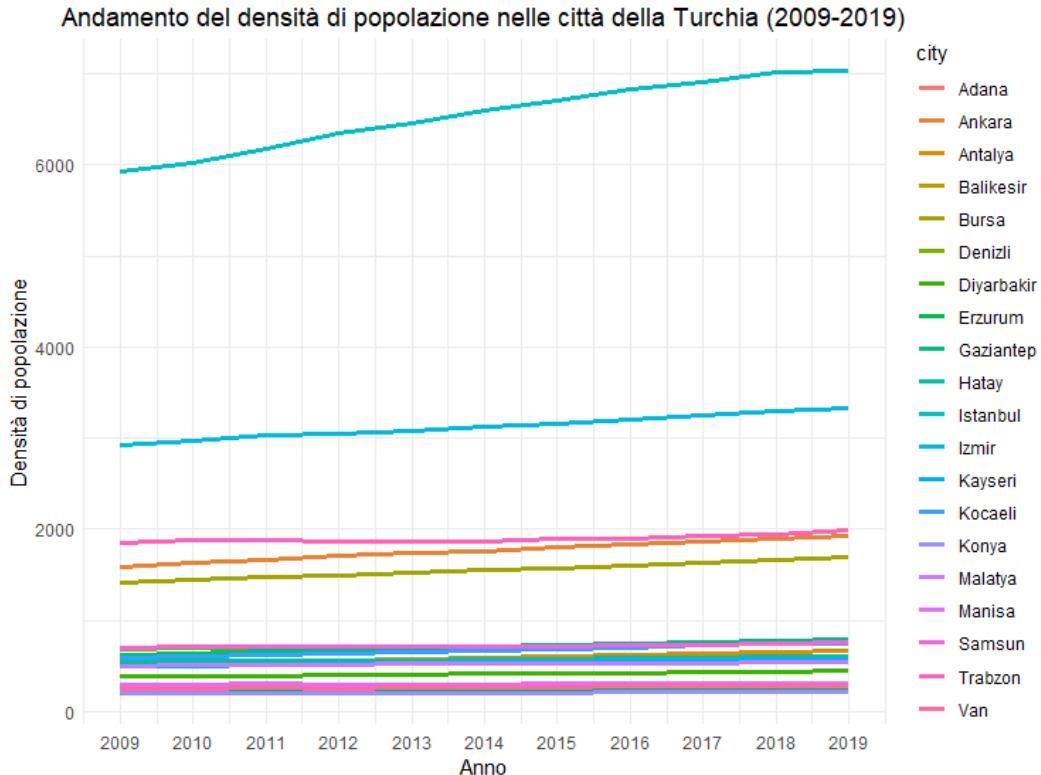


Figure 19: Confronto tra la densità di popolazione delle città turche

Tramite il confronto nella figura 19, possiamo notare come siano solo due le città turche ad essere state interessate da un aumento della densità di popolazione, mentre tutte le altre sono rimaste costanti nel tempo. Possiamo quindi concludere che non è avvenuto uno spostamento dalle aree rurali verso le grandi città, ma piuttosto ci sono stati flussi migratori da altri paesi verso la capitale turca o un alto tasso di natalità.

Analizzando i boxplot, in figura 20, per la variabile inerente alla densità di popolazione possiamo notare diverse informazioni utili.

1. Paesi con la Densità di Popolazione più Alta:

- Malta: Con una media di circa 1638.73 abitanti per km<sup>2</sup>
- Corea: Segue con una media di 1149.51 abitanti per km<sup>2</sup>
- Grecia: Con una media di 1142.95 abitanti per km<sup>2</sup>
- Türkiye (Turchia): Ha una media di 1105.33 abitanti per km<sup>2</sup>
- Romania: Con una media di 862.78 abitanti per km<sup>2</sup>

2. Paesi con la Densità di Popolazione più Bassa:

- Stati Uniti: Con una media di 161.58 abitanti per km<sup>2</sup>
- Estonia: Con una media di 131.55 abitanti per km<sup>2</sup>
- Norvegia: Ha una media di 120.18 abitanti per km<sup>2</sup>
- Finlandia: Con una media di 118.86 abitanti per km<sup>2</sup>

- Irlanda: La media è di 106.32 abitanti per km<sup>2</sup>

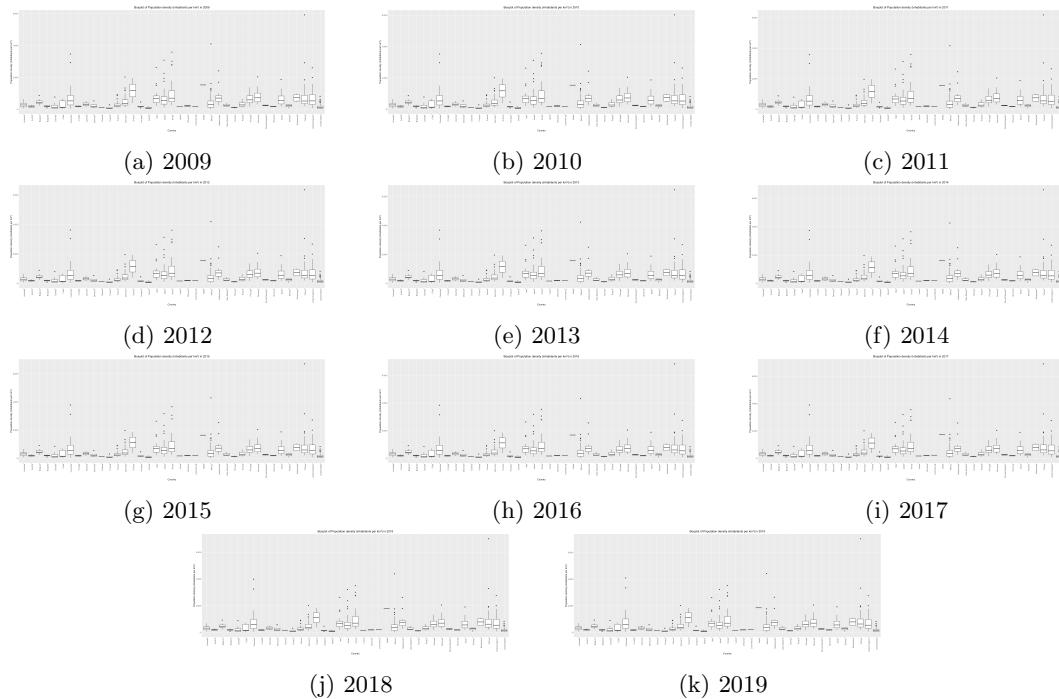


Figure 20: Boxplot Population density rilevata tra il 2009-2019

Analizzando gli outlier nei boxplot relativi alla densità di popolazione tramite lo script "script/search\_outliers.R" abbiamo prestato particolare attenzione per quanto riguarda Turchia, Messico e Corea.

Città	2009	2010	2011	2012	2013	2014	2015	2016
<b>Istanbul</b>	5923	6025	6183	6355	6463	6605	6706	6837
<b>Izmir</b>	2919	2975	3037	3049	3080	3123	3163	3206
<b>Trabzon</b>	1847	1887	1883	1867	1869	1870	1891	1895
<b>Ankara</b>	-	-	-	1708	1734	1761	1798	1840

Table 8: Valori outlier della Turchia

Città	2009	2010	2011	2012	2013	2014
Guadalajara	1307	1328	1350	1373	1395	1418
Leon	1144	1170	1193	1216	1240	1263
Mexico City	4102	4134	4165	4197	4229	4261
Tijuana	1235	1259	1288	1317	1346	1376
Tuxtla Gutierrez	1614	1643	1658	1674	1689	1704

Table 9: Valori outlier del Messico

Città	2009	2010	2011	2012	2013	2014	2015	2016
Dalseong	2799.9	2821.0	2818.0	2817.0	2815.0	2809.0	2801.0	2796.0
Seo	2775.3	2751.0	2769.0	2790.0	2806.0	2814.0	2801.0	2784.0
Seoul	3584.5	3573.0	3594.0	3613.0	3631.0	3646.0	3658.0	-
Ulsan	3014.0	2963.0	2971.0	2995.0	3023.0	3045.0	3055.0	-

Table 10: Valori outlier Corea

#### 4.6 Population, all ages, administrative data

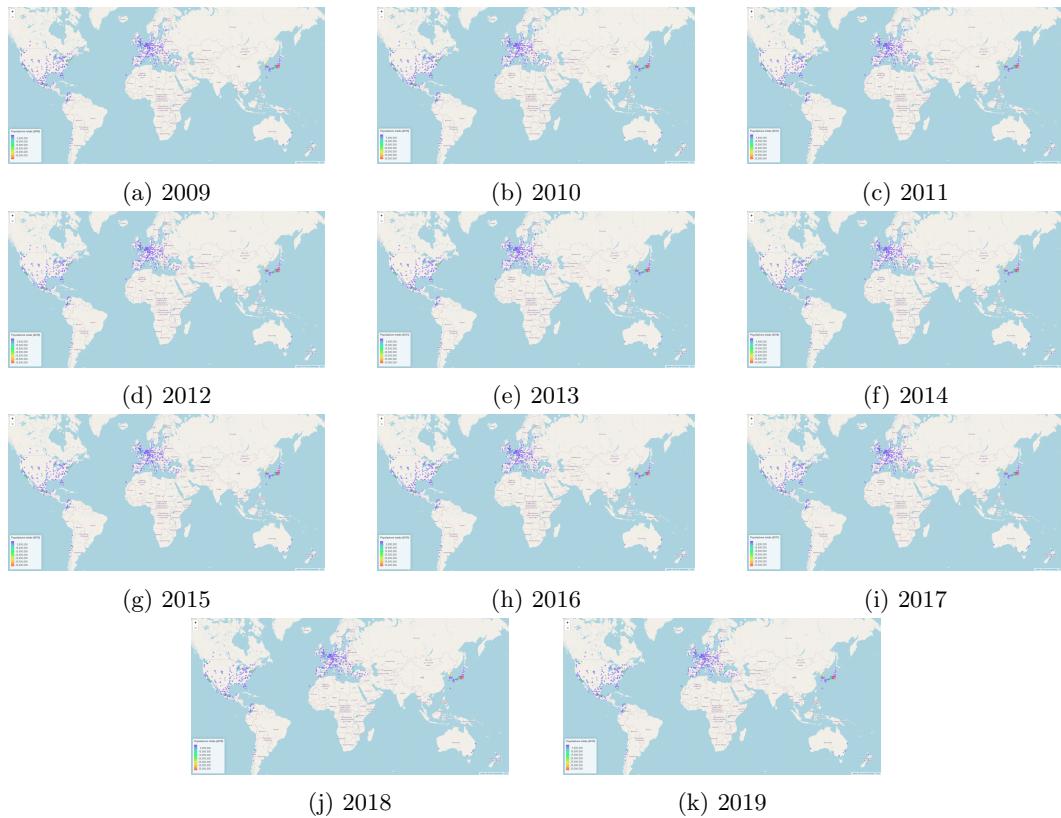
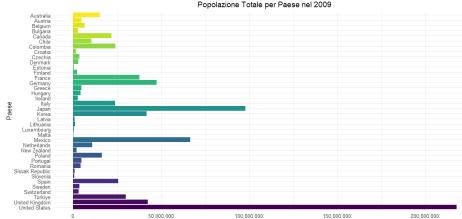
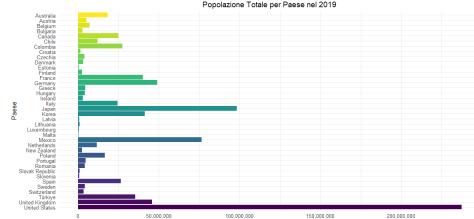


Figure 21: Population, all ages, administrative data rilevata tra il 2009-2019

Il primato per la popolazione totale dal 2009 al 2019 è da attribuire ad una singola città, ovvero Tokyo in Giappone, ribadendo l'attrattiva verso le grandi metropoli e la sua importanza come uno dei più grandi centri urbani e economici del mondo. Anche in questo caso Istanbul si pone come la città col maggiore incremento della popolazione. Mentre è la Corea il paese interessato dalla maggiore diminuzione della popolazione, a Seoul per l'esattezza, che nonostante sia una grande metropoli, potrebbe presentare questo andamento a causa di una combinazione di fattori, tra cui l'invecchiamento della popolazione, la diminuzione del tasso di natalità e la migrazione verso altre aree.



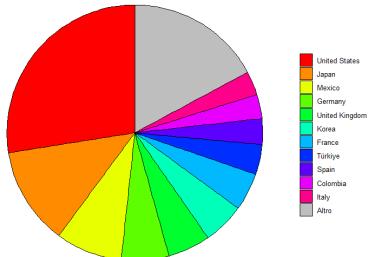
(a) Barplot della popolazione del 2009 delle nazioni OECD



(b) Barplot della popolazione del 2009 delle nazioni OECD

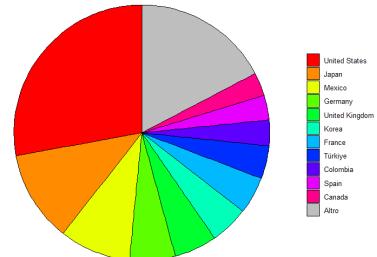
Con la figura 22a e la figura 22b, ottenute tramite lo script "script/total\_population\_barplot.R", abbiamo voluto rappresentare, con il supporto di un barplot, il costante incremento della popolazione delle nazioni OECD.

Distribuzione della popolazione per Nazione nel 2009



(a) Distribuzione della popolazione per nazione nel 2009

Distribuzione della popolazione per Nazione nel 2019



(b) Distribuzione della popolazione per nazione nel 2019

I grafici a torta del 2009 e del 2019, ottenuti con lo script "script/diagramma\_torta.R", mostrano la distribuzione della popolazione tra diverse nazioni, con una categoria "Altro" che aggredisce i dati di tutti gli altri paesi non elencati individualmente. Negli Stati Uniti si osserva una minima variazione percentuale, mantenendo la maggioranza della quota. Il Giappone e il Messico hanno visto cambiamenti significativi: una diminuzione per il Giappone e un aumento per il Messico. La categoria "Altro" è rimasta relativamente stabile, suggerendo piccoli cambiamenti distribuiti tra le altre nazioni nel decennio. Queste variazioni possono essere attribuite a dinamiche demografiche, economiche, politiche e sociali che influenzano i tassi di natalità e mortalità, i flussi migratori e altri fattori di sviluppo.

Per quanto riguarda l'analisi dei boxplot relativi alla popolazione totale nel periodo dal 2009 al 2019, è emersa una serie di risultati particolarmente rilevanti. Alcuni paesi hanno attirato la nostra attenzione in modo significativo a causa del notevole numero di outlier rilevati. In particolare, Giappone, Corea e Stati Uniti analizzandoli tramite lo script "script/search\_outliers.R" abbiamo ottenuto questi risultati:

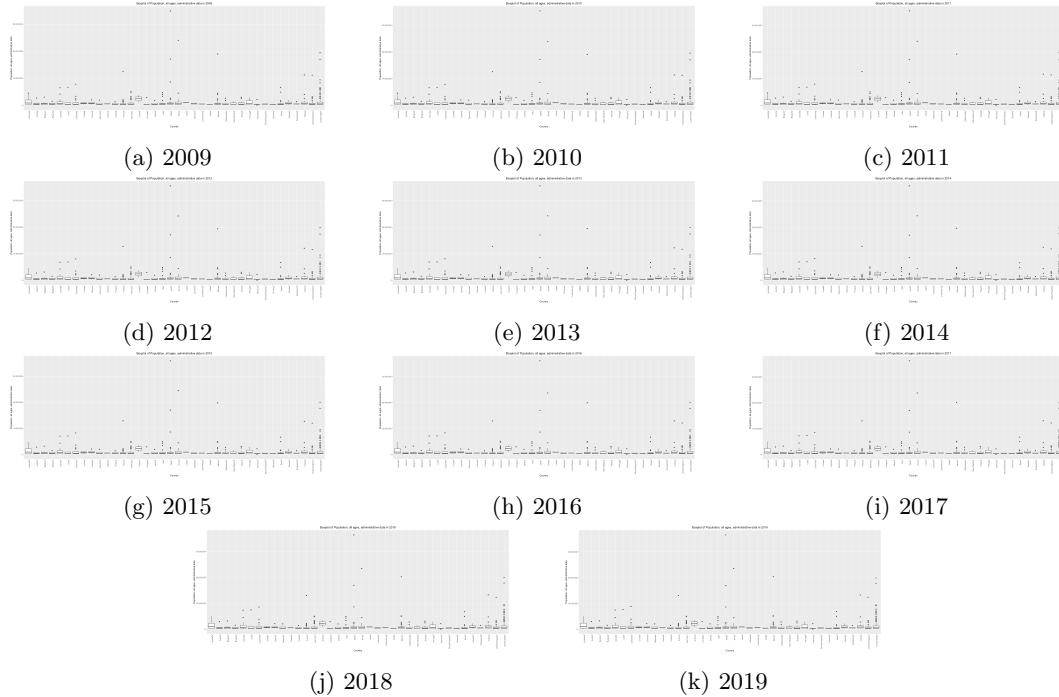


Figure 24: Boxplot Total Population rilevata tra il 2009-2019

Città	2009	2010	2011	2012	2013	2014	2015
Fukuoka	2,656,181	2,665,284	2,668,253	2,670,555	2,672,530	2,672,177	2,677,785
Nagoya	8,629,428	8,612,690	8,608,254	8,609,348	8,614,982	8,616,902	8,635,498
Osaka	17,072,745	17,155,964	17,139,209	17,113,597	17,082,471	17,043,074	17,030,960
Sapporo	2,246,856	2,246,448	2,238,288	2,227,680	2,215,848	2,203,200	2,196,509
Tokyo	34,877,463	35,406,928	35,460,778	35,487,817	35,573,773	35,698,658	35,898,687

Table 11: Valori outlier del Giappone

Città	2012	2013	2014	2015	2016	2017	2018	2019
Gimhae	4,482,500	4,488,328	4,501,965	4,515,105	4,519,437	4,518,743	4,510,286	4,494,889
Seoul	24,177,845	24,295,988	24,397,844	24,478,302	23,712,048	23,790,863	23,468,191	23,575,540

Table 12: Valori outlier della Corea

Città	2009	2010
Atlanta	4,906,016	4,968,156
Boston	4,109,103	4,147,945
Chicago	9,429,498	9,470,661
Dallas	6,505,404	6,615,837
Denver	2,509,417	2,554,588
Detroit (Greater)	4,375,250	4,355,011
Houston	5,918,515	6,039,977
Los Angeles (Greater)	16,935,262	17,080,388
Miami (Greater)	5,650,130	5,730,372
Minneapolis	3,277,805	3,302,482
New York (Greater)	19,450,319	19,575,312
Philadelphia (Greater)	6,306,927	6,338,919
Phoenix	4,153,609	-
San Diego	3,061,203	-
San Francisco (Greater)	6,122,403	-
Seattle	3,414,797	-
St. Louis	2,575,146	-
Washington (Greater)	8,299,654	-

Table 13: Valori outlier degli Stati Uniti

## 4.7 Unemployment rate

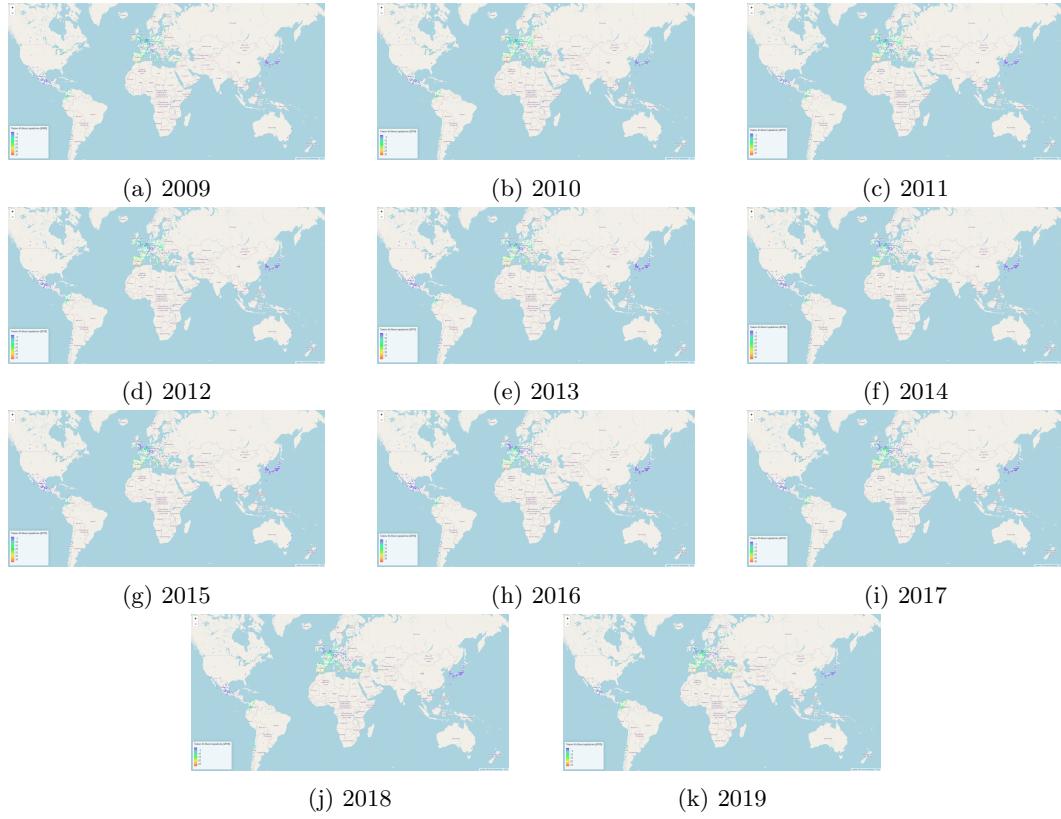


Figure 25: Unemployment rate rilevato tra il 2009-2019

Confrontando le mappe possiamo notare come sia palese che le aree urbane a Sud dei vari paesi presentino un tasso di disoccupazione maggiore rispetto a quelle del Nord. In particolare, l'Irlanda, con la città di Limerick, e poi la Spagna, con la città di Cádiz, si sono alternate ai vertici del tasso di disoccupazione. La città di Limerick è anche la città che ha diminuito maggiormente il suo tasso di disoccupazione nel corso degli anni riflettendo probabilmente il recupero economico dell'Irlanda dopo la crisi finanziaria, supportato da una forte crescita del PIL e investimenti stranieri. Mentre è l'Italia, con la città di Messina, a mostrare il maggiore incremento del tasso di disoccupazione negli anni, che potrebbe essere legato alla crisi economica che ha colpito l'Italia in quegli anni, influenzata da fattori come la recessione globale e i problemi strutturali del mercato del lavoro locale.

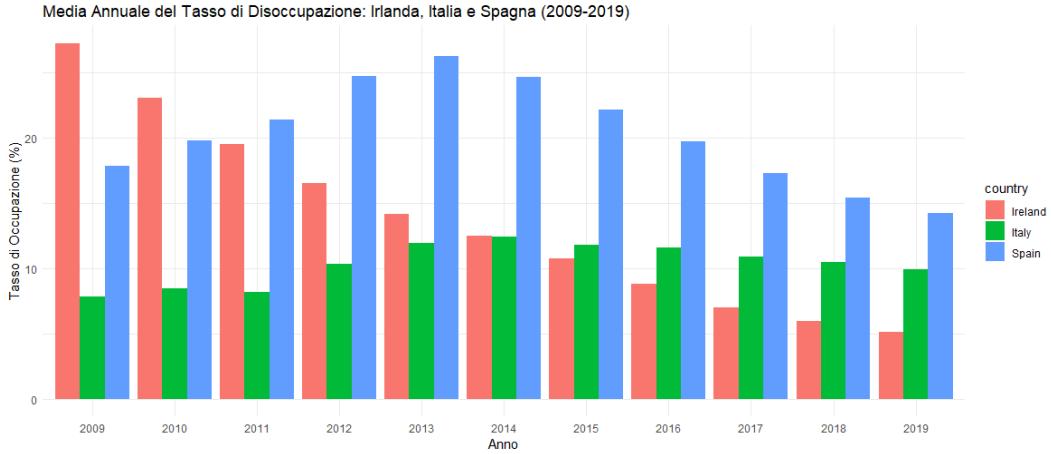


Figure 26: Confronto tra i paesi col tasso di disoccupazione migliore e peggiore

Tramite il barchart raggruppato in figura 26, ottenuto tramite lo script ”script/groupchart\\_unemployment\\_rate.R”, abbiamo voluto evidenziare le due nazioni che negli anni hanno avuto un incremento e un decremento notevole del tasso di disoccupazione, confrontandole con l’andamento della Spagna che negli ultimi anni ha presentato il maggior tasso di disoccupazione tra i Paesi OECD.

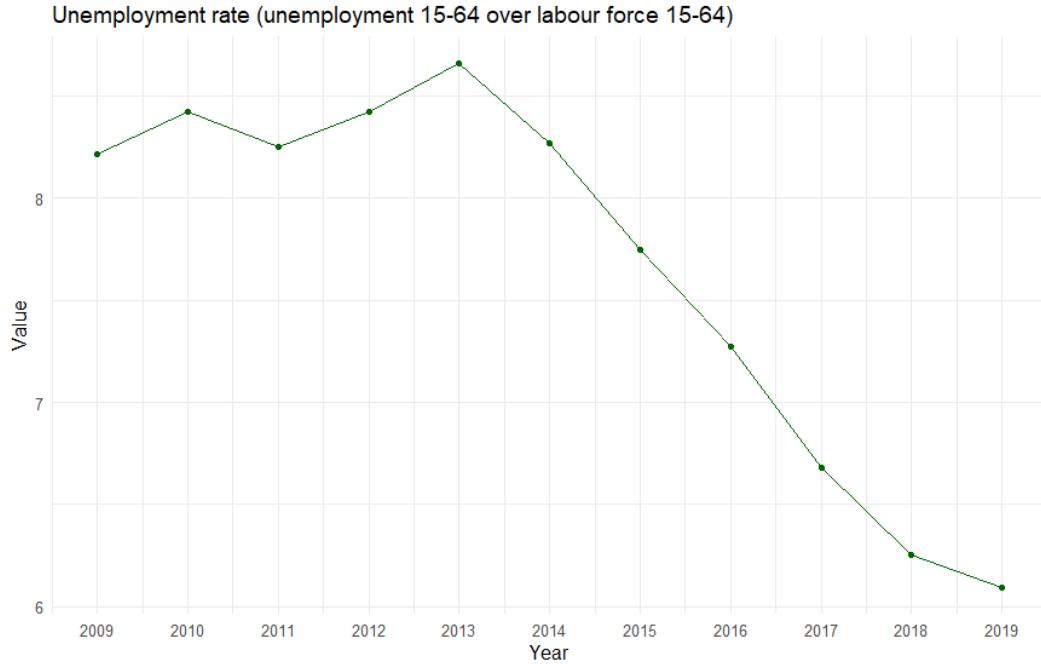


Figure 27: Evoluzione media del tasso di disoccupazione annuale dei Paesi OECD

Nella figura 27 possiamo verificare l’andamento del tasso di disoccupazione medio dei Paesi OECD, che negli anni ha continuato a diminuire, tranne tra il 2011 e il 2013.

Analizzando i boxplot per la variabile inerente al tasso di disoccupazione possiamo notare diverse informazioni utili.

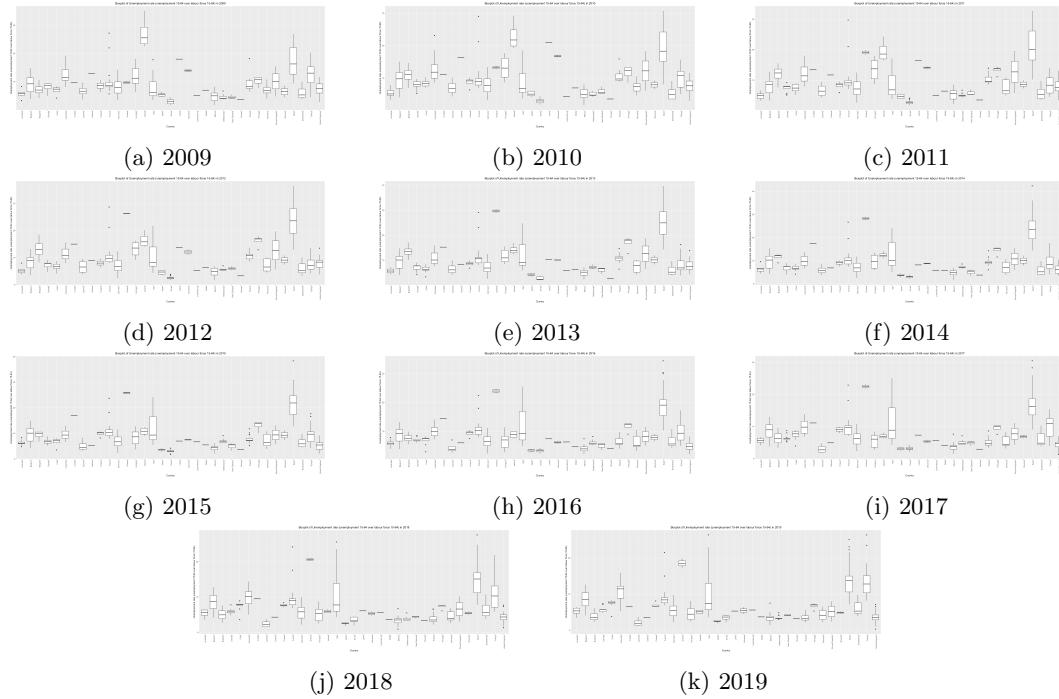


Figure 28: Boxplot Unemployment Rate rilevata tra il 2009-2019

#### 1. Ampiezza dei Boxplot:

- Paesi con boxplot ampi indicano una maggiore variabilità nel tasso di disoccupazione nel corso degli anni. Questo suggerisce una maggiore instabilità economica o cambiamenti significativi nel mercato del lavoro.
- Paesi con boxplot stretti mostrano una minore variabilità, indicando una maggiore stabilità nel tasso di disoccupazione.

#### 2. Posizione delle Mediane:

- Se la mediana si trova verso la parte superiore del box, indica che la maggior parte degli anni ha avuto un tasso di disoccupazione più alto.
- Al contrario, una mediana verso la parte inferiore indica che la maggior parte degli anni ha avuto un tasso di disoccupazione più basso.

#### 3. Outliers:

- I punti dati che appaiono come outliers separati dai boxplot indicano anni in cui il tasso di disoccupazione è stato insolitamente alto o basso. Ad esempio, un outlier alto per la Spagna potrebbe rappresentare un picco di disoccupazione durante la crisi finanziaria.

#### 4. Paesi con Elevata Variabilità:

- Spagna e Grecia: Questi paesi mostrano una notevole variabilità nel tasso di disoccupazione, con boxplot ampi. La Spagna, in particolare, potrebbe avere alcuni

degli outliers più alti, indicando anni con tassi di disoccupazione eccezionalmente elevati.

5. Paesi con Tasso di Disoccupazione Generalmente Alto:

- Grecia: Con una mediana tra le più alte, suggerendo che, per la maggior parte del periodo considerato, ha avuto un tasso di disoccupazione relativamente elevato.

6. Paesi con Stabilità nel Tasso di Disoccupazione:

- Giappone, Norvegia: Questi paesi mostrano boxplot più stretti, indicando una minore variabilità annuale nel tasso di disoccupazione.

7. Valori Specifici:

- L'outlier più alto per la Spagna indica un tasso di disoccupazione del 42,4%, che è eccezionalmente elevato rispetto ad altri anni e paesi.
- La Grecia potrebbe avere una mediana intorno al 22,5%, indicando un livello medio elevato di disoccupazione.

Analizzando gli outlier nei boxplot relativi al tasso di disoccupazione tramite lo script "script/search\_outliers.R" abbiamo prestato particolare attenzione per quanto riguarda la Spagna, Italia e Francia.

Città	2014	2015	2016	2017	2018	2019
Cádiz	42.4	38.4	34.5	30.5	27.6	25.1
Córdoba	-	-	30.4	28.3	-	23.1
Granada	-	-	30.0	-	-	22.3

Table 14: Valori outlier della Spagna

Città	2009	2018	2019
Palermo	17.8	-	-
Messina	-	25.6	26.4

Table 15: Valori outlier dell'Italia

Città	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Fort-de-France	21.2	21.0	21.2	23.0	19.6	18.2	18.0	18.1	17.6	15
Saint-Denis	29.0	29.7	28.7	29.1	26.7	23.7	22.4	22.9	24.2	21.5

Table 16: Valori outlier della Francia

## 5 Risoluzione domande di ricerca

Una volta definite le caratteristiche del nostro dataset, possiamo passare a cercare una risposta alle domande che ci siamo posti nella Sezione 2.

### 5.1 Domanda di ricerca 1

La prima domanda di ricerca mira ad individuare l'esistenza di un possibile legame significativo tra il PIL pro capite e l'inquinamento di un'area geografica. Per fare ciò ci siamo affidati alla statistica descrittiva bivariata. Innanzitutto, tramite lo script "script/1\_research\_question.R", per ottenere una misura quantitativa della correlazione tra le variabili abbiamo considerato la covarianza o correlazione campionaria, che permettono di vedere se vi è una relazione lineare tra le variabili. Abbiamo applicato lo script ad ogni anno tra il 2009 e il 2019, rilevando un andamento pressochè identico tranne che per gli ultimi due anni.

Anno	Covarianza campionaria	Correlazione campionaria	$D^2$
2009	-40684.96	-0.43	0.1849
2010	-42204.71	-0.42	0.1764
2011	-47629.27	-0.44	0.1936
2012	-43275.53	-0.44	0.1936
2013	-45495.38	-0.45	0.2025
2014	-46093.28	-0.45	0.2025
2015	-46809.64	-0.47	0.2209
2016	-47618.52	-0.48	0.2304
2017	-44478.65	-0.46	0.2116
2018	-40759.16	-0.41	0.1681
2019	-37278.88	-0.39	0.1521

Table 17: Covarianza e correlazione tra GDP pro capite e Mean Exposure to a PM2.5

Una covarianza negativa, come quelle individuate, tra il PIL pro capite e l'esposizione media a PM2.5 suggerisce che in generale, nei luoghi con un PIL pro capite più alto, si tende ad avere livelli più bassi di esposizione a PM2.5, e viceversa. In altre parole, sembra esserci una tendenza per cui un aumento del benessere economico (misurato dal PIL pro capite) è associato a una diminuzione dell'inquinamento atmosferico (misurato dall'esposizione a PM2.5).

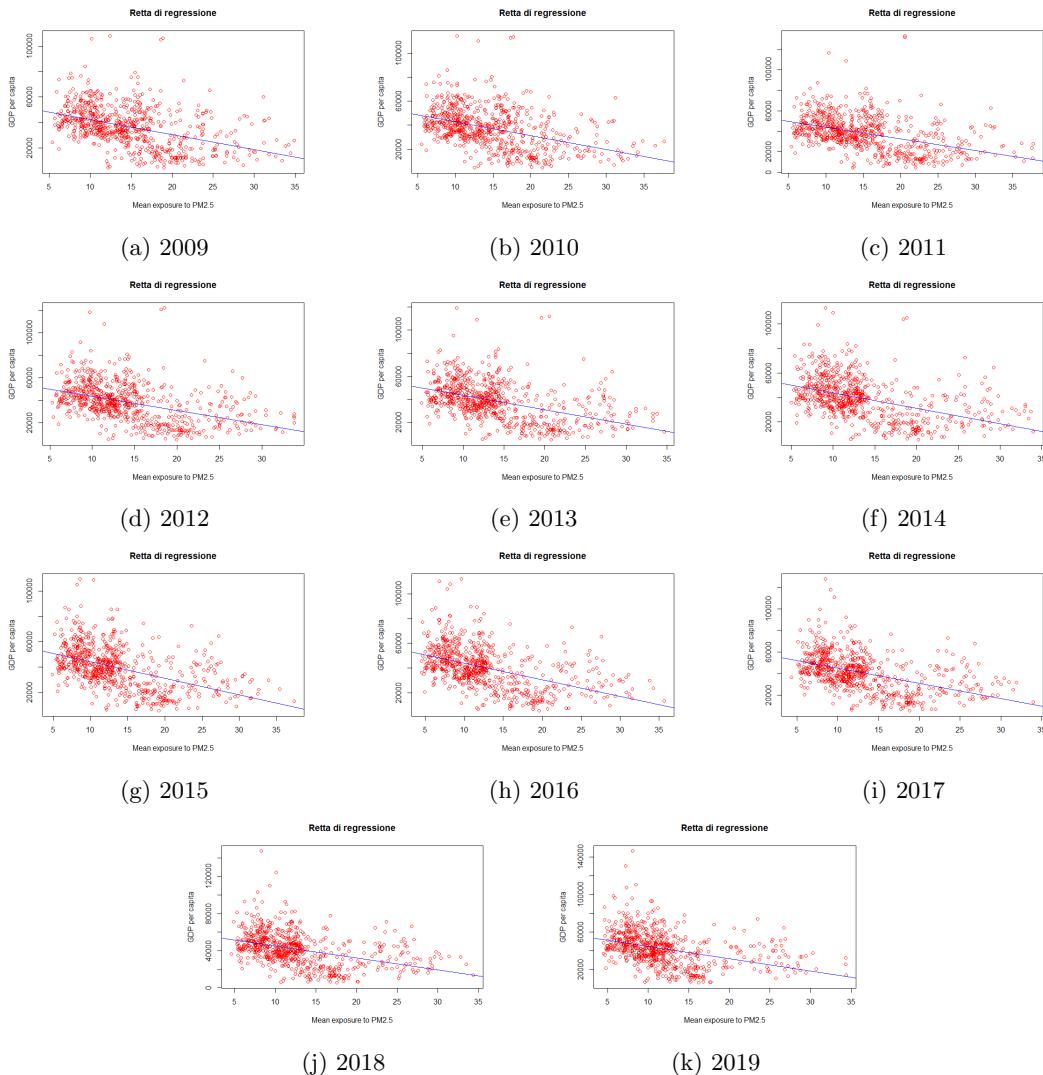
Tuttavia i valori assunti dalla covarianza sono relativamente grandi, ma la sua interpretazione è complessa perché dipende dalle scale di misura delle variabili. Le unità di misura (dollari per il PIL e  $\mu\text{g}/\text{m}^3$  per PM2.5) influenzano direttamente la grandezza della covarianza, rendendo difficile fare confronti diretti con altre covarianze senza standardizzazione. Inoltre la covarianza è molto influenzata dalla presenza di outlier, che come abbiamo visto nella sezione 4, sono molti.

Il coefficiente di correlazione conferma una correlazione negativa moderata tra il PIL pro capite e l'esposizione a PM2.5. Questo suggerisce che, in media, un aumento nel PIL pro capite è associato a una diminuzione nei livelli di PM2.5, sebbene la relazione non sia particolarmente forte.

A differenza della covarianza, il coefficiente di correlazione è una misura standardizzata che non dipende dalle unità di misura delle variabili. Pertanto, fornisce un'indicazione più chiara della forza e della direzione della relazione tra le due variabili.

In conclusione, i nostri dati indicano una relazione inversa moderata tra il benessere economico di un luogo (PIL pro capite) e il livello di inquinamento atmosferico (esposizione a PM2.5). Questa relazione è logicamente coerente, poiché spesso le aree più ricche possono permettersi migliori misure di controllo dell'inquinamento.

Una volta definita questa relazione moderata tra le due variabili, abbiamo voluto rappresentare la retta che interpola i nostri punti, in modo che ci possa aiutare a prevedere il PIL pro capite di una città, basandoci sul suo livello di inquinamento. Per far ciò, abbiamo esteso lo script "script/1\_research\_question.R", realizzando uno scatterplot per ogni anno.

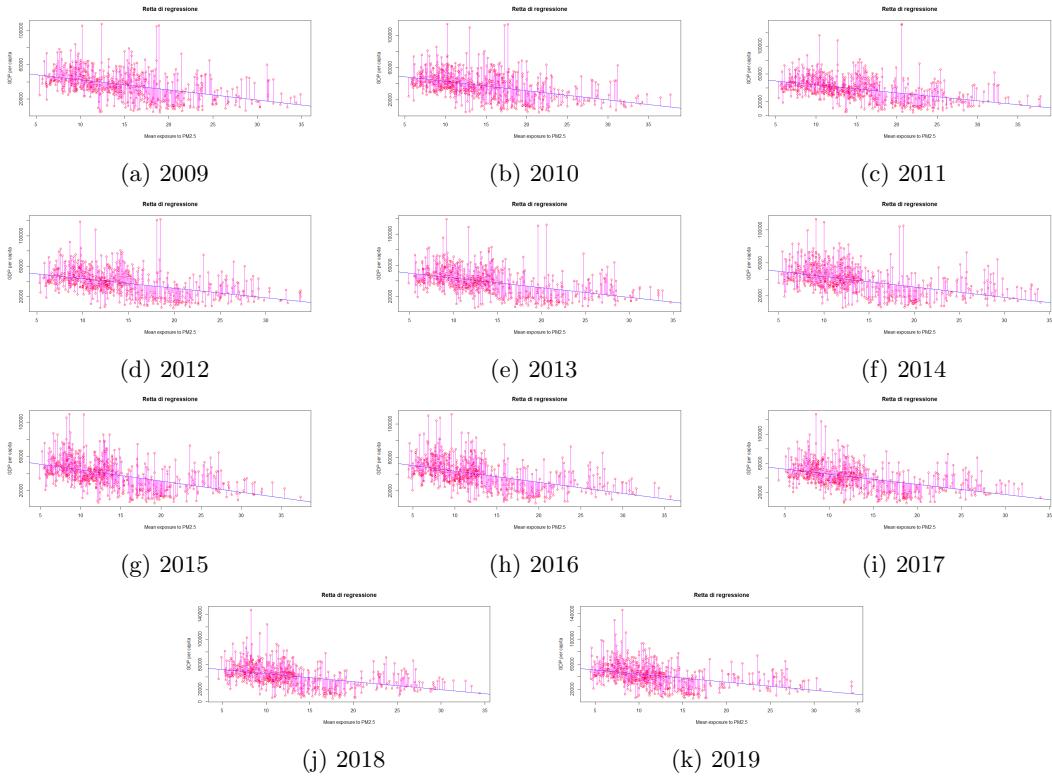


Come possiamo osservare, gli scatterplot, e la rispettiva linea che interpola i dati, sono molto simili tra loro nei diversi anni analizzati. Andiamo a vedere più nello specifico le rette ottenute, mostrando i parametri  $\beta$  (coefficiente angolare) e  $\alpha$  (intercetta) che costituiscono l'equazione relativa alla retta  $y = \alpha + \beta x$ , ottenuta tramite il metodo dei minimi quadrati.

Anno	Intercetta	Coefficiente angolare
2009	54132.11	-1178.19
2010	54251.88	-1151.12
2011	55665.43	-1145.03
2012	56265.11	-1268.66
2013	56092.44	-1246.83
2014	56215.02	-1260.93
2015	57264.90	-1304.95
2016	57159.53	-1336.09
2017	58910.61	-1397.19
2018	58062.57	-1292.98
2019	57433.82	-1302.79

Table 18: Intercetta e coefficiente angolare delle rette di regressione

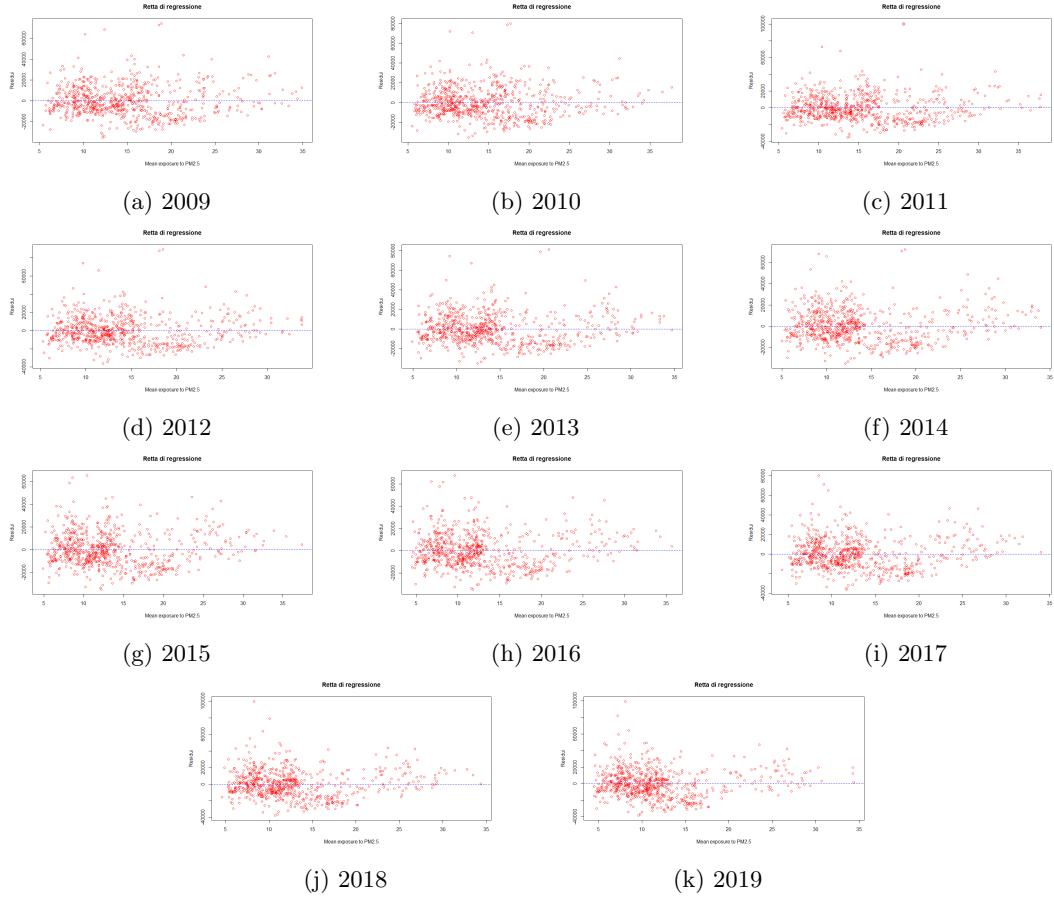
A questo punto, verifichiamo anche i residui delle rette di interpolazione, ovvero quanto si discostano i valori osservati dai valori stimati con la retta di regressione.



Nei grafici possiamo vedere che i segmenti verticali congiungono i punti reali con quelli delle rette di regressione, mostrando la grande differenza con i valori stimati. Come possiamo vedere le rette non hanno interpolato perfettamente i dati, vista anche la poca correlazione individuata precedentemente tra le due variabili.

Un esame più accurato del modo con cui la retta di regressione interpola i dati e di come i residui si dispongano intorno alla retta interpolante influenzandone la posizione, può

essere ottenuto attraverso il diagramma dei residui che è un grafico in cui i valori dei residui sono posti sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse. Individuiamo coi quali sono i valori che influenzano maggiormente la retta.



Tuttavia nelle regressioni, sia semplice che multipla, gioca un ruolo fondamentale il coefficiente di determinazione ( $D^2$ ) per vedere se interpola bene i nostri dati. Più ( $D^2$ ) è vicino ad 1, più la retta interpola bene i nostri dati, mentre più ci avviciniamo a 0 e meno la retta interpola i nostri dati. Nel caso della regressione lineare semplice, il coefficiente di determinazione è il quadrato del coefficiente di correlazione, individuato precedentemente nella tabella 17. Infatti, come possiamo vedere, si tratta di valori molto bassi, che indicano una scarsa capacità della retta di adattarsi ai nostri dati. Quindi possiamo concludere che nonostante ci sia una correlazione moderata ciò non implica necessariamente una relazione causale diretta tra queste variabili. Altri fattori non considerati potrebbero influenzarle entrambe.

## 5.2 Domanda di ricerca 2

In questa sezione, esploriamo come la forza lavoro, il tasso di disoccupazione e la popolazione totale influenzano congiuntamente il PIL pro capite di una città attraverso l'uso della regressione lineare multipla. Questo metodo ci permette di comprendere le interazioni complesse tra queste variabili, fornendo una rappresentazione più accurata e dettagliata della realtà economica urbana. L'analisi non solo rivela le dinamiche sottostanti il PIL pro capite, ma offre anche spunti preziosi per la formulazione di politiche economiche informate e mirate.

La covarianza e la correlazione, in questo caso, forniscono due matrici simmetriche; la matrice delle covarianze contiene sulla diagonale principale la varianza delle singole colonne del data frame, mentre la matrice delle correlazioni contiene il numero 1 sulla diagonale principale. La matrice di correlazione evidenzia tutte le correlazioni lineari tra tutte le coppie di variabili quantitative. Analizziamo, quindi, con lo script "script/2\_research\_question.R" questi due valori per gli anni tra il 2009 e il 2019.

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	231749951.88 / 1	2989000178.45 / 0.15	-11614.39 / -0.17	5376114316.7 / 0.13
<b>Forza lavoro</b>	2989000178.45 / 0.15	1691083002150.35 / 1	-98689.04 / -0.02	3033177694108.65 / 0.87
<b>T. di disoccupazione</b>	-11614.39 / -0.17	-98689.04 / -0.02	20.07 / 1	-137484.31 / -0.01
<b>Popolazione totale</b>	5376114316.7 / 0.13	3033177694108.65 / 0.87	-137484.31 / -0.01	7160815241442.53 / 1

Table 19: Covarianza e correlazione campionaria tra le variabili in esame (2009)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	246757810.17 / 1	2888286263.82 / 0.14	-13492.81 / -0.19	5277662622.33 / 0.12
<b>Forza lavoro</b>	2888286263.82 / 0.14	1704157574242.73 / 1	-149845.57 / -0.02	3075864328137.73 / 0.87
<b>T. di disoccupazione</b>	-13492.81 / -0.19	-149845.57 / -0.02	21.28 / 1	-262629.32 / -0.02
<b>Popolazione totale</b>	5277662622.33 / 0.12	3075864328137.73 / 0.87	-262629.32 / -0.02	7303112365886.39 / 1

Table 20: Covarianza e correlazione campionaria tra le variabili in esame (2010)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	270373736.4 / 1	2863594739.46 / 0.13	-15827.8 / -0.2	5226369990.15 / 0.12
<b>Forza lavoro</b>	2863594739.46 / 0.13	1727407967687.04 / 1	-197025.53 / -0.03	3109016337848.63 / 0.87
<b>T. di disoccupazione</b>	-15827.8 / -0.2	-197025.53 / -0.03	22.13 / 1	-386563.96 / -0.03
<b>Popolazione totale</b>	5226369990.15 / 0.12	3109016337848.63 / 0.87	-386563.96 / -0.03	7406279224805.17 / 1

Table 21: Covarianza e correlazione campionaria tra le variabili in esame (2011)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	260737557.59 / 1	2967253450.19 / 0.14	-15231.83 / -0.18	5474998190.28 / 0.12
<b>Forza lavoro</b>	2967253450.19 / 0.14	1782824161174.99 / 1	-209205.44 / -0.03	3176515335235.19 / 0.87
<b>T. di disoccupazione</b>	-15231.83 / -0.18	-209205.44 / -0.03	28.54 / 1	-391063.34 / -0.03
<b>Popolazione totale</b>	5474998190.28 / 0.12	3176515335235.19 / 0.87	-391063.34 / -0.03	7497188672910.89 / 1

Table 22: Covarianza e correlazione campionaria tra le variabili in esame (2012)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	255499909.23 / 1	3207300569.72 / 0.15	-16320.2 / -0.18	5856142835.2 / 0.13
<b>Forza lavoro</b>	3207300569.72 / 0.15	1816412707949.73 / 1	-321825.23 / -0.04	3222288919260.56 / 0.87
<b>T. di disoccupazione</b>	-16320.2 / -0.18	-321825.23 / -0.04	31.75 / 1	-607565.29 / -0.04
<b>Popolazione totale</b>	5856142835.2 / 0.13	3222288919260.56 / 0.87	-607565.29 / -0.04	7592645855239.79 / 1

Table 23: Covarianza e correlazione campionaria tra le variabili in esame (2013)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	255543861.7 / 1	3324268973.37 / 0.15	-14492.29 / -0.17	6045267662.91 / 0.14
<b>Forza lavoro</b>	3324268973.37 / 0.15	1854822981841.3 / 1	-305183.34 / -0.04	3276189261861.56 / 0.87
<b>T. di disoccupazione</b>	-14492.29 / -0.17	-305183.34 / -0.04	28.87 / 1	-543572.2 / -0.04
<b>Popolazione totale</b>	6045267662.91 / 0.14	3276189261861.56 / 0.87	-543572.2 / -0.04	7700778134573.02 / 1

Table 24: Covarianza e correlazione campionaria tra le variabili in esame (2014)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	239854254.46 / 1	3568550208.4 / 0.17	-11665.48 / -0.15	6525862090.53 / 0.15
<b>Forza lavoro</b>	3568550208.4 / 0.17	1886887803907.34 / 1	-289942.11 / -0.04	3319237133630.91 / 0.86
<b>T. di disoccupazione</b>	-11665.48 / -0.15	-289942.11 / -0.04	24.44 / 1	-508553.25 / -0.04
<b>Popolazione totale</b>	6525862090.53 / 0.15	3319237133630.91 / 0.86	-508553.25 / -0.04	7815013829720.3 / 1

Table 25: Covarianza e correlazione campionaria tra le variabili in esame (2015)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	249919208.31 / 1	3717460325.93 / 0.17	-11137.12 / -0.15	6761807719.49 / 0.15
<b>Forza lavoro</b>	3717460325.93 / 0.17	1931513947640.87 / 1	-233071.33 / -0.04	3359979578924.78 / 0.86
<b>T. di disoccupazione</b>	-11137.12 / -0.15	-233071.33 / -0.04	21.4 / 1	-380201.14 / -0.03
<b>Popolazione totale</b>	6761807719.49 / 0.15	3359979578924.78 / 0.86	-380201.14 / -0.03	7856329927263.6 / 1

Table 26: Covarianza e correlazione campionaria tra le variabili in esame (2016)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	264507802.67 / 1	3723826357.92 / 0.16	-10612.99 / -0.15	6778497232.57 / 0.15
<b>Forza lavoro</b>	3723826357.92 / 0.16	1965739465368.13 / 1	-182588.76 / -0.03	3414488379124.77 / 0.86
<b>T. di disoccupazione</b>	-10612.99 / -0.15	-182588.76 / -0.03	18.47 / 1	-270289.09 / -0.02
<b>Popolazione totale</b>	6778497232.57 / 0.15	3414488379124.77 / 0.86	-270289.09 / -0.02	7944926790694.31 / 1

Table 27: Covarianza e correlazione campionaria tra le variabili in esame (2017)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	278954565.15 / 1	3780641194.08 / 0.16	11830.95 / -0.17	6850677655.25 / 0.14
<b>Forza lavoro</b>	3780641194.08 / 0.16	2023509896872.67 / 1	-155829.2 / -0.03	3480478376362.46 / 0.86
<b>T. di disoccupazione</b>	11830.95 / -0.17	-155829.2 / -0.03	17.11 / 1	-194788.41 / -0.02
<b>Popolazione totale</b>	6850677655.25 / 0.14	3480478376362.46 / 0.86	-194788.41 / -0.02	8014673716967.47 / 1

Table 28: Covarianza e correlazione campionaria tra le variabili in esame (2018)

	<b>PIL pro capite</b>	<b>Forza Lavoro</b>	<b>T. di disoccupazione</b>	<b>Popolazione totale</b>
<b>PIL pro capite</b>	285469095.03 / 1	3775288646.01 / 0.16	-15216.91 / -0.22	6756608435.8 / 0.14
<b>Forza lavoro</b>	3775288646.01 / 0.16	2068631726905.31 / 1	-131368.64 / -0.02	3540605716964.36 / 0.86
<b>T. di disoccupazione</b>	-15216.91 / -0.22	-131368.64 / -0.02	17.39 / 1	-100355.7 / -0.01
<b>Popolazione totale</b>	6756608435.8 / 0.14	3540605716964.36 / 0.86	-100355.7 / -0.01	8110294586384.02 / 1

Table 29: Covarianza e correlazione campionaria tra le variabili in esame (2019)

Dal contenuto delle tabelle possiamo comprendere come i due parametri abbiano valori molto simili nel corso degli anni. Inoltre, il PIL pro capite ha relazioni molto deboli con il resto delle variabili, con una correlazione inversa con il tasso di disoccupazione. Possiamo notare comunque una correlazione molto forte tra la forza lavoro e la popolazione totale, che è una relazione abbastanza intuibile.

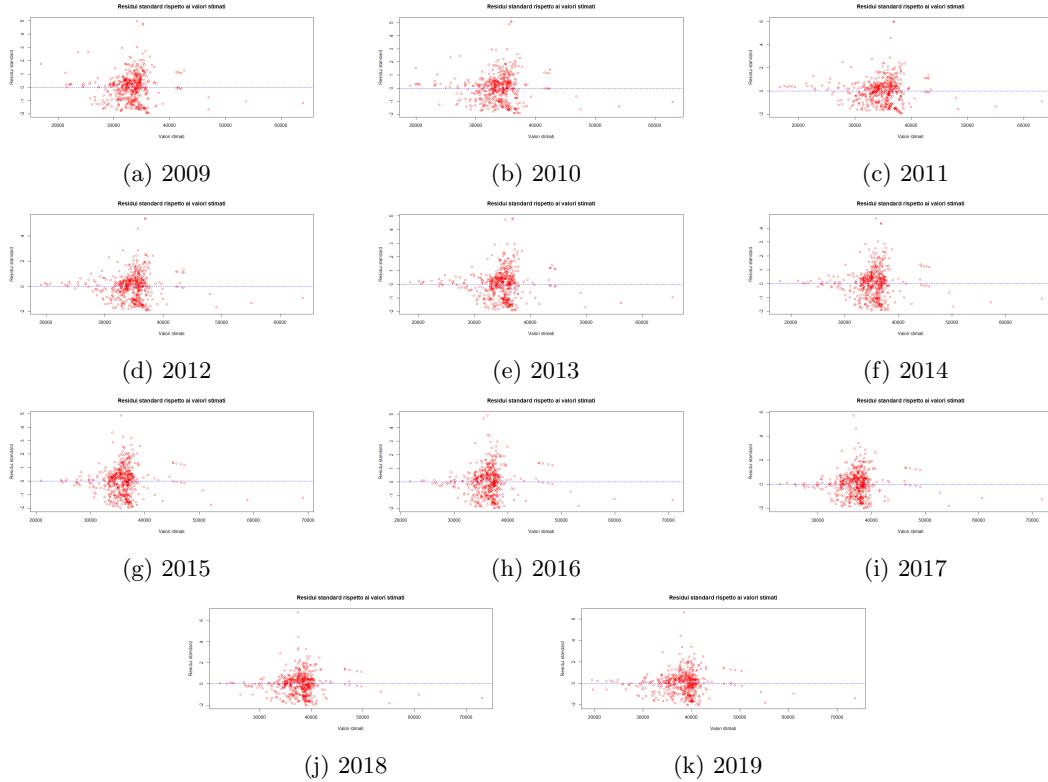
Definito ciò, verifichiamo se comunque, tramite queste variabili insieme, riusciamo a definire un modello di regressione lineare multipla che si adatti ai dati, stimando sempre col metodo dei minimi quadrati l'intercetta e i coefficienti angolari relativi a ciascuna variabile.

Dai valori riportati nella tabella 30, possiamo osservare, innanzitutto, che il PIL pro capite ha valori molto alti, quando le altre variabili sono tutte poste a 0. Ciò potrebbe indicare che il PIL pro capite ha una base alta indipendentemente dalle altre variabili considerate nel modello. Per quanto riguarda l'impatto delle variabili sulla retta di regressione, possiamo notare come sia la popolazione totale sia la forza lavoro abbiano dei valori simili, prossimi a 0, che suggeriscono la presenza di una relazione positiva, sebbene molto piccola, tra queste variabili e il PIL pro capite. D'altro canto, vi è una forte relazione negativa tra il tasso di disoccupazione e il PIL pro capite, suggerendo che aumenti in questa variabile possono avere un impatto significativamente negativo sull'economia della città.

Anno	Intercetta	Coefficiente angolare 1	Coefficiente angolare 2	Coefficiente angolare 3
2009	36,750	0.00002	570.2	0.00169
2010	37,940	0.00004	622.5	0.00157
2011	39,460	0.00003	701.1	0.00153
2012	37,980	0.0001	521.9	0.00143
2013	38,050	0.00008	496.9	0.00154
2014	38,230	0.00009	483.8	0.00155
2015	38,190	0.00013	455.7	0.0016
2016	38,780	0.00015	500.3	0.0016
2017	39,680	0.00017	556.8	0.00155
2018	40,760	0.0002	675.8	0.00148
2019	42,090	0.00020	863.2	0.00143

Table 30: Intercetta e coefficienti angoli della retta di regressione multipla

Una volta calcolati i valori dell'intercetta e dei coefficienti angolari, è possibile osservare gli scostamenti (residui) tra i valori osservati e i corrispondenti valori stimati ottenuti mediante la regressione lineare multipla. Per osservare meglio la differenza, andremo a realizzare un grafico in cui i residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) con il metodo dei minimi quadrati.



I grafici dei residui standardizzati mostrano che i residui sono distribuiti in modo apparentemente casuale attorno allo zero, senza mostrare schemi distinti che suggeriscano relazioni non lineari non catturate dal modello. La varianza dei residui appare costante attraverso i

valori stimati, indicando l'assenza di eteroschedasticità e suggerendo che l'omogeneità della varianza è una condizione soddisfatta. Tuttavia, si notano alcuni punti dati che si discostano significativamente dalla concentrazione centrale, suggerendo la presenza di potenziali outliers che potrebbero influenzare l'analisi.

Tuttavia come già accennato per la regressione lineare, uno dei fattori principali che può dirci se la retta di regressione multipla interpola bene i dati è il coefficiente di determinazione.

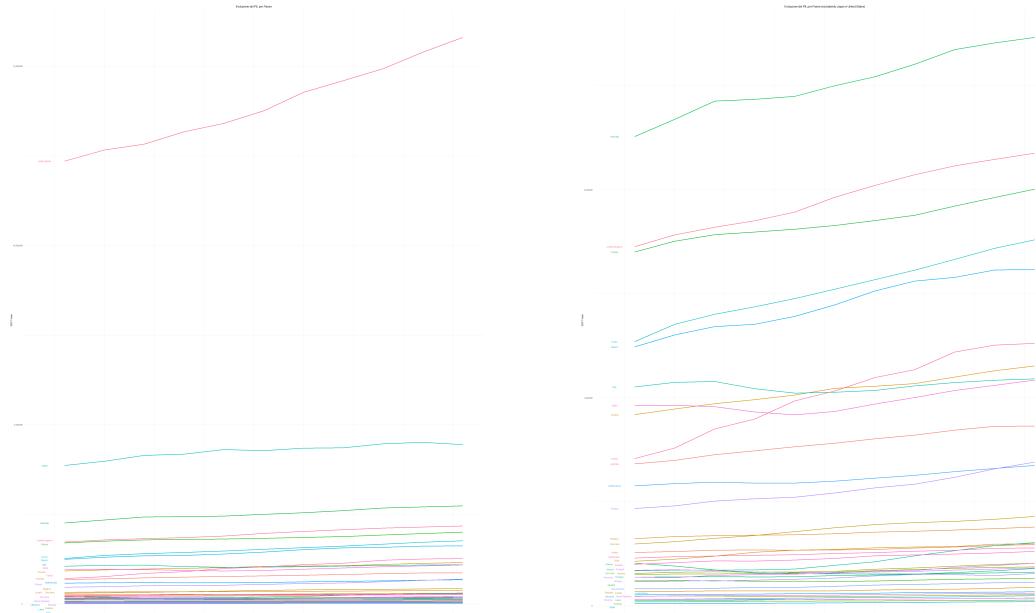
Anno	Coefficiente di determinazione
<b>2009</b>	0.05094615
<b>2010</b>	0.05324529
<b>2011</b>	0.05775944
<b>2012</b>	0.0488034
<b>2013</b>	0.05285102
<b>2014</b>	0.04977821
<b>2015</b>	0.0493919
<b>2016</b>	0.05020506
<b>2017</b>	0.04847611
<b>2018</b>	0.05352436
<b>2019</b>	0.06962778

Table 31: Coefficiente di determinazione della retta di regressione multipla

Come possiamo vedere, ( $D^2$ ) presenta dei valori molto bassi, che indicano una scarsa capacità della retta di adattarsi ai nostri dati. Quindi possiamo concludere che il PIL pro capite non è legato alle variabili forza lavoro, il tasso di disoccupazione e la popolazione totale prese in esame, come si potrebbe diversamente pensare.

### 5.3 Domanda di ricerca 3

Il nostro dataset e l'analisi svolta ci permettono di gettare luce sull'evoluzione del Prodotto Interno Lordo (PIL o GDP) nel corso del tempo, concentrandoci su diverse città in diverse nazioni. Questo approccio ci consente di scrutare le tendenze a lungo termine e di identificare le fluttuazioni economiche che si sono verificate in queste città nel contesto dei rispettivi paesi. Per prima cosa abbiamo deciso di analizzare graficamente attraverso un plot l'evoluzione del Prodotto Interno Lordo (PIL) per paese nel corso degli anni e ciò l'abbiamo fatto tramite lo script: "script/plot\_GDP\_variazione\_annua.R" ottenendo questo risultato:



(a) Evoluzione generale del PIL di tutti i Paesi OECD

(b) Evoluzione generale del PIL di tutti i Paesi tranne Stati Uniti e Giappone

Per comprendere al meglio l'evoluzione nel tempo del PIL nelle varie città abbiamo utilizzato lo script “script/3\_research\_question.R” che ci ha aiutato a esaminare come il Prodotto Interno Lordo (PIL) delle diverse città sia cambiato nel corso del tempo, consentendoci di ottenere una visione dell'evoluzione economica nel corso degli anni. I risultati ottenuti forniscono una panoramica chiara delle variazioni percentuali annuali, rivelando chiaramente se una città ha sperimentato una crescita economica positiva o una contrazione. Ad esempio, nel 2010, la città di Greater Melbourne ha registrato un aumento del PIL del 0.75% rispetto all'anno precedente, indicando una crescita economica modesta ma positiva.

Tra le varie città spicca quella di Villavicencio, che nel 2011 ha registrato un aumento del PIL del 42%. In base ad alcune ricerche sulla situazione economica e sociale di quell'anno, abbiamo indentificato diversi fattori che potrebbero aver contribuito all'aumento del PIL. In particolare, la costruzione della strada Villavicencio-Yopal è stata un fattore determinante nell'aumento del PIL di Villavicencio. Questa strada ha ridotto il tempo di percorrenza tra Villavicencio e Bogotà da 12 a 5 ore, rendendo la città più accessibile e attraente per le imprese e gli investitori. L'aumento della produzione petrolifera ha anche contribuito all'aumento del PIL di Villavicencio. Il dipartimento del Meta ha registrato una crescita significativa della produzione petrolifera negli ultimi anni. Nel 2011, la produzione petrolifera del Meta è stata di 1,6 milioni di barili al giorno, in aumento rispetto ai 1,2 milioni di barili al giorno del 2010. Infine, l'aumento del turismo ha contribuito a stimolare l'economia locale. Villavicencio è una destinazione turistica popolare, grazie alla sua posizione strategica, ai suoi paesaggi naturali e alle sue attrazioni culturali. Nel 2011, la città ha ricevuto 1,2 milioni di turisti, in aumento rispetto ai 1 milione di turisti del 2010. È importante notare che questi sono solo alcuni dei fattori che potrebbero aver contribuito all'aumento del PIL di Villavicencio tra il 2010 e il 2011.

Tramite alcune ricerche, la città di Campeche risulta aver registrato nel 2015 una con-

trazione del PIL del 38%, che potrebbe essere dovuto principalmente alla caduta dei prezzi del petrolio. Campeche è una città portuale situata sulla costa del Golfo del Messico, ed è un importante centro per l'estrazione e la raffinazione del petrolio. Nel 2014, i prezzi del petrolio erano in aumento, raggiungendo un picco di oltre 110 dollari al barile. Tuttavia, nel 2015, i prezzi del petrolio iniziarono a scendere, raggiungendo un minimo di circa 30 dollari al barile. Questa caduta dei prezzi ha avuto un impatto negativo sull'economia di Campeche, in quanto ha ridotto i ricavi dell'industria petrolifera. Altri fattori che potrebbero aver contribuito al calo del PIL a Campeche nel 2015 includono:

- La recessione economica globale, che ha ridotto la domanda di beni e servizi da parte delle imprese e dei consumatori.
- La violenza legata al narcotraffico, che ha creato incertezza e ha dissuaso gli investimenti.

In aggiunta, abbiamo deciso di calcolare la variazione percentuale media annuale del PIL per ogni città sempre tramite lo script “script/3\_research\_question.R” ottenendo, per le città con una crescita maggiore, questi risultati:

1. Cork (Irlanda): Crescita media annuale del 10.39%.
2. Gaziantep (Turchia): Crescita media annuale del 9.11%.
3. Limerick (Irlanda): Crescita media annuale del 7.57%.
4. Webb (USA): Crescita media annuale del 7.04%.
5. Weld (USA): Crescita media annuale del 6.84%.

Le percentuali di crescita possono riflettere il successo di politiche economiche, l'attrazione di investimenti, o il rapido sviluppo di settori industriali o tecnologici chiave.

Per quanto riguarda le città che invece hanno mostrato una diminuzione, abbiamo ottenuto:

1. Thessaloniki (Grecia): Decrescita media annuale del -2.40%.
2. Athina (Grecia): Decrescita media annuale del -2.25%.
3. Carmen (Messico): Decrescita media annuale del -1.38%.
4. Campeche (Messico): Decrescita media annuale del -1.38%.
5. Groningen (Paesi Bassi): Decrescita media annuale del -0.97%.

Queste città hanno sperimentato un declino economico. Ciò può essere dovuto da difficoltà economiche a livello nazionale, cambiamenti demografici, declino di settori economici importanti, o sfide nel mercato globale.

Concentrandoci sulla variabile ”GDP (Million USD, constant prices, constant PPP, base year 2015)”, abbiamo analizzato ulteriormente il dataset per ottenere informazioni dettagliate.

La Media e la Mediana del PIL ci forniscono una panoramica delle dimensioni economiche delle città nel dataset per ogni anno. Queste statistiche ci mostrano come la media sia

aumentata da circa 47,204 milioni di USD nel 2009 a 59,301 milioni di USD nel 2019, mentre la mediana, che fornisce una misura più stabile del "valore centrale", è variata da circa 17,690 milioni di USD nel 2009 a 21,162 milioni di USD nel 2019.

La Variazione Percentuale Media Annuale del PIL ci mostra come, in media, il PIL delle città sia cambiato di anno in anno. Questa metrica ha rivelato un tasso di crescita medio più alto nel 2015 (circa 4.09%) e un tasso di crescita medio più basso nel 2012 (circa 1.06%).

Abbiamo, poi, esaminato la Variazione e la Dispersione del PIL. Analizzando la variazione del PIL dal 2009 al 2019 è stata di 12,097.25 milioni di USD, indicando che, in media, le città hanno sperimentato un incremento del PIL in questo periodo. Tuttavia, la mediana della variazione, pari a 3,409.00 milioni di USD, suggerisce che la maggior parte delle città ha vissuto un aumento più moderato del PIL.

Per quanto riguarda la Dispersione del PIL (Deviazione Standard) nel 2009 era di 117,231.67 milioni di USD, mentre nel 2019 è aumentata a 148,442.10 milioni di USD. Questo incremento nella deviazione standard indica una maggiore disparità nel PIL tra le città nel 2019 rispetto al 2009. Tale aumento della variabilità può essere interpretato come un segno di crescita economica differenziata tra le città, con alcune aree che hanno registrato una crescita più rapida rispetto ad altre.

Successivamente, concentrandoci sull'anno 2019, abbiamo esaminato la distribuzione del PIL sia per paese che per città. Questo ci ha rivelato come gli Stati Uniti abbiano registrato la media del PIL più alta nel 2019, seguiti da Corea, Australia e Giappone. Allo stesso modo, alcune città come Tokyo, New York (Greater), Los Angeles (Greater) e Seoul hanno mostrato PIL notevolmente elevati. Queste informazioni mettono in evidenza notevoli differenze nella distribuzione del PIL, sottolineando le sfide legate alle disparità economiche globali e alla crescita economica equilibrata.

In conclusione, l'analisi del PIL delle città nel periodo 2009-2019 rivela una crescita economica generale, ma anche un aumento delle disparità economiche tra le città. Mentre alcune aree hanno sperimentato una rapida crescita economica, altre sono rimaste più indietro, evidenziando la necessità di politiche mirate per promuovere uno sviluppo economico più equilibrato e inclusivo.

#### 5.4 Domanda di ricerca 4

L'analisi dei cluster per raggruppare le nazioni OECD, basandosi su un sottoinsieme delle variabili, ovvero il PIL pro capite, il tasso di disoccupazione, la densità di popolazione e l'inquinamento, riveste un'importanza cruciale. Questo approccio non solo illumina le differenze e somiglianze socio-economiche tra le nazioni, ma è anche fondamentale per la formulazione di politiche pubbliche mirate e strategie di sviluppo sostenibile. Attraverso questa analisi, possiamo ottenere preziose intuizioni sulle dinamiche globali, facilitando una migliore comprensione e cooperazione internazionale. Questa analisi verrà effettuata per le osservazioni dell'anno più recente, ovvero il 2019.

Tramite lo script "script/domanda4/dataset\_clustering.R" abbiamo ottenuto un dataset contenente per ogni Paese le medie delle variabili selezionate per il clustering per l'anno 2019. Una volta ottenuto questo dataset possiamo notare come le variabili scelte abbiano delle scale di valori molto diverse tra di loro.

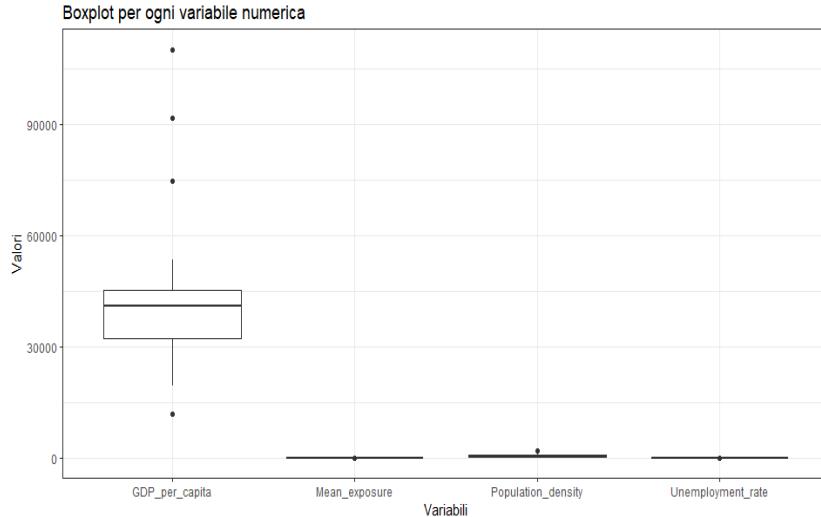


Figure 34: Differenti scale per le variabili

Gli algoritmi di clustering sono sensibili alle differenze di scala tra le variabili, per questo tramite lo script "script/scaling.R", abbiamo applicato una standardizzazione ai dati trasformando tutte le variabili in modo che abbiano media 0 e deviazione standard 1, permettendo così a ciascuna variabile di contribuire equamente al processo di clustering.

Una volta standardizzati i dati, una prima analisi che possiamo effettuare, per creare i cluster, è utilizzare le misure di distanza, utilizzata per quantificare la differenza tra due individui, nel nostro caso nazioni. Per calcolare la distanza esistono diverse metriche possibili, ma tramite lo script "script/domanda4/distance\_cluster.R" abbiamo optato per quella Euclidea, perché si adatta bene a tutti i metodi di clustering. Inoltre, questa metrica necessita di dati scalati, proprio come i nostri. La matrice delle distance può essere consultata nel file di testo "misurazioni/matrix\_distance.txt".

Successivamente, andiamo a ricavarci un'altra misura fondamentale per le fasi successive, ovvero la misura di non omogeneità totale, come parametro di confronto con i cluster ottenuti. Uno dei metodi per calcolare la misura di non omogeneità statistica all'interno di un insieme utilizza i quadrati delle distanze euclidee. Lo script utilizzato è "script/domanda4/distance\_cluster.R". Il nostro insieme di dati standardizzato ha ottenuto una misura di non omogenità interna pari a 132.

Ottenuta questa misura possiamo procedere con il clustering. Ne esistono principalmente di tre tipi, ma noi ne useremo solo due, ovvero: i cluster gerarchici e non. Il metodo di enumerazione completa è computazionalmente impraticabile.

#### 5.4.1 Cluster gerarchici

Lo scopo dei metodi gerarchici è avere una visione di insieme e creare il dendogramma. Sarà compito nostro tagliare il dendogramma e determinare il numero ottimale di cluster. Esistono due tipologie di cluster gerarchici.

- Agglomerativi, in cui inizialmente tutti gli individui vengono assegnati ad un unico cluster e poi man mano vengono uniti con quelli più vicini, fino a raggrupparli tutti in unico cluster.

- Divisivi, che prevedono il processo inverso, ovvero da un unico cluster si raggiungono quelli singoli.

Noi ci concentriamo su 5 metodi che appartengono alla classe dei cluster gerarchici agglomerativi e che utilizzano la matrice delle distanze. Vedremo i dendogrammi che generano, disegnando anche dei rettangoli intorno ai cluster, per semplificare l'analisi, individuati in base all'altezza alla quale si opera il taglio del dendrogramma oppure in base al numero k di cluster che si vogliono ottenere attraverso la funzione. I vari metodi verranno attuati tramite le funzioni dello script "script/domanda4/clustering\_gerarchico.R".

Il primo metodo che andremo ad applicare è il metodo del **legame singolo**, che utilizza la minima distanza tra gli individui per raggruppare. Il dendrogramma ottenuto è il seguente:

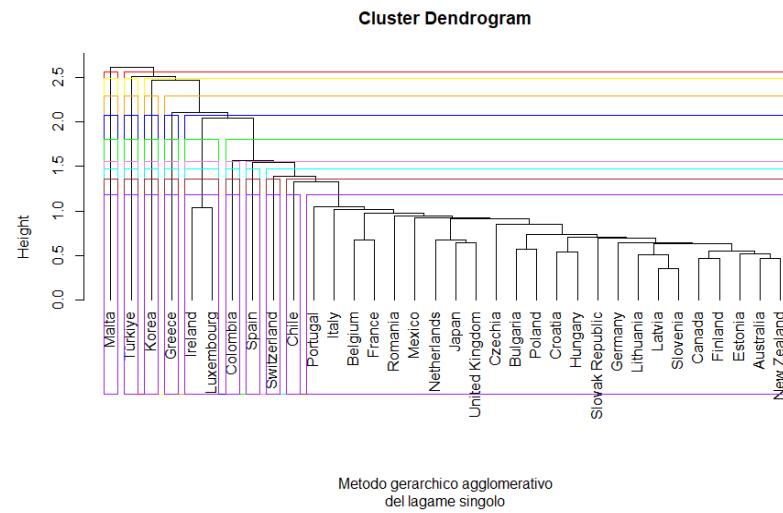


Figure 35: Cluster ottenuti col metodo del legame singolo

Il secondo metodo è quello del **legame completo**, che è definito come la massima tra tutte le distanze che si possono calcolare tra ogni individuo. Il dendrogramma ottenuto è il seguente:

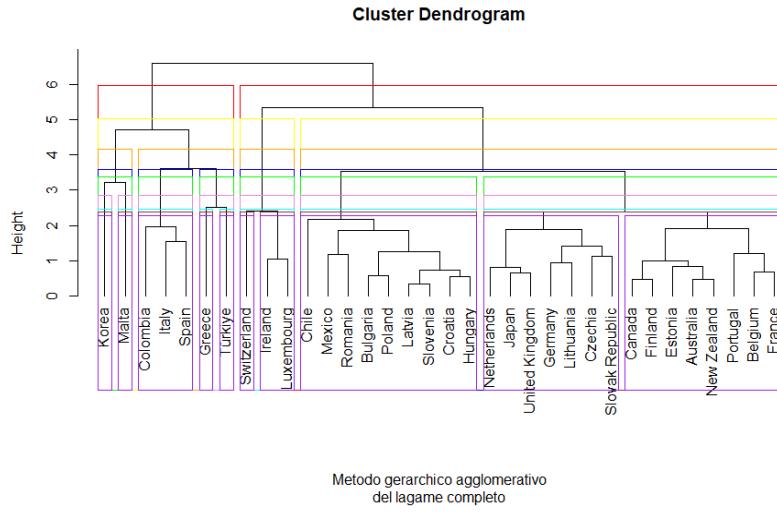


Figure 36: Cluster ottenuti col metodo del legame completo

Il terzo metodo è quello del **legame medio**, ogni cluster è unito al suo cluster più vicino sulla base della media delle distanze tra tutti i punti nel primo cluster e tutti i punti nel secondo cluster. Il dendogramma ottenuto è il seguente:

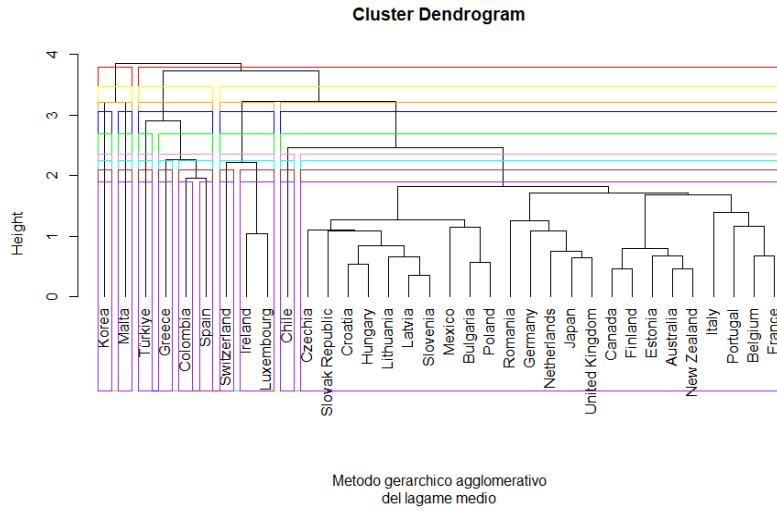


Figure 37: Cluster ottenuti col metodo del legame medio

Il quarto è il metodo del **centroide**, dove due cluster vengono uniti in base alla minima distanza tra i loro centroidi, ovvero i punti che rappresentano il centro medio di ciascun cluster. Il dendogramma ottenuto è il seguente:

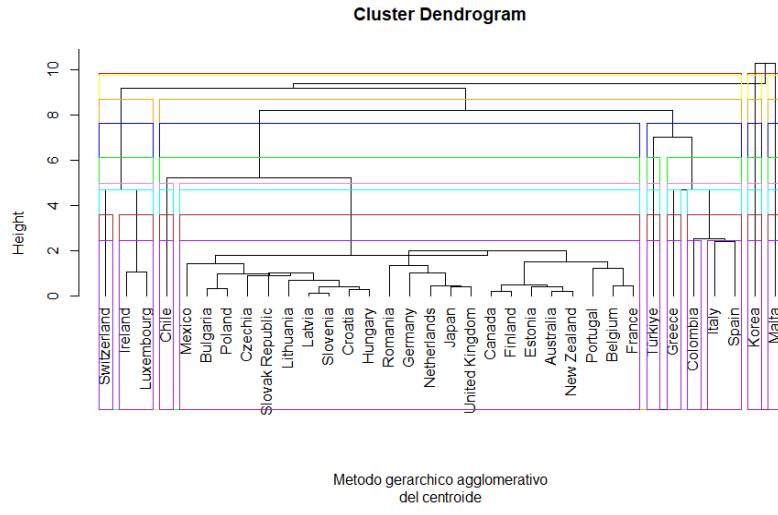


Figure 38: Cluster ottenuti col metodo del centroide

Il quinto metodo è quello della **mediana**, due cluster vengono fusi basandosi sulla distanza tra le loro mediane, che sono calcolate come il punto centrale di ogni cluster dopo ogni fusione. Il dendogramma ottenuto è il seguente:

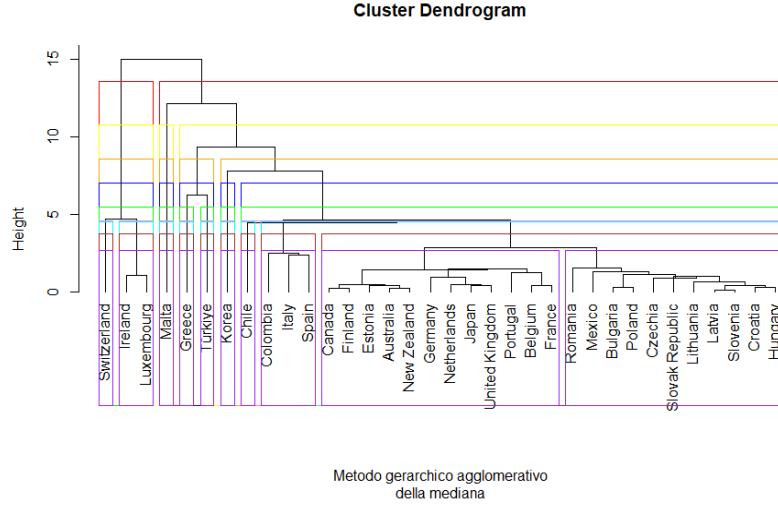


Figure 39: Cluster ottenuti col metodo della mediana

Il metodo del legame singolo e del legame completo, sono generalmente meno preferibili, in casi come il nostro, a causa della loro sensibilità agli outlier, presenti nel nostro dataset originale, e della tendenza a formare catene (nel caso del single linkage) o cluster disomogenei (nel caso del complete linkage). Anche il metodo del centroide e della mediana potrebbero essere influenzati dagli outlier. Il metodo del legame medio fornisce un equilibrio tra sensibilità agli outlier e capacità di formare cluster compatti, anche in caso di dati non standardizzati.

Per determinare il numero ottimale di cluster definiti col metodo del legame medio, inizialmente applichiamo il metodo dello screeplot per ottenere una comprensione visuale

della struttura dei dati. Successivamente, basandoci sulla scelta preliminare del numero di cluster, esaminiamo la non omogeneità statistica all'interno dei cluster e tra i vari cluster. Questa analisi viene estesa anche ai cluster adiacenti per assicurarci di aver identificato la configurazione più appropriata in termini di omogeneità interna ed esterna.

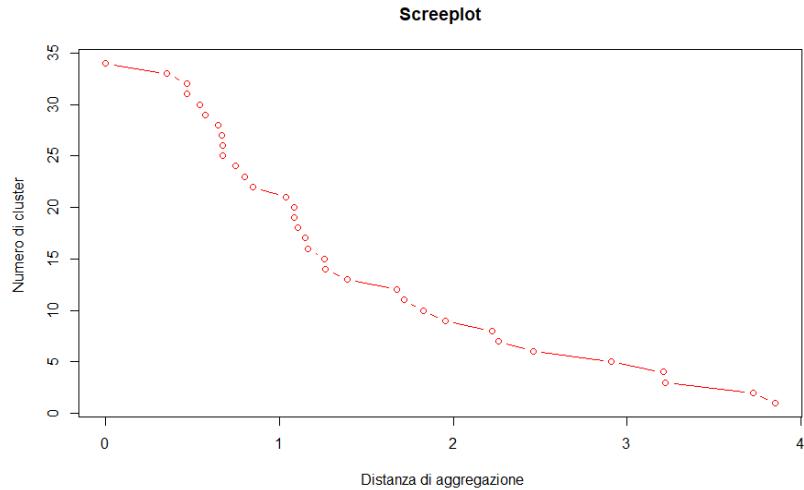


Figure 40: Screeplot applicato al clustering ottenuto col metodo del legame medio

Lo screeplot ottenuto con lo script "script/domanda4/clustering\_gerarchico.R", suggerisce di considerare una suddivisione in 7 gruppi, dal momento che nel passaggio da 8 a 7 si registra un consistente incremento della distanza di aggregazione. Da questo punto, possiamo partire per la nostra analisi successiva, utilizzando le misure di non omogeneità statistica, accennata già precedentemente. Questa misura ci descrive in modo più dettagliato la bontà di un clustering, in quanto i cluster devono essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster e massimizzare la misura di non omogeneità statistica tra i gruppi.

Tramite un'altra funzione dello script "script/domanda4/clustering\_gerarchico.R", abbiamo effettuato diverse prove, partendo da 7 come numero di cluster. Il valore della non omogeneità tra i gruppi è ottenuto sottraendo alla non omogeneità statistica interna dell'insieme originale (calcolata precedentemente) la somma dei valori di non omogeneità interna ai gruppi. I risultati ottenuti sono illustrati nella seguente tabella:

k	Within	Between	Between/Total
<b>8</b>	38.0016	93,9984	0,71
<b>7</b>	40.78829	91,21171	0,69
<b>6</b>	45.69154	86,30846	0,65
<b>5</b>	51.05097	80,94903	0,61
<b>4</b>	56.19789	75,80211	0,57
<b>3</b>	78.67849	53,32151	0,40

Table 32: Valori di non omogeneità intera ed esterna per i cluster ottenuti col legame medio

I clustering con un k pari a 8, 7 e 6 presentano molti singleton cluster, il che è poco

utile per la nostra analisi; mentre il clustering con  $k$  pari a 3 presenta un valore di non omogeneità interna superiore a quella esterna, che va contro i nostri obiettivi. Un  $k$  pari a 5 sembra la scelta migliore in base ai valori ottenuti per la non omogeneità statistica, lo screeplot e un numero moderato di singleton cluster. Con un  $k$  pari a 5 otteniamo il seguente raggruppamento:

1. Corea;
2. Malta;
3. Turchia, Grecia, Colombia, Spagna;
4. Svizzera, Irlanda, Lussemburgo;
5. Australia, Belgium, Bulgaria, Canada, Chile, Croatia, Czechia, Estonia, Finland, France, Germany, Hungary, Italy, Japan, Latvia, Lithuania, Mexico, Netherlands, New Zealand, Poland, Portugal, Romania, Slovak Republic, Slovenia, United Kingdom.

L'ultimo gruppo come possiamo notare è molto più popoloso e si tratta per la maggior parte di paesi del Centro ed Est Europa. In aggiunta, nel clustering con  $k=4$ , come possiamo notare nella figura 37, i due singleton cluster vengono fusi.

#### 5.4.2 Cluster non gerarchici

L'obiettivo dei metodi non gerarchici è quello di ottenere un'unica partizione degli  $n$  individui di partenza in cluster. A differenza dei metodi gerarchici, in tali tecniche è consentito riallocare gli individui già classificati ad un livello precedente dell'analisi. Esistono moltissime tecniche non gerarchiche, ma il più utilizzato in letteratura è il k-means, che è quello che utilizzeremo noi. Tale metodo partiziona i dati in  $k$  cluster predeterminati minimizzando la somma dei quadrati delle distanze di ogni punto dal centroide del proprio cluster, aggiornando iterativamente i centroidi fino a quando la soluzione non converge su una configurazione stabile. Quindi, richiede che il numero di cluster sia specificato a priori e fornisce in output un'unica partizione. Per determinare il numero ottimale di cluster su cui applicare il k-means, ci affidiamo ad un metodo grafico, ovvero l'Elbow Point. Questo metodo si basa sulla creazione di un grafico che mostra la variazione della somma dei quadrati all'interno dei cluster al variare del numero di cluster. Ottenuto il grafico, generalmente, dovrà scegliere il numero di cluster nel punto in cui il grafico inizia a mostrare un appiattimento, ovvero il punto di gomito. Questo punto indica che aumentare il numero di cluster oltre quella soglia non porta a un miglioramento significativo della somma dei quadrati interni. Applichiamo questo metodo al nostro dataset tramite lo script "script/domanda4/cluster\_non\_gerarchico.R".

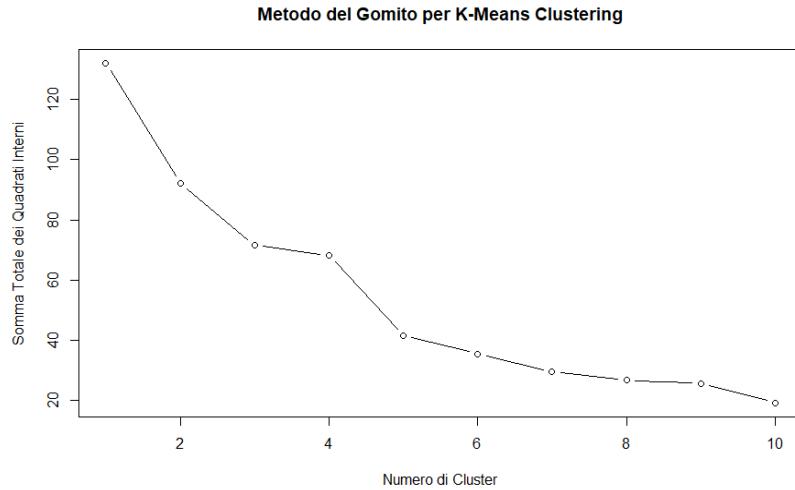


Figure 41: Grafico dell’Elbow Point

Possiamo notare che il grafico inizia a mostrare un appiattimento quando il numero di cluster è pari a 5. Proprio il valore di  $k$  che abbiamo individuato nella Sezione 5.4.1. Abbiamo quindi applicato il metodo del k-means con un  $k$  pari a 5, tramite il comando “`kmeans(cluster[-1], center = 5, iter.max=10, nstart=1)`”, ottenendo i seguenti risultati:

1. Colombia, Grecia, Italia, Spagna;
2. Bulgaria, Croazia, Repubblica Ceca, Ungheria, Lettonia, Polonia, Romania, Repubblica Slovacca, Slovenia;
3. Corea, Turchia, Svizzera;
4. Giappone, Lussemburgo, Malta, Paesi Bassi, Portogallo;
5. Australia, Belgio, Canada, Cile, Estonia, Finlandia, Francia, Germania, Irlanda, Lituania, Messico, Nuova Zelanda, Regno Unito.

#### 5.4.3 Confronto metodi

La suddivisione ottenuta col metodo del kmeans ha ottenuto un valore pari al 62%, dato dalla divisione tra il valore di non omogeneità tra cluster e il valore di non omogeneità interno all’insieme originale. Molto simile al valore ottenuto per quanto riguarda il metodo del legame medio, come possiamo osservare nella tabella 32. Anche se i Paesi raggruppati nei vari cluster sembrano differire. Infatti, non abbiamo più cluster singleton e c’è una maggiore distribuzione dei Paesi tra i vari cluster.

Si può osservare come i risultati del clustering riflettano raggruppamenti coerenti e sensati dal punto di vista socio-economico. Ad esempio, il cluster comprendente Italia, Spagna e Grecia, paesi che hanno affrontato sfide economiche simili durante periodi di crisi, suggerisce una somiglianza nelle loro condizioni economiche. Allo stesso modo, la presenza di paesi dell’Europa orientale come Bulgaria, Croazia, Repubblica Ceca, Ungheria, Lettonia, Polonia, Romania, Repubblica Slovacca e Slovenia nello stesso gruppo potrebbe riflettere

somiglianze storiche e socio-culturali. Gli altri raggruppamenti più diversificati e, soprattutto l'ultimo, più ampi, comprendono Paesi che potrebbero condividere particolari aspetti economici e di sviluppo.

## 5.5 Domanda di ricerca 5

Prima di approfondire la quinta domanda del nostro studio, abbiamo condotto diverse analisi su vari campioni per identificare quelli più adatti ai nostri obiettivi. Inizialmente, abbiamo tentato di utilizzare il dataset così come era, senza modificare i dati. Tuttavia, data la natura realistica del dataset, non abbiamo trovato immediatamente dati adeguati ai nostri scopi. Alla fine, per analizzare in modo più generale la situazione mondiale e non limitarci a una singola nazione, abbiamo scelto di usare come campioni le medie nazionali di un anno per la variabile “Esposizione media della popolazione alle PM2.5”. Ci siamo quindi posti questa domanda: “Come è cambiata l'esposizione media della popolazione alle PM2.5 dal 2009 al 2019, e quali conclusioni possiamo trarre in termini di qualità dell'aria e politiche ambientali nei diversi paesi?” Per rispondere a questa domanda, abbiamo adottato il seguente approccio:

1. **Selezione del Campione:** Abbiamo calcolato la media annuale delle misurazioni di PM2.5 per ogni nazione per l'anno 2009.
2. **Adattamento alla Distribuzione Normale:** Abbiamo utilizzato il test del chi-quadrato per verificare se la distribuzione delle medie di PM2.5 per il 2009 seguisse una distribuzione normale.
3. **Stima Puntuale:** Tramite il metodo dei momenti, abbiamo stimato i parametri ( $\mu$  e  $\sigma^2$ ) della distribuzione normale per il 2009.
4. **Stima Intervallare:**
  - Calcolo dell'intervallo di confidenza per  $\mu$  con varianza non nota.
  - Calcolo dell'intervallo di confidenza per  $\sigma^2$  con valore medio non noto.
5. **Confronto tra Due Popolazioni:** Abbiamo ripetuto tutti i passaggi per il campione relativo ai dati del 2019, effettuando poi un confronto tra le due popolazioni.
6. **Verifica delle Ipotesi:** Abbiamo formulato un'ipotesi adeguata e, tramite un test unilaterale destro, abbiamo verificato questa ipotesi.

### 5.5.1 Selezione del Campione

Per affrontare la nostra domanda di ricerca, abbiamo selezionato un campione rappresentativo dell'esposizione media della popolazione alle PM2.5 in ogni paese nel 2009. Questa scelta è stata fondamentale per valutare l'impatto ambientale e le politiche sulla qualità dell'aria a livello nazionale. Il campione è stato determinato aggregando i dati di PM2.5 per ogni città, calcolando così la media per l'anno 2009 a livello nazionale. Questa aggregazione consente di confrontare i diversi paesi e di analizzare le tendenze generali della qualità dell'aria, tutto ciò è stato elaborato attraverso lo script 'script/domanda5/script1.R'.

### 5.5.2 Adattamento alla Distribuzione Normale

Per valutare la normalità della distribuzione delle medie di PM2.5 del 2009, abbiamo applicato il test del chi-quadrato. Utilizzando lo script 'script/domanda5/script2.R', abbiamo calcolato il valore del chi-quadrato come 2 con 2 gradi di libertà. I valori critici associati, per un livello di significatività del 5%, sono 0.0506 e 7.378. Dal momento che il nostro valore chi-quadrato calcolato rientra tra questi limiti, non rifiutiamo l'ipotesi nulla, concludendo quindi che i dati possono essere considerati normalmente distribuiti. L'output dello script supporta questa conclusione:

```
Valore Chi2 calcolato: 2
Valore critico basso (0.025): 0.0506356159685798
Valore critico alto (0.975): 7.37775890822787
Gradi di libertà: 2
L'ipotesi di normalità non può essere rifiutata.
```

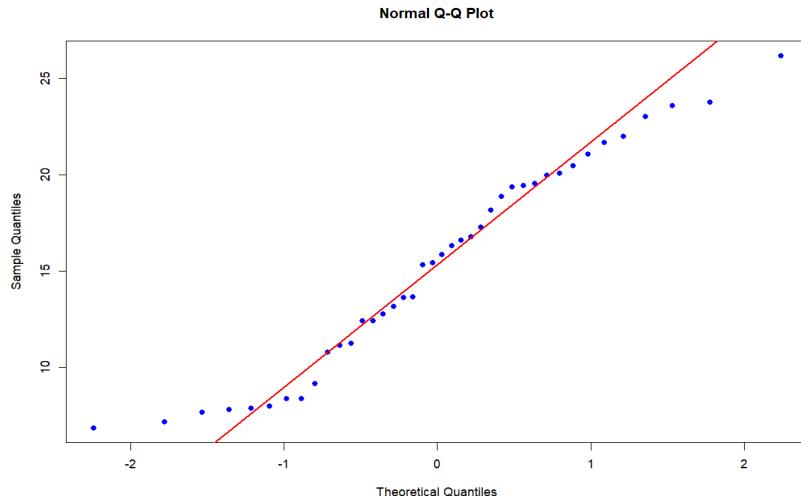


Figure 42: Normal Q-Q Plot dati del 2009

Il Q-Q plot conferma ulteriormente questa analisi. Nel grafico, i quantili del campione si allineano efficacemente con la linea rossa dei quantili teorici della distribuzione normale, indicando che la distribuzione dei dati non si discosta significativamente dalla normalità.

### 5.5.3 Stima Puntuale dei Parametri

Nel terzo passaggio, abbiamo impiegato il metodo dei momenti per stimare i parametri della distribuzione normale. Questo metodo sfrutta i momenti del campione per stimare i corrispondenti parametri della popolazione. Basandoci sui dati dell'esposizione media a PM2.5 del 2009, abbiamo ottenuto i seguenti risultati tramite lo script 'script/domanda5/script3.R':

- **Stimatore della Media ( $\mu$ ):** Lo stimatore della media è risultato essere 15.32. Questo valore rappresenta la media di esposizione a PM2.5 tra le nazioni nel nostro campione.

- **Stimatore della Varianza ( $\sigma^2$ ):** Lo stimatore della varianza è risultato essere 29.18. Questo valore indica la variabilità delle esposizioni a PM2.5 tra le nazioni nel campione.

Questi estimatori sono fondamentali per caratterizzare la distribuzione delle medie di esposizione a PM2.5 nel 2009 e saranno utilizzati nelle analisi successive.

#### 5.5.4 Stima Intervallare

##### **Intervallo di Confidenza per $\mu$ con Varianza NON Nota**

Per la stima dell'intervallo di confidenza della media di esposizione alle PM2.5 nel 2009, abbiamo impiegato la distribuzione t di Student, adeguata quando la varianza della popolazione non è nota. Con un livello di confidenza del 95%, l'intervallo di confidenza calcolato, tramite lo script "script/domanda5/script4.R", è il seguente:

- Intervallo inferiore: 13.57
- Intervallo superiore: 17.07

Questo intervallo riflette la stima della media vera con un grado di confidenza del 95%. Ciò significa che, se ripetessimo questo esperimento molte volte, il 95% degli intervalli di confidenza calcolati conterebbe il vero valore medio di PM2.5.

##### **Intervallo di Confidenza per $\sigma^2$ con Valore Medio NON Noto**

Per la varianza della popolazione, abbiamo calcolato l'intervallo di confidenza utilizzando la distribuzione chi-quadrato, che è appropriata dato che la varianza della popolazione è sconosciuta e la media non è presa in considerazione in questo calcolo. L'intervallo di confidenza al 95% per la varianza è stato stimato tramite lo script "script/domanda5/script5.R" e risulta essere compreso tra:

- Intervallo inferiore: 20.08
- Intervallo superiore: 49.34

Questo indica che, con un livello di confidenza del 95%, possiamo essere ragionevolmente sicuri che la vera varianza della popolazione delle medie di esposizione alle PM2.5 si trovi all'interno di questo intervallo.

#### 5.5.5 Confronto tra Due Popolazioni

##### **Selezione e Analisi del Campione 2019**

Abbiamo replicato il processo del 2009 per il campione del 2019, tramite lo script "script/domanda5/script1\_2019.R", concentrando l'analisi sull'esposizione media a PM2.5. I dati sono stati aggregati per paese per ottenere le medie annuali.

##### **Verifica della Normalità per il 2019**

Con un test del chi-quadrato, effettuato con lo script "script/domanda5/script2\_2019.R", abbiamo valutato la normalità delle medie di esposizione a PM2.5 del 2019. Il valore calcolato del Chi2 è risultato essere 2, che si colloca all'interno dell'intervallo definito dai valori critici basso (0.0506) e alto (7.3778) per un livello di confidenza del 95%. Questo risultato indica che non c'è una ragione statistica per rifiutare l'ipotesi di normalità dei dati per il 2019. L'output ottenuto dallo script è il seguente:

```

Valore Chi2 calcolato: 2
Valore critico basso (0.025): 0.0506356159685798
Valore critico alto (0.975): 7.37775890822787
Gradi di libertà: 2
L'ipotesi di normalità non può essere rifiutata.

```

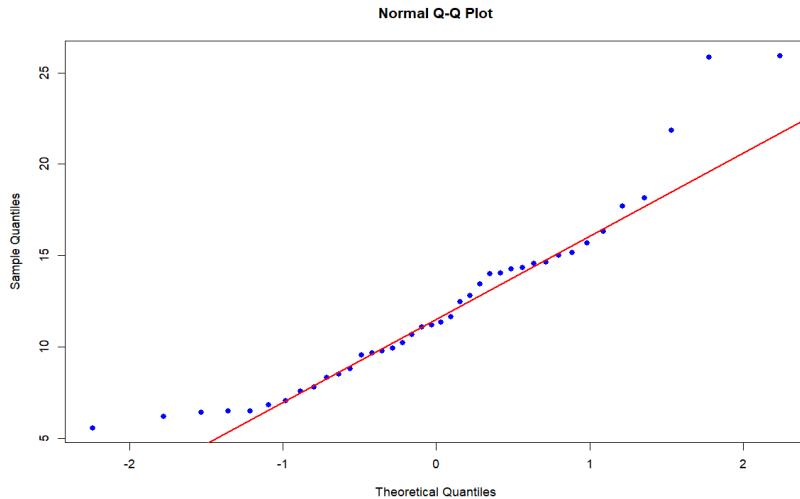


Figure 43: Normal Q-Q Plot dati del 2019

Il Q-Q Plot mostra un'alta corrispondenza con una distribuzione normale teorica, supportando ulteriormente la normalità.

#### **Stima puntuale per il campione del 2019**

Per l'anno 2019, abbiamo applicato il metodo dei momenti, tramite lo script "script/domanda5/script3\_2019.R", per stimare i parametri della distribuzione normale delle medie di esposizione a PM2.5. Gli stimatori risultanti forniscono una media campionaria ( $\mu$ ) di 12.18 e una varianza campionaria ( $\sigma^2$ ) di 23.91. Questi valori suggeriscono che, nel 2019, la media di esposizione a PM2.5 è diminuita rispetto al valore stimato per il 2009, e anche la variabilità (varianza) tra le diverse località è cambiata.

#### **Stima intervallare per la media del campione del 2019**

Abbiamo calcolato l'intervallo di confidenza al 95% per la media di esposizione a PM2.5 nel 2019, tramite lo script "script/domanda5/script4\_2019.R". La media campionaria è di 12.18, con un intervallo di confidenza che va da 10.60 a 13.76. Questo intervallo mostra dove ci aspettiamo che cada la vera media della popolazione con un alto livello di confidenza. La comparazione di questo intervallo con quello del 2009 può aiutarci a comprendere le variazioni nella media di esposizione a PM2.5 tra i due anni.

#### **Stima intervallare per la varianza del campione del 2019**

Per la varianza del campione del 2019 relativa all'esposizione a PM2.5, l'intervallo di confidenza al 95% è stato stimato tra 16.46 e 40.43, applicando lo script "script/domanda5/script5

\_2019.R”. Questo intervallo fornisce una stima dell’effettiva variabilità della media di esposizione a PM2.5 tra i paesi nel campione per quell’anno. Confrontando questo intervallo con quello calcolato per il 2009, si possono trarre conclusioni sulla consistenza o sul cambiamento nella variabilità delle esposizioni nel corso del decennio.

### **Confronto delle Esposizioni a PM2.5 tra il 2009 e il 2019**

L’analisi statistica ha permesso di determinare un intervallo di confidenza al 95% per la differenza delle medie di esposizione a PM2.5 tra il 2009 e il 2019. I risultati, ottenuti con lo script ”script/domanda5/script6.R”, indicano che questo intervallo si colloca tra 0.88 e 5.40. Il fatto che l’intervallo non includa lo zero suggerisce una differenza statisticamente significativa tra le medie dei due anni. Poiché l’intervallo è completamente positivo, ciò implica che la media di esposizione a PM2.5 nel 2009 era significativamente più alta rispetto al 2019. Questo suggerisce una riduzione dell’esposizione a PM2.5 nel corso del decennio.

#### **5.5.6 Verifica delle Ipotesi**

Abbiamo testato le seguenti ipotesi riguardo all’esposizione a PM2.5 nel 2009:

- Ipotesi Nulla (H0): La media dell’esposizione a PM2.5 nel 2009 è uguale o inferiore a  $10 \mu\text{g}/\text{m}^3$ .
- Ipotesi Alternativa (H1): La media dell’esposizione a PM2.5 nel 2009 è maggiore di  $10 \mu\text{g}/\text{m}^3$ .

Nella verifica delle ipotesi, effettuata tramite lo script ”script/domanda5/script7.R”, sull’esposizione a PM2.5 nel 2009, è stato impiegato un test unilaterale destro per determinare se la media dell’esposizione superasse un valore di riferimento di  $10 \mu\text{g}/\text{m}^3$ . L’ipotesi nulla (H0) affermava che la media era uguale o inferiore a  $10 \mu\text{g}/\text{m}^3$ , mentre l’ipotesi alternativa (H1) sosteneva che fosse maggiore di  $10 \mu\text{g}/\text{m}^3$ . Utilizzando un livello di significatività del 5%, il test ha generato una statistica di 11.92, indicando una notevole discrepanza tra la media campionaria e il valore di riferimento. Il p-value associato, estremamente basso (circa 1.02e-23), ci ha permesso di rifiutare con certezza l’ipotesi nulla.

Questo risultato dimostra significativamente che la media dell’esposizione a PM2.5 nel 2009 era superiore a  $10 \mu\text{g}/\text{m}^3$ , confermando la congettura di un problema ambientale nell’anno in questione. La metodologia adottata per questo test ha incluso il calcolo diretto della statistica del test e del p-value attraverso i dati di media e deviazione standard del 2009, ottenuti dalle precedenti analisi.

In conclusione, la verifica delle ipotesi fornisce una base scientifica solida per affermare che l’esposizione a PM2.5 nel 2009 superava un livello che potrebbe essere considerato preoccupante. Questa analisi sottolinea l’importanza delle misure ambientali implementate per ridurre l’esposizione a PM2.5 e la necessità di continue indagini e azioni in questo ambito.

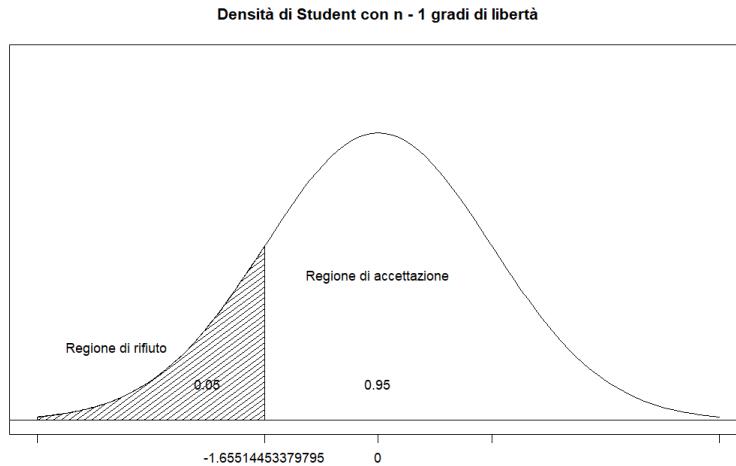


Figure 44: Densità di Student

La figura 44 mostra la distribuzione  $t$  di Student rispetto al valore critico, ottenuta tramite lo script "script/domanda5/script8.R". Nel grafico "Densità di Student con  $n-1$  gradi di libertà", la curva rappresenta la distribuzione della variabile casuale  $T$  che segue una distribuzione  $t$  di Student con i gradi di libertà del nostro campione. La regione ombreggiata a sinistra del valore critico rappresenta l'area di rifiuto per il nostro test unilaterale destro. Se la statistica del test calcolata cade in questa regione, come indicato dal valore di 11.92 ben oltre il limite sinistro del grafico, ciò fornisce una forte evidenza per rifiutare l'ipotesi nulla e accettare l'ipotesi alternativa che la media sia significativamente superiore al valore di riferimento. La parte non ombreggiata a destra rappresenta la regione di accettazione dell'ipotesi nulla.

#### 5.5.7 Conclusioni

Le analisi statistiche condotte indicano una riduzione significativa dell'esposizione media della popolazione alle PM2.5 dal 2009 al 2019. La differenza nelle medie di esposizione, sostenuta da un intervallo di confidenza al 95% e dalla verifica delle ipotesi, suggerisce che le politiche ambientali e le iniziative attuate in questo decennio hanno contribuito positivamente alla qualità dell'aria. Questi risultati sono un segnale incoraggiante dell'efficacia degli sforzi per ridurre l'inquinamento e possono fungere da incentivo per ulteriori ricerche e azioni mirate nel campo delle politiche ambientali.