

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA
SCUOLA DI ECONOMIA E STATISTICA

CORSO DI LAUREA MAGISTRALE IN
SCIENZE STATISTICHE ED ECONOMICHE



OPEN DATA SCIENCE PER IL
CAMBIAMENTO CLIMATICO GLOBALE E
IL VIRTUAL WATER TRADE

RELATORE: Prof. Antonio Candelieri
CORRELATORE: Dott. Andrea Ponti

TESI DI LAUREA DI:
Davide Luperi
MATRICOLA N. 826249

ANNO ACCADEMICO 2021/2022

Ringraziamenti

Prima di iniziare con l'esposizione della tesi vorrei esprimere i miei ringraziamenti per le persone che mi sono state vicine in questo percorso.

Innanzitutto, ringrazio il mio relatore, il Professore Antonio Candelieri, per avermi guidato durante il lavoro e per i preziosi consigli che mi hanno permesso di realizzare il mio lavoro.

Ringrazio inoltre i miei colleghi di corso con i quali ho condiviso i momenti di studio ma anche di svago durante questi anni. Il sostegno e l'affetto reciproco che ci ha accompagnato in questi anni è stato fondamentale per il raggiungimento dei miei obiettivi.

Ringrazio infine la mia famiglia per avermi sempre supportato e incoraggiato in ogni situazione. A loro va la mia più profonda e sincera gratitudine per avermi aiutato nei momenti difficili e per avermi sempre esortato a dare il meglio.

Sommario

Il lavoro ha come argomento principale la disponibilità di acqua dolce nel mondo. Questo tema viene affrontato analizzando i dati relativi alle variabili ambientali ed al virtual water trade, due tra i fattori principali che determinano la quantità di risorse idriche in ogni regione a livello globale. Le analisi statistiche che sono state implementate si basano su open data, dati disponibili al pubblico divulgati da enti ufficiali.

Il motivo che mi ha spinto ad affrontare questo tema è l'importanza che esso ricopre nella vita di tutti i giorni di milioni di persone. Inoltre mi interessava analizzare da un punto di vista spaziale le variabili che maggiormente influiscono sulla disponibilità idrica per realizzare un lavoro completo che unisse le informazioni di diversi ambiti all'interno di un tema comune.

Lo scopo della tesi è di implementare una serie di analisi di dati, sfruttando le competenze in ambito statistico e di data science, con l'obiettivo di studiare i fenomeni che regolano la presenza di acqua sul territorio a livello sia geografico sia temporale. Uno degli obiettivi principali è stato quello di fornire all'utente una serie di piattaforme in cui visualizzare le diverse analisi condotte sotto forma di grafici interattivi.

La ricerca è stata condotta utilizzando tecniche di aggregazione dei dati da un punto di vista spaziale e temporale e usando strumenti statistici quali test t di student e la regressione lineare. Inoltre è stato fatto ampio uso del pacchetto Shiny sul server R per la creazione di dashboard interattive.

I risultati comprendono sette dashboard mediante cui l'utente può visualizzare le analisi effettuate durante il lavoro. Nel dettaglio, le dashboard comprendono sia visualizzazione dei dati a livelli più o meno aggregati, sia l'applicazione e il risultato di strumenti statistici per la previsione di determinati fenomeni.

Il lavoro risulta molto utile nell'ambito dello studio delle risorse idriche sul territorio in quanto permette una visualizzazione contemporanea dei risultati ottenuti sia in ambito dello studio delle variabili ambientali, sia in ambito del virtual water trade.

Indice

Elenco delle figure	IV
Elenco delle tabelle	VI
1 Introduzione	1
1.1 Obiettivi	4
2 Presentazione dei dati	5
2.1 Dati relativi ai fattori ambientali	5
2.1.1 Temperatura	8
2.1.2 Precipitazioni	9
2.1.3 Evapotraspirazione potenziale ed Evapotraspirazione reale	10
2.1.4 Umidità del suolo	13
2.2 Dati relativi al Virtual Water Trade	15
2.2.1 Uso dell'acqua e stress	16
2.2.2 Produzione e consumo di prodotti alimentari	17
3 Analisi esplorativa	18
3.1 Dati relativi ai fattori ambientali	18
3.1.1 Distribuzione delle variabili	20
3.1.2 Relazioni tra le variabili	24
3.1.3 Aggregazione per nazione	28
3.2 Dati relativi al Virtual Water	30
3.2.1 Uso dell'acqua e stress	30
3.2.2 Produzione e consumo di prodotti alimentari	36
4 Creazione dashboard interattive e analisi statistiche	42
4.1 Mappe delle variabili	43
4.2 Mappe delle differenze rispetto la media di riferimento	44
4.3 Grafici serie storiche	46
4.4 Mappe basate sul test T di Student	49
4.5 Mappe per serie storiche per nazioni	53
4.6 Previsione variabili ambientali	55
4.7 Tabelle Virtual Water Trade	61
Riferimenti bibliografici	64

Elenco delle figure

2.1	Temperatura media 2017	8
2.2	Precipitazioni totali 2017	9
2.3	Evapotraspirazione potenziale totale 2017	11
2.4	Evapotraspirazione Reale totale 2017	12
2.5	Umidità delo suolo media 2017	13
3.1	Distribuzione della temperatura	20
3.2	Distribuzione delle precipitazioni	20
3.3	Distribuzione delle precipitazioni in scala logaritmica	21
3.4	Distribuzione dell'evapotraspirazione potenziale	22
3.5	Distribuzione dell'evapotraspirazione reale	22
3.6	Distribuzione dell'umidità del suolo	23
3.7	Correlazione tra le variabili	24
3.8	Distribuzione temperatura con focus su EP pari a 0	25
3.9	Andamento temporale della temperatura media	26
3.10	Andamento temporale dell'evapotraspirazione media	26
3.11	Scatter plot tra temperatura e Evapotraspirazione potenziale	27
3.12	Nodi non associati ad alcuna nazione	28
3.13	Nodi nel Mar Caspio	29
3.14	Stress idrico per nazione nel 2017	32
3.15	Percentuale agricoltura per nazione nel 2017	33
3.16	Percentuale industria per nazione nel 2017	34
3.17	Percentuale uso domestico per nazione nel 2017	34
3.18	Numero di Missing values per variabile nella produzione	37
3.19	Consumo pro-capite di carne bovina nel 2017	38
3.20	Consumo pro-capite di carne suina nel 2017	39
3.21	Consumo pro-capite di pollame nel 2017	39
3.22	Consumo pro-capite di latte nel 2017	40
3.23	Consumo pro-capite di uova nel 2017	40
3.24	Numero di Missing values per variabile nella produzione	41
4.1	Esempio dashboard mappe delle variabili	43
4.2	Esempio dashboard delle differenze rispetto alla media di riferimento	45
4.3	Dashboard grafici serie storiche 1	47
4.4	Dashboard grafici serie storiche 2	48

4.5	Dashboard test significatività differenza tra medie	52
4.6	Dashboard serie storiche analisi ambientali per nazione	54
4.7	Scatter plot tra Temperature ed Evap. potenziale	55
4.8	Scatter plot con modello regressione lineare	59
4.9	Dashboard Previsione con modello regressione lineare	60
4.10	Dashboard Previsione con modello regressione lineare	62

Elenco delle tabelle

3.1	Composizione dei dataset	19
3.2	Composizione dei dataset aggregati annualmente	19
3.3	Variabili ambientali per nazione per anno	29
3.4	Percentuale utilizzo di acqua	30
3.5	Stress idrico, frequenza delle osservazioni per anno	31
3.6	Percentuale utilizzo di acqua per settore	32
3.7	Utilizzo acqua, frequenza delle osservazioni per anno	33
3.8	Litri di acqua impiegati per prodotto alimentare	35
3.9	Dataset produzione prodotti alimentari	36
3.10	Top 10 nazioni per quantità prodotta - parte 1	37
3.11	Top 10 nazioni per quantità prodotta - parte 2	38
4.1	Dataset variabili ambientali per nazione	53
4.2	Coefficienti modello	58

Capitolo 1

Introduzione

L'acqua è essenziale per la vita umana e per sostenere l'agricoltura, l'industria e l'ecosistema terrestre. Tuttavia, nonostante il fatto che la Terra sia composta principalmente da acqua, solo una piccola percentuale di essa è disponibile come acqua dolce utilizzabile per scopi umani.

Inoltre, la disponibilità di acqua dolce nel mondo non è distribuita in modo uniforme, con alcune regioni che affrontano la scarsità d'acqua e altre che dispongono di quantità eccessive. La crescente domanda di acqua da parte di una popolazione globale in aumento, l'agricoltura intensiva, l'industria e il cambiamento climatico, stanno mettendo a dura prova le risorse idriche del mondo, portando a una maggiore competizione per l'acqua e a una potenziale crisi idrica in alcune parti del mondo.

La disponibilità d'acqua nel mondo è il filo conduttore del lavoro svolto e questo argomento è stato affrontato studiando due temi fondamentali: il cambiamento climatico globale e il virtual water trade.

Il tema del cambiamento climatico globale è diventato sempre più importante negli ultimi anni e rappresenta una sfida sempre più decisiva per il nostro tempo. Siccità, alluvioni, impoverimento del suolo e aumento delle temperature sono solo alcuni dei termini con cui entriamo in contatto con sempre maggiore frequenza. Il cambiamento climatico è un tema vastissimo che ricopre molte aree di studio, dalla meteorologia, alle scienze ambientali, fino alla fisica e alla statistica e lo studio di questo tema è di importanza fondamentale poiché influisce su decisioni politiche e geopolitiche di scala globale.

Con studio del cambiamento climatico si intende l'analisi dei modelli climatici passati e attuali e la proiezione delle tendenze climatiche future, nonché lo studio delle cause e delle conseguenze del cambiamento climatico e degli impatti potenziali sui sistemi umani e naturali.

Capire il rapporto causa-effetto che governa i cambiamenti dei diversi aspetti ambientali quali ad esempio la temperatura o le precipitazioni, è l'obiettivo principale degli studiosi che tramite le conoscenze e i dati a loro disposizione cercano di trovare delle risposte a dei problemi estremamente complessi.

A questo fine, un approccio molto utile che verrà presentato e approfondito in questo documento è quello di analizzare le variabili ambientali.

Per variabile ambientale si intende un qualsiasi fattore fisico, chimico o biologico che può influire sull'ambiente e sugli organismi che vi vivono. Queste variabili possono includere

temperatura, precipitazioni, umidità, luce solare, tipo di suolo e livelli di inquinamento. Possono anche includere fattori legati alle attività umane, come l'uso del suolo, la densità di popolazione e le emissioni di gas serra.

Queste variabili vengono spesso misurate e monitorate per capire come stanno cambiando nel tempo e come stanno incidendo su diversi sistemi ecologici. Sono anche utilizzate in modelli per prevedere come diversi fattori possono interagire e influire sull'ambiente in futuro.

Un secondo tema molto importante che verrà trattato, in stretta relazione con il cambiamento climatico globale, è quello del Virtual Water Trade.

Il concetto di Virtual Water Trade si riferisce al movimento o migrazione dell'acqua tra regioni del mondo sotto forma di prodotti.

Questa nozione è stata introdotta per spiegare il commercio globale di prodotti che richiedono un grande consumo di acqua durante la fase di produzione come cibo e prodotti agricoli. L'idea è che quando un paese importa prodotti, sta anche importando l'acqua utilizzata per produrli. Allo stesso modo, quando un paese esporta prodotti sta anche esportando l'acqua utilizzata per la produzione.

Si parla dunque di migrazione di acqua virtuale, ovvero di trasferimento tra aree del mondo di acqua che viene effettivamente utilizzata per la produzione di prodotti.

In questo modo il Virtual Water Trade consente ai paesi che hanno scarsità d'acqua di importare prodotti che richiedono acqua per essere prodotti, invece di utilizzare le proprie scarse risorse idriche. Inoltre, i paesi che hanno un eccesso di acqua possono esportare prodotti che richiedono acqua per produrli, il che può essere una fonte di reddito e può anche aiutare a ridurre gli sprechi d'acqua.

Il concetto di Virtual Water Trade è relativamente nuovo e si è sviluppato per aiutare i paesi e le organizzazioni a comprendere meglio le risorse idriche richieste per produrre diversi beni e servizi.

La relazione che esiste tra i due temi, lo studio delle variabili ambientali e il Virtual Water Trade, risulta di enorme importanza nella comprensione l'uno dell'altro.

Le variabili ambientali che possono essere influenzate dal Virtual Water Trade includono la disponibilità e la qualità dell'acqua, le precipitazioni e l'uso e l'umidità del suolo. Ad esempio, l'esportazione di prodotti ad alto consumo d'acqua da una regione con risorse idriche limitate può mettere sotto pressione le risorse idriche locali e può portare all'estrazione eccessiva della falda acquifera.

Viceversa, l'andamento delle variabili quali precipitazione ed evapotraspirazione¹ (verranno approfondite in seguito) determina la disponibilità di acqua sul territorio, e di conseguenza l'implementazione di politiche relative al Virtual Water Trade.

Come descritto, risulta dunque fondamentale studiare e approfondire il legame tra questi due temi per poter andare incontro alle esigenze delle società nelle varie regioni del mondo.

Al fine di seguire un approccio analitico incentrato sull'analisi dei dati, è necessario avere a disposizione i dataset contenenti le informazioni a cui si è interessati. Nello specifico è desiderabile che i dati siano il più puliti e completi possibile.

Viene quindi introdotto il tema della Open Data Science.

¹Per evapotraspirazione si intende la perdita di acqua dalla superficie del suolo a causa dell'evaporazione superficiale e della traspirazione da parte delle piante

Con Open Data Science ci si riferisce alla pratica di rendere disponibili al pubblico dati, algoritmi e modelli per promuovere la collaborazione, la trasparenza e la riproducibilità. Ciò può includere la pubblicazione di dataset, la condivisione di codici e software e la fornitura di documentazione. L'obiettivo della Open Data Science è rendere più facile per ricercatori, studiosi e il pubblico in generale l'accesso e l'uso dei dati e dei modelli.

Per Open Data si intende tutti quei dati che sono liberamente disponibili al pubblico per l'uso e l'analisi. Ciò significa che chiunque può accedere ai dati e utilizzarli per i propri scopi, senza richiedere il permesso o dover pagare. Gli open data possono provenire da una varietà di fonti, come agenzie governative, organizzazioni senza scopo di lucro e aziende private. Possono includere dati da esperimenti scientifici, dati censuari, immagini e molto altro ancora.

Per utilizzare gli Open Data, una volta che si ha accesso ai dataset, è necessario scaricarli e iniziare a lavorare con essi. E' importante controllare la licenza dei dati e i termini d'uso prima di utilizzarli, per essere sicuri di essere conformi alle leggi. E' inoltre importante tenere conto della qualità dei dati, della completezza, dell'accuratezza.

In questo lavoro vengono utilizzati Open Data relativi ai due temi principali dell'analisi, il cambiamento climatico globale e il Virtual Water Trade, derivanti da diverse fonti ed integrarli tra di loro.

La scelta delle fonti dati da utilizzare e la verifica delle informazioni contenute in essi non è da sottovalutare in quanto spesso si può incorrere in un problema comunemente noto in ambito statistico come *Garbage in, Garbage out*.

Il concetto di questa espressione è che se i dati in ingresso utilizzati per l'analisi sono di scarsa qualità, anche i risultati e le conclusioni tratte da tale analisi saranno di scarsa qualità. Sottolinea l'importanza di garantire che i dati utilizzati per l'analisi siano precisi, rilevanti e imparziali per produrre risultati significativi.

A seguito della raccolta dei dati, della loro verifica e dopo aver effettuato un controllo dell'affidabilità della fonte, è possibile svolgere il lavoro di analisi.

Questo si divide in diverse fasi, che possono essere sintetizzate in:

- Pulizia dei dati (*data cleaning*)
- Analisi esplorativa
- Scelta del miglior approccio statistico da utilizzare (a seconda dello scopo dell'analisi)
- Ottimizzazione e implementazione
- Verifica dei risultati

Lo scopo principale di un lavoro di analisi di dati è quello di comprendere, aggregare e sintetizzare i dati in input in per fornire dei risultati che siano chiari e statisticamente affidabili.

Il software impiegato per compiere le analisi è R [5], un software molto utilizzato in ambito di data science e particolarmente consigliato per gestire i dataset.

Uno dei vantaggi del software è quello di essere open source, quindi diversi utenti esperti nel settore possono implementare e sviluppare delle funzioni, aggregare in pacchetti e renderle disponibili a qualsiasi utente, il quale può scaricare il pacchetto e utilizzare le funzioni per il proprio lavoro

Per la visualizzazione di un importante mole di dati risulta essere molto utile il pacchetto *R Shiny* [2], che consiste in un insieme di funzioni che permettono la creazione di dashboard interattive per la visualizzazione dei dati.

Come sarà approfondito in seguito, un metodo di visualizzazione dei dati in maniera interattiva risulta molto utile nel momento in cui si lavora con molti dati e le informazioni contenute nei risultati sono varie e diverse tra di loro.

In questo modo, è inoltre possibile per l'utente visualizzare i risultati dell'analisi in modo indipendente a seconda delle esigenze.

1.1 Obiettivi

L'obiettivo della tesi è di effettuare un analisi approfondita di dati riguardanti le variabili ambientali e il virtual water trade, studiando le relazioni che esistono tra i dati, nonché comprendere i rapporti tra i due temi in termini di risorse idriche sul territorio.

Nel dettaglio il primo scopo è di analizzare il tema del cambiamento climatico globale e l'importanza che ha assunto negli ultimi anni nella vita quotidiana, in particolare gli effetti dei cambiamenti climatici sull'acqua disponibile sul territorio.

Questo può essere effettuato studiando le variabili ambientali come fattori che influiscono sull'ambiente e utilizzare questi dati per comprendere il rapporto causa-effetto che governa i cambiamenti dei diversi aspetti ambientali.

Le analisi sono effettuate sia a livello geografico sia a livello temporale. Le mappe create nel lavoro saranno utili per capire ed interpretare le variabili ambientali, come influiscono sul territorio e le relazioni esistenti tra di esse. Considerando le variabili da un punto di vista temporale, le serie storiche presentate sono fondamentali per comprenderne l'evoluzione nel tempo.

Inoltre, le analisi esposte sono effettuate sia a livello locale, sia a livello nazionale: aggregare i dati a livello di nazione risulta utile per stimare le risorse idriche interne di ciascuno stato e come nel futuro le risorse idriche potrebbero variare.

Per quanto riguarda il Virtual Water Trade, il fine è di approfondirne il concetto e spiegare come questo movimento di acqua virtuale tra regioni del mondo, sotto forma di prodotti, possa essere utilizzato per gestire le risorse idriche a livello globale.

Infine lo scopo più ambizioso è di analizzare la relazione tra i due temi e come le variabili ambientali possano influenzare il Virtual Water Trade, per ottimizzare le risorse idriche a livello globale e nazionale.

L'obiettivo finale è dunque quello di fornire all'utente degli strumenti interattivi mediante cui è possibile visualizzare i risultati dell'analisi in maniera indipendente e di proporre spunti di riflessione da cui possono partire ulteriori approfondimenti nei diversi ambiti.

Uno dei fini della tesi è proprio quello di fornire non solo ad esperti e studiosi, ma anche a persone non coinvolte in questo ambito di ricerca, informazioni in maniera chiara e sintetica, risultato di un lavoro di analisi del dato che comprende processi statistici e di machine learning. Informazioni che l'utente può estrarre autonomamente in base alle proprie esigenze.

L'utente ha così la possibilità di relazionarsi e interagire con il mondo della data analyst e data science pur non possedendo conoscenze statistiche avanzate.

Capitolo 2

Presentazione dei dati

In questo capitolo verranno presentati i dati utilizzati per l'analisi, nello specifico si farà riferimento alle fonti, alla composizione dei dataset ed alle informazioni contenute in essi. Dopo aver fornito una spiegazione teorica dei dati, saranno presentati dei grafici con l'obiettivo di poter meglio comprendere la struttura dei dataset e le variabili presenti.

2.1 Dati relativi ai fattori ambientali

Questo paragrafo fornisce le informazioni circa i dati relativi alle variabili ambientali.

Come premessa è doveroso informare come, in questo settore, il flusso di dati presente è molto abbondante, in quanto dati relativi a precipitazioni o temperature, per esempio, sono raccolti con elevata frequenza e da diversi istituti.

Nonostante ciò, è risultato complesso trovare una struttura dati che soddisfacesse i requisiti. In particolare, la fonte deve essere autorevole, quale un'agenzia governativa, un'università o un ente di ricerca

Inoltre i dati devono avere una profondità storica notevole: per intercettare l'andamento dei diversi fattori è necessario che lo storico dei dati a disposizione sia abbastanza profondo da consentire una corretta analisi. Oltre a ciò, è desiderabile che i dati siano completi sia da un punto di vista temporale che geo-spaziale.

Infine, dovendo sintetizzare ed aggregare diverse variabili, i diversi dataset devono poter comunicare tra di loro, ovvero deve essere possibile effettuare un'analisi che coinvolga più variabili senza perdita di informazioni.

La fonte trovata che soddisfa tutti i requisiti è il database "**Global (land) precipitation and temperature: Willmott and Matsuura, University of Delaware**" [12]

Gli autori di questo database sono Cort J. Willmott, professore presso l'università di Delaware, e Kenji Matsuura professore presso l'università di Kyoto

Come spiegato accuratamente dagli autori, il database consiste in una serie di dataset relativi a Temperatura e Precipitazioni su griglia (ogni osservazione è relativa ad un punto geografico). Nello specifico, i dataset forniscono stime puntuali della temperatura media mensile dell'aria (T , °C) e delle precipitazioni totali mensili (P , mm) sulla superficie terrestre su una griglia di punti aventi misure $0,5^\circ$ latitudine e $0,5^\circ$ longitudine.

Ogni osservazione presente nel dataset avrà dunque come valori per la latitudine e longitudine, il centro del quadrato della griglia. Considerando per esempio l'interpolazione per il quadrato della griglia che va da $20,5^{\circ}$ a 21° di latitudine e 60° e $60,5^{\circ}$ di longitudine, l'osservazione avrà valorizzati i campi latitudine e longitudine con i valori $21,75^{\circ}$ e $60,25^{\circ}$ che corrisponde alle coordinate del centro del quadrato in questione di griglia (chiamato nodo).

Il professore Kenji Matsuura precisa come vengono creati i dataset, ovvero quali sono fonti e come vengono interpolati i dati:

Le stime a griglia si basano principalmente sui record delle stazioni di osservazione (centraline meteo) che sono stati compilati, per la maggior parte, da diverse fonti pubblicamente disponibili come il set di dati del Global Historical Climatology Network (GHCN2) (Peterson e Vose, 1997), il Global Historical Climatology Network Monthly (GHCNM) set di dati versione 3 (GHCN3) Lawrimore et al., 2011), archivio Daily Global Historical Climatology Network (GHCN-Daily) (Menne et al., 2012) e Global Surface Summary of Day (GSOD)

Per interpolare i record delle stazioni a ciascun nodo della griglia, è stato utilizzato L'algoritmo di interpolazione spaziale di Shepard (1968), modificato per l'uso sulla superficie quasi sferica della Terra (Willmott et al., 1985). Sono state inoltre utilizzate ulteriori informazioni quali un modello digitale di elevazione (DEM), altre informazioni climatologiche e un tasso di decadenza atmosferica medio vicino alla superficie per la temperatura (6.0°C/Km)

Per ulteriori approfondimenti si rimanda al lavoro completo. [16]

Inoltre, viene precisato come ogni osservazione su cui abbiamo i dati, corrispondente ad un punto sulla mappa, non rappresenta una stazione o centralina meteorologica effettivamente esistente in quel preciso luogo, in cui si sono rilevati i valori, ma si tratta di un nodo della griglia creato ad hoc i cui valori di riferimento per le diverse variabili sono stimati con i diversi modelli.

Dunque, i dati reali provenienti dalle centraline, sono stati precedentemente aggregati attraverso dei modelli, descritti in seguito, e forniti all'utente già pronti per l'utilizzo.

I punti di forza del lavoro svolto dai due professori, come loro stessi tengono a precisare, sono i seguenti:

- Ciascuno dei valori mensili su griglia di precipitazione e temperatura è una stima puntuale locale con una risoluzione di $0,5$ gradi di longitudine-latitudine. Ciò implica che i valori all'interno della griglia, non variano con la latitudine.
- Le mappe globali si basano su una proiezione cartografica di uguale area. Ciò consente confronti significativi di una variabile climatica tra le aree.
- Le interpolazioni spaziali da cui derivano i valori sono tutte a base sferica e quindi non contengono bias di proiezione cartografica. Gli interpolatori sono stati migliorati nel corso degli anni con ulteriori variabili indipendenti (oltre a latitudine e longitudine) come elevazione (ad esempio, Willmott e Matsuura, 1995) e campi climatologici (ad esempio, Willmott et al., 1995).

Le limitazioni sono invece:

- La dipendenza delle interpolazioni dalla copertura spaziale delle reti di registrazione delle stazioni .

- I dati sulle precipitazioni derivano dall'osservazione delle precipitazioni "grezze".

I dataset estratti dai database di Willmott e Matsuura contengono informazioni circa diverse variabili ambientali:

1. Temperatura.
2. Precipitazioni.
3. Evapotraspirazione Potenziale.
4. Evapotraspirazione Reale.
5. Umidità del suolo.

Per ogni variabile è stato utilizzato un database contenente 118 dataset, in cui ogni dataset racchiude le informazioni di un anno di storico, i quali vanno dal 1900 al 2017.

In ogni dataset sono presenti i valori mensili per l'anno considerato.

Essendo dati artificiali, creati dunque mediante l'utilizzo di modelli statistici aggregando diverse fonti, non sono presenti valori mancanti.

2.1.1 Temperatura

I dati sono stati presi dal sito riportato nella bibliografia [14]

Per creare il database riguardante le temperature, come descritto precedentemente, i professori Cort J. Willmott e Kenji Matsuura, hanno utilizzato diverse fonti di dati (diverse rilevazioni atmosferica operate da centraline) che sono state aggregate ed usate per una interpolazione spaziale.

Nello specifico, i dati di partenza sono stati puliti eliminando medi con elevata quantità di missing values ed eseguendo un controllo sugli outliers per poi passare alla fase di interpolazione, in cui, utilizzando diversi modelli, è stato possibile estrarre l'informazione circa la temperatura centrata del nodo della griglia.

Per ulteriori approfondimenti si rimanda alla pagina web citata.

Come detto in precedenza, ogni database relativo ad una variabile si compone di 118 dataset, uno per ogni anno dal 1900 al 2017.

Ogni dataset riguardante la variabile delle temperature si compone di 85794 righe e 15 colonne. Ogni riga corrisponde ad un nodo nella griglia.

Le colonne fanno riferimento alle seguenti misure:

- Le prime due colonne rappresentano le variabili relative alla *longitudine* e alla *latitudine* del nodo
- Le dodici colonne dalla terza alla quattordicesima contengono i valori per ogni mese da gennaio a dicembre
- L'ultima colonna corrisponde alla media della temperatura annuale in quel nodo, ovvero alla media delle osservazioni mensili contenute nelle colonne 3-14.

Di seguito viene mostrata una mappa in cui viene mostrata la temperatura media per l'anno 2017 per ogni nodo in gradi centigradi ($^{\circ}\text{C}$)

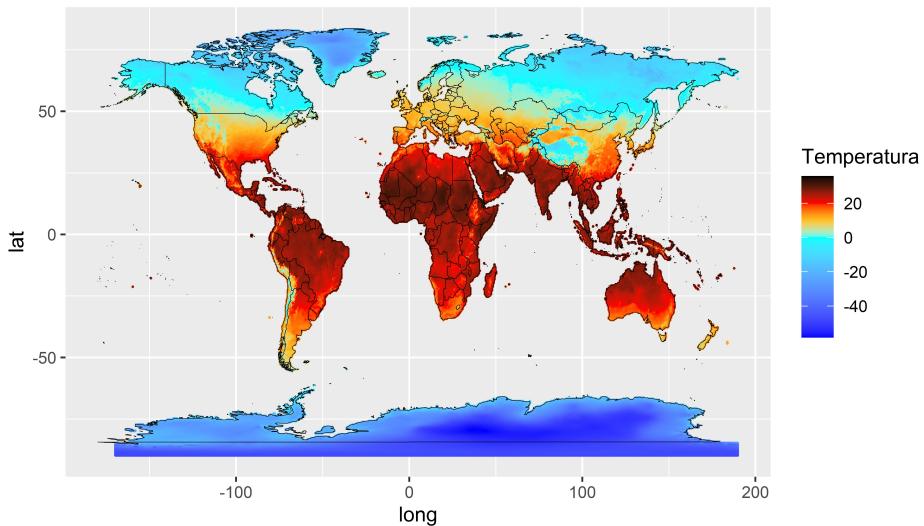


Figura 2.1: Temperatura media 2017

2.1.2 Precipitazioni

I dati provengono dal database presente in bibliografia [15]

Il database relativo alle precipitazioni è stato creato in maniera analoga a quello delle temperature: sono state utilizzate diverse fonti di dati i quali sono stati successivamente uniti mediante una interpolazione spaziale.

Anche in questo caso il procedimento ha riguardato la pulizia dei dati in input dai valori non considerati realistici (dovuti anche ad errori strumentali) e dai valori mancanti.

Nella pagina web citata è presente una descrizione dettagliata dell'intero processo.

Il database relativo alle precipitazioni è composto da 118 dataset, con profondità temporale dal 1900 al 2017

Ogni dataset è costituito da 85794 righe e 15 colonne ed ogni riga equivale ad un nodo nella griglia.

Le colonne racchiudono le seguenti informazioni:

- Le prime due colonne coincidono con le variabili relative alla *longitudine* e alla *latitudine* del nodo
- Le dodici colonne dalla terza alla quattordicesima contengono i valori per ogni mese da gennaio a dicembre
- L'ultima colonna corrisponde alla somma delle precipitazioni annuali in quel nodo, ovvero alla somma delle osservazioni mensili contenute nelle colonne 3-14.

Di seguito viene mostrata una mappa in cui vengono mostrate le precipitazioni totali per l'anno 2017 per ogni nodo in millimetri (mm)

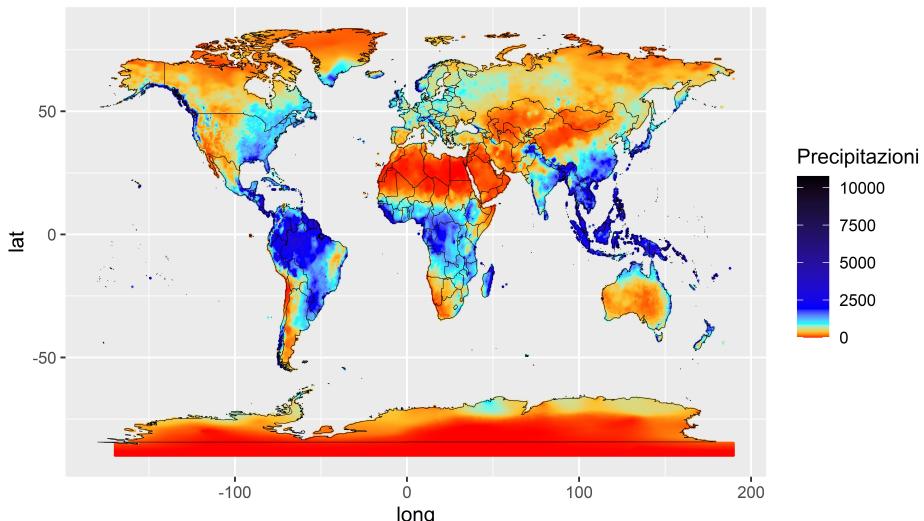


Figura 2.2: Precipitazioni totali 2017

2.1.3 Evapotraspirazione potenziale ed Evapotraspirazione reale

Viene in questo paragrafo introdotta una misura poco conosciuta e che merita un approfondimento, ovvero quella dell'evapotraspirazione.

L'evapotraspirazione è il processo attraverso cui l'acqua viene trasferita dalla superficie terrestre all'atmosfera attraverso gli effetti combinati dell'evaporazione dalla superficie dei corpi d'acqua e delle piante, e della traspirazione delle piante.

Questa variabile risulta molto importante nello studio del ciclo d'acqua in quanto misura la quantità d'acqua che una regione perde per cause naturali.

I più importanti fattori che influiscono l'intensità dell'evapotraspirazione [11] sono:

- Temperatura: temperature più calde aumentano l'energia disponibile per l'evaporazione e la traspirazione
- Radiazione solare: più radiazione solare significa più energia per l'evaporazione e la traspirazione
- Vento: la velocità del vento può influire sull'evapotraspirazione trasportando il vapore d'acqua lontano dalla superficie della Terra
- Umidità: maggiore umidità significa che c'è più vapore d'acqua disponibile per l'evaporazione e la traspirazione

Inoltre possono essere incidere anche variabili quali il tipo di vegetazione, l'uso del suolo e l'altitudine.

Per quantificare l'evapotraspirazione diventa necessario distinguere due casi in cui è possibile misurarla:

1. Nel primo caso, l'evapotraspirazione viene misurata come quantità totale di acqua che potrebbe potenzialmente essere estratta date variabili quali temperatura, vento ecc.; si parla in questo caso di *Evapotraspirazione Potenziale*.
2. Nel secondo caso, l'evapotraspirazione viene misurata come quantità totale di acqua che effettivamente viene estratta dalla superficie: si parla in questo caso di *Evapotraspirazione Reale*.

Nel dettaglio, come spiegato dal NOAA (National Centers for Environmental Information) [3], L'evapotraspirazione potenziale è la domanda o la quantità massima di acqua che verrebbe evapotraspirata se fosse disponibile acqua a sufficienza (dalle precipitazioni e dall'umidità del suolo). L'evapotraspirazione reale è la quantità di acqua effettivamente evapotraspirata ed è limitata dalla quantità di acqua disponibile sulla superficie. L'evapotraspirazione reale è sempre minore dell'evapotraspirazione potenziale.

Willmott e Matsuura hanno creato diversi database relativi a queste due misure cambiando in ognuno il parametro di riferimento per la Capacità di ritenzione idrica del suolo, misurata in mm. Questo parametro entra nel calcolo dell'evapotraspirazione e dell'umidità del suolo e può essere visto come la massima quantità d'acqua che il suolo può trattenere.

In questo lavoro vengono utilizzati i database in cui il parametro assume valore pari a 150mm [17], in quanto sono i più completi in termini di profondità temporale

I database [18] sono composti ognuno da 118 dataset, con profondità temporale dal 1900 al 2017. Ogni dataset è costituito da 85794 righe e 14 colonne ed ogni riga equivale ad un nodo nella griglia.

- Le prime due colonne coincidono con le variabili relative alla *longitudine* e alla *latitudine* del nodo
- Le dodici colonne dalla terza alla quattordicesima contengono i valori per ogni mese da gennaio a dicembre

Una quindicesima colonna è stata aggiunta manualmente come somma delle colonne 3-14 in modo da rappresentare la somma dell'evapotraspirazione (potenziale o reale) durante l'anno.

Di seguito vengono riportati i grafici per l'anno 2017, in cui viene mostrata la somma dell'evapotraspirazione potenziale e reale nei nodi della griglia.

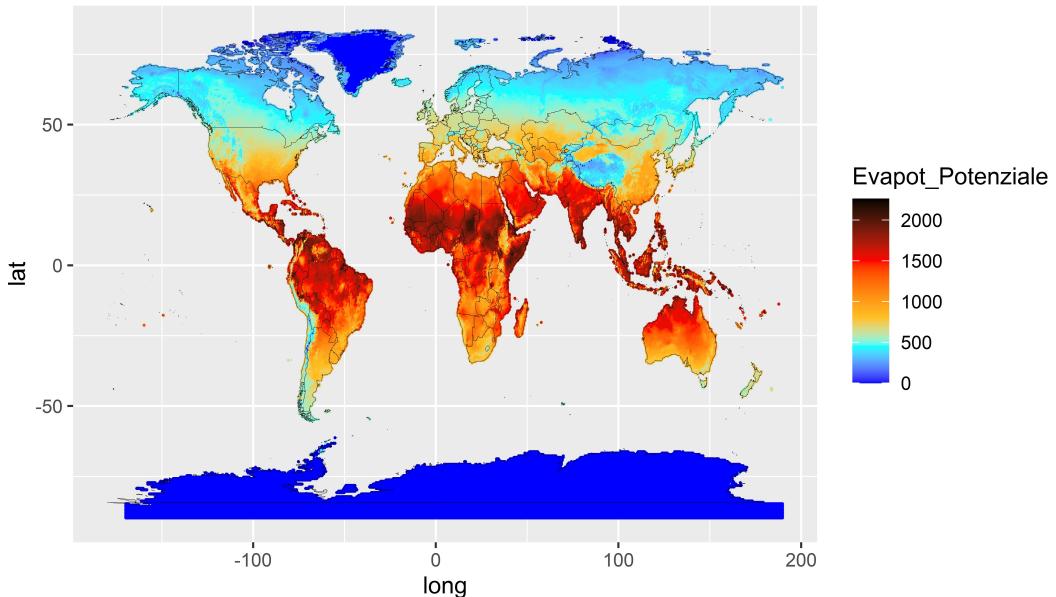


Figura 2.3: Evapotraspirazione potenziale totale 2017

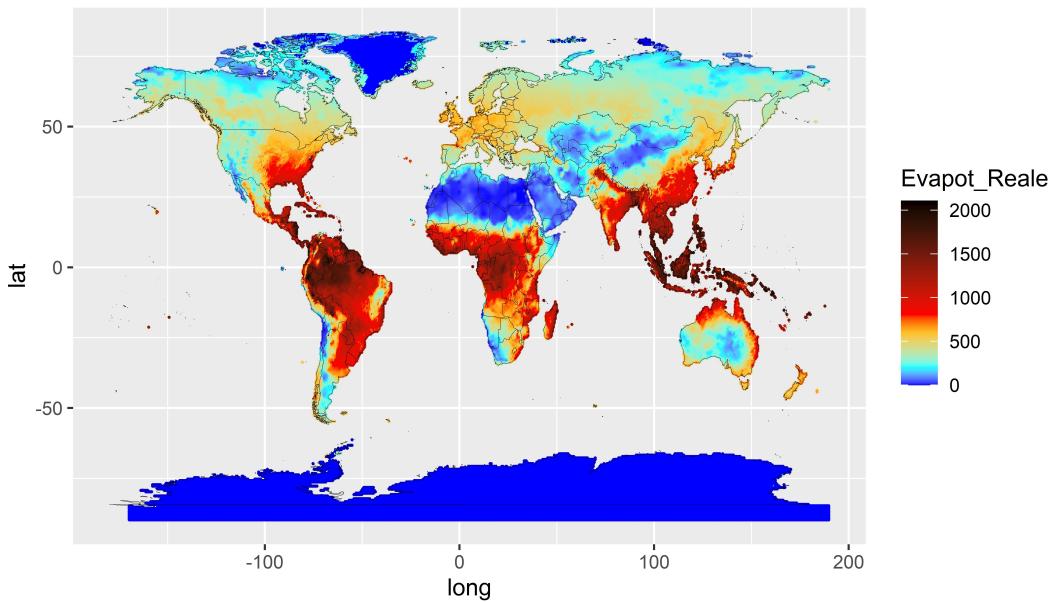


Figura 2.4: Evapotraspirazione Reale totale 2017

Da una prima analisi visiva delle mappe è possibile notare quanto è stato detto nella presentazione delle due variabili.

L'evapotraspirazione potenziale risulta più alta intorno all'equatore (dove le temperature sono più alte e l'irraggiamento solare è maggiore) e diminuisce in proporzione alla diminuzione della distanza dai poli, in cui è praticamente nulla.

Si può inoltre constatare come questa mappa risulti molto simile alla mappa create per le temperature medie dello stesso anno. Risulta quindi già ipotizzabile una forte correlazione tra l'evapotraspirazione potenziale e la temperatura, che confermerebbe come la temperatura sia una delle variabili più importanti che condizionano l'evapotraspirazione potenziale.

Osservando la mappa che mostra l'evapotraspirazione reale totale nel 2017, risulta evidente come quest'ultima sia fortemente influenzata dalla mappa circa le precipitazioni: essendo l'evapotraspirazione reale l'acqua effettivamente evapotraspirata dalla terra, non può mai essere superiore alle precipitazioni.

Dunque, in zone in cui l'evapotraspirazione potenziale è alta ma le precipitazioni sono scarse, come per esempio nel deserto del Sahara, l'evapotraspirazione reale è limitata dalla quantità d'acqua effettivamente disponibile e risulta essere molto minore dell'evapotraspirazione potenziale.

Non a caso, i valori più alti per questa variabile si sono registrati, nell'anno 2017, nelle zone più piovose del pianeta

2.1.4 Umidità del suolo

L'ultima informazione che è stata estratta dai database di Kenji Matsuura e Cort J. Willmott è l'umidità del suolo.[18]

Questa variabile non verrà utilizzata nelle successive analisi approfondite, ma risulta molto utile per studiare la quantità di acqua presente nel terreno

Come è stato detto nel precedente paragrafo, questi valori sono stati stimati valorizzando il parametro "Capacità di ritenzione idrica del suolo" con il valore 150 mm. Ciò significa che un nodo avrà valore 150 quando l'umidità del suolo stimata è massima, mentre avrà valori prossimi allo zero in condizioni di terreno estremamente secco.

Il database circa l'umidità del suolo comprende 118 dataset, che coprono gli anni dal 1900 al 2017

Ogni dataset ha 85794 righe e 14 colonne ed ogni riga equivale ad un nodo nella griglia.

Le colonne sono così strutturate:

- Le prime due colonne equivalgono alle variabili circa la *longitudine* e alla *latitudine* del nodo.
- Le dodici colonne dalla terza alla quattordicesima includono i valori per ogni mese da gennaio a dicembre.

Anche in questo caso è stata aggiunta una quindicesima colonna come media delle colonne 3-14 in modo da rappresentare la media dell'umidità del suolo durante l'anno.

La seguente mappa presenta l'umidità del suolo media nel 2017

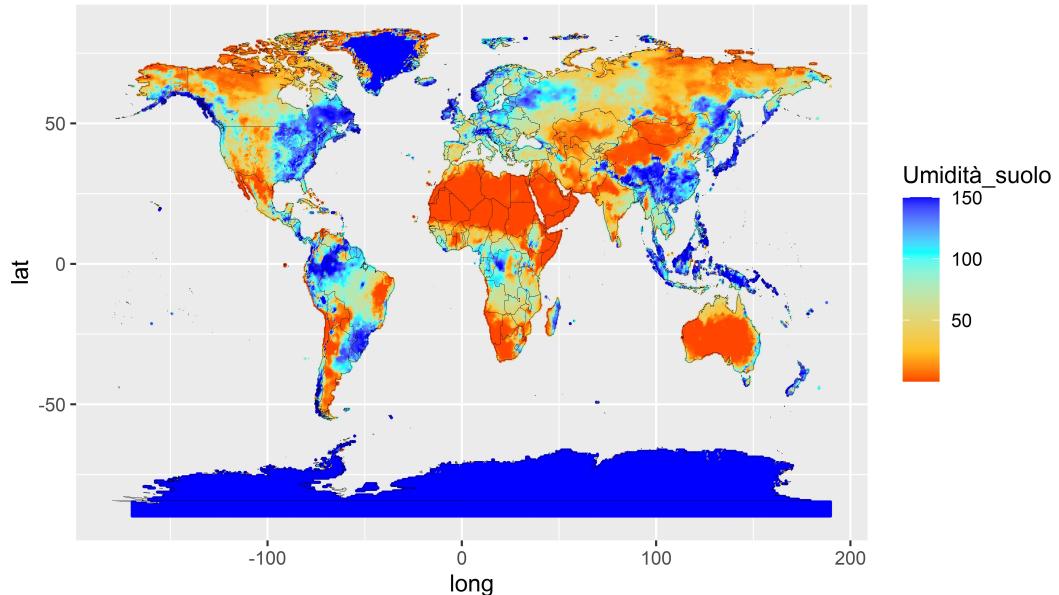


Figura 2.5: Umidità del suolo media 2017

L'utilità di questa variabile risiede nel sintetizzare le altre variabili: a terreni secchi (umidità bassa) corrispondono precipitazioni basse ed evapotraspirazione alta, viceversa terreni umidi coincidono con precipitazioni elevate e bassi livelli di evapotraspirazione.

Vengono presentati degli esempi per aiutare a comprendere alcuni pattern specifici:

- Come primo caso viene fatto un focus sul deserto del Sahara, come effettuato in precedenza: qui l'umidità del suolo è bassa, come ci si aspetterebbe, poiché le precipitazioni sono estremamente rare. Inoltre l'evapotraspirazione potenziale molto alta (molto più alta in termini numerici rispetto alle precipitazioni) fa sì che l'evapotraspirazione reale sia pressoché uguale alle precipitazioni. In altre parole, a causa delle alte temperature e dell'irradiazione solare nel deserto, l'evapotraspirazione potenziale è così elevata da far evapotraspirare la totalità dell'acqua che piove. La conseguenza è una desertificazione del terreno, traducibile un'umidità del suolo molto bassa.
- Andando ad osservare le aree in prossimità dell'equatore è possibile notare come queste abbiano un'umidità del suolo molto alta. Questo nonostante l'evapotraspirazione potenziale sia elevata. Il motivo è che in queste aree (in particolare nella foresta amazzonica, e nel sud-est asiatico) le precipitazioni superano di molto l'evapotraspirazione potenziale. L'evapotraspirazione reale è quindi decisamente vicina all'evapotraspirazione potenziale (essendoci disponibilità d'acqua fornita dalla precipitazioni) ma nonostante questo, le abbondanti precipitazioni permettono di avere un deficit idrico molto positivo. Ciò implica che una buona quantità d'acqua rimane nel suolo, rendendo il terreno umido.
- Esistono infine molti casi che non sono considerati estremi come i due punti precedentemente descritti. L'Europa presenta ad esempio dei valori nella media sia per le precipitazioni che per l'evapotraspirazione, di conseguenza l'umidità del suolo rientra nei valori medi.

N.B. *Si vuole precisare come, in questa prima fase di analisi preliminare, si utilizzino delle mappe i cui valori fanno riferimento al solo anno 2017. Ciò è possibile in quanto i valori delle variabili non subiscono grosse variazioni tra i diversi anni, quindi quello che è possibile dedurre in questa fase può essere generalizzato per tutto l'orizzonte temporale. Si puntualizza inoltre che si tratta di una prima visualizzazione dei dati avente il solo fine di migliorarne l'interpretabilità, spiegando tramite mappe dei concetti più astratti.*

2.2 Dati relativi al Virtual Water Trade

In questo paragrafo verranno descritti ed illustrati i dati utilizzati nel lavoro riguardanti il fenomeno del Virtual Water Trade.

Il termine Virtual Water (acqua virtuale) si riferisce all'acqua utilizzata nei processi di produzione di merce agricola o industriale. Per Virtual Water Trade si intende il commercio di acqua virtuale, ovvero il commercio dell'acqua impiegata nei processi di produzione del prodotto tramite appunto il commercio dello stesso.

La necessità di analizzare fenomeno deriva da una volontà di analizzare i flussi d'acqua nelle varie zone del pianeta, considerando non solo gli aspetti naturali, ma anche quelli commerciali.

L'obiettivo è quello di quantificare i flussi del Virtual Water Trade tra nazioni, ovvero studiare il traffico dell'acqua virtuale tra i diversi paesi del mondo.

In particolare è fondamentale focalizzarsi su quei paesi che già soffrono di scarsità d'acqua per cause naturali per studiare come questi si comportano nell'ambito in oggetto.

In particolare, se un paese esporta un prodotto ad alta intensità d'acqua (ovvero per il quale è richiesta una grossa quantità d'acqua per produrlo) in un altro paese, esporta acqua in forma virtuale.

In questo modo alcuni paesi supportano altri paesi nei loro bisogni idrici. Per i paesi con scarsità d'acqua potrebbe essere interessante raggiungere la sicurezza idrica importando prodotti ad alta intensità idrica invece di produrli internamente utilizzando le proprie risorse idriche. In modo opposto, i paesi ricchi d'acqua potrebbero trarre profitto dalla loro abbondanza di risorse idriche producendo prodotti ad alta intensità d'acqua per l'esportazione.

Il commercio virtuale di acqua tra nazioni e persino continenti potrebbe quindi essere utilizzato come uno strumento per migliorare l'efficienza dell'uso globale dell'acqua e raggiungere la sicurezza idrica nelle regioni povere d'acqua del mondo [4].

Per questo argomento, la fonte dei dati che verranno utilizzati nell'analisi è il sito www.ourworldindata.org

OurWorldinData è un sito Web che fornisce una panoramica accessibile e completa dei dati sullo sviluppo globale. Il sito Web è gestito dall'Università di Oxford e si basa su una ricerca condotta dal Global Change Data Lab.

Il sito web copre una vasta gamma di argomenti relativi allo sviluppo globale, tra cui la salute, l'istruzione, la povertà, la diseguaglianza e l'ambiente. Fornisce visualizzazioni di dati e strumenti interattivi che semplificano l'esplorazione e la comprensione dei dati. Il sito Web include anche un blog in cui ricercatori ed esperti condividono le loro opinioni sui dati e sulle loro implicazioni per lo sviluppo globale.

I dati del sito Web provengono da fonti affidabili come la Banca mondiale, le Nazioni Unite e altre organizzazioni internazionali. I dati vengono inoltre regolarmente aggiornati per riflettere le informazioni più recenti disponibili.

Our World in Data mira a rendere i dati sullo sviluppo globale più accessibili e comprensibili per un vasto pubblico, inclusi responsabili politici, ricercatori e il pubblico in generale. Il sito Web è una risorsa preziosa per chiunque sia interessato a comprendere lo stato del mondo e i progressi compiuti verso gli obiettivi di sviluppo globale.

Si tratta dunque di una fonte molto affidabile che può essere utilizzata per analisi analoghe a quella in questione.

Nello specifico, saranno impiegati due tipi di dati:

Nel primo caso, i dati estratti dalla fonte sono relativi al consumo dell'acqua, alla capacità ed allo stress acquifero a cui le nazioni sono sottoposte

Nel secondo caso i dati utilizzati sono relativi al consumo ed alla produzione di prodotti ad alta intensità d'acqua.

2.2.1 Uso dell'acqua e stress

In questo paragrafo verranno presentati i dati pertinenti alla disponibilità ed al consumo d'acqua nelle diverse parti del mondo presenti nel database [7]

Da questa pagina sono stati estratti cinque dataset, relativi a diversi temi:

- Il primo dataset ha come oggetto la percentuale di prelievi (consumo) d'acqua per ogni nazione. La percentuale è calcolata come prelievi di acqua dolce sul totale delle risorse interne. Da questo dataset si possono ricavare le informazioni circa le nazioni per cui è più alto lo stress idrico.

Il dataset ha profondità temporale dal 1962 al 2017, con frequenza quinquennale: per ogni nazione ci saranno dunque i dati relativi alla percentuale di acqua utilizzata in quell'anno, per ogni anno compreso nel dataset (1962, 1967 ..).

Nei primi anni persistono molti valori mancanti, in quanto era difficile andare a stimare questa misura nelle nazioni meno avanzate. Negli ultimi anni il dataset risulta completo praticamente per ogni nazione.

- In altri tre dataset è presente la percentuale di utilizzo dell'acqua a seconda del tipo di utilizzo: nel primo dataset viene indicata la percentuale di acqua disponibile impiegata nel settore agricolo; nel secondo dataset le percentuali fanno riferimento all'acqua impiegata nei processi industriali; nel terzo dataset si visualizza la quota d'acqua impiegata per famiglie e servizi pubblici.

Anche in questo dataset i dati sono disponibili dal 1967 al 2017 con frequenza quinquennale e presentano molti valori mancanti negli anni meno recenti.

Si precisa come la somma delle percentuali sopra presentate per ogni nazione e per ogni anno di storico abbia come risultato 100%.

- L'ultimo dataset, che verrà utilizzato successivamente nelle analisi, riporta il consumo di acqua adoperata per produrre un chilogrammo di prodotto alimentare.

Questo dataset è molto utile poiché permette di avere, per ogni prodotto agro alimentare, una stima precisa dell'acqua utilizzata per produrre un kg di prodotto. Integrando questa informazione con la produzione dei prodotti in termini quantitativi nelle diverse nazioni, si può ottenere la quantità d'acqua effettivamente impiegata in quel determinato campo

2.2.2 Produzione e consumo di prodotti alimentari

In questa sezione sono presentati i dati inerenti la produzione e il consumo di carne, uova e latte.

Lo scopo finale di analizzare questi dati è quello di determinare, per ogni nazione, il consumo e la produzione di determinati alimenti e di conseguenza calcolare l'export e l'import dei vari prodotti.

La fonte dati utilizzata è il sito web ourworldindata.org, nello specifico le pagine "Meat and Dairy production" [6] e "World Population Growth" [8]

Questo secondo dataset contiene informazioni circa la popolazione globale per ogni nazione, per ogni anno dal 1950. Questa informazione sarà necessaria successivamente in quanto, avendo dati pro capite, per avere il totale per nazione serve utilizzare il dato circa la popolazione della nazione in quell'anno.

Nella prima pagina è invece possibile ottenere i dati riguardanti il consumo e la produzione di determinati alimenti. In questo modo, aggregando questa informazione con quella descritta del paragrafo precedente riguardo la quantità di acqua che è necessaria per produrre gli alimenti, è possibile stimare l'export e l'import di acqua virtuale, ovvero capire gli effetti del Virtual Water Trade sulla nazione in esame.

I dataset utilizzati sono diversi e contengono informazioni circa la produzione e il consumo pro capite per ogni nazione e per ognuno dei seguenti prodotti:

- Carne (con differenziazione tra pollame, bovini e ovini)
- Latte
- Uova

Anche in questo caso, le osservazioni meno recenti sono quelle maggiormente prive di dati.

Questo non è un problema in quanto l'analisi che sarà fatta nei capitoli successivi non riguarda un trend storico, ma tende più al voler osservare ad oggi come si sta comportando una nazione relativamente al Virtual Water Trade.

Verranno dunque impiegati i dati più recenti per i quali si dispone della quasi totalità delle informazioni.

Capitolo 3

Analisi esplorativa

In questo capitolo verrà presentata l'analisi esplorativa dei dati a disposizione.

L'analisi esplorativa dei dati (EDA) è un passo importante in un lavoro di analisi dati, in quanto consente di ottenere una migliore comprensione dei propri dati. Visualizzando e riepilogando i dati, è possibile identificare modelli, tendenze e valori anomali che potrebbero non essere immediatamente evidenti osservando i dati grezzi.

Questo può portare a ulteriori ricerche e test di ipotesi, oltre ad aiutare a identificare eventuali errori o problemi con il processo di raccolta dei dati. Inoltre, l'EDA può anche aiutare a identificare eventuali presupposti o distorsioni nei dati, il che è importante per garantire la validità di qualsiasi conclusione tratta dai dati.

Nel complesso, l'EDA è un passaggio cruciale che aiuta a comprendere meglio i dati e a prendere decisioni più coerenti nei successivi sviluppi.

3.1 Dati relativi ai fattori ambientali

In questo paragrafo verrà presentata l'analisi esplorativa dei dati relativi ai fattori ambientali presentati nel capitolo 2.

Prima di visualizzare i grafici è opportuno studiare la composizione dei dataset per capire come sono strutturati e come possono essere aggregati.

Come è stato detto, i dati per ognuna delle cinque variabili ambientali sono stivati in un database, ognuno dei quali ha al suo interno 118 dataset, uno per ogni anno dal 1900 al 2017.

Tutti i dataset hanno la stessa struttura all'interno del database, ovvero le prime due colonne rappresentano le coordinate del nodo, le altre colonne fanno riferimento al valore della variabile nel mese, mentre l'ultima colonna contiene le informazioni circa la media o la somma nell'anno della variabile in quel nodo (somma o media dei 12 mesi).

I dataset hanno dunque la seguente struttura:

Poiché quello che interessa analizzare è il trend annuale dei fattori, per ogni dataset di ogni database (dunque per ogni variabile) sono state estratte solo le prime due colonne più l'ultima: in questo modo, per ogni nodo sono state prese solo le informazioni annuali della variabile desiderata.

Longitudine	Latitudine	Gen	Feb	...	Dic	Somma/Media
92.25	20.75	<i>val1</i>	<i>val2</i>	...	<i>val12</i>	somma/media (<i>val1</i> , <i>val12</i>)
-11.75	-89.75			...		
-40.75	75.25			...		

Tabella 3.1: Composizione dei dataset

Per ogni variabile, i dati sono stati poi uniti in un unico dataset, in cui le righe corrispondono sempre ai nodi della griglia, mentre le colonne corrispondono agli anni dal 1900 al 2017, oltre alle due colonne relative alla localizzazione geografica del nodo.

Ogni dataset creato avrà dunque 85794 righe e 120 colonne, le prime due che attinente le coordinate (longitudine e latitudine) e le altre 118 relative agli anni dal 1900 al 2017: per ogni nodo sono state dunque tenute i soli valori annuali della variabile.

I dataset creati hanno perciò la seguente struttura:

Longitudine	Latitudine	1900	1901	...	2017
92.25	20.75	<i>val1900</i>	<i>val1901</i>	...	<i>val2017</i>
-11.75	-89.75			...	
-40.75	75.25			...	

Tabella 3.2: Composizione dei dataset aggregati annualmente

I dataset creati come nella tabella 3.2, saranno dunque cinque, uno per ogni variabile ambientale

La fase successiva ha il focus sulla distribuzione delle cinque variabili. Verranno presi in considerazioni i valori per ogni anno per ogni nodo, in modo da visualizzare come sono distribuite le variabili.

3.1.1 Distribuzione delle variabili

Vengono ora mostrati dei grafici a barra (istogrammi) per comprendere la distribuzione dei valori di ognuna delle cinque variabili ambientali

- Temperatura

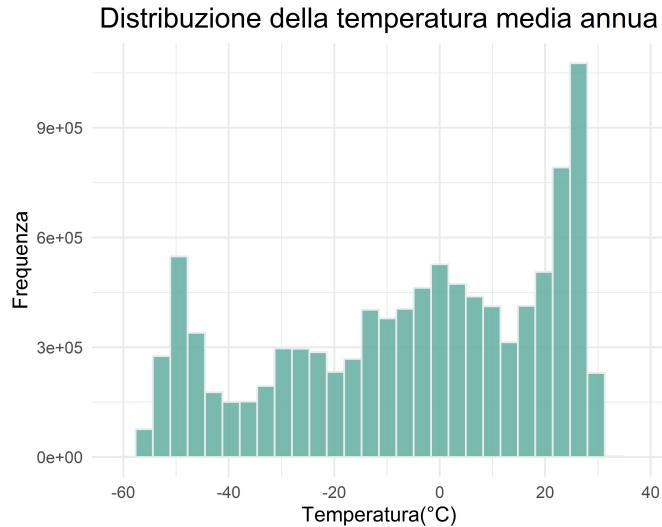


Figura 3.1: Distribuzione della temperatura

Non si notano particolari criticità per la distribuzione di questa variabile: non sono presenti valori estremi né pattern specifici da analizzare.

- Precipitazioni

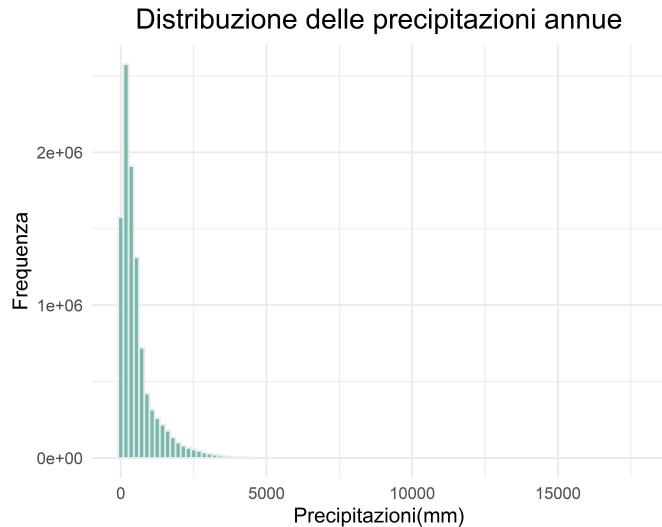


Figura 3.2: Distribuzione delle precipitazioni

Andando ad osservare la distribuzione delle precipitazioni, notiamo come sia evidente una coda a destra della distribuzione. Ciò è dovuto al fatto che questa variabile presenta molti valori concentrati in un intervallo che va da 0 e 5000 mm, mentre esistono alcuni valori che invece sono molto alti.

Questi non sono però da considerarsi outlier: è normale avere alcuni nodi in riferimento a delle specifiche coordinate che presentano valori alti delle precipitazioni. Questi valori non devono assolutamente essere esclusi, risulteranno invece fondamentali nello studio delle precipitazioni totali delle nazioni.

Per avere una visione più chiara dell'andamento delle precipitazioni è possibile visualizzare la distribuzione della variabile in scala logaritmica.

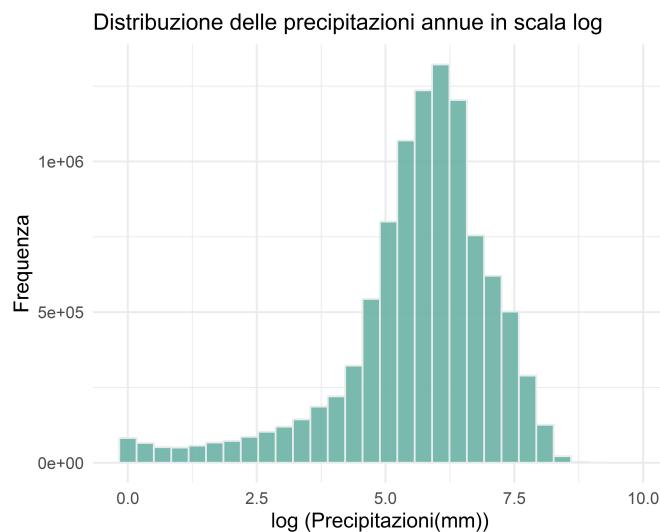


Figura 3.3: Distribuzione delle precipitazioni in scala logaritmica

Come è possibile osservare, la distribuzione risulta molto più uniforme e soprattutto non compaiono valori particolarmente alti,

- Evapotraspirazione potenziale

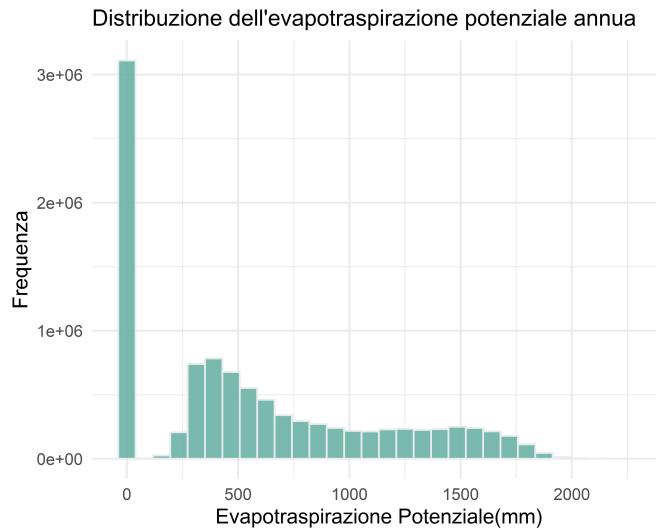


Figura 3.4: Distribuzione dell'evapotraspirazione potenziale

Si nota in questo caso un numero elevato di osservazioni con valore pari a 0. Si tratta di valori rilevati in posti specifici quali Antartica o Groenlandia. Verrà approfondita in seguito la natura di queste osservazioni

- Evapotraspirazione reale

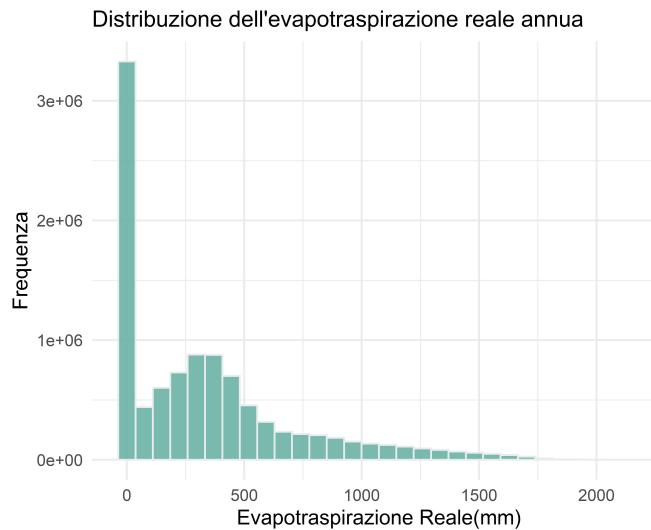


Figura 3.5: Distribuzione dell'evapotraspirazione reale

Anche per questa variabile sono presenti molto valori vicini allo zero, logica conseguenza dei valori vicini allo zero dell'evapotraspirazione potenziale

- Umidità del suolo

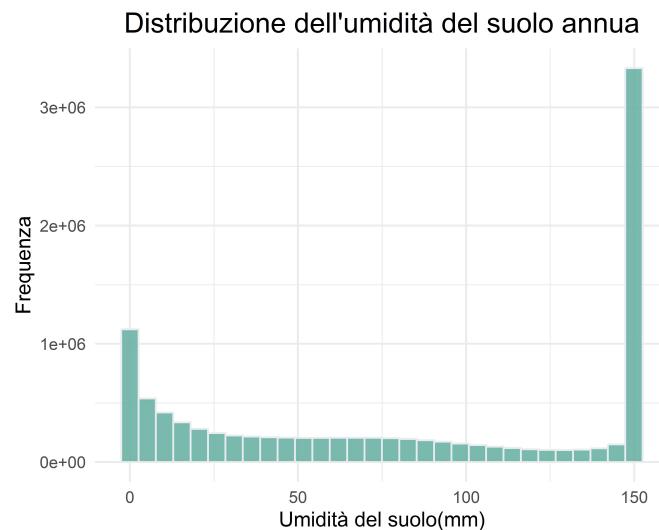


Figura 3.6: Distribuzione dell'umidità del suolo

Per questa variabile troviamo una distribuzione che va approfondita.

Esistono molti valori che sono vicini al valore massimo della variabile (150): è stato velocemente verificate che queste osservazioni sono le stesse che hanno evapotraspirazione pari a 0. Non trattandosi una variabile che verrà successivamente utilizzata, ma che viene sono mostrata per meglio interpretare i dati, non è importante approfondire questi specifici dati

Viceversa, il picco di osservazioni in prossimità del limite inferiore dovrebbe coincidere con i nodi posizionati in zone aride del pianeta.

3.1.2 Relazioni tra le variabili

In ottica di approfondire le tematiche aperte nel paragrafo precedente, in questa fase vengono analizzate le relazioni che esistono tra le variabili al fine di capire come queste si comportino.

Come primo punto viene presentato un grafico riportante la correlazione tra le variabili:

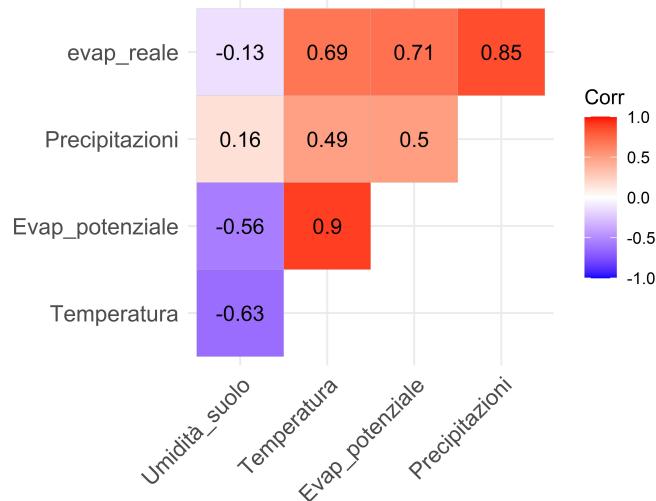


Figura 3.7: Correlazione tra le variabili

La formula utilizzata per calcolare il coefficiente di correlazione (noto anche come coefficiente di Pearson è:

$$Corr(x, y) = Cov(x, y) / (std(x) * std(y))$$

Dove:

- x e y sono due variabili.
- $Cov(x, y)$ è la covarianza tra x e y .
- $std(x)$ e $std(y)$ sono le deviazioni standard di x e y .

Il coefficiente di correlazione va da -1 a 1. Un valore di 1 indica una perfetta correlazione positiva, un valore di -1 indica una perfetta correlazione negativa e un valore di 0 indica nessuna correlazione.

Si può notare quanto dedotto nei capitoli precedenti: il fenomeno dell'evapotraspirazione potenziale è fortemente correlato alla temperatura. Sappiamo inoltre dalla teoria che la correlazione si risolve in causalità in quanto dalla letteratura si apprende che la temperatura è uno dei fattori principali che agisce sul valore dell'evapotraspirazione.

Inoltre, si nota come l'evapotraspirazione reale sia fortemente correlata alle precipitazioni. In questo caso non si tratta di un rapporto causa-effetto: l'evapotraspirazione reale è limitata alle precipitazioni, ovvero non può evapotraspirare più di quanto che precipita.

Infine, in linea con quanti ci si aspetta dalla teoria, l'umidità del suolo sembra dipendere positivamente dalle precipitazioni e negativamente dell'evapotraspirazione (e temperatura).

Risulta di particolare interesse studiare la relazione che esiste tra Temperatura e Evapotraspirazione potenziale. Per svolgere questa operazione è necessario analizzare i dati tenendo in considerazione le criticità osservate nella distribuzione dei dati per l'evapotraspirazione potenziale.

Come primo passo è utile osservare in che modo di distribuisce la temperatura in quelle osservazioni in cui l'evapotraspirazione è pari a 0, che compongono il 31% circa del totale delle osservazioni.

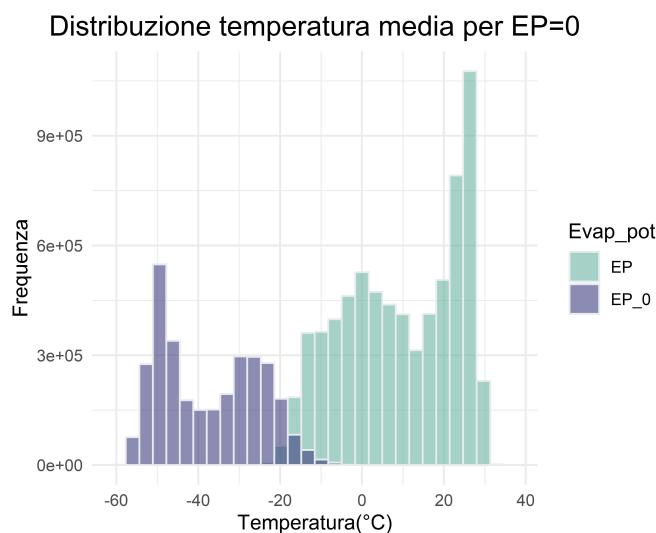


Figura 3.8: Distribuzione temperatura con focus su EP pari a 0

Da questo grafico è possibile notare come le osservazioni per cui l'evapotraspirazione potenziale è pari a 0 hanno valori molto bassi per la temperatura. Sembra esserci un soglia intorno ai -20 °C sotto la quale l'evapotraspirazione potenziale diventa pari a 0. Sotto una certa temperatura, l'effetto dell'evapotraspirazione potenziale si annulla.

Continuando nell'analisi della relazione tra queste due variabili è utile osservare il trend che queste hanno nel tempo. In questa fase vengono considerate solo le osservazioni che hanno un'evapotraspirazione potenziale diversa da 0 in quanto, se si considerasse l'intero dataset, avremmo una fetta considerevole di osservazioni (31%) che sarebbero da rumore nelle analisi.

Per ognuna delle due variabili viene effettuata la media annuale. I seguenti grafici aiutano a comprendere l'andamento temporale delle variabili.

Ai grafici di scatter plot è stato aggiunto una linea per interpolare l'andamento calcolata utilizzando un polinomio di terzo grado.

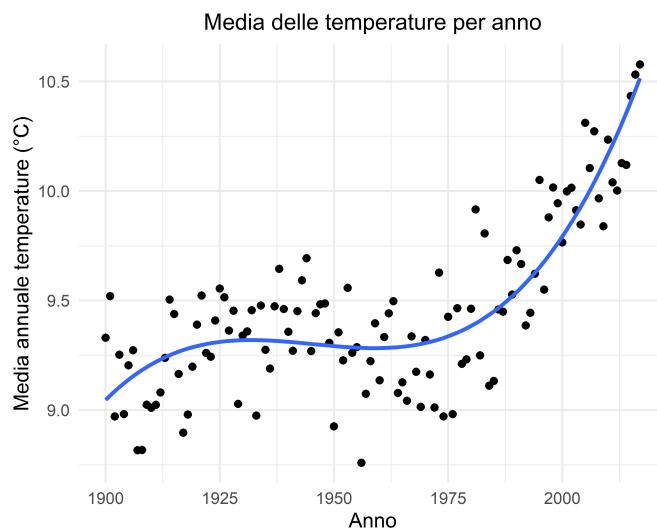


Figura 3.9: Andamento temporale della temperatura media

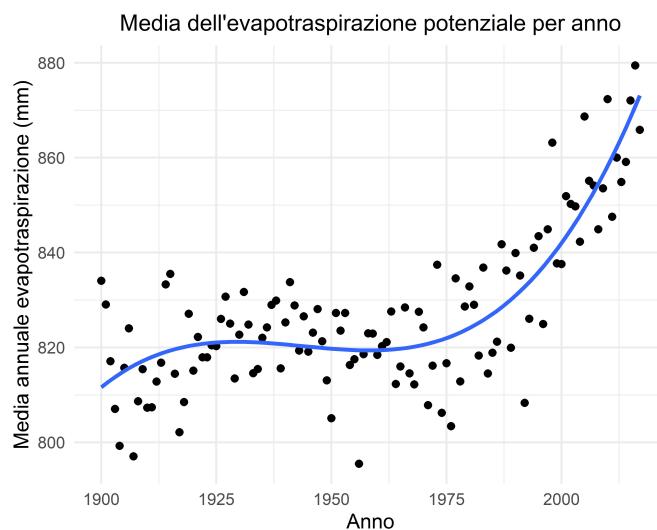


Figura 3.10: Andamento temporale dell'evapotraspirazione media

Come è evidente dalle immagini presentate, le due variabili sono fortemente correlate, e viene consolidata la teoria di una relazione causa-effetto tra temperatura e Evapotraspirazione.

Questa evidente correlazione è inoltre visibile dal seguente grafico (la linea che interpola i dati corrisponde ad un polinomio di primo grado).

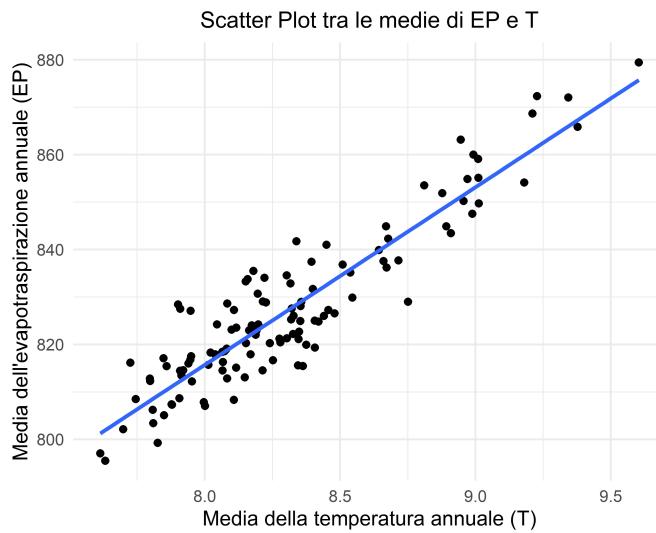


Figura 3.11: Scatter plot tra temperatura e Evapotraspirazione potenziale

Si riscontrare inoltre, una cambio netto nella tendenza temporale delle due misure: fino all'anno 1980 l'andamento non presenta grandi fluttuazioni, e la media risulta stabile.

Dall'anno 1980 è chiaro un cambio di trend verso un aumento significativo dei valori delle due variabili.

Questo argomento verrà approfondito in seguito.

Infine, è stato effettuato un test statistico sulla correlazione tra temperatura ed evapotraspirazione potenziale, considerando anche in questo caso solo le osservazioni aventi evapotraspirazione potenziale diversa da 0.

Il test effettuato è il *test di correlazione di Pearson* è un test statistico che valuta la forza della relazione lineare tra due variabili continue. Verifica l'ipotesi nulla che non vi sia alcuna correlazione tra le due variabili contro l'ipotesi alternativa che vi sia una correlazione diversa da zero. Restituisce una statistica test, un valore p e un intervallo di confidenza per il vero coefficiente di correlazione.

Viene mostrato l'output del test effettuato su R

Pearson's product-moment correlation

```
data: df_corr$media_temp and df_corr$media_p_evap
t = 25.017, df = 116, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8846031 0.9427351
sample estimates:
      cor
0.9184948
```

Il test porta a rifiutare l'ipotesi nulla di assenza di correlazione tra le variabili

3.1.3 Aggregazione per nazione

Dopo aver esplorato i dati grezzi, una fase importante da eseguire è quella di aggregare i dati per nazione: è necessario associare ad ogni nodo della griglia la nazione all'interno di cui questo nodo ricade.

Questa operazione è fondamentale per poter mettere in comunicazione i dati relativi ai fattori ambientali con quelli del Virtual Water Trade.

Avendo a disposizione per ogni nazione le informazioni circa le variabili ambientali tramite i nodi che cadono all'interno di quella nazione, è possibile aggregare i dati geograficamente al fine di avere i dati non più associati ai nodi, ma alle nazioni.

Per ogni nazione sarà dunque possibile studiare l'andamento e il comportamento delle variabili ambientali e capire come ciò influisca sul Virtual Water trade.

Come primo passo di questo processo serve vedere se esistono alcuni nodi che non compaiono all'interno di alcuna nazione e come questi sono distribuiti nel globo.

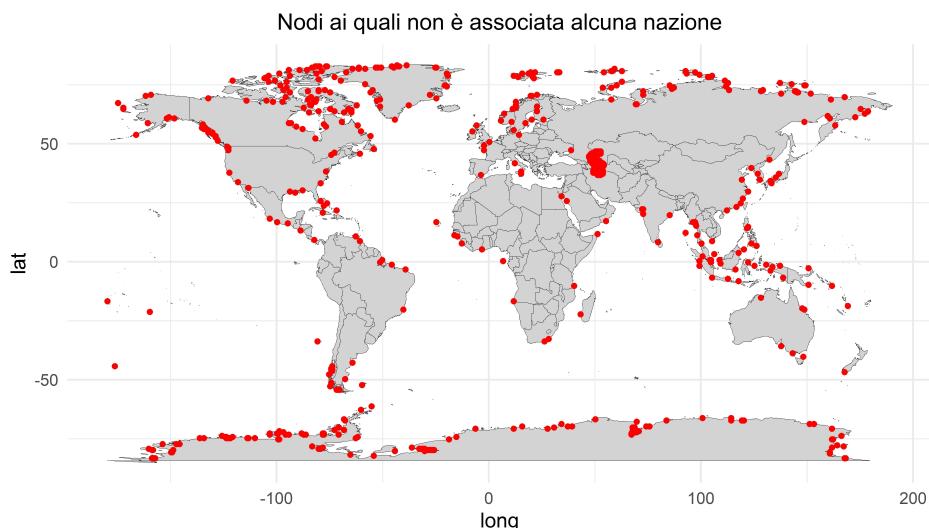


Figura 3.12: Nodi non associati ad alcuna nazione

I nodi le cui coordinate non rientrano all'interno di alcuna nazione sono 633 (0.7%) e si tratta per la maggior parte di osservazioni poste vicino alla costa. Ricordando che le osservazioni iniziali erano approssimate a determinate latitudini e longitudini, alcuni nodi cadono al di fuori del territorio, ovvero in mare.

Inoltre, una buona parte di questi missing values è dovuto al fatto che esistono osservazioni sul mar Caspio, come riportato nella successiva mappa, le quali non devono essere giustamente prese in considerazione.

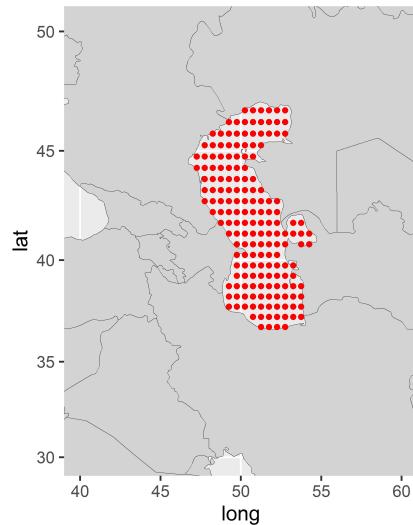


Figura 3.13: Nodi nel Mar Caspio

Tolti quei pochi nodi sopra analizzati, ad ogni nodo del dataset è possibile associare una nazione. Ogni nazione avrà dunque un numero di nodi proporzionale all'ampiezza del suo territorio.

Per avere delle informazioni sulle variabili ambientali aggregate per nazione e per anno, è sufficiente applicare la media dei valori che corrispondono ai nodi all'interno nazione per la variabile interessata per ogni anno.

Il dataset creato ha dunque la seguente struttura:

Nazione	1900	1901	...	2017
Italia	<i>val1900</i>	<i>val1901</i>	...	<i>val2017</i>
Brasile			...	
...			...	

Tabella 3.3: Variabili ambientali per nazione per anno

Il risultato di questo lavoro di aggregazione geografica è avere per ogni nazione, per ogni variabile, la media annua della variabile sui nodi che cadono all'interno della nazione.

3.2 Dati relativi al Virtual Water

In questo paragrafo verrà presentata l'analisi esplorativa relativa ai dati relativi al Virtual Water Trade. In questa fase sono stati estratti diversi dataset, di diversa natura e con scopi diversi, per questo motivo le analisi effettuate su i diversi dataset saranno diverse l'una dall'altra.

Prima di esplorare graficamente i dati, è necessario focalizzarsi sull'importanza e sul ruolo dei dati all'interno del lavoro.

Come è stato precedentemente detto, i dati utilizzati in questo capitolo hanno un orizzonte temporale variabile, indicativamente di 60-70 anni dal 2017 al 1960 e la frequenza degli anni presenti nei diversi dataset non è uniforme. Inoltre, per i dati meno recenti si riscontra un numero di osservazioni molto limitato (l'informazione è presente solo per poche nazioni), come è possibile osservare nei paragrafi successivi.

Inoltre l'impiego di questo insieme di dati, non è finalizzato allo studio di un trend temporale, bensì ad avere un giudizio sulla performance delle diverse nazioni nel campo del Virtual Water Trade. Per questo, per avere una fotografia attuale, è necessario che i dati impiegati siano quanto più recenti possibili.

Per questi due motivi, le analisi principali che verranno fatte saranno incentrate sui dati relativi all'anno più recente, ovvero al 2017, che oltre ad essere l'ultimo anno disponibile è condiviso tra tutti i dataset e, come sarà mostrato in seguito, è l'anno in cui si ha un maggior numero di osservazioni.

3.2.1 Uso dell'acqua e stress

Il paragrafo in oggetto ha come scopo quello di presentare le analisi per i 3 dataset estratti dalla pagina di **outworldindata**[7].

Prelievi acqua sul totale

Il dataset contiene le informazioni circa la percentuale di acqua utilizzata da ogni nazione, per ogni anno disponibile, sul totale delle risorse idriche interne.

Il dataset è composto da 1152 righe e 4 colonne:

Nazione	Anno	Percentuale	Indice
<i>nazione 1</i>	<i>anno 1</i>	<i>percentuale 1</i>	<i>indice 1</i>
..

Tabella 3.4: Percentuale utilizzo di acqua

Non sono presenti valori nulli

Nel dataset è presente una colonna *indice* che esprime un grado di stress idrico in base al valore della percentuale di acqua utilizzata:

- Se la percentuale supera *80%* l'indice assume valore 5, ovvero *stress molto elevato*. Esistono casi in cui la percentuale supera il *100%*: si tratta di casi in cui vengono utilizzate risorse idriche non rinnovabili, quali estrazione insostenibile da fonti acquifere o produzione di acqua dole tramite processo di desalinizzazione.
- Indice ha valore 4, pari a *stress alto* se la percentuale è compresa tra *40%* e *80%*.
- Con valori tra il *20%* e *40%*, l'indice ha valore 3, ovvero *stress medio*.
- Percentuali tra il *10%* e *20%* l'indice assume valore 2, *stress basso*
- Il valore *stress molto basso*, pari a 1, è presente se la percentuale è sotto al *10%*

Ogni riga presenta i valori relativi alla percentuale e all'indice di stress per ogni nazione per ogni anno.

Le nazioni presenti nel dataset sono 190, mentre ogni nazione è presente nella seguente frequenza nel dataset:

Anno	Frequenza
1962	1
1967	2
1972	16
1977	35
1982	55
1987	89
1992	120
1997	143
2002	159
2007	167
2012	175
2017	190

Tabella 3.5: Stress idrico, frequenza delle osservazioni per anno

Come era stato anticipato, il numero di nazioni (osservazioni) per le quali è maggiore tanto quanto è più recente l'anno considerato.

Risulta importante esplorare i dati da un punto di vista geografico, viene dunque presentata una mappa che riporta per ogni nazione l'indice di stress idrico nel 2017.

Indice di stress idrico per nazione

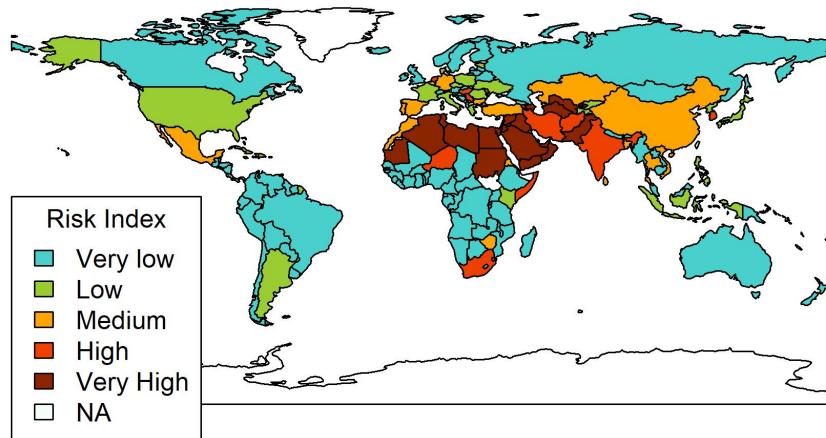


Figura 3.14: Stress idrico per nazione nel 2017

Utilizzo dell'acqua per settore

Il secondo dataset che viene presentato contiene le informazioni circa l'utilizzo d'acqua da parte delle nazioni, per ogni anno disponibile.

Il dataset è composto da 1109 righe e 5 colonne: per ogni nazione, per ogni anno, sono presenti le percentuali dell'impiego dell'acqua in 3 settori: agricoltura, industria, uso domestico.

La somma di questi tre settori dovrebbe essere 100, in quanto è il totale dell'acqua estratta ed utilizzata.

Nazione	Anno	Perc Agricoltura	Perc Industria	Perc uso domestico
<i>nazione 1</i>	<i>anno 1</i>	<i>perc ind 1</i>	<i>perc agr 1</i>	<i>perc uso dom 1</i>
..

Tabella 3.6: Percentuale utilizzo di acqua per settore

Come prima operazione è stata aggiunta una sesta colonna, *Somma percentuali* in cui è stata calcolata la somma delle tre percentuali al fine di verificare se il totale fosse 100.

Delle 1109 righe iniziali, 1033 presentano valori completi (ovvero tutte le percentuali sono valorizzate) e la somma pari a 100%.

Dopo aver pulito il dataset è possibile osservare il numero di osservazioni per ogni anno presenti.

Anno	Frequenza
1965	1
1967	1
1972	10
1977	19
1982	31
1987	64
1992	101
1997	134
2002	148
2005	1
2007	155
2012	170
2014	6
2017	190

Tabella 3.7: Utilizzo acqua, frequenza delle osservazioni per anno

Come è possibile notare, anche in questo caso il numero più alto di osservazioni si ha in corrispondenza dell'anno più recente, il 2017, mentre andando indietro nel tempo il numero di nazioni per cui si hanno i dati diminuisce.

Le prossime mappe, una per settore, mostrano le percentuali di utilizzo d'acqua per settore nel 2017 per ogni nazione.

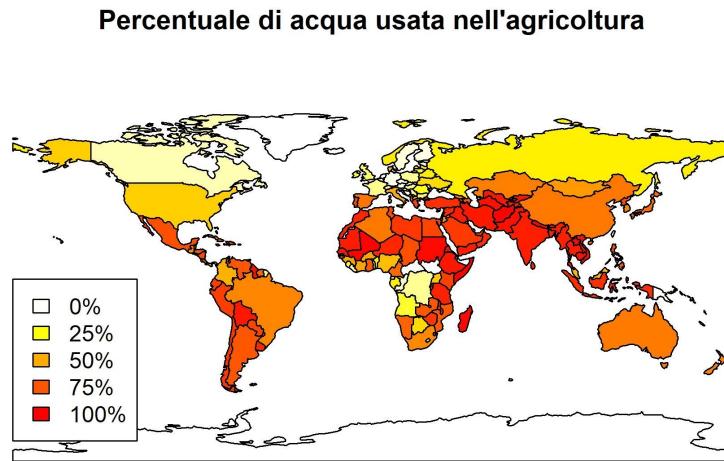


Figura 3.15: Percentuale agricoltura per nazione nel 2017

Percentuale di acqua usata nell'industria

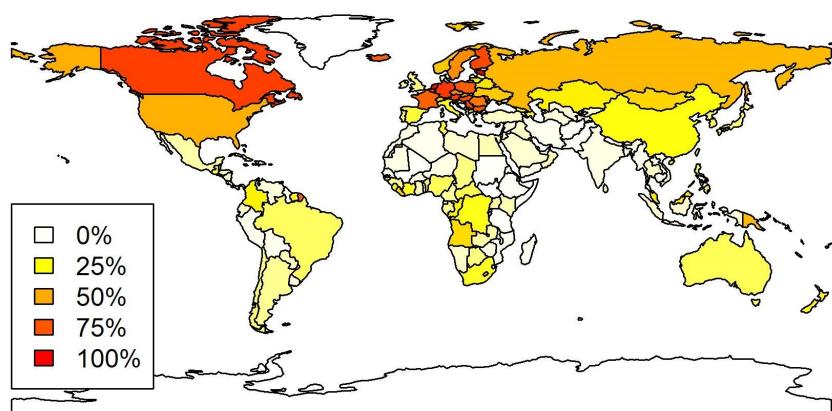


Figura 3.16: Percentuale industria per nazione nel 2017

Percentuale di acqua usata nell'uso domestico

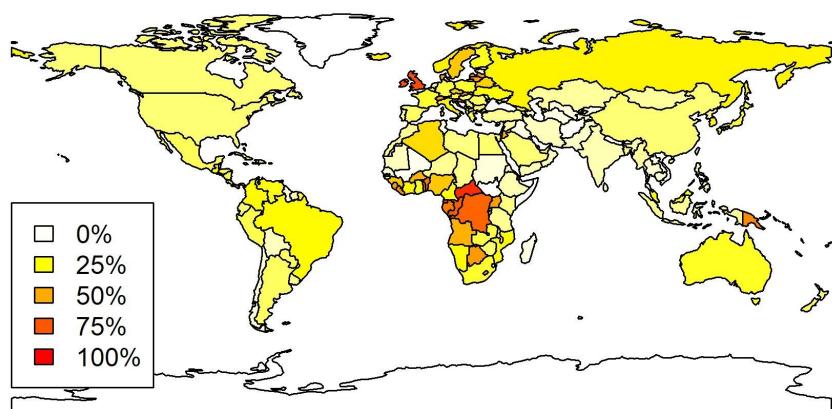


Figura 3.17: Percentuale uso domestico per nazione nel 2017

Consumo d'acqua per prodotto alimentare

L'ultimo dataset estratto da analizzare in questa sezione è quello relativo all'utilizzo d'acqua per prodotto alimentare.

Le informazioni contenute in esso sono fondamentali in quanto, aggregando questi dati con quelli relativi alla produzione e consumo di determinati prodotti, è possibile calcolare l'impiego d'acqua per nazione per prodotto.

Il dataset è molto semplice: presenta due colonne la prima contiene il prodotto alimentare, la seconda i litri d'acqua stimati per produrne un chilogrammo.

Il dataset presenta 36 righe, una per ogni prodotto: nella seguente tabella verranno visualizzate le righe relative solo ai prodotti per i quali si hanno i dati circa la produzione e il consumo.

Prodotto	Litri per kg di prodotto
Latte	628,2
Uova	577,7
Pollame	660
Ovini	1795,8
Bovini	2082,8

Tabella 3.8: Litri di acqua impiegati per prodotto alimentare

3.2.2 Produzione e consumo di prodotti alimentari

Il focus di questo paragrafo sarà incentrato sull'analisi esplorativa dei dati relativi alla produzione e consumo di prodotti alimentari quali uova carne e latte.

I dataset utilizzati sono divisi in 3 gruppi, produzione, consumo pro-capite e popolazione.

Aggregando queste informazioni è possibile avere un quadro a livello sia nazionale sia temporale della produzione e del consumo di questi prodotti.

Le analisi che vengono fornite servono per avere una visione sulla composizione dei dati. In aggiunta, sul sito da cui i dati sono estratti, ourworldindata.org, alla pagina *"Meat and Dairy production"* [6], è possibile utilizzare dei grafici interattivi per visualizzare i dati nel modo in cui l'utente preferisce.

Prima di analizzare i dataset, viene fatta una premessa: i dati di interesse per l'analisi sono quelli relativi all'anno 2017 per due ragioni.

Come descritto nel paragrafo precedente, lo scopo dell'analisi di questi dati non è quello di individuare un trend o studiare l'andamento temporale del virtual water trade. Questo lavoro può essere approfondito dal lettore direttamente nelle fonti dati citate. Lo scopo è invece quello di analizzare il comportamento di una nazione in termini di Import-Export di acqua virtuale tramite il commercio di beni.

Inoltre, avendo deciso di estrarre i dati più recenti nella precedente analisi, quelli in corrispondenza del 2017, non avrebbe senso in questa fase considerare un orizzonte temporale diverso, in quanto non sarebbe possibile effettuare un analisi incrociata tra le due informazioni.

Produzione di prodotti alimentari

Nella prima sezione vengono analizzati i dati relativi alla produzione di prodotti alimentari per nazione.

I dataset a disposizione sono 3: il primo fa riferimento alla produzione di carne, che viene divisa tra produzione bovina, suina e di pollame; il secondo dataset contiene i dati circa la produzione di latte, mentre il terzo riguarda la produzione di uova.

I dataset hanno tutti una profondità temporale che va dal 1961 al 2020.

Poiché ai fini dell'analisi è necessario focalizzarsi solo sull'anno 2017, verranno analizzati solo i dati relativi a quell'anno.

I tre dataset vengono aggregati tra di loro per nazione, formando un dataset di dimensione 200 righe per 6 colonne. Ogni riga rappresenta una nazione.

Nazione	Prod. suina	Prod. bovina	Prod. pollame	Prod. latte	Prod. uova
<i>Nazione 1</i>	<i>valore 1</i>	<i>valore 2</i>	<i>valore 3</i>	<i>valore 4</i>	<i>valore 5</i>
..

Tabella 3.9: Dataset produzione prodotti alimentari

I valori delle produzioni dei prodotti sono espressi in tonnellate.

Vengono successivamente analizzati i valori mancanti per ogni variabile, ovvero il numero di nazioni per cui non è presente il dato nel 2017

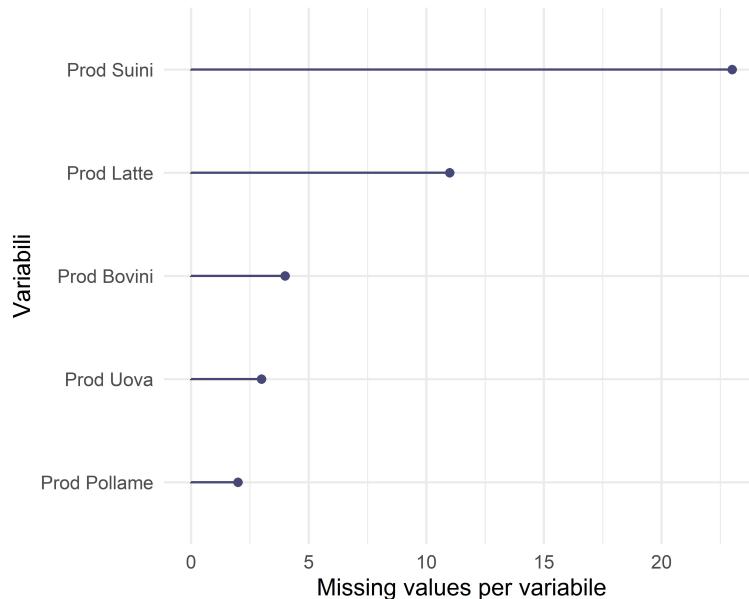


Figura 3.18: Numero di Missing values per variabile nella produzione

L'alto numero di nazioni per le quali non si hanno dati relativi alla produzione di carne suina, è dovuto a motivi religiosi.

L'analisi prosegue analizzando la classifica delle nazioni con il valore più alto per ogni variabile, ovvero osserviamo quali nazioni producono la maggior quantità dei diversi prodotti.

Rank	Bovini		Suini		Pollame	
	Nazione	Quantità(t)	Nazione	Quantità(t)	Nazione	Quantità(t)
1	Usa	11907239	Cina	54517969	Usa	21914241
2	Brasile	9550000	Usa	11610981	Cina	19486789
3	Cina	6346200	Germania	5505572	Brasile	14165530
4	Argentina	2844511	Spagna	4298789	Russia	4542244
5	India	2485396	Brasile	3824682	India	3805233
6	Pakistan	2079000	Vietnam	3733349	Messico	3249022
7	Australia	2068616	Russia	3515740	Indonesia	3218172
8	Messico	1926901	Francia	2147550	Giappone	2214911
9	Russia	1569267	Canada	2127600	Turchia	2192351
10	Francia	1432780	Polonia	2032900	Argentina	2161370
Perc.	60%		78%		62%	

Tabella 3.10: Top 10 nazioni per quantità prodotta - parte 1

Rank	Latte		Uova	
	Nazione	Quantità(t)	Nazione	Quantità(t)
1	India	176284776	Cina	30962891
2	Usa	97787291	Usa	6350756
3	Pakistan	52482000	Indonesia	4995639
4	Cina	34644927	India	4847500
5	Brasile	34576557	Brasile	3065010
6	Germania	32626482	Messico	2771198
7	Russia	30179241	Giappone	2601173
8	Francia	26509065	Russia	2518746
9	Nuova Zelanda	21461603	Turchia	1205075
10	Turchia	20699894	Thailandia	1085000
Perc.	63%		72%	

Tabella 3.11: Top 10 nazioni per quantità prodotta - parte 2

Come è possibile osservare, per ogni alimento è stato riportata anche la percentuale della produzione delle top 10 nazioni sulla produzione totale mondiale.

Consumo pro-capite di prodotti alimentari

Il secondo insieme di dataset comprendono le informazioni circa il consumo pro-capite dei prodotti alimentari in esame. Anche in questa sezione i dataset sono tre, il primo è relativo al consumo di carne, il secondo al consumo di latte e il terzo al consumo delle uova.

I dati fanno riferimento all'anno 2017 come è stato spiegato in precedenza.

Per vedere la distribuzione del consumo pro capite dei diversi alimenti verranno presentate delle mappe in cui il colore della nazione sarà dipendente dal consumo pro capite per alimento.

Consumo di carne bovina pro-capite (kg)

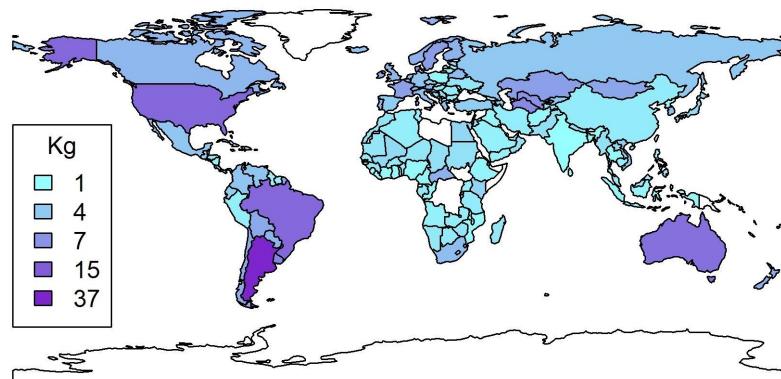


Figura 3.19: Consumo pro-capite di carne bovina nel 2017

Consumo di carne suina pro-capite (kg)

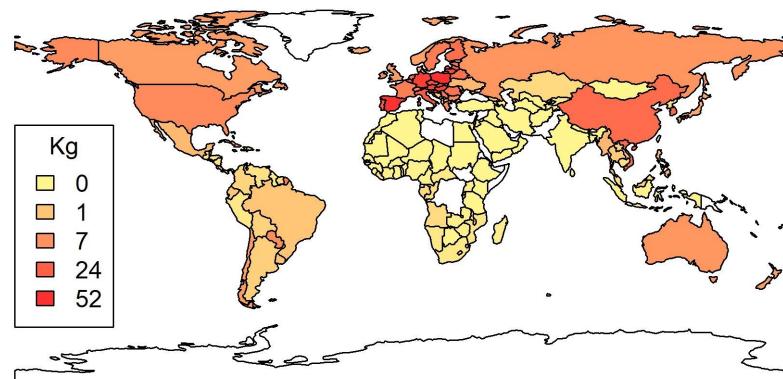


Figura 3.20: Consumo pro-capite di carne suina nel 2017

Consumo di pollame pro-capite (kg)

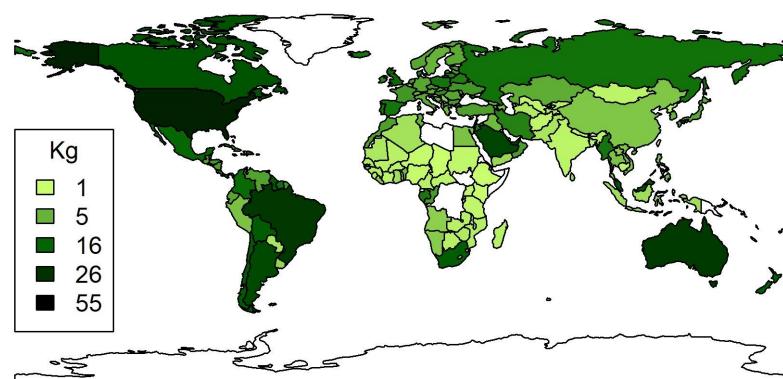


Figura 3.21: Consumo pro-capite di pollame nel 2017

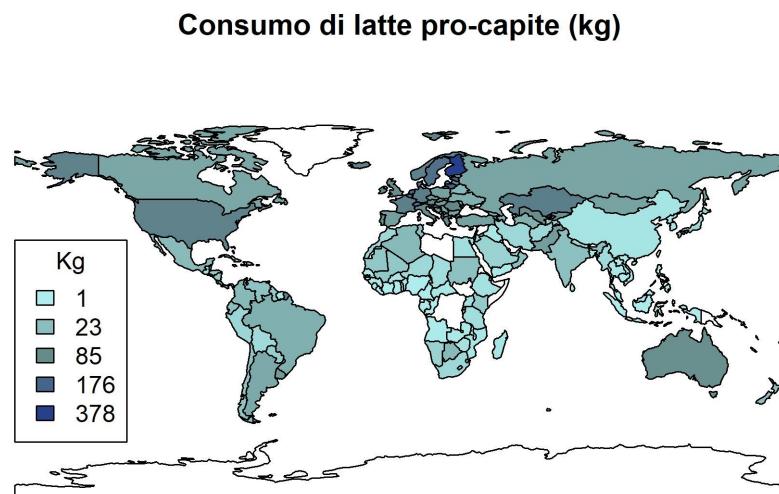


Figura 3.22: Consumo pro-capite di latte nel 2017

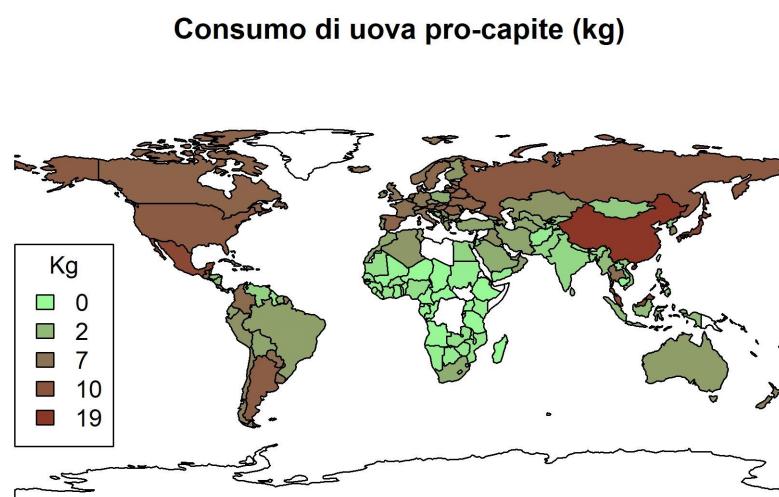


Figura 3.23: Consumo pro-capite di uova nel 2017

I grafici presentati presentano una legenda che permette di interpretare il grafico. Per ogni variabile, sono mostrati in legenda i colori relativi a determinati percentili, così suddivisi: il primo valore rappresenta il valore minimo, il secondo il 25esimo percentile, il terzo la mediana, il quarto il 75esimo percentile e l'ultimo valore è pari al valore massimo.

I tre dataset vengono successivamente aggregati in modo da avere un'unica tabella che contiene le informazioni, per ogni nazione, del consumo pro capite di ogni prodotto.

Il numero di nazioni per le quali non si hanno informazioni circa il consumo pro capite è visualizzabile nel seguente grafico:

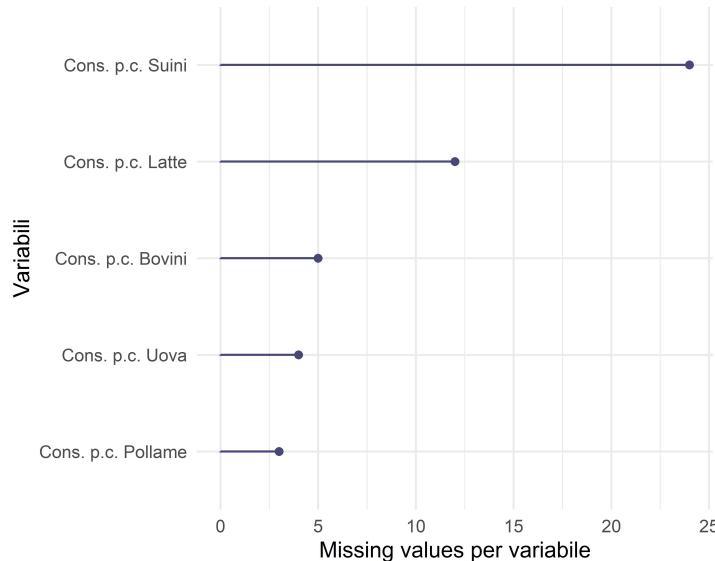


Figura 3.24: Numero di Missing values per variabile nella produzione

Come è possibile notare, il numero per ogni variabile di nazione senza dati per il consumo è molto vicino a quello riguardante la produzione.

Popolazione residente

L'ultimo dataset estratto dal sito ourworldindata.org è un dataset contenente le informazioni circa la popolazione residente per ogni nazione del mondo. Anche in questo caso vengono utilizzati i dati relativi al 2017.

Il dato è presente per 251 osservazioni: sono presenti anche i dati per diverse aggregazioni di stati (ad esempio sommati per continenti) che fanno diventare il numero delle osservazioni superiore al numero delle nazioni effettivamente esistenti.

Come detto in precedenza questo dataset viene esclusivamente utilizzato per avere il consumo totale per alimento per nazione, in quanto l'informazione circa il consumo è presente a livello pro capite: moltiplicando il dato pro capite per la popolazione residente è possibile avere il consumo totale per nazione, che può essere comparato con il dato circa la produzione.

Capitolo 4

Creazione dashboard interattive e analisi statistiche

In questo capitolo vengono mostrate e spiegate le dashboard interattive create per questo progetto tramite il pacchetto Shiny [2]. Per dashboard interattiva si intende una pagina creata per visualizzare i dati in maniera interattiva in cui sono presenti diversi oggetti che si dividono in due categorie.

- La prima categoria di oggetti comprende strumenti con cui l'utente può interagire per modificare la visualizzazione dei dati nella dashboard. Si tratta per esempio di cursori temporali, testi di input, bottoni, elenchi a selezione multipla e molti altri. Tramite questi strumenti si possono selezionare le variabili e il range temporale di interesse.
- La seconda categoria di oggetti raggruppa grafici, mappe, o altri oggetti di visualizzazione dei dati che cambiano a seconda dei comandi imposti dall'utente negli strumenti interattivi.

Il funzionamento da parte dell'utente consiste nella manipolazione degli strumenti interattivi al fine di visualizzare negli strumenti di visualizzazione del dato, le informazioni di interesse.

L'utilizzo di questo pacchetto è molto utile quando si lavora con moli di dati che contengono al loro interno innumerevoli informazioni, come nel lavoro esposto, in quanto permette di poter analizzare i dati in maniera pulita senza dover affrontare un lavoro di pulizia e manipolazione del dato che è stata effettuata a priori.

Inoltre, come detto nell'introduzione, uno dei fini di questo elaborato è quello di fornire all'utente uno strumento di analisi di dati che possa essere utilizzato in autonomia, ed è proprio a tal scopo che sono state create delle dashboard per permettere un'analisi individuale a chiunque voglia approfondire le nozioni qui presentate.

4.1 Mappe delle variabili

La prima dashboard che è stata creata e che viene introdotta in questo capitolo ha come oggetto le variabili ambientali e la loro rappresentazione su mappa.

L'obiettivo è quello di creare un interfaccia con cui l'utente può relazionarsi che mostri una mappa globale delle osservazioni per ogni anno e per ogni variabile: l'utente dunque avrà la possibilità di visualizzare una mappa in cui ogni nodo ha un intensità di colore pari al valore che la assume per una determinata variabile, in un determinato anno.

I dataset utilizzati nella dashboard sono 5, uno per ogni variabile, e hanno la struttura descritta nel capitolo circa l'analisi esplorativa. Per ogni variabile è stata tenuta la solo informazione annuale, che rappresenta la media annua (per la temperatura e l'umidità del suolo) o la somma annua (per le precipitazioni, l'evapotraspirazione potenziale e evapotraspirazione reale).

I dataset sono dunque composti da 85794 osservazioni e 120 variabili: le osservazioni rappresentano i nodi, le variabili i valori assunti dalla variabile nel nodo, oltre a due variabili che rappresentano la latitudine e longitudine dello stesso.

In seguito viene mostrato un esempio della dashboard il cui utilizzo è molto semplice. Dal pannello di selezione si sceglie la variabile, mentre dal cursore si indica l'anno: in seguito la mappa a fianco mostrerà il valore per ogni nodo della variabile selezionata nell'anno scelto.

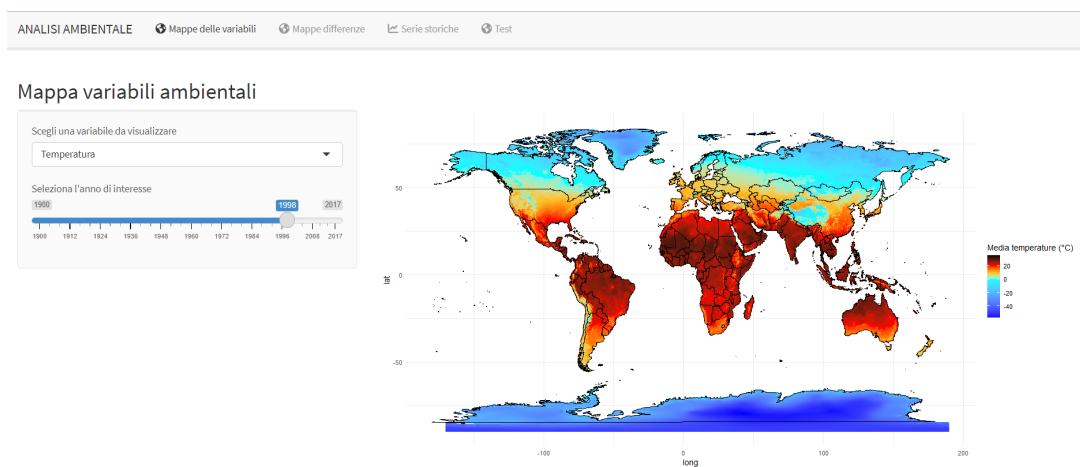


Figura 4.1: Esempio dashboard mappe delle variabili

In questo caso la variabile scelta è la temperatura e l'anno è il 1998. La legenda a destra aiuta a capire il valore assunto dalla variabile nei diversi nodi nel mondo.

4.2 Mappe delle differenze rispetto la media di riferimento

La seconda dashboard presenta un'analisi sulle variabili ambientali che si basa sulla differenza del valore assunto in una anno, rispetto alla media calcolata su un periodo di riferimento

Lo scopo di questa pagina è di fornire una dashboard che mostri lo scostamento dalla media di riferimento delle variabili per ogni nodo della mappa, per ogni anno.

I dataset utilizzati sono sempre uguali per tutte le dashboard, ovvero per ogni variabile si tiene conto della media annuale (o somma a seconda della variabile) su ogni nodo della mappa.

Il primo passo per la creazione della dashboard è stato quello di individuare un periodo di riferimento su cui calcolare la media che fosse lo stesso per tutte le variabili. Il periodo scelto è stato il trentennio che va dal 1950 al 1979, per due principali motivi:

1. Osservando i grafici di pagina 26 che mostrano l'andamento medio annuo della temperatura e dell'evapotraspirazione potenziale, è facile notare come ci sia una netto cambio di trend intorno all'anno 1980, preceduto da anni in cui il valore medio delle variabili era rimasto piuttosto stabile. Gli anni 50-79 rappresentano dunque un buon periodo di riferimento su cui calcolare le medie.
2. L'utilizzo della media calcolata sul totale degli anni avrebbe appiattito le differenze: la media di riferimento della temperatura sarebbe più alta, portando ad avere delle mappe con la quasi totalità di valori negativi per gli anni precedenti al 1980.

Successivamente è stata calcolata la media delle diverse variabili sui 30 anni selezionati, in modo da avere una media fissa per ognuna delle cinque variabili.

L'ultimo passo è stato quello di calcolare la differenza tra il valore che assunto da un nodo per una variabile in un determinato anno e il valore medio di riferimento della variabile in quel nodo.

Questa operazione viene eseguita dal codice che crea le dashboard per ogni variabile e anno.

La struttura della dashboard è simile a quella della dashboard precedente. Tramite un pulsante è possibile selezionare una delle cinque variabili e mediante un cursore temporale si può fissare l'anno su cui è desiderata l'informazione

Una volta impostati i pulsanti come desiderato l'immagine a fianco mostra una mappa del mondo in cui ogni nodo è colorato a seconda del valore che assume la differenza tra il valore della variabile e il valore medio di riferimento.

La mappa mostra dunque per ogni nodo del globo, se la variabile ambientale selezionata ha assunto nell'anno scelto valori più alti o più bassi rispetto alla media e di quanto il valore si discosta dalla media.

In seguito è riportato un esempio della dashboard.

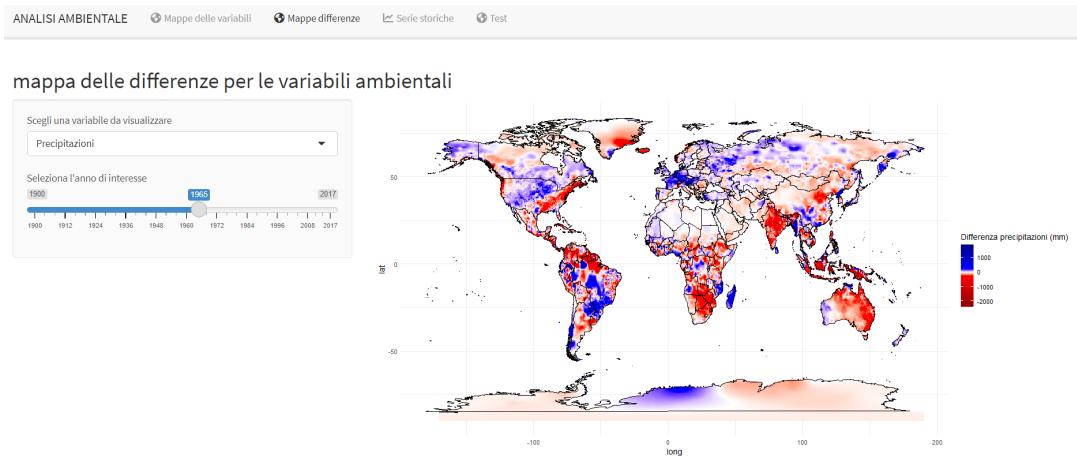


Figura 4.2: Esempio dashboard delle differenze rispetto alla media di riferimento

L'esempio riportato prende in esame la dashboard avendo selezionato la variabile precipitazioni per l'anno 1965.

Si nota come sono state usate le tonalità di blu, per evidenziare una differenza positiva mentre il rosso per le differenze negative. Le zone tendenti al bianco registrano differenze quasi nulle.

Il colore assunto dal nodo in base al valore della differenza, in questo modo è facile vedere come ci siano aree e regioni del mondo a tendenza positiva e altre a tendenza negativa per ogni anno.

Nell'immagine riportata si nota ad esempio come in Australia le precipitazioni medie per il 1965 siano state più basse della media, mentre in centro Europa ci sia stato una media annuale più alta rispetto alla media.

Questa seconda dashboard è stata creata per poter avere un'informazione più dettagliata e in parte diversa rispetto alla prima. La prima dashboard risulta utile se lo scopo è quello di confrontare il valore della variabile nei diversi luoghi del pianeta, tenendo fisso l'anno pre-stabilito. La seconda dashboard ha invece lo scopo di evidenziare le differenze per capire se in un anno, i valori per una variabile sono stati più alti o più bassi della media. Si tratta dunque di un punto di osservazione diverso, difficilmente catturabile alla prima dashboard.

4.3 Grafici serie storiche

La dashboard in oggetto in questo paragrafo centra l'attenzione sullo sviluppo di serie storiche annuali.

Nei paragrafi precedenti sono stati illustrati degli strumenti per poter interagire con i dati e creare delle mappe ad hoc mettendo in risalto la componente spaziale dei dati. Risulta necessario però approfondire anche la storicità dei dati, ovvero creare una dashboard le cui informazioni sono focalizzate sull'evoluzione temporale delle variabili.

La finalità della dashboard risulta dunque essere quello di poter osservare, per ogni nodo della mappa, l'evoluzione temporale delle diverse variabili mediante dei grafici che rappresentano le serie storiche.

I dataset impiegati sono gli stessi delle dashboard precedenti e comprendono la totalità delle informazioni aggregate annualmente per ogni variabile e ogni nodo.

La dashboard presenta una struttura diversa dalle precedenti e viene di seguito descritta.

Gli strumenti con cui l'utente può interagire sono due.

Il primo è un elenco delle cinque variabili, elenco dal quale è possibile selezionare una o più variabili per le quali sono desiderati i grafici successivamente descritti.

Il secondo è una mappa del mondo interattiva, con cui l'utente può interagire andando a cliccare con il cursore su un nodo della mappa (nodo per cui sono presenti i dati nei dataset). Una volta cliccato il nodo della mappa le coordinate di questo verranno visualizzate al di sotto della mappa.

Come detto in precedenza lo scopo della mappa è quello di fornire le informazioni circa l'evoluzione temporale delle variabili sui singoli nodi e quest'ultimo strumento descritto rappresenta un modo molto efficace e puntale per andare a selezionare delle coordinate geografiche per le quali sono richiesti i grafici delle serie storiche.

L'utilizzo della mappa è descrivibile in pochi passi

1. Selezionare una o più variabile di interesse
2. Scegliere il nodo cliccando sulla mappa con il cursore e visualizzare le coordinate del nodo
3. Visualizzare le serie storiche delle variabili che appaiono sotto

Nella figura 4.3 è possibile osservare un esempio della prima parte della dashboard in cui sono visibili l'elenco a scelta multipla, in cui sono selezionate tutte le variabili, e la mappa interattiva, in cui è già selezionato un nodo.

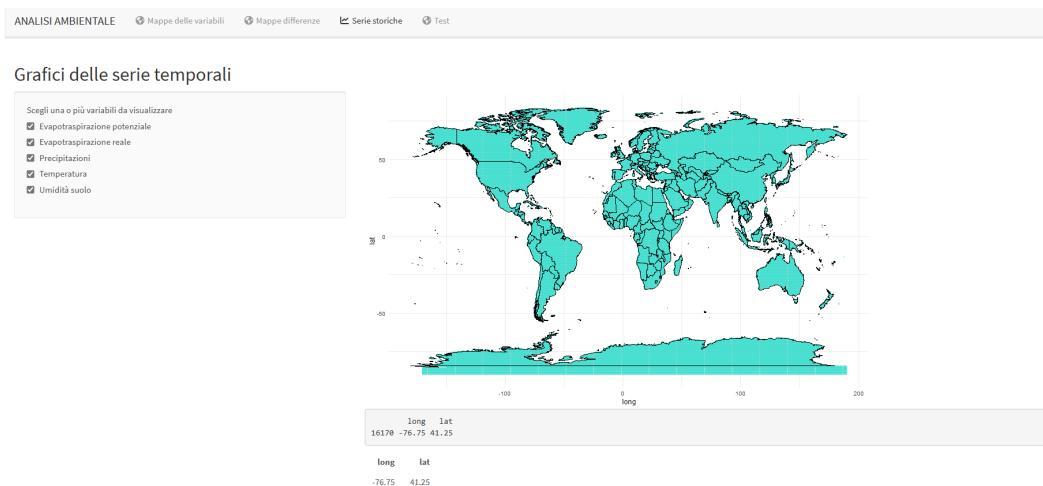


Figura 4.3: Dashboard grafici serie storiche 1

L’immagine successiva (figura 4.4) mostra 3 grafici relativi alle serie storica delle variabili selezionate. Sono state appositamente selezionate tutte le variabili nell’esempio avere una visuale completa del funzionamento della dashboard.

Il numero di grafici è dovuto ad una scelta tecnica: i valori delle variabili circa l’evapotraspirazione potenziale, evapotraspirazione reale e precipitazioni, hanno la stessa scala di misura, ovvero i millimetri (di acqua) per questo è stato ritenuto opportuno inserirli nello stesso grafico. Inoltre, poiché uno degli scopi dell’analisi è di capire l’andamento delle risorse idriche da un punto di vista ambientale, avere sullo stesso grafico la serie storica annuale delle variabili che influiscono la quantità di acqua sul territorio è molto utile per analizzarle in maniera coordinata.

Il secondo grafico contiene le informazioni circa la serie storica della temperatura, ed essendo questa misurata in gradi °C è stata raffigurata singolarmente.

Per lo stesso motivo è stato creato un grafico apposta per l’umidità del suolo: essendo valorizzata con un range da 0 a 150 (mm) deve essere rappresentata a parte rispetto alle altre variabili.

Le serie storiche, oltre a presentare i valori puntuali che rappresentano la somma (o la media) per ogni anno nel nodo (raffigurati dai punti), sono arricchite di una linea che è relativa alla media mobile su 5 anni.

Questa scelta è stata voluta per una questione visiva e grafica: osservando la linea della media mobile è più facile intuire la presenza di trend.



Figura 4.4: Dashboard grafici serie storiche 2

4.4 Mappe basate sul test T di Student

La quarta dashboard presenta un interfaccia utente in grado di mostrare le mappe delle variabili in base al risultato di un test statistico.

Questa dashboard risulta concettualmente più complessa delle precedenti e necessita di un approfondimento teorico prima di essere presentata.

I motivi iniziali che hanno spinto alla realizzazione di questa dashboard risalgono alle figure 3.9 e 3.10 del capitolo precedente, ovvero quelle rappresentanti gli andamenti temporali della temperatura e dell’evapotraspirazione a livello globale. Come già osservato nel capitolo 3, era evidente un aumento delle temperature su scala globale dall’anno 1980, ed un conseguente aumento dell’evapotraspirazione potenziale.

Sorge dunque la necessità di analizzare nel dettaglio e da un punto di vista statistico questa situazione, in particolare si è voluta approfondire l’analisi circa questa inversione di trend.

L’interrogativo che è stato posto e al quale si è cercato di dare una risposta è se esistesse un effettivo cambiamento nel comportamento delle variabili ambientali negli ultimi anni rispetto al passato, ovvero se effettivamente ci fosse un’evidenza statistica significativa che confermasse il cambio di trend osservato.

Lo scopo dell’analisi è stato quello di osservare due periodi temporali e di verificare se per ogni nodo ci fosse un’evidenza statistica di un aumento o diminuzione dei valori nel nodo. Trattandosi di un’analisi che coinvolge tutte le variabili il fine è stato quello di testare, per ogni nodo, la presenza sia di un possibile aumento sia di una possibile diminuzione dei valori tra il periodo temporale meno recente e quello più recente.

Come primo passo sono stati scelti i due periodi di riferimento da confrontare: il primo fa riferimento al range temporale 1950-1979, il secondo è relativo agli ultimi 30 anni di storico, ovvero dal 1988 al 2017.

La scelta dei due periodi è la seguente: il periodo temporale meno recente è stato preso di 30 anni affinché non fosse troppo breve e in modo che fossero considerati gli anni precedenti all’inversione di trend, mentre il periodo più recente corrisponde ai 30 anni più recenti di storico.

Successivamente sono state calcolate le medie e le deviazioni standard sui i 30 anni per i due periodi, per ogni nodo per ogni variabile. Ogni nodo ha dunque l’informazione, per ogni variabile, circa la media e la deviazione standard calcolate sulle 30 osservazioni dal 1950 al 1979 e circa la media e la deviazione standard calcolate sulle 30 osservazioni dal 1988 al 2017.

L’analisi continua con la scelta di un test statistico per confortare le medie testare la presenza di un aumento o diminuzione dei valori significativa. Il test scelto è il test t-Student.

Il test t-Student è un test che viene utilizzato per confrontare la differenza tra due medie e valutare se questa differenza è statisticamente significativa. Questo test si basa sulla distribuzione t-Student.

Il test t-Student è particolarmente utile quando i campioni sono piccoli (meno di 30) o quando non si conosce la deviazione standard della popolazione. In questi casi, il test utilizza la deviazione standard delle due medie campionarie per stimare la deviazione standard della popolazione.

Per eseguire il test t-Student, bisogna prima calcolare la differenza tra le due medie. Successivamente è necessario calcolare l'errore standard della differenza tra le medie utilizzando le deviazioni standard delle due medie campionarie e le dimensioni dei rispettivi campioni. Infine, si calcola il valore t utilizzando la formula $t = (x_1 - x_2) / s$, dove x_1 e x_2 sono le due medie e s è l'errore standard della differenza.

Il valore t che si ottiene rappresenta il rapporto tra la differenza tra le due medie e l'errore standard della differenza. Più il valore t è grande, più la differenza tra le due medie è significativa. Per determinare se la differenza tra le due medie è statisticamente significativa, devi confrontare il valore t con una distribuzione t critica. Se il valore t è maggiore del valore t critico per il livello di significatività scelto (ad esempio, 0,05), allora puoi rifiutare l'ipotesi nulla e affermare che le due medie sono significativamente diverse. Se il valore t è minore del valore t critico, allora non puoi fare alcuna affermazione riguardo alla differenza tra le due medie.[1]

In maniera analoga è possibile osservare il p-value osservato, rappresentante la probabilità di osservare una statistica di test estrema o più estrema di quella ottenuta dal campione (t), dato che l'ipotesi nulla è vera (assenza di differenza tra le medie), e confrontalo con un p-value soglia, 0.05. Se il p-value è minore della soglia, allora l'ipotesi nulla è da rifiutare e si può optare per l'ipotesi alternativa che, a seconda del test può affermare o una differenza significativa delle medie, oppure che una media è significativamente maggiore dell'altra.

Quanto descritto può essere sintetizzato dalle seguenti formule:

La differenza tra le due medie:

$$\text{Differenza tra le medie} = \mu_1 - \mu_2$$

dove μ_1 e μ_2 sono le due medie che vuoi confrontare.

L'errore standard della differenza tra le medie:

$$\text{Errore standard della differenza} = \frac{s_1}{\sqrt{n_1}} + \frac{s_2}{\sqrt{n_2}}$$

dove s_1 e s_2 sono le deviazioni standard delle due medie campionarie e n_1 e n_2 sono le dimensioni dei campioni corrispondenti.

Il valore t :

$$t = \frac{\text{Differenza tra le medie}}{\text{Errore standard della differenza}}$$

Il p-value è dato da:

$$p = P(|T| > t)$$

dove T è la statistica di test t-student e t è il valore ottenuto dalla statistica di test calcolato sui dati del campione. $P(|T| > t)$ rappresenta la probabilità che la statistica di test sia più estrema o uguale a t se l'ipotesi nulla è vera.

Dopo aver presentato il test statistico è possibile applicarlo ai dati e ciò è stato fatto nel seguente modo.

Per ogni variabile si sono considerate le due medie e le due deviazioni standard per ogni nodo ed è stato costruito un algoritmo che procede seguendo questi step (μ_1 fa riferimento alla media 1960-1979, μ_2 è la media del range temporale 1988-2017)

- Come passo viene effettuato un test t-student con ipotesi alternativa

$$\mu_2 > \mu_1$$

e ipotesi nulla

$$\mu_1 = \mu_2$$

e soglia del p-value al 0.05. Se il p-value ha valore inferiore al 0.05, il nodo presenta una variazione significativa verso l'alto dei valori della variabile in esame

- Se il p-value del primo passo è maggiore di 0.05, viene effettuato un test t-student con ipotesi alternativa

$$\mu_2 < \mu_1$$

e ipotesi nulla

$$\mu_1 = \mu_2$$

e soglia del p-value al 0.05. Se il p-value ha valore inferiore al 0.05, il nodo presenta una variazione significativa verso il basso dei valori.

- Se anche in questo caso il p-value è maggiore di 0.05 il nodo presenta valori delle medie che accettano l'ipotesi nulla

$$\mu_2 = \mu_1$$

contro l'ipotesi alternativa

$$\mu_2 \neq \mu_1$$

ad un livello di significatività del 90%.

All'interno di ogni variabile, è stata creata una nuova variabile categoriale *risultato test* per ogni nodo che riassumesse l'esito del test effettuato sulle coordinate del nodo. I possibili valori che può assumere la nuova variabile sono:

- **HI** quando il p-value risulta minore di 0,05 al primo passo dello step del test, ovvero quando viene rifiutata l'ipotesi nulla

$$\mu_1 = \mu_2$$

a fronte dell'ipotesi alternativa

$$\mu_2 > \mu_1$$

. Nei casi (nodi) in cui si verifica questa condizione, la media della variabile calcolata sugli anni più recenti (μ_2) risulta significativamente maggiore della media calcolata sul periodo meno recente μ_1 .

- **LO** quando il p-value risulta minore di 0,05 al secondo passo, ovvero quando viene rifiutata l'ipotesi nulla

$$\mu_1 = \mu_2$$

e si opta per l'ipotesi alternativa

$$\mu_2 < \mu_1$$

. Ciò significa che la media della variabile calcolata sugli anni più recenti (μ_2) risulta significativamente minore della media calcolata sul periodo meno recente μ_1 .

- **NULL** se vengono accettate le ipotesi nulle dei test effettuati ai primi due passi. In questo caso viene accettata l'ipotesi nulla

$$\mu_1 = \mu_2$$

al 90%, ovvero la differenza tra le medie risulta essere significativamente nulla.

Viene riportato l'esempio di utilizzo della dashboard nella seguente immagine.

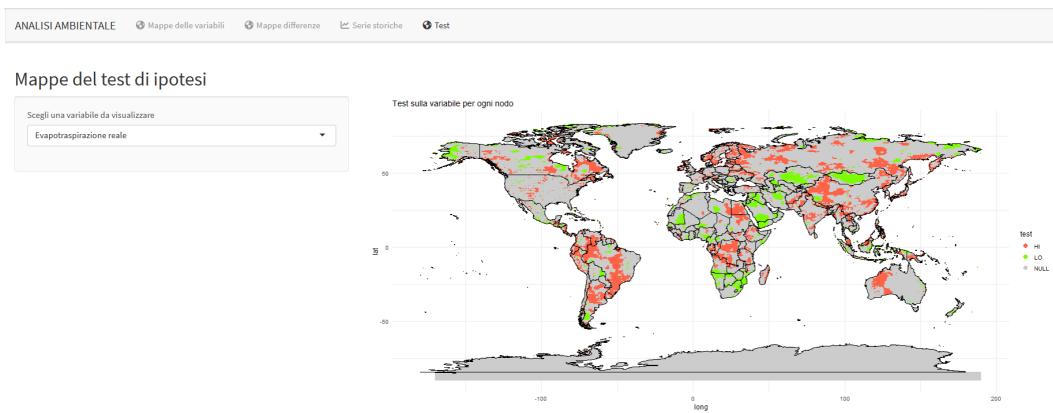


Figura 4.5: Dashboard test significatività differenza tra medie

Nell'esempio riportato la variabile considerata è l'evapotraspirazione reale. Osservando la mappa si nota come ogni nodo sia colorato in modo da assumere il colore in corrispondenza dell'esito del test sulle medie del nodo.

Il colore rosso indica un aumento significativo nella media, il colore verde una diminuzione significativa della media, mentre il grigio rappresenta l'ultimo caso, ovvero una differenza pari a 0 tra le due medie.

L'unico elemento interattivo della dashboard è l'oggetto a sinistra, tramite cui si può selezionare la variabile ambientale di interesse.

4.5 Mappe per serie storiche per nazioni

In questo paragrafo viene presentata una dashboard interattiva in cui il focus sulle variabili ambientale si sviluppa a livello nazionale.

L'obiettivo della dashboard è di poter visualizzare per ogni nazione e per ogni variabile ambientale, una serie storica del valore della variabile nella nazione.

L'esigenza di creare una dashboard che comprenda un'analisi aggregata per nazione nasce dall'idea del lavoro di poter eseguire un analisi incrociata tra l'ambito relativo alle variabili ambientali e il virtual water trade.

In ambito di analisi ambientali, risulta dunque necessario non solo analizzare i singoli nodi della mappa, ma aggregare i dati dei fattori ambientali per nazione in modo da osservarne l'andamento nel tempo su tutta la nazione nel complesso.

Poiché uno degli scopi del lavoro è quello di studiare i deficit di risorse idriche nelle diverse nazioni è dunque fondamentale possedere uno strumento mediante cui è possibile visualizzare l'andamento delle variabili che influiscono sull'approvvigionamento idrico a livello nazionale.

Dopo aver analizzato i nodi le cui coordinate non rientrano all'interno di alcuna nazione, è possibile aggregare i dati nel seguente modo.

Ad ogni nodo viene associata la nazione all'interno di cui le coordinate del nodo cadono. Avendo l'informazione relativa alla nazione, viene calcolata per ogni nazione la media di ogni variabile per ogni anno. Il dataset creato è così composto:

Nazione	Anno	media t.	media p.	media ep.	media er.	media us.
Norvegia	1900	val t	val p	val ep	val er	val us
Norvegia	1901
..
Svezia	1900	val t	val p	val ep	val er	val us
Svezia	1901
..

Tabella 4.1: Dataset variabili ambientali per nazione

Il risultato di questa aggregazione di dati è un valore medio annuale per ogni nazione per ogni variabile che fornisca l'informazione a livello nazionale e non più a livello di singolo nodo. Il dataset ha 13929 righe e 7 colonne.

Questa informazione risulta molto utile poiché, oltre a poter visualizzare dei trend storici come verrà illustrato successivamente nella dashboard, permette di confrontare i valori medi delle variabili ambientali con i valori dell'effetto del virtual water trade sulle risorse idriche interne alla nazione.

La dashboard in oggetto del paragrafo ha la seguente struttura:

Mediante due strumenti a sinistra è possibile selezionare la variabile ambientale e la nazione di interessa e destra verrà mostrato il grafico della serie storica annuale della variabile in quella nazione.

Di seguito viene mostrato un esempio della dashboard, in cui è stata selezionata la Norvegia e l'evapotraspirazione potenziale come variabile

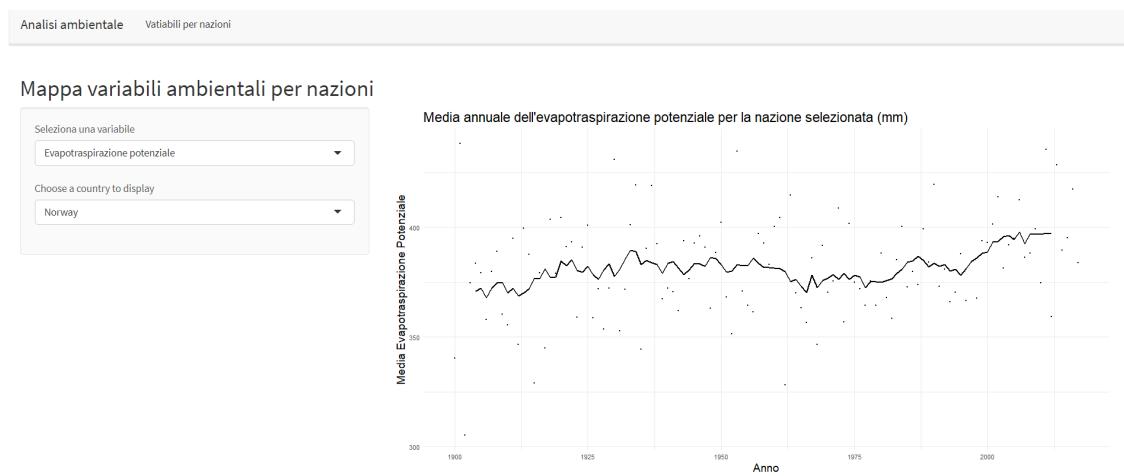


Figura 4.6: Dashboard serie storiche analisi ambientali per nazione

Si nota come, oltre all'informazione puntuale circa il valore nell'anno, viene visualizzata anche un linea che rappresenta la media mobile decennale per meglio osservare i trend.

In aggiunta, è possibile utilizzare questa dashboard per effettuare un analisi incrociata con le informazioni circa i singoli nodi che cadono all'interno della nazione. Ad esempio, utilizzando la terza dashboard si può osservare il valore della variabile nel tempo in un singolo nodo. Il grafico può essere comparato con il medesimo grafico aggregato a livello nazionale ed è possibile confrontare le due serie storiche.

4.6 Previsione variabili ambientali

La sesta dashboard ha l'obiettivo di mostrare la previsione del valore dell'evapotraspirazione potenziale al variare della temperatura media in una nazione.

L'analisi sulle variabili temporali si conclude con l'impostazione di un modello previsorio che permette di stimare la variazione dell'evapotraspirazione potenziale, sulla base di possibili variazioni nella temperatura.

Come è stato anticipato, la temperatura è uno dei fattori ambientali che maggiormente influenza l'evapotraspirazione, la quale è responsabile della perdita d'acqua che avviene in maniera naturale. Risulta dunque fondamentale approfondire il rapporto tra queste due variabili proprio per studiare come un cambiamento nella temperatura possa influenzare l'evapotraspirazione in ottica di verificare le risorse idriche di un paese.

Prima di esporre la dashboard è necessario effettuare un approfondimento sui dati e sul modello previsionale impiegati in questa analisi.

Come primo passo, partendo dai dati di input, è stata estratta l'informazione circa la nazione all'interno di cui ogni nodo della griglia cade, in maniera analoga alla dashboard precedente, ottenendo così il dataset come descritto nella tabella 4.1.

Delle variabili vengono tenute solo quelle di interesse, ovvero la media della temperatura e la media dell'evapotraspirazione potenziale.

Osservando lo scatter plot tra le variabili otteniamo il seguente risultato:

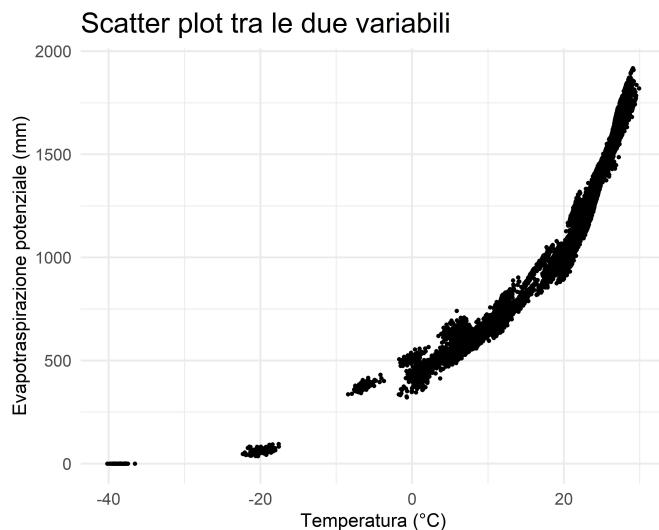


Figura 4.7: Scatter plot tra Temperature ed Evap. potenziale

Si nota come, per il gruppo di osservazioni avente temperatura molto bassa, circa sotto i -30 gradi, l'evapotraspirazione potenziale sia pari a 0. Queste osservazioni verranno escluse dal dataset su cui verrà creato e testato il modello in quanto risulterebbero solo di disturbo non fornendo informazioni utili al modello poiché è facile stimare con errore minimo che ad osservazioni con temperatura minore di -30 gradi, l'evapotraspirazione potenziale sia pari a 0.

Successivamente, deve essere scelto il modello previsionale per l'analisi, ovvero un modello statistico che sfrutta i dati e strumenti matematici per fare previsioni su eventi futuri, basandosi sui dati storici e sui modelli di relazione tra le variabili coinvolte. Questo insieme di modelli viene chiamato in termini accademici machine learning.

Gli algoritmi di machine learning si dividono in parametrici e non parametrici.

Gli algoritmi di machine learning non parametrici sono progettati per funzionare bene quando è disponibile una grande quantità di dati e quando ci sono relazioni complesse tra le variabili indipendenti e dipendenti. Sono spesso utilizzati quando la relazione tra le variabili non è facilmente descritta da una semplice formula o equazione matematica.

Tuttavia, se si ha solo una variabile indipendente, è probabile che ci sia una semplice relazione lineare tra quella variabile e la variabile dipendente. In tali casi, è spesso più appropriato utilizzare un metodo parametrico come la regressione lineare per modellare la relazione.

I metodi non parametrici, come random forest e support vector machine, sono più adatti alle situazioni in cui ci sono molteplici variabili indipendenti e le relazioni tra le variabili sono più complesse. Questi metodi possono catturare relazioni non lineari tra le variabili indipendenti e dipendenti e spesso forniscono previsioni migliori quando sono coinvolti molti fattori.

Dunque, gli algoritmi di machine learning non parametrici potrebbero non essere adatti per i set di dati con una sola variabile indipendente perché la relazione tra le variabili indipendenti e dipendenti è probabile che sia semplice e facilmente modellata da metodi parametrici come la regressione lineare. I metodi non parametrici sono più utili per set di dati più complessi con molte variabili indipendenti.

Per questo motivo nell'analisi di questo capitolo è stato scelto un modello parametrico per prevedere l'evapotraspirazione potenziale (variabile dipendente) un base al valore assunto dalla temperatura (variabile indipendente). Nello specifico il modello testato è un modello di regressione lineare [9].

La regressione lineare è un metodo statistico utilizzato per modellare la relazione tra una variabile dipendente Y e una o più variabili indipendenti X_1, X_2, \dots, X_p . L'obiettivo della regressione lineare è trovare la migliore relazione lineare tra le variabili, che può essere utilizzata per prevedere il valore della variabile dipendente per un dato insieme di valori delle variabili indipendenti.

Il modello di regressione lineare assume che la relazione tra le variabili indipendenti X_1, X_2, \dots, X_p e la variabile dipendente Y possa essere descritta da un'equazione lineare della forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

dove β_0 è il termine di intercetta o costante, $\beta_1, \beta_2, \dots, \beta_p$ sono i coefficienti o pendenze, e ϵ è il termine di errore che rappresenta la variabilità nella variabile dipendente che non è spiegata dalle variabili indipendenti.

I coefficienti $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sono stimati utilizzando il metodo dei minimi quadrati, che prevede di trovare i valori dei coefficienti che minimizzano la somma degli errori quadrati tra i valori previsti di Y e i valori effettivi di Y nel dataset. L'equazione per l'estimatore dei minimi quadrati dei coefficienti è:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

dove $\hat{\beta}$ è il vettore degli stimatori dei coefficienti, X è la matrice delle variabili indipendenti, Y è il vettore dei valori della variabile dipendente, X^T è la traspota della matrice X , e $(X^T X)^{-1}$ è l'inverso del prodotto di matrici $X^T X$.

Una volta stimati i coefficienti, possono essere utilizzati per fare previsioni per nuovi valori delle variabili indipendenti. Il valore previsto di Y per un dato insieme di valori delle variabili indipendenti $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ è:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

Applicando il modello lineare alle variabili a noi disposizioni abbiamo che Y corrisponde all'evapotraspirazione potenziale (EP) e X_1 è l'unica variabile indipendente e corrisponde alla temperatura (T)

Come nel caso in esame, quando si ha una sola variabile indipendente x e si vuole utilizzare x con diversi gradi, si può utilizzare la regressione lineare polinomiale.

La regressione lineare polinomiale [10] è una forma generalizzata di regressione lineare che prevede una relazione polinomiale tra la variabile indipendente x e la variabile dipendente y . In questo caso, il modello di regressione lineare polinomiale con grado k è definito come:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

Una volta stimati i coefficienti di regressione, è possibile utilizzare il modello di regressione lineare polinomiale per fare previsioni su nuove osservazioni di x .

Una volta stabilito il modello da utilizzare per la previsione, è necessario stabilire la variabile k , ovvero il grado del polinomio dell'equazione sopra descritta.

Per stabilire il grado k della regressione lineare polinomiale, si può utilizzare il metodo del model selection, ovvero la selezione del modello migliore tra diversi modelli candidati, ovvero tra i diversi modelli polinomiali di diverso grado.

Lo strumento di model selection utilizzato è lo step-forward, un metodo di selezione delle variabili che parte dal modello più semplice (ovvero una regressione lineare con una sola variabile indipendente) e aggiunge una variabile alla volta, selezionando la variabile che migliora di più la bontà di adattamento del modello.

Il metodo dello step-forward crea dunque una serie di modelli con diverse variabili, e seleziona il miglior modello sulla base di una misura imposta, come l'AIC.

L'Akaike Information Criterion (AIC) è un criterio di selezione del modello utilizzato per confrontare modelli statistici alternativi e determinare quale di essi fornisce la migliore descrizione dei dati.

L'AIC tiene conto sia della bontà di adattamento del modello ai dati (ovvero la capacità del modello di descrivere i dati osservati) sia della sua complessità (ovvero il numero di parametri del modello). L'AIC cerca di selezionare il modello che fornisce la migliore bontà di adattamento utilizzando il minor numero di parametri.

La formula per il criterio di selezione del modello AIC è:

$$AIC = -2 \ln(L) + 2k$$

dove L è la verosimiglianza del modello e k è il numero di parametri del modello

Il modello con il valore più basso dell'AIC è considerato il modello migliore. Quando si confrontano più modelli, il modello con la differenza minima di AIC rispetto al modello migliore è considerato un modello accettabile.

Dopo aver introdotto il modello parametrico da un punto di vista teorico il lavoro prosegue applicando il modello ai dati.

Come accennato all'inizio del paragrafo, il dataset utilizzato è quello descritto in tabella 4.1, compreso delle sole colonne circa la nazione, l'anno, la media telle temperature e la media dell'evapotraspirazione potenziale.

Delle 13924 righe di cui è composto il dataset vengono escluse, come spiegato in precedenza, quelle righe aventi media della temperatura minore di -30°C, ottenendo un dataset di 13806.

Il dataset viene dunque diviso tra test set e training set in proporzione 30% e 70%. Il training set verrà utilizzato per allenare il modello e dunque per individuare il modello migliore mediante l'AIC e per stimare i parametri del modello. Il test set serve per testare il modello, ovvero applicando il modello creato ai dati del test set si testa se il modello è performante e se non presenta problemi di overfitting.

Il modello ottenuto utilizza la variabile temperatura fino al quarto grado, ed è il seguente:

$$\hat{EP} = \hat{\beta}_0 + \hat{\beta}_1 T_1 + \hat{\beta}_2 T_2 + \hat{\beta}_3 T_3 + \hat{\beta}_4 T_4$$

I coefficienti presentano i seguenti valori:

Coefficiente	Stima	p-value
β_0	456,44	0
β_1	0,01	0
β_2	18,46	0
β_3	0,001	0
β_4	-0,11	0

Tabella 4.2: Coefficienti modello

La misura usata per valutare le performance del modello è il MAE (Mean Absolute Error), che rappresenta la media delle differenze assolute tra i valori previsti dal modello e quelli effettivi.

Il MAE ottenuto sul training set è pari a 33,54 , mentre sul test set si ottiene un MAE di 34,22. I risultati sono molto soddisfacenti in quanto il modello non presenta problemi di overfitting.

Viene di seguito mostrato un grafico che ai dati del training set sovrappone le previsioni del modello su tutto l'orizzonte di dati di input.

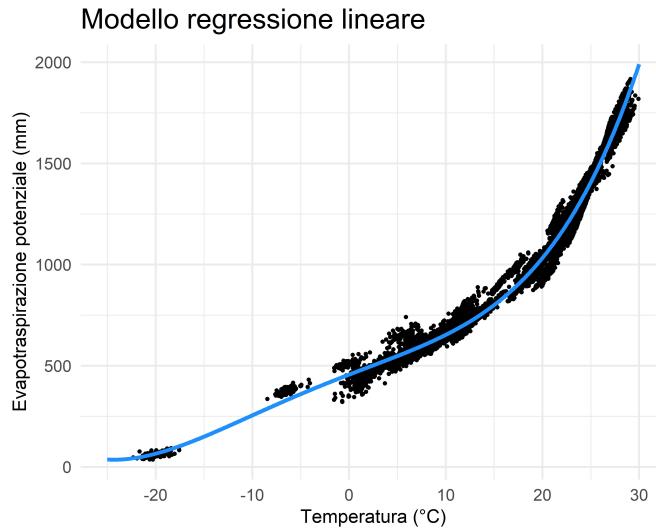


Figura 4.8: Scatter plot con modello regressione lineare

Dopo aver stabilito il modello da utilizzare per questa analisi, è possibile presentare la dashboard in oggetto.

La dashboard interattiva presenta due oggetti con l'utente può interagire: il primo si tratta di un elenco di nazioni in cui l'utente può scegliere la nazione di interesse, il secondo è un cursore che l'utente può utilizzare per cambiare la temperatura media e effettuare previsioni.

L'ultimo elemento è una tabella in cui sono presenti 4 informazioni: la temperatura media degli ultimi 10 anni, l'evapotraspirazione potenziale media degli ultimi 10 anni, la temperatura modificata mediante il cursore e la previsione dell'evapotraspirazione potenziale sulla nuova temperatura.

L'utilizzo della dashboard è il seguente:

Una volta selezionata la nazione, si hanno le informazioni circa la media la temperatura media degli ultimi 10 anni e l'evapotraspirazione potenziale media degli ultimi 10 anni della nazione. Il cursore permette di aumentare o diminuire la temperatura media (terza colonna) ottenendo una nuova temperatura. L'ultima colonna fornisce una stima dell'evapotraspirazione potenziale effettuata sulla temperatura modificata.

Viene di seguito fornito un esempio di utilizzo.

Modello previsionale



Figura 4.9: Dashboard Previsione con modello regressione lineare

Come si nota dall'esempio, la 'Nuova temperatura modificata' risulta essere la somma tra la 'Temperatura media ultimi 10 anni' e la variazione della temperatura imposta con il cursore. L'ultima colonna mostra la previsione ottenuta utilizzando il modello precedentemente descritto.

4.7 Tabelle Virtual Water Trade

L'ultima dashboard creata contiene le informazioni circa il fenomeno del Virtual Water Trade.

Lo scopo della dashboard è quello di fornire uno strumento di visualizzazione puntuale dei dati di questo fenomeno per poter combinare questa informazione con quelle circa i fattori ambientali.

Per ogni nazione è dunque possibile osservare, nel 2017, i flussi in ingresso e in uscita dell'acqua virtuale sotto forma di alimenti.

I dataset utilizzati in questa dashboard sono quelli relativi alla produzione dei diversi alimenti, al consumo pro-capite degli stessi, alla popolazione residente per nazione e al consumo di litri d'acqua necessario per la produzione di un chilogrammo di ogni alimento.

Come primo passo è stato calcolato il consumo nazionale di ogni alimento moltiplicando il dato circa il consumo pro-capite per la popolazione residente.

Successivamente, per ogni nazione, è stata calcolata la differenza tra produzione e consumo per ogni elemento, ottenendo dunque un deficit, positivo o negativo. Un deficit positivo indica una produzione maggiore del consumo e di conseguenza l'esportazione di una parte di produzione per l'alimento in esame. Viceversa, una differenza negativa indica un consumo maggiore della produzione e dunque la necessità di importare il prodotto.

Per ogni nazione, una volta ottenuto il valore della differenza tra produzione e consumo per ogni alimento, questo dato è stato moltiplicato per la corrispettiva quantità d'acqua necessaria a produrre un chilogrammo del prodotto, come descritto nella tabella 3.8 a pagina 35.

Dunque, il risultato di questa aggregazione di dati è l'informazione circa l'import o export di acqua virtuale per ogni prodotto per ogni nazione.

Successivamente, sommando a livello nazionale tutti gli elementi di import ed export di acqua virtuale (quindi sommando tutti i dati per ogni alimento) è possibile ottenere la somma di acqua virtuale importata o esportata dalla nazione nel 2017.

Questi dati sono necessario per poter avere un'idea del comportamento della nazione in ambito di virtual water trade, ovvero quantificare l'acqua virtuale importata o esportata.

La dashboard è così costruita:

L'utente ha la possibilità di interagire con un menù a tendina da cui è possibile selezionare la nazione di interesse.

A destra è presente una mappa che non varia e che riporta l'indice di rischio per quanto riguarda l'uso delle risorse idriche intere. Questa mappa è stata aggiunta per poter avere l'informazione circa la situazione idrica dei vari paesi e poter subito analizzare come i paesi si comportano. Si tratta di un oggetto che rende la dashboard più utile e che permette di effettuare un'analisi immediata all'utente.

In basso vengono visualizzati i dati relativi al virtual water trade per la nazione selezionata, nell'anno 2017.

Come descritto nel commento sottostante, i numeri fanno riferimento a litri d'acqua. Numeri positivi indicano una esportazione di acqua virtuale (produzione > consumo) mentre numeri negativi indicano un importazione di acqua virtuale (produzione < consumo).

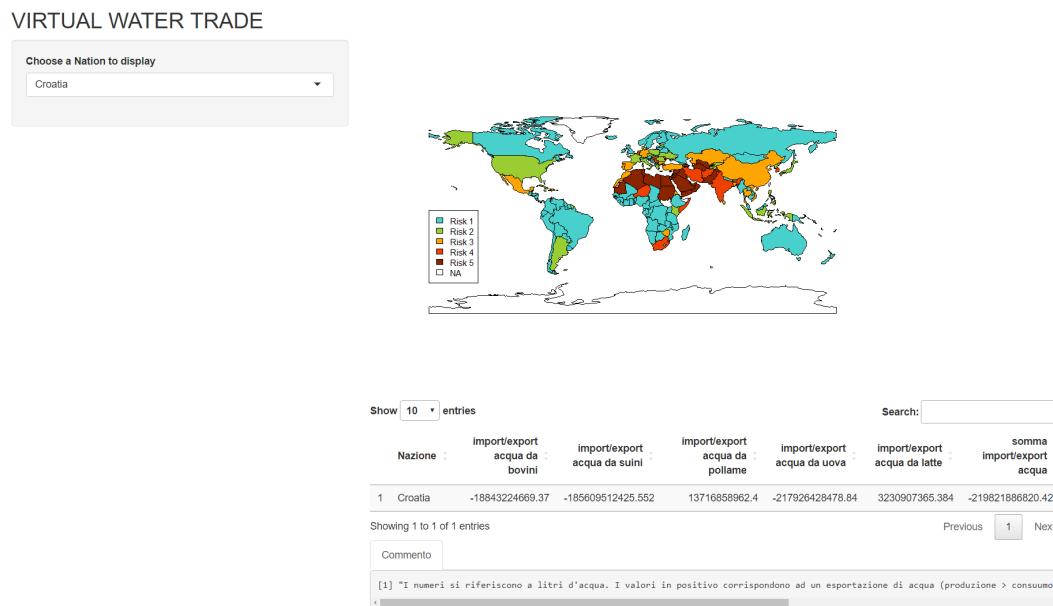


Figura 4.10: Dashboard Previsione con modello regressione lineare

Nell'esempio della dashboard qui riportato viene selezionata la Croazia come nazione, ed è possibile osservare come la Croazia nel 2017 abbia importato acqua virtuale dagli alimenti in esame, in quanto l'ultima colonna, quella riferita alla somma tra import ed export delle diverse variabili, ha valore negativo.

Conclusioni

In considerazione della vastità degli argomenti trattati in questa tesi, è possibile trarre alcune conclusioni che riflettono l'importanza della ricerca e le sue implicazioni per la comprensione del tema trattato.

Nel lavoro esposto sono stati presentati i temi del cambiamento climatico globale mediante l'analisi di variabili ambientali e del virtual water trade, settore fondamentale per lo studio dei flussi d'acqua ai nostri giorni.

Mediante diverse analisi si è potuto studiare come le variabili ambientali siano relazionate tra di loro e capire i rapporti causa-effetto che esistono tra esse.

Inoltre l'analisi congiunta dei due temi dell'analisi ha permesso di ottenere una profonda e precisa immagine della situazione riguardante la disponibilità di risorse idriche nel globo.

Un risultato importante è stata la costruzione di un modello che permette di prevedere la variazione della variabile evapotraspirazione potenziale a seguito di una variazione della variabile temperatura.

Mediante diversi processi di analisi di dati, che comprendono aggregazioni tra di essi e l'uso di test statistici, è stato possibile esaminare in maniera approfondita i dati in tutti gli aspetti di interesse, dall'aspetto geografico a quello temporale.

Infine, è stato portato a termine uno degli scopi principali della tesi, ovvero la creazione di dashboard interattive che permettano all'utente di esplorare i risultati del lavoro svolto in maniera autonoma e indipendente.

I limiti del lavoro che possono servire per possibili approfondimenti e lavori futuri sono un uso limitato dei dati circa il virtual water trade: è possibile estendere l'analisi utilizzando dati provenienti da fonti diverse che coinvolgano differenti processi produttivi al di fuori della filiera alimentare.

Il lavoro affrontato in questa tesi ha reso possibile ottenere una chiara prospettiva delle risorse idriche nel globo, sotto molti punti di vista. In relazione a ciò, il progetto può essere esteso, ad esempio mediante un possibile utilizzo di modelli previsionali meteorologici circa le precipitazioni.

In conclusione il lavoro svolto fornisce all'utente una chiara e approfondita analisi delle risorse idriche globali mediante lo studio dei temi del cambiamento climatico globale e del virtual water trade che è stato possibile realizzare attraverso l'uso di open data.

Bibliografia

- [1] Ruth Cano-Corres, Javier Sánchez-Álvarez e Xavier Fuentes-Arderiu. «The effect size: beyond statistical significance». In: *Ejifcc* 23.1 (2012), p. 19.
- [2] Winston Chang et al. *shiny: Web Application Framework for R*. R package version 1.7.2. 2022. URL: <https://CRAN.R-project.org/package=shiny>.
- [3] National Centers for Environmental Information. *Potential Evapotranspiration*. URL: <https://www.ncei.noaa.gov/access/monitoring/dyk/potential-evapotranspiration>.
- [4] Arjen Y Hoekstra e Pin Q Hung. «Virtual water trade». In: *Proceedings of the international expert meeting on virtual water trade*. Vol. 12. 2003, pp. 1–244.
- [5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: <https://www.R-project.org/>.
- [6] Hannah Ritchie, Pablo Rosado e Max Roser. «Meat and Dairy Production». In: *Our World in Data* (2017). <https://ourworldindata.org/meat-production>.
- [7] Hannah Ritchie e Max Roser. «Water Use and Stress». In: *Our World in Data* (2017). <https://ourworldindata.org/water-use-stress>.
- [8] Max Roser e Lucas Rodés-Guirao. «Future Population Growth». In: *Our World in Data* (2013). <https://ourworldindata.org/future-population-growth>.
- [9] Kurt Schmidheiny e Universität Basel. «The multiple linear regression model». In: *Short Guides to Microeometrics, Version 20* (2013), p. 29.
- [10] George AF Seber e Alan J Lee. «Polynomial regression». In: *Linear Regression Analysis* (2003), pp. 165–185.
- [11] G Tsakiris e HJEV Vangelis. «Establishing a drought index incorporating evapotranspiration». In: *European water* 9.10 (2005), pp. 3–11.
- [12] C. J. Willmott e K. Matsuura. *Terrestrial Air Temperature and Precipitation*. URL: http://climate.geog.udel.edu/~climate/html_pages/download.html#ghcn_T_P_clim3.
- [13] C. J. Willmott e K. Matsuura. *Terrestrial Air Temperature and Precipitation*. URL: http://climate.geog.udel.edu/~climate/html_pages/download.html#ghcn_T_P_clim3.
- [14] C. J. Willmott e K. Matsuura. *Terrestrial Air Temperature and Precipitation*. URL: http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalTsT2017.html.

BIBLIOGRAFIA

- [15] C. J. Willmott e K. Matsuura. *Terrestrial Air Temperature and Precipitation*. URL: http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalTsP2017.html.
- [16] C. J. Willmott e K. Matsuura. *Terrestrial Air Temperature and Precipitation*. URL: <https://climatedataguide.ucar.edu/climate-data/global-land-precipitation-and-temperature-willmott-matsuura-university-delaware>.
- [17] C. J. Willmott e K. Matsuura. *Terrestrial Water Budget Data Archive: Monthly Time Series (1900 - 2017)*. URL: http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalWbTs2017.html.
- [18] C. J. Willmott e K. Matsuura. *Terrestrial Water Budget Data Archive: Monthly Time Series (1900 - 2017)*. URL: http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalWbTs2017.html.