POLITECNICO
MILANO 1863
DIPARTIMENTO DI
INGEGNERIA CIVILE E AMBIENTALE

Analisi statistica della relazione tra consumo di suolo e danni da disastri ambientali in Europa (2006–2018)

TESI DI LAUREA MAGISTRALE IN
ENVIRONMENTAL ENGINEERING
INGEGNERIA PER L'AMBIENTE E IL TERRITORIO

Autore: **Flavio Carta**

Matricola: 233423

Relatore: Ilenia Epifani

Co-relatore: Arianna Azzellino

Anno Accademico: 2024-25

Abstract

Negli ultimi decenni l'Europa ha registrato un aumento della frequenza e dell'intensità degli eventi estremi, in un contesto caratterizzato da cambiamenti climatici e crescente pressione antropica sul territorio. In tale scenario, il consumo di suolo è frequentemente indicato come fattore di vulnerabilità potenzialmente in grado di amplificare gli impatti dei disastri naturali.

Il presente lavoro analizza la relazione tra consumo di suolo e danni da disastri ambientali in Europa nel periodo 2006–2018, con l'obiettivo di valutare il peso relativo dei fattori territoriali (impermeabilizzazione del suolo e popolazione) rispetto a quelli climatici (temperature e precipitazioni) nel determinare decessi e perdite economiche. L'analisi si basa sull'integrazione di diversi dataset: EM-DAT per la registrazione dei disastri e dei relativi danni, GDIS per la geolocalizzazione sub-statale degli eventi, dati Copernicus/EEA sull'impermeabilizzazione del suolo, informazioni demografiche provinciali e variabili climatiche (temperature e precipitazioni).

I risultati mostrano che, alla scala nazionale, il consumo di suolo non presenta un effetto diretto e statisticamente robusto sulla probabilità di danno. L'elevata correlazione del danno a scala nazionale con la popolazione e la correlazione di quest'ultima con la percentuale di superficie urbanizzata riducono la capacità di isolare un contributo autonomo dell'impermeabilizzazione. Anche le variabili climatiche risultano solo parzialmente significative nel contesto aggregato alla scala nazionale. Alla scala provinciale si osserva un miglioramento della capacità discriminatoria dei modelli (AUC compresa tra 0,63 e 0,68), ma l'effetto del consumo di suolo rimane debole rispetto a quello di altri fattori.

Nel complesso, l'analisi evidenzia che la scala spaziale di osservazione, la qualità dei dati e la struttura di correlazione tra le variabili influenzano in modo determinante l'interpretazione delle relazioni tra pressione antropica e danni da disastri ambientali, confermando però che il consumo di suolo ne rappresenti un fattore rilevante.

Abstract in English

In recent decades, Europe has experienced increasing exposure to extreme climatic events, in a context characterized by climate change and a progressive increase in anthropogenic pressure on land. In this scenario, land consumption is often considered a factor that can amplify the impacts of natural disasters, both by increasing exposure and by altering local hydrological and microclimatic balances. Quantifying its contribution in relation to other factors therefore requires an integrated analysis based on observed data.

This thesis analyses the relationship between land consumption and disaster-related damages in Europe over the period 2006–2018, with the objective of evaluating the relative role of territorial factors — soil sealing and population — compared to climatic factors — temperature and precipitation — in determining human losses and economic damages. The study integrates several harmonized European datasets: EM-DAT for recording events and impacts, GDIS for sub-national georeferencing, Copernicus/EEA data on soil sealing, provincial demographic information, and climatic variables.

At the national scale, soil sealing does not show a statistically robust direct effect, partly due to its strong correlation with population which is instead a very strong predictor for the damages. At the provincial scale, the discriminatory ability of the models improves, although the effect of soil sealing remains moderate with respect to other factors of variability. The results show that spatial scale and data quality are key to understanding how human activity relates to disaster damage, with finer models confirming soil sealing as a predictor of such damage.

Indice

1 Introduzione

"*Disastro: una grave interruzione del funzionamento di una comunità o di una società, a qualsiasi scala, dovuta a eventi pericolosi che interagiscono con condizioni di esposizione, vulnerabilità e capacità, e che porta a una o più delle seguenti conseguenze: perdite e impatti umani, materiali, economici e ambientali*" (UNGA, 2016). Questa è la definizione scientificamente condivisa di disastro. Tale definizione rende chiare tre caratteristiche fondamentali dei disastri:

a) un disastro consiste in un danno, misurabile, all'umanità o all'ambiente;
b) un disastro è causato da un evento esterno al sistema;
c) l'evento esterno costituisce una condizione necessaria ma non sufficiente: il legame tra l'evento esterno e il verificarsi del disastro, è costituito dalle condizioni dell'area che ne è colpita: la sua esposizione al rischio e la sua vulnerabilità.

Il tipo specifico di danno, la sua intensità, così come il tipo di evento esterno che causa il disastro, nonché i significati di esposizione e vulnerabilità, cambiano in base al tipo di disastro analizzato, ma il concetto di legame tra questi fattori risulta comune a ogni tipo di evento disastroso.

Nel caso specifico dei disastri naturali, che saranno oggetto di questo elaborato, l'evento esterno è costituito da uno o più eventi meteorologici, o dalle loro conseguenze, mentre le caratteristiche di esposizione e vulnerabilità della zona sono parzialmente di origine naturale e parzialmente di origine antropica. Tutte queste caratteristiche, così come i pesi che le legano al verificarsi di un disastro naturale, possono variare a livello spaziale, temporale, e a seconda della scala analizzata.

L'Integrated Research on Disaster Risk (IRDR) suddivide i disastri naturali, ossia quella categoria di disastri il cui evento scatenante è di origine naturale, in queste categorie,

"

a) *Geofisico: pericolo originato da processi geologici.*
b) *Idrologico: pericolo causato dal verificarsi, dal movimento e dalla distribuzione di acque dolci e salate, sia superficiali sia sotterranee.*
c) *Meteorologico: pericolo causato da fenomeni meteorologici estremi di breve durata, a scala micro- o mesoscala, che si sviluppano e persistono da pochi minuti a pochi giorni.*
d) *Climatologico: pericolo causato da processi atmosferici persistenti, a scala meso- o macroscopica, associati a variabilità climatica che va da quella intra-stagionale fino a quella multidecennale.*

e) *Biologico: pericolo causato dall'esposizione a organismi viventi e/o alle loro sostanze tossiche (ad esempio veleno o muffe), oppure a malattie trasmesse da vettori da essi veicolate*. Esempi *includono fauna e insetti velenosi, piante tossiche, fioriture algali e zanzare portatrici di agenti patogeni quali parassiti, batteri o virus (ad esempio la malaria).*
f) *Extraterrestre: pericolo causato da asteroidi, meteoroidi e comete quando transitano in prossimità della Terra, entrano nell'atmosfera terrestre e/o colpiscono la superficie, oppure da variazioni delle condizioni interplanetarie che influenzano la magnetosfera, ionosfera e termosfera terrestri.*

" (Integrated Research on Disaster Risk, 2014)

Il presente elaborato si concentra sui disastri di tipo idrologico, climatologico, e meteorologico, data la loro grande diffusione di questi disastri, soprattutto negli ultimi anni, e le loro sostanziali somiglianze nei meccanismi di generazione, essendo tutti e tre legati a fenomeni atmosferici.

L'estensione spaziale di questo elaborato è l'Europa continentale, grazie alla quantità e all'omogeneità dei dati esistenti circa tale realtà, rese possibili grazie ai numerosi tentativi di armonizzazione di tali dati operati dall'Unione Europea.

1.1. Evoluzione dei danni da disastri naturali

In base alla definizione di disastro, si osserva che esso è direttamente correlato al verificarsi di un danno. Per poter parlare di eventi disastrosi è, quindi, necessario definire delle metriche di misurazione di tali danni, che siano condivise, comparabili, e facilmente misurabili. Nonostante l'enorme importanza posseduta dalla conservazione della biodiversità, essa non è molto considerata nelle pubblicazioni ufficiali. Questo accade poiché la definizione di indicatori capaci di calcolare le perdite, richiederebbe un lavoro sito specifico, e l'ottenimento di una notevole quantità di dati. Per questo motivo, le metriche di danno più comuni a livello internazionale ed europeo sono quelli che colpiscono più direttamente l'uomo: il numero di decessi, e i danni economici patiti.

Secondo analisi condotte a scala europea, grazie ai dati del dataset CATDAT da parte della EEA, si osserva che, negli ultimi anni si sono verificati eventi naturali disastrosi, capaci di produrre intensi danni economici. Nello specifico il 2021, il 2022 e il 2002 sono gli anni in cui si sono concentrati i maggiori danni economici del periodo considerato (che arriva fino al 1980), rispettivamente stimati attorno a 63, 55 e 47 miliardi di euro.

A livello, invece, di decessi, lo stesso studio stima per esempio che, in est Europa, tra il 2020 e il 2023 si sono verificati circa 9000 decessi a causa di questi disastri, che superano, in un intervallo temporale molto più breve, i circa 6000 stimati tra il 2000 e il 2009 nella stessa area

Nello stesso studio si può osservare una caratteristica che complica l'analisi delle tendenze relative ai danni ambientali, la loro disomogeneità spaziale e temporale. Come è stato osservato, sia il 2002 che il 2021 sono stati anni in cui i disastri ambientali hanno provocato costi elevati. Tuttavia, proprio a causa della scarsa linearità di questi eventi, anni come il 2006, o il 2020, che ha preceduto l'anno in cui si è rilevato il massimo attuale, hanno

registrato danni relativamente ridotti (nel 2006 circa 5 miliardi, nel 2020 circa 15 miliardi di euro stimati). (European Environmental Agency, 2024)

Sempre per la stessa caratteristica del dato bisogna anche osservare, come accade per esempio in questo studio, a scala globale, circa i decessi causati dalle basse temperature, che un numero superiore di eventi, non sempre si accompagna ad un numero superiore di danni. In questo studio, infatti, si osserva come tra il 2000 e il 2009 siano avvenuti circa 80 eventi, ed hanno causato circa 6800 decessi, mentre tra 2010 e 2019 si son verificati circa 115 eventi, ed hanno portato a circa 4500 decessi. (Ocal, 2025)

Nonostante questo fatto, causato dalla diversa gravità degli eventi, è necessario ricordare che ad ogni evento dannoso, soprattutto se riportato in un database internazionale, corrisponde un danno, anche se, in taluni casi, esso potrebbe non essere riportato.

1.2. Evoluzione degli eventi climatici

Il disastro naturale è identificato dal danno che produce, ed è per questo importante studiarli e quantificarli. La causa del disastro stesso è, invece, un fenomeno meteorologico, e dunque l'analisi delle tendenze di accadimento di questi eventi non può prescindere dall'analisi delle tendenze delle loro cause. In questi decenni l'intensità e la frequenza dei fenomeni meteorologici estremi stanno variando, a causa del cambiamento climatico.

I disastri naturali che sono oggetto di questo studio sono causati direttamente, dalla quantità e dalla frequenza delle precipitazioni, dall'andamento delle temperature, e dalla velocità dei venti. Sebbene tutti questi fenomeni naturali siano, per loro natura, soggetti a variazioni locali difficilmente modellabili, presentano un andamento medio predicibile nel lungo periodo. Tale andamento medio mostra, per tutte le metriche elencate, evidenti segni di variazioni rispetto all'epoca preindustriale. Tali variazioni sono spesso in senso peggiorativo, comparando quindi un incremento del rischio, e risultano causate da attività umane.

Parlando della principale, e più evidente, variazione, ossia la temperatura media annua globale, è evidente una tendenza all'incremento. Tale tendenza ha portato il 2024 ad essere l'anno più caldo mai registrato dal 1850, con una differenza di temperatura superficiale, rispetto alla temperatura media in tutto il ventesimo secolo, di oltre 1,2°C. Inoltre, i 10 anni più caldi coincidono con il decennio 2015-2024 (il dato presentato non include il 2025). (Dahlman, 2025)

Questo incremento di temperatura media causa modifiche in numerosi equilibri climatici, fisici e biologici, che possono riflettersi in disastri ambientali sotto diverse forme. Alcuni esempi sono:

a) Causa ondate di calore più intense e più durature, che mettono in pericolo la salute delle categorie più fragili, o degli abitanti delle zone più calde del pianeta;
b) Modifica il ciclo idrologico, causando un maggior scioglimento dei ghiacciai, con conseguente liberazione di grandi quantità di acqua, ed innalzamento del livello medio dei mari, causando dissesti nelle zone costiere;

c) Modifica il clima, incrementando la probabilità di eventi meteorologici estremi, per durata o intensità, che andrebbero ad impattare su territori non abituati a simili eventi, causando dunque danni notevolmente superiori alle rilevazioni storiche;
d) Modifica la produttività dei terreni, in quanto al variare delle condizioni climatiche, varia la produttività delle specie vegetali che vi sono coltivate, causando dunque danni economici potenzialmente elevati.

Questi mutamenti possono combinarsi con altri, anche da loro stessi causati, secondo metriche non lineari, e non facilmente prevedibili. L'azione di più forzanti ambientali sul medesimo territorio, nello stesso momento o in periodi successivi, può portare ad un incremento dei danni patiti dal territorio medesimo.

L'impatto del cambiamento climatico non è omogeneo, ma è il prodotto di numerose specificità locali. Uno dei continenti su cui l'incremento di temperatura risulta generalmente maggiore è l'Europa, nello specifico l'area sud-orientale. In generale la temperatura media in Europa è cresciuta di oltre 2°C rispetto ai livelli preindustriali, il che causa un incremento nel numero di eventi estremi, e una serie di modifiche del clima, che causano una frequenza e un'intensità di disastri in costante aumento.

Negli ultimi decenni si osserva un incremento documentato della frequenza e dell'intensità di diversi fenomeni climatici estremi, e dei disastri da essi causati, sebbene con marcata variabilità spaziale e temporale. Essi riempiono sempre maggiormente i principali mezzi di informazione, e svuotano sempre maggiormente le casse pubbliche. Tutti gli scenari futuri, relativi al cambiamento climatico, stimano un incremento ulteriore delle problematiche, da qui ai prossimi anni, almeno fino a fine secolo. L'unico fattore in discussione, che distingue l'una o l'altra proiezione, riguarda l'entità di questi impatti, la quale può essere mitigata unicamente con politiche urgenti, di lungo periodo, e frutto di accordi multilaterali.

1.3. Definizione delle caratteristiche di esposizione

Come precedentemente spiegato, la presenza del solo evento esterno non è sufficiente a produrre come risultato un evento disastroso. Per comprendere la presenza, e l'intensità, del disastro naturale eventualmente verificatosi, è necessario che l'evento climatico impatti su un'area, e interagisca con le sue caratteristiche di esposizione, vulnerabilità, e capacità di adattamento. Tali tre caratteristiche sono definite dall' Intergovernamental Panel for Climate Change (IPCC) come:

"

a) *Esposizione: la presenza di persone, mezzi di sussistenza, specie o ecosistemi, funzioni, servizi e risorse ambientali, infrastrutture, oppure beni economici, sociali o culturali, in luoghi e contesti che potrebbero essere colpiti negativamente.*
b) *Vulnerabilità: la propensione o predisposizione a subire effetti negativi. La vulnerabilità comprende una varietà di concetti ed elementi, tra cui la sensibilità o suscettibilità al danno e la mancanza di capacità di farvi fronte e di adattarsi.*

c) *Capacità di adattamento: la capacità di sistemi, istituzioni, esseri umani e altri organismi di adeguarsi a potenziali danni, di cogliere opportunità oppure di rispondere alle conseguenze (MA, 2005).*

" (Intergovernmental Panel on Climate Change, 2022)

Nonostante la chiarezza delle definizioni, l'ottenimento di indicatori affidabili, ed utili, per definire queste caratteristiche, è complicato. La definizione di tali indicatori, infatti, richiede un'ottima conoscenza dei meccanismi che legano l'accadimento di un evento estremo, al verificarsi del disastro corrispondente. Questi meccanismi hanno, spesso, una natura molto locale, molto rapida, e dunque relativamente poco nota. Inoltre, per utilizzare utilmente un dato indicatore, specialmente in un contesto internazionale come quello europeo, è necessario che esso sia ben noto in tutto il territorio, e su una scala spaziale e temporale adeguata.

In generale la definizione di indicatori di esposizione risulta più semplice rispetto alle altre. Questo in quanto l'esposizione di un territorio consiste unicamente della presenza, in quel territorio, di persone, o beni, che possano astrattamente risultare danneggiati da un evento estremo. Di conseguenza i principali indicatori di esposizione sono la densità di popolazione e l'uso del suolo. (European Environment Agency, 2024)

Gli indicatori relativi alla vulnerabilità di una data area dipendono maggiormente dal tipo di disastro interessato, e dunque dal tipo di evento avverso che ne è causa. Essi sono spesso meno facilmente identificabili e misurabili rispetto agli indicatori di esposizione, oltre che essere più variegati e spesso non calcolati in maniera omogenea su vasti territori. Esempi di tali indicatori possono essere l'età media della popolazione, la sua ricchezza, l'età e lo stato di conservazione degli edifici, ed altre caratteristiche specifiche capaci di determinare, a parità di soggetti o beni esposti, un diverso rischio di danno.

Infine, ottenere indicatori numerici affidabili e condivisi che definiscano le capacità di adattamento di una determinata area geografica risulta sostanzialmente impossibile. Tali capacità, infatti, sono il frutto dell'efficacia ed efficienza di decisioni di natura organizzativa e politica, che devono rivelarsi di lungo periodo e vasto impatto. Come proxy di questa informazione è unicamente possibile valutare l'evoluzione, in un lungo arco temporale, di altri indicatori antropici, come il grado di urbanizzazione.

Per queste complicazioni, l'elaborato non prenderà in considerazione questi ultimi due aspetti, concentrandosi unicamente sui due principali indicatori di esposizione: popolazione e uso del suolo.

1.3.1. Popolazione

Nell'ambito europeo, i dati demografici sono raccolti in modo sistematico e armonizzato da Eurostat, sulla base delle comunicazioni annuali degli Stati membri e di revisioni periodiche basate sui censimenti. Ciò rende la popolazione un indicatore particolarmente adatto ad analisi comparative a scala nazionale e subnazionale. Inoltre, sebbene la popolazione sia ufficialmente fornita come dato aggregato (su scala comunale, provinciale, regionale e statale), si realizzano modelli di distribuzione della popolazione sul territorio, per poter mappare con precisione ancora superiore l'esposizione specifica di porzioni della popolazione

stessa. Esempi di questi modelli, che tuttavia non verranno utilizzati nel presente elaborato, sono le GEOSTAT population grids, prodotte da GISCO (servizio GIS di Eurostat).

La distribuzione della popolazione in Europa è fortemente eterogenea sia tra gli Stati membri sia all'interno di ciascun Paese. A fronte di una densità media europea pari a circa 106 abitanti per km² nel 2020, e valori medi statali che variano tra circa 10 e circa 500 abitanti per km², le aree urbane maggiori presentano valori di densità nettamente superiori, spesso eccedenti i 2 000 abitanti per km², mentre vaste aree rurali mostrano densità molto più basse. (Eurostat, 2024)

Dal punto di vista temporale, la popolazione non subisce variazioni brusche nel breve periodo, ma può registrare cambiamenti significativi nel medio e lungo termine, legati a dinamiche demografiche strutturali e a fenomeni migratori. A livello europeo la popolazione è, mediamente, cresciuta di 1,3 milioni di persone l'anno dal 1960 al 2025, portando il totale da 354 a 450 milioni di persone.

Questo incremento di popolazione non è causato da un incremento delle nascite. Una serie di fattori, infatti, ha fatto diminuire il numero delle nascite, a livello europeo, da oltre 6,5 milioni di bambini nel 1961, a 3,88 milioni nel 2023. Da questo fatto consegue che un ruolo di notevole importanza nella variazione della popolazione è l'immigrazione, sia interna che esterna. (Eurostat, 2024)

Si osserva, infatti, anche negli anni più recenti (come, per esempio, comparando la popolazione del 2021 con quella del 2011) che molte aree rurali (come aree dell'est Europa, del sud Italia e vasti territori in Spagna) tendano verso un leggero spopolamento, a favore delle città più importanti.

Questi fenomeni portano con sé modifiche al profilo di rischio territoriale che colpiscono, con potenziali implicazioni sul profilo di rischio delle aree interessate. Esse, infatti, possono andare incontro ad abbandono, ad incuria, e dunque può perdersi la funzionalità di sistemi di prevenzione dei disastri naturali, siti in quelle aree.

Eseguire proiezioni accurate e a lungo termine circa l'andamento della popolazione nelle varie zone d'Europa risulta complesso, a causa delle differenti ipotesi circa la crescita demografica delle aree. Si prevede, comunque, una minore crescita demografica interna rispetto alle aspettative e alle tendenze di metà e fine secolo scorso, a cui potrà contrapporsi l'impatto sei fenomeni migratori interni ed esterni all'Unione Europea, causati da squilibri politici e da cambiamenti climatici.

1.3.2. Uso del suolo

L'utilizzo che viene fatto del territorio, quindi del suolo, può influire, positivamente o negativamente, sul livello di esposizione e sul profilo di rischio della popolazione residente nell'area interessata, o in aree ad essa connesse. La modifica di un'area naturale in area forestale, o in campo coltivato a monocultura, o, peggio, la sua cementificazione, altera completamente le capacità di quel volume di suolo di contribuire alla limitazione dei danni causati da eventi estremi, e dunque incrementa il rischio del verificarsi di disastri naturali.

La variazione di uso del suolo, sebbene costituisca un aspetto fondamentale relativamente alla variazione dell'esposizione, e della vulnerabilità, delle aree circostanti, risulta spesso complessa da sintetizzare e definire in variabili specifiche. Essa, infatti, è una caratteristica legata al territorio, ed al tempo, e può assumere molte forme. Sono esempi di variazione d'uso del suolo una variazione colturale di un'area agroforestale, così come la trasformazione di un'area in area urbana in area verde, o la costruzione di un'infrastruttura (come una ferrovia, una strada, o un parcheggio), oltre che logicamente la costruzione di un edificio. L'impatto che una variazione di uso del suolo produce, in termini di rischio, varia a seconda del tipo di variazione stessa, oltre che delle caratteristiche dell'area coinvolta.

La trasformazione che più riduce le capacità di un suolo di prevenzione di danni ambientali è la sua impermeabilizzazione permanente. Un suolo impermeabilizzato, per qualsiasi causa, perde completamente le sue caratteristiche e le sue funzionalità. Diviene inadatto, per esempio, a produrre nutrienti o altre utilità, a sostenere un ecosistema, a contenere carbonio, e a filtrare e trattenere acqua. Queste modifiche alterano, dunque, l'equilibrio dell'area interessata e di quella circostante, incrementando per esempio, la velocità del deflusso idrico, e modificando il microclima a favore di temperature maggiori. Ciò incrementa la probabilità del verificarsi di fenomeni erosivi, franosi, o inondazioni, a valle della zona interessata, oltre che l'aumento di problematiche legate al caldo nell'area stessa (Commissione europea, 2012). Tuttavia, la quantificazione empirica di tale relazione su scala ampia non è univoca, poiché il consumo di suolo risulta frequentemente correlato ad altri fattori territoriali, in particolare alla densità di popolazione, rendendo complessa l'identificazione di un contributo autonomo.

Inoltre, dal momento che la causa prima di impermeabilizzazione del suolo è l'utilizzo antropico diretto, mediante la costruzione di edifici, l'impermeabilizzazione incrementa anche l'esposizione ad eventuali danni ambientali nell'area. Tuttavia, l'incremento demografico non è sempre sufficiente a giustificare l'incremento di suolo impermeabilizzato a livello europeo (per esempio, tra il 1990 e il 2015 la popolazione europea è aumentata del 7% mentre il suolo consumato è aumentato del 18% (Melchiorri, 2018))

A causa di questa molteplicità di fattori di rischio, oltre che della semplicità di definizione, e di misurazione, dell'indicatore, spesso la quantificazione del suolo impermeabilizzato può essere utilizzata come proxy di misure più complesse legate all'uso del suolo, o alla sua variazione. Questo approccio verrà seguito anche nel presente elaborato.

Il grado di impermeabilizzazione di suolo è, inoltre, molto legato alla crescita economica della regione interessata. Dal momento che globalmente non vi sono previsioni di limitare la crescita economica nel corso dei prossimi decenni, sarà probabilmente molto difficile, mantenendo gli attuali livelli di urbanizzazione, contenere i rischi ambientali legati alla stessa.

1.4. Sforzi internazionali

Negli ultimi anni, l'Unione Europea ha progressivamente inserito il contrasto al cambiamento climatico e ai suoi effetti tra le proprie priorità. Tale orientamento è stato formalizzato con l'adozione del "European Green Deal" nel 2019, che definisce un quadro strategico per affrontare le sfide climatiche e ambientali (Commissione europea, 2019). In questo quadro, l'adattamento al cambiamento climatico rappresenta un pilastro delle politiche europee di gestione del rischio.

La base giuridica delle politiche europee sul tema di tale adattamento è la Legge europea sul clima, adottata nel 2021 (…). La normativa impone ai singoli Stati, di incrementare, continuamente, le proprie capacità di adattamento, riducendo al contempo le vulnerabilità locali. La normativa prevede, inoltre, l'adozione comunitaria di una strategia coerente con l'Accordo di Parigi e l'istituzione di un processo di monitoraggio periodico dei progressi compiuti. (UE, 2021).

Un ruolo centrale è svolto anche dalla produzione e dalla diffusione di conoscenza scientifica e dati. In tale ambito, la Commissione europea finanzia il Copernicus Climate Change Service (C3S), che fornisce informazioni autorevoli sul clima passato, presente e futuro, e l'iniziativa Destination Earth (DestinE), finalizzata allo sviluppo di gemelli digitali della Terra in grado di simulare scenari climatici e valutare i rischi a diverse scale spaziali e temporali. Questi strumenti costituiscono una base informativa essenziale per le valutazioni del rischio climatico e per il supporto alle decisioni politiche.

Parallelamente, gli Stati membri dell'UE sono tenuti a rendicontare periodicamente le proprie azioni di adattamento nell'ambito del Regolamento sulla governance dell'Unione dell'energia e dell'azione per il clima. I risultati di tali processi di reporting sono sintetizzati in rapporti regolari dell'Agenzia europea dell'ambiente, che forniscono un quadro aggiornato dei progressi compiuti e delle criticità ancora presenti a livello nazionale e dell'Unione nel suo complesso (EEA, 2022; EEA, 2023).

1.5. Scopo ed organizzazione dell'elaborato

Il presente elaborato si colloca nel contesto dei cambiamenti climatici descritto e si pone il seguente obiettivo: analizzare l'importanza relativa dei fattori ambientali esterni, quali precipitazioni e temperature, rispetto ai fattori territoriali di esposizione, quali popolazione e suolo impermeabilizzato, nella determinazione di decessi e/o danni economici a seguito di disastri ambientali di tipo idrologico, meteorologico, climatologico.

L'analisi si concentra sul territorio degli Stati appartenenti a EEA37, e sugli anni compresi tra il 2006 e il 2018.

Il lavoro si articola come segue. Nel Capitolo 2 sono descritti i materiali e le basi dati utilizzate, con particolare attenzione alle caratteristiche informative e ai limiti dei dataset. Il Capitolo 3 illustra i metodi adottati per la costruzione delle variabili e per l'analisi statistica. Nel Capitolo 4 sono presentati i risultati dell'analisi esplorativa, mentre il Capitolo 5 riporta

i modelli stimati a scala nazionale e provinciale. Il Capitolo 6 conclude il lavoro discutendo criticamente i risultati ottenuti e le principali limitazioni dell'analisi.

2 Dati

Per analizzare il legame tra presenza umana e danni causati da disastri naturali sono necessarie due tipologie di informazioni: da un lato dati sulla pressione antropica, misurabile attraverso l'impermeabilizzazione del suolo e la popolazione, e dall'altro dati sugli impatti effettivamente prodotti dagli eventi estremi.

La stima dell'impermeabilizzazione del suolo e della popolazione a scala provinciale europea è relativamente agevole, grazie alla disponibilità di dataset armonizzati a livello comunitario. Più complesso risulta invece ottenere informazioni affidabili, omogenee e comparabili sui danni causati da disastri naturali. Sebbene i mezzi di informazione riportino frequentemente cifre relative a tali eventi, i progetti che raccolgono, verificano e uniformano sistematicamente questi dati su larga scala sono rari.

A livello nazionale ed europeo esistono mappe dettagliate di vulnerabilità ai rischi naturali (a titolo esemplificativo, rischio di alluvione, frana o incendio). Tuttavia, tali strumenti non sono generalmente basati su danni effettivamente osservati, bensì su modelli di rischio costruiti combinando: fattori locali di esposizione, modelli climatici o idrometeorologici, da cui deriva la stima dell'hazard, e modelli fisici e statistici che collegano hazard ed esposizione. Il risultato è un indicatore di vulnerabilità teorica, che non riflette necessariamente l'intensità o la distribuzione dei danni realmente registrati.

L'obiettivo di questo studio è differente: verificare in quale misura i danni osservati siano associati al consumo di suolo in una determinata area. Ciò richiede dataset che memorizzino eventi realmente accaduti, con un livello di dettaglio spaziale almeno sub-statale e con informazioni strutturate sugli impatti umani ed economici.

A scala europea, le banche dati con tali caratteristiche sono limitate. Tra le principali si annoverano:

- CATDAT (Catastrophes Database), sviluppato presso il Karlsruhe Institute of Technology (KIT), con l'obiettivo di ricostruire e uniformare informazioni storiche su decessi, danni economici, caratteristiche fisiche degli eventi e impatti infrastrutturali, integrando fonti eterogenee quali archivi storici, report governativi, dati assicurativi e letteratura tecnica. La componente più completa riguarda gli eventi sismici; secondo la descrizione originale, CATDAT contiene oltre 12.200 terremoti, normalizzati e verificati tramite un processo di cross-checking (Daniell, Vervaeck & Khazai, 2011). Pur rappresentando una fonte di riferimento per studi storici sull'impatto dei terremoti, CATDAT non è un database open-access e le versioni liberamente disponibili risultano aggregate a livello nazionale, non sufficienti per analisi sub-statali.

- NatCatSERVICE, database globale delle catastrofi naturali sviluppato dalla compagnia di riassicurazione Munich Re, che include oltre 30.000 eventi registrati dal 1974 ad oggi, comprendendo terremoti, alluvioni, tempeste, incendi, frane e altri hazard naturali. Il database integra fonti eterogenee (rapporti governativi, dati assicurativi, media, agenzie internazionali) con finalità prevalentemente assicurative. Nonostante la ricchezza informativa, l'accesso completo richiede licenza commerciale e le informazioni pubblicamente disponibili sono generalmente presentate in forma aggregata, risultando quindi non adeguate ad analisi territoriali di dettaglio.
- EM-DAT (Emergency Events Database), database internazionale open-access gestito dal Centre for Research on the Epidemiology of Disasters (CRED), che fornisce informazioni coerenti e standardizzate sulle principali categorie di disastro a livello globale.

Considerati i vincoli di accessibilità, comparabilità e granularità spaziale, EM-DAT rappresenta la fonte primaria adottata nel presente lavoro. Nei paragrafi seguenti vengono descritte in dettaglio le caratteristiche del database e le modalità con cui esso è stato integrato con le altre basi dati utilizzate.

2.1. Dataset dei danni ambientali EM-DAT

Il principale database utilizzato per rappresentare gli impatti dei disastri naturali è EM-DAT: The International Disaster Database, sviluppato dal *Centre for Research on the Epidemiology of Disasters (CRED)* ed istituito nel 1988 con il supporto della Commissione Europea. EM-DAT è una banca dati internazionale che raccoglie informazioni relative a disastri naturali e tecnologici in tutto il mondo, integrando fonti ufficiali quali bollettini istituzionali, report governativi, pubblicazioni tecniche e comunicati di agenzie internazionali.

Per poter essere inserito in EM-DAT, un evento disastroso deve soddisfare almeno una delle seguenti condizioni:

a) aver causato almeno dieci morti;
b) aver colpito almeno cento persone;
c) aver richiesto la dichiarazione di stato di emergenza da parte delle autorità competenti;
d) aver comportato la richiesta di assistenza internazionale.

Questi criteri di inclusione garantiscono che gli eventi registrati nel database presentino una significativa rilevanza in termini di impatti sociali o economici.

2.1.1. Struttura informativa del dataset

Ogni evento presente in EM-DAT è corredato da un insieme di variabili descrittive, tra cui:

a) Uno Stato in cui si è verificato;
b) Un anno in cui si è verificato;

c) Una classificazione a cinque livelli (si veda sezione 1.1.2);
d) Se vi è stata dichiarazione dello stato di emergenza;
e) Se vi è stata richiesta di intervento internazionale.

Oltre a queste informazioni generali, il database può includere variabili aggiuntive relative agli impatti e alle caratteristiche dell'evento, quali:

a) Indicatori di danno umano, comprendenti:
   A. Decessi causati direttamente dal disastro (inclusi casi di morte da complicazioni causate dal disastro) e persone dichiarate disperse;
   B. Persone colpite, vale a dire persone danneggiate dal disastro in quanto:
      a) Feriti a causa dell'evento;
      b) Sfollati a causa dell'evento;
      c) Dichiarati genericamente come "colpiti" dalle fonti a cui EM-DAT fa riferimento.
b) Indicatori di danno economico, espressi in migliaia di dollari statunitensi, attualizzati all'anno corrente (quindi, nel caso di questa tesi, il 2025), divisi tra:
   a) Danno totale stimato causato dal disastro;
   b) Danno riconosciuto dalle compagnie di riassicurazione;
   c) Spesa necessaria per la ristrutturazione.
c) Informazioni relative alla magnitudo dell'evento, espresse in unità di misura dipendenti dalla tipologia del fenomeno (ad esempio scala Richter per i terremoti);
d) Localizzazione spaziale più precisa:
   A. Zona, a livello regionale (AID 2) o comunale (AID 3);
   B. Coordinate.
e) Localizzazione temporale più precisa:
   A. Data di inizio/fine;
   B. Ora di accadimento.

Gli indicatori di danno umano ed economico costituiscono le variabili di interesse principale ai fini del presente studio, in quanto permettono di quantificare gli impatti effettivamente osservati degli eventi estremi.

2.1.2. Classificazione dei disastri

EM-DAT organizza gli eventi secondo una classificazione gerarchica articolata su cinque sottogruppi di disastri naturali. Al livello più generale, i disastri si distinguono in due grandi categorie: disastri naturali e disastri tecnologici. Ciascuno dei due gruppi è ulteriormente suddiviso in sottogruppi, ognuno dei quali si compone di tre diversi livelli di tipi. La struttura gerarchica completa adottata dal database è riportata in Figura 2.1, mentre in Figura 2.2 è illustrata la suddivisione dei soli disastri naturali nei rispettivi sottotipi. Nel

presente studio l'analisi è limitata ai disastri naturali; ulteriori suddivisioni di livello inferiore non sono state considerate, in quanto avrebbero comportato una frammentazione eccessiva dell'informazione disponibile.

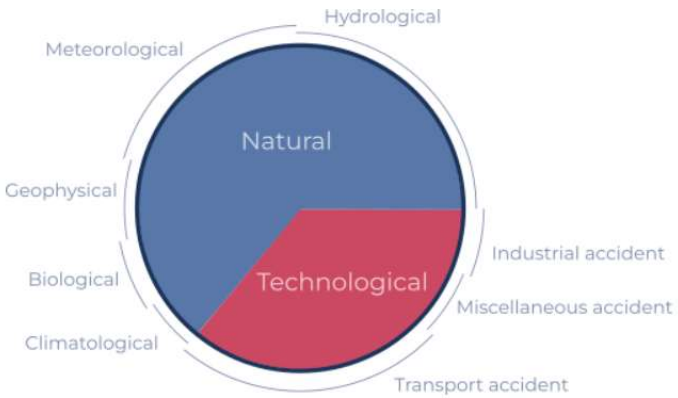

Figura 2.1: Tipi e sottotipi di disastro presenti in EMDAT, con relativa percentuale di presenza totale (fonte: www.emdat.be)

Figura 2.2: Divisione dei disastri naturali in sottotipi e gruppi operata da EM-DAT. Fonte: www.emdat.be

2.1.3. Risorse informative e qualità dei dati

I dati contenuti in EM-DAT provengono da una pluralità di fonti ufficiali e non ufficiali, tra cui bollettini governativi, rapporti della protezione civile, agenzie internazionali, media e documentazione tecnica. Le informazioni raccolte vengono periodicamente confrontate e verificate dal CRED, al fine di garantire coerenza e attendibilità.

La presenza di un numero elevato di fonti rappresenta un punto di forza del database, in quanto riduce la probabilità che eventi rilevanti non vengano registrati. Tuttavia, tale eterogeneità può comportare differenze metodologiche nella stima degli impatti, in particolare

per quanto riguarda i danni economici, che possono essere calcolati secondo criteri differenti a seconda della fonte.

Un ulteriore elemento critico riguarda la disponibilità degli indicatori di danno. Non tutti gli eventi presentano informazioni complete relative a decessi, popolazione colpita o danni economici. In diversi casi, i valori risultano mancanti a causa dell'assenza di dati nelle fonti originarie o della presenza di informazioni discordanti. Questo aspetto assume rilievo metodologico nella fase di analisi statistica, poiché limita la possibilità di utilizzare congiuntamente tutti gli indicatori disponibili.

Infine, la granularità spaziale originaria del database è limitata al livello statale. Per analisi a scala sub-nazionale, come quelle condotte nel presente studio, è necessario integrare EM-DAT con sistemi di geo-referenziazione dedicati, descritti nella sezione successiva.

L'effettiva distribuzione degli eventi nel periodo 2006–2018, la composizione per tipologia di disastro e la disponibilità degli indicatori di danno sono oggetto della successiva analisi esplorativa (Capitolo 4), volta a valutare l'idoneità del dataset all'impiego nei modelli statistici adottati.

2.2. Dataset di geo-referenziazione dei disastri GDIS

Uno dei maggiori difetti del dataset EM-DAT è la sua mancanza di informazione spaziale dettagliata: tutti i disastri sono identificati infatti unicamente dallo Stato e dall'anno di accadimento. Tale granularità risulta insufficiente per analisi territoriali che richiedono precisione sub-statale, come quelle condotte nel presente lavoro a scala provinciale. Infatti, all'interno di un medesimo Stato, le condizioni insediative, i rischi idrogeologici, climatologici, geofisici, e in generale le condizioni, possono variare notevolmente. Per questo è fondamentale potersi affidare ad un sistema per geo-referenziare in maniera più accurata i disastri.

Il database "GDIS: *Global Disasters Identifier System"* è un'elaborazione del database EM-DAT, resa disponibile in formato geospaziale attraverso il portale ufficiale NASA SEDAC (*Socioeconomic Data and Applications Center*). Esso è liberamente scaricabile e visualizzabile tramite strumenti GIS come Google Earth, consentendo l'associazione degli eventi disastrosi a coordinate geografiche e a unità amministrative più granulari rispetto alla singola scala statale. (Rosvold & Buhaug, 2021)

Lo scopo di GDIS è favorire l'utilizzo facilitato dei dati EM-DAT per analisi spaziali a scala sub-nazionale. Il risultato è stato ottenuto sfruttando parte delle informazioni circa la posizione contenute in EM-DAT stesso, verificandole, e trasformandole in dati geo-referenziati. Il database GDIS consiste di 38.953 punti georeferenziati, in WGS 84 (EPSG:4326), a cui corrispondono 9.924 disastri naturali, avvenuti tra il 1960 e il 2018, e registrati in EM-DAT. Ciascun punto corrisponde a un riferimento spaziale in coordinate geografiche WGS 84 (EPSG:4326) collegato a un singolo evento o a frazioni di esso. Per ogni punto georeferenziato sono disponibili i seguenti attributi principali, riportati nella Tabella 2.1.

Tabella 2.1: parametri noti per ogni punto in GDIS

Nome del parametro	Descrizione
Disasterno	Codice identificativo del disastro (usato per collegare GDIS a EM-DAT)
year	Anno in cui il disastro si è verificato
country	Stato in cui il disastro si è verificato
iso3	Codice iso3 dello Stato in cui il disastro si è verificato
level	Livello di precisione geografica della specifica area di accadimento (1 per la regione, 2 per la provincia, 3 per il comune)
adm1	Regione in cui il disastro è accaduto (sempre nota)
adm2	Provincia in cui il disastro è accaduto
adm3	Comune in cui il disastro è accaduto

La creazione di GDIS si basa sulle informazioni di localizzazione disponibile in EM-DAT e su procedure di verifica automatica e manuale che consentono di attribuire coordinate più precise agli eventi disastrosi. Tuttavia, è importante sottolineare che la precisione spaziale di ciascun punto in GDIS dipende strettamente dalla qualità delle informazioni originarie, cioè dalla precisione della localizzazione fornita da EM-DAT: per alcuni eventi, in particolare quelli con descrizioni generiche o prive di coordinate, la precisione non supera il livello regionale.

Inoltre, in alcuni casi, un singolo evento registrato in EM-DAT come avvenuto in un determinato Stato in un determinato anno può essere associato in GDIS a più punti georeferenziati, che condividono lo stesso codice. Ciò si verifica per fenomeni estesi sul territorio, come siccità o ondate di calore che coinvolgono più aree amministrative. Viceversa, eventi privi di sufficienti informazioni di localizzazione non sono rappresentati in GDIS.

Inoltre, dal momento che EM-DAT viene periodicamente aggiornato, mentre GDIS rappresenta uno snapshot elaborato a una data specifica, possono verificarsi situazioni in cui un evento di EM-DAT può venire aggiunto, o rimosso, senza che tale variazione sia applicata a GDIS.

2.3. Dataset di consumo di suolo

Un suolo è considerato impermeabilizzato (o "sigillato") quando la sua superficie viene coperta o modificata da materiali artificiali tali da comprometterne significativamente la naturale permeabilità, ossia la capacità di infiltrare acqua, scambiare gas con l'atmosfera, ospitare organismi viventi e svolgere le sue funzioni ecologiche.

In termini formali, la European Environment Agency (EEA) definisce il concetto di soil sealing come «*la modifica della natura del suolo in modo che si comporti come un mezzo impermeabile (ad esempio la compattazione da macchinari agricoli)*» oppure «*la copertura della superficie del suolo da materiali impermeabili artificiali quali cemento, asfalto, plastica*».

A partire da questa definizione, appare evidente che un suolo su cui insistono edifici, infrastrutture, strade o vaste aree pavimentate perde la capacità di trattenere acqua, di assorbire e filtrare infiltrazioni, di supportare biomassa e biodiversità, e di fungere da "tampone" idrico e biologico per l'ecosistema circostante.

Questa perdita di funzione ha rilevanza nel contesto dei danni ambientali per vari motivi:

a) Maggiore deflusso superficiale e rischio idraulico: quando il suolo non assorbe l'acqua, essa resta in superficie e può scorrere rapidamente verso valle, aumentando portate, picchi di piena e rischio di allagamenti.

b) Riduzione della ricarica delle acque sotterranee: un suolo impermeabilizzato impedisce l'infiltrazione dell'acqua nel sottosuolo, compromettendo le falde, il bilancio idrico e la disponibilità d'acqua nei periodi siccitosi.

c) Perdita di funzioni ecosistemiche e biodiversità: la copertura artificiale impedisce la vita del suolo (microrganismi, lombrichi, radici…) e affievolisce processi vitali come il ciclo dei nutrienti, la produzione di biomassa, la capacità di degradazione e immagazzinamento delle sostanze.

d) Compromissione della fertilità e perdita di suolo agricolo: quando aree fertile per l'agricoltura vengono urbanizzate e impermeabilizzate, si riduce la superficie produttiva disponibile, con conseguenti implicazioni per la sicurezza alimentare.

e) Incremento dell'esposizione al rischio: dal momento che l'impermeabilizzazione di suolo coincide spesso con densificazione urbana e maggiore concentrazione di popolazione e infrastrutture, in caso di eventi geologici o idraulici avversi (come alluvioni o esondazioni), il potenziale danno umano ed economico tende ad aumentare.

Per queste ragioni, la valutazione del legame tra suolo impermeabilizzato e danni ambientali non è solo opportuna, ma decisiva: l'impermeabilizzazione rappresenta una modalità diretta con cui l'intervento antropico altera il sistema terra-suolo-acqua-ecosistema, potenzialmente incrementando la vulnerabilità dell'ambiente e delle comunità agli eventi naturali avversi, e riducendo la capacità di resilienza e adattamento del territorio.

Per questa tesi, la quantificazione del suolo impermeabilizzato, è stata assunta dal dataset europeo "Imperviousness in Europe" (European Environment Agency (EEA), 2025). Tale dataset è rielaborazione del prodotto "High Resolution Layer: Imperviousness" (HRL: IMP) (European Environment Agency (EEA); Copernicus Land Monitoring Service, 2025) sviluppato nell'ambito del programma Copernicus Land Monitoring Service.

L'HRL: IMP fornisce, su base raster, la densità di impermeabilizzazione del suolo (Imperviousness Density, IMD), espressa come percentuale di superficie coperta da materiali artificiali su celle di 20 m × 20 m (ridotte a 10 m × 10 m a partire dal 2018). I dati sono disponibili per gli anni 2006, 2009, 2012, 2015 e 2018 (oltre al 2021, attualmente in fase di validazione).

Il dato di impermeabilizzazione è stimato sulla base di immagini satellitari multispettrali, prese dai satelliti Sentinel-2. A partire da tali dati, opportunamente sintetizzati, vengono calcolati opportuni indici, necessari a riconoscere la copertura del suolo in ogni pixel. Il più significativo tra gli indici calcolati è il Normalized Difference Vegetation Index (NDVI).

$$NDVI = \frac{Rifl.(NIR) - Rifl.(rosso)}{Rifl.(NIR) + Rifl.(rosso)}$$

Nella formula precedente il numeratore è la riflettanza della cella di territorio analizzata (riflettanza BOA) nella fascia dell'infrarosso vicino (NIR), a cui viene sottratta la riflettanza, sempre stimata come a livello del suolo, nella fascia del rosso

Questo indice è correlato direttamente alla presenza e alla produttività della vegetazione (le foglie, infatti, generalmente presentano elevati livelli di riflettanza nella banda dell'infrarosso vicino, mentre non ne presentano in quella del rosso visibile). Per tale motivo, pur essendo semplice da calcolare, risulta molto utilizzato nel telerilevamento. Tuttavia, questo indicatore non è sufficiente ad indicare la copertura di suoli non vegetati, che possono essere suoli nudi, rocce affioranti, o suoli realmente impermeabilizzati, e per tale motivo vengono calcolati ulteriori indici legati all'urbanizzazione. Grazie all'evoluzione rapida della tecnologia, i dati calcolati vengono inseriti in modelli di machine learning, che tra il 2006 e il 2021 sono andati incontro a progressiva evoluzione.

Per ulteriori informazioni si rimanda a questo elaborato tecnico: (Copernicus Land Monitoring Service, 2021)

Oltre al suolo impermeabilizzato ogni tre anni, vengono fornite mappe che mostrano l'evoluzione dell'impermeabilizzazione tra un triennio e il successivo. Oltre alla grezza differenza tra le due mappe di stato, attraverso analisi delle tendenze temporali degli indici utilizzati, si valuta se una data variazione di impermeabilizzazione di suolo risulta sensata, o è solo frutto di piccoli errori di calcolo.

L'EEA ha operato un'ulteriore trasformazione del dataset, che consiste nell'aggregazione spaziale dei dati originali raster, all'interno delle unità amministrative della classificazione NUTS 2021 (province o regioni), calcolando diversi indicatori sintetici:

a) Percentuale di area impermeabilizzata totale, ottenuta come media spaziale dei dati di percentuale di impermeabilizzazione;
b) km^2 di area impermeabilizzata totale in un dato anno;
c) m^2 pro capite di area impermeabilizzata totale, calcolata come rapporto tra km^2 di area impermeabilizzata e popolazione dell'area stessa;
d) Differenze (o variazioni) assolute di impermeabilizzazioni nei periodi 2006/2009, 2009/2012, 2012/2015, 2015/2018, e 2006/2018. Sono ottenute dalla sottrazione dei valori assoluti dei due anni;
e) Differenze (o variazioni) di impermeabilizzazioni pro capite nei periodi 2006/2009, 2009/2012, 2012/2015, 2015/2018, e 2006/2018. Sono ottenute dalla sottrazione dei valori pro capite dei due anni di interesse;
f) Variazione relativa percentuale di impermeabilizzazione, nei periodi 2006/2009, 2009/2012, 2012/2015, 2015/2018, e 2006/2018. Sono ottenute mediante il rapporto tra la variazione assoluta calcolata tra i due anni (punto d), e la superficie impermeabilizzata dell'anno all'inizio del periodo.

2.4. Database di popolazione

La densità di popolazione in una determinata area geografica può acuire i danni provocati da un disastro ambientale che si abbatta su tale territorio, specialmente in termini di numero di decessi e di persone colpite. Per questo motivo, risulta necessaria una fonte affidabile che fornisca la popolazione delle varie province europee nei vari anni di interesse.

La fonte utilizzata è il dataset *"Population on 1 January by age group, sex and NUTS 3 region"* (Eurostat, 2024), che fornisce la popolazione residente, suddivisa per fascia d'età e sesso, nelle varie province europee, dal 1960 al 2023.

Benché l'informazione disponibile sia ben più dettagliata per classe di età e genere, ai fini del presente studio è stata utilizzata unicamente la popolazione totale per ciascuna area geografica nel periodo 2006-2018, in quanto non è stato ritenuto che la struttura per sesso o età della popolazione costituissero fattori direttamente rilevante nell'ambito dell'analisi proposta.

I dati presenti sono stimati modificando il dato dell'anno precedente con le nascite, i decessi, e i trasferimenti, comunicati ad Eurostat da parte dei singoli Stati. Ciclicamente, ogni dieci anni, gli Stati hanno l'obbligo di eseguire un censimento della popolazione, che costituisce il dato dell'anno. Nell'arco temporale di interesse è avvenuto un censimento nel 2011.

I dati sono aggregati a livello provinciale (NUTS 3) e fenomeni come l'ingresso di nuovi Stati all'interno dell'Unione Europea, o modifiche nei confini amministrativi (fusioni o suddivisioni di province) sono la causa della presenza di dati mancanti o discontinuità nei primi anni della serie temporale.

Inoltre, il dataset esplicita gli anni in cui ci sono stati problemi nella comunicazione dei dati da parte dei governi, o in cui sono variati i meccanismi di calcolo dei flussi di popolazione per determinate province.

2.5. Confini delle province europee

L'unità utilizzata per le analisi sono le aree NUTS (*Nomenclature of Territorial Units for Statistics*). Esse sono suddivisioni del territorio operate dall'Unione Europea espressamente per fini statistici. La classificazione è stata creata a seguito del regolamento (CE) N. 1059/2003 del Parlamento europeo e del Consiglio del 26 maggio 2003.

Sebbene non siano completamente scollegate dalle divisioni amministrative nazionali, le unità NUTS possono risultare differenti da queste ultime. Un elemento fondamentale per la loro definizione è infatti la popolazione che vi risiede.

I livelli della classificazione NUTS sono quattro:

a) NUTS0, codice di due lettere coincidente sempre con uno Stato;

b) NUTS1, codice di tre caratteri, che indica aree vaste di uno Stato, la cui popolazione si aggira tra 3 e 7 milioni di abitanti;

c) NUTS2, codice di quattro caratteri, che indica aree talvolta equivalenti alle regioni, la cui popolazione si aggira tra 800 mila e 3 milioni di abitanti;

d) NUTS3 codice di cinque caratteri, che indica aree la cui popolazione si aggira tra 150 mila e 800 mila abitanti. Queste aree possono essere equivalenti a province o, in alcuni casi, a città molto densamente popolate.

I limiti demografici non devono essere intesi come vincoli rigidi, ma come intervalli di riferimento utilizzati per garantire una comparabilità statistica tra le unità territoriali.

La fonte dalla quale sono stati ottenuti i confini è GISCO (*Geographic Information System of the Commission*), servizio di Eurostat deputato alla produzione e alla diffusione di dati cartografici relativi all'Europa. Nello specifico sono stati utilizzati i confini amministrativi aggiornati al 2021 (Eurostat (GISCO), 2021). Non è stata presa in esame la versione ulteriormente aggiornata (NUTS 2024) in quanto le differenze tra essa e la versione 2021, in termini di Stati appartenenti all'Unione Europea, sono minime (la principale è l'eliminazione del Regno Unito), e le trasformazioni tra NUTS 2021 e NUTS 2024 non risultano pienamente documentate in modo chiaro ai fini della presente analisi.

I dati cartografici sono referenziati in EPSG:4326 (sistema geografico WGS84, standard).

Di ogni provincia sono noti i confini geolocalizzati, il nome, e tre variabili di classificazione del territorio (European Commission, 2019)

a) Se l'area è montuosa, che si divide tra:
    a) più del 50% dell'area della regione è montuosa;
    b) più del 50% della popolazione vive in aree montuose della regione;
    c) più del 50% dell'area della regione è montuosa e più del 50% della popolazione vive in aree montuose della regione;
    d) Nessuna delle precedenti
b) Se l'area è costiera, che si divide tra
    a) La regione comprende una costa;
    b) meno del 50% della popolazione vive entro 50 km da una costa;
    c) Nessuna delle due precedenti.
c) Se l'area è urbanizzata, che si divide in:
    a) Più del 80% della popolazione vive in città, oppure è presente una città con almeno 500 mila abitanti;
    b) Tra il 50% e l'80% della popolazione vive in città, oppure è presente una città con almeno 200 mila abitanti:
    c) Meno del 50% della popolazione vive in città.

Anche questi dati, che dovrebbero essere riferiti al solo anno 2021, sono stati considerati come rappresentativi dell'intero periodo. Dalla stessa fonte sono stati scaricati anche i

confini delle regioni, e degli Stati, per poter eseguire un'analisi anche a quei livelli. Regioni e Stati, tuttavia, non presentano nessuna caratterizzazione in merito a urbanizzazione, montuosità, o vicinanza alla costa.

2.6. Dataset climatici di precipitazioni e temperature

La superficie impermeabilizzata e le caratteristiche proprie della provincia sono indicatori che si suppone, e si vuole testare, essere legati ai danni causati da disastri naturali. Per rendere questo test più significativo si è scelto di includere nell'analisi una componente di informazioni logicamente correlate ai disastri naturali stessi, essendo parte della loro causa. Esse sono le informazioni meteorologiche. Per mantenere la struttura stabilita, ossia dati a scala annuale e provinciale, sono stati presi in considerazione i seguenti tre indicatori medi, alla medesima scala:

- temperatura media;
- precipitazione giornaliera massima nell'anno;
- numero medio di giorni di pioggia.

Sono stati scelti questi indicatori in quanto capaci di fornire una sintesi della situazione meteorologica dell'anno di interesse e logicamente correlati ai principali tipi di eventi avversi censiti.

Le precipitazioni, presenti sia come precipitazione totale annua sia come giorni di pioggia, possono essere correlate ai fenomeni di tipo idrologico, mentre le temperature medie possono essere correlate ai fenomeni meteorologici. Infatti, i disastri del primo tipo sono sostanzialmente costituiti da piene fluviali e da smottamenti, entrambi fenomeni legati alla presenza di precipitazioni intense. Al contrario, i fenomeni del secondo tipo comprendono principalmente eventi siccitosi, temperature estreme (che possono essere causate sia da caldi che da freddi estremi indistintamente) e incendi boschivi, spesso correlati a periodi secchi e caldi.

Una limitazione contenuta nell'utilizzo di questi indicatori è che essi, soprattutto a questa scala spaziale e temporale, non sono in grado di descrivere i fenomeni disastrosi. Infatti, gli eventi naturali avversi sono causati da una concomitanza di fattori, con caratteristiche molto più istantanee e molto più locali rispetto alla scala utilizzata. Nonostante ciò, è stato comunque deciso di includerli nell'analisi, considerate le possibili correlazioni con il verificarsi di eventi dannosi e con la loro gravità.

La fonte per questi dati è, in tutti i casi, Copernicus (Haylock, et al., 2008) (Royal Netherlands Meteorological Institute (KNMI), 2024)

Sono stati presi i dati di precipitazioni giornaliere e temperatura media giornaliera, disponibili in formato raster georeferenziato in EPSG:4326 (WGS84), coprente l'intera Europa (longitudine da -25 a 40 ° est, latitudine da 35 a 71 ° nord).

3 Metodi

Il presente capitolo illustra il quadro metodologico adottato per integrare e armonizzare le basi dati descritte nel Capitolo 2, trasformandole in un dataset analitico coerente sotto il profilo spaziale e temporale e adeguato alla successiva fase di analisi statistica e modellazione.

Per una descrizione dettagliata dei metodi statistici presentati in questo capitolo si rimanda a (Geoffrey Norman, 2000)

Tutte le analisi statistiche della tesi sono state implementate usando pacchetti del software statistico R (R Core Team, 2025).

In primo luogo, viene descritta la costruzione del dataset finale e la definizione delle unità di osservazione, con particolare attenzione alla procedura di geolocalizzazione degli eventi e alla costruzione delle tabelle a diversa scala territoriale (Sezione 3.1).

Successivamente sono illustrati i criteri di elaborazione degli indicatori climatici e territoriali e le trasformazioni adottate per garantirne coerenza, confrontabilità e idoneità all'analisi quantitativa (Sezione 3.2).

Una fase preliminare di stima mediante modello lineare viene quindi introdotta sia come strumento operativo per l'imputazione dei dati mancanti sia come supporto esplorativo alla modellazione successiva (Sezione 3.3).

Il capitolo prosegue con l'analisi esplorativa dei dati, condotta mediante stima kernel delle distribuzioni, analisi della varianza e tecniche di clustering, al fine di evidenziare differenze tra territori e possibili strutture latenti nei dati (Sezione 3.4).

Infine, viene presentata la regressione logistica quale principale strumento inferenziale del lavoro. Dopo la specificazione formale del modello, sono discussi l'interpretazione dei coefficienti e le principali assunzioni, nonché i criteri di valutazione delle prestazioni del modello e la scelta della soglia di classificazione tramite curva ROC e validazione Leave-One-Out.

3.1. Costruzione del dataset finale e unità di osservazione

L'analisi statistica condotta nel presente elaborato richiede la costruzione di una base dati coerente dal punto di vista spaziale e temporale, ottenuta mediante integrazione e rielaborazione delle fonti descritte nel Capitolo 2.

Il principale dataset relativo ai disastri naturali, EM-DAT, localizza ciascun evento a livello statale e annuale. Sebbene tale struttura sia adeguata a confronti tra Stati, l'analisi a livello

statale può risultare riduttiva, in quanto all'interno di uno stesso Stato possono coesistere condizioni climatiche, morfologiche e insediative significativamente differenti.

I disastri naturali presentano spesso un impatto fortemente localizzato, sia in termini di cause sia di effetti. Per questo motivo, si è ritenuto opportuno procedere a una geolocalizzazione più fine degli eventi, con l'obiettivo di attribuire, ove possibile, ciascun disastro a una specifica provincia.

3.1.1. Geolocalizzazione degli eventi

La prima fonte utilizzata per la localizzazione sub-statale è il database GDIS (Global Disasters Identifier System), che rappresenta una rielaborazione geospaziale di EM-DAT e associa a molti eventi coordinate geografiche e unità amministrative di livello inferiore.

Tuttavia, la copertura di GDIS non è completa: alcuni eventi presenti in EM-DAT non risultano georeferenziati; in altri casi, un singolo evento può essere associato a più punti spaziali, qualora abbia interessato diverse aree amministrative. Inoltre, la precisione della localizzazione dipende strettamente dalla qualità delle informazioni originarie contenute in EM-DAT.

Per aumentare la quota di eventi attribuibili a scala sub-statale, è stata quindi utilizzata anche l'informazione testuale contenuta nel campo "Location" di EM-DAT. Mediante procedure di interpretazione automatica del testo, integrate con l'informazione relativa allo Stato di occorrenza, è stato possibile associare molte descrizioni testuali al corrispondente codice amministrativo. Lo schema di questo processo di geolocalizzazione è in Figura 3.1.

Il risultato di tale procedura è un insieme di eventi classificati, quando l'informazione disponibile lo consente, a livello provinciale.

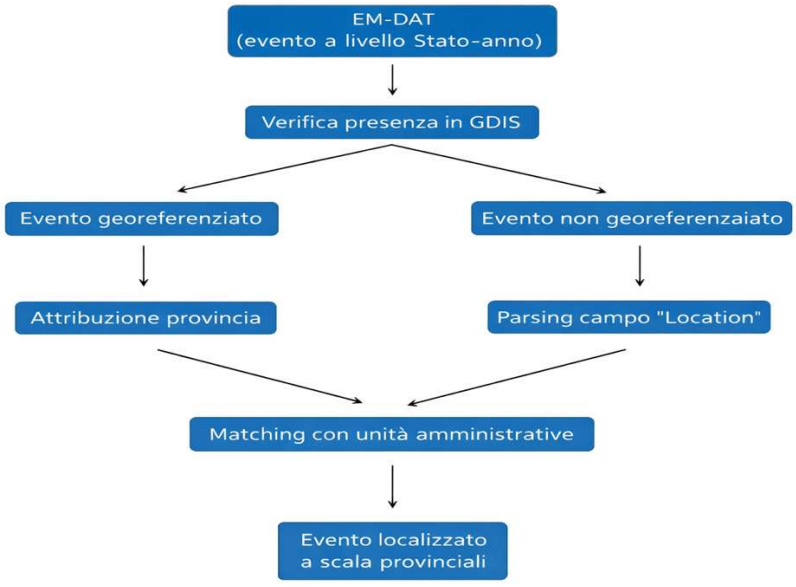

Figura 3.1: Schema del processo di geolocalizzazione degli eventi a scala provinciale mediante integrazione tra EM-DAT, GDIS e informazioni testuali contenute nel campo "Location"

Una volta completata la fase di geolocalizzazione, gli eventi sono stati aggregati secondo due livelli territoriali:

- scala nazionale (Stato–anno);
- scala provinciale (Provincia–anno).

L'unità statistica adottata nell'analisi non coincide con il singolo evento, bensì con la coppia territorio–anno.

Per ciascuna unità territoriale i e anno di osservazione t sono stati calcolati:

- numero totale di disastri (N_{it});
- totale dei decessi (D_{it});
- indicatori climatici (${X_{it}}^{(clima)}$);
- indicatori territoriali, popolazione, area, percentuale di suolo impermeabilizzato, ${X_{it}}^{(territorio)}$.

Quindi l'unità osservata può essere definita sinteticamente nel seguente modo:

$$O_{it} = (N_{it}, D_{it}, {X_{it}}^{(clima)}, {X_{it}}^{(territorio)})$$

Ai fini della modellazione inferenziale, è stata definita una variabile binaria Y_{it} indicante la presenza di almeno un evento con danni nell'unità territoriale considerata:

$$Y_{it} = 1 \; se \; N_{it} > 0 \; mentre \; Y_{it} = 0 \; se \; N_{it} = 0$$

Questa variabile costituisce la risposta nei modelli di regressione logistica descritti nella sezione 3.4.
Con questa formalizzazione, teniamo sempre esplicitamente conto del fatto che i dataset risultanti presentano una struttura panel spazio-temporale, cioè ciascuna unità territoriale è osservata ripetutamente lungo l'intervallo temporale 2006–2018, con scala annuale t = 2006, ..., 2018.

3.1.2. Costruzione delle tabelle finali

Il risultato operativo della fase di integrazione e geolocalizzazione è costituito dalle tabelle che rappresentano la base dati per le successive analisi statistiche e per la modellazione inferenziale. Sono state prodotte tre tabelle, tra loro strutturalmente simili ma riferite a scale territoriali differenti: una tabella a scala statale, ottenuta mediante trasformazione dei dati EM-DAT grezzi, senza procedere a localizzazione sub-statale; una tabella a scala regionale; una tabella a scala provinciale, costruite a partire dalla suddivisione degli eventi EM-DAT in base alla regione o alla provincia interessata, secondo la procedura descritta nella Sezione 3.1.1.
Per ciascuna unità territorio–anno O_{it}, le tabelle contengono le seguenti informazioni:
Stato di riferimento (non utilizzato come variabile esplicativa nella tabella a scala statale, al fine di evitare basi di training costituite da una sola osservazione per Stato);
anno di riferimento;

percentuale di suolo impermeabilizzato nell'area considerata, opportunamente linearizzata area e popolazione dell'unità territoriale;
indicatori climatici, rappresentati da temperatura media, numero di giorni di pioggia e precipitazioni totali;
numero totale di disastri verificatisi e totale dei deceduti nell'area di interesse. Nel caso della tabella a scala statale tale valore è ottenuto mediante conteggio diretto degli eventi EM-DAT, mentre nella tabella a scala provinciale deriva dalla procedura di attribuzione territoriale precedentemente descritta;
numero di disastri, e totale dei deceduti relativi ai soli eventi classificati come idrologici e meteorologici, in quanto rappresentano la tipologia prevalente nel dataset considerato.

Le due tabelle così costruite costituiscono la base dati analitica utilizzata nelle successive fasi di analisi esplorativa e modellazione. La loro struttura omogenea consente il confronto tra differenti livelli territoriali e permette di verificare la robustezza dei risultati rispetto alla scala di aggregazione adottata.

3.2. Elaborazione degli indicatori climatici e territoriali

Una volta costruite le tabelle alle diverse scale territoriali, si è proceduto all'elaborazione delle variabili esplicative e a una fase preliminare di analisi dei dati, finalizzata a verificarne coerenza, distribuzione e principali relazioni.

Questa fase ha lo scopo di preparare la successiva modellazione inferenziale, assicurando che gli indicatori climatici e territoriali siano correttamente definiti, confrontabili tra territori e coerenti con la struttura spazio-temporale del dataset.

3.2.1. Calcolo degli indicatori climatici

Allo scopo di inserire nelle tabelle finali indicatori climatici rappresentativi dell'area di interesse, sono stati utilizzati i dati in formato raster relativi a temperatura e precipitazioni, opportunamente proiettati nel sistema di riferimento europeo ETRS89 / LAEA Europe, coerente con la cartografia amministrativa adottata nello studio. È stato preliminarmente verificato che i layer raster risultassero spazialmente sovrapponibili ai poligoni amministrativi di Stati e Province.

I dati climatici sono stati raggruppati per anno e, per ciascun anno, sono stati calcolati tre indicatori:

a) precipitazioni massime annuali;
b) temperatura media annuale;
c) numero di giorni con precipitazioni maggiori di zero, considerati giorni di pioggia.

Una volta generati i raster annuali relativi a tali indicatori, il layer vettoriale dei poligoni amministrativi (Province e Stati) è stato rasterizzato sulla medesima griglia spaziale, al fine di velocizzare il successivo calcolo aggregato. L'aggregazione è stata quindi effettuata mediante sovrapposizione tra raster climatici e raster amministrativi, calcolando, per ciascuna unità territoriale i e anno t:

- la media dei valori nel caso della temperatura;
- la somma nel caso delle precipitazioni totali;
- il conteggio nel caso dei giorni di pioggia.

Formalmente, per un generico indicatore climatico C, il valore territoriale può essere espresso come:

$$C_{it} = \frac{\sum_{j \in A_{it}} C_{jt}}{n_{it}},$$

dove A_{it} rappresenta l'insieme delle celle raster appartenenti al territorio i nell'anno t, ed n_{it} il numero di celle considerate (nel caso della media).

Nel corso della procedura sono stati generati alcuni valori mancanti, riconducibili unicamente a singole celle costiere trattate in modo differente nelle due rasterizzazioni. Tali celle, ricoprendo una superficie marginale rispetto all'area complessiva delle unità territoriali, sono state considerate non significative e pertanto escluse dal calcolo.

3.2.2. Indicatori territoriali e trasformazioni

Gli indicatori territoriali includono:

- area dell'unità amministrativa;
- popolazione residente;
- percentuale di suolo impermeabilizzato.

Per l'indicatore di impermeabilizzazione è stata considerata la variazione nel periodo 2006–2018, coerentemente con l'intervallo temporale del dataset. La variabile è stata linearizzata al fine di facilitarne l'inclusione nei modelli statistici e ridurre possibili effetti di non linearità.

È stata inoltre analizzata la correlazione tra le variabili territoriali, in particolare tra superficie impermeabilizzata, area e popolazione, al fine di individuare eventuali fenomeni di collinearità.

3.3. Stima e valutazione di un modello lineare

L'utilizzo di un modello statistico è uno strumento utile sia per l'esplorazione preliminare delle relazioni tra variabili, sia per l'imputazione di dati mancanti. In presenza di due variabili, fra le quali vi sia, plausibilmente, una relazione funzionale, è possibile stimare i parametri di tale relazione a partire da osservazioni congiunte e, successivamente, sfruttare la relazione e una delle due variabili per stimare valori mancanti dell'altra.

La forma funzionale più semplice di tale relazione è quella lineare. In forma più generale, nel caso multivariato, il modello lineare può essere espresso come:

$$Y_{it} = \beta_0 + \sum_{k=1}^{p} \beta_k X_{it}^{(k)} + \epsilon_{it},$$

dove:

Y_{it} rappresenta la variabile di interesse per l'unità territoriale (i) nell'anno (t);

$X_{it}^{(k)}$ sono le variabili esplicative indipendenti;

ϵ_{it} è il termine di errore.

Allo scopo di stimare una simile relazione sono tecnicamente sufficienti due osservazioni delle variabili X e delle corrispondenti Y. Sebbene due sia il numero minimo, è sempre preferibile disporre di un numero maggiore di osservazioni, allo scopo di valutare l'affidabilità del modello, rendendolo meno sensibile a singole osservazioni anomale.

Nel presente elaborato, il modello di regressione lineare è stato utilizzato per imputare dati mancanti della popolazione, e della percentuale di suolo impermeabilizzato, relativi ad alcuni anni e ad alcune province. Come variabile indipendente *X* è stata scelta il tempo, sempre disponibile per l'intero periodo considerato.

Nonostante non esistano ragioni fisiche per supporre un andamento lineare di tali variabili nel tempo, si è supposto che il ridotto arco temporale e l'assenza di eventi cataclismatici al suo interno, potessero limitare l'errore introdotto dall'ipotesi di linearità. Inoltre, in alcuni casi, in particolare per la popolazione, il modello è stato utilizzato anche in estrapolazione temporale. Sebbene tale pratica sia generalmente sconsigliata, essa è stata ritenuta accettabile alla luce del ridotto arco temporale e degli indicatori di affidabilità utilizzati.

Nel contesto del presente studio, il modello lineare è stato utilizzato principalmente come strumento operativo e diagnostico, e non come modello inferenziale definitivo.

3.3.1. Valutazione del modello

Coefficiente di determinazione R^2

Il coefficiente di determinazione R^2 è uno dei coefficienti più utilizzati per valutare l'affidabilità di modelli basati sui dati. La sua formula è:

$$R^2 = 1 - \frac{\sum(Y_{it} - \hat{Y}_{it}^{stimato})^2}{\sum(Y_{it} - media(Y_{it}))^2}$$

L'indice misura la quota di variabilità delle Y originali appresa dal modello. Il valore dell'indice è compreso tra 0 e 1. Un valore dell'indice prossimo a 0 indica un modello indistinguibile dal modello costante (che basa la spiegazione di Y sulla sola sua media empirica), mentre un valore prossimo a 1 indica che la quasi totalità della variabilità del fenomeno osservato è stata catturata dal modello utilizzato.

Radice dell'errore quadratico medio relativo ($RMSE_{rel}$)

Questo indice è concettualmente differente dal precedente. Esso infatti misura, come il nome suggerisce, l'errore medio commesso dal modello su osservazioni note. L'errore quadratico medio (RMSE) è definito come:

$$RMSE = \sqrt{\frac{1}{N}\sum(Y_{it} - \hat{Y}_{it})^2}$$

Dove N è la numerosità totale del campione, Y_{it} è il termine effettivamente misurato nello Stato (i) al tempo (t), mentre $\hat{Y}_{it}$ è il termine stimato, dal modello nel medesimo Stato e anno.

Per rendere l'indicatore confrontabile tra variabili di diversa scala, è stato utilizzato l'RMSE relativo:

$$RMSE_{rel} = \frac{RMSE}{\bar{media}(Y_{it})}$$

Questo indice esprime l'errore medio in proporzione al valore medio della variabile osservata. Ad esempio, un valore pari a 0.05 indica un errore medio pari al 5% del valore medio della variabile Y.

3.4. Metodi di analisi esplorativa

Un'analisi esplorativa dei dati ha l'obiettivo di descrivere la distribuzione delle principali variabili e individuare eventuali differenze sistematiche tra territori. A tale scopo sono stati utilizzati strumenti di stima non parametrica, test di confronto tra gruppi e tecniche di classificazione.

3.4.1. Kernel Density Estimation (KDE)

La Kernel Density Estimation (KDE) è una tecnica utilizzata per stimare la densità di probabilità di una variabile continua, a partire da un campione di osservazioni. Tale metodo consente di ottenere una rappresentazione liscia della distribuzione empirica dei dati, senza assumere una forma parametrica predefinita a priori.

Nel grafico risultante, l'asse delle ascisse rappresenta la variabile continua analizzata, mentre l'asse delle ordinate rappresenta la densità di probabilità stimata. I valori assoluti della densità non hanno un'interpretazione diretta in termini di probabilità puntuale, l'informazione rilevante è contenuta unicamente nella forma complessiva della curva, che consente di individuare le aree di maggiore concentrazione delle osservazioni e di evidenziare eventuali strutture interne, come asimmetrie o multimodalità.

Il metodo si basa sulla somma di contributi individuali associati a ciascuna osservazione del campione. In corrispondenza di ogni valore osservato viene definita una funzione kernel, identica per tutte le osservazioni e centrata sul valore stesso. Le funzioni kernel sono normalizzate in modo che l'integrale di ciascun contributo sull'intero dominio sia pari a $1/N$, dove N è la numerosità del campione.

Formalmente, data una variabile continua osservata su un campione $\{x_1, x_2, \dots, x_N\}$, la stima della densità in un punto x è definita come:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

dove $K(\cdot)$ è la funzione kernel e h è il parametro di banda (bandwidth), che regola il livello di *smoothing* della curva stimata. Un valore di h più elevato produce una curva più liscia ma potenzialmente meno sensibile alle variazioni locali, mentre un valore più piccolo rende la stima più aderente ai dati ma anche più irregolare.

La densità stimata finale è ottenuta come somma di tutti i contributi kernel. Di conseguenza, nelle regioni del dominio in cui le osservazioni sono più concentrate il valore della densità risulta più elevato, mentre tende a zero nelle regioni scarsamente popolate. Essendo la somma di funzioni normalizzate, l'integrale della densità stimata sull'intero dominio è pari a uno, in accordo con la definizione di densità di probabilità.

In questa tesi, la KDE è stata utilizzata per confrontare visivamente le distribuzioni di alcune variabili tra Stati, evidenziando eventuali differenze nella forma, nella dispersione e nella presenza di code asimmetriche.

3.4.2. ANOVA

L'Analisi della Varianza (ANOVA) è una tecnica statistica utilizzata per verificare la presenza di differenze significative tra le medie di una variabile quantitativa tra due o più gruppi definiti da variabili categoriche. In questo lavoro, l'ANOVA è stata impiegata per valutare in che misura la variabilità di alcune variabili climatiche e socio-territoriali possa essere spiegata dall'appartenenza a differenti Stati.

L'idea alla base dell'ANOVA consiste nel confrontare la variabilità tra i gruppi con la variabilità interna ai gruppi stessi. Se le medie dei gruppi differiscono in modo rilevante rispetto alla variabilità osservata all'interno di ciascun gruppo, si può concludere che esiste evidenza statistica a favore dell'ipotesi di differenze tra gruppi.

È importante sottolineare inoltre che l'ANOVA presuppone, in forma classica, l'indipendenza delle osservazioni, l'omoschedasticità e la normalità dei residui. Nel contesto del presente studio, tali assunzioni sono state considerate in modo critico, utilizzando l'ANOVA principalmente come strumento comparativo ed esplorativo piuttosto che come unico criterio decisionale.

Formalmente, nel caso di un'ANOVA a una via cioè di un solo fattore che può fare la differenza in media fra G gruppi, il modello statistico è:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

dove:

Y_{ij} è il valore osservato della variabile di interesse,

μ è la media complessiva dei valori osservati,

α_i è l'effetto associato al gruppo i-esimo,

ε_{ij} è il termine di errore, assunto con media nulla e varianza costante,

L'ipotesi nulla del test ANOVA è che tutte le medie di gruppo siano uguali:

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_G = 0$$

contro l'ipotesi alternativa H_1 che afferma che almeno una media differisce dalle altre.

Il principio alla base dell'ANOVA consiste nella scomposizione della variabilità totale della variabile risposta in due componenti:

- variabilità tra i gruppi (between-group variability),
- variabilità all'interno dei gruppi (within-group variability).

Il confronto tra queste due componenti avviene tramite la statistica F, definita come il rapporto tra la varianza between-group e la varianza within-group:

$$F = \frac{SS_{between}/(G-1)}{SS_{within}/(N-G)}$$

dove $SS_{between}$ è la somma dei quadrati tra i gruppi e SS_{with} la somma dei quadrati entro gruppi. Valori elevati della statistica F indicano che le differenze tra i gruppi sono grandi rispetto alla variabilità interna ai gruppi, e dunque ci sono dei gruppi statisticamente differenti dagli altri. Come in tutti i test statistici, alla statistica test è associata la statistica p-value che rappresenta il pià piccolo livello di significatività per cui con i dati a disposizione si rifiuta H_0; operativamente rappresenta la probabilità di osservare un valore della statistica test F almeno altrettanto estremo di quello calcolato, sotto l'ipotesi nulla di uguaglianza delle medie tra i gruppi. Valori piccoli del p-value forniscono evidenza contro l'ipotesi nulla, suggerendo la presenza di differenze tra gruppi; valori elevati indicano che i dati osservati sono compatibili con l'ipotesi di uguaglianza delle medie.
Oltre al test di significatività, in questo lavoro l'ANOVA viene utilizzata anche come strumento descrittivo per quantificare la quota di variabilità spiegata dalla variabile di classificazione. A tal fine viene utilizzato l'indice η^2(eta quadrato), definito come:

$$\eta^2 = \frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{SS}_{\mathrm{total}}}$$

dove $\mathrm{SS}_{\mathrm{total}}$ è la somma dei quadrati totale. Tale indice assume valori compresi tra 0 e 1 e fornisce una misura diretta della proporzione di varianza della variabile risposta spiegata dalla classificazione adottata.

È importante sottolineare che l'ANOVA presuppone:

- indipendenza delle osservazioni,
- normalità dei residui,
- omoschedasticità (varianza costante all'interno dei gruppi).

Nel contesto del presente studio, tali assunzioni sono state considerate in modo critico, utilizzando l'ANOVA principalmente come strumento comparativo ed esplorativo piuttosto che come unico criterio decisionale. Infatti tali ipotesi possono risultare solo parzialmente soddisfatte, in particolare per quanto riguarda l'indipendenza spaziale delle osservazioni.

Per questo motivo, i risultati dell'ANOVA vengono interpretati con cautela e considerati come indicazioni descrittive piuttosto che come evidenze causali.

L'ANOVA rappresenta quindi uno strumento utile per sintetizzare e quantificare le differenze tra unità geografiche, fornendo una prima valutazione della rilevanza territoriale delle variabili considerate, che verrà successivamente approfondita mediante modelli statistici più articolati.

3.4.3. Clustering mediante algoritmo *k*-means

Il *k*-means è un algoritmo di raggruppamento non supervisionato utilizzato per suddividere un insieme di osservazioni in un numero prefissato di gruppi (cluster), sulla base della loro somiglianza rispetto a un insieme di variabili quantitative.

Nel contesto di questo lavoro, il *k*-means è impiegato come strumento esplorativo per identificare tipologie territoriali omogenee sulla base di variabili ambientali, climatiche e socio-territoriali, senza imporre a priori una suddivisione amministrativa.

Dato un insieme di osservazioni $\{x_1, x_2, \dots, x_N\}$, ciascuna rappresentata come vettore in uno spazio multidimensionale, l'algoritmo k-means procede come segue:

1. viene fissato a priori il numero di cluster k;
2. vengono inizializzati k centroidi nello spazio delle variabili;
3. ogni osservazione viene assegnata al cluster al cui centroide risulta più vicino, secondo una misura di distanza (generalmente la distanza euclidea);
4. i centroidi vengono ricalcolati come la media delle osservazioni appartenenti a ciascun cluster;
5. i passaggi 3 e 4 vengono ripetuti iterativamente fino alla convergenza, ovvero fino a quando le assegnazioni non cambiano più, o la variazione risulta trascurabile.

L'algoritmo al punto 2. minimizza la somma delle distanze quadratiche intra-cluster, formalmente:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2$$

dove C_j è il *j*-esimo cluster, μ_j il relativo centroide e $\|\cdot\|$ rappresenta la distanza euclidea.

Un aspetto cruciale nell'applicazione del *k*-means è la scelta del numero di cluster k, che non viene determinato automaticamente dall'algoritmo. In questo studio, la scelta di k è guidata da criteri esplorativi, quali:

- l'analisi della varianza intra-cluster al variare di k (metodo del "gomito"),
- la leggibilità e interpretabilità dei cluster ottenuti,
- la coerenza dei risultati con la struttura territoriale e climatica nota dell'area di studio.

Poiché il *k*-means si basa su distanze nello spazio delle variabili, i risultati dell'algoritmo risultano fortemente influenzati dalle scale di misura delle variabili considerate. Per questo motivo, prima dell'applicazione del clustering, le variabili quantitative vengono

standardizzate, in modo da garantire che ciascuna contribuisca in maniera comparabile alla definizione delle distanze.

I cluster ottenuti mediante *k*-means non hanno senso proprio, ma sono una partizione dei dati dipendente dalle variabili considerate e dal valore di k. In questo lavoro, essi vengono interpretati come tipologie territoriali caratterizzate da combinazioni simili di condizioni climatiche, ambientali e insediative.

L'utilizzo del *k*-means presenta alcune limitazioni rilevanti:

- la necessità di fissare a priori il numero di cluster;
- la sensibilità all'inizializzazione dei centroidi;
- la dipendenza dalla scelta delle variabili incluse nell'analisi.

Nel contesto di questa tesi, il *k*-means viene quindi utilizzato come metodo esplorativo di classificazione territoriale, finalizzato a sintetizzare la complessità multidimensionale delle variabili analizzate e a individuare gruppi omogenei di aree, che possono successivamente essere messi in relazione con la distribuzione e la frequenza dei disastri naturali.

3.5. Modelli di regressione logistica

I modelli di regressione logistica, o modelli logit, sono usati per analizzare la relazione tra una variabile dipendente binaria e un insieme di variabili esplicative, che possono essere sia quantitative sia categoriali. In questo lavoro, tali modelli sono utilizzati per studiare i fattori climatici e socio-territoriali associati alla probabilità di accadimento di disastri naturali.

A differenza del modello lineare, la regressione logistica è appropriata quando la variabile risposta è di natura binaria. Nel caso in esame, come già riportato sopra, la variabile risposta Y_{it} è definita come indicatore dell'avvenimento di almeno un disastro naturale nel territorio ie nell'anno t.

3.5.1. Specificazione del modello

Dato un insieme di covariate $X_1, X_2, \dots, X_p$, il modello logit esprime la probabilità che l'evento di interesse si verifichi nel territorio i nell'anno t come:

$$P(Y_{it} = 1 | X_{1;\,it}, \dots, X_{p;\,it}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^{p} \beta_k X_{k;\,it})}}$$

Equivalentemente, tramite la funzione logit, il modello può essere scritto in termini di log-odds:

$$log\left(\frac{P(Y_{it} = 1)}{1 - P(Y_{it} = 1)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \, X_{k;\,it}$$

dove:

- Y_{it} è la variabile binaria indicante l'occorrenza di un disastro,
- β_0 è l'intercetta, legata alla probabilità, stimata, base, senza nessun predittore,

- β_k è il coefficiente di regressione associato alla variabile esplicativa X_k e, per $k = 1, \dots, p$.

Nel contesto della presente tesi, il modello logit è utilizzato come strumento inferenziale per analizzare l'associazione tra caratteristiche territoriali, climatiche e socio-ambientali e la probabilità di occorrenza di disastri naturali. L'obiettivo non è fornire una previsione puntuale degli eventi, bensì valutare la direzione e l'intensità delle relazioni statistiche tra le variabili considerate e l'evento di interesse.

3.5.2. Interpretazione e assunzioni del modello

Nel modello logit, i coefficienti di regressione β_k non rappresentano variazioni dirette di probabilità, ma variazioni dei log-odds dell'evento. Il segno di ciascun coefficiente indica se l'aumento della variabile esplicativa è associato a un incremento o a una riduzione della probabilità di occorrenza del disastro.

In particolare, l'esponenziale del coefficiente $\exp(\beta_k)$ rappresenta l'odds ratio associato a un incremento unitario della variabile X_k, mantenendo costanti le altre variabili. Un valore di $\exp(\beta_k) > 1$ indica un aumento delle odds dell'evento all'aumentare della variabile esplicativa X_k; un valore inferiore a 1 indica una diminuzione.

Il modello consente di analizzare simultaneamente l'effetto di più covariate, controllando per la presenza delle altre, e confrontando l'importanza relativa dei diversi fattori X inclusi.

L'applicazione della regressione logistica si basa sulle seguenti assunzioni:

- corretta specificazione funzionale del modello;
- indipendenza delle osservazioni;
- assenza di multicollinearità eccessiva tra le variabili esplicative;

Nel caso in esame, l'ipotesi di indipendenza può risultare parzialmente violata a causa della presenza di dipendenze spaziali e temporali tra le osservazioni. Inoltre, i risultati della stima vengono interpretati come associazioni statistiche e non come relazioni causali.

È importante sottolineare che i modelli logit:

- non modellano direttamente la frequenza o l'intensità dei disastri, ma solo la loro occorrenza,
- sono sensibili alla definizione della variabile binaria di risposta.

Infine, la presenza di classi sbilanciate, ossia un numero relativamente ridotto di osservazioni con evento rispetto a quelle senza evento, può influenzare l'interpretazione delle probabilità stimate.

3.5.3. Valutazione del modello e scelta della soglia

Il risultato del modello logistico è una probabilità stimata dell'evento. Vi sono due modi principali per interpretare il risultato di un modello logistico.

Il primo consiste nell'analizzare direttamente le probabilità stimate, valutando la capacità del modello di assegnare probabilità maggiori alle osservazioni in cui l'evento si è

effettivamente verificato. Questo approccio consente di valutare la capacità discriminatoria del modello stesso, e dunque la sua affidabilità.

Il secondo modo in cui un modello logistico può essere interpretato consiste nel trasformare la probabilità in una predizione binarie, inserendo una soglia di classificazione. Se la probabilità stimata è superiore alla soglia, l'evento viene predetto come verificato; in caso contrario, come non verificato. La scelta della soglia è quindi fondamentale per l'utilizzo pratico del modello.

Una scelta normale è quella del valore 0,5, assegnando dunque 0 hai i risultati che presentano probabilità di accadimento minore di 0.5, e 1 a quelli con probabilità maggiore di 0.5. Tuttavia, la soglia dipende strettamente dalle caratteristiche delle variabili in gioco. Nell'analisi di eventi rari è probabile che servano soglie inferiori a 0,5, proprio per la caratteristica rarità dell'evento.

La procedura corretta per la stima di una soglia ottimale parte dall'applicazione del modello su una serie di dati di test, sui quali dunque si conoscano sia i valori di tutte le covariate che il risultato effettivamente prodotto, ma che non siano stati utilizzati precedentemente per la stima dei parametri del modello.

Nel presente elaborato, per stimare la soglia ottimale e valutare le prestazioni del modello è stato utilizzato il metodo Leave-One-Out (LOO), particolarmente adatto in presenza di un numero limitato di osservazioni.

Tale metodo consiste nel ripetere n volte (con n che può arrivare fino al numero totale N di dati presenti) la calibrazione di un modello, utilizzando come dati di training tutti i dati conosciuti tranne uno, che varia di volta in volta. Il dato escluso in calibrazione è poi utilizzato come dato di test, per quella iterazione del metodo. Si ottengono, quindi, tanti modelli, ragionevolmente molto simili tra loro, nessuno dei quali ha utilizzato per la propria calibrazione il proprio dato di test. È quindi possibile considerare tutti i risultati del test come parte di un unico test, performato su un campione di n dati.

Per spiegare il metodo di scelta della soglia migliore, nonché gli indici di affidabilità di un modello logistico, è necessario introdurre alcuni concetti.

Il primo concetto fondamentale è quello di matrice di confusione. Essa è la matrice, calcolabile sia dal campione di training, che dal campione di test (scelta preferibile), che conta le osservazioni, in base alla previsione del modello e all'effettiva realizzazione dell'evento, classificandole come segue:

- veri positivi (TP),
- falsi positivi (FP),
- veri negativi (TN),
- falsi negativi (FN),

e organizzando i conteggi in forma matriciale. Un esempio di matrice di confusione è riportato in Tabella 3.1.

Tabella 3.1: esempio di matrice di confusione

	Evento non verificato	Evento verificato
Evento non predetto (P(Y|X) <= soglia)	Veri negativi (TN)	Falsi negativi (FN)
Evento predetto (P(Y|X)>soglia)	Falsi positivi (FP)	Veri positivi (TP)

A partire dalla matrice di confusione è possibile calcolare molti indici, ognuno dei quali rappresenta un aspetto della precisione del modello. Gli indici che entrano in gioco nella stima della soglia ottimale sono la sensibilità, e 1-specificità. Entrambe esprimono, rispettivamente, la probabilità stimata che il modello classifichi come avvenuto un evento avvenuto, e la probabilità stimata che il modello classifichi erroneamente come avvenuto un evento non realmente avvenuto. Formalmente si hanno le seguenti definizioni:

$$Sensibilità = \frac{TP}{TP + FN},$$
$$1 - Specificità = 1 - \frac{TN}{TN + FP}\ .$$

La curva ROC (Receiver Operating Characteristic) è ottenuta calcolando sensibilità e 1 −specificità per un ampio insieme di soglie comprese tra 0 e 1.

La forma della curva rappresentata varia a seconda delle capacità di discriminazione del modello (ossia della sua probabilità di attribuire valori di probabilità maggiori a dati coincidenti con eventi realmente avvenuti). Tuttavia, ci sono tre punti fissi:

a soglia uguale a 0, tutti gli eventi sono predetti come avvenuti, e dunque la sensibilità sarà pari ad 1, non esistendo eventi predetti come non realizzati. Allo stesso modo, anche il secondo indicatore sarà pari ad 1, per la medesima ragione.

A soglia uguale a 1, tutti gli eventi sono predetti come non avvenuti; dunque, la sensibilità sarà pari a 0, e stesso dicasi per il secondo indicatore, dal momento che la specificità risulterà pari ad 1.

Una distribuzione di probabilità casuale, dunque totalmente priva di capacità discriminante, genererà, su quel grafico, una curva simile alla bisettrice. Tanto maggiore sarà la capacità di discriminare del modello, tanto più la curva supererà la bisettrice.

Le informazioni fornite dalla curva ROC sono riassunte dall'area sotto la curva ROC che fornisce una misura sintetica della capacità discriminatoria: valori di questa area prossimi a 0,5 indicano assenza di potere discriminante, mentre valori molto elevati cioè prossimi a 1 indicano elevata capacità di classificazione. Questo indice viene comunemente denominato AUC (Area Under the Curve).

Inoltre, dal momento che questo sistema testa i risultati del modello per differenti soglie di probabilità, è possibile utilizzarlo per determinare una soglia ottimale. In questo elaborato, la soglia ottimale è stata individuata massimizzando l'indice:

$$Sensibilità + Specificità - 1$$

che corrisponde alla massima distanza dalla bisettrice nel grafico ROC.

3.5.4. Selezione dei predittori

I metodi per selezionare, a partire da un modello iniziale, un insieme di predittori significativi, allo scopo di ridurne la complessità senza perdere eccessivamente in capacità esplicativa, sono molteplici.

Il primo e più immediato criterio risulta consiste nella scelta di predittori non collineari. Un predittore è collineare con altri se può essere approssimativamente espresso come combinazione lineare di essi, risultando quindi fortemente spiegato dagli altri regressori presenti nel modello. La presenza di predittori collineari tende a rendere instabili i coefficienti stimati, aumentando la varianza delle stime e introducendo informazioni ridondanti che contribuiscono unicamente ad accrescere la complessità del modello senza apportare reale contenuto informativo.

Per misurare la collinearità di un predittore rispetto agli altri si utilizza l'indice Variance Inflation Factor (VIF), calcolato come:

$$VIF_j = \frac{1}{1 - R_j^2}$$

dove R_j^2 è l'R^2 (indice presentato in precedenza) del modello lineare che ha come variabile dipendente (risposta) il j° esimo predittore X_j e come regressori (predittori) tutti le altre variabili indipendenti.

In genere, un valore di VIF superiore a cinque può essere ritenuto sintomo di collinearità eccessiva della variabile in questione. Per questo, in tali casi la variabile viene rimossa e il modello ricalcolato.

Una seconda scrematura delle variabili può essere effettuata attraverso l'indice AIC (Akaike Information Criterion), che combina in un'unica metrica due aspetti distinti del modello:

- la capacità predittiva del modello, misurata attraverso la funzione di verosimiglianza o il suo logaritmo;
- la complessità del modello stesso, stimata tramite il numero dei predittori inclusi.

Nel contesto della regressione logistica, si assume che la variabile risposta segua una distribuzione bernoulliana, i cui parametri dipendono dai predittori tramite la funzione logistica. Per ciascuna osservazione campionaria $(X_{it}\ ;\ Y_{it})$, viene dunque calcolata $p\ (Y_{it} \mid \sum_{k=1}^{p} \beta_k X_{k;\,it})$. Le probabilità così calcolate vengono poi, in base alla proprietà dell'estrazione di eventi indipendenti, moltiplicate fra loro o, equivalentemente, ne viene calcolato il logaritmo e sommate (in scala logaritmica). Il valore così calcolato è la verosimiglianza, nel primo caso, o log verosimiglianza, nel secondo.

Tornando all'indicatore AIC di un determinato modello esso si calcola come

$$AIC = -2\,log\,(L) + 2p$$

dove L è la verosimiglianza e p è il numero di parametri.

L'indice AIC non fornisce una valutazione assoluta della qualità di uno specifico modello, ma permette di confrontare le prestazioni di modelli alternativi, simili (nel senso che per esempio condividono parte dei medesimi predittori e tentano di predire la medesima variabile) e stimati sugli stessi parametri. Una riduzione dell'indice AIC rispetto a un modello di base è tipica di modelli più semplici e non significativamente meno precisi, o viceversa più precisi ma non significativamente più complessi del modello base stesso.

Per questo motivo è possibile utilizzare un metodo basato sull'indice AIC per verificare la rilevanza dei singoli predittori in un modello. Il metodo è quello della selezione a ritroso. Esso si basa sul confrontare, di ciclo in ciclo, l'indice AIC del modello utilizzato, con l'indice di un modello simile, sui medesimi dati, ma privo di un predittore. Se l'indice AIC di uno dei modelli ottenuti risulta significativamente inferiore di quello del modello originale, quel predittore viene eliminato, e il ciclo ricomincia.

L'ultimo metodo che è stato utilizzato è il più semplice Likelihood Ratio Test (LRT). Questo test verifica l'impatto della rimozione di uno specifico predittore dal modello, confrontando la verosimiglianza del modello completo con quella di un sotto-modello ridotto, stimato sui medesimi dati ma privo del predittore sotto esame.

La statistica di test è data da:

$$\Lambda = -2\big(\log\, L_{ridotto} - \log\, L_{completo}\big),$$

e, sotto opportune condizioni, la sua distribuzione asintotica è chi-quadro con gradi di libertà pari alla differenza nel numero di parametri tra i due modelli.

Lo svantaggio di questo metodo rispetto al precedente è che non penalizza esplicitamente l'incremento di complessità del modello, mentre il vantaggio del metodo è che la distribuzione asintotica della statistica LRT è nota; quindi, questo metodo consente di valutare formalmente la significatività del contributo di una singola variabile, con un test di ipotesi.

Per questo motivo, nel presente lavoro, il test LRT è stato utilizzato come ultimo passaggio di selezione delle singole covariate, una volta applicati e superati i criteri di multicollinearità e di confronto tramite AIC.

4 Analisi esplorativa dei dati

Il presente capitolo riporta i risultati dell'analisi empirica condotta sui dati descritti nel Capitolo 2 e modellati secondo l'impostazione metodologica illustrata nel Capitolo 3.

L'esposizione segue un percorso progressivo: si parte da un'analisi esplorativa complessiva del campione di disastri considerato, utile a descriverne le principali caratteristiche e a evidenziarne eventuali criticità dei dati, quali la presenza di valori mancanti, per poi passare all'analisi delle variabili territoriali e demografiche e, infine, all'esame della variabilità osservata tra Stati.

L'obiettivo è valutare preliminarmente in quale misura il consumo di suolo, unitamente ad altre caratteristiche territoriali e climatiche, risulti associato ai danni osservati in occasione di disastri naturali nel periodo 2006–2018.

Le evidenze emerse in questo capitolo costituiscono la base per le successive analisi modellistiche presentate nel Capitolo 5.

4.1. Dati di danno ambientale

L'analisi esplorativa ha l'obiettivo di descrivere la struttura del campione di disastri naturali considerato e di valutare le caratteristiche degli indicatori di danno, al fine di orientare le successive scelte modellistiche.

Il dataset analizzato comprende 476 disastri naturali avvenuti in Europa nel periodo 2006–2018.

4.1.1. Numero e tipologia dei disastri

Per seguire gli scopi di questo studio è stata utilizzata una sezione del database EM-DAT, ossia quella relativa ai disastri:

- ✓ classificati come naturali,
- ✓ avvenuti in Europa,
- ✓ avvenuti tra il 2006 e il 2018.

Il numero totale di dati utili presenti nel database è di 476. La Figura 4.1 mostra il numero di disastri, per tipo, nell'area e nel periodo oggetti di studio.

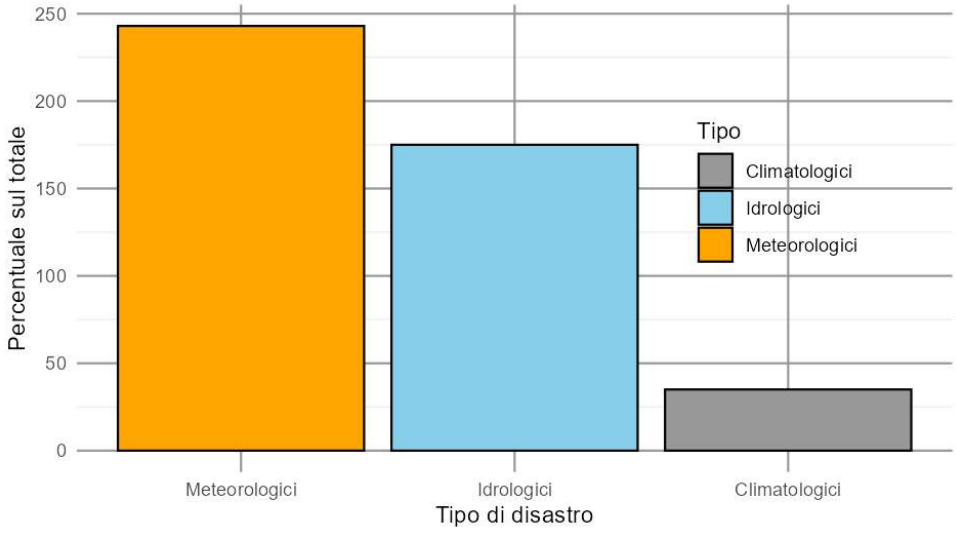

Figura 4.1: numero di disastri naturali censiti da EM-DAT, nell'area e nel periodo di interesse, per tipo.

Il tipo di disastro più presente, che costituisce circa il 50% dei dati disponibili, è quello meteorologico, seguito da quello idrologico, che ricopre circa il 40% dei dati. Questo fatto è spiegato dal ridotto numero di eventi geofisici (terremoti ed eruzioni vulcaniche) e dai ridotti danni a persone o cose, degli eventi climatologici (siccità e incendi).

4.1.2. Distribuzione spaziale e temporale

La Figura 4.2 mostra il numero di disastri naturali censiti per Stato nel periodo 2006-2018. Si osserva una marcata eterogeneità nella distribuzione spaziale degli eventi. Gli Stati europei con il maggior numero di disastri censiti sono l'Italia (primo Stato, con il numero maggiore di terremoti), Francia e Germania. Essi sono Stati economicamente avanzati, quindi con sistemi di monitoraggio ben consolidati, e più densamente popolati, e di grandi dimensioni. Al contrario gli Stati con meno eventi censiti, Lussemburgo, Islanda, Svezia, sono Stati di dimensioni minori, o con minore densità di popolazione, e dunque minori probabilità di subire danni rilevanti da eventi naturali avversi. Il numero di disastri è espresso in termini assoluti e non normalizzato rispetto alla superficie territoriale o alla popolazione; pertanto, le differenze osservate devono essere interpretate con cautela, in quanto riflettono sia l'effettiva frequenza degli eventi sia la diversa dimensione e struttura demografica degli Stati.

Il numero medio di disastri per Stato nel periodo considerato è pari a 15,6, con una deviazione standard pari a 12, a conferma dell'elevata dispersione tra Paesi.

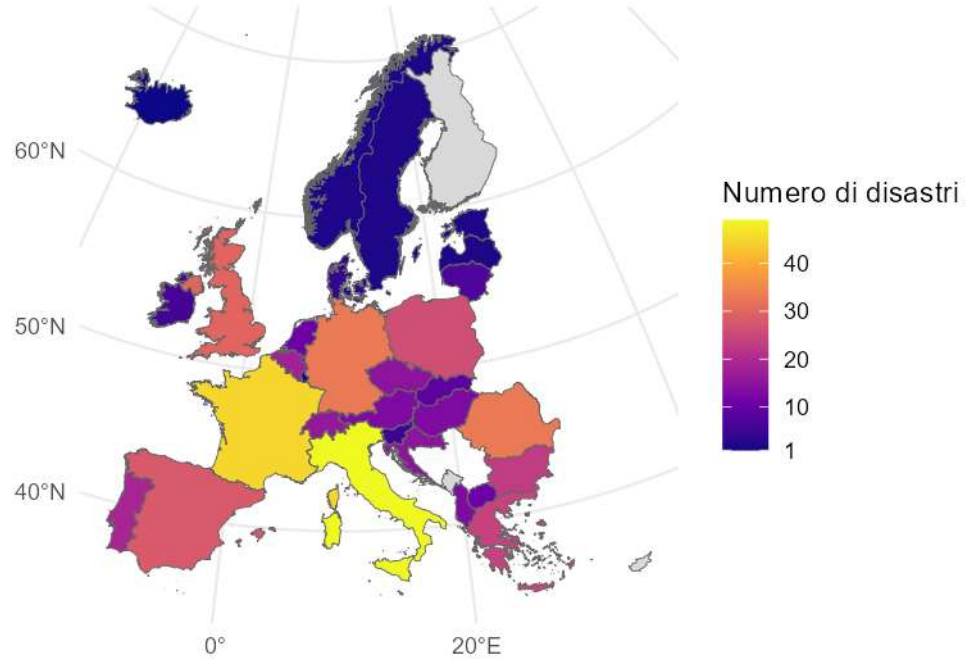

Figura 4.2: Numero di disastri naturali censiti da EM-DAT, per Stato, nel periodo e area di interesse.

A livello cronologico (come illustrato in Figura 4.3), si osserva una grande variabilità, con un minimo molto evidente nell'anno 2011 (con 13 disastri), mentre nel 2007 ne è registrato il massimo (55). Non è evidenziata alcuna tendenza temporale.

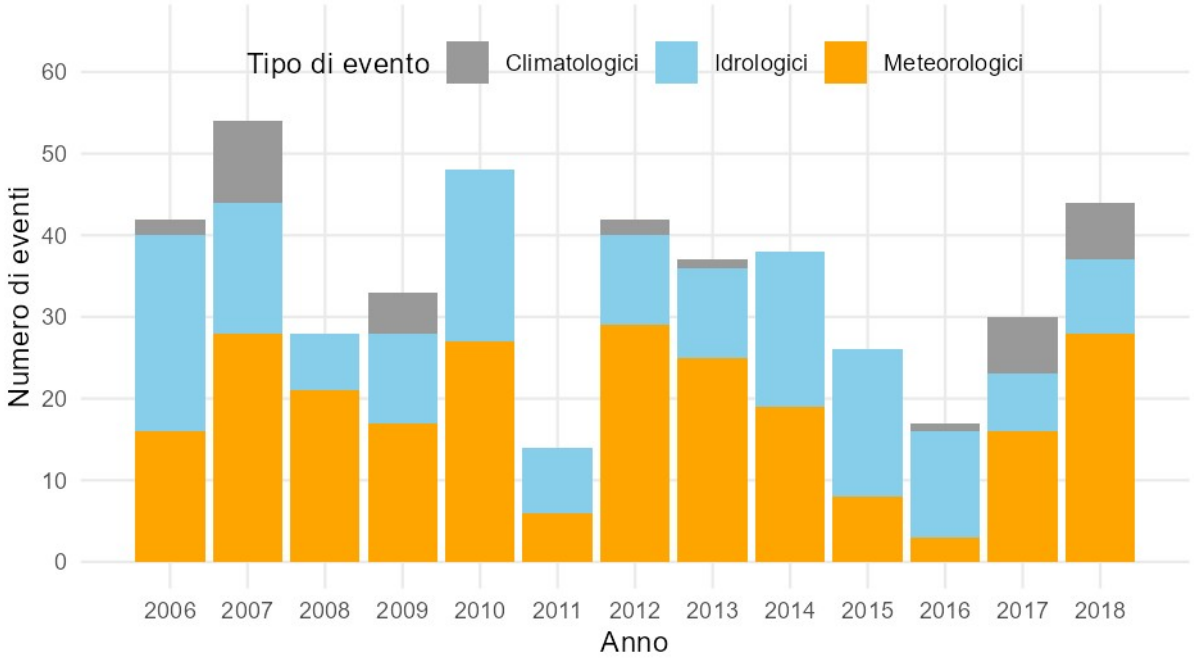

Figura 4.3: numero di disastri utili registrati in EM-DAT per anno

La variabilità osservata, sia a livello spaziale sia temporale, conferma la natura disomogenea dei fenomeni analizzati e suggerisce cautela nell'interpretazione di eventuali confronti diretti tra Stati o tra singoli anni.

4.1.3. Sintesi degli indicatori di danno

Gli indicatori di danno presenti all'interno del database EM-DAT sono i seguenti:

a) morti totali: Numero totale di persone decedute a causa diretta del disastro. Include sia i morti immediati sia quelli deceduti successivamente a seguito delle conseguenze dirette dell'evento (ad esempio ferite o condizioni sanitarie deteriorate).

b) Feriti: Numero di persone che hanno riportato lesioni fisiche tali da richiedere assistenza medica immediata come conseguenza diretta del disastro.

c) Affetti: Numero di persone che hanno subito impatti diretti non letali, tra cui:

    a. evacuazione o sfollamento;

    b. perdita dell'abitazione;

    c. perdita dei mezzi di sussistenza;

d) senzatetto: Numero di persone la cui abitazione è stata distrutta o resa inabitabile, costringendole a cercare un alloggio temporaneo o permanente.

e) Totale persone colpite: pari alla somma tra feriti, affetti e senzatetto. Grazie alla presenza di questa, le tre variabili precedenti non saranno prese in considerazione separatamente.

f) Danni economici totali: Stima del valore economico complessivo dei danni diretti causati dal disastro, espressa in dollari statunitensi correnti. Include:

    a. danni a edifici residenziali e infrastrutture;

    b. perdite nel settore agricolo e industriale;

    c. distruzione di beni materiali.

Tuttavia, non sono riportati obbligatoriamente per tutti i disastri. Al contrario, per una quantità significativa di disastri questi indicatori non sono presenti, o perché non riportati dalle fonti ufficiali, o a causa della presenza di informazioni discordanti. A livello generale, il 32% dei 476 disastri analizzati non presenta il dato di decessi, il 50% quello della popolazione colpita, e il 68% quello dei danni economici.

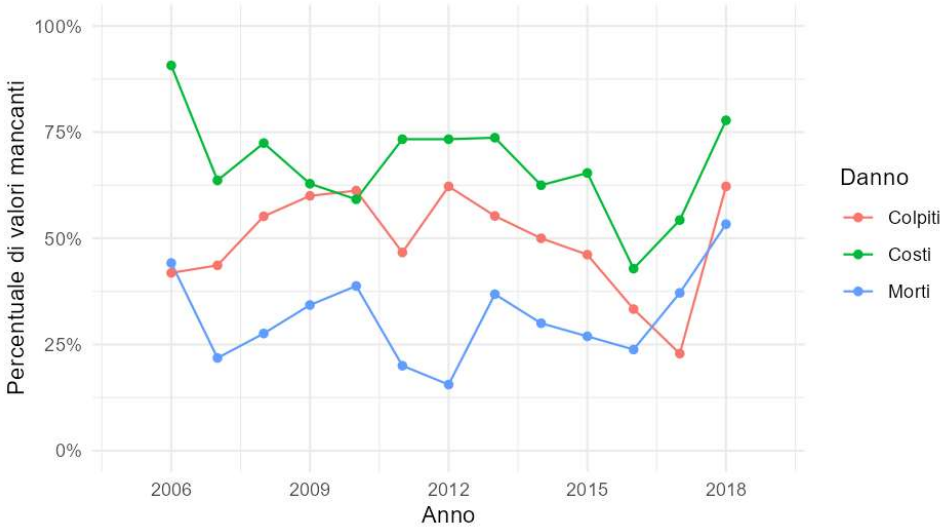

Figura 4.4: percentuale di dati mancanti per anno, per indicatore di danno

In generale, il 12 % degli eventi riportati, non presenta alcun dato circa i danni che ha causato, il 43 % presenta il dato di un solo indicatore, il 27% presenta dati di due indicatori, e solo il 17% presenta dati per tutti e tre gli indicatori.

Un'analisi della percentuale di dati mancanti per anno, in ciascuno dei tre indicatori, è visibile in Figura 4.4. Non si vedono tendenze di crescita o decrescita di tale numero, ma l'assenza di dati sembra essere un problema permanente.

È dunque insensato tentare di dedurre l'indicatore mancante a partire dai dati degli indicatori presenti, in quanto a tal fine si potrebbero utilizzare meno del 45% dei dati, e non risolverebbe il problema del 12% di dati, che non presenta alcuna indicazione.

4.1.4. Magnitudo e durata

Il database può includere informazioni in merito alla magnitudo del disastro in diverse unità di misura, a seconda del tipo di evento dannoso che si verifica. La presenza o meno di queste informazioni dipende, come per tutte le altre informazioni specifiche circa ogni disastro, alla pluralità e alla concordanza delle fonti, istituzionali e non istituzionali, che forniscano dati su un determinato disastro. La Tabella 4.1 mostra come questa informazione sia distribuita a seconda del tipo di evento.

Si nota come, in generale, l'indicazione circa la magnitudo non sia presente per la maggior parte dei disastri. Inoltre, essa viene espressa, logicamente, in unità di misura molto differenti, in base al tipo di disastro in questione. Questo rende la magnitudo un'informazione non utilmente sfruttabile in fase di aggregazione dei dati.

Tabella 4.1: presenza di indicazioni circa la magnitudo dei disastri per tipo di evento

Sottogruppo di disastro	Climatologici	
Tipo di disastro	Siccità	Incendi
Dati senza magnitudo	5	30
Sottogruppo di disastro	Idrologici	
Tipo di disastro	Alluvione e Piena	Frana
Dati senza magnitudo	112	2
UdM della magnitudo	km2	
Dati con magnitudo	61	
Sottogruppo di disastro	Meteorologici	
Tipo di disastro	Tempesta	Temperature estreme
Dati senza magnitudo	94	57
UdM della magnitudo	km/s	°C
Dati con magnitudo	44	48

Esiste un secondo indicatore che può essere utilizzato come proxy di gravità e può essere calcolato per molti dei disastri. Esso è la loro durata in giorni, ottenuta semplicemente sottraendo la data iniziale dalla data finale dei dati che le riportano. Questo indicatore è noto per circa l'85% dei disastri, e presenta l'andamento illustrato in Figura 4.5**.** La durata mediana risulta pari a 2 giorni, con un intervallo interquartile pari a 5, e un massimo di 131 giorni, evidenziando una distribuzione fortemente asimmetrica.

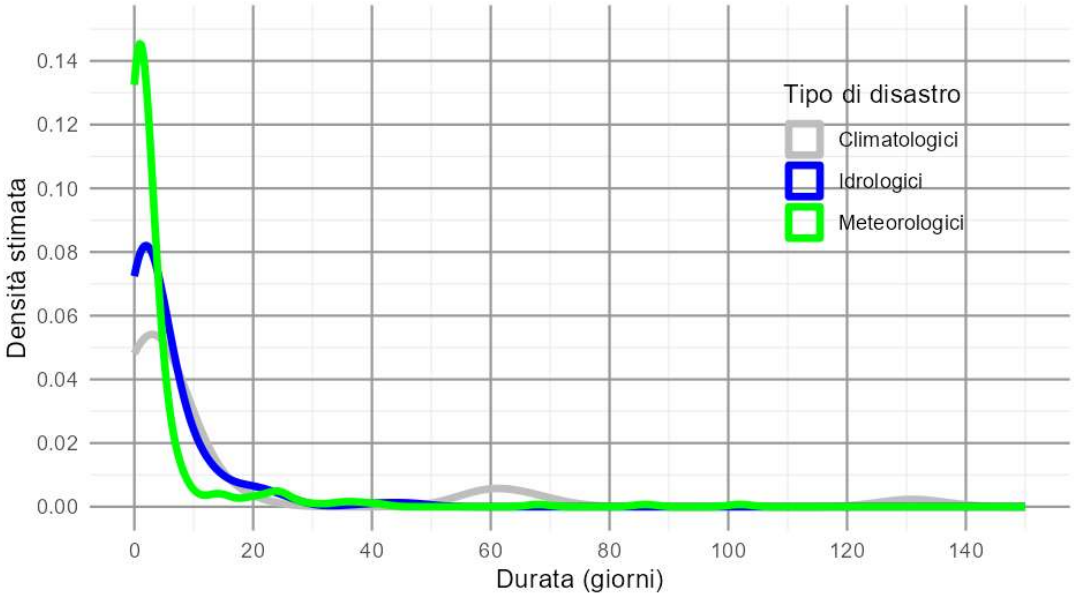

Figura 4.5: Densità stimata della durata degli eventi causanti disastri naturali, metodo kernel (KDE)

Si osserva che la durata degli eventi varia a seconda del tipo di evento. Se, per tutti i tipi il 75% degli eventi ha una durata compresa tra 0 e 6 giorni, i valori massimi variano molto, a causa di singoli eventi. La durata degli eventi meteorologici, per esempio, è spesso tra 0 e 3 giorni, avendo quindi un intervallo interquartile anche minore della media, ma l'evento più lungo registrato è durato 86 giorni, e consiste in un periodo di freddo intenso avvenuto in Polonia. Allo stesso modo, gli eventi climatologici, hanno quartili pari a 1 e 6 giorni, essendo quindi leggermente più lunghi della media, ma l'evento più lungo registrato ha durata 131 giorni.

Proprio per questi motivi, utilizzare la durata degli eventi come indice della loro gravità è un metodo non sempre efficace. Molti eventi, come le piene o le alluvioni, sono dannosi proprio in virtù della loro breve durata, mentre altri eventi, come i fenomeni siccitosi, sono dannosi a causa della loro estrema lunghezza.

Queste osservazioni rendono, in definitiva, non utile per questa analisi, considerare un indicatore di gravità oggettiva dell'evento dannoso. Esso rimane fondamentale in caso di analisi più ristrette, ad un unico tipo di evento, e per questo sarebbe augurabile un miglioramento della qualità dei dati in merito.

4.1.5. Localizzazione degli eventi

Il dataset GDIS, come precedentemente descritto, è un dataset la cui unica utilità consiste nel poter meglio localizzare gli eventi disastrosi contenuti in EM-DAT. Tuttavia, esso si limita a geolocalizzare solo una parte degli di tali eventi, sia perché si limita a quelli precedenti al 2018, sia perché, parte degli eventi attualmente presenti in EM-DAT sono stati inseriti dopo l'uscita dell'ultima versione di GDIS.

Come si vede in Figura 4.6, in cui si confrontano il numero di eventi memorizzati in EM-DAT, con il numero di eventi localizzati in GDIS per anno, questo problema è sistemico consistente, e non accenna a diminuire con il tempo.

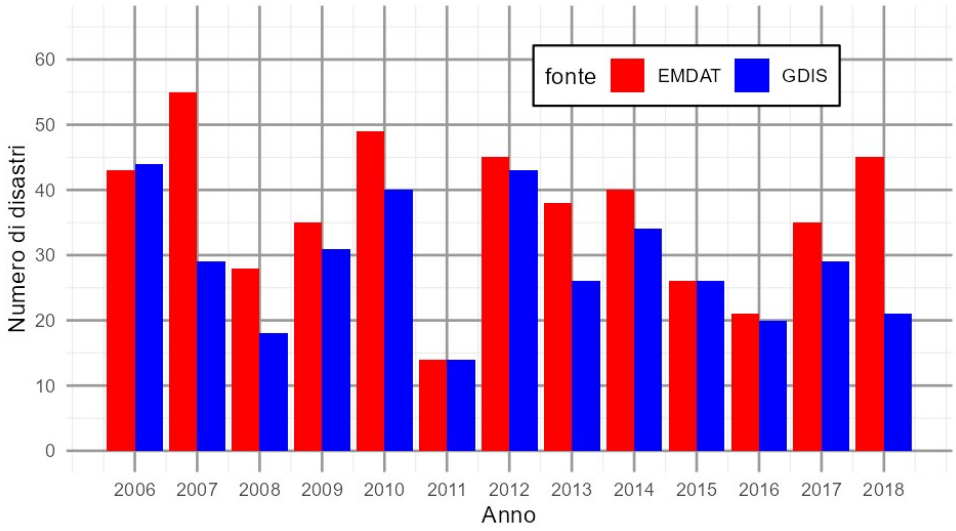

Figura 4.6: confronto tra numero di disastri presenti in EM-DAT e riportati in GDIS, per anno

Nella zona e nel periodo di interesse, i 2344 dati presenti nel sottoinsieme di GDIS analizzato, riguardano unicamente 375 eventi, contro i 473 presenti in EM-DAT, per lo stesso periodo e area geografica. Ulteriore problematica circa la geo referenziazione di EM-DAT, applicabile per esempio attraverso GDIS, è la difficile scomposizione die dati di danno. Tali dati (costi, popolazione colpita, e decessi) sono, infatti, noti sottoforma di un unico valore, valido per l'intero disastro, in tutto lo Stato che ha colpito. Non è immediato, dunque, dividere questo valore se si analizza che il disastro ha colpito differenti regioni o province.

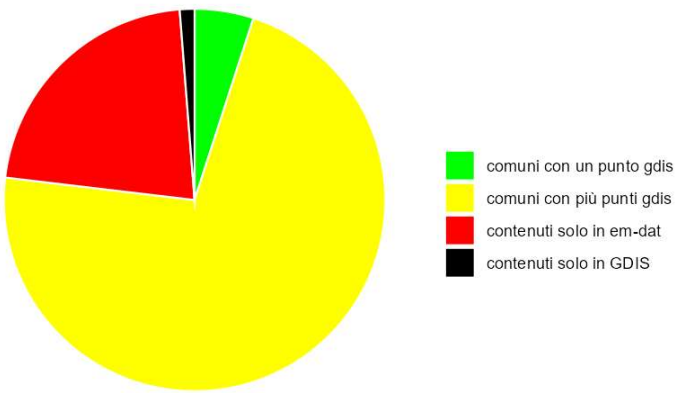

Figura 4.7: percentuale di disastri, in relazione alla loro presenza su EM-DAT e GDIS

La portata di questo problema è visibile in Figura 4.7, in cui si vede la percentuale di dati in base alla presenza di dati in EM-DAT e in GDIS.

I risultati sono i seguenti:

- EM-DAT include 476 eventi distinti, che avvengono in Stati appartenenti all'area di interesse
- GDIS include 2340 records, corrispondenti a 375 eventi distinti
- Il totale di identificativi univoci presenti nei due dataset è 478.

La distribuzione è così articolata:

- 370 eventi sono presenti in entrambi i dataset e possono essere confrontati
- 103 eventi sono presenti solo in EM-DAT
- 3 eventi sono presenti solo in GDIS.

Di questi 370 eventi condivisi, soltanto 24 presentano una corrispondenza a uno a uno tra EM-DAT e GDIS (ovvero un singolo record GDIS per evento EM-DAT).

Nei restanti casi, GDIS rappresenta l'evento attraverso più località, rendendo necessario suddividere le stime di impatto di EM-DAT fra i diversi punti, secondo un criterio da definire.

Si è inoltre verificato che gli eventi contrassegnati dallo stesso nome in EM-DAT e GDIS condividono anche tipologia di disastro, anno e Paese, che costituiscono le uniche ulteriori variabili comuni tra GDIS ed EM-DAT. Non sono stati riscontrati casi di eventi con denominazioni differenti ma coincidenti per tutte e tre queste caratteristiche. Di conseguenza, l'associazione tra record GDIS ed eventi EM-DAT è stata effettuata esclusivamente sulla base del nome dell'evento, ritenuto identificativo univoco nel contesto dei dataset analizzati.

In conclusione, per massimizzare il numero di dati geo localizzabili ai quali si ha accesso, sono stati utilizzati i dati contenenti in GDIS così come sono, ed è stata tentata una localizzazione di quei disastri memorizzati in EM-DAT, ma senza una controparte in GDIS, attraverso l'uso delle informazioni memorizzate in "Location" di EM-DAT stesso, come descritto nella sezione 3.1.1.

Questa procedura ha portato alla localizzazione di altri 88 disastri, localizzando così tutti i 476 disastri contenuti in EM-DAT, relativamente al periodo 2006/2018 in Europa.

4.2. Impermeabilizzazione del suolo

4.2.1. Scelta delle variabili di stato

Il secondo versante in cui è stato necessario muoversi, è stata la scelta di uno, o pochi, indicatori di impermeabilizzazione di suolo, da associare alle province, e il calcolo dei medesimi indicatori per il territorio statale.

Gli indicatori fra cui scegliere consistono in area impermeabilizzata totale, percentuale di suolo impermeabilizzato, e area impermeabilizzata pro capite. Tali indicatori presentano una variabilità temporale molto ridotta, causata dalla durata, di fatto perpetua, dei fenomeni insediativi. Non sorprende, quindi, la mancanza di differenze significative tra le densità stimate dei dati del 2006 e quelli del 2018, per le tre variabili, come visibile in Figura 4.8.

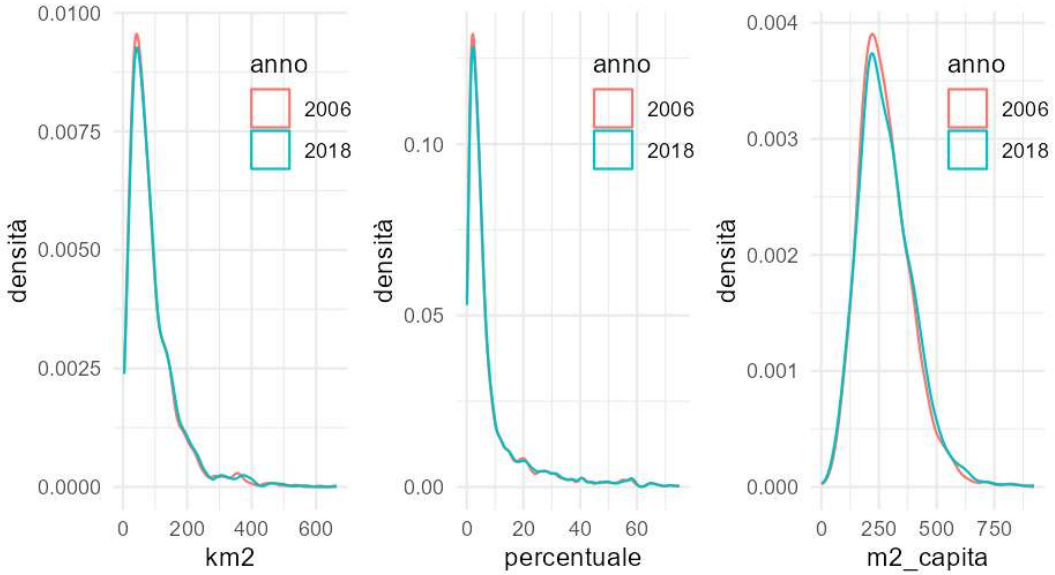

Figura 4.8: Confronto dei dati di impermeabilizzazione del suolo tra 2006 e 2018

Per tale motivo, per questa analisi preliminare, sono stati utilizzati i dati di un unico anno, ossia il 2006, mentre per la produzione della tabella finale e le analisi successive, sono stati linearizzati i dati posseduti.

Allo scopo di completare l'informazione, sempre in questa fase preliminare, è stato deciso di considerare, anche la variazione di queste variabili tra il 2006 e il 2018, per verificare sinteticamente la portata dei cambiamenti in atto.

Un aspetto rilevante emerso in fase esplorativa riguarda la relazione tra gli indicatori di impermeabilizzazione e alcune caratteristiche strutturali delle unità territoriali, in particolare superficie e popolazione. La verifica di tali correlazioni risponde a due motivazioni.

In primo luogo, data l'elevata eterogeneità nella superficie delle unità NUTS3 — variabile da 14 km² a circa 100.000 km² — è necessario accertare se eventuali relazioni osservate non siano semplicemente un effetto di scala territoriale.

In secondo luogo, poiché l'impermeabilizzazione del suolo è strettamente connessa ai processi di urbanizzazione e alla presenza umana, è plausibile attendersi una correlazione con la popolazione residente. L'analisi preliminare di tali relazioni consente inoltre di valutare possibili problemi di collinearità nelle successive specificazioni modellistiche.

In Tabella 4.2 sono riportate le correlazioni tra i tre indicatori di impermeabilizzazione e superficie e popolazione (dell'anno 2006, per coerenza con le variabili di impermeabilizzazione del suolo). Considerata la forte asimmetria nella distribuzione di area e popolazione, sono state calcolate anche le correlazioni sulle corrispondenti trasformazioni logaritmiche.

Tabella 4.2: correlazione tra area, popolazione, e variabili di impermeabilizzazione di suolo

variabile	Correlazione con area	Correlazione con $\log_{10}$ (area)	Correlazione con popolazione 2006	Correlazione con $\log_{10}$ (popolazione 2006)
km^2 imp. 2006	0,322	0,527	0,785	0,762
m^2 capita imp. 2006	0,426	0,506	-0,281	-0,298
% suolo imp. 2006	-0,258	-0,737	0,143	0,142
km^2 diff 2018-2006	0,262	0,429	0,652	0,573
m^2 capita diff 2018-2006	0,067	0,230	-0,172	-0,211
% diff 2018-2006	-0,215	-0,496	0,185	0,143

Come atteso, l'area impermeabilizzata in termini assoluti (km^2) risulta positivamente correlata con la superficie e con la popolazione. Al contrario, la percentuale di suolo impermeabilizzato mostra una correlazione negativa con la superficie, che risulta più marcata se si considera il logaritmo dell'area stessa.

Una motivazione per questo fatto è la natura non casuale delle divisioni amministrative. Si può infatti osservare che le aree più piccole coincidono spesso con grandi metropoli, molto urbanizzate, mentre quelle più vaste rappresentano province meno densamente popolate, più agricole, e dunque con una minore percentuale di suolo impermeabilizzato.

Interessante è la correlazione positiva fra area della provincia e metri quadri pro capite. Questo significa che le province di maggiori dimensioni possono presentare un consumo di suolo accessorio alla popolazione più che proporzionale all'incremento della popolazione stessa. Al contrario, aree di dimensioni ridotte, e dunque con popolazione concentrata, presentano un utilizzo del territorio, ironicamente, più razionale. Parte di questo fenomeno è, probabilmente, attribuibile alla crescente necessità di costruire vie di comunicazione che colleghino aree lontane fra loro, necessità che, nelle aree di dimensioni maggiori, si manifesta in maniera quasi indipendente dal numero effettivo ci persone che usufruiscono effettivamente di quell'infrastruttura.

Per quanto riguarda le variazioni (differenze) 2006–2018, le correlazioni risultano coerenti: l'aumento in km² è più elevato nelle province più grandi e popolose, mentre la variazione percentuale tende a mantenere una relazione negativa con la superficie.

4.2.2. Livello iniziale (2006) e variazione nel periodo 2006–2018 (analisi congiunta)

Cominciando ad analizzare l'impermeabilizzazione del suolo a livello provinciale, si nota che essa varia molto tra province. Presenta un minimo attorno all'1%, e dei massimi che superano il 65%. I valori minimi vengono, tendenzialmente, assunti da province molto vaste e poco popolate, mentre quelli massimi da aree NUTS definite all'interno di grandi centri urbani. Ciò è stato già precedentemente osservato dalla forte correlazione negativa tra superficie della provincia e percentuale di suolo impermeabilizzato.

In Figura 4.9 si osserva lo scutterplot di queste due variabili. È stata eseguita una classificazione tramite il metodo k-means, in quattro classi (numero trovato come ottimale con il metodo del gomito), ed è stata stimata una retta di regressione. Si nota che, nella definizione delle classi, la differenza di percentuale non ha giocato alcun ruolo, vista la sua ridotta variabilità rispetto alla percentuale iniziale. Si nota altresì come, tendenzialmente, province scarsamente urbanizzate a livello percentuale (vale a dire, sovente, province di dimensioni maggiori, vista l'elevata correlazione negativa precedentemente osservata, tra percentuale di impermeabilizzazione e superficie), tendano a non subire intensi fenomeni di urbanizzazione a breve termine, viceversa, province più urbanizzate inizialmente tendano a peggiorare ulteriormente la loro situazione con facilità.

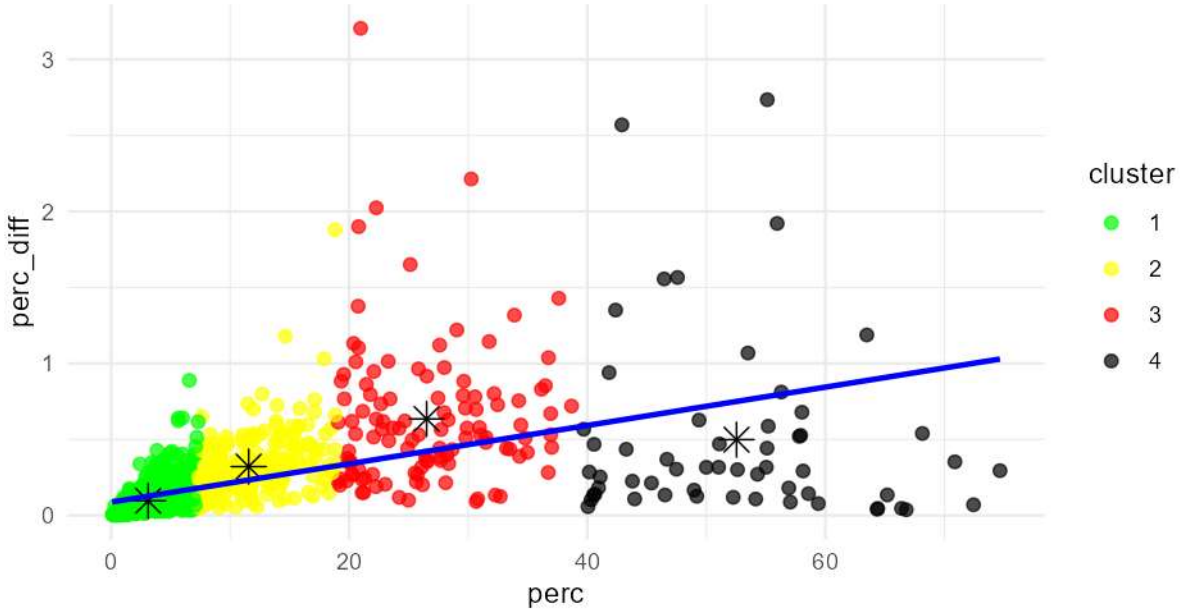

Figura 4.9: scatter plot percentuale e differenza, con classi kmeans e retta di regressione

In base alla stessa classificazione è stata rappresentata una mappa, visibile in Figura 4.10.

Si nota, sommariamente visto le dimensioni medie delle province rispetto alla scala, che la maggioranza delle unità NUTS3 (province) risulta molto poco impermeabilizzata (cioè ricade in quella che è stata identificata come classe 1), mentre solo le aree prossime a grandi centri urbani (come, per esempio, le zone di Londra, Parigi, Lisbona, Milano) cadono nei cluster superiori. La maggiore presenza di aree ad alta impermeabilizzazione in unità territoriali di piccola estensione è coerente con la relazione negativa tra percentuale impermeabilizzata e superficie.

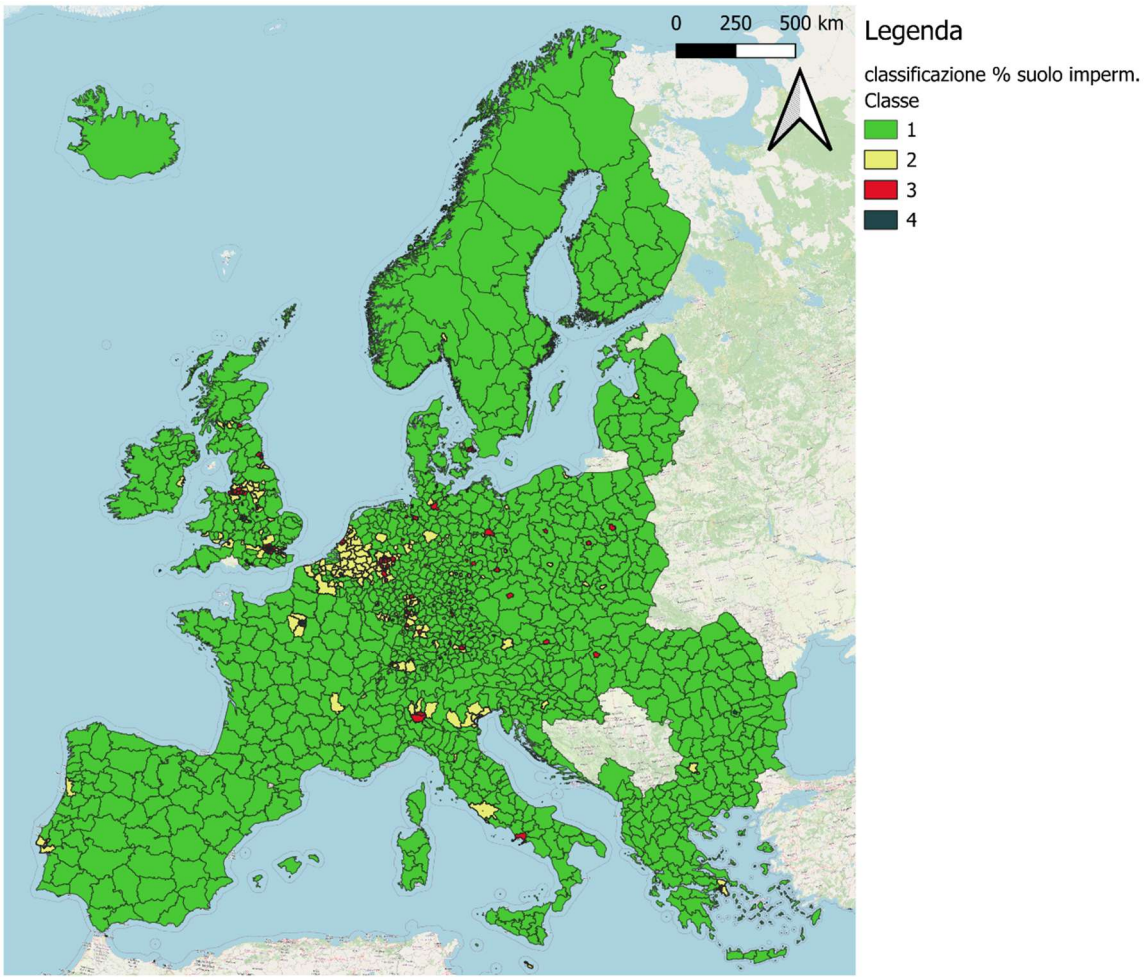

Figura 4.10: mappa della classificazione delle aree per percentuale di suolo impermeabilizzato.

Per una lettura e una descrizione più dettagliata, la Figura 4.11 Riporta un ingrandimento della mappa precedente, mostrante la sola Italia. Si osserva come la maggior parte delle province della penisola rientrano nella classe di impermeabilizzazione più bassa (classe 1). La Pianura Padana rientra prevalentemente nella classe intermedia 2, mentre l'area di Milano ricade nella classe più elevata 3. Sebbene non siano visibili aree in classe 4, le quali coincidono con grandi metropoli europee, come Parigi e Londra (come visibile nella mappa in figura precedente), si osserva esattamente quello che intuitivamente ci si attenderebbe. Non piacevole sorpresa è, però, osservare in classe 3 Napoli, e in classe 2 le province di Roma, Latina, Lecce e Ragusa.

Va ricordato che, l'appartenenza di una provincia ad una determinata classe, ed in particolare alla classe 1, non significa che tutto il territorio della stessa risulti scarsamente impermeabilizzato, ma si tratta di una media, per di più basata sui dati stessi. L'impermeabilizzazione rimane un valore numerico singolo, che nulla dice su come e su dove sia avvenuta questa impermeabilizzazione. Si ricorda infine che l'appartenenza a una determinata classe deriva da un indicatore medio a livello di unità amministrativa e non descrive la distribuzione intra-provinciale dell'impermeabilizzazione; pertanto, valori elevati possono riflettere un'urbanizzazione concentrata su porzioni limitate del territorio.

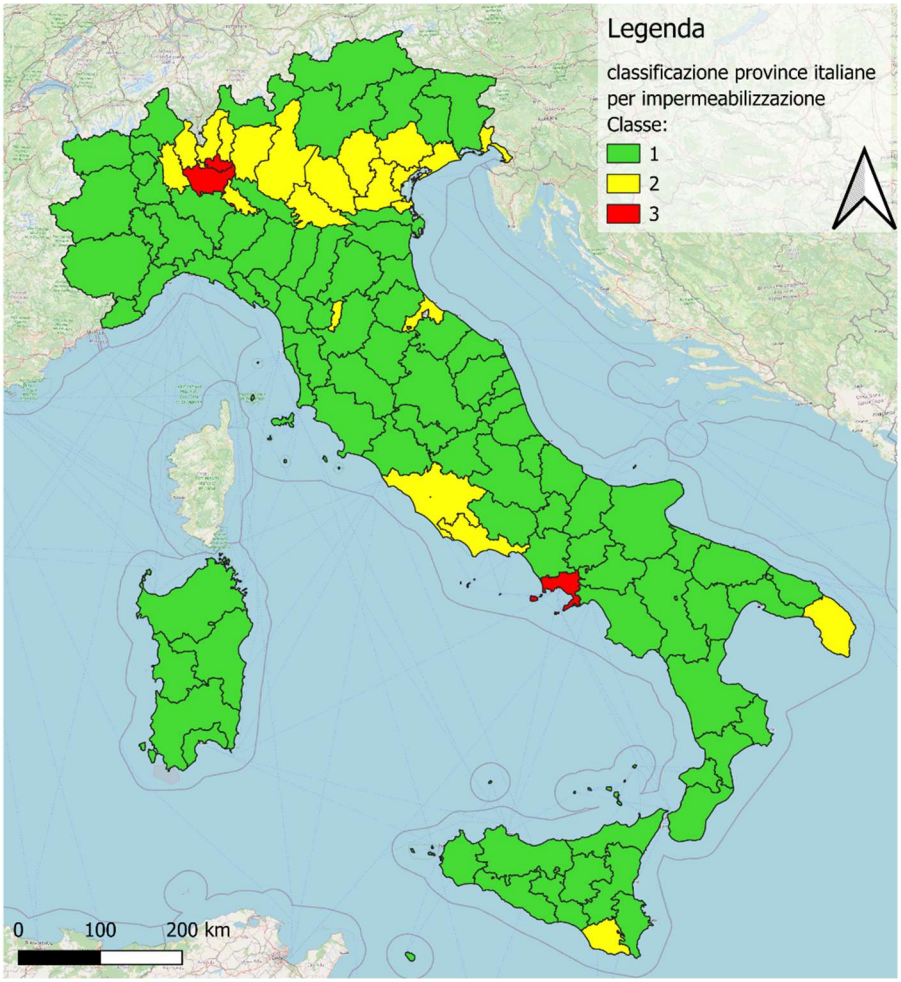

Figura 4.11: Ingrandimento della mappa precedente, con riferimento all'Italia

4.3. Analisi della popolazione

La popolazione rappresenta una variabile strutturale di primaria importanza nell'analisi dei danni da disastro, in quanto direttamente connessa sia all'esposizione al rischio sia alla misura degli impatti osservati. In questa sezione si analizzano i livelli di popolazione a scala provinciale, nonché le variazioni intervenute nel periodo 2006–2018, con particolare attenzione alla qualità e completezza del dato.

4.3.1. Imputazione dei dati mancanti

Se a livello statale non si rilevano dati mancanti né dati contrassegnati come provvisori, a livello provinciale e regionale la situazione risulta differente. Nello specifico per alcune province si osserva una carenza di dati nei primi anni del periodo considerato, verosimilmente riconducibile a modificazioni territoriali successive. Inoltre, negli anni 2011-2012 una quota significativa di osservazioni risulta contrassegnata come incerta, probabilmente a causa di cambiamenti di definizioni amministrative. Il numero di province con dato di popolazione mancante o contrassegnato da flag, rispetto al numero totale delle unità territoriali considerate (1388 province complessive), è mostrato in Figura 4.12.

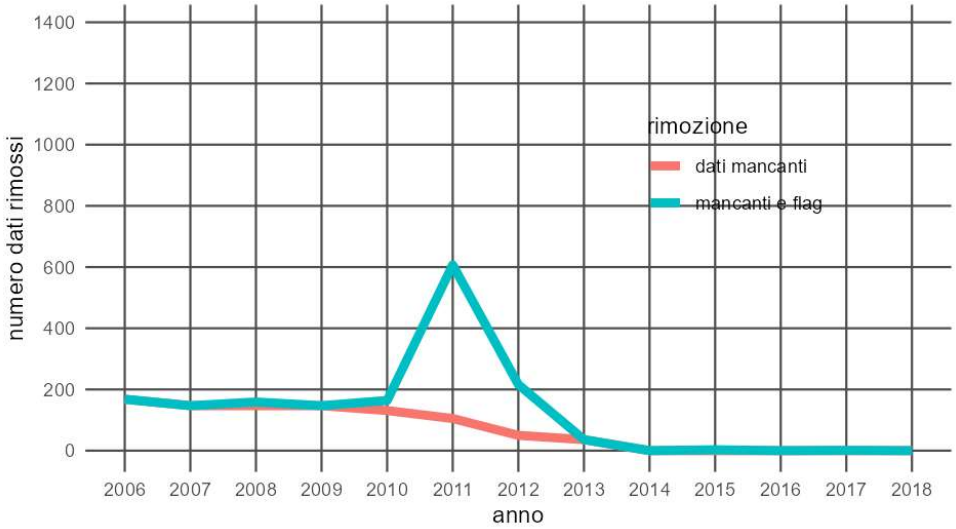

Figura 4.12: numero di province con dato mancante, o con flag, per anno, rispetto al totale

Come si osserva dal grafico, il fenomeno è concentrato in specifici anni, con un picco nel 2011 pari a 600 province, mentre risulta trascurabile negli anni successivi al 2014. Complessivamente, la quota di osservazioni coinvolte nel processo di imputazione rappresenta circa 12% del totale delle osservazioni provincia–anno nel periodo 2006–2018.

Sebbene il numero di province interessate non sia elevato in termini relativi, la variabile popolazione riveste un ruolo centrale nel presente studio, in quanto utilizzata per ripartire a livello sub-statale i dati di danno disponibili unicamente a scala nazionale. È stato pertanto necessario gestire tali mancanze al fine di preservare la completezza del campione. È stato quindi deciso di considerare come mancanti tutti i dati contrassegnati come non affidabili (ossia i dati rappresentati nel grafico in Figura 4.12 dalla linea azzurra). Per non dover eliminare dall'analisi le corrispondenti province completamente, è stato deciso di imputare i dati mancanti mediante un modello di regressione lineare semplice, stimato separatamente per ciascuna provincia.

Tale scelta è giustificata dalla natura relativamente regolare delle dinamiche demografiche nel breve periodo. Nel periodo considerato (2006–2018), pari a 13 anni complessivi, la popolazione provinciale mostra, infatti, andamenti generalmente monotoni e privi di discontinuità strutturali rilevanti. Inoltre, anche dopo l'esclusione delle osservazioni non affidabili,

tutte le province rimaste presentano un minimo di cinque dati disponibili validi, un numero indicato come regola pratica sufficiente per stimare un trend lineare.

I risultati dei modelli, sui dati noti, sono rappresentati in Figura 4.13. Sono stati utilizzati gli indicatori R^2, e RMSE relativo, descritti nella sezione 3.3.1.

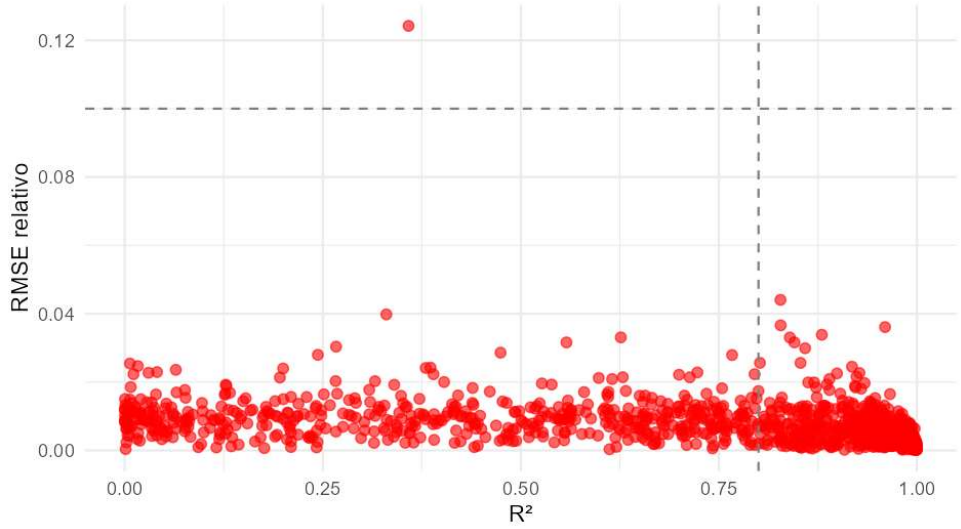

Figura 4.13: Valutazione dei modelli lineari per l'imputazione dei dati di popolazione mancanti

Il coefficiente di determinazione R^2 misura la capacità del modello lineare di spiegare la variabilità osservata nella serie temporale provinciale. I risultati mostrano una certa eterogeneità: circa il 60% dei modelli riesce a spiegare più dell'80% della variabilità dei dati, evidentemente scarsa, mentre una quota ridotta dei modelli provinciali (5% circa) mostra valori R^2 prossimi a zero, suggerendo un andamento sostanzialmente costante attorno al valore medio.

Il secondo indicatore (RMSE relativo), indica quanto l'errore commesso risulta significativo rispetto alla media dei valori noti, e i risultati in questo caso sono buoni, quasi tutti i modelli presentano errori minori del 4%, con un solo modello che arriva al 12%. Questa osservazione controbilancia parzialmente i risultati non incoraggianti dell'R^2.

Nonostante la variabilità dei risultati di validazione, si è scelto di adottare comunque il metodo regressivo di imputazione dei dati mancanti, in quanto una stima, anche sommaria, della popolazione in tutte le province e in tutti gli anni, potrebbe risultare rilevante per i passaggi futuri. L'adozione di modelli più complessi avrebbe richiesto dati aggiuntivi (non disponibili), con benefici attesi forse limitati rispetto alla finalità dell'imputazione, che rimane puramente ricostruttiva. Si sottolinea inoltre che i modelli provinciali di regressione lineare sono stati utilizzati unicamente per stimare i valori mancanti o contrassegnati da flag, mentre i dati affidabili disponibili sono stati mantenuti inalterati.

A titolo esemplificativo, in Figura 4.14 si riportano due esempi di come il modello semplice predice i dati, a sinistra uno dei casi peggiori, mentre a destra uno dei migliori. Si fa presente che, in entrambi i casi mostrati, il modello di fatto non serva, in quanto sono presenti dati per tutti gli anni.

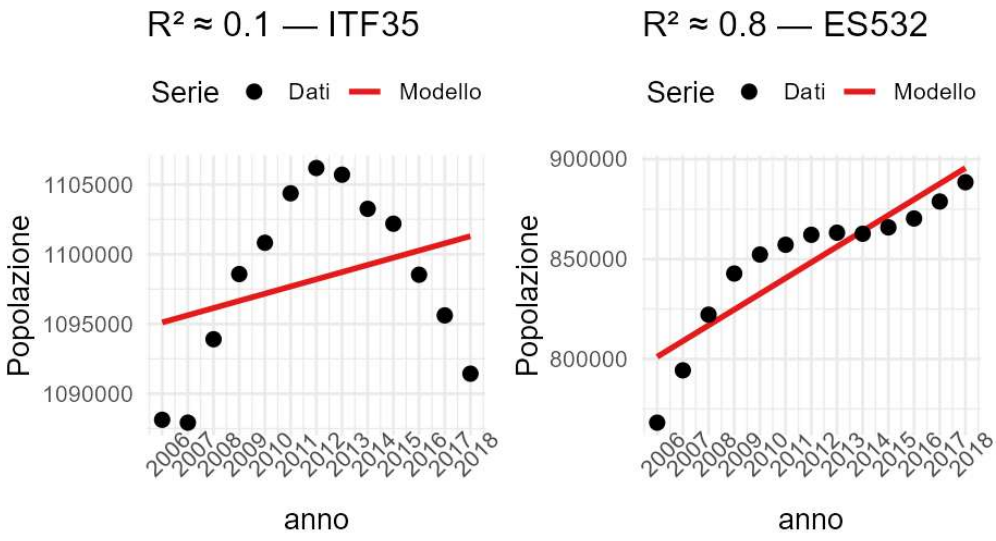

Figura 4.14: Esempio di due linearizzazioni di popolazione, una delle peggiori e una delle migliori

4.3.2. Analisi dei dati assoluti

I dati di popolazione presentano una distribuzione fortemente asimmetrica, caratterizzata da un grande numero di province con popolazione relativamente bassa, e un ridotto numero con popolazione anche molto elevata. A titolo esemplificativo si riportano, in Figura 4.15, boxplot e distribuzioni stimate per la popolazione, a scala provinciale, del 2006 e del 2018.

Nel campione, la popolazione provinciale varia da circa 15 mila abitanti nel cantone Appenzell Innerrhoden, Svizzera) fino a circa 6,5 milioni (nella provincia di Madrid, Spagna). Per fornire un ordine di grandezza complessivo, La popolazione media si aggira tra i centomila e il milione di abitanti per provincia, con un'elevata variabilità (la variazione standard del 2006 = 405 988 e del 2018 = 435 118) dovuta sia alle differenze di livello insediativo sia, soprattutto, alla forte eterogeneità di superficie delle unità NUTS3. Inoltre, le differenze nelle distribuzioni del 2006 e del 2018 risultano sostanzialmente trascurabili, a causa del ridotto arco temporale considerato; le variazioni di popolazione nel tempo sono oggetto discusse di un'analisi separata nella successiva Sezione 4.3.3.

Per tenere conto della diversa estensione territoriale, si considera la densità di popolazione (abitanti per km²), calcolata come rapporto tra popolazione e superficie provinciale. Nel 2006 la densità mediana in Europa è compresa tra 10 e 100 abitanti per km²; gli Stati del Nord Europa presentano in prevalenza densità inferiori a 10 abitanti per km², mentre le aree costiere e il Centro Europa mostrano valori tipicamente più elevati (100–1000 ab/km²). Le

principali aree metropolitane (ad es. Parigi, Milano, Londra, Berlino) superano i 1000 ab/km². La distribuzione spaziale di tale indicatore è rappresentata in Figura 4.16.

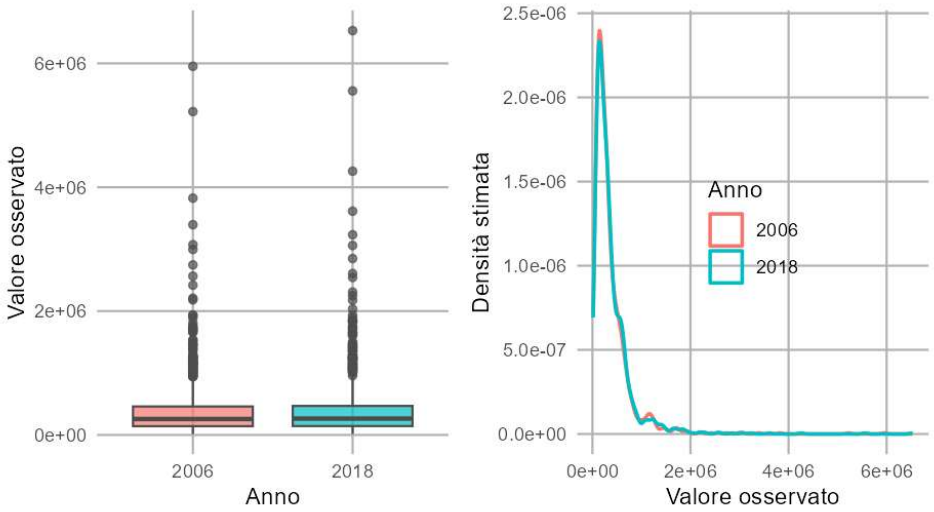

Figura 4.15: confronto tra boxplot e densità empiriche della popolazione, anni 2006 e 2018, scala naturale

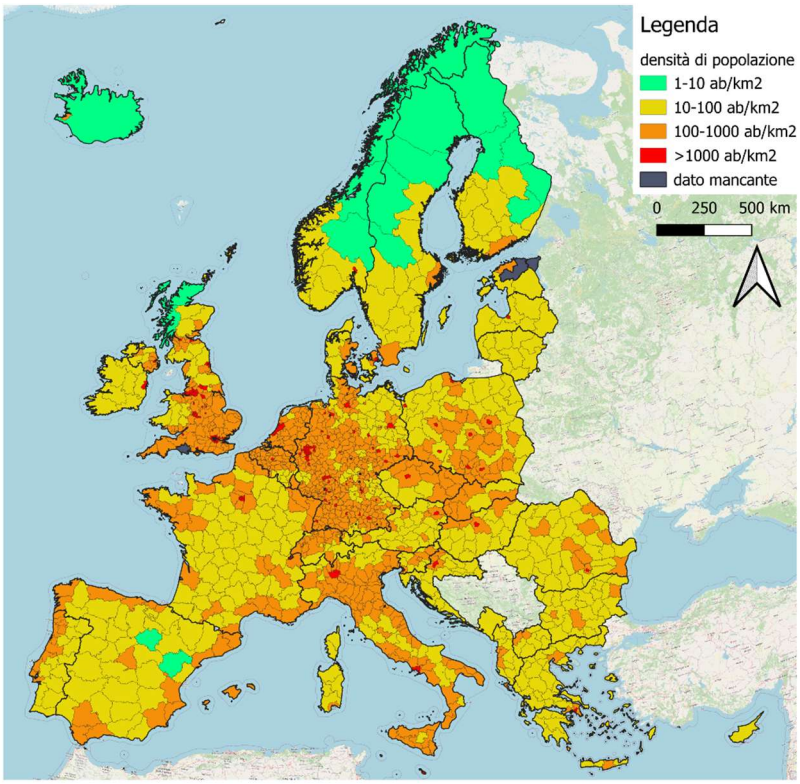

Figura 4.16: mappa della densità di popolazione del 2006

.

4.3.3. Analisi delle differenze di popolazione

Più interessante, si ritiene, è osservare l'evoluzione della popolazione fra gli anni, attraverso le differenze. Se è vero che esse si mantengono mediamente molto ridotte, esse presentano notevoli outliers, sia positivi che negativi.

Per osservare la distribuzione delle differenze si mostra, in Figura 4.17, il grafico logaritmico con segno. Tale rappresentazione è stata adottata in quanto le differenze assumono sia valori positivi che negativi, ma in valore assoluto sono sempre maggiori o uguali a 1. Valori negativi delle differenze si riferiscono esattamente a differenze negative di pari valore, su cui è stata applicata una trasformazione logaritmica analoga a quelle positive. Sull'asse delle ordinate stati riportati i valori in scala naturale per facilitarne la lettura.

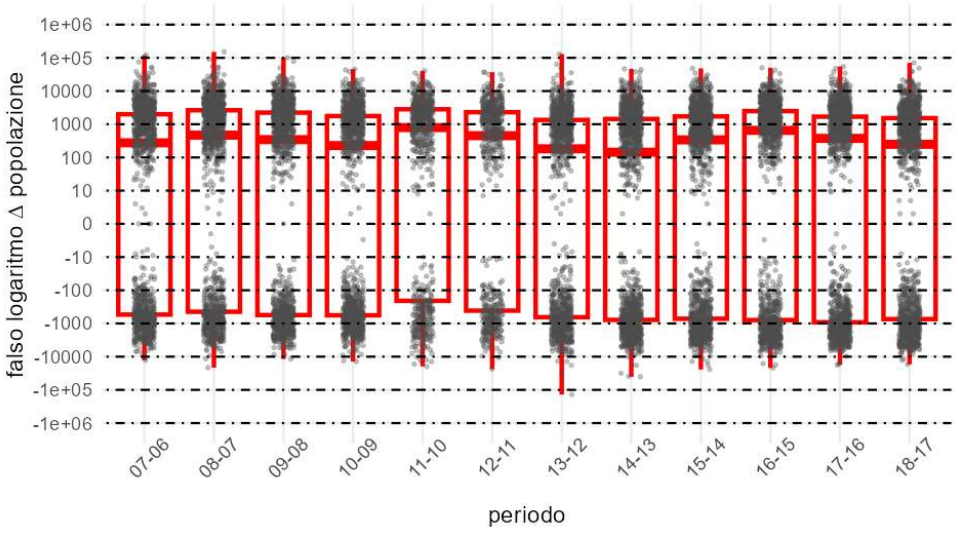

Figura 4.17: Boxplot del logaritmo con segno delle differenze di popolazione tra un anno e il precedente.

Dall'analisi emerge che la maggior parte delle province presenta variazioni annuali comprese tra poche centinaia e alcune migliaia di unità. Tuttavia, sono presenti casi estremi caratterizzati da variazioni dell'ordine delle decine di migliaia di abitanti in un singolo anno, e ci sono anche poche province che presentano piccole variazioni, in positivo o in negativo. Il solo periodo con variazioni più significative è quello tra il 2011 e il 2014, già discusso come critico nella Sezione 4.3.1. Nel complesso, la distribuzione appare approssimativamente simmetrica, pur in presenza di una lieve tendenza positiva, coerente con una crescita media della popolazione nel periodo considerato.

Per valutare l'effettivo impatto della variazione di popolazione sulla provincia, è stata calcolata la variazione relativa di popolazione, tra 2006 e 2018, rispetto a quella del 2006. La sua densità è visibile in Figura 4.18.

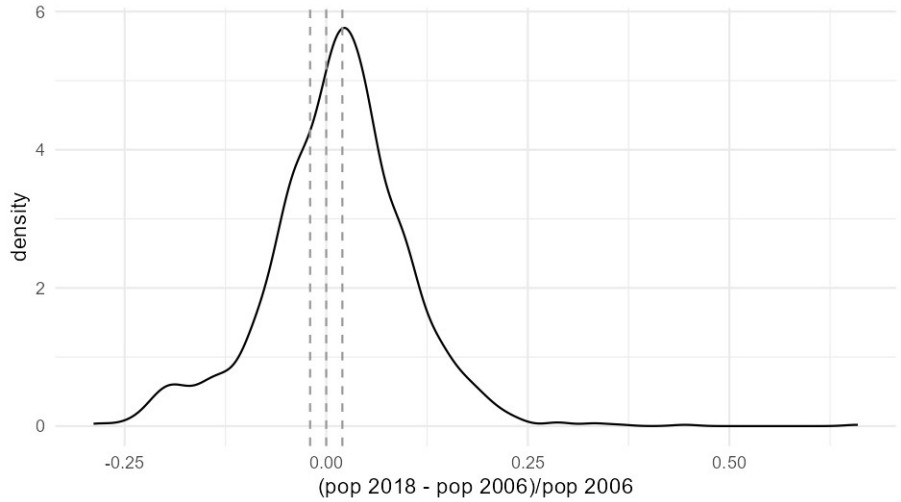

Figura 4.18: Densità stimata della variazione di popolazione nelle province europee, tra 2006 e 2018, relativa alla popolazione del 2006

Si osserva come la variazione di popolazione sia un fenomeno spazialmente molto eterogeneo in Europa. La provincia che ha sperimentato la maggiore riduzione di popolazione (che si trova in Albania), ha visto la sua popolazione ridursi del 28%, mentre quella che ha visto il maggior incremento (che si trova in Romania) la ha incrementata del 68%, ed è l'unica provincia con incrementi maggiori del 40%. I tre quarti delle province hanno visto variazioni tra -3% e +6% della loro popolazione nel 2006, con una mediana pari a +1,4% e una media di 1,7%.

Per meglio visualizzare le variazioni di popolazione, sono state rappresentate in mappa. È stata calcolata la variazione di popolazione, tra 2006 e 2018, relativa alla popolazione del 2006. La densità stimata di questa variabile è visibile in Figura 4.19.

A fini descrittivi, le province con variazione compresa tra −2% e +2% sono state considerate sostanzialmente stabili.

Dalla distribuzione spaziale emerge che le aree con maggiore contrazione demografica si concentrano prevalentemente nell'Europa orientale, nel Mezzogiorno italiano e in alcune regioni della Spagna. Al contrario, incrementi più marcati si osservano in parti della Francia, nel Nord Italia e in diverse aree del Nord Europa.

Le cause sottostanti tali dinamiche non rientrano nell'oggetto specifico del presente stu-dio; tuttavia, la marcata eterogeneità territoriale evidenziata costituisce un elemento ri-levante per le analisi successive

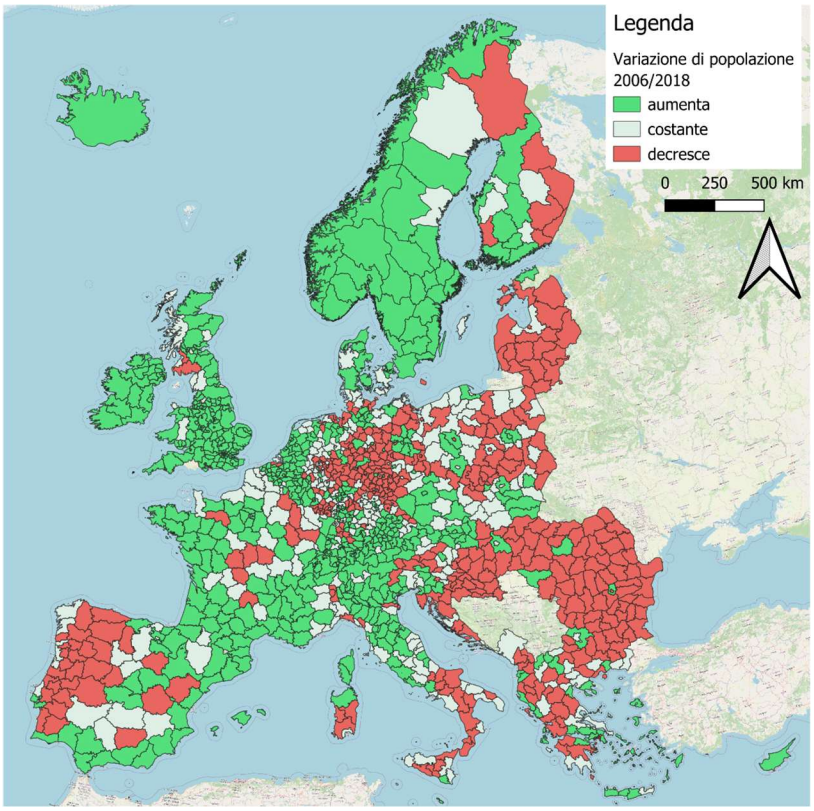

Figura 4.19: mappa della variazione di popolazione tra 2006 e 2018

4.4. Differenze tra Stati

Lo Stato a cui un'area appartiene non rappresenta una variabile esplicativa come le altre presenti nei dataset, ma sintetizza un insieme ampio di caratteristiche strutturali, alcune misurabili (a titolo esemplificativo, ma non esaustivo, le caratteristiche meteo climatiche del suo territorio), altre difficilmente quantificabili (sempre a titolo esemplificato, come l'efficacia delle politiche di prevenzione da dissesto idrogeologico e gestione del rischio). Risulta pertanto rilevante valutare in quale misura le variabili considerate siano associate allo Stato di appartenenza.

Il metodo più diretto per valutare la relazione tra i predittori numerici e la variabile Stato è l'indice $\hat{\eta}^2$, come descritto nella Sezione 3.4.2. Tale misura quantifica la quota di variabilità totale spiegata dall'appartenenza allo Stato nell'ambito di un modello ANOVA.

La Tabella 4.3riporta i valori dell'indice $\hat{\eta}^2$ per le principali variabili considerate alle diverse scale territoriali (provinciale, regionale e statale). In tutti i test ANOVA eseguiti, l'effetto Stato si è rivelato sempre come estremamente significativo (p-value sempre minore di 0,001); per tale motivo i valori puntuali dei test non sono riportati.

Tabella 4.3: $\hat{\eta}^2$ del predittore stato per gli altri predittori

Variabile	Scala provinciale	Scala regionale	Scala statale
Temperatura media	0,66	0,72	0,98
Precipitazioni totali	0,34	0,44	0,76
Giorni di pioggia	0,60	0,65	0,79
Densità di popolazione	0,08	0,11	0,99
Percentuale di suolo impermeabilizzato	0,19	0,20	0,98

L'effetto dello Stato sulle variabili di temperatura media e giorni di pioggia è molto significativo anche a scala provinciale. La densità di popolazione e gli indicatori di impermeabilizzazione presentano, viceversa, una maggiore variabilità all'interno delle province del singolo Stato.

La situazione a scala regionale (non utilizzata nelle analisi successive) risulta sistematicamente intermedia tra quella a scala provinciale e quella a scala statale. Ciò è coerente con il fatto che successive aggregazioni territoriali riducono la variabilità interna a ciascuno Stato.

A scala statale, per tutte le variabili, la variabilità è dominata dallo Stato stesso. Le differenze temporali, su un arco temporale così ridotto, risultano infatti trascurabili rispetto alle differenze strutturali tra Paesi.

La variabile che mostra la maggiore variazione di percentuale di varianza spiegata dal predittore Stato, tra scala provinciale e statale, è la densità di popolazione. Per tale motivo, nelle Figura 4.20 e Figura 4.21 sono riportati i boxplot di tale variabile rispettivamente a scala provinciale e a scala statale. Data la grande differenza fra minimo e massimo, la variabile è stata rappresentata in scala logaritmica.

A scala provinciale si osserva un'elevata variabilità della densità di popolazione tra le diverse province del medesimo Stato. Pur essendo presenti differenze nelle mediane tra Stati, quelli più grandi mostrano boxplot più estesi e presenza di outlier. Diversamente, a scala statale, si osserva unicamente la variabilità interannuale della densità di popolazione, che — dato il ridotto numero di anni considerati — risulta sostanzialmente trascurabile rispetto alle differenze tra Stati.

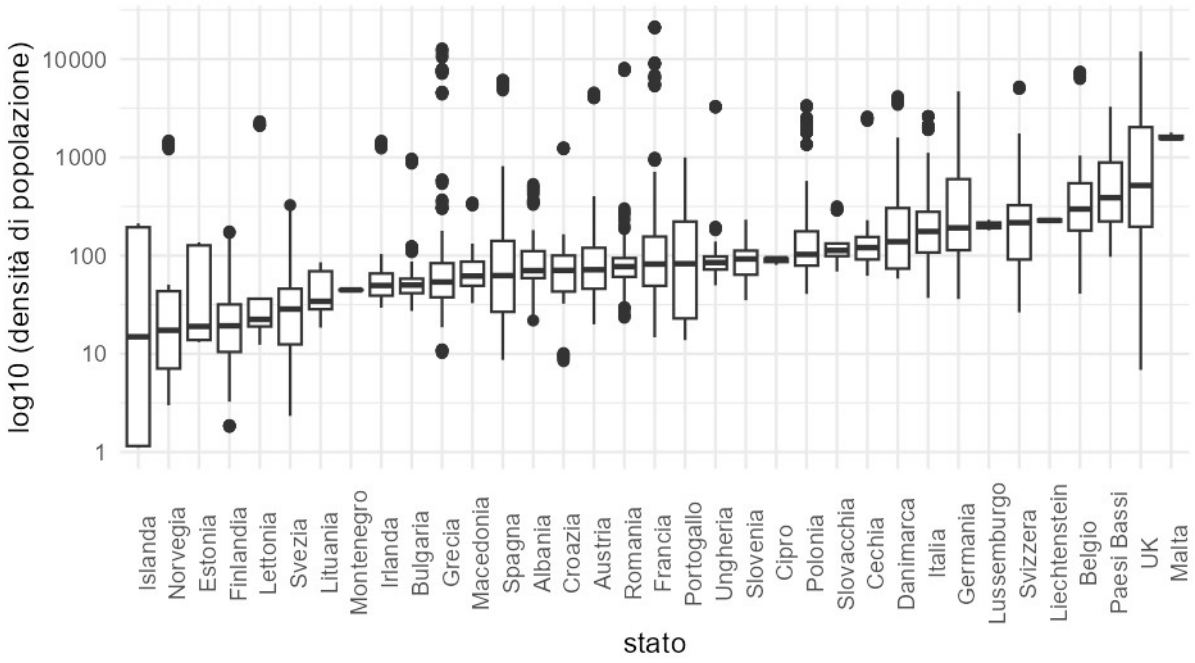

Figura 4.20: boxplot del logaritmo della densità di popolazione per Stato, a scala provinciale

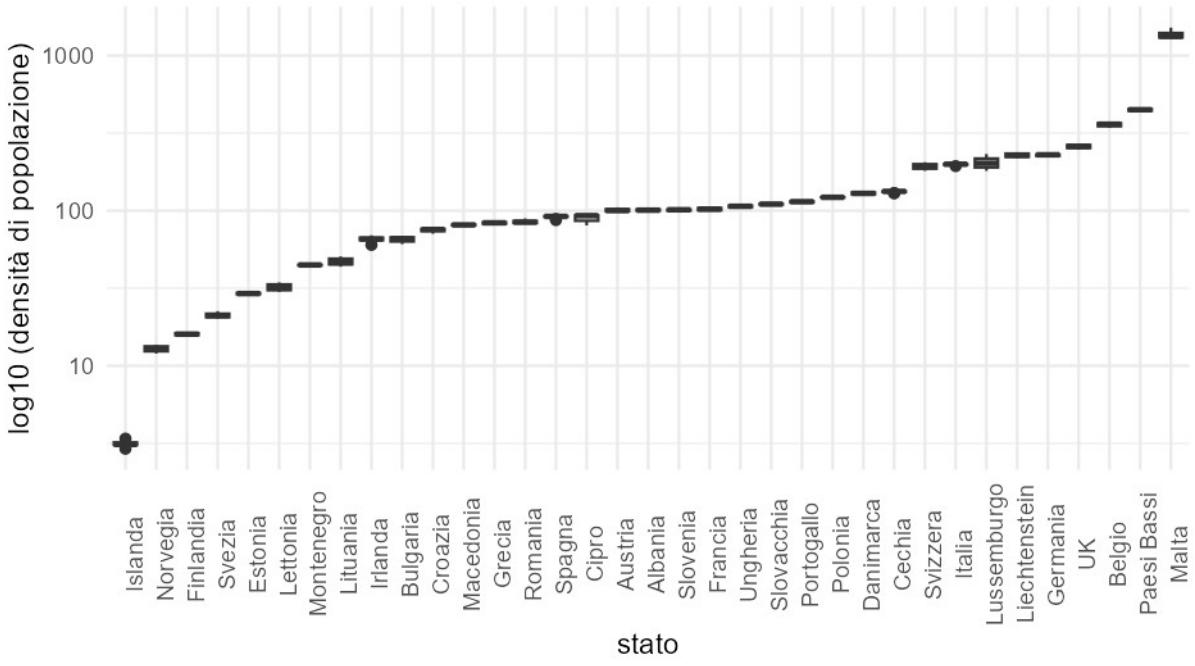

Figura 4.21: boxplot del logaritmo della densità di popolazione per Stato, scala statale

Le evidenze esplorative presentate in questo capitolo motivano l'analisi modellistica sviluppata nel Capitolo 5, nella quale si valuterà la probabilità di osservare eventi con danno controllando simultaneamente per fattori climatici e territoriali.

A dispetto della limitata variabilità delle covariate tra gli anni, la situazione degli Stati, in termini di accadimento di disastri, risulta invece fortemente eterogenea. Una sintesi grafica della situazione è presente in Figura 4.22.

Si nota che il 2011 è nettamente l'anno con minor presenza di dati, mentre il 2010 è quello con la maggiore. Ciò era già emerso dall'analisi del numero complessivo di eventi. Si nota anche che in cinque Stati (Cipro, Finlandia, Liechtenstein, Montenegro e Malta) non sono stati registrati eventi dannosi, mentre, in altri sei Stati (Lettonia, Norvegia, Svezia, Estonia, Islanda, Lussemburgo) il numero complessivo di eventi è inferiore a tre.

La carenza di dati, nello specifico di costi economici, rappresenta una criticità strutturale del dataset preso in esame, più che una caratteristica dei disastri stessi. EM-DAT, infatti, non registra mai valori nulli per gli indicatori di danno, rendendo difficile tra, eventuali disastri che non abbiano fortunatamente portato ad un certo tipo di danno, e disastri in cui questa informazione manca del tutto per carenze nelle fonti. Allo stesso modo, per quanto l'intervallo temporale sia ridotto, si ritiene improbabile che in interi Stati non sia mai accaduto alcun evento di rilievo; pertanto, l'assenza di eventi registrati in determinate aree potrebbe riflettere limiti nella documentazione o il mancato rispetto dei criteri di inclusione previsti dal database.

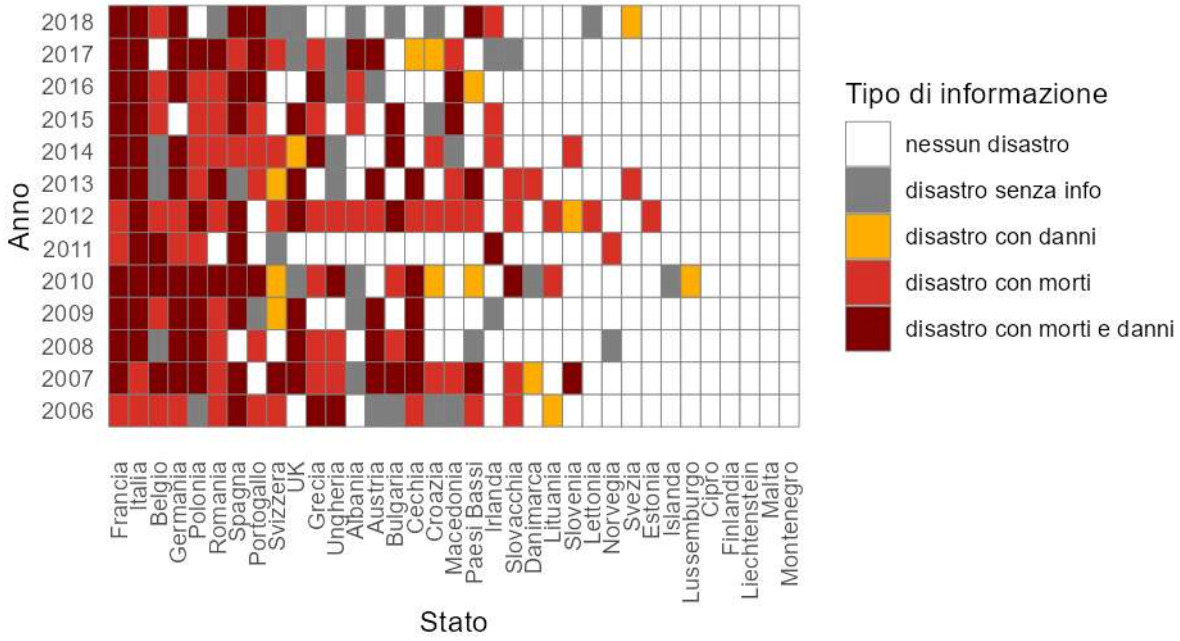

Figura 4.22: sintesi della presenza di dati di disastro, per stato e per anno

Infine, le evidenze esplorative presentate in questo capitolo motivano l'analisi modellistica del Capitolo 5, in cui si valuterà la probabilità di osservare eventi con danno controllando simultaneamente per fattori climatici e territoriali.

5 Modelli di regressione logistica

L'obiettivo di questo capitolo è valutare quantitativamente la relazione tra consumo di suolo e probabilità di occorrenza di eventi con danni, controllando simultaneamente per fattori climatici e demografici.
L'analisi è condotta su due differenti scale spaziali:

- scala nazionale, con unità di osservazione Stato–anno;
- scala provinciale, con unità di osservazione Provincia–anno.

In entrambi i casi, la variabile risposta è binaria e indica la presenza di almeno un evento con danni registrati nell'unità spaziale e temporale considerata.

5.1. Modello a scala nazionale

L'analisi prende avvio dalla scala nazionale, assumendo come unità di osservazione la coppia Stato–anno. I valori osservati non evidenziano criticità di multicollinearità tali da impedire l'inclusione simultanea delle covariate nel modello.

Si specifica inizialmente un modello completo, comprensivo di variabili territoriali e climatiche, che viene poi semplificato e valutato in termini di performance al fine di ottenere una formulazione più parsimoniosa.

5.1.1. Variabili utilizzate

I primi modelli stimati a scala nazionale sono modelli di regressione logistica, finalizzati a valutare la rilevanza di vari predittori nel determinare il verificarsi di eventi disastrosi.

Sono stati stimati modelli relativi alla presenza di disastri totali, e ai soli disastri meteorologici o idrologici, in quanto le ultime due categorie di eventi sono le più osservate nel periodo considerato.

La variabile risposta è binaria e indica la presenza o assenza di almeno un evento nell'unità Stato–anno. La scelta di una variabile dicotomica è motivata dall'elevata asimmetria e incompletezza degli indicatori di danno, che rendono poco robusta una modellazione diretta dell'intensità del danno.

I predittori inizialmente considerati sono:

- anno, per intercettare eventuali tendenze temporali;

- superficie dello Stato, come proxy dell'esposizione territoriale al rischio. Come visibile in Figura 5.1 vi sono enormi differenze tra minimo e massimo della superficie degli Stati. Per questo motivo ne è stato considerato il logaritmo.
- densità di popolazione, quale indicatore di vulnerabilità antropica, coerentemente con la definizione di disastro adottata nell'Introduzione; anche in questo caso, per l'elevata variabilità della densità di popolazione tra Stati, da 2 ab/km² (in Islanda) a 1500 ab/km² (a Malta), come si vede in Figura 4.21, di essa è stato considerato il logaritmo.
- variabili climatiche, in particolare giorni di pioggia, precipitazioni massime e temperatura media, per verificare il legame tra verificarsi di un evento e condizioni ambientali.

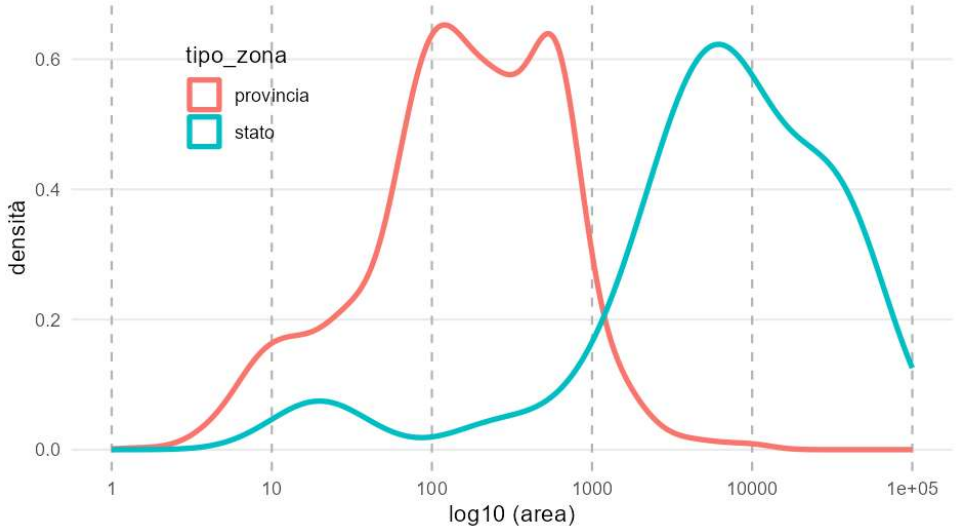

Figura 5.1: densità stimata del logaritmo della superficie di Stati e province

La percentuale di suolo impermeabilizzato non è stata inclusa nel modello a scala statale, poiché si è rivelata fortemente correlata con la densità di popolazione (correlazione superiore al 90%), rendendo difficile distinguere i due effetti.

Il test di collinearità delle covariate ha prodotto, per tutti i predittori, valori dell'indice VIF ampiamente inferiori alla soglia critica di 5 (il valore massimo 2,8 è stato osservato per la temperatura media). Da ciò consegue che non vi siano basi per accettare una collinearità tra le variabili, che dunque esse siano differenti e possano, in principio, essere utilizzate tutte.

Tutti i predittori numerici sono stati standardizzati, ossia è stata sottratta la media campionaria e sono stati divisi per la deviazione standard, allo scopo di rendere confrontabili i coefficienti stimati, anche se applicati a variabili con scale differenti.

5.1.2. Stima del modello completo

Il primo modello stimato a scala nazionale è un modello di regressione logistica della forma:

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 Anno_t + \beta_2 \log(\text{Area}_i) + \beta_3 \log(\text{dens pop}_{it}) + \beta_4 Temp_{it} + \beta_5 PrecMax_{it} + \beta_{26} GiorniPioggia_{it}$$

dove p_{it} rappresenta la probabilità che nello Stato i, nell'anno t, si verifichi almeno un evento con danni.

Innanzitutto, è stata eseguita una valutazione del tipo di modello, mediante il metodo Leave-One-Out (LOO) descritta nella sezione 3.5. I risultati di tale valutazione sono riassunti nella Tabella 5.1, che riporta, per ciascuna tipologia di evento considerata, la percentuale di osservazioni con evento, la soglia di classificazione ottimale, l'indice ROC (AUC) e la relativa matrice di confusione.

Tabella 5.1: sintesi valutazione dei modelli logit sul verificarsi degli eventi, scala stato

Modello	Eventi totali			Eventi meteorologici			Eventi idrologici		
% eventi / totale	48%			34%			26%		
Soglia ottimale	49%			33%			23%		
Indice AUC	0,85			0,81			0,82		
Matrice di confusione soglia ottimale		Non avv.	Avve-nuto		Non avv.	Avve-nuto		Non avv.	Avve-nuto
	Non pred.	191	50	Non pred.	235	37	Non pred.	247	25
	Pre-detto	44	170	Pre-detto	66	117	Pre-detto	88	95

Tutti i modelli presentano capacità predittive buone, con valori dell'indice AUC compresi tra 0,81 e 0,85. Come atteso, data la rarità degli eventi (evidenziata dalla percentuale di osservazioni con evento riportata nella prima riga della tabella), le soglie di classificazione ottimali risultano minori di 0,5, per le sottocategorie di disastri. Fa eccezione il modello relativo agli eventi totali, per i quali la frequenza osservata è maggiore e la soglia ottimale si avvicina a 0,5. Le matrici di confusione mostrano un equilibrio soddisfacente tra sensibilità e specificità, pur evidenziando — coerentemente con la struttura dei dati — una maggiore difficoltà nella corretta classificazione degli eventi rari.

La Tabella 5.2 riporta invece le stime dei coefficienti di regressione del modello completo, le rispettive deviazioni standard e il p-value del likelihood ratio test, per ciascun predittore. Seguendo le convenzioni del software R L'indicazione è stata data scrivendo, dopo la deviazione stimata:

- Niente per p-value superiori a 10%
- . per p-value compresi tra 5% e 10%
- * per indicare p-value compresi tra 5% e 1%
- ** per indicare p-value compresi tra 1% e 0,1%
- *** per indicare p-value minori di 0,1%.

Inoltre, per facilitarne la lettura, è stata utilizzata una scala di colori nella tabella:

- Verde chiaro per effetti di riduzione della probabilità minori di $-0{,}2$
- Arancione per effetti di incremento della probabilità maggiori di 0,2.
- Grigio chiaro per effetti trascurabili (compresi tra $-0{,}2$ e 0,2).

Tabella 5.2: tabella riassuntiva dei modelli di presenza disastri, scala annuale. Per ciascun predittore sono riportati il coefficiente stimato β e, tra parentesi, la corrispondente deviazione standard. Il segno del coefficiente indica la direzione dell'associazione con la probabilità di occorrenza dell'evento, mentre la sua entità misura la variazione dei log-odds associata a un incremento unitario del predittore.

Predittore	modello: eventi totali	modello: eventi meteorologici	modello: eventi idrologici
Anno	−0,18 (0,11)	−0,35 (0,11) *	−0,14 (0,12)
Temperatura media	0,42 (0,19) ***	-0,23 (0,12) .	0,75 (0,22) ***
Prec. Massime	0,19 (0,11)	0,11 (0,12)	0,53 (0,12) .
Giorni di pioggia	0,05 (0,16) *	0,01 (0,17) **	0,05 (0,18)
Log_{10} (area Stato)	2,11 (0,23) ***	1,48 (0,11) ***	1,79 (0,21) ***
Log_{10} (Densità di popolazione)	1,38 (0,22)***	1,54 (0,20) ***	0,50 (0,22) *

Dall'analisi dei coefficienti emergono alcune evidenze rilevanti.

Non si rilevano radicali differenze tra le predizioni degli eventi totali (che includono eventi idrologici, meteorologici e, in minor quantità, eventi climatologici) e quelle relative alle singole sottocategorie, che sono più rare. Per questo motivo, nel proseguo del capitolo verranno considerati unicamente gli eventi totali.

Il modello relativo agli eventi totali evidenzia una lieve riduzione del numero di disastri nel tempo, tuttavia non statisticamente significativa. Allo stesso modo, probabilmente a causa della scala spaziale ampia alla quale si sta modellizzando il fenomeno, anche il numero di giorni di pioggia e le precipitazioni massime non presentano effetti significativi. Risultano invece significativi gli incrementi associati:

- all'aumento della temperatura media;
- all'aumento della superficie dello Stato considerato;
- all'aumento della densità di popolazione.

Tali effetti appaiono coerenti con la natura del fenomeno analizzato, in quanto:

- un incremento della temperatura è correlato a una maggiore probabilità di verificarsi di eventi dannosi tipo incendi boschivi, siccità e ondate di calore;
- un incremento della superficie della zona interessata comporta una maggiore probabilità che si verifichino eventi estremi all'interno della stessa;
- un incremento della densità di popolazione comporta una crescita della vulnerabilità antropica della zona considerata, e di conseguenza, una maggiore probabilità che condizioni naturali avverse si traducano in un evento dannoso rilevabile.

5.1.3. Selezione dei predittori e modello finale

Poiché nel modello completo alcune variabili risultavano poco significative, è stata applicata una procedura di selezione tramite l'algoritmo di backward selection descritto nella Sezione 3.5.4.

Le variabili eliminate sono state, nell'ordine: precipitazioni massime, anno, e numero di giorni di pioggia. Tale procedura ha comportato una lieve riduzione dell'indice AIC (da 438 a 436), accompagnato da una sostanziale semplificazione del modello. Il modello finale è presentato in Tabella 5.3.

Tabella 5.3: presentazione del modello finale a scala statale

Predittore	Stima dei beta	SE	p-value L.R.T.
Temperatura media	0,30	0,15	$3*10^{-4}$
Log_{10} (Area Stato)	2,02	0,21	$4*10^{-27}$
Log_{10} (densità di popolazione	1,47	0,20	$1*10^{-17}$

In generale si osserva che le variabili climatiche, inclusa la temperatura media (rimasta nel modello), risultano poco significative, o con effetti marginali rispetto alle variabili territoriali. Questo risultato può essere ricondotto alla scala spaziale e temporale dell'analisi, molto vasta. Un disastro naturale è, infatti, un fenomeno locale, e tentare di caratterizzarlo a scala statale e annuale può risultare complesso. Va inoltre ricordato che gli eventi sono registrati in funzione del danno prodotto e non esclusivamente dell'intensità del fenomeno climatico che li ha generati. Per questo, fattori di tipo antropico risultano molto più rilevanti rispetto a fattori di tipo climatico.

Gli effetti, tuttavia, risultano coerenti con l'interpretazione sostanziale del fenomeno: un incremento della temperatura media, della superficie dello Stato e della densità di popolazione è associato a un aumento della probabilità di verificarsi di un evento dannoso. In particolare, l'effetto associato alla superficie risulta più marcato rispetto a quello della densità di popolazione, mentre l'effetto della temperatura appare più contenuto.

Il modello è stato valutato mediante la stessa procedura LOO come i precedenti. Esso presenta una soglia ottimale di 0,5, un indice AUC di 0,85 e la matrice di confusione mostrata in Tabella 5.4.

Tabella 5.4: matrice di confusione del modello finale a scala statale

	Disastro non avvenuto	Disastro avvenuto
Disastro non predetto	193	54
Disastro predetto	42	166

Nel complesso, il modello presenta prestazioni molto simili a quelle del modello completo, cioè calcolato sfruttando la totalità dei predittori, indicando che la semplificazione non comporta una perdita sostanziale di capacità predittiva.

Infine, dal momento che sulle 220 unità statistiche "Stato-anno" in cui si è osservato almeno un evento, in 169 casi si sono registrate vittime, è stato stimato anche un modello per la probabilità di un disastro con decessi. Tale modello risulta molto simile, in termini di variabili selezionate, coefficienti e performance, a quello qui mostrato, e per tale motivo non viene approfondito.

5.2. Modelli a scala provinciale

Si procede ora alla stima dei modelli a scala provinciale, assumendo come unità di osservazione la coppia Provincia–anno. Il passaggio a un livello territoriale più fine consente di verificare se le relazioni osservate a scala nazionale si mantengano o si modifichino al variare della granularità spaziale.

5.2.1. Variabili utilizzate

Le variabili considerate sono analoghe a quelle utilizzate a scala statale, con l'adattamento all'unità territoriale provinciale. L'unità di osservazione è la coppia Provincia–anno.

Oltre alle variabili già introdotte (anno, superficie, densità di popolazione, indicatori climatici), a scala provinciale risulta un predittore significativo anche la percentuale di suolo impermeabilizzato. Diversamente da quanto osservato a scala statale, la correlazione tra densità di popolazione e impermeabilizzazione risulta meno marcata, rendendo possibile la coesistenza delle due variabili nel modello.

Anche in questo caso, per le variabili con ampia variabilità tra unità territoriali è stato considerato il logaritmo, al fine di ridurne l'asimmetria e stabilizzare le stime.

Come per la scala nazionale, le variabili quantitative sono state standardizzate per consentire una migliore confrontabilità dei coefficienti stimati.

5.2.2. Modello senza il predittore Stato con consumo di suolo linearizzato

Sono stati stimati due modelli a scala provinciale: uno che non include lo Stato tra i predittori e uno che lo include come variabile categoriale.

È ipotizzabile che il modello che includa lo Stato tra i predittori risulti più accurato rispetto a quello che non lo include, in quanto in grado di intercettare differenze strutturali tra Paesi. Tuttavia, è anche plausibile che esso risulti complessivamente meno accurato rispetto al modello stimato a scala statale, a causa della maggiore numerosità delle osservazioni e della conseguente maggiore variabilità dei dati.

Per questo motivo vengono presentate entrambe le specificazioni, al fine di valutare come l'inclusione dello Stato modifichi l'importanza e la significatività delle altre variabili.

Si presenta innanzitutto il modello senza Stato. Come per il modello a scala nazionale, i risultati della cross-validazione Leave-One-Out sono riportati in Tabella 5.5.

Tabella 5.5: tabella delle prestazioni LOO del modello senza Stato a scala provinciale

Modello	Eventi totali			Eventi meteorologici			Eventi idrologici		
% eventi / totale	45%			34%			14%		
Soglia ottimale	43%			32%			12%		
Indice AUC	0,65			0,67			0,63		
Matrice di confusione soglia ottimale		Non avv.	Avve-nuto		Non avv.	Avve-nuto		Non avv.	Avve-nuto
	Non pred.	5816	2660	Non pred.	5798	1406	Non pred.	9687	1125
	Pre-detto	4022	5325	Pre-detto	5948	4671	Pre-detto	5608	1403

Rispetto alla scala nazionale, questi modelli sono meno capaci di discriminare casi in cui un disastro sia accaduto da casi in cui non lo sia. A differenza dei modelli a scala di stato, tendono più spesso ad attribuire come avvenuto un disastro non realmente verificatosi. Questa loro caratteristica può essere spiegata dal fatto che, affinché un evento dannoso rientri in EM-DAT, e dunque sia conosciuto a questo report, è necessario che il danno prodotto superi determinate soglie. È possibile, dunque, supporre che siano accaduti più eventi di portata minore, non di taglia sufficiente per rientrare all'interno del dataset, ma comunque in grado di generare effetti più contenuti. Tuttavia, in assenza di dati specifici a riguardo, tale ipotesi rimane non verificabile.

Dal momento che i modelli dimostrano una capacità discriminante (l'indice AUC risulta sempre superiore a 0,5), è opportuno analizzare le stime dei coefficienti di regressione visibili in Tabella 5.6.

Tabella 5.6: coefficienti del modello a scala provinciale senza Stato. Concettualmente simile alla Tabella 5.2

Predittore	modello: eventi totali	modello: eventi meteorologici	modello: eventi idrologici
Anno	-0,441 (0,016)***	-0,531 (0,017)***	-0,148 (0,022)***
Temperatura media	0,088 (0,021)***	-0,108 (0,023)***	0,195 (0,029)***
Prec. massime	0,220 (0,016)***	0,115 (0,017)***	0,262 (0,019)***
Giorni di pioggia	0,137 (0,021)***	0,127 (0,022)***	-0,017 (0,029)**
Percentuale di suolo impermeabilizzato.	-0,547 (0,034)***	-0,635 (0,037)***	-0,163 (0,051)***
$\log_{10}$ (superficie)	0,212 (0,028)***	0,139 (0,030)	0,222 (0,039)***
Log_{10} (dens pop.)	0,628 (0,038)***	0,743 (0,041)***	0,141 (0,052)

Si osserva che quasi tutte le variabili, con due sole eccezioni, risultano molto significative per tutti i modelli considerati, e nella maggior parte dei casi i loro contributi sono coerenti tra le differenti tipologie di evento.

Il modello evidenzia una tendenza alla riduzione della probabilità del verificarsi di eventi disastrosi nel tempo. Tuttavia, la maggior parte della letteratura scientifica e il senso comune concordano sulla tendenza opposta. A titolo esemplificativo si cita l'European Climatic Risk Assessment:
" *L'Europa è il continente che si sta riscaldando più rapidamente al mondo. Il caldo estremo, un tempo relativamente raro, sta diventando più frequente, mentre i modelli di precipitazione stanno cambiando. I rovesci intensi e altri eventi di precipitazione estrema stanno aumentando in gravità, e negli ultimi anni si sono verificate in diverse regioni inondazioni catastrofiche*." (European Environment Agency, 2024)

Di conseguenza, è plausibile che l'effetto di riduzione del numero di eventi disastrosi nel tempo sia legato alla qualità e alla tipologia di dati utilizzati.

La temperatura media pare avere un effetto generalmente debole, ma peggiorativo (positivo), sulla probabilità di accadimento di eventi dannosi. Il risultato è coerente sia con la natura media dell'indicatore, che non cattura la natura locale del disastro che tenta di prevedere, che con quella di una serie di eventi dannosi (quali incendi e siccità) che sono legati ad un incremento delle temperature.

Allo stesso modo le precipitazioni massime presentano un effetto positivo, ma di entità leggermente superiore a quello delle temperature, coerentemente sia con la scala del modello, che con la predominanza delle precipitazioni nei disastri idrologici e meteorologici, che costituiscono la maggior parte degli eventi presenti nel dataset. Infine, anche l'effetto del numero di giorni di pioggia risulta coerente con quanto appena affermato, anche se è generalmente di piccola entità, Questo può essere motivato sia dalla grande quantità di fattori, locali e contingenti, che trasformano un evento atmosferico in un evento dannoso, che dalla rilevanza, in questo processo, più dell'intensità puntuale delle precipitazioni che del numero complessivo di giorni piovosi.

L'effetto positivo della superficie provinciale e della densità di popolazione sul rischio risulta coerente con quanto osservato alla scala di Stato. Come ci si poteva attendere, un aumento dell'area della provincia è associato a una maggiore probabilità che si verifichi almeno un evento di disastro, mentre un incremento della densità di popolazione rende più probabile che un evento naturale su una determinate area provochi danni rilevabili.

Infine, molto significativa è l'osservazione di una correlazione negativa tra percentuale di suolo impermeabilizzato e probabilità di verificarsi di danni ambientali. Tale risultato non deve essere interpretato come un qualche effetto protettivo dell'impermeabilizzazione del suolo, ma può riflettere la tendenza maggioritaria ad impermeabilizzato suoli in località meno soggette ad eventi dannosi. Va inoltre ricordato che gli effetti stimati delle singole variabili si intendono "a parità di condizioni". Una porzione rilevante dell'influenza del consumo di suolo è probabilmente stata assorbita dalla densità di popolazione e dalla

superficie, introducendo verosimilmente come risultato un effetto protettivo non realmente esistente.

In questo caso, la scelta di un modello più semplificato con meno covariate non è giustificata, dal momento che ogni variabile risulta molto significativa e la capacità discriminante del modello completo non risulta particolarmente elevata. Inoltre, vista la maggior variabilità spaziale, la correlazione tra i vari predittori risulta sempre modesta. Il test di multicollinearità restituisce valori pari a 5,4 per il logaritmo della densità di popolazione e 3,7 per la percentuale di suolo impermeabilizzato (che dunque leggermente rimangono correlate), mentre per le altre variabili i valori risultano minori di 2. Sebbene un risultato superiore a 5 in questo test sia generalmente considerato critico, sia l'importanza della variabile evidenziata dal p-value del test del rapporto di verosimiglianza, che l'entità del suo coefficiente stimato, significativamente superiore alla deviazione standard stimata, suggeriscono che essa risulti comunque informativa.

La procedura di selezione backward delle variabili, basata sul confronto degli indici AIC, non suggerisce l'eliminazione di alcun predittore, poiché la rimozione di ciascuna variabile tenderebbe a ridurre la verosimiglianza del modello senza apportare miglioramenti significativi in termini di parsimonia.

5.2.3. Modello senza predittore Stato con dinamica del consumo di suolo

Per poter meglio analizzare l'impatto della variazione di consumo di suolo, non ben visibile considerando unicamente l'impermeabilizzazione di suolo linearizzato, viste le piccole variazioni, è stato stimato un ulteriore modello logistico che presenta le medesime variabili climatiche (numero di giorni di pioggia, precipitazioni massime annue, temperatura media), l'area, in forma logaritmica, e due variabili per il consumo di suolo: percentuale di suolo impermeabilizzato nel 2006 e differenza di percentuale impermeabilizzata tra 2006 e 2018. Per poter considerare la popolazione, senza introdurre ulteriore collinearità tra essa e la percentuale di suolo impermeabilizzato, la variabile logaritmo della densità di popolazione precedentemente utilizzata è stata sostituita con il semplice logaritmo della popolazione. Vista la scarsa differenza fra la prevedibilità dei vari tipi di disastri, anche in questo caso è stato unicamente calcolato e validato il modello sugli eventi totali. Il modello, sottoposto a cross validazione con il metodo LOO come i precedenti, presenta una soglia ottimale di 0,40, un indice AUC di 0,66. La matrice di confusione, corrispondente alla soglia ottimale, sui dati di validazione secondo il metodo LOO, è riportata in Tabella 5.7. Anche in questo caso il modello fatica a interpretare correttamente i casi in cui un disastro non si sia verificato, mentre interpreta molto meglio i casi in cui esso effettivamente si verifica.

Tabella 5.7: matrice di confusione modello 4

	Evento non avvenuto	Evento avvenuto
Evento non predetto	5106	2185
Evento predetto	4732	5800

I risultati della stima dei parametri sono riportati in Tabella 5.8.

Tabella 5.8: parametri del modello con percentuale di suolo impermeabilizzato nel solo 2006, e delta 2006/2018

Predittore	Beta	SE	P value ANOVA
Anno	-0,44	0,016	$2*10^{-100}$
Temperatura media	0,16	0,020	$8*10^{-8}$
Giorni di pioggia	0,23	0,016	$9*10^{-62}$
Precipitazioni massime	0,19	0,020	$6*10^{-15}$
Log_{10} (popolazione)	0,16	0,016	$2*10^{-19}$
% di suolo imp. nel 2006	-0,31	0,022	$1*10^{-32}$
Δ % di suolo imp. 2006/2018	0,17	0,021	$8*10^{-18}$

Tutti i predittori risultano presentare un effetto significativamente diverso da 0, con deviazione standard stimata ben minore dell'effetto stesso. La rimozione di uno qualsiasi dei predittori ridurrebbe le capacità discriminanti del modello, come testimoniato sia dai p-value (tutti notevolmente ridotti), sia da un test di selezione variabili. Non si rilevano collinearità.

L'effetto di riduzione legato all'anno risulta coerente con il modello precedente. Dal momento che si presenta in maniera netta a scala provinciale, ma non è stato rilevato in maniera parimenti chiara a scala statale (con riferimento Tabella 5.2) la sua giustificazione deve essere un progressivo miglioramento della localizzazione dei dati, ossia una tendenza dei disastri degli anni più recenti ad essere localizzati in un numero minore di province.

L'influenza delle variabili climatologiche risulta coerente con quanto osservato per i modelli precedenti, anzi la temperatura media risulta avere un effetto significativamente superiore.

La popolazione presenta un effetto di incremento della probabilità del verificarsi di eventi estremi, coerentemente con l'incremento della vulnerabilità che la popolazione implica. Inoltre, realisticamente popolazione e percentuale di suolo impermeabilizzato assumono anche una componente dell'effetto originario dell'area, rimossa in questo modello in quanto irrilevante.

La percentuale di suolo impermeabilizzato nel 2006 presenta un effetto di riduzione della probabilità di accadimento di disastri naturali, sebbene meno pronunciato di quanto non avvenisse nel modello precedente. Questo fenomeno è dovuto al fatto che nel dataset di analisi le province di maggiori dimensioni sono quelle che correlano maggiormente con la popolazione e con la frequenza dei disastri, ma sono anche quelle che presentano le minori percentuali di suolo impermeabilizzato. Diversamente, l'incremento della percentuale di suolo impermeabilizzato è correlato ad un incremento di rischio, significativo quanto lo risultano gli incrementi di popolazione e di temperatura media.

Il odd ratio negativo della percentuale di suolo impermeabilizzato nel 2006 sul rischio disastri, risulta coerente con quanto si osserva nelle mappe in Figura 5.2 e in Figura 4.10, relative rispettivamente al numero di disastri registrato per provincia e all'impermeabilizzazione di suolo nel 2006 per provincia. Risulta da queste mappe che la maggioranza delle aree colpite da un numero significativo di eventi (tutta l'Italia, e gran parte dell'Europa dell'est, per esempio) corrispondono ad aree dal consumo di suolo relativamente ridotto (province classificate in classe 1 nella mappa del consumo di suolo), mentre aree come Londra e molte aree tedesche, ad elevato consumo di suolo, registrano un numero molto minore di eventi.

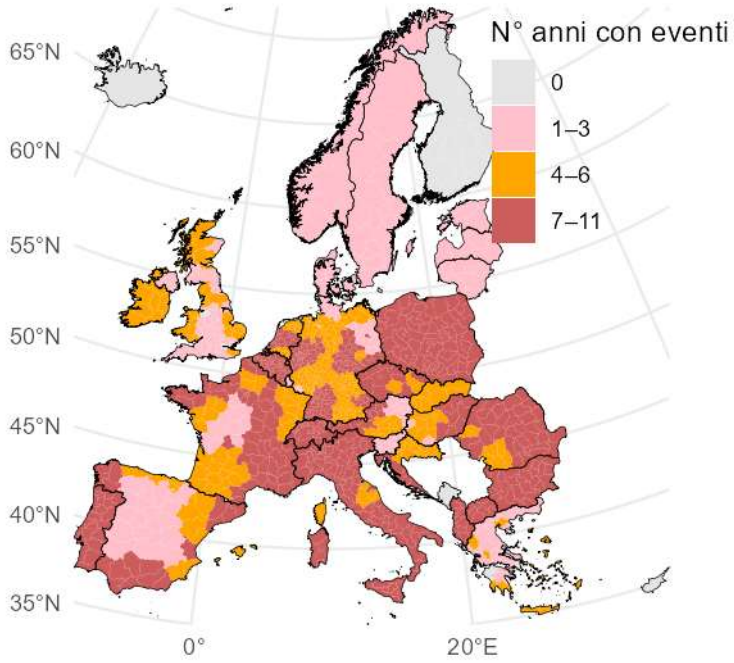

Figura 5.2: mappa che mostra il numero di anni con disastri per ogni provincia

5.2.4. Modello con il predittore Stato

Come precedentemente accennato, vista la presenza di una variabile informativa come lo Stato che, in quanto fattoriale, può racchiudere al suo interno un gran numero di informazioni differenti, è stato stimato un modello logistico sui medesimi dati, che includesse tale informazione.

Visto che non si erano riscontrate variazioni significative delle prestazioni tra modelli relativi a disastri totali, meteorologici e idrologici, è stato stimato unicamente il modello più generale relativo agli eventi totali.

Il modello presenta una soglia di classificazione ottimale del 44%, in linea con la sostanziale parità tra eventi avvenuti e non avvenuti (il 45% delle righe del dataset presenta almeno un evento). L'indice AUC risulta significativamente superiore al precedente 0,65 per il modello senza Stato, attestandosi a 0,76. La matrice di confusione, calcolata sui dati di test ottenuti mediante procedura LOO, è riportata nella Tabella 5.9. Anche in questo caso, la validazione è avvenuta con il metodo LOO.

Il modello mostra una migliore capacità di identificare i casi in cui un disastro non avviene, anche se rimane una tendenza all'errore tutt'altro che trascurabile. Anche in questo caso risulta più frequente la classificazione come "evento avvenuto'' di casi in cui l'evento non è stato registrato, rispetto all'errore opposto.

Tabella 5.9: matrice di confusione del modello provinciale con Stato

	Evento non avvenuto	Evento avvenuto
Evento non predetto	6607	2139
Evento predetto	3143	5843

I coefficienti stimati del modello vengono presentati in Tabella 5.10, con esclusione dei coefficienti associati ai singoli Stati.

Tabella 5.10: coefficienti numerici del modello provinciale con Stato

Predittore	Stima dei beta	SE	p-value L.R.T.
Stato	Non univoco	Non univoco	0
Anno	-0,47	0,01	0
Temperatura media	-0,03	0,03	$3*10^{-8}$
Prec. massime	0,13	0,02	$1*10^{-63}$
Giorni di pioggia	0,43	0,03	$1*10^{-14}$
Log_{10} (area provincia)	0,15	0,04	$2*10^{-10}$
Log_{10} (densità di popolazione)	0,28	0,04	$6*10^{-13}$
Percentuale di suolo impermeabilizzato	-0,13	0,04	$2*10^{-64}$

Gli effetti delle variabili quantitative risultano tutti coerenti, ma generalmente di entità più contenuta, rispetto a quanto osservato nel modello precedente. Probabilmente ciò è dovuto all'inserimento della variabile Stato, che ha assorbito una parte consistente dell'effetto delle singole variabili.

Le sole variabili continue con effetti ancora significativamente differenti da 0 sono l'anno e la percentuale di suolo impermeabilizzato, con effetto migliorativo (segno negativo), mentre giorni di pioggia, precipitazioni massime e superficie provinciale, con effetto peggiorativo (segno positivo).

L'effetto della temperatura media, molto significativa a scala statale, è qui sostanzialmente sparito, probabilmente assorbito con maggior efficacia dalla variabile Stato. Infatti, come già osservato nella Tabella 4.3, la temperatura media è la variabile maggiormente spiegata dallo Stato a entrambe le scale considerate.

Le variabili risultano tutte molto significative, e la rimozione di una qualsiasi di esse provocherebbe una significativa riduzione della verosimiglianza del modello, come indicato dai p-value mostrati.

I coefficienti stimati per ciascuno Stato, rispetto allo stato di riferimento Italia, sono rappresentati in Figura 5.3, insieme all'incertezza stimata di ciascuno di essi, in termini di intervalli di confidenza al 95%.

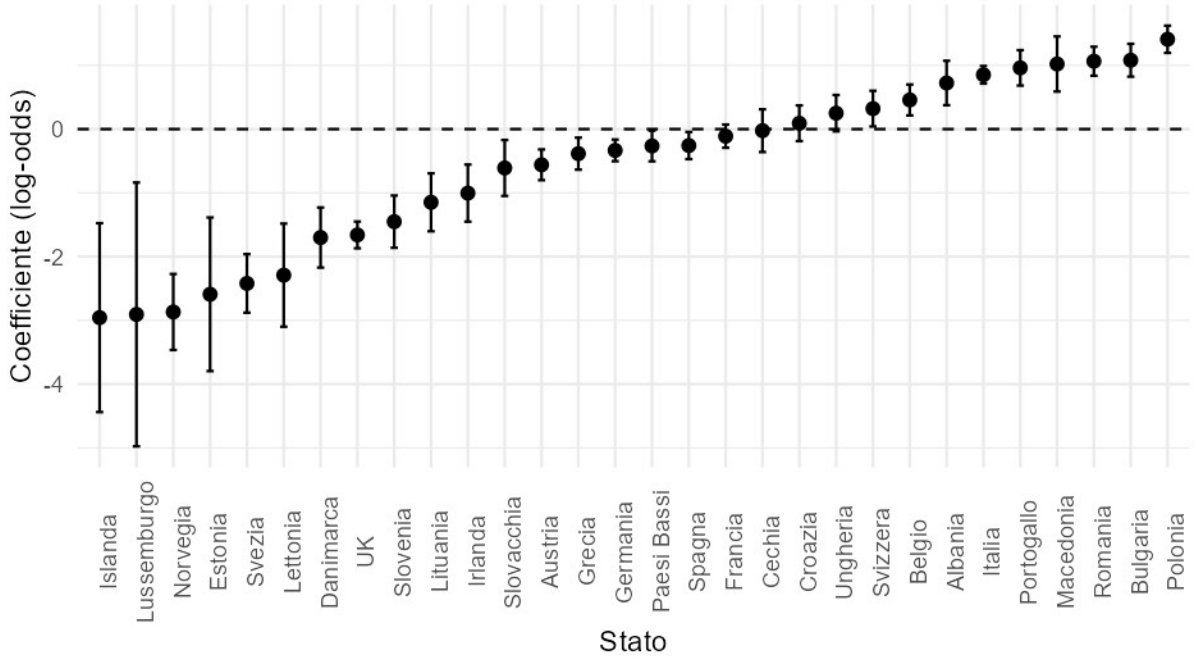

Figura 5.3: stima ed intevalli di confidenza 95% dei coefficienti beta associati agli Stati

Si osserva che la conoscenza dello Stato in cui il disastro può accadere ricopre un ruolo rilevante. Prendendo l'Italia come Stato di riferimento, la maggior parte degli Stati risulta, a parità di condizioni, associata a una minore probabilità di eventi dannosi rispetto all'Italia., con alcune eccezioni per le quali il modello stima un rischio più elevato.

A causa della ridotta dimensione del campione, sia in termini temporali che di numero di disastri, molti Stati presentano coefficienti con ampio margine di errore. Tali Stati, infatti, sono caratterizzati da un numero minore di province e soprattutto di eventi osservati. Per lo stesso motivo, dal grafico sono stati esclusi quei cinque Stati (Cipro, Finlandia, Liechtenstein, Montenegro e Malta) che non presentano disastri nel periodo di interesse, i quali, per tale ragione, presentavano incertezza molto elevata sui coefficienti.

5.2.5. Specificazioni alternative e risultati non conclusivi

Analogamente a quanto fatto nelle sezioni precedenti è stato anche stimato un modello con il predittore Stato insieme alle sole variabili relative alla percentuale di suolo impermeabilizzato nel 2006 e alla differenza tra 2006 e 2018. Le prestazioni di tale modello risultano simili a quelle del modello descritto in Sezione 5.2.4. Tuttavia, entrambe le variabili di impermeabilizzazione non sono risultate statisticamente significative. Per questa ragione, questa specificazione non è stata ulteriormente approfondita.

Sono stati anche stimati modelli di regressione lineare aventi il numero di decessi come variabile risposta, sia a scala nazionale, sia a scala provinciale. Tuttavia, le prestazioni di tali modelli sono risultate poco significative, con indici di determinazione R^2 minori del 2%.

A scala statale questo risultato può essere imputato alla grande variabilità del numero di decessi fra i diversi anni e i diversi Stati. Inoltre, la presenza di un numero alto di dati mancanti - pari al 32% dei disastri – e le grandi differenze nel numero di decessi registrati (come evidenziato dal boxplot sul lato sinistro della Figura 5.4 - rendono molto complessa una modellizzazione affidabile di tale indicatore. Considerando che, a questa scala, il dato è osservato direttamente, in quanto riportato dal dataset, e che il numero di decessi è il più noto degli indicatori di danno, nonché il più chiaramente definibile senza ambiguità di stime, tentare di modellizzare altri indicatori di danno (quali persone colpite o danni economici) risulterebbe ancora più problematico.

Il computo a scala provinciale risulta ulteriormente difficile, poiché tali dati non sono originariamente disponibili, ma sono stati stimati sulla base delle informazioni di localizzazione e del dato complessivo del disastro. Anche questi dati risultano molto variabili tra province e anni, come si osserva sul lato destro della Figura 5.4. Inoltre, a causa del metodo di ottenimento, descritto nella Sezione 3.1.2 e basato sulla popolazione dell'area, essi hanno in parte perso il significato originario di "numero di decessi", divenendo un indicatore ancor meno interpretabile.

Con riferimento ai boxplot già menzionati, va segnalato che il 62,8% e il 63,3% di dati rispettivamente a scala statale e provinciale non sono presenti in tale rappresentazione, in quanto non riportanti alcun decesso.

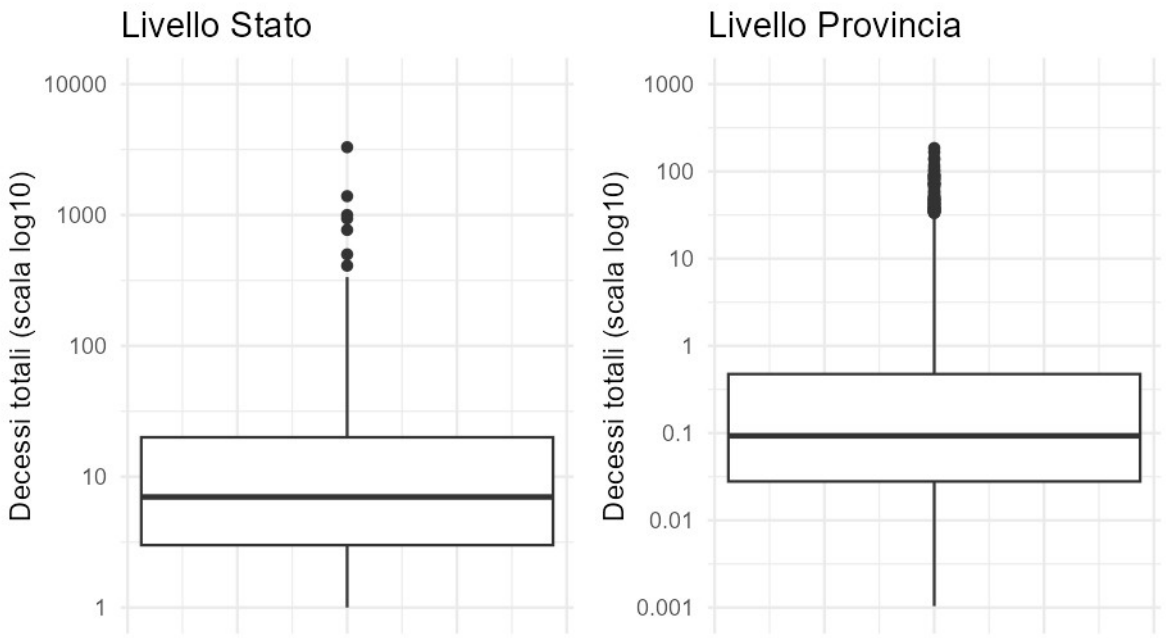

Figura 5.4: boxplot dei decessi totali e stimati, a scala statale e provinciale, in scala logaritmica a causa della differenza di valori.

Conclusioni

Sebbene il disastro naturale sia percepito come evento raro ed estremo, l'analisi a scala europea mostra una distribuzione spaziale e temporale non uniforme, con concentrazioni in specifici Stati e periodi, che ridimensionano l'idea di una rarità omogeneamente distribuita sul territorio.

A scala statale, la variabilità tra Stati risulta sistematicamente superiore alla variabilità nel tempo all'interno dello stesso Stato. Questo suggerisce che le differenze strutturali territoriali pesano maggiormente delle oscillazioni annuali nel periodo considerato, almeno con riferimento agli indicatori utilizzati.

I fenomeni che determinano il verificarsi di un evento dannoso si sviluppano tuttavia su scale spaziali e temporali più ristrette rispetto a quelle adottate nell'analisi. Di conseguenza, indicatori climatici aggregati su base annuale e statale colgono solo parzialmente la dinamica degli eventi estremi. In questo contesto, le variabili climatiche mostrano un contributo limitato nei modelli stimati, mentre risultano più rilevanti le variabili legate all'esposizione territoriale.

Per ragioni analoghe, la modellazione quantitativa dell'intensità del danno risulta complessa, anche a causa dell'elevata asimmetria e incompletezza dei dati disponibili; per questo motivo l'analisi si è concentrata sulla probabilità di occorrenza di eventi con danno, anziché sulla stima dell'ammontare dei danni stessi.

A scala provinciale emerge una maggiore variabilità territoriale e aumenta il numero di variabili statisticamente rilevanti. La capacità discriminatoria del modello risulta leggermente inferiore rispetto alla scala nazionale (AUC compresa tra 0,63 e 0,68), ma migliora sensibilmente includendo l'informazione relativa allo Stato di appartenenza della provincia.

L'effetto dell'impermeabilizzazione del suolo risulta articolato. Il livello iniziale (2006) presenta un'associazione negativa con la probabilità stimata di evento dannoso, verosimilmente legata alla distribuzione territoriale degli eventi rispetto alle aree maggiormente urbanizzate. Diversamente, l'incremento della percentuale di suolo impermeabilizzato nel periodo 2006–2018 risulta positivamente associata al rischio, con un effetto di entità comparabile a quello osservato per popolazione e temperatura media nel modello provinciale senza controllo per Stato. Tuttavia, tale evidenza non si mantiene con la stessa stabilità al variare della specificazione del modello, suggerendo cautela nell'interpretazione di un effetto causale diretto del consumo di suolo nel periodo considerato.

In generale, l'analisi evidenzia come la scala spaziale di osservazione, la qualità dei dati disponibili e la struttura di correlazione tra le variabili influenzino in modo determinante l'interpretazione delle relazioni tra pressione antropica e danni da disastri ambientali. In

questo quadro, il consumo di suolo — in particolare nella sua componente dinamica — entra nei modelli stimati e mostra associazioni significative con la probabilità di evento dannoso a scala provinciale, confermandone la rilevanza tra i fattori territoriali di esposizione. I risultati suggeriscono quindi che la trasformazione del territorio costituisca un elemento non trascurabile nell'analisi del rischio, pur all'interno di un sistema complesso in cui fattori climatici, demografici e strutturali interagiscono tra loro.

Infine, la disponibilità di dati pubblici armonizzati e geolocalizzati sui danni risulta limitata. EM-DAT rappresenta la principale fonte accessibile e comparabile a scala europea, pur essendo focalizzata sugli eventi di maggiore rilevanza e presentando alcune criticità in termini di completezza e dettaglio spaziale. L'analisi si è pertanto basata su tale fonte, consapevoli che essa non intercetta la totalità degli eventi minori.

Indice delle figure

Indice delle tabelle

Bibliografia

Centre for Research on the Epidemiology of Disasters (CRED). (2025, 10 10). *EM-DAT: The International Disaster Database*. (U. c. (UCLouvain), Editor) Retrieved from EM-DAT: https://www.emdat.be

Commissione europea. (2012). *Orientamenti in materia di buone pratiche per limitare, mitigare e compensare l'impermeabilizzazione del suolo.* Lussemburgo: Commissione europea. Retrieved from https://op.europa.eu/it/publication-detail/-/publication/e9a42c93-0825-4fc0-8032-a5975c8df3c0/language-en

Copernicus Land Monitoring Service. (2021). *Algorithm Theoretical Basis Document – High Resolution Layer Imperviousness 2021.* Luxembourg. Retrieved from https://land.copernicus.eu/en/technical-library/algorithm-theoretical-basis-document-high-resolution-layer-imperviousness-2021

Dahlman, R. L. (2025). *Climate change: global temperature.* Retrieved from Climate.gov: https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

Daniell, J. E., Vervaeck, A., & Khazai, B. (2011). The CATDAT damaging earthquakes database. *Natural Hazards and Earth System Sciences*, pp. 2235–2251. doi:10.5194/nhess-11-2235-2011

European Commission. (2019). *Methodological manual on territorial typologies.* Luxembourgh. Retrieved from https://ec.europa.eu/eurostat/documents/3859598/9530302/KS-GQ-19-008-EN-N.pdf

European Environment Agency (EEA). (2025, 10 10). *Imperviousness in Europe.* Retrieved from https://www.eea.europa.eu/en/analysis/maps-and-charts/imperviousness-in-europe

European Environment Agency (EEA); Copernicus Land Monitoring Service. (2025, 11 10). *High Resolution Layer – Imperviousness (IMD).* Retrieved from https://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness

European Environment Agency. (2024). *European Climate Risk Assessment.* Copenhagen: European Environment Agency. Retrieved from https://www.eea.europa.eu/en/analysis/publications/european-climate-risk-assessment

European Environment Agency. (2024). *European Climate Risk Assessment.* Publications Office of the European Union. doi:10.2800/8671471

European Environmental Agency. (2024). *Economic losses and fatalities from weather- and climate-related extremes.* Copenhagen. Retrieved from https://www.eea.europa.eu/en/analysis/publications/economic-losses-from-climate-extremes

Eurostat (GISCO). (2021). *Territorial units for statistics (NUTS) — GISCO geospatial dataset.* Retrieved from https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/territorial-units-statistics

Eurostat. (2024). *Fertilitiy statistics*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Fertility_statistics

Eurostat. (2024). *Population on 1 January by age group, sex and NUTS 3 region".* Retrieved from https://ec.europa.eu/eurostat/databrowser/view/demo_r_pjangrp3/default/table

Eurostat. (2024). *Population statistics at regional level.* Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?oldid=587819

Geoffrey Norman, D. l. (2000). *Biostatistica.* Zanichelli.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, D20119. doi:10.1029/2008JD010201

Integrated Research on Disaster Risk. (2014). *Peril Classification and Hazard Glossary.* Beijing: Integrated Research on Disaster Risk. Retrieved from https://council.science/wp-content/uploads/2019/12/Peril-Classification-and-Hazard-Glossary-1.pdf

Intergovernmental Panel on Climate Change. (2022). *AR6 Working Group II – Glossary.* Geneva: Intergovernmental Panel on Climate Change. Retrieved from https://www.ipcc.ch/report/ar6/wg2/

M., O. (n.d.). A global perspective on cold wave disasters. *Sci Rep 16, 1537*. doi:https://doi.org/10.1038/s41598-025-31547-4

Melchiorri, M. e. (2018). *'Unveiling 25 Years of Planetary Urbanization with Remote Sensing: Perspectives from the Global Human Settlement Layer.*

Ocal, M. (2025, 12 8). A global perspective on cold wave disasters. *Nature*. Retrieved from https://www.nature.com/articles/s41598-025-31547-4?

R Core Team. (2025). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.* Vienna. Retrieved from URL https://www.R-project.org/.

Rosvold, E., & Buhaug, H. (2021). *Global Disasters Identifier System (GDIS), Version 1.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved from https://doi.org/10.7927/ZZ3B-8Y61

Royal Netherlands Meteorological Institute (KNMI). (2024). *E-OBS: Daily gridded meteorological data for Europe.* Retrieved from https://surfobs.climate.copernicus.eu/dataaccess/access_eobs.php

Shapiro, S., & Wilk, M. (1965). *An Analysis of Variance Test for Normality (Complete Samples).* Biometrika 52.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.