# Analisi dei dati Spaziali per le Applicazioni Economiche

Corso di Laurea Magistrale in: Metodi Quantitativi per le Valutazioni economiche e finanziarie

prof. Gennaro Punzo

a cura di Punzo G., Bruno E.

31 ottobre 2023

# Indice

# Introduzione

Cambiamento climatico, migrazioni, produzione alimentare, criminalità, reddito, uso del suolo, disoccupazione e crisi pandemica sono solo alcuni esempi delle numerose problematiche delle società contemporanee nelle quali lo *spazio geografico* svolge un ruolo fondamentale. Si tratta di una moltitudine di fenomeni di varia natura – economica, sociale, ambientale, culturale e, persino, antropologica – che presentano una forte connotazione spaziale dalla quale non è possibile prescindere qualora si intenda fornire una rappresentazione più possibile aderente alla realtà. Questi fenomeni, localizzati in spazi geografici, si manifestano con intensità variabili da luogo a luogo e spesso sono legati da forme di dipendenza reciproca tra aree limitrofe, in linea con il principio di Tobler secondo cui 'tutti gli eventi sono legati tra loro, ma eventi vicini sono più strettamente correlati di quelli lontani'. La sfida di pensare in maniera spaziale e di concepire i fenomeni alla luce della loro collocazione e posizione nello spazio può offrire spunti di riflessione originali e a soluzioni innovative che potrebbero non emergere da altre prospettive.

Con specifico riferimento ai fenomeni di carattere economico, ancor prima degli anni Sessanta, gli studiosi avevano ravvisato l'opportunità di integrare la dimensione geografica nello sviluppo di teorie e approcci concettuali al fine di modellare anche la complessità delle realtà locali e dei loro meccanismi di funzionamento. Tuttavia, solo nei decenni più recenti, si è assistito a un processo di 'centralizzazione' della dimensione spaziale, per lungo tempo collocata in subordine rispetto alla dimensione temporale, anche nell'analisi di un ampio spettro di tematiche tipiche della statistica economica, dal mercato del lavoro alle innovazioni tecnologiche, dalla convergenza economica alle economie di agglomerazione, dalle dinamiche dei prezzi alla deprivazione socio-economica. Il ruolo della dimensione spaziale si è consolidato e reso imprescindibile nell'analisi statistico-economica in quanto i modelli teorici hanno gradualmente sostituito la concezione di un agente economico statico per abbracciare l'idea di un agente che opera in un ambiente complesso e prende decisioni in funzione del contesto in cui si trova, in particolare, in relazione ad altri agenti con preferenze, decisioni e comportamenti propri.

Questo ha comportato, da un lato, un notevole progresso delle tecniche quantitative per dati georeferenziati, e, dall'altro, un consolidamento di queste tecniche in un corpus organico di metodologie spaziali strettamente allineate con i paradigmi teorici in continua evoluzione. L'evoluzione rapida delle tecnologie informatiche, dei sistemi informativi geografici, del Global Positioning System (GPS) e degli innumerevoli strumenti di visualizzazione grafica ha, nel contempo, consentito avanzamenti significativi anche nel novero degli studi empirici di natura economico-territoriale a diversi livelli di dettaglio geografico. La modellizzazione dei processi spaziali sottostanti ai fenomeni economici richiede il passaggio dallo spazio reale, popolato di individui, famiglie, imprese e istituzioni, a uno spazio astratto, costituito da primitive geometriche (punti, linee e poligoni) e da numeri che ne riproducono sia gli attributi che la loro posizione in un sistema di coordinate geografiche. Il primo requisito fondamentale per la modellazione degli eventi economici da una prospettiva spaziale è la definizione della prossimità spaziale, solitamente codificata attraverso una matrice (detta, per l'appunto, dei pesi spaziali) che esprime le relazioni di vicinato tra le unità e ne quantifica l'intensità. La prossimità spaziale è da intendersi non necessariamente in senso geografico di contiguità o di distanza fisica tra le unità, ma anche in senso economico (i.e., scambi commerciali, flussi di pendolarismo per motivi di lavoro) oppure in termini di caratteristiche sociali, culturali, linguistiche, istituzionali.

L'uso di tecniche che tengono conto della struttura spaziale dei dati consente di trattare adeguatamente i problemi legati all'eterogeneità spaziale, che identifica situazioni di non uniformità distributiva di un fenomeno in un stesso territorio, e all'autocorrelazione spaziale, che valuta i livelli di similarità o dissimilarità nella distribuzione spaziale del fenomeno e, quindi, le possibili relazioni multidirezionali tra ciò che avviene in un luogo e quanto accade in luoghi limitrofi. Questa è una proprietà intrinseca ed esclusiva dei dati spaziali, in base alla quale le caratteristiche di un fenomeno non sono spiegate unicamente da determinanti interne al luogo in cui il fenomeno si manifesta, ma anche da quanto si verifica nell'insieme degli altri luoghi più o meno vicini. La consapevolezza di questa peculiarità dei dati spaziali e della sua importanza per l'affidabilità delle stime ha ispirato la comunità scientifica, a livello nazionale e internazionale, a sviluppare una gamma di modelli statistici ed econometrici spaziali. Questi modelli sono in parte mutuati e adattati dalla statistica ed econometria tradizionali, ma sono in grado di individuare, valutare e trattare metodologicamente la presenza e l'entità degli effetti spaziali nella loro complessità, e il ruolo che essi esercitano nella configurazione reale del fenomeno. Questi progressi metodologici hanno portato a una proliferazione di applicazioni empiriche, anche di interesse multidisciplinare.

Se le origini della statistica spaziale risalgono ai primi anni Sessanta grazie agli sforzi di un ingegnere minerario sudafricano di nome Krige, che lavorò allo sviluppo di tecniche di previsione della posizione dei minerali nelle formazioni geologiche, e ai lavori in ambito geostatistico di Matheron, le origini dell'econometria spaziale rimandano al 1979 con la pubblicazione di Paelnick e Klaassen dal titolo emblematico di Spatial Econometrics. Negli anni successivi, la ricerca metodologica in entrambe le discipline è proseguita con successo, talvolta con sovrapposizioni ma anche con importanti tratti distintivi, per consolidarsi nel corso degli anni Ottanta e Novanta, ed è stata tale da conferire a ciascuna disciplina una propria identità. Questo percorso di ricerca è tuttora molto vivo e intenso, ben lungi dall'essere esaurito, nel quale si è innestata una proficua comunità di studiosi che ha dato origine a una produzione scientifica eloquente e consistente anche dal punto di vista quantitativo.

# Capitolo 1

# Dati spaziali

## 1.1 Tipi di spaziali

Si definiscono dati spaziali i dati che contengono una componente geografica ovvero dati che collegano l'attributo descrittivo dell'entità con una specifica posizione geografica. Ad esempio, l'affermazione "La temperatura alle ore 14:00 del 15 dicembre 2022 alla latitudine 48° 15' Nord, longitudine 16° 21' 28" Est è di 6,7°C" lega la posizione (e il tempo) all'attributo della temperatura atmosferica. La posizione geografica è, quindi, essenziale nella definizione dei dati spaziali e permette di distinguere questi dati da altre tipologie di dati non spaziali. Qualora il dato riferisse solo attributi e non la componente geografica sarebbe non spaziale anche se relativo a unità di osservazione spazialmente definite. Ad esempio, il tasso di disoccupazione rilevato nelle province italiane non sarebbe un dato spaziale se non fosse corredato da informazioni sulla posizione geografica delle province. Pertanto, dati avulsi dal contesto spaziale non sono classificabili come dati spaziali in quanto l'analisi dei dati spaziali richiede informazioni sia sugli attributi che sulle posizioni, indipendentemente da come gli attributi siano stati misurati.

Esiste una duplice visione dei dati spaziali a seconda che i fenomeni oggetto di misurazione siano osservabili, almeno teoricamente, su un continuo spaziale o si originano in punti discreti dello spazio. I primi sono fenomeni spaziali continui poiché si manifestano in modo fluido nello spazio senza interruzioni evidenti. Ad esempio, l'altitudine, l'umidità, la temperatura, la pressione atmosferica sono esempi di fenomeni spaziali continui in quanto la loro intensità varia da un punto all'altro dello spazio geografico. Similmente, le precipitazioni (pioggia, neve, ...) e l'inquinamento atmosferico (ozono, particelle sospese, ...) si verificano nel continuo spaziale poiché la quantità di precipitazioni e la concentrazione di inquinanti nell'aria possono essere

misurate in diversi punti di un'area geografica e configurate su una mappa continua delle precipitazioni o della qualità dell'aria. I secondi sono fenomeni spaziali discreti poiché si riferiscono a eventi, caratteristiche o elementi che accadono in modo separato e distintivo nello spazio. Si tratta di fenomeni che si manifestano solo in punti specifici o regioni distinte piuttosto che in modo continuo su tutto lo spazio. Ad esempio, la distribuzione di alberi in una foresta, gli edifici in una città, i disoccupati di una data regione sono esempi di fenomeni spaziali discreti poiché gli oggetti occupano posizioni discrete nello spazio e possono essere enumerati o individuati separatamente.

Sebbene la differenza tra fenomeni spaziali continui e discreti riguardi la concezione dello spazio geografico come coperto da superfici continue piuttosto che insieme di oggetti discreti, spesso tale distinzione dipende anche dal contesto di riferimento. Inoltre, se è vero che fenomeni spaziali continui come la temperatura e le caratteristiche del suolo sono osservabili ovunque sulla superficie terrestre, è altrettanto vero che le relative variabili sono spesso discretizzate. La temperatura, ad esempio, è campionata in un insieme di luoghi e spesso rappresentata graficamente tramite linee (isoterme). Le caratteristiche del suolo sono campionate in un insieme di posizioni discrete e rappresentate come un campo che varia nel continuo. In questi casi, la continuità spaziale è stata rappresentata attraverso un processo di discretizzazione.

La distinzione tra fenomeni spaziali continui e discreti è importante ai fini della classificazione di due diverse tipologie di dati spaziali denominati, per l'appunto, continui e discreti. In particolare, secondo il tipo di discontinuità, i dati spaziali discreti si classificano in:

- dati puntuali che rappresentano una posizione specifica come la sede di un'azienda o un punto di interesse su una mappa;
- dati lineari (dati di flusso origine-destinazione) rappresentano elementi come strade, fiumi o linee di confine;
- dati poligonali o areali che si riferiscono ad aree delimitate come confini amministrativi, zone di interesse o regioni geografiche.

### 1.1.1 Dati puntuali

I dati puntuali riguardano entità spaziali ovvero oggetti fissi con una posizione specifica nello spazio. I dati puntuali sono zero-dimensionali, privi di dimensioni fisiche, rappresentati da una coppia di coordinate (x, y) nel piano bidimensionale o da una terna di coordinate (x, y, z) in uno spazio tridimensionale. Un punto è indivisibile, non ha estensione, né altezza, né larghezza. I

dati puntuali possono rappresentare una vasta gamma di informazioni, come ad esempio:

- posizione di oggetti, punti di interesse ed eventi: posizione di negozi, uffici, monumenti, distributori automatici, incidenti stradali, eventi naturali, avvistamenti di specie animali;
- monitoraggio e tracciamento in tempo reale tramite dispositivi mobili o sensori;
- campionamento geografico come, ad esempio, la posizione di punti di rilevamento per studi ambientali, indagini sul territorio o rilevamenti geologici.

I dati puntuali possono essere raccolti attraverso osservazioni dirette o strumenti di rilevamento come GPS (Global Positioning System), sensori geografici, rilevamenti sul campo, fonti di dati satellitari o fonti statistiche ufficiali. Questi dati possono essere analizzati utilizzando software specializzati come i Sistemi Informativi Geografici (GIS) che permettono, tra l'altro, la visualizzazione grafica delle aree, la creazione di mappe interattive, la realizzazione di analisi spaziali, l'aggregazione e la comparazione dei dati. Le analisi possono riguardare l'identificazione di pattern o cluster, l'interpolazione per punti non campionati, l'analisi delle distanze, l'analisi di hotspot, outlier e così via. Questi dati sono utilizzati in molteplici settori quali il monitoraggio ambientale, la pianificazione urbana, la gestione delle risorse, la logistica, l'analisi di mercato e altri.

### 1.1.2 Dati lineari

I dati lineari rappresentano entità o fenomeni che seguono una forma lineare o sono caratterizzati da una connessione sequenziale di punti nello spazio geografico (flussi origine-destinazione). Nel primo caso, i dati lineari descrivono elementi geografici come strade, fiumi, ferrovie, confini, percorsi o qualsiasi altro elemento dalla forma lineare. Nel secondo caso, rilevano, ad esempio, flussi di persone, merci, capitali, informazioni o conoscenze da un punto di origine a un punto di destinazione e sono rilevanti negli studi di pianificazione dei trasporti, migrazione della popolazione, viaggi di lavoro, comportamento di acquisto, flussi di merci, trasmissione di informazioni e conoscenze attraverso lo spazio. In entrambi i casi, i dati lineari descrivono elementi che hanno una lunghezza e una direzione all'interno di un sistema di riferimento spaziale. Le linee, infatti, sono entità geometriche unidimensionali per cui possono anche includere attributi associati a specifici segmenti della linea

che riguardano la forma, lunghezza, direzione, distanza, intersezioni o connettività tra le linee al fine di comprendere meglio le relazioni spaziali tra gli elementi. Ad esempio, le analisi di rete riguardano la valutazione dei percorsi ottimali, la determinazione delle distanze o dei tempi di percorrenza tra i punti di interesse; le analisi di connettività concernono l'identificazione delle connessioni tra le linee per determinare, ad esempio, i punti in cui le strade si intersecano; le analisi di flusso permettono la valutazione dei movimenti lungo le linee, i.e. il flusso di traffico lungo una strada o il flusso di acqua lungo un fiume; le analisi di densità consentono l'identificazione di zone in cui le linee sono più dense o sparse per comprendere, ad esempio, la densità della rete stradale in un'area urbana. I dati lineari sono utilizzati in diversi settori, come la pianificazione del trasporto, la gestione delle reti di distribuzione e delle risorse idriche, l'analisi di reti sociali e molto altro.

### 1.1.3 Dati areali

I dati poligonali o areali rappresentano aree o superfici geografiche delimitate da confini o poligoni. Questi dati descrivono entità geografiche come paesi, regioni, province, comuni, quartieri e così via. I dati areali sono costituiti da poligoni che definiscono i confini dell'area ognuno dei quali è associato a coordinate geografiche o coordinate cartesiane che indicano la posizione nel sistema di riferimento spaziale. Le analisi dei dati areali si basano di solito sulla mappatura tematica che mostra le aree e gli attributi associati ad esse e riguardano le analisi di valutazione della distribuzione spaziale degli attributi (es. per identificare aree di alta o bassa densità di popolazione) e delle relazioni spaziali tra le aree per identificare, ad esempio, processi di autocorrelazione spaziale. I dati areali sono fondamentali per comprendere le caratteristiche e le dinamiche delle diverse aree geografiche e supportano la pianificazione territoriale, l'analisi socio-economica, la gestione dell'ambiente e delle risorse naturali e così via.

I dati areali si riferiscono a una situazione in cui il fenomeno oggetto di studio non varia nel continuo spaziale, ma presenta valori solo all'interno di un insieme fisso di zone che coprono l'area di studio. Quest'ultima può essere costituita da un reticolo regolare (es. pixel nel telerilevamento) o da un insieme di unità areali irregolari (es. tratti di censimento). È possibile, pertanto, distinguere due tipologie di partizioni del territorio: le aree regolari e le aree irregolari. Le aree regolari, che trovano applicazione, ad esempio, nelle immagini rilevate da satellite o da aereo (fotogrammetria), sono ottenute da una suddivisione della superficie in poligoni di forma regolare come quadrati, rettangoli ed esagoni che, a loro volta, diventano unità areali di base per l'analisi dei fenomeni socio-economici. Le unità areali sono costituite da celle

regolari di un reticolo, referenziate da un sistema di coordinate geografiche come latitudine e longitudine o da altri sistemi di riferimento spaziale nei quali le intensità dei fenomeni corrispondono alle colorazioni assunte dai pixels delle immagini rilevate. Le aree irregolari di cui le aree amministrative (es. regioni, province, comuni) sono un sottoinsieme ricorrono di frequente come unità di rilevazione delle informazioni territoriali.

## 1.2 Acquisizione e gestione dei dati spaziali

File shp, dbf, shx..

# Capitolo 2

# Matrici dei pesi spaziali

Le matrici dei pesi spaziali (**W**) rappresentano uno strumento metodologico utilizzato per modellare la struttura spaziale dei dati. La loro finalità è quella di fornire una rappresentazione semplificata delle relazioni di vicinanza e interconnessione tra le unità presenti nello spazio geografico.

Sia n il numero di unità spaziali. La matrice dei pesi spaziali è una matrice quadrata nxn positiva, simmetrica e non stocastica, dove ogni riga e colonna rappresentano un'unità geografica (regione, provincia, comune, ...). Ogni elemento ($w_{ij}$) della matrice rappresenta un peso ovvero una misura di vicinato o connessione tra ciascuna coppia di unità. I pesi wij sono assegnate in base a regole predefinite che definiscono le relazioni spaziali tra le unità. In generale, gli elementi $w_{ij}$ possono essere di due tipi: binari, se riflettono solo la presenza o l'assenza di connessione, oppure continui, se esprimono anche l'intensità delle relazioni spaziali. In ogni caso, per convenzione, gli elementi sulla diagonale hanno tutti valore nullo. Esistono diverse tipologie di matrici dei pesi spaziali, che possono essere raggruppate in quattro principali categorie:

- **matrici di contiguità**: basate sulla contiguità spaziale, cioè sulla condivisione di confini tra unità geografiche. Gli elementi della matrice indicano se due unità sono confinanti o meno;

- **matrici basate su distanze**: basate su misure di distanza (euclidea o geodetica) per il calcolo dei pesi tra le unità. Gli elementi della matrice riflettono la distanza spaziale tra le unità;

- **matrici definite sulla base di riorganizzazioni territoriali**: basate su modifiche o riorganizzazioni del territorio, come la creazione di nuove suddivisioni amministrative;

- **matrici ottenute da analisi statistiche**: queste matrici sono create utilizzando metodi statistici per identificare le relazioni spaziali significative tra le unità geografiche. Gli elementi della matrice riflettono l'incidenza o la forza di tali relazioni.

## 2.1 Matrici di contiguità: criteri e ordini

Il primo gruppo comprende matrici la cui costruzione si basa esclusivamente sulla contiguità geografica tra coppie di unità. In queste matrici, l'elemento generico ($w_{ij}$) rappresenta la vicinanza o l'assenza di vicinanza tra l'unità i e l'unità j secondo una logica binaria. In particolare, assume il valore 1 se le due unità sono contigue e 0 altrimenti. È importante sottolineare che le matrici di contiguità binaria sono simmetriche, poiché se l'unità i è contigua all'unità j, allora l'unità j sarà contigua all'unità i. La proprietà della simmetria di queste matrici è vantaggiosa per semplificare i relativi calcoli e gli algoritmi correlati. Le matrici di contiguità binaria fanno riferimento al paradigma degli scacchi, da cui mutuano il criterio della torre (rook), dell'alfiere (bishop) e della regina (queen). Nel caso della contiguità della torre, due unità sono considerate contigue se condividono una porzione di confine, riflettendo il movimento rettilineo della torre negli scacchi, che può avvenire orizzontalmente o verticalmente. Analogamente, prendendo spunto dall'alfiere negli scacchi, due unità sono considerate contigue secondo il criterio dell'alfiere se presentano un punto o vertice in comune, poiché l'alfiere può muoversi solo lungo le diagonali. La contiguità della regina rappresenta la forma più ampia di contiguità, poiché include sia la contiguità della torre che quella dell'alfiere. La regina, infatti, può muoversi in tutte le direzioni: orizzontale, verticale e diagonale. Pertanto, due unità sono considerate contigue secondo il criterio della regina se condividono una parte di confine o anche solo un vertice.

La matrice di contiguità binaria precedentemente descritta è classificata come matrice di primo ordine, in quanto cattura la vicinanza tra unità direttamente adiacenti in base ai criteri illustrati in precedenza. È possibile ricorrere a matrici di contiguità binaria di ordine superiore quando specifiche esigenze dell'analisi o della scala spaziale considerata lo richiedono. Una matrice di contiguità spaziale di secondo ordine definisce la contiguità tra unità che sono distanti al massimo due passi (salti) l'una dall'altra nello spazio di riferimento. In generale, una matrice di contiguità spaziale di ordine r rappresenta la contiguità tra unità che sono distanti al massimo (r – 1) passi l'una dall'altra. Le matrici di contiguità spaziale di ordini superiori consentono di definire relazioni di contiguità più estese nello spazio, includendo un

numero maggiore o minore di unità vicine a distanze crescenti, a seconda che ci si sposti dal mare verso la terra o, al contrario, dalla terra verso il mare. La scelta dell'ordine di contiguità dipende dall'applicazione specifica e dalla scala spaziale considerata.

Una matrice di contiguità di secondo ordine potrebbe essere usata, ad esempio, nelle analisi dei flussi migratori tra diverse regioni, nelle analisi della diffusione delle malattie, negli studi della pianificazione urbana. Tale matrice cattura la contiguità tra le unità che sono a una distanza massima di un salto. Ciò consente di identificare le aree che sono interconnesse da flussi migratori indiretti, facilitando la comprensione delle dinamiche migratorie. Nel contesto dell'epidemiologia, una matrice di contiguità di secondo ordine può fornire informazioni cruciali sulla propagazione delle malattie utili per la pianificazione di interventi preventivi. Nell'ambito della pianificazione urbana, una matrice di contiguità di secondo ordine permette di analizzare la connettività tra quartieri o zone urbane, facilitando la valutazione dell'accessibilità e della connettività urbana e contribuendo alla pianificazione efficiente dello sviluppo urbano.

Le matrici di contiguità di ordine superiore al primo possono essere definite in modo da includere anche gli ordini inferiori. In tal caso, la matrice di contiguità di secondo ordine rappresenta la contiguità tra le unità adiacenti direttamente (primo ordine) e le unità che sono a una distanza massima di un salto (secondo ordine). In modo simile, la matrice di contiguità di terzo ordine rappresenta la contiguità tra le unità adiacenti direttamente (primo ordine), quelle a una distanza massima di un salto (secondo ordine) e quelle a una distanza massima di due salti (terzo ordine). L'inclusione degli ordini inferiori nelle matrici di contiguità di ordine superiore consente di ottenere una rappresentazione più completa delle relazioni di contiguità nello spazio geografico. Questa strategia permette di considerare le connessioni indirette e le interazioni a distanze maggiori, ampliando la comprensione delle dinamiche spaziali e fornendo un quadro più dettagliato delle relazioni tra le unità geografiche. Tuttavia, l'inclusione degli ordini inferiori nelle matrici di contiguità aumenta la complessità computazionale e può richiedere risorse aggiuntive per l'analisi e l'interpretazione dei risultati. La scelta di utilizzare tali matrici dipende dalle specifiche esigenze dell'analisi e dalla scala spaziale considerata oltre che dalla disponibilità dei dati e delle risorse computazionali adeguate.

Le matrici di contiguità di ordine superiore si prestano a una gamma di applicazioni. Ad esempio, le matrici di contiguità di ordine tre trovano applicazione nell'analisi dei flussi commerciali, permettendo di identificare le connessioni indirette e le relazioni commerciali tra regioni adiacenti a una distanza massima di due salti. Una matrice di contiguità di quarto ordine

potrebbe essere utile per esaminare la diffusione di elementi culturali tra diverse città, come gli stili architettonici o le pratiche artistiche, in modo da modellare il processo di diffusione culturale. Nel contesto dei trasporti, una matrice di contiguità di quinto ordine potrebbe evidenziare le connessioni indirette tra città o porti marittimi, inclusi i flussi di trasporto tra diverse città intermedie. Nella pianificazione delle reti di distribuzione, una matrice di contiguità di sesto ordine potrebbe aiutare a identificare i punti di distribuzione interconnessi da rotte indirette, compresi i punti intermedi di trasbordo o stoccaggio, contribuendo alla progettazione efficiente delle reti di distribuzione.

È importante sottolineare che la configurazione delle matrici di contiguità di ordine superiore dipende, oltre che dai criteri utilizzati per determinare la contiguità, anche dalle specifiche caratteristiche geografiche delle unità geografiche in esame. Ad esempio, nelle aree costiere, la presenza del mare influisce sulla configurazione delle relazioni di contiguità. Infatti, se ci si sposta dalle aree interne verso le aree costiere, le matrici di contiguità di ordine superiore riflettono questa situazione, mostrando una maggiore connessione tra le unità adiacenti dirette e connessioni via via minori all'aumentare dell'ordine di contiguità (quindi, a distanze maggiori). Al contrario, se ci si sposta dalle aree costiere verso le aree interne, le matrici di contiguità di ordine superiore mostrano una connessione via via maggiore all'aumentare dell'ordine di contiguità.

### 2.1.1 Limiti della contiguità binaria

Le matrici di contiguità binaria, pur essendo uno strumento ampiamente utilizzato nell'analisi spaziale, presentano una serie di limitazioni, tra cui:

- eccessiva semplificazione della realtà: le matrici di contiguità binaria riducono la complessità delle connessioni spaziali alle sole informazioni di vicinanza tra le unità, esprimendo solo la presenza o l'assenza di adiacenza geografica senza considerare la natura o l'intensità di tali relazioni;

- ipotesi di mera contiguità: tali matrici si basano sull'assunzione che le unità geografiche siano contigue solo se sono adiacenti direttamente (primo ordine) o indirettamente (ordini superiori), senza considerare altre forme di connessione come l'accessibilità attraverso strade o altre infrastrutture;

- mancata considerazione della distanza: queste matrici trattano tutte le unità adiacenti allo stesso modo, senza considerare la reale distanza

che le separa. Questo limite può influire sulla rappresentazione delle relazioni di vicinanza e connettività, soprattutto in contesti in cui la distanza riveste un ruolo significativo nell'interazione tra le unità geografiche;

- ignoranza della morfologia del territorio: le matrici di contiguità binaria non considerano le caratteristiche morfologiche del territorio, come montagne, fiumi e altre barriere naturali, compromettendo una realistica rappresentazione delle relazioni spaziali. Tale limite si verifica soprattutto nel caso di contesti geografici complessi o accidentati per i quali non vengono considerate le difficoltà o facilitazioni nel movimento attraverso il territorio;

- dipendenza dalla definizione e lunghezza dei confini: qualsiasi modifica o discrepanza nella definizione dei confini può influire sulla rappresentazione delle relazioni spaziali. Inoltre, tali matrici non tengono conto della diversa lunghezza dei confini che un'unità geografica condivide con altre unità. Ciò implica che le unità geografiche, indipendentemente dalla lunghezza dei loro confini, vengano trattate in modo uniforme, senza tener conto delle loro diverse estensioni che potrebbero influenzare la natura e l'intensità delle relazioni spaziali. Si consideri, ad esempio, un'unità geografica che confina con due altre unità: con la prima condivide un confine lungo 30 km, mentre con la seconda condivide un confine lungo 70 km. È ragionevole ipotizzare un potenziale di interazione maggiore con la seconda unità, anche solo per via della maggiore estensione del confine che le separa. Considerare le informazioni sulla lunghezza dei confini delle unità geografiche si rivela di fondamentale importanza per comprendere la natura e l'intensità delle relazioni di contiguità tra di esse. Ignorare queste differenze significherebbe trascurare dettagli importanti per l'analisi spaziale e potrebbe condurre a una visione distorta della complessità delle dinamiche territoriali;

- mancata considerazione dei pesi: le matrici di contiguità binaria forniscono informazioni solamente sulla presenza o l'assenza di contiguità tra le unità geografiche, senza considerare l'intensità o la forza di tali relazioni. Ciò costituisce uno svantaggio in contesti in cui è fondamentale considerare la gravità o l'intensità delle interazioni tra le unità, limitando la capacità di analizzare e comprendere appieno la complessità delle dinamiche che caratterizzano tali interazioni.

## 2.2 Matrici di distanza

Il secondo gruppo di matrici è costituito da matrici basate sulla distanza fisica ($d_{ij}$) tra le unità. A differenza delle matrici di contiguità binaria, che si basano sulla presenza o assenza di confini condivisi, le matrici di distanza utilizzano la distanza tra le unità come indicatore della loro relazione. Tali matrici consentono di superare il limite della contiguità binaria poiché tengono conto della distanza effettiva tra le unità geografiche. Ciò permette di catturare in modo più preciso le relazioni di vicinanza e accessibilità, in accordo con la prima legge della geografia di Tobler, secondo cui le unità geograficamente più vicine realizzano una maggiore interazione rispetto a unità più distanti. L'utilizzo di matrici basate sulla distanza richiede la scelta del tipo di distanza più opportuno. Le metriche comuni includono la distanza euclidea, la distanza di Manhattan e la distanza di Minkowski. Tuttavia, esistono anche altre metriche specifiche che possono essere adottate in base alle esigenze del contesto specifico.

### 2.2.1 La distanza euclidea

La distanza euclidea è una metrica di distanza ampiamente adottata per calcolare la distanza tra due punti spazialmente definiti dalle rispettive coordinate geografiche (latitudine e longitudine). Il calcolo della distanza euclidea si basa sul teorema di Pitagora per cui dati due punti $(i, j)$ con coordinate $(x_1, y_1)$ e $(x_2, y_2)$ nel piano bidimensionale, la distanza euclidea ($d_{ij}$) è data dalla seguente relazione:

$$d_{ij} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.1}$$

Tale formulazione può essere estesa in spazi di dimensioni superiori, come, ad esempio, lo spazio tridimensionale o n-dimensionali.

La distanza euclidea soddisfa diverse proprietà che la rendono una metrica utile in varie applicazioni:

- è sempre non negativa: $d_{i,j} \geq 0$;
- è uguale a zero solo se i punti sono coincidenti: $d_{i,j} = 0$ se e solo se $x_1 = x_2$ e $y_1 = y_2$;
- è simmetrica: $d[(x_1, y_1); (x_2, y_2)] = d[(x_2, y_2); (x_1, y_1)]$;
- soddisfa la disuguaglianza triangolare:
  $d[(x_1, y_1); (x_3, y_3)] \leq d[(x_1, y_1); (x_2, y_2)] + d[(x_2, y_2); (x_3, y_3)]$.

## 2.3 Distanza di Manhattan

A differenza della distanza euclidea, la distanza di Manhattan considera solo gli spostamenti orizzontali e verticali, senza tener conto degli spostamenti diagonali. La denominazione deriva dalla somiglianza con il percorso che una persona dovrebbe seguire per spostarsi tra due punti all'interno di un reticolo stradale ortogonale, tipico della griglia di strade di Manhattan. Secondo questa metrica, la distanza ($d_{i,j}$) tra due punti (i, j) con coordinate $(x_1, y_1)$ e $(x_2, y_2)$ nel piano bidimensionale è data dalla somma delle differenze in valore assoluto delle coordinate orizzontali e verticali:

$$d_{i,j} = |x_2 - x_1| + |y_2 - y_1| \tag{2.2}$$

Questa metrica è particolarmente utile in applicazioni in cui gli spostamenti sono vincolati a percorsi rettilinei, come nella pianificazione urbana, nella logistica o nell'analisi dei flussi di traffico. Anche tale formulazione può essere estesa in spazi di dimensioni superiori, come, ad esempio, lo spazio tridimensionale o n-dimensionali.

La distanza di Manhattan soddisfa le seguenti proprietà:

- è sempre non negativa: $d_{i,j} \geq 0$;
- è uguale a zero solo se i punti sono coincidenti: $d_{i,j} = 0$ se e solo se $x_1 = x_2$ e $y_1 = y_2$;
- non è simmetrica: $d[(x_1, y_1); (x_2, y_2)] \neq d[(x_2, y_2); (x_1, y_1)]$;
- soddisfa la disuguaglianza triangolare:
  $d[(x_1, y_1); (x_3, y_3)] \leq d[(x_1, y_1); (x_2, y_2)] + d[(x_2, y_2); (x_3, y_3)]$.

## 2.4 Distanza di Minkowski

La distanza di Minkowski è una metrica di distanza generalizzata tale da includere sia la distanza euclidea che la distanza di Manhattan come casi particolari. La formula per la distanza di Minkowski ($d_{i,j}$) tra due punti (i, j) con coordinate $(x_1, y_1)$ e $(x_2, y_2)$, con il parametro p che determina il tipo di distanza:

$$d_{i,j} = (|x_2 - x_1|^p + |y_2 - y_1|^p)^{1/p} \tag{2.3}$$

Il valore di p può essere regolato per adattarsi a contesti ed esigenze specifiche. Se $p = 1$, la distanza di Minkowski coincide con la distanza

di Manhattan; se $p = 2$, la distanza di Minkowki coincide con la distanza euclidea. Per altri valori di p, la distanza di Minkowski tiene conto di combinazioni pesate delle differenze assolute delle coordinate lungo ciascuna dimensione. Per $p > 2$, la distanza di Minkowski tende a penalizzare le differenze più grandi tra le coordinate dei punti. La distanza di Minkowski soddisfa i requisiti di non negatività, identità nel caso di punti coincidenti e disuguaglianza triangolare.

Nell'ambito dell'analisi dei dati spaziali, nelle matrici di distanza i pesi sono determinati dal reciproco della distanza. In tal modo, in linea con la prima legge della geografia, si realizza una relazione inversamente proporzionale tra distanza e intensità della connessione per cui quanto più due unità sono vicine tra loro, maggiore sarà il peso associato a tale collegamento. Unità molto distanti avranno pesi via via più bassi o addirittura nulli, indicando una minore o un'assenza di connessione spaziale. Questo significa che due unità molto vicine avranno un peso alto, mentre unità distanti avranno un peso basso.

# Capitolo 3

# Autocorrelazione spaziale

Il concetto di autocorrelazione spaziale, noto anche come associazione spaziale, riveste un ruolo di primaria importanza nell'ambito delle statistiche spaziali e deriva direttamente dalla *prima legge della geografia* di Tobler secondo cui: "everything is related to everythingelse, but near things are more related than distant things"[1] (Tobler, 1979).

Nella statistica classica l'analisi della correlazione ha come assunto di base l'indipendenza tra le unità sulle quali sono rilevate le variabili. Tale assunto viene meno quando si lavora con dati territoriali in quanto le unità osservate potrebbero non essere indipendenti, ma influenzarsi reciprocamente. Questo concetto è fondamentale per comprendere la distribuzione e la struttura dei dati geospaziali, come ad esempio le variazioni di temperatura, la diffusione di malattie o la distribuzione di risorse naturali.

Una volta definito come le osservazioni nello spazio possano essere connesse tra loro tramite l'utilizzo di una matrice dei pesi spaziali, è possibile ottenere una misura che valuti la relazione tra unità spaziali vicine. L'autocorrelazione spaziale è una misura statistica che consente di valutare se le osservazioni in punti diversi all'interno di un'area presentano una qualche forma di correlazione o dipendenza spaziale misurando l'orientamente e l'intensità della relazione di una determinata variabile rispetto a quella stessa variabile osservata in punti vicini. Proprio per sottolineare che si valuta la relazione di ogni variabile con se stessa si preferisce il termine autocorrelazione a quello di correlazione. Tale concetto, dunque, deriva dalla presenza di relazioni funzionali tra quanto accade in un punto dello spazio e quanto accade in altri punti. Il fenomeno, inoltre, si caratterizza per la multidirezionalità delle relazioni di interdipendenza spaziale (al contrario della dipendenza tem-

[1] Tutto è correlato con tutto e le cose più vicine sono maggiormente correlate rispetto alle coselontane.

porale che è solo unidirezionale), vale a dire che le aree considerate possono influenzarsi vicendevolmente.

In particolare, si parla di:

- autocorrelazione spaziale posita, o attrazione, quando valori simili di una variabile tendono a raggrupparsi in prossimità l'uno dell'altro,a formare cioè cluster più o meno grandi.

- autocorrelazione spaziale negativa, o repulsione, quando valori simili di una variabile tendono adessere separati o dispersi sul territorio.

- assenza di autocorrelazionespaziale, o indipendenza spaziale, quando la distribuzione nello spazio di una variabile è casuale(Boots e Getis, 1988).

Si tratta di valutare se e come una variabile osservata in due località (generalmente prossime in termini geografici) possa assumere valorisimili o dissimili con una certa puntualità. Tale fenomeno, quindi, scaturisce dalla presenza, più o meno costante, di elementi di dipendenza del valore assunto da una data variabile, osservata su una determinata unità territoriale, nei confronti di quello che la stessa variabile assume in altre unità vincolate alla precedente per motivi generalmente di prossimità geografica.

L'analisi volta a misurare l'autocorrelazione spaziale è comunemente indicata come "analisi esplorativa dei dati spaziali", e viene effettuata utilizzando un insieme di tecniche statistiche che consentono di analizzare e descrivere la distribuzione spaziale di una variabile, di identificare eventuali outliers e localizzazioni atipiche, di verificare l'esistenza di cluster di aree con comportamenti simili. Ciò che rende il concetto di autocorrelazione così importante è che molte delle caratteristiche economiche, sociali e ambientali di una popolazione risultano spazialmente associate. Ad esempio, particolari valori di reddito medio, del livello di istruzione,del tasso di occupazione, non sono distribuiti in maniera indipendente sul territorio,ma al contrario tendono a concentrarsi in aree particolari (Levine, 1999). Anche gli shock specifici di una regione influenzano non solo il rispettivo mercato del lavoro,ma si estendono anche alle aree adiacenti attraverso meccanismi di propagazione o diffusione spaziale (Molho, 1995). In tal modo, quindi, anche la produzione di un'area è influenzata da quella nelle aree limitrofe, nonché dalla disponibilità dei propri input e di quelli delle aree circostanti.

Di seguito sono descritte le principali misure di autocorrelazione spaziale.

## 3.1 Misure di autocorrelazione globale

Le misure, o indici, di autocorrelazione spaziale globale sono utilizzati nell'analisi spaziale per valutare se c'è una correlazione complessiva tra le osservazioni in uno spazio geografico. L'obiettivo di queste misure è determinare se le osservazioni simili (o dissimili) sono più propense a localizzarsi in prossimità le une delle altre di quanto ci si aspetterebbe in caso di distribuzione casuale. Questo concetto è fondamentale per capire se ci sono tendenze o modelli nello spazio che possono avere impatti significativi su vari aspetti, come pianificazione urbana, epidemiologia, o qualsiasi altro campo in cui la distribuzione geografica dei dati è rilevante.

### 3.1.1 Moran's I

L'Indice di Moran (Moran 1950) è una misura di autocorrelazione spaziale globale utilizzata nell'ambito delle statistiche spaziali per valutare se le osservazioni in una determinata regione geografica mostrino una correlazione o dipendenza significativa su scala globale. L'indice presenta una formalizzazione simile a quella coefficiente di correlazione lineare di Pearson tra due variabili. La differenza cruciale è che la correlazione non è calcolata tra variabili diversi, ma fa riferimento ad una stessa variabile osservata in punti diversi nello spazio. La dimensione spaziale è inclusa tramite una matrice dei pesi spaziali ($W$). La statistica I di Moran assume la forma descritta nell'equazione seguente:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.1}$$

dove:

**n** rappresenta il numero totale di osservazioni (unità spaziali);

$x_i$ è il valore della variabile di interesse (x) assunto dalla i-esima unità spaziale;

$x_j$ è il valore della variabile di interesse (x) assunto dalla j-esima unità spaziale;

$\bar{x}$ è la media aritmetica dei valori assunti dalla variabile x su tutte le unità spaziali;

$w_{ij}$ è il generico elemento della matrice dei pesi spaziali che riflette la relazione di vicinato/prossimità tra le unità i e j;

La somma dei pesi spaziali, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, è solitamente indicata con il termine $S_0$. Esprime un fattore di scala che permette di confrontare le diverse statistiche tra di loro, e questo indipendentemente dal fatto che la matrice dei pesi sia standardizzata o meno. Quando la matrice dei pesi spaziali è standardizzata per riga, questo elemento è semplicemente uguale alla numerosità totale ($S_0 = n$) e il primo termine a destra dell'equazione 3.1 è, pertanto, uguale all'unità.

In genere l'indice di Moran assume valori compresi tra -1 e 1. In particolare:

- valori positivi dell'indice di Moran denotano autocorrelazione spaziale positiva, tanto più forte man mano che l'indice si avvicina al valore +1.
- valori negativi dell'indice suggeriscono la presenza di autocorrelazione spaziale negativa, tanto più forte al tendere dell'indice tende -1.
- valori prossimi a 0 indicano l'assenza di autocorrelazione spaziale nelle osservazioni considerate.

La statistica-test I di Moran standardizzata si distribuisce asintoticamente secondo una Normale con valore atteso:

$$E_N(I) = -\frac{1}{n-1} \tag{3.2}$$

All'aumentare del numero di osservazioni e, quindi, di aree interessate, $E(I)$ tende a zero.

L'indice I di Moran può essere utilizzato per svolgere test di ipotesi finalizzato a determinare se l'autocorrelazione spaziale osservata in un set di dati è statisticamente significativa o se potrebbe essere il risultato del caso.

$$H_0 : I = 0$$
$$H_1 : I \neq 0$$

La statistica test di riferimento per n sufficientemente grande si distribuisce secondo una normale standardizzata:

$$Z(I) = \frac{I - E_N(I)}{\sqrt{Var_N(I)}} \tag{3.3}$$

La regola decisionale suggerisce di rifiutare l'ipotesi nulla se la statistica z, in valore assoluto, è maggiore del valore critico a una soglia $\alpha$ precedentemente definita. Le regole decisionali, in funzione dell'ipotesi alternativa,

suggeriscono di rifiutare l'ipotesi nulla a favore della presenza di autocorrelazione spaziale negativa se il valore di t è negativo e sufficientemente lontano da zero (o inferiore al valore critico), mentre l'autocorrelazione spaziale è positiva se il valore di t è positivo e sufficientemente lontano da zero (o superiore al valore critico). Il rifiuto dell'ipotesi nulla $H_0$ in favore di quella alternativa $H_1$ comporta che l'autocorrelazione spaziale nel set di dati è statisticamente significativa, indicando la presenza di una struttura spaziale reale. In termini dell'indice di Moran, il valore di Moran's I è diverso da quello che ci si attenderebbe sotto l'ipotesi di casualità. La validità del test si basa sull'ipotesi di normalità sotto l'ipotesi nulla di assenza di autocorrelazione spaziale. È possibile che questa ipotesi di normalità non sia rispettata. In questo caso, è preferibile cambiare l'approccio del test statistico e optare per un approccio di permutazioni.

Per rappresentare graficamente l'autocorrelazione spaziale è utile associare all'indice di Moran il **Moran Scatter Plot** (o diagramma di dispersione di Moran). Il Moran Scatter Plot è un grafico cartesiano che riporta sull'asse delle ascisse i valori della variabile di interesse (x) normalizzata, mentre sull'asse delle ordinate vengono rappresentati i relativi ritardi spaziali (Wx) anch'essi normalizzati. L'indice I di Moran è il coefficiente angolare della relazione lineare tra le due variabili riportate sugli assi del Moran scatterplot.

Se i punti sono dispersi fra i quattro quadranti questo indicherà assenza di autocorrelazione (il coefficiente angolare è zero e I=0). Se, invece, esiste una chiara relazione, il Moran Scatterplot potrà essere utilizzato per distinguere diverse tipologie di autocorrelazione spaziale. Il Moran Scatter Plot è infatti diviso in quattro quadranti distinti, ognuno dei quali suggerisce diverse forme di autocorrelazione spaziale:

**Nord-Est** : In questo quadrante (superiore destro) le unità geografiche con valori alti tendono ad essere vicine ad altre unità geografiche con valori alti. Questa disposizione suggerisce autocorrelazione positiva, indicando la presenza di cluster di valori simili nello spazio. La relazione è di tipo ALTO-ALTO.

**Nord-Ovest** : In questo quadrante (superiore sinistro) le unità geografiche con valori alti tendono ad essere vicine ad altre unità geografiche con valori bassi suggerendo la presenza di autocorrelazione negativa. La relazione è di tipo ALTO-BASSO.

**Sud-Est** : Qui (quadrante inferiore destro), le unità geografiche con valori bassi tendono ad essere vicine ad altre unità geografiche con valori alti.

Questo suggerisce una forma di autocorrelazione negativa, in cui vi è una relazione di tipo BASSO-ALTO.

**Sud-Ovest** : In questo quadrante (inferiore sinistro) le unità geografiche con valori bassi tendono ad essere vicine ad unità geografiche con valori altrettanto bassi. Questa disposizione suggerisce autocorrelazione positiva, indicando la presenza di cluster di aree caratterizzati da valori simili (e bassi) della variabile di interesse. La relazione è di tipo BASSO-BASSO.

Lo Scatterplot di Moran ha anche l'importante funzione di mettere in evidenza i possibili casi limite (outliers) perché possano essere eventualmente esclusi dalla analisi.

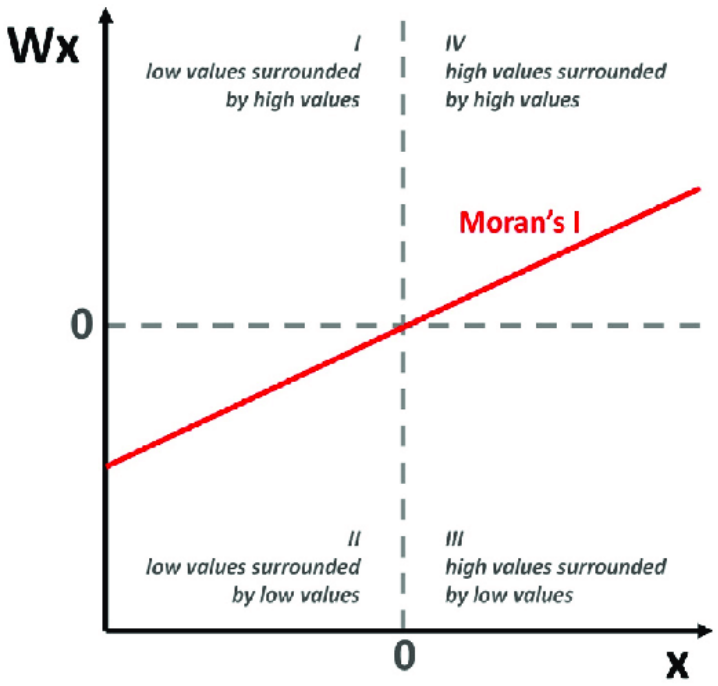


Figura 3.1: Moran Scatter Plot

Riportati su rappresemtazione cartografica (mappa) i risultati del Moran Scatterplot è possibile distinguere le aree con diverse tipi di relazione e, quindi, con diverse forme di autocorrelazione spaziale e verificare se le aree accomunate dallo stesso tipo di correlazione sono contigue e formano, dunque, dei cluster spaziali.

### 3.1.2 Geary's C

Una misulta di autocorrelazione spaziale alternativa all'indice I di Moran è l'indice C di Geary. A differenza dell' I di Moran, la cui formulazione si basa sulla somma del prodotto degli scarti dalla media, la C di Geary si basa sulla somma delle differenze al quadrato tra coppie di dati della variabile x tra tutte le aree in esame come misura della covariazione. A differenza di I enfatizza le differenze in valore tra aree e non la co-variabilità rispetto al valore medio. Inoltre, I è più stabile "globalmente", mentre C è molto più sensibile alle differenze in piccoli intorni di aree

$$C = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.4}$$

C assume valori compresi tra 0 e 2 ed è centrata su 1, pertanto:

- valori di C compresi tra 0 e 1 denotano autocorrelazione spaziale positiva, tanto più forte man mano che l'indice si avvicina al valore 0.
- valori di C compresi tra 1 e 2 dell'indice suggeriscono la presenza di autocorrelazione spaziale negativa, tanto più forte al tendere dell'indice a 2.
- valori prossimi a 1 indicano l'assenza di autocorrelazione spaziale nelle osservazioni considerate.

La statistica-test C di Geary standardizzata si distribuisce asintoticamente secondo una normale con valore atteso:

$$E_N(C) = 1 \tag{3.5}$$

Anche l'indice G di Geary può essere utilizzato per sottoporre a verifica l'ipotesi di presenza di autocorrelazione spaziale attraverso il seguente sistema di ipotesi:

$$H_0 : C = 0$$
$$H_1 : C \neq 0$$

La significatività dei risultati ottenuti può essere testata confrontando la distribuzione empirica e la distribuzione teorica per mezzo della seguente statistica test che segue una distribuzione normale standardizzata:

$$Z(C) = \frac{I - E_N(C)}{\sqrt{Var_N(C)}} \tag{3.6}$$

Come accade nel caso del test di ipotesi costruito per l'indice I di Moran, il rifiuto dell'ipotesi nulla $H_0$ in favore dell'ipotesi alternativa $H_1$ indica che l'autocorrelazione spaziale nel set di dati è statisticamente significativa. Questo suggerisce che esiste una struttura spaziale reale all'interno dei dati, ovvero che le osservazioni non sono distribuite casualmente nello spazio.

## 3.2 Misure di autocorrelazione locale

Spesso si rileva utile associare all'indice di autocorrelazione globale un indice di autocorrelazione locale in grado di misurare l'interdipendenza per ognuna delle aree in esame. Le statistiche per la misurazione del grado di autocorrelazione spaziale a livello locale consentono di individuare il contributo di ciascuna area rispetto al comportamento globale della distribuzione dando, in tal modo, la possibilità distudiare le correlazioni non solo tra i territori, ma anche all'interno di essi. Focalizzando l'attenzione su ogni singola area, gli indicatori locali possono essere impiegati nell'individuazione di forme di clusterizzazione attorno a punti specifici,risultando utili quando le statistiche "globali" non possono essere utilizzate per tali scopi. Similmente a quella globale, si parla di autocorrelazione spaziale locale positiva (o attrazione) quando, riferendosi ad un intorno e non più alla globalità dell'area di studio, valori simili di una variabile tendono a raggrupparsi in prossimità l'uno dell'altro; viceversa, si parla di autocorrelazione spaziale locale negativa (o repulsione) quando valori simili di una variabile tendono ad essere dispersi sul territorio.

Si tratta di statistiche che permettono di ricavare informazioni utili sulle proprietà locali, che dipendono dalle caratteristiche della zona, piuttosto che dall'andamento generale della distribuzione. Queste statistiche si basano sul confronto tra le proprietà delle distanze tra punti vicini di una distribuzione teorica con quelli trovati in una distribuzione osservata. In generale, un indicatore locale di associazione spaziale (Local Indicator of Spatial Association o LISA) è una qualsiasi statistica che permette di descrivere il grado di somiglianza, o differenza, di ciascun evento rispetto agli eventi più prossimi, ed è caratterizzata da due proprietà:

- ad ogni unità territoriale nell'area di studio è associata a una misura del livello di autocorrelazione spaziale rispetto al suo intorno immediato. Ciò consente di identificare come ciascuna area specifica interagisce con le aree circostanti.

- la somma degli indicatori LISA per ciascuna area è proporzionale all'indice globale di autocorrelazione spaziale.

L'indice LISA più comunemente utilizzato è rappresentato dalla versione locale della statistica I di Moran. La formulazione per l'i-esima unità è la seguente:

$$I_i = n\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{j=1;j\neq i}^{n} w_{ij}(x_j - \bar{x}) \tag{3.7}$$

$I_i$ si distribuisce asintoticamente come una normale con media:

$$E(I_i) = \frac{-\sum_{j=1}^{n} w_{ij}}{n-1} \tag{3.8}$$

# Capitolo 4

# Modelli Spaziali

Nel contesto dei dati spaziali, le ipotesi dei minimi quadrati ordinari (OLS) potrebbero non essere realistiche, ovvero i residui non correlati potrebbero non essere più validi a causa delle potenziali relazioni e interazioni tra le unità spaziali (Fisher e Wang, 2011; Lesage e Pace, 2009; 2010). A causa di queste caratteristiche specifiche dei dati spaziali, le stime OLS possono essere distorte o incoerenti e il modello può soffrire di errata specificazione. Pertanto, è essenziale modellare in modo appropriato i diversi tipi di effetti di interazione spaziale che possono spiegare le relazioni tra le unità spaziali, ovvero perché i valori osservati in un'area dipendono dai valori nelle località vicine (Plummer, 2010).

La classificazione dei modelli di econometria spaziale si basa su tre fondamentali tipi di interazioni spaziali:

- `Interazione Endogena`: tale tipo di interazione spaziale riflette le situazioni in cui le decisioni un'unità spazialmente identificata, come un agente o, più in generale, un'area geografica, sono influenzati direttamente dalle decisioni delle unità spaziali circostanti. Questo fenomeno riflette una dipendenza spaziale intrinseca tra le unità e implica che le azioni di una specifica unità possono avere un impatto significativo su quanto si osserva nelle unità adiacenti, e viceversa.

- `Interazione Esogena`: questo tipo di interazione si osserva quando le decisioni di un'unità spaziale sono influenzate dalle caratteristiche osservabili delle unità circostanti.

- `Interazione tra i termini di errore`: si verifica quando le unità spaziali vicine hanno errori sistematici simili che influenzano la variabile dipendente. Tale effetto interazione può derivare, ad esempio, da variabili latenti non osservabili che sono correlate spazialmente.

L'integrazione di tali interazioni spaziali, sia considerate singolarmente che congiuntamente, consente di derivare, a partire dal modello di regressione lineare (modello OLS [1]), formulazioni più complesse in grado di catturare la struttura spaziale intrinseca nei dati.

Per selezionare il modello che meglio si adatta al fenomeno oggetto di studio, in letteratura sono state proposte due strategie distinte.

- `Specific-to-General`: L'approccio classico, noto come "specific-to-general," considera come punto di partenza la stima di un modello di regressione lineare non spaziale (OLS) e successivamente, sulla base di test Lagrange Multiplier (LM) e Likelihood Ratio (LR), valuta se vi sia la necessità di includere uno o più effetti di interazione spaziale. In altre parole, si parte da un modello relativamente semplice e si esamina se l'introduzione di effetti di interazione spaziale migliora significativamente la capacità di spiegare il fenomeno in esame.

- `General-to-Specific`: L'approccio "general-to-specific," al contrario, valuta inizialmente un modello più generale e complesso, considerando tutte le possibili interazioni spaziali. Successivamente vengono imposti vincoli/rimossi i tipi di interazione spaziale che non risultano essere significativi. In questo modo, si semplifica il modello iniziale fino a ottenere una specificazione più parsimoniosa che mantenga solo le interazioni spaziali rilevanti.

In notazione matriciale il modello OLS è scritto come:

$$Y = \alpha\iota + X\beta + \varepsilon \tag{4.1}$$

dove:

$\mathbf{Y}$ è il vettore $N \times 1$ della variabile dipendente (N è il numero di osservazioni);

$\mathbf{X}$ è la matrice $N \times K$ delle variabili indipendenti (K è il numero di variabili indipendenti);

$\beta$ è il vettore $K \times 1$ dei parametri associati alle $K$ variabili esplicative;

$\varepsilon$ è il vettore $N \times 1$ dei termini di errore i.i.d. (indipendenti ed identicamente distribuiti) con media zero e varianza costante $\sigma^2$

Di seguito saranno presentati i modelli econometrici spaziali seguendo l'approccio specific-to-general, quindi partendo dal modello OLS.

[1] l termine "OLS" fa riferimento al metodo utilizzato per stimare i parametri di un modello di regressione lineare, ovvero il metodo dei minimi quadrati ordinari (OLS, Ordinary Least Squares)

## 4.1 Spatial Lag Model

Il modello Spatial Lag (conosciuto anche come modello SAR – Spatial AutoRegressive model) è un'estensione dell'OLS idoneo a catturare l'effetto di interazione spaziale endogena. Questo equivale a considerare la media dei valori delle unità "vicine"(c.d. ritardo o lag spaziale) come variabile addizionale nella regressione. L'introduzione della variabile dipendente ritardata spazialmente (indicata con $WY$) ha lo scopo di misurare l'effetto contagio (spillover) determinato dai valori che la variabile Y assume nelle unità statistiche localizzate nelle aree vicine (Cuffaro et al. 1999).

La forma del modello Spatial Lag può essere espressa come segue:

$$Y = \alpha\iota + \rho WY + X\beta + \varepsilon \tag{4.2}$$

dove:

**W** è la matrice $N \times N$ dei pesi spaziali che riflette la relazione di vicinato/prossimità tra le unità spaziali;

**WY** il termine $WY$ rappresenta il *ritardo spaziale* della variabile dipendente $Y$ e riflette l'effetto di interazione endogena;

$\rho$ è il parametro, chiamato coefficiente spaziale autoregressivo, che riflette l'intensità dell'interazione spaziale endogena, ovvero della relazione tra $Y$ e il suo lag spaziale.

In generale, il parametro $\rho$ è definito nell'intervallo $(w_{min}^{-1}; w_{max}^{-1})$ dove $w_{min}$ e $w_{max}$ rappresentano gli autovalori minimi e massimi della matrice $W$. Tuttavia, nella pratica, il suo dominio è spesso ristretto per ragioni computazionali e interpretative ad un intervallo specifico. In particolare, se la matrice $W$ è standardizzata per riga[2] il parametro $\rho$ generalmente assume valori compresi tra $[-1; 1]$.

Si distinguono i seguenti casi:

- $\rho > 0$: indica una dipendenza spaziale positiva. Significa che un aumento nei valori di una variabile in un'unità spaziale tende ad essere associato a un aumento nei valori delle unità spaziali vicine. In altre parole, le unità geografiche tendono a influenzarsi positivamente a vicenda. Questo può riflettere situazioni in cui le aree vicine si supportano reciprocamente.

[2]La standardizzazione per riga della matrice dei pesi spaziali comporta che la somma dei pesi per ogni unità spaziale sia uguale a 1.

- $\rho < 0$: indica una dipendenza spaziale negativa. Significa che un aumento nei valori di una variabile in un'unità spaziale tende ad essere associato a una diminuzione nei valori delle unità spaziali vicine. In altre parole, le unità geografiche tendono a influenzarsi negativamente a vicenda. Questo può riflettere situazioni in cui le aree vicine sono in competizione diretta.

- $\rho = 0$: indica l'assenza di effetto interazione endogena e il modello Spatial Lag si riduce ad un modello di OLS.

### 4.1.1 Data generating process

Nel contesto dei modelli spaziali, in particolare di quelli che considerano l'effetto interazione endogeno, è necessario distringuere il modello strutturale e il modello in forma ridotta o processo generativo dei dati.

- L'equazione in forma strutturale di un modello rappresenta la relazione teorica tra le variabili endogene e le variabili esogene in un modello. In altre parole, la forma strutturale descrive il "modello comportamentale" o il legame concettuale tra le variabili.Questa forma mette in relazione tutte le variabili coinvolte senza esplicitare il modello rispetto alla componente endogena.

- L'equazione in forma ridotta di un modello, invece, considera la variabile endogena espressa in funzione delle sole variabili esogene.

Prendiamo come esempio l'equazione 4.2. Scritta in questa forma rappresenta il modello strutturale per un modello Spatial Lag Model, ovvero offre una visione teorica e concettuale delle relazioni tra variabili coinvolte. Tuttavia, è utile pensare al processo di generazione dei dati sottostante e considerare l'equazione il forma ridotta del modello. Risolvendo la 4.2 per la variabile endogena $Y$, si ottiene l'equazione in forma ridotta per il modello Spatial Lag:

$$Y = (I - \rho W)^{-1}(\alpha\iota + X\beta) + (I - \rho W)^{-1}\varepsilon \tag{4.3}$$

## 4.2 Spatial Error Model

Il modello Spatial Error (conosciuto anche come modello SEM) è un'estensione dell'OLS idoneo a catturare l'effetto di interazione spaziale tra i termini di errore. In questo caso l' autocorrelazione spaziale non entra come variabile addizionale, ma influisce sulla struttura della covarianza dei disturbi casuali. Pertanto, la matrice W è inserita esclusivamente nel termine di errore.

La forma del modello Spatial Error può essere espressa come segue:

$$\begin{aligned} Y &= \alpha\iota + X\beta + u, \\ u &= \lambda W u + \varepsilon \end{aligned} \tag{4.4}$$

dove:

**u** è il vettore $N \times N$ dei termini di errore autocorrelati;

**Wu** il termine $Wu$ riflette l'effetto di interazione tra i termini di errore;

$\lambda$ è il parametro, chiamato coefficiente di autocorrelazione spaziale, che riflette l'intensità dell'interazione spaziale tra i termini di errore, ovvero dell'autocorrelazione spaziale tra i residui.

Così come accade per il coefficiente spaziale autoregressivo $\rho$, se la matrice $W$ è standardizzata per riga, anche il parametro $\lambda$ generalmente assume valori nell'intervallo $[-1; 1]$.

Si distinguono i seguenti casi:

- $\lambda > 0$: indica una dipendenza spaziale positiva nei termini di errore. Questo significa che le unità geografiche vicine tendono ad avere residui simili o correlati positivamente. In altre parole, se un'area ha un residuo alto, le aree circostanti tendono a presentare residui altrettanto alti.

- $\lambda < 0$: i indica una dipendenza spaziale negativa nei termini di errore. In altre parole, se un'area ha un residuo alto, le aree circostanti tendono a presentare residui bassi e viceversa.

- $\lambda = 0$: indica l'assenza di dipendenza spaziale tra i termini di errore e il modello Spatial Error equivale al modello OLS.

Come affermato da Anselin e Bera (1998), gli effetti di interazione spaziale tra i termini di errore possono essere interpretati come un disturbo (e il parametro $\lambda$ come un parametro di disturbo) nel senso che riflette l'autocorrelazione spaziale negli errori di misurazione o nelle variabili omesse (cioè, le variabili "ignorate" che si propagano attraverso le unità spaziali delle osservazioni). In altre parole, caratterizzano situazioni in cui le determinanti della variabile dipendente omesse dal modello sono spazialmente autocorrelate, o con una situazione in cui gli shock non osservati seguono un modello spaziale.

## 4.3 Lagrange Multiplier Tests

Nell'ambito dell'econometria spaziale, i test Lagrange Multiplier (LM) rivestono un ruolo fondamentale nella verifica della presenza di dipendenza spaziale nei dati. Questi test sono utilizzati per determinare se le interazioni spaziali, in particolare quella endogena e quella tra i termini di errore, sono significative e se esiste una struttura spaziale non spiegata nei modelli OLS. Si distinguono:

- Lagrange Multiplier test per il ritardo spaziale, ovvero per l'interazione spaziale endogena $LM_{lag}$
- Lagrange Multiplier test per la correlazione spaziale dei termini di errore $LM_{err}$

### 4.3.1 LM test for spatial lag dependence

Attraverso il test $LM_{lag}$ si esamina se le unità geografiche vicine hanno un impatto significativo sulla variabile dipendente osservata nell'unità spaziale di interesse. In particolare, si sottopone a verifica la significatività del parametro $\rho$, testando così la presenza di un effetto di interazione spaziale endogeno, ovvero di autocorrelazione spaziale nella variabile dipendente.

Il sistema di ipotesti del test $LM_{lag}$ è:

$$H_0 = \rho = 0$$
$$H_1 = \rho \neq 0$$

La statistica test utilizzata segue una distribuzione chi-quadrato con 1 grado di libertà $\sim \chi^2(1)$ e assume la seguene formulazione:

$$LM_{lag} = \frac{1}{H}\left(\frac{e'Wy}{e'eN^{-1}}\right) \tag{4.5}$$

dove $H = (WX\hat{\beta})'[I - X(X'X)^{-1}X'](WX\hat{\beta})\hat{\sigma}^{-2} + tr(W'W + W^2)$

Ad esempio se si considera un livello di significatività del 95%, il valore critico del test è pari a $\chi^2_{0.05;1} = 3,84$. Pertanto, si rifiuta l'ipotesi nulla se $LM_{lag} > 3,84$ e si conclude che l'autocorrelazione spaziale è presente nella variabile dipendente. Pertando, bisogna tenerne conto incorporando l'effendo interazione endogeno nel modello e procedere alla stima di un modello Spatial Lag.

### 4.3.2 LM test for spatial error dependence

Il test $LM_{err}$ esamina la presenza di dipendenza spaziale nei residui del modello. Verifica se ci sono errori residui correlati spazialmente, il che significa che gli errori residui in un'unità geografica sono influenzati dalle unità geografiche circostanti.

Si sottopone a verifica la significatività del parametro $\lambda$ testando così la presenza dell'effetto interazione tra i termini di errore.

Il sistema di ipotesti del test $LM_{err}$ è:

$$H_0 = \lambda = 0$$
$$H_1 = \lambda \neq 0$$

La statistica test di riferimento, così come la statistica test $LM_{lag}$, segue una distribuzione chi-quadrato con 1 grado di libertà $\sim \chi^2(1)$ e assume la forma:

$$LM_{err} = \frac{1}{C}\left(\frac{e'We}{e'eN^{-1}}\right)^2 \tag{4.6}$$

dove $C = tr(W'W + W^2)$

Ad esempio se si considera un livello di significatività del 99%, il valore critico del test è pari a $\chi^2_{0.01;1} = 6,63$. Pertanto, si rifiuta l'ipotesi nulla se $LM_{err} > 6,63$ e si conclude che l'autocorrelazione spaziale è presente nei residui del modello OLS bisogna procedere alla stima del modello Spatial Error.

### 4.3.3 LM test e scelta del modello

I risultati dei test LM consentono di discriminare tra la specifica che potrebbe essere più adatta a modellare la dipendenza spaziale nei dati In particolare, il principio che guida la semplice scelta tra modello spatial lag e modello spatial error può essere così riassuto:

- $\rho = 0$ e $\lambda = 0 \rightarrow$ se si rifiuta $H_0$ del test $LM_{lag}$ e si rifiuta $H_0$ del test $LM_{err}$, il modello da preferire è il modello OLS senza effetti spaziali
- $\rho \neq 0$ e $\lambda = 0 \rightarrow$ se si rifiuta $H_0$ del test $LM_{lag}$ e non si rifiuta $H_0$ del test $LM_{err}$, il modello da preferire è il modello spatial lag (o SAR)

- $\rho = 0$ e $\lambda \neq 0 \rightarrow$ se non si rifiuta $H_0$ del test $LM_{lag}$ e si rifiuta $H_0$ del test $LM_{err}$, il modello da preferire è il modello spatial error (o SEM)

Tuttavia, ci possono essere situazioni in cui entrambi i test LM sono significativi, ovvero $\rho \neq 0$ e $\lambda \neq 0$. In questo caso, è possibile:

- Confrontare i valori dei due test. Se il valore della statistica test del test $LM_{lag}$ è ¿ al valore della statistica test $LM_{err}$, può essere preferibile considerare il modello SAR. Viceversa, se il il valore della statistica test del test $LM_{err}$ è ¿ al valore della statistica test $LM_{lag}$, il modello SEM potrebbe essere più appropriato.

- Considerare le versioni robuste dei test $RLM_{lag}$ e $RLM_{err}$

- Considerare la possibilità di stimare un modello autoregressivo generale (Spatial Autoregressive Combined model – SAC) che tenga conto di ulteriori complessità spaziali nei dati incorporando sia l'effetto interazione endogeno che quello tra i termini di errore.

## 4.4 Spatial Autoregressive Combined Model

Il modello SAC (Spatial Autoregressive Combined)[3] è utilizzato nel caso in cui entrambi i parametri $\rho$ e $\lambda$ risultano essere significativamente diversi da 0. Questo modello incorpora sia la variabile dipendente spazialmente ritardata ($WY$) sia il termine di errore spazialmente autocorrelato ($Wu$) e, pertanto, consente di modellarle simultaneamente l'effetto interazione endogeno e tra i termini di errore. Il modello SAC assume la seguente formulazione:

$$\begin{aligned} Y &= \alpha\iota + \rho WY + X\beta + u \\ u &= \lambda Mu + \varepsilon \end{aligned} \tag{4.7}$$

dove:

**W e M** sono le matrici dei pesi spaziali predefinite e non-stocastiche di dimensione $N \times N$. In molte applicazioni le due matrici sono uguali $W = M$.

Gli altri termini della 4.7 sono stati già introdotti e spiegati in riferimento ai modelli presentati nelle sezioni precedenti.

[3]Anche noto come modello SARAR o modello Kelejian-Prucha dal nome dei suoi principali fautori.

## 4.5 Altri modelli spaziali: i modelli con effetti interazione esogeni

Fino ad ora, sono stati discussi modelli che tengono conto della dipendenza spaziale tra le unità attraverso l'introduzione del ritardo spaziale della variabile dipendente, il termine di errore autocorrelato o entrambi questi elementi. Tuttavia, la dipendenza spaziale può essere introdotta in un modello anche attraverso il ritardo spaziale dei regressori. Tale approccio si adatta a situazioni in cui si ritiene che siano le variabili indipendenti ad essere correlate spazialmente, ovvero che i valori dei regressori di una data unità siano influenzati dai valori che gli stessi regressori assumono nelle unità circostanti. Considerare il ritardo spaziale dei singoli regressori tra le esplicative del modello equivale a modellare l'effetto di interazione spaziale esogeno. Di seguito saranno presentati tre modelli per dati spaziali che considerano l'effetto interazione esogeno singolarmente, in combinazione con l'effetto interazione endogeno o con l'effetto interazione tra i termini di errore.

### 4.5.1 Spatially Lagged-X Model

Il pù semplice modello spaziale, noto come *spatially lagged X model* (o SLX) estende l'OLS incorporando il ritardo spaziale dei singoli regressori (LeSage e Pace 2009). Ciò deriva dall'assunzione che il valore della variabile dipendente in un'area specifica (i) sia soggetto all'influenza non solo delle caratteristiche osservate nell'area stessa, ma anche di quelle riscontrate nelle unità spaziali prossime. In altri termini, il modello SLX riconosce che le dinamiche della variabile in esame non sono determinate solamente dalle sue specifiche condizioni locali, ma sono interconnesse alle condizioni delle aree circostanti.

Il modello SLX assume la seguente formulazione:

$$Y = \alpha\iota + X\beta + WX\theta + \varepsilon \tag{4.8}$$

dove:

**WX** rappresenta il *ritardo spaziale* delle $K$ variabili esplicative $X$ e riflette l'effetto di interazione esogeno;

$\theta$ è il vettore $K \times 1$ dei parametri associati ai ritardi spaziali delle K variabili esplicative. Tali parametri misurano l'impatto marginale sulla variabile dipendente Y dell'i-esima unità spaziale esercitato da ciascuna variabile esplicativa osservate nelle aree vicine all'area i.

Gli altri termini della 4.8 sono stati già introdotti e spiegati in riferimento ai modelli presentati nelle sezioni precedenti.

### 4.5.2 Spatial Durbin Model

Il modello Spatial Durbin (SDM) è un modello spaziale autoregressivo che, oltre al ritardo spaziale della variabile dipendente, include il ritardo spaziale dei regressori. Di seguito sono riportate sia forma strutturale del modello SDM (4.9) che la relativa forma ridotta o processo generativo dei dati (4.10).

$$Y = \alpha\iota + \rho WY + X\beta + WX\theta + \varepsilon \tag{4.9}$$

$$Y = (I - \rho W)^{-1}(\alpha\iota + X\beta + WX\theta + \varepsilon) \tag{4.10}$$

Tutti i termini della 4.9 sono stati già introdotti e spiegati in riferimento ai modelli presentati nelle sezioni precedenti.

Il modello SDM ricopre un ruolo importante nell'ambito dei modelli spaziali sia per le sue importanti proprietà, che saranno approfondite successivamente, sia perchè attraverso l'imposizione di restrizioni ai parametri è possibile ottenere tutte e tre le specifiche spaziali più semplici (ovvero quelle che includono una sola forma di interazione spaziale: SAR, SEM e SLX).

- imponendo la restrizione $\rho = 0$ il modello SDM risulta in un modello SLX;
- vincolando $\theta = 0$ il modello SDM risulta in un modello SAR;
- infine, imponendo $\theta = -\rho\beta$ si dimostra che si ottiene un modello SEM.

### 4.5.3 Spatial Durbin Error Model

Il modello Spatial Durbin Error (SDEM) è sviluppato a partire dai modelli SEM e SLX, dovuto all'interazione spaziale tra gli errori e alle interazioni spaziali esogene e assume la forma:

$$\begin{aligned} Y &= \alpha\iota + X\beta + WX\theta + u, \\ u &= \lambda Wu + \varepsilon \end{aligned} \tag{4.11}$$

Tutti i termini della 4.11 sono stati già introdotti e spiegati in riferimento ai modelli presentati nelle sezioni precedenti.

## 4.6 Likelihood Ratio Test

A seguito di questa panoramica dei modelli spaziali è necessario introdurre uno dei principali criteri che possono essere utilizzati per identificare quello più adatto ai dati analizzare.