

STIMA DI INDICATORI DI DOTAZIONE INFRASTRUTTURALE SULLA BASE DI
MODELLI DI DISAGGREGAZIONE SPAZIALE

Claudio MAZZIOTTA¹, Francesco VIDOLI²

¹ Università degli Studi Roma Tre, Dipartimento di Istituzioni pubbliche, Economia e Società, via G. Chiabrera 199, 00145 Roma, c.mazziotta@uniroma3.it

² Società per gli Studi di Settore S.p.A., via M. Maggini 47, 00147 Roma, fvidoli@sose.it

SOMMARIO

Il paper ha per obiettivo l'approfondimento di una linea di ricerca precedentemente avviata, avente ad oggetto la predisposizione ed applicazione di modelli in grado di disaggregare sotto il profilo spaziale variabili disponibili ad un livello territoriale più aggregato. Il modello, improntato all'approccio di Chow-Lin, è applicato alla disaggregazione territoriale degli indicatori infrastrutturali sintetici, sulla base dell'ipotesi che le variabili di generazione delle infrastrutture siano in grado di spiegare la dotazione infrastrutturale sia a livello aggregato (regioni) che a livello disaggregato (province).

Due le direzioni di approfondimento perseguite in questo lavoro: da un lato, l'introduzione di variabili più puntuali riguardo alle caratteristiche fisiche ed ai livelli di concentrazione-urbanizzazione del territorio; d'altro lato, l'introduzione nel modello di variabili intervallate, al fine di ottenere una ulteriore verifica di robustezza dei risultati ottenuti.

Il primo approfondimento, pur ottenendo un miglioramento nell'adattamento complessivo del modello, conferma in buona sostanza che la distribuzione territoriale della dotazione infrastrutturale non è conforme ai fattori di generazione che teoricamente ne dovrebbero determinare i livelli nelle diverse aree del paese. Ad analoga conclusione si giunge a seguito del secondo approfondimento, che conferma la stabilità del modello stesso anche in presenza della modifica casuale dei dati di base.

1. INTRODUZIONE

In un precedente lavoro (Mazziotta e Vidoli, 2009, ripreso e modificato in Vidoli e Mazziotta, 2010) è stato fatto un tentativo di applicare alcuni modelli basati sull'approccio di Chow-Lin per giungere all'articolazione spaziale di indicatori di dotazione infrastrutturale: in particolare, a partire da indicatori *noti* ad un determinato livello territoriale (regioni, nel nostro caso), è stata applicata una procedura volta a derivare indicatori *incogniti* ad un livello territoriale più dettagliato (province, nel nostro caso). L'approccio modellistico adottato (richiamato al successivo paragrafo 2) si basa sull'utilizzazione di alcuni regressori (variabili di natura sia economica che demografica), ai quali viene attribuito un ruolo determinante quali fattori di generazione delle infrastrutture. In tal modo i risultati del modello, oltre a presentare interesse sotto l'aspetto più strettamente statistico, assumono il significato di una verifica di congruità tra offerta e domanda di infrastrutture ad un livello territorialmente disaggregato.

Preso atto che i risultati raggiunti nel citato lavoro indicano una sostanziale difformità tra la configurazione territoriale attesa (in termini di dotazione infrastrutturale) e quella risultante dal modello¹, nel presente lavoro si intende procedere lungo una duplice linea di approfondimento:

- i) in primo luogo, verificare l'opportunità di introdurre tra i regressori variabili che tengano conto in misura più puntuale delle specificità dei territori considerati (le province italiane), con particolare riferimento alle caratteristiche orografiche ed ai livelli di urbanizzazione;
- ii) in secondo luogo, estendere l'applicazione del modello a variabili intervallate, in modo da ottenere livelli minimi e massimi di stima della variabile dipendente (l'indicatore infrastrutturale a livello provinciale) e verificarne la congruità (e quindi la robustezza) rispetto alla stima base.

Il paper è così strutturato: nel paragrafo 2 si richiamano i tratti essenziali della metodologia di tipo Chow-Lin utilizzata per la stima del modello di disaggregazione spaziale; nel paragrafo 3 si illustrano i risultati ottenuti a seguito del tentativo di migliorare la significatività statistica ed economica delle variabili assunte quali regressori del modello; nel paragrafo 4 si illustrano i risultati ottenuti dalla particolare analisi di robustezza effettuata attraverso l'applicazione del modello a variabili intervallate; nel paragrafo 5 è riportato un breve commento conclusivo.

¹ In effetti, il confronto è effettuato tra gli indicatori di dotazione infrastrutturale risultanti dal modello di disaggregazione spaziale e gli analoghi indicatori disponibili (fonte ISTAT, 2006) al livello territoriale desiderato (province).

2. IL MODELLO DI DISAGGREGAZIONE SPAZIALE UTILIZZATO

Il modello utilizzato, costruito sulla scia dei lavori di Bollino e Polinori (2007) e di Polasek e Sellner (2008), ha per obiettivo di verificare se e in quale misura i metodi di disaggregazione spaziale permettano di ricostruire in maniera apprezzabilmente puntuale l'indice di dotazione infrastrutturale a livello disaggregato sulla base del corrispondente indicatore a livello territorialmente superiore. Nel caso specifico qui considerato ciò significa verificare se, una volta disponibili gli indicatori infrastrutturali *a livello regionale* e dopo aver constatato la significatività della relazione che lega tali indicatori alle variabili demografiche ed economiche che ne determinano la generazione, i corrispondenti indicatori infrastrutturali *a livello provinciale* possano anch'essi essere "spiegati" dai medesimi fattori di generazione, ossia di "domanda" delle infrastrutture stesse.

La verifica in questione è agevolata dalla disponibilità dei livelli provinciali di infrastrutturazione pubblicati dall'ISTAT (ISTAT, 2006): il confronto tra questi indicatori, assunti come "veri", e gli indicatori "stimati", ottenuti come risultato del modello di disaggregazione spaziale applicato, consente di registrare la conformità dell'offerta provinciale di infrastrutture ai corrispondenti fattori di domanda nel caso l'accostamento delle due serie di indicatori (*veri* e *stimati*) sia buono, oppure di constatarne la difformità nel caso l'accostamento sia limitato.

Il modello adottato (per il cui dettaglio si rinvia al citato Vidoli e Mazziotta, 2010) applica l'approccio di Chow-Lin (1971) alla dimensione spaziale (anziché a quella temporale, di più comune utilizzo), sulla base delle ipotesi di:

- i) *structural similarity*: il modello aggregato e quello disaggregato sono strutturalmente simili, il che comporta che i parametri di regressione siano gli stessi;
- ii) *error similarity*: gli errori spazialmente correlati presentano la stessa struttura sia a livello aggregato sia disaggregato;
- iii) *reliable indicators*: le variabili assunte quali regressori presentano un buon potere predittivo a livello sia aggregato che disaggregato.

Partendo da una relazione econometrica lineare del tipo:

$$y_d = X_d \beta_d + \varepsilon_d \quad [1]$$

dove:

y_d è un vettore ($N \times 1$) di osservazioni dell'indicatore sintetico infrastrutturale a livello disaggregato;

X_d è una matrice ($N \times k$) di osservazioni delle k variabili esplicative a livello disaggregato;

N è il numero di unità territoriali disaggregate (province, nel nostro caso),

e sfruttando le proprietà sopra richiamate, si giunge, sulla scia dei lavori di Bollino e Polinori (2007) e di Polasek e Sellner (2008), a stimare la variabile incognita (indicatore infrastrutturale) a livello provinciale nella forma seguente²:

$$\hat{y}_d = \underbrace{\hat{R}_N^{-1} X_d \hat{\beta}_a}_{1^\circ \text{termine}} + \underbrace{\hat{\Sigma}_d C' (C \hat{\Sigma}_d C')^{-1} (y_a - C \hat{R}_N^{-1} C' X_a \hat{\beta}_a)}_{2^\circ \text{termine}} \quad [2]$$

in cui:

C è una matrice di dimensione $(n * N)$, dove n è il numero delle regioni, capace di trasformare le osservazioni disaggregate in osservazioni aggregate. Il generico elemento di tale matrice, nel caso di trasformazione con operatore media aritmetica, è definibile come:

$$C_{i,j} = \begin{cases} \frac{1}{k}, & \text{se provincia } i \in \text{ regione } j, \text{ con } k = \text{numero di province } \in \text{ regione } j \\ 0, & \text{altrimenti} \end{cases}$$

$R_N = I - \rho_d W_N$ è una matrice che rielabora una matrice di pesi spaziali W_N ed un parametro *spatial lag* $\rho \in [0,1]$. In tal modo l'apporto delle X_d risulta filtrato attraverso la componente spaziale, che opera sostanzialmente in modo proporzionale alla distanza;

$y_a = \sum y_d$ rappresenta l'indicatore infrastrutturale a livello aggregato (regionale);

$\beta_d = \hat{\beta}_a$ secondo le ipotesi di *structural similarity*, per cui $y_a = X_a \beta_d + \epsilon_a$.

In sostanza, il primo termine della [2] rappresenta la stima *naïve* del vettore incognito y_d , mentre nella seconda parte dell'equazione l'errore di stima a livello aggregato viene distribuito attraverso la “*gain projection matrix*” (Goldberger, 1962).

$$G = \hat{\Sigma}_d C' (C \hat{\Sigma}_d C')^{-1} \quad [3]$$

Questa quantità dipende in modo cruciale dal parametro *spatial lag* ρ_a a livello aggregato; il parametro ρ_d e la matrice W_N permettono quindi che la $1/N$ parte del residuo a livello aggregato non venga assegnata in parte uguale a tutte le province, ma che venga filtrata attraverso la distanza spaziale tra le stesse.

3. APPLICAZIONI E RISULTATI DEL MODELLO

L'applicazione del modello richiede la disponibilità di informazioni relative a: i) indicatore sintetico infrastrutturale a livello regionale (y_a nell'equazione [2]); ii) variabili demografiche ed economiche correlate con i fabbisogni infrastrutturali, disponibili a livello provinciale e regionale (rispettivamente, X_d e X_a nell'equazione [2]).

Nell'applicazione precedentemente effettuata (Vidoli e Mazziotta, 2010) la variabile dipendente y_a era stata identificata nell'indicatore infrastrutturale sintetico rappresentativo dei

² Le variabili riferite al fenomeno aggregato sono caratterizzate dal pedice “a” (*aggregated*), mentre le variabili riferite al fenomeno disaggregato spazialmente sono caratterizzate dalla lettera “d” (*disaggregated*).

trasporti terrestri (i cui indicatori elementari sono forniti da ISTAT, 2006), costruito in un precedente lavoro attraverso il ricorso alla procedura Benefit of the Doubt – BoD (Mazziotta e Vidoli, 2009); mentre per le variabili indipendenti X la scelta era caduta, nell’ambito di un più ampio *set* iniziale, su tre variabili, ritenute *proxy* dello sviluppo economico (prodotto interno lordo), della densità demografica (quota di popolazione residente in comuni con più di 50 mila abitanti) e dell’offerta turistica (ricettività alberghiera), assunti quali fattori rilevanti di generazione “potenziale” della dotazione infrastrutturale.

Pur in presenza di risultati statisticamente soddisfacenti in ordine alla stima del modello applicato, nel presente lavoro si è ritenuto tuttavia opportuno tentare di approfondire l’esame delle variabili idonee a rappresentare i fattori di domanda potenziale della dotazione infrastrutturale. Tale approfondimento è volto, in particolare, ad acquisire variabili che introducano nel modello informazioni relative sia alle caratteristiche fisiche dei territori considerati, sia ai fenomeni di concentrazione e urbanizzazione, nella convinzione che questa duplice dimensione debba essere rappresentata in un modello che grande importanza assegna alla differenziazione spaziale. Naturalmente, l’esigenza di disporre di tali variabili a livello sia regionale che provinciale costituisce un vincolo non indifferente alla selezione delle variabili, dato che quelle teoricamente più idonee non sempre sono reperibili ai livelli di disaggregazione desiderati.

Dopo alcuni tentativi, operati nell’ambito del modello a livello regionale, la scelta è caduta sulle seguenti variabili, che, oltre ad essere disponibili ai due livelli territoriali richiesti (regionale e provinciale), presentano il duplice pregio di essere rappresentative del fenomeno che si intende quantificare e di consentire un buon accostamento statistico del modello:

- quale *proxy* del livello produttivo, si è preferito assumere la quota di valore aggiunto dei settori extragricoli dell’unità territoriale considerata rispetto al totale nazionale. Questa variabile sostituisce la precedente, più generica, del PIL;
- per la concentrazione (e, per estensione, per l’urbanizzazione) si è mantenuta la precedente variabile della quota di popolazione residente in comuni con più di 50 mila abitanti, ma ad essa è stata affiancata la variabile della densità demografica;
- per tener conto della dimensione fisica dei territori considerati – oltre che di quella antropizzata, cui si riferiscono le variabili sopra richiamate – sono state introdotte due variabili, precedentemente non considerate: l’estensione territoriale complessiva e una *proxy* della conformazione orografica, identificata nell’ammontare di popolazione residente in comuni montani.

Per la verità, si è anche tentato di introdurre una variabile rappresentativa del diverso livello di efficienza riscontrabile sul territorio con riferimento alla dotazione infrastrutturale; al riguardo, si è fatto ricorso ad una stima ricavata dall’accostamento tra livello di infrastrutturazione fisica e spesa complessiva per infrastrutture (Golden e Picci, 2005), che dovrebbe avere la capacità di individuare le situazioni di inefficienza (quando non di vera e

propria corruzione) e dunque consentire di introdurre nel modello un fattore correttivo del *mismatch* tra offerta e domanda di infrastrutture. Ma questo tentativo non ha dato risultati statisticamente apprezzabili e di conseguenza la variabile non è stata inclusa nel modello.

In sintesi, le variabili inserite nel modello sono le seguenti.

Sigla	Identificazione	Unità di misura
VA	Valore aggiunto dei settori extra-agricoli rispetto al totale nazionale	Quota percentuale
Resid. 50	Popolazione residente in comuni con più di 50 mila abitanti sul totale della popolazione	Quota percentuale
Sup. terr.	Superficie territoriale	kmq
Densità	Popolazione residente su superficie territoriale	Abitanti per kmq
Pop. montagna	Popolazione residente in comuni di montagna	Numero di abitanti

La fonte dei dati è sempre ISTAT, ad eccezione del valore aggiunto dei settori extra-agricoli a livello provinciale, di fonte Istituto Tagliacarne.

Le relazioni di ciascuna variabile indipendente con l'indicatore sintetico infrastrutturale (per brevità, definito con la sigla BoD) sono desumibili dalla Fig. 1, mentre i risultati della stima del modello a livello regionale sono riportati nella Tabella 1. Essi mostrano un buon accostamento complessivo, nonché un miglior adattamento delle variabili introdotte.

Figura 1: Variabili di regressione e relativi scatterplot

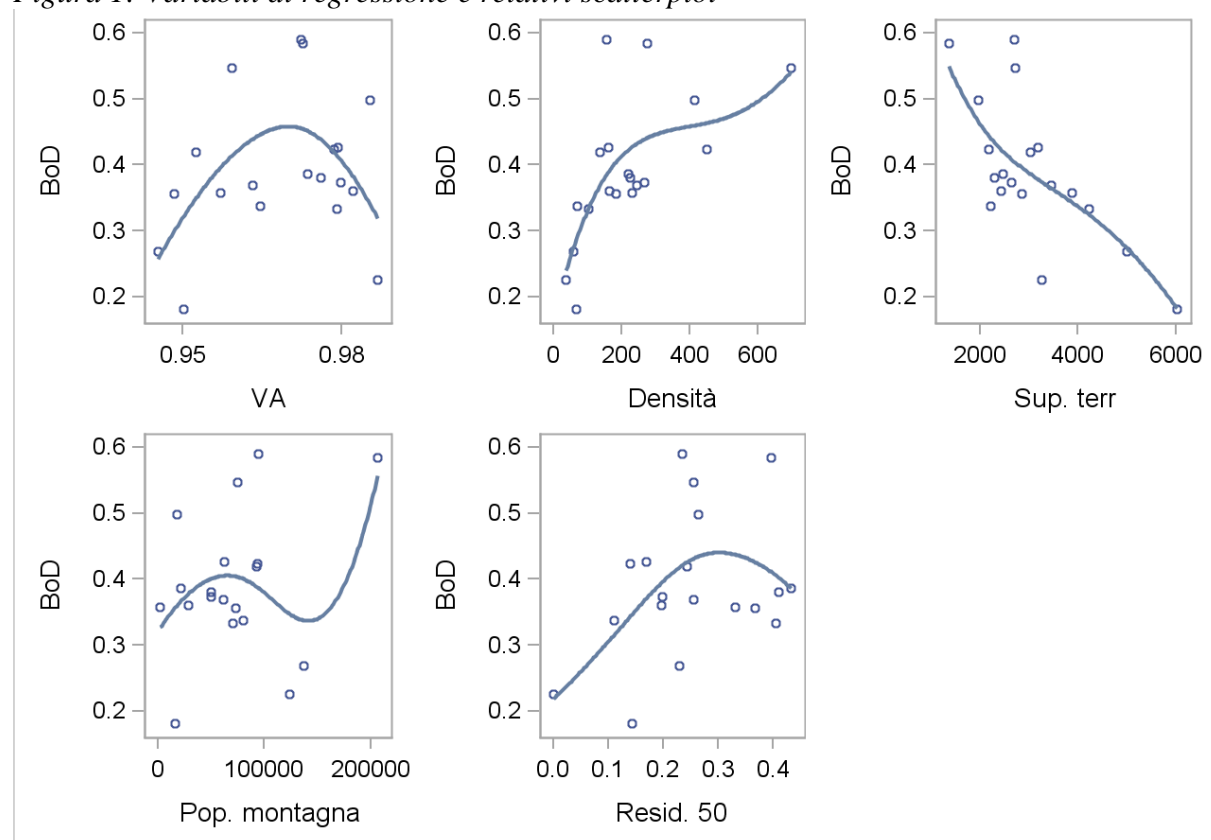


Tabella 1 – Risultati della stima del modello a livello regionale

Variabile	Stima dei parametri	Errore standard	Valore t	Pr > t	Stima standardizzata
Densità	0,00033253	5,86E-05	5,67	<,0001	0,25484
Pop. montagna	5,07E-07	1,94E-07	2,61	0,0215	0,11961
Resid. 50	0,26073	8,79E-02	2,96	0,011	0,17863
VA	0,31381	5,52E-02	5,68	<,0001	0,74477
Sup. terr	-0,00003293	9,73E-06	-3,39	0,0049	-0,26539

R-quadro 0,992.

In conformità della struttura del modello, ed in particolare grazie alle ipotesi di *structural similarity* e di *error similarity*, la soluzione a livello regionale permette di disporre dei parametri β_{α} e σ_{α}^2 , che, introdotti nel modello disaggregato, consentono di ottenere la stima dell'indicatore infrastrutturale sintetico (BoD) a livello provinciale. Il confronto dei dati così *stimati* con quelli *veri* (anch'essi di fonte ISTAT) consente, come già detto, di dare un giudizio sulla conformità della distribuzione territoriale delle infrastrutture (di trasporto terrestre) rispetto ai potenziali fattori di generazione.

Nonostante il miglioramento del modello grazie alle nuove variabili considerate, lo scostamento tra dati *stimati* (a livello provinciale) e dati *veri* risulta ancora di notevole entità, come si può verificare dalla Figura 2, in cui è evidente lo scostamento dei risultati del modello dai dati veri (scostamento dalla bisettrice), e dalla Figura 3, che riporta le differenze per ciascuna provincia tra i due livelli dell'indicatore (*vero* e *stimato*).

Figura 2: Distribuzione delle province secondo la relazione tra l'indicatore infrastrutturale vero (BoD score) e quello stimato (estimated BoD score) attraverso il modello di disaggregazione.

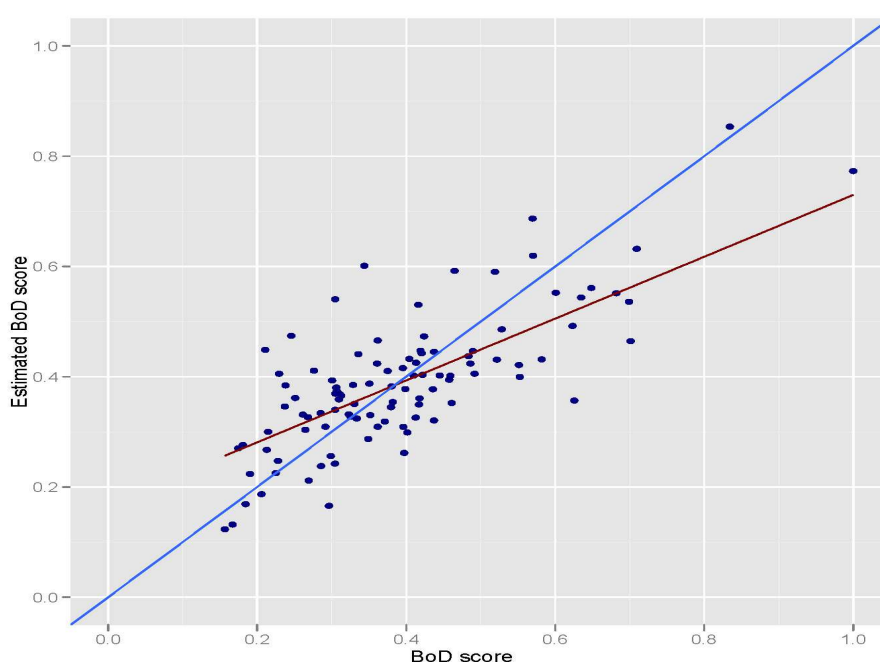
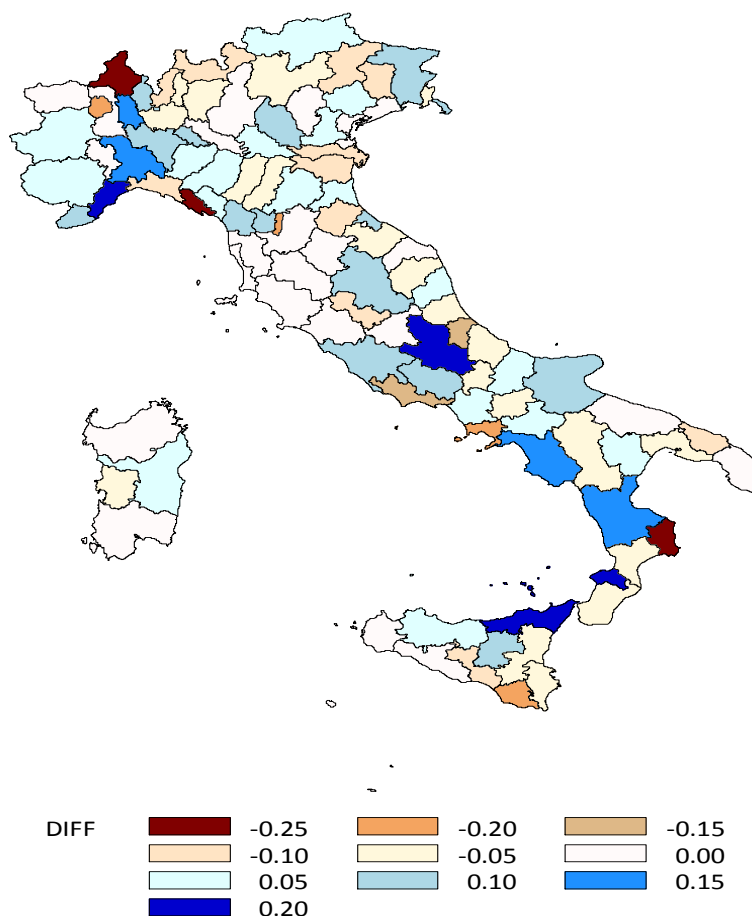


Figura 2: Differenze tra gli indicatori infrastrutturali veri e quelli stimati attraverso il modello, entrambi a livello di provincia.



Occorre peraltro considerare che tali differenze identificano scostamenti tra le due serie di variabili che, pur se rilevanti in valore assoluto – di conseguenza anche in termini di posizione nelle rispettive graduatorie (rango) –, non necessariamente comportano uno spostamento sostanziale in termini di appartenenza alla parte alta o bassa o intermedia del posizionamento complessivo delle province. Ciò appare evidente se si considerano, da un lato, i dati assoluti riportati nella tavola A.1 (in Appendice) e, dall'altro, l'attribuzione delle province a classi di appartenenza, rappresentate dai quintili della distribuzione (cfr. Figure A.1 e A.2 in Appendice). Un salto nella posizione in graduatoria non necessariamente si ripropone come salto (o, almeno, non della stessa intensità) tra le classi di appartenenza.

Si prenda, esemplificativamente, il caso di Messina, che presenta una differenza molto sensibile tra i livelli dei due indicatori (circa il 50%). Ciò fa slittare questa provincia dal 10° posto di una graduatoria al 40° dell'altra; in termini di classe di appartenenza, tuttavia, la differenza è assai meno sensibile, facendosi registrare lo slittamento di appena una classe tra una classifica e l'altra. A maggior ragione questo ragionamento vale nei casi in cui le differenze tra gli indicatori siano meno rilevanti.

In definitiva, si può ritenere che i risultati ottenuti attraverso il modello di disaggregazione utilizzato, anche nella versione migliorata sotto il profilo della rilevanza delle variabili indipendenti, confermano sostanzialmente che lo stock infrastrutturale relativo al settore dei trasporti (terrestri) non si distribuisce sul territorio in maniera conforme ai fabbisogni espressi dai fattori socio-economici presenti a livello locale. D'altra parte, l'esame del posizionamento delle province in termini di classi di appartenenza, anziché di rango in graduatoria, pur confermando le differenze tra le due serie di indicatori, ne rende meno drastica l'intensità effettiva sotto il profilo sostanziale.

4. MODELLI DI DISAGGREGAZIONE SPAZIALE IN PRESENZA DI DATI INTERVALLATI

4.1 Gli aspetti metodologici

La seconda linea di approfondimento del presente lavoro, come già accennato, è consistita in una ulteriore analisi di robustezza dei risultati ottenuti, perseguita attraverso l'introduzione nel modello in esame di dati intervallati e la conseguente verifica dello scostamento dei risultati ottenuti rispetto alla versione base di cui al precedente paragrafo 3.

Dal punto di vista metodologico, si può constatare che in una generica analisi di regressione i dati sono solitamente rappresentati come vettori quantitativi; tuttavia, non è raro nella pratica applicativa essere in presenza di intervalli di valori.

Gli intervalli di valori emergono in situazioni pratiche, come ad esempio la registrazione mensile dell'intervallo di temperature nelle stazioni meteorologiche, nei prezzi delle azioni finanziarie o, come nel nostro caso, in presenza di fenomeni rilevanti ad un livello territoriale molto disaggregato, stimati spesso in modo statisticamente poco affidabile³.

L'analisi di dati intervallati (o simbolici, da cui l'acronimo SDA, *Symbolic Data Analysis*, su cui si vedano Book e Diday, 2000 e Billard e Diday, 2002) è la famiglia di tecniche statistiche che permette di analizzare dati intervallati nel momento in cui i valori osservati appartengono all'insieme dei numeri reali \mathbb{R} .

Billard e Diday (2000) presentano il primo tentativo di stima di un modello di regressione lineare per un insieme di dati intervallati; il loro modello (CM, *Centre Method*) consiste nella stima di un modello di regressione lineare nei punti medi dei valori intervallati assunti dalle variabili per poi applicare tale modello al limite inferiore e superiore delle variabili esplicative al fine quindi di prevedere, rispettivamente, il limite inferiore e superiore della variabile dipendente.

³ Un'altra fonte di dati intervallati è l'aggregazione di grandi basi dati in un numero ridotto di gruppi, le cui proprietà sono descritte da variabili intervallate.

Il metodo *MinMax* (Billard e Diday 2002) suggerisce di stimare i limiti inferiori e superiori degli intervalli utilizzando diversi vettori di parametri, il che equivale a supporre l'indipendenza tra i valori dei limiti inferiori e superiori degli intervalli.

Lima Neto e De Carvalho (2008) sviluppano le metodologie precedentemente presentate suggerendo di utilizzare un nuovo approccio al problema chiamato CRM (*Centre and Range Method*), basato su due distinti modelli di regressione lineare; il primo stima le relazioni lineari nei punti medi degli intervalli, mentre il secondo sull'intervallo (*over the ranges*), in modo da ricostruire i limiti dell'intervallo di valori della variabile dipendente in modo più efficiente⁴.

La presentazione formale di questo metodo oltrepassa lo scopo di questo lavoro; ne presentiamo, comunque, le notazioni principali. Siano Y^r e X_j^r ($j=1, 2, \dots, p$), rispettivamente, le variabili quantitative che assumono il valore nella metà dell'intervallo assunto dalle variabili simboliche e Y^c ed X_j^c le variabili quantitative che assumono valore nel punto medio dell'intervallo.

Questo significa che ogni errore di stima ε_i può essere separato in due termini: la differenza tra due vettori ($Y^r, Y_{estimated}^r$) e tra ($Y^c, Y_{estimated}^c$) secondo le seguenti relazioni lineari:

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \dots + \beta_p^c x_{ip}^c + \varepsilon_i^c \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \dots + \beta_p^r x_{ip}^r + \varepsilon_i^r \end{aligned}$$

Questi approcci, per contro, risolvono il problema di stima con tecniche di ottimizzazione lineare non considerando gli aspetti probabilistici legati ai modelli di regressione. Questo rende impossibile dunque utilizzare le usuali tecniche di inferenza per le stime dei parametri o degli intervalli di confidenza.

I modelli lineari generalizzati rappresentano una sintesi importante dei modelli di regressione, consentendo una vasta gamma di tipi di dati risposta e di variabili esplicative. Questi modelli, basati sulla famiglia di distribuzioni esponenziali, rappresentano uno strumento molto importante di analisi grazie alla loro flessibilità e alla concreta applicabilità nella pratica. In particolare, Iwasaki e Tsubaki (2005) hanno introdotto una classe di modelli lineari generalizzati bivariati (*BGLMs*) basata sulla famiglia di distribuzioni esponenziali bivariate con una richiesta di analisi dei dati meteorologici; anche Lima Neto *et al.* (2009) hanno considerato i modelli BGLM come un importante strumento per risolvere problemi con dati intervallati presentando un modello basato sulla distribuzione gaussiana bivariata⁵.

⁴ Nella nostra applicazione abbiamo usato questo metodo grazie al codice *R* inviatoci dagli autori stessi.

⁵ Non è scopo di questo lavoro proporre un altro interessante approccio al problema di regressione in presenza di dati affetti da rumore o intervallati che presentino determinate caratteristiche geometriche (ad esempio quando le variabili possono essere visti come variabili *fuzzy*); questi metodi rientrano nella categoria della *ε -Support Vector Regression approach (ε -SVR)* (Carrizosa, Gordillo and Plastria, 2007) che risolve il problema di regressione attraverso tecniche *fuzzy* di ottimizzazione.

4.2 Applicazione del modello per dati intervallati alla ricostruzione degli indicatori infrastrutturali provinciali

Questa applicazione, ancora in fase di studio per la corretta fase di disaggregazione dell'errore di stima a livello aggregato⁶, nasce dall'ipotesi che le informazioni da utilizzare per stimare i livelli a livello disaggregato, specie quando da un punto di vista spaziale si scenda ad un dettaglio molto fine, siano affette da errore o comunque da un certo grado di approssimazione. Questo ci porta a proporre di introdurre tecniche statistiche che leghino in un primo tempo i livelli di un certo fenomeno y_d a variabili indipendenti X_d intervallate e che, in un secondo momento, permettano, sempre sotto le ipotesi base del modello di Chow-Lin, di stimare i livelli disaggregati minimi e massimi.

A partire dai dati a livello regionale considerati in Tabella 1, per simulare la presenza di dati approssimati abbiamo aggiunto un rumore casuale uniforme (vedi Figura A.3 in Appendice) sia sulle y_d che sulle X_d allo scopo di ottenere dei dati a livello provinciale (e quindi poi regionale) nella forma:

$$\begin{aligned} y_d^{N,low} &= y_d + Unif(-0.1, 0)y_d \\ y_d^{N,high} &= y_d + Unif(0, 0.1)y_d \end{aligned} \quad [4]$$

e

$$\begin{aligned} X_d^{N,low} &= X_d + Unif(-0.1, 0)X_d \\ X_d^{N,high} &= X_d + Unif(0, 0.1)X_d \end{aligned} \quad [5]$$

Si è quindi considerato il modello CRM di regressione a livello aggregato (regionale):

$$\begin{aligned} BoD_a^{N,low} &= f(VA_a^{N,low}, Densita_a^{N,low}, Superf_a^{N,low}, Pop_montagna_a^{N,low}, Quota50_a^{N,low}) \\ BoD_a^{N,high} &= f(VA_a^{N,high}, Densita_a^{N,high}, Superf_a^{N,high}, Pop_montagna_a^{N,high}, Quota50_a^{N,high}) \end{aligned}$$

ottenendo quindi due funzioni di regressioni stimate.

A questo punto, quindi, il percorso di disaggregazione territoriale può ripercorrere le fasi delineate nei modelli Chow-Lin spaziali standard, questa volta però separatamente per i livelli minimi ed i livelli massimi.

I risultati ottenuti confermano anche in questo caso le conclusioni cui si è giunti a seguito dell'applicazione del modello base, ossia l'evidente differenza tra i valori *veri* e quelli *stimati* attribuibili alle province italiane con riferimento alla loro dotazione di infrastrutture di trasporto terrestre (cfr. Tabella 2 e Figura 3).

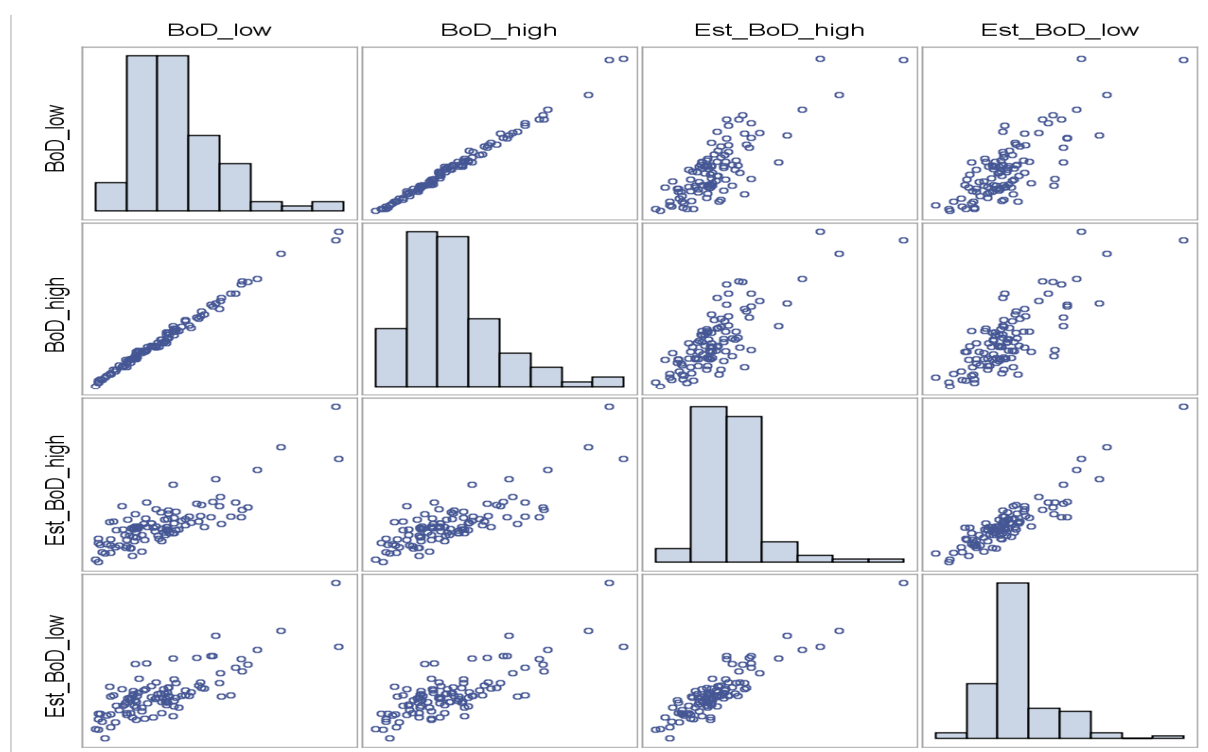
La sostanziale stabilità dei risultati ottenuti applicando il modello di disaggregazione spaziale in precedenza descritto si conferma anche a seguito della verifica effettuata per mezzo di un particolare indice di robustezza spaziale (IRS), costruito e utilizzato in un precedente lavoro

⁶ In particolare si ipotizza una distribuzione dell'errore di stima sia sui livelli minimi che massimi tale che $E(\varepsilon_d)=0$, $Cov(\varepsilon_d)=\sigma_d^2$, ipotesi non ancora dimostrata sui modelli CRM.

Tabella 2 – Applicazione del modello per dati intervallati – stime low ed high

Regione	Ymin	Y_est_min	Ymax	Y_est_max
ABRUZZO	0,568	0,379	0,594	0,408
BASILICATA	0,256	0,250	0,270	0,284
CALABRIA	0,389	0,367	0,447	0,402
CAMPANIA	0,509	0,499	0,596	0,580
EMILIA ROMAGNA	0,367	0,396	0,408	0,436
FRIULI VENEZIA GIULIA	0,487	0,449	0,531	0,518
LAZIO	0,345	0,351	0,397	0,396
LIGURIA	0,544	0,450	0,615	0,491
LOMBARDIA	0,383	0,472	0,442	0,524
MARCHE	0,326	0,376	0,382	0,421
MOLISE	0,337	0,375	0,357	0,417
PIEMONTE	0,393	0,337	0,452	0,388
PUGLIA	0,328	0,336	0,377	0,388
SARDEGNA	0,178	0,202	0,182	0,224
SICILIA	0,332	0,371	0,370	0,413
TOSCANA	0,362	0,409	0,414	0,449
TRENTINO ALTO ADIGE	0,216	0,172	0,235	0,196
UMBRIA	0,304	0,288	0,352	0,317
VALLE D'AOSTA	0,215	0,306	0,236	0,340
VENETO	0,350	0,403	0,376	0,442

Figura3: Confronto tra livelli dell'indicatore infrastrutturale sintetico (BoD) veri e stimati nell'ipotesi di applicazione del modello per dati intervallati



(Mazziotta e Vidoli, 2009), cui si rinvia per maggiori dettagli. Qui basti ricordare che tale indicatore è il risultato del prodotto di due matrici: la prima (matrice di contiguità, W) individua la contiguità territoriale delle unità (le province, nel nostro caso) tra loro; la seconda (matrice di transizione, T) evidenzia sia le differenze nei ranghi sia da quale unità e verso quale altra unità questa differenza si sia manifestata. Moltiplicando la matrice $I-W$ per la T si ottiene un indice che, confrontando i ranghi occupati dalle province nelle due situazioni considerate (indicatori infrastrutturali *veri* e *stimati*), evidenzia i cambiamenti nei ranghi che hanno interessato solamente unità spazialmente non contigue. L'indicatore IRS tra due ordinamenti in ranghi (R_0 e R_1), in forma algebrica, può essere scritto come:

$$IRS_{R_0, R_1} = \frac{\sum_{i,j} T_{i,j}(1 - w_{i,j})}{Max_I} \quad [6]$$

Va osservato che il massimo dell'indice proposto (max_I) equivale alla situazione peggiore dal punto di vista della conformità dei due ordinamenti, ovvero a quella in cui l'unità i che era al primo posto dell'ordinamento R_0 si ritrovi all'ultimo posto dell'ordinamento R_1 , e così via, questo tante volte quante sono le unità non contigue, $(n-1) * (n-2) * \dots$, ovvero tante volte quante si ha un valore maggiore di zero nella matrice $T(I-W)$.

Come si vede dalla Tabella 3, l'introduzione nel modello dei dati intervallati non sembra aver prodotto grandi differenze in termini di stabilità dei risultati: simili infatti risultano i livelli fatti registrare dall'IRS nelle diverse ipotesi considerate, e simile altresì risulta il numero di variazioni medie nei ranghi tra unità non appartenenti alla stessa area.

Tabella 3: Indice di robustezza spaziale (IRS) degli indicatori infrastrutturali stimati (modello base) rispetto a quelli veri, low ed high.

	IRS	Conteggio cambi extra area	Numeratore IRS	Denominatore IRS	Cambio medio extra - area del ranking
Bod vero	0,263	75	1284	4875	17,12
low	0,307	81	1541	5022	19,02
high	0,284	80	1421	5000	17,76

5. COMMENTO CONCLUSIVO

Obiettivo di questo lavoro era l'approfondimento di una linea di ricerca già avviata in precedenti lavori, avente per oggetto la predisposizione ed applicazione di modelli in grado di disaggregare sotto il profilo spaziale variabili disponibili ad un livello territoriale più aggregato. I modelli in questione sono improntati all'approccio di Chow-Lin, trasferito nell'ambito dell'analisi spaziale dal più consueto campo dell'analisi delle serie temporali. In particolare, il modello utilizzato si pone sulla scia delle precedenti elaborazioni di Bollino e

Polinori (2007) e di Polasek e Sellner (2008) e viene qui sperimentato con riferimento alla disaggregazione territoriale degli indicatori infrastrutturali sintetici relativi alla categoria dei trasporti terrestri. Sulla base delle ipotesi di similarità caratteristiche dell'approccio di Chow-Lin, si tratta quindi di verificare che le variabili che determinano la dotazione infrastrutturale a livello regionale siano in grado di determinare tale dotazione anche a livello provinciale.

Rispetto a precedenti applicazioni, in questo lavoro due sono state le linee di approfondimento: in primo luogo, l'introduzione, tra i regressori del modello, di variabili che tenessero conto in misura più puntuale delle specificità dei territori considerati, in particolare per ciò che riguarda le caratteristiche fisiche e i livelli di concentrazione-urbanizzazione; in secondo luogo, l'introduzione nel modello di variabili intervallate, in modo da ottenere livelli minimi e massimi dell'indicatore infrastrutturale a livello provinciale e verificarne la robustezza rispetto alla stima base.

La prima modifica, relativa alle nuove variabili indipendenti, ha avuto per risultato un miglioramento nella capacità esplicativa del modello (a livello regionale), che tuttavia non sembra avere influito in misura rilevante sulla distanza, tuttora persistente, tra i livelli *veri* e quelli *stimati* degli indicatori infrastrutturali a livello provinciale. Sembra dunque confermarsi il risultato che la distribuzione territoriale della dotazione infrastrutturale non è conforme ai fattori di generazione che teoricamente ne dovrebbero determinare i livelli nelle diverse aree del paese. A questo riguardo, tuttavia, occorre considerare che la difformità tra le due serie di indicatori si attenua se, anziché alle posizioni delle province in graduatoria, si fa riferimento al loro inserimento in classi omogenee di dotazione infrastrutturale: gli spostamenti da una classe all'altra sono in questo caso sono più contenuti di quanto non facessero supporre i cambiamenti di posizionamento in graduatoria.

Quanto alla seconda modifica, l'introduzione di dati intervallati nel modello di regressione utilizzato (ottenuti per mezzo dell'aggiunta di un rumore casuale uniforme sulle variabili considerate nel modello) non sembra aver prodotto risultati che facciano ritenere instabile il modello utilizzato: tali risultati, infatti, mantengono più o meno lo stesso scarto, rispetto alla soluzione del modello base, dell'intervallo imposto ai dati di ingresso del modello. L'applicazione di uno specifico indice di robustezza spaziale messo a punto in precedenti lavori (IRS) conferma la sostanziale stabilità dei risultati ottenuti. Ciò comporta, ancora una volta, la conferma dell'ipotesi di rilevante difformità tra domanda e offerta di infrastrutture (di trasporto terrestre) al livello territorialmente disaggregato delle province.

Riconoscimenti

Il lavoro è stato condotto in stretta collaborazione tra i due Autori. Quanto alla redazione del testo, C. Mazziotta ha scritto i parr. 1, 2, 3; F. Vidoli ha scritto i parr. 4 e 5.

Riferimenti bibliografici

- Billard, L., Diday, E. (2000), Regression Analysis for Interval-Valued Data, *Proceedings of the Seventh Conference of the International Federation of Classification Societies*, Springer-Verlag.
- Billard, L., Diday, E. (2002), Symbolic regression analysis, in *Classification, Clustering and Data Analysis, Proceedings of the Eighteenth Conference of the International Federation of Classification Societies*, Springer, Poland.
- Billard, L., Diday, E. (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley.
- Bollino C. A., Polinori P. (2007), Ricostruzione del valore aggiunto su scala comunale e percorsi di crescita a livello micro-territoriale: il caso dell'Umbria, *Scienze Regionali, Italian Journal of Regional Science*, n. 2.
- Book, H.H., Diday, E. (2000), *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer-Verlag.
- Carrizosa E., Gordillo J., Plastria F. (2007), "Support Vector Regression for imprecise data", Technical Report MOSI/35, MOSI Department, Vrije Universiteit Brussel.
- Chow G. C., Lin, A. (1971) Best linear unbiased interpolation, distribution, and extrapolation of time series by related series, *The Review of Economics and Statistics*, 53(4).
- Golden M., Picci L. (2005), Proposal for a New Measure of Corruption, Illustrated with Italian Data, *Economics and Politics*, vol. 17, n. 1.
- ISTAT, Le infrastrutture in Italia, Un'analisi provinciale della dotazione e della funzionalità, *Collana di Informazioni*, n. 6.
- Iwasaki, M., Tsubaki, H. (2005), A bivariate generalized linear model with an application to meteorological data analysis, *Statistical Methodology* 2.
- Lima Neto, E.A., De Carvalho, F.A.T. (2008), Centre and range method to fitting a linear regression model on symbolic interval data, *Computational Statistics and Data Analysis*, 52.
- Lima Neto, E.A., De Carvalho, F.A.T., Cordeiro, G.M, Anjos, U.U., Costa, A.G. (2009), Bivariate Generalized Linear Model for Interval-Valued Variables, *Proceedings of the 2009 IEEE International Joint Conferences on Neural Networks*. IEEE.
- Mazziotta C., Mazziotta M, Pareto A., Vidoli F. (2010), La costruzione di un indicatore sintetico ponderato di dotazione infrastrutturale: metodi e applicazioni a confronto, *Rivista di Economia e Statistica del territorio*, n. 1.
- Mazziotta C., Vidoli F. (2009a), La costruzione di un indicatore sintetico ponderato. Un'applicazione della procedura Benefit of Doubt al caso della dotazione infrastrutturale in Italia, *Scienze Regionali, Italian Journal of Regional Science*, n. 1.
- Mazziotta C., Vidoli F. (2009b), Robustezza e stabilità spaziale di indicatori di dotazione infrastrutturale: una verifica per le province italiane, in *Federalismo, integrazione europea e crescita regionale, Atti della XXX Conferenza Italiana di Scienze Regionali*, Firenze.
- Polasek W., Sellner R. (2008), *Spatial Chow-Lin methods: Bayesian and ML forecast comparisons*, Rimini Centre for Economic Analysis (RCEA), working paper 38-08.
- Vidoli F., Mazziotta C. (2010), Spatial Composite and Disaggregate Indicators: Chow-Lin Methods and Applications, *Proceedings of the 45th Scientific Meeting of the Italian Statistical Society*, Padua..

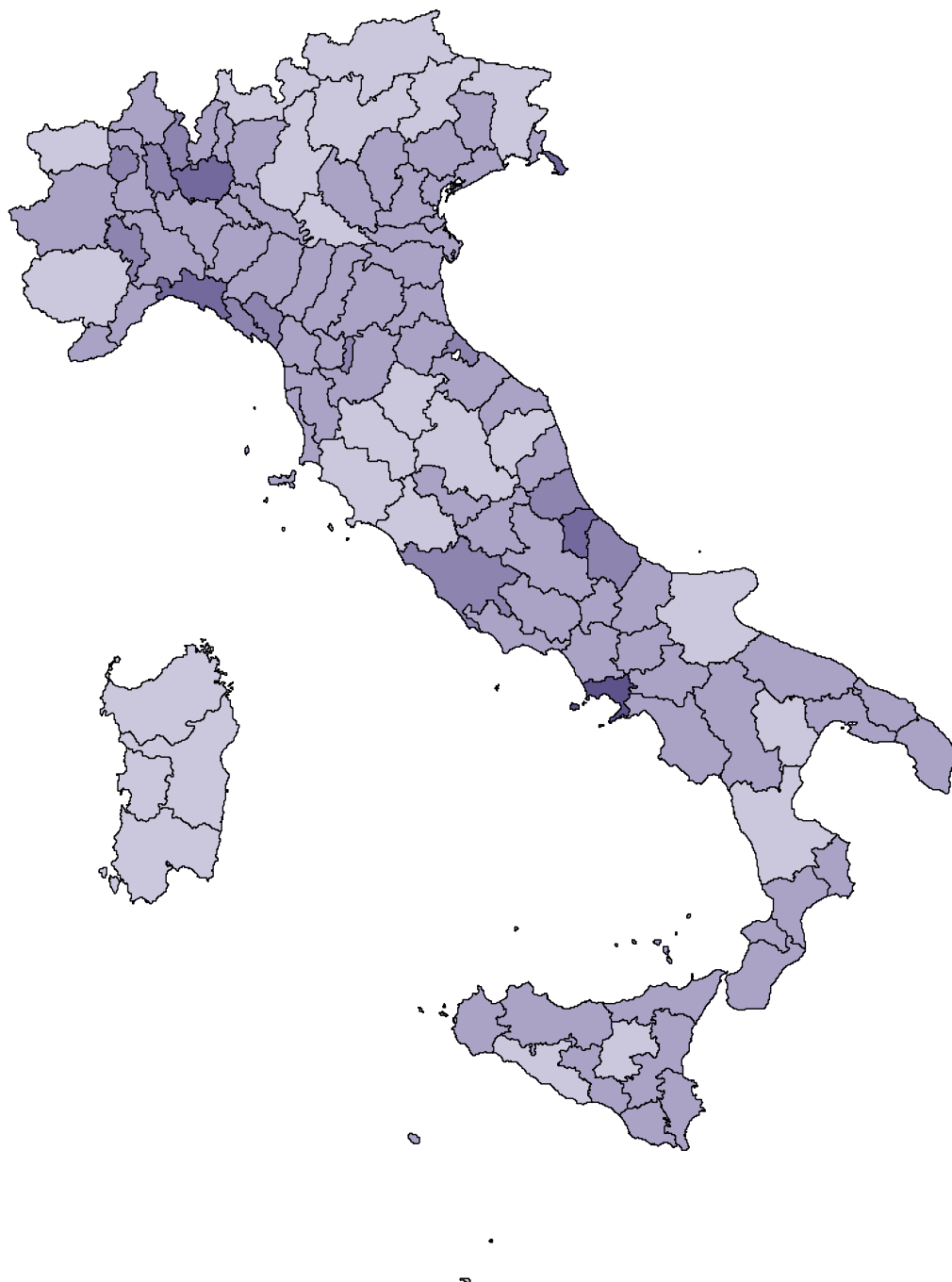
APPENDICE

Tavola A.1

Provincia	Indice infrastrutturale	Indice infrastrutturale stimato
AGRIGENTO	0,285	0,275
ALESSANDRIA	0,553	0,394
ANCONA	0,422	0,414
AOSTA	0,225	0,225
AREZZO	0,292	0,282
ASCOLI PICENO	0,436	0,364
ASTI	0,528	0,518
AVELLINO	0,458	0,399
BARI	0,352	0,356
BELLUNO	0,181	0,266
BENEVENTO	0,375	0,424
BERGAMO	0,329	0,402
BIELLA	0,305	0,513
BOLOGNA	0,417	0,357
BOLZANO	0,206	0,181
BRESCIA	0,299	0,287
BRINDISI	0,336	0,429
CAGLIARI	0,185	0,191
CALTANISSETTA	0,307	0,388
CAMPOBASSO	0,362	0,334
CASERTA	0,490	0,423
CATANIA	0,311	0,368
CATANZARO	0,419	0,482
CHIETI	0,519	0,558
COMO	0,362	0,477
COSENZA	0,397	0,250
CREMONA	0,382	0,359
CROTONE	0,246	0,489
CUNEO	0,286	0,212
ENNA	0,325	0,250
FERRARA	0,237	0,333
FIRENZE	0,330	0,344
FOGGIA	0,297	0,182
FORLI'	0,305	0,394
FROSINONE	0,418	0,328
GENOVA	0,710	0,787
GORIZIA	0,416	0,484
GROSSETO	0,190	0,196
IMPERIA	0,601	0,490
ISERNIA	0,313	0,340
LA SPEZIA	0,344	0,570
L'AQUILA	0,701	0,494
LATINA	0,276	0,423
LECCE	0,404	0,390
LECCO	0,424	0,457
LIVORNO	0,437	0,453
LODI	0,486	0,389
LUCCA	0,492	0,408
MACERATA	0,262	0,289
MANTOVA	0,334	0,284
MASSA CARRARA	0,552	0,504

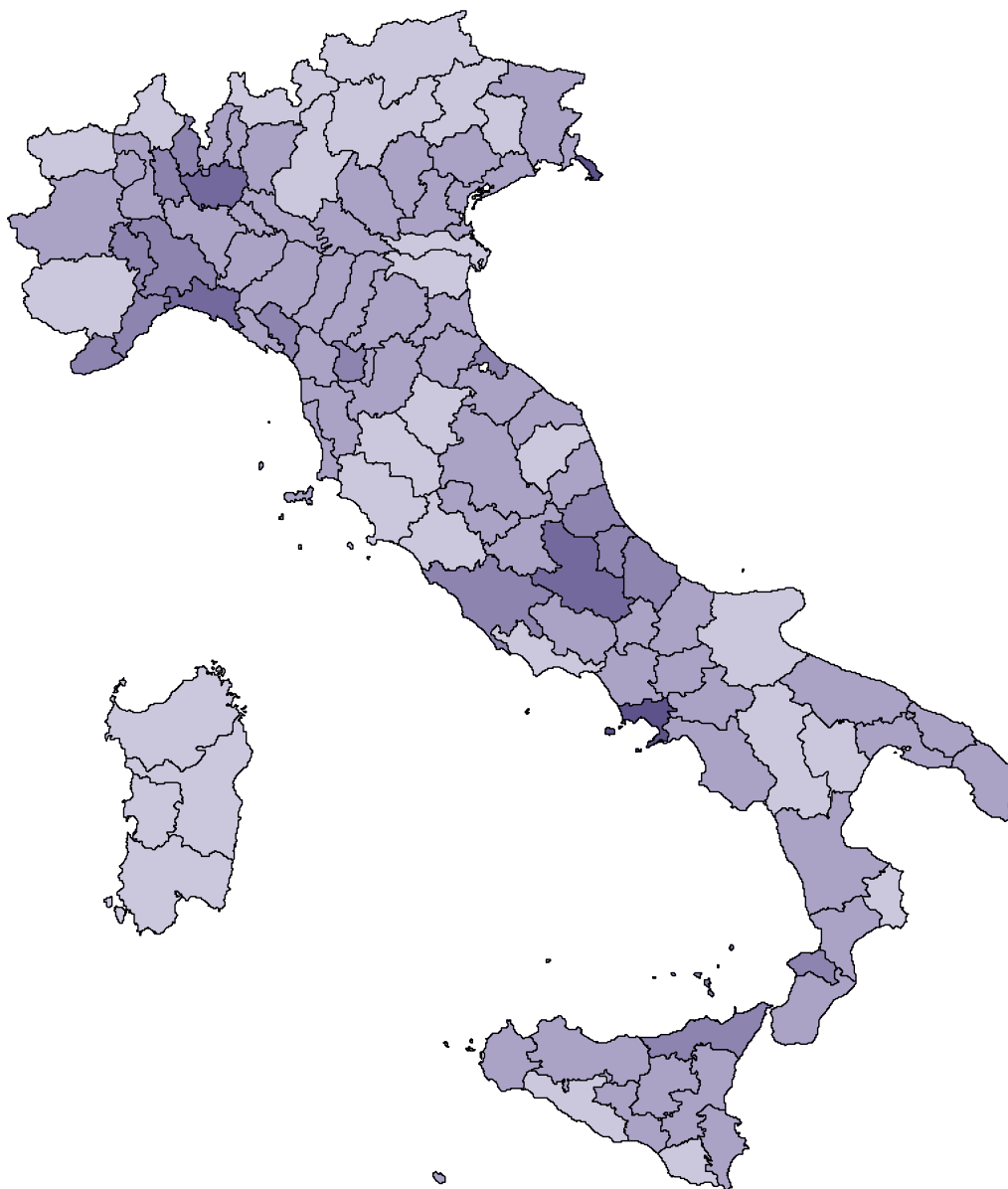
Provincia	Indice infrastrutturale	Indice infrastrutturale stimato
MATERA	0,268	0,320
MESSINA	0,626	0,417
MILANO	0,834	0,866
MODENA	0,351	0,379
NAPOLI	0,975	1,179
NOVARA	0,699	0,558
NUORO	0,167	0,105
ORISTANO	0,215	0,288
PADOVA	0,484	0,435
PALERMO	0,402	0,332
PARMA	0,372	0,314
PAVIA	0,396	0,318
PERUGIA	0,305	0,229
PESARO E URBINO	0,323	0,376
PESCARA	0,570	0,710
PIACENZA	0,380	0,331
PISA	0,310	0,314
PISTOIA	0,522	0,430
PORDENONE	0,230	0,354
POTENZA	0,269	0,217
PRATO	0,465	0,656
RAGUSA	0,238	0,425
RAVENNA	0,459	0,423
REGGIO CALABRIA	0,414	0,439
REGGIO EMILIA	0,301	0,372
RIETI	0,305	0,304
RIMINI	0,649	0,568
ROMA	0,582	0,505
ROVIGO	0,251	0,361
SALERNO	0,437	0,310
SASSARI	0,157	0,139
SAVONA	0,682	0,490
SIENA	0,213	0,215
SIRACUSA	0,306	0,360
SONDRIO	0,175	0,261
TARANTO	0,396	0,426
TERAMO	0,570	0,599
TERNI	0,361	0,437
TORINO	0,413	0,369
TRAPANI	0,399	0,383
TRENTO	0,228	0,253
TREVISO	0,445	0,372
TRIESTE	1,000	0,886
UDINE	0,349	0,271
VARESE	0,635	0,556
VENEZIA	0,410	0,431
VERBANO CUSIO OSSOLA	0,211	0,440
VERCELLI	0,421	0,411
VERONA	0,461	0,368
VIBO VALENTIA	0,623	0,439
VICENZA	0,380	0,378
VITERBO	0,265	0,285

FiguraA.1: Indicatore infrastrutturale sintetico (BoD) nel settore dei trasporti terrestri, valori stimati attraverso il modello di disaggregazione spaziale



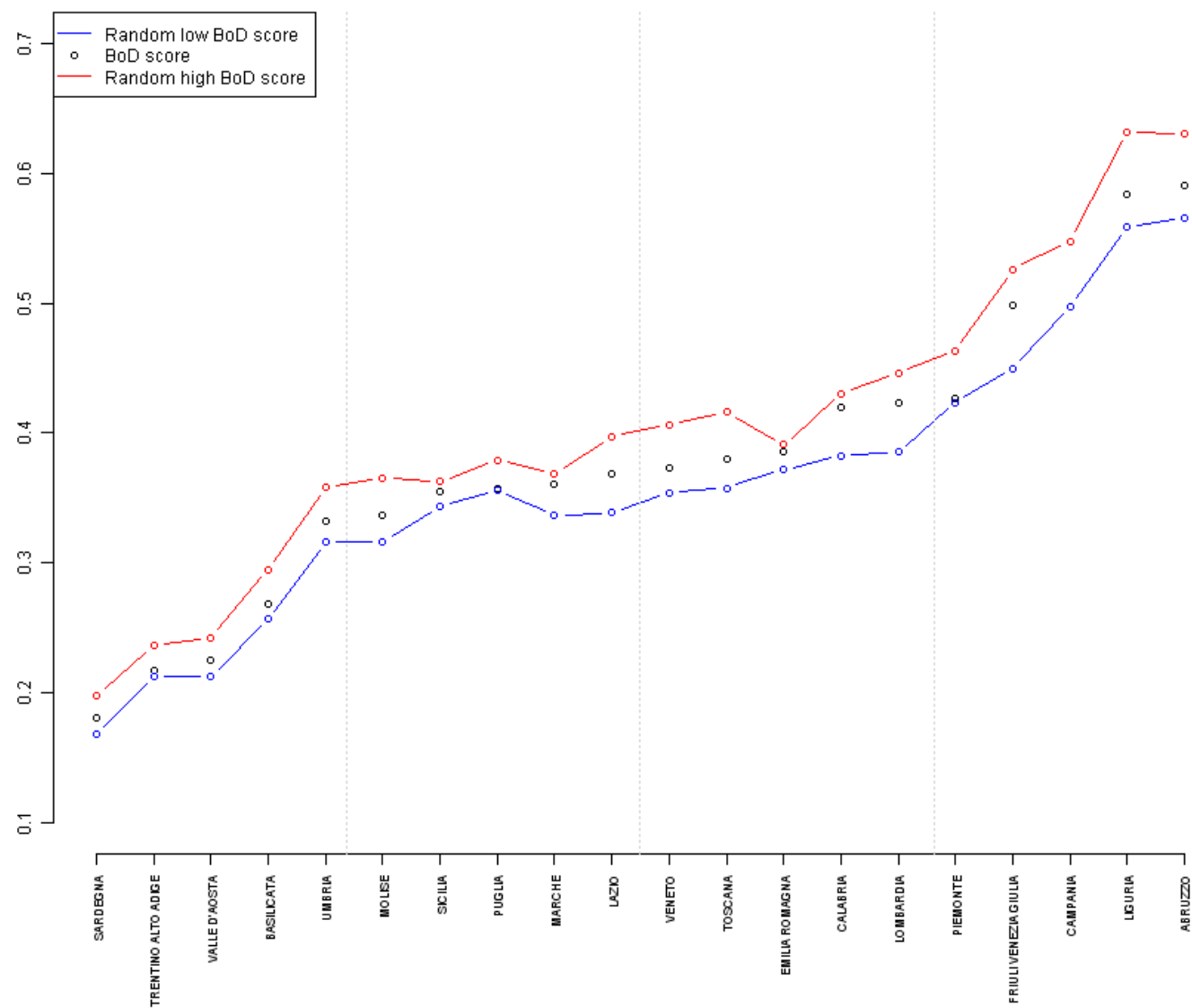
Distribuzione delle province secondo l'appartenenza a classi di dotazione infrastrutturale omogenee (quintili), indicatore sintetico da 0 a100, colore di intensità crescente dalla classe più bassa.

FiguraA.2: Indicatore infrastrutturale sintetico (BoD) nel settore dei trasporti terrestri, valori veri (sintesi su dati elementari ISTAT)



Distribuzione delle province secondo l'appartenenza a classi di dotazione infrastrutturale omogenee (quintili), indicatore sintetico da 0 a100, colore di intensità crescente dalla classe più bassa.

Figura A.3: Dati intervallati rispetto all'indicatore sintetico infrastrutturale(BoD), regioni italiane.



ABSTRACT

The present paper is focused to improve and verify methods and results of our previous work on spatial models, with specific application to the Italian infrastructure endowment, analyzed at the disaggregated territorial level (Italian “province”). We based our analysis on Chow-Lin models, extended to the spatial sphere. The independent variable is a composite indicator of infrastructure endowment, built up by a Benefit of the Doubt (BoD) approach; the dependents ones are some relevant factors that determine the transport demand level: production, urbanisation, population distribution and density, and so on.

Two improvement directions have been pursued in this paper: the first one regards the introduction of some new variables in the model, more adapted for describing the characteristics, both physical and socio-economic, of the considered areas. The second one concerns the introduction of the approach based on the *Symbolic Data Analysis*, aimed to testing the stability and the robustness of the model in presence of stochastic noise of the database. Both improvements produce the same conclusion: the robustness of the model is confirmed; therefore, considering the distance that the model indicates between the two series of infrastructure indicators – the “effective” data and the “estimated” data – we can confirm the persistence of a serious mismatch between infrastructure supply and demand in the Italian “province”.