

## “LOOKING FOR BVER INTO THE GRASS”. DATABASE PUBBLICI E MODELLI PREDITTIVI IN ARCHEOLOGIA: IL PROGETTO BVER IN DIALOGO CON GNA

### 1. INTRODUZIONE

L'area di studio del “Progetto ricerche e scavi nel Basso Verbano - BVER” (Fig. 1), focalizzato sulle evidenze di epoca romana, tardoantica e altomedievale, corrisponde alle sponde meridionali del lago Maggiore e alle aree adiacenti al fiume Ticino comprese negli odierni comuni di Angera, Sesto Calende, Vergiate, Somma Lombardo e Golasecca sulla sponda orientale, Arona e Castelletto sopra Ticino su quella occidentale. La storia degli studi relativi a questo territorio nell'ambito dell'archeologia classica e medievale vede la presenza di numerosi lavori, che possiedono però un approccio settoriale e affrontano principalmente aspetti legati a range cronologici molto specifici, come le fortificazioni, le necropoli o gli insediamenti di epoca altomedievale in rapporto a preesistenze di epoca tardo antica o a fondazioni *ex-novo* (DE MARCHI 1999). Alcuni studi hanno messo in luce come questa zona sia stata un crocevia commerciale in epoca romana, in quanto snodo tra la pianura padana e l'Europa centrale, indagando i possibili tracciati viari romani che da *Mediolanum* conducevano al Verbano e da qui, procedendo attraverso percorsi pedemontani o via lago, raggiungevano le zone dell'odierna Svizzera (DOLCI 2003).

Questo panorama di studi offre un'ottima base per lo sviluppo degli obiettivi del progetto: comprendere le dinamiche di popolamento su lunga durata, dall'età del Ferro all'epoca medievale, per individuare e analizzare le scelte umane compiute in relazione a fattori ambientali e ad esigenze socio-culturali, che gioco forza mutano nel passare dei secoli. In parallelo, il progetto si propone di comprendere se vi siano aspetti di continuità o discontinuità nelle scelte insediative di lungo periodo. Per poter meglio indagare le dinamiche insediative antropiche in rapporto con l'ambiente naturale, nelle prime fasi del progetto si è scelto di testare l'applicabilità di un approccio quantitativo al territorio, sull'esempio della letteratura esistente per altre aree (CARRER 2013; BRANDOLINI, CARRER 2020). Come dataset di riferimento per le analisi spaziali è stato scelto il Geoportale Nazionale per l'Archeologia (GNA, <https://ica.cultura.gov.it/geoportale-nazionale-per-larcheologia/>), integrato con dati provenienti da spoglio bibliografico e ricerche d'archivio.

L'areale preso in esame si presenta oggi come frutto di un lungo processo di antropizzazione, documentabile soprattutto a partire dagli anni '50 del secolo scorso, che ha reso difficile la ricostruzione e l'interpretazione

organica di quello che doveva essere l'assetto paesaggistico storico nella sua interezza. Il paesaggio attuale è caratterizzato *in primis* da un'urbanizzazione aggressiva tra Milano e il lago Maggiore, che vede anche un'intensa opera di realizzazione di infrastrutture per la connessione dei centri urbani, con la creazione e l'ampliamento della rete autostradale e ferroviaria, e la presenza di infrastrutture aeroportuali, tra cui l'importante scalo di Malpensa. Lo sfruttamento del lago come polo turistico ha portato ad un'intensificarsi dell'edilizia per la realizzazione di strutture di ricezione a ridosso delle rive del bacino, già interessate anche da interventi per il contenimento e la regolarizzazione delle sponde. A questo quadro si vanno ad aggiungere attività di disboscamento su lungo periodo finalizzate all'impianto di coltivazioni intensive, con un consistente impatto sullo sfruttamento del suolo. Queste dinamiche hanno fortemente inciso sulla visibilità del record archeologico: l'edificazione di nuove infrastrutture e l'urbanizzazione hanno infatti permesso di ampliare la conoscenza archeologica del territorio grazie ad interventi di archeologia preventiva e di emergenza in anni recenti. Per la seconda metà del XX secolo, che rappresenta il momento di maggiore espansione edilizia e infrastrutturale dell'area, per molti siti si hanno invece solo notizie frammentarie, quando non la totale assenza di dati.

## 2. MATERIALI E METODI

### 2.1 *Geoportale Nazionale per l'Archeologia*

Il GNA costituisce il punto di raccolta e condivisione online dei dati esito di tutte le indagini archeologiche condotte sul territorio italiano, con accesso libero e apertura al riuso e all'integrazione dei dati da parte di tutti gli utenti (CALANDRA 2022; ACCONCIA *et al.* 2024). Il progetto ha rivolto la sua attenzione alla standardizzazione dei dati, con la costruzione di un template GIS, avente un sistema gerarchico, che richiede l'inserimento dapprima di dati generici, per arrivare poi a quelli di dettaglio (GABUCCI 2024). L'architettura del database si è trovata a dover superare soprattutto problemi di aggiornabilità, rendendo obbligatorio l'inserimento di una documentazione di scavo il più possibile completa. La maggior criticità del database risiede tuttavia nella presenza di numerosi dati pregressi, ossia di tutte quelle relazioni depositate in formato cartaceo, che devono ancora essere digitalizzate e poi caricate sul portale (BOI 2024); questo processo risulta farraginoso e comporta delle lacune che auspicabilmente saranno colmate solo nel corso di diversi anni.

### 2.2 *Modelli predittivi induttivi*

Per l'analisi preliminare del territorio del Basso Verbano si è scelto di utilizzare la tecnica della modellazione predittiva di tipo induttivo, in primo

luogo per indagare se tale approccio metodologico fosse utile per la comprensione dei contesti archeologici presenti nell’area ed eventualmente applicabile anche ad un’area di maggiore estensione.

Dal punto di vista strettamente teorico un modello predittivo è uno strumento che permette di individuare un pattern di relazioni all’interno di un campione analitico, generalizzandolo in seguito ad una popolazione più ampia (VERHAGEN, WHITLEY 2020). Un modello predittivo di tipo induttivo, detto anche “data-driven”, analizza le relazioni tra una variabile dipendente, in questo caso un pattern di siti, e diverse variabili indipendenti (CONOLLY, LAKE 2006; ALBERTI *et al.* 2018) che si ipotizza possano influire sulla presenza o assenza dei siti. Fin dalle prime applicazioni di questa tecnica in ambito archeologico (KVAMME 1988) è stato evidente il suo forte potere analitico, in aggiunta alla sua funzione propriamente predittiva. Questo tipo di modello è infatti utilizzabile anche per indagare il rapporto degli insediamenti umani con un preciso ambito territoriale, sottolineando le connessioni tra azione antropica e ambiente naturale. Tale approccio rende la modellazione predittiva uno strumento rilevante, seppur non risolutivo, per l’indagine delle connessioni profonde tra le varie entità, umane e non umane, che strutturano i processi evolutivi dei paesaggi archeologici.

La metodologia predittiva è stata però anche soggetta ad una serie di critiche in ambito archeologico, legate soprattutto alla solidità statistica dei dati utilizzati e a una serie di limitazioni, spesso insormontabili, nella scelta delle variabili indipendenti (WHEATLEY, GILLINGS 2002; VAN LEUSEN *et al.* 2005; VERHAGEN, WHITLEY 2020). Solitamente si propende infatti per la scelta di variabili che esprimono precise caratteristiche fisiche di un territorio, come l’altitudine, l’esposizione solare, la pendenza dei versanti, etc., poiché questi elementi sono facilmente misurabili nel paesaggio attuale e si presuppone che non abbiano subito forti modificazioni rispetto all’assetto dello stesso territorio in antico. La natura statistica del processo di elaborazione del modello porta quindi spesso ad escludere tutte quelle variabili di tipo socioculturale, che sicuramente sono state decisive per le scelte insediative umane nel passato, ma che non sono quantificabili con una precisione tale da essere utilizzabili per i calcoli previsti dalla metodologia predittiva. Lo stesso dato archeologico soffre di diversi bias legati soprattutto ai processi formativi, alle dinamiche tafonomiche e alla visibilità del record stesso, che non permettono di considerarlo come un campione statistico ottimale. Una soluzione proposta per il superamento delle criticità legate alla modellazione prognostica prevede un approccio etnoarcheologico (CARRER 2013; CROCE *et al.* 2025). In questa proposta si cerca di risolvere le criticità dell’uso del record archeologico come campione di riferimento attraverso la modellazione di evidenze moderne e ben

documentate, correlabili dal punto di vista funzionale ad evidenze antiche. Questo approccio non sembra essere possibile per il caso del Basso Verbano, data l'eterogeneità funzionale delle evidenze e l'ampio intervallo cronologico di riferimento. Anche per questa ragione, si è deciso di implementare il calcolo di un modello predittivo induttivo non etnoarcheologico per l'area di pertinenza del progetto BVER. In primo luogo, per poter valutare dal punto di vista pratico la validità di un tale approccio in relazione ai fini del progetto, ma anche per ottenere dei dati utili ad una riflessione più ampia sulla persistenza di questo tipo di metodologia all'interno dell'attuale panorama della ricerca archeologica.

### 2.3 Manipolazione dei dati

La prima fase della ricerca, in vista del calcolo del modello predittivo, è consistita nella raccolta dei dati relativi ai siti archeologici conosciuti, con il censimento dei siti databili tra l'età del Ferro e il Medioevo in una porzione circoscritta di territorio. L'area di riferimento è stata definita utilizzando confini naturali, come i corsi d'acqua e le linee di spartiacque, per limitare al minimo l'aleatorietà della scelta. La sua forma risulta vagamente triangolare (Fig. 1), con la punta a meridione, definita dallo spartiacque Ticino-Agogna a W e il corso del torrente Arno a E e con la base disegnata da una linea che corre sulla sponda orientale del Verbano, fino a Ispra, e continua sulla sponda SW del lago di Varese. La raccolta dati si è inizialmente concentrata sui contenuti della piattaforma GNA (accesso: ottobre 2024), che non si presentano tuttavia omogenei, con evidenti lacune corrispondenti ai territori di vari comuni probabilmente per la maggior parte dei casi dovute alla mancata digitalizzazione delle relazioni di scavo, alla totale assenza della documentazione o a errori di posizionamento. Il dato, in questi casi, è stato integrato mediante spoglio bibliografico dei notiziari della Soprintendenza della Lombardia e del Piemonte e la consultazione delle carte archeologiche. Il dataset così ottenuto conta 844 siti, con cronologie comprese tra la Preistoria e l'età medievale.

I dati sono stati gestiti tramite piattaforma QGIS (<https://www.qgis.org/>), mantenendo come base per le tabelle di riferimento l'impostazione del template di GNA, per facilitare un caricamento futuro di dati aggiuntivi sulla piattaforma stessa. In questa fase i dati sono stati raffinati e suddivisi per tipologie strutturali e cronologiche semplificate, in modo da agevolare il successivo processo di analisi. La seconda fase di elaborazione dei dati ha visto l'impiego della piattaforma GRASS GIS (<https://grass.osgeo.org/>) e ha previsto la manipolazione statistica degli stessi in ambiente R (<https://www.r-project.org/>), attraverso l'interfaccia di RStudio (<https://posit.co/products/open-source/rstudio/>). La scelta di GRASS è stata determinata dalla presenza di strumenti di analisi specifici, relativi soprattutto alla

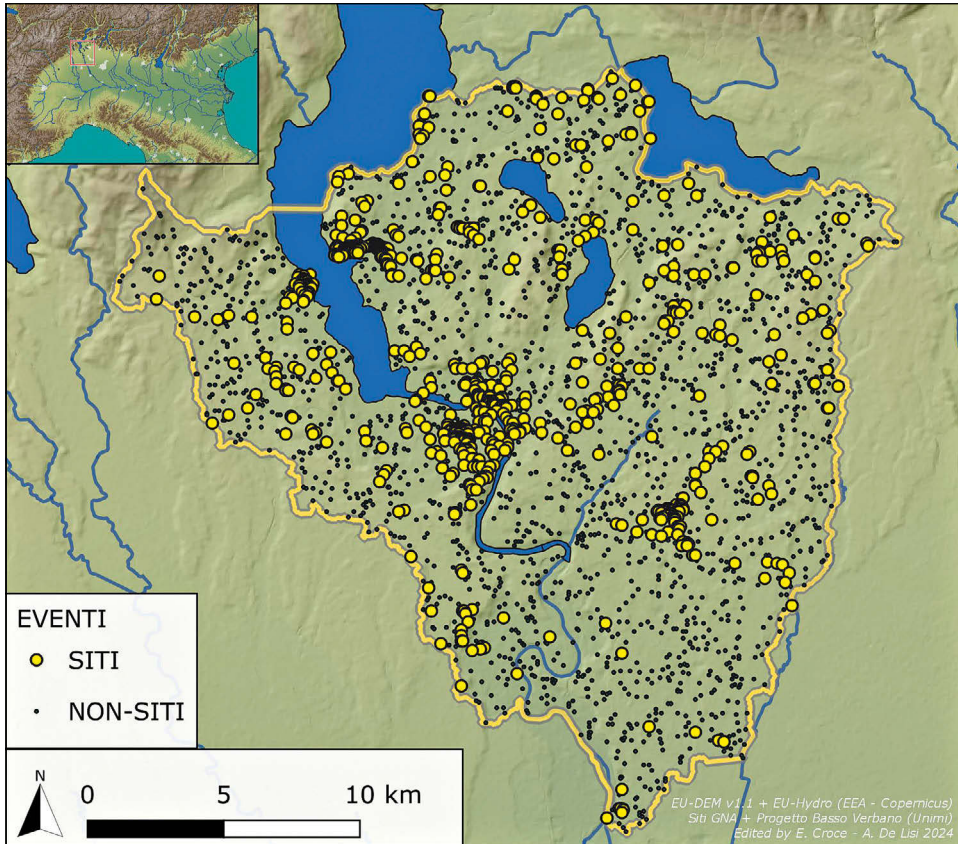


Fig. 1 – Posizionamento dell’area di studio del progetto Basso Verbano e del record di eventi (siti e non-siti) utilizzati come variabile dipendente per il calcolo di regressione logistica multivariata alla base del modello predittivo.

Map Algebra, oltre che dalla possibilità di far interagire direttamente i dati spaziali in ambiente R. Un evidente problema del dato presente in GNA è legato all’aggregazione dei siti. Visti gli obiettivi primari del database, legati a dinamiche di tutela e valorizzazione del patrimonio, il dataset tiene traccia di tutti gli interventi in modo separato, senza ricondurre quelli plurimi in un singolo sito ad una scheda comune. Questo tipo di aggregazione dei dati dal punto di vista statistico può portare a sovrastimare o sottostimare alcune evidenze.

Per attenuare questo problema si è deciso di rasterizzare il record di siti, ricampionandolo ad una risoluzione di 50m, coerentemente con la cartografia raster utilizzata per le successive analisi geostatistiche. In

NOME FILE	TIPO	DESCRIZIONE	K-S TEST
<b>SITI_PIXEL</b>	dipendente (1)	Siti archeologici	
<b>BVER_RANDOM</b>	dipendente (0)	Punti Casuali	
<b>DEM</b>	indipendente	Altitudine	< 2.22e-16
<b>SLOPE_50</b>	indipendente	Pendenza dei versanti	0.0026
<b>WALK_NF</b>	indipendente	Distanza di costo anisotropica dal Verbano/Ticino	< 2.22e-16
<b>DEM_EASTERNESS</b>	indipendente	sin(aspect)	0.1357
<b>DEM_NORTHERNESS</b>	indipendente	esposizione est-ovest cos(aspect)	< 2.22e-16
<b>SLOPE_50_CURVATURE</b>	indipendente	esposizione nord-sud Curvatura dei versanti	0,0004
<b>TOPIDX_50</b>	indipendente	Topographic Wetness Index	0.1784
<b>TPI_9</b>	indipendente	Topographic Position Index	< 2.22e-16

Tab. 1 – Elenco delle variabili utilizzate per il calcolo della regressione logistica multivariata.

questo modo, da un totale di 844 evidenze, si è passati ad un campione finale di 701 punti di interesse. Per la validazione statistica del dato è stato creato un record di punti casuali, in rapporto di 3:1 rispetto al numero dei siti, avendo cura di evitare sovrapposizioni con i siti noti e i corpi d'acqua maggiori. Questa proporzione è stata scelta in modo arbitrario, sulla base di esperienze pregresse, data la relativa assenza in letteratura di indicazioni specifiche per lo studio dei contesti archeologici. I punti casuali, insieme ai siti archeologici ricampionati, vanno a costituire il dataset degli eventi (Fig. 1), suddiviso tra le categorie di siti e non-siti. Per la posizione di ogni evento si registra il corrispettivo valore delle mappe raster di riferimento, andando a comporre un dataset delle variabili indipendenti utilizzato poi per il calcolo del modello. Queste ultime sono derivate dal modello digitale del terreno European Digital Elevation Model (EU-DEM v1.1, European Environment Agency - Copernicus programme; <https://ec.europa.eu/eurostat/web/gisco/geodata/digital-elevation-model/>), ricampionato ad una risoluzione di 50 m.

La scelta di un DEM a scala sovranazionale, piuttosto che l'utilizzo di dati regionali o nazionali, si è resa necessaria per il posizionamento dei siti in esame in un'area di confine, sia regionale che nazionale, che avrebbe determinato possibili errori di calcolo nelle superfici derivate da diversi raster altitudinali locali, soprattutto lungo i bordi degli stessi. Queste mappe derivate contengono informazioni relative a diverse caratteristiche fisiche del territorio (Tab. 1), come l'esposizione solare, espressa con seno

e coseno dell'aspect (KING *et al.* 1999; OLAYA 2009), la pendenza dei versanti in gradi, la curvatura dei profili, il “topographic wetness index” e “topographic position index” calcolato usando la deviazione standard dell'elevazione (DE REU *et al.* 2013); la distanza di costo anisotropica dal lago Maggiore e dal corso del Ticino esprime invece la stima del tempo di percorrenza a piedi del territorio, a partire da un elemento di interesse (NETLER, MITASOVA 2008).

#### 2.4 Calcolo del modello

Per il calcolo del modello è stato utilizzato il protocollo operativo definito da CROCE, CARRER (2024) omettendo la sua premessa etnoarcheologica (CARRER 2013). Il codice utilizzato e i risultati completi del calcolo sono disponibili in open access ([https://doi.org/10.13130/RD\\_UNIMI/2WL9VZ](https://doi.org/10.13130/RD_UNIMI/2WL9VZ)). Il modello finale rappresenta l'elaborazione spaziale del risultato di una regressione logistica multivariata, in cui i siti, insieme ad un record di punti generati in modo in casuale (non-siti), costituiscono la variabile dipendente e i valori delle mappe raster (Tab. 1) nella posizione di ogni singolo evento rappresentano le variabili indipendenti.

Prima di effettuare il calcolo di regressione è necessario testare la reale indipendenza delle variabili utilizzate, attraverso il calcolo della collinearità, effettuato utilizzando l'indice di correlazione di Pearson (DORMANN *et al.* 2013; ALBERTI *et al.* 2018). I risultati (Fig. 2) vengono restituiti in una scala da 1 a -1, dove 0 significa completa assenza di correlazione. Il dato utilizzato rappresenta le caratteristiche geomorfologiche di un territorio e quindi risulta poco probabile registrare una totale assenza di correlazione tra di esse (cfr. TOBLER 1970). Viene quindi usata una soglia a  $\pm 0.70$ , sull'esempio di ALBERTI *et al.* (2018), per la valutazione della collinearità. Tutte le variabili hanno valori al di sotto della soglia.

In seguito, per valutare la significatività delle variabili ai fini della modellazione, i valori di tutte le variabili indipendenti relativi alla posizione dei siti vengono confrontati, attraverso il test di Kolmogorov-Smirnov, con i valori delle stesse nella posizione dei non-siti (CONOLLY, LAKE 2006; KVAMME 2020). Questo test, confrontando la distribuzione cumulativa dei valori di due campioni, permette di valutare la loro ipotetica appartenenza ad una stessa popolazione statistica. In caso di uniformità della distribuzione dei valori tra siti e punti casuali non sarebbe possibile supporre che la variabile in esame influenzi in qualche modo il posizionamento dei siti. Per aumentare la robustezza del risultato il test viene effettuato in bootstrapping (CARLSON 2017), ripetendolo per 10.000 volte su diversi sottocampioni del dato in esame. Utilizzando un valore di soglia (p-value) di 0.05 è possibile rigettare l'ipotesi di appartenenza alla stessa popolazione per tutte le variabili tranne due (Easternness, TWI; Tab. 1). Un'ulteriore fase di validazione delle variabili

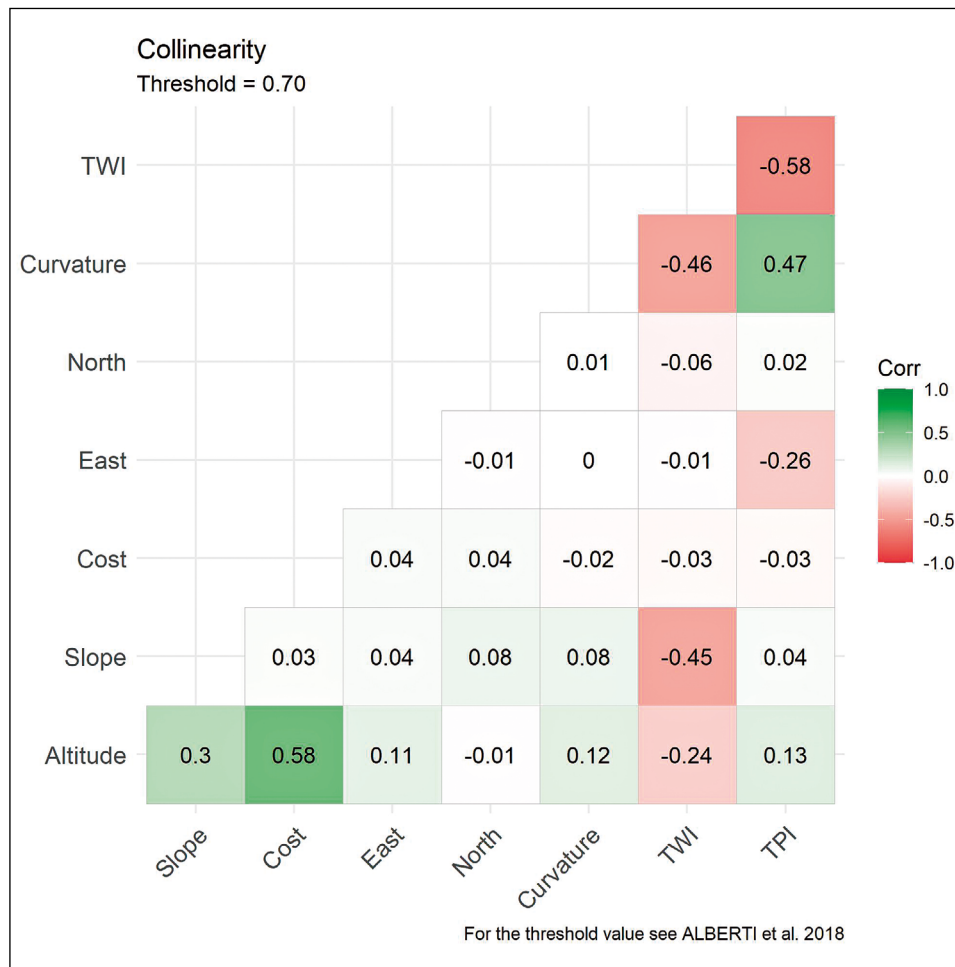


Fig. 2 – Risultato del calcolo della collinearità per le variabili indipendenti utilizzate per il calcolo di regressione logistica multivariata. Il valore di soglia è stato stabilito a 0.70 sulla base di ALBERTI et al. 2018.

ha visto l’implementazione di una regressione logistica univariata per ogni singola variabile, utile a testare la presenza di relazioni di tipo non lineare che potrebbero compromettere l’accuratezza del modello. Le variabili che erano state già scartate sulla base del test di Kolmogorov-Smirnov sono anche quelle che restituiscono una curva di regressione univariata poco significativa. Tutte le altre variabili sono quindi utilizzate per un calcolo di regressione logistica multivariata.

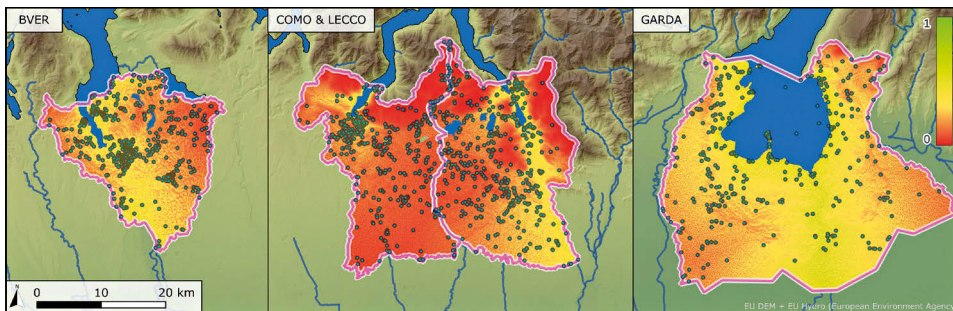


Fig. 3 – Tavola di confronto delle superfici predittive calcolate utilizzando i risultati della regressione logistica multivariata del dataset del Basso Verbano (BVER). Le aree di controllo comprendono territori del basso Lario e del basso Garda.

Per identificare la migliore combinazione di variabili è stato utilizzato il metodo della “stepwise model selection” (CARRER 2013; VENABLES *et al.* 2024), che permette di valutare le prestazioni di un modello attraverso l’AIC (Akaike Information Criterion) e l’AIC di Schwartz. Quest’ultimo, noto anche come “Bayesian Information Criterion” (BIC), perché vincolato dal numero di osservazioni (CLAESKENS, JANSEN 2015), permette di ottenere la combinazione che restituisce il risultato più parsimonioso, cioè che spiega il fenomeno utilizzando il minor numero di variabili possibile. Sulla base di questa selezione, il calcolo di regressione utilizzato per la definizione finale del modello comprende soltanto le variabili che descrivono l’altitudine, la distanza anisotropica di costo dal Ticino/Verbano, la northernness e il topographic position index.

I coefficienti della regressione logistica multivariata, tramite gli strumenti di Map Algebra di GRASS, sono applicati alle mappe raster che descrivono le variabili selezionate, permettendo di ottenere la cosiddetta “superficie predittiva” (Fig. 3, sinistra). Questa mappa esprime a livello spaziale i risultati della regressione e viene resa graficamente su una scala di valori normalizzati tra 0 e 1 (CARRER 2013). In questo modo il valore di ogni cella, che rappresenta una porzione di territorio quadrata di 50 m di lato, esprime a livello probabilistico la sua coerenza con il pattern insediativo descritto dal modello.

## 2.5 Validazione interna

I risultati della regressione vanno validati per determinare il corretto funzionamento e il potere predittivo del modello (Fig. 4). In prima battuta si analizza il grado di multicollinearità delle variabili utilizzate nella regressione, cioè quanto esse si influenzino reciprocamente nella determinazione del risultato (MIDI *et al.* 2010), attraverso il calcolo del Variance Inflation

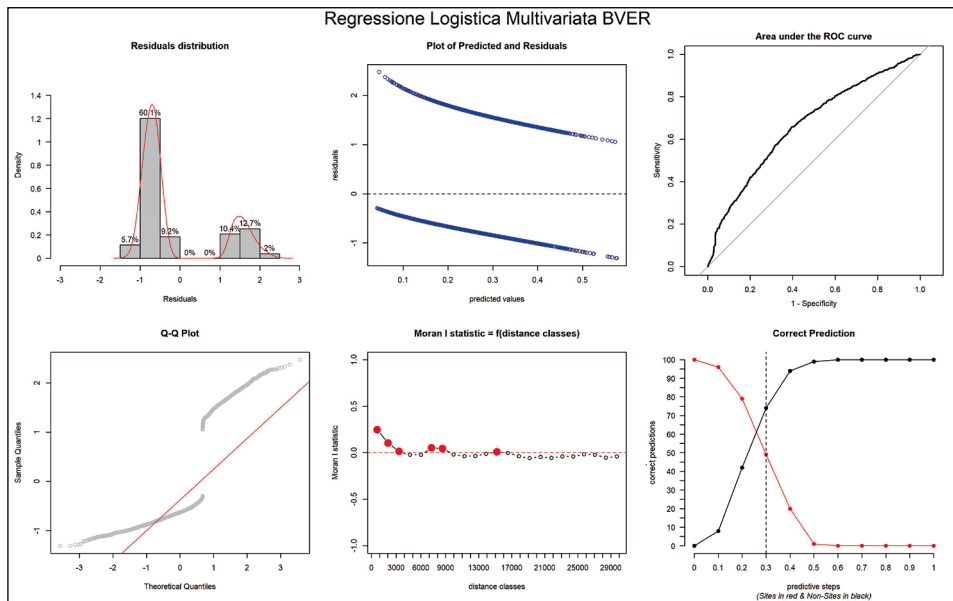


Fig. 4 – Risultati grafici dell’analisi dei residui, calcolo della I di Moran, area sotto la curva ROC e soglia di predizione ottimale per il modello predittivo del Basso Verbano calcolato per il record completo di siti.

GLM(FORMULA = NAME ~ DEM + WALK\_NF + DEM\_NORTHERNESS + TPI\_9, FAMILY = BINOMIAL(LOGIT), DATA = TAB)

	Estimate	Std. Error	z value	p-value	VIF
<b>INTERCEPT</b>	0.2976	0.2499	1.191	0.233685	
<b>DEM</b>	-0.003802	0.001078	-3.526	0.000422	1.627025
<b>WALK_NF</b>	-0.000135	0.00002257	-5.980	0.0000000223	1.600141
<b>DEM_NORTHERNESS</b>	-0.227	0.06431	-3.529	0.000417	1.001526
<b>TPI_9</b>	0.4136	0.08171	5.062	0.000000415	1.032011
<b>AIC</b>	2992.2				
<b>AUC</b>	0.6661				

Tab. 2 – Risultati della regressione logistica multivariata effettuata con le variabili selezionate attraverso la stepwise model selection.

Factor (VIF). I valori ottenuti (Tab. 2) sono abbondantemente sotto la soglia critica proposta in letteratura in ambito archeologico (ALBERTI *et al.* 2018) per cui si esclude la presenza di tale fenomeno in misura tale da inficiare il risultato del calcolo. Per valutare il potere discriminatorio del modello viene utilizzato il test dell’area sotto la curva ROC (HOSMER *et al.* 2013; ALBERTI

ETÀ DEL FERRO					
	Estimate	Std. Error	z value	p-value	VIF
INTERCEPT	-1.1502253	0.6883411	-1.671	0.09472	
DEM	0.0109844	0.0034657	3.169	0.00153	2.237725
SLOPE_50	-0.1232231	0.0304685	-4.044	0.0000525	1.191186
WALK_NF	-0.0012443	0.0001494	-8.330	< 2e-16	2.130355
DEM_NORTHERNESS	0.6561684	0.1427957	4.595	0.00000432	1.026266
TPI_9	0.4766682	0.1787869	2.666	0.00767	1.018798
AIC	602.89				
AUC	0.8793				
ETÀ ROMANA					
	Estimate	Std. Error	z value	p-value	VIF
INTERCEPT	-0,6265	0,1072	-5.844	0,0000000051	
WALK_NF	-0,0001837	0,00002721	-6.749	0,000000000149	1.002365
DEM_NORTHERNESS	-0,6182	0,1023	-6.044	0,0000000015	1.002365
AIC	1343.2				
AUC	0.6848				
TARDOANTICO					
	Estimate	Std. Error	z value	p-value	VIF
INTERCEPT	-0.0340337	0.3723704	-0.091	0.92718	
WALK_NF	-0.0003826	0.0001266	-3.022	0.00251	---
AIC	98.753				
AUC	0.7425				
MEDIOEVO					
	Estimate	Std. Error	z value	p-value	VIF
INTERCEPT	-1.2719	0.1397	-9.102	< 2e-16	
DEM_NORTHERNESS	-0.4590	0.1826	-2.514	0.01193	1.000006
TPI_9	0.6698	0.2330	2.874	0.00405	1.000006
AIC	386.33				
AUC	0.6493				

Tab. 3 – Risultati delle regressioni logistiche multivariate effettuate sui subset cronologici del record di eventi principale.

*et al.* 2018), che restituisce un valore di 0.66, classificato come “scarso” in letteratura (HOSMER *et al.* 2013). Per attestare la significatività della distribuzione dei valori predittivi assegnati ai siti, essa viene comparata con la distribuzione dei valori assegnati ai non-siti, attraverso il test di Kolmogorv-Smirnov in bootstrapping, con le stesse modalità impiegate per la valutazione

<b>BVER</b>	<b>&gt;0.1</b>	<b>&gt;0.2</b>	<b>&gt;0.3</b>	<b>&gt;0.4</b>	<b>&gt;0.5</b>	<b>&gt;0.6</b>	<b>&gt;0.7</b>	<b>&gt;0.8</b>	<b>&gt;0.9</b>
% Area	0.92	0.60	0.27	0.06	0.01	0	0	0	0
% Siti	0.96	0.79	0.49	0.21	0.01	0	0	0	0
Kvamme's Gain	0.04	0.24	0.45	0.71	0.00	NaN	NaN	NaN	NaN
<b>ETÀ DEL FERRO</b>	<b>&gt;0.1</b>	<b>&gt;0.2</b>	<b>&gt;0.3</b>	<b>&gt;0.4</b>	<b>&gt;0.5</b>	<b>&gt;0.6</b>	<b>&gt;0.7</b>	<b>&gt;0.8</b>	<b>&gt;0.9</b>
% Area	0.46	0.34	0.26	0.19	0.12	0.07	0.03	0.01	0
% Siti	0.95	0.92	0.86	0.77	0.61	0.34	0.12	0.02	0
Kvamme's Gain	0.52	0.63	0.70	0.75	0.80	0.79	0.75	0.50	NaN
<b>ETÀ ROMANA</b>	<b>&gt;0.1</b>	<b>&gt;0.2</b>	<b>&gt;0.3</b>	<b>&gt;0.4</b>	<b>&gt;0.5</b>	<b>&gt;0.6</b>	<b>&gt;0.7</b>	<b>&gt;0.8</b>	<b>&gt;0.9</b>
% Area	0.90	0.59	0.27	0.09	0	0	0	0	0
% Siti	0.97	0.78	0.54	0.31	0	0	0	0	0
Kvamme's Gain	0.07	0.24	0.50	0.71	NaN	NaN	NaN	NaN	NaN
<b>TARDOANTICO</b>	<b>&gt;0.1</b>	<b>&gt;0.2</b>	<b>&gt;0.3</b>	<b>&gt;0.4</b>	<b>&gt;0.5</b>	<b>&gt;0.6</b>	<b>&gt;0.7</b>	<b>&gt;0.8</b>	<b>&gt;0.9</b>
% Area	0.76	0.55	0.37	0.19	0	0	0	0	0
% Siti	1.00	0.79	0.62	0.50	0	0	0	0	0
Kvamme's Gain	0.24	0.30	0.40	0.62	NaN	NaN	NaN	NaN	NaN
<b>MEDIOEVO</b>	<b>&gt;0.1</b>	<b>&gt;0.2</b>	<b>&gt;0.3</b>	<b>&gt;0.4</b>	<b>&gt;0.5</b>	<b>&gt;0.6</b>	<b>&gt;0.7</b>	<b>&gt;0.8</b>	<b>&gt;0.9</b>
% Area	0.97	0.58	0.22	0.06	0.01	0	0	0	0
% Siti	0.99	0.83	0.40	0.10	0.00	0	0	0	0
Kvamme's Gain	0.02	0.30	0.45	0.40	-Inf	NaN	NaN	NaN	NaN

Tab. 4 – Calcolo del Kvamme's Gain per tutti gli intervalli predittivi dell'intero record dei siti e delle singole selezioni cronologiche.

preventiva delle variabili indipendenti. Il risultato ( $< 2.22-16$ ) ci permette di rigettare l'ipotesi di appartenenza delle due distribuzioni ad una medesima popolazione statistica.

Il potenziale predittivo del modello viene calcolato anche attraverso il calcolo del Kvamme's Gain (KVAMME 1988), che compara la percentuale di area predetta con la percentuale di siti correttamente posizionati al suo interno. Questo test restituisce risultati compresi tra -1 e 1, dove 0 rappresenta una predittività nulla, 1 un alto valore predittivo e -1 un valore predittivo inverso. Nel nostro caso (Tab. 4) viene effettuato per diversi step predittivi cumulativi in cui può essere suddivisa la superficie predittiva. Sulla base di precedenti esperienze (CROCE 2022; CROCE *et al.* 2025) si considera un valore di 0.6 come soglia di predittività ottimale. Nel caso attuale soltanto lo step di valori  $\geq 0.4$  risulta avere una predittività accettabile. Confrontando la percentuale di siti e non-siti correttamente categorizzati per ogni step predittivo è possibile anche stabilire la soglia di corretto funzionamento del modello.

Nel caso in esame la soglia è stabilita allo step di  $\geq 0.3$  ma il modello risulta essere operativo soltanto fino allo step di  $\geq 0.5$ . Questo dato risulta coerente con il risultato del Kvamme's Gain, che restituisce un valore predittivo ottimale solo per lo step di valori di 0.4. Questi dati, insieme ai risultati del calcolo dell'area sotto la curva ROC, ci portano ad affermare che il modello ha uno scarso potere predittivo.

## 2.6 Validazione esterna

I dati ottenuti dalla regressione logistica multivariata possono essere applicati a mappe raster che descrivono le stesse variabili in territori diversi da quello utilizzato per il calcolo del modello. Implementando i risultati della regressione logistica multivariata su un'area esterna a quella utilizzata per la sua creazione, contenente siti archeologici noti, è possibile testare in modo empirico il potere predittivo del modello. Le aree scelte per questa operazione sono i dintorni di Como e Lecco e l'area a S del lago di Garda, comprendente l'anfiteatro morenico benacense e parte della pianura circostante (Fig. 3). Queste aree hanno caratteri morfologici analoghi a quelli del Basso Verbano, determinati dalla presenza di sistemi lacustri di origine glaciale al margine dell'alta pianura padana.

Estraendo da GNA un record di evidenze comprese in queste aree, rasterizzato e ricampionato allo stesso modo di quello del Basso Verbano, e comparando la distribuzione dei loro valori predittivi con i valori dei siti utilizzati per il calcolo della regressione, è possibile valutare la capacità predittiva del modello al di fuori della sua area originale. Il risultato (Fig. 5) dimostra che il modello riesce a predire, entro i limiti di performance sopra descritti, soltanto i siti dell'area gardesana. Al contrario, la capacità predittiva per l'area lariana risulta del tutto inadeguata, con una distribuzione dei valori sotto la soglia di predittività ottimale stabilita per il modello originale.

## 2.7 Subset cronologico

Vista la performance di basso livello dimostrata dal modello calcolato sull'intero record di siti archeologici del Basso Verbano, soprattutto se comparata con i risultati di modelli ottenuti in altre aree applicando la stessa metodologia (CARRER 2013; CROCE 2022; CROCE *et al.* 2025), si decide di approfondire l'analisi del dato. Il dataset utilizzato è composto da elementi eterogenei per modalità di raccolta dei dati, tipologia strutturale e collocazione cronologica. Analizzare tutte le sottocategorie generabili dal filtraggio incrociato del dato esula dalle prospettive di studio attuali, che intendono effettuare solo un'analisi preliminare dei dati e dei metodi utilizzabili. Per il presente ci si è quindi concentrati soltanto sul sezionamento cronologico del dato archeologico. Si ipotizza infatti che dal punto di vista diacronico

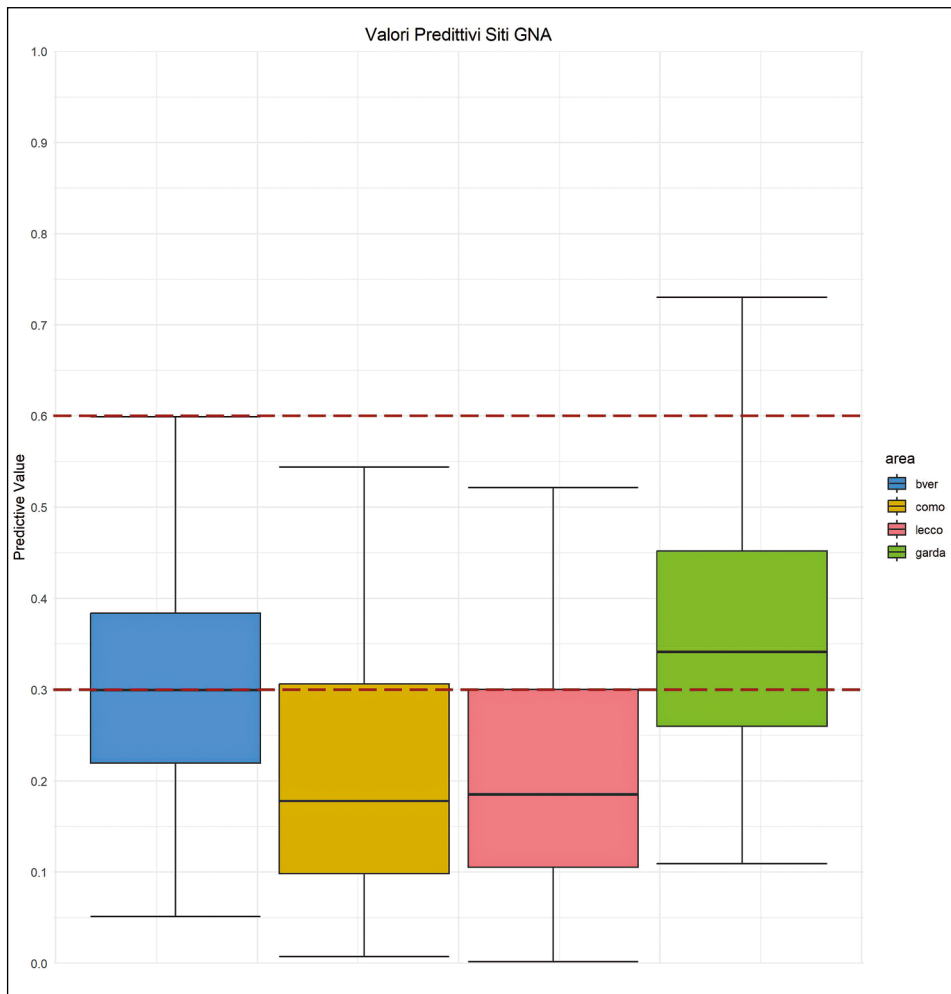


Fig. 5 – Plot dei valori predittivi dei siti del Basso Verbano, comparati con i valori dei siti dell’area lariana e gardesana. Le superfici predittive di riferimento sono state calcolate utilizzando i risultati del calcolo di regressione logistica multivariata effettuato per l’area del Basso Verbano.

possano esistere differenti dinamiche di popolamento, esprimibili tramite modelli distinti e con selezione differenziata delle variabili.

I dati sono stati suddivisi in quattro sottocampioni cronologici, corrispondenti ad età del Ferro, età romana, Tardoantico e Medioevo. I risultati (Tab. 3) dimostrano che sussistono effettivamente modalità diverse di approccio al territorio per i periodi designati, descrivibili con modelli

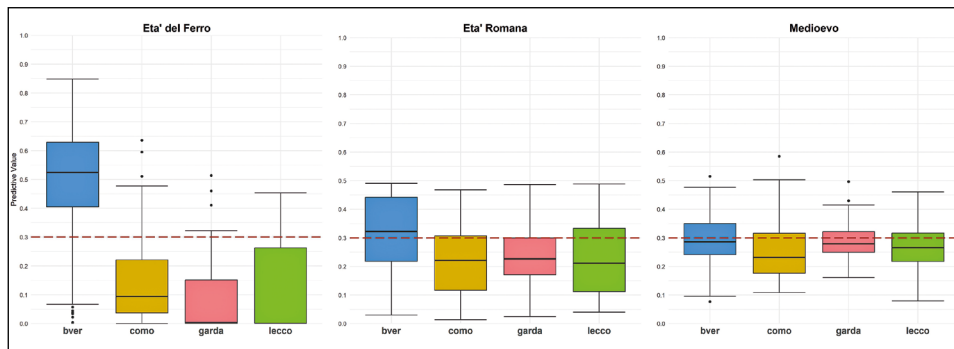


Fig. 6 – Valori predittivi dei siti del Basso Verbano e delle aree di controllo del Lario e del Garda, relativi alle superfici predittive calcolate per differenti intervalli cronologici. I dati relativi al periodo Tardoantico sono stati omessi in quanto non sono disponibili in GNA evidenze riferibili a questa classificazione cronologica per l'area lariana e gardesana.

differenti. Si notano anche delle sostanziali differenze a livello di performance interna dei vari modelli. Il modello più accurato e con migliore potenziale predittivo è quello ascrivibile al popolamento dell'età del Ferro, comprendente una ridotta percentuale dei siti del record di base (28,1%). Il modello relativo all'età romana, che utilizza la maggior percentuale di siti del record generale (45,6%) è quello che restituisce anche i risultati più simili al modello di base. I modelli del Tardoantico e del Medioevo sono basati su un minore numero di siti (rispettivamente 3,4% e 12,5% del totale) e restituiscono quindi dei risultati meno solidi dal punto di vista statistico. Per quanto riguarda il potere predittivo, nessun modello risulta capace di individuare in modo accurato il pattern di siti coevi nelle aree di controllo (Fig. 6).

### 3. DISCUSSIONE

I risultati delle analisi condotte sui siti del Basso Verbano evidenziano i limiti della metodologia di modellazione predittiva di tipo induttivo applicata al record archeologico, anche in presenza di dataset di dimensioni consistenti. Per poter precisare in modo migliore l'affidabilità di questo campione statistico andrebbero analizzate in modo più approfondito le modalità di raccolta del dato, che in molti casi sembra suddiviso in cluster forse attribuibili più alle dinamiche della ricerca sul campo e della gestione dei dati pubblici, piuttosto che ad un'effettiva agglomerazione del popolamento antico. Tuttavia, un'analisi del genere richiederebbe uno sforzo che esula dallo scopo specifico della ricerca qui presentata, dovendosi relazionare con tematiche molto più ampie e relative allo stesso impianto teorico alla base di GNA.

Per quanto concerne i risultati dell'analisi qui proposta, il sezionamento cronologico dimostra comunque che l'uso diacronico del territorio potrebbe essere considerato esso stesso una variabile di importanza fondamentale per la modellazione statistica in ambito archeologico. Il buon funzionamento del modello creato a partire dai siti dell'età del Ferro comparato con i modelli relativi alle epoche successive permette di ragionare sulla validità delle variabili quantitative in relazione alle reali dinamiche di popolamento e pone delle questioni sulla possibile assenza di variabili rilevanti. Inoltre, a prescindere dalla performance predittiva dei singoli modelli è interessante evidenziare come anche soltanto la stessa selezione delle variabili sia un dato utilizzabile per l'interpretazione delle dinamiche diacroniche di interrelazione tra umani e ambiente.

### 3.1 *La scelta delle variabili*

La ricerca qui presentata ha fatto uso di un dataset comprendente un elevato numero di siti, che dovrebbero rappresentare il livello massimo della conoscenza archeologica dell'area, o quantomeno il livello massimo della conoscenza resa pubblica al momento dell'analisi. Il risultato ottenuto tramite l'analisi di questo dato ha una scarsa performance a livello statistico e dimostra l'inadeguatezza di un dataset archeologico eterogeneo, seppur numeroso e ben distribuito sul territorio, come campione di riferimento per il calcolo di modelli predittivi di tipo induttivo. Il subset del dataset a livello cronologico restituisce però almeno un risultato positivo, relativo al pattern dei siti databili all'età del Ferro. Le variabili indipendenti non possono rappresentare tutti i fattori determinanti nelle scelte insediative umane, ma almeno nel caso dell'età del Ferro sembrano rappresentare un campione significativo di questi fattori, capace di descrivere in modo abbastanza accurato il pattern insediativo. Per quanto riguarda le epoche successive invece, pur mutando la selezione delle variabili, le loro interazioni si rivelano troppo deboli per spiegare in modo solido le dinamiche insediative antropiche.

Questa carenza non può essere data dall'inadeguatezza dello strumento statistico, come dimostrano i buoni risultati dell'applicazione della tecnica all'età del Ferro. Sicuramente la natura stessa della variabile dipendente, cioè i siti noti, influenza il risultato, essendo affetta dai bias sopra descritti. Ma dobbiamo ipotizzare che queste criticità influenzino il record di siti in maniera uniforme e quindi, alla luce dei risultati ottenuti per l'età del Ferro, non possiamo ascrivere ad esse la bassa performance del modello. Riguardo la scelta delle variabili indipendenti, è stato dimostrato come l'assenza anche di una singola variabile può influire in modo determinante sulla qualità dei risultati di un modello predittivo di tipo induttivo (CROCE *et al.* 2025). La bassa performance dei modelli successivi all'età del Ferro potrebbe quindi essere

correlabile con l’assenza di una variabile determinante per la strutturazione antropica del territorio.

Anche solo ad un’analisi superficiale del contesto storico-archeologico locale possiamo affermare con certezza che almeno una variabile è stata sicuramente esclusa: la viabilità. Allo stato attuale della ricerca risulta infatti ancora difficoltoso ricostruire precisamente l’entità di questo elemento per tutta l’area in analisi. Del percorso che da *Mediolanum* doveva condurre al lago Maggiore si sono riconosciuti sul terreno labili tratti viari, frutto di indagini d’emergenza nei comuni di Somma Lombardo (SIMONE 1985) e Rho, altri sono solo supposti per la presenza di necropoli romane come ad Arsago Seprio (BINAGHI 1993). Anche le fonti scritte risultano inadeguate, concentrandosi maggiormente sulle descrizioni di Milano e dintorni. A questi dati parziali va ad aggiungersi l’impossibilità di riconoscere o ricostruire un percorso antico partendo dalla cartografia attuale, dato il forte livello di antropizzazione dell’area. Soltanto l’intensificarsi di indagini di terreno o tramite remote sensing potrebbe risolvere questo nodo interpretativo.

### 3.2 *Predittività e interpretazione del territorio*

L’applicazione dei risultati del modello ad altre aree, soprattutto per l’età del Ferro, dimostra che la validazione positiva interna di un modello non si accompagna in modo obbligato alla possibilità di generalizzare i suoi risultati. Se si sposta invece l’attenzione dalla capacità predittiva al potenziale analitico dei modelli, si nota come sia possibile utilizzare i risultati dei calcoli di regressione per indagare le dinamiche insediative locali.

Le variabili selezionate per l’età del Ferro descrivono un’occupazione a ridosso delle più importanti fonti idriche della zona, che sono probabilmente anche utilizzate come via di comunicazione preferenziale. A questo parametro si aggiungono caratteristiche correlate con l’esposizione, l’altitudine e la pendenza dei versanti, che determinano la scelta di luoghi favorevoli all’insediamento in termini di facilità di accesso e di condizioni ottimali di irraggiamento solare e umidità. In epoca romana le scelte sembrano mutare sensibilmente, pur con un risultato finale probabilmente inficiato dall’assenza della variabile che descrive la viabilità. La via d’acqua del Ticino/Verbanò rimane comunque significativa, probabilmente con un duplice valore di fonte d’approvvigionamento idrica e via commerciale per l’Europa centrale, ma è affiancata solo dall’esposizione solare NS. In questo contesto sembrano venire meno tutte le variabili correlate con l’accessibilità e l’altitudine. Per quanto riguarda il Tardoantico, il campione di siti si presenta decisamente sottorappresentato in GNA e non riteniamo sia utilizzabile per una disamina di tipo interpretativo. Per l’epoca medievale la distanza dalle vie d’acqua non è più un fattore determinante, ma i siti vanno invece a posizionarsi all’interno di un sistema in cui il rilievo assume la posizione

dominante. Il pattern insediativo medievale sembra quindi prediligere delle posizioni che favoriscono la visibilità sul territorio e la difendibilità, in una strutturazione che sembra andare a confermare quanto si evince dalla letteratura di riferimento.

Il risultato dell'analisi risulta quindi in linea con le interpretazioni correnti delle dinamiche di popolamento, derivate dall'analisi storico-archeologica delle evidenze. Lo scarso potere predittivo viene quindi bilanciato dall'apparente coerenza interpretativa delle associazioni di variabili selezionate per i singoli periodi. In quest'ottica i modelli di questo genere possono ancora rappresentare uno strumento utile, seppur non risolutivo, per l'indagine dei paesaggi archeologici.

### 3.3 GNA: limiti e potenzialità

I risultati della ricerca hanno evidenziato come il dataset messo a disposizione dal GNA sia uno strumento dal potenziale enorme per l'analisi geospaziale, in quanto fornisce un record di dati archeologici esteso a livello nazionale e, auspicabilmente, continuamente aggiornato con i risultati di tutti gli ambiti della disciplina archeologica, dagli approcci preventivi fino alla ricerca accademica. Va comunque precisato quanto il dataset GNA, allo stato attuale dello sviluppo, seppur efficace per fini di tutela e conservazione (che rimangono il suo scopo primario), abbia delle lacune e ridondanze che andrebbero colmate o aggirate per permettere un suo uso in ambiti di ricerca geostatistica e di archeologia del paesaggio. La possibilità di rieditare il dato all'interno di GNA o di rilavorarlo direttamente sulla piattaforma avendo modo di distinguere un sito come entità organica, composta da differenti rinvenimenti ed evidenze singole, collegabili a interventi separati, agevolerebbe l'interpretazione statistica del dato e permetterebbe anche un suo migliore utilizzo ai fini della ricerca e della disseminazione.

Un ultimo aspetto da non sottovalutare è una certa aleatorietà del trattamento del dato cronologico. In GNA, al momento della consultazione, risultava obbligatorio inserire solo un dato cronologico nominale (ad es. "età del Ferro", "Epoca Romana", etc.), che si basa su una valutazione spesso soggettiva e non ancorata a dati quantitativi. In questo scenario molti siti tardoantichi sono stati classificati come appartenenti generalmente all'epoca romana, senza possibilità di riformattare il dato nel caso non siano presenti altre attribuzioni cronologiche più precise. Nel momento in cui scriviamo vengono modificate le modalità di inserimento dei dati cronologici in GNA: un deciso miglioramento che porterà in futuro ad una maggiore fruibilità del dato per scopi geostatistici. Si auspica che questa modifica sia applicata non solo ai nuovi dati inseriti, ma che si propaghi in modo capillare anche al dato preesistente. Una delle potenzialità di GNA nell'ambito dell'analisi spaziale

risiede infatti nella possibilità di avere a disposizione un dataset completo e affidabile, che non necessiti di ulteriore revisione da parte di chi si occupa della ricerca.

#### 4. CONCLUSIONI

I risultati delle ricerche effettuate a livello geostatistico nell’area del Basso Verbano sembrano confermare le critiche portate alla metodologia di modellazione predittiva di tipo induttivo, soprattutto per quanto concerne l’inaffidabilità del record archeologico come campione statistico e la scelta delle variabili. Tuttavia, se di questa metodologia si sottolinea soprattutto il potenziale interpretativo a livello locale, piuttosto che la capacità di generalizzazione dei risultati, appare chiaro che si è di fronte ad uno strumento con delle notevoli potenzialità nell’ambito dell’indagine del rapporto tra esseri umani e ambiente naturale nel passato, soprattutto in una prospettiva di ricerca diacronica. Sottolineare i limiti e le potenzialità di questo strumento deve essere uno sprone verso l’implementazione di nuovi approcci che, pur rimanendo nell’alveo dell’analisi quantitativa, possano dialogare in modo più completo con la complessità dei paesaggi archeologici.

I dati del Basso Verbano hanno rivelato come molte dinamiche occupazionali umane, mutevoli nel corso dei secoli, siano comprensibili e interpretabili anche con un approccio strettamente quantitativo. Si dovrà però avere cura di selezionare correttamente le variabili in gioco e di calibrare gli strumenti sulle specificità locali dei contesti di studio, quantificando anche l’importanza di fenomeni del tutto antropici, come le infrastrutture viarie e le dinamiche socioculturali. Appare quindi chiaro che una modellazione con finalità prettamente predittive, in ambito archeologico, debba essere abbandonata in favore di altri approcci che abbiano come scopo primario la comprensione delle relazioni complesse che strutturano l’evoluzione del paesaggio.

ENRICO CROCE, AMEDEO DE LISI

Dipartimento di Beni Ambientali e Culturali

Università degli Studi di Milano

enrico.croce@unimi.it, amedeo.delisi@unimi.it

#### *Ringraziamenti*

Gli autori hanno contribuito in maniera paritaria alla stesura del presente articolo. Le ricerche sono state condotte e finanziate nell’ambito dei seguenti progetti di ricerca dell’Università degli Studi di Milano, Dipartimento di Beni Ambientali e Culturali: Progetto ricerche e scavi nel Basso Verbano (finanziamento PSR Linea 4, P.I. prof. Emanuele Intagliata); Progetto RuRES: “Rural Resilience. Decentralised Landscapes and Ecological Strategies of Non-elite Groups in Cisalpine Gaul” (PRIN 2022, P.I. prof. Lorenzo Zamboni).

## BIBLIOGRAFIA

- ACCONCIA V., BOI V., FALCONE A., DI COCCO I., SERLORENZI M. 2024, *Dati aperti in archeologia: una riflessione sullo stato dell'arte nell'ambito del Ministero della Cultura*, «Archeologia e Calcolatori», 35.2, 29-38 (<https://doi.org/10.19282/ac.35.2.2024.04>).
- ALBERTI G., GRIMA R., VELLA N. 2018, *The use of geographic information system and 1860s cadastral data to model agricultural suitability before heavy mechanization. A case study from Malta*, «PLoS ONE», 13 (<https://doi.org/10.1371/journal.pone.0192039>).
- BINAGHI L. 1993, *Arsago Seprio, via Roma, Necropoli a incinerazione*, «Notiziario della Soprintendenza Archeologica della Lombardia», 71.
- BOI V. 2024, *Il Geoportale Nazionale per l'Archeologia (GNA). Standardizzazione e apertura dei dati* ([https://doi.org/10.60974/GNA\\_02](https://doi.org/10.60974/GNA_02)).
- BRANDOLINI F., CARRER F. 2020, *Terra, silva et paludes. Assessing the role of alluvial geomorphology for late-Holocene settlement strategies (Po Plain, N Italy) through Point Pattern Analysis*, «Environmental Archaeology» (<https://doi.org/10.1080/14614103.2020.1740866>).
- CALANDRA E. 2022, *Il Geoportale Nazionale per l'Archeologia*, in *L'archeologia preventiva nel quadro del recovery plan*, Roma, Accademia dei Lincei, 71-77.
- CARLSON D.L. 2017, *Quantitative Methods in Archaeology Using R*, Cambridge, University Press.
- CARRER F. 2013, *An ethnoarchaeological inductive model for predicting archaeological site location: A case-study of pastoral settlement patterns in the Val di Fiemme and Val di Sole (Trentino, Italian Alps)*, «Journal of Anthropological Archaeology», 32, 54-62 (<https://doi.org/10.1016/j.jaa.2012.10.001>).
- CLAESKENS G., JANSEN M. 2015, *Model selection and model averaging*, in J.D. WRIGHT (ed.), *International Encyclopedia of the Social & Behavioral Sciences*, 2<sup>nd</sup> edition, vol. 15, Elsevier, 647-652.
- CONOLLY J., LAKE M. 2006, *Geographical Information Systems in Archaeology*, Cambridge, Cambridge University Press.
- CROCE E. 2022, *Archeologia d'alta quota alle sorgenti del Brembo*, PhD Thesis, Università di Trento.
- CROCE E., CARRER F. 2024, *SBC Predictive Model (v1.0.1) - Dataset*.
- CROCE E., CARRER F., ANGELUCCI D.E. 2025, *Ethnoarchaeological inductive predictive model: A field test in the Italian Alps*, «Journal of Archaeological Methods and Theory», 32, 43 (<https://doi.org/10.1007/s10816-025-09712-w>).
- DE MARCHI M. 1999, *Insedimenti longobardi e castelli tardoantichi tra Ticino e Mincio*, in G.P. BROGIOLO (ed.), *Le fortificazioni del Garda e i sistemi di difesa dell'Italia settentrionale tra Tardoantico e Altomedioevo. 2° Convegno archeologico del Garda (Gardone Riviera 1998)*, Quingentole (MN), SAP Società Archeologica, 109-136.
- DE REU J., BOURGEOIS J., BATS M., ZWERTVAEGHER A., GELORINI V., DE SMEDT P., CHU W., ANTROP M., DE MAEYER P., FINKE P., VAN MEIRVENNE M., VERNIERS J., CROMBÉ P. 2013, *Application of the topographic position index to heterogeneous landscapes*, «Geomorphology», 186, 39-49 (<https://doi.org/10.1016/j.geomorph.2012.12.015>).
- DOLCI M. 2003, *Perviae paucis Alpes, viabilità romana attraverso i valichi delle Alpi Centrali*, Oxford, Archaeopress.
- DORMANN C.F., ELITH J., BACHER S., BUCHMANN C., CARL G., CARRÉ G., GARCÍA MARQUÉZ J.R., GRUBE B., LAFOURCADE B., LEITÃO P.J., MÜNKEMÜLLER T., MCCLEAN C., OSBORNE P.E., REINEKING B., SCHRÖDER B., SKIDMORE A.K., ZURELL D., LAUTENBACH S. 2013, *Collinearity: A review of methods to deal with it and a simulation study evaluating their performance*, «Ecography», 36, 27-46 (<https://doi.org/10.1111/j.1600-0587.2012.07348.x>).

- GABUCCI A. 2024, *Un template QGIS al servizio del Geoportale Nazionale per l'Archeologia (GNA)*, 1-5 ([https://doi.org/10.60974/GNA\\_05](https://doi.org/10.60974/GNA_05)).
- HOSMER D.W., LEMESHOW S., STURDIVANT R. 2013, *Applied Logistic Regression*, III, Hoboken, John Wiley & Sons.
- KING D., BOURENNANE H., ISAMBERT M., MACAIRE J.J. 1999, *Relationship of the presence of a non-calcareous clay-loam horizon to DEM attributes in a gently sloping area*, «Geoderma», 89, 95-111 ([https://doi.org/10.1016/S0016-7061\(98\)00124-4](https://doi.org/10.1016/S0016-7061(98)00124-4)).
- KVAMME K.L. 1988, *Development and testing of quantitative models*, in W.J. JUDGE, L. SEBASTIAN (eds.), *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modeling*, Washington, US Government Printing Office, 325-428.
- KVAMME K.L. 2020, *Analysing regional environmental relationship*, in M. GILLINGS, P. HACIGÜZELLER, G. LOCK (eds.), *Archaeological Spatial Analysis*, Oxon-New York, Routledge, 212-230.
- VAN LEUSEN M., DEEBEN J., HALLEWAS D., ZOETBROOD P., KAMERMANS H., VERHAGEN P. 2005, *A baseline for predictive modelling in the Netherlands*, in P.M. VAN LEUSEN, H. KAMERMANS (eds.), *Predictive Modeling for Archaeological Heritage Management: A Research Agenda*, Amersfoort, Rijksdienst voor het Oudheidkundig Bodemonderzoek, 27-94.
- MIDI H., SARKAR S.K., RANA S. 2010, *Collinearity diagnostics of binary logistic regression model*, «Journal of Interdisciplinary Mathematics», 13, 253-267 (<https://doi.org/10.1080/09720502.2010.10700699>).
- NETLER M., MITASOVA H. 2008, *Open Source GIS: A Grass GIS approach*, III, New York, Springer.
- OLAYA V. 2009, *Basic land-surface parameters*, in T. HENGL, H.I. REUTER (eds.), *Geomorphometry. Concepts, Software, Applications*, Elsevier, 142-169.
- SIMONE L. 1985, *Strada romana*, «Notiziario della Soprintendenza Archeologica della Lombardia», 56.
- TOBLER W.R. 1970, *A computer movie simulating urban growth in the Detroit region*, «Economic Geography», 46, 234-240 (<https://doi.org/10.2307/143141>).
- VENABLES W.N., SMITH D.M., R CORE TEAM 2024, *An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics*. Electronic edition (<http://cran.r-project.org/doc/manuals/R-intro.html>).
- VERHAGEN P., WHITLEY T.G. 2020, *Predictive spatial modelling*, in M. GILLINGS, P. HACIGÜZELLER, G. LOCK (eds.), *Archaeological Spatial Analysis*, Oxon-New York, Routledge, 231-246.
- WHEATLEY D., GILLINGS M. 2002, *Spatial Technology and Archaeology: The Archaeological Applications of GIS*, London, CRC Press.

## ABSTRACT

The work presented in this paper investigates the settlement dynamics of the Lower Verbano area between the Iron Age and the Middle Ages. The known archaeological sites, obtained from the public GNA database, were used along with a series of physical characteristics of the territory as variables for the calculation of an inductive predictive model. The results demonstrate the analytical potential of this methodology in the field of archaeological landscape analysis, highlighting also its shortcomings in strictly predictive terms. The model created for the general record of archaeological sites in the GNA proves to be underperforming from a predictive standpoint, whereas heterogeneous results were obtained when the sample of sites was selected chronologically. The model calculated for the Iron Age shows a high discriminatory power and, when compared to the models for the subsequent periods, under-

scores the importance of the selection of variables for this methodological approach. The use of data from a public database revealed several deficiencies in the management of complex territorial information inherent in its structure, but also highlighted the intrinsic potential of such a tool, which we hope to fully develop in the future.