MapPapers 4-I, 2011, pp.68-76
doi:10.4456/MAPPA.2011.07

Modelli matematici per la determinazione del potenziale archeologico

Bini D., Dubbini N., Steffè S.

Imitando il modo di procedere nella pratica dell'equipe archeologica e geologica per la determinazione del potenziale archeologico, i modelli di tipo 'page rank', con i quali attualmente vengono classificate le pagine web dai motori di ricerca, sembrano essere adatti per la determinazione del potenziale archeologico dell'area urbana pisana. Questo perché l'importanza di un ritrovamento non è assoluta, ma dipende in maniera sostanziale dai ritrovamenti 'vicini', che gli attribuiscono importanza: è lo stesso criterio che viene implementato nel caso del page rank. In questo report quindi descriveremo brevemente il modello page rank classico, mostrando come potrebbe essere adattato al problema della determinazione del potenziale archeologico. Inoltre presenteremo anche un modello, piu semplice, basato su uno smoothing dei dati disponibili, che servirà da confronto con il modello basato sul page rank. I due diversi modelli verranno testati mediante simulazioni.

Keywords: page rank, potenziale archeologico, teorema di Perron-Frobenius, matrici sparse, smoothing

Introduzione

Dalle discussioni avute all'interno del nostro gruppo, nelle quali l'equipe matematica si è confrontata con i gruppi archeologico e geologico, è emerso un parallelo tra i criteri che si usano nella pratica per attribuire il potenziale archeologico e i criteri secondo cui viene assegnata l'importanza delle pagine web negli algoritmi dei motori di ricerca.

Cerchiamo di rendere palese questo parallelo. Durante la determinazione del potenziale archeologico i dati di natura geo-morfologica, e di natura archeologica, vengono integrati per assegnare le unità stratigrafiche, i.e. le tracce delle azioni umane o degli eventi naturali vengono progressivamente accorpate con un processo di sintesi interpretativa (attività, gruppi di attività) fino a definire una carta dei ritrovamenti di periodo che consenta, con un lavoro di analisi, la realizzazione della carta di potenziale per un dato periodo storico. Infine i diversi periodi storici vengono 'sovrapposti' per arrivare al risultato finale. Già il processo di interpretazione delle unità stratigrafiche consiste nel dedurre da una serie di dati archeologici e geologici una categorizzazione stratigrafica sulla base delle relazioni spaziali e funzionali che ci sono tra i vari ritrovamenti. Questo processo diventa sempre più evidente man mano che si procede con la determinazione delle attività e dei gruppi di attività, e poi con la costruzione della mappa di potenziale per periodo archeologico. Per esempio un'attività stratigrafica semplice potrebbe essere l'erezione di una colonna, che quindi viene dedotta a partire dai resti della colonna stessa, dalla presenza di un basamento, dalla presenza di una preparazione pavimentale nei dintorni. A loro volta queste attività possono portare alla definizione di un gruppo di attività, come per esempio una domus: più colonne, insieme a strutture pavimentali, muri, finestre, ecc, adeguatamente collocate le une rispetto alle altre, portano alla deduzione della presenza di una domus. A loro volta più domus, insieme ad abitazioni ed elementi di altro tipo – come templi, strade, o discariche per i rifiuti, ecc. - fanno sì che l'archeologo ne deduca la presenza di un complesso abitativo.

Il nodo cruciale di queste analisi dal punto di vista astratto, e quindi lasciando da parte il know how e l'esperienza del personale archeologico e geologico, è l'individuazione delle *relazioni* che intercorrono tra i vari ritrovamenti, sia dal punto di vista *spaziale* (cioè la collocazione nello spazio), sia dal punto di vista *funzionale* (cioè qual è o potrebbe essere la loro funzione). Detto in altre parole, la presenza di un particolare tipo di ritrovamento vicino ad un ritrovamento già scoperto potrebbe rafforzare o attenuare la probabilità che questi vadano a formare una struttura di livello più complesso, e quindi rafforzare o attenuare il potenziale archeologico dell'area stessa. Questo è appunto il criterio alla base degli algoritmi di page ranking, nei quali ogni pagina web attribuisce importanza alle pagine web a cui punta (mediante un link, potremmo quindi dire 'vicina'), e a sua volta assume importanza dalle pagine web dalle quali riceve un link.

Il report è così organizzato. Nella sezione 2 descriveremo brevemente i due modelli principalmente utilizzati in letteratura per la determinazione del potenziale archeologico, quello basato sulla map algebra, e quello basato sulla regressione. Nella sezione 3 presenteremo il modello basato sul page rank, partendo da quello classico, indicando le modifiche che andranno fatte per adattarlo alla determinazione del potenziale archeologico, e mostrando alcune simulazioni. Nella sezione 4 presenteremo un modello per la determinazione del potenziale archeologico alternativo, più semplice, basato sullo smoothing, con alcune simulazioni, che servirà da confronto con il modello basato sul page rank. Infine, nella sezione 5 faremo alcune considerazioni circa il confronto tra i vari modelli e procederemo alle conclusioni.

2. Modelli esistenti in letteratura

2.1 Approccio basato sulla map algebra

Uno dei modi utilizzati in letteratura, probabilmente il più semplice, per predire il potenziale archeologico consiste in un modello predittivo capace di generare una decision rule. Gli input per determinare una tale regola possono essere ad esempio la conformazione del terreno (pianura/pendenza), la presenza di fonti d'acqua vicine, o il tipo di terreno. Caratteristiche come queste posso essere combinate in regole del tipo

$$(pendenza \geq 10) \cap (distanza\ \ fonte \geq 1 \quad km) \cup (suolo \neq A)$$

per predire la presenza o l'assenza di un sito archeologico. Alcune varianti di questo modo di procedere possono includere l'assegnare dei 'pesi' alle diverse condizioni, in modo tale da dare più importanza ad alcune e meno ad altre. Allo stesso tempo, si possono utilizzare dei test di significatività per valutare se il modello predittivo proposto può essere associato alla presenza di siti archeologici entro un certo intervallo di confidenza. Per questo tipo di modelli si veda (Cumming 1997; Wheatley 2002).

I modelli basati su regole siffatte sono molto facili da implementare, ma forniscono risultati del tipo on/off – per esempio presenza o assenza di un sito archeologico - e non vanno molto oltre la semplice giustapposizione di alcune semplici regole. Detto in altre parole, non sfruttano affatto la potenza di un modello matematico o la capacità di calcolo di un computer.

2.2 Approccio basato sulla regressione

Un altro modo di procedere in letteratura per la determinazione del potenziale archeologico (Wheatley 2002) è basato sull'applicazione della regressione lineare (o logistica). Questo genere di approccio deriva dell'esigenza di rispondere ad alcune questioni alle quali il metodo descritto precedentemente non può rispondere. Per esempio:

• In che modo un predittore influenza il modello?

• In che modo possono essere predette grandezze *continue* anziché discrete?

La regressione lineare può essere usata per rispondere a questo genere di domande, ed è inoltre particolarmente facile sia da implementare, sia da un punto di vista matematico. Senza entrare nei dettagli, tutti le regressioni lineari producono equazioni del tipo

$$y = a + b_1 x_1 + \ldots + b_k x_k ,$$

dove y è la variabile che deve essere predettta (per esempio il potenziale archeologico), e $x_1, \ldots, x_k$ sono gli input della regressione lineare. In altre parole, stimando i coefficienti $a, b_1, \ldots, b_k$ sulla base dei dati disponibili, si trova un valore di y, che sarà quello utilizzato per fare predizione. Anche nei modelli regressivi si possono introdurre molte varianti (regressioni singole o multiple, di tipo diverso dal lineare, statistiche, ecc.). Per un maggiore approfondimento sulla regressione lineare, anche in relazione all'archeologia, il lettore può consultare (Shennan 1997; Wescott 2000).

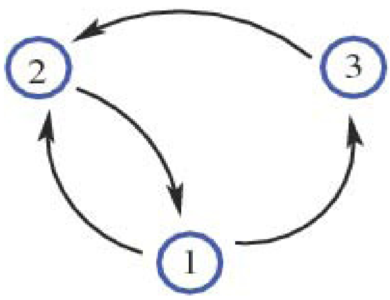
Se gli approcci basati sulla regressione lineare hanno il vantaggio di utilizzare delle variabili per predirne altre, il modello che implementano è troppo semplice, e non riesce a tenere conto della grande complessità di cui bisogna tenere conto nella determinazione del potenziale archeologico. Questo è talmente vero che tutt'ora i modelli basati sulla regressione lineare (o logistica) non sono a volte nemmeno preferiti a quelli basati sulla map algebra.

3. Determinazione del potenziale basata sul page rank

3.1 Modello classico di page rank

In questo paragrafo descriviamo le idee generali e alcuni dettagli dei modelli matematici più usati per attribuire un valore di importanza alle pagine presenti sul Web indipendentemente dal valore del loro contenuto e *in funzione unicamente delle interconnessioni tra le pagine*. Per maggiori dettagli ci si riferisce a (Langville 2006).

Assumiamo di avere *n* pagine web e numeriamole con gli interi da *1* a *n*. Per descrivere il World-Wide Web è utile usare un grafo orientato in cui i nodi rappresentano le pagine presenti sul Web e gli archi orientati descrivono le connessioni di tali pagine. Più precisamente un arco collega il nodo *i* col nodo *j* se la pagina *i* ha un link che punta alla pagina *j*. Ad esempio se il nostro WWW fosse fatto da 3 pagine in cui la pagina 1 punta alla 2 e alla 3, la pagina 2 punta alla 1 e la 3 punta alla 2, allora il grafo sarebbe



Un grafo orientato può essere univocamente descritto da una matrice di adiacenza $H = h_{ij}$ di dimensione $n \times n$ in cui $h_{ij} = 1$ se c'è un arco orientato che collega il nodo *i* col nodo *j* (se la pagina *i* contiene un link alla pagina *j*); mentre $h_{ij} = 0$ altrimenti. La matrice di adiacenza associata al grafo di sopra è

$$H = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Il criterio usato per determinare l'importanza delle pagine web, può essere riassunto nella seguente: una pagina *i* che punta altre pagine, ad esempio $j_1, \dots, j_k$, distribuisce la sua importanza in parti uguali alle pagine $j_1, \dots, j_k$, e quindi dà *1/k* della sua importanza alle pagine che punta.

In questo modello, se denotiamo con $d_i = \sum_{i=1}^{n} h_{ij}$ supponendo $d_i \neq 0$ per *i=1,...,n,* e se indichiamo con w_j l'importanza della pagina *j*, vale allora

$$w_j = \sum_{i=1}^{n} w_i \frac{h_{ij}}{d_i}, \qquad j = 1, \dots, n.$$

Ad esempio nel caso del grafo d'esempio si ha

$$\begin{cases} w_1 = w_2 \\ w_2 = 1/2 w_1 + w_3 \\ w_3 = 1/2 w_1 \end{cases}$$

Si osserva che questo non è altro che un problema di autovalori e autovettori formulato nel seguente modo. Posto $e = (1,1,\dots,1)^T$, d=He, e *D=diag(d)*, si ha

$$w^T M = w^T, \qquad M = D^{-1} H,$$

dove $w^T = (w_1, \dots, w_n)$. In questa formulazione naïve però si riscontrano alcuni problemi tecnici, che elenchiamo:

1. Cosa succede se $d_i = 0$ per qualche *i*? Questo succede nei casi in cui ci sono pagine che non puntano a nulla. Il problema non è insolito, infatti ci possono essere pagine, quali un file postscript, che non hanno link a nulla. I nodi che hanno questa caratteristica sono chiamati dangling nodes
2. Esiste sempre una soluzione?
3. La soluzione è unica (a meno di multipli scalari)?
4. La soluzione è positiva?
5. Come si può calcolare?

Si osserva che i dangling nodes sono individuati per il fatto che essi corrispondono alle righe di *H* con tutti gli elementi nulli. Per far fronte al problema dei dangling nodes si introduce una leggera modifica al modello. Più precisamente si sostituisce la matrice iniziale di adiacenza *H* con una nuova matrice $\hat{H}$ che coincide con *H* dappertutto eccetto che nelle righe tutte nulle in cui gli elementi di $\hat{H}$ vengono posti tutti uguali a *1*. Dal punto di vista modellistico è come assumere salomonicamente che un documento che nel modello originale non cita nessun altro documento nel web, nel nuovo modello modificato va a citare tutti i documenti presenti. Quindi distribuisce *1/n* della sua importanza uniformemente a tutti. La matrice $\hat{H}$ viene quindi scritta come

$$\hat{H} = H + u e^T$$

dove *u* è il vettore con componente *1* in corrispondenza dei dangling nodes e con componente *0* altrove. In seguito denoteremo con *M* la matrice $M = \hat{D}^{-1}\hat{H}$, dove

$$\hat{D} = diag(\hat{d}), \; \hat{d} = \hat{H}e.$$

Possiamo inoltre dare subito risposta affermativa alla domanda 2 osservando che *Me=e* e quindi *1* è autovalore, e quindi *w* è un qualsiasi autovettore sinistro corrispondente all'autovalore *1*. Per rispondere alle altre domande dobbiamo riportare alcuni risultati classici della teoria di Perron-Frobenius delle matrici non negative.

Teorema: *Sia A una matrice $n \times n$ di elementi non negativi. Allora esiste un autovalore λ di A tale che $\lambda = \rho(A) \geq 0$. Esistono un autovettore destro x e sinistro y corrispondenti a λ con componenti non negative. Se inoltre A è irriducibile allora λ è semplice e gli autovettori x e y hanno componenti positive. Se infine A ha elementi positivi allora λ è l'unico autovalore di modulo massimo.*

Si osserva che in base al teorema di Perron-Frobenius ogni soluzione ha sempre componenti non negative come è giusto che sia. Però la sola condizione di non-negatività non garantisce l'unicità della soluzione (a meno di multipli scalari). Mentre con la condizione di irriducibilità la soluzione è unica. È facile costruire reti di pagine interconnesse che hanno una matrice di adiacenza riducibile. Quindi il modello così come è stato introdotto non è ancora adeguato. Si osserva ancora che nel caso di matrici irriducibili e non negative possono esistere altri autovalori che hanno lo stesso modulo del raggio spettrale. Questo crea dei seri problemi dal punto di vista algoritmico. Per far fronte ai problemi discussi, la matrice M viene sostituita con la matrice

$$A = \gamma M + (1-\gamma) e v^T, \qquad 0 < \gamma < 1,$$

dove *v* è un arbitrario vettore a componenti negative tale che $v^T e = 1$, detto vettore di personalizzazione e γ è un parametro, di solito si sceglie $\gamma = 0.85$. In questo modo la matrice *A* ha elementi positivi. La soluzione quindi esiste, è unica (a meno di multipli) e $\rho(A)$ è l'unico autovalore di modulo *1*. Dal punto di vista modellistico è come se l'importanza di una pagina fosse ripartita in due parti: una frazione γ viene distribuita in base ai link come nel modello originale; la frazione complementare *1-*γ viene distribuita a tutte le altre pagine secondo un criterio dato dal vettore *v*. Se ad esempio *v=(1/n)*e allora la distribuzione è fatta in modo uniforme a tutte le pagine del Web.

3.2 Il modello page rank nella determinazione del potenziale archeologico

Per applicare un modello di tipo page rank alla determinazione del potenziale archeologico, suddivideremo il sottosuolo dell'area urbana pisana in $m \times n \times p$ celle tridimensionali in cui le prime due coordinate *i,j* definiscono gli elementi di una sezione orizzontale del terreno e la terza coordinata *k* definisce il deposito considerato. L'obiettivo è dunque assegnare a ciascuna cella *(i,j,k)* un potenziale archeologico espresso da un numero reale non negativo x_{ijk}, per *i=1...,m, j=1,...,n, k=1,...,p*. Una delle condizioni che vorremmo soddisfatte dalla soluzione del nostro modello è che

$$x_{ijk_1} \le x_{ijk_2} \qquad se \;\; k_1 \le k_2$$

La spiegazione di questa proprietà risiede nel fatto che il potenziale archeologico di ogni cella *(i,j,k)* va più propriamente interpretato come il potenziale che si ottiene scavando *dalla superficie, in verticale, fino a quella cella*. Per questo motivo il potenziale archeologico, andando in profondità, non può che aumentare.

La principale modifica da apportare al page rank per adattarlo ad un modello per la determinazione del potenziale archeologico riguarda il criterio per definire la 'vicinanza' tra celle, e cioè per definire l'influenza di ogni singola cella sulle altre celle. Nel modello del page rank classico la zona di influenza di una pagina web era definita dai link che partivano dalla pagina stessa. Nel definire quindi i criteri di influenza delle celle, terremo presenti le seguenti considerazioni, scaturite da approfondite discussioni con le equipe archeologica e geologica.

• Costruire un modello tipo page rank in cui ciascun nodo del grafo, che corrisponde ad una cella di coordinate (i,j,k), viene individuato da un intero *r* tale che $1 \le r \le mnp$. Il numero *r* viene determinato da un ordinamento delle terne *(i,j,k)* di tipo lessicografico;

• Formare una matrice $N \times N$ con *N=mnp*, $H = (h_{rs})$ tale che h_{rs} è la parte di importanza che la cella *r* trasferisce alla cella *s*. Il valore di h_{rs} è compreso tra *0* e *1*;

• Usare l'informazione archeologica disponibile, se disponibile, per ogni cella in modo duplice: da una parte, in *modo relativo* per costruire gli elementi della matrice *H* che regola il trasferimento di importanza delle celle; dall'altra, in *modo assoluto*, dando importanza alla specifica cella in cui è presente un ritrovamento archeologico;

• La parte relativa alla costruzione della matrice che regola il trasferimento di importanza viene svolta in base alle categorie che saranno utilizzate per classificare i ritrovamenti archeologici. In particolare, la categoria caratterizza sia la 'geometria' che i valori della distribuzione di importanza. Ad esempio, una cella che contiene ritrovamenti relativi a una strada può interessare celle contigue disposte lungo una retta;

• Nel costruire la matrice *H*, si mantiene la condizione di normalizzazione $\sum_{j=1}^{N} h_{ij} = 1$, che dice che tutta l'importanza di una cella si conserva nella distribuzione: l'importanza non si amplifica nè si riduce;

• Nella seconda parte in cui si vuole attribuire ad una (o più) componente x_k del potenziale archeologico un valore più alto d_k, si può procedere nei seguenti modi:

  o Si può forzare quella componente ad assumere il valore assegnato. In questo caso il sistema $x^T H = x$ con la condizione di normalizzazione $\sum x_i = 1$ e la condizione $x_k = d_k$ diventa un sistema lineare non omogeneo e sovradeterminato, con più equazioni che incognite. Va quindi trattato con tecniche dei minimi quadrati;

  o Si puo mantenere un approccio omogeneo rinunciando alla normalizzazione (stocasticità) della matrice, cioè scalando le righe di *H* col fattore assegnato. In questo caso la matrice *H* viene sostituita da *A=DH*, $D = diag(d_1, ..., d_N)$ e il problema è ricondotto a calcolare il vettore di

Perron di A, cioè quel vettore non negativo x tale che $x^T A = \rho x^T$, dove $\rho > 0$ è il raggio spettrale di A. Infatti, non essendo più stocastica, la matrice A non è detto abbia raggio spettrale *1*. In questo approccio, scalando le righe pesiamo diversamente l'importanza che ciascuna cella distribuisce alle altre. Si osserva che la condizione che

$\sum_{i=1}^{N} d_{ii} = 1, d_{ii} \geq 0$, non garantisce che il raggio spettrale di A rimanga *1*;

o Si può mantenere ancora un approccio omogeneo scalando le colonne di H col fattore assegnato. In questo caso la matrice H viene sostituita da $A=HD$. In questo approccio viene scalata (amplificata o compressa) l'importanza che riceve ciascuna cella;

• Usare l'informazione geologica in modo binario, cioè considerare o escludere delle celle. Alternativamente usare un peso compreso tra *0* e *1* in modo moltiplicativo;

• La condizione che il potenziale archeologico cresca con l'aumentare della profondità del terreno può essere realizzato in almeno due modi:

o imponendo dei link nel grafo che collegano la cella *(i,j,k)* con la cella *(i,j,k+1)*; in questo modo le celle profonde ricevono più importanza di quelle superficiali;

o adottando come soluzione per il potenziale archeologico le quantità $y_{ijk} = \sum_{h=1}^{k} x_{ijh}$, dove x_{ijh} è il vettore di Perron di H.

3.3 Simulazioni

In attesa del lavoro da condurre insieme agli archeologi per assegnare, in seguito alla categorizzazione dei ritrovamenti, la distribuzione e la geometria dei valori di importanza che un ritrovamento in una cella assegna alle altre celle, in questo paragrafo descriveremo alcune versioni dell'algoritmo di page rank adattato alla determinazione del potenziale archeologico. Questi algoritmi saranno implementati in una versione semplificata, per condurre delle simulazioni.

Per questo motivo in questo report supponiamo che il nostro modello sia monodimensionale. Dopo aver numerato da *1* a *N* le celle del problema tridimensionale, definiamo la matrice $A = a_{ij}$ di dimensione N tale che a_{ij} è la parte di potenziale archeologico che la cella j dà alla cella i. Gli elementi di A sono non negativi e tali che $\sum_{i} a_{ij} = 1$, cioè la matrice A è stocastica. La matrice A viene definita dalle informazioni disponibili. Il problema da risolvere, nella sua *formulazione non omogenea*, diventa dunque

$$Ax = x$$
$$x_j = b_j, \qquad j \in \Delta$$

dove Δ è l'insieme degli indici delle celle su cui si hanno informazioni del potenziale archeologico e b_j sono i valori di tale potenziale. Nella sua *formulazione omogenea* invece il problema assume una delle due seguenti forme

$$DAx = \lambda x, \qquad ADx = \lambda x$$

dove λ è l'autovalore di modulo massimo rispettivamente di DA e di AD, mentre D è la matrice diagonale i cui elementi diagonali danno un peso al potenziale archeologico della cella corrispondente.

Quindi il problema non omogeneo consta di un sistema sovradeterminato da risolvere, ad esempio con la tecnica dei minimi quadrati. Sostituendo i valori noti delle x_j diventa un sistema lineare in N equazioni e N-d incognite, dove d è la cardinalità di Δ. Nel modello omogeneo si ha invece la risoluzione del problema agli autovalori.

Se non si hanno informazioni archeologiche la matrice iniziale dei pesi viene assunta uguale a $1/N \cdot e^T$ cioè la matrice di elementi tutti uguali a *1/N*. La motivazione è che, non essendoci informazione, ciascun elemento dà importanza (molto trascurabile poiché è *1/N*) a tutti gli altri. Questa ipotesi è più vantaggiosa rispetto ad assumere come matrice dei pesi la matrice identità, che sembrerebbe piu naturale, in cui ogni elemento dà importanza solo a se stesso. Il vantaggio di scegliere la matrice dei pesi con tutti gli elementi uguali a *1/N* sta nel fatto che la matrice è positiva e quindi vale il teorema di Perron-Frobenius, che garantisce l'esistenza e unicità della soluzione x. Tra l'altro, in questo caso di mancanza di informazioni, x è il vettore di componenti tutte uguali a *1/N* (se normalizziamo con somma a *1*). È quindi ragionevole dal punto di vista modellistico che tutte le celle abbiano potenziale archeologico uguale tra di loro e quindi trascurabile. Se fossimo partiti dalla matrice identica non ci sarebbe stata unicità della soluzione e il problema sarebbe posto in modo non consistente.

Quando qualche informazione archeologica è presente procediamo su due livelli, come già descritto in precedenza. Da una parte formiamo la matrice A assegnando il valore dei pesi. In secondo luogo invece assegnamo un valore di importanza del ritrovamento in sé: questo viene fatto mediante una moltiplicazione per una matrice diagonale D contenente su ogni elemento della diagonale il valore di importanza associato al ritrovamento nella corrispondente cella. Questa moltiplicazione può essere fatta a destra o a sinistra della matrice A, a seconda che si voglia scaricare il peso di tale valore sulle connessioni che partono o che arrivano dalle celle. Infine si trova l'autovalore di Perron associato alla matrice dei pesi moltiplicata (a destra o a sinistra) per la matrice dei valori di importanza.

In questa simulazione abbiamo scelto *n=100*, ed abbiamo inserito ritrovamenti nelle celle *15, 37, 39, 68*, con importanza rispettivamente di *3, 1.5, 1.7, 2*. Per quanto riguarda la distribuzione dei pesi, abbiamo simulato il caso in cui

• il ritrovamento della cella 15 dia importanza alle celle *3, 4, 5, 6, 7, 8* in misura *1/6*;

• il ritrovamento della cella *37* dia importanza alle celle *45, 47, 49, 51, 53, 55, 57, 59* in misura *1/8*;

• il ritrovamento della cella *39* dia importanza alle celle *46, 48, 50, 52, 54, 56, 58, 61* in misura *1/8*;

• il ritrovamento della cella *68* dia importanza alle celle *13, 14, 15, 16, 25* in misura *1/5*;

Risolvendo i problemi sovradeterminato ed agli autovalori si ottengono i valori del potenziale archeologico come in figura 1, dove abbiamo riportato contemporaneamente le soluzioni ottenute per la matrice stocastica (cioè senza tenere conto dell'importanza dei ritrovamenti, il problema non omogeneo), e per la moltiplicazione a destra e a sinistra per la matrice diagonale con i valori di importanza dei ritrovamenti. Osserviamo che nel caso di matrice stocastica la distribuzione del potenziale archeologico risultante dalla simulazione viene distribuita essenzialmente secondo i pesi assegnati ad ogni cella.

Nel caso in cui assegnamo importanza ai ritrovamenti invece le cose cambiano notevolmente. Considerando la scalatura a sinistra - cioè il caso in cui l'importanza del ritrovamento viene 'scaricata' sui pesi ricevuti - si nota un innalzamento generale del potenziale, accentuato nelle zone in cui i ritrovamenti sono avvenuti: la differenza tra le zone di potenziale archeologico più alto e le zone con potenziale archeologico più basso è molto più marcata che nel caso stocastico. Se facciamo riferimento a questo secondo caso il potenziale più alto risulta nelle celle *13, 14, 15, 16, 25*, le quali ricevono importanza in maniera più 'concentrata' dalla cella *68*, che distribuisce un valore di importanza uguale a *2*. Considerando la scalatura a destra - cioè il caso in cui l'importanza del ritrovamento viene 'scaricata' sui pesi inviati - l'iterazione tra i pesi e l'importanza dei ritrovamenti sembra essere più rilevante nell'assegnare il potenziale archeologico finale. Se facciamo riferimento a questo terzo caso infatti il potenziale più alto risulta nella cella 68, la quale oltre ad avere un valore di importanza del ritrovamento uguale a *2* distribuisce il suo peso in maniera più 'concentrata' rispetto agli altri ritrovamenti.

4. Modello alternativo

In questa sezione descriviamo un modello alternativo per la determinazione del potenziale archeologico a partire dai dati disponibili a partire dagli scavi e dai ritrovamenti già disponibili. Proponiamo tale modello con lo scopo di fornire un 'modello base', più semplice rispetto a quello basato sul page rank, con il quale testare la bontà stessa del metodo basato sul page rank.

Questo modello alternativo infatti non è costruito su considerazioni relative alla pratica archeologica, ma consta di una 'semplice' approssimazione dei dati esistenti: si basa semplicemente sul principio secondo cui il potenziale archeologico tende ad aumentare procedendo da un punto dove questo è più basso ad un punto dove questo è più alto, e viceversa. In questo contesto abbiamo utilizzato la funzione matlab csaps, che fornisce un'approssimazione dei dati mediante splines cubiche.

Nella simulazione che mostriamo abbiamo scelto *n=100,* ed inserito ritrovamenti nelle celle *15, 37, 39, 68*, con importanza rispettivamente di 3/10, 1.5/10, 1.7/10, 2/10. Le celle e le importanze relative dei ritrovamenti sono le stesse della simulazione con i modelli basati sul page rank, per facilitare poi il confronto tra i metodi. La soluzione del problema data dall'approssimazione dei dati forniti è mostrata in figura 2.

Osserviamo che, essendo l'unica condizione che conta in questo caso l'importanza del ritrovamento, lo smoothing non fa altro che costruire una curva adattandola ai punti nei quali i valori (di importanza) sono noti. Per questa simulazione abbiamo utilizzato peraltro uno smoothing vero e proprio (questa possibilità si può modulare mediante un parametro della funzione csaps di Matlab), e non una interpolazione, così da poter supporre i valori di importanza del ritrovamento (che verranno assegnati dall'equipe archeologico-geologica) suscettibili di un certo errore.

5. Comparazione tra modelli e conclusioni

In questa sezione finale proponiamo un confronto dei diversi modelli utilizzati, quelli basati sul page rank e quello basato sullo smoothing. Questo confronto è fatto nella convinzione che i modelli presenti in letteratura, basati sulla map algebra o sulla regressione, non sono adeguati alla determinazione del potenziale archeologico come pensato per il progetto MAPPA, per i motivi già citati nella sezione dedicata alla letteratura esistente. Per eseguire il confronto tra i vari modelli faremo riferimento alla figura 3, la quale riporta contemporaneamente le simulazioni effettuate nelle sezioni precedenti.

Una prima, fondamentale, differenza tra il modello basato sullo smoothing e quello basato sul page rank riguarda la possibilità di 'distribuire' importanza di una cella ad altre celle, cosa possibile soltanto nel modello basato sul page rank. Come abbiamo già avuto occasione di far notare, questo è un modo per assegnare una certa 'probabilità di importanza' in celle dove non c'è stato alcun ritrovamento. Nel processo basato sullo smoothing - ma nemmeno nei modelli citati dalla letteratura, quali quello basato sulla map algebra o quello basato sulla regressione lineare - questa possibilità non è contemplata. Possiamo infatti notare, guardando la figura come il modello basato sullo smoothing (curva in verde) concentri il potenziale archeologico soltanto nelle zone circostanti ai ritrovamenti, mentre i modelli basati sul page rank concentrano il potenziale archeologico anche intorno alle celle che acquisiscono importanza dalle celle con i ritrovamenti.

Un'altra importante differenza riguarda il carattere essenzialmente relativo del modello basato sul page rank, rispetto al carattere essenzialmente assoluto del modello basato sullo smoothing.

Nel modello basato sul page rank, oltre ai pesi, che rappresentano la porzione della propria importanza che ogni cella distribuisce alle altre celle, abbiamo assegnato l'importanza di ogni ritrovamento attribuendo al ritrovamento stesso un valore che ne rappresentasse l'importanza in sé. Questo valore però non è assoluto, bensì relativo rispetto ai valori assegnati agli altri ritrovamenti.

Nel modello basato sullo smoothing invece il primo problema che si pone è quello di assegnare un valore assoluto per il potenziale di ogni cella dove questo sia noto. Questo, oltre ad essere una valutazione in più da fare rispetto al modello basato sul page rank, è anche poco corretto dal punto di vista teorico e/o generale. Infatti il potenziale archeologico, e con esso alcune caratteristiche che concorrono a determinarlo (quali la densità dei ritrovamenti, o la 'rarità' di questi), è di per sé relativo alla zona in esame, ed anche alla dimensione di questa.

5.1 CONCLUSIONI

In conclusione, abbiamo presentato due diversi modelli per stimare il potenziale archeologico a partire dai ritrovamenti già rinvenuti, nella convinzione che i modelli presenti in letteratura, basati fondamentalmente sulla map algebra o sulla rergressione, non siano adeguati per via della loro estrema semplicità.

Il lavoro preliminare da fare per testare sui dati reali questi modelli sarà quello di associare una 'distribuzione di importanza' ad ogni categoria di ritrovamento che sarà indicata dall'equipe archeologica: in queste distribuzioni metteremo l'informazione che ogni ritrovamento dà su ciò che potrebbe essere trovato nei dintorni. Infine le informazioni di tipo geologico, che abbiamo detto verranno prese in considerazione in maniera binaria, non sono state incluse nelle simulazioni, in quanto l'implementazione di questa cosa a livello modellistico è immediata.

Il modello sul quale punteremo l'attenzione sarà quello basato sul page rank, perché permette di prendere in considerazione una serie di caratteristiche che ricalcano ciò che nella pratica gli archeologi fanno per determinare il potenziale archeologico. Il modello basato sullo smoothing invece, il quale prende in considerazione soltanto i ritrovamenti senza dare la possibilità di 'distribuire' l'importanza di questi, servirà da confronto. Abbiamo già messo in evidenza alcuni vantaggi dei modelli basati sul page rank, ma riteniamo che questi vantaggi siano in realtà molti di più di quelli evidenziati in questa fase. Pensiamo ad esempio al fatto che il modello di page rank verrà implementato prima tenendo conto dei ritrovamenti in un singolo periodo, e poi per 'sommare' tutti i vari periodi, ed alle diverse interazioni che possono avere i ritrovamenti di periodi diversi; oppure pensiamo al fatto che la 'distribuzione' di importanza può avvenire a livello di celle, ma anche a livello di oggetti (che quindi sarebbero gruppi di celle), e cioè a vari livelli di complessità.

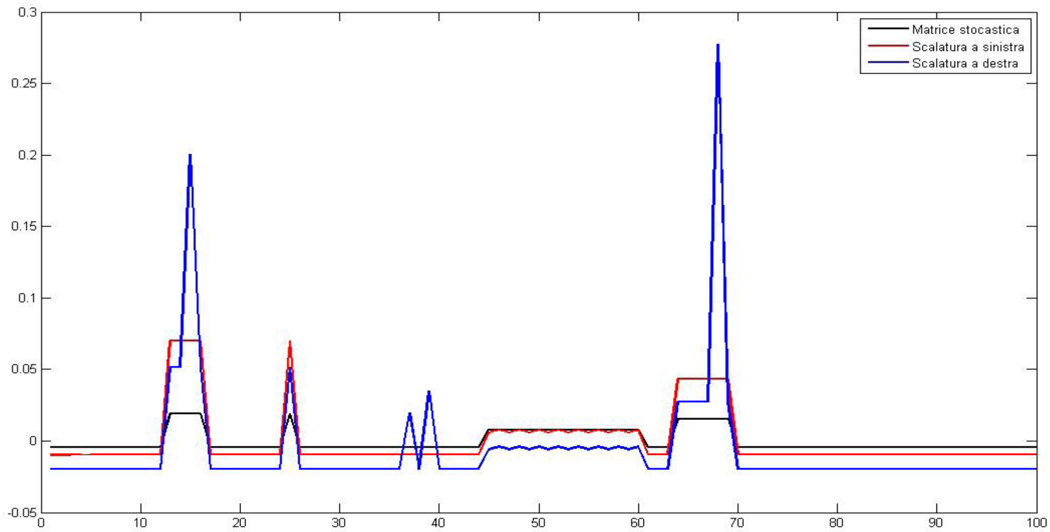


Fig.1 Stime del potenziale archeologico come ottenute dai modelli basati sul page rank.

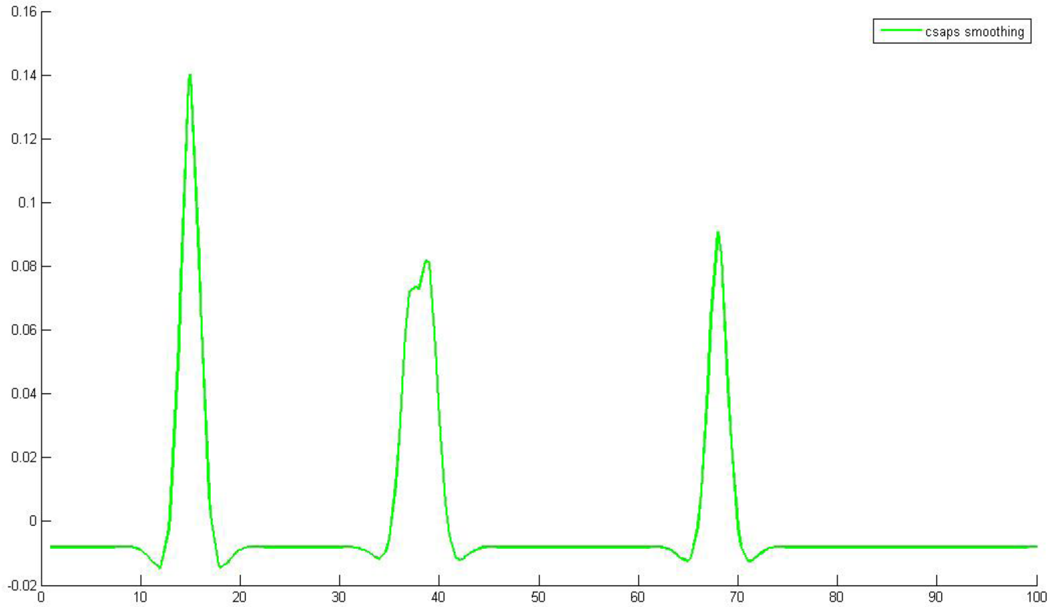


Fig.2 Stima del potenziale archeologico come ottenuta dallo smoothing, implementato dalla funzione matlab csaps.

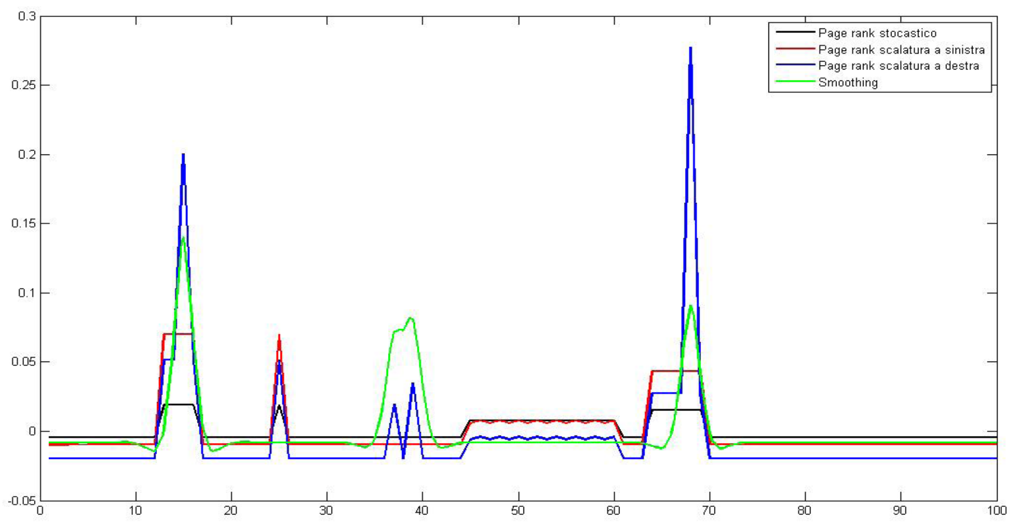


Fig.3 Stime del potenziale archeologico come ottenute da tutti e quattro i modelli implementati.

Bibliografia

CUMMING P.A.B. 1997, *An assessment and enhancement of the sites and monuments record as a predictive tool for cl-tural resouce management, development control and academic research,* Ph.D. thesis, University of Southampton.

SHENNAN S.J. 1997, *Quantifying archaeology*, Edinburgh Univesity Press.

LANGVILLE A.N., MEYER C.D. 2006, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press.

WESCOTT K.L., BRANDON R.J. 2000, *Practical applications of GIS for archaeologists - A predictive modeling kit*, Taylor and Francis.

WHEATLEY D., GILLINGS M. 2002, *Spatial technology and archaeology - the archaeological applictions of GIS, Taylor and Francis.*