Analisi della distribuzione di radon indoor con tecniche di Machine Learning

Pegoretti S.[1] e Verdi L.[2]

[1]Università degli Studi di Trento - Dipartimento di Fisica - via Sommarive 14, 38050 Povo (TN)

[2]APPA Bolzano – via Amba Alagi 5, 39100 Bolzano, luca.verdi@provincia.bz.it

RIASSUNTO

Nell'ambito dell'individuazione delle radon prone areas, risulta importante disporre di uno strumento di previsione affidabile, soprattutto nel caso in cui la variabile di interesse sia la concentrazione di radon "indoor", ovvero quella misurata all'interno degli edifici. In queste situazioni, infatti, le interazioni con le caratteristiche dell'abitazione e una opportuna caratterizzazione della stessa risultano fondamentali.

In questo lavoro vengono messi a confronto (sia a livello globale che locale) i risultati che si sono ottenuti ricorrendo a due differenti approcci per la stima della concentrazione di radon indoor in localizzazioni non campionate: da un lato, un modello di tipo geostatistico basato su kriging ordinario (OK), dall'altro un modello basato su tecniche di Machine Learning (ML).

Nel primo caso, si è considerato un modello volto alla trattazione specifica e raffinata (attraverso la modellizzazione della struttura variografica) della sola componente spaziale del fenomeno in esame; nel secondo caso, invece, un modello in grado di sfruttare anche informazioni aggiuntive contenute nelle variabili secondarie che caratterizzano il contesto abitativo nel quale la misura è stata condotta.

Dall'analisi delle localizzazioni per le quali si sono ottenuti gli errori maggiori, emerge come l'approccio ML esaminato fornisca i risultati migliori, avendo accesso all'informazione contenuta in variabili come "contatto con il terreno" e "materiale da costruzione" che rendono peculiari i singoli casi esaminati.

INTRODUZIONE

Come spesso accade per i problemi di tipo ambientale, l'importanza del fenomeno che si intende analizzare va di pari passo con la sua complessità, rendendo quindi difficile tanto una sua comprensione esaustiva e dettagliata quanto una sua efficace modellizzazione. Peculiarità del fenomeno radon indoor, rispetto alla situazione in cui questo gas possa disperdersi liberamente in atmosfera, è la stretta e complessa interazione con l'edificio nel quale il gas si può accumulare. Benché i principali meccanismi di diffusione siano oggi noti e correttamente modellizzati (AA.VV., 1998), questo purtroppo non si traduce in previsioni del valore di concentrazione misurato in ambienti chiusi altrettanto affidabili.

Con la convinzione che per una descrizione esaustiva ed efficace del fenomeno la sola analisi spaziale (di tipo geostatistico) non risulti sufficiente, soprattutto in relazione all'influenza non trascurabile delle caratteristiche dell'edificio sede della misura, è sembrato in questo contesto interessante verificare tale assunzione nella pratica, applicando due approcci differenti al medesimo dataset operativo costituito da misurazioni reali.

L'APPROCCIO GEOSTATISTICO

Sulla base delle misurazioni georeferenziate che costituiscono il dataset operativo, sono inizialmente state condotte analisi geostatistiche convenzionali, che prevedono essenzialmente: i) uno studio esplorativo variografico volto alla ricerca della presenza di una struttura spaziale nei dati; ii) una successiva modellizzazione del variogramma sperimentale e infine iii) il suo utilizzo per una stima del valore di concentrazione in localizzazioni non campionate mediante un algoritmo di kriging — nel caso specifico, Kriging Ordinario (OK) (N. A. C. Cressie, 1993, E. H. Isaaks *et. al*, 1989).

L'APPROCCIO MACHINE LEARNING: w*k*-NN

Nell'ambito della discriminazione statistica, il metodo del "vicino più prossimo" (Nearest Neighbor, NN) rappresenta una delle tecniche più intuitive e semplici. È un metodo non-parametrico mediante il quale una nuova osservazione viene classificata ricorrendo alle osservazioni che costituiscono il dataset di riferimento, e in particolare a quella che le è "più vicina": in base alle covariate disponibili, alla nuova osservazione verrà assegnata la classe che compete all'osservazione del dataset di riferimento che più le "assomiglia" — la determinazione del grado di somiglianza si basa sulla misura di un qualche tipo di distanza (che andrà opportunamente definita) in uno spazio multidimensionale.

Una prima estensione dell'idea appena descritta, ormai ampiamente usata nella pratica, è quella che prende il nome di *k*-nearest neighbor (*k*-NN); in questo caso, non viene considerato solo il primo vicino, ma i

primi *k* vicini (osservazioni simili a quella nuova). Di conseguenza, con un sistema 'a votazione' verrà scelta la classe di appartenenza della nuova osservazione. Il parametro *k* deve essere definito dall'utente.

Il package "kknn" per l'ambiente statistico R (R, 2008) che è stato utilizzato per questo lavoro risulta particolarmente interessante in questo contesto applicativo poiché consente di implementare questa tecnica aggiungendo delle potenzialità che sono sembrate utili, e in particolare i) la possibilità di utilizzare l'approccio anche per la regressione (e non solo per la classificazione) e ii) l'introduzione di uno schema di pesatura basato su funzioni a kernel.

Questa ulteriore estensione della tecnica si fonda sull'idea che le osservazioni del dataset di riferimento che sono particolarmente vicine alla nuova osservazione dovrebbero ricevere un peso maggiore nella fase di decisione rispetto a quelle che risultano invece più distanti. Per raggiungere l'obiettivo proposto, la distanza sulla quale la ricerca dei vicini si basa deve essere successivamente trasformata in una misura di somiglianza che possa essere utilizzata come peso in fase di assegnazione della classe (o stima, nel caso della regressione) per la nuova osservazione.

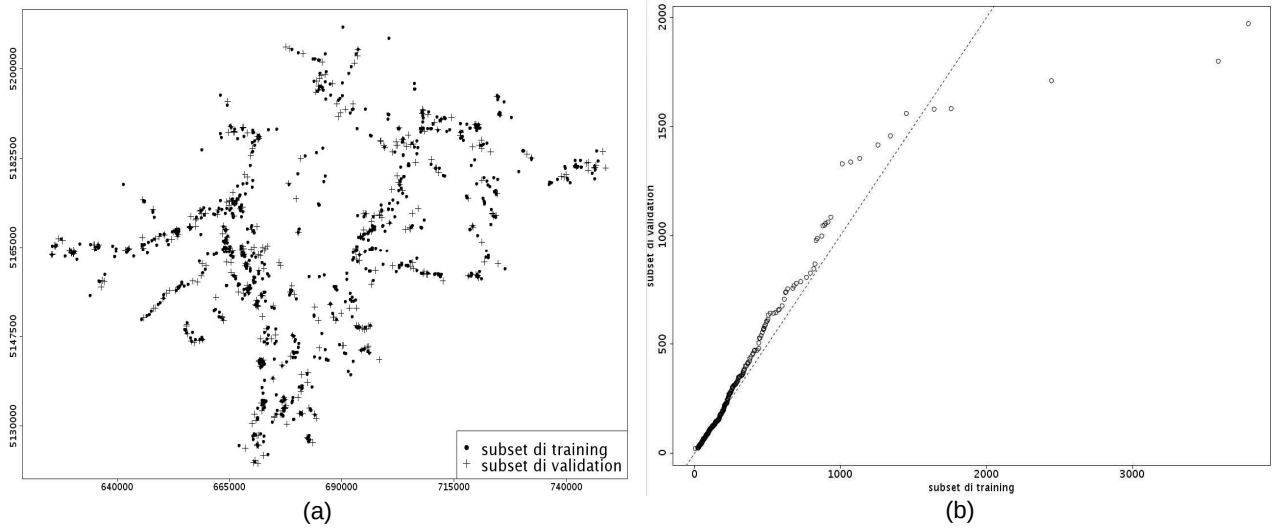
Uno degli obiettivi di questo nuovo approccio è anche quello di poter disporre di un metodo che risulta, in un certo senso, indipendente da una scelta infelice del parametro *k*, che può portare a grandi errori di misclassificazione; infatti, il numero di primi vicini viene implicitamente mascherato nei pesi: se *k* è troppo grande, viene corretto a un valore più basso in maniera automatica — se questo fosse il caso, un piccolo numero di vicini cui viene associato un peso elevato dominerà i rimanenti, che non avranno di conseguenza una forte influenza sulla stima in quanto verrà loro assegnato un peso inferiore.

DESCRIZIONE DELLA BANCA DATI

I dati cui il presente lavoro fa riferimento sono stati raccolti in successive campagne di misura condotte sull'intero territorio altoatesino, in modo da ottenere una copertura spaziale piuttosto omogenea e rappresentativa di tutta l'area geografica in esame. I dati fanno riferimento al valore di concentrazione di attività di radon misurata all'interno degli edifici (radon indoor) mediante dosimetri passivi a stato solido, esposti tipicamente per un semestre (estivo o invernale). Al fine di ottenere un valore più affidabile, i dati di concentrazione hanno subito una procedura di correzione per un fenomeno noto come "saturazione da sovrapposizione" (Verdi *et al.*, 2005), la cui influenza si manifesta in maniera sensibile sui valori più elevati, che in ambito preventivo vanno pertanto trattati con maggior attenzione.

Ogni singola misura che compone il dataset di riferimento è stata accuratamente georeferenziata ed è accompagnata da una ricca serie di informazioni secondarie codificate in variabili sia di tipo qualitativo che quantitativo, volte alla descrizione del contesto geografico ed edilizio che caratterizza ogni singola misura. Risultano così disponibili informazioni relative alle proprietà del terreno e dell'edificio sede della misura, come il piano di esposizione, la qualità degli infissi, il tipo di materiale da costruzione, la tipologia dell'edificio, ecc.

Figura 1 - Confronto tra le distribuzioni spaziali e di probabilità per i due subset impiegati.



(a) (b)

In (a), un post-plot relativo alla distribuzione spaziale per il subset di training e quello di validation (le coordinate sono espresse in metri); in (b), un qq-plot per un confronto visivo tra le rispettive distribuzioni di probabilità empiriche.

Nello specifico, il dataset operativo è stato costruito pre-selezionando le sole misure condotte al piano zero e nel semestre invernale, al fine di garantire una maggior uniformità dei dati. Si è quindi operata una

successiva suddivisione del dataset in due subset, uno cui ricorrere per la determinazione dei parametri che andranno a caratterizzare i differenti modelli implementati (subset di "training") e uno cui far riferimento per la fase di confronto (subset di "validation"). Si sottolinea come si abbia prestato attenzione a ottenere due subset entrambi rappresentativi della distribuzione spaziale dei campionamenti che caratterizza quello di partenza. Alcuni parametri statistici relativi ai citati subset sono riportati in tab. 1, mentre la fig. 1 mostra un confronto visivo tra le rispettive distribuzioni spaziali (post-plot) e distribuzioni di probabilità (qq-plot).

Tabella 1 - Parametri statistici relativi ai subset impiegati

subset	*min*	*1 quartile*	*mediana*	*media*	*3 quartile*	*max*	*s.q.m.*	*N*
training	3	69	120	229	240	3794	352	820
validation	22	71	124	240	276	1937	305	409

I DIFFERENTI APPROCCI A CONFRONTO

Le stime ottenute sulla base dei due approcci presi in considerazione, che riassumendo sono:

a) Kriging Ordinario — approccio geostatistico, che considera solo l'informazione contenuta nella distribuzione spaziale dei campionamenti sfruttando il modello di variogramma;
b) Weighted *k*-Nearest Neighbor (w*k* -NN) — approccio Machine Learning, che oltre alla componente spaziale considera anche le informazioni contenute nelle covariate che accompagnano ogni singola misura;

sono state messe a confronto sia da un punto di vista globale, ovvero analizzando i residui delle stime stesse, sia da un punto di vista più locale, ovvero analizzando alcune localizzazioni di particolare interesse.

Si sottolinea come sia stato possibile condurre questi tipi di analisi sfruttando il dataset 'validation', che ha le utili caratteristiche di:

- avere una serie di campionamenti con una distribuzione spaziale sul territorio dell'Alto Adige che riproduce quella del dataset operativo — in questo modo, è possibile mettere a confronto i vari approcci in tutti i diversi contesti geografici che caratterizzano il territorio in esame;
- avere, per ogni localizzazione di stima, anche il valore reale di concentrazione misurato — così da poter confrontare gli approcci non solo tra loro, ma anche con la situazione reale.

ANALISI GLOBALE

Le analisi e i confronti sono stati eseguiti ricorrendo a grafici che riportano i valori fittati e dei residui in funzione del reale valore di concentrazione, a grafici proporzionali relativi alla distribuzione spaziale dei residui stessi e ai parametri statistici riportati in tab. 2 relativi al fit lineare basato su minimi quadrati per scatter plot del tipo 'valori fittati' vs. 'valori reali'.

Tabella 2 - Parametri statistici relativi alle analisi globali sui residui

	intercetta	*pendenza*	R^2	*RMSE*
OK	172±9	0.22±0.02	0.192	276
wk-NN	187±7	0.16±0.02	0.154	280

Si può concludere che da un punto di vista globale, ricorrendo a un dataset indipendente di validazione, l'approccio di ML e quello di tipo geostatistico danno risultati paragonabili tra loro, benché il modello w*k*-NN abbia accesso a informazioni aggiuntive (relative alla parte antropogenica) rispetto al modello di OK, che prende in considerazione solo la componente spaziale del fenomeno radon indoor. Inoltre, i valori dei residui, sia in sovra- che in sotto-stima, si distribuiscono uniformemente su tutto il territorio esaminato.

ANALISI LOCALE

Accanto alle analisi di tipo globale appena descritte, è sembrato opportuno e interessante indagare più nello specifico il comportamento dei differenti approcci controllando puntualmente alcune particolari localizzazioni; l'idea è stata quella di poter notare alcune differenze in situazioni che si potrebbero definire "difficili" per i modelli in esame.

La scelta dei punti da destinare a questo tipo di analisi è stata fatta sulla base dei punti più distanti dalla bisettrice nei grafici 'valore fittato' vs. 'valore reale', ovvero, localizzazioni per le quali gli errori, sia in sovrastima che in sottostima, sono risultati maggiori. Interessante infine notare come tali punti siano comuni a tutti gli approcci sotto esame, indice del fatto che le situazioni per così dire anomale o difficili (piuttosto

comuni per la banca dati da cui le misure sono state estratte) sono tali indipendentemente dall'approccio scelto, configurandosi quindi come una caratteristica intrinseca del fenomeno radon indoor (almeno in relazione alla situazione che caratterizza l'Alto Adige).

Mediante strumenti esplorativi di tipo GIS, si sono analizzate caso per caso le situazioni appena descritte, valutando anche le caratteristiche specifiche dei campionamenti che costituiscono il vicinaggio di stima. In tutti i casi presi in considerazione, sia per quanto riguarda le sovra-stime che le sotto-stime:

- in generale, l'approccio di ML porta a stime che si avvicinano di più al valore realmente misurato rispetto all'approccio dell'OK;
- i casi anomali indagati si caratterizzano per intorni con edifici tipicamente in contatto con il terreno e costruiti con sassi: queste sono peculiarità che hanno manifestato, anche in relazione ad altre analisi condotte, una influenza non trascurabile sul valore di concentrazione misurato;
- "istruendo" il modello con informazioni aggiuntive relative alla parte antropogenica del fenomeno, si ottengono delle stime che più si avvicinano alla situazione reale, a sostegno dell'ipotesi che la sola componente spaziale non sia sufficiente per una modellizzazione esaustiva del fenomeno radon indoor — questi risultati possono essere mascherati o quantomeno poco evidenti in analisi di tipo globale, ma risultano invece significativi se si focalizza l'attenzione sulle situazioni più complesse e anomale.

CONCLUSIONI

Ricorrendo a un dataset operativo comune ricavato da quello generale costituito da misure di concentrazione di attività di radon indoor, abbiamo confrontato i risultati ottenuti ricorrendo a due differenti approcci per la stima in localizzazioni non campionate: da un lato un modello di tipo geostatistico basato su Kriging Ordinario (OK), dall'altro un modello basato su tecniche di Machine Learning (ML). Nel primo caso, quindi, un modello volto alla trattazione specifica e raffinata della componente spaziale del fenomeno in esame; nel secondo, un modello in grado di aggiungere, accanto alla componente spaziale (cui non viene però dedicata una trattazione geostatistica), anche informazioni aggiuntive contenute in variabili secondarie che caratterizzano il contesto nel quale la misura è stata effettuata.

Dalle analisi statistiche e spaziali condotte mettendo a confronto le stime ottenute su localizzazioni per le quali sono disponibili anche i reali valori di concentrazione, si può concludere che:

- limitando il confronto a rappresentazioni visuali e spaziali dei residui e a semplici analisi statistiche sugli stessi, globalmente l'approccio w*k*-NN e quello di tipo geostatistico danno risultati paragonabili tra loro;
- focalizzando invece l'attenzione su specifiche localizzazioni che hanno mostrato elevati valori del modulo del residuo, in tutti i casi l'approccio ML porta a stime che rispecchiano meglio il valore realmente misurato: il fornire al modello informazioni aggiuntive sembra quindi manifestare in maniera più evidente la sua efficacia e potenzialità in situazioni "anomale" o quantomeno di più complessa modellizzazione.

Risulta infine interessante notare come i casi che sono stati oggetto di analisi specifiche siano tutti caratterizzati da intorni costituiti principalmente da edifici in contatto con il terreno e costruiti con sassi: queste sono infatti caratteristiche legate a variabili che, in base ad analisi condotte in ambiti differenti, sono risultate come le più efficaci in relazione alla loro capacità predittiva sul valore di concentrazione misurato.

Bibliografia

AA.VV. Radon and Its Decay Products in Indoor Air. Environmental Science and Technology. W. Nazaro and A. V. Nero, Jr (1988). ISBN 0-471-62810-7

E. H. Isaaks e R. M. Srivastava. An Introduction to Applied Geostatistics. Applied Geostatistics Series. Oxford University Press (1989)

N. A. C. Cressie. Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, inc. (1993). ISBN 0-471-00255-0

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). ISBN 3-900051-07-0. URL http://www.R-project.org

Verdi L., Pegoretti S. e Ferrarini M., *Saturation on LR-115 Radon Detectors: Potential and Real Effects on Radon Mapping*, in International Conference, 4th Dresden Symposium: Survey of Geo-Hazards, 26–30 settembre 2005, Dresda, Germania.