AESTIMUM 48, luglio 2006: 19-38

Claudio Acciani*
Gaetano Gramazio**

**Ricercatore presso il Dipartimento di Economia e Politica agraria, Estimo e Pianificazione rurale (DEPAR) Università degli Studi di Bari*
e-mail:claudio.acciani@agr.uniba.it

***Informatico, libero professionista*
e-mail:gramazio.gaetanol@tiscali.it

L'Albero di Decisione quale nuovo possibile percorso valutativo[1]

In the real estate evaluations, both in urban and rural one, the several parameters methodology is assuming a more and more thickness role, considering that the determination of the value is reached analysing the real contribution of the single variables they define. This procedure supposes that information (data) is detailed as much as possible, assuring therefore, the objectivity of the entire operation. This work examines the possibility of using this method which exploit the "knowledge" contained in these data, through classification processes, typical of the Data Mining.
Decisions Tree (D.T.), in fact, concurs to preview, through iterative processes, the most probable the real estate values, chancing variables intensity. Comparing to the regressive techniques, by now consolidated, the D.T. allows us to verify the different sequences by which variables participate to the price making.

Parole chiave: Data Mining, Albero di decisione, valutazioni immobiliari

1. Introduzione

Nel corso del tempo, l'Estimo si è arricchito di nuove e più attendibili procedure di valutazione; una svolta è dovuta, senza alcun dubbio, all'introduzione delle stime pluriparametriche e, in particolare, alla regressione lineare multipla (Milano 1968; Simonotti 1988, 1989, 1991; Grillenzoni & Grittani 1994; Acciani 1996, 2004).

Il grande salto di qualità delle stime, grazie a quest'ultima, è da ricondurre a due motivi principali:

- la possibilità di verificare l'attendibilità del modello di stima attraverso opportuni test statistici;
- la maggiore aderenza ai fatti del mercato, da cui non si dovrebbe mai prescindere.

A tale proposito, dobbiamo ricordare che la stima pluriparametrica si pone in alternativa alla capitalizzazione dei redditi; infatti, in presenza di beni di confronto piuttosto dissimili dal bene da valutare, la procedura da utilizzare dovrebbe consi-

[*] Il presente lavoro fa parte della ricerca svolta con i Fondi di Ateneo 2005 "*Data Mining: gli Alberi Decisionali come supporto per la valutazione dei beni immobiliari*" nell'ambito delle attività del Centro Universitario di Studi e Ricerche del Mercato Fondiario. È frutto del lavoro comune dei due autori. C. Acciani ha curato la stesura dei paragrafi 1, 2, 3, 5 e 6; G. Gramazio il paragrafo 4. Gli autori ringraziano il Prof. G. Fratepietro per gli utili suggerimenti. Ovviamente la responsabilità dello scritto resta a carico degli autori.

derare un parametro economico in grado di tenere conto delle numerose caratteristiche differenziali oppure, ancora meglio, la procedura deve consentire di apprezzare il peso esercitato da ogni singola caratteristica differenziale sul prezzo.

È questa, per l'appunto, la logica alla base delle stime pluriparametriche, che poi è la logica che sul mercato porta alla formazione dei prezzi.

Non è pensabile, infatti, che i diversi livelli di prezzo siano influenzati da diversi livelli di beneficio fondiario (nozione del tutto sconosciuta ai comuni operatori di mercato); è logicamente pensabile, invece, che i prezzi siano il portato dell'apprezzamento di una serie di caratteristiche oggettive dei beni (per i terreni agricoli, l'ampiezza, la distanza dal centro abitato, la fertilità del terreno, la forma, il numero dei corpi, il sesto e l'età delle piante, ecc.) che sono quelle, poi, che vengono scelte come variabili esplicative nei modelli di stima a più parametri.

L'unico limite che osta all'applicazione di questi modelli è rappresentato dalla numerosità del campione di riferimento: pari a non meno di 4-5 volte il numero di variabili indipendenti.

Ma, un altro limite potrebbe essere rappresentato dalla non fedele aderenza del modello pluriparametrico a quei mercati estremamente complessi, come quello degli appartamenti per civile abitazione e dei terreni agricoli.

Ci chiediamo, cioè, se non sia possibile un ulteriore passo in avanti che ci consenta di definire non un unico modello, ma modelli diversi, ognuno dei quali più rappresentativo di un segmento di mercato.

La possibilità appena ventilata è offerta, oggi, da una nuova scienza, l'intelligenza artificiale.

2. Le nuove possibilità di indagine offerte dalla Intelligenza Artificiale

Sebbene si tratti di una disciplina relativamente giovane (i primi esperimenti hanno avuto inizio a partire dagli anni quaranta[1]), l'Intelligenza Artificiale trova oggi diversi campi di applicazione sia nella fase della ricerca sia nella vita di tutti i giorni.

È difficile definire correttamente questa disciplina; universalmente è riconosciuta come quella branca della scienza informatica e ingegneristica che si occupa di riprodurre mediante computer le capacità intellettive della mente. È altrettanto vero che, poiché si tratta di una scienza moderna, e quindi in forte e rapida evoluzione, una precisa definizione rischierebbe di relegarla entro limiti forse troppo stretti; si tratta, infatti, di una scienza di tipo "trasversale" che attraversa e interloquisce con altre discipline, ai più, apparentemente slegate: logica, informatica, matematica, psicologia, neurologia, psicoscienze (Schinardi 2001).

---

[1] C. Shannon e A. Turing realizzarono, a partire dal 1943, i primi programmi per il gioco degli scacchi; A. Minski e D. Edmons, nel 1951, il primo computer neurale. Il primo programma, Logic Theorist, capace in qualche modo di ragionare, è messo a punto da A. Newell e H. Simon. Ma la vera e propria svolta si ha nel 1956 quando J. McCartney organizza un convegno sull'argomento Intelligenza Artificiale (A.I. acronimo inglese).

È possibile ricondursi a due scuole di pensiero che distinguono una I.A. Forte da una Debole; la prima vede le macchine simili all'uomo, in grado, quindi, di sostituirsi ad esso ("macchine che pensano"); la seconda vede il computer che agisce, si comporta "come se" fosse un uomo, in definitiva una macchina che possa operare al suo posto, non in grado, comunque, di sostituirsi completamente all'essere umano (Mello 2002).

Tra le diverse branche che costituiscono l'I.A., troviamo la Machine Learning (apprendimento automatico) e nell'ambito di questa è collocato il Data Mining, inteso come processo che utilizza una o più tecniche per estrarre, dai dati, la conoscenza in termini di associazione, classificazione, regole, sequenze ripetute.

La diversità delle tecniche adoperate dipende essenzialmente dagli algoritmi utilizzati per realizzarle e la scelta di quale adoperare è in funzione degli obiettivi da raggiungere oltre che dai dati di cui si dispone. Le tecniche più adoperate sono: Clustering, Reti Neurali (R.N.), Alberi di Decisione, Individuazione e Associazione.

Il Clustering (unsupervised classification – classificazioni non supervisionate) o analisi dei cluster o, ancora, analisi di raggruppamento è una tecnica di analisi multivariata dei dati volta alla selezione e raggruppamento di elementi omogenei in un insieme di dati, che consiste nella segmentazione di un database in sottoinsiemi (i cluster).

Tutte le tecniche di clustering si basano sul concetto di distanza tra due elementi. Infatti, la bontà delle analisi ottenute dagli algoritmi di clustering dipende essenzialmente da come è stata definita la distanza, concetto fondamentale dato che gli algoritmi di clustering raggruppano gli elementi a seconda della stessa e quindi l'appartenenza o meno ad un insieme dipende da quanto l'elemento preso in esame è lontano dall'insieme. Le tecniche di clustering si possono basare principalmente su due filosofie: dal basso verso l'alto (*bottom-up*) e dall'alto verso il basso (*top-down*).

La prima prevede che, inizialmente, tutti gli elementi siano considerati cluster a sé e poi l'algoritmo provvede ad unire i cluster più vicini. L'algoritmo continua ad unire elementi al cluster fino ad ottenere un numero prefissato di cluster oppure fino a che la distanza minima tra i cluster non supera un certo valore.

Dall'alto verso il basso: all'inizio tutti gli elementi sono un unico cluster e poi l'algoritmo inizia a dividere il cluster in tanti cluster di dimensioni inferiori. Il criterio che guida la divisione è sempre quello di cercare di ottenere elementi omogenei. L'algoritmo procede fino a che non ha raggiunto un numero prefissato di cluster. Questo approccio è anche detto gerarchico.

La tecniche di clustering vengono utlizzate generalmente quando si hanno tanti dati eterogenei e si è alla ricerca di elementi anomali. Tali tecniche rientrano nella categoria dei meccanismi di classificazioni non supervisionate che individuano e metteno in evidenza le somiglianze tra oggetti; questi ultimi vengono rappresentati come punti in uno spazio multidimensionale, in cui ogni dimensione corrisponde a una caratteristica di interesse. Fare clustering significa raggruppare gli oggetti in un ridotto numero di insiemi che caratterizzino al meglio la popolazione degli stessi (Gori 2003).

In conclusione la tecnica di clustering si può definire come una classificazione non supervisionata che consente, quindi, di effettuare operazioni di segmentazio-

ne sui dati; non c'è una ricerca di previsione, ma si cerca una ripartizione ottimale del campione. Proprio perché non viene individuata alcuna variabile "obiettivo", non è previsto alcun approfondimento supervisionato.

Le tecniche di analisi delle Associazioni consentono di individuare delle regole nelle occorrenze concomitanti di due o più eventi, permettendo, quindi, lo studio di fenomeni simultanei o in sequenza degli eventi. Non c'è un criterio standard per definire interessante una regola, in quanto si possono fare valutazioni sia sul supporto che sulla confidenza e comunque occorre una buona conoscenza del fenomeno studiato per dividere regole imprevedibili e interessanti da regole note che non portano alcuna informazione aggiuntiva.

Le Reti Neurali supervisionate e gli Alberi di Decisione (A.D.) consentono, al contrario, di effettuare delle classificazioni, conseguenza di una conoscenza acquisita (tramite addestramento – training set), per classificare nuovi oggetti o per prevedere nuovi eventi. In realtà, le reti neurali presentano degli aspetti, almeno ai fini della previsione, alquanto lacunosi nel senso che non permettono di individuare tutti quei passaggi che portano poi al risultato finale né, tanto meno, di facilitare l'interpretazione degli elementi che compongono la rete stessa. Le R.N., infatti, non esplicitano alla fine alcun risultato circa l'effettivo uso delle variabili esplicative in quanto i risultati restano intrappolati nei pesi sinaptici e quindi di difficile interpretazione (Gastaldi 2002).

E, sempre a proposito della differenza tra questi due metodi, un altro aspetto favorevole all'uso dell'A.D. è che mentre le R.N. devono essere ogni volta "tarate", le regole induttive estrapolate dagli A.D. restano valide nel tempo e possono essere utilizzate anche per altri prodotti affini (Gastaldi 2002).

3. Gli Alberi di Decisione

Per prevedere nuovi eventi attraverso regole ottenute tramite processi induttivi, adopereremo un modello di Albero Decisionale; è una tecnica, come detto, molto usata per la segmentazione di oggetti in classi. Un Albero si compone di: a) NODI che contengono i nomi delle variabili indipendenti; b) ARCHI che sono etichettati con i possibili valori delle variabili indipendenti; c) FOGLIE che rappresentano le classi, cioè una collezione di osservazioni raggruppate in base ai valori di una variabile indipendente, e sono unite ai nodi tramite gli archi. Un Albero di Decisione prende, in entrata, un elemento x descritto da un insieme di caratteristiche (coppie valore-attributo) ed emette, in uscita, una *decisione* che può essere sotto forma binaria, nominale, quantitativa. Per essere poi interpretato, l'albero può essere trasformato da una sequela di regole (if > then) che descrivono, in pratica, il percorso. Tutto ciò è regolato da un algoritmo che, in definitiva, effettua un'analisi discriminante in quanto analizza le variabili/attributi predittive cercando di individuare i legami tra esse esistenti, spiegando, così, il comportamento della variabile da osservare, la variabile dipendente.

L'A.D. permette quindi di correlare, a determinate caratteristiche di un oggetto, il suo più probabile comportamento, la sua più probabile risposta.

In questo studio, pertanto, si è fatto un tentativo circa l'utilizzo del metodo per la determinazione del più probabile valore di mercato di beni immobiliari, cercando di individuare quelle regole induttive in grado di enfatizzare, o meno, la partecipazione di alcuni caratteri propri del bene da stimare.

La filosofia di applicazione consiste nel determinare quelle regole che associano le variabili indipendenti (nel nostro caso le caratteristiche degli appartamenti oggetto di compravendita), ai corrispondenti valori unitari; è per questo che la scelta di una analisi discriminante come quella degli alberi decisionali appare tra le più interessanti.

La base, lo ripetiamo, è quella di associare le caratteristiche del bene al loro "comportamento" (prezzo previsto); per intenderci: un appartamento di una certa superficie, dotato di accessori, facente parte di un complesso residenziale, dotato di posto-auto, con impianto di riscaldamento autonomo, "come si comporta?", cioè qual è il prezzo che riesce a spuntare sul mercato e quali sono le variabili che prioritariamente determinano tale prezzo? E in quale modo?

Evidenti appaiono le analogie con il settore commerciale o del marketing quando, osservando il profilo del cliente, quindi le sue caratteristiche, si vuole prevederne il comportamento (eventuale affiliazione o altro).

Per la costruzione dell'albero di decisione, il primo passo che compie l'algoritmo consiste nel selezionare la variabile che più di ogni altra assume la funzione discriminante, vale a dire quella che presenta l'errore di classificazione più basso che, in termini pragmatici, significa un più basso rischio di errore. Dove, per errore di classificazione più basso, si intende la capacità, da parte di una variabile, di suddividere (discriminare) l'insieme di partenza delle osservazioni in sottoinsiemi il più possibile omogenei dal punto di vista dei valori assunti. La tendenza ad avere delle classi più omogenee in fase di costruzione dell'albero garantisce un tasso di errore più basso in fase di utilizzo dello stesso (ad es. durante la classificazione di una nuova osservazione).

4. L'algoritmo M5'

L'algoritmo M5'[2] realizza modelli multivariati in quanto formati da varie formule di regressione. Lo scopo dell'analisi è quello di costruire un modello in grado di stabilire una relazione tra la variabile dipendente del training set e i rispettivi valori delle variabili indipendenti. Il software in dotazione[3] ci permette di costruire due modelli di albero: a) nelle foglie possiamo trovare dei valori numerici che rappresentano il valore calcolato sulla base delle osservazioni del campione, che raggiungono la foglia ed è questo il caso degli Alberi di Regressione; b) un modello di regressione lineare che produce il valore delle osserva-

[2] L'algoritmo utilizzato è l'M5', versione ottimizzata da Witten (1997) sulla base dell'algoritmo M5 di Quinlan (1992).

[3] Weka, scaricabile gratuitamente dal sito internet: http://www.cs.waikato.ac.nz/ml/weka/

zioni che raggiungono la foglia, caso dell'Albero Modello (modello adoperato in questa analisi).

Quando si utilizzano strutture ad albero in problemi di tipo decisionale, nei nodi sono contenuti i confronti tra variabili che permettono di decidere, in base al risultato ottenuto, il ramo da seguire e quindi il test da eseguire successivamente. Nelle foglie, invece, è contenuto il valore che risulta dai vari confronti effettuati (nel caso degli Alberi di Regressione) o direttamente la formula di regressione da applicare (nel caso degli Alberi Modello).

Un'ultima considerazione relativa all'algoritmo M5' riguarda il fatto che esso esegue una ricerca "greedy" per eliminare quelle variabili indipendenti che contribuiscono in maniera affatto non significativa alla realizzazione del modello; in alcuni casi, addirittura, M5' può eliminare tutte le variabili, lasciando solo la costante.

Gli alberi modello sono una tecnica di rappresentazione della conoscenza strutturata in maniera gerarchica, in cui le variabili del sistema sono processate all'interno dei nodi. Solitamente viene effettuato un confronto di attributi con valori costanti, mentre nelle foglie sono contenute formule di regressione. Gli alberi modello nascono dalla combinazione di due tecniche ben note dell'Intelligenza Artificiale: gli alberi di decisione e le formule di regressione lineare. L'unione di queste due tecniche favorisce la costruzione di alberi con un numero di nodi mediamente più basso, garantendo comunque tassi di errore più bassi se rapportati a quelli ottenuti con gli alberi di decisione.

Tabella 1
Dati meteorologici

Aspetto generale	Temperatura	Giocare
Soleggiato	Caldo	No
Soleggiato	Caldo	No
Nuvoloso	Caldo	Sì
Piovoso	Mite	Sì
Piovoso	Fresco	Sì
Piovoso	Fresco	No
Nuvoloso	Fresco	Sì
Soleggiato	Mite	No
Soleggiato	Fresco	Sì
Piovoso	Mite	Sì
Soleggiato	Mite	Sì
Nuvoloso	Mite	Sì
Nuvoloso	Caldo	Sì
Piovoso	Mite	No

La procedura adottata nella costruzione di un albero modello è la stessa degli alberi di decisione (differiscono unicamente per il fatto che negli alberi modello le variabili sono di tipo numerico): si seleziona una variabile indipendente da collocare nel nodo radice e si creano tanti rami quanti sono i suoi possibili valori.

In questo modo il set di osservazioni viene diviso in sottoinsiemi, uno per ogni valore dell'attributo. A questo punto il processo continua ripetendo la stessa procedura in ogni nodo dell'albero (che in questo contesto rappresentano i punti dell'algoritmo in cui viene effettuata la divisione in sottoinsiemi), usando solo le osservazioni raggiunte da quel ramo. In un qualunque punto dell'albero, se tutte le osservazioni contenute in un nodo hanno la stessa classificazione, termina lo sviluppo di quella parte dell'albero.

In questa fase della procedura, dato un insieme di dati, non resta che determinare l'attributo su cui operare.

Consideriamo, a titolo di esempio, un insieme di dati meteorologici (tabella 1):

I possibili nodi radice danno luogo ai seguenti sottoalberi:

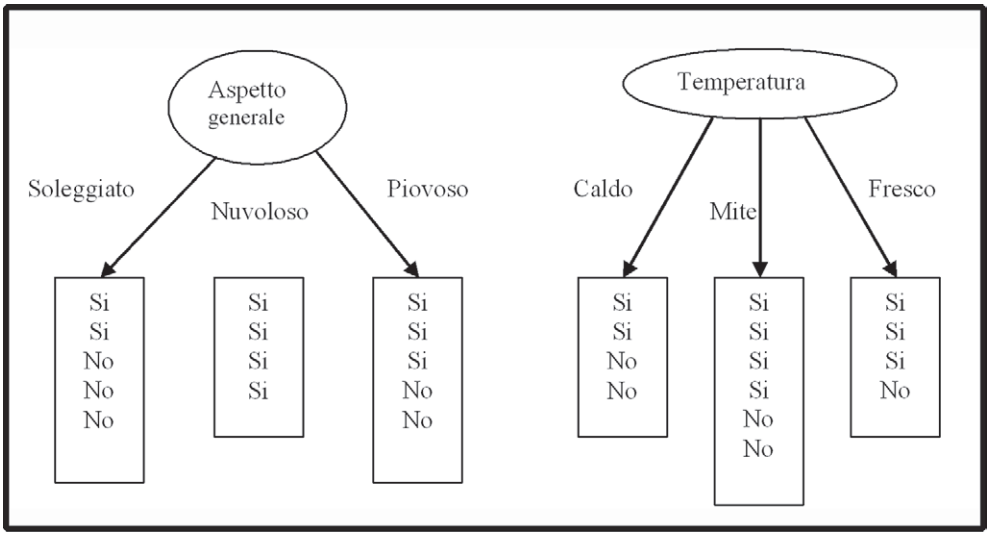

Fig. 1. Sottoalberi dati meteorologici

A questo punto bisogna decidere qual è l'attributo che meglio discrimina l'insieme. Le foglie contengono le varie risposte. Quando una foglia contiene una sola classe (cioè solo risposte "Sì" o solo risposte "No"), non possono essere effettuate altre separazioni ed il processo ricorsivo lungo quel ramo termina.

Dal momento che per ragioni di efficienza si preferisce costruire alberi della dimensione minore possibile, è fondamentale la ricerca di nodi foglia che contengono elementi della stessa classe in modo da terminare prima possibile la generazione di nodi. In questo caso l'attributo candidato è Aspetto Generale.

Pertanto l'albero diventa:

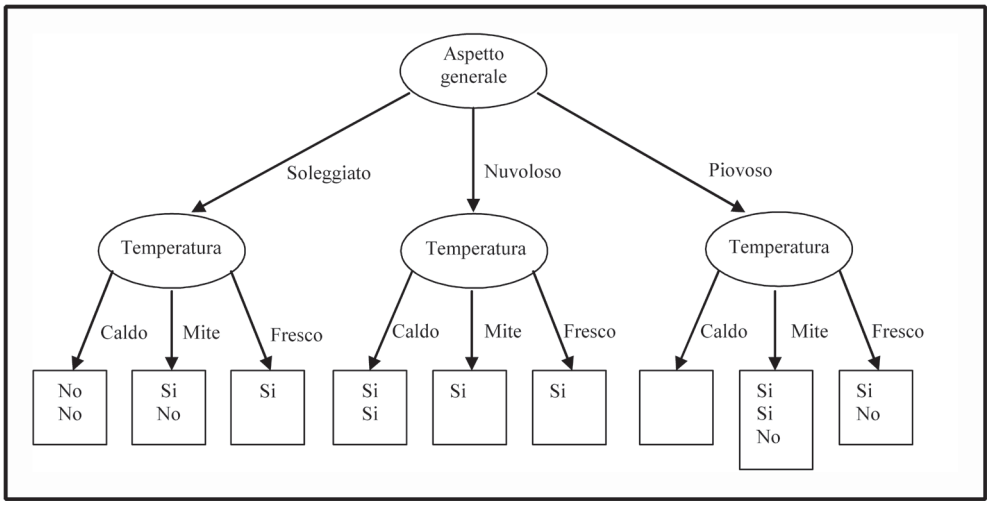

Fig. 2. Albero delle Decisioni

In questo modo abbiamo ottenuto una rappresentazione ad albero dei dati contenuti in tabella. La caratteristica di tale albero è quella di essere il più piccolo possibile tra quelli ottenibili con quel numero di parametri e di osservazioni. Solitamente i parametri sono più di due e lo studio su quello più discriminante viene effettuato su tutti i parametri ad ogni passo di generazione dell'albero.

Con variabili di tipo numerico, invece, l'algoritmo di costruzione dell'albero mira ad individuare un valore soglia in base al quale dividere le istanze che passano per quel nodo. Un caso particolare è costituito da variabili che possono assumere solo i valori 0 ed 1, in cui il valore discriminante viene fissato pari a 0,5.

Da un punto di vista più tecnico, questa operazione ha lo scopo di individuare le variabili a più alto "contenuto informazionale", cioè quelle che hanno una maggiore rilevanza nel modello; in altri termini, più una variabile è vicina al nodo radice, maggiore è il peso che tale attributo assume in fase decisionale.

Nelle foglie, inoltre, sono contenuti altri due valori numerici: il primo (un numero intero) indica il numero di osservazioni che passano in quel nodo, mentre il secondo numero rappresenta lo scarto quadratico medio relativo dato dalla seguente formula:

$$\frac{\sqrt{\sum_{i=1..m}\left(Y_i - y_i\right)^2 / m}}{\sqrt{\sum_{i=1..n}\left(Y_i - \bar{y}\right)^2 / n}} \qquad (1)$$

dove gli indici m ed n indicano rispettivamente il numero di osservazioni che passano per quel nodo, e quindi sono coinvolte nel calcolo dell'errore, ed il numero

totale delle osservazioni. Inoltre Y_i è il valore reale dell'osservazione i-esima, y_i è il valore predetto dell'osservazione i-esima dal modello lineare nella foglia e $\bar{Y}$ è la media aritmetica delle osservazioni. È un indice dell'errore, presente nella foglia, normalizzato rispetto alla dispersione dei dati dell'intero campione.

Il programma usato per eseguire gli esperimenti (WEKA) è una raccolta di strumenti software realizzati in Java e consente di agire su parametri che determinano la dimensione finale dell'albero e l'uso o meno di ulteriori procedure di raffinamento dell'albero ottenuto. In particolare il primo parametro permette di modificare il fattore di *pruning* dell'albero. Per *pruning* dell'albero si intende l'operazione di "potatura" per ridurre le dimensioni dello stesso, qualora si preferisca avere un modello di dimensioni più modeste e strutturalmente più semplice (conseguentemente diminuisce anche l'accuratezza del modello elaborato). Un fattore di *pruning* pari a 0 implica l'assenza della procedura di potatura, per cui in questo caso l'albero avrà le dimensioni massime. Un fattore maggiore di 0 (ad esempio 1, 2, 10, ecc.) indica la misura con cui viene applicato il *pruning* e può condurre in casi estremi ad un unico ramo (ossia, un unico modello di regressione).

L'altro parametro, invece, detto di *smoothing*, controlla una procedura messa a disposizione da WEKA anch'essa eseguita dopo la costruzione dell'albero, ma che agisce essenzialmente sulle formule di regressione contenute nelle foglie. Il processo di *smoothing* viene eseguito per compensare la netta discontinuità che ne deriverebbe inevitabilmente tra modelli lineari adiacenti presenti nelle foglie dell'albero; in particolar modo per alcuni modelli costruiti su un numero ristretto di osservazioni. Pertanto, durante tale processo, le equazioni lineari adiacenti vengono aggiornate in modo che tutti i valori in output ai modelli lineari abbiano valori più vicini.

Il modo più semplice per comprendere questo fenomeno della frammentarietà consiste nel vedere ogni formula di regressione lineare come una funzione che produce in uscita un sottoinsieme continuo dei valori reali. La continuità dei valori ottenibili implica la possibilità di avere una certa uniformità nei valori risultanti e quindi nelle stime. Se la formula utilizzata è unica, questa uniformità è garantita, mentre nel caso di più formule si possono presentare casi in cui i valori in uscita non garantiscono la continuità delle stime. Questa discontinuità nei valori significa che il modello non è in grado di rappresentare tutti i valori possibili, il che si traduce inevitabilmente in una ridotta attendibilità dello stesso.

Per questa ragione è possibile intervenire con la procedura di *smoothing* che esegue un percorso a ritroso dell'albero (cominciando la scansione dalle foglie e risalendo fino al nodo radice) e, misurando l'errore ottenuto, effettua una correzione sui valori soglia stabiliti nei nodi, in modo da ridurre l'eventuale discontinuità generata dalle formule contenute nelle foglie. Generalmente WEKA applica di default questa opzione e nel nostro caso si è preferito mantenerla attivata.

5. Un caso concreto

Il campione sottoposto ad analisi si compone di 137 osservazioni (tabella 2), riferite ad altrettanti casi di compravendita di appartamenti avvenuti nella città di

Tabella 2
Campione di 137 osservazioni

N.	Prezzo	Zona	Sup.	Piano	Età	Stato fabbr.	Stato appart.	Locat.	Risc.	Int.	Bagni	Port.	Posti auto	Riv. est.	Data
1	905	0,77	137	0,94	2	2	1	0	2	0	1	0	0	0	1
2	1027	0,77	67,9	0,97	1	2	1	0	2	0	1	0	0	0	1
3	1094	0,69	132,5	0,98	1	1	0	1	2	0	2	1	0	0	1
4	1632	0,83	95	1,05	2	2	1	1	1	0	1	0	0	0	1
5	1495	1,00	133,75	0,94	1	1	0	1	2	0	2	1	0	1	1
6	1483	0,73	144,55	1	1	2	1	1	2	0	2	0	0	0	1
7	797	0,67	68	1	0	1	1	0	0	0	1	0	0	0	1
8	1244	0,67	116,25	1	1	1	1	1	0	0	1	0	0	0	1
9	1328	0,72	70	0,80	1	1	1	0	2	0	1	0	0	0	1
10	2085	1,00	54,5	0,96	0	2	2	0	0	0	1	0	0	0	1
11	1006	1,00	132,5	0,90	0	1	1	1	2	0	2	0	0	0	1
12	1467	0,83	88	1,05	1	1	1	1	2	0	1	0	0	0	1
13	1876	0,80	129,4	0,96	2	2	1	1	2	0	2	0	1	0	1
14	1156	0,77	134	0,94	1	1	2	0	1	0	2	0	1	0	1
15	963	0,73	122,25	0,96	0	1	1	0	2	0	1	0	0	0	1
16	1530	0,89	192,75	0,98	1	1	0	1	2	0	2	0	0	0	1
17	1519	0,80	123,75	1	1	2	1	0	2	0	2	0	0	1	1
18	1044	0,73	96,5	0,90	0	1	1	0	0	0	1	0	0	0	1
19	1219	0,80	89	0,90	1	1	1	0	2	0	1	0	0	0	1
20	1623	1,00	70	0,98	0	2	2	1	0	0	1	0	0	0	1
21	1295	0,69	167,5	1,05	2	2	2	0	2	0	2	0	1	0	1
22	1158	0,69	136	0,94	2	2	1	0	1	0	1	0	2	0	1
23	1007	0,69	128,2	0,90	2	1	1	1	2	0	2	0	0	1	1
24	1228	0,77	92,5	0,70	0	1	1	0	2	0	1	0	0	0	1
25	1154	0,54	156	0,90	1	1	2	0	2	0	2	0	0	1	1
26	2167	0,82	150	1	1	2	2	1	2	1	2	0	0	0	1
27	743	0,77	157,5	0,97	0	0	0	1	0	0	1	0	0	0	1
28	1410	0,80	186,85	1	2	2	1	0	1	0	2	1	1	0	1
29	1558	0,80	126	1,05	2	1	1	0	1	0	1	0	1	0	1
30	1932	1,00	106,9	1	1	1	1	0	1	0	1	1	0	0	1
31	1668	1,00	96	1	0	2	2	1	2	0	1	0	0	0	1
32	1206	0,69	137	0,96	2	2	1	0	2	0	1	0	1	0	1
33	1064	0,77	67,5	0,70	1	1	1	1	0	0	1	0	0	0	1

N.	Prezzo	Zona	Sup.	Piano	Età	Stato fabbr.	Stato appart.	Locat.	Risc.	Int.	Bagni	Port.	Posti auto	Riv. est.	Data
34	1242	0,69	116,75	0,98	2	2	1	1	2	0	2	0	1	0	1
35	2439	1,00	123	1	1	2	0	1	0	0	2	0	0	1	1
36	1807	1,00	332	0,94	0	2	0	1	1	0	2	0	0	1	1
37	945	0,50	100,5	1	3	2	2	0	2	0	2	0	0	0	1
38	1380	0,72	144,9	1	0	2	0	1	2	0	1	1	0	0	1
39	1616	0,82	61,25	0,94	2	1	0	0	1	0	1	0	1	0	1
40	1508	0,77	125	1	1	1	1	1	2	0	2	0	1	0	1
41	1192	0,54	67,105	1,05	3	2	2	1	2	0	1	0	1	0	1
42	1246	0,67	80,25	1,05	1	1	0	1	2	0	1	0	0	0	1
43	861	0,67	30	0,90	0	1	2	1	0	0	1	0	0	0	1
44	1789	0,67	85,75	1	3	1	1	0	2	0	1	0	0	0	1
45	1538	1,00	136	1	1	1	2	1	2	0	2	0	0	0	1
46	1476	0,83	75,25	0,94	1	2	1	1	2	0	1	0	0	0	1
47	1282	0,77	72,5	0,80	1	1	1	0	0	0	1	0	0	0	1
48	1865	0,77	90	0,94	3	2	2	1	2	0	1	0	0	0	1
49	1732	0,77	98,4	1	3	2	2	1	2	0	1	0	0	0	1
50	1184	0,69	121,25	1	2	2	1	0	2	0	1	0	0	0	1
51	2452	1,00	79	1,05	1	1	2	1	1	0	1	0	0	1	1
52	1108	0,73	51,25	0,90	0	1	2	1	0	0	1	0	0	0	1
53	1261	1,00	115	0,90	0	1	0	1	2	1	1	0	0	0	1
54	2636	1,00	129	1	1	2	1	1	2	0	1	0	0	1	2
55	1032	0,69	25	0,90	2	2	2	1	2	0	1	0	0	1	2
56	1755	0,82	124,5	0,96	1	1	1	0	2	0	2	0	0	0	2
57	1296	0,73	68,5	0,90	0	1	2	0	2	0	2	0	0	0	2
58	1492	1,00	136,75	0,90	0	1	2	0	2	0	2	0	0	0	2
59	1385	0,80	88,75	0,97	1	1	1	1	0	0	1	0	0	0	2
60	1492	0,82	218,25	1	1	1	2	0	2	0	2	0	0	0	2
61	1006	0,67	41,75	0,97	0	1	1	1	0	0	1	0	0	0	2
62	1516	0,69	122	1	2	1	2	1	2	0	2	0	2	0	2
63	1540	0,69	126	0,94	3	2	2	1	2	0	2	0	1	0	2
64	1398	0,69	114,5	0,98	1	1	1	0	2	0	1	0	0	0	2
65	1811	0,80	77	0,96	1	1	1	0	2	0	1	0	0	0	2
66	1483	0,77	97,5	0,94	1	1	1	0	2	0	1	0	0	0	2
67	1601	0,77	106,45	1	1	1	2	0	1	0	1	0	0	1	2
68	1335	0,77	97,85	0,90	1	1	2	0	1	0	1	0	0	0	2

N.	Prezzo	Zona	Sup.	Piano	Età	Stato fabbr.	Stato appart.	Locat.	Risc.	Int.	Bagni	Port.	Posti auto	Riv. est.	Data
69	1497	0,77	46,75	0,70	1	1	1	0	2	0	1	0	0	0	2
70	1138	0,77	49,2	0,80	0	1	1	0	0	0	1	0	0	0	2
71	1201	0,77	66,6	0,90	3	2	2	0	2	0	1	0	0	0	2
72	734	1,00	45	1	0	0	0	1	0	0	1	0	0	0	2
73	1316	0,67	120	1	1	2	1	1	2	0	1	0	0	0	2
74	1833	1,00	150	0,98	1	1	2	1	2	0	2	0	0	0	2
75	1344	1,00	98	1	0	1	2	1	2	0	1	0	0	0	2
76	2066	0,77	72,5	0,90	1	2	2	1	2	0	1	0	0	0	2
77	2121	0,77	91,3	0,90	2	1	2	1	2	0	1	0	0	0	2
78	1696	0,77	109,6	0,90	1	1	2	0	1	0	1	0	1	0	2
79	1472	0,89	30	0,90	0	2	2	1	0	0	1	0	0	0	2
80	1565	0,89	66	1	0	1	1	1	0	0	1	0	0	0	2
81	918	0,72	90	0,90	0	2	1	0	0	0	1	0	0	0	2
82	1253	0,72	127,75	0,94	1	1	1	1	1	0	1	0	0	0	2
83	1881	1,00	70	0,97	0	1	2	1	2	0	1	0	0	0	2
84	1720	1,00	130	0,98	1	1	0	1	0	0	1	0	0	1	2
85	1067	0,73	75	0,80	1	0	0	1	0	0	1	0	0	0	2
86	913	0,50	115	1,05	2	2	2	0	2	1	2	0	1	0	2
87	1663	0,67	105,25	0,98	1	2	0	1	0	1	1	0	0	0	2
88	1578	0,77	136,25	0,94	2	2	1	0	2	0	2	0	0	0	2
89	1177	0,69	131,65	0,98	2	2	1	1	1	0	2	0	1	0	2
90	1043	0,82	70	0,90	0	0	0	1	0	0	1	0	0	0	2
91	1010	0,50	102	1	2	2	2	0	2	0	2	0	1	0	2
92	1449	0,89	144,9	0,90	1	2	0	1	2	0	1	0	0	0	2
93	1682	0,82	61,25	0,96	2	2	0	0	1	0	1	0	1	0	2
94	3296	1,00	145	0,96	1	2	0	1	1	1	2	1	0	0	2
95	3303	1,00	145	0,96	1	2	0	1	1	1	1	1	0	0	2
96	1691	0,80	110	0,98	2	2	2	1	2	1	1	0	0	0	2
97	1711	0,83	90	1	1	2	0	0	0	1	1	0	0	0	2
98	2256	0,83	164	0,98	2	2	1	1	2	0	2	0	0	0	2
99	3340	1,00	199,25	0,98	2	2	2	1	1	0	3	1	0	0	2
100	1445	0,67		0,94	1	2	1	1	2	1	2	0	1	0	2
101	1662	0,83	96,25	1,05	2	2	1	1	1	0	1	0	0	0	2
102	1814	0,83	71,65	0,94	2	2	2	1	2	0	1	0	0	0	2
103	1617	1,00	136	1	0	2	0	1	2	0	2	0	0	0	2

N.	Prezzo	Zona	Sup.	Piano	Età	Stato fabbr.	Stato appart.	Locat.	Risc.	Int.	Bagni	Port.	Posti auto	Riv. est.	Data
104	1481	0,77	135	1	2	1	1	0	2	0	2	0	1	0	2
105	725	0,67	40	1	0	0	0	1	0	0	1	0	0	0	2
106	1295	1,00	115	1	0	1	0	1	2	0	1	0	0	0	2
107	1709	0,67	86	0,94	1	1	1	1	2	0	1	0	0	1	2
108	2250	0,77	88	0,94	1	1	0	1	0	0	1	0	0	0	2
109	1235	0,67	85	1	1	1	0	1	2	0	1	0	0	0	2
110	1834	1,00	332,5	0,94	0	2	0	1	2	0	2	0	0	1	2
111	2583	1,00	120	1	0	2	0	1	0	0	2	0	0	1	2
112	934	0,77	139,2	0,94	1	2	1	0	2	0	2	0	0	0	2,5
113	1156	0,77	140,5	0,94	1	2	1	0	2	0	1	0	0	0	2,5
114	861	0,67	65	1	0	1	1	0	0	0	1	0	0	0	2,5
115	1234	0,67	121,5	1	1	2	0	1	0	0	1	0	0	0	2,5
116	1296	0,67	115,75	1	2	2	0	1	0	0	1	0	0	0	2,5
117	2861	1,00	154,5	0,96	3	2	2	0	0	0	1	0	0	0	2,5
118	1063	0,77	67,9	0,97	0	2	1	0	2	0	1	0	0	0	2,5
119	1595	0,83	87,75	1,05	1	2	1	1	2	0	1	0	0	1	2,5
120	1911	0,80	133,3	0,96	2	2	2	1	2	0	1	0	1	1	2,5
121	1562	1,00	133,75	0,94	1	1	0	1	2	0	2	1	0	1	2,5
122	2133	1,00	150	0,98	1	1	1	1	1	1	2	0	0	1	2,5
123	1360	0,47	99,25	1	0	1	1	1	2	0	1	0	0	0	2,5
124	1650	0,77	136,25	0,94	1	2	1	0	2	0	3	0	0	0	2,5
125	1848	1,00	130	0,98	0	2	1	1	0	0	2	0	0	1	2,5
126	1966	1,00	70	0,97	0	2	2	1	2	0	1	0	0	0	2,5
127	1538	0,89	30	0,90	0	2	2	1	0	0	1	0	0	0	2,5
128	1843	0,77	109,6	0,96	1	2	2	0	1	0	2	1	0	0	2,5
129	2234	0,77	91,3	0,94	2	2	2	0	2	1	1	0	1	0	2,5
130	2071	0,77	74,5	0,94	1	2	1	1	2	0	1	0	0	0	2,5
131	1132	0,69	132,5	0,98	1	1	0	1	2	1	2	1	0	0	2,5
132	986	0,73	70	1	0	0	0	1	0	0	1	0	0	0	2,5
133	1286	0,77	117,4	0,90	2	2	2	1	2	0	2	0	0	1	2,5
134	1297	0,77	92,5	0,70	0	2	1	0	1	0	1	0	0	0	2,5
135	1186	0,54	156	0,90	2	1	2	0	2	0	2	0	0	1	2,5
136	1600	0,80	123,75	1	2	2	1	0	2	0	2	0	0	1	2,5
137	1572	0,72	192,75	0,98	1	1	0	1	2	0	2	0	0	0	2,5

Bari nel periodo compreso tra il gennaio 2002 e il giugno 2004, e 14 variabili indipendenti; la quantificazione delle variabili esplicative è riportata, in dettaglio, nell'Allegato 1.

Per la quantificazione dei diversi livelli delle variabili Età, Stato del Fabbricato, Stato dell'Appartamento, Data di compravendita, è stata utilizzata l'ipotesi di scarti costanti tra un livello e l'altro, in funzione dell'apprezzamento del mercato.

La tecnica di analisi utilizzata è quella dell'Albero Modello, algoritmo M5' con due soluzioni di *pruning*: uguale a 0, per ottenere l'albero nella sua massima estensione; uguale a 1,5, per contenere la ramificazione dell'albero in un limite praticamente accettabile.

Con quest'ultima soluzione, in altri termini, abbiamo tentato di operare una segmentazione più contenuta del campione iniziale, per ottenere un conseguente più ridotto numero di modelli di regressione.

Nella Fig. 3 e nella tabella 3 sono riportati i risultati relativi alla prima soluzione: l'albero risulta estremamente ramificato, tanto da individuare ben 21 diversi modelli di regressione, 13 dei quali, però, relativi a subcampioni con un numero di osservazioni inferiore o uguale a 5.

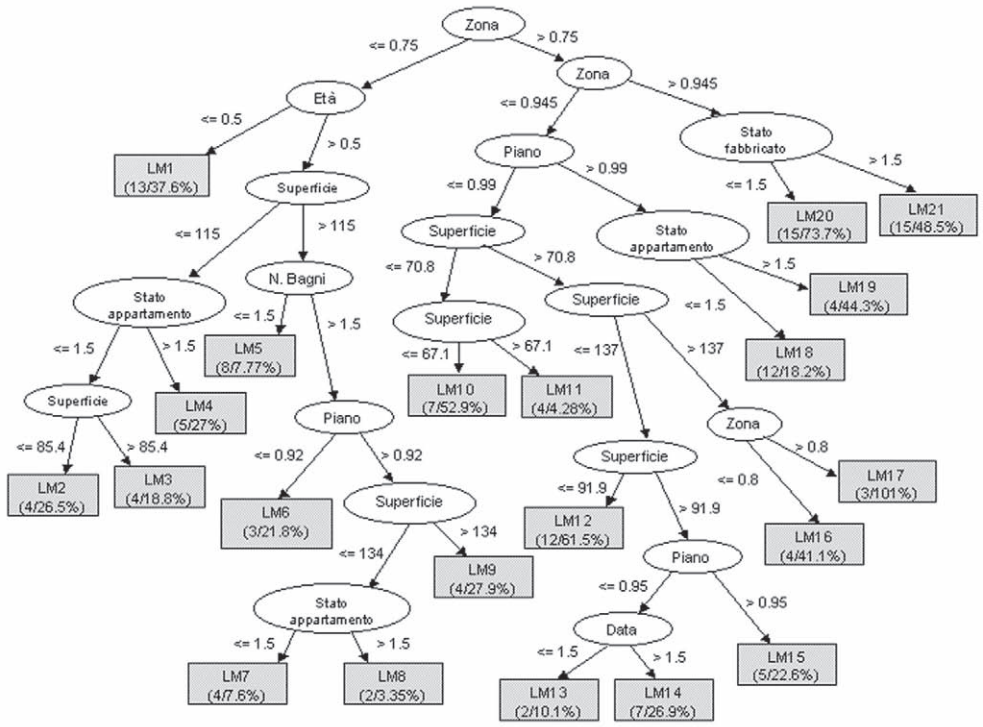

Fig. 3. Albero con 21 ramificazioni

La seconda soluzione, invece, conduce ad un albero con quattro ramificazioni e quattro modelli (Fig. 4 e tabella 4). La prima variabile discriminante è rappresentata dalla *zona* con un limite di 0,75; successivamente le osservazioni con valore inferiore o uguale a detto limite vengono ulteriormente segmentate in funzione della variabile *età* (con un limite di 0,5), mentre quelle con valore superiore a 0,75 vengono suddivise nuovamente in funzione della variabile zona, con un limite di 0,945.

I subcampioni così ottenuti presentano una numerosità di osservazioni abbastanza congrua, da un minimo di 13 ad un massimo di 60, e hanno condotto a modelli di regressione con all'interno variabili diverse sia come numero che come qualità. Infatti, i primi due modelli sono caratterizzati dalla presenza di 9 variabi-

Tabella 3
Modelli relativi ai 21 subcampioni

n.	Costante	Zona	Superficie	Piano	Età	Stato Fabbricato	Stato Appartamento	Presenza Locatario	Riscaldamento	N. Bagni	Portiere	Rivestimento esterno	Data
LM1	264	825	0,974		65,1	38		55,4	98,2	-62,5	70,1		32,8
LM2	409	939	0,557		55,3	38		31,7	23,3	-35,7	70,1		32,8
LM3	435	939	0,37		55,3	38		31,7	23,3	-35,7	70,1		32,8
LM4	70	1400	0,557		55,3	38		31,7	23,3	-35,7	70,1		32,8
LM5:	82,7	999	1,85		55,3	62,4	28,1	108	23,3	-60	70,1		32,8
LM6	-122	1270	2,03		55,3	58,1	38,7	86,9	23,3	-55,7	70,1		32,8
LM7	-194	1400	1,89		55,3	58,1	41,8	86,9	23,3	-55,7	70,1		32,8
LM8	-196	1400	1,92		55,3	58,1	42,1	86,9	23,3	-55,7	70,1		32,8
LM9	-207	1420	1,92		55,3	58,1	39,6	86,9	23,3	-55,7	70,1		32,8
LM10	-385	1210		566	65,5	63	20,2		-15,5		97,8		114
LM11	-464	1290		566	65,5	63	20,2		-15,5		97,8		114
LM12	-123	622	-1,36	917	65,5	63	20,2		-15,5		97,8		183
LM13	-587	1150	-1,32	1010	65,5	63	20,2		-15,5		97,8		141
LM14	-593	1150	-1,21	1010	65,5	60,2	20,2		-15,5		97,8		141
LM15	-594	1150	-1,32	1040	65,5	63	20,2		-15,5		97,8		141
LM16	-1140	1970	-1,46	870	65,5	63	20,2		-15,5		97,8		143
LM17	-1150	2000	-1,46	870	65,5	63	20,2		-15,5		97,8		143
LM18	193	755	-0,27	397	65,5	96	64	-15,5			97,8		76
LM19	102	755		397	65,5	140	74,7	-15,5			97,8		76
LM20	-858	1040		898	360	280	58,9		-25,8		245		56,6
LM21	-601	1040	-0,59	898	311	280			-25,8		579	139	56,6

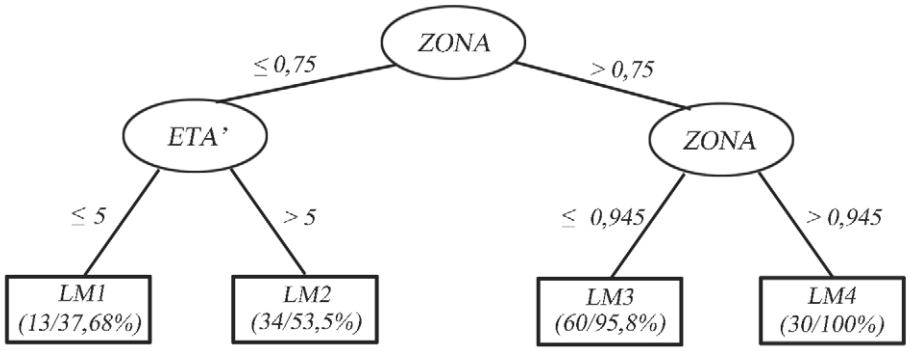

Fig. 4. Albero con 4 ramificazioni

li esplicative (oltre alle due variabili discriminanti, zona ed età, troviamo lo *stato del fabbricato*, la *superficie*, la *presenza del locatario*, la tipologia dell'*impianto di riscaldamento*, il numero dei *bagni*, la *presenza del portiere* e la *data della compravendita*), mentre gli altri due da 5 variabili (oltre alla variabile discriminante, la zona, troviamo *l'età dell'immobile*, lo stato del fabbricato, la presenza del portiere e la data di compravendita). Dalla Fig. 4 si osserva che nelle quattro foglie dell'albero, ciascuna individuata da un modello lineare (LM), sono contenuti: a) il numero delle osservazioni che le raggiungono; b) il valore dello scarto quadratico medio percentuale.

L'Albero Modello nel complesso presenta un Coefficiente di Correlazione pari a 0,684, laddove si intende che il 68.4% è la misura della correlazione statistica tra i valori osservati e quelli calcolati (grafico 1). Tale coefficiente varia da 0 a ± 1: 1 per valori perfettamente correlati, -1 quando sono correlati negativamente; assume valore uguale a 0 quando non esiste alcuna correlazione.

Tabella 4
Modelli relativi ai 4 subcampioni

Variabili	LM1	LM2	LM3	LM4
Costante	452	180	729	-56,2
Zona	550	1250	751	1040
Superficie	1,11	0,636		
Età	42,3	42,3	60,6	371
Stato fabbricato	38	38	65,6	474
Locatario	61,5	35,2		
Riscaldamento	103	25,7		
N. Bagni	-64,5	-36,8		
Portiere	70,1	70,1	101	141
Data	32,8	32,8	19,3	19,3

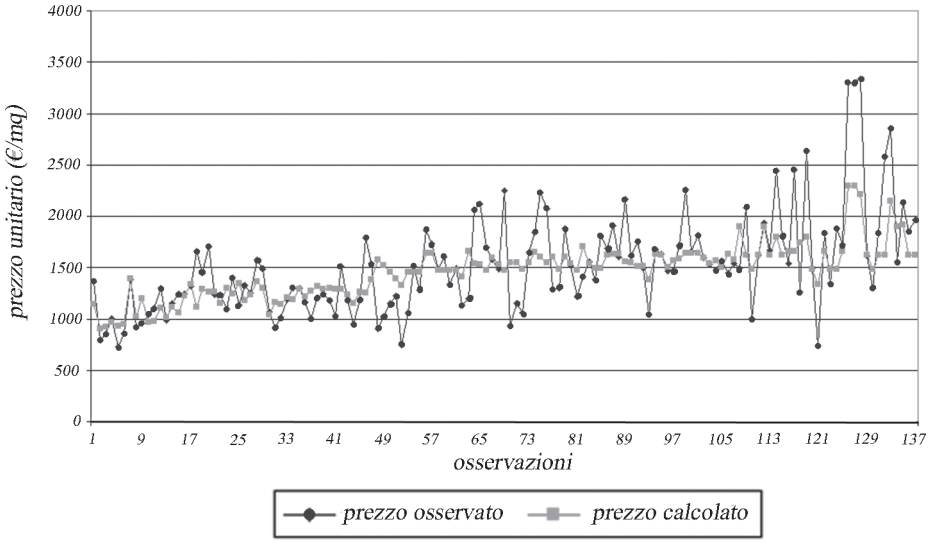

Grafico 1. Correlazione tra valori osservati e valori calcolati

Un altro punto che ora vogliamo affrontare riguarda l'opportunità di utilizzare questa tecnica di analisi in sede di valutazioni immobiliari.

Tale opportunità ovviamente non potrà che scaturire dalla maggiore attendibilità dei risultati di stima rispetto a quella derivante dall'uso delle tecniche regressive tradizionali (il riferimento è ai modelli di regressione lineare multipla).

A tale fine abbiamo messo a confronto gli errori percentuali assoluti medi di ogni sub campione, ottenuti utilizzando il modello di regressione lineare unico e il modello derivante dell'albero (tabella 5).

I risultati evidenziano senza alcun dubbio la maggiore attendibilità dei modelli dell'albero, in quanto caratterizzati sempre da un errore più contenuto rispetto a quello del modello unico.

Ciò vuol dire che il grado di accostamento dei valori calcolati ai prezzi di mercato è di gran lunga più elevato quando si fa riferimento ai modelli dell'albero.

Nella tabella 6 si riportano i valori (€/m^2) di appartamenti standard[4] per ogni subcampione determinati con le due tecniche di analisi; quella tradizionale che fa

[4] Per i primi due subcampioni (1 e 2) la variabile zona presenta un uguale valore, pari a 0,67, mentre i valori del parametro età sono pari, rispettivamente, a 0 e 2. Per i subcampioni 3 e 4, la variabile zona è pari a 0,80 e 1 rispettivamente. I valori delle altre variabili, per i quattro subcampioni, sono i seguenti: superficie di 90 m^2 piano, 0,90; portiere, posti auto, presenza del locatario, rivestimento esterno, presenza di interno, 0; bagni, stato del fabbricato e dell'appartamento, 1; riscaldamento e data della compravendita, 2.

Tabella 5
Errori percentuali assoluti medi

	Albero di Decisione	Analisi di Regressione
Subcampione 1	11,6	22,6
Subcampione 2	12	19
Subcampione 3	16,9	17,7
Subcampione 4	6,5	20,6
Totale	12,9	19,1

Tabella 6
Prezzi unitari calcolati per appartamenti standard (€/mq)

	Albero di Decisione	Analisi di Regressione
Subcampione 1	1.165,50	728,60
Subcampione 2	1.277,50	1.079,40
Subcampione 3	1.549,20	1.376,50
Subcampione 4	2.238,40	1.833,50

riferimento ad un unico modello di regressione[5] e quella che qui si suggerisce, relativa all'albero di decisione[6].

Gli scarti di non lieve entità dei risultati ottenuti dovrebbero rappresentare un ulteriore elemento di considerazione nei confronti della tecnica relativa all'Albero Decisionale.

6. Conclusioni

La possibilità, avanzata in premessa, di trovare un nuovo percorso valutativo caratterizzato, ovviamente, da un maggior grado di attendibilità dei risultati, non apparirebbe velleitario, almeno con riferimento all'esempio concreto che è stato considerato.

Il mercato degli immobili urbani in Bari, infatti, caratterizzato da non lievi differenziazioni nel livello delle quotazioni, conseguenza di una estrema variabilità nelle caratteristiche immobiliari, ben si presterebbe ad essere indagato attraverso modelli del tipo qui suggerito: l'albero delle decisioni.

[5] Equazione di regressione: y= −944.3+2285 (zona) +175.4 (età) +142 (stato fabbricato) + 259.7 (interno) + 239 (portiere)

[6] Si rileva, inoltre, la costante superiorità dei prezzi calcolati con la tecnica decisionale rispetto agli altri; sarà questo un ulteriore elemento di riflessione per approfondimenti successivi.

Il salto in avanti nella qualità delle stime immobiliari, con tutte le cautele che devono ovviamente caratterizzare le nuove sperimentazioni, potrebbe assumere carattere di certezza, se supportato da ulteriori applicazioni in altri segmenti del mercato immobiliare.

Il presente lavoro ha voluto rappresentare, per l'appunto, un primo tentativo sulla strada, non priva di difficoltà, che deve portare ad un continuo miglioramento delle stime.

La nostra speranza è di avere sollecitato la curiosità nei cultori della materia e che ulteriori e ben più insigni approfondimenti seguano a questa breve nota, per colmare le tante lacune qui presenti.

Bibliografia

Acciani C. 1996. L'Analisi di Regressione Multipla nelle valutazioni immobiliari. *Genio Rurale* 12: 23-31

Acciani C. & Bozzo F., 2004. *Il mercato immobiliare urbano nella città di Bari*. Bari, Grafiche Eurostampa

Barrai I., 1984. *Metodi di regressione e classificazione in biometria*. Bologna Edagricole. pp. 65-94

Berenson M.L. & Levine D. M. 1993. *Statistica per le scienze economiche*. Zanichelli, Bologna, pp. 549-688

Berloco A.D,. Fratepietro G. & Grittani G. 1991. La valutazione a più parametri: dalla teoria alla prassi. *Genio Rurale* 10: 15-20

Coletta A, 1997. La valutazione a più parametri: un nuovo approccio operativo basato sull'impiego delle reti neurali artificiali. *Genio Rurale* 2: 33-40

Del Giudice V.& Amabile R. 1996. Reti neurali nelle valutazioni estimative ed economiche. *Genio Rurale* 5: 3-7

Gastaldi T. 2002. Data Mining: Alberi Decisionali e Regole Induttive per la profilazione del cliente. *Scienze & Business*, Anno IV, n. 5-6: 16-25

Gori A. 2003. Data Mining, in www.sdipisistemi.it

Grillenzoni M. & Grittani G. 1994. *Estimo – teoria, procedure di valutazione e casi applicativi*. Bologna, Calderini. pp. 71-80

Mello P. 2002. Intelligenza Artificiale, in Dizionario Interdisciplinare di Scienza e Fede. Sito Internet: www.lia.deis.unibo.it/Courses/AI/articoli

Milano G. 1968. L'analisi della regressione nella valutazione dei fondi rustici. *Annali della Facoltà di Agraria dell'Università di Bari* XXII: 443-424

Milano G., Bruno R., Fratepietro G.& Bozzo G. 1978. Correlazione fra caratteristiche intrinseche e prezzi di mercato dei fondi rustici in periodo breve. *Annali della Facoltà di Agraria dell'Università degli Studi di Bari* XXX: 431-460

Morano P. 2001. Un modello di regressione di outlier per l'analisi del mercato immobiliare. *Genio Rurale - Estimo e Territorio* 10: 19-35

Quinlan J.R.1992. Learning with continous classes, in Procedings AI '92 (Adams & Sterling Eds.), 343-348, Singapore: World Scientific, 1992.

Roiger R.J. & Geatz M.W. 2003. *Introduzione al Data Minino*. McGraw-Hill, Milano, pp. 3-91

SAS Institute Inc., SAS/STAT, Cary, NC, USA, 1987

Schinardi R. 2001. Analisi introduttiva all'Intelligenza Artificiale, Sito internet www.alpha01.unito.it/personalpages/cerruti/studenti.html

Simonotti M. 1988. L'analisi di regressione nelle valutazioni immobiliari, in Studi di Economia e Diritto. *Bollettino degli Interessi Sardi* 3: 369-401

Simonotti M. 1989. *Fondamenti di metodologia estimativa*. Liguori Editore, Napoli, pp. 237-273

Simonotti M. 1991. Un'applicazione dell'analisi di regressione multipla nella stima di appartamenti. *Genio Rurale* 2: 9-15

Walker R., Teoria e sistemi di Intelligenza Artificiale, Sito internet: www.gral.ip.rm.cnr.it/r.walker/naplestexts.

Witten I.H. & Frank E. 2000. *Data Mining: practical machine learning tools and techniques with Java implementations*. Department of Computer Science, University of Waikato, Morgan Kaufmann Publishers, pp. 70-71 e seguenti

www.cs.waikato.ac.nz/ml/weka/

www.dizionarioinformatico.com