Consiglio Nazionale delle Ricerche

Prototipo per l'estrazione di informazione geografica da dati epidemiologici

Roberto della Maggiore, Roberto Fresco

Nota Tecnica
CNUCE-B4-2000-019

CNUCE

Pisa

Consiglio Nazionale delle Ricerche

Prototipo per l'estrazione di informazione geografica da dati epidemiologici

Roberto della Maggiore, Roberto Fresco

Nota Tecnica
CNUCE-B4-2000-019

email: Roberto.dellaMaggiore@cnuce.cnr.it

CNUCE

Via Alfieri 1, Area della Ricerca San Cataldo
56010 GHEZZANO (PI)
Tel. +39 050 315 2215 - Fax +39 050 31380091/2

Indice

Struttura del prototipo

Le informazioni relative al luogo di residenza dei soggetti partecipanti all'indagine sono state integrate in un sistema informativo geografico (GIS) nell'ambito del quale, ogni soggetto e' stato localizzato esattamente sul territorio attraverso il proprio luogo di residenza (georeferenziazione).
Per la classificazione dei soggetti si e' fatto ricorso alla logica fuzzy, secondo la quale un soggetto può appartenere contemporaneamente (con percentuali diverse) a più di una delle classi definite.
Il prototipo sviluppato e' essenzialmente costituito da due fasi:

1) definizione delle funzioni di appartenenza legate alla fuzzificazione dei dati epidemiologici;
2) uso del modello per la selezione di dati dal database epidemiologico con conseguente formazione di strati informativi GIS, utilizzabili per le successive analisi di tipo geostatistico.

Il software con funzionalità GIS utilizzato e' ArcView GIS 3.1 della ESRI™, mentre l'applicazione per la valutazione fuzzy dei soggetti e' stata realizzata con il prodotto Microsoft™ Visual Basic 6.0, collegato ad ArcView attraverso il meccanismo di interazione DDE ([DemaRF2000]), tipico del sistema operativo Microsoft Windows.

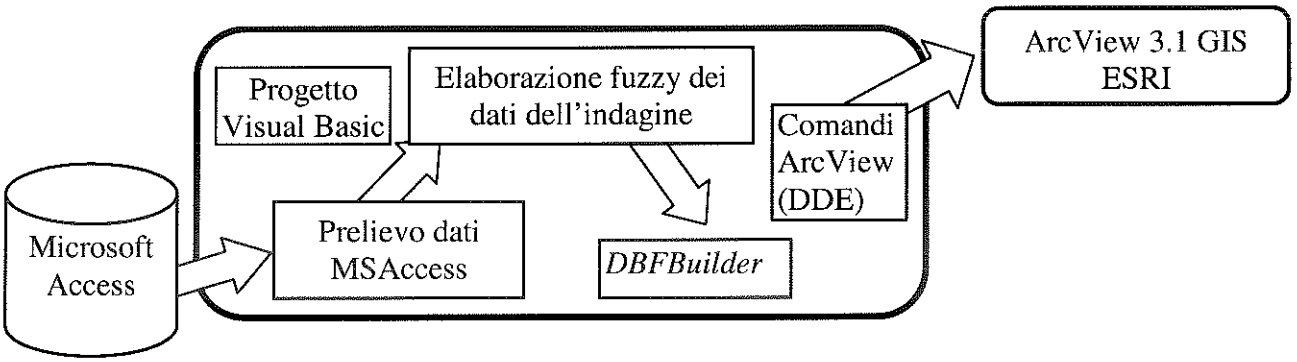


Fig.1: Struttura del prototipo

Cenni di logica fuzzy.

L'incertezza si riferisce alla nostra imperfetta ed inesatta conoscenza dell'ambiente che ci circonda. Si possono distinguere ([Rossiter95]) due classi di incertezza:

- nei dati, ovvero alle osservazioni che noi possiamo fare dei fenomeni umani, sociali e territoriali;
- nelle decisioni, ovvero al modo con cui sono analizzate le osservazioni effettuate in relazione al fatto che possiamo non essere sicuri circa le conclusioni tratte dai dati.

Un primo passo nell'utilizzo dei dati e' costituito dalla loro classificazione.
Quando occorre effettuare una classificazione di dati generalmente si definisce un numero finito di classi ed ogni elemento appartiene ad una sola classe (tipicamente si dice che un elemento appartiene/non appartiene ad una classe); si ha quindi una classificazione netta, ben distinta (*crisp*) degli elementi.
L'incertezza della classificazione sta nella impossibilità, in alcuni casi, di identificare con esattezza la classe alla quale l'elemento appartiene. Allora accade che il risultato a posteriori rappresenti una "stima" della compatibilità dell'elemento alla classe, ovvero l'appartenenza può essere stabilita nella classe con il piu' alto grado di vicinanza semantica. Possiamo parlare di piena appartenenza dell'elemento, attraverso una funzione che assume il valore uno per questa classe e zero per le altri classi. Nel nostro caso, come avviene nella classificazione fuzzy, può essere invece significativo considerare il concetto di *grado di appartenenza* alle classi ([De Bruin2000]) basata sul concetto degli insiemi fuzzy. Nel modello degli insiemi fuzzy la funzione di assegnazione alla classe attribuisce ad ogni elemento un grado di appartenenza nell'intervallo continuo [0,1] per ogni insieme che e' stato definito.

Questo grado di appartenenza (membership) corrisponde al grado con il quale l'elemento e' compatibile al concetto rappresentato da quell'insieme. Per questo motivo gli insiemi fuzzy permettono la rappresentazione di classi definite in modo impreciso ovvero tramite concetti espressi in modo linguistico. Per esempio se affermiamo che un soggetto fuma un numero eccessivo di sigarette ciò che rimane 'fuzzy' e' quello che noi intendiamo per il termine 'eccessivo'.
Il nome 'fuzzy' deriva da una teoria matematica (fuzzy set theory) sviluppata da Lofti.A. Zadeh presso l'università della California nel 1965. Venne così proposta la logica fuzzy il cui principio base e' contrassegnato dagli insiemi fuzzy.
Se X e' l'universo di possibili elementi allora un insieme fuzzy A in X e' rappresentato come un insieme di coppie ordinate:

$$A = \{x, \mu_A(x) | \, x \in X\}$$

in cui[1]:

- x e' un generico elemento dell'insieme X; per esempio x e' un fumatore di sigarette nell'insieme X di tutti i fumatori;
-

[1] La notazione usata per rappresentare l'insieme fuzzy, ovvero che ad ogni elemento x di X corrisponda un valore $\mu_A(x)$, può dar luogo all'idea di un rappresentazione per un numero *discreto* di valori; invece va osservato che, in relazione al dominio X dei possibili valori, la notazione implica una *continuità* di valori per cui ha senso per ogni elemento x_i di X calcolare il suo grado di appartenenza $\mu_A(x_i)$ ad X.

- $0 \leq \mu_A(x) \leq 1$ e' la funzione di appartenenza di x in A. Intuitivamente 1 = totalmente nell'insieme, 0 = totalmente escluso dall'insieme. La funzione di appartenenza mappa ogni elemento di X in un valore di appartenenza compreso tra 0 e 1. Tale funzione esprime anche la misura di quanto l'elemento x sia compatibile con il concetto fuzzy di A. Nell'esempio di sopra A potrebbe essere relativo al concetto vago di numero di sigarette *eccessivo*. Gli insiemi numerici della logica booleana (crisp sets) permettono soltanto i valori di 1/0, corrispondenti a vero/falso, in/out, giusto/sbagliato e così via, ovvero valori ben definiti; le rappresentazioni grafiche piu' comuni di queste funzioni possono essere triangolari e trapezoidali.

L'idea degli insiemi fuzzy e' quella di non chiedere se x appartiene alla classe A_i ma piuttosto quella di chiedere quale sia il grado di appartenenza di x relativamente a tutte le classi A_i per tutti i possibili valori di i.
Esemplifichiamo il concetto degli insiemi fuzzy; prendiamo l'insieme delle persone e classifichiamole in base all'altezza in metri secondo l'impostazione della logica booleana:

B (persone basse) con altezza < 1.65,
M (persone medie) con altezza compresa tra 1.65 e 1.75,
A (persone alte) con altezza > 1.75.

Si noti che abbiamo dovuto scegliere valori ben precisi per discriminare l'appartenenza agli insiemi suddetti. E' molto importante rilevare che la transizione da un insieme all'altro e' brusca: ad esempio il valore altezza = 1.75 non appartiene ad A mentre vi appartiene il valore quasi identico altezza = 1.7501. Questa situazione e' contraria al senso comune che certamente non classificherebbe in categorie diverse persone la cui altezza differisce di un millimetro o di una sua frazione.
Esprimendo per mezzo di insiemi fuzzy, possiamo avere, come esempio, una caratterizzazione di questo tipo:

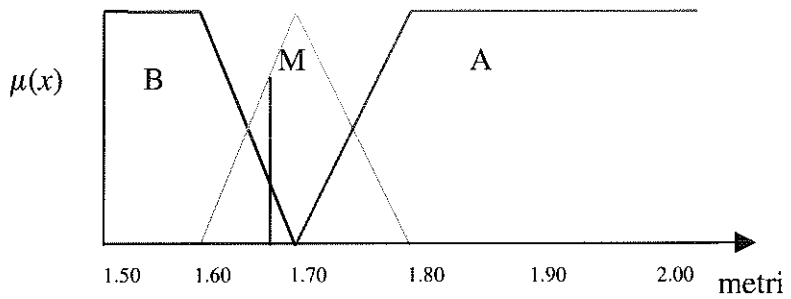


Fig.2: insiemi fuzzy per la classificazione delle altezze

dove le funzioni di appartenza (membership) alle tre classi (Bassa, Media ed Alta) sono definite nel seguente modo:

$$\mu_B(x) = \begin{cases} 1 & x \leq 1.60 \text{ cm} \\ -\frac{x}{0.1} + \frac{1.6}{0.1} + 1 & 1.60 < x < 1.70 \text{ cm} \\ 0 & x \geq 1.70 \text{ cm} \end{cases}$$

$$\mu_A(x) = \begin{cases} 0 & x \leq 1.70 \text{ cm} \\ \frac{x}{0.1} - \frac{1.7}{0.1} & 1.70 < x < 1.80 \text{ cm} \\ 1 & x \geq 1.80 \text{ cm} \end{cases}$$

$$\mu_M(x) = \begin{cases} 0 & x \le 1.60 \text{ cm} \\ \frac{x}{0.1} - \frac{1.6}{0.1} & 1.60 < x \le 1.70 \text{ cm} \\ -\frac{x}{0.1} + \frac{1.7}{0.1} + 1 & 1.70 < x < 1.80 \text{ cm} \\ 0 & x \ge 1.80 \text{ cm} \end{cases}$$

Una persona di altezza pari a 1.67 m ha un grado di appartenenza all'insieme B pari a 0.3 e all'insieme M appartiene per un grado pari a 0.8, come evidenziato in fig.2, dove si notano delle intersezioni tra i vari insiemi; quindi la logica fuzzy non e' binaria (appartiene/non appartiene) bensì *multivalore*. Rileviamo inoltre da quest'esempio che non c'e', come nella logica classica, una transizione immediata da un insieme a quello adiacente. Ad esempio una persona alta m 1.79 ha membership $\mu_M(x)$, $\mu_A(x)$ rispettivamente vicine a 0 e 1 e poco nulla cambia se l'altezza fosse m 1.81; questa rappresentazione consente di percepire meglio il concetto (linguistico) di quanto un soggetto sia *alto* rispetto ad un altro.
Usando quindi la teoria degli insiemi fuzzy si ha la possibilità di modellare quella che e' l'incertezza linguistica dei dati proprio perche' si possono modellare gli insiemi fuzzy in modo che essi corrispondano a variabili linguistiche (di solito aggettivi, come basso, alto, caldo ecc.) tipiche del linguaggio naturale.

L'elaborazione dei dati attraverso gli insiemi fuzzy

Il problema classico nella teoria della logica fuzzy e' quello di combinare gli insiemi fuzzy in un modo che sia:
matematicamente consistente;
corrispondente all'idea con cui noi formuliamo i concetti linguistici.
La soluzione consiste nella formulazione di **regole** del tipo *if...then...*; la parte che segue immediatamente l'if costituisce la *premessa* della regola e quella che segue immediatamente il then ne e' la *conclusione*[2]. Il mapping in senso matematico dell'input all'output e' così rappresentato dalla struttura della regola, nella cui premessa e conclusione si utilizzano i concetti ed operatori linguistici corrispondenti ad altrettanti insiemi fuzzy ([Fresco2000]). Un sistema fuzzy e' un sistema in cui input ed output sono insiemi fuzzy e l'output viene calcolato, in funzione dell'input, applicando le regole. Il funzionamento di un sistema fuzzy comprende le seguenti fasi:

a. fuzzificazione dell'input;
b. inferenza fuzzy;
c. composizione degli output fuzzy;
d. defuzzificazione.

[2] Nei problemi di logica classica le conoscenze relative ad un dominio possono essere rappresentate da espressioni logiche come le implicazioni del tipo A →B che equivalgono a regole tipo *If A Then B* dove A e B sono le variabili che costituiscono, rispettivamente, la premessa e la conclusione della regola. Con espressioni logiche come questa e' possibile effettuare deduzioni ovvero inferenze sulla verità di altre proposizioni logiche. Analogamente anche nella logica fuzzy si possono definire regole che determinano processi inferenziali.

Fuzzificazione

In generale gli input di un sistema non sono fuzzy ma determinati valori numerici (*crisp*) di grandezze fisiche, economiche ecc. come per esempio 25 gradi di temperatura, 1.70 m di altezza e così via. La fuzzificazione ([Cammarata97]) e' un processo di trasformazione dei valori numerici o crisp dell'input nei corrispondenti insiemi fuzzy input; per far questo occorre definire le funzioni di appartenenza (membership) relative alle variabili in input. Poiche' si hanno anche variabili in output (generalmente corrispondenti alle conclusioni delle regole e con un proprio dominio di valori ammissibili) e' necessario effettuare la fuzzificazione anche per l'otuput ovvero stabilire le funzioni di appartenenza che consentono di trasformare i valori crisp per le variabili di output in valori fuzzy.

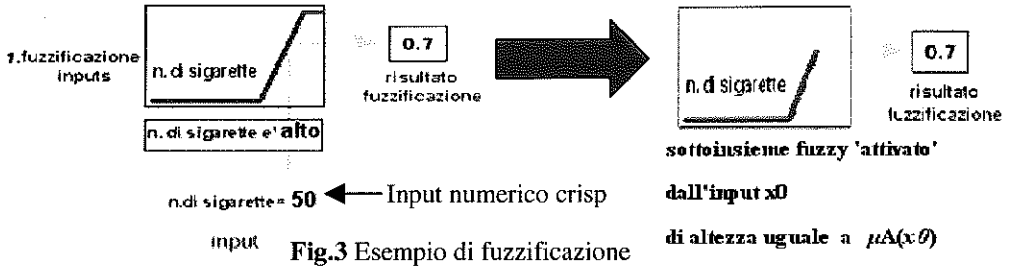


Fig.3 Esempio di fuzzificazione

Dato un valore x_0 (crisp) dell'universo del discorso e un insieme fuzzy A, e' possibile calcolare il valore $\mu_A(x_0)$ e si ottiene come risultato un sottoinsieme A' di A (si dice che A' e' attivato) avente come ordinata massima $\mu_A(x_0)$. In generale un valore numerico (nella fig.3 il n. di sigarette) attiva diversi insiemi fuzzy, se questi sono sovrapposti, e i corrispondenti gradi di verità sono generalmente diversi. Riassumiamo il concetto di fuzzificazione attraverso i seguenti punti:

- gli insiemi fuzzy descrivono concetti vaghi, incerti (altezza media, n.di sigarette alto); un insieme fuzzy ammette la possibilità di parziale appartenenza in esso (m 1.68 e' altezza un poco bassa e quasi media);
- il grado con cui un oggetto appartiene ad un insieme fuzzy e' denotato da un valore di appartenenza dell'intervallo [0,1] (m.1.67 all'insieme B per un grado 0.3 e all'insieme M per 0.8);
- una funzione di appartenenza associata ad un insieme fuzzy mappa un valore di input all'appropriato valore di membership.

Inferenza fuzzy

Come abbiamo detto sopra, il punto chiave della logica fuzzy e' quello di effettuare un *mapping* da uno spazio di input ad uno spazio di output e il meccanismo per far ciò e' quello di definire una lista di regole *If...Then* e l'ordine di valutazione delle regole non e' importante. Va osservato, inoltre, che le regole stesse si riferiscono alle variabili e aggettivi che descrivono quelle variabili; prima di costruire un sistema che interpreta le regole occorre definire tutti i termini linguistici (ovvero definire gli insiemi fuzzy che li descrivono) che prevediamo di usare nelle regole stesse; questa fase come, già detto, e' definita come fuzzificazione; sia i termini linguistici in input che quelli in output vanno espressi tramite delle opportune funzioni di membership ([Cammarata97]). I vari termini sono collegati nella stessa regola tramite gli operatori logici NOT, OR, AND. Nella figura successiva vediamo un diagramma per il processo di inferenza fuzzy. A sinistra si evidenzia una descrizione generale di un sistema fuzzy e a destra uno specifico esempio (la determinazione del grado di esposizione al fumo passivo):

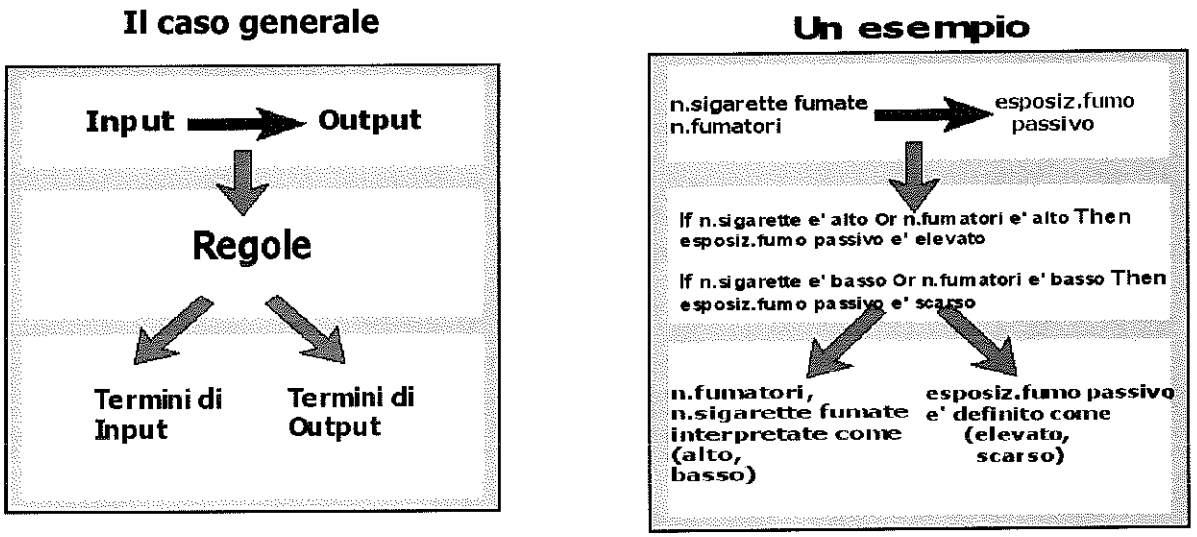


Fig.4 Relazione tra Input ed Output nel processo di inferenza fuzzy

Per riassumere la figura precedente possiamo dire che l'inferenza fuzzy e' un metodo che interpreta i valori in input e, in base ad un insieme di regole, assegna valori all'insieme di output.
Le regole fuzzy sono del tipo: ***if x e' A then y e' B*** dove A e B sono termini linguistici definiti da insiemi fuzzy sui domini X e Y, rispettivamente la parte if della regola "x e' A" e' la premessa o antecedente (calcolo di $\mu_A(x)$) e la parte Then della regola "y e' B" e' la conclusione o conseguente. Un esempio di una regola può essere

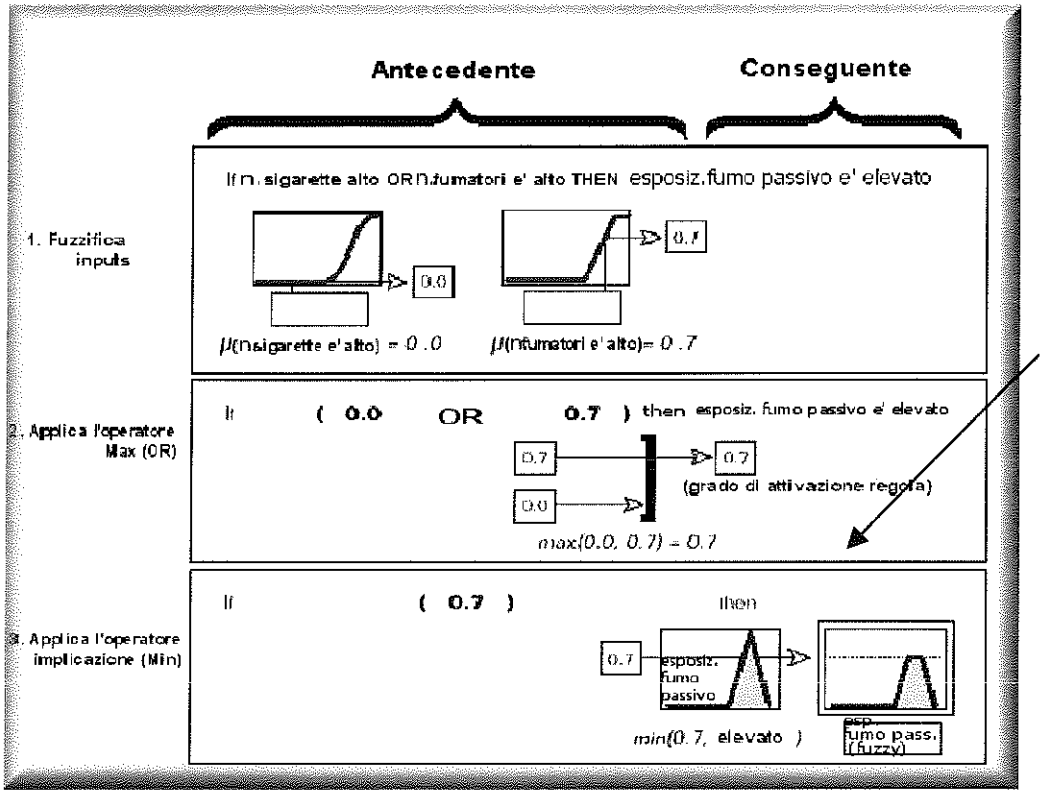
If n. di sigarette e' alto Then Esposizione fumo passivo e' elevato

Si noti che *alto* e' rappresentato come un numero tra 0 e 1 e così l'antecedente e' interpretato come un singolo valore tra 0 e 1; dall'altro canto *elevato* e' rappresentato come un insieme fuzzy e così la conclusione e' un assegnamento poiché l'intero insieme B e' assegnato alla variabile di output y.
Il valore di $\mu_B(x)$ viene calcolato in funzione di $\mu_A(x)$ e costituisce ciò che e' definito come procedimento di implicazione.
E' possibile avere un antecedente che coinvolge più di un insieme fuzzy, usando i connettivi logici AND, OR, NOT; in tal caso il risultato dell'antecedente e' sempre un singolo valore dell'intervallo [0,1]. Un esempio di una regola di questo tipo e'

If n. di sigarette e' alto OR n. di fumatori e' alto Then Esposizione fumo passivo e' elevato

In tal caso si definiscono gli operatori fuzzy ([Fresco2000]):
OR: max $(\mu_A(x), \mu_B(x))$
AND: min $(\mu_A(x), \mu_B(x))$
dove A e B insiemi fuzzy dell'antecedente cioè n.sigarette e n. di fumatori.
Il conseguente della regola specifica un insieme fuzzy che e' assegnato come output; il conseguente viene alterato dall'antecedente mediante la *fase di implicazione* che modifica quell'insieme fuzzy in base al grado specificato dall'antecedente.

Uno dei modi più semplici di modificare l'insieme fuzzy output è quello di usare la funzione minimo (l'insieme fuzzy e' *tagliato* come mostrato nella figura successiva):
L'ottenimento di un singolo valore come risultato della valutazione dell'antecedente di una regola, a partire dagli inputs, si dice *attivazione* della regola e il valore ottenuto e' il *grado di attivazione* della regola.



Taglio della funzione output della conclusione della regola

Fig.5 Esempio di processo di inferenza fuzzy

Gli stessi input sono applicati a più regole contemporaneamente e l'attivazione delle regole creano, per inferenza, uno o piu' insiemi fuzzy output.
Il processo di inferenza di regole If...Then coinvolge tre fasi distinte:

- si fuzzificano gli inputs: si valutano tutte le asserzioni presenti nell'antecedente determinando un grado di membership compreso tra 0 e 1;
- si applicano gli operatori fuzzy agli antecedenti composti da piu' parti: se l'antecedente e' composta da più parti, allora vengono applicati gli operatori logici fuzzy in modo da attribuire a tutto l'antecedente un unico valore compreso tra 0 e 1. Questo e' definito come il grado di attivazione della regola; se c'è una sola parte nell'antecedente, allora la sua fuzzificazione e' il grado di attivazione della regola; generalmente si assume, come grado di attivazione di una regola (detto anche "alfa cut"), il minimo dei gradi membership degli insiemi fuzzy dell'antecedente. L'ordine di valutazione delle regole non e' rilevante;
- si applica il metodo di implicazione: si usa il grado di attivazione, prodotto dagli antecedenti, per determinare la forma geometrica dell'insieme fuzzy di output. Il conseguente di una regola fuzzy assegna un intero insieme fuzzy all'output. Questo insieme e' rappresentato da una funzione di appartenenza che indica le caratteristiche del conseguente. Come detto anche i termini linguistici output hanno una propria funzione di appartenenza (ovvero una forma geometrica che li rappresenta).Se l'antecedente e' solo parzialmente vero (il grado di attivazione e' un valore minore di 1) allora la forma dell'insieme output fuzzy viene modificato in relazione al metodo di implicazione usato (generalmente viene usato la funzione del minimo).

Composizione e defuzzificazione

Nel sistema fuzzy abbiamo generalmente più di una regola; l'output di ogni regola è un insieme fuzzy come visto nella figura precedente. Se abbiamo generalmente più insiemi fuzzy output, occorre comporre questi diversi insiemi (ovvero combinare tutte le regole) in un unico insieme fuzzy output. La composizione ([Cammarata97]) è il processo per mezzo del quale gli insiemi fuzzy, ottenuti dal processo di implicazione per ogni regola, sono combinati in un singolo insieme fuzzy.Il risultato della composizione e' un insieme fuzzy per ogni variabile di output; uno dei metodi generalmente usato e' quello della somma, ovvero gli insiemi fuzzy vengono sommati o uniti (nel senso dell'OR logico):

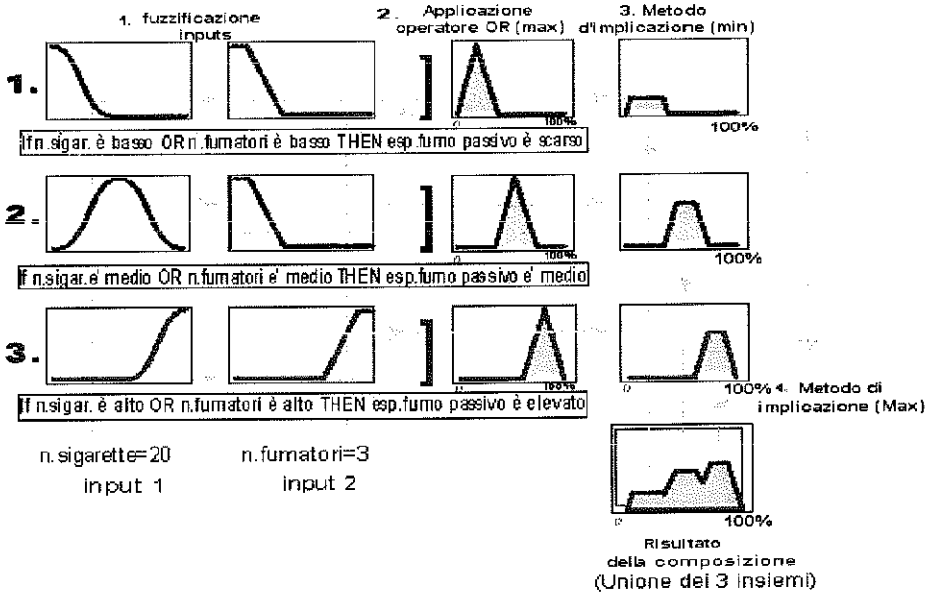


Fig.6 Esempio di processo di inferenza e composizione

Può però essere necessario determinare come risultato finale opportuni valori numerici (crisp) di 'sintesi' dell'insieme fuzzy output. L'operazione che fornisce questi valori numerici e' denominata *defuzzificazione* ([Cammarata97]). Essa consiste nel determinare il valore numerico più rappresentativo dell'insieme finale output.

Uno dei metodi più usati e' il calcolo del centroide che fornisce l'ascissa del baricentro della figura solida delimitata dall'insieme fuzzy output:

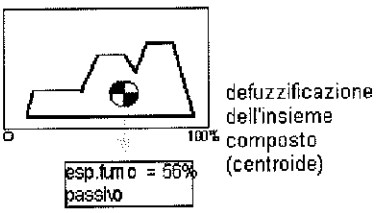


Fig.7 Risultato della defuzzificazione: l'insieme fuzzy che si *defuzzifica* e' quello risultato dalla fase di composizione della fig. precedente.

Modellazione dei dati

Avendo a disposizione i dati del questionario compilato dai soggetti partecipanti all'indagine, e' stato predisposto un prototipo software che consente di valutare, adottando la logica fuzzy, quei fattori che sono stati ritenuti di maggior rilievo per poter considerare un soggetto come effettivamente rappresentativo del luogo in cui si trova la sua abitazione, nel seguente modo:

- si definiscono criteri di **esclusione/inclusione** (base) dei soggetti dalla totalità dei partecipanti;
- sugli inclusi si opera una **modellazione fuzzy** per valutare il grado di affidabilità geografica.

I soggetti partecipanti furono circa 2800 e quindi un campione di popolazione abbastanza consistente, considerando l'estensione territoriale a cui si fa riferimento (prossimità della Tosco Romagnola); data la notevole dispersione di caratteristiche personali, e' stata fatta una selezione preliminare dei partecipanti di cui si vuole valutare il contributo informativo geografico.

Dai dati dell'indagine quindi si e' deciso di determinare una **base** di soggetti considerando criteri sottrattivi, in relazione a determinati gruppi di domande del questionario.

Per l'elaborazione dei dati in un sistema fuzzy e' essenziale disporre di:
- dati in input in formato numerico
- regole che consentono di valutare i dati

Per quanto riguarda il primo punto abbiamo osservato che i dati del questionario dell'indagine sono espressi secondo proprietà od espressioni linguistiche (come ad esempio i nomi delle sostanze aeroinquinanti, presenza di impianti di ventilazione nel luogo di lavoro ecc.) oppure secondo valori numerici (es. numero di sigarette fumate in media in un giorno, il numero di componenti della famiglia che fumano ecc.); per poter avere a disposizione dei dati in formato numerico, necessari all'elaborazione fuzzy, e' stato necessario utilizzare i valori numerici e/o associare ai termini linguistici dei pesi numerici, in un modo tale da sintetizzare l'informazione disponibile in un parametro numerico necessario alla definizione di insiemi fuzzy.
Per il secondo punto sono state definite regole che permettono di avere in output un valore espresso in percentuale che esprime l'attendibilità del soggetto in relazione al proprio luogo di residenza.

I criteri di esclusione

Il prototipo permette di escludere totalmente dalla valutazione alcune categorie di soggetti, come spiegato di seguito:

- Soggetti con malattie pregresse (maschi specialmente) ovvero patologie contratte durante il periodo bellico e/o servizio militare; nel questionario sono, a tal proposito, presenti le seguenti domande:

20. Ha fatto il servizio militare : └─┘
1.Si
2.No

Se si alla 20
20.a Ha contratto malattie durante il servizio di leva : └─┘
1. No
2. Malattie polmonari
3.Altre malattie

21. Ha subito prigionie durante il periodo bellico : └─┘
1.Si
2.No

Se si alla 21
21.a Ha contratto malattie durante il periodo di prigionia : └─┘
1. No
2. Malattie polmonari
3. Altre malattie

Si ritiene che l'insorgenza di malattie nel passato possa non essere significativa in relazione al luogo dove il soggetto alla data dell'indagine abitava. Nell'applicazione si ha la possibilità così di escludere coloro che hanno contratto malattie durante il servizio militare e/o il periodo bellico;

- Fumatori: si considera che il fumo sia un potenziale fattore di confondimento a causa della quantità e qualità delle sostanze inalate in modo diretto dai soggetti, indipendentemente dal luogo dove la persona risiede. Le relative domande del questionario sono:

56. Fuma sigarette attualmente
1. No
2. Si

se si come
1. Regolarmente
2. Occasionalmente

se fuma regolarmente sigarette
56.a Quante sigarette fuma al giorno : └─┘└─┘

Quindi e' stata introdotta la possibilità di escludere i fumatori che hanno dichiarato di fumare regolarmente;

- Ex-fumatori: alcuni studi (v. [Fresco2000]) ritengono che il soggetto che smette di fumare possa ridurre significativamente i danni respiratori in modo da riacquistare, dopo un certo numero di anni, la funzionalità polmonare simile ad un soggetto non fumatore (dal punto di vista statistico e' stato stabilito almeno un periodo di sette anni); nel questionario le domande pertinenti questo aspetto sono:
-

57. Se ora non fuma sigarette, le ha mai fumate in passato
1. No
2. Si

se si come
1.Regolarmente
2.Occasionalmente

57.d A che eta' ha iniziato a fumare :└__┴__┘

57.e Mediamente quante sigarette fumava al giorno :└__┴__┘

57.f Eta' ha smesso di fumare :└__┴__┘

Nell'applicazione si ha la possibilità di eliminare dalla valutazione fuzzy quei soggetti che hanno smesso di fumare da meno di un certo numero di anni (il valore di default e' sette);

- Soggetti che hanno un legame con lo stesso luogo di residenza per periodi limitati di tempo: un aspetto interessante da considerare, per l'attendibilità geografica del soggetto, e' sicuramente la durata della residenza del soggetto nella casa dove stabilmente abita. Il problema e' qui quello di stabilire dei range di significatività in base agli anni.
  Poiche' l'indagine e' stata svolta su base familiare, al capofamiglia e' stato chiesto di specificare da quanto tempo tutti i componenti della famiglia risiedono in quel luogo, caratterizzato da un proprio indirizzo sul territorio.
  Nel questionario le domande relative sono:

68. Da quanto tempo └__┴__┴__┴__┴__┴__┴__┘

1. Sempre
2. Meno di 2 anni
3. 2-5 anni
4. Piu' di 5 anni

I soggetti che risiedono da un periodo superiore a 5 anni sono sicuramente quelli che portano un contributo informativo piu' radicato sul territorio, invece per le altre fasce (2-5 anni e meno di 2 anni) si può avere incertezza; nell'applicazione si può decidere di includere od escludere quei soggetti che rientrano in queste categorie. Poiche' l'informazione sul tempo di residenza e' associata alla singola famiglia, e' ovvio che, decidendo di escludere le famiglie per le quali il capofamiglia ha dichiarato di abitare per es. da meno di 2 anni in quella casa, verranno esclusi tutti i componenti relativi.

Criteri per la modellazione fuzzy dei dati di input e output

Sul sottoinsieme di soggetti rimanenti dopo l'utilizzo dei criteri di esclusione, per i quali e' rilevante valutare il contributo informativo geografico (base), il propotipo consente di eseguire l'elaborazione su di essi attraverso la:
definizione dei parametri per la fuzzificazione delle variabili in ingresso;
selezione delle informazioni da restituire con strati GIS.

Per il primo punto abbiamo preso in considerazione le sezioni del questionario riguardanti principalmente il lavoro, il fumo, l'esposizione quotidiana agli aeroinquinanti.
Per ognuna di queste informazioni e' stato definito un parametro numerico che sintetizza l'esposizione; la determinazione di questo numero e' stata effettuata utilizzando i dati più significativi presenti sul questionario, ai fini dell'influenza sulla qualità e la quantità dell'esposizione. I parametri numerici sono l'ingresso *crisp* per il processo di fuzzificazione dei dati. Da notare che per uno stesso valore crisp si possono avere piu' valori relativi alle diverse funzioni membership della variabile linguistica che si sta considerando, un valore per ogni insieme fuzzy identificato da una funzione di membership. La determinazione di questi numeri e' relativa ai seguenti aspetti:

➧ Esposizione ad agenti sul posto di lavoro: al soggetto e' stato chiesto di indicare quali sono le sostanze con le quali abitualmente e' a contatto sul posto di lavoro:

Agenti

1. Silice
2. Asbesto
3. Berillio
4. Talco
5. Grafite (Carbone)
6. Vetro
7. Lana artificiale
8. Polveri di legno
9. Polveri di ferro
10. Polveri di zinco
11. Polveri(cadmio, cromo, piombo, vanadio, nichel)
12. Esalazioni auto
13. Altri fumi
14. Solventi
15. Insetticidi, fertilizzanti, antiparassitari

A queste sostanze si e' attribuito un indice di pericolosità in un intervallo di 1 a 12; alla sostanza più pericolosa per la salute e' stato associato il valore più alto.
Un altro dato utile per determinare l'importanza del lavoro per il soggetto e' rappresentato dagli anni di esposizione ad un certo agente. In questo caso e' importante rilevare che ci possono essere
state esposizioni multiple, ovvero a più agenti nel corso degli anni:

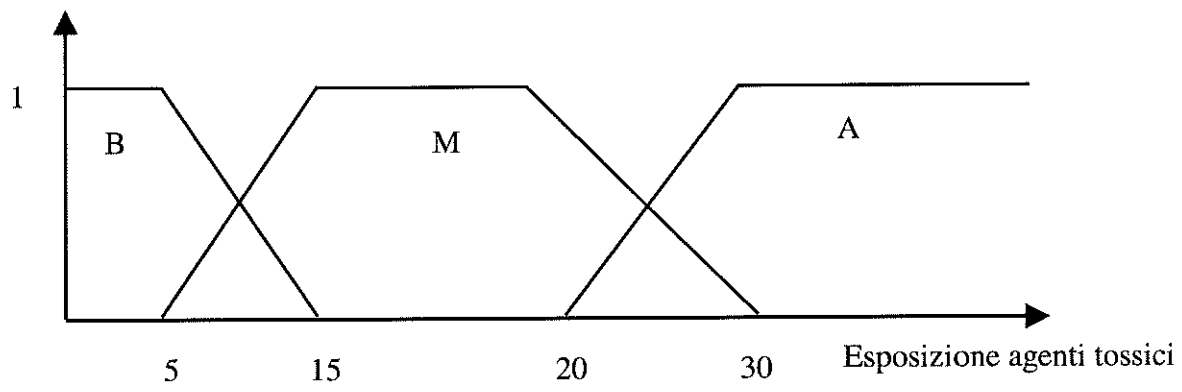
66.m Ha mai avuto nel suo lavoro una esposizione regolare ad uno degli agenti dell'elenco precedente:

Agenti	**Industria**	**Posizione Lavorativa**	**Esposizione per Anni**	**Esposizione Ultimo Anno**
⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋

Gli anni di esposizione sono stati ricavati dall'indicazione esplicita nel questionario (esposizione per anni).
Il parametro numerico e' stato posto uguale al rapporto:

$$\frac{\text{indice di pericolosità dell'agente} * \text{n. di anni di esposizione all'agente}}{12}$$

Se lo stesso soggetto ha dichiarato esposizioni multiple il valore di sopra viene sommato per ogni esposizione. In base ai valori numerici possibili sono stati stabiliti i tre termini linguistici di pericolosità agenti bassa, media ed alta. Sulle ascisse del grafico della funzione membership sottostante sono indicati i delimitatori numerici delle tre categorie.



➧ Caratteristiche dell'ambiente di lavoro: nel questionario sono presenti informazioni riguardanti la qualità dell'ambiente di lavoro:

66.d Indichi le caratteristiche degli ambienti di lavoro ultimo e/o precedenti:

Polveri	1=Assenti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	2=Scarsi	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	3=Abbondanti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
Fumi	1=Assenti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	2=Scarsi	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	3=Abbondanti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
Gas	1=Assenti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	2=Scarsi	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋
	3=Abbondanti	⌊__⌋	⌊__⌋	⌊__⌋	⌊__⌋

66.f Nell'ultimo ambiente di lavoro esiste o esisteva un impianto di ventilazione :
1. No
2. Centrale ad aria forzata
3. Centrale di altro tipo
4. Limitato a qualche stanza

La presenza di polveri, gas e fumi vari contribuisce all'esposizione nociva per la salute del soggetto e quindi la sua attendibilità legata al posto di residenza può diminuire. La presenza di un impianto di ventilazione può contribuire a ridurre l'esposizione stessa.
Il parametro definito in base a queste informazioni segue queste considerazioni:

1. alla presenza di polveri, fumi e gas si assegna un peso = 1, 2, 3 rispettivamente;
2. alla loro quantità: scarsa = ½ ;
3. alla loro quantità: abbondante = 1 ;
4. alla presenza di un impianto di ventilazione si assegna un fattore (di riduzione) = ½ (di default);

Quindi si calcolano i seguenti valori:

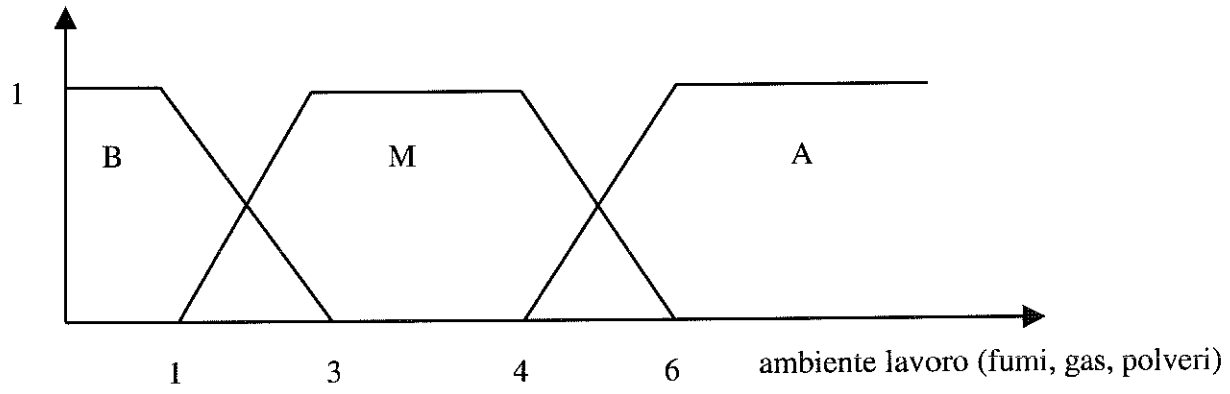
presenza_polveri * quantità; (A)

presenza_fumi * quantità; (B)

presenza_gas * quantità; (C)

Se e' presente l'impianto di ventilazione i valori (A), (B) e (C) vengono moltiplicati per il fattore di riduzione sopra specificato.

Quindi il parametro pericolosità_ambiente_di_lavoro viene posto uguale ad (A) + (B) + (C). Questi valori rappresentano gli input crisp per le funzioni membership dei termini linguistici (bassa, media, alta) relativi alla variabile pericolosità_ambiente_di_lavoro:



➧ Fumo passivo: e' ritenuto uno dei principali fattori di inquinamento indoor (degli ambienti confinati); l'esposizione a questo elemento può pregiudicare l'attendibilità del contributo informativo del soggetto in relazione al proprio posto di residenza.
Le domande del questionario che seguono sono state utilizzate nell'applicazione:

64. E' esposto abitualmente al fumo di sigaretta di altre persone:
␣ 1.Si
2.No

se si alla 64.
64.a Dove

In **casa** : ␣ Numero fumatori : ␣␣ Ore : ␣␣

Giorni/settimana : ␣␣

Posto di lavoro : ␣ Numero fumatori : ␣␣ Ore : ␣␣

Giorni/settimana : ␣␣

Altri ambienti : ␣ Numero fumatori : ␣␣ Ore : ␣␣

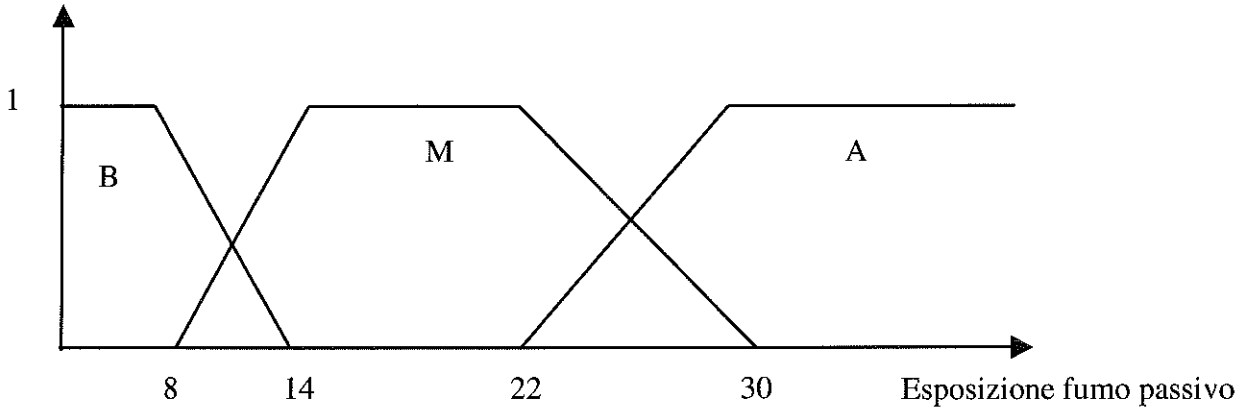
Giorni/settimana : ␣␣

L'esposizione a fumo passivo e' considerata in tre ambienti:

1. in casa;
2. nell'ambiente di lavoro;
3. in altri ambienti confinati.

Nel questionario vi sono informazioni sul numero di fumatori nei tre ambienti sopra menzionati, sul numero di ore al giorno e numero di giorni della settimana per cui vi e' esposizione.
Per la fuzzificazione vengono calcolati i valori intermedi:
A = (n. di ore al giorno di esposizione in casa) *(n. di giorni_settimana) * (n. di fumatori in casa);
B = (n. di ore al giorno di esposizione al lavoro) *(n. di giorni_settimana) * (n. di fumatori al lavoro);
C = (n. di ore al giorno di esposizioni in altri ambienti) *(n. di giorni_settimana) * (n. di fumatori in altri amb);

Al parametro di esposizione al fumo passivo viene assegnato il valore cumulativo di questi valori (A+B+C); si ottengono così i valori crisp per le tre funzioni membership dei tre termini basso, medio ed alto per il fumo passivo:



➧ Fumatori ed ex-fumatori: se i fumatori non sono stati esclusi attraverso la selezione della base dei soggetti il sistema dà la possibilità di stimare per ogni soggetto fumatore la quantità di fumo a cui e' stato esposto. Da notare che sia un fumatore o ex-fumatore può essere esposto anche al fumo passivo.
Le domande del questionario utilizzate sono:

56. Fuma sigarette attualmente 1. No
2. Si

se si come
1. Regolarmente
2. Occasionalmente

se fuma regolarmente sigarette
56.a Quante sigarette fuma al giorno : ␣␣

Nel caso il soggetto abbia smesso di fumare si considera invece la parte del questionario:

57. Se ora non fuma sigarette, le ha mai fumate in passato : ␣ 1. No
2. Si

se si come ␣ 1.Regolarmente
2.Occasionalmente

57.d A che eta' ha iniziato a fumare : ␣␣

57.e Mediamente quante sigarette fumava al giorno : ␣␣

57.f Eta' in cui ha smesso di fumare : ␣␣

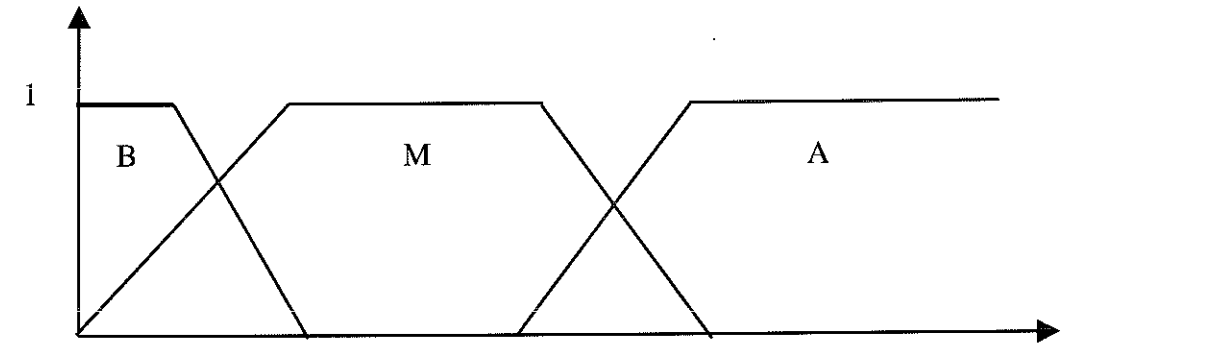
La differenza tra l'anno in cui il soggetto ha smesso di fumare e l'anno in cui ha iniziato stabilisce il numero di anni in totale di fumo *attivo*.
Tra i parametri usati dagli epidemiologi per valutare il fumo c'e' il valore cosiddetto **packyears** che costituisce una sorta di *integrale* di quanto un soggetto abbia fumato in passato:

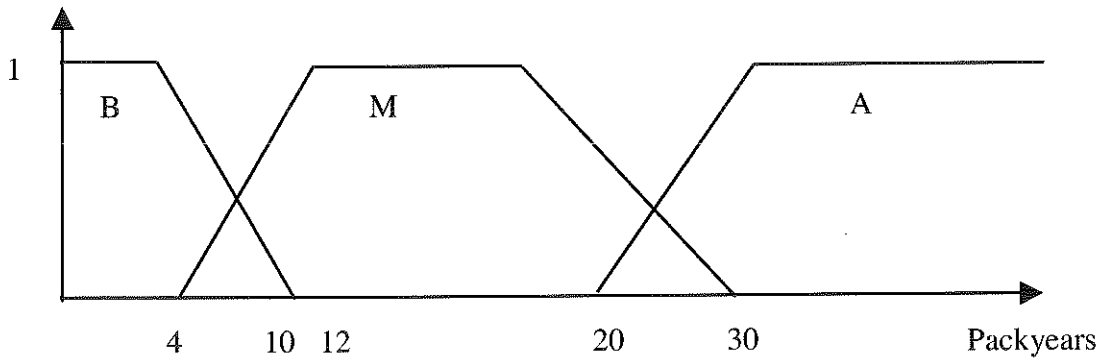
$$\frac{[(\text{n. di anni di fumo attivo}) * (\text{n. medio di sigarette fumate al giorno})]}{20}$$

Il **parametro** numerico crisp per il **fumo** assume, a seconda dei casi, i valori:

- per i soggetti fumatori il n. medio di sigarette fumate giornalmente;
- per i soggetti che hanno smesso di fumare il packyears.

Le funzioni di appartenenza per il fumo attivo sono state, determinate secondo tre insiemi fuzzy associati ai termini basso, medio, alto, come segue:





➧ Valutazione dell'esposizione giornaliera agli aero-inquinanti: si prendono in considerazione gli spostamenti quotidiani del soggetto sul territorio; si vogliono quantificare i fattori dell'ambiente outdoor che limitano l'informazione che il soggetto porta sul proprio luogo di residenza. Le parti del questionario utilizzato sono:

VALUTAZIONE ESPOSIZIONE GIORNALIERA AGLI AERO-INQUINANTI

81. In quale maniera si reca al lavoro o a scuola

1. A piedi
2. In bicicletta
3. In motociclo
4. In autovettura
5. In autobus
6. In treno
7. In taxi

82. Quanti minuti impiega per recarsi al lavoro o a scuola ⌊__⌋⌊__⌋⌊__⌋⌊__⌋	1. <15 2. 15-30 3. 30-60 4. >=60
83. Quante volte al giorno compie questo tragitto:	1. 2 2. 4 3. >4
84. Durante il percorso, quanti semafori incontra	1. Nessuno 2. Pochi 3. Molti

Il parametro crisp che rappresenta l'ingresso per le funzioni di fuzzifcazione e' calcolato in base a queste considerazioni:

1. il rischio maggiore e' attribuito al soggetto che si sposta a piedi tenendo conto del tempo impiegato. Si attribuisce un indice di pericolosità alla modalità di spostamento:
   a. a piedi 3;
   b. bici/moto 2;
   c. autobus/auto 1;

2. anche per il tempo impiegato si attribuisce un indice di pericolosità:
   a. < 15 minuti 1;
   b. 15-30 minuti 2;
   c. 30-60 minuti 3;
   d. > 60 minuti 4;

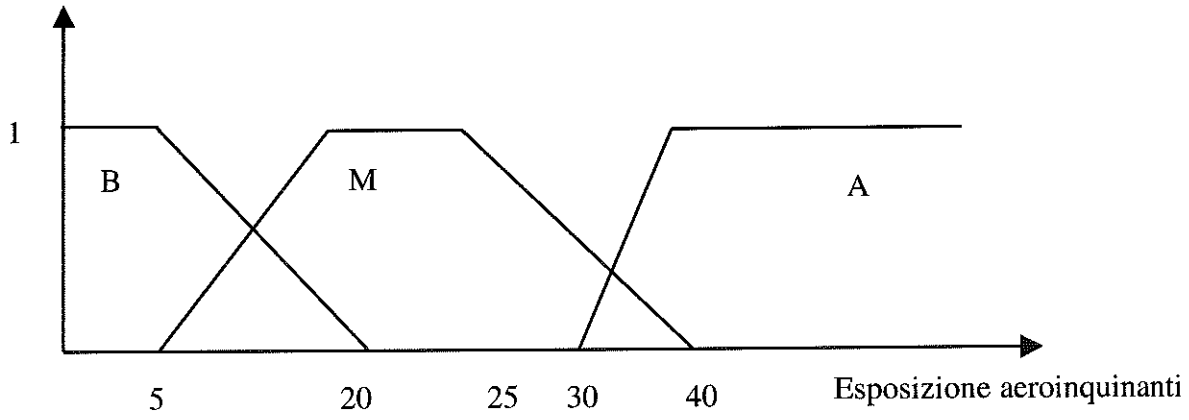
3. per il numero di volte di effettuazione del tragitto si stabiliscono i seguenti pesi:
   a. due volte 2;
   b. quattro volte 4;
   c. > quattro volte 5;

4. i semafori vengono considerati come un valore booleano (si/no): se sono presenti viene
   stabilito un parametro = 1, se sono assenti il parametro = 0.

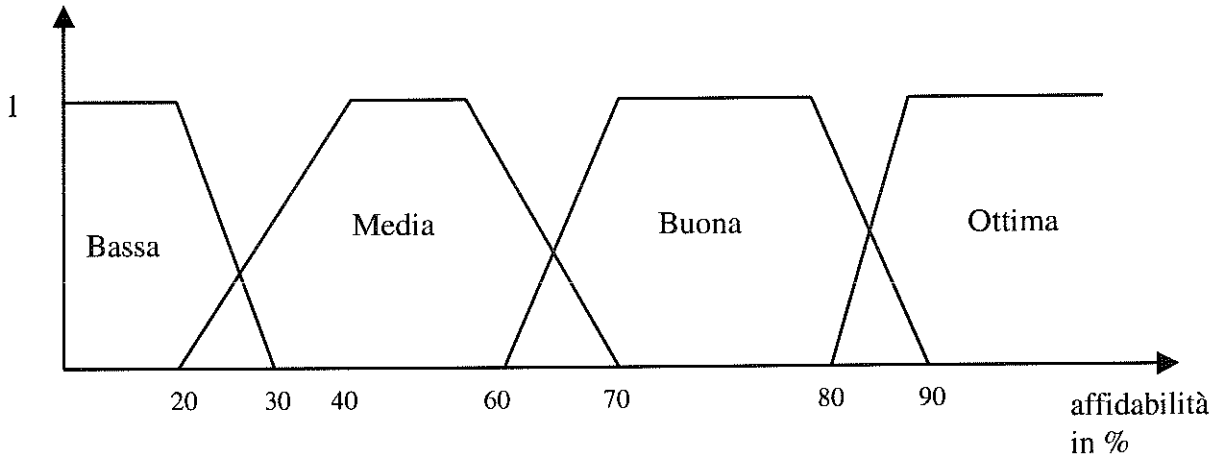
In totale il parametro aeroinquinanti assume il valore:

[(Indice_modo_spostamento * indice_tempo_impiegato) + flag_semaforo] *n. di_tragitto.

Le tre funzioni membership che quantificano l'esposizione agli aeroinquinanti, secondo i termini bassa, media, alta sono rappresentate nella figura seguente:



Le classi di output previste per l'affidabilita' dei soggetti sono quattro:
affidabilità bassa;
affidabilità media;
affidabilità buona;
affidabilità ottima.
Le quattro funzioni di appartenenza fuzzy sono così rappresentate:



Regole

I concetti linguistici possono essere legati usando degli operatori espressi in termini di gradi di appartenenza ad insiemi fuzzy in modo da formare regole del tipo If...Then; ci sono molte combinazioni possibili con differenti proprietà desiderabili. In generale gli operatori più usati sono i seguenti:

- AND: il minimo dei due gradi di appartenenza: min $(\mu_A(x), \mu_B(x))$
- OR: il massimo dei due gradi di appartenenza: max $(\mu_A(x), \mu_B(x))$
- NOT: il complemento del grado di appartenenza su [0,1]: $\mu_{\neg A}(x) = 1 - \mu_A(x)$.
-

dove A e B rappresentano gli insiemi fuzzy corrispondenti a concetti linguistici (es. esposizione fumo passivo e' *alta*, quantita di sigarette fumate bassa).
In genere la determinazione delle regole in un sistema fuzzy non e' un operazione semplice, poiché i domini essendo relativi a concetti 'fuzzy' sono espressi con espressioni linguistiche e quindi solitamente l'esperienza, il buon senso e la conoscenza dei fatti

modellati, da parte di chi predispone il sistema ([Cammarata97]), possono suggerire l'insieme delle regole per inferire i concetti espressi dal conseguente delle regole stesse.
Le regole formulate nel prototipo operano sui fattori di confondimento che introducono rumore alterando il contributo informativo geografico dei soggetti.
Gli operatori usati nel prototipo seguono lo schema matematico descritto sopra; inoltre poiché la conclusione delle regole ha insiemi output diversi e' necessaria una fase di composizione che aggreghi gli outputs di ogni regola in un singolo insieme fuzzy. Il risultato del processo della composizione e' un insieme fuzzy ottenuto attraverso l'operatore unione applicato a tutti gli insiemi fuzzy output ottenuti per inferenza da ogni regola. L'affidabilità geografica dei soggetti e' stata ottenuta in termini percentuali attraverso il processo di defuzzificazione attuato con il metodo del centroide,
Dall'analisi dei dati sono stati individuati due gruppi di fattori *di confondimento* su cui le regole sono state sviluppate:

- fattori più gravi che incidono sull'attendibilità geografica:
  A. esposizione agenti tossici
  B. esposizione fumo passivo
  C. fumo

- fattori meno gravi che incidono sull'attendibilità geografica:
  D. ambiente lavoro (presenza fumi, gas, polveri)
  E. esposizione aeroinquinanti

Questi fattori sono corrispondenti ai criteri utilizzati per la modellazione dei dati in input come descritti nel paragrafo precedente.
Nel seguito sono descritte le regole usate nel prototipo per l'estrazione del contributo geografico dai dati epidemiologici dell'indagine. Si noti come i termini linguistici siano riferiti ai corrispondenti insiemi fuzzy descritti nei paragrafi precedenti. I fattori sopra esposti rappresentano gli input crisp che sono fuzzificati e utilizzati in tal modo come antecedenti nelle regole.
Ogni soggetto *presenta* in tal modo al sistema inferenziale i propri dati e questi sono valutati passando per ogni regola presente nel sistema, fino a trovare quella che contempla il suo caso.
Va rilevato quindi che ogni soggetto viene valutato attraverso almeno una delle regole e nessuno di essi 'sfugge' all'elaborazione proprio perché le regole sono stabilite (ciò vale in generale per un sistema fuzzy qualsiasi) in modo da contemplare, in base al contesto in cui si opera, i casi possibili. Se consideriamo il numero degli input e quello degli output certamente il numero di tutte le combinazioni possibili può essere elevato; in generale non tutte queste sono necessarie per la risoluzione del problema ([Cammarata97]) ed in base al dominio del problema e all'esperienza in merito di coloro che predispongono il sistema si stabiliscono i casi da contemplare. Di sotto sono elencate e descritte le regole del nostro prototipo. Con le lettere A, B, C e D facciamo riferimento, per praticità di notazione, ai fattori precedentemente elencati:

Regola 1

If A e' alto OR
B e' alto OR
C e' alto THEN
Affidabilità BASSA

A, B, C sono i fattori assunti come i più gravi

se uno almeno tra essi e' alto l'affidabilità del soggetto e' bassa. Per la spiegazione delle regole consideriamo dapprima solo i tre fattori A,B,C.
I valori crisp del soggetto relativi ai fattori menzionati sopra sono fuzzificati e i valori ottenuti sono esaminati dalla struttura della 1ª regola: se un soggetto non rientra con i suoi valori (nessuno dei tre fattori A,B,C è alto) allora può essere valutato da una delle successive regole.

Regola 2

If [A e' medio AND B e' basso AND C e' basso)] OR
[B e' medio AND A e' basso AND C e' basso] OR
[C e' medio AND A e' basso AND B e' basso]
OR
[A è basso AND
B è basso AND
C è basso] AND
[D e' medio OR E e' medio] THEN Affidabilità BUONA

} 1ª parte della regola 2

Rispetto alla regola precedente si assume che l'affidabilità diminuisca a BUONA qualora almeno **uno solo** dei tre fattori più gravi (A, B o C) e' di livello medio e gli altri due sono bassi oppure quando A, B e C sono tutte e tre bassi e almeno uno dei fattori meno gravi (D o E) e' di livello medio.
Il soggetto se nella regola 1 non presentava valori alti allora necessariamente potrebbe presentare, nell'ambito dei fattori A,B,C uno i di livello Medio e gli altri due di livello basso. E questo caso e' contemplato appunto dalla prima parte della 2ª regola. Se nemmeno questo è il caso allora si passa alla 3ª regola.

Regola 3

If [A e' medio AND B e' medio] OR
[(B e' medio AND C e' medio] OR
[(A e' medio AND C e' medio)]
OR
[(D e' alto OR E e' alto) AND
([A è basso AND
B è basso AND
C è basso])]
THEN Affidabilità MEDIA

} 1ª parte della regola 3

Rispetto alla regola precedente si assume che l'affidabilità diminuisca a MEDIA qualora almeno **due solamente** dei tre fattori più gravi (A, B o C) sono di livello medio e l'altro rimanente e' basso oppure quando A, B e C sono tutte e tre bassi **insieme** al fatto che almeno uno dei fattori meno gravi (D o E) e' di livello alto.
Il soggetto se non rientra nella regola 1 e 2 allora può presentare almeno 2 tra i fattori più gravi pari a livello medio.
Se nemmeno questo è contemplato allora i tre fattori A,B,C sono bassi e si considera la 4ª regola.

Regola 4

IF [A e' basso AND
B e' basso AND
C e' basso AND
D e' basso
E e' basso]
THEN Affidabilità OTTIMA

Se tutti i fattori A,B,C,D,E sono di livello basso allora l'attendibilità del soggetto e' ottima perché meno influenzato dai fattori di confondimento.
Se il soggetto non rientra nei casi delle regole precedenti allora presenta valori crisp tutti e tre bassi (A,B,C) ed entrano in gioco (nella seconda parte delle regole 2, 3, 4) i due fattori meno gravi (D,E): nella 4ª regola si suppone tutte e due bassi. Se D,E non sono ambedue bassi si procede a ritroso nelle regole.
Infatti mantenendo l'ipotesi che A,B,C siano bassi tutti e tre insieme, allora se D ed E non sono bassi può succedere che uno almeno tra D ed E sia alto come dice la 3ª regola. Se ciò non e' accade, mantenendo sempre l'ipotesi che A,B,C siano bassi, può succedere che *almeno uno* tra D ed E sia di livello medio come

descritto nella seconda regola. Queste note spiegano la consistenza delle regole adottate, ma è chiaro che queste ultime possono essere ampliate e/o modificate tenendo conto di osservazioni da parte di esperti del settore che possono contribuire a raffinare il modello stesso.

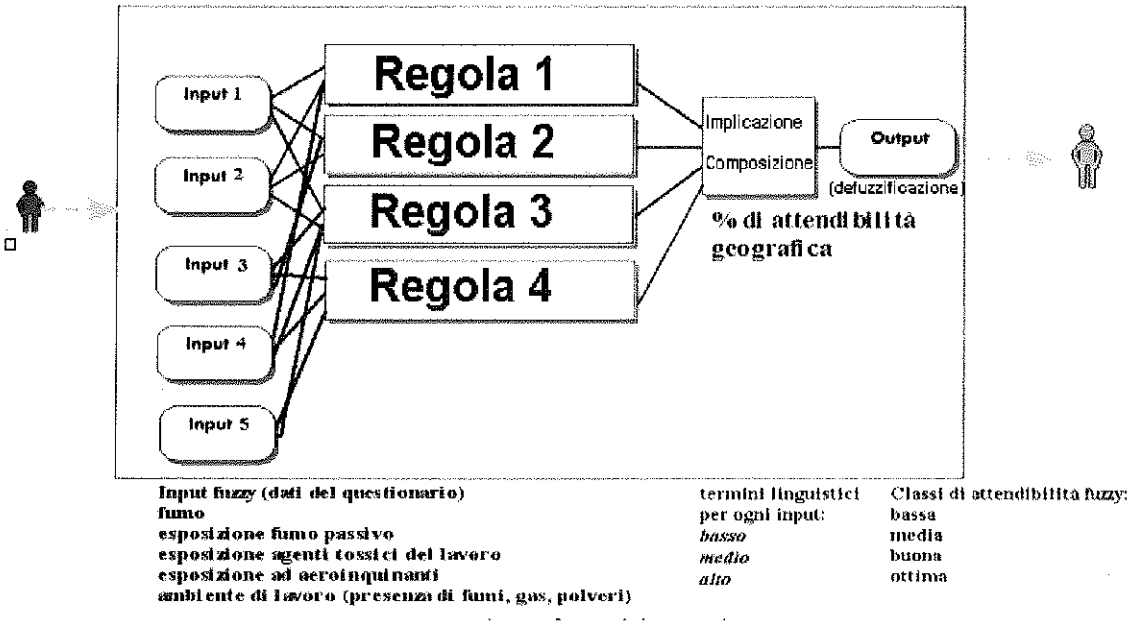


Fig. 8 Schema fuzzy del prototipo

Modellazione dei risultati: composizione e valori crisp in uscita

Dopo la valutazione dell'antecedente delle varie regole del sistema e' necessario valutare la conclusione della regola (con il procedimento di *implicazione* realizzato con il criterio del minimo, v.fig.5); per ogni regola, il grado di attivazione dell'antecedente viene confrontato (come valore dell'asse y) con la funzione di appartenenza dell'output relativa al conseguente della regola stessa, cosicché viene effettuato il *taglio* della funzione di appartenenza di output.
Poiché nel sistema (v. fig.8) ci sono quattro regole le quattro funzioni di appartenenza dell'output, tagliate con la fase di implicazione, vengono aggregate in un'unica forma mediante la fase di *composizione* (criterio dell'unione, v.fig.6). Per la determinazione dell'affidabilità geografica del soggetto, in termini percentuali, si effettua la *defuzzificazione* mediante il calcolo del centroide, ovvero si calcola una sorta di baricentro della forma geometrica della funzione composta e se ne prende la proiezione sull'asse x (v. fig.7); si stabilisce così il valore crisp (ovvero non fuzzy) in uscita che indica la percentuale di affidabilità geografica del soggetto.

Schema del prototipo

Il prototipo sviluppato e descritto in questa nota, e' essenzialmente composto dall'insieme di funzionalità offerte da due ambienti applicativi diversi; uno e' quello offerto dal linguaggio Visual Basic che consente di costruire applicazioni in ambiente Windows e l'altro dal software ArcView GIS della ESRI; l'analisi dei dati spaziali condotta soltanto con i software di tipo GIS, in alcuni casi, non fornisce tutte le risposte necessarie per la comprensione del territorio e quindi si rende necessaria l'estensione delle funzionalità presenti nei programmi GIS commerciali mediante applicazioni esterne che interagiscono con il software (in questo caso ArcView). Non essendo presenti i meccanismi della logica fuzzy nelle funzioni di ArcView e' stato necessario sviluppare un'applicazione che prevedesse le fasi tipiche di un processo di inferenza fuzzy e l'aggancio dei risultati con ArcView ([DemaRF2000]) costituisce di fatto il prototipo in totale che consente l'integrazione tra i due ambienti applicativi.

Operazioni fondamentali del sistema fuzzy nel prototipo

Per quanto riguarda l'elaborazione fuzzy, l'applicazione sviluppata esegue sui dati i seguenti passi:

- fuzzificazione degli inputs: valore crisp (non fuzzy) ⟶ valore fuzzy, tramite la definizione di insiemi fuzzy associati ai termini linguistici come definito nei paragrafi precedenti;
- inferenza fuzzy: definizione di regole e predicati tramite operatori fuzzy;
- composizione degli insiemi output fuzzy ottenuti dall'inferenza in un unico insieme fuzzy output;
- defuzzificazione: valore fuzzy ⟶ valore crisp (con il metodo del centroide)

Flusso dell'elaborazione tra l'applicativo Visual Basic e il software ArcView

Riepilogando quanto detto nei paragrafi precedenti il prototipo consente di effettuare una selezione di una base di soggetti, ovvero sono stabiliti dei criteri di esclusione adottati sulla totalità dei soggetti quali:

- soggetti con malattie pregresse
- fumatori
- ex-fumatori da meno di un certo numero di anni
- soggetti che abitano nello stesso luogo da meno di un certo numero di anni.

Questi criteri permettono di stabilire la base di soggetti su cui si applica l'elaborazione fuzzy.
Dopo la selezione della base si possono scegliere quali tra i seguenti *input* si vuole fuzzificare in base ai dati del questionario:

- fumo: n. sigarette (fumo attuale) e packyear (fumo passato)
- ambiente di lavoro (presenza di fumi, gas, polveri ed impianto di ventilazione)
- esposizione ad agenti tossici nel luogo di lavoro
- esposizione fumo passivo
- esposizione giornaliera ad agenti aeroinquinanti

Nell'applicazione sono proposte già forme delle funzioni di appartenenza fuzzy di default che però e' possibile modificare per ogni categoria succitata.
Prima di avviare l'elaborazione si specificano le caratteristiche richieste per i dati in output:

- selezione del livello di percentuale di affidabilità geografica desiderata;
- modellazione delle funzioni membership per l'elaborazione fuzzy (mediante oggetti grafici presenti nella finestra di interazione del programma con l'utente).

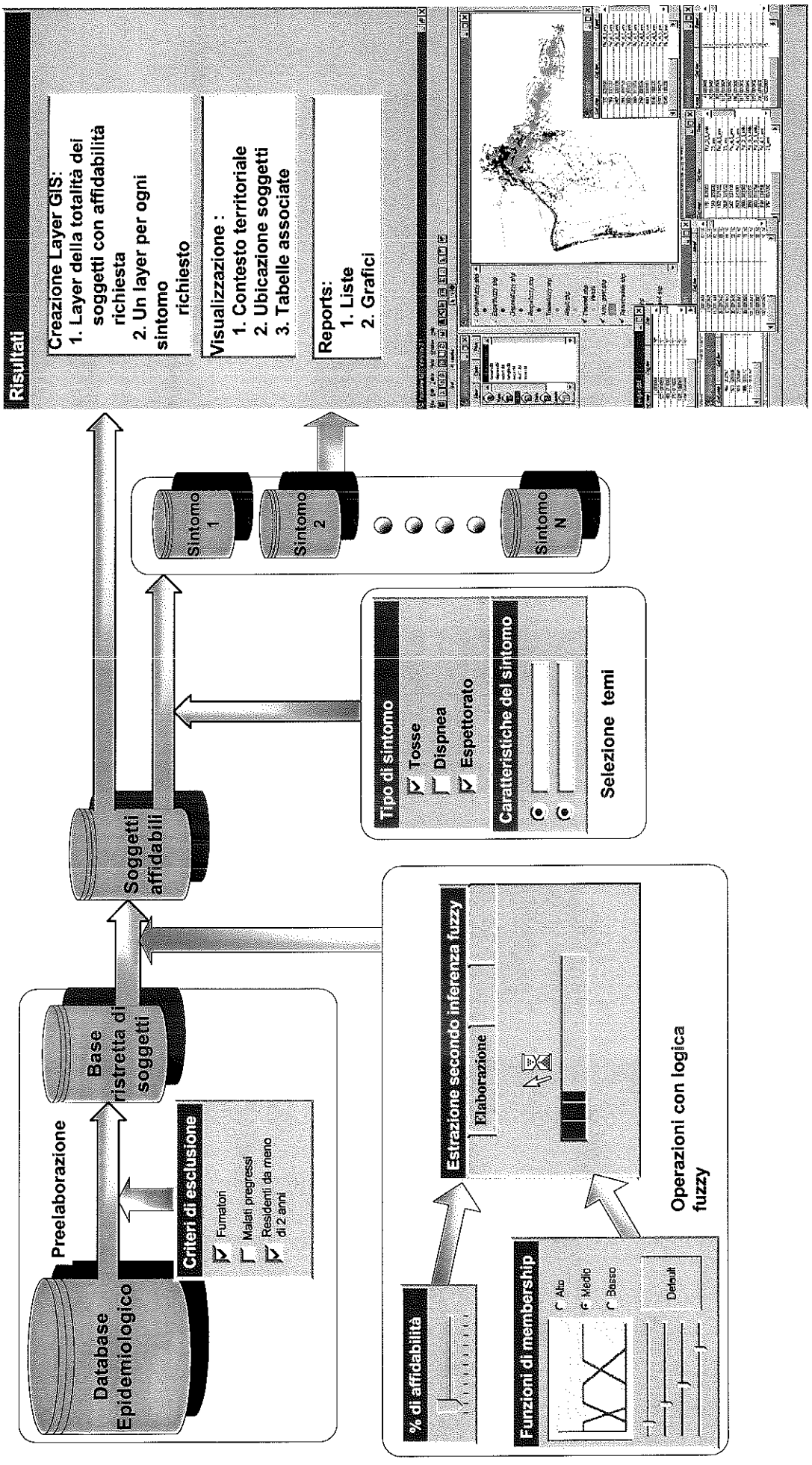
Con l'elaborazione fuzzy si determina un insieme di soggetti che rispettano le caratteristiche impostate. Su questo insieme e' possibile effettuare un'ulteriore selezione dei dati in base ad altre informazioni tra cui i sintomi rilevati con il questionario epidemiologico ovvero dalla totalità dei soggetti estratti secondo inferenza fuzzy e' possibile determinare quanti di questi presentano alcune caratteristiche rispetto ad alcuni fenomeni (per es. i sintomi), ottenendo in tal modo le tabelle relative. Con queste tabelle e' possibile rappresentare sotto forma di strati GIS la distribuzione geografica dei luoghi di residenza dei soggetti che soddisfano i criteri impostati; è possibile, per esempio, osservare sia la localizzazione geografica di tutti i soggetti valutati per inferenza fuzzy sia quelli ottenuti da quest'ultima secondo i sintomi.
E' possibile, quindi, effettuare una selezione dei layer GIS da riportare nel progetto ArcView tra i seguenti:

- tosse
- espettorato
- dispnea
- altre patologie respiratorie
- contaminazione ambientale del luogo di residenza.

Il prototipo si avvale di un progetto ArcView che :

- e' predisposto a chiamare in esecuzione il programma Visual Basic;
- riceve dall'applicazione i dati necessari alla visualizzazione di tabelle e temi inerenti ai dati epidemiologici.

Nella figura successiva viene mostrato il diagramma funzionale del prototipo in relazione alle fasi effettuate per arrivare alla produzione degli strati GIS.



Flusso logico del funzionamento del prototipo

Schema del prototipo

Dettagli funzionali del prototipo

Il progetto ArcView contiene una vista (View1) con i seguenti temi di contesto:

- tema dei venuti che visualizza sotto forma di punti le abitazioni (**indirizzi**) dei soggetti; va tenuto presente che uno stesso punto di questo tema (e di tutti gli altri temi puntuali presenti nel progetto) può rappresentare una pluralità di soggetti (soggetto singolo, soggetti che appartengono alla stessa famiglia e quindi condividono l'indirizzo di residenza, oppure soggetti che appartengono a famiglie diverse ma che risiedono nello stesso luogo).
  Questo tema rappresenta il risultato della geocodifica ovvero la localizzazione su una mappa dell'indirizzo di residenza dei soggetti partecipanti all'indagine;
- tema delle mezzerie stradali per Pisa e della rete stradale per Cascina;
- tema poligonale del fiume Arno (parte di Pisa);
- tema degli edifici di Pisa.

Dai dati epidemiologici sono state estratte tutte le famiglie partecipanti classificate secondo un codice (cod_famigl) utilizzabile come aggancio al tema puntuale degli indirizzi. Nel progetto viene visualizzata la suddetta tabella ed utilizzata per mettere in relazione i soggetti, caratterizzati dal codice della famiglia di appartenenza, al tema degli indirizzi; in tal modo viene stabilito un riferimento indiretto alle coordinate geografiche dell'indirizzo del soggetto.

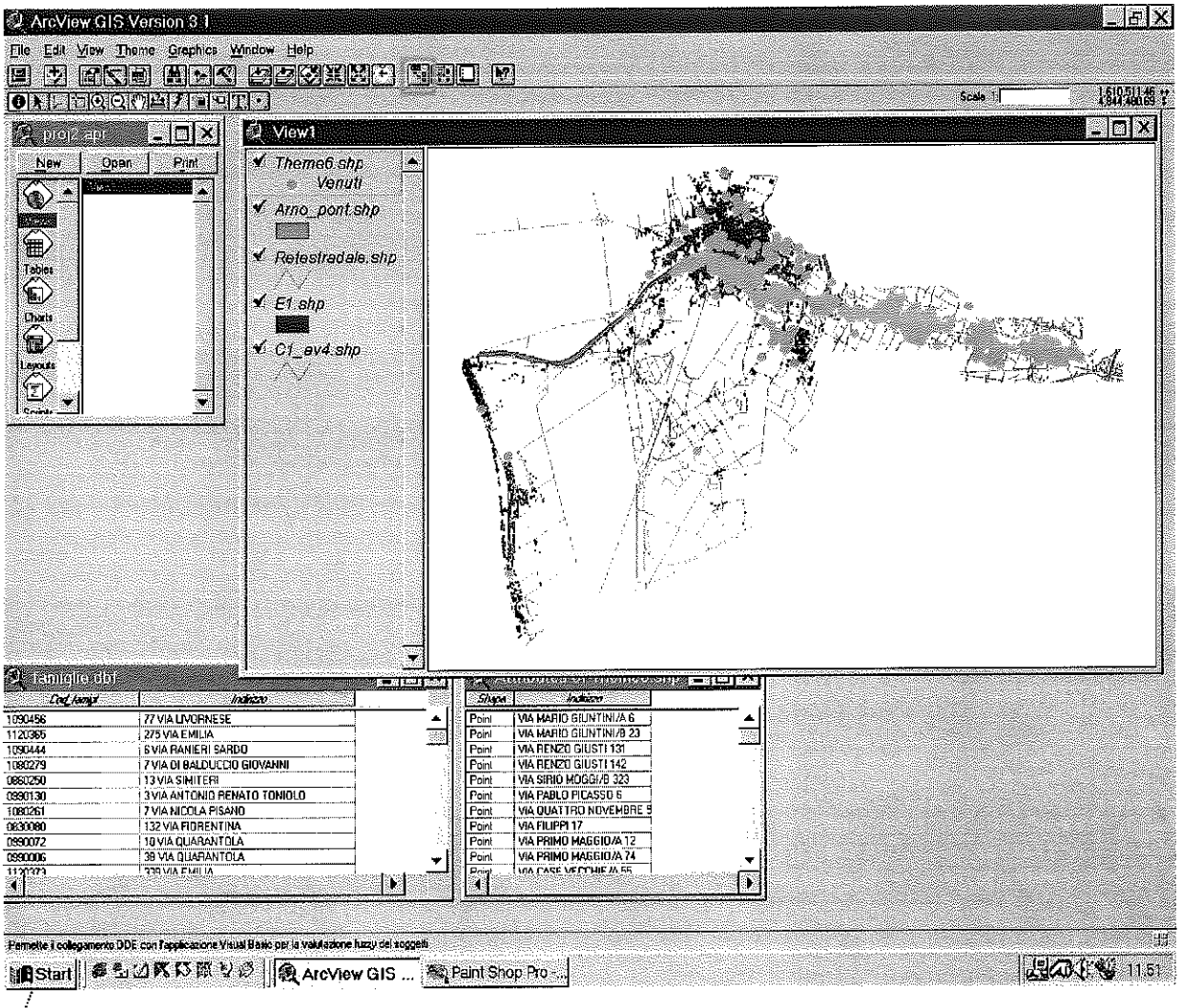


Fig.9 Il bottone evidenziato nel quadrato arancione e' il punto di esecuzione dell'applicazione VB

La pressione sul pulsante indicato in Fig.9 attiva l'esecuzione di uno script Avenue che avvia l'applicazione di Analisi Fuzzy.
Viene proposto un menu' organizzato con controlli di tipo Tab:

- la prima scheda Base consente all'utente di stabilire i criteri di esclusione dei soggetti dall'analisi;
- la seconda scheda Parametri consente la possibilità di definire le funzioni di appartenenza fuzzy per i dati numerici *crisp* di ingresso, ottenuti come specificato precedentemente;
- la terza scheda Output dà la possibilità di definire quali temi devono essere generati in ArcView, di stabilire la percentuale di attendibilità dei soggetti desiderata e di avviare l'elaborazione fuzzy dei dati in accordo ai parametri definiti nella scheda precedente;
- la quarta scheda visualizza le scelte effettuate nella scheda Base, i parametri impostati per l'analisi fuzzy e le tematiche dei layers che verranno creati nel software GIS;
- la quinta scheda visualizza fornisce i risultati quantitativi dell'operazione, indicando il numero di soggetti con il grado di affidabilità richiesta per ogni layer considerato.

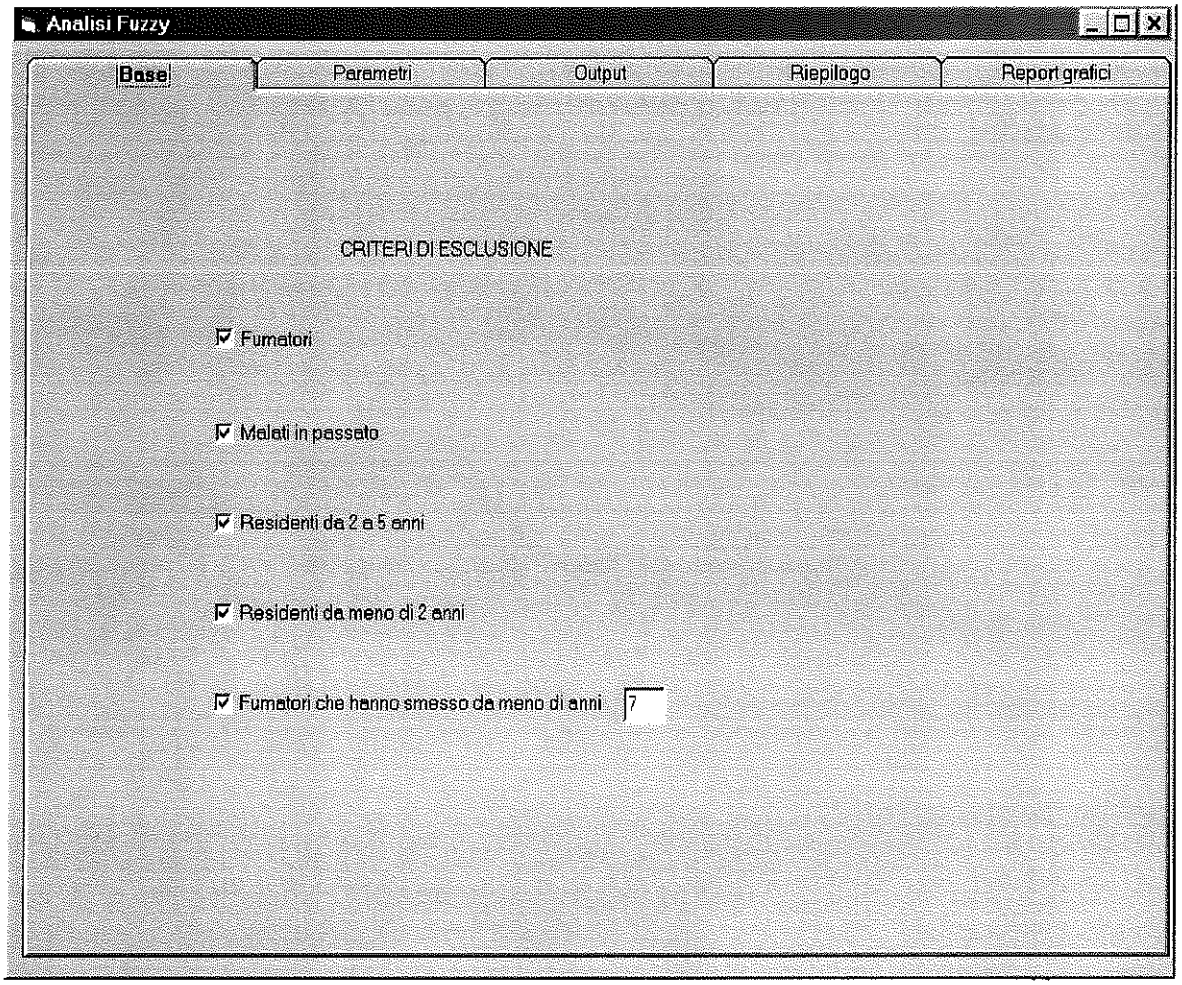


Fig.10 La *base* per escludere i soggetti: le categorie di soggetti non selezionati tramite questa scheda vengono inclusi (per esempio i soggetti residenti da piu' di 5 anni sono sempre considerati nell'indagine proprio perche' rappresentano il nucleo di informazione radicata al territorio più antica). Un soggetto viene escluso se rientra in almeno una delle categorie che vengono attivate in questa scheda.

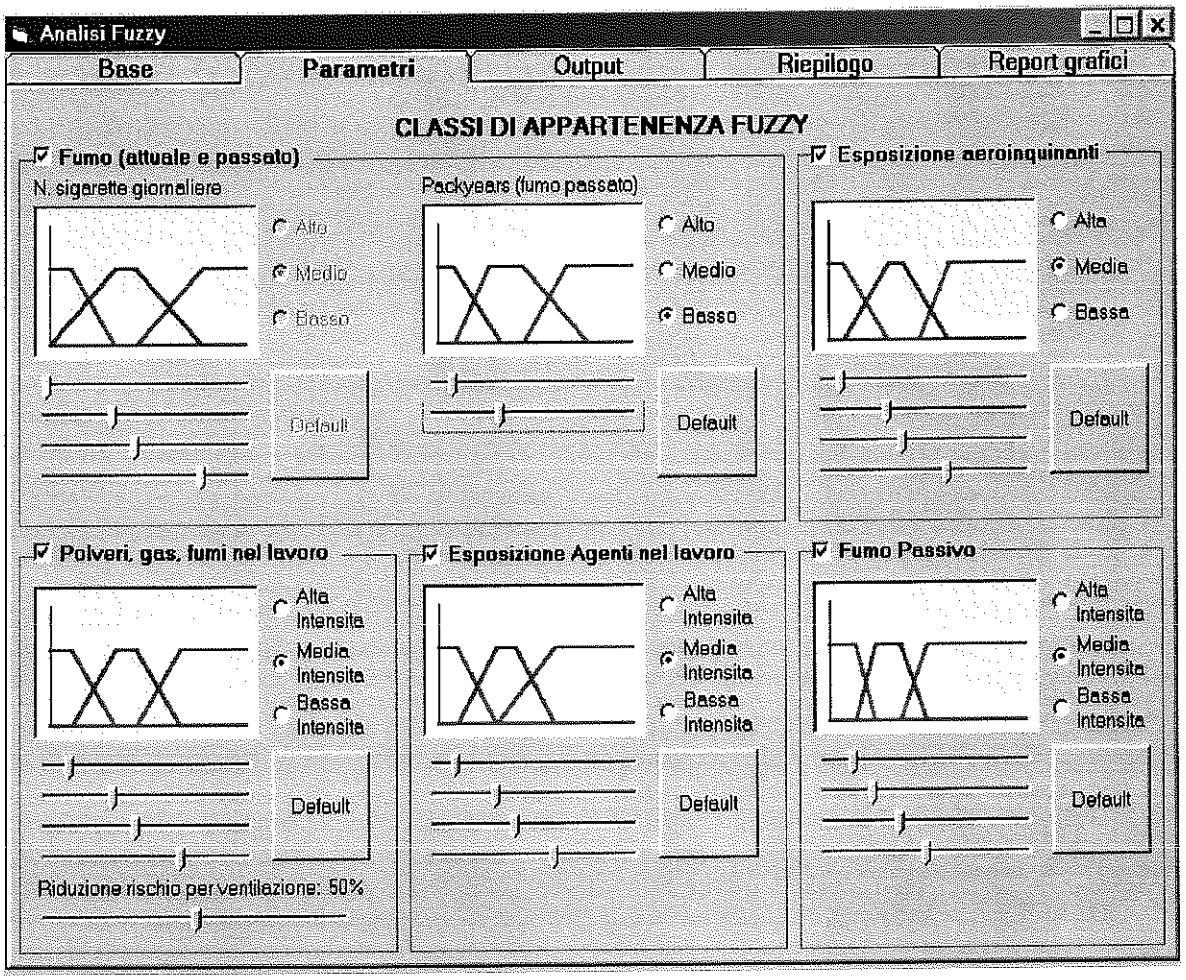


Fig.11 La scheda Parametri: In questa scheda viene presentata inizialmente una funzione di default (corrispondente a quelle presentate nel paragrafo per i tre insiemi fuzzy. La possibilità di variare la 'forma' ovvero i valori che definiscono le tre funzioni per ogni variabile e' data dall'utilizzo con il mouse degli slider presenti sotto ogni grafico. Si seleziona quale termine (basso, medio, alto) si vuol modificare e si interagisce con gli slider sottostanti. E' presente, per la variabile linguistica ambiente di lavoro, la possibilità di variare anche il fattore di riduzione rischio dovuto alla presenza di impianto di ventilazione. Le checkbox delle variabili da fuzzificare permettono all'utente di scegliere nella propria analisi quali dati fuzzificare.

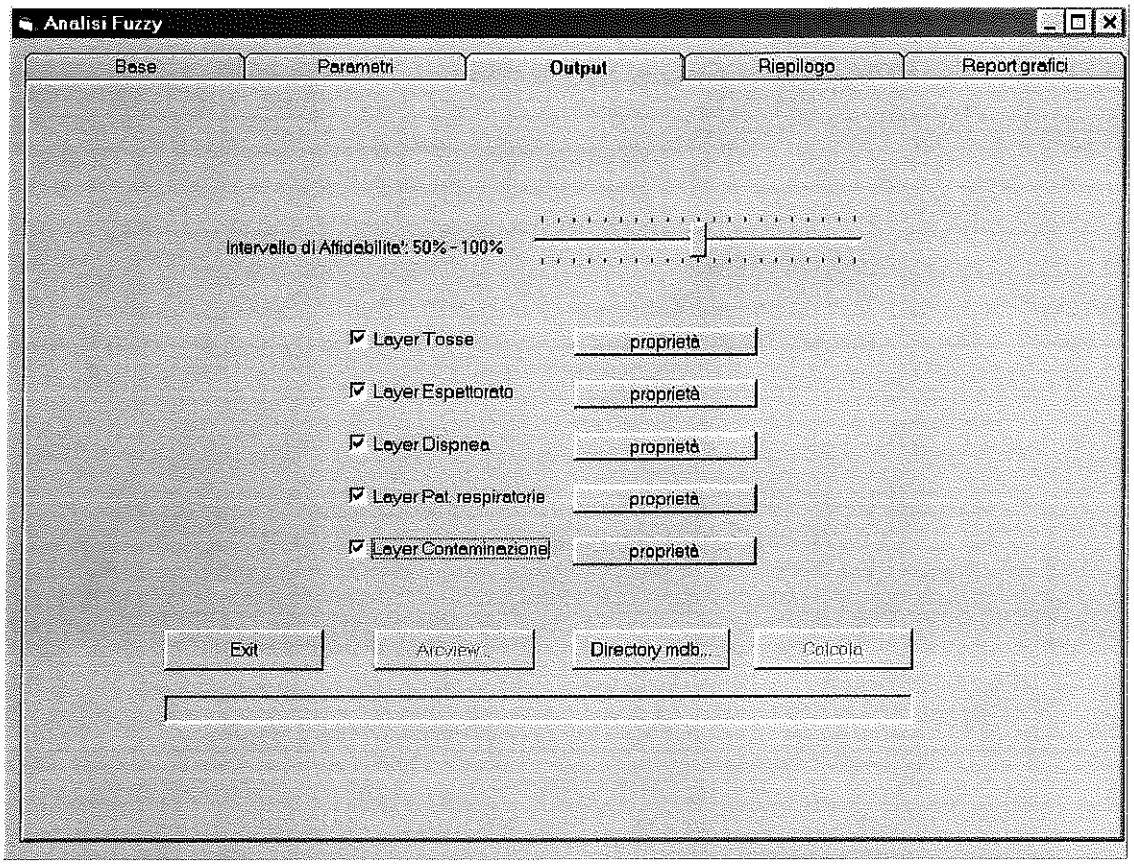


Fig.12 La scheda Output: in questa sezione si specifica in base a quali caratteristiche devono essere selezionati i soggetti da estrarre dal database: per ogni sintomo richiesto devono essere specificati i criteri di estrazione (pulsante proprietà); si può agire su uno slider per stabilire quale percentuale di affidabilità geografica dovranno avere i soggetti selezionati; il sistema provvede a creare un layer di punti (corrispondenti agli indirizzi dei soggetti) per ogni sintomo selezionato. La scelta dei criteri per i sintomi non influenza il processo elaborativo fuzzy e stabilisce unicamente il filtro finale da applicare ai soggetti.

I bottoni nella parte bassa della figura attivano le seguenti funzioni:

- Calcola: attiva le procedure di fuzzificazione secondo i parametri stabiliti e la creazione delle tabelle dei risultati; inizialmente e' disattivato;
- Directory mdb: specifica il path per individuare il database MS Access; una volta selezionato l'archivio dei dati si attiva il pulsante Calcola;
- ArcView: riporta i risultati come temi (e corripondenti tabelle) nel progetto ArcView 3.1;
- Exit: chiude il collegamento DDE [DemaRF2000] con l'applicativo GIS e termina il programma.

La scheda riepilogo fornisce il quadro sinottico delle scelte effettuate nelle sezioni precedenti del programma. Può essere visualizzato prima di avviare il calcolo per controllare le impostazioni ed eventualmente modificarle.
E' presente un pulsante per mandare in stampa questo riepilogo di informazioni.

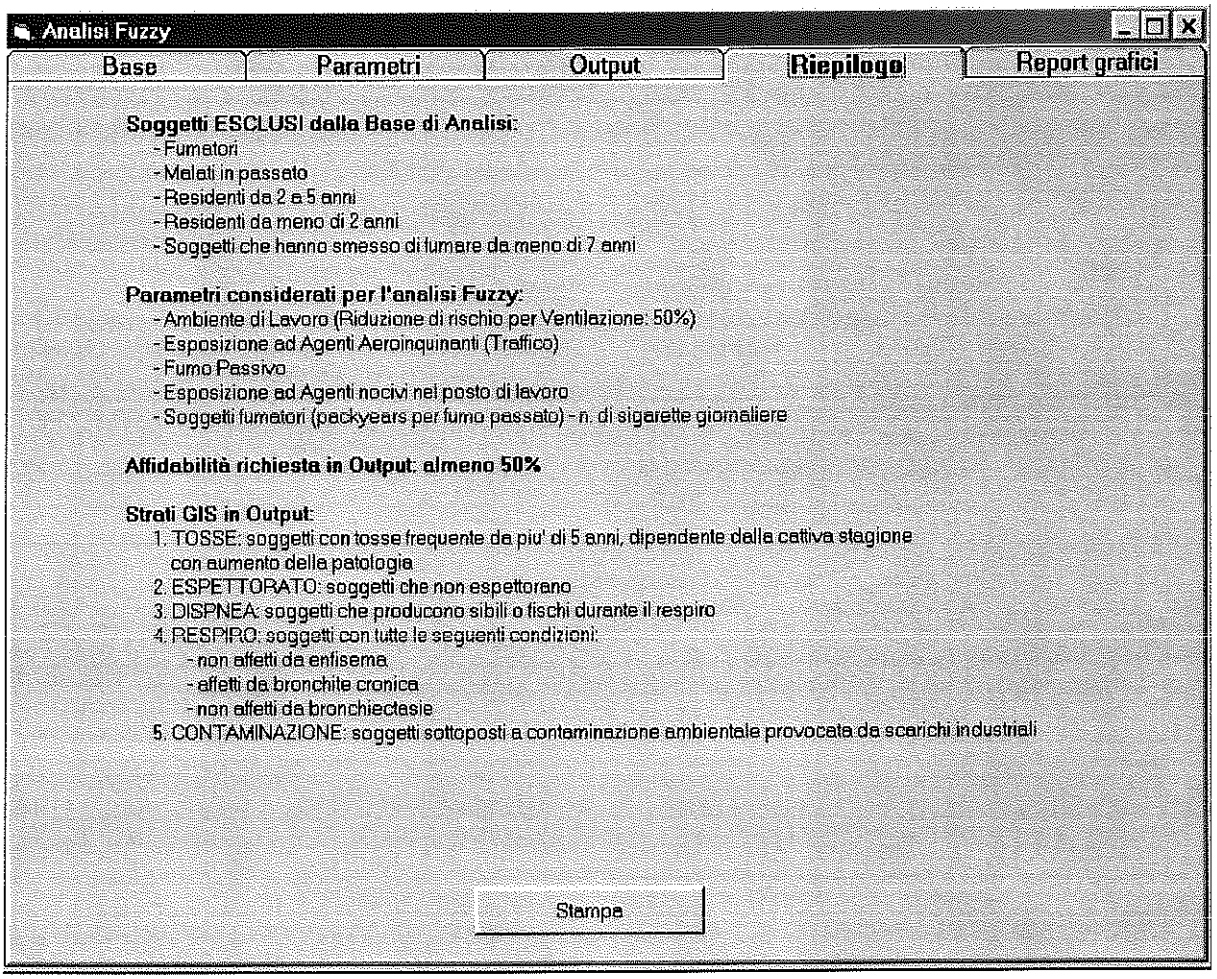
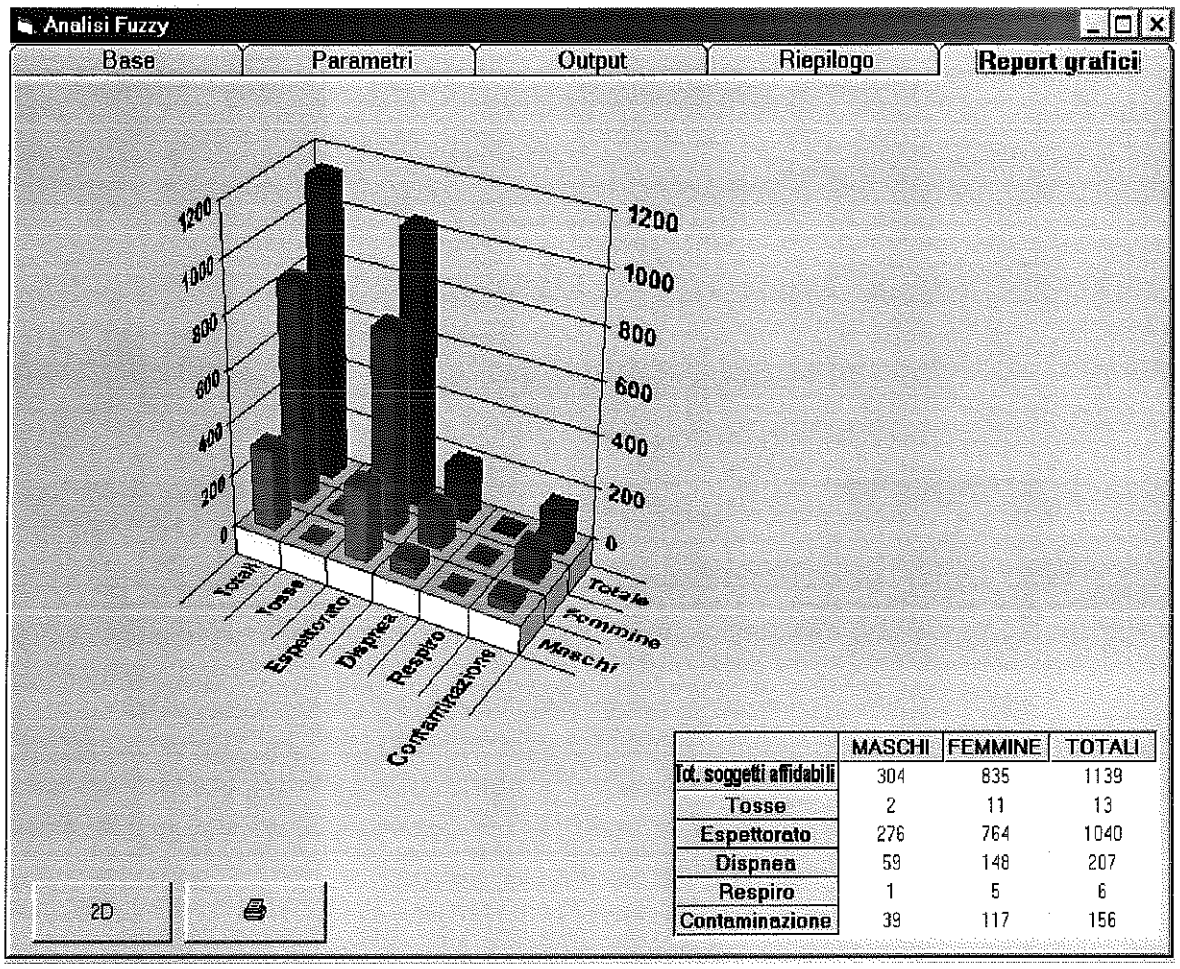


Fig.13 Il riepilogo delle scelte

Nella sezione Report grafici del programma e' possibile dopo che e' stata svolta l'elaborazione vedere quanti soggetti sono stati selezionati dai dati epidemiologici, in relazione all'affidabilità richiesta e ai layer selezionati. La barra blu evidenzia il totale tra maschi e femmine. E' possibile switchare tra i due tipi di grafici e di mandare in stampa quello attualmente visibile sullo schermo. La tabella riporta il numero esatto dei soggetti selezionati per ogni layer.



	MASCHI	FEMMINE	TOTALI
Tot. soggetti affidabili	304	835	1139
Tosse	2	11	13
Espettorato	276	764	1040
Dispnea	59	148	207
Respiro	1	5	6
Contaminazione	39	117	156

Fig.14 Le informazioni fornite in questa sezione sono visibili per una prima valutazione quantitativa dei risultati che si ottengono con i parametri impostati

Cliccando sul pulsante Arcview nella sezione Output si avvia il procedimento di creazione dei temi nel sistema GIS e si passa il controllo ad ArcView; osserviamo che dopo c'è la possibilità di ritornare all'applicazione di Analisi Fuzzy mantenendo gli stessi parametri impostati precedentemente, cosicché si possono modificare alcune impostazioni per poter effettuare i confronti con l'elaborazione precedente.

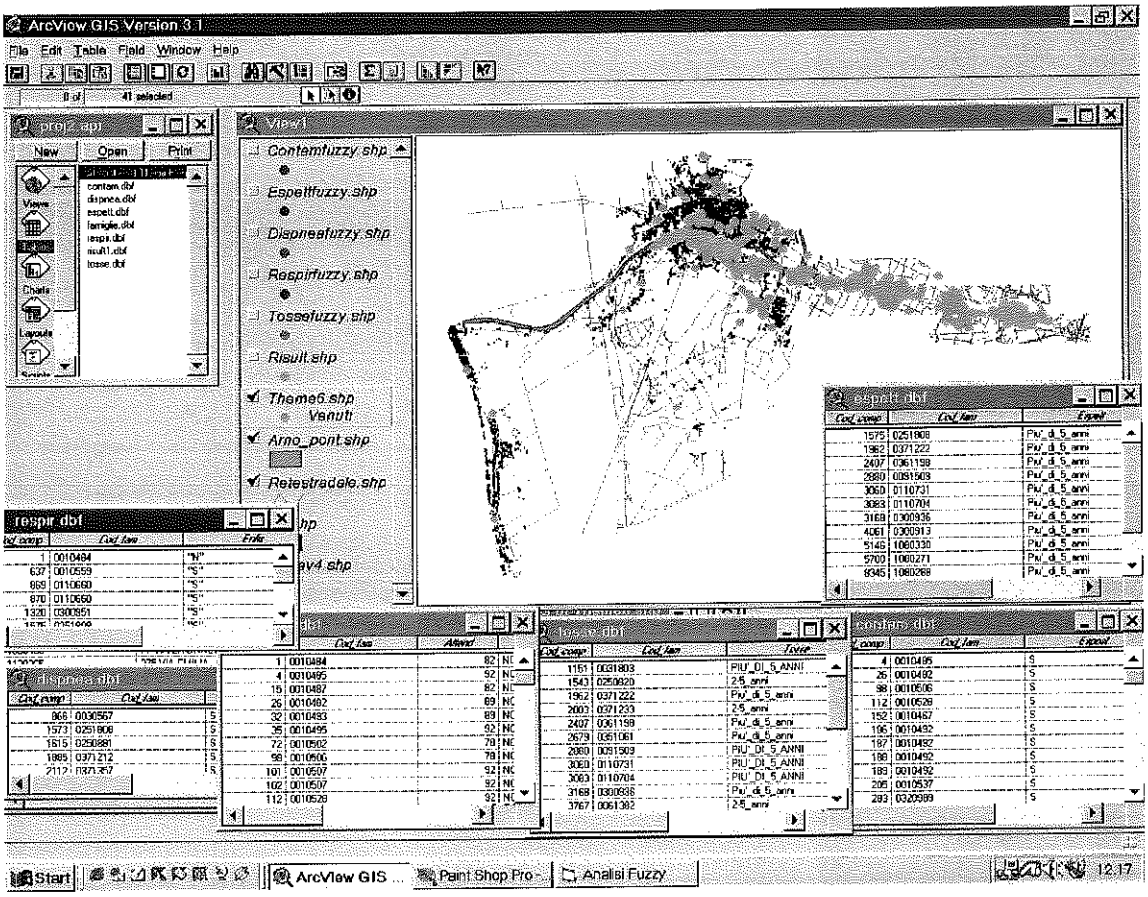


Fig. 15 Visualizzazione dei temi in Arcview.

Nel progetto ArcView vengono generati i seguenti temi:

- Risult.shp: il tema dei soggetti fuzzificati secondo i valori indicati nella sezione parametri specificato nella sezione Output (v. fig.12); vengono visualizzati i punti coincidenti con gli indirizzi delle famiglie di appartenenza dei soggetti;
- Tossefuzzy.shp, respirfuzzy.shp, dispneafuzzy.shp, espettfuzzy.shp, contamfuzzy.shp sono temi puntuali relativi agli indirizzi dei soggetti che rientrano nella percentuale di affidabilità geografica specificata e che hanno dato come risposte nelle parti del questionario (tosse, dispnea, espettorato ecc.) quelle stabilite nelle proprietà dei layer della sez. Output; ognuno di questi è un sottoinsieme di Risult.shp.

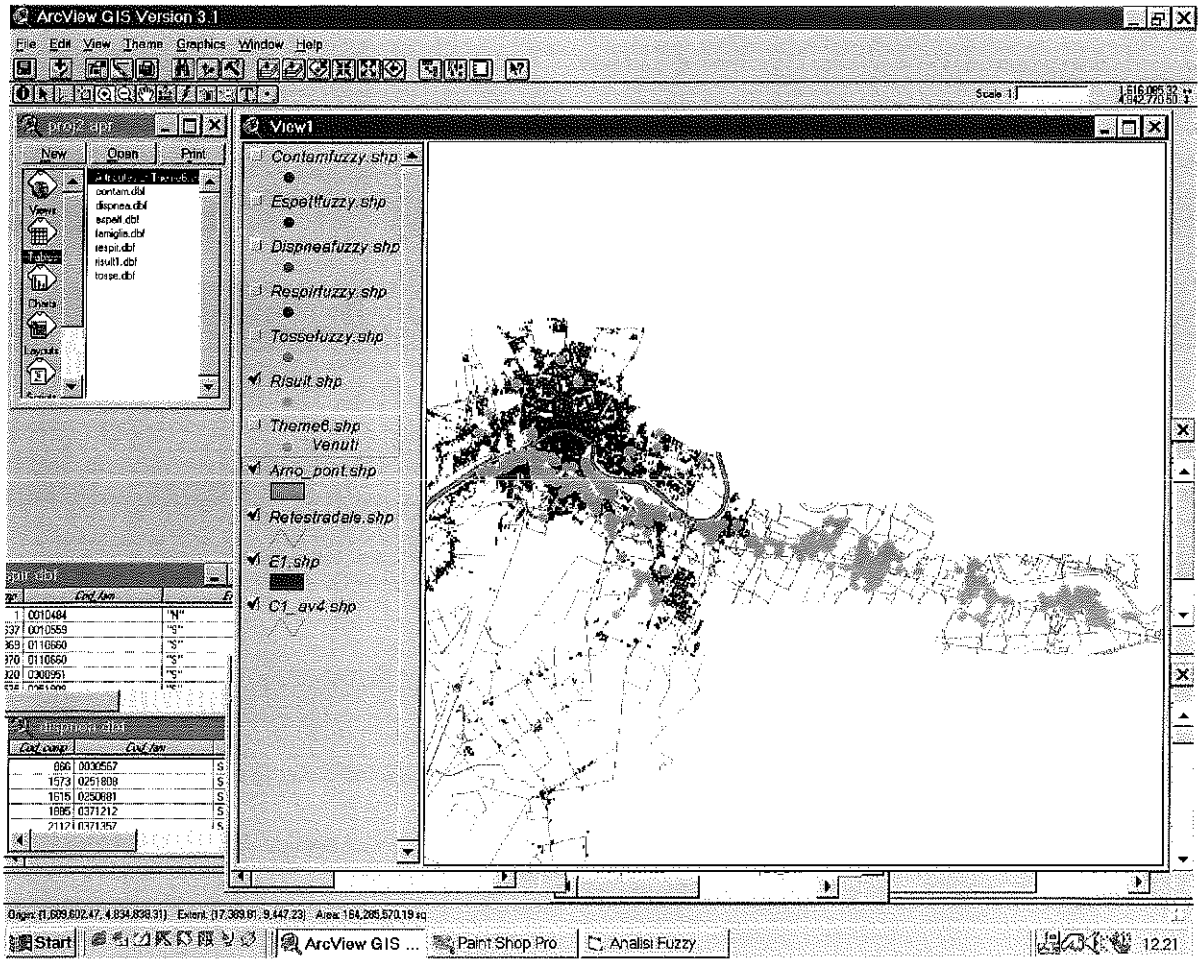


Fig. 16: La visualizzazione del tema Risult.shp evidenziato con i punti arancione.

Bibliografia

[DemaRF2000]: R. della Maggiore, R. Fresco: Integrazione di ArcView GIS (ESRI) in ambiente applicativo Microsoft Windows NT , Nota interna CNUCE B4 -2000 - 018.

[De Bruin2000]: S. De Bruin – Querying probabilistic land cover data using fuzzy set theory, I.J.G.I.S. Volume 14 N.4 ISSN 1365-8816 June 2000.

[Fresco2000]: R. Fresco: I sistemi informativi geografici e lo studio di fenomeni epidemiologici – tesi di laurea, Università degli studi di Pisa, Facoltà di Scienze M.F.N., a.a 1998-1999.

[GatrellLöy98]: Gatrell, Löytönen et al.: GISRUK 98 – Peaks and Troughs: Space-Time Cluster detection in rare diseases <http://www.lancs.ac.uk/postgrad/sabel/gisruk98.html>

[Cammarata97] : S.Cammarata: I sistemi a logica fuzzy: come rendere intelligenti le macchine, Etas Libri RCS, Ottobre 1997;

[Rossiter95]: David G. Rossiter: Uncertainty – Part 5 in Land Evaluation – Cornell University Notes – 1995 <http://www.scas.cornell.edu/landeval/le_notes/s494ch5p.htm>