UNIVERSITÀ DEGLI STUDI DI NAPOLI

FEDERICO II

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELLE TECNOLOGIE DELL'INFORMAZIONE

Corso di Laurea Magistrale in Informatica

Metodi di apprendimento non supervisionato per l'analisi di mappe di estinzione in astrofisica

Relatori

Prof. Francesco Isgrò

Dr. Massimo Brescia

Correlatore

Prof.ssa Paola Festa

Candidato

Francesco Esposito

N97000183

ANNO ACCADEMICO 2014/2015

Ringraziamenti

Desidero ringraziare tutte le persone che sono state importanti in ogni tappa di questo lungo viaggio che va a concludersi.

Ringrazio il Prof. Isgrò e la Prof.ssa Festa per il prezioso contributo e la disponibilità mostrata in questi mesi.

Ringrazio le splendide persone conosciute all'Osservatorio Astronomico che mi hanno accompagnato, con la loro professionalità ed allegria, in questo percorso: Massimo Brescia, la cui stima mi ha permesso, in questi anni, di crescere dal punto di vista umano e professionale, nonché Stefano e Giuseppe, sempre disponibili ad aiutarmi. Ringrazio tutti loro per avermi mostrato come svolgere con passione questo lavoro e aver reso questo periodo bello ed indelebile nella mia memoria.

Un ringraziamento speciale inoltre va al Dr. Alcalà per il fondamentale aiuto offerto, senza il quale questo lavoro non sarebbe stato possibile.

Ringrazio infine la mia famiglia per avermi supportato in questi anni rendendo possibile il conseguimento dei miei obiettivi.

1 Introduzione

L'image processing è un insieme di tecniche informatiche particolarmente importante in settori multi-disciplinari in cui la speculazione scientifica sia basata sull'analisi di immagini ad altissima risoluzione. Inoltre, riveste un ruolo particolarmente cruciale laddove non esista una verità assoluta cui far riferimento, ma che sia solo frutto di ipotesi e congetture scientifiche basate su modelli teorici e/o simulazioni. In quest'ambito rientra perfettamente l'insieme delle problematiche di tipo astrofisico. Come noto, infatti, in Astronomia le osservazioni producono immagini e/o spettri in cui risulta estremamente complesso identificare i vari tipi di oggetti con una sufficiente affidabilità e precisione, a causa del rapporto segnale/rumore particolarmente limitato. Ad aggravare ulteriormente il problema, contribuisce il fatto che le immagini astronomiche forniscono naturalmente una proiezione bidimensionale delle regioni di cielo, confondendone la prospettiva rispetto alla profondità, o distanza interstellare tra gli oggetti rivelati.

In ambito astronomico una particolare importanza riveste il concetto di estinzione galattica, una sorta di "effetto cataratta" naturale, forzatamente sovrapposta alla strumentazione osservativa, che diminuisce fortemente la nitidezza delle immagini osservate. Le cause sono molteplici, dipendenti sia da fattori locali (presenza dell'atmosfera per le osservazioni da terra), sia da fattori esterni (effetti di foreground/background noise tra l'osservatore e la sorgente, dovuti alla presenza caotica di nubi sovrapposte di gas e polveri), particolarmente rilevanti nelle regioni di formazione stellare.

Le tecniche tradizionalmente impiegate per la rivelazione dell'estinzione si basano su conteggi statistici delle sorgenti brillanti presenti nelle regioni osservate e relative sogliature arbitrarie dei livelli di flusso registrato nei pixel delle immagini.

Il principale scopo di questo lavoro consiste dunque nell'esplorazione originale di tecniche auto-adattive (machine learning non supervisionato) applicate all'image processing, con lo scopo di ottenere risultati scientifici quantomeno paragonabili alle tecniche tradizionali nella rivelazione dell'estinzione interstellare in immagini astronomiche. Il valore aggiunto deriva quindi dalla maggiore precisione nel circoscrivere regioni a densità di estinzione variabile e dall'eliminazione di qualunque meccanismo arbitrario di scelta di soglie di flusso e di conteggio delle regioni di estinzione.

D'altronde, nei test eseguiti, l'evidente complessità delle immagini astronomiche ha dimostrato come metodi considerati tradizionali nel campo dell'image processing falliscano l'obiettivo. Tecniche basate su filtraggio del gradiente e maschere di convoluzione, come Canny, Sobel, Prewitt e Roberts, hanno infatti mostrato l'estrema difficoltà nell'individuare correttamente gli edge e nella ricostruzione della distribuzione di densità di estinzione. Anche successivi test effettuati su tecniche basate su logica fuzzy, hanno evidenziato difficoltà oggettive nel superare il problema. La logica conseguenza è stata quindi esplorare soluzioni basate sul paradigma non supervisionato del Machine Learning. In particolare, il clustering multi-stadio, ossia basato sull'applicazione a cascata di algoritmi di clustering, ha permesso di ottenere i migliori risultati, circoscrivendo al meglio lo spazio dei parametri specifico per il tipo di dati. Il miglior modello è quindi risultato la rete

SOM (Self Organizing Map) + K-Means, mappando le immagini in uno spazio composto da elementi statistici (gradiente, entropia e deviazione standard) e informazioni derivanti direttamente dai dati stessi (valori dei pixel relativi al flusso).

2 Ricostruzione di mappe di estinzione in Astrofisica

In Astronomia, si definisce "estinzione" il fenomeno di assorbimento e deflessione caotica (*scattering*) di radiazione elettromagnetica nei gas e nelle polveri cosmiche interposte tra una sorgente emittente e l'osservatore. In generale, per le stelle che giacciono in prossimità del piano della Via Lattea, ad una distanza di alcune migliaia di *parsec* (pc) dalla Terra, l'estinzione nella banda del visibile è dell'ordine di 1.8 magnitudini per *kiloparsec* (Whittet 2003).

Tale fenomeno è dovuto sia alla presenza dell'atmosfera terrestre, sia al cosiddetto mezzo interstellare (ISM), che si interpongono tra l'osservatore e l'oggetto osservato. Il primo contributo può essere evitato effettuando costose osservazioni dallo Spazio, mentre il secondo è purtroppo un elemento non immediatamente eliminabile perché intrinseco del mezzo interstellare (ISM), il cui effetto è l'introduzione di un livello estremamente variabile di attenuazione della luce (*foreground*/*background noise*) con effetto moltiplicativo rispetto al rapporto segnale-rumore delle immagini astronomiche.

In particolare, l'estinzione non è facilmente quantificabile in quanto le immagini astronomiche sono bi-dimensionali (mancano cioè dell'informazione in profondità), per cui la quantità effettiva di attenuazione della luce dovuta all'estinzione non ha a priori caratteristiche immediatamente distinguibili tra fotoni provenienti dall'ISM

diffuso nella regione d'interesse e quello presente in regioni remote poste davanti o dietro la zona osservata.

Inoltre, essendo noto che la luce blu sia molto più attenuata di quella rossa, l'estinzione causa il cosiddetto *arrossamento* delle sorgenti osservate. Quest'ultimo fenomeno è dovuto alla polvere interstellare che assorbe e riflette le onde di luce blu molto più di quelle di luce rossa, facendo apparire più rossi gli oggetti. Un effetto quindi del tutto analogo al cielo arrossato al tramonto visibile sulla Terra.

La quantità di estinzione è, come anticipato, estremamente variabile, dipendendo dalla natura delle zone di formazione stellare. Ad esempio alcune regioni del centro galattico presentano anche 30 magnitudini di estinzione nella banda ottica, che corrisponde in pratica al fatto che circa 1 fotone può arrivare all'occhio dell'osservatore per ogni 10^{12} fotoni emessi dalla sorgente.

All'interno delle immagini astronomiche la relazione tra l'estinzione totale A_V, misurata in magnitudini, e la cosiddetta *column density* (colonna di densità) degli atomi di idrogeno N_H, misurata in cm^{-2}, mostra l'interazione tra il gas e le polveri nel mezzo interstellare. Da studi di regioni stellari effettuati con la spettroscopia ultravioletta e ai raggi X, risulta nota la relazione $\frac{N_H}{A_V}$, per la quale abbiamo usato due varianti nel calcolo delle mappe di estinzione utilizzate in questo lavoro:

$$\begin{cases} \left[\frac{N_H}{A_V}\right]_A = 1.37\times10^{21} \\ \left[\frac{N_H}{A_V}\right]_C = 1.87\times10^{21} \end{cases} \quad [cm^{-2}mag^{-1}] \qquad (1)$$

L’equazione (1) mostra due varianti del rapporto di polvere/gas intergalattico, come indicate, rispettivamente da Alcalà 2016 ($\left[\frac{N_H}{A_V}\right]_A$) e da Cambrèsy 1999 ($\left[\frac{N_H}{A_V}\right]_C$).

Tradizionalmente si suole misurare la curva di estinzione di una sorgente (di solito una stella) attraverso il confronto tra il suo spettro e quello osservato per una stella simile di cui sia noto il flusso senza estinzione. È anche possibile utilizzare un modello teorico di spettro per una regione con una certa densità molecolare. Attraverso questi modelli e i dati osservativi, è possibile ottenere una buona approssimazione della quantità media di peso molecolare dei componenti che risiedono nel mezzo interstellare (principalmente idrogeno, elio e residui di trizio) e quindi tenerne conto nella misurazione delle mappe d’estinzione di regioni galattiche. In particolare, nel presente lavoro abbiamo tenuto conto dei seguenti parametri:

$$\begin{cases} [\mu]_A = 2.02 \\ [\mu]_C = 2.4 \end{cases} \tag{2}$$

L’equazione (2) mostra due varianti del termine peso molecolare medio corretto, rispettivamente, per abbondanza d’idrogeno molecolare e trizio ($[\mu]_A$, Alcalà 2016) e per abbondanza di elio ($[\mu]_C$, Cambrèsy 1999);

$$\frac{m_H}{M_\odot} = \frac{1.66}{1.98} \times 10^{-57} [M_\odot] \tag{3}$$

l’equazione (3) mostra il fattore di correzione dovuto al peso molecolare dell’idrogeno, espresso in masse solari (Cambrèsy 1999).

Quello di cui ci siamo occupati in questo lavoro riguarda primariamente la rivelazione e misurazione di vaste mappe d'estinzione di una delle più grandi nebulose molecolari giganti vicine, denominata regione *Lupus*.

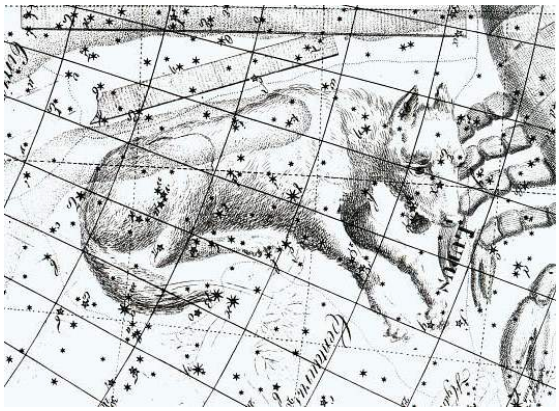


Figura 1 - Raffigurazione mitologica della costellazione Lupus (fonte: Bode, Uranographia)

Le nebulose di *Lupus* compongono una delle principali regioni di formazione stellare (*low-mass star forming region*) in prossimità del Sole (circa 200 pc), facente parte dell'associazione *Scorpius-Centaurus* (Figura 2) e comprendente circa un centinaio di stelle, fra le quali molte doppie e blu.

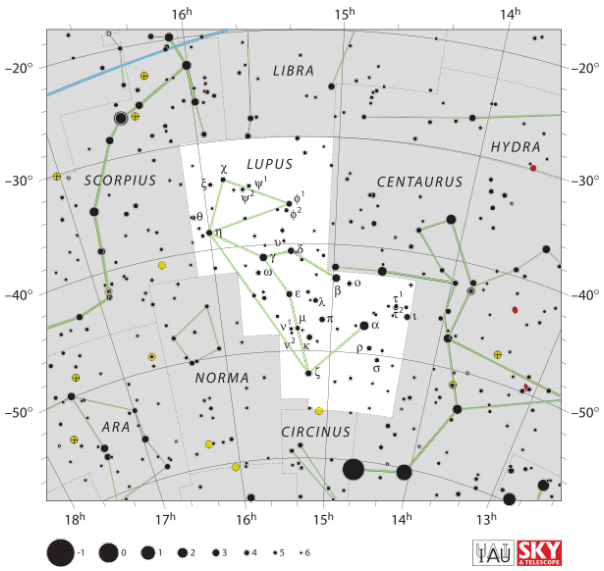


Figura 2 - Localizzazione di Lupus (Sinnott and Fienberg, IAU and Sky & Telescope magazine)

Queste nebulose sono suddivise per convenzione in 6 principali zone di formazione stellare, come mostrato in Figura 3.

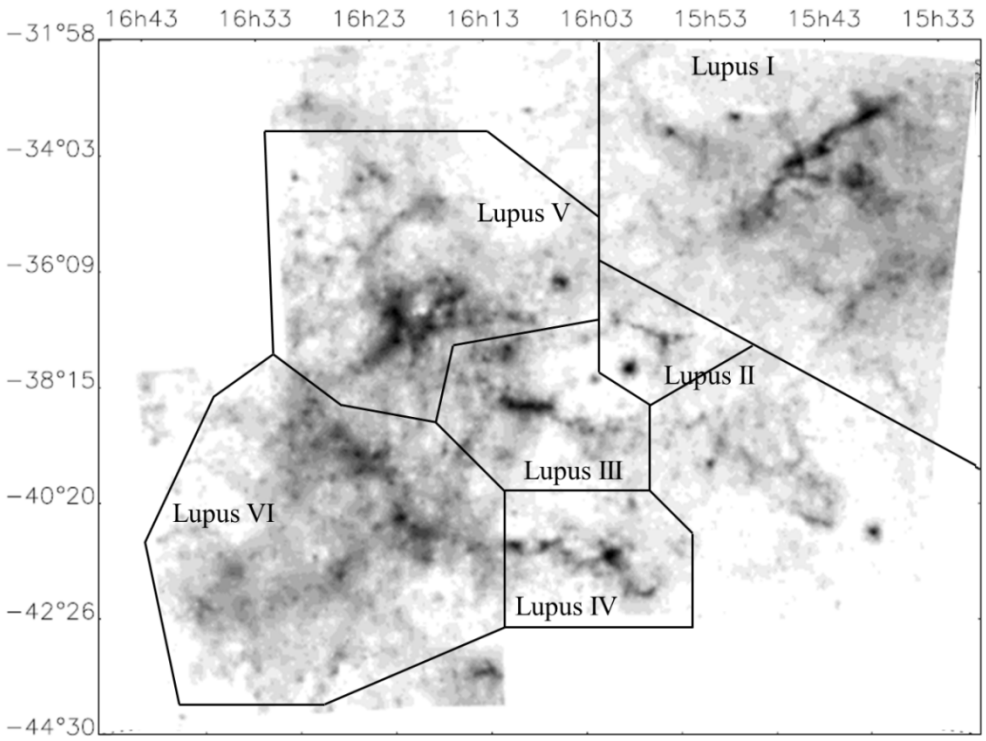


Figura 3 – Mappa d'estinzione di Lupus osservata nella banda B (coordinate J2000; Cambrèsy 1999)

Nel lavoro di riferimento (Cambrèsy 1999) la mappa d'estinzione misurata si riferisce ad immagini osservate con lo strumento USNO-Precision Measuring Machine (USNO-PMM). di cui un esempio è mostrato in Figura 3.

La misurazione è stata effettuata attraverso un metodo di conteggio di stelle usando una griglia adattiva formata da una decomposizione basata su filtri wavelet applicata ad immagini ottiche ad alta risoluzione (Cambrèsy 1999) di cui in Figura 4 è mostrata la regione riferita a Lupus I.

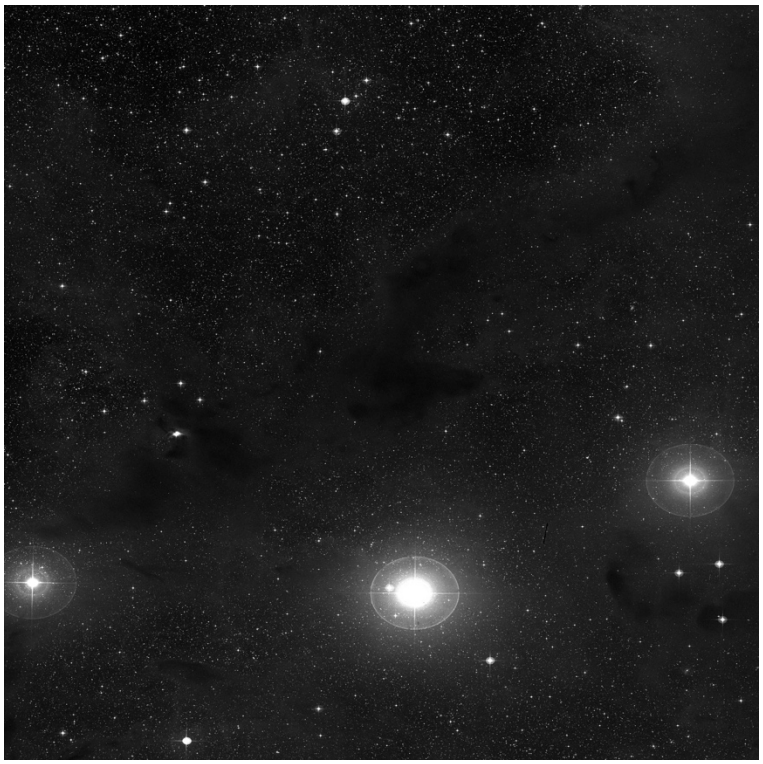


Figura 4 - Immagine osservata da USNO-PMM di Lupus I nella banda B (coordinate J2000, 5x5 gradi DEC, 7000 x 7000 pixel di risoluzione)

La ricostruzione della distribuzione e della misurazione del valore massimo dell'estinzione nelle nebulose risulta particolarmente utile alla stima delle loro masse totali. Il metodo basato sul conteggio di stelle con filtraggio adattivo, usato su larga scala (~250 gradi quadrati), è un sistema tradizionale di misura nel settore, in grado di ottenere una stima accurata della massa di regioni ad alta complessità molecolare. In particolare questo metodo viene calato in un contesto nel quale si assume che lo spettro di massa delle nebulose interstellari abbia una struttura frattale (Elmegreen & Falgarone 1996). Tali strutture possono essere caratterizzate da una relazione lineare tra il raggio di un cerchio e la massa contenuta in un diagramma log-log. Tra le varie definizioni di frattale, quella presa in considerazione è espressa da $M \propto L^D$, dove M è la massa, L è il raggio del cerchio e D la dimensione frattale della nebulosa. La massa misurata è, infatti, contenuta in un cilindro di raggio di base L ed altezza

indefinita H (indefinita dato che la profondità della nebulosa non è uniforme rispetto alla superficie della base del cilindro). Nel nostro caso, si è interessati alla relazione diretta tra la massa e l'estinzione. Poiché quest'ultima è correlata alla grandezza H, che individua la profondità delle nebulose, quello che si cerca è una relazione tra la massa e A_V (o H), avendo L indefinita.

Il logaritmo della massa varia linearmente con l'estinzione in un intervallo di magnitudini d'estinzione:

$$logM = logM_{tot} + slope \times A_V \quad (4)$$

dove il temine *slope* indica il fattore di compensazione tra la variazione di massa ed estinzione (Figura 5). L'andamento mostrato in figura mostra la relazione tra gli iso-contorni d'estinzione ed il logaritmo della massa contenuta nei vari iso-contorni. Come si vede la relazione è lineare fino a circa $A_V \leq 5.5$. Per estinzioni superiori la massa inizia a diminuire drasticamente, poiché il conteggio di stelle tende a sottostimare l'estinzione. Per cui, per zone ad alta estinzione, la bassa densità di stelle richiede un'area più ampia per poter conteggiare un numero di stelle sufficiente a stimare il valore d'estinzione.

Questo risultato è compatibile con la struttura frattale della nebulosa, ipotizzando che $A_V \propto logH$, cioè che la densità della materia segua una legge di potenza, che corrisponde appunto a quanto usato nei modelli teorici di nebulose interstellari (Cambrèsy 1999).

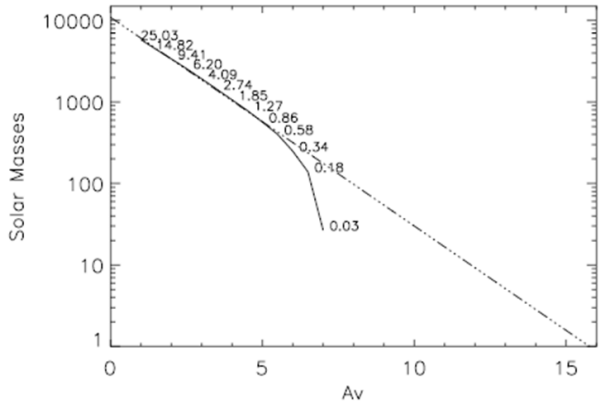


Figura 5 - Regressione tra masse solari ed iso-contorni di estinzione, Cambrèsy 1999

Il metodo descritto porta alla ricostruzione di mappe di estinzione come quella mostrata in Figura 6, dove i livelli di grigio si riferiscono ai diversi iso-contorni di estinzione nelle varie regioni. I livelli più scuri indicano un livello di estinzione crescente, ossia dove la variazione in magnitudine indotta dal contributo d'estinzione è maggiore.

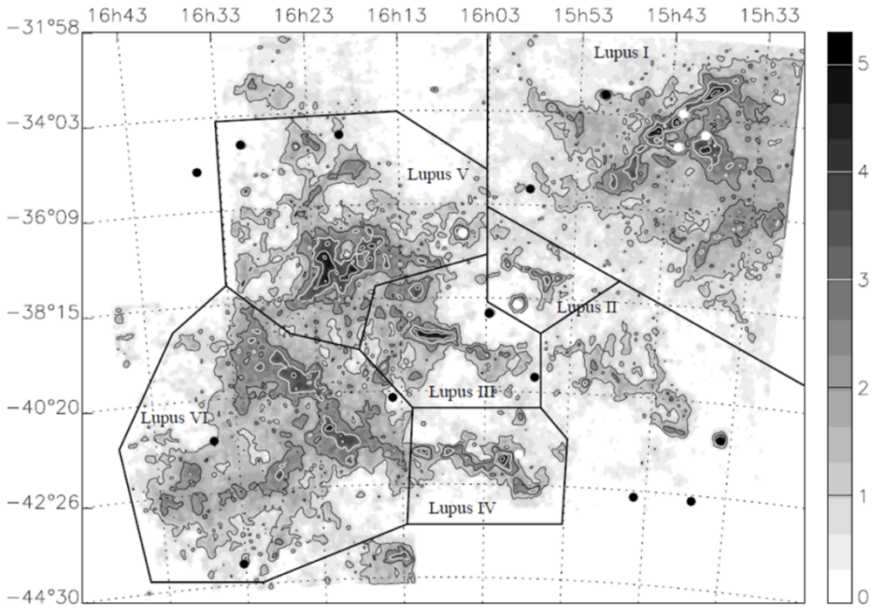


Figura 6 - Mappa d'estinzione della nebulosa Lupus in banda B, con i livelli degli iso-contorni di estinzione, misurati dal metodo tradizionale (Cambrèsy 1999)

Circa il calcolo della massa totale dell'estinzione, assumendo un rapporto gas/polvere, suggerito dall'equazione (1), si usa la seguente relazione (Dickman 1978):

$$M(A_V) = l_{pix}^2 \frac{N_H}{A_V} \mu \frac{m_H}{M_\odot} \sum_i A_V(i) \qquad (5)$$

dove l_{pix}^2 indica l'area del pixel che tenga conto della risoluzione angolare dello strumento e della distanza della nebulosa dalla Terra (che è di circa 150 pc), mentre $\sum_i A_V(i)$ rappresenta la somma dei valori di magnitudine d'estinzione contenuta nei pixel della regione d'interesse.

La principale e inevitabile fonte d'incertezza sulla determinazione della massa deriva dall'approssimazione della distanza della nebulosa. Per cui in linea teorica si può considerare che una sottostima di 0.5 magnitudini d'estinzione implichi una riduzione della massa totale di un fattore ~2.

Con riferimento all'equazione (4), i valori di distanza, estinzione, M_{tot} e *slope* sono tabulati nel lavoro di Cambrèsy 1999 e riportati di seguito nella seguente tabella.

Regione	***distanza (pc)***	$\boldsymbol{A_V^m}$	$\boldsymbol{A_V^e}$	$\boldsymbol{M_{tot}}$	**slope**
Lupus I	150	5.3	7.1	10^4	-0.56
Lupus II	150	3.8	5.7	80	-0.33
Lupus III	150	4.9	7.6	1150	-0.40
Lupus IV	150	5.3	7.0	630	-0.40
Lupus V	150	5.2	10.6	2500	-0.32
Lupus VI	150	4.8	7.0	10^4	-0.57

Tabella 1 - Proprietà della regione Lupus. Le masse totali sono espresse in masse solari, A_V^m indica i valori di estinzione misurata dal metodo tradizionale, mentre A_V^e indica le estinzioni stimate dall'equazione (4), assumendo una struttura frattale.

I metodi di Machine learning descritti in questo lavoro sono stati usati per ottenere una ricostruzione delle mappe d'estinzione alternativa e con cui confrontare direttamente i risultati con il metodo tradizionale, sia in termini di calcolo dell'estinzione sia dell'impatto sulla stima della massa delle regioni interstellari d'interesse.

È importante sottolineare che la valutazione comparativa tra il metodo tradizionale ed i metodi proposti in questo lavoro non possono trovare un riscontro oggettivo, dato che, come spesso accade in Astronomia, il risultato "vero" non esiste. La valenza scientifica dei metodi proposti può essere acclarata in termini di verosimiglianza tra le misure ottenute e quelle presenti in letteratura, ottenendo indirettamente una conferma delle concrete potenzialità dei metodi semi-automatici proposti, alternativi ai metodi tradizionali che, viceversa, richiedono un notevole sforzo in termini computazionali e di lavoro umano.

3 Strumenti utilizzati

In letteratura gli algoritmi di edge detection sono tradizionalmente classificati in due categorie a seconda che si basino sulla derivata prima o seconda della funzione $f(x, y)$ che descrive l'andamento del livello di grigio. Gli edge si configurano infatti come bruschi cambiamenti nell'intensità dei pixel cui corrispondono punti di massimo della derivata prima o attraversamenti dello zero (*zero-crossing*) della derivata seconda (Figura 7).

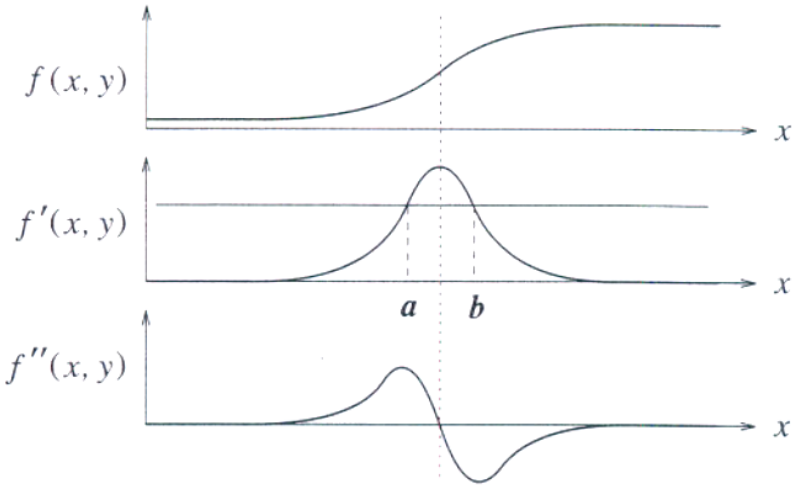


Figura 7 - Identificazione di un edge tramite derivata prima e seconda

L'attività di edge detection può quindi essere ricondotta alla ricerca di picchi nella derivata prima, eventualmente al di sopra di una predeterminata soglia, o zero-crossing della derivata seconda (Marr & Hilderth 1980). Tuttavia quest'ultima, sebbene abbia una forte risposta anche ai dettagli più fini, è solita produrre "doppi" edge e per questo raramente è usata in maniera diretta in attività di edge detection (Bin & Samiei Yeganeh 2012).

3.1 Metodi di edge detection basati sulla derivata prima

Nel campo dell'image processing si è soliti utilizzare un'implementazione della derivata prima basata sul modulo del gradiente $G(x, y)$, ovvero il vettore, mostrato nell'equazione (6) le cui componenti sono le derivate parziali nelle diverse direzioni.

$$G(x, y) = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \end{bmatrix} \tag{6}$$

Il modulo del gradiente può invece essere definito come da equazione (7):

$$|G(x,y)| = \sqrt{{G_x}^2 + {G_y}^2} \approx |G_x| + |G_y| \tag{7}$$

L'approssimazione discreta del gradiente è ottenuta attraverso la convoluzione di maschere, o *kernel*, che vengono fatte scorrere sull'immagine. Il risultato O della convoluzione nel punto (i,j) è ottenuto come somma dei prodotti del valore del kernel stesso per i pixel sottostanti, come mostrato nell'equazione (8). In questo tipo di approccio, come vedremo, le caratteristiche del kernel stesso determinano eventuali proprietà peculiari dell'algoritmo.

$$O(i,j) = \sum_{p=1}^{m} \sum_{l=1}^{n} I(i+p-1, j+l-1) K(p,l) \tag{8}$$

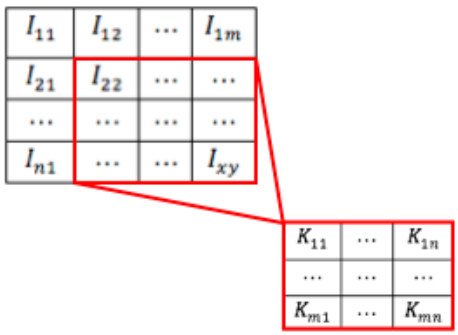


Figura 8 – Matrice dei pixel dell'immagine (a sinistra) e kernel di convoluzione (a destra)

Tra i primi operatori di questo tipo proposti in letteratura figurano gli operatori di Roberts (1963) di cui, di seguito, sono proposti i relativi kernel di convoluzione:

$$G_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad G_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{9}$$

Chiaramente questi operatori fanno uso della differenza tra pixel adiacenti lungo la diagonale ed, essendo una maschera 2×2, il gradiente di $f(x,y)$ non è

approssimato nel punto (x, y) come ci si aspetterebbe, bensì nel punto interpolato $\left[x + \frac{1}{2}, y + \frac{1}{2}\right]$ (Jain et al. 1995).

Un modo per evitare ciò, nonché per ridurre la sensibilità delle maschere al rumore, consiste nell'utilizzo di kernel 3×3 come avviene con gli operatori di Prewitt (1970) e Sobel (1968), i cui kernel sono mostrati rispettivamente dalle equazioni (10) e (11).

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad G_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{10}$$

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{11}$$

Questi ultimi in particolare risultano tra i più utilizzati e sono caratterizzati da una maggior enfasi data al valore dei pixel vicini al centro della maschera.

3.2 Il metodo di Canny

Il lavoro proposto da Canny (1986) è una pietra miliare nel campo dell'edge detection avendo definito una tecnica che risulta tra i migliori metodi *general purpose*. Il suo approccio analitico si basa sulla massimizzazione delle performance in relazione a tre criteri da lui stesso enunciati:

- Buona individuazione: deve esistere un'alta probabilità di individuare punti di edge reali (veri positivi) ed una bassa probabilità di identificare come edge punti che in realtà non lo sono (falsi positivi);

- Buona localizzazione: i punti individuati come edge devono essere il più vicino possibile al centro dell'edge reale;
- Unicità della risposta: deve fornire una singola risposta in corrispondenza di un edge reale.

L'algoritmo può essere riassunto nei seguenti step:

1. *Smoothing dell'immagine con filtro gaussiano;*
2. *Calcolo della magnitudo e direzione del gradiente;*
3. *Soppressione dei non-massimi (analisi dei massimi locali);*
4. *Sogliatura con Isteresi.*

L'applicazione di un filtro gaussiano all'immagine nella prima fase dell'algoritmo ha lo scopo di provare a rimuovere, almeno in parte, il rumore presente nell'immagine. L'equazione (12) mostra il tipo di filtro utilizzato, i cui effetti derivano dalla scelta del parametro σ che determina l'ampiezza della gaussiana e di conseguenza l'entità della regolarizzazione.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{12}$$

Sull'immagine così ottenuta si calcola il gradiente con uno degli operatori analizzati nel paragrafo precedente, dopodiché si procede ad una fase di soppressione dei non-massimi in modo da ottenere edge più sottili. Il procedimento consiste nell'analisi dell'intorno 3×3 di ogni pixel valutandone il gradiente ed eliminando tutti quei punti che non siano massimi locali lungo la direzione del gradiente stesso. L'ultimo step applica il concetto di isteresi utilizzando due soglie, T_1 e T_2 con $T_1 > T_2$. La soglia T_1 è utilizzata come una tradizionale soglia del gradiente, ovvero

etichettando come edge tutti quei pixel il cui modulo del gradiente risulti maggiore. Se tale valore è invece inferiore a T_1 ma superiore a T_2 il corrispondente pixel è candidato ad essere edge e lo diventa solo se tra i suoi 8 vicini ve ne sia almeno uno già sicuramente etichettato come tale.

3.2.1 Scelta dei parametri

Da quanto descritto è evidente come la scelta dei valori soglia influenzi le prestazioni dell'algoritmo. Il concetto di soglia, sia intesa come parametro di un algoritmo sia come tecnica di segmentazione a sé stante (*thresholding)*, è da sempre oggetto di ricerca. Tra le diverse tecniche proposte nel corso degli anni il metodo di Otsu (1979) si è attestato come la miglior soglia in senso statistico e come metodo stabile per quanto riguarda la segmentazione di immagini. La soglia t è calcolata come il valore per cui le due classi di pixel, risultanti dalla sogliatura, abbiano varianza minima internamente alla singola classe e conseguentemente varianza massima tra classi. La soglia così calcolata può essere usata per il metodo di Canny ottenendo performance migliori rispetto a soglie calcolate tramite altri metodi statistici (Fang et al. 2009). Se t è la soglia calcolata tramite il metodo di Otsu, le soglie T_1 e T_2, utilizzate nell'algoritmo di Canny per la fase di isteresi, saranno rispettivamente pari a t e $t/2$.

3.3 Fuzzy Edge Detection

La teoria degli insiemi fuzzy, introdotta da Zadeh (1965), è un'estensione della convenzionale teoria degli insiemi che si pone l'obiettivo di manipolare la verità parziale allo scopo di modellare l'incertezza. Volendo fornire un esempio pratico

potremmo dire che, nell'accezione classica di insieme, è possibile etichettare un pixel come "scuro" definendo una soglia e stabilendo l'appartenenza a tale insieme a seconda che il valore del pixel sia inferiore o superiore. Logica conseguenza di questo approccio è che un pixel apparterrà esclusivamente o all'insieme dei pixel "scuri" o all'insieme dei pixel "chiari" (Figura 9a).

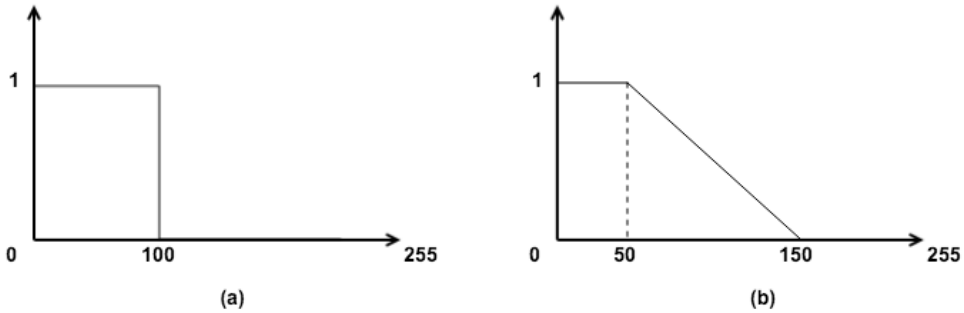


Figura 9 - Appartenenza ad un insieme crispy (a) e ad un insieme fuzzy (b)

Ma se invece stabilissimo due soglie potremmo affermare che tutti i pixel al di sotto della prima soglia appartengono per intero all'insieme dei pixel scuri ed analogamente che tutti quelli superiori alla seconda appartengono all'insieme dei pixel chiari. Tutti i pixel intermedi apparterranno invece all'insieme dei pixel scuri, secondo un grado m stabilito tramite una funzione di appartenenza (Figura 9b) ed all'insieme dei pixel chiari secondo un grado $1 - m$.

Si evince quindi che la corretta definizione delle funzioni di appartenenza ha un ruolo chiave in sistemi di questo tipo, poiché la loro forma determina il risultato dei processi volti alla rappresentazione di un determinato input nel dominio degli insiemi fuzzy, nonché la riconversione per l'ottenimento del risultato finale, processi noti rispettivamente come *fuzzification* e *defuzzification*.

Questa capacità di trattare dati ambigui ha portato all'applicazione della logica fuzzy al campo dell'image processing dando origine a quello che è noto come *Fuzzy Image Processing* (FIP), definito come la collezione di approcci che rappresentano, processano ed interpretano le immagini e le loro features come insiemi fuzzy (Tishoosh 1997).

In Figura 10 è mostrata la struttura di un classico *Fuzzy Inference System* (FIS) applicato ad un'immagine in input.

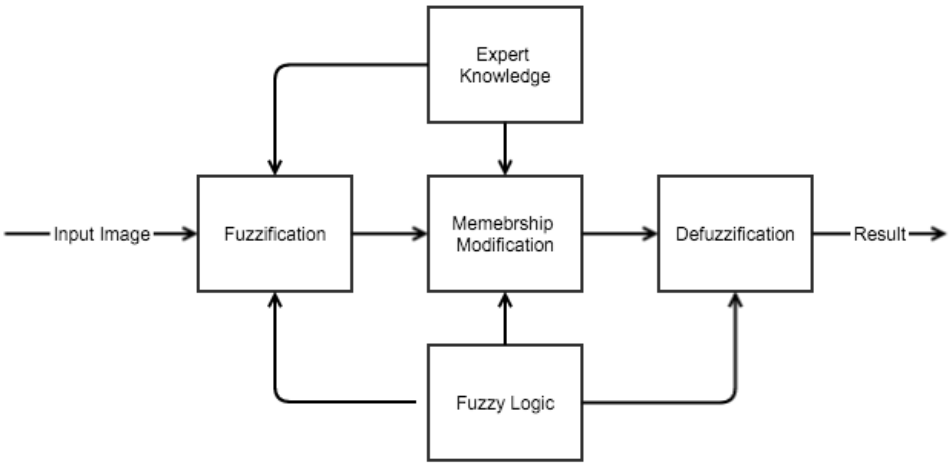


Figura 10 - Struttura generale del Fuzzy Image Processing

L'idea alla base di questo tipo di logica è semplice e si avvicina al modo naturale di intendere la conoscenza, tanto da poter esprimere l'appartenenza ad un determinato insieme fuzzy in termini di regole nella forma *If – Then* utilizzando variabili linguistiche. Ciò permette di rappresentare la conoscenza del dominio di un esperto, la quale può essere codificata ed utilizzata nella fase di fuzzification nonché nella fase di inferenza, cioè quando avviene la modifica del grado di appartenenza di una variabile ad un insieme.

Ovviamente approcci fuzzy sono stati proposti anche nell'ambito dell'edge detection, con ottimi risultati come mostrato da Kaur et al. (2010). Come avviene in molti approcci di questo tipo, anche in questo caso viene fatta scorrere sull'immagine una finestra 3×3 cui vengono applicate 16 regole fuzzy che stabiliscono la presenza di un edge in quella determinata regione. Le performance di questo tipo di approccio calano però se applicate ad immagini rumorose, come evidenziato da Haq et al. (2015) dove viene proposto un approccio simile ma più robusto rispetto al rumore.

Una finestra 3×3 scorre sull'immagine analizzando di volta in volta i valori P_j degli otto pixel intorno a P (Figura 11a) e tali valori sono successivamente processati ottenendo $\Delta P_j = |P_j - P|$ (Figura 11b).

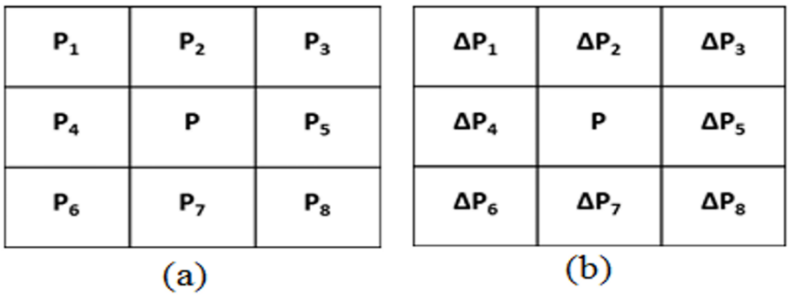


Figura 11 - Maschere utilizzate nel sistema di Fuzzy Edge Detection

Di seguito sono invece proposte le funzioni di appartenenza scelte sulla base di valutazioni empiriche che hanno determinato l'utilizzo di una funzione trapezoidale per l'input ed una gaussiana per l'output. Quest'ultima particolarmente adatta in quanto smooth e diversa da 0 in tutti i punti.

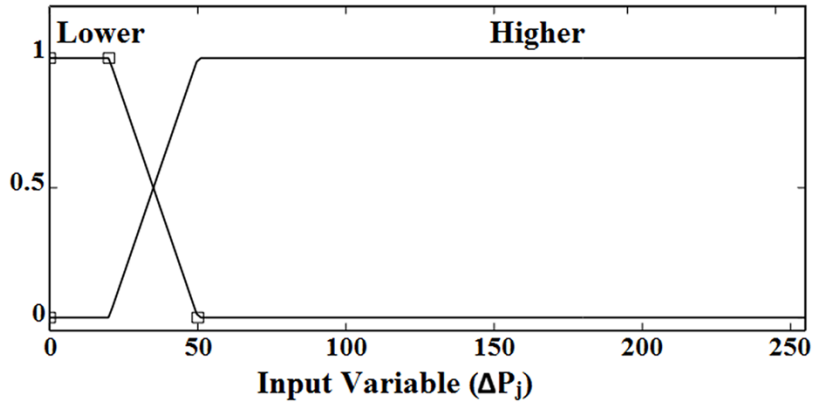


Figura 12 - Funzione di appartenenza per la variabile in input

I valori per la funzione trapezoidale per la variabile *Lower* sono [0 0 20 50] mentre quelli per la variabile *Higher* [20 50 255 255].

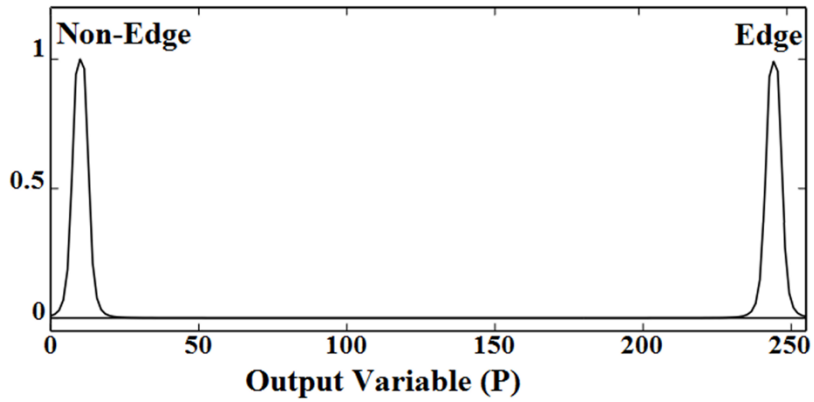


Figura 13 - Funzione di appartenenza per la variabile in output

Per quanto riguarda l'output, i valori per la funzione della variabile *Non-Edge* sono [3.5 10] mentre per la variabile *Edge* sono [3.5 245].

L'immagine seguente mostra invece la base di conoscenza fuzzy, ovvero il set di regole che determina la trasformazione del grado di appartenenza della variabile di output rispetto agli insiemi fuzzy, in questo caso Edge e Non-Edge. Questi valori sono combinati, secondo determinate regole, in modo da avere un output unico ottenibile tramite diversi metodi, tra cui il calcolo del centroide che si attesta come una delle tecniche più efficaci ed efficienti (Leekwijck 1999).

Rules	Input Variables								Output Variable
	ΔP_1	ΔP_2	ΔP_3	ΔP_4	ΔP_5	ΔP_6	ΔP_7	ΔP_8	P
1	Higher	Higher	None	None	None	None	None	Lower	Edge
2	Higher	None	None	High	None	None	None	Lower	Edge
3	None	Higher	Higher	None	None	None	None	Lower	Edge
4	None	None	None	Higher	None	Higher	None	Lower	Edge
5	Higher	Higher	None	None	None	None	Lower	None	Edge
6	Higher	None	None	Higher	None	None	Lower	None	Edge
7	None	Higher	Higher	None	None	None	Lower	None	Edge
8	None	None	None	Higher	None	Higher	Lower	None	Edge
9	Higher	Higher	None	None	Lower	None	None	None	Edge
10	Higher	None	None	Higher	Lower	None	None	None	Edge
11	None	Higher	Higher	None	Lower	None	None	None	Edge
12	None	None	None	Higher	Lower	Higher	None	None	Edge

Figura 14 - Set di regole per il sistema di Fuzzy Edge Detection. Quelle relative all'output non-edge sono il complementare di quelle mostrate.

3.4 Clustering

L'image segmentation consiste nel partizionamento di un'immagine in gruppi di pixel tali che, pixel appartenenti alla stessa partizione condividano delle proprietà. I pixel delle regioni così formate risultano quindi simili rispetto ad una o più caratteristiche, mentre quelli delle regioni adiacenti differiscono tra loro in relazione alle medesime proprietà. Lo scopo finale è quindi quello di ottenere una rappresentazione dell'immagine che sia in qualche modo più semplice da analizzare (Bishop 2006) e risulta evidente come l'image segmentation sia strettamente correlata al dominio funzionale del clustering (Brescia 2012).

Il clustering è il processo di raggruppamento (segmentazione nel caso di immagini) di oggetti in un insieme finito di classi che siano rappresentative per il problema affrontato. Le relazioni che intercorrono tra gli oggetti sono solitamente rappresentate in termini di prossimità tra gli stessi, da intendersi come distanza tra coppie di punti nello spazio dei parametri. Quest'informazione è nella maggior parte dei casi l'unico input di un algoritmo di clustering che determina l'appartenenza di

un oggetto ad un cluster piuttosto che ad un altro (Jain & Dubes 1988). Si parla cioè di apprendimento non supervisionato, poiché nessuna conoscenza sulle classi è nota a priori. Una tassonomia dei metodi di clustering largamente accettata, sebbene di alto livello, li divide in due principali categorie: clustering gerarchico e clustering partizionale.

Il clustering gerarchico raggruppa i dati attraverso una sequenza di partizionamenti che può partire da un singolo cluster contenente tutti i dati, detto approccio top-down o divisive, oppure da un cluster per ogni punto, noto come approccio bottom-up o agglomerative. I risultati di un clustering gerarchico sono solitamente descritti in una struttura ad albero che prende il nome di dendogramma (Figura 15) la cui radice rappresenta l'intero dataset mentre ogni foglia rappresenta un oggetto. Ogni livello interno rappresenta invece un partizionamento per cui il risultato finale può essere ottenuto tagliando suddetta struttura in un determinato punto.

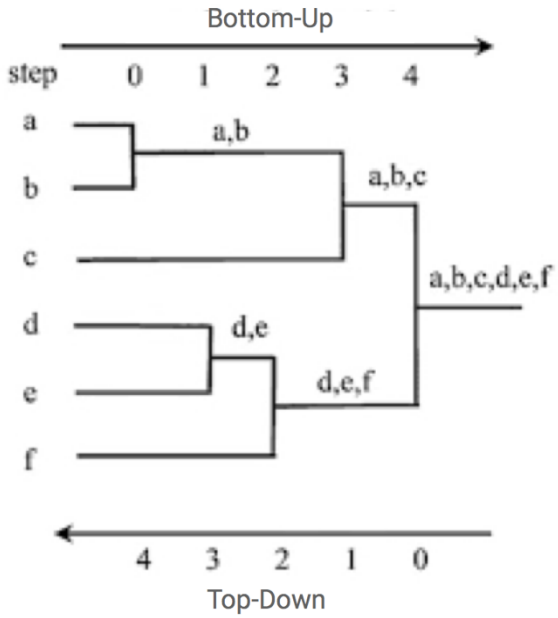


Figura 15 - Dendogramma con partizioni ottenute tramite metodi agglomerative e divisive

Il clustering partizionale invece assegna gli oggetti ai cluster senza avvalersi di una struttura gerarchica. Il partizionamento ottimale non può ovviamente essere ottenuto esaminando tutte le possibilità, per cui si cerca euristicamente di ottenere l'approssimazione migliore minimizzando una funzione di costo. Un criterio ampiamente utilizzato è l'errore quadratico definito come:

$$J(\Gamma, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{i,j} \left\| x_j - \ m_i \right\|^2 \quad (12)$$

dove

$\Gamma = una\ partizione$

$M = \ [m_1, \dots, m_k] = \ centroidi\ dei\ cluster$

$x_j = oggetto\ del\ dataset$

$N = numero\ di\ oggetti\ nel\ dataset$

Il più diffuso algoritmo di clustering che operi sulla base di questo criterio è il K-Means (MacQueen 1967).

3.4.1 K-Means

L'algoritmo inizia calcolando un insieme di k partizioni iniziali, in modo casuale o secondo un criterio specificato, dopodiché ne calcola i centroidi $M = [m_1, \dots, m_k]$. Ogni oggetto x del dataset è assegnato al cluster C il cui centroide risulta più vicino, ovvero:

$$for\ j = 1, \dots, N,\ \ i \neq w\ and\ i = 1, \dots, k$$

$$x_j \in C_w\ if\ \|x_j - m_w\| < \|x_j - m_i\|$$

I centroidi sono quindi ricalcolati in base alle nuove partizioni ottenute ed il procedimento è reiterato fin quando non vi siano più cambiamenti nei cluster.

L'algoritmo è molto semplice e veloce e si adatta bene a gran parte dei problemi di clustering, il che lo rende, nella maggior parte dei casi, una buona scelta preliminare quando si opera su problemi di questo tipo. Il K-Means, soffre però di problemi ben noti in letteratura. Il principale risiede nell'identificazione dell'insieme di partizioni iniziale, problema per cui non esiste un metodo efficiente ed universale. È inoltre sensibile al noise ed agli outliers. Questi ultimi, sebbene molto distanti dal centroide più vicino, sono comunque forzatamente inclusi nel cluster relativo, distorcendo di conseguenza la forma del cluster stesso. (Xu 2005)

Per quanto riguarda il problema dell'inizializzazione, diverse proposte sono state fatte in letteratura nel corso degli anni, alcune di queste con il minimo comun denominatore di utilizzare un approccio basato sull'esecuzione di diversi stadi di clustering. In alcuni casi queste tecniche si traducono nell'utilizzo di una serie di K-Means i cui risultati sono combinati secondo vari criteri (Bradley & Fayyad 1998; Likas et al. 2003), ma di particolare interesse è la proposta contenuta in Vesanto & Alhoniemi (2000) che, oltre ad intervenire sul problema delle partizioni iniziali,

riduce contemporaneamente il disturbo indotto da noise ed outliers. Tale proposta si basa sull'utilizzo di una Self Organizing Map (SOM) e, per comprendere meglio i vantaggi che questo tipo di approccio può offrire, il paragrafo seguente chiarirà prima di tutto il funzionamento di tale tipo di rete.

3.4.2 Self Organizing Map

Una SOM (Kohonen 2001) è costituita da due strati, dei quali uno di input e l'altro di output. I neuroni dei due strati sono completamente connessi tra loro, mentre i neuroni dello strato di output sono connessi, ciascuno, con un vicinato di neuroni secondo un sistema di inibizione laterale. I pesi dei collegamenti intra-strato dello strato di output, o strato di Kohonen, non sono soggetti ad apprendimento ma sono fissi e positivi. Ogni neurone dello strato di Kohonen riceve uno stimolo pari alla sommatoria degli input moltiplicati per il rispettivo peso sinaptico.

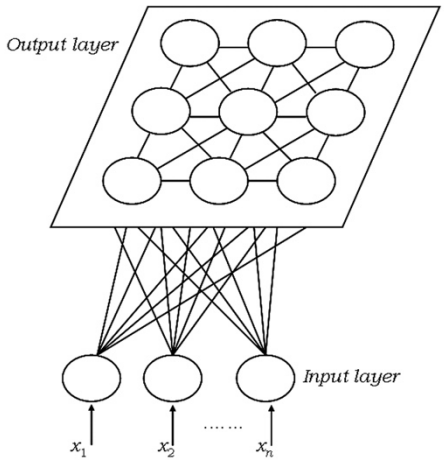


Figura 16 - Architettura di una rete SOM

I pesi sinaptici del neurone che risulta "vincente" (*Best Matching Unit,* BMU), per un determinato input fornito alla rete sono invece aggiornati come mostrato nell'equazione (13) e quindi avvicinati, nello spazio dei parametri, al vettore in input in modo che il neurone vincente sia ulteriormente sensibilizzato al riconoscimento di un determinato pattern.

$$W(k,j)_{new} = W(k,j)_{old} + \eta * (X(k) - W(k,j)_{old}) \quad (13)$$

dove:

$W(k,j) = peso\ sinaptico\ del\ collegamento\ tra\ input\ k\ e\ neurone\ vincente$;

$\mathrm{X(k)} = input\ k_{esimo}\ dell'input\ pattern$;

$\eta = tasso\ di\ apprendimento\ (learning\ rate), nel\ range\ ]0,\ 1[.$

Generalmente anche i pesi dei neuroni vicini al neurone vincente sono soggetti ad apprendimento, con un learning rate decrescente a seconda della distanza da esso. Di solito il decadimento del learning rate avviene secondo la funzione definita a "cappello messicano". L'apprendimento esteso ai neuroni vicini favorisce il formarsi di "bolle di attivazione" che identificano input simili.

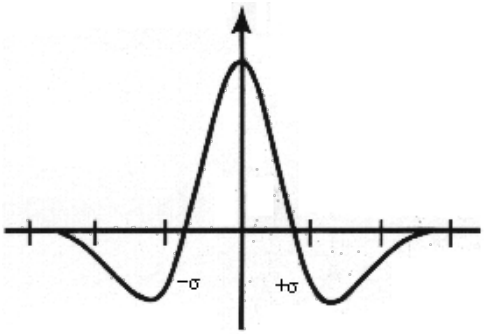


Figura 17 - Decadimento del learning rate secondo la funzione a cappello messicano

Si verifica inoltre che, elementi vicini tra loro sulla mappa siano anche vicini nello spazio dei pesi rendendo la SOM un potente strumento di visualizzazione dei dati. Lo strato di Kohonen è infatti in grado di rappresentare, rispettando la topologia, dati multi-dimensionali su una mappa a due o tre dimensioni. Lo standard per la valutazione e l'interpretazione di una SOM è la unified distance matrix (U-Matrix). Siano:

n, $neurone\ della\ mappa$

$NN(n)$, $insieme\ dei\ nodi\ adiacenti\ ad\ n\ sulla\ mappa$

$w(n)$, $vettore\ dei\ pesi\ associato\ al\ neurone$ n

$\|w(n), w(m)\|$, $distanza\ euclidea\ tra\ i\ vettori\ dei\ pesi\ dei\ neuroni\ n\ ed$ m

$U_{height(n)}$, $valore\ (peso)\ associato\ al\ nodo$ n

Ad ogni nodo dello strato di Kohonen verrà assegnato un peso secondo la seguente formula:

$$U_{height(n)} = \sum_{m \in NN(n)} \|w(n), w(m)\| \quad (14)$$

Il valore così calcolato diventa identificativo della distanza di un nodo da tutti i suoi vicini più prossimi ed è visualizzabile utilizzando una heat map in cui colori chiari rappresentano neuroni vicini tra loro nello spazio dei pesi, viceversa colori scuri rappresentano neuroni lontani tra loro (Moutarde & Ultsch 2005). Tipicamente si usa rappresentare la mappa con una scala di grigi e, per aumentare ulteriormente il grado di interpretabilità della U-Matrix, è possibile sovrapporre ad ogni nodo un colore che ne identifichi il cluster di appartenenza. Ovviamente, i nodi, su cui non sia sovrapposto alcun colore, non sono mai risultati vincenti per qualche pattern input.

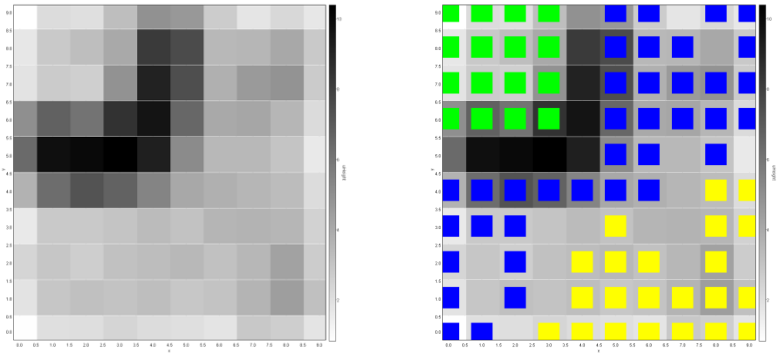


Figura 18 - Esempi di U-Matrix

Per valutare la qualità di una mappa ottenuta ci si può basare sui criteri suggeriti da Kiviluoto (1996):

- Qual è il grado di continuità della mappa topologica?
- Qual è la risoluzione della mappa?
- La topologia della mappa riflette la probabilità di distribuzione dello spazio dei dati?

Per quanto concerne il terzo punto, in letteratura sono presenti diversi esempi di SOM con struttura incrementale in grado di preservare la topologia dello spazio dei dati mentre la quantificazione delle prime due proprietà può essere ottenuta tramite il calcolo dell'errore di quantizzazione e dell'errore topografico.

L'errore di quantizzazione viene utilizzato per calcolare la similarità tra i pattern assegnati al medesimo nodo. Di seguito è mostrata la formula per il calcolo del suddetto errore, che corrisponde alla media delle distanze di ogni pattern dal vettore dei pesi del nodo che lo identifica.

$$QE = \frac{1}{N}\sum_{i=1}^{N}\|\overrightarrow{w_{BMUi}} - \overrightarrow{x_i}\| \quad (14)$$

dove:

$\overrightarrow{w_{BMUi}} = vettore\ dei\ pesi\ dell'\ i_{esimo}$ BMU

$N = numero\ di\ pattern\ che\ compongono\ il\ dataset$

$\overrightarrow{x_i} = i_{esimo}\ vettore\ in\ input\ rappresentato\ dal\ BMU\ considerato$

L'errore topografico viene utilizzato per calcolare la dissimilarità tra i pattern assegnati a nodi differenti. Anche per questo errore di seguito viene fornita la formula.

$$TE = \frac{1}{N}\sum_i^N u(\overrightarrow{w_i}) \tag{15}$$

dove

$N = numero\ di\ pattern\ che\ compongono\ il\ dataset$

$$u(\overrightarrow{x_i}) = \begin{cases} 1, & se\ primo\ e\ secondo\ BMU\ dell'i_{esimo}\ pattern\ non\ adiacenti \\ 0\ altrimenti \end{cases}$$

La formula (15) corrisponde quindi alla media del numero di volte in cui, per uno stesso pattern, primo e secondo BMU non siano adiacenti sulla griglia dello strato di Kohonen, dove il primo e secondo BMU sono intesi in termini di vicinanza del relativo vettore dei pesi rispetto al pattern di riferimento.

3.4.3 Two-Stage Clustering

Alla luce di quanto detto, è evidente che algoritmi quali SOM e K-Means, possano trarre vantaggio da un approccio a doppio stadio. Una rete SOM infatti, per sua natura non è in grado di produrre un corretto raggruppamento dei nodi della mappa limitandosi, quasi sempre, a produrre una corrispondenza 1:1 tra nodi e cluster.

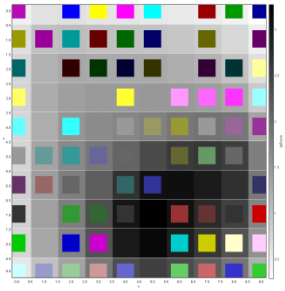


Figura 19 - U-Matrix senza corretto partizionamento dei nodi

È tuttavia in grado di convertire dati complessi in una forma più semplice riducendone l'elevata dimensionalità e configurandosi quindi come ottimo primo step di una tecnica a due stadi. I nodi della rete sono infatti medie locali dei dati e, per questo, meno sensibili al rumore rispetto ai dati in input. Anche gli outliers non risultano un problema in quanto sono, per definizione, pochi punti che non incidono sul risultato finale.

Una tecnica di clustering al secondo stadio, come ad esempio il K-Means, può quindi beneficiare di questi vantaggi lavorando sui neuroni dello strato di Kohonen piuttosto che sul dataset iniziale e raggrupparli in cluster che siano meglio rappresentanti delle strutture presenti nei dati in input.

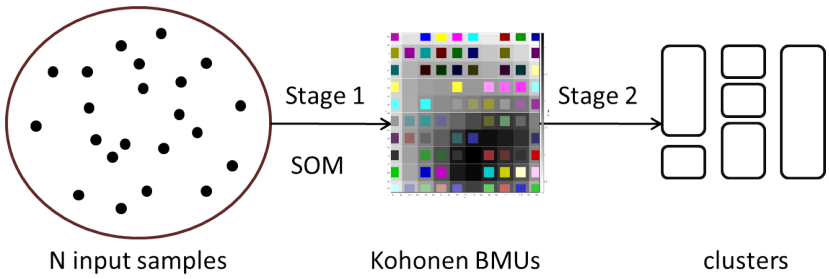


Figura 20 - Procedura del two-stage clustering

È dimostrato come l'utilizzo di un metodo a doppio stadio basato su SOM e K-Means produca risultati migliori rispetto ad un K-Means utilizzato singolarmente (Chi & Yang 2008). Ovviamente resta l'esigenza della scelta del parametro *k*

iniziale, sebbene le migliori prestazioni computazionali permettano più efficientemente di testare il metodo, aumentando il parametro di volta in volta fino ad un criterio di arresto. In alternativa è possibile utilizzare, al secondo stadio, un metodo che non necessiti di una conoscenza pregressa sul numero di cluster.

Hamel & Brown (2011) ad esempio propongono un metodo per migliorare l'interpretazione della U-Matrix immaginando i nodi della stessa come vertici di un grafo in cui delle componenti connesse (CC) siano identificative di cluster. Il procedimento con cui le CC sono identificate si basa sul concetto che, per ognuna di esse, esisterà un nodo, definito interno ad una CC, il cui gradiente, inteso come livello di grigio, sarà inferiore rispetto a quello di tutti gli altri. Per ogni nodo della mappa quindi verranno valutati i gradienti di tutti i suoi nodi adiacenti. Se il nodo esaminato non è quello con il minimo gradiente, allora ci si potrà muovere lungo un percorso attraversando di volta in volta il nodo con gradiente minimo, seguendo i suoi vicini e connettendo tutti i nodi attraversati. All'attraversamento di un nodo interno ad una CC, la procedura termina e si passa ad esaminare un nodo successivo. Mostrando le CC così create, al di sopra della U-Matrix, l'appartenenza di un nodo ad un cluster piuttosto che ad un altro, diventa evidente, come è possibile notare nell'esempio proposto dagli stessi autori e mostrato di seguito.

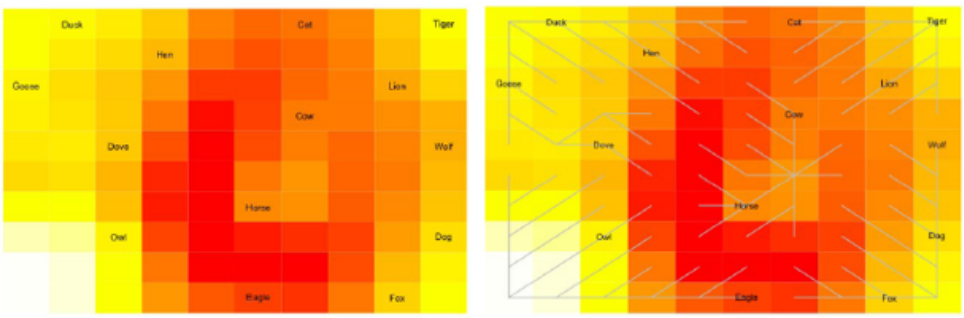


Figura 21 - U-Matrix con componenti connesse evidenziate

3.4.4 Validazione del clustering

Tra i criteri più noti per la valutazione di un clustering figurano l'analisi della Silhouette e l'indice di Davies-Bouldin. Entrambi gli approcci, definiti criteri interni in quanto calcolati sulla base dei dati in input, valutano il rapporto tra la distribuzione *intra-cluster* e le distanze *inter-cluster*. Mentre però la Silhouette utilizza le distanze medie tra punti, l'indice di Davies-Bouldin utilizza i centroidi dei cluster per effettuare la valutazione.

Empiricamente è stato riscontrato che i risultati ottenuti utilizzando un criterio basato sulla Silhouette risultano più accurati rispetto all'utilizzo dell'indice di Davies-Bouldin, chiaramente ad un costo computazionale notevolmente superiore (Petrovic 2006).

4 Il Metodo di cluster analysis proposto

Dopo aver analizzato le prestazioni ottenute con metodi più o meno classici di edge detection, l'attenzione è stata rivolta a metodi di Machine Learning. Più in particolare si è analizzata la possibilità che metodi di clustering, date le caratteristiche peculiari che esibiscono, possano, lavorando nel giusto spazio dei parametri, identificare dei confini per la regione di estinzione diversi da quelli ottenibili con un semplice valore soglia.

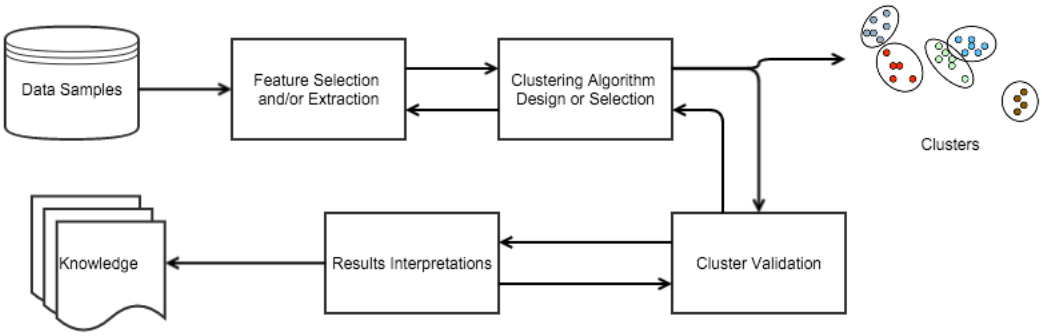


Figura 22 - Procedura tradizionale di cluster analysis

Seguendo fedelmente il flow chart di una tipica procedura di cluster analysis, riportata in Figura 22, sarà in primo luogo affrontata la questione relativa alla scelta dello spazio dei parametri, ovvero la feature selection ed extraction. Analizzeremo successivamente i criteri per cui alcuni algoritmi di clustering sono stati giudicati più promettenti di altri ed infine forniremo una descrizione del metodo proposto corredata dai criteri per la validazione ed interpretazione del clustering ottenuto.

Nel descrivere alcuni aspetti del metodo si farà riferimento, nel corso della trattazione, all'immagine riportata in Figura 23, le cui caratteristiche sono già state illustrate nel Capitolo 2.

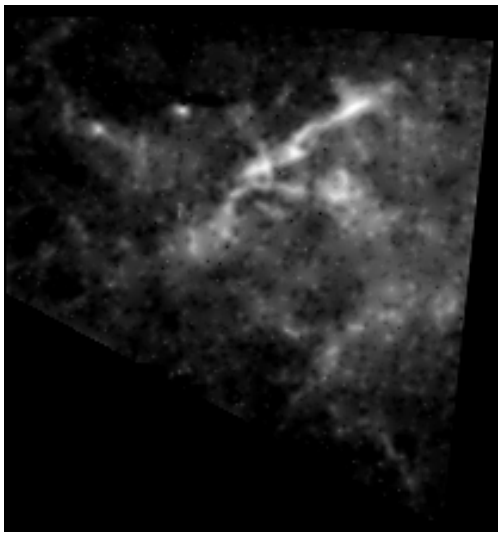


Figura 23 - Mappa di estinzione della regione Lupus I

4.1 Feature Selection ed Extraction

Come è noto gli algoritmi di machine learning, ed in particolare gli algoritmi non supervisionati, devono gran parte della loro efficacia e versatilità alla scelta dello spazio dei parametri sul quale lavorano, ovvero l'insieme di features con cui si sceglie di rappresentare ogni oggetto presente nello spazio dei dati (Jain et al. 1999). Procedure di *feature selection* e *feature extraction* figurano infatti come primi step della maggior parte dei processi di data mining. Per feature selection si intendono quegli algoritmi volti alla selezione del miglior set di features da utilizzare, ovvero quell'insieme di caratteristiche rilevanti per la definizione della classe oggetto. Ci si riferisce invece alla feature extraction come al processo di creazione di nuove features basato su trasformazione o combinazione delle features esistenti (Jain et al. 2000). Nell'ambito dell'image processing l'uso di features estratte a partire dall'immagine è largamente preferito ad un approccio basato sui valori dei pixel (Viola & Jones 2001). La motivazione principale risiede nel fatto che alcune features possono essere scelte ad-hoc per codificare la conoscenza intrinseca di un dominio.

Le tecniche di feature selection sono solitamente divise in tre categorie. I *filter methods* sono tecniche basate sulla misura di importanza che ogni feature assume nello spazio dei parametri di partenza (Sanchez-Marono et al. 2007), permettendo quindi di determinare quali siano le feature più rilevanti per ottenere un determinato risultato. Questo tipo di analisi è indipendente dall'algoritmo di machine learning utilizzato, tuttavia ogni feature è valutata separatamente trascurando il contributo offerto dal loro utilizzo combinato. I *wrapper methods* operano invece selezionando di volta in volta un sottoinsieme dello spazio dei parametri e valutando i risultati ottenuti applicando un algoritmo di machine learning (Kohavi & John 1997). Il vantaggio di questi metodi è la miglior interazione tra feature e algoritmo utilizzato. Sebbene ciò si ottenga al costo di un elevato carico computazionale, dovuto alla ripetizione del processo di selezione di un sottoinsieme di feature ed esecuzione dell'algoritmo fin quando non sia raggiunto il risultato ottimale. Negli *embedded methods* invece il sottoinsieme ottimale è cercato direttamente durante l'esecuzione dell'algoritmo di machine learning, come ad esempio nei Random Forest (Breiman 2001). Come i *wrapper* includono quindi l'interazione con il metodo ma in un modo più efficiente (Vapnik 1998).

L'obiettivo di questo lavoro spinge, in prima istanza, all'identificazione di features in grado di valutare l'appartenenza di un pixel ad un edge. Nel capitolo 3 è stato descritto come la letteratura sull'argomento abbia abbondantemente dimostrato l'efficacia dell'utilizzo del gradiente nella definizione degli edge. Quindi tale feature trova naturale impiego nel metodo proposto. Le restanti features sono invece state scelte da un insieme di classiche misurazioni statistiche quali: media, deviazione standard, range, skewness, kurtosis, ed entropia. La selezione è avvenuta tramite un

approccio di *backward elimination*, partendo cioè dall'intero set di features e valutando i risultati ottenuti rimuovendone alcune ad ogni iterazione.

Come si ha avuto modo di constatare nei capitoli precedenti, il carico informativo estratto da un singolo pixel non è sufficiente se non messo in relazione con quello dei suoi vicini in una finestra di dimensioni predefinite. Per questo motivo l'algoritmo utilizzato calcola le features di un pixel $[x, y]$ prendendo in analisi una finestra 3×3 intorno ad esso.

4.1.1 Pre-Processing dei dati

Sebbene sia acclarato che l'obiettivo da conseguire è l'identificazione di un set di features tale che il carico informativo apportato sia massimo rispetto al problema affrontato, è altrettanto valido l'approccio secondo cui, sulla base di considerazioni a priori, si effettua una trasformazione dei dati volta alla massimizzazione dell'informazione estraibile. Quest'affermazione è sicuramente vera nell'ambito dell'image processing in cui tipicamente i due approcci coesistono in una pipeline che parte con una delle tante tecniche di image enhancement proposte in letteratura, scelta secondo il dominio di interesse, per poi proseguire con l'estrazione delle features selezionate.

In una mappa di estinzione è lecito ipotizzare che pixel caratterizzati da alta estinzione siano, con buona probabilità, appartenenti alla nube mentre, più verosimilmente, un possibile miglioramento può essere apportato ai margini, dove l'estinzione è più bassa. Fatta questa considerazione possiamo quindi assumere di effettuare un miglioramento del contrasto nell'immagine senza perdita di informazione e sfruttare i vantaggi che una regione ad alta intensità, più o meno

uniforme, può portare. Di seguito è riportata l'immagine ottenuta tramite la tecnica di miglioramento del contrasto che prevede la saturazione di una percentuale di pixel al livello più alto e più basso di intensità di grigio[1].

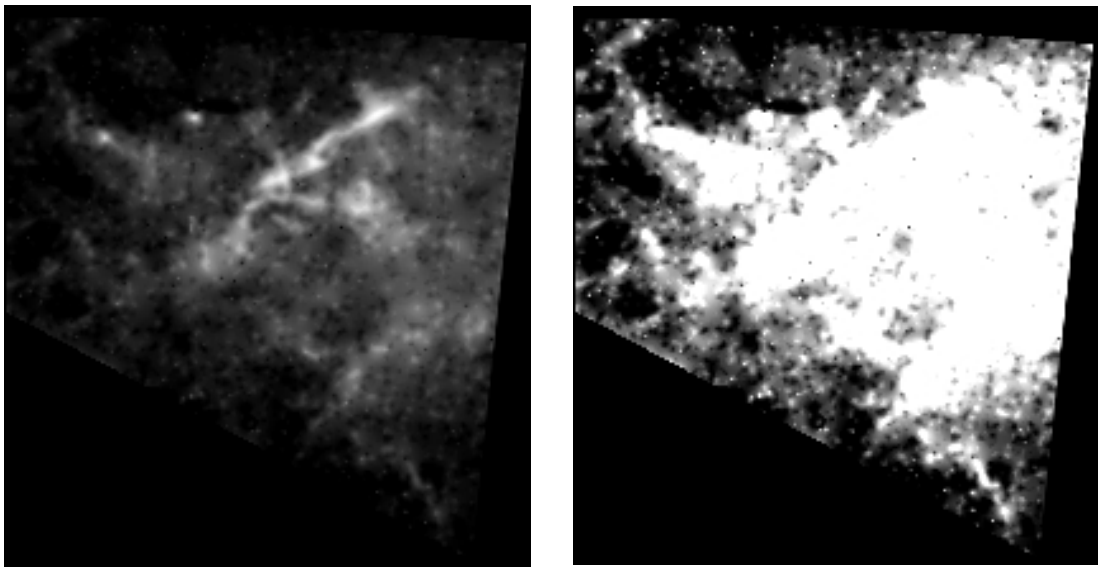


Figura 24 - Mappa di estinzione della regione Lupus I prima e dopo la fase di pre-processing

4.2 Il metodo generale

Alla luce di quanto discusso nei precedenti paragrafi, il metodo proposto può essere riassunto nel diagramma di flusso mostrato in Figura 25.

[1] Funzione imadjust di Matlab (it.mathworks.com/help/images/ref/imadjust.html)

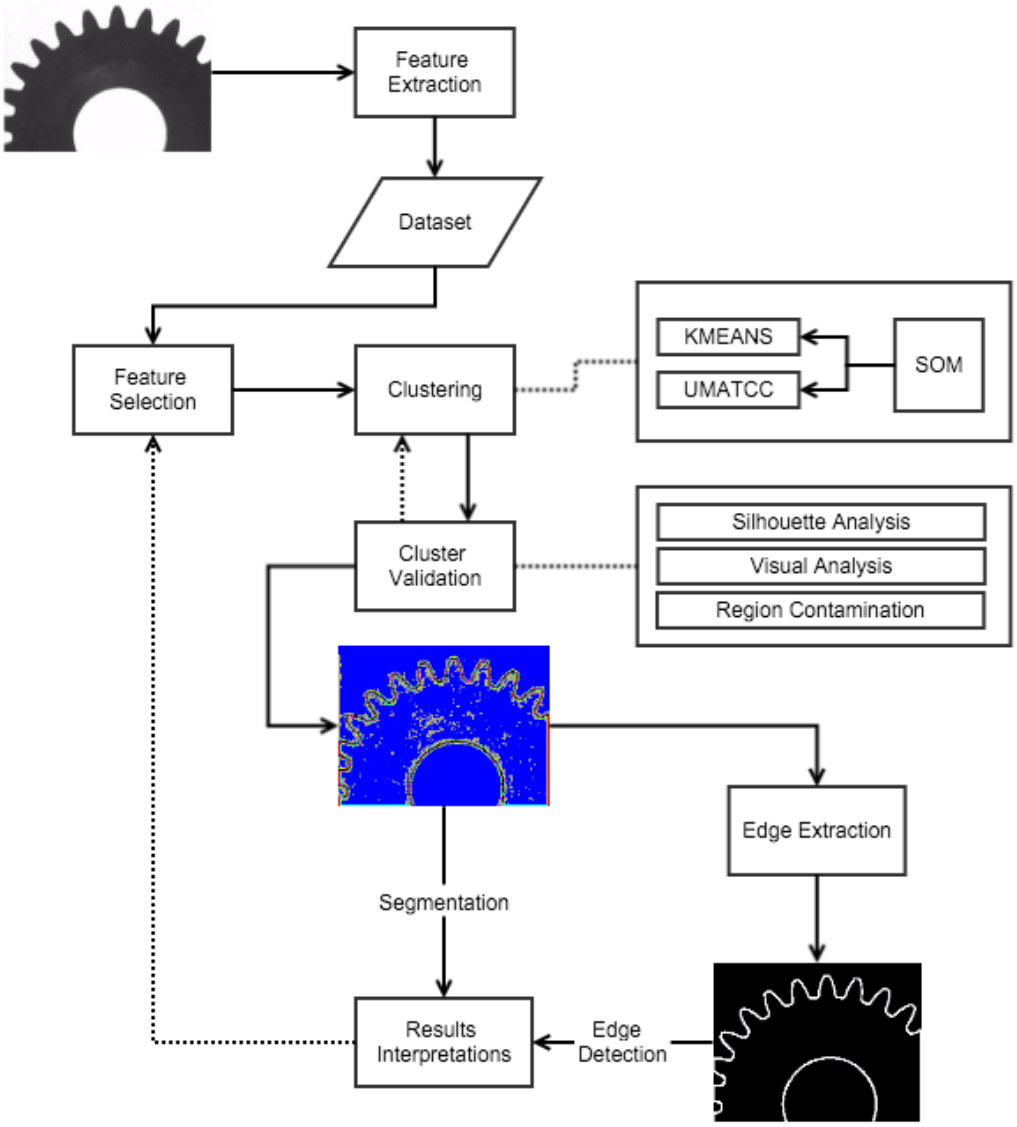


Figura 25 - Diagramma di flusso del metodo proposto

Tramite il processo di feature extraction otteniamo, per ogni pixel, un vettore $n-dimensionale$ che lo identifica nello spazio dei parametri. Per quanto riguarda l'edge detection i risultati proposti in questo lavoro sono stati ottenuti con il set di feature che ha mostrato maggior capacità di modellazione del problema. Nella

fattispecie le feature più adatte sono risultate essere gradiente, entropia e deviazione standard.

	Gradiente	**Entropia**	**Dev. Standard**
Pixel 1	*Gradiente_Pixel_1*	*Entropia_Pixel_1*	*Dev.Standar._Pixel_1*
Pixel 2	*Gradiente_Pixel_2*	*Entropia_Pixel_2*	*Dev.Standard_Pixel_2*
…	…	…	…
…	…	…	…
Pixel n	*Gradiente_Pixel_n*	*Entropia_Pixel_n*	*Dev.Standard_Pixel_n*

Figura 26 - Dataset dopo la feature extraction

Il dataset così ottenuto diventa input per una procedura di clustering. Successivamente a seconda dei parametri richiesti dal metodo utilizzato, può essere necessaria una fase di validazione del clustering ottenuto.

Nel paragrafo precedente sono stati mostrati due indici, Silhouette e Davies-Bouldin, comunemente utilizzati a tale scopo. Misure di questo tipo risultano necessarie soprattutto per quei metodi che richiedano in input il numero di cluster attesi poiché, solitamente, il processo di selezione del valore ottimale per tale parametro si basa su una di queste metriche. Nel caso proposto, non essendoci vincoli dal punto di vista computazionale, la scelta è ricaduta sull'uso della Silhouette, privilegiando quindi una maggior accuratezza. Il valore iniziale del parametro k, numero di cluster attesi, può essere scelto in base ad eventuali conoscenze dei dati in input. Successivamente si può procedere incrementando tale valore ad ogni iterazione e calcolando, al termine di ogni esecuzione dell'algoritmo, l'indice di valutazione selezionando, infine, il risultato per cui tale valore risulta massimo. In alternativa il processo può essere reso più efficiente seguendo un

approccio *greedy* arrestando le iterazioni quando il nuovo valore della Silhouette calcolato è minore di quello precedente. I risultati proposti sono stati ottenuti seguendo questo criterio.

4.2.1 Analisi di mappe di estinzione

Il metodo descritto nel paragrafo precedente è un metodo di natura generale che può essere applicato senza una particolare conoscenza del dominio applicativo. Analizzando però le caratteristiche dell'immagine elaborata ci si può rendere conto di eventuali carenze e tarare il metodo per l'ottenimento di risultati ottimali. La tecnica proposta infatti, di cui la Figura 27 mostra un esempio, soffre di una fondamentale carenza quando applicata a mappe di estinzione.

Figura 27 – Esempio di Edge Detection su una mappa di estinzione (SOM+K-means)

Sebbene infatti vengano identificati degli edge, le feature utilizzate non sono sufficienti per valutare l'estinzione presente all'interno di una determinata regione circoscritta. Ovvia conseguenza di questa valutazione è l'aggiunta di un'ulteriore dimensione allo spazio dei parametri, ovvero l'estinzione stessa. Va considerato però

che tale feature non ha alcuna relazione con pixel appartenenti ad edge, come invece avviene per le altre. Il clustering ottenuto, ed in particolare le metriche utilizzate per valutarlo, potrebbero quindi risentirne e, per tale motivo, è necessario stabilire un diverso criterio di valutazione che permetta inoltre di stabilire un criterio d'arresto per il calcolo del parametro *k* in input ad algoritmi come il K-Means. È stato inoltre verificato che la rimozione di una delle feature utilizzate per l'individuazione degli edge, in particolare la deviazione standard, media questo effetto senza alterare i risultati.

Poiché l'uso di metriche interne, come quelle viste al paragrafo 3.4.4, non assicura comunque l'efficacia del metodo, può essere conveniente stabilire un criterio che si basi direttamente sui risultati ottenuti in relazione al contesto di applicazione. Nella fattispecie, appurato che, come sarà mostrato a breve, l'aggiunta dell'estinzione come feature ed il variare del numero di cluster attesi non influenzino la stabilità degli edge individuati, possiamo indicare un criterio basato sulla distribuzione dei pixel caratterizzati da alta estinzione.

Per esporre più chiaramente il criterio possiamo riferirci alla Figura 28, in cui i pixel facenti parte dei cluster di edge sono mostrati in rosso mentre, i restanti, sono mostrati in una scala di grigi che esprime il valore medio dell'estinzione dei pixel appartenenti al relativo cluster. Pixel neri fanno quindi parte del cluster caratterizzato dal più basso valore medio di estinzione mentre, al contrario, pixel bianchi fanno parte del cluster con estinzione più alta.

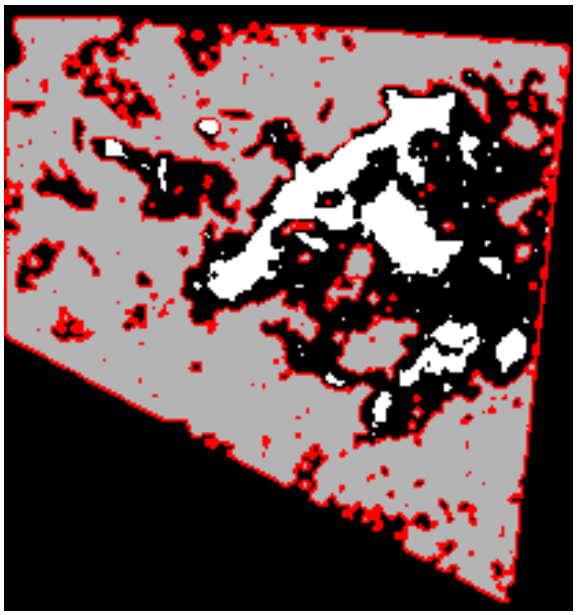


Figura 28 - Edge-cluster e pixel ad alta estinzione (in rosso i pixel di edge)

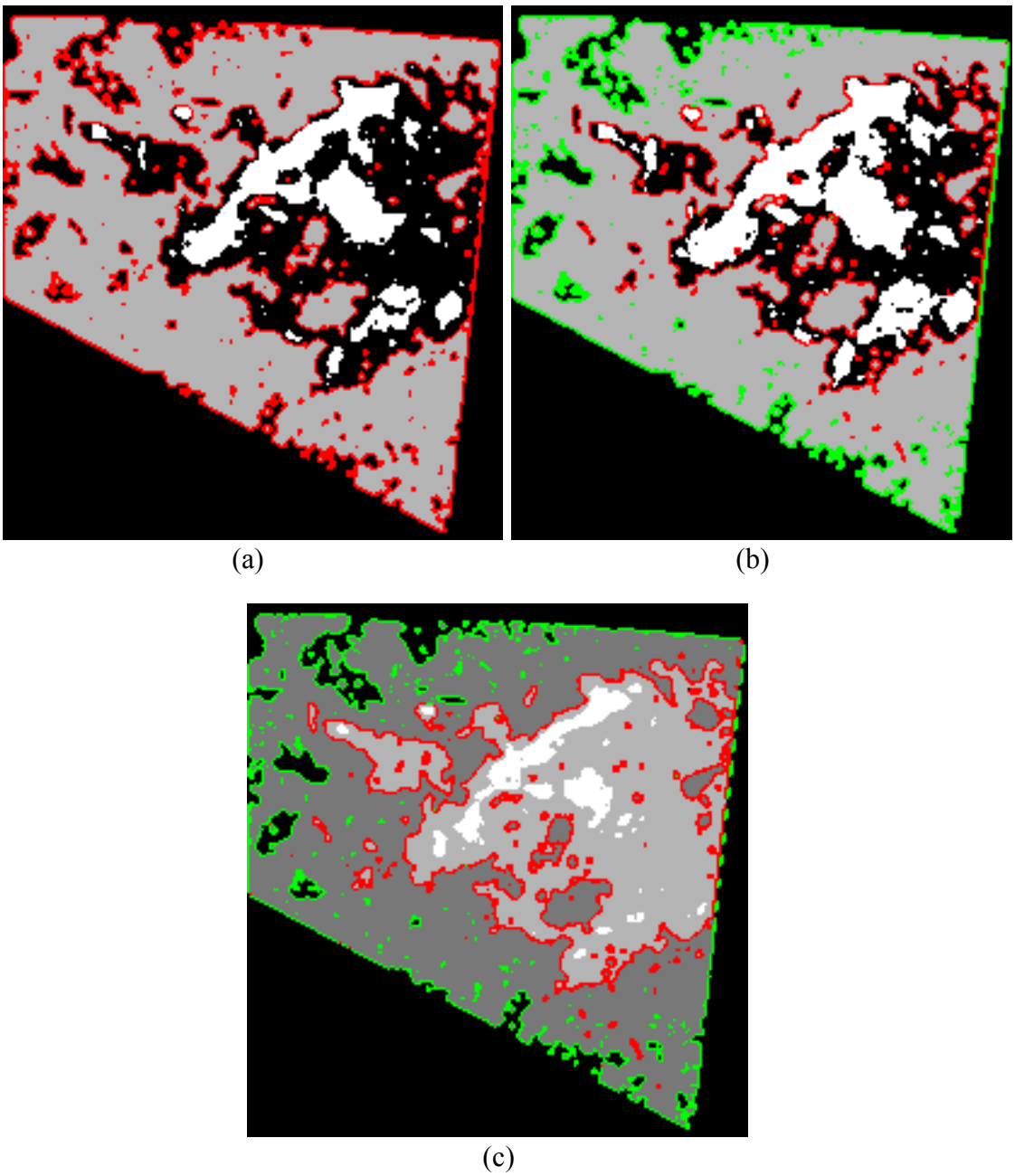
A questo punto un criterio semplicistico per determinare il risultato finale potrebbe essere quello di attribuire alla nube tutte le zone chiuse che abbiano al loro interno almeno un pixel ad alta estinzione, ovvero pixel bianchi. Un metodo più robusto per raffinare il risultato può essere però ottenuto definendo un indice di contaminazione delle regioni[2] (*Region Contamination Index – RCI)* in grado di esprimere il grado di coesistenza, in una stessa regione, tra pixel ad alta e bassa estinzione, rispettivamente P_H e P_L (Equazione 16).

$$RCI = 1 - \frac{|P_H - P_L|}{P_H + P_L} \tag{16}$$

Tale indice assume valore 0 quando non esistono regioni caratterizzate dalla co-presenza di pixel bianchi (alta estinzione) e neri (bassa estinzione), mentre assume valore massimo, cioè 1, quando questi sono presenti in egual numero. Un risultato di

[2] Indice definito solo per le regioni che presentano al loro interno almeno un pixel ad alta o bassa estinzione

clustering può quindi ritenersi soddisfacente quanto più l'RCI risulta basso (Figura 29).



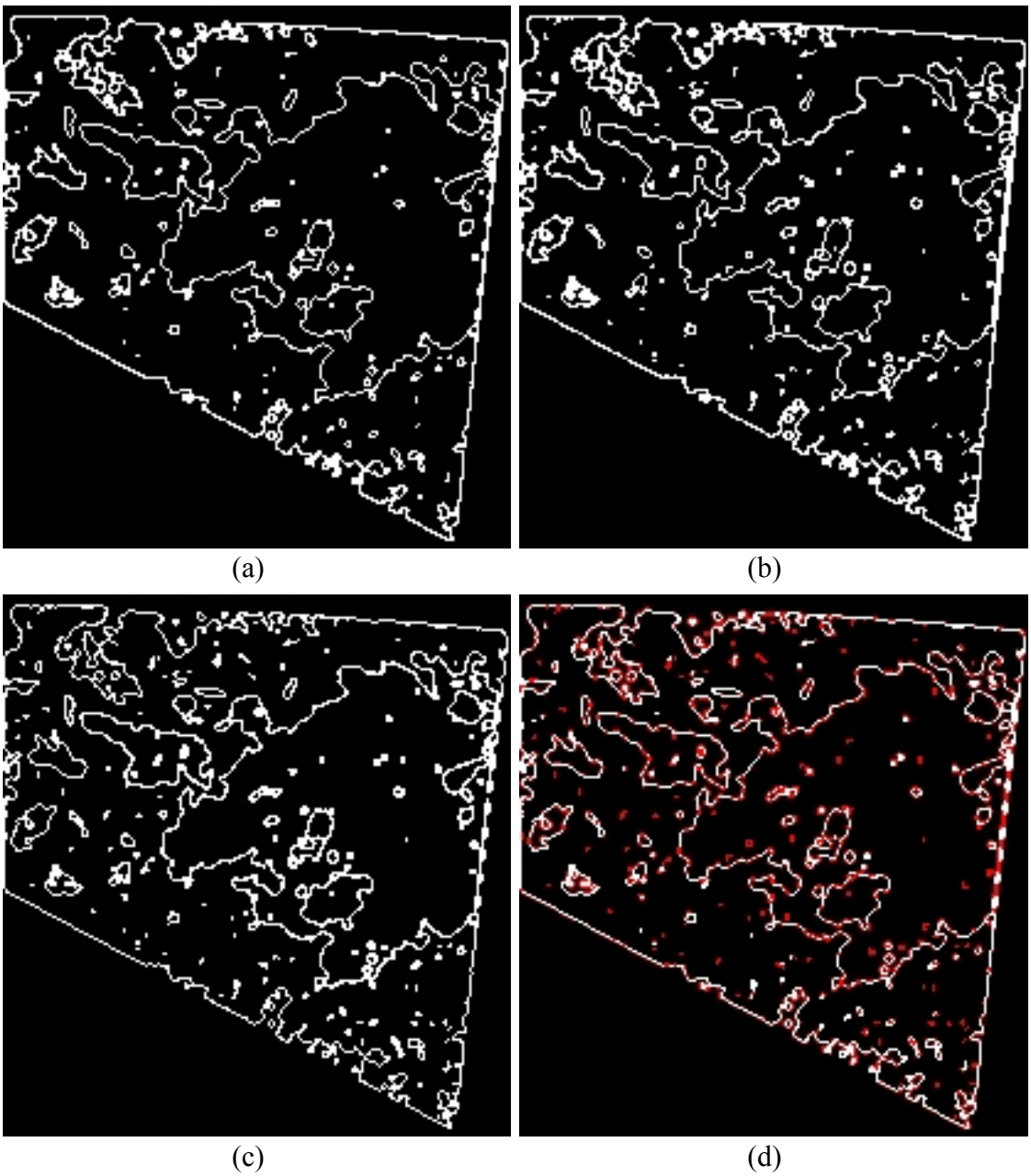
(a) (b)

(c)

Figura 29 - Criterio per un corretto clustering della mappa di estinzione. Iterazioni per k=4 (a), k=5 (b) e k=6 (c) con RCI rispettivamente pari a 0.57, 0.81 e 0

Mostriamo inoltre quanto precedentemente affermato, ovvero che nonostante l'aggiunta di un parametro non correlato all'individuazione degli edge, quale

l'estinzione, nonché al variare del numero di cluster, gli edge individuati mantengono un elevato grado di stabilità.



(a) (b)

(c) (d)

Figura 30 - Stabilità degli edge al variare del numero di cluster. Iterazioni per k=4 (a), k=5 (b) e k=6 (c) e confronto (d) in cui pixel bianchi sono individuati in tutti e tre i casi precedenti.

In Figura 30 sono mostrati, da sinistra verso destra, gli edge ottenuti con un numero di cluster atteso pari a 4, 5 e 6, numero per il quale il criterio d'arresto viene raggiunto, come mostrato in Figura 29(c). Nell'ultima immagine i pixel di edge sono

mostrati in bianco se riconosciuti come tali in tutti e tre i casi, altrimenti in rosso. Un'analisi visiva ci suggerisce come permangano in bianco tutte le principali strutture individuate mentre i pixel rossi siano per la maggior parte pixel spuri. Volendo analizzare numericamente il risultato possiamo affermare che il 54% dei pixel di edge viene riconosciuto in tutti i casi e la percentuale sale al 95% se consideriamo anche i pixel che vengono riconosciuti in due casi su 3.

Dal risultato finale è infine possibile ricavare una maschera binaria che identifica la nube di estinzione. L'estrazione di tale maschera avviene sulla base dell'estinzione media, indicata dal livello di grigio, dei cluster interni alle regioni delimitate da edge. In Figura 31, ad esempio, è mostrata una maschera ricavata prendendo in considerazione le regioni al cui interno ricadano pixel appartenenti ai due cluster con estinzione media più elevata.

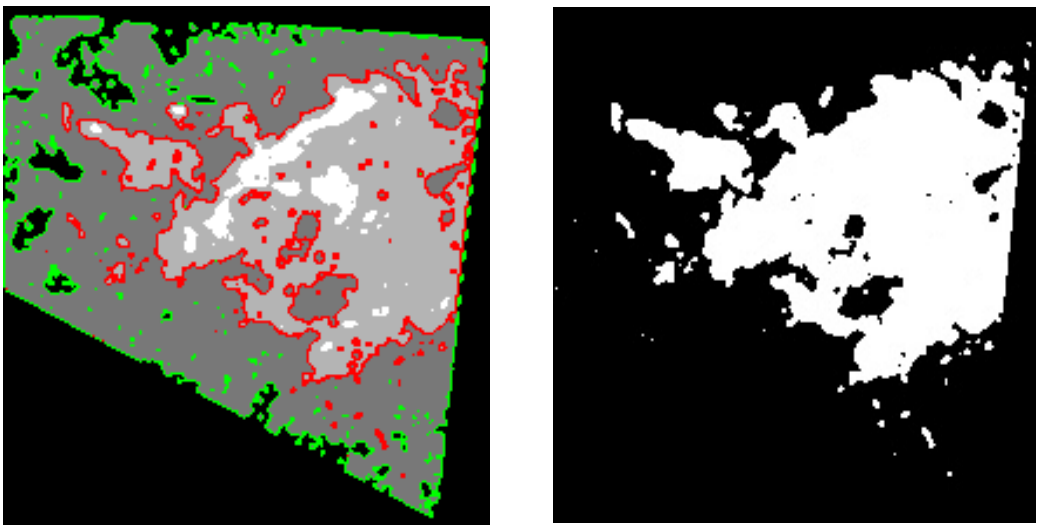


Figura 31 - Maschera binaria della nube di estinzione ricavata considerando le regioni delimitate da edge in cui sono presenti pixel appartenenti ai due cluster con valore medio di estinzione (livello di grigio) più elevato

In realtà dal risultato del metodo proposto è possibile estrarre anche indicazioni aggiuntive. Nel calcolo della contaminazione, su cui si basa il criterio di validazione,

è infatti presa in considerazione solo la coesistenza di pixel bianchi e neri, relativi ai cluster con estinzione media, rispettivamente massima e minima. Per cui non esiste vincolo che limiti la coesistenza di altre differenti gradazioni all'interno della medesima regione. Queste differenti gradazioni possono quindi essere estratte e prese in considerazione individualmente, come mostrato in Figura 32.

Figura 32 - Maschera di una regione interna ad alta estinzione

5 Analisi dei risultati

L'analisi delle performance qualitative dei metodi selezionati è stata condotta attraverso due fasi principali. Inizialmente abbiamo utilizzato un set di immagini sintetiche, facilmente reperibili su internet, con le quali siamo stati in grado di confermare la scelta dei metodi e di ridurre lo spazio dei parametri utile sia all'edge detection che alla segmentazione tramite clustering. Successivamente, abbiamo applicato i metodi ad immagini astronomiche reali, relative sia a osservazioni dirette,

sia a mappe di estinzione di regioni di cielo, derivate dalle immagini osservate mediante applicazione di metodi statistici.

In sintesi, gli esperimenti condotti sono stati organizzati secondo la procedura seguente:

- ***analisi dei metodi e dello spazio dei parametri su immagini sintetiche***: edge detection e segmentazione tramite clustering su immagini sintetiche, al fine di caratterizzare i metodi proposti e lo spazio dei parametri;
- ***analisi di mappe d'estinzione***: edge detection e segmentazione tramite clustering su immagini reali relative a mappe d'estinzione calcolate da un metodo tradizionalmente impiegato in Astrofisica, al fine di analizzare e ottimizzare lo spazio dei parametri confrontandoci con i risultati noti in letteratura relativi al calcolo dell'estinzione;
- ***estrazione di mappe d'estinzione***: edge detection e segmentazione tramite clustering su immagini reali direttamente osservate, al fine di analizzare la capacità dei metodi proposti e dello spazio dei parametri identificato per l'estrapolazione delle regioni attribuite al contributo di estinzione.

Come ragionevole attendersi, gli esperimenti su immagini reali hanno confermato la necessità di dotarsi di uno spazio di parametri più complesso, includendo features statistiche, sufficienti ai fini dell'edge detection, nonché direttamente derivate dai dati per la segmentazione, evidenziando buone prestazioni in entrambi i casi. In particolare, siamo riusciti ad ottenere degli edge che ben definiscono zone circoscritte della nube di estinzione, laddove alcuni tra i più noti algoritmi

tradizionali di edge detection, nonché algoritmi innovativi basati su logica Fuzzy, falliscono.

Sono state inoltre fin da subito chiare le potenzialità di un approccio basato su una parametrizzazione auto-adattiva dei pixel rispetto ad approcci basati su una caratterizzazione statica, potenzialità che si traducono principalmente nel vantaggio di non dover imporre livelli di soglia arbitrari.

Questa considerazione ci ha permesso di valutare una tecnica capace di aumentare l'informazione estraibile dal risultato ottenendo, in maniera automatica, non solo dei confini della nube, ma anche delle regioni interne che si distinguono per livelli di estinzione caratteristici. I risultati, valutati sia dal punto visuale che statistico, sono perfettamente in linea con quelli ottenuti tramite tecniche tradizionali, confermando quindi l'efficacia del metodo automatico nell'ottenere un risultato quantomeno comparabile, senza fare ricorso ad intervento manuale.

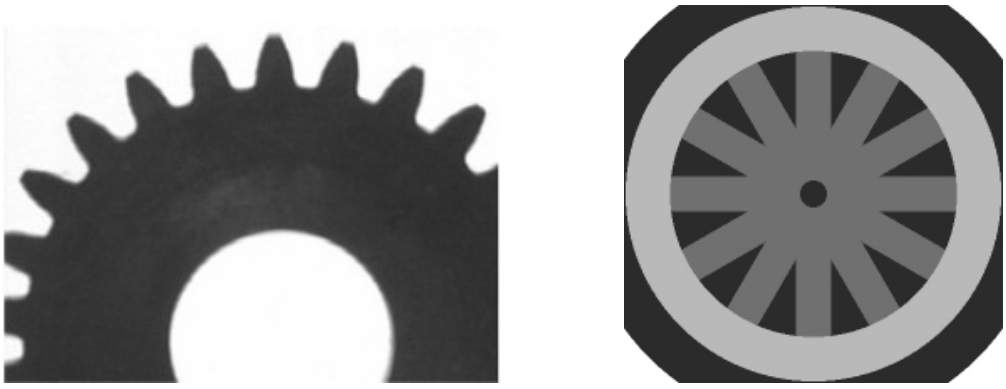
Va sottolineato che la maschera della nube di estinzione ottenuta tramite il nostro metodo non è, nella maggior parte dei casi, ottenibile tramite sogliatura del livello di estinzione. Ciò vuol dire che, qualora successivi studi approfonditi riscontrassero un miglioramento dei risultati, il metodo proposto si dimostrerebbe non solo più performante dal punto di vista computazionale, ma anche necessario in quanto in grado di fornire un risultato altrimenti non ottenibile.

Poiché il risultato ottenuto è strettamente correlato allo spazio dei parametri scelto, abbiamo investigato, successivamente, la possibilità di intervenire su quest'ultimo affinché il metodo fosse in grado di elaborare una mappa di estinzione direttamente dall'immagine osservata. Ciò coniugando quindi il duplice obiettivo

scientifico di raffinamento delle mappe di estinzione e di rivelazione dell'estinzione stessa. Quest'ultimo ottenibile attraverso osservazioni ad alta risoluzione nella banda B dell'infrarosso.

5.1 Analisi su immagini sintetiche

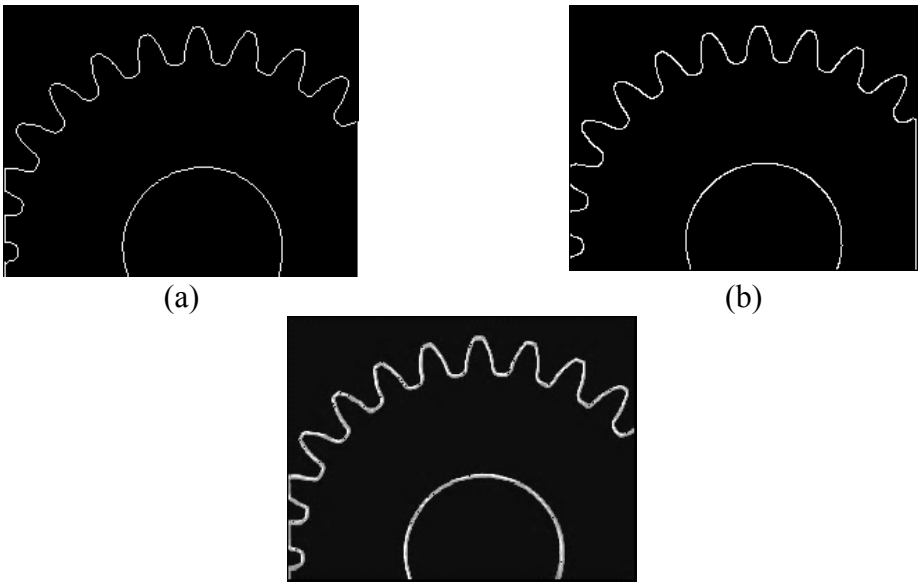
Gli esperimenti sono stati condotti sulle immagini mostrate in Figura 33. Trattandosi di edge detection, lo spazio dei parametri utilizzato è quello descritto in Figura 26.



(a) (b)

Figura 33 - Immagini generiche utilizzate per la validazione del metodo

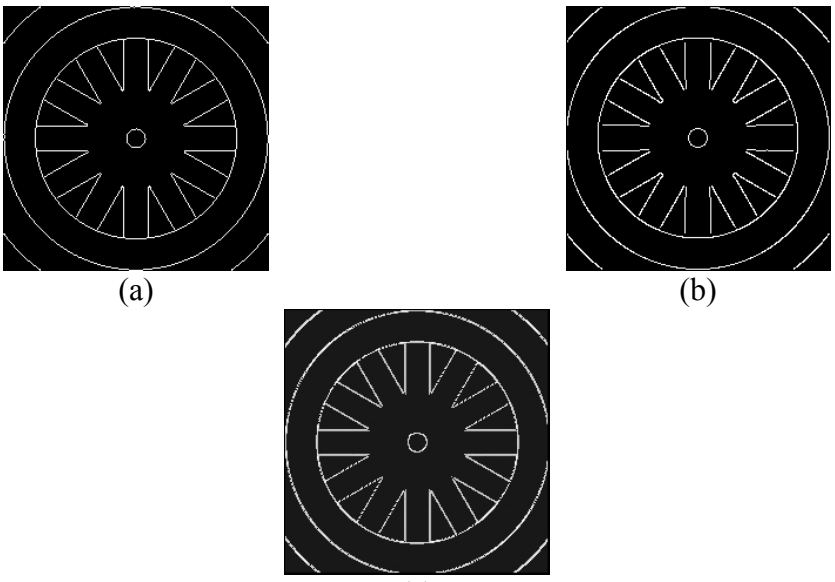
Sono immagini di carattere generale ed i metodi tradizionali di edge detection sono correttamente in grado di individuarne i contorni, sebbene non sempre risultino edge chiusi e siano soggetti, in taluni casi, ad un settaggio manuale dei parametri. In Figura 34 e Figura 35 sono proposti, da sinistra verso destra, i risultati relativi, rispettivamente, agli operatori di Sobel, metodo di Canny e metodo Fuzzy. Per quanto riguarda la scelta dei parametri, negli operatori di Sobel, ove l'utilizzo di una soglia euristica, come di default implementato in MATLAB, non sia risultata soddisfacente, questa è stata settata in modo manuale. Per Canny invece, la soglia euristica è stata sostituita, quando necessario, con la soglia di Otsu, utilizzata come descritto al paragrafo 3.2.1.



(a) (b)

(c)

Figura 34 - Risultati dell'utilizzo degli algoritmi di edge detection sulla prima immagine generica con operatori di Sobel (a), metodo di Canny (b) e metodo Fuzzy (c)



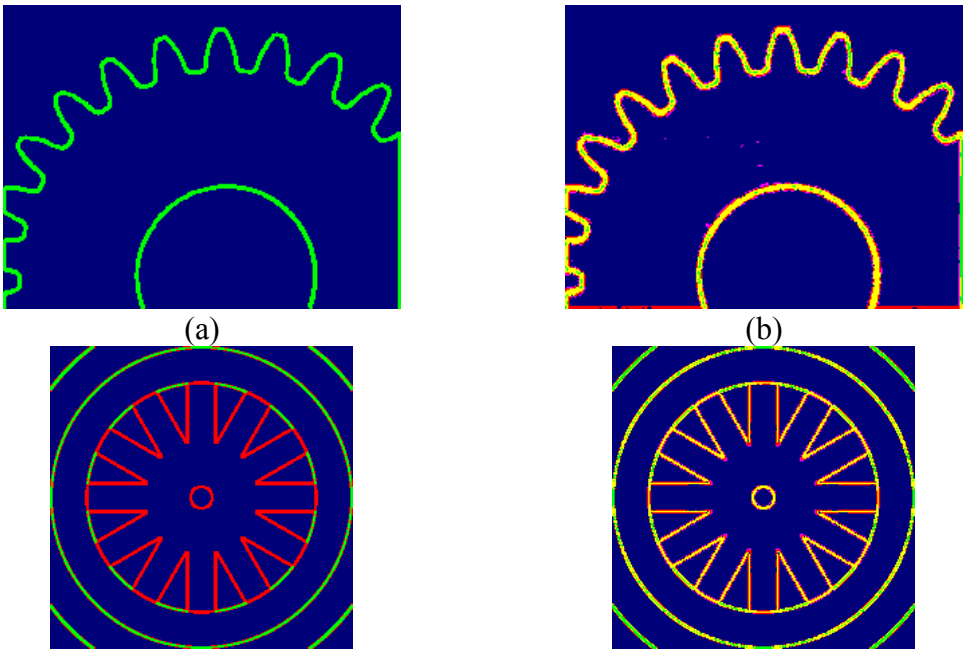
(a) (b)

(c)

Figura 35 - Risultati dell'utilizzo degli algoritmi di edge detection sulla seconda immagine generica con operatori di Sobel (a), metodo di Canny (b) e metodo Fuzzy (c)

Successivamente, lo stesso esperimento di edge detection è stato eseguito addestrando una rete SOM sull'immagine di Figura 33(a), la cui mappa di Kohonen risultante è stata processata con le due tecniche proposte, K-Means e UMAT-CC, con i risultati mostrati in Figura 37(a) e Figura 37(b). L'estrazione degli edge è stata

eseguita in base all'identificazione dei cluster con numero medio di oggetti più basso. La Figura 36 mostra invece il risultato della fase intermedia di clustering da cui effettuare l'estrazione degli edge.



(a) (b)

Figura 36 - Risultati del clustering applicato alle immagini sintetiche, utilizzando K-Means (a) e UMAT-CC (b)

(a) (b)

Figura 37 - Risultati dell'edge detection con il metodo proposto sulla prima immagine generica, utilizzando K-Means (a) e UMAT-CC (b)

La medesima rete, non riaddestrata, è stata quindi utilizzata anche sulla seconda immagine, Figura 33(b), ed a seguito del post-processing con i due metodi sopracitati, si ottengono i risultati mostrati in Figura 38.

(a) (b)

Figura 38 - Risultati dell'edge detection con il metodo proposto sulla seconda immagine generica, utilizzando K-Means (a) e UMAT-CC (b)

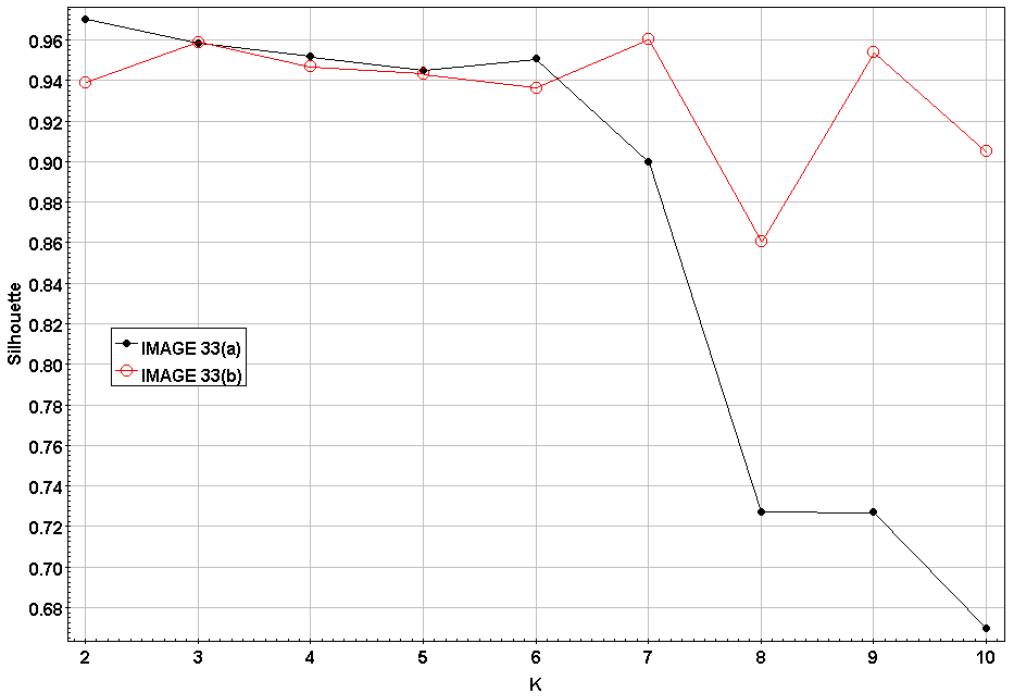


Figura 39 – Scelta del valore di K per il K-means in base al metodo Silhoulette nelle 2 immagini sintetiche utilizzate (vedi Figura 33). I valori scelti in base alla tecnica greedy sono K=2 per l'immagine 33(a) e K=3 per l'immagine 33(b).

In primo luogo è possibile constatare che i risultati sono soddisfacenti in tutti i test condotti. Inoltre, l'aver ottenuto buoni risultati sull'immagine Figura 33(b), avendo utilizzato una rete addestrata su un'immagine differente, Figura 33(a), conferma l'effettiva capacità dello spazio dei parametri selezionato di modellare il problema generale di individuazione di edge.

Va sottolineato però che tali feature non si distribuiscono in maniera binaria, cioè con valori elevati per i pixel di edge e valori nettamente più bassi nei pixel circostanti, ma tendono piuttosto a diminuire gradualmente quanto più ci si allontana dall'edge stesso. Questo comportamento si riflette sulla distribuzione dei nodi della mappa di Kohonen (Figura 40), che non sarà ovunque caratterizzata da nette zone separate e, in generale, ciò si traduce con l'identificazione di edge più spessi. Inoltre, metodi come l'UMAT-CC, sensibili alla struttura della mappa stessa, tendono a risentirne maggiormente, come evidenziato dai risultati lievemente peggiori confrontando fra loro le 2 immagini di Figura 37 e Figura 38.

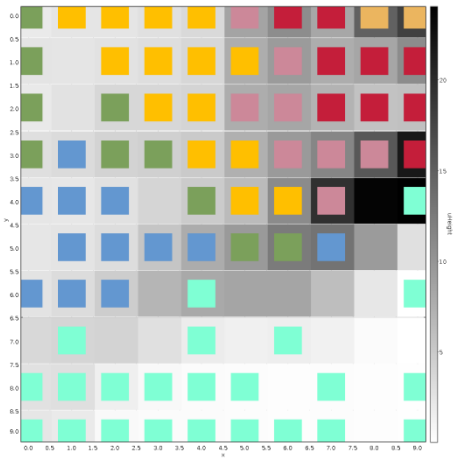


Figura 40 - Nella U-Matrix la distribuzione dei nodi, con gradazioni di grigio, non evidenzia zone nettamente separate se non nei pixel di background i cui relativi nodi sono collocati nella parte inferiore

Questi risultati evidenziano che lo spazio dei parametri scelto, riferito all'edge detection, si è rivelato efficace in senso generale, quantomeno sulle immagini sintetiche. Inoltre il metodo basato su SOM+K-means ha fornito il miglior risultato rispetto all'alternativa SOM+UMAT-CC.

5.2 Analisi di mappe di estinzione

Come anticipato in precedenza, questo paragrafo è dedicato all'analisi dei risultati del miglior metodo metodo proposto (SOM+K-means) applicato a immagini astronomiche, con particolare riferimento alle mappe di estinzione. Ossia, applichiamo il metodo a mappe d'estinzione, ottenute mediante un metodo tradizionale in Astrofisica su immagini astronomiche osservate.

Prendendo come riferimento le mappe d'estinzione di alcune sotto-regioni della nebulosa di Lupus (Figura 3), abbiamo confrontato le prestazioni di edge detection tra il nostro metodo proposto e i 3 metodi di Sobel, Canny e basato su fuzzy logic (Figura 41, Figura 42, Figura 43). In particolare abbiamo anche effettuato il pre-processing basato sul miglioramento del contrasto nella mappa d'estinzione originaria (come descritto nel paragrafo 4.1.1).

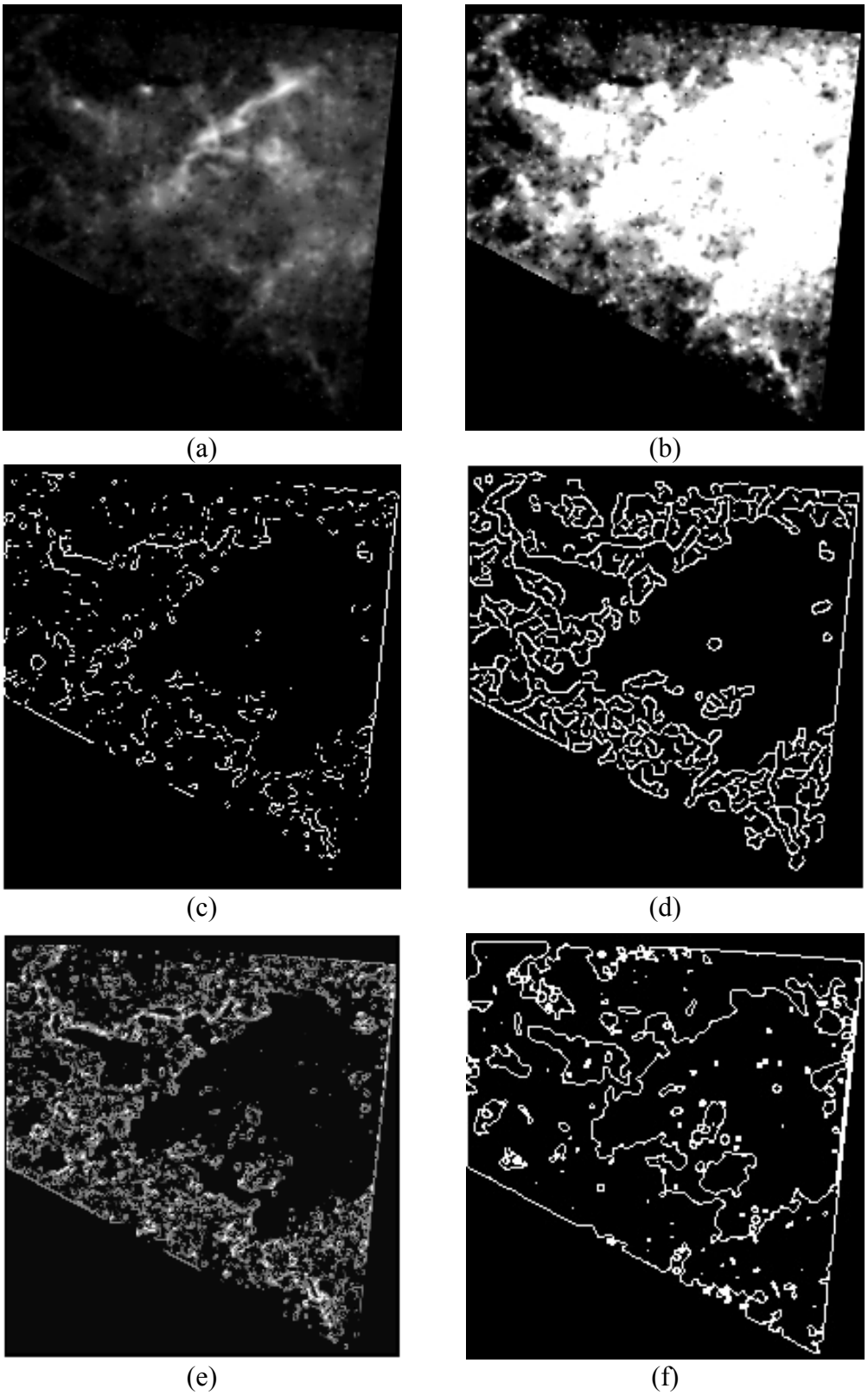


Figura 41 - Mappa di estinzione della regione Lupus I (a) con contrasto migliorato (b) ed il risultato di edge detection ottenuto tramite Sobel (c), Canny (d), metodo Fuzzy (e) e metodo proposto (f)

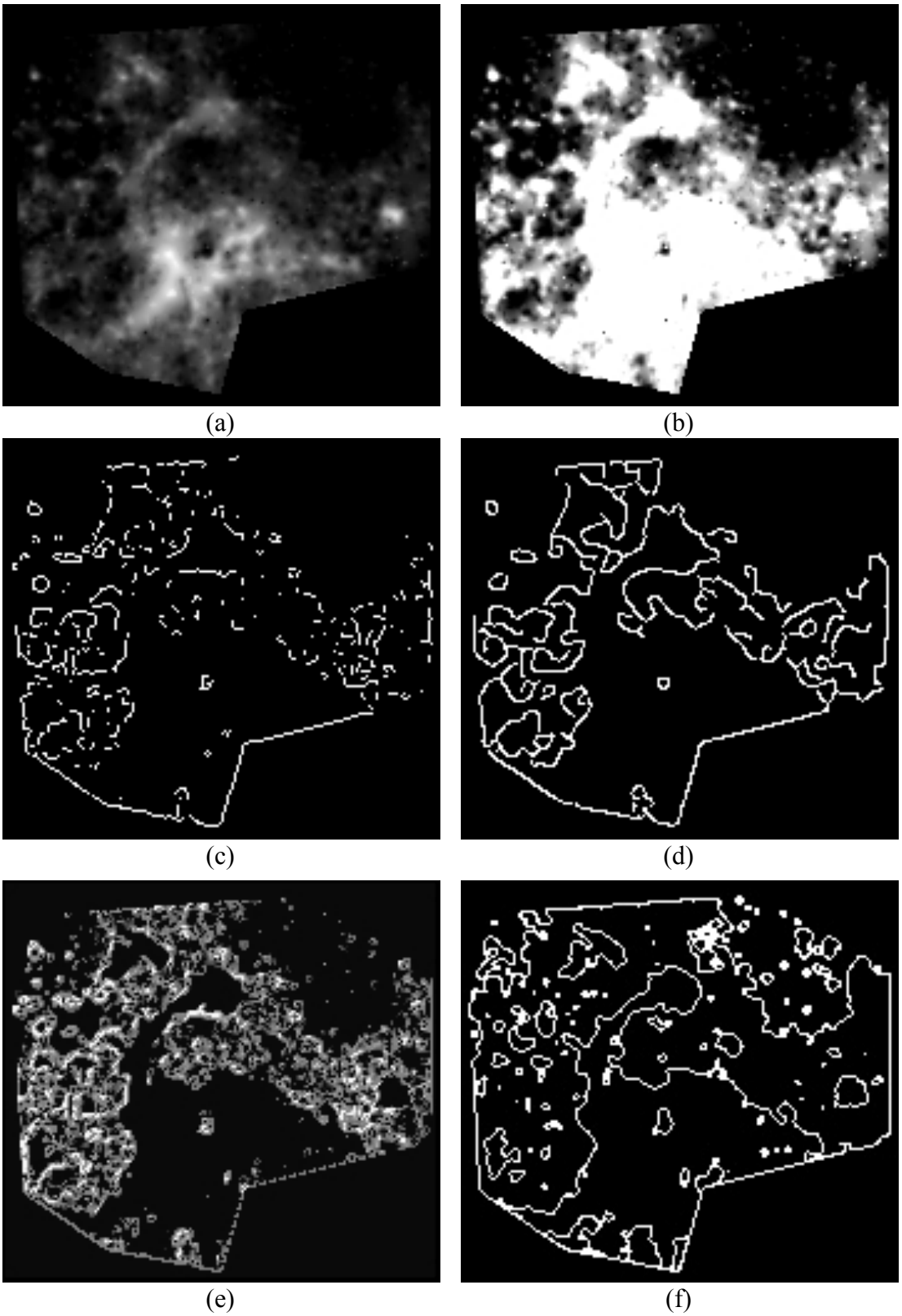


Figura 42 - Mappa di estinzione della regione Lupus V (a) con contrasto migliorato (b) ed il risultato di edge detection ottenuto tramite Sobel (c), Canny (d), metodo Fuzzy (e) e metodo proposto (f)

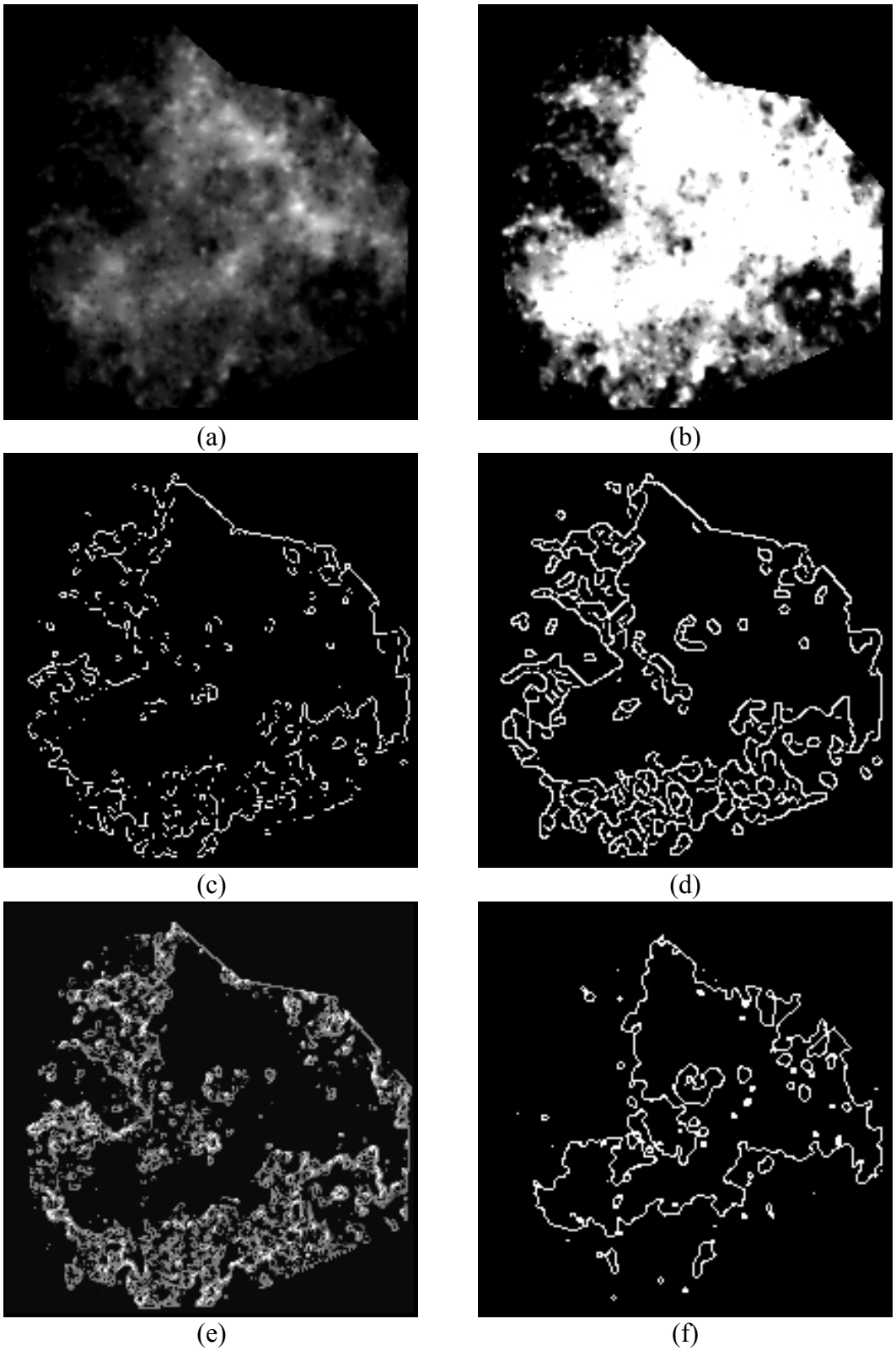


Figura 43 - Mappa di estinzione della regione Lupus VI (a) con contrasto migliorato (b) ed il risultato di edge detection ottenuto tramite Sobel (c), Canny (d), metodo Fuzzy (e) e metodo proposto (f)

Il pre-processing dell'immagine causa inevitabilmente un effetto fuorviante ai fini dell'edge detection, in quanto le zone caratterizzate da alta estinzione risultano più omogenee, eliminando quindi la possibilità di individuare erroneamente edge al loro interno. Tuttavia nella restante porzione di immagine i metodi presi come termine di paragone presentano ancora risultati non soddisfacenti, presentando un numero eccessivo di edge che non formano altrettante regioni chiuse. Al contrario, rispetto agli altri metodi analizzati, gli edge ottenuti tramiti il metodo SOM-K-means risultano in grado di delimitare delle precise regioni circoscritte.

Come successivo esperimento, a seguito della sostituzione nello spazio dei parametri, introdotta nel paragrafo 4.2.1, abbiamo sperimentato l'uso del valore del pixel come ulteriore feature al posto della standard deviation. Nel caso specifico abbiamo dunque sostituito tra i parametri il valore dell'estinzione al fine di ottenere un potenziamento espressivo dei risultati ottenuti.

A seguito dell'applicazione del metodo di clustering SOM+K-means, in parallelo rispetto al puro edge detection, abbiamo focalizzato l'attenzione anche alla possibilità di isolare le regioni più dense in termini di estinzione dal resto dell'immagine. Ossia ottenere una maschera della nube di estinzione (Figura 44, Figura 45 e Figura 46) nonché sue eventuali sotto-regioni caratteristiche (Figura 47). Queste ultime sono caratterizzate da un'evidente differenza rispetto all'area circostante, in termini di densità di estinzione, sebbene non circoscrivibile da un edge ben definito.

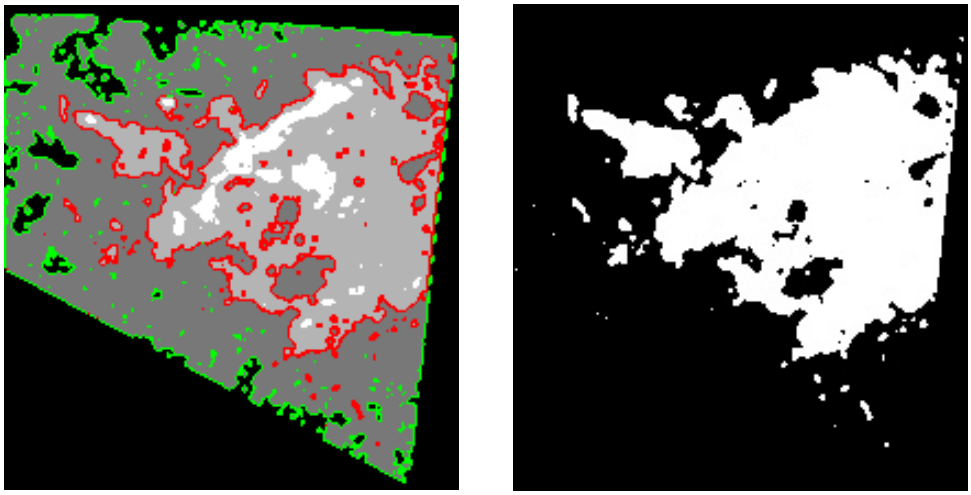


Figura 44 - Clustering della mappa di estinzione della regione Lupus I (a sinistra) e relativa maschera della nube di estinzione (a destra)

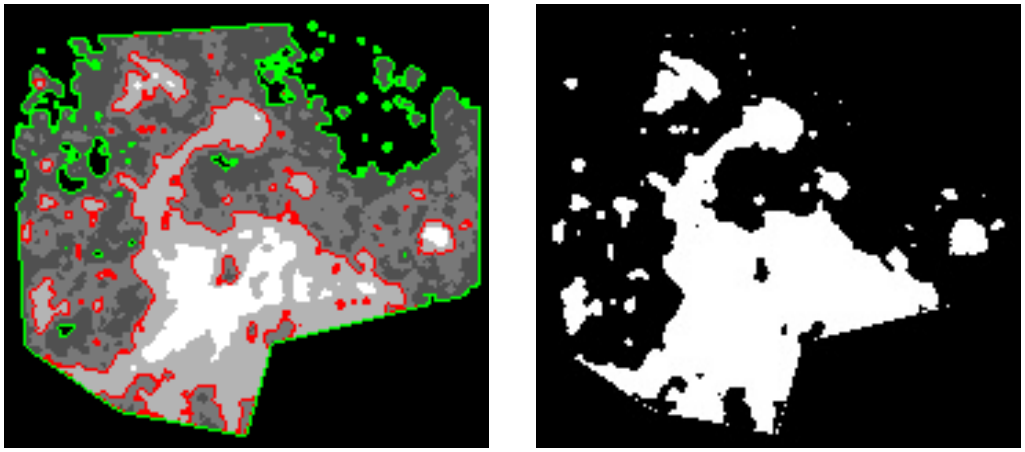


Figura 45 - Clustering della mappa di estinzione della regione Lupus V (a sinistra) e relativa maschera della nube di estinzione (a destra)

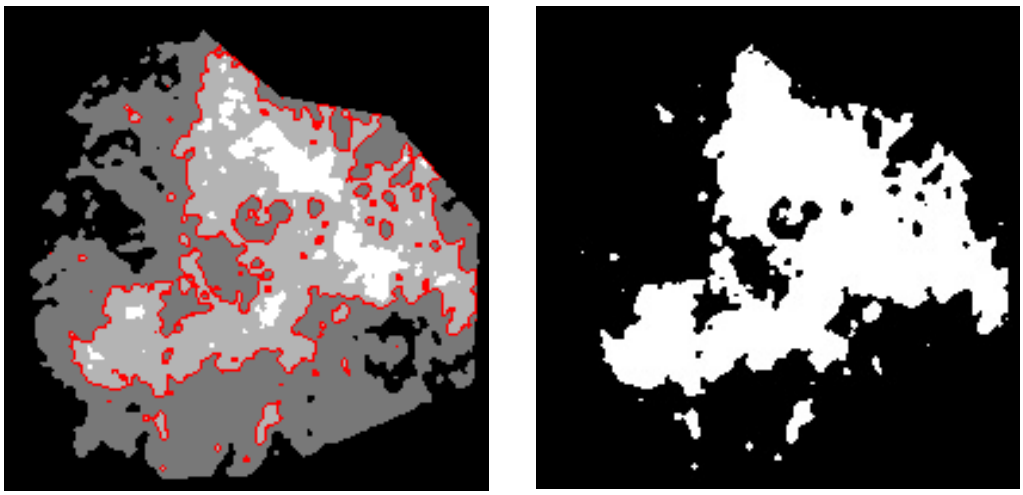
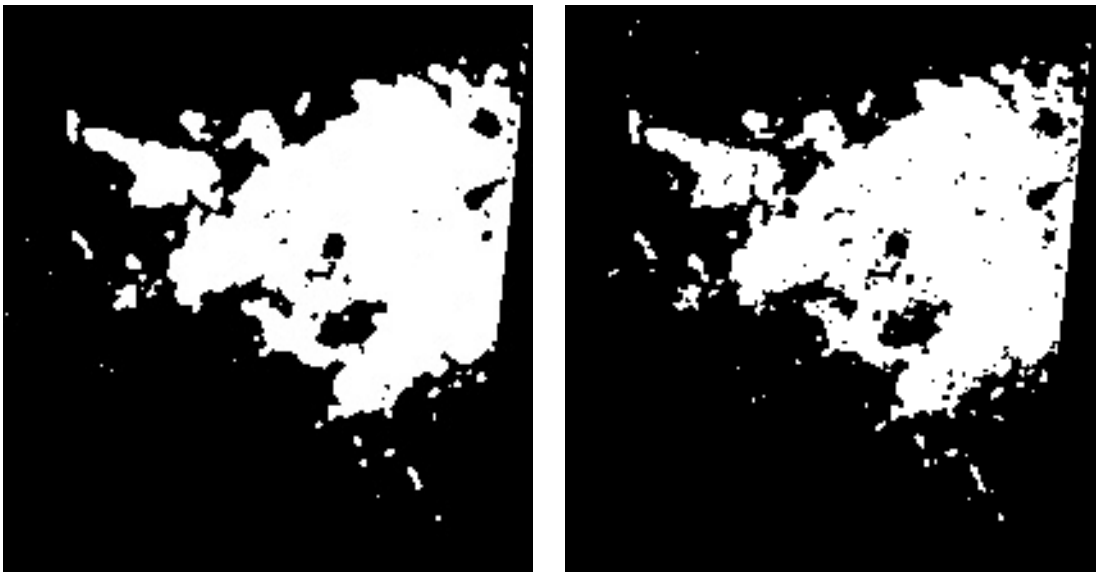


Figura 46 - Clustering della mappa di estinzione della regione Lupus VI (a sinistra) e relativa maschera della nube di estinzione (a destra)

(a) (b) (c)

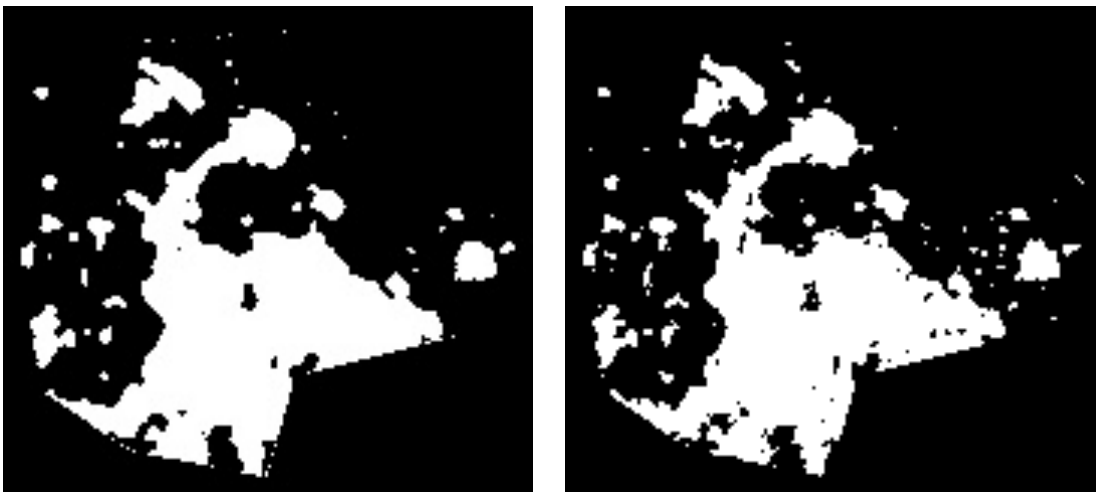
Figura 47 - Sotto-regioni nelle nubi di estinzione di Lupus I (a), Lupus V (b) e Lupus VI (c)

Le maschere così estratte possono essere confrontate con le maschere derivanti dall'applicazione del metodo astrofisico tradizionale. Tale confronto non può però essere completamente oggettivo poiché, come descritto nel capitolo 2, tradizionalmente i margini della nube di estinzione sono ottenuti con una sogliatura arbitraria. Per i confronti proposti tale soglia sarà settata in modo tale da avere la minor percentuale possibile di pixel non in comune con la maschera ottenuta tramite clustering, facendo in modo, cioè, che le due maschere, risultino più simili possibile. Tale confronto è mostrato nelle figure successive (da Figura 48 a Figura 53).



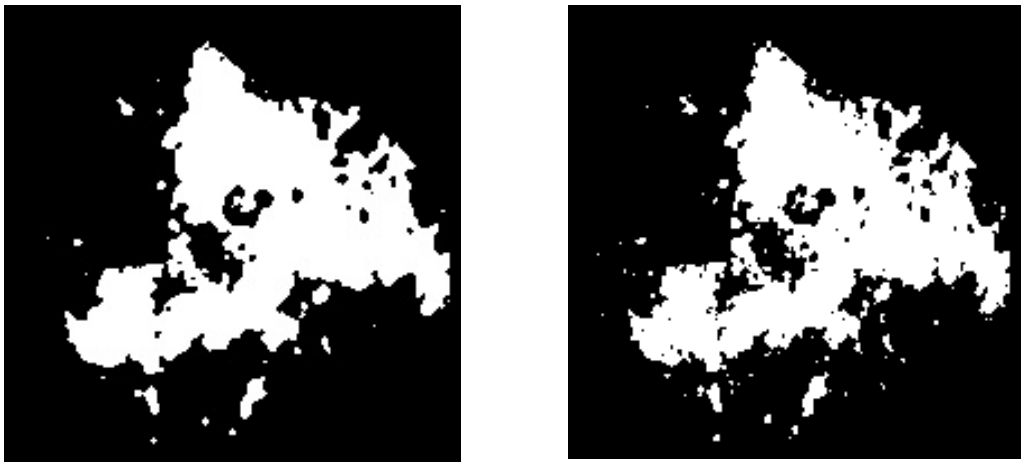
(a) (b)

Figura 48 - Maschere della nube di estinzione Lupus I (a) SOM+K-means, (b) sogliatura



(a) (b)

Figura 49 - Maschere della nube di estinzione Lupus V (a) SOM+K-means, (b) sogliatura

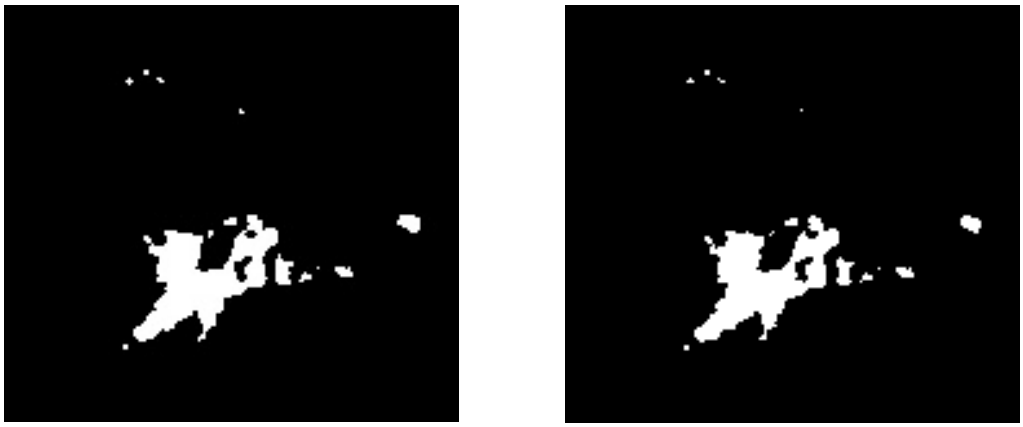


(a) (b)

Figura 50 - Maschere della nube di estinzione Lupus VI (a) SOM+K-means, (b) sogliatura

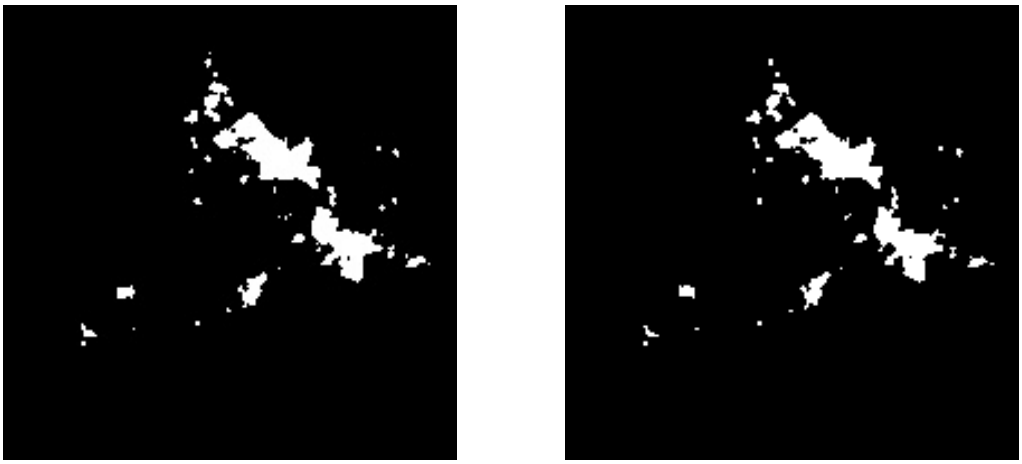
(a) (b)

Figura 51 - Maschere della sotto-regione interna di Lupus I (a) SOM+K-means, (b) sogliatura



(a) (b)

Figura 52 - Maschere della sotto-regione interna di Lupus V (a) SOM+K-means, (b) sogliatura



(a) (b)

Figura 53 - Maschere della sotto-regione interna di Lupus VI (a) SOM+K-means, (b) sogliatura

Le immagini seguenti mostrano invece il dettaglio dei pixel aggiunti e rimossi tra le maschere ottenute dai due metodi a confronto. In particolare i pixel appartenenti esclusivamente alla maschera ottenuta tramite clustering sono mostrati in verde, viceversa i pixel appartenenti solo alla maschera risultante dalla sogliatura sono mostrati in rosso (da Figura 54 a Figura 59).

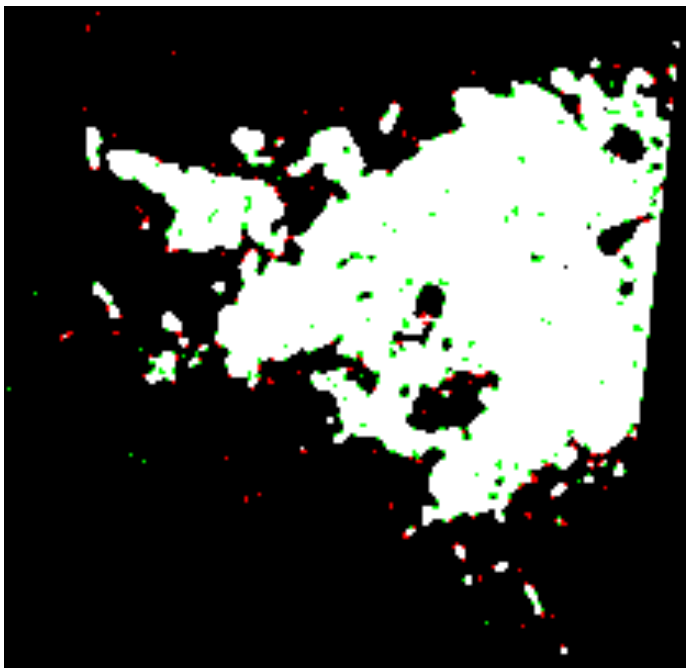


Figura 54 - Confronto tra maschere della nube di estinzione della regione di Lupus I. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

Figura 55 - Confronto tra maschere della nube di estinzione della regione di Lupus V. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

Figura 56 - Confronto tra maschere della nube di estinzione della regione di Lupus VI. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

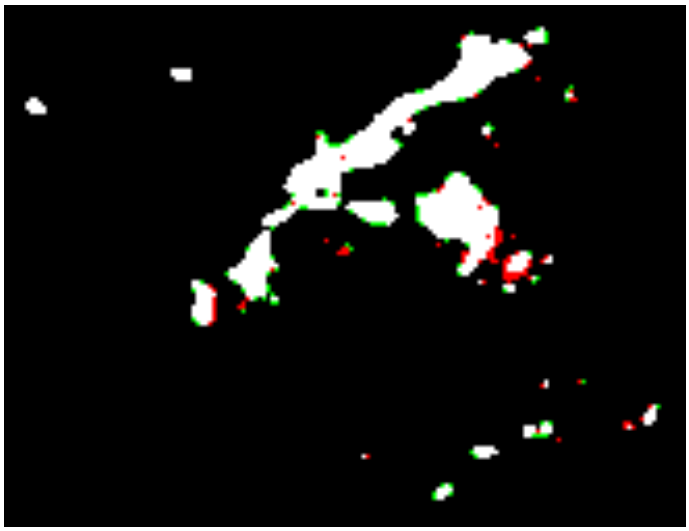


Figura 57 - Confronto tra maschere della nube di estinzione della sotto-regione di Lupus I. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

Figura 58 - Confronto tra maschere della nube di estinzione della sotto-regione di Lupus V. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

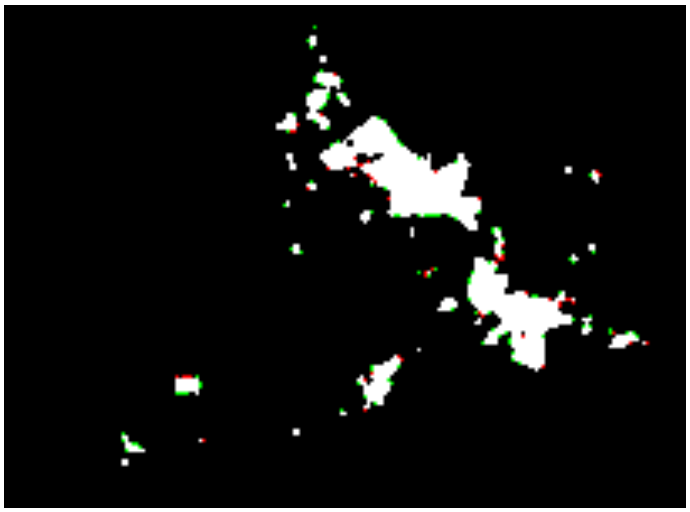


Figura 59 - Confronto tra maschere della nube di estinzione della sotto-regione di Lupus VI. In evidenza i pixel aggiunti (verde) e rimossi (rosso) tramite il metodo proposto

Come accennato nel capitolo 2, la mancanza di una *groundtruth* non permette di quantificare numericamente eventuali miglioramenti apportati all'identificazione dell'estinzione rispetto ai metodi tradizionali. È tuttavia possibile, alla luce dei risultati ottenuti, effettuare alcune considerazioni in base alle quali ipotizzare l'efficacia del metodo descritto.

Un aspetto che caratterizza il nostro risultato è quello di non essere assimilabile ad una sogliatura del livello di estinzione. I differenti criteri che regolano i meccanismi di discriminazione dei pixel appartenenti alla nube permettono di includerli, o escluderli, non solo in base al valore di flusso, ma anche in base a caratteristiche come la densità dell'area circostante.

Si noti infatti come molti pixel rossi, cioè che a differenza della sogliatura il nostro metodo non etichetta come appartenenti alla nube, siano pixel spuri. Sono cioè distribuiti in maniera isolata e lontani dalla regione ad alta estinzione, ragion per cui è ragionevole ipotizzare che non ne facciano parte (Figura 60).

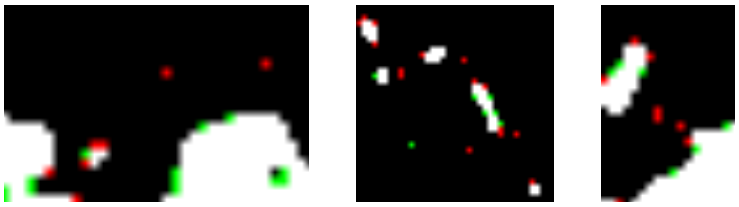


Figura 60 - Esempi di pixel spuri rimossi dalla nube di Lupus I

Per lo stesso motivo il metodo proposto è inoltre in grado di mantenere separate strutture che una sogliatura deve necessariamente includere (Figura 61).

Figura 61 – Altro esempio di pixel spuri in Lupus I. I pixel rossi aggreganti due sotto-strutture mostrano come una maschera ottenuta tramite un livello di sogliatura sia impossibilitata ad effettuare una discriminazione in grado di separarle.

Inoltre i pixel verdi e rossi che si collocano lungo gli edge non sono distribuiti uniformemente, non causando quindi un preciso allargamento o restringimento della regione ad alta estinzione, cosa che avrebbe indotto un malfunzionamento del metodo proposto.

Le considerazioni fin qui fatte sono valide anche per il confronto tra le maschere delle sotto-regioni, mostrate in Figura 57, Figura 58 e Figura 59. Queste aree sono però caratterizzate dall'avere un valore di estinzione nettamente più elevato rispetto a quello riscontrabile ai confini della nube e sono immerse in zone con flusso molto più omogeneo. In questa situazione il valore dell'estinzione diventa la feature dominante che contraddistingue l'area, delimitando quindi una regione molto più simile a quella ottenibile tramite sogliatura.

In Tabella 2 sono riportati i valori relativi alla massa della nube calcolati, come descritto nel capitolo 2, utilizzando i parametri proposti in (Cambresy 1999) ed (Alcalà 2016), rispettivamente M_C ed M_A. Il parametro A_V indica invece il livello di estinzione utilizzato per la sogliatura.

	M_C *(Clustering vs thresholding)*	M_A *(Clustering vs thresholding)*	A_V *(thresholding)*
Lupus I	$\sim +1.09\%$	$\sim +1.10\%$	0.98
Lupus V	$\sim +0.15\%$	$\sim +0.15\%$	0.85
Lupus VI	$\sim +0.7\%$	$\sim +0.71\%$	1.01
Lupus I (sub-reg)	$\sim +0.32\%$	$\sim +0.32\%$	2.5
Lupus V (sub-reg)	$\sim +2.28\%$	$\sim +2.30\%$	2.5
Lupus VI (sub-reg)	$\sim +2.97\%$	$\sim +2.96\%$	2.3

Tabella 2 – Confronto in percentuale di masse solari tra le regioni d'estinzione estratte con metodo proposto (Clustering) e metodo tradizionale (Thresholding)

	N° di pixel aggiunti/ totale	*N° di pixel rimossi/totale*	*Variazione del numero di pixel*	*Estinzione media dei pixel aggiunti*
Lupus I	499/12112	221/12112	+2.30%	~0.89
Lupus V	293/6582	224/6582	+1.05%	~0.85
Lupus VI	376/10883	205/10883	+1.57%	~0.94
Lupus I (sub-reg)	110/1251	100/1251	+0.80%	~2.44
Lupus V (sub-reg)	53/1306	13/1306	+3.06%	~2.45
Lupus VI (sub-reg)	113/1552	54/1552	+3.80%	~2.24

Tabella 3 - Lupus I, statistiche su pixel della nube identificata tramite il metodo proposto. Le colonne 2 e 3 si riferiscono, rispettivamente, al rapporto tra pixel verdi e rossi rispetto al totale di pixel nella maschera relativa al metodo di sogliatura. La colonna 4 riassume la variazione complessiva delle precedenti due colonne.

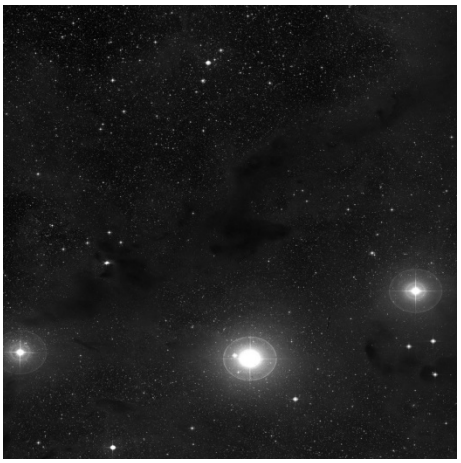
I risultati mostrati su diverse regioni della nebulosa di Lupus, sebbene ragionevolmente variabili in termini di estinzione, background noise e flusso delle masse di polveri e gas, dimostrano una valida robustezza e capacità di rivelazione degli edge e della densità di estinzione, rispetto alle misure presenti in letteratura. Ciò quindi conferma la validità scientifica del metodo, oltre al valore aggiunto del sistema di sogliatura automatica. Infatti la Tabella 2 evidenzia un generale accordo tra il nostro metodo e quello tradizionale nel valutare il contributo di massa dell'estinzione. Mentre la Tabella 3 giustifica il lieve apporto di massa misurato dal nostro metodo in termini di quantità di pixel aggiuntivi dotati di un livello di estinzione congruo rispetto alle regioni locali circostanti.

5.3 Estrazione di mappe di estinzione

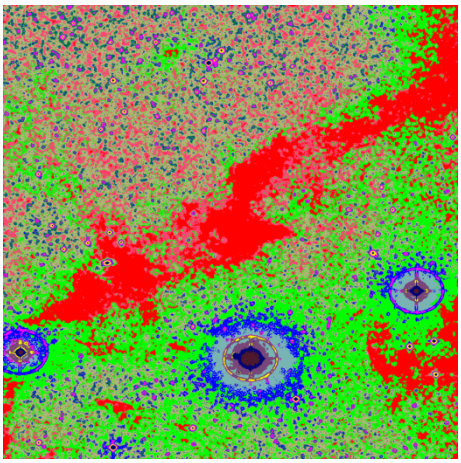
I risultati ottenuti utilizzando le mappe d'estinzione hanno dimostrato la capacità del metodo di clustering proposto (SOM+K-means) nell'individuare regioni ad alta estinzione, rispetto alle aree circostanti e nel circoscriverle attraverso l'identificazione di edge ben definiti.

Tale esito, a priori non scontato, ci ha consentito di provare ad esplorare il problema più difficile, ma nel contempo più interessante dal punto di vista astrofisico. Ossia, riuscire a rivelare regioni di estinzione a partire dalle immagini astronomiche direttamente osservate. La differenza principale, rispetto alle immagini utilizzate in precedenza, consiste nel fatto che in questo caso il valore del pixel non è l'estinzione, ma il flusso fotonico misurato. Ci proponiamo quindi di verificare la possibilità di rimpiazzare completamente il metodo tradizionale di sogliatura, basato sul conteggio statistico delle stelle, con il metodo proposto.

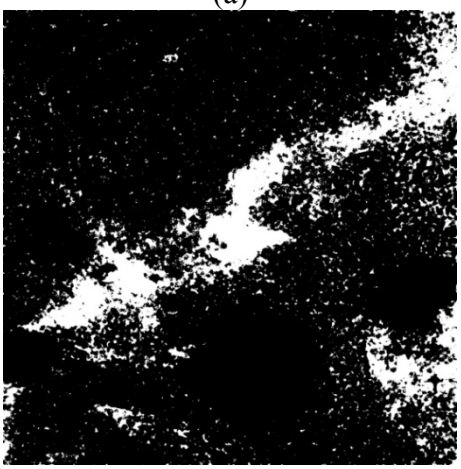
Naturalmente, nel caso del metodo proposto, non è possibile calcolare il valore puntuale dell'estinzione, quanto piuttosto raggiungere un risultato intermedio. Ossia individuare automaticamente un profilo di segmentazione di una regione con il quale avere indicazioni preziose in termini di distribuzione della densità dell'estinzione. Tale informazione permetterebbe di effettuare un ulteriore raffinamento del meccanismo di sogliatura, localizzando in modo più puntuale le zone entro cui calcolare le soglie.



(a)



(b)



(c)

(d)

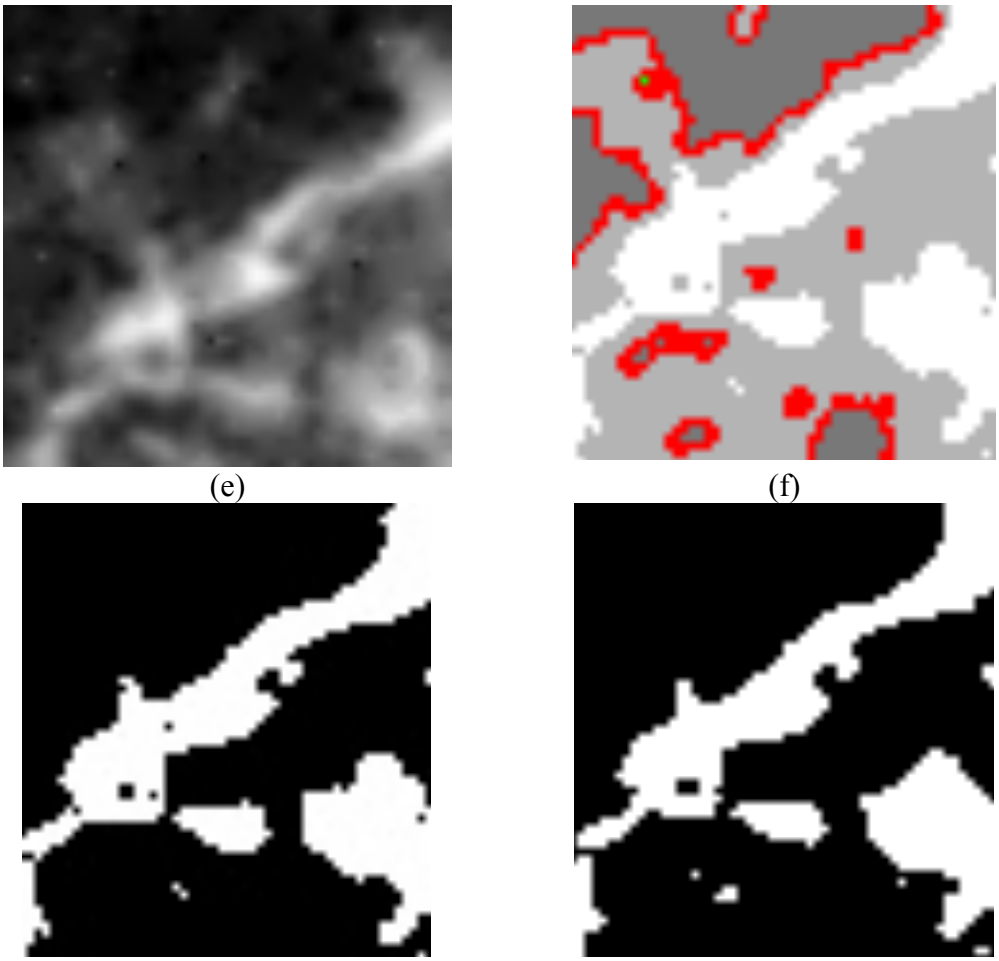


Figura 62 - Confronto tra metodi nella regione di Lupus I. Sono mostrate: in (a) l'immagine osservata originaria, in (b) il risultato del clustering, (c) la mappa binaria ricavata dal clustering, (d) l'immagine osservata riportata alla risoluzione della mappa d'estinzione, (e) la mappa d'estinzione ottenuta dal metodo tradizionale, (f) il risultato del clustering sulla mappa d'estinzione, (g) la corrispondente maschera binaria ricavata dal metodo proposto sulla mappa d'estinzione, (h) la maschera binaria ricavata dalla sogliatura del metodo tradizionale sulla mappa d'estinzione. Il metodo proposto si riferisce al clustering con il modello SOM + K-means.

L'analisi visuale dei risultati ottenuti (Figura 62, pannelli c e g), mette in evidenza una relazione spaziale tra le regioni di estinzione individuate nei due esperimenti mostrati, confermando, seppur in modo preliminare, la possibilità di estrarre la nube di estinzione in modo diretto dall'immagine osservata.

Il procedimento seguito è consistito nei seguenti passi:

1. applicazione del metodo proposto (SOM+K-means) all'immagine osservata ad alta risoluzione (Figura 62, pannello a); il risultato

dell'operazione di clustering ha portato all'estrazione di una serie di clusters evidenziati in colori diversi (Figura 62, pannello b);

2. creazione della maschera binaria, isolando il cluster corrispondente alla regione interessata dalla nube d'estinzione (Figura 62, pannello c);
3. downgrade della risoluzione dell'immagine originaria ad alta risoluzione (Figura 62, pannello d), riportandola alla risoluzione della mappa d'estinzione ottenuta dal metodo tradizionale. Ciò ha permesso di associare in modo biunivoco il valore d'estinzione ad ogni pixel del cluster individuato nell'immagine osservata;
4. calcolo massa d'estinzione relativa alle regioni individuate dal clustering sull'immagine osservata scalata in risoluzione e relativi confronti con la massa d'estinzione calcolata sulla mappa ottenuta dal metodo tradizionale (Figura 62, pannello h) e quella calcolata mediante applicazione del metodo proposto sulla mappa d'estinzione (Figura 62, pannello g).

In particolare, il doppio confronto menzionato al punto 4 della procedura doveva permettere di confrontare, in termini di stima della massa d'estinzione, sia il metodo proposto rispetto a quello tradizionale, agendo direttamente sull'immagine osservata; sia le variazioni indotte dal metodo proposto applicandolo sull'immagine osservata e sulla mappa d'estinzione. Quest'ultimo confronto ha una validità intrinseca dato che si era già verificata una corrispondenza del clustering con il metodo tradizionale sulla mappa d'estinzione relativamente alla stima della massa.

Naturalmente, nel caso dell'applicazione del metodo proposto all'immagine osservata originaria, non essendo più interessati all'individuazione di edge, è stato

necessario specializzare la fase di feature selection, apportando opportune modifiche allo spazio dei parametri. Inizialmente abbiamo utilizzato l'entropia. Tale feature, infatti, risulterebbe teoricamente in grado di descrivere l'andamento del flusso. A valle di una serie di test, la grandezza ottimale della finestra individuata (21 x 21 pixel) ha permesso di analizzare regioni più ampie rispetto al singolo pixel, adattandosi quindi a contesti ad elevatissima risoluzione. Questo comportamento, unito all'ipotesi che in regioni di formazione stellare vi sia una distribuzione omogenea di sorgenti, ha favorito la capacità di caratterizzare le regioni ad alta estinzione come quelle a più bassa entropia. Tale feature è infatti teoricamente rappresentativa di un flusso particolarmente costante e, per la definizione stessa di estinzione, basso, a causa del forte assorbimento di fotoni. Un caso particolare da affrontare è rappresentato da regioni che presentano flusso costante ma elevato, ad esempio il centro di una stella particolarmente grande. Queste regioni esibiscono comunque bassa entropia e quindi, per ottenere una individuazione più precisa, è necessario introdurre una feature in grado di tener conto dell'intensità del flusso. A tale scopo, lo spazio dei parametri ha richiesto l'aggiunta della media del flusso di una finestra di pixel, di dimensione pari a 5 x 5 (scelta su base empirica).

	Clustering sulla mappa d'estinzione		Clustering sull'immagine osservata		
	M_C *Clustering vs thresholding*	M_A *Clustering vs thresholding*	M_C *Clustering vs thresholding*	M_A *Clustering vs thresholding*	A_V *(thresholding)*
Lupus I (sub-reg)	~ +0.32%	~ +0.32%	~ -0.81%	~ -0.81%	2.5

Tabella 4 – Confronto in percentuale di masse solari tra le regioni d'estinzione estratte con metodo proposto (Clustering) e metodo tradizionale (Thresholding), alternando l'applicazione del metodo di clustering alla mappa d'estinzione e all'immagine osservata originaria.

L’esito del doppio confronto, effettuato al passo 4 della procedura su una regione d’esempio (Lupus I), ha fornito il risultato atteso, cioè che la stima della massa d’estinzione, rispettivamente estrapolata dal metodo tradizionale a sogliatura arbitraria, dal metodo di clustering applicato sull’immagine osservata e dal metodo di clustering applicato alla mappa d’estinzione, rimane sempre entro lo stesso ordine di grandezza, confermando quindi la validità qualitativa del metodo proposto.

6 Conclusioni

Questo lavoro, di carattere fortemente multi-disciplinare, consiste nello studio di applicabilità di tecniche non supervisionate di machine learning all’image processing, focalizzando l’attenzione su problemi di edge detection e clustering in ambito astrofisico. Lo scopo consiste dunque nella possibilità di rivelare l’estinzione in immagini monocromatiche ad alta risoluzione di nebulose (nella fattispecie le nebulose di Lupus, una delle principali regioni di formazione stellare in prossimità del Sole), ossia il noto fenomeno di assorbimento e *scattering* di radiazione elettromagnetica nei gas e nelle polveri cosmiche interposte tra una sorgente emittente e l’osservatore. In particolare si vuole ottenere risultati scientificamente almeno comparabili con le tecniche tradizionali note in letteratura, guadagnando primariamente in termini di automatizzazione del metodo. Unendo infatti le note proprietà di auto-adattamento ai dati, auto-correlazione nello spazio dei parametri, riduzione delle dimensionalità del problema e preservazione della topologia dei metodi di machine learning non supervisionato, si vuole da un lato migliorare la capacità di circoscrivere e delimitare le sotto-regioni a densità di estinzione variabile e dall’altro sostituire con tecniche automatiche il meccanismo arbitrario di scelta di

soglie di flusso e di conteggio delle regioni di estinzione. Obiettivi che abbiamo verificato non risultare raggiungibili con le tecniche tradizionali basate sul filtraggio del gradiente e maschere di convoluzione, come Canny, Sobel, Prewitt e Roberts, o meno tradizionali basate sulla logica fuzzy, principalmente a causa del rapporto segnale/rumore estremamente basso nelle immagini astronomiche. Viceversa, il metodo di clustering multi-stadio, ossia basato sul clustering con algoritmi di machine learning a cascata, e in particolare la rete SOM + K-Means ha permesso di ottenere i migliori risultati, circoscrivendo al meglio lo spazio dei parametri specifico per il tipo di dati in esame. Dal confronto con i risultati del metodo tradizionale, presenti in letteratura, in termini di misurazione della massa dell'estinzione e conformazione delle mappe d'estinzione in varie sotto-regioni, si è evidenziata una maggiore precisione nel circoscrivere le diverse distribuzioni di densità di estinzione, mantenendo valori di massa congruenti. Questo risultato ci ha consentito di estendere l'indagine, potendo valutare le capacità di rivelazione diretta dell'estinzione attraverso l'applicazione dei metodi di clustering direttamente sulle immagini originali, relative cioè alle osservazioni astronomiche, piuttosto che sulle mappe d'estinzione derivate.

Un ulteriore sviluppo del lavoro proposto consisterà dunque nell'approfondimento della rivelazione diretta dell'estinzione tramite i metodi presentati in altre nebulose, raffinando la scelta dello spazio dei parametri, e l'analisi dei parametri astrofisici derivati (ad es. la correlazione massa/luminosità), al fine di verificare la qualità del risultato in termini prettamente scientifici.

7 Bibliografia

Alcalà, J., 2016. *Comunicazione privata*. INAF, Osservatorio Astronomico di Capodimonte.

Bin, L., Samiei yeganeh, M., 2012. *Comparison for Image Edge Detection Algorithms*. IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661, Vol. 2, Issue 6.

Bishop, C.M., 2006. *Pattern recognition and machine learning.* Springer ISBN 0-387-31073-8.

Bradley, P., Fayyad, U., 1998. *Refining initial points for K-means clustering*. Proc. 15th Int. Conf. Machine Learning, pp. 91–99.

Breiman, L., 2001. *Random Forests*. Machine Learning, Springer Eds., 45, 1, pp. 25-32.

Brescia, M., 2012. *New Trends in E-Science: Machine Learning and Knowledge Discovery in Databases.* Horizons in Computer Science Research, chapter 1679, NOVA Publishing.

Cambrèsy, L., 1999. *Mapping of the extinction in giant molecular clouds using optical star counts*. A&A, 345, pp. 965-976.

Canny, J., 1986. *A Computational Approach to Edge Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6), 679-6987.

Chi, S-C., Yang, C-C., 2008. *A Two-stage Clustering Method Combining Ant Colony SOM and K-means*. Journal of Information Science and Engineering 24, 1445-1460.

Dickman, R.L., 1978. AJ, 83, 363.

Elmegreen, B.G., Falgarone, E., 1996. ApJ, 471, 816.

Fang, M., Yue, G., Yu, Q., 2009. *The Study on An Application of Otsu Method in Canny Operator.* Proceedings of the 2009 International Symposium on Information Processing (ISIP'09) Huangshan, P. R. China, August 21-23, pp. 109-112.

Haq, I., Anwar, S., Shah, K., Khan, M.T., Shah, S.A., 2015. *Fuzzy Logic Based Edge Detection in Smooth and Noisy Clinical Images.* PLoS ONE 10(9): e0138712. doi: 10.1371/journal.pone.0138712.

Hamel, L., Brown, C.W., 2011. *Improved interpretability of the unified distance matrix with connected components*. Proceedings of the 2011 International Conference on Data Mining.

Jain, R., Kasturi, R., Schunck, B.G., 1995. Machine Vision. Published by McGraw-Hill, Inc., ISBN 0-07-032018-7.

Jain, A., Dubes, R., 1988. *Algorithms for Clustering Data*. Cliffs, NJ: Prentice-Hall.

Jain, A., Murty, M., Flynn, P., 1999. *Data clustering: A review*. ACM Comput. Surv., vol. 31, no. 3, pp. 264–323.

Jain A.; Duin R.; Mao J. (2000). *Statistical pattern recognition: A review*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37.

Kaur, E.K., Mutenja, V., Gill, E.I.S., 2010. *Fuzzy logic based image edge detection algorithm in MATLAB*. International Journal of Computer Applications; 1(22):55–58, doi: 10.5120/442-675.

Kiviluoto, K., 1996. *Topology preservation in self-organizing maps.* Proceedings of the International Conference on Neural Networks. 294-299.

Kohavi, R., John, G.H., 1997. *Wrappers for feature subset selection*. Artificial Intelligence 97, 273–324.

Kohonen, T., 2001. *Self-Organizing Maps*. 3rd ed., Springer.

Leekwijck, W.V., Kerre. E.E., 1999. *Defuzzification: criteria and classification*. Fuzzy Sets and Systems. 108(2):159–178. doi: 10.1016/s0165-0114(97)00337-0 .

Likas, A., Vlassis, N., Verbeek, J., 2003. *The global K-means clustering algorithm*. Pattern Recognition, vol. 36, no. 2, pp. 451–461.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* Proc. 5th Berkeley Symp., vol. 1, 1967, pp. 281–297.

Marr, D., Hildreth, E., 1980. *Theory of Edge Detection*. Proceedings of the Royal Society B.

Moutarde, F., Ultsch, A., 2005. *U*F Clustering: A new performant cluster-mining method on segmentation of self-organizing map*. Proceedings of WSOM '05, September 5-8, Paris, France, 25-32.

Otsu, N.A., 1979. *Threshold Selection Method from Gray-Level Histograms*. IEEE Trans. on System, Man, and Cybernetics. vol.9, no.1, pp.62-66.

Petrović, S., 2006. *A comparison between the Silhouette index and the Davies-Bouldin index in labeling IDS clusters*. Proceedings of the 11th Nordic Workshop on Secure IT-systems, NORDSEC 2006, Linkoping, Sweden, pp. 53-64.

Prewitt, J.M.S., 1970. *Object enhancement and extraction*. Lipkin, B.S., Rosenfeld, A. (eds.) Picture Processing and Psychopictorics, pp. 75-149. Academic Press, New York.

Roberts, L.G., 1963. *Machine Perception of Three Dimensional Solids.* MIT Lincoln Laboratory Report, TR 315.

Sanchez-Marono, N., Alonso-Betanzos, A., Tombilla-Sanroman, M., 2007. *Filter methods for Feature Selection - A comparative study*. Intelligent Data Engineering and Automated Learning 4881, 178–187.

Sobel, I., 1968. *An isotropic 3x3 image gradient operator.* Talk at the Standford, Artificial Intelligence Project (SAIP).

Tizhoosh, H.R., 1997. Fuzzy Image Processing, Springer-Verlag.

Vapnik, V.N., 1998. *Statistical Learning Theory*. John Wiley and Sons, NY USA.

Vesanto, J., Alhoniemi, E., 2000. *Clustering of the Self-Organizing Map*. IEEE Transactions on neural networks. Vol. 11, No. 3, 586-600.

Viola, P., Jones, J.M., 2001. *Rapid Object Detection using a Boosted Cascade of Simple Features*. IEEE CVPR.

Xu, R., Wunsch, D., 2005. *Survey of clustering algorithms.* IEEE Trans Neural Networks, 16:645–678.

Whittet, D.C.B., 2003. *Dust in the Galactic Environment*. Series in Astronomy and Astrophysics (2nd ed.). CRC Press. p. 10. ISBN 0750306246.

Zadeh, L., 1965. Fuzzy Sets. Information and control. 8. 338-358.