QdG
Anno 2011_Numero 94

Implementazione di una nuova procedura per caratterizzare la forma di particelle mediante misure al CAMSIZER e algoritmi di clustering

Quaderni di Geofisica

94

Istituto Nazionale di Geofisica e Vulcanologia

Quaderni di Geofisica

Segreteria di Redazione
Francesca Di Stefano - coordinatore
Tel. +39 06 51860068
Fax +39 06 36915617
Rossella Celi
Tel. +39 06 51860055
Fax +39 06 36915617

redazionecen@ingv.it

ISSN 1590-2595
Anno 2011_Numero 94

Implementazione di una nuova procedura per caratterizzare la forma di particelle mediante misure al CAMSIZER e algoritmi di clustering

A new procedure for the characterization of the shape of particles by CAMSIZER measurements and cluster algorithms

Maria Deborah Lo Castro[1], Daniele Andronico[1], Carmelo Cassisi[2], Placido Montalto[1], Michele Prestifilippo[1]

[1]**INGV** (Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Catania - Osservatorio Etneo)
[2]**Università degli Studi di Catania** (Dipartimento di Matematica e Informatica)

In copertina Alcuni dei materiali utilizzati negli esperimenti di misurazione con CAMSIZER (particolare)

Implementazione di una nuova procedura per caratterizzare la forma di particelle mediante misure al CAMSIZER e algoritmi di clustering

In questo lavoro viene illustrata la fase di calibrazione di una nuova procedura mirata alla caratterizzazione della forma di particelle piroclastiche. Questa ricerca, finanziata dalla sezione INGV di Catania con fondi derivanti dal "Progetto Giovani", è stata condotta in collaborazione con la ditta Retsch Technology di Haan (Germania). L'innovazione di tale tecnica è infatti legata all'utilizzo dello strumento CAMSIZER (sviluppato proprio nella sede tedesca) che permette di acquisire importanti informazioni riguardo la taglia e la forma di un numero molto elevato di particelle (centinaia di migliaia). Unitamente a tale strumento sono stati utilizzati anche algoritmi di *clustering* e classificazione mediante i quali sono stati ottenuti dei raggruppamenti di individui sulla base delle loro caratteristiche morfologiche.
La fase qui descritta è stata condotta esclusivamente su materiali standard aventi geometrie regolari, quali cubi, sfere e cilindri e costituisce un primo stadio di validazione della metodologia che verrà estesa, in futuro, alle particelle di cenere vulcanica che, come è noto, sono caratterizzate da forme irregolari.

In this work we present the calibration phase of a new procedure for the characterization of the shape of pyroclastic particles. This research has been granted by INGV of Catania, with funds deriving from the "Progetto Giovani", in collaboration with Retsch Technology in Haan. The innovation of this procedure arises from the use of CAMSIZER (an instrument developed by the German leader company). This instrument permits to obtain very important information both on size and shape parameters of a high number of particles (hundreds of thousands data). Moreover, we used clustering and classification algorithms in order to group particles according to their morphologic characteristics.
This calibration phase has been tested only on standard materials with regular geometries such as cubes, spheres and cylinders. In the future we will apply this methodology to volcanic ash particles that, as well-known, are characterized by irregular morphologies.

Introduzione

La forma è una caratteristica molto importante che influenza le proprietà ed i comportamenti fisici di materiali di diversa natura. Anche in ambito vulcanologico, lo studio della forma delle particelle vulcaniche emesse durante le eruzioni esplosive, permette di trarre informazioni sia sull'origine e sui meccanismi di frammentazione delle particelle stesse, ma anche sui processi post-eruttivi quali alterazione, deposizione e trasporto [Riley et al., 2003]. Inoltre, la forma influenza il tempo di residenza delle particelle in atmosfera ed è quindi un parametro utilizzato in alcuni modelli di dispersione delle ceneri vulcaniche [Scollo et al., 2008].
La quantificazione della forma di una particella non è una materia di semplice trattazione e gran parte delle tecniche utilizzate in vulcanologia, basate generalmente sull'analisi di immagine mediante microscopi, permettono di analizzare soltanto numeri limitati di particelle. Grazie alle nuove tecnologie introdotte in questo campo di analisi, esistono delle strumentazioni avanzate capaci di analizzare una mole di dati sempre maggiore. Tra queste strumentazioni va annoverato il CAMSIZER (www.retsch-technology.com), uno strumento che permette di ottenere informazioni sulla taglia e la forma di ogni singola particella componente un campione di materiale incoerente, molto utilizzato in campo industriale per il controllo di qualità di materiali più disparati [Lo Castro e Andronico, 2008].
L'obiettivo che ci proponiamo è quello di riuscire a suddividere un campione di piroclastiti, caratterizzato da particelle aventi diverse forme, in classi distinte rappresentate da indi-

vidui aventi simili caratteristiche morfologiche. Per raggiungere tale obiettivo, in questo lavoro presentiamo la fase di calibrazione di una tecnica che è il risultato dell'integrazione dei dati ottenuti dal CAMSIZER, con una successiva analisi di clustering e di classificazione. In particolare, verranno descritti gli strumenti (CAMSIZER e software di clustering), gli esperimenti di calibrazione e validazione condotti su materiali di forma nota ed infine i risultati ottenuti.

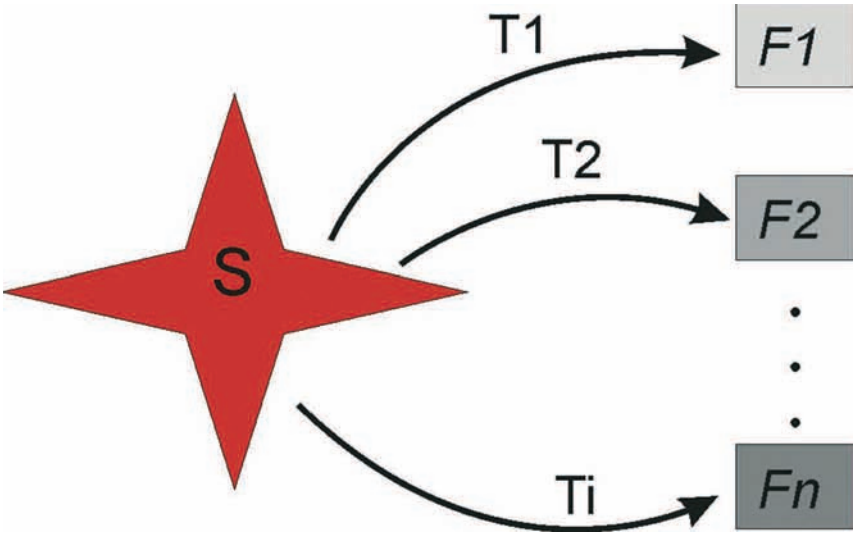


Figura 1 Schema di caratterizzazione di una forma S in base alle features descrittive F.
Figure 1 Scheme illustrating the characterization of a generic shape S according to a series of features F.

1. Definizione di forma

Il concetto di forma è di così semplice comprensione ed utilizzo nella vita comune, quanto difficile risulta invece la sua definizione e rappresentazione. Le definizioni più comuni in letteratura si basano spesso sulla nozione dell'invarianza delle proprietà di un oggetto alle trasformazioni geometriche base (traslazione, rotazione e fattore di scala) [Dryden e Mardia, 1998]. In base a questa definizione, per descrivere una forma, bisogna considerare degli appositi "descrittori", cioè dei set numerici con diversi gradi di complessità tali che i descrittori di forme distinte debbano essere sufficientemente dissimili tra loro in modo da discriminare ogni singola forma [ISO 9276-6 2003].

Data una specifica forma *S*, è possibile identificare una serie di misure e proprietà che la caratterizzano che sono definite *features*. Per esempio, una forma può essere caratterizzata in base al valore della sua area, al perimetro, al numero di cavità o di estremità, ecc. . Il processo di caratterizzazione di una forma implica quindi una serie di trasformazioni *Ti* tali che la forma possa essere rappresentata da una serie di misure scalari o *features* F_k con $k = 1,2....n$ le quali possono essere raggruppate in un vettore $F = (F1, F2,....Fn)$ (Figura 1) [Costa e Cesar Jr, 2001]. Le features possono essere scalari o vettori e devono essere tali da enfatizzare le proprietà di interesse e godere di un forte potere discriminativo. Per esempio, se si volessero caratterizzare dei poligoni, la features relativa al numero di lati sarebbe molto più significativa di quella relativa al numero di cavità.

1.1 Metodi di caratterizzazione della forma

I metodi per descrivere la forma possono essere classificati in base a diversi criteri e principalmente si dividono in:

a) **metodi qualitativi**: si basano su descrizioni abbastanza soggettive che fanno riferimento all'apparenza visiva di una data particella. Ad esempio, si possono avere "particelle arrotondate", "sub angolari" ed "angolari" riferendosi generalmente a carte comparative (Figura 2).
b) **metodi quantitativi**: si basano su valori numerici che possono essere calcolati dalle immagini delle particelle o da particolari proprietà fisiche delle stesse mediante operazioni matematiche o numeriche.

In questa trattazione considereremo esclusivamente i metodi quantitativi basati sull'analisi di immagine.

L'analisi di immagine è una tecnica molto versatile che trova applicazione in svariate discipline e consiste nella manipolazione ed analisi di informazioni scientifiche rappresentate come immagini. Data una particella reale (e quindi tridimensionale), per poter fare l'analisi di immagine occorrono dei dispositivi di ingresso, capaci di raccogliere immagini (generalmente una telecamera, una fotocamera, microscopi o scanner). Le immagini acquisite, che saranno immagini bidimensionali della forma reale di partenza, verranno tradotte in forma digitale, cioè leggibile da un computer al quale è installato un software capace di effettuare l'analisi dell'immagine acquisita e di poter restituire in uscita una serie di informazioni quali ad esempio i parametri dimensionali e della forma (Figura 3).

Ci sono diversi sistemi e strumenti per effettuare l'analisi di immagine. L'analisi al microscopio (stereoscopico o a scansione elettronica, SEM) è stata finora la tecnica di riferimento, in quanto permette di misurare in modo diretto la taglia e la forma delle particelle. Ciononostante, questa tecnica manuale comporta molte ore di lavoro ed è spesso soggetta a poca oggettività di misura. I più moderni sistemi di tipo automatico permettono di analizzare in modo più preciso, oggettivo e veloce decine di migliaia di particelle alla volta. Tra queste metodologie possiamo distinguere tra:

a) **Analisi di Immagine di tipo statico**, in cui le particelle stazionano su una slitta in movimento che viene inquadrata da una telecamera e da un microscopio (Figura

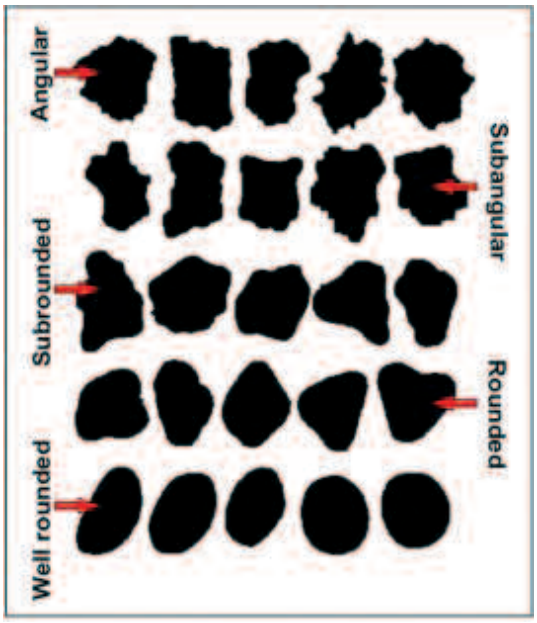


Figura 2 Carta comparativa di Russell, Taylor e Pettijohn per la descrizione qualitativa della forma [modificata da Muller, 1967].
Figure 2 Comparative chart of Russell, Taylor and Pettijohn for the qualitative characterization of shape [modified after Muller, 1967].

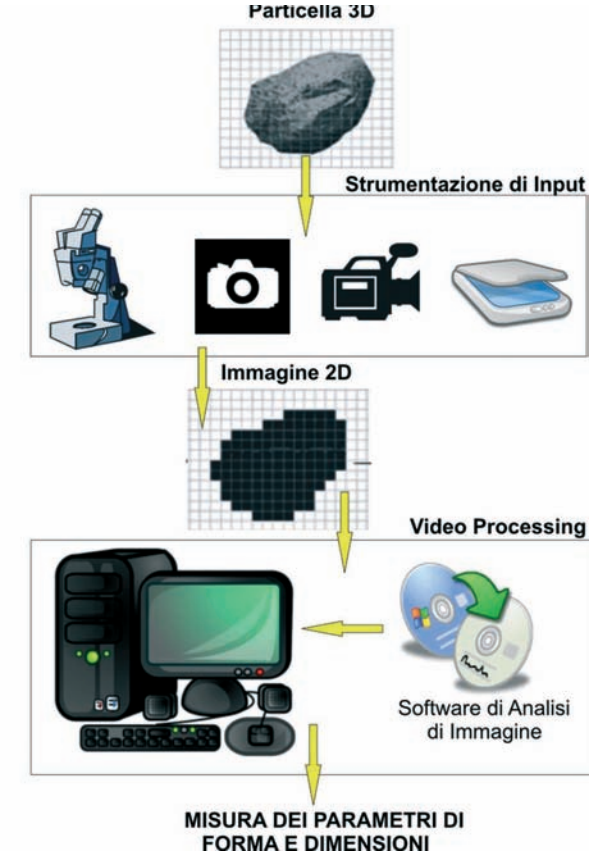


Figura 3 Schema della strumentazione necessaria per l'analisi di immagine.
Figure 3 Scheme showing the basic instruments for image analysis.

4a). Con questo metodo si possono trattare solo un numero ristretto di dati e le particelle sono orientate in funzione della loro base.

b) **Analisi di Immagine di tipo dinamico**, in cui le particelle si muovono lungo un corridoio per poi cadere in flusso, dotate di una certa velocità di caduta v, all'interno di una camera di misura dove vengono inquadrate da una o più telecamere (Figura 4b). In questo modo le particelle cadono secondo delle orientazioni arbitrarie.

2. Metodologia

La procedura da noi proposta si basa sui risultati delle misure ottenute al CAMSIZER e da successiva analisi di clustering e classificazione con determinati algoritmi.
Il primo step consiste nell'effettuare la misura di un campione al CAMSIZER dalla quale si ottengono dei file di output contenenti diverse informazioni relative ai parametri dimensionali e di forma delle particelle che compongono il campione scelto per l'analisi. Per estrapolare le informazioni da questi file sono stati utilizzati dei software esterni (**Alp-reader** e **Contproc**, forniti dalla ditta Retsch Technology) che hanno permesso di estrarre le features descrittive delle particelle.
Queste features sono state successivamente utilizzate come dati di input all'interno degli algoritmi di clustering e classificazione necessari per raggruppare un dato campione in base alle forme caratteristiche.
Lo schema del processo di analisi è mostrato in Figura 5.
Di seguito verranno descritti in dettaglio gli strumenti e le procedure utilizzate in questo lavoro.

2.1 CAMSIZER

Il CAMSIZER ® è uno strumento da laboratorio costruito dalla Retsch Technology (www.retschtechnology.com) che misura e analizza simultaneamente le dimensione e la forma di particelle solide incoerenti in un intervallo compreso tra 30 μm e 30 mm, sfruttando l'analisi di immagine di tipo dinamico. Con questo strumento è possibile ottenere un numero di dati statisticamente più attendibile ed affidabile rispetto alle misure ottenute dai comuni microscopi.
Lo strumento (Figura 6) è costituito da un corpo centrale provvisto di un imbuto (*funnel*) in cui viene posto il campione di materiale da misurare. Le particelle componenti il campione scorrono lungo un piatto vibrante (*feeder*) fino a quando, giunte all'estremità dello stesso, iniziano a cadere nella camera di misura. All'interno della camera di misura ogni particella, illuminata da una luce bianca parallela, viene ripresa da due telecamere digitali, una per i clasti di dimensioni maggiori (*CCD-Basic*) e l'altra per quelli più piccoli (*CCD-Zoom*). Le immagini registrate dalle telecamere, che rappresentano la proiezione dell'ombra di ogni particel-

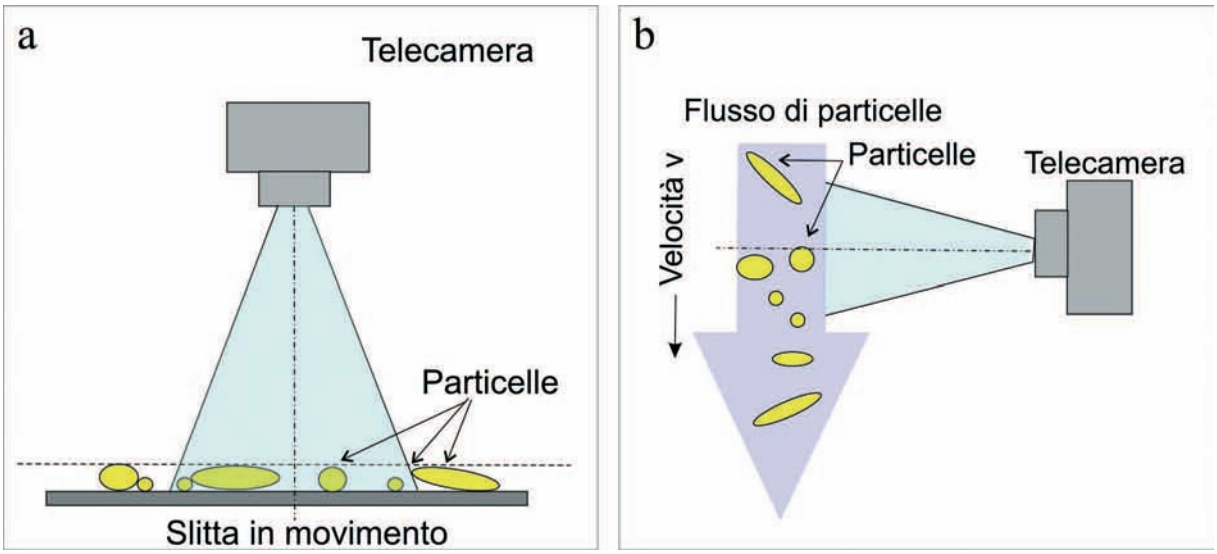


Figura 4 Schemi illustranti diverse metodologie di Analisi di Immagine: a) statica e b) dinamica.
Figure 4 Different Image Analysis methodologies: a) static and b) dynamic.

la, vengono elaborate in tempo reale da un software installato su un computer collegato allo strumento. La proiezione di ogni particella viene scansionata secondo 64 direzioni di misura diverse con un'altissima risoluzione, in modo da poter ottenere la misura precisa di diverse grandezze quali area, perimetro, misure di lunghezza e particolari parametri morfologici.

Per definire la dimensione (x) di una particella, il software utilizza diversi modelli: xc_{min}, x_{area}, xFe_{min}, xFe_{max}, xMa_{min}, xMa_{rec}, x_{length}. Ogni modello permette di misurare proprietà diverse di una particella a cui corrispondono diverse distribuzioni cumulative del volume di particelle (*Q3-distribution*).

I risultati finali possono essere rappresentati sia mediante una curva cumulativa che attraverso altri tipi di grafici (istogrammi e gaussiane) (Figura 7). Infine è possibile visualizzare delle tabelle e dei *reports* riassuntivi dei valori e dei parametri misurati sia delle dimensioni che delle forme delle particelle costituenti il campione analizzato.

I risultati di misura vengono salvati in una serie di file nella directory specifica di salvataggio dei dati (CAMDAT). In particolare, i file di output sono:

Raw data file - *.RDF: file nativo del software del CAMSIZER che contiene tutte le informazioni relative al processo di misura. Viene letto dal software del CAMSIZER.

File Excel - *.XLE (formato inglese): tabella con i risultati di alcuni parametri impostati in precedenza relativi ad ogni classe granulometrica.

*File *alp*: contiene tutte le informazioni dei parametri dimensionali e di forma relativi ad ogni singola particella misurata. Viene letto da software esterni opportunamente programmati (Alpreader).

*File *kon*: contiene informazioni di altri parametri particolari che si riferiscono al contorno delle particelle. Anche questa estensione può essere letta da appositi software esterni (Contproc).

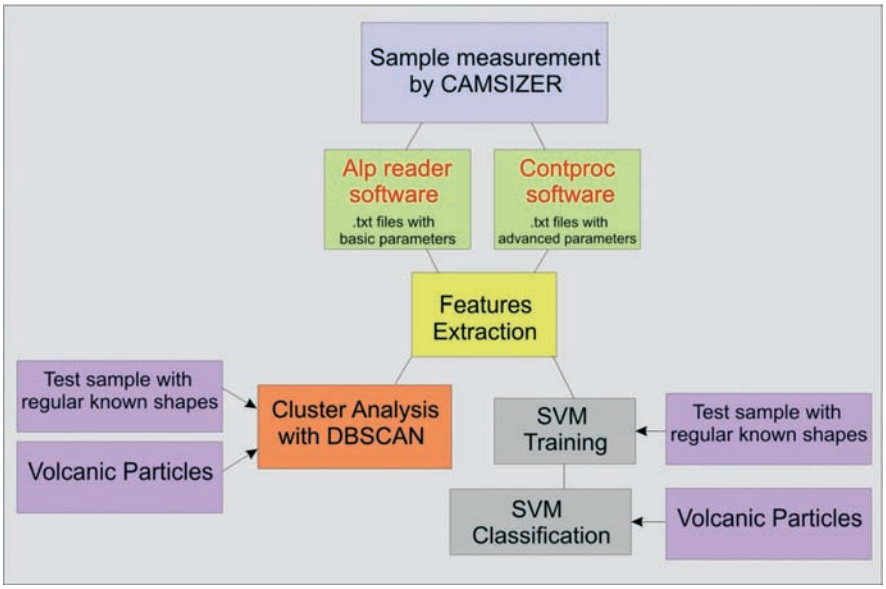


Figura 5 Schema illustrante la metodologia di analisi utilizzata.
Figure 5 Scheme of the methodology used in the research.

2.2. Features: parametri di forma

Una semplice tecnica per la caratterizzazione della forma di una particella è quella di utilizzare il rapporto di due misure dimensionali (x_i e x_j) ottenendo il cosiddetto *Conventional Shape Descriptor* (S_{ij}) [Hentschel et al. 2003]:

$$S_{ij} = x_i / x_j \qquad (1)$$

In base ai tipi di misure dimensionali scelte, ogni parametro di questo tipo sarà più sensibile a particolari aspetti della forma. Bisogna scegliere in modo appropriato i parametri di forma in modo che essi siano in grado di descrivere al meglio e senza ridondanze le caratteristiche specifiche che si vogliono evidenziare. Nel 2003, Hentschel e coautori hanno applicato l'analisi di clustering alle diverse combinazioni di parametri di forma ottenuti da grandezze dimensionali, in modo da identificare un set ridotto di parametri che permettessero di descrivere in modo parsimonioso la forma di una particella. Da questa ricerca è emerso che la forma, per un determinato range di polveri e materiali sciolti commerciali, può essere efficientemente descritta da due *Conventional shape descriptor* che sono il rapporto di aspetto (AR) per la stima dell'allungamento della particella e il fattore di forma (FF), definito spesso anche sfericità o circolarità, per l'irregolarità dei contorni [Hentschel e Page, 2003].

I parametri descritti corrispondono anche a quelli normalmente utilizzati in vulcanologia per la descrizione della forma delle particelle vulcaniche [Riley et al., 2003], per cui sono stati considerati anche in questa trattazione, dopo essere strati estrapolati dai file alp di uscita. È stato inoltre utilizzato un terzo parametro mirato alla descrizione dell'angolarità, ottenuto dalle informazioni del contorno della particella contenute dall'estensione *.kon* dei file di output.

Ecco di seguito una breve descrizione dei parametri di forma utilizzati:

a) Rapporto di Aspetto

$$xc_{min} \,/\, xFe_{max} \qquad (2)$$

Questo parametro descrive il rapporto tra l'ampiezza (xc_{min}) e la lunghezza (xFe_{max}) (Figura 8a) della proiezione della particella e si riferisce al grado di allungamento della particella stessa. Le particelle tozze e globulari avranno un rapporto di aspetto più vicino all'unità, mentre le particelle allungate presenteranno valori inferiori.

b) Sfericità (**Circolarità** *sensu* Cox [1927]; **Form Factor** *sensu* Kuo et al. [1998])

$$\frac{4\pi A}{P^2} \qquad (3)$$

Questo parametro, definito come il rapporto dell'area di proiezione di una data particella (A) rispetto al perimetro della stessa (P) (Figura 8b) si riferisce al grado di sfericità della particella che può essere quantificato esattamente comparando la superficie di una particella con quella di una sfera avente lo stesso volume [Blott e Pye, 2008; Wadell, 1932; Wentworth, 1933]. Dato che l'area di superficie e il volume sono parametri difficili da misurare, sono stati proposti dei metodi alternativi su misure di particelle in 2 dimensioni, ed in questo caso si potrebbe più propriamente parlare di ~~CIRCOLARITÀ~~. Spesso tale parametro viene confuso con il grado di arrotondamento (*roundness*) di una particella ma,

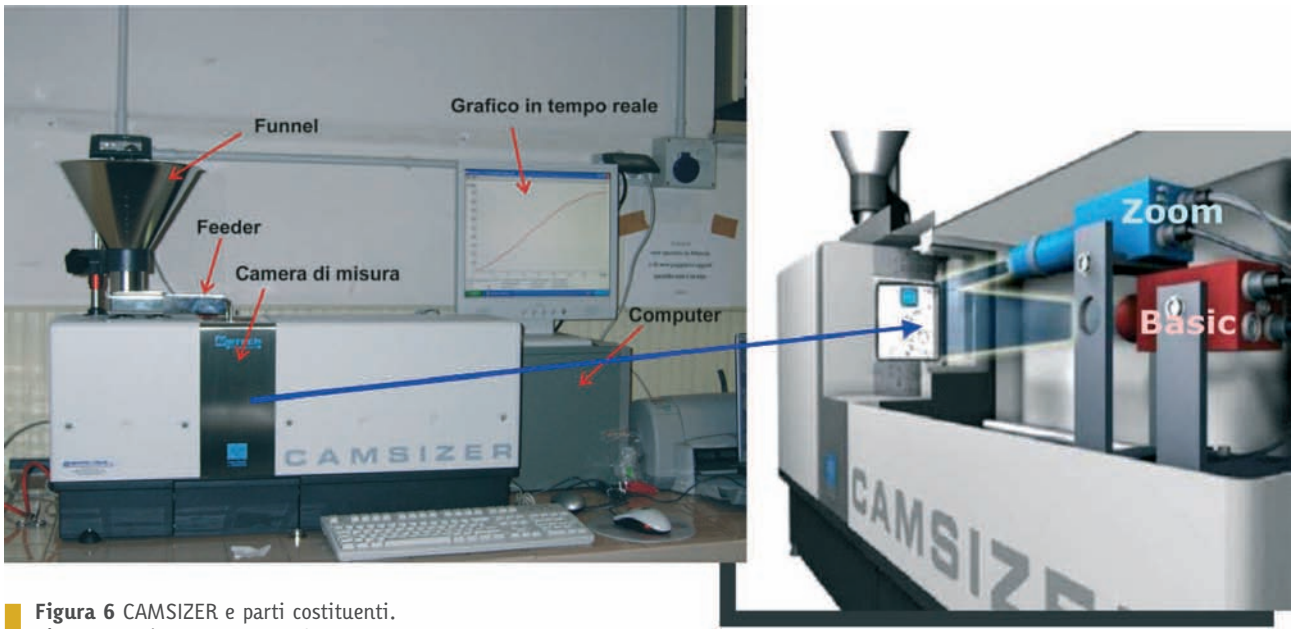


Figura 6 CAMSIZER e parti costituenti.
Figure 6 Main components of CAMSIZER.

nonostante i due concetti siano correlabili, sono effettivamente diversi. Se si considerano ad esempio un cubo e un dodecaedro regolare (12 facce), quest'ultimo presenta una sfericità maggiore rispetto al cubo ma una rotondità nulla, perché le facce formano tra loro degli angoli diversi da zero. Una sfera è invece perfettamente arrotondata perché è costituita da un raggio di curvatura costante in ogni direzione [Blott e Pye, 2008].
Quindi, ragionando in 2D, un cerchio perfetto ha un valore di circolarità pari a 1, mentre oggetti con forme più irregolari avranno valori inferiori in quanto una forma irregolare sarà caratterizzata da un aumento del perimetro.

c) Angolarità

$$E_{polygon} = \frac{\Sigma_i h_i \frac{\pi - \alpha_i}{\pi}}{\Sigma_i h_i} \qquad (4)$$

Parametro che fa riferimento al contorno di una particella e al grado di irregolarità della stessa. Grazie ad un software esterno (Contproc), implementato dalla Retsch Techgnology ed ancora in fase di sperimentazione, è stato possibile estrapolare una serie di parametri riferibili al contorno della particella, salvati in output nel file *.kon*.
In particolare abbiamo considerato il parametro Epolygon (4) [Zilly, 2005] che tiene conto del valore medio degli apici del poligono rilevante (angoli convessi), definito come quel poligono risultante dal *best-fit* di un dato contorno, moltiplicato per l'altezza relativa h (Figura 8c). Tale valore dipende dal numero di spigoli del poligono: un contorno perfettamente arrotondato (cerchio) avrà un valore Epolygon = 0, mentre uno appuntito darà un valore Epolygon = 1.

2.3. Clustering: PyDBSCAN

Con il termine *clustering* si indica il processo mediante il quale è possibile raggruppare oggetti in base a caratteristiche comuni (*features*). Ogni individuo x_i (i=1…m) di una data popolazione X, di cardinalità m, viene singolarmente caratterizzato da n features y_j (j=1..n), che possono essere considerate le coordinate della posizione dell'i-esimo individuo in uno spazio n-dimensionale. Individui caratterizzati da varia-

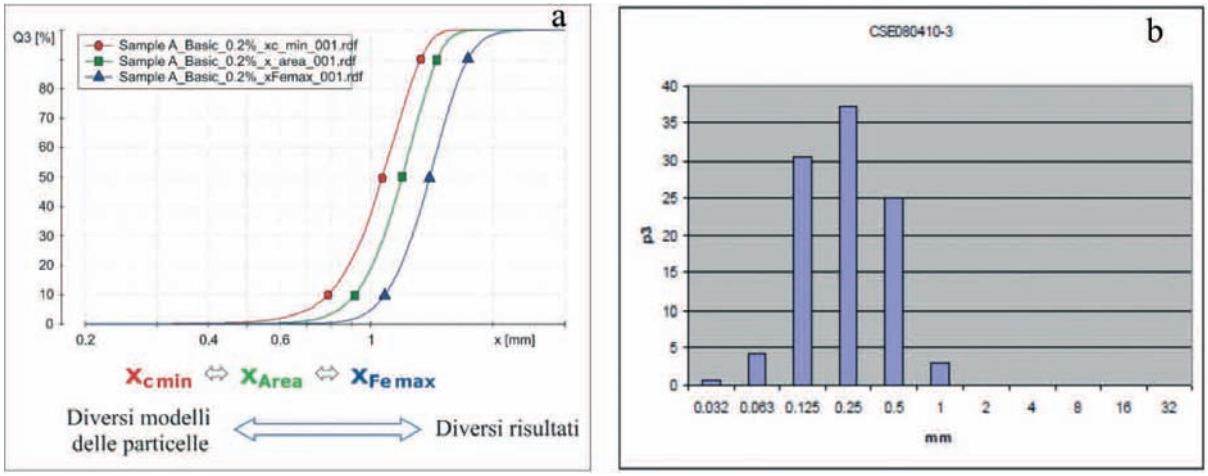


Figura 7 Alcune rappresentazioni grafiche ottenute dai risultati di una misura al CAMSIZER: a) curve cumulative; b) istogramma.
Figure 7 Output graphic representations obtained from a CAMSIZER measurement: a) cumulative curves; b) histogram.

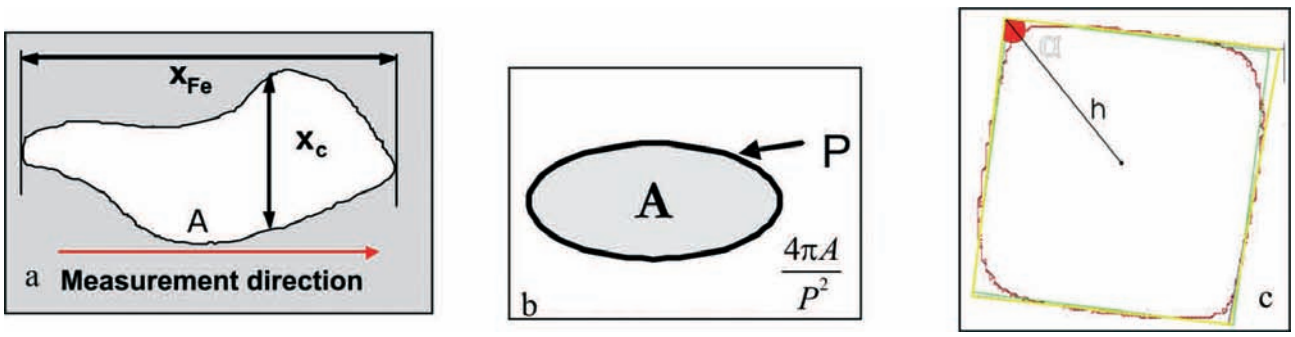


Figura 8 Schemi illustranti i parametri di forma utilizzati nel lavoro: a) rapporto d'aspetto; b) sfericità; c) angolarità.
Figure 8 Schemes describing the shape parameters used: a) aspect ratio; b) sphericity; c) angularity.

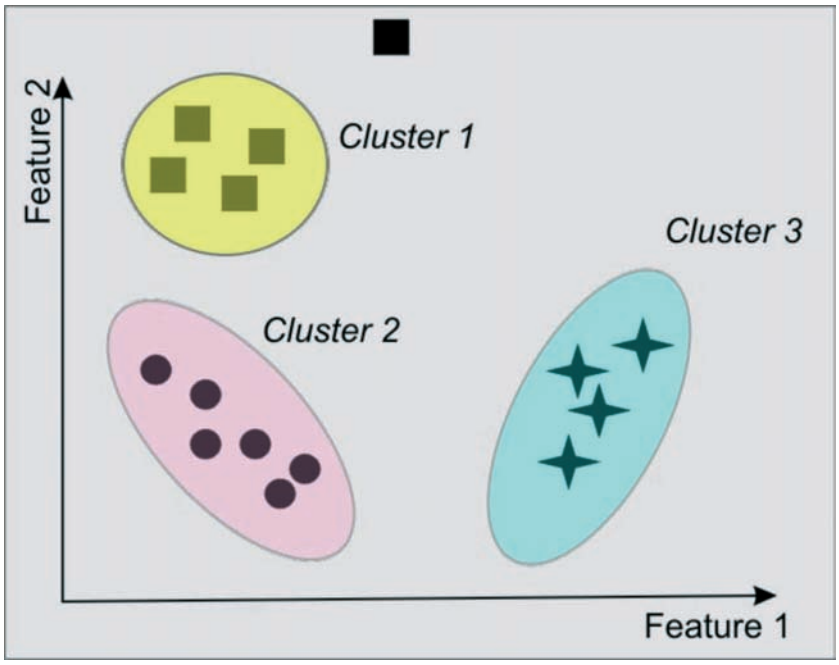


Figura 9 Schema illustrante il concetto di clustering.
Figure 9 Scheme showing the basics of cluster analysis.

bili simili giaceranno vicini all'interno di questo spazio.
Il software utilizzato per il *clustering* è il *PyDBSCAN* [Cassisi et al., 2011] fondato sul concetto di densità che sfrutta l'algoritmo *DBSCAN* (*Density Based Spatial Clustering on Application with Noise*) [Ester et al., 1996], basato sull'idea che oggetti che formano regioni dense possono essere raggruppati in *cluster*. Gli oggetti sono **punti** in uno spazio *d*-dimensionale (R^d) nel quale viene definita una funzione di distanza tra due punti *p, q: dist(p,q)*. Viene definito ε-*neighbourhood* di un punto *p* l'insieme di punti N_ε che ricadono nel cerchio di raggio ε e centro *p*. Se $|N_\varepsilon| \geq MinPts$ allora *p* viene chiamato *core point*. Tutti i punti in N_ε sono **direttamente raggiungibili per densità** (*Directly density-reachable*) da *p* (Figura 10a). Un punto *q* è **raggiungibile per densità** (*Density-reachable*) da un punto *p* se esiste una catena di punti $q_1,\ldots,q_n$, $q_1 = p$, $q_n = q$ tale che per ogni *i*, q_i+1 è direttamente raggiungibile per densità da q_i, per $1 \leq i \leq n$ (Figura 10b). Un punto *p* è **connesso per densità** (*density-connected*) a un punto *q*, se esiste un punto *o* tale che sia *p* che *q* sono raggiungibili per densità da *o* (Figura 10c). Un *cluster* è un insieme massimale di punti *density-connected*.
L'algoritmo *DBSCAN* opera come segue: sceglie casualmente un punto *p* in *D* e controlla il suo vicinato N_ε. Se N_ε contiene più di *MinPts*, crea un nuovo *cluster* con *p* come *core point* e, in modo iterativo aggiunge tutti i punti direttamente raggiungibili per densità da *p*. Il processo termina quando non ci sono più punti da aggiungere al *cluster*. Verrà poi scelto in maniera casuale un nuovo punto non classificato e i passi precedentemente descritti verranno re-iterati finché non ci saranno più punti da assegnare a nessun *cluster*. Un punto in *D* è definito *outlier* se non è possibile assegnarlo a nessun *cluster*.

2.4. Classificazione con Support Vector Machines (SVM)

Le *Support Vector Machines* (SVM) [Cannata et al., 2011] sono un metodo di classificazione binaria che permette di restituire il più ampio margine di separazione tra classi di oggetti. L'idea alla base dell'algoritmo SVM è quella di utilizzare gli oggetti che stanno tra le frontiere delle varie classi per identificare l'iperpiano separatore ottimale che massimizza il margine di separazione tra le classi, chiamato *Maximum Marginal Hyperplane* (MMH) (Figure 11 e 12). Il problema del calcolo del MMH viene formulato in termini di programmazione quadratica nel seguente modo:
minimizzare:

$$W(\alpha) = -\sum_{i=1}^{l}\alpha_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \qquad (5)$$

condizionato da:

$$W(\alpha) = \sum_{i=1}^{l} y_i \alpha_i = 0$$
$$\forall i : 0 \leq \alpha_i \leq C \qquad (6)$$

dove *l* denota il numero di oggetti del *training set*, α è un vettore di *l* variabili, dove ogni componente α_i corrisponde ad un elemento del *training set* (x_i, y_i). *C* è un parametro per la gestione dell'influenza degli *outlier* (o *noise*) sul *training set*.
Trattandosi di classificatori lineari, le SVM presentano difficoltà nell'apprendimento di classi non linearmente separabili. Per ovviare a tale limitazione, vengono utilizzati diversi tipi di trasformazioni (o proiezioni) del *training set* originale, sostituendo $k(x_i, y_i)$ con una funzione kernel φ, come ad esempio il *kernel polinomiale* $(x^T x_i + 1)^p$ o il *radial basis function kernel*

$$\exp\left(-\frac{1}{2\sigma^2}\|x - x_i\|^2\right),$$

che permette di proiettare le classi in un nuovo spazio dove possono essere separate linearmente.
Le due estensioni più conosciute delle SVM al problema multi classe sono i metodi: *One-Against-One* (OAO) e *One-*

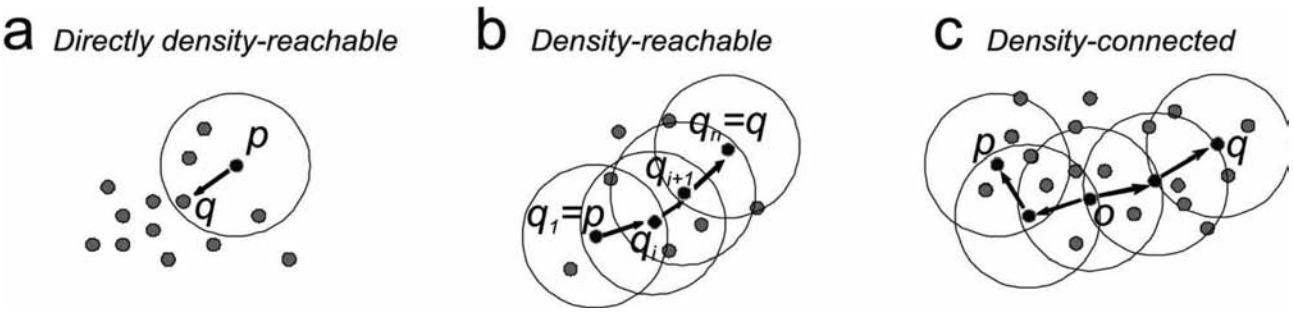


Figura 10 Alcune definizioni importanti per descrivere i punti nell'ambito di un clustering basato sulla densità. Immagine da Cannata et al. [2011].
Figure 10 Significant definitions to describe the points inside a density-based clustering. Image from Cannata et al. [2011].

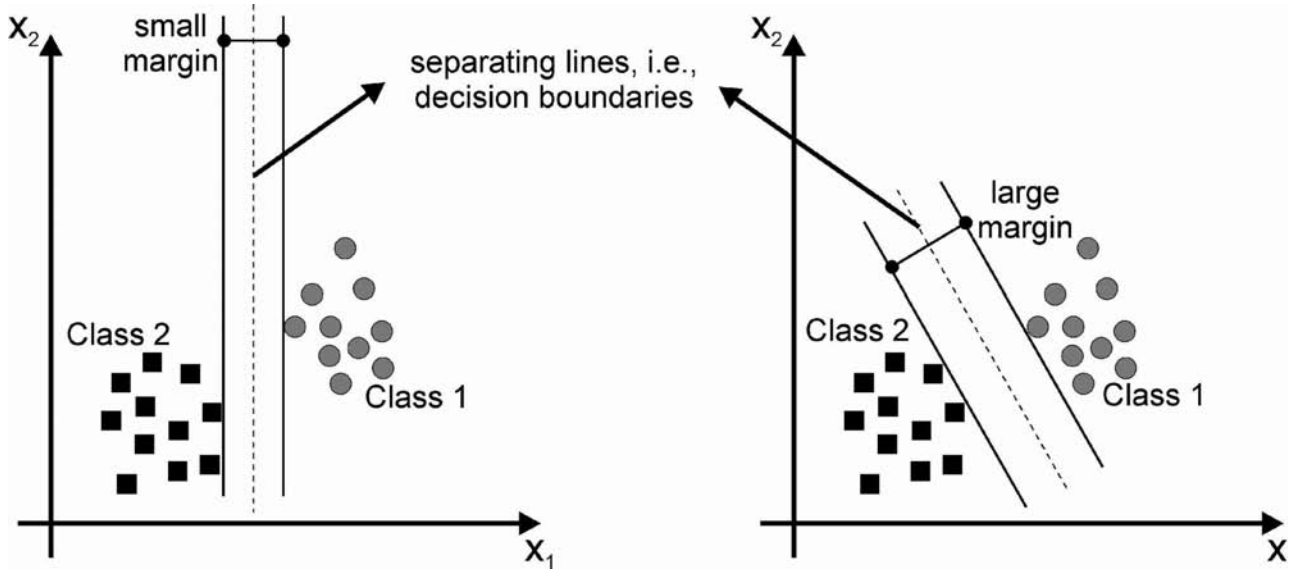


Figura 11 Due differenti rette di separazione per uno spazio contenente due classi di dati (quadrati neri e cerchi grigi). Sulla sinistra il margine di separazione risulta più stretto di quello proposto sulla destra.
Figure 11 Two different lines separating a two-class space (black squares and grey circles). On the left the separation margin is narrower than those on the right.

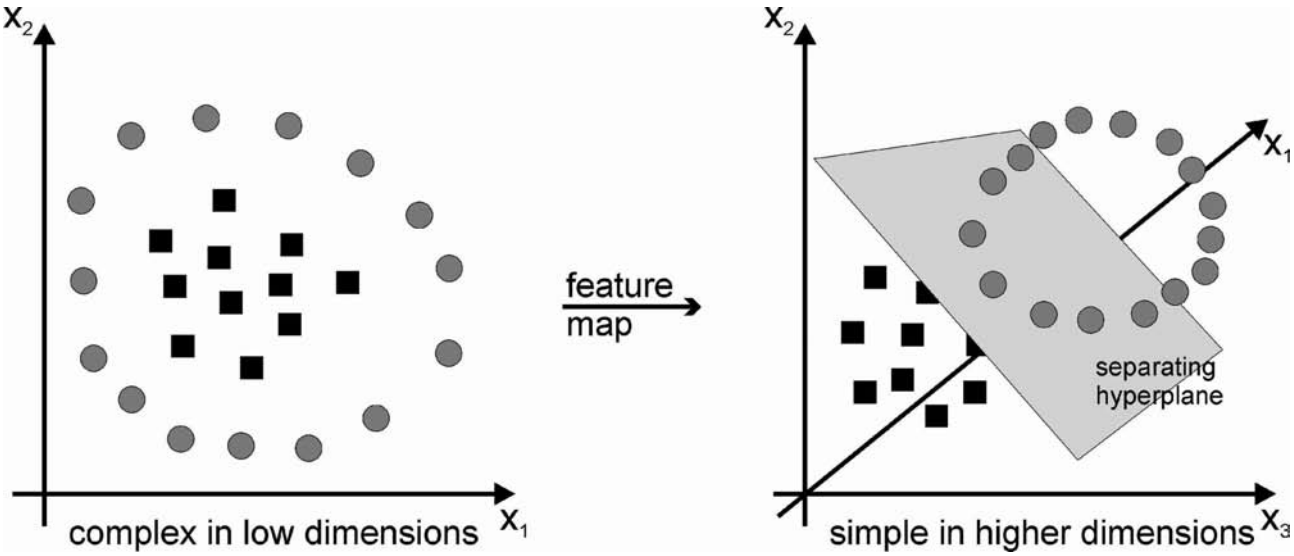


Figura 12 Due classi di dati nello spazio 2D originale (sinistra) e una probabile proiezione su uno spazio delle features di dimensionalità maggiore (destra). Immagine da Cannata et al., [2011].
Figure 12 Two data class in a 2D original space (on the left) and a possible projection on a feature space with higher dimensions (on the left). Figure from Cannata et al., [2011].

Against-All (OAA). Il primo costruisce $k(k-1)/2$ classificatori considerando ogni combinazione di classi a coppie. Il secondo costruisce k classificatori, dove l'i-esimo classificatore utilizza tutti i *patterns* appartenenti alla i-esima classe considerando gli altri come oggetti facenti parte di un'altra classe.

3. Analisi dei dati

Per questa fase di calibrazione abbiamo utilizzato esclusivamente delle forme note, precisamente sfere, cubi e cilindri, nonché le seguenti tipologie di misure:

- misure effettuate su forme create con una versione di simulazione del software del CAMSIZER;
- misure su campioni reali effettuate con lo strumento vero e proprio.

3.1 Dati simulati

Usando un software di simulazione del CAMSIZER (Figura 13) è stato possibile riprodurre diverse tipologie di misure

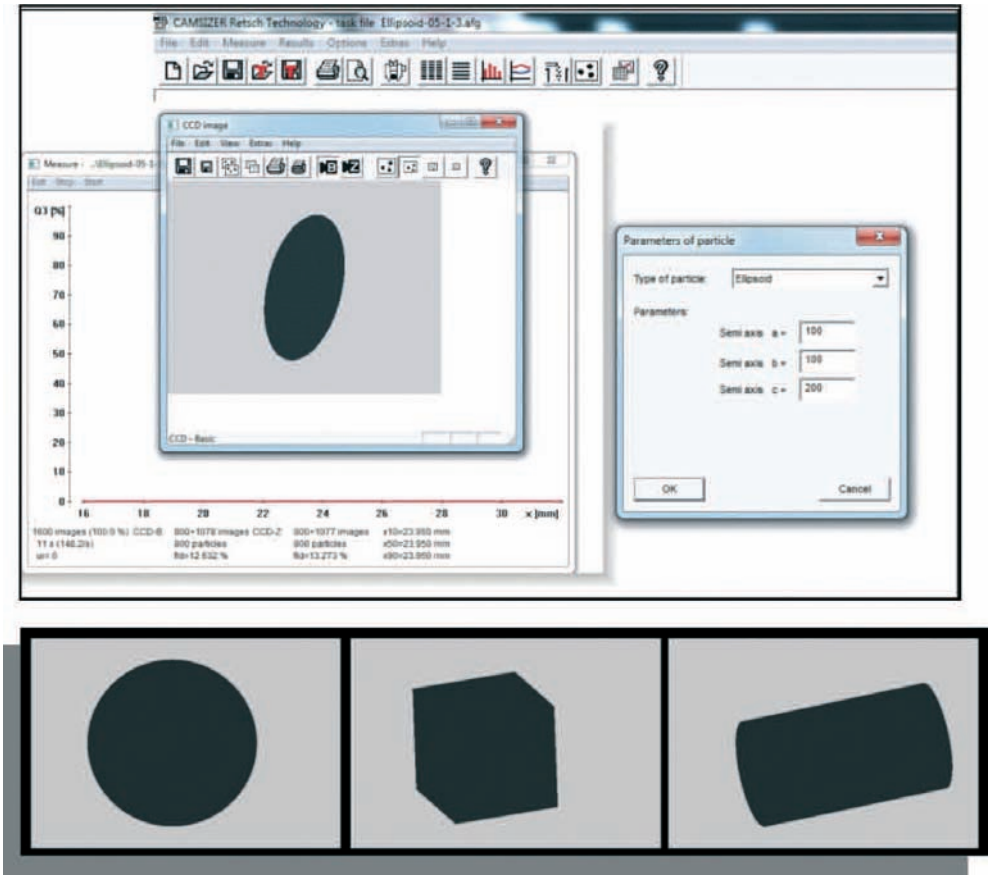


Figura 13 Interfaccia grafica del software di simulazione del CAMSIZER (in alto) e forme simulate utilizzate negli esperimenti (in basso).
Figure 13 Graphic interface of the simulation software of CAMSIZER (on top) and simulated shapes used in the experiments (bottom).

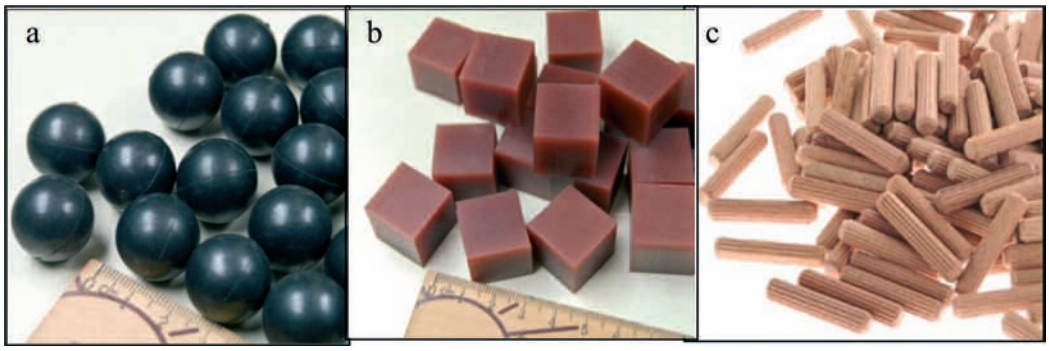


Figura 14 Materiali utilizzati negli esperimenti: a) sfere di plastica; b) cubi standard ~~in~~ gomma; c) cilindri ~~in~~ legno.
Figure 14 Material used in the experiments: a) plastic spheres; b) standard rubber cubes; c) wood cylinders.

utilizzando geometrie elementari quali sfere, cubi e cilindri. In questo modo è possibile fissare, non solo la forma della particella, ma anche il numero di campioni da analizzare, in modo da avere un controllo preciso sulle analisi svolte.
Sono state effettuate diverse misure su un numero crescente di forme simulate di cubi, cilindri e sfere, partendo da un numero di 600 campioni fino ad un totale di 1500 utilizzando il simulatore del CAMSIZER.
I file di *output* (file *alp* e *kon*) sono stati importati rispettivamente su Alp-reader e su Contproc attraverso i quali è stato possibile estrarre le features desiderate. Queste sono state inserite, a loro volta, all'interno di un'apposita tabella costituita da 4 colonne, riportanti le *label* indicanti il tipo di forma e, per ogni particella, le rispettive 3 *features* (Tabella 1). La tabella è stata poi importata all'interno del software PyDBSCAN in modo da ottenere le nubi di densità del *dataset*.

3.2 Dati reali

I campioni reali utilizzati si riferiscono a cubi di gomma con dimensioni standard (2 cm e 1.2 cm di lato), sfere in plastica (diametro di 2 cm) e cilindri in legno (lunghezza di 4 cm e diametro di 1 cm). I campioni sono stati analizzati al CAMSIZER, effettuando sia prove singole di ogni campione che prove su un totale degli stessi.

4. Risultati

4.1 Analisi di clustering su dati simulati

In Figura 15 osserviamo nello spazio tridimensionale caratterizzato dagli assi (b/l, SPHT, Epol) 3 distinte nuvole di densità e quindi 3 *cluster* che raggruppano le diverse forme analizzate, cioè sfere, cubi e cilindri. Questo significa che le features scelte sono appropriate per discriminare in modo ottimale le 3 diverse forme che si trovano separate tra loro nello spazio.
In particolare si rileva che:

- SFERE: sono raggruppate in una piccola area corrispondente al valore 1 di b/l e SPHT e 0 del parametro Epol.
- CILINDRI: si estendono in un'area compresa tra un valore di b/l compreso tra 0.5 e 0.9 ma con una percentuale maggiore attorno a b/l=0.6, e quindi risultano più allungati rispetto alle sfere. Il valore di SPHT risulta più o meno costante e compreso tra 0.75 e 0.85 mentre il valore di Epol presenta variazioni più ampie comprese tra 0.1 (particelle più arrotondate) ~~a~~ 0.5 (particelle più spigolose). Tali differenze si osservano in quanto~~,~~ il cilindro, durante la sua "caduta" nella camera di misura dello strumento, viene ripreso dalle telecamere virtuali in ogni possibile direzione e quindi può presentarsi secondo la massima o la minima area di proiezione, cioè più o meno allungato (Figura 15b).
- CUBI: presentano valori di b/l abbastanza ampi, compresi tra 0.6 e 0.9, un andamento similare per il valore di SPHT ed un valore abbastanza uniforme per quanto riguarda Epol. Ciò implica che l'angolarità è pressoché costante, dato che, in qualsiasi proiezione viene ripreso il cubo nella sua caduta, mantiene sempre una certa spigolosità, mentre gli altri parametri possono variare in base all'area di proiezione ripresa.

	b/l	SPHT	Epol
S	0.997	0.9971	0.0955
S	0.997	0.9979	0.0958
S	0.997	0.9961	0.0872
S	0.9969	0.9996	0.0871
S	0.997	10.011	0.0872
S	0.9971	0.9988	0.0954
...	...	...	...
Ci	0.4988	0.7682	0.2669
Ci	0.4987	0.7647	0.3759
Ci	0.4983	0.753	0.2591
Ci	0.4989	0.7947	0.2466
Ci	0.4984	0.7862	0.1777
Ci	0.4995	0.726	0.4422
...	...	...	...
Cu	0.7744	0.8645	0.3561
Cu	0.8413	0.9152	0.3356
Cu	0.6524	0.8538	0.4431
Cu	0.6499	0.779	0.4456
Cu	0.748	0.9173	0.3731
Cu	0.6625	0.802	0.4328
Cu	0.7045	0.8648	0.4042
...	...	...	...

Tabella 1 Tabella di input del sofware di clustering in cui vengono riportate nella prima colonna le label delle forme utilizzate (S=sfere; Ci=cilindri; Cu=cubi) e nelle altre 3 colonne i parametri di forma utilizzati (b/l=rapporto d'aspetto; SPHT= sfericità; Epol=angolarità).
Table 1 Input data interface for the cluster software. Labels of the shape typologies are shown in the first column (S=sphere; Ci=cylinders; Cu=Cubes). The other 3 columns report the shape parameters (b/l=aspect ratio; SPHT= sphericity; Epol= angularity).

Utilizzando lo stesso dataset dell'analisi di *clustering* sui dati simulati abbiamo testato un modello basato sulle Support Vector Machine (SVM) calcolando l'iperpiano ottimale capace di massimizzare i margini di separazione tra i vari *cluster* ottenuti. In Figura 16 è possibile osservare un diagramma 2D in cui è ben visibile come i 3 *cluster* vengano divisi in 3 aree ben definite.

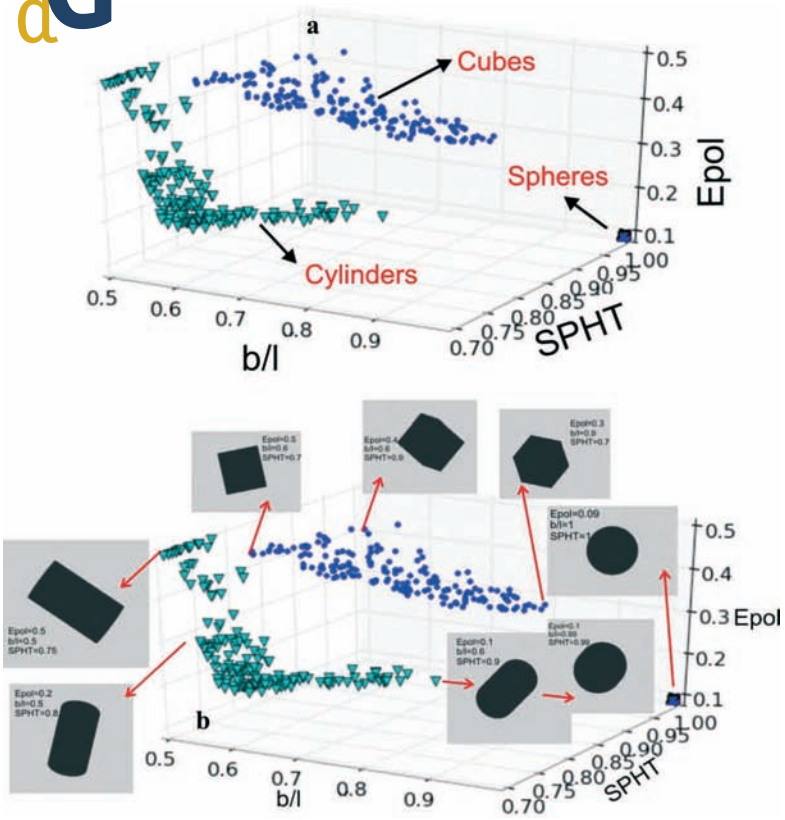


Figura 15 a) Nubi di densità relative alle 3 forme analizzate in uno spazio tridimensionale caratterizzato dalle features scelte; b) stesso diagramma a in cui vengono inserite le immagini relative alle forme studiate.
Figure 15 a) Density clusters showing the 3 different analysed shapes in a 3D space defined by the descriptive features; b) the same diagram ~~a~~ with the different shape typologies.

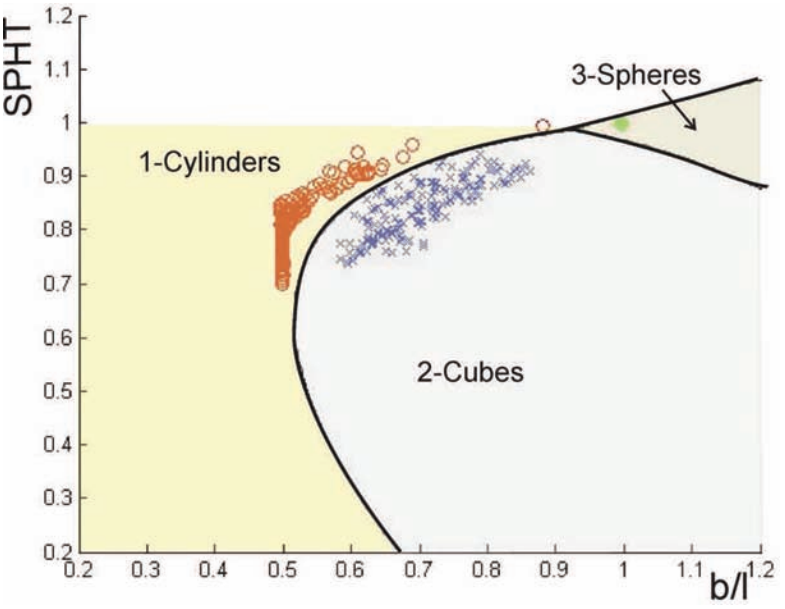


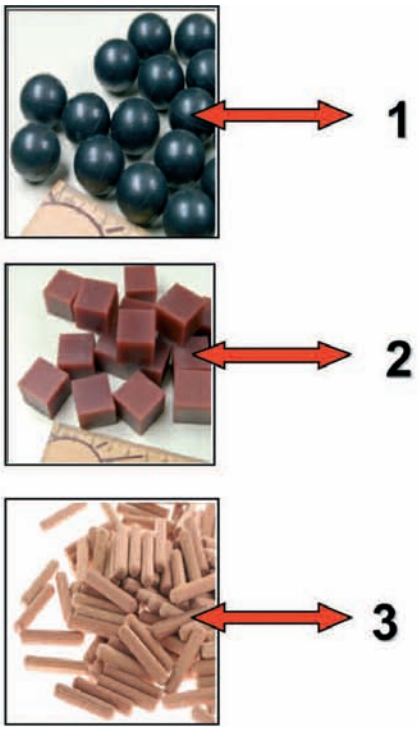
Figura 16 Diagramma bidimensionale in cui è stato riportato l'iperpiano ottimale di separazione dei 3 cluster ottenuto dall'analisi SVM.
Figure 16 2D diagram showing the optimal hyperplane separating the 3 clusters, obtained by the SVM analysis.

4.2 Validazione del sistema utilizzando i dati reali

Il controllo di qualità della classificazione costituisce una fase molto importante in quanto permette di stimare l'affidabilità degli algoritmi di classificazione utilizzati e favorire l'identificazione di eventuali errori. Per validare il sistema abbiamo effettuato la classificazione di un nuovo *data-set* caratterizzato da cubi, sfere e cilindri reali misurati con il CAMSIZER (Figura 17) e abbiamo eseguito l'indicizzazione di ogni classe. La qualità della classificazione viene quantificata utilizzando quella che viene chiamata *matrice di confusione* che restituisce una rappresentazione dell'accuratezza di classificazione statistica. In particolare, dato un oggetto appartenente al nuovo insieme di dati, la *matrice di confusione* è composta da colonne che rappresentano le istanze delle classi predette mentre le righe rappresentano le istanze delle classi reali impiegate per la validazione (Figura 17). Mentre nella diagonale troviamo il numero di oggetti classificati correttamente, gli altri elementi mostrano il numero di oggetti che vengono associate a classi di appartenenza errate.
Nel caso in oggetto, considerando la classe di indice 1, che rappresenta le sfere, il numero maggiore di campioni (619) si trova in corrispondenza della cella classificata correttamente, corrispondente al *cluster* 1 della classe predetta (le sfere del modello di *clustering*); 10 campioni sono invece classificati in modo errato, cadendo invece nel campo del *cluster* 2, quello dei cubi.
La stessa cosa avviene se consideriamo la classe 2 (cubi), cioè 365 campioni vengono classificati in modo corretto e 78 in modo errato, cadendo nel campo delle sfere.
Per la classe 3 (cilindri) si osserva come 111 campioni ricadono correttamente nella classe 3 dei cilindri mentre 11 campioni sono classificati in modo non appropriato nella classe delle sfere.

5. Conclusioni e prospettive future

In questo lavoro è stata descritta la fase di calibrazione di una nuova tecnica sviluppata per caratterizzare le famiglie di forme predominanti delle particelle costituenti un campione di cenere vulcanica, sfruttando le potenzialità del CAMSIZER.
Le forme utilizzate per la calibrazione sono riferibili a geometrie molto semplici, con dimensioni e caratteristiche note (sfere, cilindri e cubi) che sono state sia create arbitrariamente con un software di simulazione, che misurate nella realtà utilizzando dei materiali standard.
I risultati hanno permesso di osservare come questa tecnica abbia dato dei risultati positivi per le forme elementari, in quanto l'algoritmo di *clustering* è riuscito perfettamente a suddividere le 3 forme in classi distinte. Ciò è stato avvalorato anche dalla successiva fase di classificazione



		PREDICTED		
		Cluster 1	Cluster 2	Cluster 3
ACTUAL	Cluster 1	619	10	0
	Cluster 2	78	365	0
	Cluster 3	11	0	111

Figura 17 Materiali reali utilizzati nella fase di test (a sinistra) e matrice di confusione (a destra).
Figure 17 Real materials used during the testing phase (on the left, with labels) and confusion matrix (on the right).

mediante tecnica SVM, grazie alla quale siamo riusciti a trovare un iperpiano di separazione ottimale delle 3 classi identificate in precedenza.

Il passo successivo di questa ricerca sarà mirato alla realizzazione di nuovi esperimenti sui materiali vulcanici (lapilli e ceneri) che, a differenza dei materiali qui descritti, sono composti da particelle di forma assai irregolare e quindi di più difficile caratterizzazione.

Ringraziamenti

Ringraziamo Gert Beckmann, Jörg Westermann, Kai Düffels della ditta Retsch Technology (Haan) per il prezioso supporto e per la disponibilità mostrata durante le fasi di preparazione del progetto.

Bibliografia

Blott S.J. and Pye K., (2008). *Particle shape: a review and new methods of characterization and classification*. Sedimentology, 55, 31-63. doi: 10.1111/j.1365-3091.2007.00892.x.

Cannata A., Montalto P., Aliotta M., Cassisi C., Pulvirenti A., Privitera E. and Patanè D., (2011). *Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques*. Geophysical Journal International, no. doi: 10.1111/j.1365-246X.2011.04951.x.

Cassisi C., Montalto P., Pulvirenti A., Aliotta M., Cannata A., (2011). *PyDBSCAN un software per il clustering di dati*. Rapporti Tecnici INGV, n. 182.

Costa L.F. and Cesar – Jr R.M., (2001). *Shape analysis and classification: theory and practice*. CRC Press, Boca Raton.

Cox E.P., (1927). *A method of assigning numerical and percentage values to the degree of roundness*. J. Paleont., 1, pp.61-73.

Dryden I., L. and Mardia K., V., (1998). *Statistical Shape Analysis*. Wiley, Chichester.

Ester M., Kriegel H.P., Sander J., Xu X., (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

Hentschel M. L. and Page N. W., (2003). *Selection of Descriptors for Particle Shape Characterization*. Part. Part. Syst. Charact. 20 (2003) 25 ± 38.

ISO 9276-6, (2008). *Representation of results of particle size analysis* -- Part 6: Descriptive and quantitative representation of particle shape and morphology.

Kuo C-Y., Rollings R.S., Lynch L.N., (1998). *Morphological study of coarse aggregates using image analysis*. Journal

of Materials in Civil Engineering, 10(3), 135-142.

Lo Castro M.D. and Andronico D., (2008). *Operazioni di base per la misura della distribuzione granulometrica di particelle vulcaniche tramite il CAMSIZER*. Rapporti Tecnici INGV, n. 79.

Muller G., (1967). *Methods in sedimentary petrology*. Hafner, New York

Riley C.M., Rose W.I., Gregg J. S. Bluth G.J.S. (2003). *Quantitative shape measurements of distal volcanic ash*. Journal of Geophysical Research, vol. 108, no. b10, 2504, doi:10.1029/2001jb000818.

Scollo S., Folch A., Costa A., (2008). *A parametric and comparative study of different tephra fallout models*. Journal of Volcanology and Geothermal Research 176, 199–211.

Wadell H., (1932). *Volume, shape, and roundness of rock particles*. Journal of Geology 40, 443–451.

Wentworth C.K., (1933). *The shapes of rock particles: a discussion*. J. Geol., 41, 306–309.

Zilly M., (2005). *Algorithms for Optical Measurement of Granular Matter*. Tesi di laurea in Fisica. Università di Duisburg-Essen.

Indice

Coordinamento editoriale e impaginazione
Centro Editoriale Nazionale | INGV

Progetto grafico e redazionale
Daniela Riposati | Laboratorio Grafica e Immagini | INGV

Via di Vigna Murata, 605
00143 Roma
Tel. +39 06518601 Fax +39 065041181

http://www.ingv.it

Istituto Nazionale di Geofisica e Vulcanologia