Il Data Mining nello sviluppo di modelli previsionali per la gestione integrata dell'agricoltura sostenibile

Stefania D'Arpa[1], Emanuele Barca[1], Vito Felice Uricchio[1]*

Riassunto: *Traendo spunto dall'articolata letteratura scientifica in tema di applicazioni del* Data Mining *(DM) per la gestione integrata dell'informazione ambientale, nel presente lavoro viene tracciato un percorso applicativo riferito alla valorizzazione dei dati di tipo spazializzato al fine di favorire l'adozione di pratiche gestionali che facciano proprie le istanze di sostenibilità ambientale, superando le limitazioni alla produttività legate all'agricoltura puramente biologica. In particolare, partendo dalle problematiche legate all'agricoltura di precisione, si focalizza l'attenzione sulla classe di algoritmi di DM detti "supervisionati" indirizzati alla* clusterizzazione *delle superfici coltivate delle aziende agricole, illustrandone le potenzialità utili alla definizione di processi decisionali in fase di gestione e pianificazione di interventi. Tale approccio è particolarmente significativo per la gestione integrata dei dati regionalizzati con quelli di carattere ambientale riferiti in particolare alle caratteristiche topografiche e pedologiche del territorio che si vuole modellizzare ed alla natura chimico-fisica del suolo, al tipo di coltivazione avviata, alla disponibilità e qualità delle acque, nonché ad aspetti socioeconomici e congiunturali.*
Parole chiave: *Sistemi Agrometeorologici; Agricoltura di Precisione; Modelli Previsionali; Data Mining; Spatial Clustering.*

Abstract: *Starting from the literature concerning applications of Data Mining (DM) for the integrated management of environmental information, in the present work an application path referred to the smart use of spatial data is traced to encourage the adoption of sustainable practices in agriculture, overcoming the limitations to the productivity related to biological agriculture. In particular, this work focuses its attention on the class of DM algorithms called "supervised algorithms" addressed to the clustering of the cultivated area of farms, illustrating its potentiality to define the typical decision making of management and planning of interventions. This approach is particularly significant for the integrated management of regionalized and environmental data to the topographical and soil features of the area that we want to model and to the physical-chemical nature of soil, to the kind of cultivation used, to the availability and quality of the water, as well as economic and socio-economic aspects.*
Keywords: *Agrometeorological systems; Precision Farming; Predictive models; Data Mining; Spatial Clustering.*

INTRODUZIONE

È sempre più condivisa l'opinione secondo cui nella gestione delle problematiche complesse, quali quelle ambientali, agricole ed agroforestali si debbano prediligere approcci integrati piuttosto che settoriali, in cui primaria importanza venga attribuita all'analisi delle relazioni tra le molte parti che contribuiscono alla definizione complessiva dei fenomeni.

L'agrometeorologia con la complessità delle problematiche affrontate è una delle scienze che più si presta ad approcci di studio di tipo integrato ed i sistemi agrometeorologici, studiando i parametri ambientali che direttamente influenzano l'attività e la produzione agricola, sono ottimi strumenti che permettono la gestione di una delle principali priorità da conseguire in agricoltura ed in ambito eco-ambientale in generale: la sostenibilità.

Un tipico esempio di pratica agricola orientata alla sostenibilità del sistema ambientale è la cosiddetta "agricoltura di precisione".

L'agricoltura di precisione (nota anche come *precision agriculture* o *precision farming*) nasce come pratica agricola codificata negli anni '80 in seguito a ricerche ed applicazioni effettuate negli Stati Uniti ed in Australia (Colby Torbett *et al.*, 2007; Pringle and McBratney, 2004). Il paradigma applicativo dell'agricoltura di precisione è quello di una gestione delle superfici agricole aziendali che tenga conto della variabilità intrinseca e indotta del suolo e delle specifiche esigenze delle colture, al fine di incrementare la produzione, minimizzare i danni ambientali ed elevare gli standard qualitativi dei prodotti agricoli (Pierce and Nowak, 1990; National Research Council, 1997; Robert, 2002).

Uno dei più ambiziosi ed interessanti aspetti che emerge dalla filosofia di impostazione delle tecniche agricole di precisione è, quindi, il tentativo di coniugare due obiettivi apparentemente inconciliabili: la massimizzazione della produttività

* Corresponding Author e-mail: stefania.darpa@ba.irsa.cnr.it
[1] Consiglio Nazionale delle Ricerche - Isituto di Ricerca sulle Acque CNR-IRSA.

riducendo, al contempo, sia i costi ambientali sia quelli economici.
Per perseguire tali obiettivi è necessaria una dettagliata conoscenza di parametri colturali, topografici e meteo-ambientali. Possedendo tali informazioni, infatti, è possibile operare la razionalizzazione delle principali fasi che intervengono nel processo di coltivazione e ribaltare l'assunto dell'agricoltura biologica in base al quale l'applicazione di metodologie sostenibili influenzi negativamente l'andamento di resa delle colture. Tali metodologie infatti, operando la scelta di non utilizzare antiparassitari e/o fito-farmaci ma antagonisti biologici di parassiti ed infestanti, rinunciano consapevolmente ad una parte della resa colturale.
La fertilizzazione e l'irrigazione, in particolare, sono due momenti cruciali che intervengono nel processo di coltivazione
Ottimizzare tali fasi razionalizzando, ad esempio, l'utilizzo di fertilizzanti ed acqua può essere un utile strumento per migliorare qualitativamente e quantitativamente le rese colturali, riducendo l'impatto ambientale.
Possedendo dettagliate conoscenze di tutti i parametri che intervengono in tali fasi, le aziende agricole, utilizzando l'agricoltura di precisione, possono implementare modelli numerici contenenti una griglia di dettaglio in cui ad ogni punto sono associati un grande numero di attributi quali: posizione geografica, quota, esposizione, irradiazione, contenuto idrico del suolo, contenuto di sostanza organica, etc. Tali informazioni concorrono a formare un *dataset* di dati eterogenei in grado di pilotare macchine fertilizzatrici e sistemi automatici di irrigazione distribuendo sulla superficie coltivata, opportunamente partizionata in aree omogenee, la giusta quantità di acqua e fertilizzante. Per condurre un tale tipo di gestione (che viene non per caso detta "di precisione") è ovviamente necessaria un'approfondita analisi spazio-temporale dei dati disponibili, ovvero, diventa necessario conoscere ad ogni ciclo di ferti-irrigazione, la variazione spaziale di contenuto idrico e di sostanza organica del suolo, e, al contempo, la quantità di irraggiamento e di umidità già assorbita dalla coltura al fine di stimare le quantità da fornire e prevedere i periodi di raccolta.
La quantità di informazioni da gestire per applicare in maniera più efficace e corretta questo tipo di pratica agricola, dunque, è davvero imponente e richiede l'applicazione di specifiche metodologie di analisi dei dati, quali quelle di *Data Mining* (Hand *et al.*, 2001), in cui l'estrazione di *patterns* significativi da grandi masse di dati, attraverso l'implementazione di algoritmi, può rivelarsi un passaggio estremamente efficace per produrre conoscenza utile ai processi decisionali coinvolti nell'iter di applicazione dell'agricoltura di precisione.
Obiettivo del presente lavoro è illustrare come gli algoritmi di DM implementati nell'analisi di grandi *dataset* come quelli che provengono dai monitoraggi a monte dell'agricoltura di precisione possono, attraverso il riconoscimento di *patterns* significativi, facilitare e valorizzare il processo di sviluppo di modelli previsionali utili alla gestione sostenibile dell'attività agricola.
Di seguito sarà proposta una metodologia speditiva e supervisionata per procedere alla clusterizzazione di un'area sottoposta a coltivazione sotto l'egida dell'agricoltura di precisione.

MATERIALI E METODI

Agricoltura di precisione e problematiche connesse

L'agricoltura di precisione può essere descritta come un processo articolato in alcune fasi:
- la raccolta dei dati (in genere estremamente eterogenei);
- l'elaborazione e la definizione di un modello spaziale dell'area interessata;
- la gestione automatizzata dei cicli che concorrono alla coltivazione attraverso una mappa delle prescrizioni;
- aggiornamento dei dati (Fig. 1)

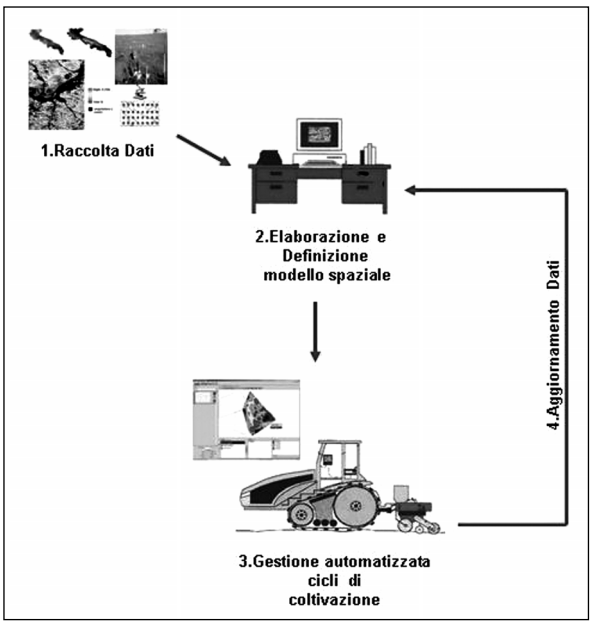

Fig. 1 - Schematizzazione fasi in agricoltura di precisione.
Fig. 1 - Scheme of the phases of precision farming.

Come è facile intendere, si tratta di una procedura di tipo iterativo poiché ad ogni intervento sul campo le informazioni disponibili saranno aggiornate ed il modello spaziale dell'area interessata eventualmente rimodulato in base agli aggiornamenti introdotti.

In questa prima fase si distingueranno due tipologie di dati:

– dati puramente spaziali (coordinate, tutti i dati estraibili da un DEM oppure un DTM) che sono invarianti rispetto al tempo;
– dati spazio-temporali (pioggia, umidità dell'aria, evapotraspirazione, umidità del suolo, contenuto di sostanza organica, etc) che sono quelli che necessitano di essere costantemente aggiornati.

Dal punto di vista dei costi di gestione, la fase di raccolta dati rappresenta un onere del processo descritto, anche se, evidentemente i costi possono differire a seconda dei casi.

I costi relativi alla disponibilità dei dati puramente spaziali sono ovviamente da considerarsi costi *"una tantum"* solo alla partenza del processo, mentre i dati spazio-temporali necessitano di un continuo aggiornamento e quindi, rappresentano costi che si rinnovano nel tempo.

La raccolta dei dati utili si può avvantaggiare di tecniche di acquisizione diretta o differita.

L'acquisizione diretta può avvenire tramite sensori multiparametrici, installati su autoveicoli dotati di GPS e di una guida per effettuare "strisciate" rettilinee, con la capacità di misurare i parametri di interesse in tempo reale con un elevato livello di dettaglio spaziale.

Fra le tecniche di acquisizione dati differita importanti fonti di informazioni ci provengono dal *remote data sensing* e dalle immagini telerilevate catturate attraverso sensori presenti su satelliti o aerei che consentono di valutare caratteristiche chimico-fisiche dei suoli e/o parametri fisiologici dei tessuti vegetali.

Una volta acquisite tali informazioni di dettaglio l'area coltivata, viene suddivisa in aree omogenee rispetto ai parametri monitorati e vengono sintetizzate delle "mappe di prescrizione" che, noti i parametri del tipo di coltivazione interessata, contengono le strategie di intervento, da abbinare a ciascuno dei *cluster* individuati nell'area, espresse in un formato numerico adeguato a pilotare macchinari *computer aided*. L'obiettivo di tale sforzo è quello di uniformare l'intera area agricola nel rispetto dei parametri colturali, fornendo il giusto apporto di risorse in ogni punto dell'area ed ottenere, alla fine, la massima resa.

Il DM e la *cluster analysis* supervisionata

Nella fase di monitoraggio i dati vengono organizzati secondo l'usuale struttura delle matrici *casi x variabili*, in cui la riga rappresenta la batteria di misure effettuata in corrispondenza di un punto dello spazio, mentre le colonne rappresentano tutti i valori misurati di una specifica grandezza (es. la concentrazione di nitrato nel suolo). L'approccio consueto quando si vogliono individuare aree omogenee dal punto di vista spaziale è quello dell'applicazione della geostatistica (Chiles, and Delfiner, 1999).

La geostatistica è una disciplina che studia la variabilità spaziale di grandezze che presentano dei *patterns* di autocorrelazione, ossia che obbediscano alla cosiddetta "prima legge della geografia", secondo la quale due punti spazialmente prossimi mostrano in genere di avere attributi molto più simili rispetto a punti spazialmente distanti (Tobler, 1970). Tale studio del comportamento spaziale viene condotto mediante variografia o "analisi strutturale" che attraverso l'osservazione del comportamento dei punti campionati, attribuisce un modello teorico in grado di descrivere la variazione media dei valori assunti dalla variabile, in coppie di punti generici in funzione della distanza (Chiles, and Delfiner, P., 1999).

La geostatistica offre, dunque, un quadro metodologico sicuramente potente e flessibile per condurre interpolazioni di variabili autocorrelate in punti non campionati ma necessita, per contro, di approfondite conoscenze possedute solo da tecnici altamente specializzati (ad es. la teoria delle variabili regionalizzate Matheron, 1965).

La proposta avanzata in tale lavoro è quella di sostituire, per i fini di cui ci si propone, la geostatistica con la *cluster analysis* supervisionata (Raykov and Marcoulides, 2008) che è più semplice dal punto di vista metodologico e non necessita conoscenze matematiche particolarmente approfondite anche se si tratta di una metodologia di statistica multivariata.

I metodi supervisionati sono strutturati nel seguente modo:

il set di dati raccolti viene suddiviso in parti prive di sovrapposizioni (subsets): $<D_1, D_2>$; il primo subset, D_1, viene utilizzato nella prima fase di *training* in cui il sistema viene appunto addestrato a riconoscere i patterns significativi sulla base di esempi dati. Il secondo subset, viene utilizzato nella seconda fase di validazione in cui, il modello addestrato nella fase di *training*, fa delle previsioni che vengono confrontate con il subset D_2, attraverso indicatori di aderenza (*fitting*); se il modello supera i controlli di aderenza viene accettato e utilizzato per fare previsioni altrimenti si ritorna alla fase di addestramento e così via.

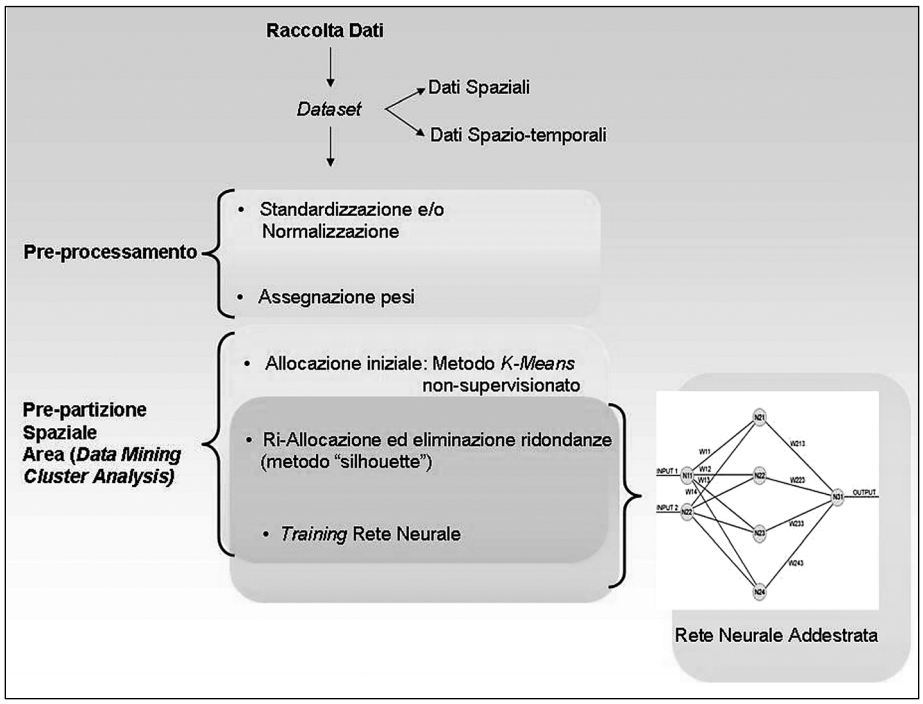

Fig. 2 - Struttura della metodologia proposta.
Fig. 2 - Structure of the proposed methodology.

L'approccio previsto nel presente lavoro propone la creazione di un modello costruito con metodo supervisionato in cui, nella prima fase di training l'applicazione di un algoritmo di DM non supervisionato, un *clustering*, fornisce una prima approssimazione della partizione dell'area di interesse. Nella seconda fase, l'intervento di un operatore corregge le attribuzioni di punti dello spazio a *cluster* ritenute inesatte e procede, nel caso, eliminando *cluster* ridondanti. Questa seconda fase può essere facilmente condotta attraverso strumenti visuali, ad esempio su mappe e cartografia GIS.
L'algoritmo prescelto per condurre l'approssimazione della partizione della prima fase dell'area è il classico algoritmo *k-means* (MacQueen, 1967), basato sulla minimizzazione delle varianze interne ai singoli *cluster* e sulla massimizzazione delle differenze tra le varianze inter-cluster. Questo algoritmo possiede un apprezzabile pregio, che lo rende particolarmente adatto al nostro obiettivo di pre-partizione dell'area; è in grado infatti di omogeneizzare in maniera ottimale i singoli cluster, popolandoli iterativamente di individui estremamente simili tra loro e al contempo di differenziare significativamente fra i vari cluster. Lo svantaggio nell'uso del metodo proposto risiede nell'inserimento fra i dati di ingresso, anche del numero di *cluster*; operazione questa che presuppone da parte dell'operatore una conoscenza a priori del numero di parti in cui suddividere l'area di interesse. Tale svantaggio si può superare definendo un numero di cluster superiore a quello realistico in modo che l'operatore, nella fase di training della clusterizzazione potrà, in base alla propria conoscenza del territorio, accorpare tutti i cluster ridondanti.
La definizione del numero iniziale di cluster utile per innescare il metodo k-*means* viene suggerito in letteratura dalla seguente formula:

$$k \approx (n/2)^{1/2}$$

dove n rappresenta il numero di record da cui è composto il dataset (Mardia *et al*., 1979).
Tuttavia l'applicazione di tale formula, a causa del campionamento di dettaglio effettuato, fornirebbe un numero di cluster iniziale irrealistico e, di conseguenza, appesantirebbe il volume di calcoli da eseguire. Nel presente lavoro si propone un iter alternativo basato su un concetto di buon senso: poiché la clusterizzazione avviene nello spazio reale, si definisce la "parcella minima" come l'unità areale minima da gestire e si pone il numero iniziale di cluster k uguale a

$$k \approx Area\ Tot./parc.\ min$$

dove *Area Tot*. è l'area di interesse e *parc. min* è il valore della "parcella minima" che l'operatore si farà carico di definire.

Nella seconda fase di perimetrazione e riallocazione dei cluster, bisognerà far attenzione a vincolare tali raggruppamenti in modo geograficamente coerente, ossia facendo in modo che punti prossimi cadano all'interno degli stessi cluster. Questo effetto si può ottenere in vari modi, ad esempio, semplicemente utilizzando nella matrice *dati x variabili* anche le coordinate geografiche.

Nella fase di riallocazione ed eliminazione supervisionata dei cluster ridondanti, la metodologia proposta utilizza il "metodo *silhouette*", così come documentato in letteratura, attraverso il quale è possibile interpretare e validare l'analisi cluster condotta (Rousseeuw, 1987).

La fase di riallocazione supervisionata dei cluster si riferisce ad una situazione in cui l'algoritmo non è "a regime" e necessita di una fase di *training* per addestrare la rete neurale sottostante. Quest'ultima, nelle successive applicazioni, provvederà a fornire in maniera del tutto automatizzata l'esatta partizione dell'area indagata.

Va inoltre ricordato che al fine di non far sbilanciare l'analisi verso dati che possiedono valori assoluti più grandi, tutto il *dataset* necessità di *standardizzazione* (o normalizzazione) un'operazione che ha il doppio pregio di rendere le misure adimensionali e di farle variare grossomodo tra -1 e +1 (Bernstein and Bernstein, 2003). È possibile, attribuire in un secondo momento ai dati così riscalati un fattore moltiplicativo w_i detto peso, compreso tra 0 ed 1 e tale che $\sum w_i = 1$ in modo da conferire nell'analisi più importanza ad una certa variabile piuttosto che ad un'altra ovvero creare ove necessario una gerarchia di importanza delle variabili.

Tutti i dati raccolti durante le campagne di campionamento concorrono alla definizione dei raggruppamenti, per cui il tipo di analisi oltre che spaziale è anche dinamica. Nel caso in discussione sono almeno due le clusterizzazioni sovrapposte da realizzare che, come già detto, riguardano le aree omogenee dal punto di vista del contenuto della sostanza organica nel suolo e quelle omogenee per contenuto idrico.

La metodologia fin qui descritta può dunque essere presentata come una procedura strutturata in fasi, in cui ogni fase prevede un ventaglio di possibili azioni da intraprendere secondo una successione temporale; dalle azioni intraprese derivano delle nuove situazioni che consentono di decidere quale, tra il ventaglio di azioni alternative disponibili, è quella da intraprendere, tornando eventualmente sui propri passi se nuove informazioni sul territorio fossero disponibili (Fig. 2).

DISCUSSIONE E CONCLUSIONI

Nell'ambito della gestione integrata ed in particolare in quello dell'agricoltura di precisione il valore aggiunto apportato dall'uso delle tecniche di DM deriva dall'efficienza e dall'accuratezza con cui le sue metodologie ed i suoi algoritmi riescono a gestire i *dataset* facilitando l'identificazione delle zone omogenee rispetto a certe variabili target all'interno di un'area molto più ampia che è, nel caso in specie, la superficie agricola aziendale.

Nel presente lavoro si propone una metodologia, inseribile in un *Decision Support System* (DSS), per la partizione della superficie di un'azienda agricola in aree (*cluster*) omogenee dal punto di vista del contenuto idrico del suolo e del contenuto di materia organica attraverso l'applicazione di un metodo di *clustering* supervisionato organizzato in due fasi: una prima in cui viene eseguita una pre-approssimazione della partizione; a seguire una seconda fase in cui l'intervento manuale di un operatore, unicamente dotato buona conoscenza del territorio, corregge eventuali errori della prima fase e soprattutto, elimina i cluster ridondanti definiti sempre nella precedente fase. Al momento, le variabili considerate sono esclusivamente di tipo numerico, per la precisione, "ad intervalli" (Bernstein and Bernstein, 2003) che sono le variabili più semplici da sottoporre al tipo di analisi proposto. In un futuro sviluppo del presente lavoro, sarà proposto un adattamento della prima fase di partizione approssimata della metodologia, che possa coinvolgere anche variabili di tipo non numerico e non ordinale (categoriale sconnesso) più complicate da trattare. Il pregio della metodologia proposta rispetto ad altre risiede nella sua facilità di applicazione che la rende di facile gestione anche per operatori a digiuno di conoscenze nel campo della statistica matematica e della geostatistica.

BIBLIOGRAFIA

Bernstein S., Bernstein R., 2003. Statistica descrittiva. McGraw-Hill.

Chiles J.P., Delfiner P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley Inter-science.

Colby Torbett J., Roberts R.K., Larson J.A., English B.C., 2007. Perceived importance of precision farming technologies in improving phosphorus and potassium efficiency in cotton production Precision Agriculture 8: 127–137

Hand D., Mannila H., Smyth P., 2001. Principles of Data Mining. The MIT Press, Cambridge Massachusetts London England.

MacQueen J.B., 1967. Some Methods for classifi-

cation and Analysis of Multivariate Observations. In Proc. of the 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press.: 281-297.

Mardia K., Kent J.T., Bibby J., 1979. Multivariate analysis. Academic Press.

Matheron G., 1965. Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature, Masson, Parigi, Francia.

National Research Council, 1997. Precision Agriculture in the 21st Century, Geospatial and Information technologies in Crop Management. National Research Academy Press: 16-19.

Pierce F.S., Nowak P., 1990. Aspects of precision agriculture. Advances in Agronomy, 67: 1-86.

Robert P.C., 2002. Precision Agriculture: a challenge for crop nutrition management. Plant and Soil, 247, 1: 143-149.

Pringle M. J., Mcbratney A. B., 2004. Field-Scale Experiments for Site-Specific Crop Management. Part II: A Geostatistical Analysis, Precision Agriculture, 5: 625-645.

Raykov T., Marcoulides G.A., 2008. An Introduction to Applied Multivariate Analysis - Taylor & Francis group.

Rousseeuw P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20: 53-65.

Tobler W., 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(2): 234-240.