NOTA METODOLOGICA

L'INDICATORE DI DINAMISMO STRATEGICO: UN'APPLICAZIONE DELLE METODOLOGIE ACM E *RANDOM FOREST* ALLE PRIME DUE EDIZIONI DEL CENSIMENTO PERMANENTE DELLE IMPRESE[1]

Le due edizioni del Censimento permanente delle imprese (riferite, rispettivamente al 2018 e al 2022) contengono ampie informazioni sulle scelte strategiche delle unità produttive. L'indicatore di dinamismo strategico ha l'obiettivo di sintetizzare tale insieme informativo considerando allo stesso tempo un ventaglio molto ampio di scelte, relative a condizioni operative, andamento del fatturato, relazioni produttive con altre imprese ed enti, fonti di finanziamento, gestione del personale, transizione digitale, scelte di investimento e, infine, criticità sofferte e strategie adottate.

Il confronto tra i risultati delle due rilevazioni consente quindi di studiare le relazioni esistenti tra le variabili (ad esempio i cambiamenti prodotti dall'emergenza sanitaria *COVID-19* nelle strategie delle imprese), attraverso modelli di comportamento che consentano una rappresentatività generale del fenomeno esaminato. Tuttavia, la struttura del questionario e il campione delle due edizioni non sono gli stessi (il disegno campionario non prevede l'estrazione di un panel di unità); di conseguenza, per garantire la possibilità di un confronto temporale tra i risultati quale quello utilizzato in questo Rapporto, si rende necessario il ricorso a una procedura di omogeneizzazione.

A tal fine si utilizza un approccio sequenziale ("*Tandem approach")* di tecniche di analisi dei dati che realizzano ordinamenti e classificazioni multidimensionali: a) modelli e metodi fattoriali; b) metodi di *clustering*, che forniscono una classificazione automatica (non supervisionata) con l'obiettivo di individuare tipi o gruppi, ottimali secondo una prescelta funzione obiettivo.

a) *Analisi delle corrispondenze multiple*. Con riferimento ai modelli fattoriali, il primo passaggio è costituito dallo studio delle relazioni attraverso un'analisi delle corrispondenze multiple (ACM), una tecnica di analisi statistica multivariata a carattere esplorativo volta ad analizzare l'esistenza di schemi di associazione tra variabili qualitative, attraverso l'identificazione di uno spazio "ottimale", di dimensione ridotta, sintesi dell'informazione strutturale contenuta nei dati originari. In particolare, questa tecnica si applica quando si è interessati a estrarre dai dati l'informazione utile, in termini di similarità tra gli elementi appartenenti a ciascuno dei due insiemi. Tale similarità si osserva attraverso la rappresentazione fattoriale della configurazione o forma delle nuvole dei punti, associate a tali insiemi. Il pattern è costituito dall'insieme delle distanze riprodotte su un piano fattoriale e fornisce, a un tempo, una visione sintetica e globale delle relazioni tra i punti (volta cioè a comprendere le relazioni strutturali presenti nel fenomeno) e una lettura analitica sui particolari aspetti di queste relazioni (volta a descrivere ciascuna relazione strutturale).

b) *Clustering*. Il secondo passaggio consiste in una procedura di clusterizzazione articolata nelle seguenti fasi:

1) individuazione della matrice di dati e standardizzazione delle variabili;

1 Nota metodologica redatta da Stefano De Santis.

2) scelta dei criteri di classificazione da applicare ai dati (agglomerativo o scissorio);
3) valutazione del risultato ottenuto, consolidamento delle partizioni e interpretazione della tassonomia ottenuta.

Nell'applicare tale metodologia alle due edizioni del Censimento permanente sulle imprese, il punto 2 è stato preceduto da una fase esplorativa, realizzata mediante una serie di *k-means*, con numero di gruppi variabile da 9 a 2, ognuno dei quali ottimizzato con una serie di *random starts* (in ragione di 100). La partizione ottimale è risultata essere costituita da 5 gruppi, che sono stati preliminarmente valutati per verificare l'esistenza di partizioni dei suddetti elementi in specifiche "classi di equivalenza" multidimensionali. Per limitare gli effetti delle scelte preliminari e dei vincoli che sia le procedure gerarchiche sia quelle non gerarchiche impongono al risultato di una classificazione automatica, si è optato per una tecnica di classificazione "mista", realizzata mediante:

- produzione di una classificazione fine con un numero elevato di classi provvisorie (rapporto unità/nuclei 1:100), ottenuta mediante un algoritmo non gerarchico (*k-means* – distanza euclidea);
- definizione della tassonomia finale mediante applicazione di un metodo gerarchico (distanza di Ward) valutando convenientemente il salto ottimale (criterio del salto minimo) al fine di ottenere il minimo numero di gruppi con massima omogeneità interna; l'esame del dendrogramma permette infatti di conoscere la similarità tra i nuclei della classificazione fine, ricavati nella fase precedente;
- consolidamento della tassonomia finale mediante una procedura non gerarchica a centri mobili che ottimizza, attraverso una riclassificazione di tutti gli elementi, il risultato della classificazione gerarchica. Questo consolidamento può solo migliorare le classi già ottenute: se gli elementi fossero già adeguatamente classificati non si otterrebbe nessuno spostamento da un gruppo a un altro e il risultato non cambierebbe.

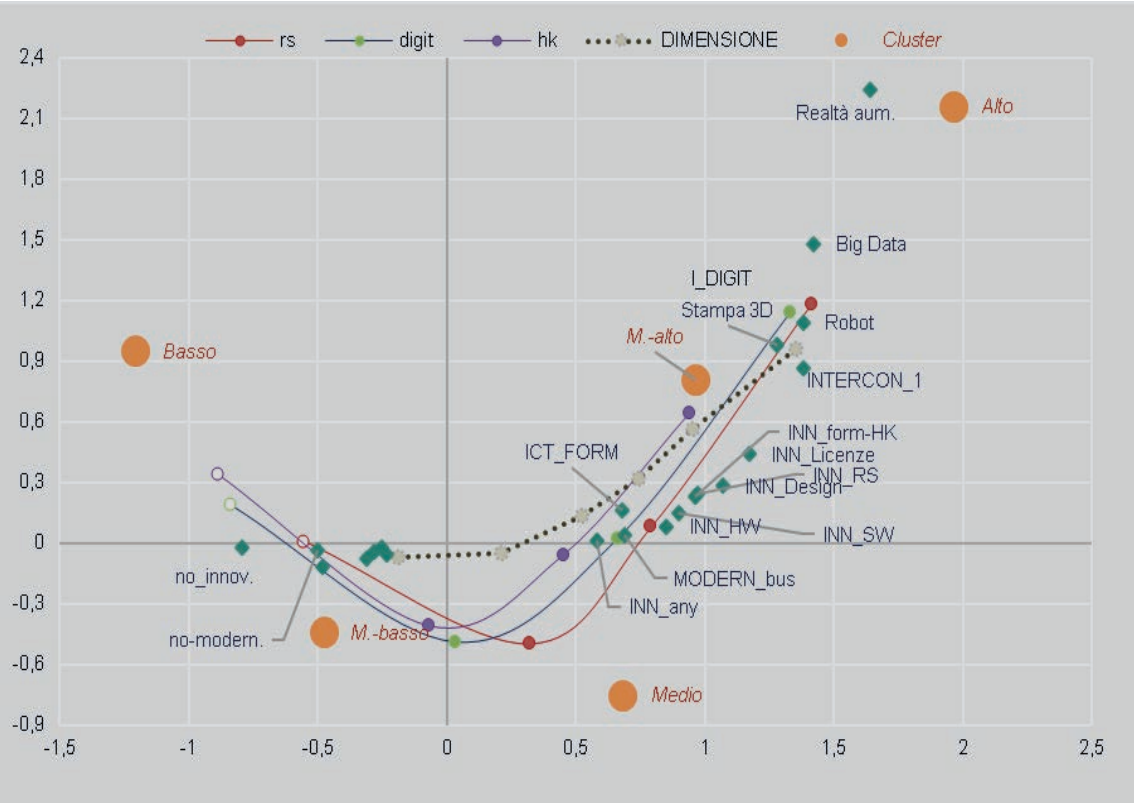
Il numero dei gruppi ottenuto e il loro significato economico sono risultati coerenti con quelli emersi dalla procedura esplorativa. La strategia mista, tuttavia, risulta migliore in termini di incremento del rapporto tra la varianza inter-gruppo e quella totale della seconda classificazione rispetto a quella preliminare. Data la natura gerarchica del sistema di clusterizzazione, si è sfruttata l'informazione del dendrogramma conservando anche un "taglio" superiore in corrispondenza della tassonomia a 3 gruppi, in modo da avere diversi livelli di dettaglio (a 5 e 3 gruppi, congruenti tra loro) della medesima classificazione. Infine un'ulteriore procedura di clusterizzazione "*fuzzy*" ha confermato la suddivisione della popolazione in 5 gruppi fornendo, oltre a una nuova evidenza sulla robustezza del risultato, anche informazioni relative al *degree* di appartenenza delle imprese a ciascun gruppo.

La clusterizzazione si è limitata ai primi due fattori, dato l'elevato potere risolutorio del piano (prime due dimensioni), in ragione di due evidenze:

- i primi due fattori rappresentano quasi il 90 per cento della variabilità lineare del fenomeno complesso, essendo perciò rappresentativi di (quasi) tutto il fenomeno multivariato;
- sussiste una corrispondenza tra forma della nuvola dei punti e strutturazione dei dati nella matrice, struttura peraltro abbastanza ricorrente, rappresentata da una forma paraboloide della nuvola dei punti ("effetto Guttman") che evidenzia una disposizione degli elementi di riga e colonna lungo un unico *continuum*; tale struttura rivela sia l'esistenza di una relazione tra i caratteri, sia quella di un primo fattore dominante, nonché di assi successivi che ne rappresentano funzioni d'ordine superiore (il secondo fattore è una funzione di secondo grado, il terzo di terzo grado e così via).

Nelle mappe fattoriali sottostanti sono rappresentate la distribuzione dei punti modalità e i baricentri dei 5 cluster individuati sulla base dei dati analizzati (Figura 1), in cui la disposizione dei punti anticipa la forma paraboloide tipica dell'effetto Guttman (che sarà poi evidente nella rappresentazione dello spazio delle unità). L'interpretazione dei fattori conduce all'individuazione della dimensione strutturale latente.

Figura 1 - Rappresentazione delle variabili (punti modalità) sulla mappa fattoriale e suddivisione in classi di dinamismo. Anno 2018

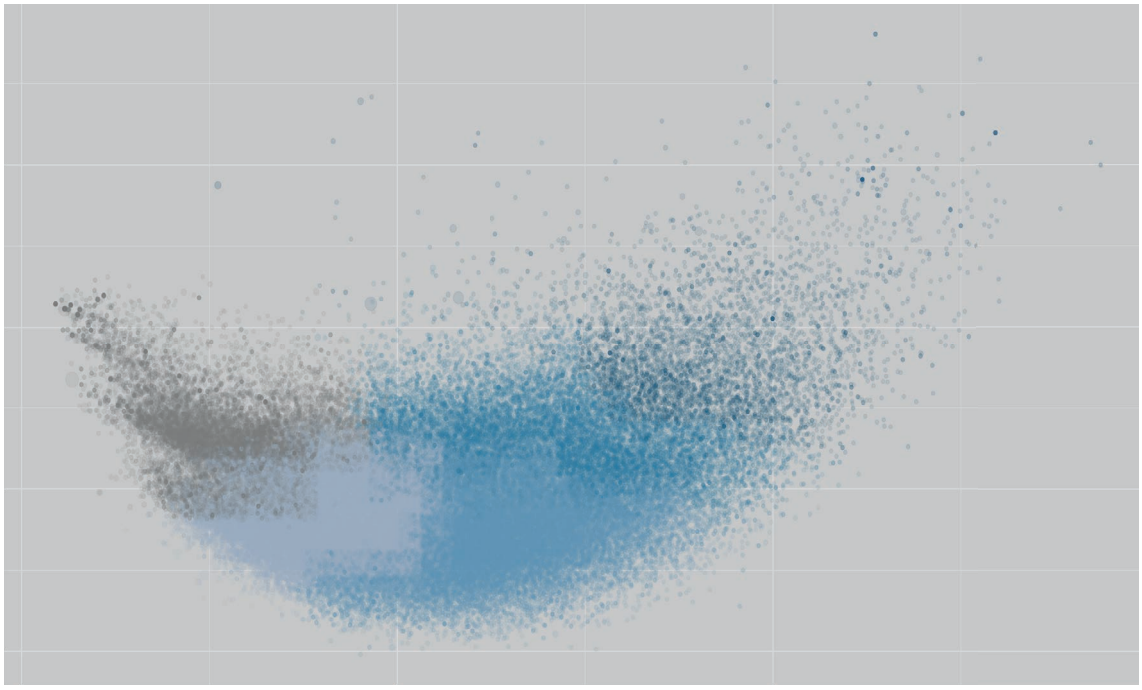


Fonte: Elaborazioni su dati Istat, Censimento permanente delle imprese

Il *continuum* dei punti (sia modalità, sia unità, qui riassunti dai baricentri dei cluster) evidenzia il progressivo passaggio delle imprese da livelli di dinamismo bassi o nulli a livelli via via più elevati: il primo fattore contrappone quindi, ai propri estremi, le unità non dinamiche rispetto a quelle più dinamiche, graduando in tutto il *continuum* le imprese per livelli crescenti di dinamismo. Il secondo fattore, essendo una potenza di ordine 2 del primo ("effetto Guttman"), fornisce informazioni circa il carattere esponenziale di questo passaggio.

Dall'esame della successiva Figura 2 è evidente la disposizione paraboloide dei punti unità (imprese), congruente con quanto visto per la precedente rappresentazione dei punti modalità (strategie). Si ricorda, *inter alia*, che le rappresentazioni dello spazio delle modalità e delle unità costituiscono un duale, essendo cioè la rappresentazione del medesimo fenomeno.

Figura 2 - Rappresentazione delle imprese (punti unità) sulla mappa fattoriale e suddivisione in classi di dinamismo. Anno 2018



Fonte: Elaborazioni su dati Istat, Censimento permanente delle imprese

Tuttavia, uno dei problemi delle classificazioni *data-driven*, che ne limita l'utilizzo soprattutto in una logica longitudinale, è legato al fatto che sia la tassonomia (variabile di classificazione) sia le statistiche ottenute dal suo utilizzo (conteggio, somma, media, eccetera) sono entrambe definite dai medesimi dati. Tendono perciò a "muoversi" al variare dei dati, creando ovvie difficoltà nell'analisi diacronica, dove i confronti devono essere effettuati a invarianza di classificazioni al fine di evidenziarne l'evoluzione temporale. L'impiego di soluzioni *data-driven* richiede perciò di assolvere una serie di accorgimenti aggiuntivi per poter essere utilizzati in una classica lettura longitudinale dei fenomeni. In primo luogo, occorre verificare la natura statica o dinamica dei fenomeni che hanno generato la tassonomia; nel secondo caso la rappresentazione longitudinale va resa coerente con il quadro complessivo del cambiamento[2]. Il carattere statico del fenomeno può essere testato con apposite metodologie[3]. Nel caso sia confermata la stabilità del fenomeno, è necessario procedere per passi successivi, corrispondenti ciascuno a ogni passaggio della strategia di analisi complessa, per cercare di sterilizzare l'effetto metodologia e fare in modo che, nei confronti temporali, eventuali differenze nei dati evidenzino esclusivamente l'evoluzione del fenomeno.

Tornando alla applicazione delle fasi precedenti alla analisi multidimensionale relativa alle due edizioni del Censimento permanente sulle imprese, si è inizialmente

2 I fenomeni multivariati evolutivi sono piuttosto rari, specie nel breve-medio periodo. Il complesso delle relazioni strutturali tra variabili tende infatti ad avere un quadro generalmente stabile.

3 L'insieme delle statistiche che, con livelli di sintesi differente, riportano tali informazioni fanno parte della reportistica delle analisi *multiway*. In linea di principio, si può prendere il valore della correlazione tra intere matrici, o con maggior dettaglio tra coppie di fattori ordinate (correlazione tra i primi, secondi, terzi (e successivi) fattori delle analisi *cross-section*) come *proxy* della staticità del fenomeno. In particolare, si può analizzare il peso relativo delle modalità delle variabili (in caso di fenomeni discreti e relativa analisi delle corrispondenze multiple) sulla traccia standardizzata della matrice di varianze/covarianze per evidenze di maggior dettaglio.

costruito un longitudinale tra le domande dei due questionari, per poi verificare, per ogni singola variabile ricostruita, la presenza di eventuali variazioni. Tra le due edizioni del questionario si evidenzia qualche modifica: tuttavia, con riferimento ai quesiti selezionati per l'Analisi in componenti multiple (ACM) nel 2018, nel complesso non si segnalano grandi cambiamenti, con l'eccezione della sezione sulla Sostenibilità, soggetta invece a una più importante ristrutturazione (che verosimilmente comporta un *break* strutturale dei quesiti). Le descrittive a livello di singolo quesito non riportano sostanziali differenze tra le due edizioni.

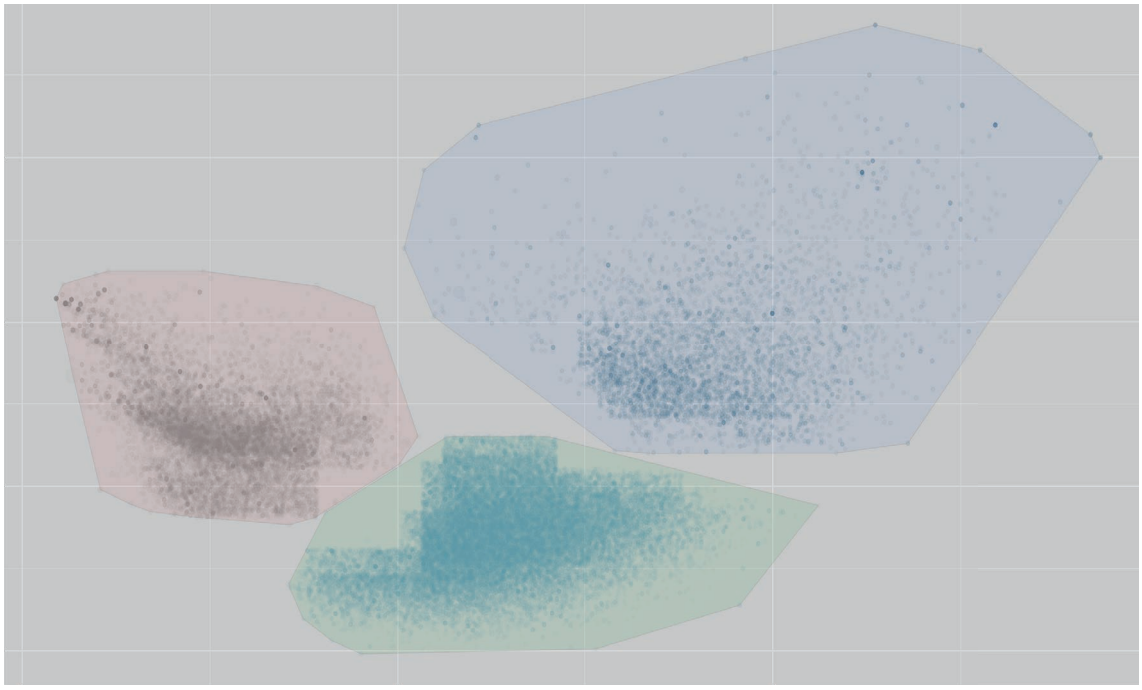
Inizialmente la procedura di ACM del 2018 è stata iterata applicandola alle "nuove" variabili presenti nel dataset comune alle due edizioni (il longitudinale 2018-2022). Prima di effettuare il confronto temporale è stato infatti necessario comparare le due soluzioni ora disponibili per il 2018 (quella effettuata in occasione della prima edizione della rilevazione e quella, effettuata sempre sulle informazioni del 2018, basata sulle variabili presenti nel longitudinale 2018-2022) per valutarne il grado di similarità. Le due ACM sono risultate sostanzialmente equivalenti (le correlazioni tra i rispettivi due fattori sono altissime: 99,4 per il primo; 94,7 per il secondo), confermando la possibilità di un confronto intertemporale.

Successivamente si è iterata anche la procedura di clusterizzazione. Tale procedura, tuttavia, nel 2018 era composta da più step, alcuni dei quali non esattamente replicabili poiché legati ai *random starts* degli algoritmi *k-means* utilizzati: cambiando il punto di partenza, cambierebbe il risultato finale. Replicare esattamente la procedura di *clustering* rappresenta tuttavia un *nonsense*. Infatti la scelta di una tassonomia in un sottospazio a *n* dimensioni, in base a qualsiasi criterio di ottimalità, corrisponde in sostanza alla definizione di soglie che definiscono la partizione del sottospazio in un determinato numero di sottoinsiemi. Non è un caso, infatti, che i termini tassonomia e partizione siano usati come sinonimi. In caso di *hard clustering* (tassonomie disgiunte, senza sovrapposizione i gruppi individuati dalla tassonomia) è perciò possibile affermare che, nel caso di analisi univariata, il *clustering* segmenta una retta; nel caso di analisi bivariata individua porzioni disgiunte di un piano; in uno spazio tridimensionale individua una partizione composta da volumi distinti; più in generale in uno spazio a *n* dimensioni individua perciò un insieme disgiunto di ipercubi.

Nel caso in esame, si tratta di una partizione del piano fattoriale individuato dalle prime due componenti, le cui soglie individuano i sottoinsiemi propri del collettivo costituiti dai cluster. Tali soglie sono scelte in base ai criteri di ottimalità sopra descritti; pertanto, in caso di invarianza del fenomeno multivariato (e della relativa soluzione fattoriale) non sarebbe giustificabile un loro cambiamento, soprattutto in un'ottica di confronto intertemporale.

L'individuazione delle soglie su un piano richiede invece la definizione dell'insieme convesso delle unità che costituiscono la frontiera di ciascun cluster (poligoni di Voronoi) e la successiva applicazione sul piano fattoriale dei "nuovi" valori del 2018 e del 2022.

Figura 3 - Tassonomia del "vecchio" dinamismo 2018: imprese a basso, medio e alto dinamismo e relativi poligoni di Voronoi



Fonte: Elaborazioni su dati Istat, Censimento permanente delle imprese

Nel fare ciò, i poligoni di Voronoi sono stati approssimati attraverso un albero decisionale (*random forest*), una metodologia di particolare efficacia nelle applicazioni "*out of sample*". Infatti, un albero di regressione realizza, esattamente come una procedura di *hard clustering*, una partizione ottimale dello spazio a *n* dimensioni (Breiman, 2001).

Poiché i cluster sono definiti unicamente sui valori delle due coordinate fattoriali, il modello basato su albero di regressione ricostruisce per approssimazione, attraverso l'elevato numero di "foglie" (555 *split* a comporre la partizione finale del piano fattoriale, con una concordanza tra valori reali e predetti del 98,1 per cento), i 5 poligoni di Voronoi corrispondenti ai cluster. Il modello applicato ai "nuovi" dati 2018 (ossia sulle coordinate fattoriali 2018 del longitudinale), evidenzia una elevata accuratezza (85,6 per cento), ma inferiore alla precedente, in virtù della correlazione più bassa tra i secondi fattori omologhi.

Reinserendo nei modelli ad albero anche le variabili di partenza utilizzate nella ACM, si ricostruisce un modello valido quanto il precedente (tasso di accuratezza: 96,5 per cento): giocano un ruolo sussidiario le variabili delle sezioni che presentano quesiti non comuni alle due edizioni del Censimento; nella ricostruzione dei poligoni di Voronoi, in particolare, tali variabili suppliscono a quelle cadute a causa della costruzione del longitudinale. È stato testato un duplice modello di albero per la definizione del modello con quale effettuare le stime *out of sample*: un primo in *cross validation*, un secondo con la classica suddivisione del campione in test e *validation set* (in ragione dell'80 e del 20 per cento), ottenendo risultati virtualmente identici, confermando la robustezza dei risultati.

A questo punto si è applicata l'ACM sulle due edizioni del Censimento. Come conseguenza della staticità del fenomeno nel tempo, i primi due fattori appaiono assai simili ma non identici, ivi inclusi i pesi delle combinazioni lineari. Al fine di sterilizzare completamente questo effetto, e di evitare anche minime differenze legate a variazioni intertemporali si sono quindi applicati i coefficienti della ACM del 2018 (*prediction out of sample* sui dati 2022). È stato infine applicato il modello dell'albero di regressione per approssimare i poligoni di Voronoi sul dataset 2022 e stimare le classi di dinamismo nella edizione 2022 del Censimento permanente.