UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI ECONOMIA " GIORGIO FUA' "

Corso di Laurea Magistrale in Economia e Management

Gli indici di deprivazione per rischio aziendale

Deprivation indicators in risk management

Relatore
Prof.ssa Maria Cristina Recchioni
Correlatore
Prof. Simone Poli

Rapporto finale di
Laura Iura

Anno Accademico 2020-2021

Gli indici di deprivazione per rischio aziendale

Laura Iura

Ottobre 2021

Indice

Introduzione

L'arrivo del nuovo codice della crisi d'impresa e dell'insolvenza (D. Lgs. 14/2019) ha responsabilizzato amministratori e sindaci delle società nell'attuazione delle procedure di allerta, in caso di subentro di una crisi, ponendo sempre più l'attenzione sulla prevenzione dell'insolvenza. Tali disposizioni legislative hanno portato quindi a voler migliorare i modelli di previsione di *default* che, già da molti anni, erano stati costruiti con lo scopo di essere implementati da analisti americani, interni all'azienda, per verificare il giusto andamento della propria gestione.

Viste le attuali necessità, non solo degli amministratori, ma anche degli stakeholders, di studiare lo stato di "salute" della stessa in letteratura nascono nuovi modelli di previsione. Tra questi, verranno evidenziati i modelli costruiti da Poli, i quali hanno lo scopo di rispondere il più possibile alle esigenze degli analisti esterni, ed in particolare, di quelli italiani. Infatti, esso permette non solo di effettuare un'analisi sulla solvibiltà delle imprese, ma anche di svolgerla partendo da un bilancio redatto in forma abbreviata, il quale viene pubblicato dalla maggioranza delle imprese nazionali.

Dato lo scopo di individuare ulteriori sistemi di indagine che integrino gli attuali modelli di previsione, essi verranno prima descritti nelle loro peculiarità e modalità di implementazione, e poi verranno applicati sui dati di partenza utilizzati nei modelli costruiti da Poli per consentirne la comparabilità.

Quindi, questi ultimi si pongono come elemento di confronto non solo in quanto più recenti e di più larga applicazione sul territorio nazionale, ma soprattuto perché ne sono stati messi a disposizione i dati alla base e quindi ciò permette di porli sullo stesso piano di quelli qui sviluppati per la valutazione delle performance.

Tra le analisi proposte, sono stati posti al centro dell'attenzione i cosiddetti indici statistici di deprivazione, ovvero strumenti di analisi tendenzialmente utilizzati in ambiti socio-sanitari per confrontare gli stili di vita tra più classi della società.
La loro applicazione all'interno di questo campo è del tutto innovativa e si caratterizzano per la elevata semplicità di studio, per la non necessità di riservare trattamenti particolari ad elementi estremi della distribuzione e, come vedremo, per il buon livello di stima delle aziende "sane" fornito.
Si vuole quindi puntare molto sul livello raggiunto di specificità (rapporto tra le aziende previste come "sane" ed il totale delle aziende "sane" osservate), piuttosto che su quello di accuratezza (rapporto tra le previsioni corrette ed il totale delle aziende analizzate) in quanto è molto importante evitare che le aziende , in realtà "sane", non vengano poste nella condizione di avviare procedure di allerta non necessarie, pur se ciò tende a penalizzare le capacità di intervento delle aziende in "crisi".
La struttura della tesi prevederà innanzitutto la trattazione dei modelli di previsione della crisi, fornendo un focus su quelli costruiti da Poli, per poi parlare degli indici di deprivazione come strumenti di analisi statistica.
Fornite le conoscenze preliminari necessarie per comprendere quanto alla base dell'analisi empirica svolta in questo lavoro, sarà possibile successivamente procedere con lo sviluppo di alcuni studi presentati ed applicati nei capitoli 3 e 4, considerando come dati di partenza quelli riportati nel paragrafo 1.3 per rendere i modelli confrontabili. Questo consentirà di fornire eventuali incipit futuri per sviluppare, con le analisi qui proposte, i modelli di previsione attualmente esistenti.
Il programma per mezzo del quale verrà implementato tutto il lavoro è *RStudio*, ovvero un software che permette di effettuare calcoli statistici anche molto complessi attraverso solo il download dei pacchetti necessari e l'inserimento di determinati comandi. Durante il proseguo della trattazione empirica verranno riportati alcuni comandi usati per lo svolgimento delle principali funzioni parte delle analisi, mentre nell'Appendice 1 si riportano i codici relativi a tutte le applicazioni statistiche.

Capitolo 1

I modelli di previsione della crisi d'impresa

La questione della previsione è un tema affrontato in letteratura ormai da molti anni e nasce in America con la costruzione dei primi modelli basati sulle peculiarità delle imprese del posto. Il più ricorrente è quello sviluppato per la prima volta negli anni '60 del '900, ovvero quello di Altman [11], ma ne esistono anche altri come il modello di Ohlson del 1980 [33] ed il modello di Zmijewski del 1984 [41]. Tuttavia, dato che l'efficacia di un modello aumenta se costruito sulla base delle caratteristiche delle aziende del paese, anche in Italia sono stati proposti vari studi quali i modelli di Ciampi del 2015 [15], di Giacosa-Mazzoleni del 2018 [19] e di Poli del 2020 [35]. In particolare, i modelli alla base del lavoro sono quelli sviluppati nel 2020 da Simone Poli, ricercatore e professore di ragioneria e di economia aziendale presso la facoltà di economia "G. Fuà" dell'Università Politecnica delle Marche.

Dato il periodo recente di costruzione di questi, la possibilità di disporre dei loro dati e la loro larga possibilità di applicazione sul territorio nazionale, nell'elaborato si partirà da tali dati per poi effettuare ulteriori sviluppi attraverso delle analisi statistiche. L'introduzione dei modelli in considerazione, infatti, ha permesso di superare alcune problematiche legate sia al tipo di dati che si debbano avere a disposizione per applicare i modelli di previsione, sia alle esigenze di tempestività di intervento

in caso di rischio di dissesto. Tutto ciò, quindi, li rende molto adatti al contesto italiano non solo per la tipologia di bilanci redatti dalla maggioranza delle aziende ma anche per conformarsi alle ultime disposizioni legislative introdotte con il nuovo codice della crisi d'impresa e dell'insolvenza.
Per comprendere quanto si ritroverà nei capitoli applicativi è quindi neessario quanto prima descriverne il funzionamento.

1.1 L'analisi di bilancio

Alla base dei modelli di previsione della crisi d'impresa si trova l'analisi di bilancio, ovvero, un'indagine di tipo quantitativo che ha come fonte di informazione la sintesi contabile dei valori derivanti dalla gestione aziendale.
Difatti, partendo dal bilancio redatto dalle imprese, si effettua una riclassificazione dei prospetti che ne fanno parte e si calcolano tutta una serie di indici in grado di informare circa l'andamento degli aspetti economici, finanziari e patrimoniali dell'azienda.
Questo tipo di analisi, appunto, ha come finalità quelle di esprimere un giudizio sullo stato di salute dell'azienda e di rendere confrontabili più realtà aziendali [34].
Inoltre, essa rende possibile, non solo l'effettuazione di comparazioni di tipo spaziale, ma anche di tipo temporale e dimensionale, grazie al calcolo di indici quozienti. Tanto è vero che l'analisi di bilancio si costituisce di svariate tipologie di indici quali: gli indici di composizione (confrontano tra loro classi di valori e corrispondenti margini di conto economico), gli indici di situazione (confrontano tra loro valori provenienti da sezioni diverse del medesimo prospetto parte del documento di bilancio), gli indici operativi (confrontano tra loro valori provenienti da differenti prospetti del medesimo bilancio) e gli indici di andamento (confrontano tra loro valori provenienti da successivi bilanci di una medesima azienda) [34].
Accanto a questi indici quoziente è poi possibile avere informazioni in valori assoluto, piuttosto che relativo, attraverso il calcolo dei margini.
L'analisi di bilancio non trova fonte in alcuna disciplina di tipo legislativo, ma varia

a seconda degli obiettivi conoscitivi che si vogliono raggiungere, dell'esperienza professionale e del periodo temporale in cui ci si trova [34].
In questa sede, essa è alla base di molteplici analisi di previsione ed è approfondita tenendo in considerazione come fonte di informazione il bilancio abbreviato, oggetto dei modelli di previsione di Poli.
Per questo motivo, la riclassificazione e l'analisi per indici trattati in questa sezione sono basati su tale tipologia di bilancio.

1.1.1 La riclassificazione dello Stato Patrimoniale abbreviato

La riclassificazione dello Stato Patrimoniale può seguire due metodologie differenti. La prima segue un criterio di tipo finanziario, ovvero le attività vengono suddivise in base al principio della liquidità, mentre le passività in base a quello dell'esigibilità [34].
Le prime, quindi, vengono riclassificate in base al periodo entro il quale si prevedere il ritorno in termini monetari e si suddividono, in ordine di tempo crescente, in:

- liquidità immediate, cioè attività liquidabili in qualsiasi momento (nel brevissimo termine, dove con brevissimo termine intendiamo anche pochi mesi);
- liquidità differite, quindi smobilizzabili nel breve termine (entro l'anno successivo) ma non nel brevissimo;
- disponibilità, comprendono le sole rimanenze di magazzino;
- attivo fisso, comprende tutte le attività liquidabili oltre l'esercizio successivo

Per quanto riguarda il lato delle fonti di finanziamento, invece, si effettua la riclassificazione in base al periodo in cui si prevede un'uscita in termini monetari [34]. In questo caso si avranno le seguenti classi:

- patrimonio netto, ovvero comprende le componenti del patrimonio;
- passivo corrente, contiene tutte le componenti di debito che scadranno entro l'esercizio successivo;

- passivo consolidato, riguarda tutti impegni di debito che prevederanno un'uscita finanziaria oltre l'esercizio successivo.

La riclassificazione dello Stato Patrimoniale di un bilancio redatto in forma abbreviata secondo il criterio finanziario è la seguente [35].

ATTIVO		**PASSIVO**	
B) Immob.		**A) P. netto**	
I - Imm. immat.	Att. fisso	I - Capitale	P. netto
II - Imm. mat.	Att. fisso	II - Ris. sovr. az.	P. netto
III - Imm. finanz.	Att. fisso	III - Riserva riv.	P. netto
Totale imm. (B)		IV - Riserva legale	P. netto
C) Att. circ.		V - Riserve stat.	P. netto
I - Rimanenze	Disp.	VI - Altre ris.	P. netto
II - Crediti		VII - Ris. cop.	
es. entro l'es. succ.	Liq. diff.	flussi fin. attesi	P. netto
es. oltre l'es. succ.	Att. fisso	VIII - U./P. a nuovo	P. netto
III - Att. finanz.		IX - Utile (perd.) es.	P. netto
non imm.	Liq. diff.	X - Ris. neg.	P. netto
IV - Disp. liq.	Liq. imm.	az. proprie in port.	P. netto
Tot. att. circ. (C)		*Totale p. netto (A)*	
		B) F. rischi/oneri	Pass. cons.
		C) TFR	Pass. cons.
		D) Debiti	
		es. entro l'es. succ.	Pass. corr.
		es. oltre l'es. succ.	Pass. cons.
		Totale debiti (D)	
TOTALE ATTIVO		*TOTALE PASSIVO*	

Essa si differenzia da quella svolta su un bilancio ordinario per tre elementi:

- la voce A) dell'attivo è accorpata nella voce CII), inserendola tra i crediti esigibili entro o oltre l'esercizio successivo a seconda del caso;

- anche la voce D) dell'attivo viene riportata tra crediti, nelle liquidità differite ovvero nell'attivo fisso;

- la voce E) del passivo la si ritrova tra i debiti esibili entro o oltre l'esercizio successivo [35].

Esiste anche la riclassifcazione di tipo funzionale, ovvero gli elementi del patrimonio vengono suddivisi in base all'area gestionale a cui fanno riferimento. In particolare si classificano le attività/passività come operative oppure accessorie [34].
Questo secondo criterio, denominato criterio della pertinenza gestionale, tuttavia, è molto poco utilizzato nell'analisi di bilancio [35].

1.1.2 La riclassificazione del Conto Economico abbreviato

Il Conto Economico che parte dal bilancio ordinario si differenzia da quello del bilancio abbreviato in quanto nel secondo caso, è possibile, raggruppare le seguenti voci:

- le voci A2 e A3;

- le voci B9(c), B9(d) e B9(e);

- le voci B10(a), B10(b) e B10(c);

- le voci C16(b) e C16(c);

- le voci D18(a), D18(b), D18(c) e D18(d);

- le voci D19(a), D19(b), D19(c) e D19(d) [35].

Vi sono tre riclassificazioni del Conto Economico, in tutti e tre i casi i costi sono suddivisi in base all'area gestionale cui si riferiscono registrando una differenza nella modalità con cui viene suddivisa l'area operativa [34]. Esse vengono qui di seguito riassunte:

1. *riclassificazione a valore della produzione e valore aggiunto.* I ricavi ed i costi operativi, classificati per natura, vengono suddivisi a seconda che questi siano "interni" o "esterni;

2. *riclassificazione a ricavi e costi della produzione venduta.* I ricavi ed i costi operativi sono distinti in base alla pertinenza gestionale;

3. *riclassificazione a costi fissi e variabili.* La riclassificazione delle voci operative viene effettuata in base alla loro variabilità rispetto al volume di produzione.

La riclassificazione del Conto Economico adottata per svolgere l'analisi di bilancio, ovvero l'anlisi per indici, è tendenzialmente quella *a valore della produzione e valore aggiunto*, in quanto l'unica che può essere svolta da un soggetto esterno [35]. Essa viene riportata nella tabella sottostante.

Valore della produzione
− Costi operativi "esterni"
= **Valore aggiunto**
− Costi per il personale
= **Margine operativo lordo (MOL/EBITDA)**
− Ammortamenti, svalutazioni e accantonamenti
= **Margine operativo netto (MON)**
± Risultato della gestione accessoria
= **EBIT**
− Interessi passivi e altri oneri finanziari assimilabili
= **Risultato ordinario (RO)**
± Risultato della gestione straordinaria
= **EBT**
− Imposte
= **RISULTATO NETTO (RN)**

1.1.3 L'analisi per indici

Sulla base dei prospetti riclassificati è possibile procedere con l'analisi per indici, la quale, prescindendo dai valori assoluti, rende comparabili i dati.
Essa si compone dei seguenti ambiti di indagine: analisi della redditività, analisi della solidità patrimoniale ed analisi della liquidità [34]. Considerando i soli indici necessari per l'implementazione dei modelli di Poli, si descrivono quelli legati al bilancio abbreviato. Per il calcolo degli indici inseriti negli altri modelli si può consultare un qualsiasi testo di analisi di bilancio.

Gli indici di redditività

Gli indici di redditività sono dati da un rapporto tra una specifica configurazione di reddito e la correlata fonte di finanziamento necessaria a generarlo. Tendenzialmente, al numeratore vengono riportati i risultati economici intermedi generati dal conto economico riclassificato, in quanto si caratterizzano per essere "depurati" da alcune voci più soggettive e/o non legate alla gestione operativa [35].
Vediamo quindi quali sono gli indici che permettono di svolgere l'analisi della redditività dell'impresa.
In primo luogo, è possibile studiare l'incidenza dei margini del conto economico riclassificato, rispettivamente, sui ricavi e sul totale dell'attivo, sostitendoli nelle seguenti espressioni [35]. Esse sono state ordinate in base alla posizione occupata dal margine stesso all'interno del conto economico e, tanto più registreranno valori alti, tanto più la redditività dell'impresa verrà considerata buona.

Incidenza del	sui ricavi	sul totale dell'attivo
Valore attuale	$\frac{\text{VA}}{\text{Ricavi}}$	$\frac{\text{VA}}{\text{TA}}$
EBITDA (o MOL)	$\frac{\text{EBITDA}}{\text{Ricavi}}$	$\frac{\text{EBITDA}}{\text{TA}}$
Reddito operativo (Margine Operativo Lordo)	$\frac{\text{RO}}{\text{Ricavi}}$	$\frac{\text{RO}}{\text{TA}}$
EBIT	$\frac{\text{EBIT}}{\text{Ricavi}}$	$\frac{\text{EBIT}}{\text{TA}}$
EBT	$\frac{\text{EBT}}{\text{Ricavi}}$	$\frac{\text{EBT}}{\text{TA}}$
Risultato netto	$\frac{\text{RN}}{\text{Ricavi}}$	$\frac{\text{RN}}{\text{TA}}$

Successivamente possono essere effettuate ulteriori analisi sulla capacità reddituale dell'impresa attraverso altri indici [35]:

- **indice di onerosità del capitale di terzi**, il quale rappresenta un'approssimazione del costo medio dell'indebitamento e si calcola attraverso il rapporto tra gli oneri finanziari ed i mezzi di terzi.

$$\frac{\text{OF}}{\text{MT}}$$

- **indice di rotazione del capitale investivo**, rappresentativo del numero delle volte che il capitale investito viene recuperato attraverso il conseguimento dei ricavi.

$$\frac{\text{RICAVI}}{\text{TA}}$$

- **indice di rotazione dell'attivo fisso**, esso esprime il numero delle volte che l'investimento effettuato sull'attivo immobilizzato viene rigenerato attraverso le vendite.

$$\frac{\text{RICAVI}}{\text{AF}}$$

- **indice di rotazione dell'attivo fisso**, ovvero si determina il numero delle volte che l'attivo immobilizzato viene riacquistato attraverso le vendite.

$$\frac{\text{RICAVI}}{\text{AC}}$$

- **indice di rotazione del magazzino**, esprime il numero delle volte che le rimanenze ruotano all'interno del magazzino e può essere anche considerato un indice di immobilizzo dello stesso.

  Infatti, tanto più alto è il valore dell'indice, tanto meno il magazzino può essere considerato obsoleto.

$$\frac{\text{RICAVI}}{\text{D}}$$

Infine, vi sono indici che analizzano la capacità dell'impresa nel coprire gli oneri finanziari con alcuni dei margini reddituali prima visti [35].

Indice	Calcolo
Indice di incidenza degli oneri finanziari	$\frac{\text{OF}}{\text{RICAVI}}$
Indice di sostenibilità degli oneri finanziari	$\frac{\text{MOL}}{\text{OF}}$
Indice di copertura degli oneri finanziari	$\frac{\text{EBIT}}{\text{OF}}$

Gli indici di solidità

L'analisi della solidità può essere svolta considerando due dimensioni differenti della stessa:

1. analisi della solidità patrimoniale.

2. analisi della solidità finanziaria.

Nel primo caso ci si riferisce ad un equilibrio che deve sussistere tra le fonti e gli impieghi, nel senso che si analizza quanto l'azienda sia capace di finanziare l'attivo attraverso passività che prevedano uno smobilizzo contemporaneo ad esso. Mentre nell'accezione più ristretta, ossia quella legata alla solidità finanziaria, ci si chiede quanto l'azienda sia in grado di finanziarsi attraverso fonti di finanziamento proprie [34]. Vediamo gli indici alla base dell'analisi di solidità dell'azienda [35].

- **Grado di autocopertura dell'attivo fisso** (analisi della solidità patrimoniale), idealmente dovrebbe assumere valori superiori all'unità, testimoniando la capacità dell'impresa di coprire con il capitale proprio, non solo l'attivo fisso, ma anche voci non immobilizzate degli impieghi. Tuttavia, non risulta essere problematica una cifra inferiore all'unità, purché non troppo bassa.

$$\frac{\text{PN}}{\text{AF}}$$

- **Incidenza del margine di struttura primario** (analisi della solidità patrimoniale), semplicemente si calcola il margine di struttura primario, ovvero si esprime il grado di autocopertura dell'attivo fisso in termini di margine e se ne calcola l'incidenza sul totale dell'attivo. Sulla base di quanto detto sopra, esso

dovrebbe avere un valore positivo ma non è ancora fonte di preoccupazione il contrario.

$$\frac{\text{PN} - \text{AF}}{\text{TA}}$$

- **Grado di copertura dell'attivo fisso** (analisi della solidità patrimoniale), in questo caso è molto importante che non si giunga ad un valore inferiore all'unità. Infatti, non è un bene che vengano usate fonti di finanziamento di breve termine per coprire gli impieghi di medio/lungo termine, in quanto si potrebbe non avere la possibilità di rinnovare tali rapporti di debito.

$$\frac{\text{PN} + \text{Pcon}}{\text{AF}}$$

- **Incidenza del margine di struttura secondario** (analisi della solidità patrimoniale), rappresenta, una volta calcolato in forma di margine piuttosto che di indice il valore precedente, l'incidenza dello stesso sul totale dell'attivo.

$$\frac{\text{PN} + \text{Pcon} - \text{AF}}{\text{TA}}$$

- **Grado di dipendenza finanziaria** (analisi della solidità finanziaria) rappresenta quanto, tra le fonti di finanziamento, viene reperito da terzi piuttosto che apportato in termini di patrimonio netto. Esso dovrebbe assumere un valore il più basso possibile.

$$\frac{\text{Pcon} + \text{Pcor}}{\text{TA}}$$

- **Indice di esigibilità del debito** (analisi della solidità finanziaria), calcola la quota di passivo corrente sul totale del capitale di terzi. Idealmente dovrebbe raggiungere un valore molto basso, in quanto le fonti di finanziamento a breve termine risultano molto più oscillanti e costose.

$$\frac{\text{Pcor}}{\text{Pcon} + \text{Pcor}}$$

Gli indici di liquidità

L'analisi della liquidità si può descrivere come la verifica della sussistenza di un equilibrio finanziario di breve periodo. Essa è molto importante in quanto permette

di verificare che non siano necessarie fonti di finanziamento infrannuali, ma anzi che l'azienda sia in grado di far fronte anche ad obblighi di brevissimo periodo [34].
Gli indici di liquidità calcolabili nel bilancio redatto in forma abbreviata sono i seguenti [35]:

- **indice di liquidità di primo livello**, tanto più è alto tante più saranno le liquidità a disposizione dell'azienda per coprire le fonti di finanziamento a breve termine, ovviamente, solo raramente si raggiunge l'unità (spesso soltanto nei modelli teorici);

$$\frac{\text{LI}}{\text{Pcor}}$$

- **incidenza del margine di tesoreria di primo livello**, anche in questo caso si calcola prima il margine piuttosto che l'indice precedente e poi se ne individua l'incidenza sul totale dell'attivo;

$$\frac{\text{LI} - \text{Pcor}}{TA}$$

- **indice di liquidità di secondo livello**, complementare ai discorsi fatti per gli indici di solidità, idealmente dovrebbe assumere un valore maggiore dell'unità;

$$\frac{\text{LI} + \text{LD}}{\text{Pcor}}$$

- **incidenza del margine di tesoreria di secondo livello**, viene calcolato il margine di tesoreria di secondo livello e poi si valuta la sua incidenza sul totale delle attività del patrimonio;

$$\frac{\text{LI} + \text{LD} - \text{Pcor}}{TA}$$

- **indice di liquidità di terzo livello**, è molto importante che non raggiunga un valore più basso dell'unità in quanto si compone anche di una voce che non sempre può essere considerata parte dell'attivo circolante, ovvero le rimanenze. Questo perché non è detto che esse escano dal magazzino in tempi brevi, quindi

è necessario che almeno una parte dello stesso venga coperto da una fonte di finanziamento di medio/lungo termine;

$$\frac{\text{AC}}{\text{Pcor}}$$

- **incidenza del capitale circolante netto**, anche denominata *incidenza del margine di tesoreria di terzo livello*, rappresenta l'incidenza del capitale circolante netto (attivo circolante - passivo corrente) sul totale delle attività.

$$\frac{\text{AC} - \text{Pcor}}{TA}$$

1.1.4 Altri indici

Nei modelli di previsione costruiti da Poli sono stati utilizzati anche indici dimensionali ed indici di maturità dell'impresa [35], ovvero:

- **dimensione 1** = ln (totale attivo);
- **dimensione 2** = ln (vendite + 1);
- **età** = ln (numero di anni dalla costituzione).

1.2 I modelli di previsione della crisi

Trattasi di un insieme coordinato di variabili indipendenti, relazioni statistiche e variabili dipendenti, dove le ultime rappresentano la previsione ottenuta sulla base della combinazione delle altre componenti e dei dati di input [14].
I primi studi di questo tipo sono avvenuti circa 100 anni fa come costruzioni di modelli attraverso i quali poter definire un'azienda, per le sue caratteristiche attuali, come un'azienda più simile a quelle a rischio di *default* piuttosto che ad altre aziende "sane".
Questo perché è molto difficile effettuare una vera e propria previsione, essendo il mondo imprenditoriale molto variabile, imprevedibile ed inserito in un contesto ambientale altrettanto incerto. Si pensi infatti che, così come la pandemia causata

dal virus $Covid-19$ non poteva essere altamente immaginabile, molti eventi sono difficili da intuire e questo si rende ancora più complicato se si tiene conto che ciò andrebbe fatto anche 3 anni prima della data che si sta prevedendo.
Nonostante ciò, sono stati costruiti dei modelli attraverso i quali poter tenere sotto controllo determinati elementi del bilancio più prevedibili per evitare l'insorgere di una crisi irreversibile [14]. La loro costruzione avviene mediante i seguenti step:

1. Identificazione dell'evento che segnala l'esistenza della crisi.

   Ciò vuol dire individuare quale degli stadi della crisi si vuole prevedere e, di conseguenza, con che tempestività si vuole agire. Secondo il modello di Guatri, tali stadi sono quattro e possono essere visionati nella figura 1.1.

   - Il primo stadio della crisi e la prima fase del declino viene denominata **incubazione**, ossia la manifestazione di tutta una serie di iniziali squilibri aziendali che rendono sempre più difficile l'attività imprenditoriale.
   - **Maturazione**. Ci troviamo di fronte ad un declino più evidente in quanto si hanno dei primi sintomi visibili dal bilancio. Infatti, si registreranno delle perdite economiche dovute alla riduzione dei ricavi e del capitale economico dell'azienda.
   - Il terzo stadio della crisi fuoriesce dalla situazione di declino per subentrare nella vera e propria crisi. Si tratta delle cosiddette **gravi ripercussioni sui flussi finanziari e sulla fiducia** che si manifestano in carenze di cassa, perdite di credito e di fiducia da parte degli stakeholders, perdita sempre più alta del capitale economico e rischio di non sopravvivenza.
   - **Conseguenze sugli stakeholder**. Siamo nell'ultimo stadio, ovvero nella situazione in cui non si è più in grado di far fronte alle obbligazioni assunte. Siamo subentrati, quindi, nella vera e prorpia definizione di insolvenza, la quale sfocierà poi nel dissesto.

   I modelli sviluppati inizialmente prendevano a riferimento come stadio di previsione quello dell'insolvenza anziché della crisi, rendendo molto difficile attuare

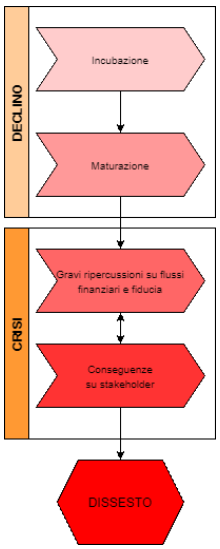


Figura 1.1: Stadi crisi d'impresa

azioni di ripresa tempestive.

Diverso è prevedere un'eventuale crisi, in quanto condizione in cui risulta ancora possibile intervenire per risollevare il business.

Infatti, se l'insolvenza si può definire come la condizione in cui l'impresa non è più capace di far fronte ai suoi debiti, la crisi invece si concretizza nella possibilità che quest'ultima si manifesti in futuro [14].

2. Definizione del campione di stima

   Ad esempio, ci si riferisce solo ad aziende con dimensioni prefissate, oppure ad aziende con una certa stabilizzazione nel mercato in quanto affermate da tempo, o ancora ad aziende che operino solo in un determinato settore.

   Ciò porterebbe alla costruzione di modelli plurimi che però abbiano come riferimento solo imprese che rispettino delle particolari caratteristiche, rendendo il risultato più attendibile e la previsione più veritiera.

   In questo modo, sarebbe possibile poter applicare il modello adeguato per la tipologia di attività imprenditoriale in questione, che tenga quindi conto del livello di rischiosità legato al tipo settore/età/dimensione/ecc...

   Non di rado infatti si considerano, per esempio, le piccole o medie imprese mol-

to più esposte al rischio di fallimento rispetto a quelle di grandi dimensioni, così come si tende a sostenere un maggior rischio quando si è nei primi anni di vita piuttisto che in periodi di stabilizzazione di un'impresa, e così via [14].

3. Scelta delle variabili indipendenti e loro ponderazione

   E' necessario capire quali sono gli indici di bilancio che abbiano un potere discriminante rispetto ad altri tali da poter classificare l'azienda come "sana" o "non sana". A tal scopo si procede con l'effettuazione di analisi statistiche, come per esempio l'analisi delle componenti principali descritta nel capitolo 3 [14].

4. Individuazione della tecnica di analisi da svolgere per effettuare la previsione

   Attualmente, le tecniche statistiche più diffuse che si utilizzano per costruire i modelli di previsione sono la regressione logistica e l'analisi discriminante multivariata [14].

5. Verifica dei risultati ottenuti

   In questo caso è molto importante effettuare una verifica su un campione della popolazione che non sia però quello su cui il modello è stato costruito cosicché si possano valutare le performance in modo più attendibile.
   Creata una matrice di confusione in cui vengono riportati i risultati delle previsioni e quelli delle osservazioni, si possono determinare i valori corrispondenti ai falsi positivi (FP), ai falsi negativi (FN), ai veri positivi (VP) ed ai veri negativi (VN).

   I primi possono essere definiti come le aziende che sono state previste dal modello come "in crisi" e che, in realtà, non lo sono. I secondi, al contrario, sono le aziende che sono state previste dal modello come "in salute" e che, in realtà, sono "in crisi". Infine, i veri positivi ed i veri negativi sono, rispettivamente, le aziende che sono state previste dal modello come "in crisi" ed "in salute" in modo corretto [35].

Questi valori alla base della valutazione delle performance del modello, permettono di calcolare i seguenti elementi [35]:

- **l'accuratezza del modello** (*accuracy*) che esprime, in termini percentuali, la capacità di prevedere in modo esatto lo stato di salute dell'azienda come "in crisi" o "non in crisi";

$$accuracy = \frac{\text{VP} + \text{VN}}{\text{TOTALE AZIENDE}} \tag{1.1}$$

- **la specificità del modello** (*specificity*) come la percentuale di probabilità di determinare in modo esatto i soggetti "sani";

$$specificity = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{1.2}$$

- **la sensitività del modello** (*sensitivity*) come la percentuale di probabilità di determinare in modo esatto i soggetti a rischio di *default*;

$$sensitivity = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{1.3}$$

- **la probabilità di errore** (*standard error*) come il numero dei soggetti determinati in modo scorretto (falsi positivi e falsi negativi).

  Difficilmente si otterrà un modello in grado di effettuare una previsione corretta al 100% dei casi, ma si può fare in modo che nel margine di errore ammesso il modello possa essere costruito con una previsione pessimista o ottimista. Nel primo caso si tenderà a preferire un modello che registri, come errori, un numero maggiore di falsi positivi, viceversa nel secondo.

  Il margine di errore può essere maggiormente compreso facendo riferimento ad un test d'ipotesi che preveda come ipotesi nulla la classificazione dei soggetti come "sani" e come ipotesi alternativa il contrario. In questo caso, i falsi positivi sarebbero associati all'errore di prima specie mentre i falsi negativi all'errore di seconda specie.

$$standard\ error = \frac{\text{FP} + \text{FN}}{\text{TOTALE AZIENDE}} \tag{1.4}$$

All'interno dello *standard error* è poi possibile individuare i seguenti tassi di insuccesso: $\frac{FN}{VP+FN}$, $\frac{FP}{VN+FP}$, $\frac{FP}{VP+FP}$, $\frac{FN}{VN+FN}$;

- **curva ROC** spiegata nel paragrafo 3.6.4 in quanto in questo capitolo vengono semplicemente presentati i modelli costruiti da Poli senza tener conto di quanto questi avrebbero potuto raggiungere in più in termini di *accuracy*.

I primi due aspetti sono posti al centro della valutazione delle performance dei modelli di previsione, in particolare la *specificity* in quanto si vuole prevedere le aziende "sane" in modo corretto, evitando che queste debbano attivare delle procedure di allerta non necessarie.

Per quanto riguarda la sensitività e la specificità del modello, è bene precisare che sono differenti da quelle presentate nel capitolo 2, in quanto queste ultime si riferiscono agli indicatori [35].

Tra i numerosi modelli di previsione confronteremo i modelli di Poli [35] con altri due molto validi: il modello di Altman [18] ed il modello proposto dai dottori commercialisti ed esperti contabili per l'individuazione degli indici di allerta.

1.2.1 Il modello di Altman

Sviluppato su un campione di 66 aziende manifatturiere statunitensi, si basa su un'analisi discriminante multivariata che ha determinato i coefficienti delle variabili indipendenti [18]. La versione qui esposta è del 1995, ma sono state effettuate numerose revisioni tanto da giungere a molteplici funzioni dello Z-Score.

Costruzione del modello

La relazione tra le variabili indipendenti e quella dipendente Z-Score è data dalla formula 1.5 dove tutte le variabili sono moltiplicate per dei coefficienti positivi, ciò vuol dire che le variabili indipendenti e la variabile dipendente hanno tra loro una polarità positiva [18].

$$Z = 6,56x_1 + 3,26x_2 + 6,72x_3 + 1,05x_4 \tag{1.5}$$

Descriviamo singolarmente il significato delle x.

$$x_1 = \frac{\text{ATTIVO CORRENTE} - \text{PASSIVO CORRENTE}}{\text{TOTALE ATTIVO}}$$

Si tratta del capitale circolante netto in senso stretto e cresce all'incrementare delle liquidità. Il suo aumento potrebbe essere considerato positivo, ma in realtà ciò potrebbe significare anche una situazione di stasi finanziaria, cioè di equilibrio apparente e formale. Questo comporta che il valore di soglia ideale può risultare molto soggettivo e perciò indefinibile per tutte le aziende in modo paritario [14].

$$x_2 = \frac{\text{UTILI NON DISTRIBUITI}}{\text{TOTALE ATTIVO}}$$

Anche in questo caso un valore ideale non esiste in quanto dipende molto dall'età dell'azienda. Infatti tanto più l'azienda è matura tanto più tenderà ad avere utili non distribuiti e viceversa. Quindi si può dire che, nel caso si stia effettuando una previsione su un'azienda matura e questo indice risulti essere basso, allora ciò significherebbe che essa non è in grado di autofinanziarsi [14].

$$x_3 = \frac{\text{EBIT}}{\text{TOTALE ATTIVO}}$$

Si tratta di un indice di redditività, ovviamente, tanto più questo è alto tanto più l'azienda può considerarsi in salute [14].

$$x_4 = \frac{\text{PATRIMONIO NETTO}}{\text{CAPITALE DI TERZI}}$$

Anche l'indice di indebitamento x_4 si muove nella stessa direzione della variabile "salute" in quanto significherebbe che l'azienda si finanzia soprattutto con capitale proprio.

Un'attenzione particolare va però tenuta su questa variabile in quanto segnala uno dei primi limiti del modello di Altman. Ovvero, essendo il modello costruito su aziende statunitensi, si tiene conto del fatto che le aziende americane molto difficilmente ricorrono al capitale di debito, ma tendono molto più a sovracapitalizzarsi.

Questo conduce all'idea che nel caso si effettuino delle previsioni su aziende italiane attraverso l'applicazione del modello di Altman, vi saranno numerose aziende che,

registrando un valore di x_4 molto più basso, verranno considerate come aziende in crisi e quindi vi saranno più falsi positivi [14].
La bontà di questo modello quindi va confrontata con il tipo di visione prima detta che si vuole avere, se pessimista piuttosto che ottimista.

Determinazione dei risultati

Calcolati tutti i valori delle variabili, questi si sostituiscono all'interno della funzione e si ottiene un numero dello Z-Score che, come si può vedere anche dalla figura 1.2, può assumere tre intervalli di valori [18].

Figura 1.2: Regola di decisione modello di Altman

- se lo Z-Score è minore di 1,10 l'impresa viene associata dal modello come a rischio di insolvenza, in quanto assume parametri similari a quelli registrati da imprese che sono poi fallite;

- se lo Z-Score è compreso tra 1,10 e 2,60 l'impresa si considera in una certa zona grigia ossia si necessita di una maggiore attenzione in quanto non vi sono elementi per classificarla correttamente. Ciò vuol dire che, piuttosto che rischiare uno stato di insolvenza, l'azienda rischia di raggiungere lo stato precedente: la crisi;

- infine se lo Z-Score è maggiore di 2,60 l'impresa viene considerata similare a quelle in "salute" perciò il modello la ritiene "sana".

Considerazioni sul modello

Come già evidenziato, il primo limite che si presenta dalla trattazione di questo modello è legato al fatto che è stato costruito sulla base di imprese manifatturiere che redigono il bilancio di tipo statunitense, comportando una scarsa capacità di adattamento al contesto italiano e a tutti i settori economici.
Inoltre, pur se lo stadio oggetto di previsione è quello della crisi, piuttosto che dell'insolvenza, e quindi adatto a possibili tentativi di ripresa, i dati sui quali si effettuano delle previsioni sono dati storici (dati di bilancio riguardanti una situazione contabile e quindi passata, perché registrata almeno sei mesi prima). Questo comporta la non tempestività dello stesso.
Si consideri poi, che nel caso debba effettuare l'analisi un soggetto esterno, questo troverà molta difficoltà nell'implementarlo in quanto vi sono parametri che difficilmente sono individuabili da un bilancio. Soprattutto se si considera che il 95% dei bilanci italiani sono redatti in forma abbreviata [35].
Infine, dai pochi valori legati alle variabili che dovrebbero essere interpretate in modo soggettivo ed in base al contesto in cui sono inserite, si ottiene un'interpretazione oggettiva che non ne tiene conto, potendo giungere a conclusioni errate.

1.2.2 Gli indici di allerta

Il consiglio nazionale dei dottori commercialisti e degli esperti contabili, in attuazione del codice della crisi d'impresa e dell'insolvenza (D. Lgs. 14/2019), ha costruito un modello di previsione della crisi in modo da poter, eventualmente, attivare le procedure di allerta ivi previste. La riforma, infatti, ha introdotto un vero e proprio obbligo di attivazione di determinate procedure qualora si riscontri la possibilità di un'eventuale insolvenza futura. Non solo, tale responsabilità è stata estesa a tutti i soggetti che tengano rapporti con la società in modo che questa possa immediatamente attivarsi per risollevare l'azienda.
Questo modello elaborato dal consiglio si costituisce della regola di analisi contenuta nella figura 1.3, che si presenta come una sorta di albero decisionale [37].

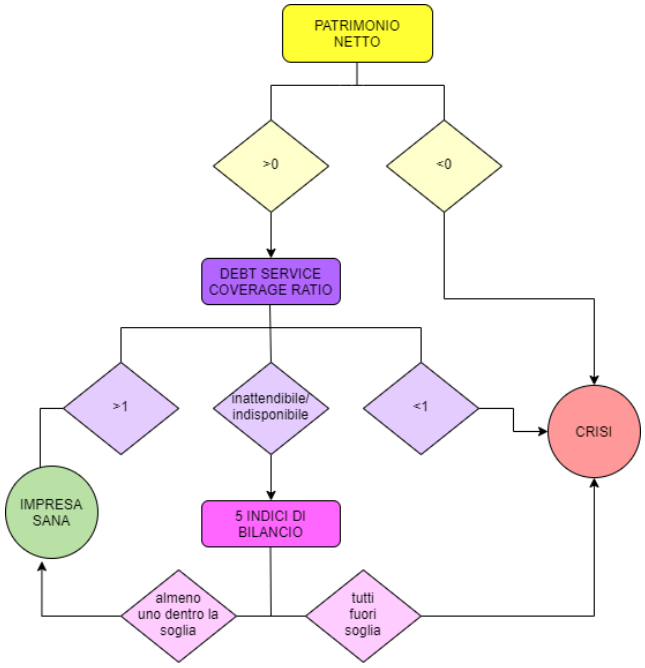


Figura 1.3: Regola di decisione del modello elaborato dal Consiglio Nazionale dei Dottori Commercialisti ed Esperti Contabili (CNDCEC)

Dallo schema si può vedere che il primo passaggio prevede il calcolo del patrimonio netto, ridotto della riserva per operazioni di copertura dei flussi finanziari attesi (non comprensiva dei crediti verso soci per versamenti ancora dovuti e dei dividendi deliberati). Nel caso questo valore sia negativo allora l'azienda verrà considerata immediatamente in crisi, altrimenti si potrà procedere con l'applicazione del modello e quindi con la determinazione dell'indice *Debt Service Coverage Ratio* (rapporto di copertura del servizio del debito). Quest'ultimo si concretizza in un'analisi del merito creditizio che fornisce informazioni in ambito di sostenibilità finanziaria del debito aziendale ed il cui calcolo deriva da dati contenuti nel budget di tesoreria.
Qualora il *Debt Service Coverage Ratio* (DSCR) assuma un valore maggiore dell'unità, allora l'impresa verrà considerata "sana", mentre nel caso contrario si dovranno attivare le procedure di allerta [14].
Tuttavia, non tutte le aziende redigono il budget di tesoreria ed in tali situazioni gli organi di controllo potrebbero non ritenere affidabili i dati a disposizione necessari

per calcolarlo, definendo quindi il DSCR inattendibile o non disponibile. In tal caso, piuttosto che fare riferimento ad un dato non affidabile si prosegue con il calcolo dei 5 indici individuati dal consiglio nazionale dei dottori commercialisti ed esperti contabili e, qualora solo alcuni dei valori siano al di fuori delle soglie previste allora si definirà l'impresa "sana", viceversa la si considererà in crisi [37].
Gli indici individuati come *indici di allerta* sono i seguenti:

- indice di sostenibilità degli oneri finanziari, ha polarità positiva con la crisi e si calcola come segue:

$$\frac{\text{interessi e altri oneri finaziari (C.17)}}{\text{ricavi delle vendite e delle prestazioni (A.1)}}$$

- indice di adeguatezza patrimoniale, ha polarità negativa con la crisi e si ottiene dal seguente rapporto:

$$\frac{\text{P.N.} - \text{cred. v/soci per vers. ancora dovuti} - \text{div. deliberati}}{\text{debiti (D)} + \text{ratei e risconti (E)}}$$

- indice di liquidità, ha polarità negativa con la crisi in quanto le passività a breve saranno sempre più elevate delle attività a breve. Il suo calcolo è dato da:

$$\frac{\text{attivo circolante} - \text{cred. es. oltre l'es. succ.} + \text{ratei e risc. (attivi)}}{\text{debiti} - \text{deb. es. oltre l'es. succ.} + \text{ratei e risc. (passivi)}}$$

- indice di ritorno liquido dell'attivo, ha polarità positiva con la crisi e si calcola in questo modo:

$$\frac{\text{risultato dell'es.} + \text{costi non monetari} - \text{ricavi non monetari}}{\text{totale attivo}}$$

- indice di indebitamento previdenziale e tributario, ha polarità positiva con la crisi e si ricava dalla frazione qui riportata:

$$\frac{\text{deb. tributari (D.12)} + \text{deb. v/istituti di previd. e assist. soc. (D.13)}}{\text{totale attivo}}$$

In questo modello le soglie dipendono dal settore di appartenenza, quindi si tiene conto della rischiosità legata al tipo di attività svolta, cosa che non accade nel modello

di Altman. Tuttavia, seppur più ponderato, in questo caso devo sapere anche il codice ATECO dell'azienda, il quale non sempre risulta disponibile.
Il fuori soglia dipende dal tipo di indice, potendo essere sopra o sotto la stessa a seconda del tipo di polarità sussistente tra questo e la crisi [37].

Punti di forza e criticità degli indici di allerta

Si tratta di un modello destinato ai soggetti interni, in quanto conseguenza di una responsabilizzazione degli stessi. Quindi si può in primo luogo dire che ciò lo rende molto inadatto per soggetti che non fanno parte dell'azienda.
Inoltre, dato che la costruzione del modello si è basata su imprese che redigono il bilancio ordinario e non su imprese che redigono il bilancio in forma abbrevita, l'applicazione dello stesso è molto rapida se si dispone della prima tipologia, risulta molto più difficile implementarlo negli altri casi in cui si hanno meno informazioni. Ad esempio:

- i costi ed i ricavi non monetari contenuti nel calcolo dell'indice di ritorno liquido dell'attivo, molto difficilmente vengono individuati da un soggetto esterno che dispone solo di un bilancio abbreviato;
- nell'indice di liquidità è necessario sapere quando debiti e crediti torneranno ad avere una manifestazione monetaria, ciò non lo si può sapere se non dalla lettura di una nota integrativa, non obbligatoria nel bilancio abbreviato.

Per quanto riguarda la tempestività, così come nel modello di Altman, la previsione ha come oggetto lo stadio della crisi e quindi una situazione in cui, se preventivata, si può intervenire ancora. Tuttavia, rispetto al precedente, permette di agire molto prima in quanto i dati su cui ci si basa riguardano una previsione di sei mesi.
La differenza, quindi, sta nel fatto che, seppur ipotetici e non verificati, i dati sono orientati ad un futuro, perciò consente di attuare un intervento efficace con più possibilità di successo.
Volendo, invece, focalizzarci sul tipo di errori che vengono commessi, anche in questo caso si registrano molti falsi positivi, manifestando una visione molto pessimista.

Avendo introdotto le limitazioni sussistenti nei modelli di previsione trattati, si può ora vedere come i modelli costruiti da Poli risultino molto più adeguati per effettuare previsioni attendibili.

1.3 Modelli per soggetti esterni e bilancio abbreviato

Come introdotto, si caratterizzano per essere molto più adeguati alle tipologie di documenti pubblicati dalle imprese italiane, in quanto implementabili anche attraverso le informazioni fornite dal bilancio redatto in forma abbreviata e quindi anche dal soggetto esterno [35].
Rispetto ai precedenti, si considera come oggetto della previsione l'attivazione di una procedura concorsuale, quindi uno stadio successivo a quello della crisi (l'insolvenza), potendo questo essere l'input per ulteriori miglioramenti successivi. Tuttavia, anche in questo caso, ci si basa su dati di budget con un orizzonte futuro di 2 anni, testimoniato dal passaggio in termini di probabilità che si effettua per la classificazione dell'azienda.
Essi adottano una metodologia statistica che spiegata più approfonditamente nei capitoli applicativi, ossia la regressione logistica, capace di adattarsi anche a distribuzioni non normali.
Sulla base delle riclassificazioni individuate all'inizio del capitolo si possono effettuare delle previsioni e calcolare gli indici necessari per l'implementazione dei modelli.
Se si considera il modello 2, non sarà neanche necessaria la riclassificazione del conto economico in quanto basterà sapere il risultato atteso. Vediamone quindi il funzionamento.

1.3.1 Definizione del campione

Il campione rappresentativo della popolazione di riferimento si compone di 2264 aziende, in particolare si costituisce di:

- uguale numero di aziende fallite e non fallite. La scelta effettuata per il settore di appartenenza è stata quella di prediligere un modello generico, cioè valido per

tutti i settori in cui sono state riscontrate numerose imprese in crisi, piuttosto che un modello specifico per solo alcuni, dove l'omogeneità potrebbe causare una minore affidabilità dello stesso;

- aziende situate all'interno di aree geografiche diverse;
- imprese con periodi di vita differenti;
- soprattutto imprese che redigono il bilancio in forma abbreviata.

Il campione è stato ulteriormente suddiviso nel seguente modo: 1584 aziende sono state considerate per costruire i 3 modelli (campione *train*), le restanti 680 (campione *test*), invece, sono state alla base della loro valutazione [35]. Grazie alla considerazione di imprese al di fuori del campione si può contare su una maggiore affidabilità del risultato ottenuto. Tuttavia, dato che le analisi empiriche trattate nei capitoli 3 e 4, sono state svolte e verificate sullo stesso campione di 2264 aziende, per rendere i confrontabili i modelli si riportano i risultati dei modello di Poli riguardanti il campione *train*.

1.3.2 Selezione delle variabili

Innanzitutto, sono stati considerati 31 indici di bilancio, dei quali è stata poi verificata la calcolabilità (in quanto per alcune aziende si registravano ricavi ed altre voci pari a 0, comportando l'inserimento di un denominatore non ammesso).
Fatto ciò, si è continuato con l'individuazione degli indicatori con la seguente regola decisionale: gli indici calcolabili sono stati considerati, gli indici non calcolabili per un numero di aziende inferiore a 8 sono stati considerati sostituendo il denominatore con il valore di 1 e gli indici non calcolabili per un numero di aziende superiore a 8 non sono stati considerati [35].
Si è giunti perciò alla considerazione dei seguenti indici di bilancio:

- incidenza del margine di struttura primario
- incidenza del margine di struttura secondario

- grado di dipendenza finanziaria
- indice di esigibilità del debito
- indice di liquidità di primo livello
- incidenza del margine di tesoreria di primo livello
- indice di liquidità di secondo livello
- incidenza del margine di tesoreria di secondo livello
- indice di liquidità di terzo livello
- incidenza del margine di tesoreria di terzo livello (o del capitale circolante netto)
- incidenza del valore attuale sul totale attivo
- incidenza dell'EBITDA (o MOL) sul totale attivo
- incidenza del reddito operativo sul totale attivo
- incidenza dell'EBIT sul totale attivo
- incidenza dell'EBT sul totale attivo
- incidenza del risultato netto sul totale attivo
- rotazione del capitale investito
- rotazione attivo circolante.

Stabilite le variabili calcolabili, sarà necessario prima procedere con l'eliminazione degli outlier, evitando di analizzare dati che influenzino negativamente i risultati dei modelli.
Successivamente, si escluderanno alcune delle variabili con lo scopo di lavorare solo con quelle che non siano correlate significativamente tra loro perché alteranti le stime dei modelli.

Infatti, una rilevante correlazione tra più variabili porta a stime poco attendibili, comportando l'esigenza di inserirne nel modello soltanto una [35].

1.3.3 Eliminazione degli outlier

Gli outlier sono dati rilevantemente diversi dagli altri e sono previsti solo per gli indici che non abbiano limiti di massimo e di minimo definiti.
La loro eliminazione può essere attuata escludendo quei dati che, avendo intervalli di variazione infiniti sia a "destra" che a "sinistra" abbiano registrato un valore inferiore o superiore ad un determinato intervallo di percentili. In questo caso sono stati esclusi quelli al di sotto del quinto percentile o al di sopra del novantacinquesimo percentile [35].

1.3.4 Costruzione dei modelli

Dopo l'eliminazione degli outlier sono state applicate tre differenti tecniche attraverso cui individuare le variabili più rappresentative della distribuzione. Quindi, sono stati costruiti altrettanti modelli di previsione della crisi all'interno dei quali definire le funzioni di regressione logistica determinanti la Y [35].
Due di queste tecniche, tra loro alternative, permettono di individuare le variabili indipendenti correlate in modo rilevante e sono le seguenti:

1. calcolo dell'indice di correlazione di Pearson (spiegato nel capitolo 3) e considerazione delle sole variabili che registrino un valore dello stesso superiore a quello prefissato. In questo caso sono state considerate correlate in modo rilevante le variabili con un valore dell'indice di correlazione superiore a $0,65$ [35];

2. calcolo del livello del *fattore di crescita della varianza* e determinazione, anche qui, di un livello di significatività della correlazione lineare. In questo caso il livello discriminante è stato fissato a 5 e sono state mantenute le variabili con un valore del fattore inferiore a questa soglia [35].

Mentre la terza tecnica - la *stepwise selection method* derivante dalla *likelihood-ratio test* - serve per identificare, tra gli indici di bilancio che sono interessati dal problema della collinearità (individuati attraverso il calcolo dell'indice di correlazione di Pearson o del livello del *fattore di crescita della varianza*), quelli che più incidono sulla previsione, ma può essere applicata anche senza dover passare prima per una delle metodologie precedenti [35]. Essa seleziona le variabili da utilizzare nella regressione logistica che registrano dei risultati più importanti dopo essere state sottoposte al *likelihood-rario test*. Quest'ultimo, detto anche *rapporto di verosimiglianza*, confronta l'accuratezza (percentuale delle previsioni corrette) tra più modelli di previsione ed è stato utilizzato per individuare quali delle variabili abbiano registrato una performance migliore [39].
Individuate le variabili, si applica il metodo della regressione logistica (descritta nel capitolo 3) per determinare il valore della dipendente Y assunta dall'azienda, si procede poi con il passaggio ad una ipotesi di probabilità attraverso la formula 1.6 [35].

$$p[Y_i] = \frac{e^{Y_i}}{1 + e^{Y_i}} \tag{1.6}$$

La probabilità, che può assumere un valore compreso nell'intervallo $[0, 1]$, permette di classificare l'azienda come "in crisi" o "non in crisi" a seconda che, rispettivamente, superi o non superi il valore soglia di $= 0,5$ (vedi figura 1.4).

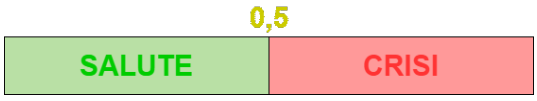


Figura 1.4: Regola decisiva modelli costruiti da Poli

Vediamo sinteticamente come nei tre modelli è stata determinata la Y.

Modello 1

Nel modello 1 [35] l'eliminazione delle variabili rilevantemente correlate è stata effettuata attraverso i seguenti procedimenti:

1. in un primo momento sono state individuate le variabili tra loro correlate attraverso la tecnica dei coefficienti di correlazione;

2. su queste è stata successivamente applicata la *stepwise*, sono quindi stati identificati gli indici di bilancio rilevanti nel modello.

Dall'analisi è stata prodotta la seguente regressione logistica.

$$Y = -7,4289 + 2,7681X_1 - 1,7203X_2 - 3,7776X_3 + 0,3488X_4 + 1,5970X_5 + 0,9271X_6 - 0,0499X_7 \quad (1.7)$$

dove

X_1 =grado di dipendenza finanziaria

X_2 =indice di liquidità di primo livello

X_3 =incidenza dell'EBIT sul totale attivo

X_4 =dimensione 1

X_5 =indice di esigibilità del debito

X_6 =incidenza del margine di tesoreria di secondo livello

X_7 =dimensione 2

Facendo riferimento al campione di costruzione del modello (campione *train*), si può riportare quanto osservato e quanto preventivato in una matrice di confusione.

classificazione osservata			
		in crisi	non in crisi
classificazione prevista	in crisi	615	230
	non in crisi	177	562

Come si può vedere, la tabella contiene quattro quadranti nei quali sono riportati i valori di:

- VN nel quadrante (non in crisi, non in crisi);
- VP nel quadrante (in crisi, in crisi);
- FP nel quadrante (in crisi, non in crisi);

- FN nel quadrante (non in crisi, in crisi).

I primi due consistono, perciò nel numero delle previsioni corrette, mentre gli ultimi due nel numero dele previsione errate.

Modello 2

La procedura del modello 2 [35], è la seguente:

1. prima è stata adottata la tecnica del fattore di crescita della varianza;
2. in un secondo momento la *stepwise selection method*.

L'analisi ha condotto al seguente risultato:

$$Y = -7,6908 + 2,6743X_1 - 2,1875X_2 - 3,5978X_3 + 0,3422X_4 + 1,3724X_5 + 0,4835X_6 - 0,0467X_7 \quad (1.8)$$

dove

X_1 =grado di dipendenza finanziaria

X_2 =indice di liquidità di primo livello

X_3 =incidenza del reddito netto sul totale attivo

X_4 =dimensione 1

X_5 =indice di esigibilità del debito

X_6 =indice di liquidità di secondo livello

X_7 =dimensione 2

Anche in questo caso si trascrive la tabella in cui si confrontano le previsioni del modello con le osservazioni effettuate, entrambe svolte sul campione *train*.

classificazione osservata			
		in crisi	non in crisi
classificazione prevista	in crisi	606	225
	non in crisi	186	567

Modello 3

Nel modello 3 [35], la procedura è stata quella di applicare la *stepwise selection method* direttamente su tutte le variabili senza effettuare prima un'analisi sulla correlazione. Il risultato ottnuto è il seguente:

$$Y = -5,2082 - 4,6406X_1 - 1,7041X_2 + 0,3393X_3 - 3,4443X_4 + 1,2377X_5 - 0,0584X_6 \quad (1.9)$$

dove

X_1 =incidenza del margine di tesoreria di primo livello

X_2 =indice di esigibilità del debito

X_3 =dimensione 1

X_4 =incidenza dell'EBIT sul totale attivo

X_5 =incidenza del margine di tesoreria di secondo livello

X_6 =dimensione 2

L'implementazione del terzo modello ha portato ai seguenti valori di osservazione e di previsione nel campione *train*.

classificazione osservata			
		in crisi	non in crisi
classificazione prevista	in crisi	586	205
	non in crisi	206	587

1.3.5 Considerazioni sui modelli

Rappresentate le matrici di confusione si possono calcolare la bontà del modello costruito, ovvero l'*accuracy* e la *specificity* con lo svolgimento delle equazioni 1.1 e 1.2.

I tre modelli costruiti da Poli hanno registrato le seguenti performance effettuando la verifica all'interno del campione *train* [35]:

	Modello 1	Modello 2	Modello 3
Accuracy	74,31%	74,05%	74,05%
Specificity	70,96%	71,59%	74,12%

Si può notare come in tutti e tre i casi vi sia un buon risultato nella previsione effettuata utilizzando il metodo della regressione logistica, ma sulla base di quanto riportato in questo lavoro si individuano eventuali elementi che possono essere fonte di miglioramenti del modello, tanto da sperare in aumenti della stessa:

- la selezione del campione può essere effettuta di volta in volta in un settore differente, potendo così giungere alla costruzione di modelli specifici che tengano conto delle caratteristiche del settore piuttosto che di un unico modello generale valido per tutti i settori;

- abbiamo visto l'elevata tempestività garantita dal modello grazie alla considerazione di dati riguardanti i 2 anni futuri. Tuttavia, l'oggetto di analisi rimane ancora uno stadio della crisi tardivo per garantire all'impresa la piena capacità di risollevarsi. Sarebbe quindi il caso di porre come oggetto dei modelli la previsione della crisi piuttosto che dell'insolvenza;

- può svolgersi l'integrazione di ulteriori tecniche di analisi con lo scopo di giungere ad *accuracy* e *specificity* più elevate.

Nel proseguo del lavoro ci si focalizza sul terzo aspetto e si sviluppano delle analisi empiriche confrontabili con quanto qui riportato.

Capitolo 2

Gli indici di deprivazione

La particolarità di questo lavoro è quella di proporre, tra gli altri, un approccio del tutto nuovo per la costruzione dei modelli di previsione della crisi, ovvero l'utilizzo degli indici di deprivazione.
La novità la si ritrova sia nell'ambito di applicazione dello stesso, in quanto solitamente adattato a studi di tipo sociale, sia nella differente prospettiva di visione, perché, piuttosto che prevenire uno stato di *default*, con questo metodo si vuole visualizzare quanto ad un'azienda manchi per essere percepita come "sana".
Ciò vuol dire costruire un modello in grado, non solo di individuare tali imprese, ma anche di cambiare l'ottica con la quale l'azienda reagisca dopo la conoscenza di tale risultato, ponendo in una condizione più ottimista quella che non venga prevista come "sana". Nello specifico, non si vuole fare in modo che questa risponda semplicemente evitando l'insolvenza, anzi ci si pone l'obiettivo di presentarle di quanto e di cosa necessiti per raggiungere una condizione minima di solvibilità.
Vista tale peculiarità, il capitolo, in un primo momento descrive il funzionamento degli indici in generale, quali strumenti di carattere statistico in quanto sintesi matematiche in grado di fornire informazioni su un dato aspetto del fenomeno, e poi termina con un focus sugli indici di deprivazione.

2.1 L'analisi per indici

I termini "indicatore" ed "indice" sono molto spesso utilizzati come sinonimi, in realtà si tratta di due elementi molto diversi in quanto, se il primo tende ad essere un mezzo molto semplice, il secondo si caratterizza per la sua articolazione in più componenti e quindi per una maggiore difficoltà di comprensione. L'indicatore, a sua volta, risulta essere un elemento differente dalla variabile quantitativa, anzi la contiene con lo scopo di rendersi confrontabile con altri indicatori [29].
Gli indici, invece, sono sempre più adottati per analizzare dimensioni complesse legate alla misurazione dell'efficacia delle politiche di bilancio e garantiscono un'analisi simultanea degli svariati elementi che li compongono, tra cui gli indicatori statistici: ottimi supporti per lo svolgimento di processi decisionali in quanto sintesi dell'andamento di un fenomeno e facilmente interpretabili [24].
L'analisi per indici si rende necessaria soprattutto nei casi di studio dei fattori latenti, cioè di avvenimenti di difficile misurazione in quanto influenzati da più agenti e comprensivi sia di aspetti positivi che di aspetti negativi. In questi casi, entrano in gioco gli indici sintetici che svolgono il ruolo di interpreti dei fattori su detti attraverso l'utilizzo di un'unica misurazione in grado di riassumere l'andamento delle performance [29].

2.1.1 L'analisi dei fenomeni d'interesse

Nell'analisi per indici si effettua, prima di tutto, l'analisi dei fenomeni di interesse che inizia con la definizione dei seguenti tre elementi [29]:

- il conceptual framework.
  Definire il conceptual framework vuol dire individuare ciò che si vuole analizzare delineando una successione di rappresentazioni non concrete, la cui applicabilità dipende da quella della teoria cui si riferisce;

- l'area di indagine.
  Essa deve essere collegata alle variabili latenti, ossia a degli elementi analizzabili

empiricamente che riflettano la natura dell'area di indagine in questione, e definita partendo dalla determinazione di quella generale per giungere a quella particolare;

- gli elementi osservabili.

  Si distinguono gli elementi osservabili direttamente esaminabili, ossia le variabili che delimitano l'area di indagine, da quelli non direttamente esaminabili, ossia gli indicatori.

  Questi ultimi devono rispettare determinati requisiti: devono essere misurabili, coerenti con gli elementi definiti precedentemente ed in grado di fare riferimento a diverse componenti non sempre correlate tra loro. Inoltre, è importante che posseggano le caratteristiche della non sostituibilità con un altro indice, la rappresentatività dell'intero fenomeno osservato, la non ambiguità, l'affidabilità, la capacità di sintesi delle aggregazioni di osservazioni parziali, la correlatività con gli obiettivi dello studio, la esaustività e la confrontabilità tra aree geografiche.

  Spesso, l'elevata complessità di ciò che viene identificato dalla variabile comporta che, perché gli indicatori rispettino i precedenti parametri, debbano essere individuati più indicatori che lo definiscano.

  Soprattutto nel caso degli indicatori predittivi, è fondamentale che gli indicatori posseggano ulteriori caratteristiche, ovvero la sensibilità, la specificità e l'affidabilità.

  Solitamente, le prime due sono in continua relazione inversa per cui nell'analisi si tende a scegliere il giusto trade-off tra le due [29].

La sensibilità

Quando si dice che un indicatore possiede il carattere della sensibilità, vuol dire che quell'indicatore sarà in grado di adattarsi ad eventuali mutazioni di luogo e di tempo dell'area di indagine in cui si inserisce.

La caratteristica della sensibilità può anche non sussistere in un determinato momento per poi generarsi in un altro e viceversa [16].

La specificità

Insieme alla sensibilità, l'indicatore viene sempre valutato anche nella sua capacità di specificità, cioè dipendere soltanto dalle mutazioni che riguardino l'avvenimento di riferimento.
La specificità risulta essere una caratteristica molto rara tra gli avvenimenti in cui vi sia un'elevata correlazione, come ad esempio tra quelli legati alla comunità [16].

L'affidabilità

Molto importante è anche la caratteristica dell'affidabilità, ovvero la capacità dell'indicatore di mantenere sempre un margine di errore similare [16].
La verifica di affidabilità degli indicatori può essere effettuata attraverso:

1. la valutazione della stabilità nel tempo;
2. la valutazione di equivalenza;
3. la valutazione della concordanza fra osservatori;
4. la valutazione della coerenza interna.

Fissati questi elementi, si stabiliscono le relazioni esistenti tra loro categorizzando così il modello che si sta sviluppando. Si formano in questo senso due parti dello stesso: quella strutturale e quella di misurazione. La prima dipende dalla relazione sussistente tra le variabili latenti, mentre la seconda da quella che lega i concetti astratti ai corrispondenti indicatori, ciò permetterà anche di quantificare il livello di affidabilità di questi ultimi.
Infine, si qualificheranno gli indicatori come costitutivi o concomitanti a seconda del tipo di legame caratterizzante il loro rapporto [29].

Il modello potrà definirsi costruito una volta passati dal concetto generale ad un indice che lo sintetizzi. Tuttavia, ciò potrebbe portare ad una perdita di informazioni e sarà tanto meno affidabile quanto più esprimerà parzialmente il concetto generale cui si riferisce [28].

2.1.2 La costruzione degli indici

Il procedimento di costruzione degli indici prevede, innazitutto, l'individuazione di una struttura teorica suddivisa in dimensioni e sub-dimensioni e la identificazione degli indicatori elementari.
Si prosegue, poi, con il controllo della completezza dei dati e con la loro trasformazione, ossia con loro eventuali standardizzazioni e polarizzazioni. La polarizzazione è una trasformazione dei dati che garantisce che l'indice e l'indicatore si muovano nella stessa direzione (polarità positiva), effettuando delle trasformazioni di segno opposto quando si è di fronte ad una polarità negativa (i due elementi si muovono in direzioni opposte) [36].
Successivamente, gli indicatori saranno sottoposti all'analisi multivariata e a quella dimensionale, alla loro ponderazione e alla loro aggregazione.
In quest'ultimo caso non si parlerà più di indicatori elementari omogenei ma di indici multidimensionali in grado di studiare più elementi contemporaneamente.
Sarà molto importante evitare che l'interpretazione di questo strumento di analisi non sia influenzata negativamente da errori aleatori eccessivi, perciò ogni singolo aspetto che lo costituisca andrà ben calibrato.
L'aggregazione potrà essere di tipo continuo, la quale costituirà gli indici continui, oppure di tipo numerabile, generando i cosiddetti indici countable. Nel nostro caso verrà posta l'attenzione sui primi fornendone di seguito alcuni esempi [28].

2.1.3 Alcuni indici continui

Il più utilizzato è la **media aritmetica di indici**.
Dato con $j = 1, 2, \ldots, N$, il calcolo della media aritmetica di indici è datta dalla

formula 2.1 [36].

$$\frac{1}{N}\sum_{j=1}^{N} Z_j \tag{2.1}$$

Le Z_j si calcolano in due modalità differenti a seconda del tipo di polarità sussistente tra la variabile e l'indice. Dati m = valore minimo ed M = valore massimo della variabile X, Z è uguale a:

- $\frac{X-m}{M-m}$ qualora la variabile abbia polarità positiva;
- $\frac{M-X}{M-m}$ qualora vi sia polarità negativa.

Segue poi l'indice denominato **media aritmetica degli z-scores**, ossia la media aritmetica della Z tradizionale [36].

$$\frac{1}{N}\sum_{j=1}^{N} Z_j \tag{2.2}$$

con $j = 1, 2, \ldots, N$

Data la variabile standard s, in questo caso la Z è uguale a:

- $\frac{X-m}{s}$ qualora la variabile abbia polarità positiva;
- $-\frac{X-m}{s}$ qualora la polarità sia negativa.

Sempre prendendo a riferimento la variabile standard tradizionale è possibile calcolare il **Mazziotta Pareto Index** del 2013 [30], il quale tuttavia effettua una penalizzazione nell'aggregazione attraverso la media aritmetica e lo si calcola come segue.

$$A \pm \frac{S_M^2}{A} = A \pm \left(\frac{S_M}{A}\right) S_M \tag{2.3}$$

dove

$A = \frac{1}{N}\sum_{j=1}^{N} Z_j$
$S_M = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(Z_j - A)^2}$
$\frac{S_M}{A}$ = coefficiente di variazione

Ovvero, ci si sposta dalla media tanto quanto è grande il coefficiente di variazione. Ciò lo si vede dal fatto che si considera la media classica $\pm$ un intervallo dato dal

prodotto tra il coefficiente di variaziazione a la deviazione standard.
La variabile Z si ottiene così:

- nel caso di polarità positiva $Z = 100 + 10\frac{X-m}{s}$;
- nel caso di polarità negativa $Z = 100 - 10\frac{X-m}{s}$

Un altro indice continuo molto utilizzato è l'**indice di Jevons** [17], caratterizzato per l'aggregazione attraverso l'utilizzo della media geometrica.

$$\prod_{j=1}^{N} Z_j^{\frac{1}{N}} = \sqrt[N]{\prod_{j=1}^{N} Z_j} \tag{2.4}$$

Dove b =valore di base
In questo caso la variabile standard è la stessa sia che vi sia polarità positiva si che sussista una polarità negativa e si ottiene svolgendo la formula $Z = \frac{X}{b}$.
Costruita la matrice $X = \{x_{ijt}\}$ di n righe (dimensione del campione pari a n), di m colonne (m indicatori) e di p strati (p anni), si possono costruire due particolari forme sintetiche dell'indice di jevons [17]:

1. L'indice sintetico statico per l'unità i al tempo t, dato da:

$$JS_{it} = \prod_{j=1}^{m} \left(\frac{x_{ijt}}{x_{bjt}} 100 \right)^{\frac{1}{m}}$$

   dove x_{bjt} è il valore base dell'indicatore j al tempo t.

2. L'indice sintetico dinamico per l'unità i al tempo t, dato da:

$$JD_{it} = \prod_{j=1}^{m} \left(\frac{x_{ijt}}{x_{ij(t-1)}} 100 \right)^{\frac{1}{m}}$$

   dove $x_{ij(t-1)}$ è il valore dell'indicatore j per l'unità i al tempo $t-1$.

Sempre effettuando l'aggregazione attraverso la media geometrica è possibile calcolare poi la **media geometrica degli indici relativi** attraverso la seguente formula [36]:

$$\prod_{j=1}^{N} Z_j^{\frac{1}{N}} = \sqrt[N]{\prod_{j=1}^{N} Z_j} \tag{2.5}$$

In questo caso la Z si differenzia a seconda del tipo di polarità sussitente tra l'indice e la variabile in tal modo:

- $Z = 1 + 198\frac{X-m}{M-m}$ se la polarità di X è positiva;
- $Z = 1 + 198\frac{M-X}{M-m}$ se la polarità di X è negativa.

Per terminare si ricorda anche il **adjusted Mazziotta Pareto index** [31], calcolato come aggregazione con la media aritmetica che penalizza la variazione degli indicatori elementari della seguente variabile Z:

- Nel caso di polarità positiva $Z = 60 + 70\frac{X-m}{M-m}$;
- Nel caso di polarità negativa $Z = 60 + 70\frac{M-X}{M-m}$.

Segue la formula del AMPI:

$$A \pm \frac{S_M^2}{A} = A \pm \left(\frac{S_M}{A}\right) S_M \tag{2.6}$$

dove

$A = \frac{1}{N}\sum_{j=1}^{N} Z_j$
$S_M = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(Z_j - A)^2}$ è la deviazione standard
$\frac{S_M}{A}$ = coefficiente di variazione

Classificati gli indici in base al metodo di aggregazione è possibile effettuare anche ulteriori suddivisioni del tipo di indicatori che li compongono, i quali possono essere sottoposti a più classificazioni sulla base dell'oggetto di osservazione [5].

2.2 Tipologie di indicatori

Considerando quanto appena detto, classifichiamo gli indicatori sulla base dei seguenti aspetti: lo scopo conseguito, il tempo di osservazione, la veridicità delle informazioni alla base della loro costruzione, la loro articolazione in base a diversi aspetti, il tipo di attività che supportano, la comprensibilità dell'informazione fornita, l'oggetto di analisi e le modalità di aggregazione.
Per quanto attiene all'ultimo aspetto abbiamo già illustrato le principali tecnice di aggregazione, perciò si può proseguire con la trattazione degli altri aspetti [29].

Le finalità

Sulla base dello scopo da raggiungere, si possono distinguere gli indicatori [29]:

- descrittivi, hanno lo scopo di informare sui comportamenti assunti dalla società e sui rapporti sussistenti al suo interno;
- esplicativi, sono orientati ad interpretare l'andamento dei fenomeni osservati per comprendere la realtà circostante;
- predittivi, di nostro interesse, mirano a prevedere ipotetiche situazioni future per poter prevenire eventuali problematiche. Si tratta, forse, di quelli di più complessa costruzione e la loro forma standard è la differenza non assoluta;
- normativi, trattasi di indicatori di supporto per processi decisionali di intervento. In questo caso è imporante inserire l'orentamento temporale o spaziale di interesse per ottenere i risulati desiderati;
- problem oriented, si muovono verso l'analisi dell'andamento sociale per ottenere informazioni su particolari caratteristiche collegatevi.

Modalità di osservazione

Si distinguono gli indicatori di stato, di tendenza, conglomerativi e deprivativi [29]. Se i primi analizzano la società in un dato istante, i secondi la studiano, invece, in più lassi temporali. Mentre per quanto riguarda gli ultimi due, la distinzione la si ritrova nel tipo di realtà osservata, se quella ottimale (indicatori conglomerativi) o quella peggiore (indicatori deprivativi).

Modalità di costruzione

Possiamo avere tre differenti tipologie di costruzione degli indicatori, le quali originano rispettivamente [29]:

- gli indicatori elementari (o semplici) che si costituiscono partendo dai dati quantitativi;

- indicatori sintetici, originati da aggregazioni dei precedenti purché caratterizzati da una relazione di omogeneità;
- indicatori compositi, ottenuti dall'aggregazione dei due precedenti e con lo scopo di ottenere analisi multidimensionali.

Gli indicatori compositi si distinguono a loro volta dai sistemi di indicatori. Mentre i primi sono in grado di interpretare una variabile latente attraverso l'utilizzo di un solo strumento di misurazione (potendo quindi effettuare confronti di vario genere), i sistemi di indicatori permettono di spiegare eventuali divergenze sussistenti in tali confronti. In questo secondo caso, quindi, si parla proprio di un sistema creato attorno ad un ragionamento e non di un semplice insieme di formule matematiche [3].
Di seguito viene riportato lo schema del sistema di indicatori in modo che sia ben comprensibile il passaggio dagli indicatori ai sistemi di indicatori.

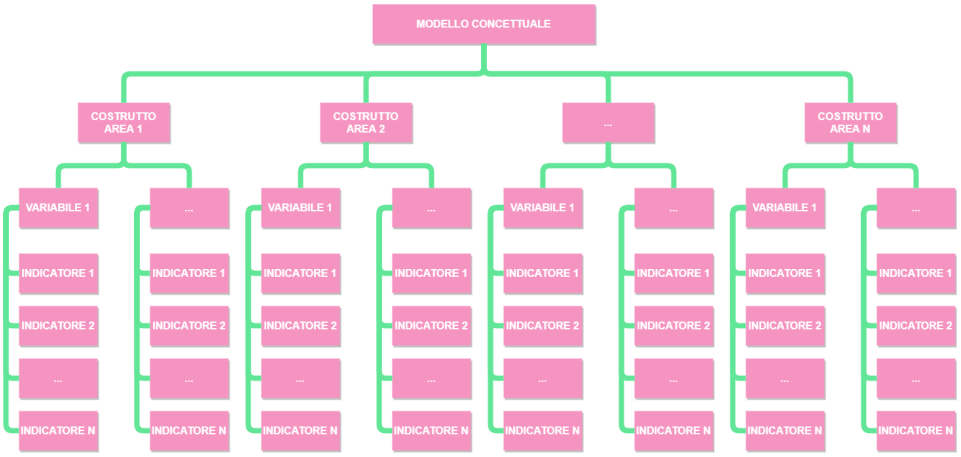


Figura 2.1: Sistema di indicatori

Capacità di stratificazione e origine dell'informazione

Gli indicatori possono essere articolati in base alla località, al periodo o alla classe sociale. Quindi si otterranno rispettivamente gli indicatori [29]:

- territoriali, rendono possibile effettuare confronti spaziali;

- temporali, confrontabili su più istanti differenti;
- individuali, permetteno di comparare più individui.

(indicatori territoriali) rendendo possibile effettuare confronti spaziali, al periodo (indicatori temporali) in modo da rendersi confrontabili su più istanti diversi o articolati in base alla classe sociale (indicatori individuali) garantendo la possibilità di effettuare confronti tra più individui.
Sulla base della tipologia fonte delle informazioni, invece, si possono avere indicatori oggettivi ed indicatori soggettivi [29].

Livelli di comunicazione

Gli indicatori sono frequentemente fonte di comunicazione per la società, tuttavia questo può orientarsi solo a particolari classi di soggetti, in quanto particolarmente complessi e di difficile comprensione per tutta la comunità, oppure a tutti gli individui in quanto di maggiore comprensibilità.
Nella pratica, quindi, si assiste ad una continua scelta tra due aspetti dell'informazione che vuole essere trasmessa: la qualità e la comprensibilità [29]. Si avranno perciò:

- indicatori molto complessi e fonte di informazioni di elevata qualità scientifica;
- indicatori meno complessi ma con maggior seguito;
- indicatori che presentano un buon bilanciamento.

Fenomeni osservati

In base all'area di interesse si possono categorizzare in:

- indicatori di sviluppo;
- indicatori sociali;
- indicatori economici;
- indicatori multidimensionali.

2.3 Gli indici di povertà

Nel lavoro svolto siamo interessati a sapere quanto le aziende possano essere considerate lontane dalla definizione di "azienda sana", evidenziando un rischio di *default* sempre più basso tanto più ci si avvicina a tale condizione .
L'analisi è molto simile a quella effettuata dagli indici di povertà qui richiamati ma, in particolare, a quella dei cosiddetti **indici di deprivazione**. Questi ultimi, sono sempre classificabili all'intero degli indici di povertà ma assumono delle caratteristiche peculiari. Prima di trattarli verrà fatto un breve richiamo ad alcune tipologie di indici di povertà principali [2].

1. **Headcount ratio (indice di diffusione)**.

   Il procedimento prevede, in primo luogo, l'individuazione di una soglia rischiosa per ognuna delle variabili e la costruzione di una matrice x_{ij} (con i = unità e j = indicatori semplici) costituita da 0 oppure da 1 a seconda che l'unità, rispettivamente, non superi o superi tale soglia. Si procede poi con la somma, spesso ponderata, delle righe e quindi dei valori registrati da ogni singola unità ottenendo il numero delle deprivazioni di ciascun individuo.

   Dopodiché si definisce un certo k, ossia un'ulteriore soglia anomala che in questo caso non riguarda le variabili ma gli indici individuati dalla somma, e si determina l'head count, cioè:

   $$\frac{\text{numero di unità con almeno k deprivazioni}}{\text{totale del campione}}$$

   Si ottiene quindi un valore (**head count ratio**) che rappresenta la percentuale di individui che non rispettino determinate caratteristiche definendosi rischiosi, ossia la percentuale di individui privi di certe caratteristiche minime.

   L'*head count* può essere anche perfezionato determinando l'*adjusted head count*, ossia moltiplicandolo per il numero delle dimensioni (delle caratteristiche osservate) [22];

2. **Income gap ratio (indice di intensità di reddito)**;

Permette di determinare quanto i soggetti con condizioni economiche scarse sono al di sotto della soglia di povertà in termini percentuali. Il calcolo consiste nel risolvere la formula 2.7 [22].

$$I = \frac{1}{q} \sum_{i=1}^{q} \frac{z - y_i}{z} \tag{2.7}$$

dove

z = soglia della povertà

q = poveri

3. **Poverty gap ratio (indice di intensità della povertà)**.

   In questo caso la media è calcolata sull'intera popolazione quindi determina quanto i ricchi dovrebbero trasmettere ai poveri perché questi raggiungano una soglia economica "decente" [22].

$$PG = \frac{1}{N} \sum_{i=1}^{q} \frac{z - y_i}{z} \tag{2.8}$$

4. **Squared poverty gap**.

   Trattasi di una media dei PG individuali ponderata per i PG stessi. Tanto più α è alto, tanto più i soggetti considerati più poveri sono al centro dell'attenzione [22].

$$P_\alpha = \frac{1}{N} \sum_{i=1}^{q} \left(\frac{z - y_i}{z} \right)^\alpha \tag{2.9}$$

5. **Sen-Shorrocks-Thon measure**.

   Permette di considerare le disparità esistenti anche tra i poveri e si compone dell'indice di Gini, dell'indice di intensità e dell'indice di diffusione. Esso può avere valore compreso nell'intervallo $[0, 1]$, dove un valore pari a 0 vuol dire che non vi sono soggetti al di sotto della soglia di riferimento [22].

$$S = H[I + (1 - i)G_q] \tag{2.10}$$

6. **Watts measure**.

   Una misura di povertà molto semplice che relaziona il reddito ad una linea di povertà [22].

$$W = \frac{1}{N}\sum_{i=1}^{q}[\ln z - \ln y_i] = \frac{1}{N\sum_{i=1}^{q}\ln\left(\frac{z}{y_i}\right)} \tag{2.11}$$

2.4 Gli indici di deprivazione

Appartengono alla famiglia degli indici di povertà ed esprimono quanto un soggetto della società o un gruppo della società stessa necessitino per raggiungere il tenore di vita minimo [22].
Possono essere calcolati sia con l'approccio continuo che con quello countable si possono qualificare come indici sia di tipo normativo che problem oriented, ma nel nostro caso saranno utilizzati con finalità soprattutto predittive. Possono inoltre definirsi come indici di stato composti da indicatori sintetici di tipo oggettivo.
L'informazione fornita mantiene un livello ben bilanciato tra qualità e comprensibilità e riguarda aspetti soprattutto di tipo economico-sociale (nel nostro caso si osserveranno le aziende perciò fenomeni di tipo economico).
In questa sede si focalizzerà l'attenzione sugli indici di deprivazione continui, solitamente utilizzati in ambito sanitario per confrontare le situazioni sanitarie nella società e nelle diverse classi sociali [22].

2.4.1 La costruzione degli indici di deprivazione

Nella costruzione degli indici di deprivazione continui, fondamentale risulta la scelta dei domini, degli indicatori semplici, le eventuali modalità di ponderazione ed aggregazione ed il tipo di ranking che viene effettuato.
Solitamente si ricorre all'uso di 7 domini per la loro costruzione: il reddito, l'occupazione, l'istruzione, i servizi, l'ambiente, la criminalità e la salute.
La costruzione di questi indici richiede di individuare i giusti criteri di ponderazione e di sintesi procedendo soggettivamente oppure attraverso procedure statistiche og-

gettive. Oltre questo, si necessita di una accurata selezione delle componenti che ne faranno parte e della giusta combinazione tra le stesse. Nel determinare i giusti indicatori elementari che li compongono si tiene conto della loro capacità interpretativa di almeno uno dei fattori latenti.
Individuati gli indicatori elementari questi vengono standardizzati attraverso il calcolo degli scarti dalla media nella popolazione in osservazione [22]. Dopodiché, si può effettuare la somma degli stessi in quanto trattasi di una metodologia di aggregazione molto elementare [25].
Un'altra modalità con cui riassumere è quella che prevede l'adozione di metodologie ordinali, ossia si passa per l'individuazione del valore di rango assunto dall'indicatore facendone poi la somma o la media dei ranghi individuati. Questo consente di ottenere indicatori comprensibili ed interpretabili ma portatori di informazioni molto limitate.
I passi da seguire nel metodo ordinale sono i seguenti [22]:

1. polarizzare gli indicatori;

2. calcolare l'indice di deprivazione sulla base degli indicatori stessi;

Per molti paesi sono disponibili indici di deprivazione aggregati. In Italia sono calcolati gli indici di deprivazione per ogni comune, rendendo possibile i confronti anche tra le regioni.

2.4.2 Un esempio di indice di deprivazione: l'indice di Caranci

Un tipo di indice di deprivazione molto usato è l'indice di deprivazione di Nicola Caranci, sviluppato nel 2009 ed espressivo della quantità di svantaggio sociale relativo. Questo indice individua la quantità di svantaggio sociale relativo tra i comuni del territorio italiano con lo scopo di identificare zone critiche da migliorare e di collegare tale svantaggio con la situazione raggiunta [22]. Esso si compone dei seguenti indicatori, tutti con polarità positiva:

- $X_1 \Rightarrow$ Percentuale di popolazione con istruzione pari o inferiore alla licenza elementare (dominio = istruzione)

$$\frac{\text{Popolazione con licenza elementare, alfabeta o analfabeta}}{\text{Popolazione di 6 anni ed oltre}} * 100 \tag{2.12}$$

- $X_2 \Rightarrow$ Percentuale di popolazione attiva disoccupata o in cerca di prima occupazione (dominio = occupazione)

$$\frac{\text{Forza lavoro - disoccupati o in cerca di prima occupazione}}{\text{Forza lavoro}} * 100 \tag{2.13}$$

- $X_3 \Rightarrow$ Percentuale di abitazioni occupate in affitto (domimnio = ambiente)

$$\frac{\text{Abitazioni occupate da persone residenti in affitto}}{\text{Abitazioni occupate da persone residenti}} * 100 \tag{2.14}$$

- $X_4 \Rightarrow$ Densità abitativa ogni 100 metri quadrati (dominio = ambiente)

$$\frac{\text{Popolazione totale}}{\text{Superficie (in mq) abitazioni occupate da persone residenti}} * 100 \tag{2.15}$$

- $X_5 \Rightarrow$ Percentuale di famiglie monogenitoriali con figli dipendenti conviventi (dominio = condizione sociale)

$$\frac{\text{Padre o madre soli con figli}}{\text{Famiglie totali}} * 100 \tag{2.16}$$

Dati questi, l'indice si calcola come segue [22]:

$$ID = \sum_{i=1}^{5} Z_i \tag{2.17}$$

Dove

$Z_i = \frac{X_i - m_i}{s_i}$

$m_i =$ media indicatore X_i

$s_i =$ deviazione standard indicatore X_i

$i = 1, 2, 3, 4, 5$

Capitolo 3

PCA, Analisi Discriminante e Modello Logit

Gli studi effettuati in questo capitolo hanno lo scopo di individuare ulteriori approcci nella costruzione dei modelli di previsione attualmente esistenti.
Si considerano come dati di partenza quelli adottati nei modelli di previsione di Poli per poi procedere con lo svolgimento di tre tipologie di analisi: l'analisi delle componenti principali, l'analisi discriminante e la regressione logistica.

3.1 Definizione del campione e delle variabili

Innanzitutto, si precisa che il campione su cui è stata effettuata l'analisi e, la successiva verifica, è costituito dalle 2264 aziende considerate nei modelli di previsione di Poli, sommando le aziende del campione *train* con quelle del campione *test* [35].
In questo caso, quindi, le performance del modello vengono analizzate sullo stesso campione alla base della costruzione dello stesso, questo la rende meno attendibile ed elemento migliorabile in eventuali verifiche successive [14].
Di queste aziende sono stati raccolti anche i dati utilizzati nei modelli di previsione costruiti da Poli, ed in particolare i valori assunti dalle seguenti variabili:

1. Incidenza del margine di struttura primario

2. Incidenza del margine di struttura secondario
3. Grado di dipendenza finanziaria
4. Indice di esigibilità del debito
5. Indice di liquidità di primo livello
6. Indice di liquidità di secondo livello
7. Indice di liquidità di terzo livello
8. Incidenza del margine di tesoreria di primo livello
9. Incidenza del margine di tesoreria di secondo livello
10. Incidenza del margine di tesoreria di terzo livello (o del capitale circolante netto)
11. Incidenza del valore attuale sui ricavi
12. Incidenza dell'EBITDA (o MOL) sui ricavi
13. Incidenza del reddito operativo sui ricavi
14. Incidenza dell'EBIT sui ricav
15. Incidenza dell'EBT sui ricavi
16. Incidenza del risultato netto sui ricavi
17. Dimensione 1
18. Dimensione 2
19. Età dell'azienda
20. Settore di appartenenza dell'attività svolta
21. Stato di fallita o non fallita (dove lo stato "fallita" corrisponde a 1 mentre lo stato di "non fallita" corrisponde a 0)

3.2 La selezione delle variabili

Una delle prime modalità di analisi è l'uso delle componenti principali.
Abbiamo visto nella costruzione dei modelli di Poli che, nella selezione delle variabili, erano state utilizzate tre metodologie differenti determinando altrettanti modelli. In questo caso si vuole considerare ancora un'altra tecnica che permetta di raggiungere tale scopo, la cosiddetta analisi delle componenti principali.
Per svolgerla si deve prima studiare la correlazione tra le variabili, perciò vengono considerate le prime 19 elencate nel paragrafo 3.1 e si effettua tale studio, dapprima, senza l'eliminazione degli outlier (quindi, analizzando tutti i dati) per poi ripetere la procedura in seguito all'esclusione dei dati estremi attraverso l'applicazione della winsorizzazione, ovvero una delle tecniche più utilizzate nel trattamento dei dati di bilancio.
In particolare, quando si effettuano delle analisi dei dati è molto frequente la presenza di dati estremi alteranti lo studio, in quanto si rischia di concentrare l'attenzione su elementi che in realtà non sono rappresentativi della distribuzione. Ciò può essere risolto attraverso la winsorizzazione e cioè l'eliminazione dei dati estremi [13].
Nelle figure 3.1 e 3.2 si riportano l'esempio di come due delle variabili registrino distribuzioni di densità molto diverse dopo l'eliminazione degli outlier.
Nei grafici a destra (densità delle distribuzioni dopo aver applicato la tecninca della winsorizzazione) si può vedere come tali distribuzioni sono molto più visualizzabili e quindi analizzabili, dimostrando che in alcuni casi tale operazione è fondamentale.

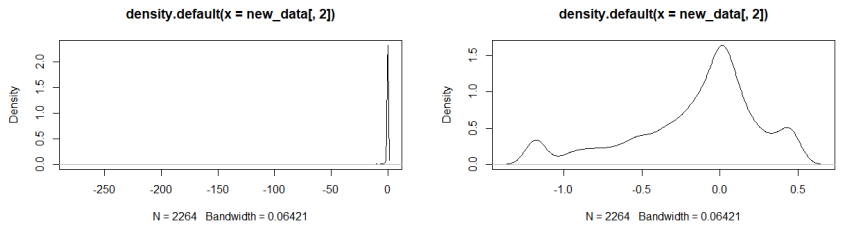


Figura 3.1: Distribuzione di densità della variabile margine di struttura primario prima e dopo la winsorizzazione

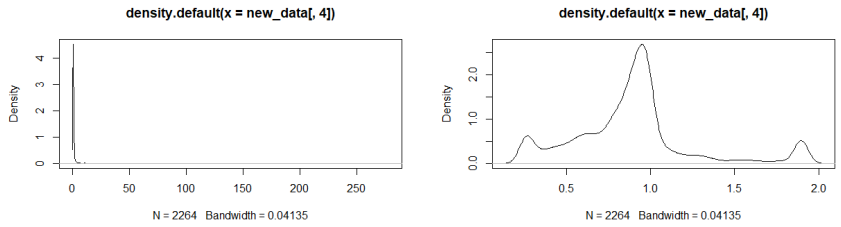


Figura 3.2: Distribuzione di densità della variabile grado di dipendenza finanziaria prima e dopo la winsorizzazione

Per evitare confusioni, si precisa che l'eliminazione degli outlier e la successiva analisi della correlazione sono state effettuate anche nel primo modello costruito da Poli, ma in questo caso le variabili di partenza sono differenti.

3.3 La correlazione di Pearson

L'indice di correlazione di Pearson, anche detto coefficiente di correlazione lineare, è una misura standardizzata che esprime la relazione di linearità tra due variabili statistiche. Essa fornisce sia la direzione che l'intesità del legame e la si calcola nel seguente modo [32]:

$$\frac{Cov(x,y)}{\sigma_x \sigma_y} \tag{3.1}$$

$Cov(x,y)$: covarianza tra le varibili x e y

σ_x: deviazione standard della variabile x

σ_y: deviazione standard della variabile y

Tale indice può assumere un valore compreso tra +1 e -1, dove gli estremi corrispondono, rispettivamente, ai casi di perfetta correlazione lineare diretta e di perfetta correlazione lineare inversa. Perfetta correlazione lineare perché, riportando tutti i dati all'interno di un grafico a dispersione, questi potrebbero essere congiunti attraverso una linea retta[32].

Qualora il coefficiente di correlazione sia pari a zero, le variabili si diranno non correlate linearmente e, in alcuni casi, potrebbero essere anche indipendenti tra loro ma, se l'indipendenza è sempre sintomo di incorrelazione, l'incorrelazione è condizione

necessaria ma non sufficiente per l'indipendenza[8].

Volendo studiare la correlazione tra le 19 variabili considerate, attrverso il comando `cor(as.matrix())`, si costruisce una matrice di correlazione, ovvero una matrice quadrata simmetrica 19x19 che riporta i valori dei coefficienti di Pearson tra le variabili. Tuttavia, data la difficoltà di riportare tutti i valori dei coefficienti in una sola immagine, sono stati inseriti nella matrice rappresentata nella figura 3.3 dei punti colorati in base al valore assunto dai coefficienti.

In particolare, sono visibili solo quelli che presentano una correlazione rilevante: tanto più vi è un valore lontano dallo 0 tanto più il colore è intenso (nel caso di correlazione negativa si visualizzano diverse gradazioni di rosso, mentre per quella positiva si hanno colori tendenti al blu).

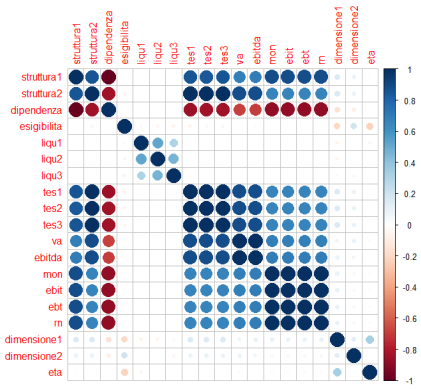


Figura 3.3: Matrice di correlazione

Con lo scopo di selezionare le variabili è stata effettuata una prova d'ipotesi sui risultati della correlazione. In generale, quest'ultima si effettua con l'analisi campionaria casuale e, fino a quando non vi siano sufficienti evidenze empiriche per rifiutare l'ipotesi formulata (ipotesi nulla), questa non verrà rigettata [32].

Per decidere è necessario stabilire un certo livello di significatività minimo tale per cui, qualora raggiunto, si possano riscontrare sufficienti evidenze empiriche in grado di dimostrare che l'ipotesi nulla potrebbe non aver luogo perché le probabilità sono molto basse (errore di prima specie) e quindi comportare la validazione dell'ipotesi alternativa.

Per determinare se le evidenze empiriche siano sufficientemente in grado di comportare il rifiuto dell'ipotesi nulla o meno si possono adottare due modalità [32]:

- si costruisce la distribuzione della statistica test in base alle ipotesi formulate e si stabiliscono i valori della stessa che avrebbero una bassa probabilità di verificarsi se l'ipotesi nulla fosse vera.
  Si decide, quindi, in base al valore assunto dallo stimatore del parametro oggetto di studio;

- determinato il minimo livello di significatività per cui l'ipotesi nulla non può essere rifiutata, si calcola quello osservato (o p-value) e si effettua quindi la decisione.

Costruita la matrice di correlazione, si effettua il test di verifica d'ipotesi sull'assenza di correlazione tra una coppia di variaibili aleatorie sugli indici, il quale viene risolto attraverso il calcolo del p-value. In particolare, in questo caso i p-value servono ad individuare se, sulla base di quanto analizzato sul campione, si possa dedurre in maniera significativa che i coefficienti di correlazione della popolazione siano diversi da zero. Questo avviene nel caso in cui si ottengano valori dei livelli di significatività osservati non rientranti nel limite prefissato, per cui si può rifiutare l'ipotesi nulla in favore dell'ipotesi alternativa [32]
Nello specifico, stabiliti come livello di significatività α=0.05, come ipotesi:

- H_0: coefficiente di correlazione $= 0$

- H_1 : coefficiente di correlazione $\neq 0$

e come area di rifiuto dell'ipotesi nulla (H_0) i valori di p-p-value inferiori a 0.05, si eseguono i seguenti comandi per ottenere l'immagine riportata nella figura 3.4:

- `as.matrix()`;

- `dev.new()`;

- `pval` $\leftarrow$ `psych::corr.test(M.adjust="none")$p`;

- `corrplot(round(pval,1),method="circle")`.

La figura riporta la matrice dei p-value osservati sulla popolazione attestanti la veridicità o meno della correlazione tra le coppie di variabili, sempre inserendo dei punti colorati in base al valore assunto dagli stessi e solo per quelli rilevanti.
Da notare che questa seconda matrice è, in valore assoluto, complementare alla precedente, attestando infatti che, qualora vi siano dei p-value elevati, la correlazione sarà assente e quindi nella matrice di correlazione la casella corrispondente sarà bianca e viceversa.

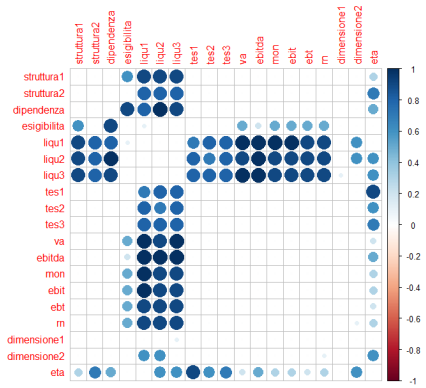


Figura 3.4: Matrice di p-value

Prima di trattare l'analisi delle componenti principali si procede con le stesse procedure di studio della correlazione sui dati ridotti di quelli estremi.

Eliminazione degli outlier e correlazione di Pearson

Il comando di *RStudio* che ha svolto la winsorizzazione è `lapply([,-c(1,21)], Winsorize)`, ciò ha permesso di originare le nuove matrici di correlazione riportate nella figura 3.5.
Si può notare come le variabili registrino in questo secondo caso una correlazione molto più alta, mentre la seconda complementare alla prima mostra p-value molto più piccoli.

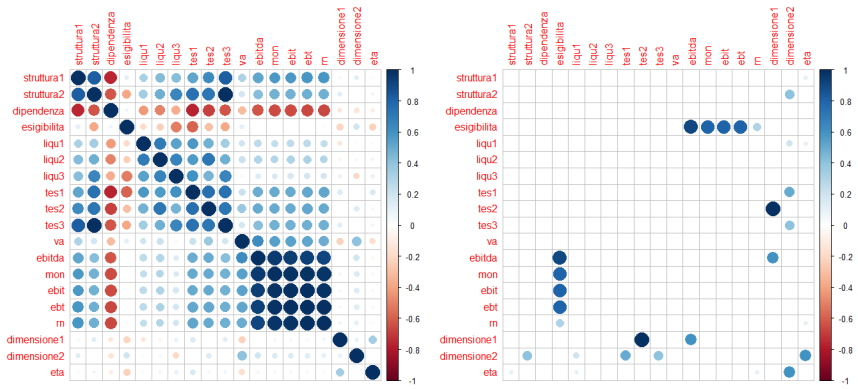


Figura 3.5: Matrici analisi correlazione con winsorizzazione

3.4 L'analisi delle componenti principali

Sulle matrici di correlazione tra le variabili è stata svolta una risoluzione matriciale con lo scopo di effettuare l'analisi delle componenti principali.

La *principal component analysis* (P.C.A.) è una procedura, applicata nella statistica multivariata, per ridurre il numero di variabili utilizzate nell'analisi dei dati ad un insieme di variabili latenti in grado di produrre la maggiore informazione possibile. Infatti, si tratta di una tecnica adottata in molti ambiti proprio perché, individuando la quantità adeguata di autovettori, permette di minimizzare il "trade-off" esistente tra la perdita di informazioni e la facilitazione dello studio [21].

Ciò avviene attraverso la determinazione di nuove variabili, tante quante sono quelle originarie ma differenti da queste pur se frutto di una loro combinazione lineare [1]. Prima di vederne il procedimento, è necessario introdurre i concetti di autovalore, autovettore e componente principale.

Autovettore e componente principale

In questo caso si parla degli autovettori della matrice delle covarianze, ovvero delle direzioni verso cui gli assi caratterizzati dalla maggiore concentrazione di varianza sono orientati. Essi, quindi, denominati anche componenti principali, sono costruiti come correlazione lineare delle variabili originarie e sono tra loro non correlate [23].

Autovalore

Ogni autovettore ha un proprio autovalore, quindi un coefficiente che rappresenta la varianza della singola componente. L'autovalore rapportato alla somma di tutti gli autovettori permette di calcolare un elemento molto importante: la percentuale di varianza totale spiegata dalla componente in questione [23].

3.4.1 Procedimento per l'analisi delle componenti principali

Il procedimento attraverso il quale effettuare l'analisi delle componenti principali è il seguente [21]:

1. standardizzazione delle variabili

   Si crea la matrice X composta dai dati di partenza dove le colonne si costituiscono delle osservazioni mentre le righe delle variabili considerate. Essa può essere rappresentata come segue.

   $$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

   Ogni valore della matrice viene poi standardizzato con lo scopo di ridurre il range di variazione dei valori e di ottenere un set di variabili con medesima scala di misura. Infatti, in certe situazioni anziché sottrarre la media, si cerca di standardizzare la matrice iniziale, soprattutto quando si stanno confrontando diverse caratteristiche che hanno differenti unità di misura o ordine di grandezza, per garantire un confronto equo tra le stesse.

   Ciò viene svolto attraverso la risuluzione della seguente equazione:

   $$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}$$

   dove $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, p$ [23];

2. analisi della correlazione

Si controlla la sussistenza di eventuali relazioni tra le variabili. Ciò viene fatto attraverso la costruzione della matrice delle covarianze e delle varianze delle variabili standardizzate, la quale corrisponde alla matrice di correlazione tra le variabili originarie X_{ij} [23];

3. identificazione delle componenti principali

   Si calcolano gli autovettori della matrice creata nel passaggio precedente e si procede con il calcolo del loro rispettivo autovalore.
   Effettuato ciò, si procede con la disposizione in ordine decrescente degli autovalori calcolati e con la selezione dei primi k auovettori corrispondenti, dove k è tendenzialmente minore di p e dipende dalla soggettività di chi compie l'analisi. Si ottengono così solo le componenti che forniscono più informazioni sulla distribuzione dei dati [23];

4. creazione della matrice dei vettori futuri

   Si costruisce una matrice in cui si riportano gli autovettori selezionati nelle colonne per ordine decrescente dei rispettivi autovalori [23];

5. riorientamento dei dati

   Viene effettuato il prodotto tra la matrice trasposta degli autovettori futuri e la matrice trasposta del data set originario. Si ottiene quindi la matrice delle componenti principali composta dai dati originali trascritti tenendo conto degli autovettori futuri come assi di orientamento [23].

3.4.2 Svolgimento dell'analisi delle componenti principali

Con lo scopo di ridurre il numero di variabili da analizzare e di mantenere solo quelle realmente portatrici di informazioni può essere quindi effettuata l'analisi delle componenti principali sulla base della matrice di correlazione.
Nel nostro caso, attraverso il comando `princomp()` si ottiene una tabella dalla quale si può vedere che vi sono 10 variabili in grado di spiegare quasi tutta la varianza, in particolare il $99,87178\%$. La tabella è la seguente.

Componenti	Deviazione standard	Proporzione di varianza spiegata	Proporzione varianza spiegata cumulata
1	3.163185	0.5266180	0.5266180
2	1.377105	0.0998115	0.6264300
3	1.263871	0.0840721	0.7105020
4	1.196181	0.0753078	0.7858100
5	1.064573	0.0596482	0.8454580
6	0.840951	0.0372210	0.8826790
7	0.821395	0.0355099	0.9181890
8	0.789436	0.0328005	0.9509900
9	0.692223	0.0252196	0.9762090
10	0.65397	0.0225090	0.9987180
11	0.13749	0.0009949	0.9997130
12	0.05082	0.0001359	0.9998490
13	0.04383	0.0001010	0.9999500
14	0.02862	0.0000431	0.9999930
15	0.00798	0.0000034	0.9999963
16	0.00674	0.0000024	0.9999987
17	0.00444	0.0000010	0.9999997
18	0.00192	0.0000002	0.9999999
19	0.00141	0.0000001	1.0000000

Ciò vuol dire poter effettuare delle analisi su 10 piani piuttosto che su 19 potendo comunque spiegare il $99,87\%$ della variabilità totale delle osservazioni. Si tratta di un livello di informazione fornita molto elevato, in quanto utilizzando circa la metà delle variabili è possibile descrivere quasi tutta la varianza.

3.4.3 Confronto con winsorizzazione

Effettuando invece l'analisi delle componenti principali sulla base della matrice di correlazione della figura 3.5, ovvero dopo aver eliminato gli outlier, si conduce ad un

risultato in cui si determina la possibilità di poter spiegare all'incirca lo stesso livello di varianza utilizzando 15 componenti piuttosto che 10. Si tratta di un numero molto più elevato che testimonia come, nel nostro caso, i valori estremi siano altamente incisivi sui risultati delle analisi che vengono svolte.

3.5 L'analisi discriminante

Un altro metodo attraverso il quale si può semplificare lo studio è l'analisi discriminante, metodo della statistica descrittiva multidimensionale che ha lo scopo di determinare delle modalità attraverso le quali suddividere le unità statistiche in più raggruppamenti attraverso una relazione funzionale di alcune variabili in uno dei g gruppi [40]. Il procedimento prevede, innanzitutto, lo svolgimento dell'analisi discriminante, la quale può essere effettuata attraverso delle assunzioni sulla distribuzione della sottopopolazione oppure non effettuando alcun tipo di ipotesi sulla stessa.
Del secondo tipo sono quelle che adottano il metodo di Fisher, uno dei primi ad approcciarsi a tale tipologia di studi. Egli effettuò una classificazione sulla base della determinazione di una funzione lineare che assuma medie il più differenti possibile per ciascuno dei g gruppi.
Questo tipo di studio viene molto applicato quando non si conosce la distribuzione del campione [21].
Quando si conosce la distribuzione della sottopopolazione, invece, è possibile effettuare un'analisi discriminante adottando il metodo della massima verosimiglianza. Ciò vuol dire effettuare una classificazione dell'osservazione x nel raggruppamento per il quale la verosimiglianza è maggiore.
Quindi, x viene considerata come parte di un gruppo k qualora la probabilità che essa appartenga a quel raggruppamento sia maggiore di quella di appartenza ad un altro [21]. L'analisi discriminante lineare assume che ogni classe sia distribuita come una normale e che abbiano tutte pari variabilità. Al contario, l'analisi discriminante quadratica, pur sempre ipotizzando una distribuzione normale delle suddivisioni, ipotizza che ognuna di queste abbia variabilità differente, comportando regole deci-

sionali non lineari [21].

Successivamente, si procede con il calcolo delle **probabilità a priori**. Infatti, vi sarà un rischio di effettuare una categorizzazione errata, il quale aumenta al crescere del numero dei raggruppamenti, definendo una probabilità di collocazione esatta a priori inversamente proporzionale a tale quantità o a qualsiasi altra regola a scelta [7].

In qualità di distribuzione uniforme discreta, la probabilità può essere calcolata come $P(x) = \pi_K = \frac{N_K}{N}$ dove N_K è il numero degli elementi osservati nel gruppo k ed N è l'ampiezza della popolazione [7].

In questo caso si ottiene un dato "grezzo" che può essere migliorato attraverso il calcolo delle **probabilità a posteriori**. In particolare, qualora vi sia effettiva attinenza alla classificazione k si otterrà una suddivisione esatta, nel caso contrario sarà sbagliata.

L'errore complessivo si concretizza nell'insieme delle suddivisioni del secondo tipo [21].

In sostanza, l'analisi discriminante si può ritenere, al pari dell'analisi delle componenti principali, un altro metodo attraverso il quale effettuare una semplificazione dei calcoli in quanto mira a ridurre lo spazio di analisi in una misura inferiore rendendo più semplice la suddivisione delle variabili [10].

Per calcolare la probabiltà a posteriori si utilizza il metodo di Bayes [26]. Esso dice che:

- $P(A|B)P(B) = P(B|A)P(A) \Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;
- $P(B) = P(B \cap A) \cup P(B \cap \overline{\mathrm{A}}) \Rightarrow P(B) = P(B|A)P(A) + P(B|\overline{\mathrm{A}})P(\overline{\mathrm{A}})$

Allora si può affermare che $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\overline{\mathrm{A}})P(\overline{\mathrm{A}})}$.

Data la funzione di densità di X condizionata a $Y = k$ $f_k(x) = Pr(X = x|Y = k)$, è possibile applicare Bayes e quindi calcolare la probabilità a posteriori $p_k(x)$ nel seguente modo:

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{g} \pi_j f_j(x)} \tag{3.2}$$

Calcolate le probabilità a posteriori sarà possibile misurare le performance del modello [21].

3.5.1 Svolgimento dell'analisi discriminante

Svolgendo l'analisi discriminante di Fisher si effettua innanzitutto una riduzione delle dimensioni da osservare, poi si calcolano le funzioni discriminanti in grado di determinare le soglie delimitative delle aree di classificazione (aree che classificano gli individui come parte di un gruppo piuttosto che di un altro).
La riduzione dimensionale avviene attraverso la considerazione della direzione (autovettore) che massimizzi le codevianze campionarie *between* (tra le medie campionare) e minimizzi quelle *within* (intorno alle medie campionarie). Ciò è dovuto al fatto che tanto più le prime risultino maggiori delle seconde, tanto più si potranno differenziare tra loro le medie campionarie stesse e quindi individuare l'elemento discriminante [27].
L'obiettivo è perciò quello di costruire le seguenti funzioni discriminanti z_i tali che il rapporto tra le codevianze campionarie between e quelle within (assunto uguale) sia massimo.

$$z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

$$z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$$

$$\cdots$$

$$z_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{np}x_p$$

Ad esempio, dati soli due gruppi si vuole fare in modo che sia massimo il rapporto:

$$\frac{\frac{[a^t(\bar{x}_1 - \bar{x}_2)]^2 n_1 n_2}{n_1 + n_2}}{a^t S a}$$

dove

$a^t = \begin{pmatrix} a_{i1} & \dots & a_{ip} \end{pmatrix}$;

$\bar{x}_1$ = media campionaria del gruppo 1;

$\bar{x}_2$ = media campionaria del gruppo 2;

n_1 = ampiezza del gruppo 1;

n_2 = ampiezza del gruppo 2;

S = covarianza campionaria dei gruppi;

$$a = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix}$$

All'interno del programma *RStudio* è stata svolta l'operazione `lda()` dove è stata indicata un'espressione in grado di mettere insieme:

- la classe di riferimento individuata;
- i valori di previsione da utilizzare;
- il data frame inclusivo dei dati;
- altri elementi quali la probabilità a priori.

Il risultato sarà dato dall'elemento discriminante sulla base del quale svolgere la formula `predict()` e quindi effettuare la categorizzazione, sarà quindi possibile valutare le performance del modello.
Applicando l'analisi discriminante al nostro caso, questa verrà svolta su quattro insiemi di dati differenti:

1. considerazione di tutti i dati, con l'esclusione della variabile "settore di appartenenza";
2. considerazione di tutti i dati, compresa anche la variabile "settore di appartenenza";
3. considerazione dei dati ottenuti dall'eliminazione degli outlier, con l'esclusione della variabile "settore di appartenenza";
4. considerazione dei dati ottenuti dall'eliminazione degli outlier, compresa anche la variabile "settore di appartenenza";

Svolgimento analisi discriminante non considerando la variabile "settore di appartenenza"

Le performance del modello vengono controllate, dapprima, attraverso la costruzione di una tabella, con il comando `xtabs()`, all'interno della quale confrontare i dati osservati ed i dati preventivati.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	752	330
	in crisi	380	802

Si svolgono poi le formule 1.1 e 1.2 ottenendo un'*accuracy* pari al $68,63958\%$ ed una *specificity* del $66,431095\%$.

$$accuracy = \frac{802 + 752}{2264} * 100 = 68,63958\%$$

$$specificity = \frac{752}{752 + 380} * 100 = 66,431095\%$$

Svolgimento analisi discriminante considerando la variabile "settore di appartenenza"

Si considera ora anche il settore di appartenenza, verificando se, tenendo conto del fatto che ogni settore abbia una rischiosità differente, si raggiunga una performance migliore. Da questa analisi si ottengono i seguenti risultati:

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	750	320
	in crisi	382	812

Svolgendo le formula 1.1 e 1.2 si ottengono un'*accuracy* pari al $68,99293\%$ ed una *specificity* pari a $66,254417\%$

$$accuracy = \frac{812 + 750}{2264} * 100 = 68,99293\%$$

$$specificity = \frac{750}{750 + 382} * 100 = 66,254417\%$$

Svolgimento dell'analisi discriminante considerando i dati ottenuti dall'eliminazione degli outlier

Effettuando le stesse analisi dopo aver eliminato gli outlier si ottiene quanto segue:

- analisi discriminante senza outlier e senza la variabile settore di appartenenza

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	808	267
	in crisi	324	865

In questo caso, si raggiungono livelli di *accuracy* e di *specificity* pari, rispettivamente al $73,89576\%$ e al $71,378092\%$

$$accuracy = \frac{865 + 808}{2264} * 100 = 73,89576\%$$

$$specificity = \frac{808}{808 + 324} * 100 = 71,378092\%$$

- analisi discriminante senza outlier e con la variabile settore di appartenenza

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	822	268
	in crisi	310	864

Applicando le formule 1.1 e 1.2 si determinano un'*accuracy* del $74,46996\%$ e una *specificity* del $72,614841\%$

$$accuracy = \frac{864 + 822}{2264} * 100 = 74,46996\%$$

$$specificity = \frac{822}{822 + 310} = 72,614841\%$$

Confronto

Effettuando l'eliminazione degli outlier, si ottiene sia nel caso in cui si tenga conto della variabile "settore di appartenenza", sia nel caso non se ne tenga conto, un netto miglioramento della bontà del modello.
Ovviamente, la considerazione del livello di rischiosità del settore di appartenenza lo rende più affidabile.

3.6 La regressione logistica

In quanto funzioni in grado di esplicare le variazioni di una determinata variabile dipendente Y per ogni valore assunto da una o più variabili indipendenti X, le funzioni di regressione sono molto utilizzate per effettuare previsioni. Tale proprietà dell'analisi in questione, nel caso di sistemi di regressione lineare, incontra tuttavia, delle limitazioni quando ci si trova di fronte a particolari distribuzioni quali la distribuzione di Poisson o la binomiale. Per questo motivo, si è sviluppato un sistema di analisi denominato **regressione logistica**, il quale è in grado di determinare una funzione che individui i valori della variabile dipendente per ogni valore assunto da quelle indipendenti, superando l'ostacolo di casistiche peculiari.
La regressione logistica, applicata anche nei modelli costruiti da Poli, viene svolta in questo lavoro con lo scopo di riscontrare delle differenze in termini di performance qualora si parta dalla considerazione di variabili differenti. In particolare, si procede con la sua implementazione sia sui dati della distribuzione originari, sia su quelli trattati con la tecnica della winsorizzazione. Soprattutto quest'ultimo risulta essere altamente confrontabile con i modelli costruiti da Poli [35], in quanto si distinguono semplicemente per le modalità di lavorazione degli outlier e per le variabili considerate.
Per poter spiegare in cosa consiste questa tipologia di analisi è necessario introdurre dapprima le analisi di regressione lineare **semplice** e **multipla** nelle quali si assume che gli errori aleatori siano pari 0.

3.6.1 La regressione lineare semplice

Per studiare il rapporto esistente tra le variabili oggetto di analisi viene solitamente utilizzata la regressione. Nella sua configurazione più semplice, essa viene definita **regressione lineare semplice** e si tratta di una specifica relazione funzionale che individua i mutamenti della variabile dipendente Y ad ogni cambiamento della variabile indipendente X. Essa può essere approssimata con l'equazione lineare sottostante, la quale rappresenta l'insieme delle coppie costituite dai valori della X e della Y dipendente dai primi.

$$Y = \beta_0 + \beta_1 X \tag{3.3}$$

β_0: intercetta della retta di regressione

β_1 : coefficiente angolare della retta di regressione

Tale funzione, graficamente, si esplica nella migliore curva di rappresentazione dei dati osservati, i quali verranno utilizzati per calcolare le stime di β_0 e β_1 attraverso il metodo dei minimi quadrati. Si otterrano quindi gli stimatori b_0 e b_1.

Il metodo dei minimi quadrati prevede che venga minimizzata la somma dei quadrati degli errori, ossia la differenza tra $Y = \beta_0 + \beta_1 X$ e $\hat{y} = b_0 + b_1 x$. Ciò vuol dire minimizzare la seguente sommatoria: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ [32]. Da ciò si ottengono le seguenti funzioni di determinazione degli stimatori dei coefficienti della retta di regressione:

$$b_0 = \bar{y} - b_1 \bar{x} \tag{3.4}$$

$$b_1 = \frac{S_(xy)}{(S_x)^2} \tag{3.5}$$

Dove:

$\bar{x}$= media campionaria della variabile x

$\bar{y}$= media campionaria della variabile y

$S_(xy)$= covarianza campionaria tra le variabili x e y

S_x= deviazione standard campionaria della variabile x

3.6.2 La regressione lineare multipla

Qualora le variabili rappresentative dei dati siano più di 2, si tende ad utilizzare il modello di **regressione lineare multipla**, ossia una relazione funzionale attraverso la quale si rende possibile determinare, apportando il minor errore possibile, i valori registrati dalla variabile dipendente Y per ogni valore assunto dalle altre variaibli indipendenti X_j, con $j = 1, 2, ..., k$. In questo caso la relazione si concretizza nella seguente equazione a più variabili:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_k \tag{3.6}$$

Per determinare il valore degli stimatori dei coefficienti della retta è necessario minimizzare SSE.

$$\min SSE = \min \sum_{i=1}^{n} (y_i - \hat{y_i}) = \min \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + ... + b_k x_{ki})]^2$$

Si ottengono quindi (nel caso di solo due variabili indipendenti):

$$b_0 = \bar{y} - b_1 \bar{x_1} - b_2 \bar{x_2} \tag{3.7}$$

$$b_1 = \frac{s_y(r_{x_1,y} - r_{x_1,x_2} * r_{x_2,y})}{s_{x_1}(1 - r^2_{x_1,x_2})} \tag{3.8}$$

$$b_2 = \frac{s_y(r_{x_2,y} - r_{x_1,x_2} * r_{x_1,y})}{s_{x_2}(1 - r^2_{x_1,x_2})} \tag{3.9}$$

Dove:

$\bar{x_1}$= media campionaria della variabile x_1

$\bar{x_2}$= media campionaria della variabile x_2

S_y= deviazione standard campionaria della variabile y

s_{x_1}=deviazione standard campionaria della variabile x_1

s_{x_2}=deviazione standard campionaria della variabile x_2

$r_{x_1,y}$=coefficiente di correlazione campionario tra le variabili x_1 e y

$r_{x_2,y}$=coefficiente di correlazione campionario tra le variabili x_2 e y

r_{x_1,x_2}=coefficiente di correlazione campionario tra le variabili x_1 e x_2

I coefficienti β_j rappresentano l'effetto delle variabili indipendenti su quella dipendente, quindi tali coefficienti sono condizionati anche dalle altre variabili [32].

3.6.3 La regressione logistica

Se il legame individuato è lineare, il sistema noto come regressione lineare può essere riprodotto nel grafico a dispersione attraverso una linea retta.
Diversa è la regressione logistica, ossia un sistema di regressione paragonabile a quello lineare multivariato ma in grado di fornire come variabile finale una categoria. In questo caso, differentemente dalle regressioni lineari, il sistema funziona anche se le variabili non siano né connesse linearmente, né normalmente distribuite né omoschedastiche [32]. Questo perché se per esempio, come in questo caso, ci si trovi di fronte ad una variabile dicotomica il codominio della funzione di regressione lineare potrebbe fornire valori appartenenti a tutto l'insieme R, non essendo in grado in questo modo di fornire alcuna classificazione [9]. Questo potrebbe comportare valutazioni di probabilità al di fuori dell'intervallo $[0, 1]$.
Mentre attraverso la regressione logistica, in quanto funzione non lineare, si otterrebbe una derivata prima subordinata alla X, ossia mutabile a seconda del valore assunto da quest'ultima [12].
Inoltre, bisogna tenere in considerazione due aspetti:

- se le variabili dipendenti dovessero essere più di 2, il sistema lineare diverrebbe inappropriato;

- nei casi precedenti, le variabili erano di tipo quantitativo e la risposta finale era una funzione dello stesso genere, nella regressione logistica i dati possono essere sia di tipo quantitativo che categoriale ma l'output è sempre del secondo genere. Per quanto riguarda la variabile qualitativa si utilizzano le variabili dummy così come accade nel modello a regressione lineare.

Si introduce quindi il procedimento adottato nella regressione logistica dove, invece che individuare il valore di Y, si costruisce un modello determinante la probabilità che Y appartenga ad una particolare categoria. Ipotizzando inizialmente la sussistenza di una sola variabile indipendente, ossia X, il sistema di regressione in questione permette di stimare la probabilità che Y possa essere classificato nella categoria 1

al variare dei valori della X, ossia $P(X) = P(Y = 1|X)$, con probabilità compresa nell'intervallo $[0, 1]$. Da questa stima sarà possibile costruire il modello grazie al quale individuare la collocazione [12].

In particolare si utilizza la seguente funzione logistica

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{3.10}$$

che approssimata diviene

$$\frac{P(X)}{1 - P(X)} = e^{\beta_0+\beta_1 X} \tag{3.11}$$

Il primo membro, chiamato **odds**, può assumere qualsiasi valore compreso nell'intervallo $[0, \infty)$. In particolare, un valore prossimo allo 0 indicherebbe una probabilità molto bassa di *default*, viceversa per valori molto alti. L'odds permette di adattare la Y, funzione dicotomica, alle caratteristiche della X, piano dimensionale [12].

Qui di seguito vengono fornite le modalità di calcolo della probabilità (formula 3.13) e del logit (formula 3.14):

$$P(X) = \frac{odds}{1 + odds} \tag{3.12}$$

$$\ln odds = \beta_0 + \beta_1 X \tag{3.13}$$

In questo caso quindi i coefficienti β_0 e β_1 corrispondono ad una trasformazione di Y e non direttamente alla probabilità, la quale aumenta o diminuisce, in modo non proporzionale, a seconda che i valori dei coefficienti siano, rispettivamente, positivi o negativi. Dalla formula 3.12, inoltre, si può vedere come il coefficiente e^{β_1} indichi la variazione dell'odds in corrispondenza di una variazione di X poiché $e^{\beta_0+\beta_1 * X} = e^{\beta_0} e^{\beta_1^X}$.

Per questo motivo, solitamente, lo studio comprende anche l'osservazione di tale valore [6]. Per quanto riguarda la stima dei coefficienti, il metodo dei minimi quadrati non è adeguato al sistema di regressione in considerazione, ma si tende ad utilizzare il metodo della massima verosimiglianza [12]. Dato che lo stimatore è uno stimatore di massima verosimiglianza si utilizza lo Z-Score.

Nel caso si abbiano più variabili indipendenti $X = x_1, X_2, ..., X_p$, le formule 3.11 e

3.12 divengono

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{3.14}$$

$$logit(P(X)) = \log \frac{P(X)}{1-P(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{3.15}$$

La regressione logistica può essere applicata anche per variabili dipendenti che si caratterizzino su più di 2 suddivisioni ma, in questi contesti, l'analisi discriminante è molto più utilizzata e semplice da svolgere [12].

3.6.4 Svolgimento della regressione logistica su tutti i dati

Come detto, data la sua capacità di prevedere i mutamenti delle variabili dipendenti al variare di quelle indipendenti, la regressione viene molto utilizzata nei modelli di previsione. Nel nostro caso, risultando molto più adatto, è stato utilizzato un sistema di regressione logistico, applicato anche nei modelli di Poli. Ora verrà sviluppato utilizzando le variabili che si stanno considerando in questo lavoro per effettuare un confronto tra i risultati e lasciare spazio ad eventuali studi futuri nel caso si giunga all'ottenimento di maggiori consapevolezze circa le informazioni che potrebbero rendere più attendibile un modello di previsione della crisi.
Svolgendo tale analisi attraverso i comando `logit` e `predict()` si ottiene una bontà della previsione del $74,64664\%$ e una specificità del $72,349823\%$.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	819	261
	in crisi	313	871

$$accuracy = \frac{871+819}{2264} = 74,64664\%$$

$$specificity = \frac{819}{819+313} = 72,349823\%$$

Volendo visualizzare la bontà del sistema, indipendentemente dalla sogglia individuata, si svolge una rappresentazione della curva ROC.

La curva ROC

La curva ROC consiste in una rappresentazione grafica delle due tipologie di errore commesse per tutte le soglie possibili che sarebbero potute essere scelte. Al di sotto della curva si trova la bontà del modello, all'aumentare dell'area aumenta anche l'affidabilità della previsione [38].

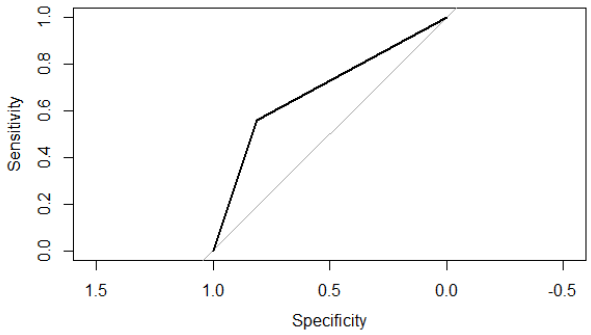


Figura 3.6: Curva roc regressione logistica

Idealmente si dovrebbe avere una curva ROC pari ad un angolo retto con vertice in alto a sinistra [4]. Nel nostro caso si ottiene quanto riportato nella figura 3.5, pari a un'area sottesa alla curva pari a $0,6842$ (figura 3.6).

3.6.5 Svolgimento della regressione logistica sui dati senza gli outlier

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	827	263
	in crisi	305	869

In questo secondo caso si ottiene una bontà del modello pari al $74,91166\%$, una specificità del $73,056537\%$ ed una curva roc con un'area sottesa alla curva pari a 0,6612 (figura 3.7).

$$accuracy = \frac{869 + 827}{2264} = 74,91166\%$$

$$specificity = \frac{827}{827 + 305} = 73,056537\%$$

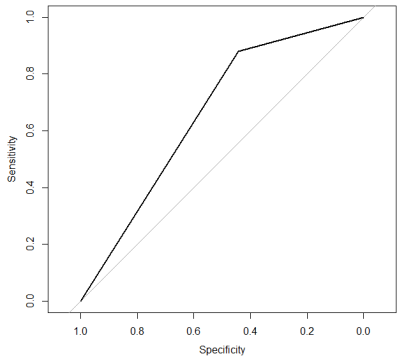


Figura 3.7: Curva roc regressione logistica con winsorizzazione

3.7 Confronto con i modelli di Poli

Ripresentando i livelli di *accuracy* ottenuti dai tre modelli di Poli e quelli ottenuti in questo paragrado si può effettuare un confronto.

Modello	*Accuracy*	*Specificity*
Poli 1	$74,31\%$	$70,96\%$
Poli 2	$74,05\%$	$71,59\%$
Poli 3	$74,05\%$	$74,12\%$
Logit su tutti i dati	$74,65\%$	$72,35\%$
Logit sui dati senza outlier	74.91%	$73,06\%$

Seppur minima, si trova una differenza di *accuracy* migliore nelle regressioni logistiche sviluppate in questa sede, mentre i livelli di *specificity* risultano essere minori rispetto al modello 3 di Poli [35] e maggiori degli altri due.

Capitolo 4

Analisi con indici di deprivazione

Come detto nell'introduzione, un metodo del tutto diverso ed innovativo attraverso il quale poter effettuare una previsione della crisi d'impresa potrebbe essere quello di costruire un modello attraverso l'utilizzo degli indici di deprivazione spiegati nel capitolo 2.
Infatti, se questi esprimono quanto ad un soggetto manchi per poter raggiungere un tenore di vita almeno "dignitoso", lo stesso può essere fatto per le aziende, ovvero si vuole individuare quanto ad un'azienda manchi per poter raggiungere una condizione "sana" dal punto di vista di solvibilità finanziaria.
In questo capitolo vedremo, quindi, come tali indici possano essere fonte di previsione della crisi d'impresa.

4.1 Procedimento per la costruzione degli indici di deprivazione

Nel capitolo 2 sono stati indicati due step per la costruzione degli indici di deprivazione, i quali si articolano poi in modo differente a seconda della tipologia di studio scelta.
In questo paragrafo, essi vengono approfonditi adattandoli alla modalità di svolgimento delle analisi empiriche utilizzate in questo lavoro.

1. studio della polarità;

   I dati rilevati si costituiscono di tutta una serie di variabili che, una volta rilevate, vanno polarizzate. Questo perché sarà necessario costruire un indice le cui componenti siano tutte riportate in modo tale da considerarsi relazionate direttamente con il rischio di *default*.

   Per fare ciò si individua innanzitutto l'elemento attraverso il quale studiare la relazione con il *default*, ossia le medie dei gruppi [20]. In particolare, si considera solo il vettore delle medie dei gruppi delle aziende che sono poi fallite. Essendo il settore una variabile paculiare, il calcolo della sua media segue un procedimento diverso ossia vengono calcolate come l'insieme dei rapporti tra ogni settore ed un settore di riferimento. Di seguito (figura 4.1) sono riportate le distribuzioni delle fallite (1) e delle non fallite (0) dove si nota che le variabili tendano ad avere più medie superiori allo 0 qualora si parli di aziende poi fallite e viceversa. Vengono riportate nella figura 4.1(a) le distribuzioni di tutti i dati, mentre nella figura 4.1(b) quelle riguardanti i soli dati non estremi. Nel secondo caso, si evidenzia una netta distinzione nella posizione delle due distribuzioni considerate [20].

   Calcolate le medie si procede con l'individuazione della polarità. Si applicano due modalità:

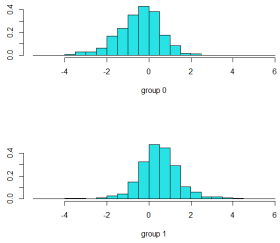
   - regressione logistica;
   - differenza tra le medie.

2. sulla base della distribuzione delle singole variabili si individuano 10 quantili;

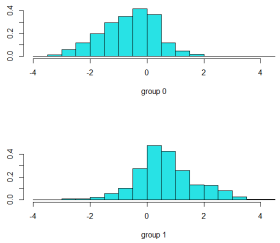
3. si considerano i valori registrati dalle aziende per ogni singola variabile e si individua in quale dei quantili determinati si inseriscono;

4. per ogni azienda si effettua una sommatoria dei quantili registrati dalle variabili;

5. si ordinano i valori delle somme in modo crescente;



(a) Distribuzione score



(b) Distribuzione score con winsorize

Figura 4.1: Grafici distribuzione score

6. si determina come percentuale discriminante il 55% e si individua quale dei valori calcolati registri tale percentuale nella distribuzione delle sommatorie;

7. si costruisce un indicatore che seleziona l'impresa come a rischio di *default* qualora superi tale valore mentre non a rischio nel caso contrario.

Di seguito vengono riportate tutte analisi svolte per raggiungere l'obiettivo finale.

4.2 Previsione con analisi su otto variabili

Dapprima si è provato a svolgere l'analisi attraverso l'utilizzo di sole 8 variabili su 19, ovvero le seguenti:

- incidenza del margine di struttura primario;
- grado di dipendenza finanziaria;
- indice di esigibilità del debito;

- indice di liquidità di primo livello;
- incidenza del margine di tesoreria di primo livello;
- incidenza del valore attuale sui ricavi;
- incidenza del reddito operativo (o MON) sui ricavi;
- incidenza dell'EBT sui ricavi.

Si chiarisce che la scelta di queste 8 variabili è stata svolta selezionando quelle che più risultassero correlate con le componenti principali ottenute nel paragrafo 3.4.

4.2.1 Studio polarità

Studio polarità con differenza tra medie

Individuate le medie dei due gruppi attraverso la funzione `colMeans()` di *RStudio*, vengono calcolate le differenze tra le medie del gruppo 0 (medie dei valori assunti dalle variabili per le aziende che poi non sono fallite) e le medie del gruppo 1 (medie dei valori assunti dalle variabili per le aziende che poi sono fallite). Qualora si registri una differenza positiva si assume quella variabile come inversamente proporzionale alla possibilità di *default*, viceversa per il caso contrario.
Si ottiene come risultato che cinque variabili sono inversamente proporzionali (incidenza del margine di struttura primario, incidenza del margine di tesoreria di primo livello, incidenza del valore attuale sui ricavi, incidenza del reddito operativo sui ricavi ed incidenza dell'EBT sui ricavi) e tre direttamente proporzionali (grado di dipendenza finanziaria, indice di esigibilità del debito e l'indice di liquidità di primo livello). Per verificare tali risultati si svolge l'analisi anche con la regressione logistica.

Studio polarità con regressione logistica

Si costruisce una funzione di regressione logistica e, sulla base del segno assunto dai coefficienti delle variabili indipendenti, si studia la loro polarità ottenendo gli

stessi risultati registrati con la differenza tra medie. Perciò, tali valori verranno poi riportati in una matrice con lo stesso segno qualora si sia giunti alla conclusione che vi sia una relazione diretta con la probabilità di *default*, mentre nel caso contrario verrano riscritti col segno opposto.

4.2.2 Costruzione indice di deprivazione

Sulla base della nuova matrice creatasi si determina, per ogni variabile, in quale dei quantili le aziende si trovano, si effettua la sommatoria e si costruisce l'indice di deprivazione che seleziona come a rischio di *default* l'azienda che registri un valore della sommatoria dei ranghi delle variabili al di sopra del 55% di questa distribuzione.

4.2.3 Analisi Risultati

Effettuando questo tipo previsione sulle aziende del campione si effettua poi il confronto con quanto osservato sulle stesse e si costruisce la tabella sottostante.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	831	417
	in crisi	301	715

Si calcolano quindi la bontà della previsione svolta attraverso tale indice di deprivazione risolvendo le formule 1.1 e 1.2. Si ottengono quindi un'*accuracy* pari al $68,28622\%$ ed una *specificity* del $73,409894\%$

$$accuracy = \frac{715 + 831}{2264} = 68,28622\%$$

$$specificity = \frac{831}{831 + 301} = 73,409894\%$$

Volendo rappresentare graficamente le performance del modello si riporta nella figura 4.2 la curva ROC spiegata nel capitolo 3, la quale registra un'area sottesa alla curva pari a 0,6829.

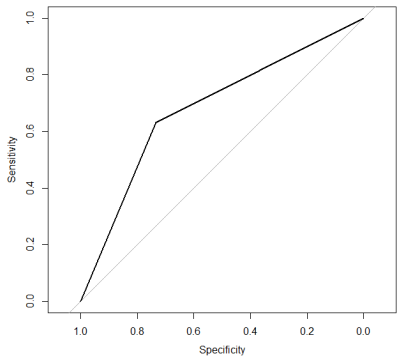


Figura 4.2: Curva ROC indice deprivazione con otto variabili

4.3 Previsione con analisi su 19 variabili

Si costruisce un indice di deprivazione seguendo gli stessi step utilizzati nel paragrafo precedente ma utilizzando tutte e 19 le variabili.

4.3.1 Indice di deprivazione con 19 variabili

In questo caso si utilizza soltanto il medoto della regressione logistica per effettuare lo stodio della polarità e, costruendo l'indice di deprivazione, si ottengono una bontà dello stesso pari al $68,41873\%$ ed una specificità del $81,095406\%$, la cui curva roc presenta un'area sottesa pari a 0,6842.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	918	501
	in crisi	214	631

$$accuracy = \frac{631 + 918}{2264} = 68,41873\%$$

$$specificity = \frac{918}{918 + 214} = 81,095406\%$$

La curva ROC raggiunge lo stesso risultato se si considera anche la variabile settore, come si può vedere dalla figura 4.3 in cui vengono riportate entrambe.

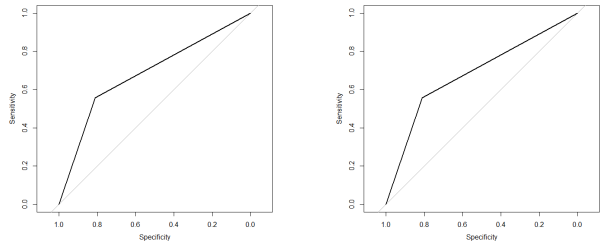


Figura 4.3: Bontà indice di deprivazione con più variabili

Tuttavia, l'*accuracy* e la *specificity* risultano essere molto ridotte rispetto al caso in cui la variabile "settore di appartenenza" non era stata considerata in quanto raggiungono rispettivamente solo il $56,89046\%$ ed il $48,939929\%$.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	554	398
	in crisi	578	734

$$accuracy = \frac{734 + 554}{2264} = 56,89046\%$$

$$specificity = \frac{554}{554 + 578} = 48,939929\%$$

4.4 Previsione con analisi sui dati ridotti degli outlier

Ripetendo le analisi svolte nei paragrafi precedenti sui dati ottenuti dall'eliminazione degli outlier si ottengono alcune inversioni delle polarità, ad esempio, se in precedenza l'indice di liquidità di primo livello veniva individuato come direttamente proporzionale alla probabilità di *default*, ora si ottiene l'esatto contrario.
Per quanto riguarda le performance del modello si registrano dei buoni risultati, tra i quali spicca il livello di specificità regsitrato nel caso di analisi di sole 8 variabili.
Ciò può essere spiegato dal fatto che alla base della costruzione dell'indice di deprivazione si trovano tutti i dati dell'anlisi, quindi l'eliminazione dei valori estremi non farebbe altro che alterarne i risultati. Di conseguenza, si ritiene migliore l'analisi

svolta sui dati del campione lasciati così come sono, senza riservare loro particolari trattamenti quale la winsorizzazione.

- Analisi di sole otto variabili

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	859	323
	in crisi	273	809

Volendo studiare le performance della previsione ottenuta dalla costruzione dell'indice di deprivazione si ottengono un'*accuracy* pari al $73,67491\%$ ed una *specificity* del $75,883392\%$.

$$accuracy = \frac{809 + 859}{2264} = 73,67491\%$$

$$specificity = \frac{859}{859 + 273} = 75,883392\%$$

La curva ROC, invece, presenta un'area sottesa alla curva di 0,7367.

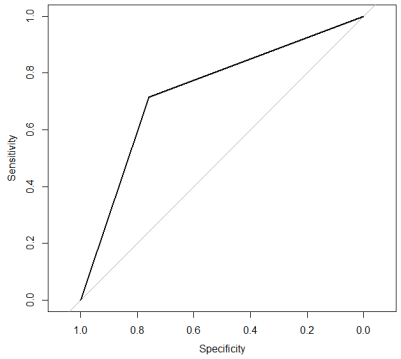


Figura 4.4: Curva roc indice di deprivazione costruito con 8 variabili e solo sui valori non estremi

- Analisi di 19 variabili

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	501	136
	in crisi	631	996

La bontà della previsione in questo caso è pari al $66,12191\%$ mentre la capacità di individuare le imprese "sane" è del $44,257951\%$.

$$accuracy = \frac{996 + 501}{2264} = 66,12191\%$$

$$specificity = \frac{501}{501 + 631} = 44,257951\%$$

La curva roc presenta in questo caso un'area sottesa alla curva pari a 0,6612.

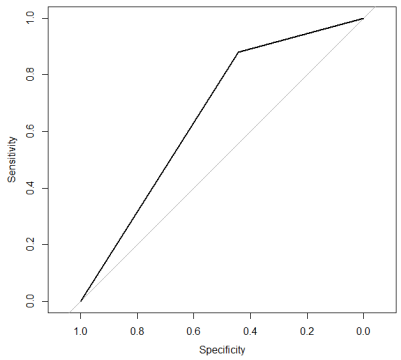


Figura 4.5: Curva roc indice di deprivazione costruito con 19 variabili e solo sui valori non estremi

- Analisi di 20 variabili

  Considerando anche la variabile settore, in questo caso si ottengono delle performance molto più buone, infatti si registrano un'*accuracy* pari al $72,87986\%$ ed una *specificity* del $68,462898\%$.

classificazione osservata			
		in salute	in crisi
classificazione prevista	in salute	775	257
	in crisi	357	875

$$accuracy = \frac{875 + 775}{2264} = 72,87986\%$$

$$specificity = \frac{775}{775 + 357} = 68,462898\%$$

Mentre la curva roc è la stessi di quella ottenuta dall'analisi che non tiene conto della variabile "settore di appartenenza".

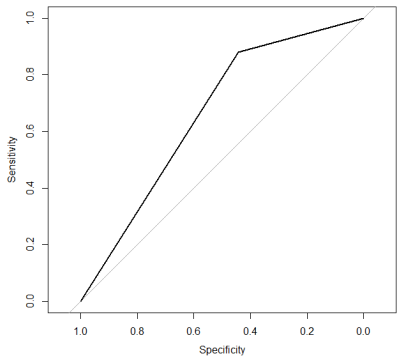


Figura 4.6: Curva roc indice di deprivazione costruito con 20 variabili e solo sui valori non estremi

4.5 Considerazioni finali

Dall'applicazione degli indici di deprivazione quale fonte di previsione per la crisi d'impresa sono stati ottenuti dei buoni risultati, dimostrando come tale nuovo approccio, solitamente ambientato in ricerche di tipo sociale e sanitario, non solo risulti altamente applicabile, ma garantisca una buona qualità dell'analisi.
In particolare, l'innovazione apportata dall'utilizzo degli indici di deprivazione la si ritrova nelle sue seguenti caratteristiche:

1. risulta essere una modalità di studio molto semplice da implementare;
2. seppure meno complessa conduce comunque a dei risultati di *accuracy* buoni;
3. raggiunge livelli di *specificity* anche più alti rispetto ad analisi più complesse quali quella di Poli [35];

4. garantisce l'ottenimento di buoni risultati anche solo utilizzando poche variabili, ad esempio 8;

5. non richiede particolari trattamenti degli outlier ma risulta soddisfacente anche solo con l'applicazione della tecnica della winsorizzazione, ovvero una delle più semplici, anzi risulta migliore se svolta sui dati del campione orginario.

Conclusione

Con l'obiettivo di individuare ulteriori modalità di studio rispetto a quelle già presenti negli attuali modelli di previsione sono state proposte le seguenti analisi statistiche: analisi delle componenti principali, analisi discriminante, regressione logistica e costruzione degli indici di deprivazione.

Per quanto riguarda l'analisi delle componenti principali si è rivelata una buona modalità di riduzione delle dimensioni da studiare e quindi di agevolazione del lavoro. In particolare, è stata applicata come esempio sulle 19 variabili in esame senza effettuare alcuna eliminazione dei valori estremi, conducendo ad una semplificazione sino a 10 variabili.

Si è invece ottenuto un risultato meno utile quando è stata applicata ai dati che erano stati sottoposti all'eliminazione degli outlier, in quanto già erano state escluse delle dimensioni di variabilità della distribuzione.

, Invece, grazie alle altre tipologie di analisi sopra dette è stato possibile costruire dei veri e propri modelli di previsione, i quali hanno raggiunto il più delle volte buoni risultati di *accuracy* e di *specificity*.

Tra i modelli di Poli [35] è stato ottenuto come miglior livello di *accuracy* il $74,31\%$. Confrontandolo con quanto ottenuto dalle analisi empiriche di questo lavoro, si evidenziano i risultati ottenuti dalla regressione logistica svolta su tutti i dati pari al $74,65\%$, dalla regressione logistica svolta sui soli dati non estremi pari al 74.91% e dall'analisi discriminante svolta sui valori non estremi considerando la variabile "settore di appartenenza" pari al $74,47\%$. Tra i quali spicca il livello di *accuracy* raggiunto dal sistema logistico applicato sui dati ridotti degli outlier. Fa pensare il

fatto che il risultato migliore sia stato ottenuto dall'applicazione di un sistema di analisi già utilizzato da uno dei modelli di Poli, del quale sono state modificate solo le variabili oggetto di studio ed il numero delle imprese oggetto del campione.
Per quanto riguarda, invece, i livelli di *specificity*, ovvero l'obiettivo principale dell'analisi svolta in questo lavoro, si hanno come valori di confronto il $74,12\%$ raggiunto dal modello 3 di Poli [35], l'$\mathbf{81,10}\%$ raggiunto dall'indice di deprivazione su tutti i dati considerando 19 variabili ed il $75,88\%$ raggiunto dall'indice di deprivazione sui dati ridotti degli outlier considerando 8 variabili.
In questo caso si nota come gli indici di deprivazione siano dei buoni metodi di analisi attraverso i quali prevedere le aziende "sane" in quanto raggiungono livelli di specificità anche pari all'$81,10\%$.
Da questo lavoro si è quindi visto che gli indici di deprivazione possano essere adattati al contesto della previsione della crisi d'impresa garantendo la costruzione di modelli con delle performance similari a quelle registrate dai modelli implementati grazie ad analisi più complesse, come ad esempio quelle di Poli [35]
In particolare, si nota dai risultati sopra esposti che, pur senza concedere particolari trattamenti agli outlier se non una semplice tecnica delle winsorizzazione, si producano delle ottime performance in termini di specificità.
Riflettendo su tali modalità di indagine e sui miglioramenti che sono stati proposti, quali l'utilizzo di indici di bilancio reperibili anche da quello abbreviato e la focalizzazione della previsione su una situazione di insolvenza futura (la crisi) piuttosto che dell'insolvenza stessa, si è voluto fornire l'incipit per lo sviluppo di analisi che possano sempre più prevenire il fallimento imprenditoriale.
Il lavoro qui svolto ha quindi lo scopo di mettere a disposizione ulteriori approcci di analisi per migliorare le attuali procedure di prevenzione del *default* a catena, da un lato, incrementando le possibilità per i soggetti esterni all'impresa di effettuare dei controlli sulla solvibilità della stessa prima di intrattenervici relazioni contrattuali, e dall'altro, allineandosi alla *ratio* degli ultimi interventi normativi.

Appendici

Appendice 1: codice di RStudio utilizzato

Codice per l'applicazione delle analisi senza lo svolgimento della winsorizzazione

$rm(list = ls())$

$library(rstudioapi)$

$current\_path < -getActiveDocumentContext()\$path$

$library(openxlsx)$

$dati = read.xlsx("Dati\_fallimenti.xlsx")$

$View(dati)$

$new\_data = data.frame(as.factor(dati[, 24]), dati[, 2 : 20], as.factor(dati[, 21]))$

$nomi = names(dati)$

$names(new\_data) = nomi[c(24, 2 : 21)]$

$View(new\_data)$

$library(DescTools)$

$plot(density(new\_data[, 2]))$

$plot(density(new\_data[, 4]))$

$library(corrplot)$

$library(psych)$

$C = cor(as.matrix(new\_data[, 2 : 20]))$

$dev.new()$

$corrplot(C, method =' number')$

$corrplot(C, method =' circle')$

$M = as.matrix(new\_data[, 2 : 20])$

$dev.new()$

$pval < -psych :: corr.test(M, adjust = "none")\$p$

$corrplot(round(pval, 1), method =' number')$

$corrplot(round(pval, 1), method = "circle")$

$library(FactoMineR)$

$library(stats)$

$library(MASS)$

$pr\_data\_new = princomp(new\_data[, 2 : 20], cor = TRUE, scores = TRUE)$

$print(summary(pr\_data\_new))$

$attributes(pr\_data\_new)$

$print((pr\_data\_new\$loadings))$

$new\_data\_scale = new\_data[, 2 : 20]$

$data\_elabora = data.frame(new\_data[, 1], new\_data\_scale)$

$names(data\_elabora) = names(new\_data[, 1 : 20])$

$lda\_new\_data = lda(STATO ., data\_elabora, prior = c(1, 1)/2)$

$y < -predict(lda\_new\_data, data\_elabora)$

$y\_oss < -data\_elabora\$STATO;$

$y\_prev < -y\$class$

$tab < -xtabs(y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$new\_data\_scale = new\_data[, 2 : 20]$

$data\_elabora = data.frame(new\_data[, 1], new\_data\_scale, new\_data[, 21])$

$names(data\_elabora) = names(new\_data)$

$lda\_new\_data = lda(STATO ., data\_elabora, prior = c(1, 1)/2)$

$y < -predict(lda\_new\_data, data\_elabora)$

$y\_oss < -data\_elabora\$STATO;$

$y\_prev < -y\$class$

$tab <- xtabs(\ y\_prev + y\_oss); tab$

$n <- nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$dev.new()$

$plot(lda\_new\_data)$

$attributes(lda\_new\_data)$

$medie\_gruppi = lda\_new\_data\$means$

$medie\_gruppi\_fall = colMeans(as.matrix(data\_elabora[data\_elabora\$STATO == 1, 2 : 20]))$

$mm = c(1, 3, 4, 5, 8, 11, 13, 15)$

$nq = 10$

$print(medie\_gruppi[, mm])$

$polarita = NULL;$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1;$

$polarita[ic] = 1;$

$((medie_gruppi[1, j] - medie\_gruppi[2, j]) > 0)\{$

$polarita[ic] = -1;$

$\}$

$\}$

$polarita\_LR = NULL;$

$peso = NULL$

$ic = 0$

$for(jinmm)\{$

$ic = ic + 1$

$aux = data.frame(data\_elabora[, 1], data\_elabora[, (j + 1)])$

$names(aux) = c('Stato', ' x')$

$out\_LR <- glm(Stato\ x, family = binomial(link = "logit"), data = aux)$

$print(out\_LR\$coefficients)$

$peso[ic] = out\_LR\$coefficients[2]$

$polarita\_LR[ic] = 1;$

$if(out\_LR\$coefficients[2] < 0)\{$

$polarita\_LR[ic] = -1;$

$\}$

$\}$

$indicatori = matrix(rep(0, length(mm) * length(data\_elabora[, 1])),$

$nrow = length(data\_elabora[, 1]), ncol = length(mm))$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1;$

$indicatori[, ic] = polarita\_LR[ic] * (data\_elabora[, j + 1])$

$\}$

$verifica = data.frame(data\_elabora[, 1], indicatori)$

$names(verifica) = names(data\_elabora[, c(1, mm)])$

$colnames(verifica)[2] = "struttura1"$

$library(dplyr)$

$my\_rank = list()$

$My\_rank\_matrix = matrix(rep(0, length(mm) * length(verifica[, 2])),$

$nrow = length(verifica[, 2]), ncol = length(mm))$

$for(jin1 : length(mm))\{$

$df1 = mutate(verifica, quantile_rank = ntile(indicatori[, j], nq))$

$my\_rank[[j]] = data.frame(df1\$quantile_rank, indicatori[, j])$

$names(my\_rank[[j]]) = c("quantile", names(verifica)[j + 1])$

$My\_rank\_matrix[, j] = df1\$quantile\_rank$

$\}$

$head\_count = NULL;$

$head\_count = rowSums(My\_rank\_matrix)$

$summary(head\_count)$

$dataframe\_final = data.frame(verifica[, 1], head\_count)$

$sortedFinal < -dataframe\_final[order(dataframe\_final[,2]),]$

$names(sortedFinal) = c('STATO','rank')$

$View(sortedFinal)$

$previsio = NULL;$

$for(iin1 : length(head\_count))\{$

$previsio[i] = 0;$

$if(head\_count[i] > (length(mm) * nq) * 0.55)previsio[i] = 1$

$\}$

$y\_oss < -dataframe\_final[,1];$

$y\_prev < -previsio$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$library(pROC)$

$roc < -roc(y\_oss\ y\_prev, plot = TRUE)$

$plot.roc(roc)$

$auc(roc)$

$mm = 1 : 19$

$soglie\_fall = medie\_gruppi[2, mm]$

$print(soglie\_fall)$

$polarita = NULL;$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1;$

$polarita[ic] = 1;$

$if((medie\_gruppi[1, j] - medie\_gruppi[2, j]) > 0)\{$

$polarita[ic] = -1;$

$\}$

$\}$

$indicatori = matrix(rep(0, length(mm) * length(data\_elabora[,1])),$

$nrow = length(data\_elabora[,1]), ncol = length(mm))$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1$

$for(iin1 : length(data\_elabora[,1]))$

$\{$

$indicatori[i, ic] = 0$

$app = polarita[ic] * (data\_elabora[i, j + 1] - 0.5 * soglie\_fall[ic])$

$if(app >= 0)indicatori[i, ic] = 1$

$\}$

$\}$

$head\_count\_1 = NULL;$

$head\_count\_1 = rowSums(indicatori)$

$dataframe\_final1 = data.frame(verifica[,1], head\_count\_1)$

$View(dataframe\_final1)$

$sortedFinal1 < -dataframe_final1[order(dataframe\_final1[,2]),]$

$names(sortedFinal1) = c('STATO','rank')$

$View(sortedFinal1)$

$previsio1 = NULL;$

$for(iin1 : length(head\_count\_1))\{$

$previsio1[i] = 0;$

$if(head\_count\_1[i] > (length(mm)) * 0.25)previsio1[i] = 1$

$\}$

$y\_oss < -dataframe_final1[,1];$

$y\_prev < -previsio1$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$library(pROC)$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

$lda\_new = lda(verifica\$STATO\ ., verifica, prior = c(1,1)/2)$

$y < -predict(lda\_new, verifica)$

$y\_oss < -verifica[,1];$

$y\_prev < -y\$class$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

$out\_LR < -glm(STATO\ ., family = binomial(link = "logit"),$

$data = data\_elabora)$

$out\_LR\_predict = predict(out\_LR, data\_elabora, type = "response")$

$View(out\_LR\_predict)$

$previsione = as.factor(ifelse(out\_LR\_predict > 0.5, "1", "0"))$

$y\_oss < -data\_elabora[,1];$

$y\_prev < -previsione$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

Codice per l'applicazione delle analisi con lo svolgimento della winsorizzazione

$rm(list = ls())$

$library(rstudioapi)$

$current\_path < -getActiveDocumentContext()\$path$

$library(openxlsx)$

$dati = read.xlsx("Dati\_fallimenti.xlsx")$

$View(dati)$

$new\_data = data.frame(as.factor(dati[, 24]), dati[, 2 : 20], as.factor(dati[, 21]))$

$nomi = names(dati)$

$names(new\_data) = nomi[c(24, 2 : 21)]$

$View(new\_data)$

$library(DescTools)$

$new\_data[, 2 : 20] < -lapply(new\_data[, -c(1, 21)], Winsorize)$

$plot(density(new\_data[, 2]))$

$plot(density(new\_data[, 4]))$

$library(corrplot)$

$library(psych)$

$C = cor(as.matrix(new\_data[, 2 : 20]))$

$dev.new()$

$corrplot(C, method =' number')$

$corrplot(C, method =' circle')$

$M = as.matrix(new\_data[, 2 : 20])$

$dev.new()$

$pval < -psych :: corr.test(M, adjust = "none")\$p$

$corrplot(round(pval, 1), method =' number')$

$corrplot(round(pval, 1), method = "circle")$

$library(FactoMineR)$

$library(stats)$

$library(MASS)$

$pr\_data\_new = princomp(new\_data[, 2:20], cor = TRUE, scores = TRUE)$

$print(summary(pr\_data\_new))$

$attributes(pr\_data\_new)$

$print((pr\_data\_new\$loadings))$

$new\_data\_scale = new\_data[, 2:20]$

$data\_elabora = data.frame(new\_data[, 1], new\_data\_scale)$

$names(data\_elabora) = names(new\_data[, 1:20])$

$lda\_new\_data = lda(STATO\ ., data\_elabora, prior = c(1, 1)/2)$

$y < -predict(lda\_new\_data, data\_elabora)$

$y\_oss < -data\_elabora\$STATO;$

$y\_prev < -y\$class$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$new\_data\_scale = new\_data[, 2:20]$

$data\_elabora = data.frame(new\_data[, 1], new\_data\_scale, new\_data[, 21])$

$names(data\_elabora) = names(new\_data)$

$lda\_new\_data = lda(STATO\ ., data\_elabora, prior = c(1, 1)/2)$

$y < -predict(lda\_new\_data, data\_elabora)$

$y\_oss < -data\_elabora\$STATO;$

$y\_prev < -y\$class$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$dev.new()$

$plot(lda\_new\_data)$

$attributes(lda\_new\_data)$

$medie\_gruppi = lda\_new\_data\$means$

$medie\_gruppi\_fall = colMeans(as.matrix(data\_elabora[data\_elabora\$STATO ==$

$1, 2 : 20]))$

$mm = c(1, 3, 4, 5, 8, 11, 13, 15)$

$nq = 10$

$print(medie\_gruppi[, mm])$

$polarita = NULL;$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1;$

$polarita[ic] = 1;$

$((medie_gruppi[1, j] - medie\_gruppi[2, j]) > 0)\{$

$polarita[ic] = -1;$

$\}$

$\}$

$polarita\_LR = NULL;$

$peso = NULL$

$ic = 0$

$for(jinmm)\{$

$ic = ic + 1$

$aux = data.frame(data\_elabora[, 1], data\_elabora[, (j + 1)])$

$names(aux) = c('Stato', ' x')$

$out\_LR < -glm(Stato\ x, family = binomial(link = "logit"), data = aux)$

$print(out\_LR\$coefficients)$

$peso[ic] = out\_LR\$coefficients[2]$

$polarita\_LR[ic] = 1;$

$if(out\_LR\$coefficients[2] < 0)\{$

$polarita\_LR[ic] = -1;$

$\}$

$\}$

$indicatori = matrix(rep(0, length(mm) * length(data\_elabora[, 1])),$

$nrow = length(data\_elabora[, 1]), ncol = length(mm))$

$ic = 0;$

$for(jinmm)\{$

$ic = ic + 1;$

$indicatori[, ic] = polarita\_LR[ic] * (data\_elabora[, j + 1])$

$\}$

$verifica = data.frame(data\_elabora[, 1], indicatori)$

$names(verifica) = names(data\_elabora[, c(1, mm)])$

$colnames(verifica)[2] = "struttura1"$

$library(dplyr)$

$my\_rank = list()$

$My\_rank\_matrix = matrix(rep(0, length(mm) * length(verifica[, 2])),$

$nrow = length(verifica[, 2]), ncol = length(mm))$

$for(jin1 : length(mm))\{$

$df1 = mutate(verifica, quantile_rank = ntile(indicatori[, j], nq))$

$my\_rank[[j]] = data.frame(df1\$quantile_rank, indicatori[, j])$

$names(my\_rank[[j]]) = c("quantile", names(verifica)[j + 1])$

$My\_rank\_matrix[, j] = df1\$quantile\_rank$

$\}$

$head\_count = NULL;$

$head\_count = rowSums(My\_rank\_matrix)$

$summary(head\_count)$

$dataframe\_final = data.frame(verifica[, 1], head\_count)$

$sortedFinal < -dataframe\_final[order(dataframe\_final[, 2]),]$

$names(sortedFinal) = c('STATO', 'rank')$

$View(sortedFinal)$

$previsio = NULL;$

$for(iin1 : length(head\_count))\{$

$previsio[i] = 0;$

$if(head\_count[i] > (length(mm) * nq) * 0.55)previsio[i] = 1$

```
    }
y_oss <-dataframe_final[,1];
y_prev <-previsio
tab <-xtabs( y_prev + y_oss);tab
n <-nrow(data_elabora)
(sum(diag(tab))/n) * 100
library(pROC)
roc <-roc(y_oss y_prev,plot = TRUE)
plot.roc(roc)
auc(roc)
mm = 1 : 19
soglie_fall = medie_gruppi[2,mm]
print(soglie_fall)
polarita = NULL;
ic = 0;
for(jinmm){
    ic = ic + 1;
    polarita[ic] = 1;
    if((medie_gruppi[1,j] - medie_gruppi[2,j]) > 0){
    polarita[ic] = -1;
    }
    }
indicatori = matrix(rep(0,length(mm) * length(data_elabora[,1])),
nrow = length(data_elabora[,1]),ncol = length(mm))
ic = 0;
for(jinmm){
    ic = ic + 1
    for(iin1 : length(data_elabora[,1]))
    {
    indicatori[i,ic] = 0
```

$app = polarita[ic] * (data\_elabora[i, j+1] - 0.5 * soglie\_fall[ic])$

$if(app >= 0) indicatori[i, ic] = 1$

}

}

$head\_count\_1 = NULL;$

$head\_count\_1 = rowSums(indicatori)$

$dataframe\_final1 = data.frame(verifica[, 1], head\_count\_1)$

$View(dataframe\_final1)$

$sortedFinal1 < -dataframe_final1[order(dataframe\_final1[, 2]),]$

$names(sortedFinal1) = c('STATO', 'rank')$

$View(sortedFinal1)$

$previsio1 = NULL;$

$for(iin1 : length(head\_count\_1)){$

$previsio1[i] = 0;$

$if(head\_count\_1[i] > (length(mm)) * 0.25) previsio1[i] = 1$

}

$y\_oss < -dataframe_final1[, 1];$

$y\_prev < -previsio1$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$library(pROC)$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

$lda\_new = lda(verifica\$STATO\ ., verifica, prior = c(1, 1)/2)$

$y < -predict(lda\_new, verifica)$

$y\_oss < -verifica[, 1];$

$y\_prev < -y\$class$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

$out\_LR < -glm(STATO\ ., family = binomial(link = "logit"),$

$data = data\_elabora)$

$out\_LR\_predict = predict(out\_LR, data\_elabora, type = "response")$

$View(out\_LR\_predict)$

$previsione = as.factor(ifelse(out\_LR\_predict > 0.5, "1", "0"))$

$y\_oss < -data\_elabora[, 1];$

$y\_prev < -previsione$

$tab < -xtabs(\ y\_prev + y\_oss); tab$

$n < -nrow(data\_elabora)$

$(sum(diag(tab))/n) * 100$

$roc < -roc(y\_oss\ y\_prev)$

$plot(roc)$

$auc(roc)$

Bibliografia

[1] Analisi delle componenti principali. http://oldwww.unibas.it/utenti/dinardo/pca.pdf, 2006.

[2] Gli indicatori statistici: concetti, metodi e applicazioni. https://flore.unifi.it/retrieve/handle/2158/328151/7940/ASTRIS_6_-_Gli_Indicatori_Statistici_-_concetti%2c_metodi_e_applicazioni.pdf, 2006.

[3] Dispense del corso di sistemi informativi statistici. http://homes.stat.unipd.it/mariobolzan/sites/homes.stat.unipd.it.mariobolzan/files/MaterialeBoccuzzoIndicatori2013_14pdf.pdf, 2010.

[4] Analisi discriminante. http://www.cs.unitn.it/~taufer/Slide-pdf/4d%20AD.pdf, 2015.

[5] Il processo di costruzione degli indicatori compositi di bes 2015. https://www.istat.it/it/files/2016/03/06-Tinto-Pres.pdf, 2016.

[6] Regressione logistica. http://www.cs.unitn.it/~taufer/Handout-pdf/4a_RLg.pdf, 2017.

[7] Analisi discriminante. https://docenti-deps.unisi.it/stefanianaddeo/wp-content/uploads/sites/35/2019/05/7.-Analisi-discriminante-SAS.pdf, 2019.

[8] Coefficiente di correlazione. https://www.jmp.com/it_it/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html/, 2020.

[9] Differenza tra regressione lineare e logistica. https://it.differencevs.com/6857994-difference-between-linear-and-logistic-regression, 2021.

[10] Introduzione alla linear discriminant analysis (lda). https://www.lorenzogovoni.com/linear-discriminant-analysis-lda/, Febbraio 2020.

[11] Edward I. Altman. Financial ratios discriminant analysis and the prediction of corporate brankruptcy. *The Journal of Finance*, 1968.

[12] Fabrizio Laurini Andrea Cerioli. *Il modello di regressione logistica per le applicazioni aziendali*. Uni.Nova, 2019.

[13] Vic Barnett. The study of outliers: Purpose and model. *Journal of the Royal Statistical Society*, 1978.

[14] Paolo Bastia. *Crisi aziendali e piani di risanamento*. Giappichelli Editore, Torino, 2019.

[15] Francesco Ciampi. Corporate governance characteristics and default prediction modeling for small enterprise: an empirical analysis of italian firms. *Journal of Business Research*, 2015.

[16] F. Delvecchio. *Scale di misura e indicatori sociali*. Cacucci Ed., Bari, 1995.

[17] W. Erwin Diewert. Axiomatic and economic approaches to elementary price indexes. *Journal of Economic Literature*, 1995.

[18] M. Peck Edward I. Altman, J. Hartzell. *Emerging markets corporate bonds: a scoring system*. Salomon Brothers Inc., New York, 1995.

[19] Alberto Mazzoleni Elisa Giacosa. *I modelli di previsione dell'insolvenza aziendale. Efficacia predittiva, limiti e prospettive di utilizzo*. Giappichelli Editore, Torino, 2018.

[20] N. Caranci G. Costa, C. Cislaghi. *Le disuguaglianze sociali di salute. Problemi di definizione e di misura.* FrancoAngeli Editore, 2009.

[21] Trevor Hastie Robert Tibshirani Gareth James, Daniela Witten. *An Introduction to Statistical Learning with Applications in R.* Springer, New York, 2013.

[22] Nikolai Genov. *Advances in Sociological Knowledge.* VS Verlag für Sozialwissenschaften, 2004.

[23] Hamit Abderamane Paul Sardini Moumtaz Razack Hamza B. Mahamat, Mathieu Le Coz. Hydrochemical and isotopic characteristics of the basement aquifer in the wadi fira area, eastern chad. *Journal of Water Resource and Protection*, 2017.

[24] Robert V. Horn. *Statistical Indicators.* Cambridge University Press, 1993.

[25] Istat. Gli indici di deprivazione per l'analisi delle disuguaglianze tra i comuni della sardegna. *Censimento popolazione e abitazioni*, 2001.

[26] J. Bibby Kanti Mardia, J. Kent. *Multivariate Analysis.* Academic Press, 1979.

[27] A. Lubischew. *On the use of discriminant functions in taxonomy.* Biometrics, 1962.

[28] S. Schifini M. Strassoldo, E. Mattioli. *Teoria dei numeri indici dei prezzi e degli indicatori economici, finanziari e sociali.* CEDAM, Padova, 1996.

[29] Filomena Maggino. *L'analisi dei dati nell'indagine statistica.* Firenze University Press, Firenze, 2005.

[30] Adriano Pareto Matteo Mazziotta. Methods for constructing composite indices: on for all or all for one? *Rivista Italiana di Economia Demografia e Statistica*, 2013.

[31] Adriano Pareto Matteo Mazziotta. *Measuring Well-Being Over Time: The Adjusted Mazziotta–Pareto Index Versus Other Non-compensatory Indices.* Springer, Rome, 2017.

[32] Thorne Betty Newbold Paul, Carlson William. *Statistica.* Pearson, Milano, 2010.

[33] J. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 1980.

[34] Guido Paolucci. *Analisi di bilancio.* Franco Angeli, Milano, 2016.

[35] Simone Poli. *I modelli di previsione della crisi d'impresa.* Giappichelli editore, Torino, 2020.

[36] M. Reudenberg. *Composite Indicators of Country Performance: A Critical Assessment.* STI Working Paper, Paris, 2003.

[37] Bruno Ricci. *Gli indicatori di crisi e di insolvenza.* Eclettica, 2020.

[38] John A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 1988.

[39] Fagan TJ. Nomogram for bayes theorem. *The New England Journal of Medicine*, 1975.

[40] Léopold Simar Wolfgang Karl Härdle. *Applied Multivariate Statistical Analysis.* Springer Verlag, 2015.

[41] M. E. Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 1984.