UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

DIPARTIMENTO DI SCIENZE DELLA VITA

Corso di Laurea Magistrale in Controllo e Sicurezza degli Alimenti

Studio preliminare dell'utilizzo di tecniche basate sull'imaging iperspettrale nel vicino infrarosso per il monitoraggio in campo di *Halyomorpha halys*

Relatore

Prof. Alessandro Ulrici

Correlatori

Dott.ssa Rosalba Calvini

Prof.ssa Lara Maistrello

Laureando

Veronica Ferrari

ANNO ACCADEMICO 2020/2021

INDICE

Scopo della tesi

Negli ultimi decenni, l'incremento delle attività antropiche e i cambiamenti climatici hanno determinato una riduzione della biodiversità nei differenti habitat. Tra le principali cause della perdita di biodiversità vi è l'introduzione di specie aliene invasive, le quali possono generare anche danni economici in quanto possono compromettere la produzione agroalimentare. Una delle specie infestanti ad elevato potenziale invasivo nel mondo, soprattutto nel nord Italia, è cimice asiatica (*Halyomorpha halys*), agente di danno di numerose piante da frutto, ortive e ornamentali.

Il monitoraggio dello stato delle coltivazioni rappresenta un punto cruciale nella gestione sostenibile del sistema agroalimentare, poiché consente l'acquisizione di informazioni riguardanti la presenza di specie infestanti al fine di adottare azioni tempestive contenendo l'infestazione ed i conseguenti danni alle coltivazioni. Generalmente, il monitoraggio viene effettuato tramite ispezione diretta in campo da parte di personale specializzato, risultando dispendioso in termini economici e di tempo per gli agricoltori.

Per favorire un'identificazione più efficiente delle specie infestanti in campo, è possibile implementare sistemi di monitoraggio automatizzati basati sull'utilizzo di telecamere spettrali. Tali strumenti consentono un monitoraggio rapido, efficiente e in tempo reale.

A causa della sua colorazione scura, *H. halys* è difficilmente distinguibile da sfondi come rami o corteccia utilizzando telecamere che lavorano nella zona del visibile, incluse le telecamere RGB. Per superare il problema dovuto al mimetismo della cimice asiatica, nel presente lavoro di tesi è stata valutata la possibilità di utilizzare tecniche basate sull'imaging iperspettrale nel vicino infrarosso per il monitoraggio in campo di *H. halys*.

A tale scopo, sono state acquisite immagini iperspettrali nel vicino infrarosso (900-1700 nm) di campioni di cimice asiatica su diverse tipologie di sfondi vegetali, tra cui corteccia, erba, foglie gialle, foglie secche, foglie verdi, rami, terreno e uno sfondo misto dato dall'unione delle diverse matrici vegetali. Questi sfondi sono stati scelti al fine di simulare l'applicazione in campo.

Sulle immagini iperspettrali è stata effettuata una prima analisi esplorativa mediante *Principal Component Analysis* (PCA): tale passaggio ha consentito di identificare per ogni immagine i pixel relativi ad *H. halys* e ai diversi sfondi. A partire da questa selezione è stato creato un dataset di spettri di riferimento per *H. halys* e gli sfondi vegetali.

Successivamente, il dataset ottenuto è stato utilizzato per lo sviluppo di modelli di classificazione atti a discriminare la cimice asiatica dai diversi sfondi. Per la classificazione è stato utilizzato l'algoritmo

Partial Least Squares Discriminant Analysis (PLS-DA) e la sua variante Soft PLS-DA, sviluppata dal gruppo di ricerca nel quale è stato svolto il presente lavoro. L'algoritmo Soft PLS-DA offre il vantaggio di limitare la presenza di falsi positivi in applicazioni pratiche, rendendo i modelli di classificazione più flessibili. Inoltre, Soft PLS-DA è stato accoppiato a metodi *sparse* di selezione delle variabili per identificare le variabili spettrali maggiormente informative per la classificazione.

La validazione dei modelli ottenuti è stata effettuata sia tramite l'applicazione dei modelli al test set esterno, sia tramite la visualizzazione delle immagini in predizione, ottenute applicando i modelli di classificazione alle immagini iperspettrali acquisite in fase sperimentale.

I risultati ottenuti da questo studio preliminare potranno essere il punto di partenza per la successiva implementazione di un sistema di imaging multispettrale basato sulle lunghezze d'onda selezionate a partire dai dati iperspettrali.

1 La cimice asiatica (*Halyomorpha halys*)

Negli ultimi decenni, gli equilibri degli ecosistemi presenti sulla Terra sono stati compromessi dall'intensificarsi delle attività antropiche, causando una riduzione della biodiversità nei differenti habitat. Tra le principali cause della perdita di biodiversità vi è l'introduzione di specie aliene invasive in nuovi ambienti, le quali possono influenzare negativamente il biota nativo attraverso predazione, parassitismo, sfruttamento delle risorse e diffusione di malattie. Infatti, a partire dalla seconda metà del ventesimo secolo, l'intensificarsi delle attività antropiche e l'avvento della globalizzazione, con il conseguente aumento di traffici commerciali, viaggi e trasporti, hanno causato un progressivo rimescolamento della componente biotica globale e l'introduzione di un alto numero di specie in nuove regioni favorendo, di conseguenza, una maggior frequenza di invasioni da parte di specie aliene.

Dato il potenziale invasivo dovuto alla presenza nella quasi totalità degli ambienti, sia naturali che antropici, numerose specie della classe insetti possono essere causa, diretta o indiretta, di ingenti danni economici sia in ambito agricolo che sanitario.

La cimice asiatica è una delle specie invasive emergenti di maggior interesse nel mondo, a causa della rapidissima espansione nel continente americano e in quello europeo e degli ingenti danni causati alle coltivazioni ortofrutticole e ornamentali.

1.1 Caratteristiche tassonomiche

Halyomorpha halys (Stål, 1855) (Heteroptera: Pentatomidae: Pentatominae), comunemente detta cimice asiatica o cimice marmorizzata grigio-marrone, è una specie originaria dell'Asia orientale (Figura 1.1).

Figura 1.1 *Esemplare adulto di cimice asiatica (https://www.homestratosphere.com).*

Gli esemplari adulti di cimice asiatica presentano dimensioni di 12-17 mm di lunghezza e 7-10 mm di larghezza. In qualità di emittero eterottero, *H. halys* possiede due paia di ali diverse: quelle anteriori sono trasformate in emielitre, le quali si adagiano sull'addome in fase di riposo, mentre quelle posteriori sono membranose. Il tegumento risulta sclerificato e marmorizzato con piccole macchie di colore marrone-grigio, come suggerisce il nome volgare. Il capo possiede un apparato boccale perforante-succhiante (*rostro*), caratterizzato da mandibole e mascelle a stiletto contenute nel labbro inferiore scanalato. Questa tipologia di apparato boccale è tipica degli insetti fitomizi, la quale consente di perforare il tessuto vegetale e succhiarne linfa e/o parenchima cellulare. L'apparato boccale possiede due paia di ghiandole salivari, principali e accessorie, le quali sono localizzate nel torace, nei pressi dell'intestino. Tali modalità di alimentazione possono causare ingenti danni alle colture ortofrutticole, come verrà illustrato nella Sezione 1.2.

La forma del corpo è tipica della famiglia Pentatomidae, caratterizzata da primo segmento del torace di forma sub triangolare, il capo piccolo e rettangolare, le antenne lunghe di colore grigiastro con bande più chiare formate da pochi articoli, occhi piccoli e sporgenti, mentre le zampe sono lunghe e sottili, adatte alla deambulazione.

La colorazione di *H. halys* è peculiare e, pertanto, fondamentale per il suo riconoscimento. Il capo tende all'ocra e presenta dei puntini neri laterali, il *labium* è giallo con gli ultimi segmenti neri, il *corium* è ocra e può presentare delle macchie rosse mentre pronoto e scutello richiamano le stesse tonalità della testa. Le membrane alari presentano dei marchi longitudinali neri sulle venature. I segmenti delle antenne presentano bande bianche alternate a bande marroni sugli ultimi due segmenti delle antenne, caratteristica distintiva della sottofamiglia. Il ventre è giallo, mentre le zampe sono color ocra, più scure della zona prossimale al torace. La morfologia descritta può essere visualizzata nella Figura 1.2.

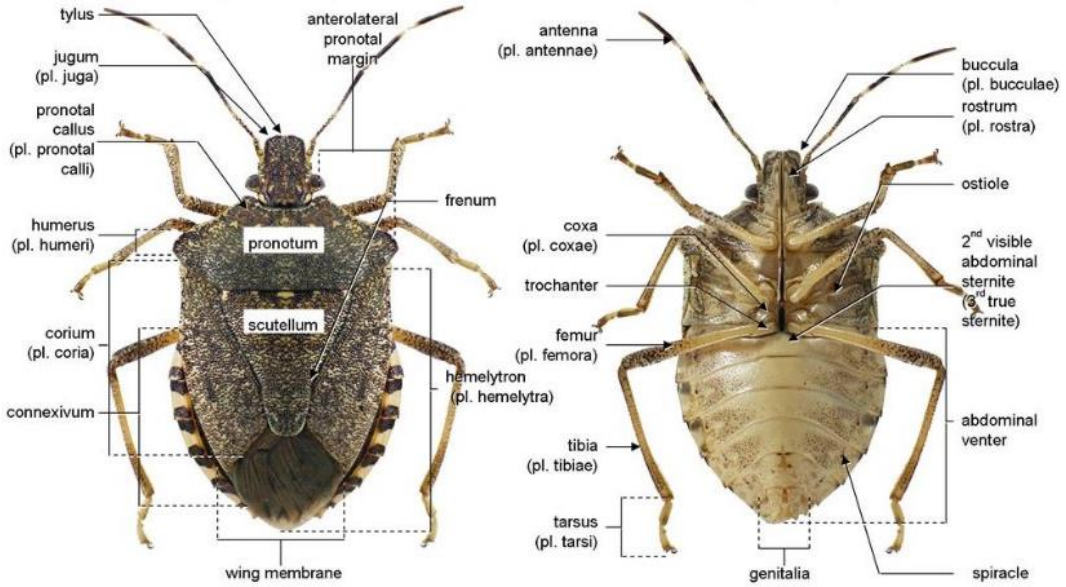

Figura 1.2 *Morfologia di un esemplare adulto di cimice asiatica (Halyomorpha halys) (https://mansanma.com/project/dichotomous-key/)*

La morfologia risulta cruciale per il riconoscimento di *H. halys* poiché difficilmente distinguibile da altre specie autoctone non invasive o non particolarmente dannose per l'agricoltura. Difatti, in Europa *H. halys* può essere confusa in particolare con *Rhaphigaster nebulosa*, specie appartenente alla stessa sottofamiglia (Pentatominae), a causa delle caratteristiche morfologiche comuni.

Nello specifico, *H. halys* e *Rhaphigaster nebulosa* sono molto simili per dimensioni e colorazione oltre che per la presenza di una spina a livello del secondo sternite, caratteristica della sottofamiglia Pentatominae. Nonostante ciò, *H. halys* può distinguersi grazie all'assenza della spina addominale che volge verso la parte anteriore, alla presenza di piccole striature brune punteggiate nell'estremità distale membranosa delle emielitre, per la morfologia del capo, allungato e triangolare, per la presenza di calli biancastri posti nella parte discale del pronoto e alla base dello scutello, per la presenza di pigmentazioni puntiformi sugli sterniti ventrali, per la presenza di tarsi delle zampe posteriori di color bianco-avorio e per la presenza di anelli gialli sul quarto segmento antennale nelle regioni basale e apicale.

1.2 Distribuzione ed impatto ecologico

H. halys è un esemplare di cimice originaria dell'Asia orientale (Cina, Giappone, Corea e Taiwan). Essa è stata intercettata per la prima volta al di fuori dell'habitat originario negli anni ′90 del secolo scorso in America settentrionale (Stati Uniti e Canada). In Europa è stata rilevata per la prima volta nel 2004 a Zurigo (Svizzera), mentre le prime segnalazioni in Italia risalgono al 2012 (Modena). Attualmente è presente in tutte le regioni italiane, isole incluse.

Ad oggi, si stima che *H. halys* abbia colonizzato circa 40 stati (Figura 1.3) e che, soltanto nel 2019, nelle aree frutticole del Nord Italia *H. halys* abbia causato un danno economico di circa 590 milioni di euro.

In ambiente urbano, *H. halys* è presente frequentemente sul verde pubblico e privato; tuttavia, data la sua tendenza ad aggregarsi per svernare in luoghi asciutti e riparati, non è raro ritrovare numerosi individui all'interno di edifici durante la stagione autunnale. Ciò genera diversi disagi legati all'elevato numero di individui presenti all'interno della stessa struttura e allo sgradevole odore che le cimici emettono se minacciate; in questi casi ricoprono il ruolo di *aesthetic pest*, risultando particolarmente fastidiosi. Inoltre, vi è il sospetto che cimice asiatica abbia un impatto sulla salute umana: sembra che la sostanza maleodorante prodotta dalla cimice sia in grado di indurre reazioni allergiche nell'uomo causando reazioni respiratorie e dermatiti da contatto.

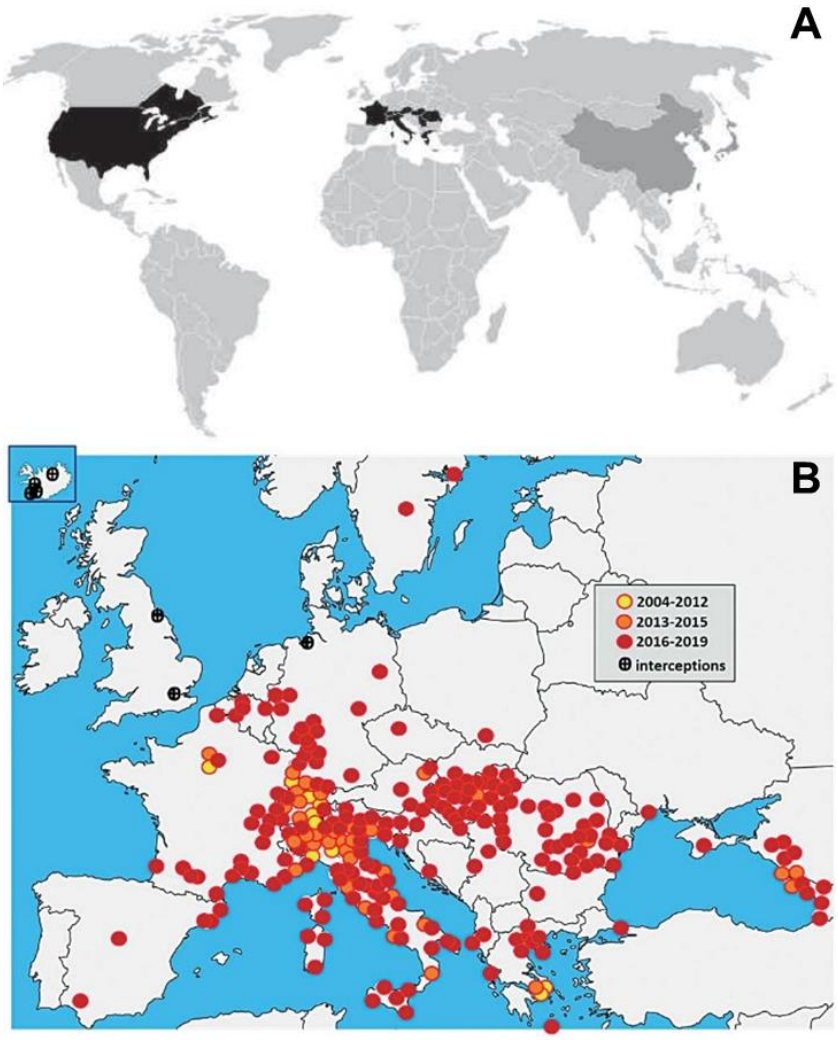

Figura 1.3 *Areale di distribuzione di Halyomorpha halys nel mondo (in grigio scuro gli stati di origine, in nero gli stati di nuova introduzione (A)e diffusione in Europa (B) (Piemontese et al., 2016, Maistrello, 2019).*

Dal punto di vista agrario, si stima che *H. halys* si nutra di un'ampia varietà di piante coltivate e spontanee (circa 300 attualmente conosciute). Seppur prediliga rosacee e fabacee, l'alimentazione della cimice asiatica comprende la maggior parte delle piante da frutto (melo, pero, pesco, albicocco, pruno, vite, nocciolo, etc.), numerosi ortaggi, leguminose, cereali come mais e sorgo ma anche diverse piante ornamentali. Per questi motivi è una specie altamente pericolosa per il potenziale danno che può determinare nei confronti di numerose piante coltivate. Sia gli esemplari adulti che gli esemplari allo stadio giovanile preferiscono nutrirsi di frutti e semi tramite punture di suzione; nonostante ciò, possono attaccare anche le altre parti epigee delle piante ospiti, provocando danni quali caduta precoce dei frutti, deformazioni e colorazioni anomale, determinando uno sviluppo stentato delle piante.

La pericolosità di questa specie è strettamente correlata alle condizioni climatiche: è stato dimostrato che nella Pianura Padana, così come nelle regioni medio-atlantiche degli Stati Uniti, *H. halys* è bivoltina ossia è presente con due generazioni all'anno. Pertanto, in queste regioni, al contrario di quanto avviene nei restanti stati

europei in cui è monovoltina, *H. halys* tende a causare maggiori danni ai raccolti. Per questo motivo, risulta fondamentale non sottovalutare l'impatto del riscaldamento globale, il quale può influenzare direttamente la capacità delle specie di ampliare il proprio areale e di stabilirsi in nuovi habitat. In particolare, l'aumento delle temperature è un fattore determinante per organismi ectotermi come gli insetti: anche un solo grado di differenza può facilitare sopravvivenza e riproduttività nelle stagioni favorevoli, così come l'aumento della resistenza alla stagione invernale.

L'incidenza e la gravità dei danni provocati da *H. halys* sulle coltivazioni variano in base alla specie colpita e alla fase fenologica in cui avviene l'attacco. Nel caso delle piante da frutto, i danni causati da *H. halys* coinvolgono l'aspetto esteriore rendendo il frutto sgradevole alla vista e non commerciabile mentre, in altri casi, il frutto subisce danni interni che ne influenzano negativamente lo sviluppo e la maturazione.

I danni vengono provocati a causa delle modalità con cui *H. halys* si alimenta (Figura 1.4): l'apparato boccale perforante-succhiante, provvisto di rostro, prevede l'inserimento degli stiletti in germogli, foglie, fiori, frutti o semi al fine di assorbire il contenuto delle cellule vegetali, ossia la linfa o parenchima cellulare. Inoltre, tale modalità di alimentazione prevede l'iniezione di saliva nel frutto la quale, essendo tossica per i tessuti vegetali, determina danni profondi e scolorimenti. I frutti colpiti possono manifestare abscissione anticipata o, se permangono sulla pianta, necrosi, suberificazioni, depressioni del tessuto vegetale, deformazioni, scolorimenti, consistenza spugnosa dei tessuti e calo di peso.

Sebbene i picchi dell'infestazione di *H. halys* si concentrino nel periodo di maturazione dei frutti, la presenza del fitofago può essere osservata a basse densità anche in altre fasi fenologiche della pianta con potenziali ripercussioni sulla produzione. In generale, le infestazioni precoci di cimice asiatica a carico dei frutteti possono causare cascola o deformazioni dei frutti mentre le infestazioni tardive sono associate a suberificazioni, necrosi e, nei casi peggiori, portano ad una deliquescenza della polpa. Oltre ai danni diretti, vi è la possibilità che le infestazioni di *H. halys* favoriscano la trasmissione di fitoplasmi e lo sviluppo di infezioni batteriche o fungine sui frutti. Inoltre, la frutta danneggiata dalla puntura di *H. halys* può avere proprietà organolettiche alterate, con conseguente compromissione della qualità della materia prima intaccata e diminuzione del valore commerciale del prodotto.

La gravità degli attacchi di *H. halys* è dovuta principalmente al suo comportamento, il quale rende difficile il controllo tramite insetticidi. Infatti, la cimice asiatica è una specie polifaga che tende a spostarsi da una pianta all'altra, anche in funzione delle esigenze dello stadio del ciclo vitale. Le piante ospiti possono essere suddivise in ospiti riproduttivi, dove si osservano tutte le fasi del ciclo vitale (dall'uovo all'esemplare adulto), e in ospiti adibiti a meri scopi alimentari, dove gli esemplari adulti possono alimentarsi. Inoltre, per effetto dei feromoni di aggregazione caratteristici della specie, gli esemplari tendono a concentrarsi in determinate zone dei frutteti. Il comportamento di diffusione è ulteriormente facilitato dalla capacità di diffusione come autostoppista, mediata e favorita dai mezzi di trasporto e dall'attività dell'uomo.

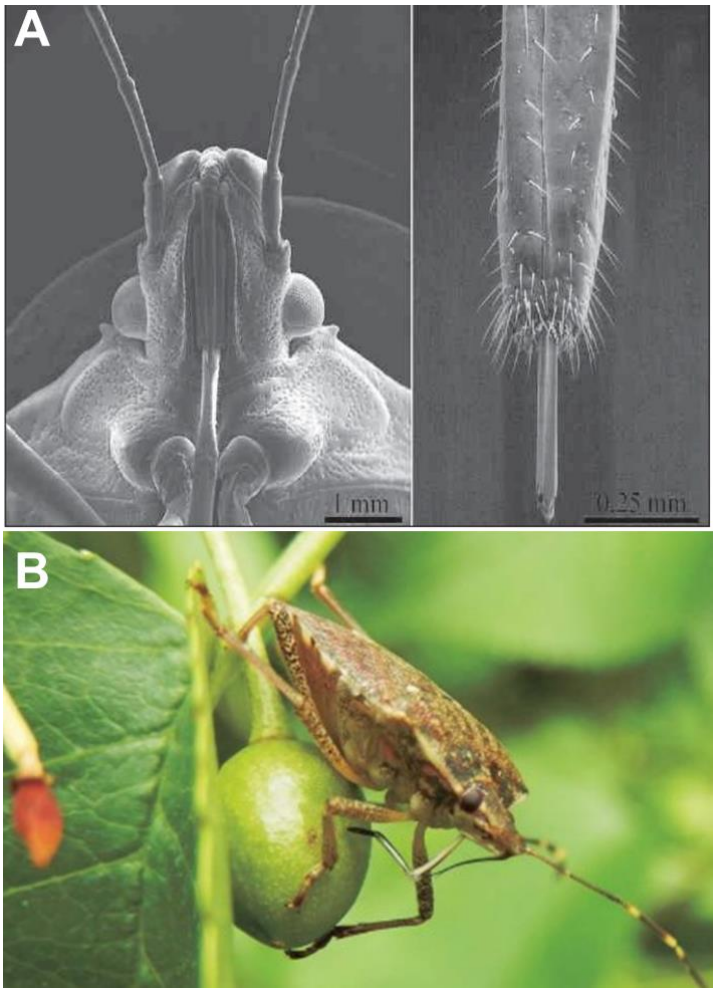

Figura 1.4 *Apparato boccale perforante-succhiante di H. halys (A) ed esemplare adulto mentre si alimenta su un frutto (B) (Piemontese et al., 2016; Maistrello, 2019).*

L'elevata mobilità, la polifagia, l'assenza di efficaci fattori di controllo delle popolazioni associati a una rilevante dannosità fanno della cimice asiatica un fitofago invasivo di rilievo a livello mondiale. Le difficoltà nel prevedere l'andamento delle popolazioni di *H. halys*, che può variare a seconda degli ambienti e delle stagioni, rendono difficoltoso il controllo di questo fitofago. A causa di questi fattori risulta complicato indirizzare efficacemente i trattamenti insetticidi: tuttora non è stata sviluppata una strategia di controllo affidabile e duratura.

Tra le produzioni più colpite nella Pianura Padana vi sono diverse cultivar di pero, specialmente quelle a maturazione precoce: secondo una stima, in alcuni ambiti il danno alla raccolta ha superato il 50%. Come anticipato, i danni riportati alle pere possono distinguersi in base alla maturità del frutto colpito: le pere intaccate nel mese di luglio presentano punture e depressioni dei tessuti vegetali superficiali mentre quelle intaccate nel mese di agosto presentano danni tissutali interni. In entrambi i casi, le malformazioni sono localizzate e caratterizzate da colore biancastro (Figura 1.5).

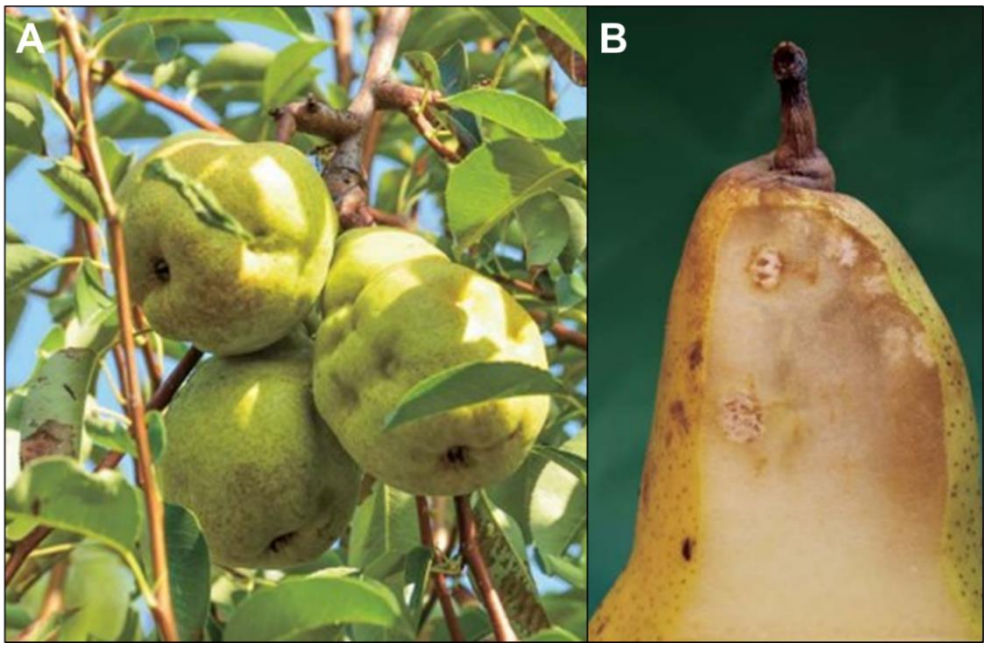

Figura 1.5 *Danni da H. halys su pero causati da punture precoci (A) e punture tardive (B) (Haye, 2019; Maistrello, 2019).*

I danni riscontrati su pero sono simili a quelli riscontrati su kiwi (Figura 1.6), produzione per la quale è stato stimato un possibile incremento dei danni a causa dell'attività di *H. halys*. Secondo le stime, l'Italia è il secondo produttore e fornitore di kiwi al mondo, con una produzione complessiva di 555 mila tonnellate nell'anno 2018; pertanto, *H. halys* ha il potenziale di causare ingenti danni economici.

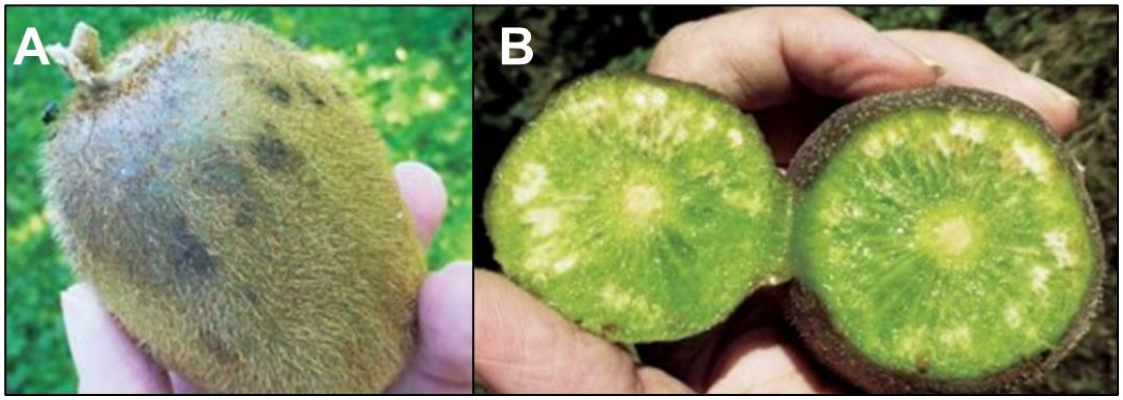

Figura 1.6 *Danni dei tessuti esterni (A) ed interni (B) causati da H. halys su kiwi (Scaccini et al., 2019).*

Un'altra tipologia di specie vegetale colpita dall'attività della cimice asiatica è il melo, anch'esso con danni diretti sul frutto. I danni osservati sulle mele sono caratterizzati da colore scuro, tendente al marrone, caratteristica che rende difficoltosa la classificazione del danno poiché può essere confusa con una patologia legata alla mancanza di calcio. In questi casi, soltanto un'analisi interna del frutto può chiarire l'origine del

danno: se il danno presente è dovuto all'attività della cimice asiatica esso si presenterà circoscritto e con la presenza di punture di suzione; viceversa, se dovuto alla patologia, risulterà esteso a tutta la polpa (Figura 1.7).

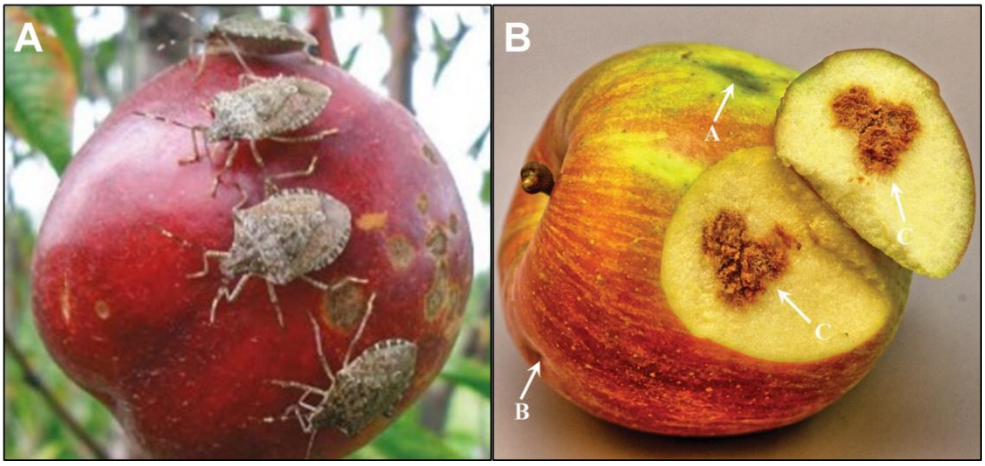

Figura 1.7 *Danni da Halyomorpha halys su mela: danni esterni (A) e interni con segnalazione dei fori di ingresso (B) (Buffington et al., 2018; Morrison III et al., 2019).*

1.3 Tecniche di monitoraggio in campo

Le conseguenze dei cambiamenti climatici e degli effetti negativi delle attività antropiche sull'ambiente hanno fatto emergere la necessità di sviluppare strategie sostenibili nella gestione degli agroecosistemi al fine di conciliare un incremento della resa e della produzione alimentare con la preservazione dell'ambiente.

Per quanto riguarda la gestione delle infestazioni, negli ultimi anni sta avendo sempre maggiore diffusione la gestione integrata delle avversità (*Integrated pest management*, IPM), che consente la salvaguardia delle produzioni agrarie tramite la razionalizzazione e l'integrazione dell'uso di tutte le possibili strategie di controllo disponibili rispettando, al tempo stesso, i principi ecologici, tossicologici ed economici.

In quest'ottica, la difesa sostenibile si basa sul concetto di gestione delle avversità con l'obiettivo ultimo di ridurre l'impatto dell'avversità a livelli tollerabili, ossia al di sotto delle soglie di danno economico individuate, spostando il focus dalla semplice presenza dell'avversità alla valutazione della densità di popolazione degli infestanti e ai possibili effetti sulle coltivazioni, considerando al tempo stesso il contesto ecologico, economico ed ambientale. A tale scopo, il monitoraggio rappresenta un punto cruciale nella gestione sostenibile del sistema agroalimentare.

Generalmente, il monitoraggio può essere effettuato tramite metodi attivi o passivi. I metodi attivi risultano particolarmente dispendiosi in termini economici e di tempo per gli agricoltori, in quanto prevedono l'ispezione diretta in campo da parte di personale specializzato. Dall'altro lato, i metodi passivi, come l'utilizzo di trappole, hanno un'efficacia minore e possono incrementare i danni alle piante limitrofe alla trappola.

I principali strumenti per il monitoraggio della cimice asiatica sono le trappole commerciali innescate con feromone di aggregazione, *frappage*, campionamenti con retino da sfalcio e controlli visuali. I primi due sono adatti alle aree esterne delle coltivazioni o frutteti per cui sono posizionate principalmente su siepi e piante di bordo. Infatti, i primi siti da monitorare a livello territoriale sono le aree di bordo dei frutteti, la vegetazione spontanea lungo i corsi d'acqua e le scoline, le siepi, le fasce boscate, ed altre eventuali colture ospiti come orticole e seminativi.

La tecnica più diffusa per il monitoraggio di cimice asiatica prevede l'utilizzo di trappole poiché più semplice da attuare e meno dispendiosa rispetto ad altri metodi di campionamento diretto. Il funzionamento delle trappole si basa sulla presenza di uno stimolo attrattivo, normalmente specie-specifico, il quale richiama gli esemplari intrappolandoli all'interno della trappola stessa. Lo stimolo attrattivo comunemente utilizzato per la cattura di cimice asiatica è una miscela di 10,11-epossi-1-bisabolen-3-olo, feromone di aggregazione prodotto dai maschi di cimice asiatica, e metil-(E,E,Z)-2,4,6-decatrienoato (o MDT), feromone di un'altra cimice (*Plautia stali*). Tale miscela compie la stessa funzione di un feromone di aggregazione essendo caratterizzato da notevole capacità attrattiva nei confronti di tutti gli individui della specie, i quali tendono ad aggregarsi sulla vegetazione limitrofa alla trappola. Nonostante ciò, tale metodo risulta problematico poiché presenta un rapporto tra numero di esemplari attirati e catturati molto basso, stimato tra il 5-10%, e, perciò, può causare danni collaterali alla vegetazione limitrofa.

I metodi di monitoraggio diretti sono diversi e possono prevedere: un'ispezione tramite metodo visivo standardizzato nel tempo (*frappage*), che consiste nel campionamento tramite battimento standardizzato di rami e nella successiva raccolta degli insetti che cadono sopra ad ampi teli o in contenitori, o il campionamento tramite retino da sfalcio (*sweep-netting*), da utilizzare con le piante erbacee standardizzando il numero di battute effettuate.

Negli ultimi anni, allo scopo di effettuare un'identificazione più efficiente delle specie infestanti in campo, sono stati implementati sistemi di monitoraggio automatizzati i quali consentono un monitoraggio rapido, efficiente e in tempo reale. L'avvento e l'uso sempre più diffuso di tecnologie dell'informazione e della comunicazione (ICT), sensori, dispositivi per la comunicazione, Internet e software per la simulazione e la gestione di grandi quantità di dati, ha concesso diverse opportunità di modernizzazione dei sistemi di monitoraggio in campo. Tali strumenti possono essere sfruttati per l'agricoltura di precisione consentendo di analizzare la variabilità spazio-temporale di diversi fattori chiave, i quali possono influenzare salubrità e produttività delle piante. Le informazioni ottenute grazie a dispositivi di monitoraggio automatizzati possono essere archiviate grazie al collegamento con piattaforme digitali di storage. È necessario specificare che, sebbene siano tecniche di monitoraggio con la potenzialità di apportare numerosi vantaggi, essendo ancora in via di sviluppo non sono tuttora state implementate al fine di monitorare la specie infestante *H. halys*.

Tra i metodi più utilizzati per il monitoraggio automatizzato in campo vi sono trappole o droni dotati di sensori per la raccolta dati, equipaggiati con hardware e software per la trasmissione dei dati raccolti ad un server remoto accessibile online, utile ai fini di archiviazione ed elaborazione delle informazioni in database

georeferenziati. Anche in questo caso le trappole presentano uno stimolo attrattivo specie specifico al fine di esercitare una selezione e attrarre soltanto gli infestanti di interesse.

I dispositivi per il monitoraggio automatizzato presentano diverse tipologie di sensori e modalità di funzionamento: tra i più specifici vi sono i sensori ottici basati su tecniche di imaging. I sensori più semplici sono costituiti da fotocamere che vengono posizionate in campo nelle zone di interesse ed acquisiscono immagini digitali a colori.

I sistemi dotati di questo tipo di sensori possono funzionare a diversi livelli di automazione: nei sistemi semiautomatici il sensore raccoglie le immagini in determinati momenti e le trasmette via internet ad un server remoto, dove l'operatore ha la possibilità di contare gli insetti direttamente guardando l'immagine da un dispositivo in tempo reale, mentre in sistemi completamente automatizzati gli insetti target dell'immagine vengono riconosciuti e contati da algoritmi di classificazione delle immagini. In questi casi generalmente vengono considerate caratteristiche discriminanti come la dimensione o la proporzione del corpo, le quali possono essere utili per differenziare la specie bersaglio.

I vantaggi offerti dai metodi di monitoraggio automatizzato sono diversi:

- i dati disponibili in tempo reale sono facilmente rappresentabili nel tempo e nello spazio e possono essere utilizzati automaticamente per l'ottimizzazione dei metodi di controllo;
- i costi di manodopera e di trasporto delle trappole sono ridotti grazie alla possibilità di effettuare controlli meno frequenti;
- il monitoraggio risulta facilitato in aree remote o inaccessibili;
- i dati acquisiti possono essere trasmessi automaticamente al server ad intervalli di tempo desiderati;
- il monitoraggio su larga scala, dove le trappole sono localizzate in un vasto territorio, risulta più efficiente;
- tale approccio consente la scelta di interventi mirati e localizzati in tempo reale.

I limiti principali delle tecniche di monitoraggio automatizzate sono di carattere economico, legati ai costi per l'adozione di software di analisi ed elaborazione dell'immagine, telecamere con sensori dalle caratteristiche idonee, accesso internet e a server per la conservazione dei dati online. Ciò nonostante, tali tecniche di monitoraggio dimostrano grandi potenzialità soprattutto in contesti di produzioni ad alto valore commerciale o ingenti costi di manodopera. Gli studi effettuati finora sull'acquisizione di immagini in campo a carico di specie infestanti prevedono l'analisi e l'elaborazione di immagini digitali a colori (immagini RGB).

Tuttavia, l'analisi delle immagini nel visibile potrebbe risultare insufficiente per la rilevazione di specie invasive caratterizzate da mimetismo, come nel caso della cimice asiatica. Al fine di superare i limiti imposti dal mimetismo di *H. halys*, nel presente lavoro di tesi è stata proposta l'acquisizione di immagini iperspettrali nel vicino infrarosso (*Near Infrared Hyperspectral Imaging*, NIR-HSI).

2 Chemiometria

La chemiometria è una branca della chimica analitica che si basa sull'utilizzo di metodi matematici e statistici per analizzare dati di natura chimica. Grazie all'utilizzo di tale disciplina è possibile estrarre le informazioni utili da un sistema chimico attraverso l'applicazione di modelli matematici per l'interpretazione dei dati. Le finalità principali della chemiometria possono essere sintetizzate nei seguenti aspetti:

- progettazione, selezione ed ottimizzazione di procedure ed esperimenti;
- estrazione delle informazioni utili da una matrice di dati attraverso metodi di analisi statistica multivariata;
- aumento della conoscenza del sistema studiato mediante rappresentazioni grafiche delle informazioni ottenute.

Negli ultimi decenni, la chemiometria si è diffusa grazie allo sviluppo di tecniche analitiche sempre più complesse, le quali richiedono l'applicazione di metodi statistici multivariati per estrarre le informazioni utili, e alla disponibilità di software di calcolo che permettono l'elaborazione di dati complessi.

Le tecniche chemiometriche si basano su un approccio statistico multivariato al fine di estrarre le informazioni utili da matrici complesse. Ciò è fondamentale poiché tramite l'applicazione di metodi univariati è possibile considerare soltanto una variabile alla volta trascurando, di conseguenza, le interazioni che si possono instaurare fra tutte le variabili. Pertanto, i metodi univariati permettono di ottenere una valutazione limitata della matrice di dati in esame, causando una perdita dell'informazione utile. Infatti, l'approccio multivariato permette di considerare sia le interazioni fra le variabili stesse sia le interazioni fra le variabili e i campioni e, pertanto, consente di visualizzare la naturale struttura dei dati al fine di estrarre la maggior parte dell'informazione utile relativa alla matrice di dati considerata.

Generalmente, i dati da analizzare vengono riportati in matrici bidimensionali di dimensioni $n \times m$, in cui n corrisponde al numero di oggetti (ossia ciò che può essere considerato come un'entità distinta e che si vuole confrontare con altre) e m corrisponde al numero di variabili che caratterizzano l'oggetto (ovvero i parametri sperimentali misurati per ogni oggetto).

Ad oggi i campi di applicazione della chemiometria sono molteplici poiché queste tecniche possono essere applicate a metodi di analisi rapidi, non distruttivi e compatibili con l'automatizzazione, come i metodi spettroscopici. La chemiometria trova impiego anche nel settore alimentare per il controllo qualità, ai fini di rilevare frodi alimentari o per il monitoraggio e la gestione dei processi produttivi lungo tutta la filiera alimentare.

Per quanto riguarda l'ambito dell'agricoltura di precisione, e più in particolare l'oggetto del presente lavoro di tesi, l'applicazione di tecniche chemiometriche consente di ricavare informazioni utili dai sistemi di monitoraggio che sfruttano l'acquisizione di immagini iperspettrali tramite l'analisi esplorativa dei dati e la formulazione di modelli di classificazione multivariati per il riconoscimento della presenza di specie invasive in campo.

2.1 Pretrattamento dei dati

Prima di applicare i metodi chemiometrici per l'analisi esplorativa o per lo sviluppo di modelli di calibrazione e/o classificazione, è necessario effettuare il pretrattamento dei dati. Questo passaggio è fondamentale poiché consente contemporaneamente di minimizzare l'effetto di sorgenti di variabilità indesiderate e migliorare l'estrazione delle informazioni utili.

Anche la scelta dei metodi di pretrattamento più opportuni risulta un passaggio determinante in quanto dipende fortemente dalla natura dei dati da indagare ed è, inoltre, necessario per evitare l'introduzione di artefatti o la rimozione di informazioni utili.

I metodi di pretrattamento dei dati hanno lo scopo di migliorare la qualità dell'informazione che si può ottenere da un'analisi esplorativa del dataset e/o ottimizzare i modelli di classificazione o regressione. Essi possono essere distinti in due principali categorie: pretrattamenti per colonna, i quali agiscono su ogni singola variabile considerando tutti gli oggetti, e pretrattamenti per riga, i quali agiscono su una riga alla volta della matrice di dati considerando, quindi, tutte le variabili di un determinato oggetto. I diversi pretrattamenti possono essere utilizzati insieme o singolarmente a seconda delle esigenze. In particolare, mentre i pretrattamenti per colonna vanno obbligatoriamente applicati prima dell'utilizzo delle tecniche chemiometriche, i pretrattamenti per riga vengono applicati quando si ha a che fare con segnali, come ad esempio gli spettri NIR.

Tra i pretrattamenti per colonna più utilizzati vi sono *mean centering* e *autoscaling*. Il pretrattamento dei dati con *mean centering* consiste nel sottrarre ad ogni variabile originale di ciascun campione ($x_{i,j}$) il valore medio della variabile stessa ($\bar{x}_j$). Pertanto, per l'oggetto i-esimo della matrice di dati, il corrispondente valore mediocentrato della variabile j-esima ($x^*_{i,j}$) sarà dato dalla seguente equazione:

$$x^*_{i,j} = x_{i,j} - \bar{x}_j \tag{2.1}$$

Questo pretrattamento di colonna viene solitamente utilizzato quando si considera un dataset di variabili della stessa natura come, ad esempio, variabili di natura spettrale. Una volta effettuato il pretrattamento, le variabili del dataset avranno valore medio uguale a 0 e la medesima varianza delle variabili originali: ciò è fondamentale affinché l'analisi esplorativa dei dati descriva le direzioni di massima varianza.

Il pretrattamento dei dati con *autoscaling*, invece, viene utilizzato quando si è in presenza di variabili di natura diversa o con scala diversa, in modo tale che abbiano a priori la stessa importanza nell'analisi dei dati. Per ogni campione, i valori delle variabili autoscalate ($x^*_{i,j}$) si ottengono sottraendo a ciascuna variabile il proprio valore medio ($\bar{x}_j$) e dividendo per la deviazione standard (s_j), come riportato nella seguente equazione:

$$x^*_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_j} \tag{2.2}$$

Pertanto, le variabili autoscalate avranno media uguale a 0 e deviazione standard unitaria, in modo tale da limitare l'influenza data dalle scale delle diverse variabili sui risultati.

I pretrattamenti di riga si utilizzano spesso nell'analisi di segnali spettroscopici al fine di eliminare informazioni non pertinenti e rumore. Gli spettri NIR possono essere influenzati dalle caratteristiche fisiche del campione: ad esempio, queste ultime possono determinare fenomeni di *scattering* per i quali la luce viene diffusa in modo diverso a seconda della granulometria dei campioni. In questi casi, i pretrattamenti di riga consentono di limitare o eliminare il contributo dello *scattering* sugli spettri in modo da permettere una migliore estrazione dell'informazione utile (informazione chimica) dal dataset. I pretrattamenti di riga più utilizzati in ambito della spettroscopia NIR (sia in riflettanza che in trasmittanza) possono essere suddivisi in metodi di correzione dello *scattering* e metodi di filtraggio, quali lo *smoothing* e le derivate.

I principali pretrattamenti di riga utilizzati nell'ambito del presente lavoro di tesi sono stati *Detrend*, *Multiplicative Scatter Correction* (MSC), *Standard Normal Variate* (SNV) e Derivate (metodo di *Savitzky-Golay*) (Figura 2.1):

- *Detrend*: pretrattamento che consente la correzione nei casi in cui è presente una deviazione costante, lineare o non lineare. Grazie al metodo dei minimi quadrati viene calcolato il polinomio di un dato ordine che meglio approssima i valori del segnale. Successivamente, i valori calcolati mediante tale polinomio vengono sottratti allo spettro originale ottenendo il segnale pretrattato. In questo modo, è possibile eliminare gli effetti dello *scattering* e ottenere un segnale non afflitto da un *offset* della linea di base, rendendo così i segnali direttamente confrontabili fra loro.
- *Multiplicative Scatter Correction* (MSC): pretrattamento atto alla correzione delle altezze dei diversi segnali e degli spostamenti della linea di base, permettendo di valutare soltanto la variazione di altezza delle diverse bande di segnale. In questo caso, considerando uno spettro di riferimento (generalmente lo spettro medio), si calcola il modello di regressione lineare dei valori di assorbanza di ciascuno spettro calcolato in funzione dei corrispondenti valori dello spettro di riferimento. Lo spettro viene quindi corretto utilizzando la pendenza e l'intercetta del modello di regressione.
- *Standard Normal Variate* (SNV): pretrattamento che consiste nella sottrazione del valore medio ad ogni segnale, il quale viene successivamente normalizzato dividendo per la deviazione standard del segnale stesso. Nella pratica, SNV è una sorta di *autoscaling* per riga, pertanto, ciascun segnale pretrattato avrà media uguale a zero e deviazione standard unitaria. Anche questo pretrattamento consiste nel minimizzare la variazione dovuta a *scattering* e, grazie ad esso, è possibile ottenere spettri confrontabili in termini di intensità o assorbanza.
- *Derivate:* le derivate utilizzate più utilizzate per il pretrattamento dei segnali sono derivata prima e derivata seconda. La derivata prima consente di rimuovere l'effetto dell'offset mentre la derivata seconda permette la rimozione sia dell'effetto dell'offset che della pendenza. In aggiunta, l'utilizzo della derivata offre la possibilità di separare meglio i contributi delle bande spettrali che sono sovrapposte nel segnale originale. Il principale svantaggio risiede nel fatto che la derivata può alterare in modo consistente il segnale e, se quest'ultimo è particolarmente affetto da rumore, essa tende ad accentuarlo. Inoltre, il pretrattamento con

derivata tende a causare uno spostamento dei picchi, il quale può determinare difficoltà interpretative degli spettri ottenuti.

Il calcolo delle derivate avviene tramite l'applicazione dell'algoritmo di Savitzky-Golay (SavGol), il quale si basa sull'utilizzo di una finestra mobile la cui ampiezza è definita da un numero dispari di punti dello spettro, definito dall'operatore. I punti presenti all'interno della finestra vengono utilizzati per calcolare il polinomio di un dato ordine che meglio li approssima. Successivamente, si calcola il valore della derivata del polinomio nel valore centrale della finestra e si ottiene così il primo valore dello spettro pretrattato. Grazie allo spostamento della finestra mobile, le stesse operazioni vengono ripetute lungo l'intero segnale ottenendo lo spettro pretrattato. Inoltre, il medesimo algoritmo può essere utilizzato per effettuare il pretrattamento *smoothing*: come suggerisce il nome, questo pretrattamento consente di ammorbidire e filtrare il segnale eliminando le oscillazioni dovute al rumore strumentale.

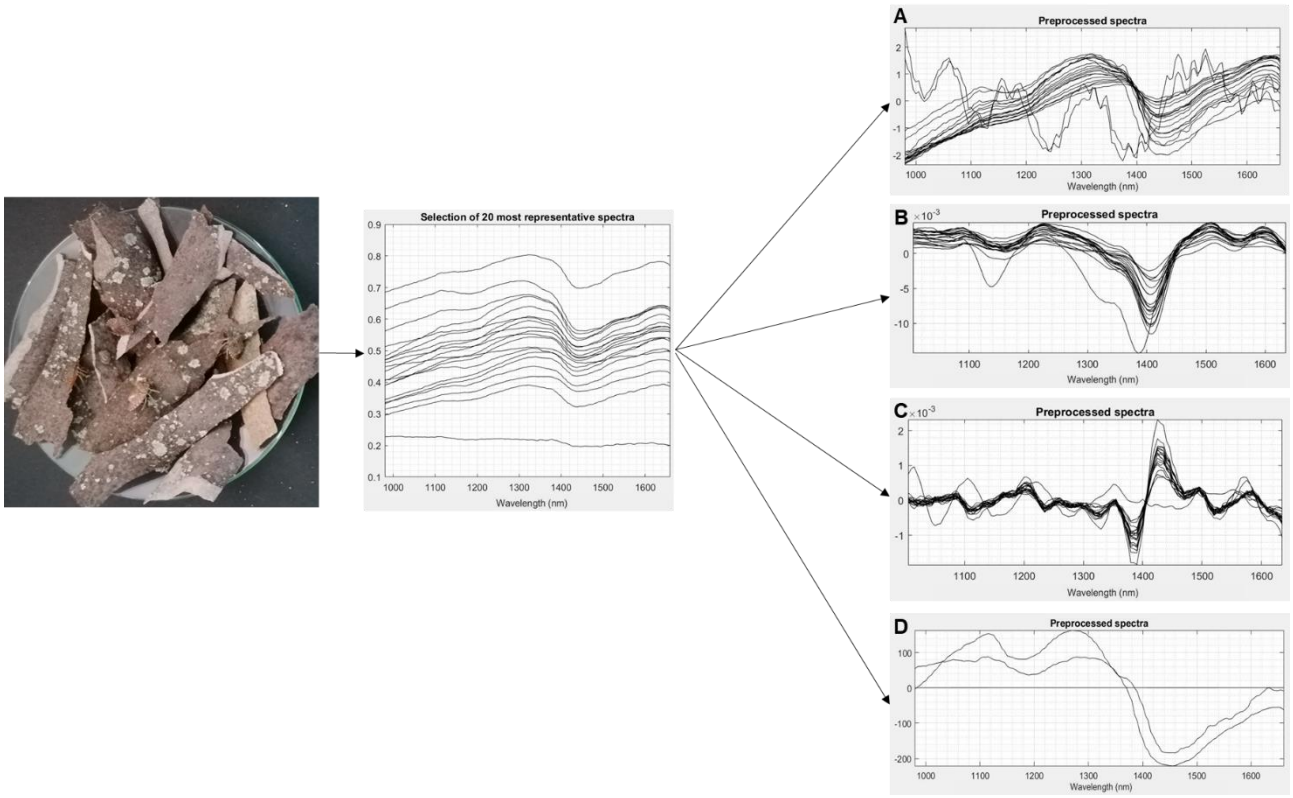

Figura 2.1 *Esempi di pretrattamenti di riga partendo dai segnali del campione Corteccia_HH_G1_a: SNV (A), derivata di primo ordine (B), derivata di secondo ordine (C) e MSC (D). Le immagini sono state ottenute tramite il software HYPER-Tools (https://www.hypertools.org).*

2.2 Analisi esplorativa dei dati

L'analisi esplorativa di un dataset è cruciale per poter carpire informazioni riguardanti la struttura dei dati. Pertanto, è importante servirsi di tecniche esplorative mirate a questo scopo considerando, allo stesso tempo, un approccio multivariato. In questo modo, è possibile visualizzare la naturale struttura dei dati senza forzare alcuna correlazione o classificazione. Pertanto, l'obiettivo ultimo dell'analisi esplorativa è trovare una relazione di qualsiasi tipo tra n oggetti e m variabili.

A tale scopo, l'analisi delle componenti principali (*Principal Component Analysis*, PCA) è una delle tecniche di statistica multivariata più diffuse e, inoltre, è tra le metodologie che più hanno influenzato le procedure di analisi e di interpretazione dei dati sperimentali. La PCA permette di visualizzare ed estrarre l'informazione utile da dataset multivariati eliminando, almeno parzialmente, il contributo del rumore, garantendo quindi una analisi più efficace dell'informazione contenuta nei dati.

L'obiettivo principale della PCA consiste, quindi, nell'estrazione di informazioni utili dal dataset e nella visualizzazione della struttura dei dati grazie a rappresentazioni grafiche, le quali permettono di valutare le relazioni esistenti tra gli oggetti e le variabili al fine di identificare gruppi di campioni con caratteristiche simili, campioni *outlier* nonché relazioni tra oggetti e variabili. Affinché ciò sia possibile, PCA consente di visualizzare l'informazione contenuta nel dataset originale in uno spazio a dimensionalità ridotta definito da variabili latenti dette componenti principali (PC), ossia un nuovo set di variabili ortogonali tra loro e calcolate come combinazione lineare delle variabili originali.

Fondamentalmente, il passaggio dallo spazio delle variabili originali allo spazio delle componenti principali comporta diversi vantaggi:

- solitamente per descrivere l'informazione utile è sufficiente un numero ridotto di PC, di gran lunga inferiore rispetto al numero delle variabili originali considerate;
- le PC non hanno alcuna correlazione e sono ortogonali tra loro; pertanto, ciascuna PC descrive un'informazione diversa rispetto alle altre;
- le PC riflettono la variabilità sistematica dei dati consentendo di ottenere l'informazione utile e, allo stesso tempo, limitano il contributo dell'errore sperimentale;
- PCA consente di ottenere una rappresentazione grafica immediata dei dati e facilmente interpretabile.

Ogni componente principale descrive una determinata percentuale di varianza del dataset originale. La prima componente principale (PC1) corrisponde alla direzione di massima varianza, ossia alla direzione di massima dispersione dei dati e, pertanto, è la direzione predominante allo scopo di descrivere il dataset considerato, sebbene non sia necessariamente l'unica. La seconda componente principale (PC2) è vincolata ad essere ortogonale alla prima e descrive la seconda direzione di massima varianza. Allo stesso modo, possono essere calcolate PC successive, le quali descrivono una percentuale di varianza residua del dataset minore rispetto alle precedenti. In generale, per la rappresentazione di un dataset sono sufficienti poche PC poiché le direzioni

di massima varianza corrispondono alle direzioni di variazione più significative; le PC che presentano valori molto bassi di varianza spiegata descrivono le variazioni dovute prevalentemente all'errore sperimentale.

Al fine di rappresentare al meglio il dataset nello spazio delle componenti principali, la scelta del numero di componenti principali è un passaggio cruciale; tale valutazione è da effettuare caso per caso considerando la percentuale di varianza totale spiegata da ogni componente principale, lo scopo per il quale vengono utilizzate le componenti principali ma anche il problema in esame.

Nel calcolo del modello PCA, il numero di PC significative può essere individuato grazie al grafico *scree* plot (Figura 2.2), nel quale vengono riportati il numero delle componenti principali (ascissa, *x*) e la percentuale di varianza catturata corrispondente (ordinata, *y*). La scelta del numero di componenti principali significative può essere effettuata dal punto di vista pratico osservando la variazione di pendenza della curva riportata nello scree plot: in questo modo, le zone a pendenza costante corrispondono alla variazione residua mentre in corrispondenza della zona in cui la curva si spezza ed è presente un "gomito" la variazione è dovuta principalmente all'informazione utile.

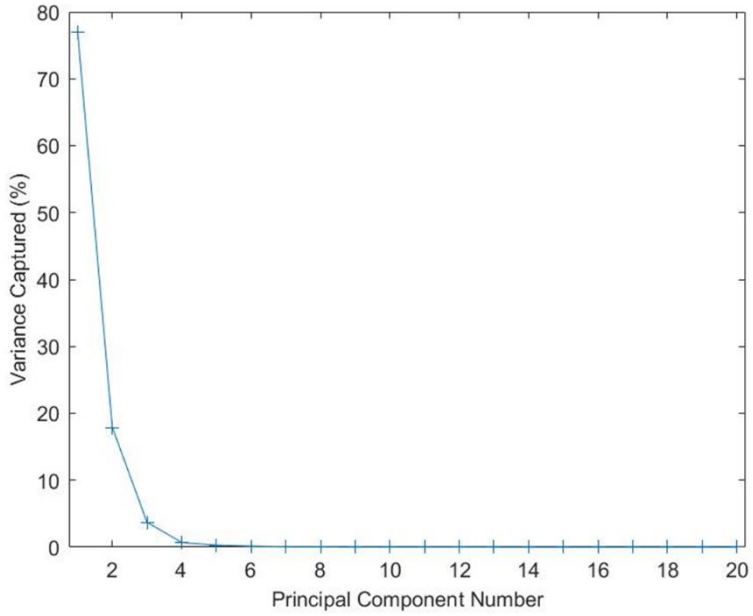

Figura 2.2 *Esempio di scree* plot

Successivamente, una volta scelto il numero di PC più adeguato, è possibile visualizzare l'eventuale presenza di campioni *outlier* grazie al grafico dei valori di T^2 di Hotelling contro i valori dei residui Q (Figura 2.3). Il valore di T^2 di Hotelling descrive l'influenza di un determinato oggetto sul modello mentre i valori dei residui Q rappresentano quanto un determinato oggetto dista dal valore che viene stimato per esso dal modello. Inoltre, al fine di determinare le motivazioni delle anomalie, è possibile visualizzare i contributi delle variabili originali al campione in esame tramite i *contribution* plot di T^2 di Hotelling e/o dei residui Q. Pertanto, il grafico in cui

vengono riportati i valori T^2 di Hotelling (ascissa, *x*) e i residui Q (ordinata, *y*) permette di identificare ed eliminare i campioni *outlier* evitando che essi possano influenzare in modo errato le direzioni delle componenti principali influendo negativamente sul modello calcolato.

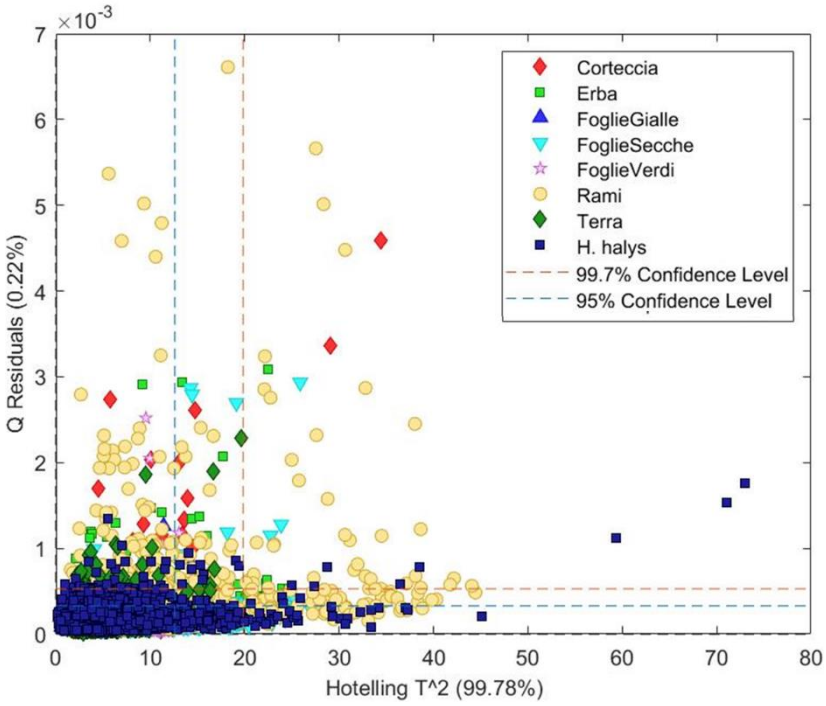

Figura 2.3 *Grafico T^2 di Hotelling e residui Q.*

Dopo aver effettuato la scelta del numero di componenti principali significative e aver rimosso eventuali campioni *outlier*, è possibile visualizzare lo spazio descritto dalle componenti principali. In primo luogo, è necessario definire ciascuna componente principale tramite due vettori:

- Il vettore degli *score*, il quale indica la posizione degli *n* oggetti lungo la componente principale e, pertanto, presenta un numero di elementi corrispondente al numero di campioni; esso si ottiene grazie alla proiezione ortogonale di ciascun oggetto lungo la componente principale. Gli score descrivono il contributo che ciascun campione determina sulla varianza totale della PC: infatti, gli oggetti più estremi (caratterizzati da valore assoluto del vettore degli score più alto rispetto agli altri), influenzano maggiormente la direzione della componente principale. I valori dei vettori degli score per ogni componente principale possono essere riportati in un grafico a dispersione, detto *score plot* (Figura 2.4), al fine di valutare la distribuzione degli oggetti, l'eventuale presenza di raggruppamenti tra campioni con caratteristiche similari oppure la presenza di *outlier*;

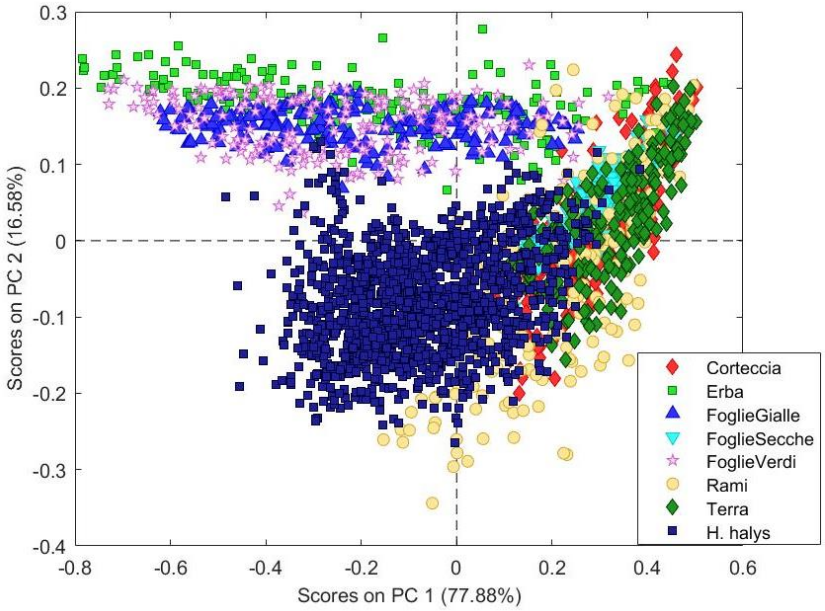

Figura 2.4 *Esempio di* score plot *PC1/PC2.*

- Il vettore dei *loading*, il quale indica il peso di ogni variabile originale nel determinare la direzione della componente principale considerata. Per ogni componente principale a e per ogni variabile x_j, il vettore dei loading si presenta come una colonna di numeri dove ciascun valore è dato dal prodotto tra un fattore di correzione α, il quale dipende a sua volta dalla scala della variabile originale, ed il coseno direttore, il quale è legato alla direzione della variabile rispetto alla componente principale considerata (Equazione 2.3).

$$\alpha(x_j)\, cos(\widehat{x_j PC_a}) \tag{2.3}$$

Come nel caso degli score, anche i valori dei vettori dei loading possono essere rappresentati in un grafico a dispersione chiamato *loading plot*. Il grafico dei loading è utile per determinare quali siano le variabili originali più influenti nel definire la direzione di una certa componente principale e per capire quali siano le relazioni tra le variabili. Ad esempio, le variabili che presentano la stessa direzione rispetto all'origine del grafico sono correlate positivamente, le variabili posizionate ortogonalmente rispetto all'origine non sono correlate tra loro mentre le variabili posizionate in posizione opposta rispetto all'origine sono correlate negativamente. Tuttavia, quando si ha a che fare con dati di natura spettrale dove le variabili sono numerose e fortemente correlate tra loro, il loading plot sotto forma di grafico a dispersione risulta di difficile interpretazione. Pertanto, in questi casi, ciascun vettore dei loading può essere rappresentato come una curva in funzione del dominio delle variabili originali, ossia delle lunghezze d'onda nel caso si considerino segnali spettrali (Figura 2.5).

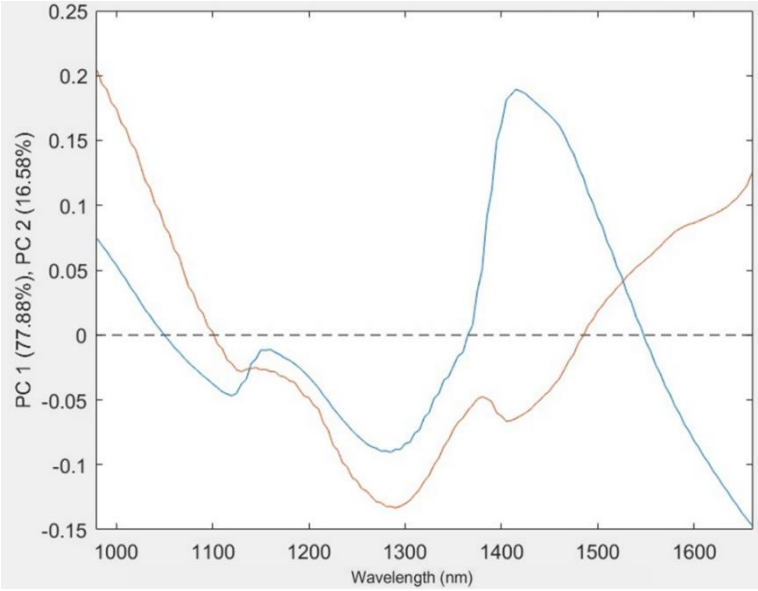

Figura 2.5 *Esempio di* loading plot *di un modello PCA calcolato a partire da spettri NIR.*

Anche nel caso dei segnali le regole per l'interpretazione dei loading rimangono le stesse: le regioni dello spettro che presentano valori di loading che si discostano maggiormente dall'origine corrispondono alle lunghezze d'onda che influenzano maggiormente la direzione della componente principale considerata.

Dato che per ciascuna componente principale può essere definito un vettore degli score, relativo alla proiezione degli oggetti lungo la PC, ed un vettore dei loading, relativo al peso di ciascuna variabile nel definire la direzione della PC, la relazione tra i due può essere definita dall'equazione 2.4:

$$t_{a,i} = p_{a,1}x_{i,1} + p_{a,2}x_{i,2} + \cdots + p_{a,j}x_{i,j} + \cdots p_{a,m}x_{i,m} \tag{2.4}$$

dove per un generico oggetto i, il valore di score lungo la a-esima PC ($t_{a,i}$) può essere calcolato dalla combinazione lineare delle variabili originali, i cui pesi corrispondono ai valori di loading per quella PC.

Passando alla notazione matriciale, un modello PCA può essere definito come segue (Figura 2.6):

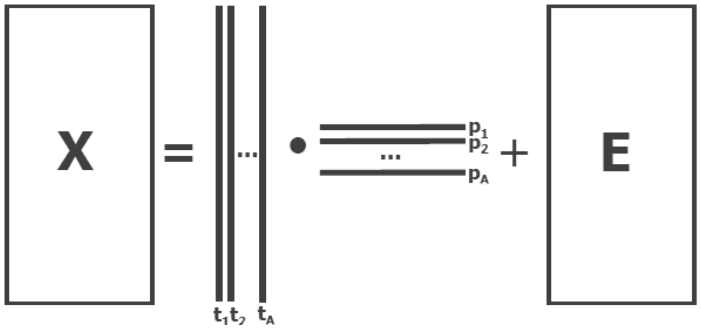

Figura 2.6 *Definizione di PCA secondo notazione matriciale.*

dove la varianza della matrice dei dati mediocentrati (*X*) viene decomposta in una parte relativa agli oggetti (matrice degli score, *T*), una parte relativa alle variabili (matrice dei loading, *P*), ed una parte relativa alla variazione residua non descritta dal modello e riconducibile al rumore (matrice dei residui, *E*).

2.3 Classificazione

Per classe si intende solitamente un insieme di oggetti con una o più proprietà in comune. Tali proprietà possono essere generalmente descritte mediante variabili matematiche e, di conseguenza, si può affermare che oggetti appartenenti alla stessa classe presentano lo stesso valore di una o più variabili discrete, oppure hanno valori simili all'interno di un range ben definito di variabili continue.

I metodi di classificazione multivariata devono consentire di riconoscere l'appartenenza di un campione ad una determinata classe in base ai valori sperimentali delle variabili misurate per quel campione. Dal punto di vista matematico, i metodi di classificazione prevedono il calcolo di modelli atti a definire le relazioni esistenti tra le variabili di natura sperimentale e una variabile qualitativa che identifica la classe. Pertanto, come per la calibrazione, la classificazione consente di definire un modello per prevedere una variabile *y* partendo da una matrice *X*: in questo caso, però, la variabile *y* rappresenta una qualità, la quale implica, a sua volta, l'appartenenza o la non appartenenza a una determinata classe. Inoltre, i metodi di classificazione necessitano di un valore di soglia (*threshold*) per ogni classe, il quale permette di definire una linea di separazione al fine di attribuire o meno un campione ad una determinata classe. A fine di determinare l'efficienza di un modello di classificazione è necessario considerare quanti campioni vengono attribuiti correttamente ad una determinata classe.

Per valutare la performance predittiva del modello è necessario effettuare una validazione tramite l'uso di un test set esterno. Per fare ciò, generalmente il dataset si suddivide nel *training set*, composto da circa i 2/3 dei campioni totali ed utilizzato per il calcolo del modello di classificazione, e nel *test set*, composto dal restante 1/3 dei campioni, il quale viene utilizzato per la validazione esterna, ovvero per valutare la reale capacità predittiva del modello.

Nello sviluppo di un modello di classificazione è generalmente necessario ottimizzare alcuni parametri, come ad esempio il numero di variabili latenti in *Partial Least Squares Discriminant Analysis* (PLS-DA), ovvero la dimensionalità ottimale del modello, o il pretrattamento ottimale dei segnali spettrali; la definizione opportuna di questi parametri è un passaggio cruciale affinché il modello ottenuto risulti efficiente. La scelta si effettua mediante cross-validazione utilizzando i medesimi oggetti del training set, impiegati per la costruzione del modello. In pratica, la cross-validazione consiste nella suddivisione del training set in un determinato numero di gruppi di cancellazione. Durante il calcolo del modello, per ciascun numero di variabili latenti il modello viene ricalcolato escludendo di volta in volta i campioni appartenenti ad un determinato gruppo di cancellazione; in seguito, il modello ottenuto viene utilizzato per prevedere i campioni esclusi. Tale procedura

viene ripetuta fintanto che non vengono esclusi tutti i campioni del training set e, infine, i risultati ottenuti in cross-validazione vengono confrontati con le classi a cui effettivamente appartengono i campioni.

I metodi di cross-validazione possono essere diversi e si distinguono principalmente in base alla modalità con cui i campioni del training set vengono suddivisi nei gruppi di cancellazione:

- metodo *Contiguous Blocks*: definisce gruppi contigui di campioni;
- metodo *Venetian Blinds*: separa gli oggetti in maniera alternata nei diversi gruppi di cancellazione, i quali risultano sparsi su tutte le righe dell'intervallo;
- metodo *Random Groups*: consente di assegnare in modo casuale i campioni a ciascun gruppo di cancellazione;
- metodo *Custom*: prevede la creazione di vettori di cross-validazione personalizzati, definiti *ad hoc* per il dataset considerato. In questo caso, l'operatore può scegliere liberamente la suddivisione nei diversi gruppi di cancellazione in base alle caratteristiche del dataset.

Mediante la cross-validazione è possibile valutare l'errore del modello in funzione del numero di variabili latenti, ed il numero ottimale di LV del modello viene scelto in corrispondenza del valore minimo di errore in cross-validazione. È importante selezionare il numero corretto di variabili latenti da includere nel modello per evitare problematiche relative ad *underfitting* o o*verfitting*. Nei casi in cui viene considerato un numero troppo elevato di variabili latenti si verifica overfitting, ossia per la creazione del modello vengono considerate sia l'informazione utile sia l'errore sperimentale; di conseguenza, il modello risultante avrà scarse performance nel prevedere campioni diversi rispetto a quelli del training set. Viceversa, considerando un numero troppo basso di variabili latenti, si verifica underfitting, ovvero non sono state incluse sufficienti informazioni per il calcolo del modello utili per ottenere buoni risultati.

Generalmente, i risultati dei modelli di classificazione vengono rappresentati sotto forma di matrice di confusione, nella quale le righe corrispondono alle classi di appartenenza degli oggetti e le colonne rappresentano le classi assegnate dal modello (Tabella 2.1). Ciascun elemento $c_{g,k}$ della matrice di confusione rappresenta il numero di campioni appartenenti alla classe *g* e predetti come appartenenti alla classe *k*. Pertanto, lungo la diagonale della matrice di confusione sono riportati i campioni che per ciascuna classe sono stati correttamente classificati dal modello.

		Classe predetta		
		Classe 1	**Classe 2**	**Classe 3**
Classe di appartenenza	**Classe 1**	23	2	5
	Classe 2	5	25	0
	Classe 3	5	3	22

Tabella 2.1 *Esempio di matrice di confusione.*

Per ciascuna classe è possibile riassumere i risultati riportati nella matrice di confusione utilizzando i seguenti parametri:

- *True positives* (TP) ossia il numero di oggetti correttamente assegnati ad una certa classe,
- *True negatives* (TN) ossia il numero di oggetti correttamente rigettati da una certa classe,
- *False positives* (FP) ossia il numero di oggetti erroneamente attribuiti ad una certa classe,
- *False negatives* (FN) ossia il numero di oggetti erroneamente rigettati dalla propria classe di appartenenza.

Inoltre, a partire dai valori di TP, TN, FP e FN di ciascuna classe è possibile calcolare diversi parametri di performance del modello che permettono di valutare in maniera coincisa i risultati della classificazione relativi ad ogni classe considerata e al modello nel complesso.

Per valutare le performance ottenute per ogni classe si considerano generalmente i seguenti parametri:

- Sensibilità (SENS): corrisponde al rapporto tra i campioni correttamente assegnati ad una determinata classe (TP) ed il numero totale degli oggetti che realmente appartengono a quella stessa classe (TP+FN) (equazione 2.5). Tale valore esprime la percentuale di oggetti appartenenti alla classe considerata correttamente assegnati dal modello.

$$SENS = \frac{TP}{TP+FN} \tag{2.5}$$

- Specificità (SPEC): corrisponde al rapporto tra il numero di campioni correttamente non assegnati ad una determinata classe (TN) ed il numero totale di oggetti che realmente non appartengono alla classe in esame (TN+FP) (equazione 2.6). Tale valore esprime la percentuale di oggetti appartenenti ad altre classi correttamente rigettati dal modello.

$$SPEC = \frac{TN}{TN+FP} \tag{2.6}$$

- Efficienza di classificazione (EFF): corrisponde alla media geometrica tra i valori di SENS e SPEC (equazione 2.7).

$$EFF = \sqrt{SENS \times SPEC} \tag{2.7}$$

I valori di SENS, SPEC e EFF possono essere calcolati in calibrazione e cross-validazione considerando i campioni del training set e, per la validazione del modello, possono essere calcolati in predizione sul test set esterno.

Al fine di valutare la qualità globale di un modello di classificazione è necessario utilizzare parametri in grado di fornire una valutazione del modello nel suo insieme e tali parametri sono in genere ottenuti a partire dalle performance calcolate per ogni singola classe. A tale scopo, vengono generalmente utilizzati i seguenti parametri:

- Accuratezza (ACC): corrisponde al rapporto tra il numero di campioni classificati correttamente ed il numero totale di campioni. Tale parametro viene calcolato senza prendere in considerazione le prestazioni delle singole classi e, nel caso in cui il numero di campioni appartenenti alle diverse classi non sia bilanciato il valore di accuratezza è maggiormente influenzato dai risultati ottenuti per la classe con il maggior numero di campioni.
- Non-error rate (NER): corrisponde alla media aritmetica della sensibilità delle classi considerate.
- Efficienza media ($MEFF$): corrisponde alla media aritmetica dell'efficienza delle singole classi.

La classificazione può essere basata su due approcci distinti: l'analisi discriminante e il modellamento di classe. L'analisi discriminante ha come obiettivo l'identificazione di una relazione matematica tra un set di variabili descrittive e una categoria di risposte (y) e per l'applicazione dei metodi di analisi discriminante è necessario definire almeno due classi. I metodi discriminanti permettono di individuare confini ben precisi che suddividono il dominio globale in un numero di regioni pari al numero di classi considerate. Tali confini vengono calcolati in modo tale da massimizzare la discriminazione dei campioni appartenenti alle classi in esame: pertanto, secondo queste modalità, ciascun campione viene attribuito forzatamente ad una delle classi considerate a seconda che si trovi al di sopra o al di sotto della linea di soglia corrispondente alla classe. Il risultato che si ottiene non può che essere binario, ossia se il campione appartiene o non appartiene alla classe. Lo svantaggio nei metodi di analisi discriminante consiste nel fatto che questo approccio non prende in considerazione l'eventuale presenza di campioni *outlier*, ovvero di campioni non appartenenti alle classi considerate per la costruzione del modello. Uno dei metodi di analisi discriminante più utilizzati è l'algoritmo PLS-DA.

Un discorso differente vale per l'approccio del modellamento di classe, che consiste nel considerare le classi singolarmente e valutare la possibilità che un campione vi appartenga o meno. Questi metodi definiscono un modello matematico in grado di descrivere le caratteristiche di ciascuna classe indipendentemente dalle altre classi in esame potendo, pertanto, essere applicati anche in situazioni in cui sia presente una sola classe di interesse come nel caso di problematiche legate all'autenticazione degli alimenti. Modellando separatamente ciascuna classe è possibile che un campione *outlier* venga correttamente non assegnato a nessuna delle classi considerate. Dall'altro lato è anche possibile che uno stesso oggetto venga riconosciuto come appartenente a più di una classe generando ambiguità; tale situazione si verifica quando le classi modellate sono almeno parzialmente sovrapposte, in quanto il modello non è orientato alla discriminazione delle categorie considerate. Uno degli algoritmi più diffusamente utilizzati nell'ambito del modellamento di classe è *Soft Independent Modelling of Class Analogy* (SIMCA).

Per sfruttare i vantaggi di entrambi gli approcci di classificazione, nel presente lavoro di tesi per lo sviluppo dei modelli di classificazione è stato utilizzato l'algoritmo Soft PLS-DA, che verrà descritto in seguito in Sezione 2.3.3.

2.3.1 PLS-DA

Partial Least Squares Discriminant Analysis (PLS-DA) è uno dei metodi di classificazione più diffusi, e si basa sull'adattamento dell'algoritmo di regressione multivariata *Partial Least Squares* (PLS) ad un contesto di analisi discriminante. Nel caso di PLS-DA i modelli calcolati si basano su un modello di regressione PLS2, dove la matrice Y è rappresentata da una matrice fittizia denominata "*dummy*", la quale codifica l'appartenenza dei campioni alle classi considerate grazie ad una codifica binaria. Nel dettaglio, la matrice Y *dummy* presenta tante colonne quante sono le classi e tante righe quanti sono i campioni. Per ogni classe considerata, la corrispondente colonna della matrice Y si presenta come un vettore binario, dove gli elementi uguali a 1 corrispondono ai campioni appartenenti alla classe in esame mentre quelli uguali a 0 identificano i campioni non appartenenti alla classe (Figura 2.7).

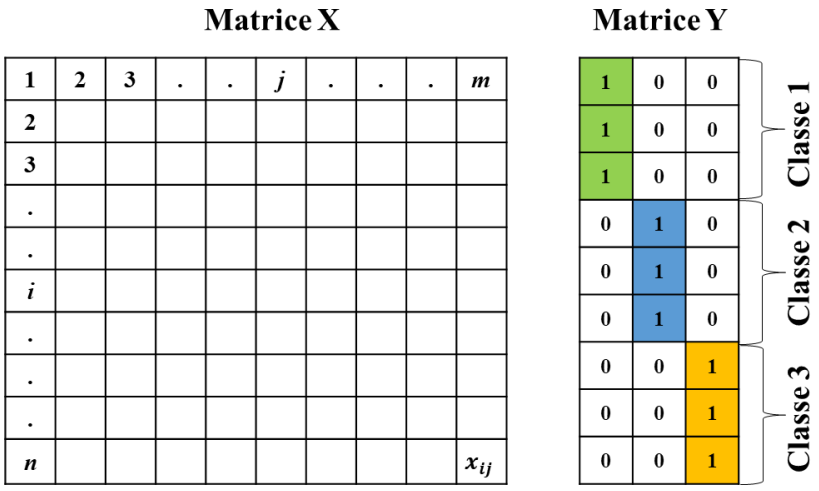

Figura 2.7 *Matrice X (variabili sperimentali) e corrispondente Y (dummy), calcolata tramite l'applicazione dell'algoritmo PLS-DA.*

Analogamente a PLS, PLS-DA prevede il calcolo di un nuovo set di variabili, definite variabili latenti (LV), che sono combinazioni lineari delle variabili originali definite in modo tale da descrivere la massima covarianza tra le due matrici X e Y. Poiché la matrice Y consiste nell'insieme dei vettori di classe, le variabili latenti di PLS-DA descrivono le direzioni lungo le quali si ha la massima discriminazione tra le diverse classi.

Una volta calcolato il modello, l'assegnazione di un campione ad una determinata classe si basa sul valore di *y* stimato dal modello ($\hat{y}$), che viene confrontato con la distribuzione dei valori di *y* stimati dal modello per i campioni appartenenti alla classe in esame. In particolare, per ogni classe i corrispondenti valori di *y* stimati dal modello non avranno mai valori esattamente uguali a 0 oppure uguali a 1, ma si avvicineranno a questi valori di riferimento. In PLS-DA si assume che la dispersione dei valori di *y* stimati per ogni classe dipenda principalmente dall'errore casuale e che, di conseguenza, seguano una distribuzione Gaussiana. Pertanto, a partire dai valori di *y* calcolati ($\hat{y}$) ottenuti dal modello in calibrazione per ciascuna classe, vengono calcolate le due distribuzioni Gaussiane dei valori di $\hat{y}$ per i campioni appartenenti e per quelli non appartenenti alla

classe considerata. Il punto in cui le due distribuzioni si incontrano, ossia dove le probabilità di appartenenza o non appartenenza alla classe sono equivalenti, viene utilizzato come valore soglia (*threshold*) per determinare l'appartenenza di un campione alla classe. Pertanto, se il valore di $\hat{y}$ per un campione incognito è superiore alla *threshold*, esso verrà assegnato alla classe corrispondente; viceversa, quest'ultimo non verrà assegnato alla classe.

2.3.2 SIMCA

Soft Independent Modelling of Class Analogies (SIMCA) è uno dei metodi di classificazione basato sull'approccio del modellamento di classe più diffusi. Questo metodo prevede la creazione di un modello PCA per ogni classe considerata: ciascun modello è caratterizzato da dimensionalità differente ed è indipendente dagli altri permettendo, di conseguenza, l'impostazione di valori di soglia distinti per ogni classe.

Il numero ottimale di componenti principali di ogni modello PCA viene individuato mediante cross-validazione ed utilizzando le statistiche di T^2 di Hotelling e dei residui Q vengono identificati eventuali campioni *outlier*, che possono essere eliminati per evitare che influenzino la direzione delle componenti principali. Su ogni componente principale del modello vengono calcolati gli score ed il relativo range, definito *normal range*, che indica una stima del range della popolazione (Figura 2.8). Come descritto in precedenza, dato che per definizione le PC sono ortogonali tra loro, il modello di ogni classe avrà dimensionalità e forma diverse a seconda del numero di PC considerate. Per definire il confine dello spazio intorno al modello di ogni classe si utilizza la statistica di Fischer, che definisce lo spazio racchiuso entro tale confine, definito SIMCA box. Pertanto, un nuovo oggetto verrà accettato dal modello della classe in esame se avrà un valore di F minore del valore critico per quella classe. Può quindi accadere che un oggetto venga accettato da più di un modello di classe oppure che venga rigettato da tutti i modelli di classe.

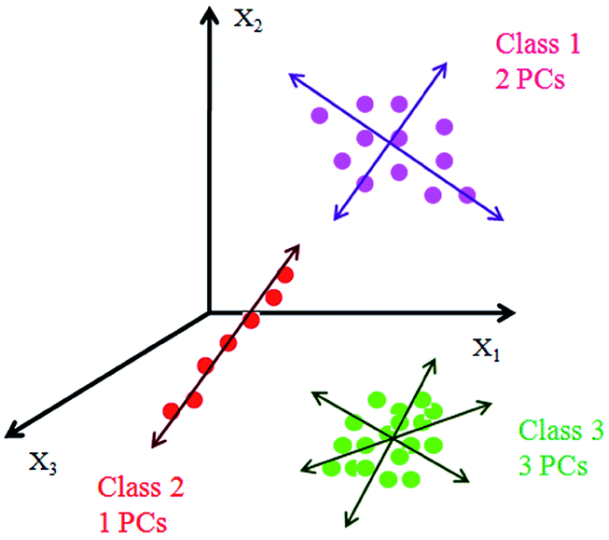

Figura 2.8 *Rappresentazione modello calcolato tramite l'applicazione dell'algoritmo SIMCA (Pirhadi et al., 2015).*

2.3.3 Soft PLS-DA

Dopo aver presentato i due diversi approcci alla classificazione possibili, risulta evidente il fatto che l'utilizzo di un metodo in grado di combinare i vantaggi di entrambi, ossia massimizzare la discriminazione tra le categorie di interesse e, al tempo stesso, riconoscere e rifiutare gli *outlier*, apporterebbe diversi benefici.

Una possibile soluzione è rappresentata da Soft PLS-DA, sviluppato dal gruppo di ricerca nell'ambito del quale è stato svolto il presente lavoro di tesi. Soft PLS-DA è un algoritmo implementato tramite una modifica dell'algoritmo PLS-DA e, pertanto, presenta il medesimo principio di base. La differenza sostanziale consiste nel fatto che l'assegnazione di un oggetto ad una determinata classe avviene fissando limiti aggiuntivi sia sui valori di *y* predetti sia sui residui Q. In questo modo, è possibile costruire un modello di classificazione massimizzando le differenze tra le classi modellate e, contemporaneamente, l'introduzione di limiti aggiuntivi consente di rifiutare campioni appartenenti a categorie non previste e di relegare questi ultimi in una categoria generale di campioni non assegnati.

In particolare, l'attribuzione di un campione ad una determinata classe deve soddisfare diversi criteri:

- il valore dei residui Q del campione deve rientrare nel limite di confidenza al 99,9 %. Tale valore è stato scelto in modo da porre limiti sufficientemente ampi da considerare il più possibile la variabilità delle diverse classi e, allo stesso tempo, essere in grado di escludere campioni con un adattamento molto basso al modello;
- i valori di *y* predetti devono rientrare nel range di accettabilità per la classe considerata. In particolare, in aggiunta al valore soglia calcolato per ciascuna classe g nella versione standard di PLS-DA ($y_{tsh1,g}$), è stato introdotto anche un limite superiore ($y_{tsh2,g}$) sui valori y predetti (equazione 2.8).

$$y_{tsh2,g} = m_{y,g} + 5 \times s_{y,g} \tag{2.8}$$

  dove $m_{y,g}$ e $s_{y,g}$ corrispondono, rispettivamente, a media e deviazione standard dei valori di *y* predetti per la classe g calcolati a partire dal training set. Pertanto, per essere assegnato alla classe g, un campione incognito deve avere un valore di *y* predetta compreso tra $y_{tsh1,g}$ e $y_{tsh2,g}$;
- nel caso in cui vi siano problemi di classificazione con più di due classi di interesse, i campioni devono essere assegnati in modo univoco ad una sola classe. In caso di ambiguità in campione viene considerato come non assegnato.

I campioni che non rientrano nei criteri precedentemente elencati non vengono assegnati a nessuna classe e vengono etichettati come *non assegnati* (NA). In questo modo, Soft PLS-DA, tramite il delineamento di limiti attorno a ciascuna classe modellata, consente di massimizzare la discriminazione tra le categorie di interesse e minimizzare la presenza di falsi positivi dovuti ad esempio alla presenza di *outlier* o alla presenza di campioni appartenenti a classi diverse rispetto a quelle considerate.

Anche nel caso di Soft PLS-DA la valutazione della performance dei modelli di classificazione può essere effettuata utilizzando i parametri precedentemente riportati.

2.4 Metodi di selezione delle variabili

Uno dei principali svantaggi riguardo l'acquisizione di immagini iperspettrali risiede nell'elaborazione di una grande quantità di dati, la quale è causa di difficoltà dal punto di vista computazionale, oltre a complicare l'estrazione delle informazioni utili. Per risolvere tali criticità è necessario individuare strumenti e tecniche chemiometriche in grado di gestire e analizzare in maniera efficiente un'elevata mole di dati.

Nel caso di sistemi che si avvalgono di immagini iperspettrali per un'attività di monitoraggio in tempo reale, una selezione delle sole variabili utili per il problema in esame permette una drastica riduzione del numero di variabili considerate ma anche del carico computazionale e, conseguentemente, dei tempi di calcolo. Inoltre, la selezione di variabili, effettuata a partire da sistemi iperspettrali, può essere un punto di partenza per l'implementazione di sistemi multispettrali specifici per il problema considerato, i quali sono molto più economici e veloci rispetto ai sistemi di tipo iperspettrale.

In questo lavoro di tesi, per identificare le regioni dello spettro NIR maggiormente informative per distinguere la cimice asiatica dai diversi sfondi vegetali sono stati utilizzati di metodi *sparse* di selezione di variabili.

I metodi *sparse* rappresentano un'estensione dei corrispettivi metodi tradizionali di classificazione o regressione in cui è possibile forzare ad essere uguali a zero i coefficienti del modello che corrispondono a variabili non rilevanti o contenenti un'eccessiva percentuale di rumore, individuando di conseguenza le regioni spettrali maggiormente significative per una specifica applicazione (Figura 2.9). La sparsità del modello, ossia il numero di variabili per cui i corrispondenti coefficienti del modello devono essere forzati ad essere uguali a zero, è un parametro che deve essere definito in maniera opportuna dall'operatore per evitare l'eliminazione accidentale di informazione utile.

Oltre alla definizione di un numero di componenti ottimali del modello, ossia di variabili latenti (LV), i metodi *sparse* richiedono anche l'ottimizzazione del livello di sparsità, che è legato al numero di variabili i cui coefficienti sono posti pari a zero nel modello. Generalmente, la combinazione ottimale di questi due parametri viene individuata in cross-validazione andando a selezionare quella combinazione che permette di minimizzare l'errore di classificazione e selezionare il minor numero di variabili, ottenendo quindi il giusto compromesso tra performance in classificazione e parsimonia del modello.

I metodi *sparse* permettono di effettuare selezione di variabili grazie all'introduzione di un termine di penalizzazione alla funzione obiettivo di un determinato metodo, ovvero alla funzione matematica che viene massimizzata o minimizzata nel calcolo di un modello. L'algoritmo *sparse* utilizzato nella fase sperimentale del presente elaborato si basa sul metodo di penalizzazione LASSO (*Least Absolute Shrinkage and Selection Operator*), che consiste nel forzare la somma dei valori assoluti di un vettore ad essere minore di un dato valore, in modo tale che i coefficienti relativi alle variabili meno significative risultino uguali a zero.

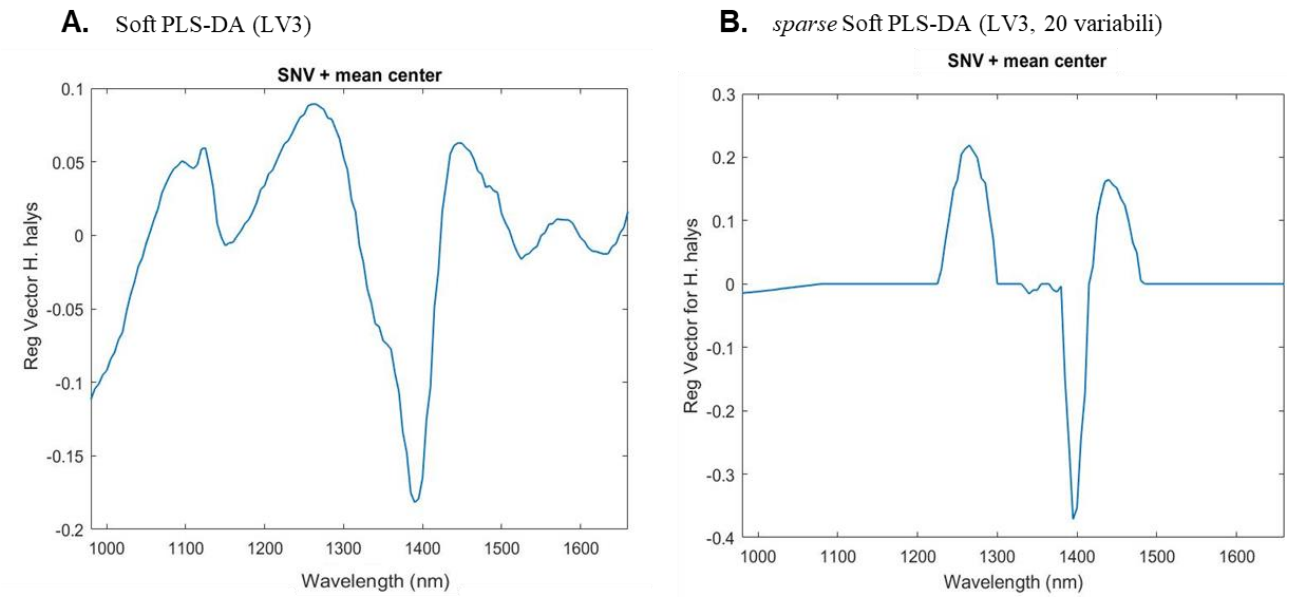

Figura 2.9 *Esempio di coefficienti di regressione di un modello Soft PLS-DA ottenuti prima (A) e dopo (B) l'applicazione di metodi sparse.*

Nel presente lavoro, l'approccio *sparse* è stato combinato con l'algoritmo Soft PLS-DA per evidenziare le regioni spettrali più influenti per la discriminazione di *H. halys* dagli sfondi vegetali considerati. L'applicazione di metodi *sparse* all'algoritmo Soft PLS-DA non modifica il principio di funzionamento di quest'ultimo: l'unica differenza è costituita dall'aggiunta di una penalizzazione (LASSO) sui loading e, di conseguenza, sui coefficienti di regressione utilizzati per prevedere i campioni incogniti. Grazie a questo abbinamento è possibile effettuare la selezione di variabili e la classificazione in un'unica fase.

3 Imaging iperspettrale nel vicino infrarosso

3.1 Cenni di spettroscopia NIR

Con il termine spettroscopia si intende la branca della fisica e della chimica che si occupa di studiare l'interazione fra radiazione elettromagnetica e materia. Secondo la meccanica quantistica, la radiazione elettromagnetica è caratterizzata da una duplice natura ondulatoria e corpuscolare.

La teoria ondulatoria considera la radiazione elettromagnetica come un'onda che si propaga nel vuoto o nei materiali. La radiazione elettromagnetica può essere descritta da un campo elettrico e uno magnetico oscillanti, sinusoidalmente in fase, perpendicolarmente fra loro e rispetto alla direzione di propagazione della luce (Figura 3.1). La radiazione trasporta quindi energia elettromagnetica, che nel vuoto si propaga con velocità costante, *c*, pari a $2{,}998 \times 10^8$ m/s.

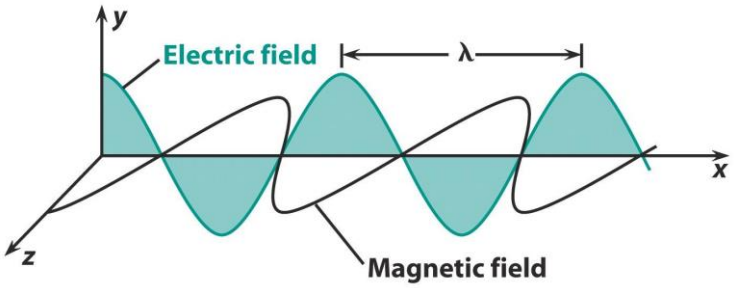

Figura 3.1 *Rappresentazione della radiazione elettromagnetica (Harris, 2010).*

La radiazione elettromagnetica può essere descritta da diversi parametri:

- lunghezza d'onda (λ), corrispondente alla distanza tra due picchi successivi dell'onda, misurata in nanometri;
- frequenza (ν), corrispondente al numero di oscillazioni complete che avvengono in un determinato punto per unità di tempo, misurata in Hertz;
- ampiezza (A), che misura la distanza tra il punto massimo dell'onda e l'asse di propagazione (intensità);
- numero d'onda ($\bar{\nu}$), corrispondente al numero di onde che passano per una certa unità di lunghezza, misurata in cm^{-1}.

Lunghezza d'onda e frequenza sono inversamente proporzionali, e la relazione tra queste due grandezze può essere espressa tramite la formula:

$$\nu = \frac{c}{\lambda} \tag{3.1}$$

La frequenza della radiazione elettromagnetica corrisponde quindi al rapporto fra velocità della luce (*c*) e lunghezza d'onda (λ).

La teoria corpuscolare definisce la radiazione elettromagnetica come un flusso di fotoni (o quanti), ossia particelle prive di massa che si propagano in linea retta nello spazio a velocità costante (pari a *c* nel vuoto). I fotoni sono descritti da un'energia (*E*) direttamente proporzionale alla sua frequenza:

$$E = h\nu \tag{3.2}$$

$$E = hc/\lambda = hc\bar{\nu} \tag{3.3}$$

L'equazione di Planck (Equazione 3.2) quantifica l'energia associata a ciascun fotone, la quale è direttamente proporzionale alla frequenza della radiazione (ν) e inversamente proporzionale alla sua lunghezza d'onda (λ); a fini di calcolo viene introdotta la costante di Planck ($h = 6{,}626 \cdot 10^{-34}\ Js$). Inoltre, dato che l'energia di ciascun fotone resta invariata per una data lunghezza d'onda della radiazione elettromagnetica, l'intensità di un fascio di luce è definita dalla quantità di fotoni trasmessi nell'unità di tempo.

A dimostrazione delle relazioni precedenti, le radiazioni elettromagnetiche caratterizzate da lunghezza d'onda minore ed elevata frequenza sono quelle che presentano energia maggiore; esse sono potenzialmente in grado di rompere i legami chimici tra le molecole. Al contrario, le radiazioni elettromagnetiche caratterizzate da elevata lunghezza d'onda sono quelle che presentano frequenza ed energia minore.

Lo spettro elettromagnetico (Figura 3.2) contiene l'insieme di tutte le lunghezze d'onda della radiazione elettromagnetica presenti in natura: pertanto, può essere suddiviso in diversi intervalli in funzione di frequenza, lunghezza d'onda, numero d'onda o energia.

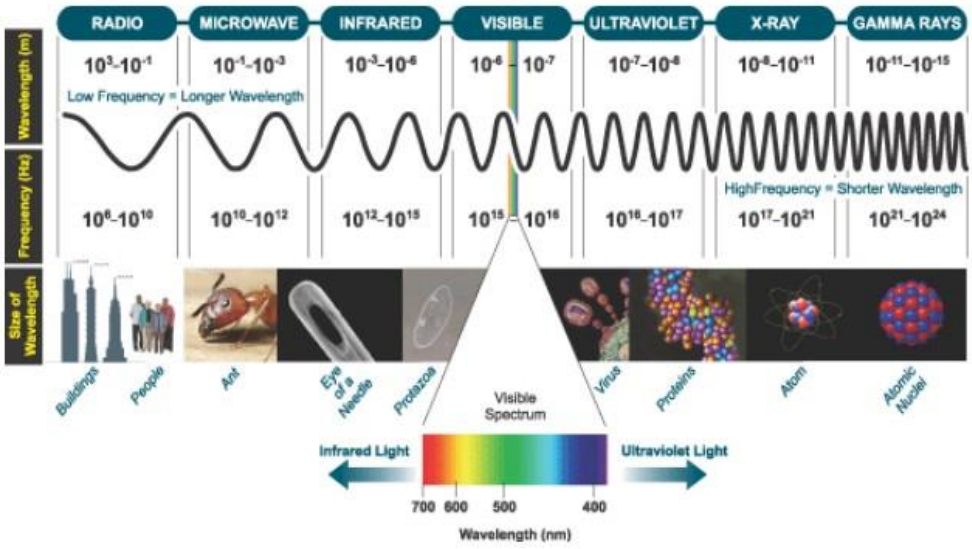

Figura 3.2 *Spettro elettromagnetico (fonte: https://slidetodoc.com/introduction-to-remote-sensing-department-of-rs-and/).*

L'interazione tra la radiazione elettromagnetica e la materia è un fenomeno quantico dipendente dalle proprietà della radiazione e dalla struttura della materia. Nel momento in cui la radiazione elettromagnetica colpisce gli

atomi di un oggetto, alcune frequenze possono essere assorbite da questi ultimi mentre altre vengono riflesse o tramesse a seconda delle caratteristiche dell'oggetto in esame.

Nel caso in cui si verifichi l'assorbimento della radiazione elettromagnetica, gli atomi della materia precedentemente allo stato energetico fondamentale acquistano l'energia necessaria per compiere il passaggio allo stato eccitato. L'energia della radiazione elettromagnetica corrisponde quindi all'energia necessaria affinché un elettrone compia il salto quantico dallo stato fondamentale allo stato eccitato. Dopo un certo periodo di tempo, l'energia acquistata viene emessa spontaneamente, sotto forma di radiazione elettromagnetica a energia inferiore (fluorescenza) e/o calore, e l'elettrone ritorna spontaneamente allo stato iniziale.

Generalmente, gli atomi si trovano allo stato fondamentale, dove occupano i livelli più bassi di energia, mentre allo stato eccitato gli atomi passano a livelli energetici superiori. I diversi livelli energetici di un atomo sono ben definiti e la transizione dallo stato fondamentale allo stato eccitato avviene soltanto se l'energia della radiazione elettromagnetica è uguale alla differenza di energia tra i livelli energetici. Pertanto, solo una radiazione caratterizzata da una specifica lunghezza d'onda ha la capacità di interagire con l'atomo affinché si verifichi il passaggio da stato fondamentale a stato eccitato.

L'interazione per le molecole risulta più complicata in quanto, oltre alle transizioni tra livelli elettronici, si verificano anche transizioni tra livelli vibrazionali e rotazionali. A causa dell'elevatissimo numero di possibili transizioni energetiche si osserva uno spettro di assorbimento continuo.

La spettroscopia sfrutta la specificità dell'interazione tra radiazione elettromagnetica e atomi o molecole per eseguire analisi qualitative e quantitative sulla composizione della materia. Considerando una sostanza che assorbe ad una certa lunghezza d'onda, il fascio della radiazione elettromagnetica in uscita avrà un'intensità minore rispetto alla radiazione incidente. Il rapporto tra radiazione in uscita e radiazione incidente prende il nome di trasmittanza (*T*):

$$T = \frac{I_\lambda}{I_{0,\lambda}} \tag{3.4}$$

dove I_λ corrisponde all'intensità della radiazione elettromagnetica in uscita mentre $I_{0,\lambda}$ all'intensità della radiazione elettromagnetica in entrata a una lunghezza d'onda definita. A seconda della natura della materia attraversata dalla radiazione, *T* può assumere valori diversi compresi tra 0 e 1: *T*=0 indica che la radiazione è stata assorbita completamente dal campione mentre *T*=1 indica che la radiazione non è stata assorbita . Grazie alla definizione di trasmittanza (*T*), può essere definito il valore di assorbanza (*A*) secondo la relazione:

$$A_\lambda = log\left(\frac{I_{0,\lambda}}{I_\lambda}\right) = -\,log\,T \tag{3.5}$$

Registrando i valori di assorbanza alle diverse lunghezze d'onda considerate si ottiene lo spettro di assorbimento del campione, che ha un andamento caratteristico in base alla natura chimica del campione stesso.

In ambito chimico, la misura dell'assorbanza (*A*) è particolarmente importante poiché, per ciascuna lunghezza d'onda, questa quantità è linearmente correlata con la concentrazione molare, secondo la legge di Lambert-Beer:

$$A_\lambda = \varepsilon_\lambda bc \tag{3.6}$$

dove ε_λ è l'assorbività molare (M^{-1} cm^{-1}), *b* è il cammino ottico (cm) e *c* è la concentrazione molare (M).

La spettroscopia infrarossa si riferisce alla regione dello spettro elettromagnetico che comprende le lunghezze d'onda da 780 nm e 1000 mm. Tale regione può essere ulteriormente suddivisa in vicino infrarosso (NIR), che comprende l'intervallo tra 780 nm e 2500 nm, medio infrarosso (MIR), che comprende l'intervallo tra 2500 nm e 25 μm e il lontano infrarosso (FIR) che comprende l'intervallo tra 25 μm e 1000 μm. È da notare che esiste una certa variabilità nella definizione esatta di tali intervalli; un esempio viene riportato in Figura 3.3.

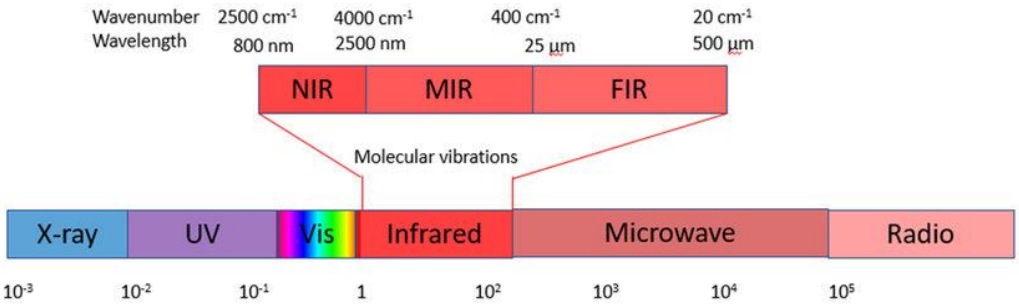

Figura 3.3 *Spettro infrarosso (Fox, 2020).*

La spettroscopia infrarossa (IR) è una spettroscopia vibrazionale: a differenza delle radiazioni UV-Vis che sono dotate di energia sufficiente per provocare transizioni elettroniche, la radiazione infrarossa riesce ad indurre solo variazioni nel moto vibrazionale e rotazionale della molecola. La radiazione infrarossa, nel caso in cui i gruppi funzionali di una molecola possiedano un dipolo, amplifica le oscillazioni che avvengono naturalmente in relazione alle variazioni delle lunghezze e degli angoli di legame.

In particolare, l'energia ceduta dalla radiazione NIR agisce sui moti vibrazionali dei legami covalenti tra gli atomi che formano le molecole. Si possono descrivere due tipologie di moti vibrazionali causati dalla radiazione elettromagnetica nel vicino infrarosso:

- *Stretching* - prevede una sorta di "stiramento" ritmico lungo l'asse di legame che comporta una variazione della distanza interatomica, la quale può aumentare o ridursi. Tale variazione può essere simmetrica, nel caso in cui i diversi legami si allunghino o accorcino contemporaneamente, oppure asimmetrica, nel caso un legame si accorci e l'altro si allunghi.
- *Bending* – vibrazione con conseguente variazione dell'angolo di legame che può essere simmetrica o asimmetrica e può avvenire sul piano del legame stesso oppure al di fuori del piano. Di conseguenza, si

possono distinguere quattro tipologie di *bending*: *scissoring*, variazione simmetrica dell'angolo di legame nel piano, *rocking*, variazione asimmetrica dell'angolo di legame nel piano, *wagging*, la variazione simmetrica dell'angolo di legame al di fuori del piano, e *twisting*, la variazione asimmetrica dell'angolo di legame al di fuori del piano.

La frequenza vibrazionale, dovuta all'assorbimento di radiazione infrarossa da parte di uno specifico gruppo funzionale di una molecola di interesse, fa riferimento alla legge fisica di Hooke. L'applicazione della legge di Hooke consente di descrivere un generico legame chimico come una molla in cui le masse oscillanti sono gli atomi interessati nel legame. A causa delle differenze a livello di massa e tipologia di legami tra gli atomi, l'assorbimento di energia da parte della molecola può subire delle variazioni. In condizioni reali, però, si verificano diverse deviazioni rispetto al modello armonico descritto in precedenza; per questi motivi è stato delineato un modello anarmonico.

Le bande di assorbimento che si osservano nella regione del NIR sono principalmente bande armoniche (*overtone*) e bande di combinazione. Gli *overtone* prevedono una transizione energetica a livelli superiori rispetto al primo livello energetico garantendo, di conseguenza, una cessione di energia multipla rispetto a quella caratteristica delle vibrazioni fondamentali. Le bande di combinazione, invece, si verificano quando l'energia assorbita dalla molecola provoca contemporaneamente la vibrazione di più legami interatomici adiacenti. Le bande di combinazione possono derivare dalla concomitanza di due modi vibrazionali fondamentali oppure da vibrazioni fondamentali e *overtone*.

La tipologia dei legami interessati nelle transizioni vibrazionali che avvengono nella regione del vicino infrarosso sono i principali legami chimici che interessano l'idrogeno (X–H), ossia i legami C−H, O−H, N−H e S−H, oltre ad altri legami quali C=O, C−Cl, C−N, C−O−O, C−N−C. Tutte le molecole contenenti idrogeno presentano uno spettro NIR misurabile, pertanto, è possibile utilizzare questa tecnica per analizzare una vasta gamma di molecole organiche.

La spettroscopia NIR offre diversi vantaggi rispetto ai metodi tradizionali utilizzati per l'analisi nel settore agroalimentare. La sua applicazione permette di effettuare analisi rapide, non distruttive e di limitare notevolmente la fase di preparazione dei campioni. Inoltre, grazie alla facilità d'uso, consente l'utilizzo da parte di operatori non specializzati. Grazie ai suoi pregi, la spettroscopia NIR si è affermata come metodo efficace per l'analisi qualitativa e quantitativa in campo agroalimentare, soprattutto nell'ambito del controllo qualità.

Per contro, a causa della presenza di bande armoniche e di combinazione, oltre che alla vasta gamma di possibili vibrazioni, gli spettri NIR sono costituiti da bande larghe e sovrapposte fra loro, caratteristica che li rende difficilmente interpretabili.

Per risolvere questi limiti è necessario adottare un approccio chemiometrico: esso consente l'estrazione dell'informazione utile contenuta negli spettri risolvendone la ridondanza e permette un'interpretazione visiva semplificata.

3.2 Imaging iperspettrale

L'imaging iperspettrale è stato sviluppato intorno agli anni '70 e le prime applicazioni prevedevano l'acquisizione di immagini satellitari. Soltanto dopo diverso tempo questa tecnica ha trovato impiego in numerosi campi, come ad esempio il settore farmaceutico, il settore agro-alimentare, l'ambito biochimico e biomedico, e tanti altri.

L'imaging iperspettrale (*Hyperspectral Imaging*, HSI) unisce i benefici delle tecniche spettroscopiche con quelli delle tecniche di imaging, dando la possibilità di ottenere informazioni sia spaziali che spettrali del campione in esame. In particolare, ciascun pixel di un'immagine iperspettrale contiene uno spettro completo acquisito in una determinata regione dello spettro elettromagnetico, come ad esempio la regione del NIR. L'imaging iperspettrale nel vicino infrarosso offre gli stessi vantaggi della spettroscopia NIR, ossia velocità e facilità di esecuzione, limitata fase di preparazione del campione e possibilità di riutilizzare quest'ultimo, essendo una tecnica non distruttiva. Tuttavia, le tecniche spettroscopiche classiche consentono l'acquisizione dello spettro in una porzione limitata del campione, mentre l'imaging iperspettrale consente di ottenere informazioni relative non solo alla composizione chimica del campione ma anche alla sua variazione sulla superficie del campione stesso.

Tale vantaggio rende questa tecnica ottimale per l'analisi di campioni eterogenei, come le matrici alimentari. Per questi motivi, l'imaging iperspettrale trova applicazioni nei sistemi industriali automatizzati al fine di effettuare un costante monitoraggio del sistema produttivo. Inoltre, queste tecniche si stanno diffondendo anche nell'ambito dell'agricoltura di precisione, che prevede l'utilizzo in campo di sistemi iperspettrali e multispettrali per valutare in tempo reale lo stato qualitativo e fitosanitario delle coltivazioni.

Nonostante ciò, l'imaging iperspettrale è affetto da due principali svantaggi. Infatti, l'acquisizione di immagini iperspettrali prevede l'elaborazione di una grande quantità di dati da cui conseguono difficoltà computazionali e di estrazione delle informazioni utili. Pertanto, l'analisi di immagini iperspettrali non può prescindere dall'applicazione di metodi chemiometrici finalizzati all'estrazione dell'informazione utile in tempi brevi.

Le immagini iperspettrali vengono dette anche *ipercubi* in quanto sono *array* tridimensionali di dati di dimensioni $\{x, y, \lambda\}$, caratterizzate da due dimensioni spaziali $\{x, y\}$, ovvero righe e colonne di pixel, e da una dimensione spettrale, λ. Questa struttura può essere interpretata come una serie d'immagini sovrapposte, ciascuna delle quali è stata acquisita ad una determinata lunghezza d'onda (λ), oppure come un insieme di spettri, ciascuno con la sua posizione precisa all'interno dell'immagine (Figura 3.4).

Ciascun pixel di una determinata immagine iperspettrale corrisponde ad uno spettro, il quale, a sua volta, può essere descritto dalla somma degli spettri dei composti puri presenti in quella specifica posizione del campione e dal rumore. Nonostante ciò, è importante non considerare un pixel come un oggetto singolo, poiché affetto negativamente da diversi fattori:

- *eterogeneità e forma della superficie*: difficilmente è possibile ottenere superfici perfettamente lisce e ciò influenza l'intensità della luce riflessa dal pixel poi trasmessa al detector;
- *profondità di penetrazione della radiazione*: la maggior parte della radiazione spettrale che riceve il detector riguarda la superficie del campione. Tuttavia, a causa della profondità di penetrazione della spettroscopia vibrazionale (NIR), la radiazione è in parte influenzata dagli strati adiacenti più interni del campione, che possono modificare l'assorbanza del pixel.
- *risoluzione*: generalmente un pixel contiene informazioni che appartengono a diversi composti chimici presenti nell'area del pixel stesso.

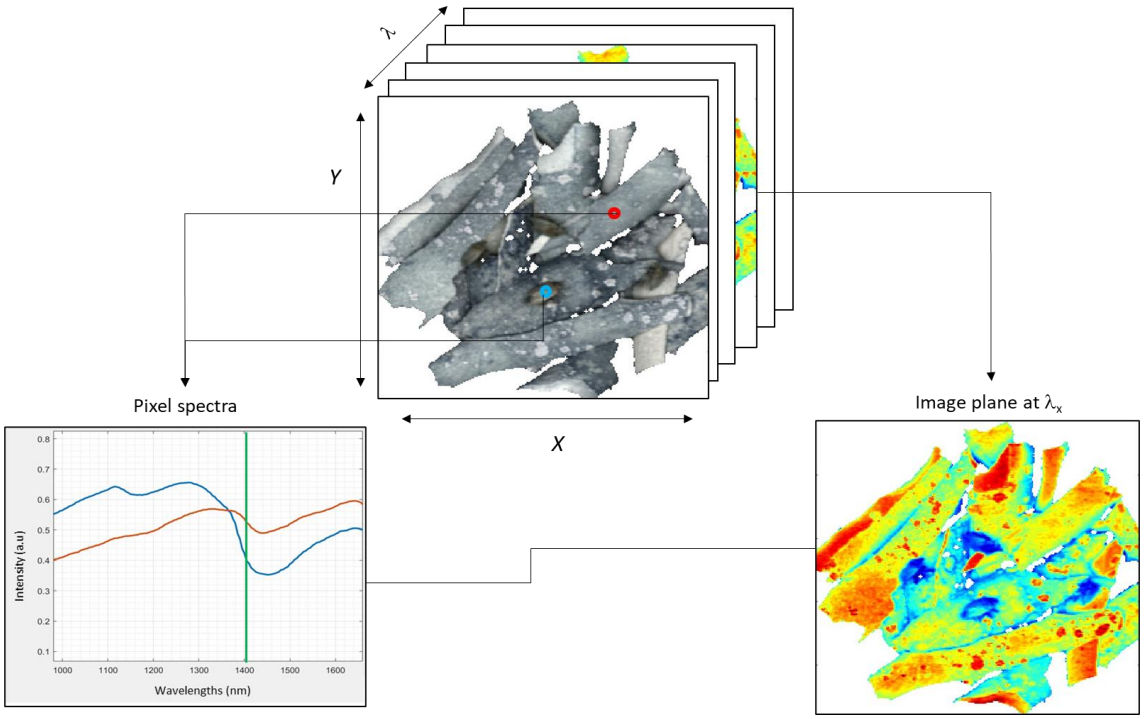

Figura 3.4 *Struttura di un'immagine iperspettrale.*

3.2.1 Metodi di acquisizione delle immagini iperspettrali

L'acquisizione degli ipercubi può essere eseguita utilizzando strumenti con diversa configurazione, tra cui *point scanning, line scanning, area scanning* e *single shot* (Figura 3.5).

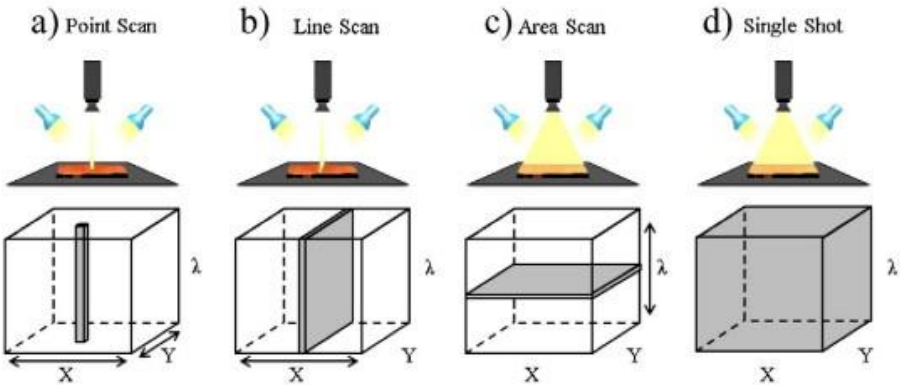

Figura 3.5 *Metodi di acquisizione delle immagini iperspettrali (Wu & Sun, 2013).*

Point scanning

La configurazione *point scanning*, detta anche *whiskerbroom*, consiste nell'acquisizione dello spettro di un pixel alla volta; l'acquisizione dell'intera superficie da analizzare avviene tramite lo spostamento del campione o del rivelatore nelle due direzioni spaziali *x* e *y*. Il vantaggio principale offerto dal metodo *point scanning* risiede nella possibilità di ottenere immagini con un'ottima risoluzione spaziale, mentre risulta svantaggioso a causa dei lunghi tempi di riposizionamento del campione o del detector, necessari al fine di garantire la ripetibilità dell'analisi.

Line scanning

La configurazione *line scanning*, detta anche *pushbroom*, consente di acquisire contemporaneamente gli spettri corrispondenti ad un'intera riga di pixel del campione. In questo modo, grazie al movimento del campione o del detector lungo una sola direzione spaziale, è possibile ottenere un'immagine iperspettrale completa. Questo metodo consente quindi di ridurre i tempi di acquisizione rispetto alla configurazione *point scanning*: non a caso è il metodo più utilizzato. Inoltre, è compatibile con sistemi di monitoraggio per il controllo qualità a livello industriale, essendo facilmente applicabile per l'analisi di prodotti su nastri trasportatori. Il principale svantaggio di questa tecnica riguarda il tempo di esposizione, che deve essere impostato ad uno stesso valore per tutte le lunghezze d'onda. Il tempo di esposizione rappresenta un punto cruciale, poiché deve essere sufficientemente ridotto a tutte le lunghezze d'onda per evitare la saturazione degli spettri, ma allo stesso tempo deve essere tale da garantire una misura spettrale sufficientemente accurata.

Il principio di funzionamento di un sistema *line scanning* prevede una sorgente luminosa, come ad esempio una lampada alogena, che illumina il campione. La radiazione della luce riflessa da parte di una sottile riga del campione viene raccolta dall'obiettivo e ogni pixel della riga viene separata nelle diverse lunghezze d'onda grazie ad un sistema di dispersione prisma-reticolo-prisma (*Prism-Grating-Prism*, PGP) mostrata in Figura 3.6. Infine, la radiazione dispersa viene proiettata sul rivelatore, il quale è costituito da un array bidimensionale di sensori in cui una dimensione rappresenta l'informazione spettrale (λ) e l'altra dimensione rappresenta l'informazione spaziale, ossia la posizione del pixel nella riga.

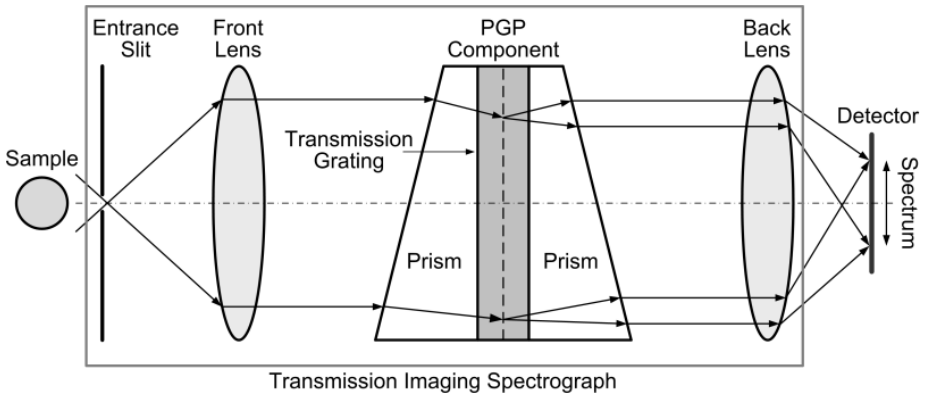

Figura 3.6 *Sistema di dispersione prisma-reticolo-prisma (PGP) (Qin et al., 2017).*

Area scanning

La configurazione *area scanning* prevede l'acquisizione di un'immagine monocromatica alla volta per ciascuna lunghezza d'onda; in seguito, le immagini acquisite per ciascuna lunghezza d'onda vengono unite a formare l'*ipercubo*. In questo caso, viene mantenuto fisso il campo visivo dell'immagine e, dato che il rivelatore è esposto volta per volta ad una lunghezza d'onda differente, è possibile impostare il tempo di esposizione ottimale per ogni lunghezza d'onda. Tuttavia, a causa della necessità di un campo visivo fisso, questo metodo risulta incompatibile per l'analisi di oggetti in movimento e per il controllo qualità in tempo reale.

Single shot

Il metodo *single shot* consente di registrare contemporaneamente le informazioni spaziali e spettrali tramite l'utilizzo di un detector in grado di rilevare un'ampia area e dotato di un'esposizione tale da permettere l'acquisizione delle immagini. Tali caratteristiche rendono il metodo idoneo all'acquisizione di immagini iperspettrali in contesti produttivi industriali. Tuttavia, per il momento, è una tecnica in fase di sviluppo iniziale, che può offrire soltanto risoluzioni limitate e range spettrali piuttosto ristretti.

3.3 Analisi multivariata delle immagini iperspettrali

Il primo passaggio nell'elaborazione delle immagini iperspettrali consiste nella trasformazione dell'immagine *raw*, in cui i dati sono espressi come conteggi di intensità rilevati dal detector, in un'immagine in cui i valori sono espressi in riflettanza. A tal fine, è necessario effettuare una calibrazione della riflettanza, operazione generalmente effettuata in automatico dai software degli strumenti di acquisizione delle immagini tramite la misurazione di un "bianco", riferimento standard con un elevato valore noto di riflettanza (ad es. spectralon o materiale ceramico), e la misurazione della corrente oscura, ovvero il rumore strumentale associato al detector che viene ottenuto durante l'acquisizione di un'immagine dopo aver oscurato l'obiettivo della telecamera. I valori del riferimento ad elevata riflettanza e della corrente oscura vengono utilizzati per calcolare i valori di riflettanza dell'immagine iperspettrale come riportato nell'equazione 3.7:

$$R = \frac{(I - I_d)}{(I_0 - I_d)} \tag{3.7}$$

dove R è il valore di riflettanza, I è il valore dell'intensità misurata dallo strumento, I_d è l'intensità della corrente oscura mentre I_0 è l'intensità misurata sul riferimento altamente riflettente.

L'analisi multivariata delle immagini (*Multivariate Image Analysis*, MIA) consiste nell'applicazione di tecniche chemiometriche per estrarre l'informazione utile dalle immagini al fine di valutare similitudini e differenze tra i pixel, quantificare o classificare regioni di interesse (*Regions of Interest*, ROI) e ridurre la dimensionalità dell'ipercubo.

Il primo passaggio di MIA è rappresentato dall'*unfolding*, che consiste nel trasformare la matrice tridimensionale dell'ipercubo in una matrice bidimensionale di dimensioni $\{(x \times y), \lambda\}$, dove il numero di righe corrisponde al numero totale di pixel dell'immagine, mentre le colonne corrispondono ai canali spettrali. Pertanto, nella matrice *unfolded* ogni riga rappresenta lo spettro associato ad un pixel (Figura 3.7).

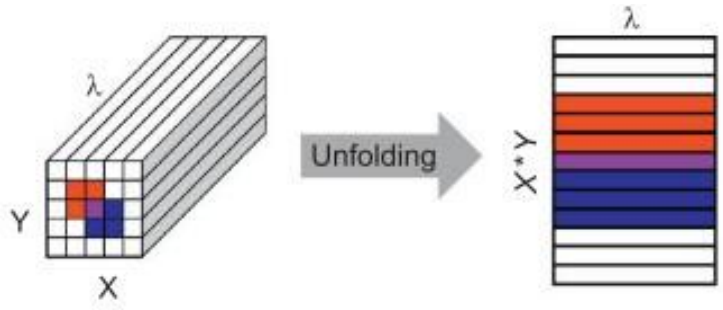

Figura 3.7 *Ipercubo sottoposto ad unfolding e a successiva analisi PCA (Gowen., 2013).*

Successivamente, dopo aver ottenuto la matrice *unfolded*, è possibile pretrattare i segnali spettrali utilizzando i pretrattamenti per riga e per colonna. In particolare, i pretrattamenti di riga sono molto utilizzati per l'elaborazione di dati di natura spettrale in quanto permettono di ridurre o rimuovere fenomeni indesiderati

come lo *scattering*, effetto dovuto all'eterogeneità della superficie dei campioni, come precedentemente descritto in Sezione 2.1.

3.3.1 Analisi esplorativa di immagini iperspettrali

L'analisi delle componenti principali (PCA, Sezione 2.2) è una delle tecniche chemiometriche maggiormente utilizzate per effettuare l'analisi esplorativa delle immagini iperspettrali, consentendo di individuare le principali fonti di variabilità e visualizzare la distribuzione qualitativa degli elementi su una singola immagine; ciò rende PCA uno strumento di analisi versatile ed essenziale per l'analisi preliminare dei dati.

In un modello PCA calcolato a partire da un'immagine iperspettrale (Figura 3.8) la matrice degli score identifica la posizione di ogni pixel su ciascuna componente principale, mentre la matrice dei loading esprime l'importanza delle variabili originali, ovvero delle lunghezze d'onda considerate, nel definire le componenti principali. In aggiunta, è possibile trasformare ciascun vettore degli score nella corrispondente immagine degli score (*score image*) per visualizzare i risultati di PCA nel dominio spaziale dell'immagine. Le score image vengono solitamente rappresentate sotto forma di immagini in scala di grigi (o pseudo-colori), in cui ogni pixel viene rappresentato da un colore che dipende dal suo valore di score per la PC considerata.

Inoltre, è possibile selezionare un gruppo di pixel nel grafico degli score visualizzando gli stessi nella score image e, viceversa, è possibile selezionare una zona della score image visualizzando i pixel corrispondenti nello score plot. Tale operazione prende il nome di *brushing* e può essere utilizzata al fine di eliminare manualmente i pixel dello sfondo che non devono rientrare nell'analisi del campione oppure per la selezione delle ROI (Figura 3.9).

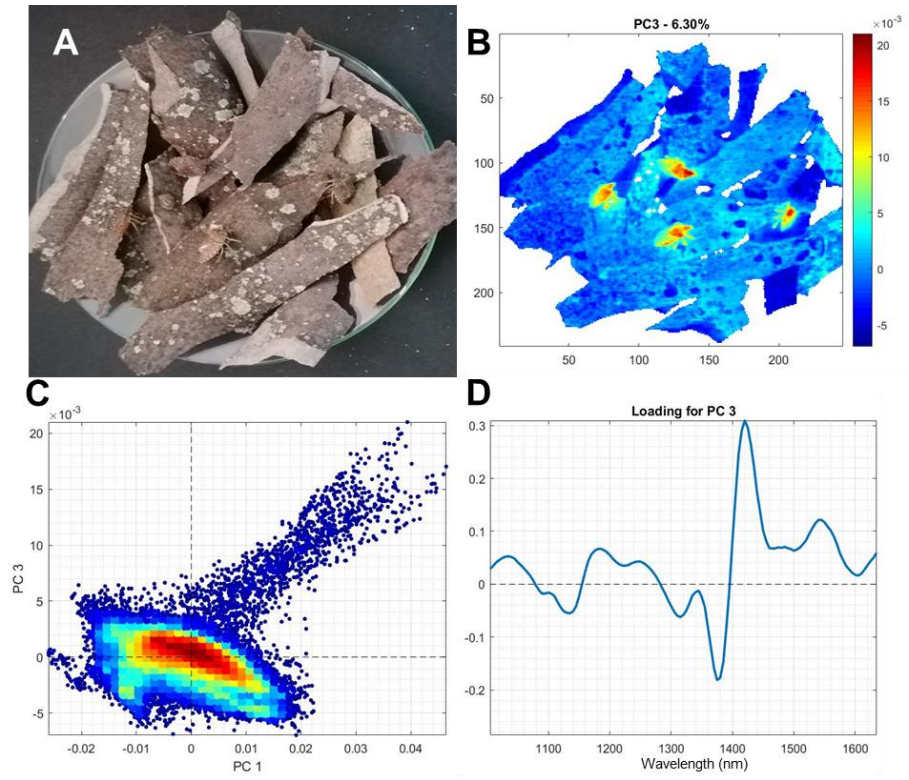

Figura 3.8 *Esempio di PCA effettuata su un'immagine del dataset (Corteccia_HH_G1_1a) di cui viene riportata l'immagine RGB del campione (A), la* score *image di PC3 (B),* score plot *di PC1 e PC3 (C) e* loading plot *di PC3 (D). I colori riportati nella* score *image di PC3 corrispondono ai valori di* score *di PC3 dei pixel mentre i colori riportati nello* score plot *si riferiscono alla densità dei pixel).*

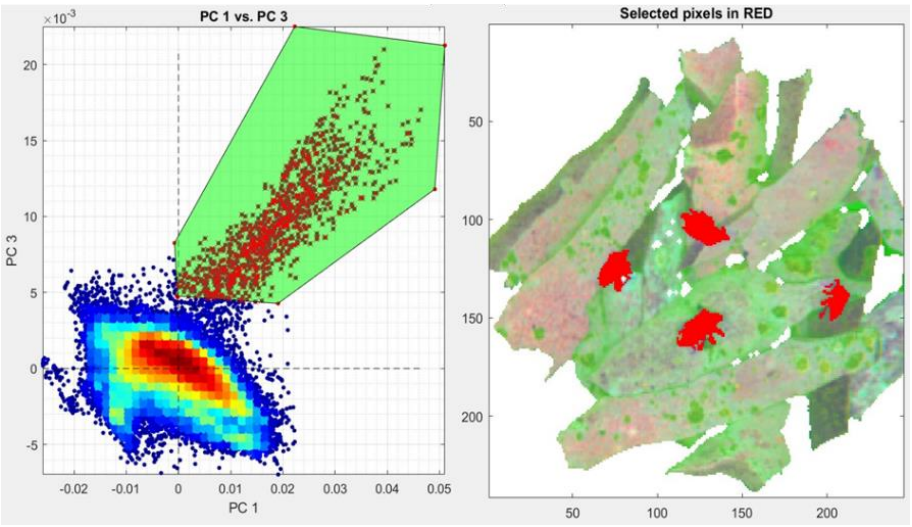

Figura 3.9 *Esempio di brushing: tramite la selezione di un cluster di pixel nello* score plot *PC1/PC3 è possibile effettuare una maschera che permette di isolare i pixel di Cimice asiatica dallo sfondo vegetale e viceversa.*

3.3.2 Sviluppo di modelli di classificazione a partire da immagini iperspettrali

Le tecniche chemiometriche possono essere applicate all'analisi delle immagini anche per lo sviluppo di modelli di classificazione, come precedentemente descritto in Sezione 2.3. A questo scopo, per prima cosa è necessario ottenere, a partire dalle immagini, un dataset di spettri di attribuzione nota appartenenti alle classi considerate e il più possibile rappresentativi della variabilità delle classi stesse. Il dataset di spettri rappresentativi di ciascuna classe può essere ottenuto andando a selezionare i pixel sulle immagini delle ROI corrispondenti alle diverse classi e calcolando per ciascuna ROI lo spettro medio oppure selezionando un numero limitato di spettri, che possono essere scelti in maniera casuale o utilizzando opportuni algoritmi capaci di identificare gli spettri più rappresentativi della ROI (ad es., algoritmo di Kennard-Stone descritto in Sezione 4.4). Il dataset così creato può essere suddiviso successivamente in training set (TR set), per calcolare il modello di calibrazione, e test set (TS set) al fine di effettuare la validazione esterna del modello.

La validazione del modello può essere effettuata anche utilizzando le intere immagini iperspettrali e andando ad applicare il modello su ciascun pixel dell'immagine stessa, valutando l'attribuzione alle classi considerate. In questo caso, è possibile visualizzare nel dominio spaziale dell'immagine originale il risultato della predizione effettuata da parte del modello di classificazione utilizzando la corrispondente immagine in predizione (*prediction image*), in cui ogni pixel viene rappresentato con un colore che corrisponde all'assegnazione effettuata dal modello di classificazione di quel determinato pixel ad una certa classe (Figura 3.10).

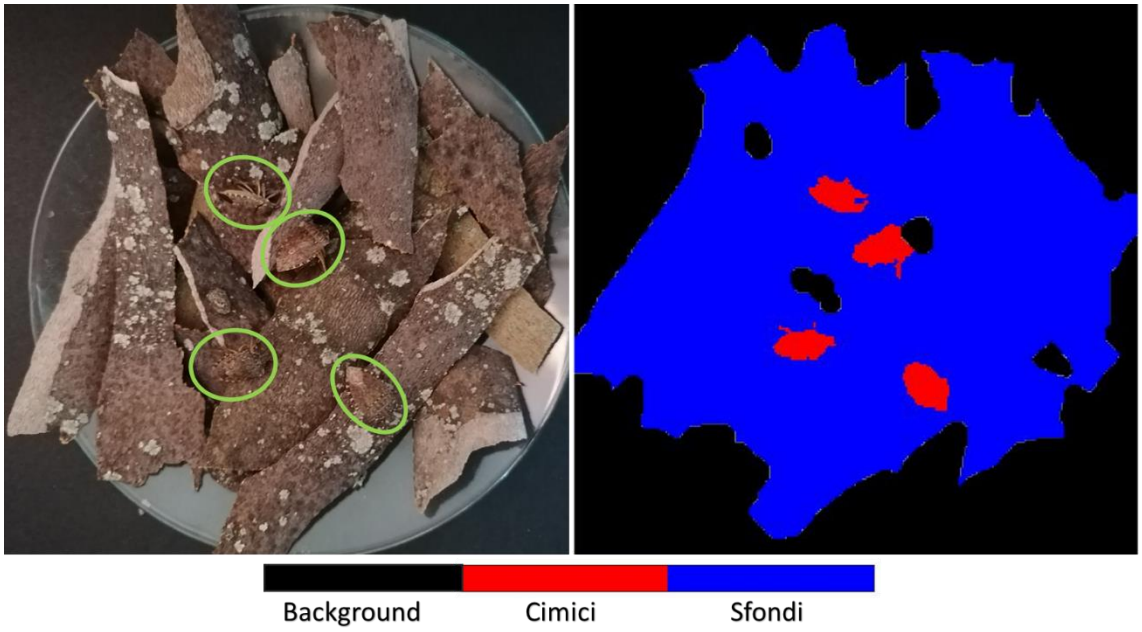

Figura 3.10 *Esempio di prediction image ottenuta tramite l'applicazione di un modello PLS-DA e corrispondente immagine RGB.*

4 Materiali e metodi

4.1 Descrizione e raccolta dei campioni

La sperimentazione del presente lavoro di tesi può essere suddivisa in due fasi principali:

- **Fase I**: acquisizione di un dataset preliminare per una prima valutazione del potenziale utilizzo del NIR-HSI per identificare *H. halys* su diversi sfondi vegetali;
- **Fase II**: acquisizione di un dataset esteso, ottenuto considerando un numero maggiore di campioni al fine di aumentare la rappresentatività dei dati acquisiti e ottimizzare la fase di implementazione dei modelli di classificazione tramite PLS-DA, Soft PLS-DA e *sparse* Soft PLS-DA.

Gli esemplari di *H. halys* analizzati sono stati forniti dal laboratorio di entomologia applicata coordinato dalla Prof.ssa Maistrello. I campioni degli sfondi vegetali sono stati raccolti nei pressi della sede universitaria (padiglione Besta - Via Amendola, 2, Località San Maurizio, Reggio Emilia) nei giorni in cui sono state effettuate le acquisizioni, ossia, in data 29/04/2021 per la Fase I e in data 01/07/2021 per la Fase II.

4.1.1 Fase I

Per lo studio di fattibilità è stato acquisito un dataset preliminare di dimensioni ridotte. Tale passaggio è stato effettuato per valutare le potenzialità del metodo considerato e definire le procedure da applicare nelle fasi successive. I campioni considerati sono composti da due gruppi di cimici differenti e otto diversi sfondi vegetali, tra cui corteccia, erba, foglie gialle, foglie secche, foglie verdi, rami, terra e uno sfondo misto che comprende i precedenti, al fine di simulare le condizioni in campo. Gli esemplari di cimice considerati sono stati denominati:

- **HH_1**: nel caso di esemplari deceduti da diversi giorni. In questo caso gli esemplari di *H. halys* sono stati consegnati dal laboratorio di entomologia già deceduti;
- **HH_2**: nel caso di cimici decedute da poche ore. In questo caso gli esemplari di *H. halys* sono stati consegnati ancora vivi dal laboratorio di entomologia e sono stati messi in freezer ad una temperatura di - 18°C per circa 30 minuti, per essere poi riportati a temperatura ambiente prima dell'analisi.

4.1.2 Fase II

Al fine di implementare e ottimizzare i modelli di classificazione per discriminare *H. halys* dagli sfondi vegetali, sono stati considerati campioni contenenti soltanto cimici decedute da poche ore e otto diversi sfondi vegetali, tra cui corteccia, erba, foglie gialle, foglie secche, foglie verdi, rami, terra e uno sfondo misto, al fine di simulare le condizioni in campo. Per il secondo campionamento sono stati consegnati dal gruppo di entomologia 20 esemplari di *H. halys* ancora vivi, che sono stati messi in freezer ad una temperatura di - 18°C

per circa 30 minuti prima e quindi riportati a temperatura ambiente dell'analisi. Le cimici sono state poi suddivise in cinque gruppi (G1, G2, G3, G4 e G5), ciascuno composto da quattro esemplari differenti. Ciascun gruppo è stato impiegato per l'acquisizione di immagini iperspettrali sui diversi sfondi vegetali. In questa fase, la scelta di considerare solamente cimici morte da poche ore (HH_2) è stata effettuata sulla base dei risultati preliminari ottenuti dalle immagini acquisite in Fase I, che hanno permesso di evidenziare differenze nella risposta spettrale tra esemplari morti da poche ore ed esemplari morti da qualche giorno, probabilmente a causa di modificazione chimiche e fisiche di questi ultimi campioni.

4.2 Descrizione della strumentazione utilizzata e procedura di acquisizione

L'acquisizione delle immagini iperspettrali è stata effettuata utilizzando un sistema iperspettrale di tipo *line scanning* NIR Desktop Spectral Scanner (DV Optics), costituito da uno spettrometro Specim N17E accoppiato ad una telecamera Xenics XEVA-1.7-320 (320 × 256 pixel), che lavora nel range spettrale compreso tra 955 e 1700 nm, con una risoluzione di 5 nm per un totale di 150 variabili spettrali. L'acquisizione delle immagini è avvenuta utilizzando uno sfondo di carta vetrata nera, la quale è caratterizzata da uno spettro con valori molto bassi e costanti di riflettanza. Inoltre, nel campo dell'immagine erano presenti anche uno standard bianco certificato ad elevato valore di riflettanza, costituito da una piastrella bianca di ceramica, e due piastrelle di ceramica con valori di riflettanza intermedi.

La calibrazione delle immagini è stata effettuata automaticamente dal software di acquisizione basandosi sul segnale della corrente oscura e sul segnale del riferimento altamente riflettente. Inoltre, al fine di ridurre la variabilità tra le immagini acquisite, è stata eseguita una calibrazione interna ulteriore, basata sui valori medi di riflettanza dello standard bianco, dello sfondo nero di carta vetrata e delle piastrelle di ceramica con valori intermedi di riflettanza.

In entrambe le fasi sperimentali, per effettuare l'acquisizione delle immagini iperspettrali i campioni sono stati allestiti su vetrini da orologio (*watch glass*) sui quali sono stati posizionati i diversi sfondi vegetali e le cimici. Inoltre, per ogni campione è stata acquisita la corrispondente immagine RGB in formato .jpg utilizzando la fotocamera digitale di uno smartphone.

4.2.1 Fase I

Al fine di effettuare uno studio preliminare sono state acquisite le immagini iperspettrali di due tipologie di campioni: la prima tipologia presenta cimici HH_1 mentre l'altra tipologia presenta cimici HH_2. Per entrambe le tipologie sono stati considerati otto sfondi vegetali differenti (corteccia, erba, foglie gialle, foglie secche, foglie verdi, rami, terra e uno sfondo misto) e per i campioni contenenti HH_1 sono state acquisite due repliche. Pertanto, nella Fase I sono state acquisite in totale 24 immagini iperspettrali, corrispondenti a 16 immagini con le cimici della tipologia HH_1 e 8 immagini delle cimici della tipologia HH_2 (Tabella 4.1).

Nella fase di calcolo dei modelli di classificazione non sono state considerate le immagini acquisite utilizzando come sfondo foglie gialle e lo sfondo misto, e neppure tutte le immagini della seconda replica (replica b) delle cimici del gruppo HH_1. Pertanto, il calcolo dei modelli di classificazione in Fase I è stata effettuata considerando 12 immagini, mentre le immagini escluse sono state utilizzate successivamente per effettuare la validazione esterna del modello tramite visualizzazione delle immagini in predizione.

n° immagine	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sfondo	Corteccia			Erba			Foglie Gialle			Foglie Secche			Foglie Verdi			Rami			Terra			Misto		
Cimice	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2	HH_1		HH_2
Replica	a	b	-	a	b	-	a	b	-	a	b	-	a	b	-	a	b	-	a	b	-	a	b	-

Tabella 4.1 *Immagini iperspettrali acquisite nella Fase I.*

Le immagini sono state codificate in modo tale da identificare lo sfondo e la categoria di cimici considerate per la costituzione del campione analizzato e l'eventuale replica (ad es., *Corteccia_HH_1_a* è il nome dell'immagine acquisita su cimici decedute da diversi giorni, sfondo corteccia e prima replica).

4.2.2 Fase II

Al fine di acquisire un dataset utile all'implementazione e all'ottimizzazione dei modelli di classificazione atti a discriminare *H. halys* dagli sfondi vegetali, è stato quindi acquisito un secondo dataset di immagini iperspettrali più ampio del precedente.

In questo caso, per ognuno degli otto sfondi vegetali considerati sono stati acquisiti cinque campioni, uno per ciascuno dei cinque gruppi di cimice. Per ogni campione sono state acquisite due immagini replicate, ciascuna ruotata di 180° rispetto all'altra. In questa fase sono state ottenute quindi 80 immagini iperspettrali (= 8 tipologie di sfondo × 5 gruppi di *H. halys* × 2 repliche). Anche in questo caso per ogni campione sono state acquisite anche le corrispondenti immagini RGB in formato .jpg utilizzando la fotocamera di uno smartphone.

Per motivi computazionali, in questa tesi è stato analizzato un sottoinsieme delle immagini acquisite nella Fase II, in particolare sono state considerate solamente le immagini corrispondenti alla prima replica degli sfondi corteccia, erba, foglie gialle, foglie secche, foglie verdi, rami e terra, per un totale di 35 immagini (Tabella 4.2). Le immagini dello sfondo misto e della seconda replica non sono state considerate per la creazione dei modelli ma soltanto per effettuare la validazione esterna tramite creazione di immagini in predizione.

n° immagine	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Sfondo	Corteccia					Erba					Foglie Gialle					Foglie Secche					Foglie Verdi					Rami					Terra				
Gruppo cimici	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5

Tabella 4.2 *Immagini iperspettrali acquisite nella Fase II ed utilizzate nella successiva fase di calcolo dei modelli di classificazione*.

Le immagini sono state codificate in modo tale da identificare lo sfondo, il gruppo di cimici di riferimento e la replica per la costituzione del campione analizzato (ad es., *Corteccia_HH_G1_a* è il nome dell'immagine acquisita su cimici del gruppo 1, su corteccia, prima replica).

4.3 Analisi ed elaborazione delle immagini

Le immagini iperspettrali acquisite sono convertite in file in formato .mat al fine di effettuare le procedure di analisi in ambiente MATLAB (R2020b, The MathWorks Inc., USA). Durante l'importazione in MATLAB delle immagini iperspettrali sono state escluse le zone iniziali e finali del range spettrale poiché affette da bassi valori del rapporto segnale/rumore. Pertanto, le immagini iperspettrali analizzate considerano un range spettrale ridotto, compreso tra 980 e 1660 nm, per un totale di 137 variabili spettrali.

In primo luogo, prima di effettuare l'analisi esplorativa dei dati è stati necessario rimuovere da ciascuna immagine i pixel dello sfondo nero e del porta-campione. Tale passaggio è stato eseguito grazie alla funzione *masking* del software HYPER-Tools (version 3.0, https://www.hypertools.org), il quale consente l'eliminazione dei pixel tramite la selezione di un valore soglia sui valori di riflettanza ad una determinata lunghezza d'onda. L'eliminazione dello sfondo nero è stata eseguita eliminando dalle immagini i pixel con un valore di assorbanza inferiore a 0,35 unità di riflettanza alla lunghezza d'onda di 1000 nm. I pixel relativi allo sfondo nero e al vetrino porta-campione residui sono stati rimossi tramite un'ulteriore procedura di *masking* effettuata manualmente utilizzando la funzione *erosion* disponibile sull'interfaccia dello stesso software.

Successivamente, è stata effettuata un'analisi preliminare di ciascuna immagine tramite PCA. Per ciascuna immagine sono stati calcolati dei modelli PCA considerando quattro diversi pretrattamenti di riga (SNV, detrend, derivata prima e derivata seconda) e mean centering come pretrattamento per colonna. Tra i modelli PCA calcolati, sono stati considerati soltanto quelli che permettevano una migliore segregazione dei pixel delle cimici rispetto ai pixel degli sfondi. Per ogni immagine, il modello PCA migliore è stato sfruttato per identificare i pixel relativi alle cimici e agli sfondi mediante *brushing* (Sezione 3.1.3). Grazie a questa operazione, per ogni immagine è stato possibile ottenere due maschere, una per identificare i pixel di *H. halys* ed una per identificare i pixel dello sfondo vegetale (Figura 4.1). Inoltre, la selezione delle diverse maschere è stata ulteriormente rifinita manualmente utilizzando le funzioni *erosion*, *open*, *close*, *filling holes* ed *elimination of regions* disponibili sull'interfaccia di HYPER-Tools.

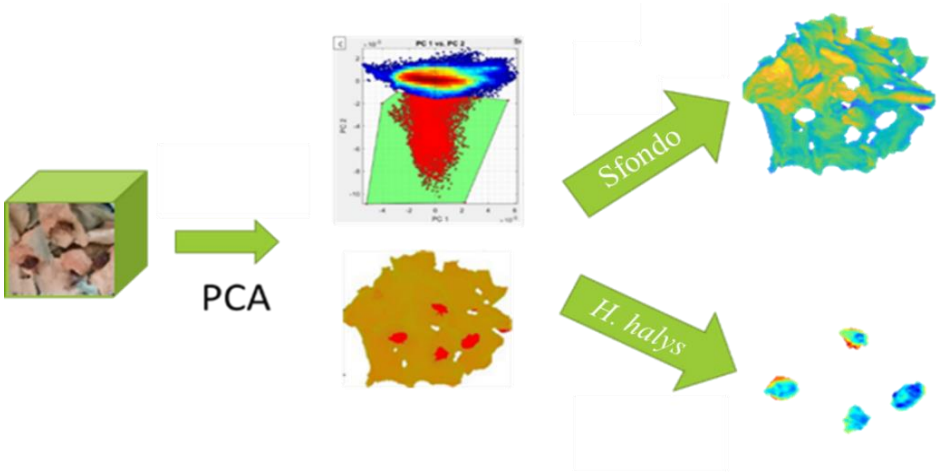

Figura 4.1 *Fasi del passaggio dalla matrice tridimensionale (ipercubo) all'ottenimento delle maschere di H. halys e degli sfondi vegetali grazie all'utilizzo di PCA*

4.4 Sviluppo e validazione dei modelli di classificazione

Per il calcolo dei modelli di classificazione a partire delle immagini iperspettrali acquisite è necessario costruire un dataset di spettri rappresentativi di ciascuna classe da modellare, in questo caso *H. halys* e i diversi sfondi vegetali. Generalmente, questa fase può essere effettuata andando a selezionare sulle immagini delle ROI corrispondenti alle diverse classi e calcolando per ciascuna ROI lo spettro medio oppure, in alternativa, si può ricorrere alla selezione di un certo numero di spettri in maniera casuale oppure utilizzando opportuni algoritmi.

Nel presente lavoro di tesi, la selezione delle ROI è stata effettuata mediante segmentazione dei pixel appartenenti ad *H. halys* e ai diversi sfondi utilizzando PCA (Sezione 4.3), mentre la selezione degli spettri rappresentativi delle due classi da utilizzare per lo sviluppo dei modelli di classificazione è stata effettuata con due approcci diversi in Fase I e in Fase II.

Nel caso delle immagini acquisite in Fase I, da ciascuna immagine sono stati selezionati in maniera casuale 100 spettri, di cui 50 riconducibili ad *H. halys* e 50 riconducibili allo sfondo vegetale. Il dataset così ottenuto (Dataset I) è composto da 1200 spettri prelevati da 12 immagini differenti, di cui 600 spettri appartenenti agli sfondi vegetali, 300 spettri di *H. halys* appartenenti alla tipologia HH_1 (esemplari deceduti da diversi giorni) e 300 spettri di *H. halys* appartenenti alla tipologia HH_2 (esemplari deceduti da poche ore).

N° immagine	1	2	3	4	5	6	7	8	9	10	11	12
Sfondo	Corteccia		Erba		Foglie Secche		Foglie Verdi		Rami		Terra	
Cimice	HH_1	HH_2	HH_1	HH_2	HH_1	HH_2	HH_1	HH_2	HH_1	HH_2	HH_1	HH_2
TR/TS	TR	TS	TS	TR	TR	TS	TS	TR	TR	TS	TS	TR

Tabella 4.3 *Suddivisione dei campioni del dataset I in training set e test set per lo sviluppo dei modelli di classificazione.*

Il Dataset I è stato suddiviso in training set (TR), utilizzato per il calcolo dei modelli di classificazione mediante PLS-DA, e test set (TS) per la validazione esterna del modello. Ulteriori dettagli sulla suddivisione del Dataset I in training set e test set e sono riportati in Tabella 4.3 ed in Tabella 4.4. Per la cross-validazione è stata effettuata una suddivisione *custom* del training set in 6 gruppi di cancellazione, in modo tale da mantenere all'interno dello stesso gruppo gli spettri appartenenti alla stessa immagine.

Dataset I			
Classe	**Spettri (tot.)**	**Spettri TR**	**Spettri TS**
Halyomorpha halys	600	300	300
HH_1	300	150	150
HH_2	300	150	150
Sfondi	600	300	300
Corteccia	100	50	50
Erba	100	50	50
Foglie Secche	100	50	50
Foglie Verdi	100	50	50
Rami	100	50	50
Terra	100	50	50

Tabella 4.4 *Suddivisione degli spettri selezionati da ciascuna immagine in training set e test set.*

Per quanto riguarda le immagini acquisite in Fase II, la selezione degli spettri più rappresentativi di ciascuna classe è stata effettuata utilizzando l'algoritmo di Kennard-Stone a partire dagli score di modelli PCA calcolati per ogni immagine considerando solamente i pixel degli sfondi o delle cimici. L'algoritmo di Kennard-Stone consente di estrarre un sottoinsieme di campioni a partire da una popolazione in modo tale che essi abbiano distribuzione il più possibile uniforme. Tale algoritmo lavora sulla matrice delle distanze tra gli oggetti della popolazione iniziale: dapprima vengono selezionati i due campioni più distanti tra loro, successivamente per

ogni oggetto non selezionato si calcolano le distanze tra i due oggetti selezionati in precedenza e si seleziona quello con distanza di separazione maggiore, dove per distanza di separazione si intende la distanza tra l'oggetto in esame e l'oggetto selezionato più vicino. Tale procedura viene ripetuta fintanto che non viene raggiunto il numero richiesto di campioni selezionati.

Più in dettaglio, per ogni immagine la selezione dei pixel appartenenti allo sfondo è stata effettuata con i seguenti passaggi:

1. Calcolo di un modello PCA, tramite pretrattamento con mean centering, considerando solamente i pixel dello sfondo e selezionando 3 PC;
2. Eliminazione dei pixel *outlier* con valori di T^2 di Hotelling e/o residui Q al di fuori del corrispondente limite di confidenza del 99.9%;
3. Calcolo di un nuovo modello PCA dei pixel dello sfondo dopo aver eliminato quelli identificati come *outlier*. Anche questo modello viene calcolato con mean centering come pretrattamento degli spettri utilizzando 3 PC;
4. Applicazione dell'algoritmo Kennard-Stone nello spazio delle PC per selezionare 200 spettri rappresentativi della variabilità dello sfondo nell'immagine considerata.

L'utilizzo dell'algoritmo di Kennard-Stone ha permesso di campionare gli spettri da utilizzare per il calcolo di modelli di classificazione in modo tale da tenere in considerazione la variabilità e l'eterogeneità di ciascuno sfondo.

Infatti, partendo dal presupposto che il modello PCA calcolato solamente sugli spettri dello sfondo utilizzando mean centering come pretrattamento ne descriva con buona approssimazione la variabilità, l'applicazione dell'algoritmo di Kennard-Stone agli score di PCA permette di selezionare i pixel in modo tale da coprire in maniera uniforme lo spazio delle PC, consentendo così di selezionare anche i pixel agli estremi delle PC stesse, che avrebbero invece minore probabilità di essere selezionati adottando un approccio casuale. Di fatto, tramite l'approccio di selezione casuale degli spettri, l'algoritmo tende con minore probabilità a considerare le zone a più bassa densità di pixel, le quali si trovano solitamente agli estremi delle PC (Figura 4.2). Tuttavia, questi pixel sono fondamentali per ottenere un dataset rappresentativo poiché permettono di considerare al meglio la variabilità descritta dalle immagini, specialmente per quanto riguarda matrici altamente eterogenee come gli sfondi vegetali. Pertanto, tralasciando tali regioni per la costruzione del dataset, si rischia di inficiare la capacità predittiva dei modelli di classificazione ottenuti.

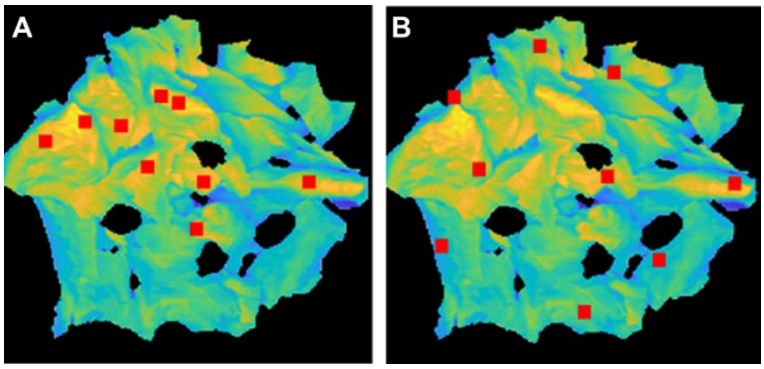

Figura 4.2 *Esempi di selezione di pixel utilizzando A) selezione casuale; B) selezione con algoritmo di Kennard-Stone.*

La stessa procedura descritta in precedenza è stata utilizzata per selezionare da ciascuna immagine 200 spettri appartenenti a *H. halys*. Pertanto, il dataset ottenuto a partire dalle immagini acquisite in Fase II (Dataset II) è costituito da 14000 spettri campionati a partire da 35 immagini, di cui 7000 appartenenti ai diversi sfondi vegetali e 7000 appartenenti agli esemplari di cimice asiatica.

Questo dataset è stato utilizzato per calcolare i modelli di classificazione utilizzando gli algoritmi Soft PLS-DA e *sparse* Soft PLS-DA. La suddivisione in training set e test set è stata effettuata considerando i cinque differenti gruppi di cimici; pertanto, per la costruzione del training set sono stati utilizzati gli spettri selezionati a partire dalle immagini contenenti i gruppi G1, G2 e G3, per un totale di 8400 spettri, mentre per il test set sono stati utilizzati gli spettri selezionati a partire dalle immagini contenenti i gruppi G4 e G5, per un totale di 5600 spettri. Maggiori informazioni sulla suddivisione del Dataset II in training set e test set sono riportate nelle Tabelle 4.5 e 4.6.

n° immagine	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Sfondo	Corteccia					Erba					Foglie Gialle					Foglie Secche					Foglie Verdi					Rami					Terra				
Gruppo cimici	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5
TR/TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS	TR	TR	TR	TS	TS

Tabella 4.5 *Suddivisione dei campioni del dataset II in training set e test set per lo sviluppo dei modelli di classificazione.*

Dataset II			
Classe	**Spettri (tot.)**	**Spettri TR**	**Spettri TS**
Halyomorpha halys	7000	4200	2800
Sfondi (tot.)	7000	4200	2800
Corteccia	1000	600	400
Erba	1000	600	400
Foglie Gialle	1000	600	400
Foglie Secche	1000	600	400
Foglie Verdi	1000	600	400
Rami	1000	600	400
Terra	1000	600	400

Tabella 4.6 *Suddivisione degli spettri selezionati da ciascuna immagine in training set e test set.*

Come metodo di cross-validazione è stata effettuata una suddivisione del training set in 3 gruppi di cancellazione in modo tale da mantenere all'interno dello stesso gruppo gli spettri selezionati a partire dalla stessa immagine.

I modelli di classificazione ottenuti a partire da entrambi i dataset sono stati calcolati testando i seguenti pretrattamenti di riga: SNV, detrend, derivata prima e derivata seconda, in aggiunta al pretrattamento di colonna mean centering. La capacità predittiva dei modelli è stata valutata calcolando i valori di sensibilità (SENS), specificità (SPEC) ed efficienza (EFF) in calibrazione, cross-validazione e predizione del test set.

Come ulteriore validazione dei modelli di calibrazione ottenuti in entrambe le fasi sperimentali, è stata effettuata la predizione a livello di pixel andando applicare i modelli calcolati direttamente sulle immagini iperspettrali e visualizzando le corrispondenti immagini in predizione (*prediction images*, Sezione 3.1.4).

5 Risultati e discussione

5.1 Fase I

5.1.1 Analisi esplorativa

L'analisi esplorativa delle immagini acquisite in Fase I è stata effettuata principalmente per verificare la possibilità di separare i campioni di *H. halys* dai diversi sfondi vegetali. Per effettuare una prima analisi esplorativa sono state unite immagini iperspettrali di diversi campioni di cimice, HH_1 o HH_2, acquisite sui diversi sfondi vegetali; ciò ha permesso la creazione di un'immagine composita. L'immagine composita risultante è stata analizzata mediante PCA effettuando il pretrattamento di colonna mean centering in combinazione con diversi pretrattamenti di riga (SNV, derivata prima e derivata seconda).

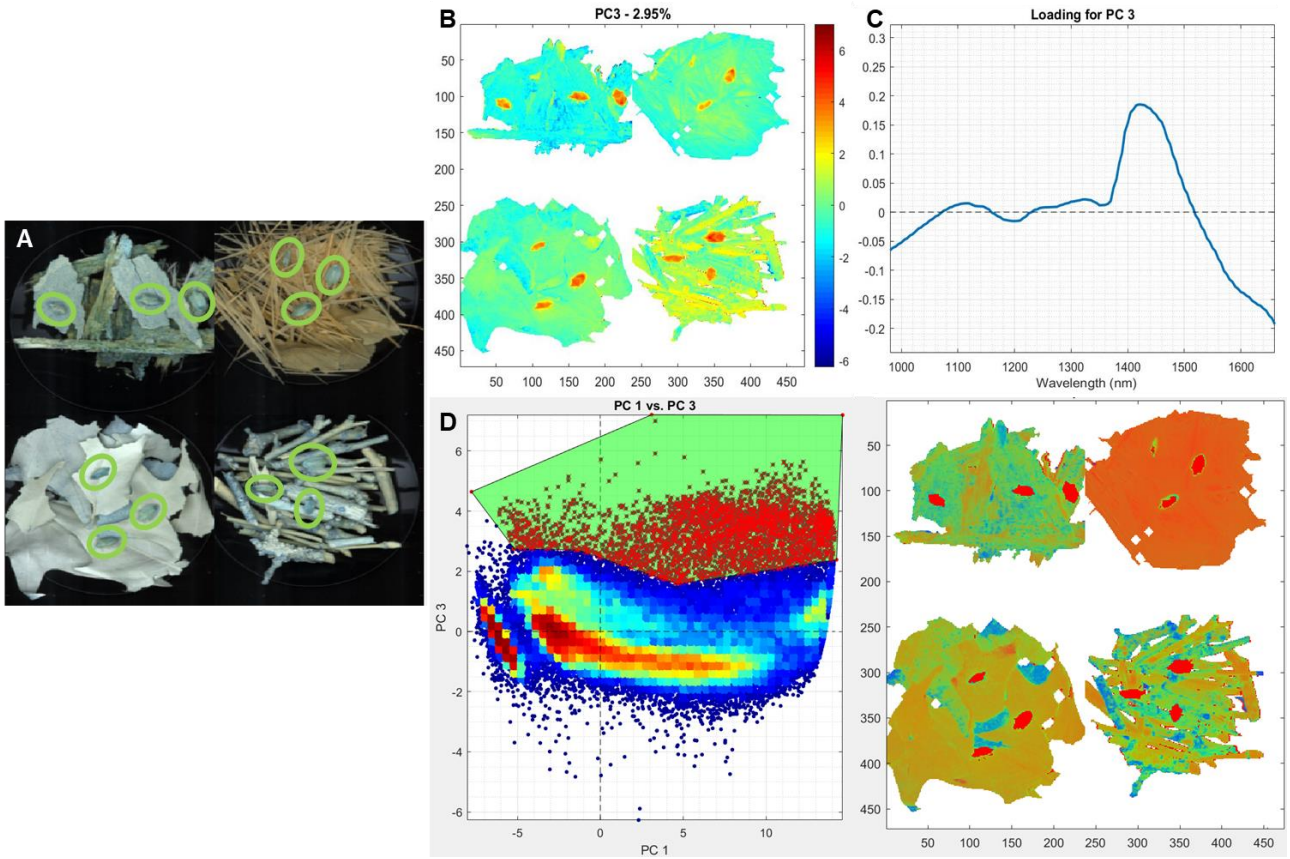

Figura 5.1 *PCA effettuata sull'immagine composita di quattro campioni di H. halys* ***HH_1*** *sugli sfondi corteccia, erba, foglie secche e rami con pretrattamento SNV e mean centering: immagine RGB composita di riferimento (A), immagine degli* score *di PC3 (B), grafico dei* loading *di PC3 (C) e selezione di pixel (brushing) dallo* score plot *di PC1 e PC3 al fine di separare H. halys.*

Tra i diversi pretrattamenti testati per l'analisi esplorativa dei campioni HH_1 sui diversi sfondi vegetali, i pretrattamenti di riga SNV e derivata prima accoppiati con mean centering hanno permesso una migliore

separazione dei pixel di cimice asiatica rispetto a quelli degli sfondi vegetali. A tal proposito, in Figura 5.1 viene riportato un esempio di immagine composita dei campioni HH_1 su sfondi corteccia, erba, foglie secche e rami. In Figura 5.1 (B) è riportata l'immagine degli score di PC3, che permette di osservare la separazione dei pixel delle cimici HH_1, le quali presentano valori positivi agli estremi della componente principale (rappresentati in colore rosso nella figura), distanziandosi dai pixel dei diversi sfondi considerati. Nonostante ciò, come visibile nello score plot di PC1 e PC3 mostrato in Figura 5.1 (D), i pixel corrispondenti ad *H. halys* non formano un cluster compatto e ben distinto dagli sfondi vegetali, anche se è possibile identificare mediante *brushing* i pixel appartenenti agli esemplari di cimice.

Il vettore dei loading di PC3, riportato in Figura 5.1 (C), consente di valutare quali regioni dello spettro influenzino la separazione dei campioni di *H. halys* dallo sfondo: in questo caso, la zona spettrale di maggior interesse è rappresentata dall'intervallo compreso tra 1400 e 1660 nm, la quale descrive il primo *overtone* del legame O-H, la banda di combinazione del legame C-H aromatico, il terzo *overtone* del legame C=O ed il primo *overtone* dello *stretching* del legame N-H.

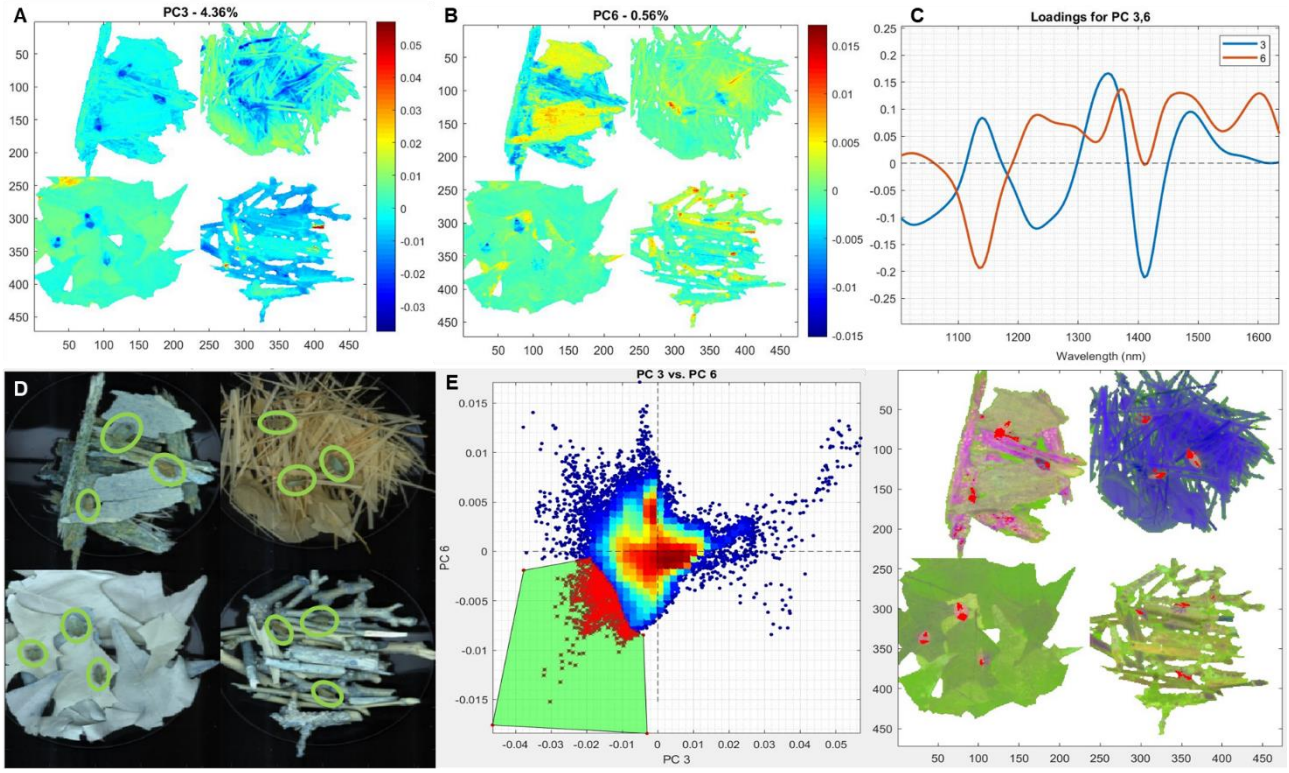

Figura 5.2 *PCA effettuata sull'immagine composita di quattro campioni di H. halys* ***HH_2*** *sugli sfondi corteccia, erba, foglie secche e rami con pretrattamento derivata prima e mean centering: immagine degli* score *di PC3 (A), immagine degli* score *di PC6 (B), grafico dei* loading *di PC3 e PC6 (C), immagine RGB composita di riferimento (D) e selezione di pixel (brushing) dallo* score plot *di PC3 e PC6 al fine di separare H. halys (E).*

Per l'analisi esplorativa delle immagini iperspettrali dei campioni HH_2 sui diversi sfondi vegetali sono stati utilizzati i pretrattamenti di riga SNV, derivata prima e derivata seconda accoppiati a mean centering, i quali hanno permesso la separazione dei pixel di cimice asiatica rispetto a quelli degli sfondi. Nella Figura 5.2 viene riportata un'immagine composita di quattro campioni HH_2 su sfondi corteccia, erba, foglie secche e rami.

In questo caso l'indagine esplorativa, effettuata tramite PCA al fine di separare i pixel appartenenti ad *H. halys* da quelli dei diversi sfondi vegetali, è risultata più difficoltosa. Innanzitutto, dal grafico degli score di PC3 e PC6 (Figura 5.2 E) è possibile separare soltanto parzialmente i pixel di cimici HH_2, i quali sono descritti da valori negativi per entrambe le componenti principali. Perciò, confrontando le Figure 5.1 (D) e 5.2 (E), la separazione dei pixel di *H. halys* risulta meno evidente per i campioni HH_2 rispetto a quella riscontrata con i campioni HH_1.

Per quanto riguarda il grafico dei loading, riportato in Figura 5.2 (C), le zone spettrali che influenzano maggiormente la separazione dei pixel dei campioni di cimice HH_2 dagli sfondi sono più difficilmente interpretabili rispetto al caso precedente, anche a causa del pretrattamento di riga utilizzato (ossia derivata prima).

Un'ulteriore differenza rispetto ai risultati ottenuti tramite l'analisi esplorativa di un'immagine composita partendo da campioni di HH_1 riguarda gli sfondi: erba, corteccia e rami, considerati nella costruzione dell'immagine composita dei campioni HH_2, non sembrano consentire una separazione ottimale dei pixel di *H. halys*.

L'analisi esplorativa mediante PCA è stata effettuata anche sul Dataset I, ovvero sul dataset di spettri di *H. halys* (HH_1 e HH_2) e sfondi vegetali campionati in maniera casuale dalle immagini acquisite nella prima fase. L'analisi di questo dataset ha permesso di chiarire alcune criticità e di confermare diverse ipotesi effettuate precedentemente. In Figura 5.3 sono riportati i risultati del modello PCA calcolato utilizzando come pretrattamento derivata prima e mean centering.

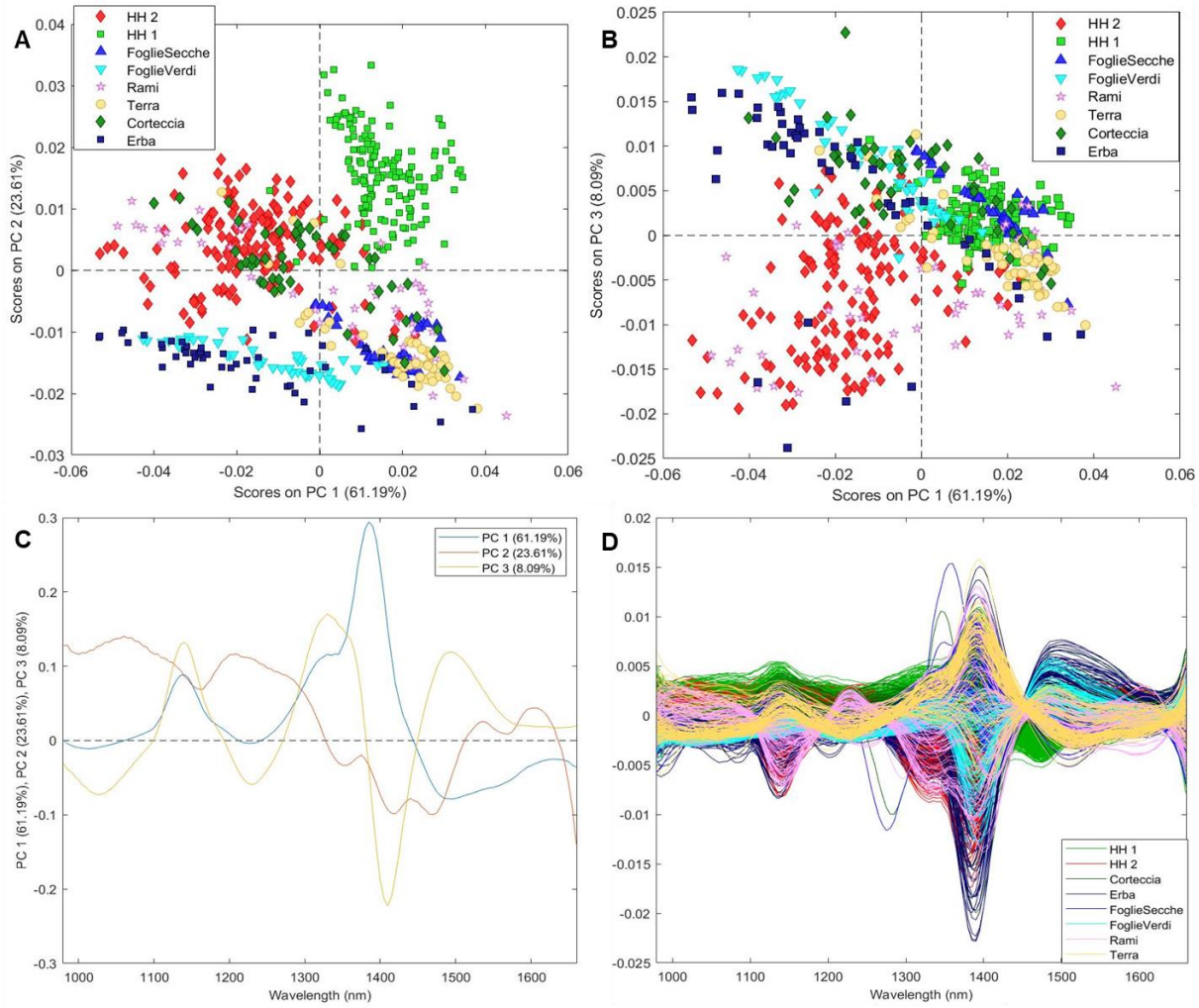

Figura 5.3 *Risultati del modello PCA calcolato a partire dal Dataset I: vengono riportati il grafico degli* score *PC1 e PC2 (A), il grafico degli* score *PC1 e PC3 (B) e il corrispondente grafico dei* loading *di PC1, PC2 e PC3 (C) insieme al grafico degli spettri pretrattati (D).*

Nel grafico degli score di PC1 e PC2 (Figura 5.3 A) è possibile visualizzare la separazione degli spettri di cimice HH_1 (quadrati verdi) rispetto agli altri spettri; infatti, gli spettri di cimice HH_1 tendono a formare un cluster a valori positivi sia per PC1 che per PC2. Si nota quindi che gli spettri dei due gruppi di cimice tendono a formare due cluster separati e che gli spettri di cimice HH_2 risultano sovrapposti agli spettri di corteccia e rami. Dallo stesso score plot è inoltre possibile osservare che i campioni di foglie verdi ed erba, descritti da valori negativi di PC2, tendono a discostarsi dagli altri sfondi formando un cluster.

Considerando invece il grafico degli score di PC1 e PC3 (Figura 5.3 B), è possibile osservare che i campioni di cimice HH_2 (rombi rossi) tendono a formare un cluster posizionato a valori tendenzialmente negativi sia per PC1 che per PC3. Tuttavia, in questo caso la separazione non è così netta come osservato per i campioni di cimice HH_1 e si ha comunque una certa sovrapposizione con alcuni spettri dello sfondo rami.

Il grafico dei loading di PC1, PC2 e PC3 (Figura 5.3 C) evidenzia le zone spettrali più rilevanti sui raggruppamenti osservati nelle prime tre PC. A tal proposito, confrontano il grafico dei loading con gli spettri pretrattati (Figura 5.3 D) risulta più semplice individuare le zone spettrali più rilevanti per definire ciascuna classe.

Per i loading di PC1, pare che le zone dello spettro più rilevanti con valori positivi di loading siano comprese nell'intervallo tra 1100 nm e 1150 nm (secondo *overtone* del legame C-H) e nell'intervallo tra 1300 nm e 1420 nm (banda di combinazione del legame C-H, primo *overtone* del legame O-H), mentre la zona più rilevante per valori negativi di PC1 sembra corrispondere all'intervallo tra 1500 e 1600 nm (primo *overtone* dello *stretching* del legame N-H). Per quanto riguarda i loading di PC2, essi dovrebbero comprendere le regioni spettrali rilevanti per la separazione degli sfondi foglie verdi ed erba rispetto agli altri campioni: in questo caso, le zone dello spettro più rilevanti sembrano essere comprese nell'intervallo 1300-1500 nm. I loading di PC3, invece, dovrebbero consentire l'identificazione delle zone utili per separare i campioni di cimice HH_2, i quali presentano valori tendenzialmente negativi di PC3, rispetto alla maggior parte degli sfondi e degli spettri di HH_1. Le zone dello spettro con valori più elevati in termini di valore assoluto di loading di PC3 sembrano quelle comprese nell'intervallo 1100-1150 nm (secondo *overtone* del legame C-H), nell'intervallo 1300-1400 nm (banda di combinazione del legame C-H), nell'intervallo 1400-1450 (primo *overtone* del legame O-H) nm e nell'intervallo 1450-1550 nm (primo *overtone* dello *stretching* del legame N-H). Nonostante ciò, l'interpretazione dei loading del presente modello PCA risulta di difficile interpretazione, anche a causa del pretrattamento di riga utilizzato (ovvero derivata prima).

Confrontando i risultati ottenuti tramite l'analisi esplorativa effettuata sulle immagini composite e sul Dataset I, è stato possibile ottenere alcune conclusioni utili. Innanzitutto, l'imaging iperspettrale nel vicino infrarosso sembra permettere, in linea generale, di distinguere l'informazione spettrale riconducibile ad *H. halys* e ai diversi sfondi vegetali. In secondo luogo, le differenze riscontrate tra cimici morte da diversi giorni (HH_1) e cimici morte da poche ore (HH_2) hanno messo in luce la necessità di prendere alcuni accorgimenti dal punto di vista sperimentale riguardo ai campioni di cimice. Infatti, essendo i campioni HH_2 più similari agli esemplari di cimice riscontabili in campo, per l'acquisizione delle immagini nella seconda fase sperimentale sono stati considerati solamente esemplari di *H. halys* morti da poco tempo. Sempre per questo motivo, i modelli di classificazione con PLS-DA ottenuti a partire dal Dataset I sono stati calcolati considerando solamente gli spettri di cimici del gruppo HH_2.

Un ulteriore punto critico, in particolar modo sui campioni di sfondi corteccia e rami, può essere rappresentato dall'influenza della regione spettrale nell'intervallo 1400-1450 nm, riconducibile alla presenza di acqua. Ciò potrebbe essere collegato al periodo di campionamento degli sfondi della Fase I (aprile 2021), che è avvenuto dopo qualche giorno di pioggia. Poiché la spettroscopia NIR è particolarmente sensibile all'influenza dell'acqua, è da tenere in considerazione che in applicazioni pratiche l'acquisizione in campo possa essere condizionata dall'effetto dell'umidità; pertanto, sarebbe preferibile minimizzarne l'effetto nel calcolo dei modelli di classificazione

5.1.2 Modelli di classificazione

I modelli di classificazione PLS-DA ottenuti a partire dai campioni del Dataset I sono stati calcolati per discriminare i campioni di *H. halys* (HH_2) dalle diverse matrici vegetali. Ai fini di calcolo sono stati testati i pretrattamenti di riga SNV, detrend, derivata prima e derivata seconda in combinazione con mean centering come pretrattamento di colonna.

Il calcolo dei modelli è stato effettuato sia considerando tutto lo spettro, sia escludendo la zona spettrale corrispondente alla regione di assorbimento dell'acqua (1350-1500 nm). Tale passaggio è stato effettuato per capire se le variazioni di umidità possano impattare negativamente l'efficienza dei modelli di classificazione, andando a minimizzare la rilevanza di zone spettrali più significative ai fini della discriminazione, oppure se effettivamente anche in questa zona dello spettro è presente informazione utile per la classificazione.

I modelli ottenuti considerando l'intero range spettrale (Tabella 5.1) sono risultati lievemente migliori rispetto a quelli ottenuti non considerando la regione di assorbimento dell'acqua (Tabella 5.2). In generale, in entrambi i casi i risultati migliori in cross-validazione ed in predizione sono stati ottenuti grazie ai pretrattamenti di riga derivata prima e detrend, accoppiati al pretrattamento di colonna mean centering. I risultati ottenuti sono abbastanza simili e per entrambi i modelli; i principali problemi nella classificazione si riscontrano per gli sfondi corteccia e rami.

		SNV + mean center		**Derivata prima + mean center**		**Detrend + mean center**		**Derivata seconda + mean center**	
		CIMICI	**SFONDI**	**CIMICI**	**SFONDI**	**CIMICI**	**SFONDI**	**CIMICI**	**SFONDI**
CV	**SENS**	0,927	0,813	0,920	0,890	0,920	0,880	0,893	0,897
	SPEC	0,813	0,927	0,890	0,920	0,880	0,920	0,897	0,893
	EFF	0,868		0,905		0,900		0,895	
PRED	**SENS**	0,887	0,894	0,920	0,890	0,873	0,903	0,867	0,899
	SPEC	0,894	0,887	0,890	0,920	0,903	0,873	0,899	0,867
	EFF	0,890		0,905		0,888		0,899	

Tabella 5.1 *Tabella riassuntiva dei valori di sensibilità (SENS), specificità (SPEC) ed efficienza (EFF) per ciascun modello PLSDA calcolato considerando le classi di cimici HH_2 e sfondi per l'intero range spettrale.*

		SNV + mean center		Derivata prima + mean center		Detrend + mean center		Derivata seconda + mean center	
		CIMICI	**SFONDI**	**CIMICI**	**SFONDI**	**CIMICI**	**SFONDI**	**CIMICI**	**SFONDI**
CV	**SENS**	0,933	0,793	0,900	0,873	0,927	0,850	0,873	0,873
	SPEC	0,793	0,933	0,873	0,900	0,850	0,927	0,873	0,873
	EFF	0,861		0,887		0,888		0,816	
PRED	**SENS**	0,840	0,929	0,887	0,881	0,907	0,873	0,860	0,884
	SPEC	0,929	0,840	0,881	0,887	0,873	0,907	0,884	0,860
	EFF	0,883		0,884		0,889		0,872	

Tabella 5.2 *Tabella riassuntiva dei valori di sensibilità (SENS), specificità (SPEC) ed efficienza (EFF) per ciascun modello PLSDA calcolato considerando le classi di cimici HH_2 e sfondi, senza considerare la regione spettrale 1350-1500 nm.*

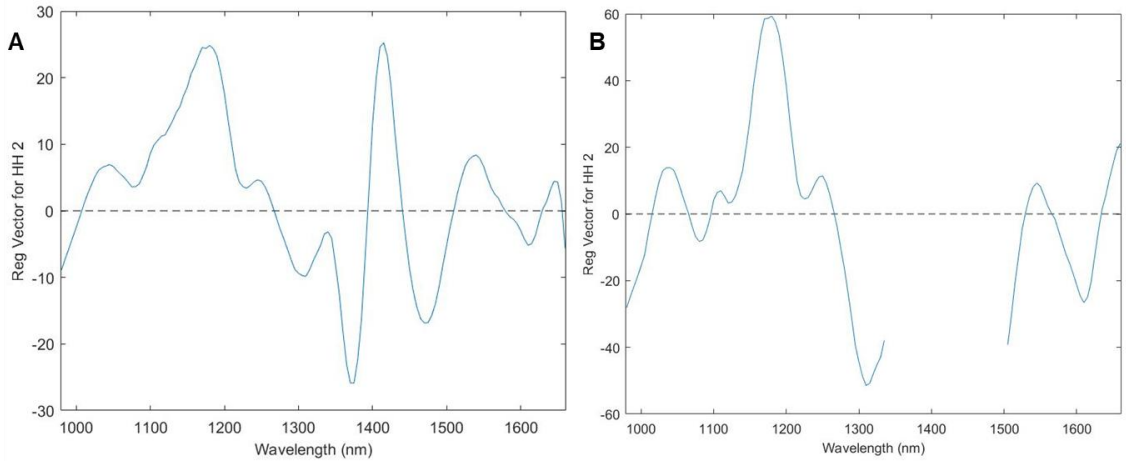

Figura 5.4 *Vettori dei coefficienti di regressione del modello PLS-DA (pretrattamento derivata prima e mean centering) per la classe cimici HH_2 considerando l'intero range spettrale (A) e senza considerare la regione spettrale 1350-1500 nm (B).*

I vettori di regressione riportati in Figura 5.4 si riferiscono al miglior modello PLS-DA (pretrattamento derivata prima e mean centering) calcolato considerando l'intero range spettrale (A) o escludendo la regione attribuibile all'assorbimento dell'acqua (B). Essi risultano utili ai fini di individuare le regioni spettrali più rilevanti ai fini della classificazione.

Nel caso del modello PLS-DA calcolato considerando l'intero range spettrale (Figura 5.4 A), le regioni più influenti per la discriminazione della classe HH_2 rispetto agli sfondi sono comprese nell'intervallo 1000-1100 nm (secondo *overtone* del legame N-H), nella zona centrata a 1200 nm (secondo *overtone* del legame C-H), nella zona centrata a 1400 nm (primo *overtone* del legame O-H) e nella zona centrata a 1450 nm (bande di combinazione dei legami C-H e C=O).

Confrontando tali risultati con il vettore di regressione del modello PLS-DA che non considera la regione spettrale tra 1350 nm e 1500 nm (Figura 5.4 B) è possibile visualizzare una variazione nell'intensità dei picchi per le regioni spettrali a 1050 nm, 1200 nm e 1280 nm (banda di combinazione del legame C-H). In questo caso, tuttavia, la capacità predittiva del modello in cross-validazione ed in predizione su un test set esterno è peggiorata rispetto ai risultati ottenuti considerando tutto lo spettro. Pertanto, la zona compresa tra 1350 nm e 1500 nm porta informazioni utili alla discriminazione tra spettri di cimice e spettri dei diversi sfondi vegetali. Questo effetto può essere dovuto al fatto che la banda di assorbimento dell'acqua è piuttosto ampia e va a coprire anche zone in cui avvengono altri assorbimenti, come quelli relativi alla banda di combinazione del legame C-H (1360-140 nm) ed al primo *overtone* dello *stretching* del legame N-H (1460-1530 nm).

Per effettuare una ulteriore validazione dei risultati relativi alla classificazione, in Figura 5.5 sono riportate le immagini in predizione ottenute applicando il miglior modello PLS-DA (ossia pretrattamento con derivata prima e mean centering considerando tutto il range spettrale) su alcune immagini iperspettrali di cimici HH_2 sugli sfondi foglie secche, foglie verdi, terra e rami. Per una migliore valutazione dei risultati della classificazione, insieme alle immagini in predizione sono state riportate anche le immagini RGB dei campioni corrispondenti.

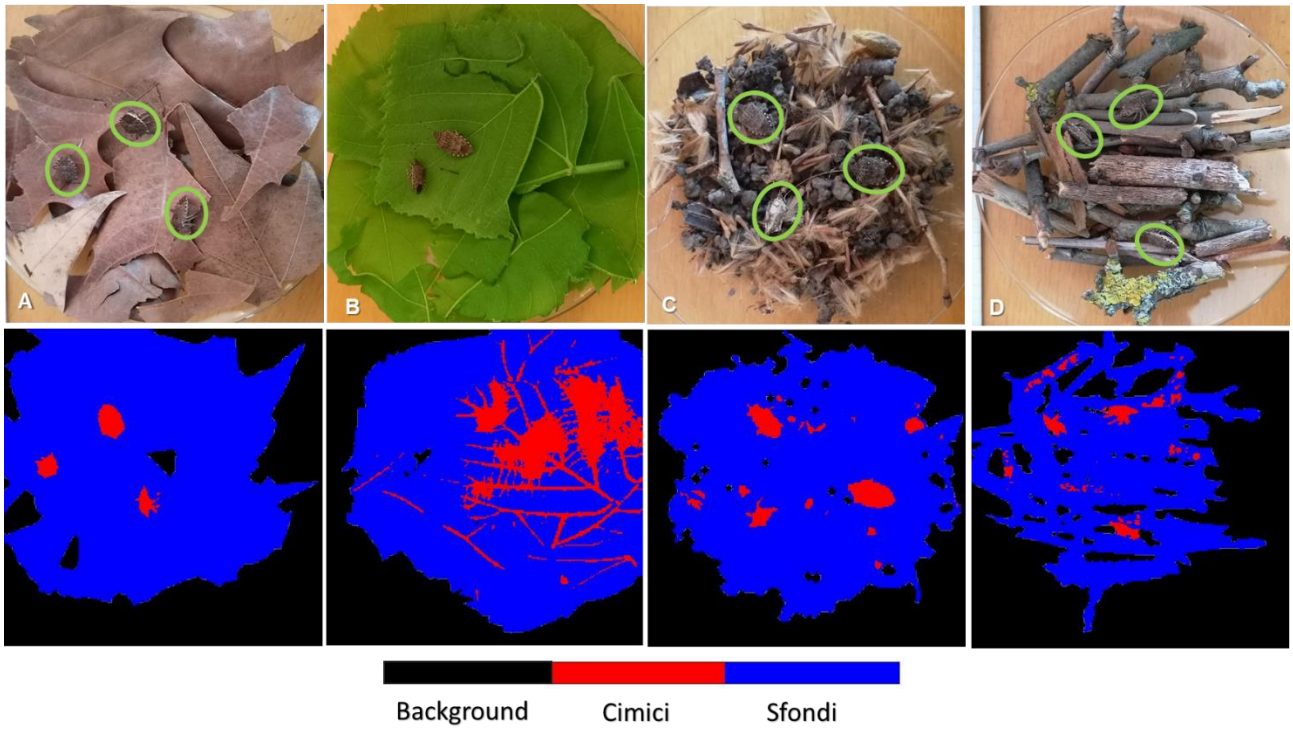

Figura 5.5 *Immagini in predizione dei campioni di cimici HH_2 e gli sfondi vegetali e le relative immagini RGB di riferimento. Le immagini in predizione si riferiscono al miglior modello PLS-DA (intero range spettrale, derivata prima e mean centering), applicato sulle immagini iperspettrali con sfondi foglie secche (A), foglie verdi (B), terra (C) e rami (D).*

Considerando le immagini in predizione riportate in Figura 5.5, i risultati migliori sono stati ottenuti sull'immagine con lo sfondo foglie secche, mentre per le immagini con gli sfondi terra e rami alcuni pixel dello sfondo sono stati predetti come appartenenti alla classe *H. halys;* nonostante ciò, nel complesso i risultati della classificazione sono buoni. Inoltre, visualizzando le immagini RGB di questi campioni si può osservare come sia piuttosto difficile identificare la presenza delle cimici su sfondi di colore marrone, a causa del mimetismo di *H. halys,* confermando quindi la necessità di ricorrere alla regione del vicino infrarosso per un'identificazione efficace di questo insetto.

Le maggiori criticità sono invece emerse per l'immagine con sfondo foglie verdi, in cui un elevato numero di pixel dello sfondo viene erroneamente assegnato alla classe *H. halys*. Da notare che la quasi totalità dei pixel dello sfondo erroneamente classificati sono localizzati in corrispondenza delle nervature delle foglie. Questo effetto può essere dovuto al fatto che all'interno degli spettri del Dataset I, utilizzati per il calcolo dei modelli di classificazione, non viene rappresentata a sufficienza la variabilità relativa ai diversi sfondi, come ad esempio la presenza di venature delle foglie.

Per risolvere questa criticità e rappresentare al meglio la variabilità dei diversi sfondi vegetali nel dataset utilizzato per il calcolo dei modelli, nella successiva fase sperimentale è stato utilizzato un diverso approccio per l'ottenimento del Dataset II, basato sull'utilizzo dell'algoritmo di Kennard-Stone a partire da modelli PCA calcolati singolarmente per ciascuna immagine considerando solamente i pixel dello sfondo o delle cimici.

5.2 Fase II

5.2.1 Analisi esplorativa

L'analisi esplorativa del Dataset II è stata effettuata ai fini di confronto per verificare un possibile miglioramento nella separazione degli spettri di *H. halys* dai diversi sfondi vegetali rispetto a quanto osservato per il Dataset I. Il modello PCA è stato calcolato utilizzando il pretrattamento detrend seguito da mean centering (Figura 5.6 A).

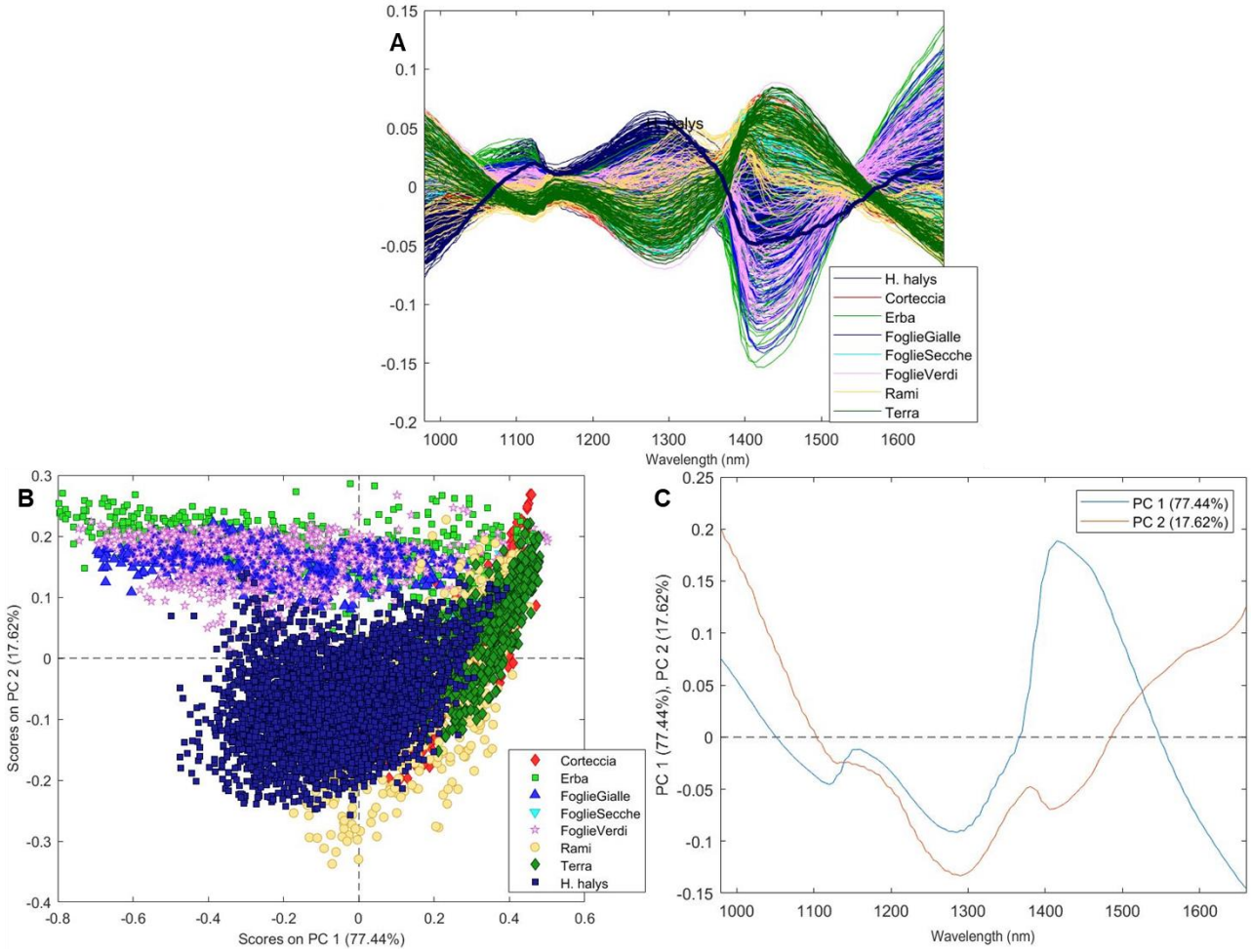

Figura 5.6 *Risultati del modello PCA relativo al Dataset II: vengono riportati i segnali pretrattati (A), il grafico degli* score *PC1/PC2 (B) ed il grafico dei* loading *di PC1 e PC2.*

Osservando il grafico degli score di PC1 e PC2 (Figura 5.6 B) è possibile distinguere due principali cluster di sfondi: gli sfondi "verdi" come foglie verdi, erba e foglie gialle che presentano valori positivi di PC2, e gli sfondi "scuri" quali corteccia, terra, foglie secche e, in parte, rami, che presentano valori positivi di PC1. Per quanto riguarda i campioni di *H. halys*, questi sono tendenzialmente caratterizzati da valori negativi di PC2; inoltre, rispetto al dataset preliminare formano un cluster più compatto (Figura 5.6 B). Anche in questo caso vi è una sovrapposizione parziale degli spettri dello sfondo rami e di *H. halys.*

Il grafico dei loading di PC1 e PC2 (Figura 5.6 C) evidenzia le zone spettrali più influenti sulla formazione dei cluster osservati nello score plot corrispondente. Per i loading di PC1, le zone dello spettro più rilevanti con valori positivi sono comprese nella zona intorno a 1000 nm (secondo *overtone* del legame N-H) e nell'intervallo 1400-1500 nm (primo *overtone* del legame O-H e primo *overtone* del legame N-H), mentre le zone più rilevanti a valori negativi di PC1 è compresa nell'intervallo 1200-1350 nm (secondo *overtone* del legame C-H) e attorno a 1650 nm (primo *overtone* del legame C-H).

Considerando i loading di PC2 è possibile identificare le regioni spettrali rilevanti per la separazione degli sfondi verdi come erba, foglie gialle e foglie verdi rispetto ai campioni di cimice asiatica. In questo caso, le zone dello spettro più rilevanti con valori positivi di PC2 sono a 1000 nm (secondo *overtone* del legame N-H) e tra i 1500 nm e i 1660 nm (primo *overtone* del legame N-H e primo *overtone* del legame C-H), mentre le zone spettrali con valori negativi di PC2 sono a 1120 nm, 1280 nm (secondo *overtone* del legame C-H) e a 1420 nm (primo *overtone* del legame O-H).

5.2.2 Modelli di classificazione

Come anticipato, il Dataset II è stato utilizzato per la costruzione di modelli di classificazione PLS-DA, Soft PLS-DA e *sparse* Soft PLS-DA. Sebbene i modelli ottenuti tramite la classificazione con PLS-DA abbiano buone performance in termini di efficienza di classificazione, l'implementazione di modelli di classificazione Soft PLS-DA e *sparse* Soft PLS-DA risulta maggiormente vantaggiosa dal punto di vista pratico, grazie alla possibilità di non assegnare eventuali spettri non appartenenti alle classi considerate e alla selezione tramite *sparse* Soft PLS-DA delle zone spettrali più utili alla classificazione.

Per il calcolo di ciascun modello sono stati considerati i pretrattamenti di riga SNV, derivata prima, derivata seconda e detrend, accoppiati al pretrattamento di colonna mean centering. In Tabella 5.3 sono riportati i principali risultati ottenuti per ciascun modello calcolato. Le successive immagini riportate per PLS-DA, Soft PLS-DA e *sparse* Soft PLS-DA si riferiscono al modello migliore, ovvero a quello più parsimonioso e con il valore di efficienza più elevato individuato in cross-validazione.

Dalla tabella riassuntiva (Tabella 5.3) dei risultati ottenuti si può evincere che il pretrattamento che permette di ottenere il modello migliore è *Standard Normal Variate* (SNV) accoppiato a mean centering. Sebbene il pretrattamento con derivata seconda consenta di ottenere risultati lievissimamente migliori in cross-validazione nel calcolo del modello *sparse* Soft PLS-DA, si preferisce utilizzare SNV in quanto permette l'utilizzo di un minor numero di variabili latenti (LV) e di variabili selezionate, oltre ad offrire un vettore dei coefficienti di regressione più facilmente interpretabile.

PLSDA								
	SNV + mean center		**Derivata prima + mean center**		**Detrend + mean center**		**Derivata seconda + mean center**	
	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI
SENS CV	0,934	0,958	0,909	0,949	0,907	0,958	0,933	0,970
SPEC CV	0,958	0,934	0,949	0,909	0,958	0,907	0,970	0,933
EFF CV	0,946	0,946	0,929	0,929	0,932	0,932	0,951	0,951
SENS Pred	0,954	0,965	0,924	0,953	0,916	0,962	0,931	0,973
SPEC Pred	0,965	0,954	0,953	0,924	0,962	0,916	0,973	0,931
EFF Pred	0,959	0,959	0,938	0,938	0,939	0,939	0,952	0,952
Soft PLSDA								
	SNV + mean center		**Derivata prima + mean center**		**Detrend + mean center**		**Derivata seconda + mean center**	
	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI
SENS CV	0,934	0,948	0,909	0,939	0,908	0,951	0,920	0,939
SPEC CV	0,958	0,934	0,948	0,909	0,961	0,910	0,976	0,934
EFF CV	0,946	0,941	0,928	0,924	0,934	0,930	0,948	0,936
SENS Pred	0,954	0,955	0,924	0,945	0,915	0,958	0,924	0,944
SPEC Pred	0,966	0,954	0,953	0,924	0,964	0,916	0,978	0,933
EFF Pred	0,960	0,954	0,939	0,934	0,939	0,937	0,950	0,939
***sparse* Soft PLSDA**								
	SNV + mean center		**Derivata prima + mean center**		**Detrend + mean center**		**Derivata seconda + mean center**	
LV	3		5		7		7	
Variabili (LV)	20		20		30		35	
Tot. variabili	60		73		123		101	
	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI	CIMICI	SFONDI
SENS CV	0,929	0,952	0,915	0,941	0,928	0,953	0,922	0,938
SPEC CV	0,961	0,929	0,964	0,918	0,960	0,928	0,976	0,934
EFF CV	0,945	0,941	0,939	0,929	0,944	0,941	0,948	0,936
SENS Pred	0,949	0,953	0,925	0,941	0,930	0,956	0,930	0,939
SPEC Pred	0,966	0,953	0,960	0,927	0,963	0,930	0,974	0,935
EFF Pred	0,957	0,953	0,942	0,934	0,946	0,943	0,952	0,937

Tabella 5.3 *Tabella riepilogativa dei risultati ottenuti per i modelli di classificazione PLS-DA, Soft PLS-DA e sparse Soft PLS-DA considerando diversi pretrattamenti.*

In Tabella 5.4 vengono riportati i valori di TP, TN, FN ed FP riferiti alla classe *H. halys*, calcolati in cross-validazione e predizione sul test set a partire dai migliori modelli PLS-DA, Soft PLS-DA e *sparse* Soft PLS-DA. In predizione i tre algoritmi di classificazione hanno fornito risultati analoghi; tuttavia, un approccio di tipo Soft PLS-DA è più indicato per applicazioni pratiche perché permette di identificare eventuali spettri *outlier*, mentre il modello *sparse* Soft PLS-DA unisce ai vantaggi di Soft PLS-DA la selezione delle variabili spettrali maggiormente utili ai fini della classificazione.

Inoltre, in Tabella 5.4 gli spettri del test set appartenenti agli sfondi e classificati erroneamente sono stati suddivisi in base alla tipologia di sfondo. Per tutti e tre i modelli la maggior parte delle classificazioni errate

avviene per gli sfondi rami e terra, ma è necessario sottolineare che si tratta di un numero molto piccolo di spettri erroneamente classificati in confronto al numero totale di spettri degli sfondi presenti nel test set.

Infine, confrontando i risultati ottenuti dai modelli PLS-DA costruiti a partire dal Dataset I e dal Dataset II è possibile notare un netto aumento dell'efficienza di classificazione, grazie all'acquisizione di un dataset esteso e maggiormente rappresentativo delle classi considerate.

PLS-DA					**Soft PLS-DA**							**sparse Soft PLS-DA**						
	TP	**TN**	**FN**	**FP**		**TP**	**TN**	**FN**	**FP**	**NA**			**TP**	**TN**	**FN**	**FP**	**NA**	
										Cimici	Sfondi						Cimici	Sfondi
CV	3921	4024	279	176	**CV**	3924	3980	276	175	0	45	**CV**	3903	3999	297	166	0	35
PRED	2671	2701	129	99	**PRED**	2671	2673	129	96	0	31	**PRED**	2656	2668	131	96	13	36
Tipologia Sfondi classificati erroneamente in predizione					**Tipologia Sfondi classificati erroneamente in predizione**							**Tipologia Sfondi classificati erroneamente in predizione**						
			FP					**FP**	**NA**						**FP**	**NA**		
Corteccia			7		**Corteccia**			6	7			**Corteccia**			7	3		
Erba			15		**Erba**			14	1			**Erba**			14	0		
Foglie Gialle			12		**Foglie Gialle**			12	7			**Foglie Gialle**			16	11		
Foglie Secche			0		**Foglie Secche**			0	2			**Foglie Secche**			0	6		
Foglie Verdi			1		**Foglie Verdi**			1	0			**Foglie Verdi**			1	0		
Rami			25		**Rami**			24	9			**Rami**			23	9		
Terra			39		**Terra**			39	5			**Terra**			35	7		

Tabella 5.4 *Valori di TP, TN, FN ed FP relativi alla classe H. halys per i modelli di PLS-DA, Soft PLS-DA e sparse Soft PLS-DA, ottenuti grazie al pretrattamento SNV e mean centering.*

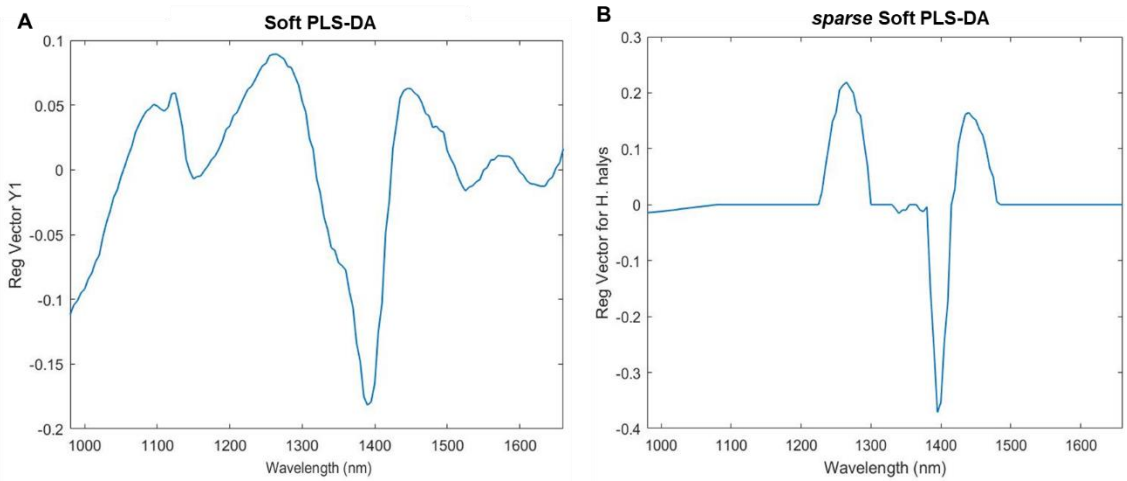

Figura 5.7 *Vettori dei coefficienti di regressione dei modelli Soft PLS-DA (A) e sparse Soft PLSDA (B) per la classe cimici H. halys.*

L'applicazione degli algoritmi Soft PLS-DA e *sparse* Soft PLS-DA ha permesso di ottenere modelli con efficienza di classificazione paragonabili ai modelli PLS-DA, massimizzando la capacità discriminante e, al tempo stesso, delineando una serie di limiti a ciascuna classe modellata in modo da recludere i campioni non appartenenti alle classi cimici e sfondi in una terza classe di campioni "non assegnati".

A tal proposito, in Figura 5.7 sono riportati i vettori di regressione per *H. halys* ottenuti per i modelli Soft PLS-DA (A) e *sparse* Soft PLS-DA (B). In quest'ultimo caso, le regioni spettrali selezionate dall'algoritmo sono comprese negli intervalli 1220-1300 nm (banda di combinazione del legame C-H), 1370-1410 nm (banda di combinazione del legame C-H e primo *overtone* del legame O-H) e 1420-1480 nm (primo *overtone* del legame O-H, terzo *overtone* dello *stretching* del legame C=O e primo *overtone* del legame N-H). Inoltre, si può osservare la presenza di un ulteriore intervallo, compreso nel range 1000-1070 nm e corrispondente al secondo *overtone* del legame N-H, che pur essendo selezionato dall'algoritmo presenta valori molto vicini allo zero risultando, pertanto, meno influente nella classificazione rispetto agli altri intervalli.

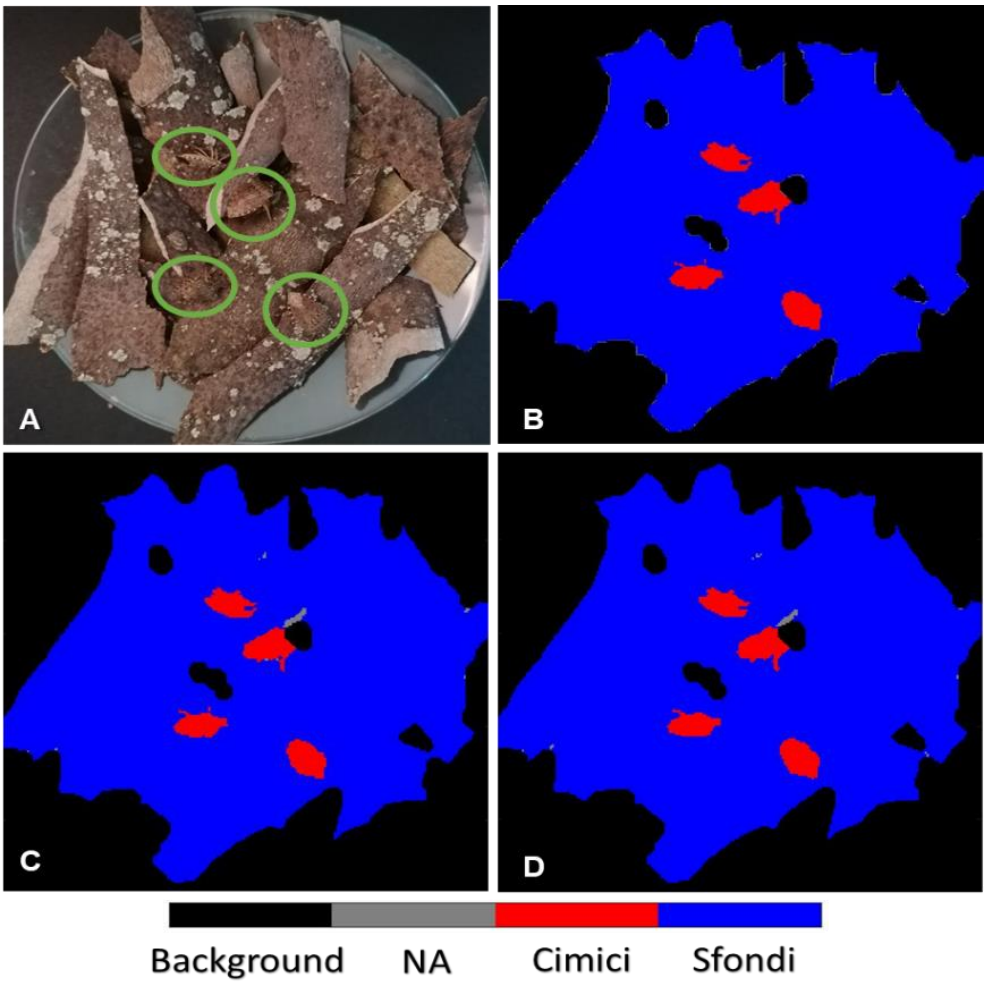

Figura 5.8 *Immagini in predizione del campione Corteccia_HH_G5 acquisito in Fase II: immagine RGB di riferimento (A), immagini ottenute grazie ai modelli PLS-DA (B), Soft PLS-DA (C) e sparse Soft PLS-DA (D) costruiti a partire dalle immagini acquisite durante la fase II.*

Infine, come ulteriore validazione dei modelli di classificazione, è stata effettuata la predizione a livello dei pixel andando ad applicare i modelli calcolati direttamente sulle immagini iperspettrali acquisite e visualizzando le corrispondenti immagini in predizione. A titolo di esempio, in Figura 5.8 sono riportate le immagini in predizione ottenute a partire dai modelli PLS-DA, Soft PLS-DA e *sparse* Soft PLS-DA per un'immagine iperspettrale con corteccia come sfondo. In questo caso per tutti e tre i metodi di classificazione i pixel relativi allo sfondo e alla cimice vengono correttamente assegnati.

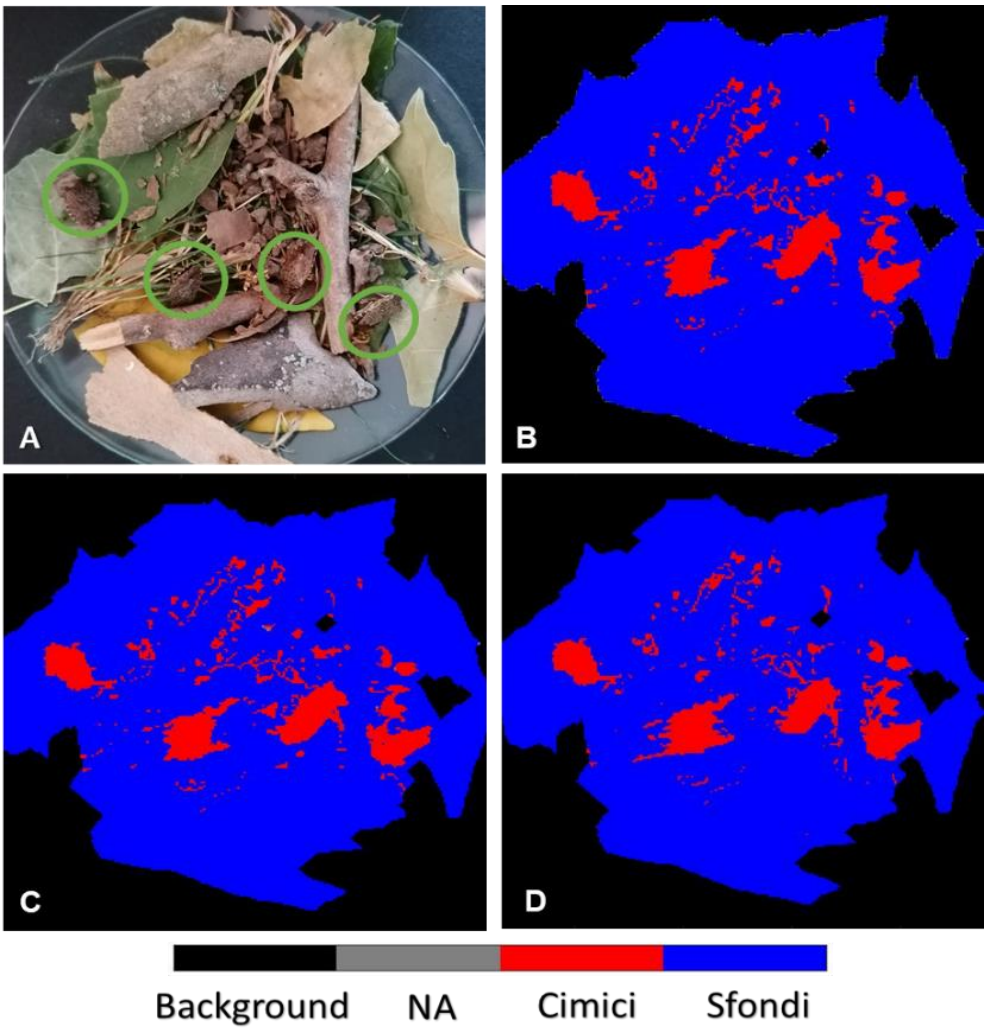

Figura 5.9 *Immagini in predizione del campione Misto_HH_G4 acquisito in Fase II: immagine RGB di riferimento (A), immagini ottenute grazie ai modelli PLS-DA (B), Soft PLS-DA (C) e sparse Soft PLS-DA (D) costruiti a partire dalle immagini acquisite durante la fase II.*

Nella Figura 5.9 sono invece riportate le immagini in predizione ottenute a partire da un'immagine iperspettrale con sfondo misto. In questo caso la classificazione risulta leggermente meno efficace in quanto una parte dei pixel dello sfondo viene attribuita alla classe *H. halys*. Tuttavia, l'immagine in predizione ottenuta a partire dal modello *sparse* Soft PLS-DA presenta un numero minore di pixel classificati erroneamente, evidenziando quindi l'efficacia del modello. Inoltre, in sviluppi futuri su può pensare di applicare operatori morfologici, quali ad esempio l'*erosion*, sulle immagini in predizione per ottenere risultati più affidabili.

Infine, i modelli sviluppati a partire dal Dataset II sono stati applicati anche alle immagini iperspettrali acquisite in Fase I. In Figura 5.10 sono riportate le immagini in predizione ottenute applicando all'immagine FoglieVerdi_HH_2, acquisita in Fase I, i seguenti modelli: PLS-DA costruito sulle immagini acquisite durante la fase I (B), PLS-DA costruito sulle immagini acquisite durante la fase II (C), Soft PLS-DA costruito sulle immagini acquisite durante la fase II (D) e *sparse* Soft PLS-DA costruito sulle immagini acquisite durante la fase II (E). In questo caso, i modelli di classificazione costruiti partendo dal Dataset II hanno consentito di

migliorare nettamente i risultati in predizione ottenuti sull'immagine iperspettrale acquisita durante la Fase I con foglie verdi come sfondo. Viceversa, per sfondi vegetali come corteccia o rami acquisiti durante la fase I non sono stati raggiunti risultati così soddisfacenti ed in questo caso l'influenza dell'umidità della matrice vegetale campionata potrebbe rappresentare una criticità.

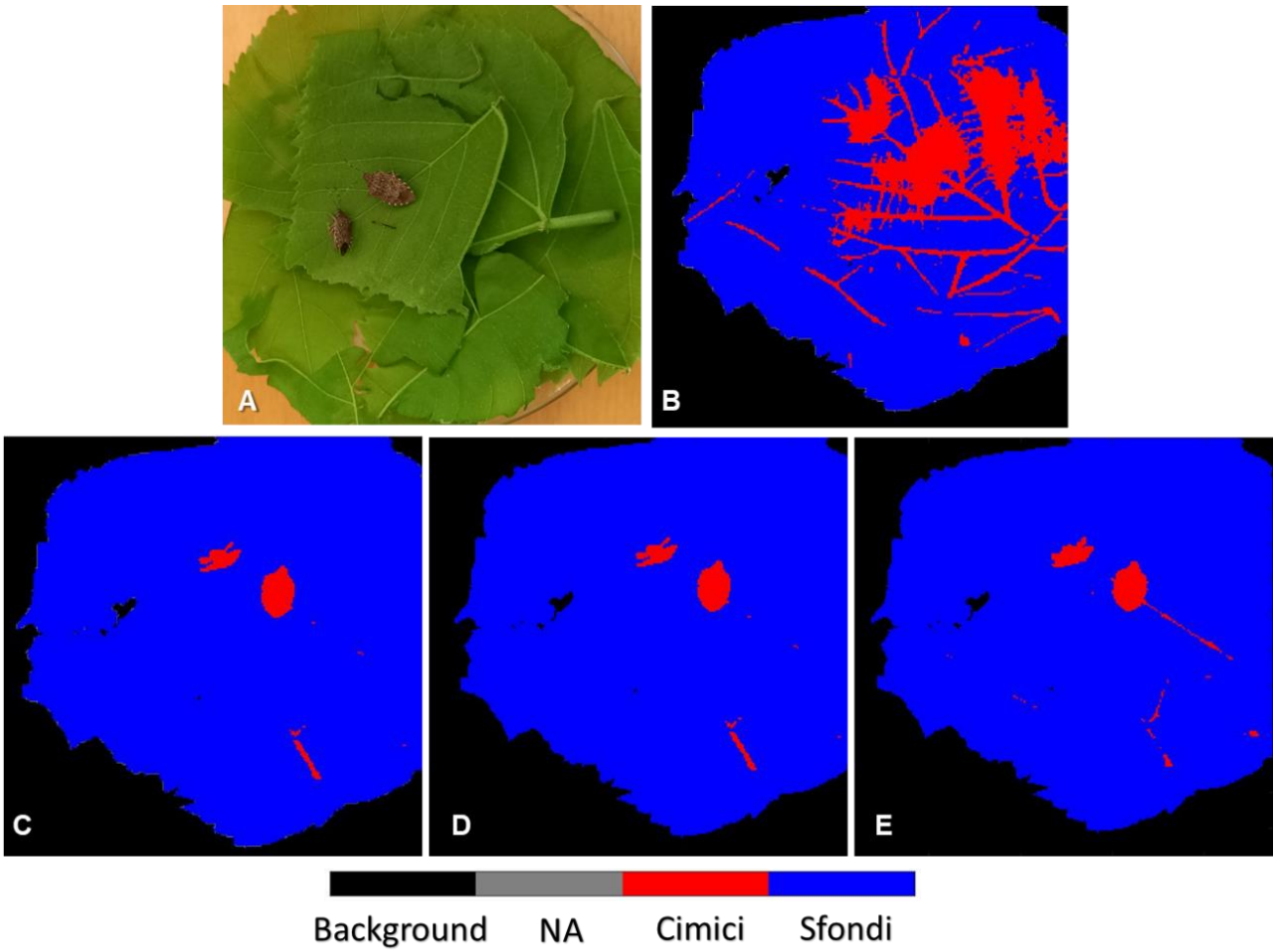

Figura 5.10 *Immagini in predizione del campione FoglieVerdi_HH_2 acquisito in Fase I: l'immagine RGB di riferimento (A), immagine ottenuta grazie al modello PLS-DA costruito sul Dataset I (B) e le immagini ottenute grazie ai modelli PLS-DA (C), Soft PLS-DA (D) e sparse Soft PLS-DA (E), costruiti utilizzando il Dataset II.*

6 Conclusioni e prospettive future

Nel presente lavoro di tesi è stato possibile valutare l'efficacia dell'imaging iperspettrale nel vicino infrarosso per identificare esemplari di cimice asiatica (*H. halys*) su diverse tipologie di sfondi vegetali, nell'ottica dello sviluppo di sistemi di monitoraggio automatizzati da utilizzare in campo.

L'imaging iperspettrale nel vicino infrarosso consente di ottenere sia l'informazione spettrale relativa alla composizione chimica sia l'informazione spaziale relativa alla distribuzione dei pixel corrispondenti ad *H. halys*, attraverso una procedura di analisi rapida e non distruttiva. Grazie ai suoi numerosi vantaggi, questa tecnica risulta avere promettenti potenzialità per la rilevazione in campo tramite l'implementazione di metodi di monitoraggio automatizzati. Per contro, ogni immagine iperspettrale è costituita da un'elevata mole di dati per cui risultano necessari metodi di analisi multivariata delle immagini selezionati *ad hoc* allo scopo di estrarre le informazioni utili.

Nel presente caso studio, l'analisi esplorativa e la successiva costruzione di modelli di classificazione a partire dalle immagini iperspettrali di esemplari di cimice asiatica su diverse tipologie di sfondi vegetali hanno consentito di ottenere risultati incoraggianti al fine di rilevare gli insetti durante il monitoraggio in campo.

In particolare, una prima fase sperimentale condotta su un numero ristretto di immagini ha evidenziato alcune criticità dal lato sperimentale e computazionale, tra cui la necessità di utilizzare per le acquisizioni esemplari di cimice deceduti da poche ore al fine di simulare al meglio le condizioni delle acquisizioni in campo e la necessità di costruire i modelli di classificazione a partire da dataset di spettri delle due classi *H. halys* e sfondi vegetali il più possibile rappresentativi della variabilità delle classi in esame.

Sulla base di quanto emerso nella prima fase sperimentale, nella seconda fase è stato acquisito un dataset esteso, costituito da un numero maggiore di immagini iperspettrali di esemplari di cimice asiatica su sfondi vegetali. Grazie ad un diverso approccio di campionamento degli spettri di *H. halys* e sfondi da utilizzare per il calcolo dei modelli di classificazione, nella seconda fase sono stati ottenuti risultati migliori. Inoltre, le immagini acquisite nella seconda fase sono state impiegate nella costruzione di modelli di classificazione calcolati con l'algoritmo Soft PLS-DA accoppiato ad un approccio di tipo *sparse* per la selezione delle variabili maggiormente rilevanti ai fini della discriminazione.

L'utilizzo dell'algoritmo *sparse* Soft PLS-DA ha permesso di ottenere modelli più parsimoniosi, ovvero che considerano un numero inferiore di variabili spettrali, e più flessibili, grazie alla possibilità di identificare eventuali spettri *outlier* pur mantenendo allo stesso tempo ottime performance di classificazione.

Al fine di ottimizzare ulteriormente l'efficienza dei modelli di classificazione in funzione di un'eventuale applicazione in campo, sarà necessario considerare all'interno dei modelli stessi l'effetto dell'umidità, ed in generale delle condizioni metereologiche, per migliorarne la robustezza.

Le zone selezionate potranno in seguito essere utilizzate per lo sviluppo di sistemi di imaging multispettrale specifici per il monitoraggio in campo di *H. halys*. Infatti, i sistemi di imaging multispettrale, considerando

soltanto determinate regioni spettrali, risultano più veloci ed economici rispetto ai sistemi iperspettrali e, quindi, più adatti per l'applicazione pratica.

Bibliografia

- Acebes-Doria, A. L., Morrison, W. R., Short, B. D., Rice, K. B., Bush, H. G., Kuhar, T. P., Duthie, C., & Leskey, T. C. (2018). Monitoring and biosurveillance tools for the brown marmorated stink bug, *Halyomorpha halys* (Stål) (Hemiptera: Pentatomidae). *Insects*, *9*(3), 1–17. https://doi.org/10.3390/insects9030082
- Amigo, J. M., Babamoradi, H., & Elcoroaristizabal, S. (2015). Hyperspectral image analysis. A tutorial. *Analytica Chimica Acta*, *896*, 34–51. https://doi.org/10.1016/j.aca.2015.09.030
- Anderson, B. E., Miller, J. J., & Adams, D. R. (2012). Irritant contact dermatitis to the brown marmorated stink bug, *Halyomorpha halys*. *Dermatitis*, *23*(4), 170–172.
- Anfora, G., Ioriatti, C., & Mazzoni, V. (2019). Strumenti per il monitoraggio e controllo di *Halyomorpha halys* basati su semiochimici e vibrazioni, ed esperienze di "citizen science." *Atti Accademia Nazionale Italiana Di Entomologia*, *Anno LXVII*, 119–123.
- Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790–3798.
- Ballabio, D., & Todeschini, R. (2009). Multivariate classification for qualitative analysis. *Infrared Spectroscopy for Food Quality Analysis and Control*, *83*, e102.
- Bariselli, M., Bugiani, R., & Maistrello, L. (2016). Distribution and damage caused by *Halyomorpha halys* in Italy. *EPPO Bulletin*, *46*(2), 332–334. https://doi.org/10.1111/epp.12289
- Barker, M., & Rayens, W. (2003). *Partial least squares for discrimination*. 166–173. https://doi.org/10.1002/cem.785
- Berg, F. Van Den, & Engelsen, S. B. (2009). *Review of the most common pre-processing techniques for near-infrared spectra*. *28*(10). https://doi.org/10.1016/j.trac.2009.07.007
- Bergmann, E. J., Venugopal, P. D., Martinson, H. M., Raupp, M. J., & Shrewsbury, P. M. (2016). Host plant use by the invasive *Halyomorpha halys* (Stål) on woody ornamental trees and shrubs. *PloS One*, *11*(2), e0149975.
- Brereton, R. G. (2003). *Chemometrics: data analysis for the laboratory and chemical plant* (Vol. 24, Issue 5). John Wiley & Sons.
- Buffington, M. L., Talamas, E. J., & Hoelmer, K. (2018). Brown Marmorated Stink Bug. American Entomologist, 64(4), 225.
- Burns, D. A., & Ciurczak, E. W. (2008). *Handbook of near-infrared analysis* (Chapter 17, pp. 356-357). Boca Raton, Florida, USA. CRC Press 978-0-8493-7393-0
- Calvini, R. (2016). *Chemometric Tools for Food Characterization Through RGB and Hyperspectral Imaging*.
- Calvini, R., Orlandi, G., Foca, G., & Ulrici, A. (2018). Developmentof a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging. *Journal of Spectral Imaging*, *7*, 1–15. https://doi.org/10.1255/jsi.2018.a13

- Calvini, R., Ulrici, A., & Amigo, J. M. (2015). Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems*, *146*, 503–511. https://doi.org/10.1016/j.chemolab.2015.07.010
- Centro Servizi Ortofrutticoli (CSO). (2020). *Estimation of damage from brown marmorated stink bug and plant pathologies related to climate change*. http://www.csoservizi.com
- Costi, E., Haye, T., & Maistrello, L. (2017). Biological parameters of the invasive brown marmorated stink bug, Halyomorpha halys, in southern Europe. *Journal of Pest Science*, *90*(4), 1059–1067. https://doi.org/10.1007/s10340-017-0899-z
- Cozzi, R., Protti, P., & Ruaro, T. (1998). *Elementi di analisi chimica strumentale*. Zanichelli.
- Davies, A. M. C., & Fearn, T. (2007). *Back to basics : removing multiplicative effects (1) The world leader in IR spectroscopy with a big reputation*. *19*(4), 2–6.
- Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, *50*, 96–106.
- FAOSTAT. (2019). *Food and Agriculture Organization Corporate Statistical Database*. http://www.fao.org/faostat/en/#data/QC/visualize
- Ficetola, G. F., Bonin, A., & Miaud, C. (2008). Population genetics reveals origin and number of founders in a biological invasion. *Molecular Ecology*, *17*(3), 773–782. https://doi.org/10.1111/j.1365-294X.2007.03622.x
- Filzmoser, P., Gschwandtner, M., & Todorov, V. (2012). *Review of sparse methods in regression and classi fi cation with application to chemometrics*. *November 2011*, 42–51. https://doi.org/10.1002/cem.1418
- Fox, G. (2020). The Brewing Industry and the Opportunities for Real-Time Quality Analysis Using Infrared Spectroscopy. *Applied Sciences*, *10*. https://doi.org/10.3390/app10020616
- Francati, S., Masetti, A., Martinelli, R., Mirandola, D., Anteghini, G., Busi, R., Dalmonte, F., Spinelli, F., Burgio, G., & Dindo, M. L. (2021). *Horticultural Entomology Halyomorpha halys (Hemiptera : Pentatomidae) on Kiwifruit in Northern Italy : Phenology , Infestation , and Natural Enemies Assessment*. *114*(July), 1733–1742. https://doi.org/10.1093/jee/toab126
- Gallagher, N. B. (2020). *Savitzky-Golay Smoothing and Differentiation Filter*. 1–2.
- Gemperline, P. (2006). *Practical guide to chemometrics*. CRC press.
- Gowen, A. (2013). *Hyperspectral Imaging and Chemometrics : A Perfect Combination for the Analysis of Food Structure , Composition and Quality Provided for non-commercial research and educational use only . Not for reproduction , distribution or commercial use . July*. https://doi.org/10.1016/B978-0-444-59528-7.00009-0
- Gowen, A. A., Marini, F., Esquerre, C., Donnell, C. O., Downey, G., & Burger, J. (2011). Analytica Chimica Acta Time series hyperspectral chemical imaging data : Challenges , solutions and applications. *Analytica Chimica Acta*, *705*(1–2), 272–282. https://doi.org/10.1016/j.aca.2011.06.031
- Grahn, H. F., & Geladi, P. (2007). Techniques and Applications of Hyperspectral Image Analysis. In *Techniques and Applications of Hyperspectral Image Analysis*. https://doi.org/10.1002/9780470010884

- *Gruppo divisionale di Chemiometria*. https://www.gruppochemiometria.it/index.php/la-chemiometria
- Hamilton, G C, Shearer, P. W., & Nielsen, A. L. (2008). Brown marmorated stink bug–a non-native insect in New Jersey. *Fact Sheet FS002. Rutgers NJAES Cooperative Extension. Rutgers University, New Brunswick, New Jersey, USA.*
- Hamilton, George C. (2009). Brown marmorated stink bug. *American Entomologist*, *55*(1), 19–20.
- Harris, D. C. (2010). *Quantitative chemical analysis*. Macmillan.
- Haye, T. (2019). Global pest status of *Halyomorpha halys* and impact of its associated parasitoids. *Atti Accademia Nazionale Italiana Di Entomologia*, *Anno LXVII*, 95–100.
- Haye, T., Gariepy, T., Hoelmer, K., Xavier, J. S., Nicolas, T., Rossi, J.-P., Streito, J.-C., Tassus, X., Desneux, N., Xavier, J. S., & Nicolas, T. (2015). Range expansion of the invasive brown marmorated stinkbug, *Halyomorpha halys*: an increasing threat to field, fruit and vegetable crops worldwide. *Journal of Pest Science*, *88*(4), 665–673. https://doi.org/10.1007/s10340-015-0670-2
- Haye, T., & Weber, D. C. (2017). Special issue on the brown marmorated stink bug, *Halyomorpha halys*: an emerging pest of global concern. *Journal of Pest Science*, *90*(4), 987–988. https://doi.org/10.1007/s10340-017-0897-1
- Haye, T, Wyniger, D., & Gariepy, T. (2014). Recent range expansion of brown marmorated stink bug in Europe. *Proceedings of the 8th International Conference on Urban Pests, 20-23 July 2014, Zurich, Switzerland*, 309–314.
- Hoebeke, E. R., & Carter, M. E. (2003). *Halyomorpha halys* (Stål)(Heteroptera: Pentatomidae): a polyphagous plant pest from Asia newly detected in North America. *Proceedings of the Entomological Society of Washington*, *105*(1), 225–237.
- Huang, D., Haack, R. A., & Zhang, R. (2011). Does global warming increase establishment rates of invasive alien species? A centurial time series analysis. *PloS One*, *6*(9), e24733.
- Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*(1), 137–148.
- Lee, D. H., Short, B. D., Joseph, S. V., Bergh, J. C., & Leskey, T. C. (2013). Review of the biology, ecology, and management of *Halyomorpha halys* (Hemiptera: Pentatomidae) in China, Japan, and the Republic of Korea. *Environmental Entomology*, *42*(4), 627–641. https://doi.org/10.1603/EN13006
- Lee, D.-H., Wright, S. E., Boiteau, G., Vincent, C., & Leskey, T. C. (2013). Effectiveness of glues for harmonic radar tag attachment on *Halyomorpha halys* (Hemiptera: Pentatomidae) and their impact on adult survivorship and mobility. *Environmental Entomology*, *42*(3), 515–523.
- Leskey, T. C., & Nielsen, A. L. (2018). Impact of the invasive brown marmorated stink bug in North America and Europe: history, biology, ecology, and management. *Annual Review of Entomology*, *63*, 599–618.
- Lima, M. C. F., Leandro, M. E. D. de A., Valero, C., Coronel, L. C. P., & Bazzo, C. O. G. (2020). Automatic detection and monitoring of insect pests—A review. *Agriculture (Switzerland)*, *10*(5). https://doi.org/10.3390/agriculture10050161

- Maistrello, L., Costi, E., Stefano, C., Vaccari, G., Bortolotti, P., Nannini, R., Casoli, L., Montermini, A., Bariselli, M., & Guidetti, R. (2016). *Halyomorpha halys* in Italy: first results of field monitoring in fruit orchards. *Integrated Protection of Fruit Crops*, *112* (January), 1–5.
- Maistrello, L, Dioli, P., Vaccari, G., Nannini, R., Bortolotti, P., Caruso, S., Costi, E., Montermini, A., Casoli, L., & Bariselli, M. (2014). *First records in Italy of the Asian stinkbug Halyomorpha halys, a new threat for fruit crops.* Alma Mater Studiorum, Universitá di Bologna. https://www.cabi.org/isc/abstract/20163169705
- Maistrello, L. (2019). *Biologia e diffusione di Halyomorpha halys, l'autostoppista invasivo che sconvolge la difesa integrata. Atti Accademia Nazionale Italiana Di Entomologia*, *Anno LXVII*, 95–100.
- Maistrello, L., Dioli, P., Bariselli, M., Mazzoli, G. L., & Giacalone-Forini, I. (2016). Citizen science and early detection of invasive species: phenology of first occurrences of *Halyomorpha halys* in Southern Europe. *Biological Invasions*, *18*(11), 3109–3116.
- Maistrello, L., Dioli, P., Dutto, M., Volani, S., Pasquali, S., & Gilioli, G. (2018). Tracking the spread of sneaking aliens by integrating crowdsourcing and spatial modeling: the Italian invasion of *Halyomorpha halys*. *BioScience*, *68*(12), 979–989.
- Maistrello, L., Vaccari, G., Caruso, S., Costi, E., Bortolini, S., Macavei, L., Foca, G., Ulrici, A., Bortolotti, P. P., & Nannini, R. (2017). Monitoring of the invasive *Halyomorpha halys*, a new key pest of fruit orchards in northern Italy. *Journal of Pest Science*, *90*(4), 1231–1244.
- Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. *Chemical Society Reviews*, *43*(24), 8200–8214. https://doi.org/10.1039/c4cs00062e
- Miller, J. N., Miller, J. C. and Miller, R. D. (2004). Statistics and Chemometrics for Analytical Chemistry. In *Technometrics* (Vol. 46, Issue 4). https://doi.org/10.1198/tech.2004.s248.
- Mobaraki, N., & Amigo, J. M. (2018). *Chemometrics and Intelligent Laboratory Systems HYPER-Tools . A graphical user-friendly interface for hyperspectral image analysis*. *172* (September 2017), 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003
- Morrison III, W. R., Blaauw, B. R., Short, B. D., Nielsen, A. L., Bergh, J. C., Krawczyk, G., Park, Y., Butler, B., Khrimian, A., & Leskey, T. C. (2019). Successful management of *Halyomorpha halys* (Hemiptera: Pentatomidae) in commercial apple orchards with an attract-and-kill strategy. *Pest Management Science*, *75*(1), 104–114.
- Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification* (Vol. 6). NIR Chichester.
- Piemontese, L., Cesari, M., Ganzerli, F., Maistrello, L., Dioli, P., Rebecchi, L., & Guidetti, R. (2016). *Specie aliene invasive: il caso della cimice bruna marmorizzata Halyomorpha halys (Heteroptera, Pentatomidae) in Italia e nel territorio modenese*.
- Pirhadi, S., Shiri, F., & Ghasemi, J. B. (2015). Multivariate statistical analysis methods in QSAR. *RSC Advances*, *5*(127), 104635–104665. https://doi.org/10.1039/C5RA10729F
- Polajnar, J., Maistrello, L., Bertarella, A., & Mazzoni, V. (2016). Vibrational communication of the brown

marmorated stink bug (*Halyomorpha halys*). *Physiological Entomology*, *41*(3), 249–259.

- Pyšek, P., Jarošík, V., Hulme, P. E., Kühn, I., Wild, J., Arianoutsou, M., Bacher, S., Chiron, F., Didžiulis, V., & Essl, F. (2010). Disentangling the role of environmental and human pressures on biological invasions across Europe. *Proceedings of the National Academy of Sciences*, *107*(27), 12157–12162.
- Qin, J., Kim, M. S., Chao, K., Chan, D. E., Delwiche, S. R., & Cho, B. K. (2017). Line-scan hyperspectral imaging techniques for food safety and quality applications. *Applied Sciences (Switzerland)*, *7*(2). https://doi.org/10.3390/app7020125
- Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639.
- Scaccini, D., Moore, L., Tirello, P., Fornasiero, D., Cecchetto, M., Duso, C., & Pozzebon, A. (2019). Andamento delle popolazioni e dannosità di Halyomorpha halys sulle colture agrarie. *Atti Accademia Nazionale Italiana Di Entomologia*, *Anno LXVII* (XXXV. Halyomorpha halys: nuove acquisizioni e applicazioni nella difesa), 113–118.
- Sciarretta, A., & Calabrese, P. (2019). *Development of Automated Devices for the Monitoring of Insect Pests*. *7*(1).
- Skoog, D. A., Holler, F. J., & Crouch, S. R. (2017). *Principles of instrumental analysis*. Cengage learning.
- Ulrici, A., Serranti, S., Ferrari, C., Cesare, D., Foca, G., & Bonifazi, G. (2013). Efficient chemometric strategies for PET-PLA discrimination in recycling plants using hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems*, *122*, 31–39. https://doi.org/10.1016/j.chemolab.2013.01.001
- Watanabe, M. (1994). Overwintering flight of brown-marmorated stink bug, *Halyomorpha mista* to the buildings. *Med. Entomol. Zool.*, *45*, 25–31.
- Weber, D. C., Morrison, W. R., Khrimian, A., Rice, K. B., Leskey, T. C., Rodriguez-Saona, C., Nielsen, A. L., & Blaauw, B. R. (2017). Chemical ecology of *Halyomorpha halys*: discoveries and applications. *Journal of Pest Science*, *90*(4), 989–1008.
- Wu, D., & Sun, D. (2013). *Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment : A review — Part I : Fundamentals*. *19*, 1–14. https://doi.org/10.1016/j.ifset.2013.04.014
- Wyniger, D., & Kment, P. (2010). Key for the separation of *Halyomorpha halys* (Stål) from similar-appearing pentatomids (Insecta: Heteroptera: Pentatomidae) occurring in Central Europe, with new Swiss records. *Mitteilungen Der Schweizerischen Entomologischen Gesellschaft*, *83*(3/4), 261–270.
- Xu, J., Fonseca, D. M., Hamilton, G. C., Hoelmer, K. A., & Nielsen, A. L. (2014). Tracing the origin of US brown marmorated stink bugs, *Halyomorpha halys*. *Biological Invasions*, *16*(1), 153–166.

Ringraziamenti

Con immensa gratitudine desidero ringraziare le persone che hanno reso possibile la mia partecipazione al progetto HALY.ID e che hanno contribuito alla realizzazione del presente lavoro di tesi. Un sentito ringraziamento al Prof. Alessandro Ulrici, relatore, per aver supervisionato il presente elaborato e per avermi concesso l'opportunità di approfondire diversi aspetti della materia. Un ringraziamento speciale alla Dott.ssa Rosalba Calvini, correlatrice, la quale mi ha supportato costantemente durante gli ultimi mesi, rendendosi disponibile al fine di chiarire qualsiasi dubbio, oltre che a fornire un grande contributo per la stesura dell'elaborato finale. Un ringraziamento alla Prof.ssa Lara Maistrello e ai suoi collaboratori del laboratorio di entomologia applicata BIOGEST-SITEIA per aver fornito i campioni utilizzati per l'acquisizione.

Inoltre, vorrei ringraziare i professori del corso di Controllo e Sicurezza degli Alimenti, i quali hanno saputo trasmettere competenze e professionalità attraverso i loro insegnamenti nonostante le difficoltà dovute all'emergenza sanitaria degli ultimi due anni.

Infine, desidero ringraziare Leo e Alice per il supporto costante, Elisa, Emma, Martina e Jacopo per avermi accompagnato durante il percorso universitario e la mia famiglia per il sostegno ricevuto.