ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

CAMPUS DI CESENA

DIPARTIMENTO DI
INGEGNERIA DELL'ENERGIA ELETTRICA E DELL'INFORMAZIONE
"GUGLIELMO MARCONI"

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA BIOMEDICA

Determinazione dell'umidità del terreno con tecnica in guida d'onda

Tesi in

Sensori e Nanotecnologie Lm

Relatore

Prof. *Marco Tartagni*

Presentata da

Leonardo Franceschelli

Correlatori:

Prof. *Luigi Ragni*

Dott.ssa *Annachiara Berardinelli*

Sessione III
Anno Accademico 2017/2018

Abstract

Il presente lavoro di tesi riguarda la creazione di un modello di calibrazione tramite regressione PLS per un sensore non invasivo in grado di rilevare il grado di umidità del terreno. Il modello è stato successivamente implementato all'interno del firmware del MCU presente all'interno del sensore.

Il modello sviluppato ha permesso di calcolare l'umidità direttamente dai dati ottenuti tramite la misurazione dell'impedenza complessa di segnali a RF riflessi dal terreno, seguendo i principi della spettroscopia d'impedenza.

Il lavoro ha avuto inizio con la raccolta di dati spettrali sperimentali, è proseguito con la creazione del modello di calibrazione e si è concluso con l'implementazione nel firmware del calcolo dell'umidità percentuale.

Indice

Introduzione

La quantità di acqua contenuta nel suolo rappresenta solo una piccola frazione dell'acqua dolce globale (0.15%), ma è di fondamentale importanza in diversi processi idrogeologici e biochimici. L'umidità del suolo, influenzandone l'evaporazione, modula le interazioni tra la superficie del terreno e l'atmosfera, agendo quindi sul clima e sulle condizioni meteorologiche. La presenza di un alto livello di umidità del suolo è un aspetto fondamentale per il corretto sviluppo e mantenimento di un ecosistema naturale: una sua mancanza porta inevitabilmente alla desertificazione. Come conseguenza di ciò, il livello di acqua nel suolo è un prerequisito della produzione primaria, permettendo uno scambio di nutrienti costante.
La monitorazione e il mantenimento di questo parametro aumentano notevolmente la produttività di questi ecosistemi e aiutano a preservare la biodiversità[1].

Il progetto alla base della tesi si pone in questo contesto applicativo, avendo il sensore su cui si è lavorato lo scopo di misurare la percentuale di umidità del terreno in maniera veloce e non invasiva, al contrario delle tecniche oggi più diffuse. Grazie al modello di calibrazione sviluppato in questo lavoro e successivamente implementato nel MCU del sensore, è possibile misurare l'umidità senza dover inserire alcun strumento nel terreno e con tempi di misurazione dell'ordine di 1-2 minuti.

L'elaborato descrive tutte le fasi del lavoro svolto, partendo dalle basi teoriche, per poi descrivere la creazione del modello di calibrazione sperimentale e la sua successiva implementazione.

Nel primo capitolo viene descritto il contesto generale, descrivendo inizialmente le definizioni di umidità e le formule con cui è possibile calcolarla, per poi fare una panoramica sulle tecniche oggi usate per la misurazione di questo parametro, sia dirette che indirette. Per ognuna di queste vengono indicati i pregi e messe in luce le problematiche.

Nel secondo capitolo vengono fornite le basi della spettroscopia d'impedenza, partendo da una descrizione di due parametri fondamentali per questa tecnica, ovvero la conducibilità e la permettività elettrica. Vengono successivamente presentate le caratteristiche della trasmissione lungo una guida d'onda e della loro riflessione sul terreno, concludendo con una trattazione del problema nel campo delle frequenze, che permette una notevole semplificazione dei calcoli.

Nel terzo capitolo viene presentata la PLS Regression, ovvero l'analisi statistica con cui si è creato il modello di regressione. Vengono delineati i principi fondanti di questa tecnica, per poi descrivere una trattazione matematica più dettagliata, compresi gli algoritmi più usati nei software di calcolo. Si descrivono successivamente i procedimenti fondamentali per la corretta creazione e validazione di un modello, descrivendo anche i parametri di maggiore interesse per la diagnostica della capacità di predizione.

Nel quarto capitolo viene descritto il sensore, partendo dai suoi principi di funzionamento e dai componenti del circuito elettronico, di cui vengono indicate le caratteristiche tecniche più importanti e il ruolo che svolgono durante il processo di misura ed elaborazione. Viene trattato successivamente il firmware del MCU, descrivendone le operazioni principale e le funzioni con le quali sono state implementate. Viene infine presentato lo script di Matlab, grazie al quale vengono letti i valori di interessi da seriale e creata un'interfaccia grafica che permette all'utente di utilizzare facilmente il sensore.

Nel quinto e ultimo capitolo vengono presentati tutti i procedimenti che hanno portato alla creazione e validazione del modello di calibrazione tramite PLSR e alla sua successiva implementazione nel firmware. Viene quindi descritta la raccolta di dati su campioni di terreno, dei quali viene misurata l'umidità grazie alla tecnica gold-standard. Questi dati sono stati usati in varie combinazioni per la creazione di modelli di regressione grazie al software Eigenvector, tra i quali è stato scelto quello con maggiori capacità predittive. Vengono infine descritte le modifiche apportate al firmware e all'intefaccia grafica di Matlab.

Il lavoro si conclude con il riassunto del lavoro svolto sul sensore e descrivendo brevemente alcuni possibili studi futuri che potrebbero migliorare questa tecnologia.

Capitolo 1

Umidità: definizioni e tecniche di misura

In questo capitolo viene introdotto quello che è il contesto applicativo del sensore di umidità, dapprima definendo la definizione di umidità e descrivendo tutte le forme in cui può essere calcolata, per poi fare una panoramica sulle tecniche principali utilizzate oggigiorno per la misura dell'umidità locale del terreno, descrivendo sia i metodi diretti che quelli indiretti.

1. Definizioni di umidità

Il contenuto di acqua nel suolo è generalmente definito come il rapporto della massa di acqua M_w e la massa di terreno secco M_s, o come il volume di acqua V_w per il volume totale dell'unità di suolo, V_t. In entrambi i casi il calcolo del valore di contenuto d'acqua dipende dalla definizione di condizione di suolo secco. Tradizionalmente si fa riferimento a una condizione standard ottenuta in laboratorio, estraendo l'acqua da un campione di terreno ponendolo in un forno a una temperatura di circa 100-110 °C (metodo termo-gravitometrico). Sebbene la scelta di questo range di temperatura sia arbitraria, è stato dimostrato che tenere il campione di terreno nel forno per un tempo adeguato e alla temperatura media di 105 °C garantisce l'evaporazione dell'acqua "libera" presente nel suolo (Romano, 1999).

Si può quindi definire il contenuto di acqua del suolo su una base volumetrica grazie a questo rapporto adimensionale:

$$\theta = \frac{V_W}{V_T}$$

ovvero il rapporto del volume di acqua nel suolo (V_W) con il volume totale del suolo (V_T). Quest'ultimo è la somma del volume delle particelle solide (V_S), il volume dell'acqua nel suolo (V_W) e il volume di aria nel suolo (V_a) [2].

In ambito chimico si preferisce spesso esprimere il contenuto d'acqua in base alle masse:

$$\theta_M = \frac{M_W}{M_S}$$

Per un terreno rigido, definendo $\rho_b = M_S/V_T$ la densità apparente e $\rho_W = M_W/M_S$ la densità dell'acqua liquida si possono mettere in relazione tra loro il contenuto di acqua volumetrico θ e il contenuto di acqua gravimetrico θ_M grazie alla seguente espressione:

$$\theta = \theta_M \frac{\rho_b}{\rho_w}$$

Il massimo contenuto di acqua nel terreno, ovvero quando tutti i pori interconnessi sono completamente pieni, viene chiamato come il contenuto di acqua del suolo alla

saturazione, θ_S, che è sempre minore o uguale della porosità del suolo, η, definita come il rapporto tra il volume dei vuoti (V_V) e il volume totale del suolo (V_T). In alcuni casi si preferisce esprimere il contenuto di acqua in termini di grado di saturazione, S, calcolato solitamente in percentuale:

$$S = \frac{V_W}{V_v} \times 100$$

Per tenere conto non solo delle interazioni con l'atmosfera, ma anche di quelle con la vegetazione, le condizioni di umidità del suolo sono caratterizzate spesso sfruttando il concetto di disponibilità di umidità del suolo per la crescita delle piante, definita "plant-available soil water content" (PAWC). Questo è stato storicamente introdotto per scopi agronomici, ma è diventato rapidamente la base per diversi modelli in grado di calcolare il bilancio idrologico e valutare le dinamiche dell'umidità[2].

Per calcolarlo si usa la differenza tra il contenuto di acqua alla "field capacity", θ_{FWC}, (limite superiore) e il contenuto d'acqua al "permanent wilting", θ_{PW}, (limite inferiore) (Denmead e Shaw, 1962). Il punto di permanent wilting rappresenta il contenuto di acqua al quale una pianta appassisce completamente e non è più in grado di recuperare la sua attività biologica quando posta in un ambiente umido.

2. Tecniche di misura dell'umidità del suolo

Esistono diversi metodi per determinare il contenuto di acqua nel suolo, specialmente nelle condizioni di campo. Questi si dividono in due famiglie, metodi diretti o metodi indiretti.

2.1 Metodi diretti

Il metodo diretto più diffuso è il metodo termo-gravitometrico, spesso assunto come una procedura standard in quanto usa strumenti precisi, oltre a essere poco costosa. Richiede tuttavia la distruzione del campione di terreno, non permettendo quindi di ripetere la misura più volte.

Questo metodo consiste nel prelevare un campione di suolo (solitamente 100 o 200 g) alla profondità appropriata, pesarlo e asciugarlo in una stufa a 105 °C. Il campione deve essere tenuto nella stufa il tempo necessario a ottenere un peso stabile, che dipende sia dalle caratteristiche del campione sia da quelle della stufa: si passa da 12 ore in una stufa a convezione forzata fino a 24 ore in una stufa più convenzionale. Dopo aver completato la fase di asciugatura il campione viene prelevato dalla stufa e pesato nuovamente. Il contenuto di acqua gravitometrico è calcolato con la formula:

$$\theta_M = \frac{(W_w + t_a) - (W_d + t_a)}{(W_d + t_a) - t_a}$$

dove w_w e w_d sono le masse del terreno umido e secco e t_a e la massa della tara.

Questo metodo è universalmente riconosciuto come il gold standard per la misura dell'umidità, essendo l'unico che fornisce una misura diretta. I metodi che verranno trattati nella prossima sezione infatti non misurano l'umidità, ma una variabile ad essa correlata[2].

2.2 Metodi indiretti

I metodi indiretti consistono nella misurazione di alcune proprietà fisiche o fisico-chimiche del suolo che sono dipendenti dal contenuto d'acqua. In genere non comprendono metodi distruttivi e usano strumenti che possono essere posti permanentemente nel suolo, o sensori remoti localizzati su piattaforme aeree o satelliti. Sono quindi metodi adatti a eseguire misure ripetute e permettono di registrare i dati in maniera automatica, ma in quasi tutti i casi richiedono la conoscenza di curve di calibrazioni accurate. I migliori metodi indiretti su scala locale sono quelli che sfruttano l'attenuazione dei raggi gamma o l'effetto di scattering dei neutroni e le tecniche di sensing basate su dielettrico. Sono state proposte anche tecniche basate sulla riflessione di raggi infrarossi o sulla risonanza magnetica nucleare(NMR) (Stingaciu e al., 2010; Blümich e al., 2011), che sono tuttavia poco utilizzate.

2.2.1 Tecniche basate su radiazioni

Il primo metodo si basa sull'attenuazione e il backscattering di un raggio collimato di raggi gamma emessi da una sorgente radioattiva, come per esempio il cesio-137(^{137}Cs) (Reginato e Van Bavel, 1964) .

Questo metodo viene molto usato, sebbene sia possibile farlo solo in condizioni di laboratorio, poiché è molto preciso ed è l'unico che permette di misurare il contenuto d'acqua in un volume piccolo, tanto da poter essere considerato un metodo di misurazione puntuale (Romano e Santini, 1974). Ha inoltre una grande risoluzione spaziale e temporale.

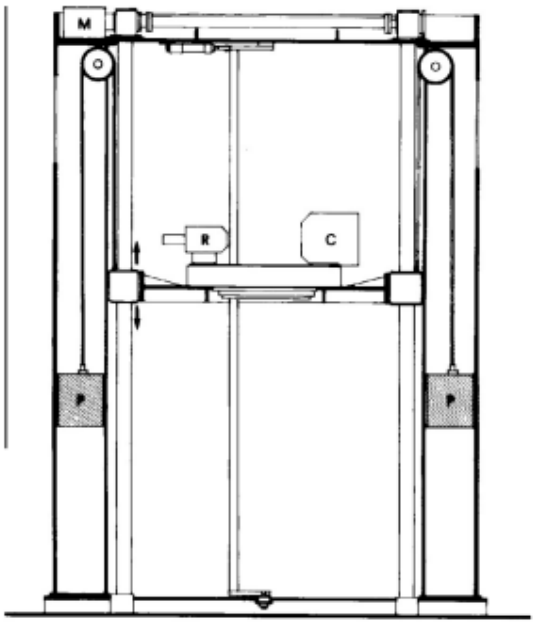

Fig. 1.1 – Layout della strumentazione a raggi gamma[2]

Lo scattering di neutroni è usato principalmente per studi sui campi e permette di determinare indirettamente il contenuto d'acqua grazie al processo di termalizzazione di neutroni ad alta energia che collidono con i nuclei atomici del suolo, specialmente quelli di idrogeno (Gardner e Kirkham, 1952). Essendo questo il principale fattore che influisce sulla perdita di energia dei neutroni veloci, il numero di impulsi dei neutroni termalizzati possono essere messi in relazione al contenuto volumetrico di acqua grazie a una curva di calibrazione. Tuttavia, il flusso di neutroni termalizzati diffusi dal suolo dipende dal valore di umidità attorno alla sonda: maggiore è il contenuto di acqua, minore è la sfera di influenza (volume misurato). A seconda del tipo di sorgente della radiazione (solitamente berillio) e del tipo di terreno questa tecnica riesce ad eseguire misure fino a 15 cm quando il suolo è ad alta saturazione,

mentre ad un basso livello di saturazione la sfera di influenza prende la forma di un ellissoide con l'asse maggiore di circa 50-60 cm. Nonostante i pericoli associati allo scattering di neutroni, dovuti principalmente ai problemi di salute dati dalla radiazione, e la difficoltà tecnica di derivare una curva di calibrazione rappresentativa, questo metodo è frequentemente utilizzato per la monitorazione verticale del suolo, essendo una tecnica non distruttiva con una buona risoluzione spaziale , specialmente ai livelli più alti di saturazione, ma con una risoluzione temporale relativamente bassa (Grayson e Western, 1998).

2.2.2 Tecniche basate su dielettrico

Le tecniche di misura indirette che si basano sulle proprietà dielettriche del suolo per ricavare il valore di umidità sono diventate sempre più popolari nelle ultime decadi, a causa dell'uso di componenti elettronici avanzati in vari strumenti. Le proprietà dielettriche di una sostanza dipendono dalla polarizzazione delle sue molecole e sono descritte dalla permettività dielettrica relativa , ε, che è una variabile adimensionale maggiore di 1e descritta come la somma di una componente reale, ε', e una immaginaria, ε'':

$$\varepsilon(\omega) = \varepsilon'(\omega) - j\left[\varepsilon''(\omega) + \frac{\sigma_{DC}}{\omega_{\varepsilon 0}}\right]$$

Dove j è l'unità immaginaria, ω è la frequenza angolare del campo elettromagnetico imposto, ε_0 (=8.854 $\times$ 10^{-12} $N^{-1}m^{-2}C^2$) è la permettività dielettrica del vuoto, e σ_{DC} è la conduttività elettrica diretta, che non dipende dagli effetti di polarizzazione ma dal trasporto di portatori di carica, come nel caso di ioni in un elettrolita.

La componente reale della permettività elettrica relativa è relazionata all'energia presente in un sistema a causa dell'allineamento dei dipoli con il campo elettromagnetico. La componente immaginaria invece è dovuta principalmente al rilassamento molecolare e a effetti di dissipazione di energia. In un sistema eterogeneo complesso come può essere un terreno non saturato, fatto in diverse proporzioni da particelle solide, aria, acqua, composti organici e minerali, è estremamente difficile interpretare il suo comportamento dielettrico, specialmente se il campo elettrico alternato imposto ha una bassa frequenza: in questi casi la conduttività elettrica DC, σ_{DC}, ha una forte influenza sulla componente immaginaria.

Tuttavia, nel range di 100 MHz-2GHz, la permettività dielettrica relativa apparente del suolo, ε'_{soil}, è maggiormente influenzata dalla permettività dielettrica dell'acqua ($\varepsilon'_{water} \cong 80$ a 20 °C), essendo questa maggiore di quella dell'aria ($\varepsilon'_{air} \cong 1$) e di quella delle particelle solide ($\varepsilon'_{solid} \cong$ 3-7). E' quindi possibile in questo range di frequenze correlare ε'_{soil} con θ attraverso una curva di calibrazione.

Alle frequenze più alte ε'_{soil} diventa quasi invariante rispetto alla frequenza del campo elettromagnetico: in questa condizione ci si può riferire a questa proprietà come la costante dielettrica apparente del suolo (κ_{soil}).

Due principali tecniche di misura sono state sviluppate sfruttando la possibilità di identificare con un errore accettabile una relazione empirica che permetta la stima del valore di θ partendo dalla misurazione di ε'_{soil}[2].

La prima è chiamata TDR (time domain reflectometry) e permette di determinare la permettività dielettrica del suolo monitorando il tempo che impiega un impulso di voltaggio a gradino a propagarsi lungo una linea di trasmissione connessa a una sonda posta nel suolo alla profondità desiderata

Fig. 1.2 – Sensore TDR con sonda

Il tempo di riflessione è influenzato dalle proprietà dielettriche del sistema che circonda la sonda, in questo caso quindi dell'acqua presente nel suolo.

Un cavo coassiale connette la sonda al sensore TDR, che fornisce l'impulso a gradino e analizza l'onda di ritorno con una risoluzione temporale dell'ordine dei nanosecondi. Attraverso lo studio del comportamento dielettrico di un ampio numero di minerali è stato visto che la migliore equazione di regressione tra θ e ε_{TDR} è una polinomiale del terzo ordine. Sebbene questa equazione non descriva in modo accurato la relazione quando ε_{TDR} tende all'unità o al valore della permettività dell'acqua pura, permette una buona stima nel range di θ= 0.05-0.60: usando questa curva di calibrazione su terreni senza un contenuto troppo alto di argilla o composti organici si può avere un errore minore di ± 0.015, mentre è riportato un errore di ± 0.035 per composti organici. Se sono richiesti valori assoluti di θ e maggiore accuratezza, è possibile creare una curva di calibrazione specifica.

In questo caso, specialmente se le misurazioni sono eseguite vicino alla superficie (dove le fluttuazioni della temperatura sono più alte), la dipendenza di ε'_{soil} dalla temperatura devono essere tenute in conto (Roth e al., 1990).

Alcuni studi hanno cercato di identificare una relazione tra θ e ε_{TDR} meno empirica e maggiormente basata sulla fisica (Dobson e al., 1985; Roth e al., 1990; Dirksen e Dasberg, 1993). L'idea sottostante a questi lavori è la determinazione della permettività di un sistema basandosi sulla permettività dei suoi componenti individuali: nel caso di un medium come il suolo è necessario conoscere il valore della porosità e le densità di aria, acqua e particelle solide; insieme ai parametri riguardanti l'orientazione degli elementi in base al campo elettromagnetico.

L'altra tecnica basata sul dielettrico stima il contenuto di acqua a livello locale considerando il suolo come un componente di un capacitore. Questa tecnica stima quindi la permettività dielettrica relativa apparente misurando il tempo di carica di un capacitore inserite nel suolo , che svolge il ruolo di medium dielettrico.

Fig. 1.3 – Sensore FDR

Sebbene siano stati fatti numerosi miglioramenti alla tecnica TDR, non c'è dubbio che i maggiori miglioramenti siano stati realizzati sfruttando sensori capacitivi. Una delle ragioni è sicuramente il loro minore costo, in termini sia di risorse finanziarie che di tempo, dando tuttavia risultati peggiori dal punto di vista dell'accuratezza e della precisione.

Il principale errore di misura associato al loro utilizzo dipende dalla frequenza del campo elettromagnetico imposto dallo strumento, che può andare da 20 MHz fino a 300 MHz.

Queste frequenze sono in ogni caso minori di quelle a cui operano i sensori TDR (fino a 2GHz), rendendo le misurazioni più sensibili alla temperatura del suolo e alla sua salinità.

Sia la tecnica TDR che quella basata sulla capacità non danno informazioni sull'umidità puntuale del terreno (solo la tecnica ad attenuazione di raggi gamma riesce a farlo) ma eseguono una media del contenuto di acqua nel volume esaminato, che dipende principalmente dalla lunghezza e dalla forma della sonda immersa nel suolo. Tuttavia in diversi campi è presente un forte interesse nel monitoraggio della variazione spaziale e temporale dell'umidità lungo un profilo del suolo (Huisman e al., 2006; Melone e al., 2006). Alcuni ricercatori hanno sviluppato una procedura inversa per stimare il profilo del contenuto d'acqua lungo una sonda TDR durante processo di infiltrazione ed evaporazione. Questo studio, nato inizialmente in laboratorio, è stato testato con successo anche in applicazioni esterne (Greco e Guida, 2008).

Inserire singole sonde a diverse profondità per ricreare un profilo di umidità non è tuttavia sempre una soluzione efficiente: per questo compito la sonda a neutroni è una tecnica molto più efficace. Sono state comunque creati alcuni sensori che permettono di monitorare il profilo del contenuto d'acqua usando una singola sonda capacitiva. Queste sonde richiedono l'utilizzo di un tubo di accesso, solitamente fatto in plastica PVC: una tecnica ottimale di trivellazione è quindi molto importante per eseguire misure con successo.

2.2.3 Tecniche basate sugli infrarossi

Numerosi studi hanno dimostrato il ruolo della spettroscopia con infrarossi nel monitoraggio dell'umidità del suolo. Molti illustrano la relazione inversa tra queste due variabili (Post e al., 2000; Galvao e al., 2001): all'aumentare del contenuto di acqua nel suolo si ha una diminuzione della sua riflettenza. Questa relazione è causata da due fenomeni diversi: le particelle di suolo coperte da un sottile film d'acqua e la presenza di acqua nella struttura di alcuni cristalli (Stoner e Baumgardner, 1981). Con il miglioramento degli strumenti di misurazione si è potuto osservare come le variazioni della riflettenza spettrale dovute all'umidità sono più pronunciate a lunghezze d'onda maggiori di 1450 nm (Weidong e al, 2002). Gli stessi studi hanno anche mostrato che ad alti contenuti di acqua il trend si modifica, ottenendo un aumento congiunto di riflettenza e umidità. E' stato determinato che questa modifica nel comportamento si ottiene attorno ai valori della capacità di campo, variando per terreni diversi, e accade prima della saturazione del segnale di riflettenza[3].

Lo studio di qualsiasi proprietà del suolo è relazionato allo studio delle zone dello spettro più sensibili alla presenza di acqua. Le frequenze vibrazionali delle molecole d'acqua oltre 2500 nm influenzano le lunghezze d'onda dell'assorbimento dell'acqua. Le bande di assorbimento con i picchi più elevati si trovano a 1450 e 1950 nm, mentre la banda più significativa per lo studio delle modifiche della riflettenza si ha nel range 2080-2320 μm. (Baumgardner e al.,1985; Galvao e al., 2001) [3].

2.2.4 Tecniche di misura idro-geofisica e multifunzionale

Altre tecniche di sensing basate su onde elettromagnetiche sono le tecniche idro-geofisiche di tomografia della resistività elettrica indiretta (ERT), radar a penetrazione (GPR) e induzione elettromagnetica (EMI). Queste tecniche di misura permettono di osservare il contenuto di un grande volume di terreno,offrendo una via intermedia tra le tecniche descritte in precedenza e quelle basate su sensori aerei (Cassiani e al., 2006; Robinson e al., 2012).

Le tecniche ERT consistono nella misurazione della conduttività elettrica del suolo (EC_b) , che viene poi relazionata al contenuto d'acqua (Samouelian e al., 2005).

Sebbene la tomografia della resistività elettrica è largamente applicata alle condizioni sperimentali, specialmente nella versione a due dimensioni, un monitoraggio ERT tridimensionale è stato sviluppato e usato con successo per studiare un processo di trasporto dei soluti (Binley e al., 1996) o il bilancio di acqua in un suolo radicato (Garrè e al. 2010). La tomografia elettrica 3D non invasiva è stata recentemente applicata nel monitoraggio delle interazioni tra diverse piante, per ottenere una maggiore conoscenza dello scambio di flussi nel sistema vegetazione-suolo-aria.

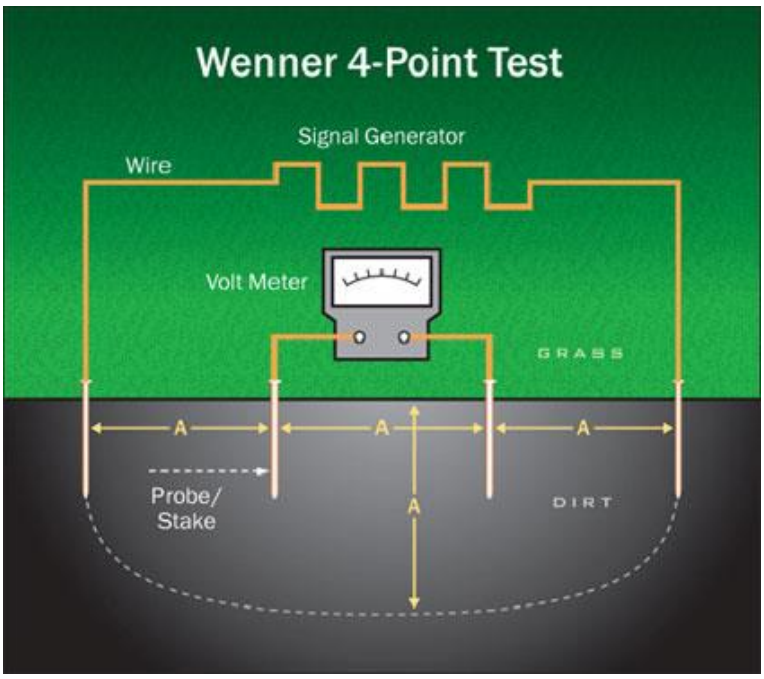

Fig. 1.4 – Schematizzazione del test ERT/SRT

Il GPR è un'altra tecnica che, sebbene sia stata inizialmente sviluppata e usata in altri settori tecnici e scientifici, è oggigiorno largamente usata per il monitoraggio dell'umidità del suolo, in alcuni casi combinata con la tecnologia ERT. Esiste anche uno studio (Brovelli e Cassiani, 2011) in cui si esegue una stima combinata della conduttività e della permettività del suolo, unendo quindi le informazioni delle

tecniche ERT e TDR.

Una tecnica indiretta alternativa è il metodo a impulsi di calore (Sayde e al., 2010), applicata in condizioni di laboratorio usando un sensing di temperatura su fibra ottica. I cavi a fibra ottica possono essere utilizzati per ottenere misure distribuite del contenuto locale di acqua in ambiente esterno, con un'accuratezza di 1 m nello spazio e di 1h nel tempo. Il principio di misurazione del sensore a impulsi di calore è basato sul monitoraggio della variazione della conduttività termica del suolo, ma un interessante sviluppo è dato dal fatto che la sonda, se opportunamente integrata con altri sensori, permette la simultanea stima della concentrazione di soluzione del suolo e perfino della densità del flusso dell'acqua (Mori e al., 2006) hanno sviluppato una sonda termica multifunzionale che comprende sei sensori, tra cui quattro termistori e quattro elettrodi di Wenner, permettendo la determinazione contemporanea di conduttività elettrica, contenuto e flusso d'acqua e capacità di calore.

L'integrazione di vari tipi di sensori in una sola unità è di particolare interesse in quanto riesce a minimizzare il disturbo per il suolo nel punto di misurazione. E' anche importante sottolineare che combinare le misurazioni indubbiamente comporta benefici nel poter valutare in maniera più efficace modelli di sistemi complessi, dove sono presenti simultaneamente flussi d'acqua, trasporti di soluti e flussi di calore nel sistema suolo-vegetazione-atmosfera[2].

Capitolo 2

Spettroscopia d‘impedenza

In questo capitolo viene presentata la spettroscopia d’impedenza, tecnica usata dal sensore per la rilevazione di parametri legati all’umidità del terreno. Inizialmente vengono presentati i concetti teorici legati alle caratteristiche di permettività e conducibilità elettrica, per passare poi alla trasmissione delle onde elettromagnetiche. Vengono infine presentati i principi alla base della misurazione della spettroscopia, specialmente riguardo a tecniche con guida d’onda.

1. Concetti teorici

La spettroscopia di impedenza è una tecnica basata sulla caratterizzazione delle proprietà elettriche dei materiali e delle interfacce tra di essi nel dominio della frequenza. All’interfaccia tra due diversi materiali i cambiamenti delle proprietà fisiche ed elettriche, insieme a quello della distribuzione eterogenea di carica, provocano una modifica dell’impedenza di un mezzo materiale, da cui è possibile ricavare una successiva caratterizzazione di molte delle proprietà elettriche..

1.1 Grandezze elettriche intrinseche dei materiali

Tra le proprietà elettriche più interessanti si trovano la conducibilità e la permittività, fortemente dipendenti dalla frequenza dello stimolo. L'indagine di queste proprietà elettriche è una metodologia di analisi percorribile per valutare il grado di umidità del terreno.

La conducibilità è una grandezza fisica è definita come l'inverso della resistività, grandezza denotata da ρ. Essa descrive come si comporta il materiale rispetto al passaggio di corrente: maggiore è il valore di resistività, maggiore sarà la difficoltà con cui una corrente elettrica potrà fluire attraverso il mezzo. Una definizione più rigorosa si può ottenere facendo riferimento al campo elettrico e alla densità di corrente che fluisce nel materiale, dato che la resistività, e quindi anche la conducibilità, è esprimibile con i rapporti delle due grandezze:

$$\rho = \frac{E}{J}$$

$$\sigma = \frac{1}{\rho} = \frac{J}{E} = \frac{IL}{SV}$$

Con J = densità di corrente, E = campo elettrico, I = corrente elettrica, L = tratto del conduttore attraversato da corrente, S = sezione conduttore, V = differenza potenziale ai capi del tratto.

La permettività è invece una grandezza fisica che esprime l'attitudine di un materiale ad immagazzinare carica elettrica qualora sollecitato da un campo elettromagnetico.

La permettività elettrica è fortemente legata alla suscettibilità elettrica, una proprietà del materiale che parametrizza la capacità di polarizzarsi una volta sottoposto ad un campo elettrico.

La polarizzazione degli atomi e delle molecole interne genera a sua volta un altro campo elettrico descritto dal vettore induzione elettrica. La permettività è facilmente esprimibile nel caso di un mezzo lineare, omogeneo e isotropo, essendo sotto queste condizioni rappresentabile come un valore scalare. La seguente relazione costituiva dei materiali descrive in forma analitica questo fenomeno:

$$\vec{D} = \varepsilon \vec{E}$$

dove $\vec{E}$ rappresenta il campo elettrico esterno, $\vec{D}$ il vettore induzione elettrica ed ε la permettività elettrica in forma scalare. Qualora invece il mezzo avesse caratteristiche anisotrope, la sua permettività si rappresenta con un tensore in cui i valori riflettono una descrizione locale dipendente dalle coordinate.

Spesso si ricorre alla definizione di permettività elettrica relativa, chiamata anche costante dielettrica e rappresentata dal simbolo ε_r , o talvolta da k, riferendosi al rapporto con la permettività del vuoto. E' descritta dalla seguente formula, dove ε_0 indica la permettività elettrica del vuoto

$$\varepsilon_r = \frac{\varepsilon}{\varepsilon_0}$$

L'esperimento classico per valutare questa grandezza è quello del condensatore a facce piane parallele: un condensatore con il vuoto tra i piatti ha una capacità pari a

$$C_0 = \varepsilon_0 \frac{A}{d}$$

dove A è l'area di un piatto e d è la distanza tra i piatti.

Se tra le due armature viene posto un altro materiale con permettività elettrica relativa maggiore di 1, il condensatore avrà una capacità C maggiore di C_0 e il rapporto fornirà una stima di ε_r

$$C = \varepsilon_0 \varepsilon_r \frac{A}{d}$$

$$\varepsilon_r = \frac{C}{C_0}$$

A differenza del vuoto, in cui la permettività è costante, per gli altri mezzi la permettività è fortemente dipendente dalla frequenza; questo comportamento è legato alla polarizzazione del materiale che non è istantanea, causando perciò uno sfasamento tra il campo elettrico esterno e l'induzione elettrica. Per descrivere questo fenomeno è utile modellare la permettività come una grandezza complessa in funzione della pulsazione. Viene rappresentata dal simbolo $\hat{\varepsilon}(\omega)$ e può essere espressa separando la parte reale da quella immaginaria

$$\hat{\varepsilon}(\omega) = \varepsilon'(\omega) - j\varepsilon''(\omega)$$

$$\hat{\varepsilon}_r(\omega) = \varepsilon_r'(\omega) - j\varepsilon_r''(\omega)$$

dove $\varepsilon_r'(\omega)$ rappresenta la costante dielettrica ed $\varepsilon_r''(\omega)$ rappresenta il fattore di perdita. La costante dielettrica corrisponde alla definizione fornita precedentemente, quindi è legata all'immagazzinamento di energia nel mezzo, mentre il fattore di perdita quantifica la dissipazione di energia nel mezzo in presenza di un campo elettrico. E' possibile infatti esprimere la parte immaginaria anche come funzione della conducibilità del mezzo, rappresentata dal simbolo σ:

$$\varepsilon'' = \frac{\sigma}{\omega}$$

$$\varepsilon_r'' = \frac{\sigma}{\omega \varepsilon_0}$$

La permettività del suolo è strettamente influenzata da fattori quali l'umidità, la "bulk density", la tessitura del terreno, la temperatura e la frequenza[4].

Umidità

Il contenuto di umidità è forse la variabile più importante, nonché la variabile di maggiore interesse ai fini di questa tesi. Gli studi disponibili mostrano che al suo aumentare anche la permettività relativa del terreno aumenta. Questo è prevedibile, in generale, dal momento che la permettività relativa del terreno asciutto è inferiore a 5, mentre quella dell'acqua può essere più di un ordine di grandezza maggiore: ci si aspetta una modifica percepibile della permettività media al variare dell'umidità del terreno. Per un basso contenuto di umidità, la parte reale della permettività del terreno ε' descrive un andamento quasi costante. Questo andamento, secondo Lundien e Wiebe (1971), deriva dalla limitata mobilità dell'acqua assorbita sulla superficie delle particelle del terreno, fenomeno per il quale la ε', nel caso di un terreno molto argilloso, descriverebbe un andamento quasi costante per un range di umidità più ampio, proprio a causa della forte tendenza a trattenere acqua. La parte immaginaria della permettività del terreno ε'', sebbene cresca anch'essa con l'aumento del contenuto di umidità, risulta soggetta ad un comportamento variabile e di difficile interpretazione. In particolare, alcuni studi effettuati da Leschanskii et. al. (1971), mostrano come i valori di ε'' cambino da un terreno argilloso ad un terreno sabbioso: valori pressoché simili a quelli di ε' per un terreno argilloso, valori dieci volte inferiori rispetto alla ε' per un terreno sabbioso. Tuttavia, va considerato come questo fenomeno sia trascurabile nel caso delle alte frequenze (ordine dei GHz). A tal proposito, la conduttività (che è direttamente proporzionale alla ε'' ad una data frequenza) presenta un comportamento altrettanto difficilmente interpretabile.

Bulk Density

Un altro parametro da considerare è quello del bulk density, indice di compattezza del suolo. La parte reale della permettività del terreno ε' e il bulk density sono legate dalla seguente relazione (Hipp, 1974):

$$\varepsilon' = (\frac{1+\rho}{2})^2$$

dove ρ è il bulk density.

Il bulk density ha inoltre un effetto indiretto importante sulla relazione tra la permettività dielettrica e il contenuto di umidità del terreno. La parte reale della permettività del terreno ε' descrive un andamento in funzione dell'umidità differente a seconda che quest'ultima sia espressa in forma gravimetrica o in forma volumetrica: quando è usata la forma volumetrica, la permettività assumerà determinati valori terranno conto del grado di compattazione del terreno.

Tessitura

Risultati di vari esperimenti indicano che, ad alte frequenze (sopra i 10 GHz), la permettività dielettrica del suolo sia quasi non influenzata dalla composizione dei terreni. A basse frequenze (minori di 10 GHz), invece, un diverso tipo di tessitura incide sulle proprietà dielettriche: in particolare la perdita dielettrica cresce in funzione della presenza di particelle argillose.

Temperatura

La permettività dielettrica del terreno è fondamentalmente indipendente dalla temperatura. Quella dell'acqua invece è calcolata in funzione della temperatura.

Dal momento che il terreno contiene quasi sempre dell'acqua (salvo che non sia essiccato in forno), la sua permettività dipende dalla temperatura. Questo è stato confermato sperimentalmente da Poe et. al. (1971) i quali osservarono che la parte reale della permettività del terreno secco non cambiava quando la temperatura cresceva da 20°C a 60°C ma cambiava quando era presente anche una piccola quantità di acqua.

Frequenza

Poiché la ε' è legata a fenomeni di tipo dipolare dovuti all'acqua contenuta nel terreno, un aumento della frequenza delle oscillazioni di un campo elettromagnetico incidente rende più difficili le interazioni dipolari, facendo diminuire gradualmente la ε'. La zona di frequenza in cui si manifesta questo fenomeno è tra 1 GHz e 50 GHz. A temperature più alte, ε' diminuisce a frequenze leggermente più alte. La parte immaginaria ε'' invece fino ai 100 MHz rimane pressoché costante perché gli effetti conduttivi sono effettivamente dominanti, mentre per frequenze più alte si ha un aumento a causa delle perdite dielettriche. Come si osserva in figura, i valori della parte immaginaria e della tangente di perdita sono piccoli per frequenze intorno ai 100 MHz e sono massimi alle basse frequenze. La parte reale invece ha un andamento quasi costante nel range 10^7 -10^9 Hz.

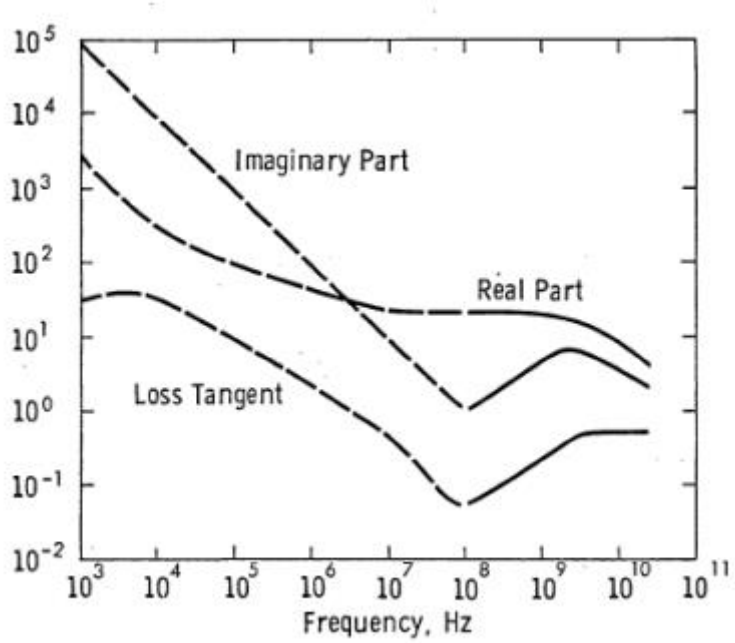

Fig. 2.1 – Parte reale, parte immaginaria e tangente di perdita relative ad un limo argilloso contenente il 15% di umidità gravimetrica. [4]

1.2 Propagazione delle onde elettromagnetiche

Le proprietà dielettriche di un mezzo influenzano la propagazione di un'onda elettromagnetica che lo attraversa: quando la radiazione incontra una discontinuità tra due materiali differenti, si modificano la velocità di propagazione e la direzione, quest'ultima nel caso in cui la propagazione non sia perpendicolare alla superficie. Inoltre l'onda incidente che incontra la discontinuità viene in parte trasmessa ed in parte riflessa. Attraverso l'analisi di queste caratteristiche si può giungere ad una valutazione della permettività complessa .

Per dare una rappresentazione analitica ai concetti appena descritti, è necessario introdurre la velocità di propagazione di un'onda v e l'indice di rifrazione n, utili nell'applicazione della legge di Snell,

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \cong \frac{c}{\sqrt{\varepsilon_r}}$$

$$n = \frac{c}{v} \cong \sqrt{\varepsilon_r}$$

$$\frac{\sin\theta_2}{sin\theta_1} = \frac{n_1}{n_2}$$

dove c è la velocità della luce nel vuoto. La terza equazione è la legge di Snell, che mette in relazione l'angolo dell'onda trasmessa θ_2 e dell'onda incidente θ_1 con gli indici di rifrazione dei relativi mezzi trasmissivi.

2. Principi di funzionamento

L'approccio generale della spettroscopia d'impedenza consiste nell'applicare uno stimolo al mezzo di analisi e osservare la risposta, la quale è solitamente mediata dall'impedenza. Il confronto tra input e output permette una definizione dell'impedenza e la caratterizzazione completa del mezzo. Qualora l'analisi preveda l'interesse per l'impedenza intrinseca di un mezzo, cioè la reazione che oppone un materiale alla propagazione del campo elettromagnetico, la tecnica di spettroscopia prevede un irraggiamento del materiale analizzato con uno stimolo elettromagnetico con caratteristiche di modulo e fase note, e la successiva analisi dell'onda riflessa o trasmessa (a seconda del setup) raccolta da un ricevitore.

Assumendo infatti che i parametri caratterizzanti il sistema di misura siano considerati costanti o stazionari e che quindi il sistema stesso sia definito lineare, è possibile garantire, in uscita a quest'ultimo, un segnale altrettanto sinusoidale. Ciò che tuttavia differenzia l'uscita dall'ingresso sono proprio i valori di modulo e fase.

Come esempio si immagini di applicare ad un materiale il segnale stimolo:

$$v(t)=V_m sin(wt)$$

alla frequenza

$$f = \frac{\omega}{2\pi}$$

E, ponendosi sotto l'ipotesi di stato stazionario, per la quale le variazioni della grandezza da misurare sono molto più lente rispetto a quello del segnale utilizzato come stimolo, la corrente misurata sarà pari a

$$i(t)=I_m sin(wt + \theta)$$

dove θ è la differenza di fase tra la tensione e la corrente.

Nel caso l'onda elettromagnetica si propaghi in una struttura guidata simile a quella presente nel sensore, è utile fare riferimento al modello di linea di trasmissione, ovvero una rappresentazione circuitale che permette di analizzare il comportamento con un approccio più semplice. Il modello da scegliere è legato al tipo di struttura e al tipo di propagazione .

Ad esempio, la guida d'onda rettangolare ha un comportamento di tipo passa alto: in questo caso il modello di una linea ideale, cioè senza perdite, è quello mostrato nella figura 2.2 [5].

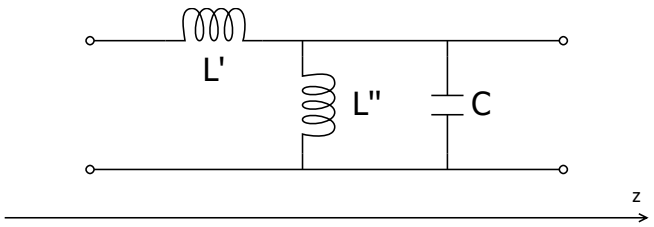

Figura 2.2 - Modello a linea di trasmissione ideale della guida d'onda rettangolare. [5]

I valori dei tre componenti sono legati alle proprietà elettriche dei materiali e alla frequenza dell'onda trasportata:

$$L' = \mu$$

$$C' = \varepsilon$$

$$L'' = \frac{c^2 \mu_0}{\varepsilon_r \omega_c^2}$$

$$\omega_c = \sqrt{\frac{1}{L''C'}}$$

dove ω_c è la frequenza di cutoff.

Attraverso questa rappresentazione è possibile definire l'impedenza caratteristica della linea come:

$$Z_c = \frac{\sqrt{\frac{L'}{C'}}}{\sqrt{1 - \left(\frac{1}{\Omega}\right)^2}}$$

dove $\Omega = {}^{f}/_{f_c}$, con $f_c = {}^{v}/_{2a}$ la frequenza di cutoff, a la dimensione maggiore e v la velocità di propagazione. Il valore dell'impedenza caratteristica, come si può notare dal modello appena presentato, dipende anche dalla permettività elettrica

complessa.

Attraverso la rappresentazione circuitale è possibile studiare il comportamento della linea mediante grandezze classiche, ovvero in termini di tensioni e correnti. In una generica posizione z lungo la linea è possibile definire:

$$V(z) = V_+ e^{-j\beta(\omega)z} + V_- e^{+j\beta(\omega)z}$$

$$I(z) = \frac{V_+}{Z_c} e^{-j\beta(\omega)z} - \frac{V_-}{Z_c} e^{+j\beta(\omega)z}$$

$$V_i(z) = V_+ e^{-j\beta(\omega)z}$$

$$V_r(z) = V_- e^{+j\beta(\omega)z}$$

$$I_i(z) = \frac{V_+}{Z_c} e^{-j\beta(\omega)z}$$

$$I_r(z) = \frac{V_-}{Z_c} e^{+j\beta(\omega)z}$$

dove $\beta = {}^{\omega}/_{v}$ è la costante di fase e v è velocità di propagazione dell'onda, mentre V_i, I_i e V_r, I_r rappresentano rispettivamente l'onda incidente e riflessa.

Quando la linea presenta una discontinuità nell'impedenza caratteristica, variando da Z_1 a Z_2, una parte dell'onda incidente viene riflessa e una parte viene trasmessa.

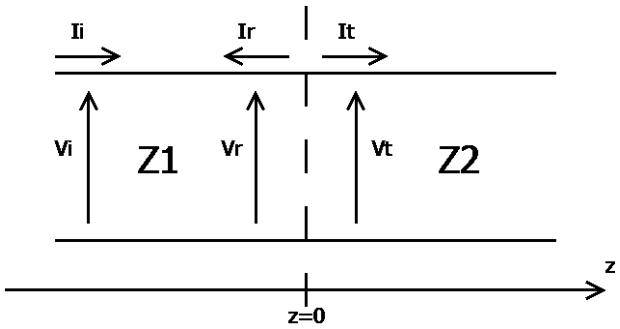

Figura 2.3 - Discontinuità in una linea d'onda. [6]

Questo è mostrato in figura 2.3, dove in ascissa $z = 0$ è presente una discontinuità, V_i, I_i rappresentano l'onda incidente; V_r, I_r l'onda riflessa e V_t, I_t l'onda trasmessa. In questa situazione è possibile definire il coefficiente di riflessione ρ e il coefficiente di trasmissione τ attraverso il rapporto delle tensioni e correnti o come rapporto tra le impedenze caratteristiche nel modo seguente:

$$\rho = \frac{V_r}{V_i} = \frac{I_r}{I_i} = \frac{Z_2 - Z_1}{Z_2 + Z_1}$$
$$\tau = \frac{V_t}{V_i} = 1 + \rho = \frac{2Z_2}{Z_2 + Z_1}$$

La relazione tra la risposta del sistema allo stimolo nel dominio del tempo e le sue proprietà è, in generale, molto complessa. A volte essa può essere priva di una soluzione in forma chiusa e quest'ultima può essere calcolata solo per via numerica.

L'analisi nel dominio delle frequenze semplifica il problema riducendo le equazioni nel dominio temporale a una rappresentazione in frequenza più immediata e di maggiore comprensione

$$V(jw) = Z(jw) * I(jw)$$

dove Z(jw) è definita come la funzione impedenza e rappresenta l'opposizione del circuito elettrico ad uno stimolo di tensione o corrente.

Essa viene comunemente rappresentata nel piano complesso attraverso la sua parte reale ed immaginaria

$$Z(w) = Z' + jZ''$$

con

$$Z' = Re(Z) = |Z|cos(wt)$$

$$Z'' = Im(Z) = |Z|sin(wt)$$

Dal risultato della risposta in frequenza dell'impedenza complessa si è in grado di ricavare le proprietà di interesse del materiale e, analizzando la risposta nel range delle microonde (1GHz – 1000GHz), si è in grado di sfruttare le proprietà della lunghezza d'onda per ridurre il rumore introdotto sulla misura dai fenomeni parassiti. Sulla base di queste considerazioni si è pensato di realizzare un sistema di spettroscopia di impedenza, allo scopo di analizzare le risposte in frequenza del terreno ricavando stime della sua umidità.

Capitolo 3

PLS Regression

In questo capitolo verrà trattato un particolare tipo di analisi multivariata, chiamato PLS-Regression, usata per modellare la relazione tra 2 matrici, X e Y. Ne vengono presentati prima gli aspetti teorici fondamentali e le assunzioni su cui si basa, per poi passare al corretto sviluppo dei modelli e agli algoritmi su cui si fonda il calcolo.

1. Considerazioni generali

La PLS-Regression(PLSR) è una generalizzazione della regressione multipla lineare (MLR). Queste tecniche nascono per trovare soluzioni ai problemi di regressione, ovvero come modellare una o più variabili Y partendo da un set di variabili predittrici X, necessità comune nell'analisi dei dati scientifici. Alcuni esempi possono essere le relazioni tra proprietà di campioni chimici(Y) e composizioni chimiche(X), la qualità e la quantità dei prodotti e il loro il processo manifatturiero, oppure proprietà chimiche o attività biologiche di un set di molecole e la loro struttura chimica(codificata da un set di variabili X).

La PLSR è di particolare interesse in quanto, a differenza della MLR, può analizzare dati con variabili X fortemente correlate, rumorose o numerose; oltre a modellare nello stesso momento diverse variabili Y. La MLR infatti lavora bene se le variabili X sono relativamente poche e non correlate (X ha rango massimo).

Con i moderni strumenti di misurazione, inclusi spettrometri, cromatografi e diversi altri sensori, le variabili X tendono a essere molte e correlate fra loro, nonché rumorose e spesso incomplete.

La PLSR, lavorando invece con variabili X numerose e correlate, ci permette di investigare problemi più complessi e analizzare i dati in una maniera più realistica [7].

Come esempio, consideriamo un esperimento caratterizzato da M osservazioni. Per ognuna di queste osservazioni vengono registrati i valori delle variabili dipendenti (Y) e delle variabili indipendenti (X) . La PLS trova quell'insieme di componenti di X, le variabili latenti, che rappresentano una decomposizione della stessa matrice e che allo stesso tempo massimizzano la loro correlazione con le variabili dipendenti Y. In formule X è decomposta nel seguente modo:

$$X = T \cdot P^T + E \qquad \text{con} \quad P \cdot P^T = 1$$

dove I è la matrice identità , T è la matrice degli scores , P quella dei loadings ed E quella dei residui. La matrice T contiene le nuove coordinate di X nel nuovo spazio descritto dalle variabili latenti. Nello stesso modo possiamo decomporre Y utilizzando la matrice degli scores T di X:

$$Y = T \cdot C^T + F \qquad \text{con} \quad C \cdot C^T = 1$$

dove C è la matrice dei loadings ed F quella dei residui. Tenendo presente le equazioni precedenti è possibile scrivere la matrice delle variabili dipendenti come:

$$Y = X \cdot W \cdot C^T + F = X \cdot B + F \qquad \text{con } B = W \cdot C^T$$

dove la matrice W viene determinata attraverso la seguente relazione:

$$T = X \cdot W$$

Lo scopo della PLS è quello di trovare le matrici T e C in modo da massimizzare la covarianza tra X ed Y [8].

1.1 Interpretazione geometrica

La PLSR, così come la regressione multipla, rappresenta un metodo di proiezione. Infatti questa tecnica può essere considerata come una proiezione dei vettori della matrice X in un sottospazio la cui dimensione è definita dalle k variabili latenti che massimizzano la covarianza tra X ed Y. Le coordinate di tale proiezione rappresentano dei buoni predittori di Y.

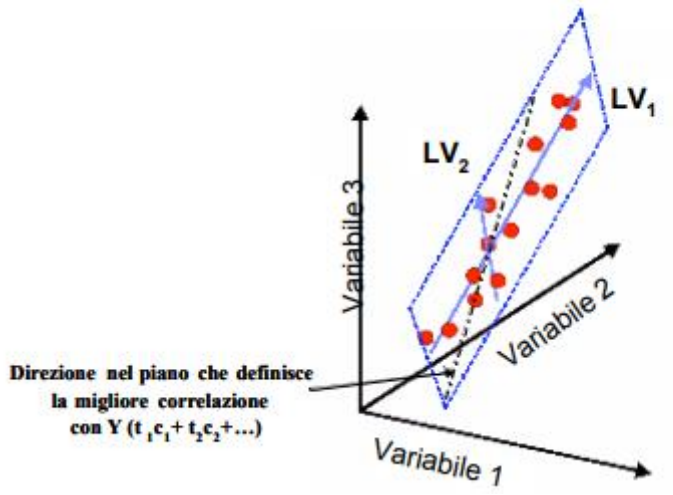

Fig. 2.1 Rappresentazione geometrica della PLSR.[8]

La direzione del piano è espressa come la pendenza p_{ak} di ogni componente della PLS nel piano rispetto a ciascuna coordinata, x_k. Questa pendenza è data dal coseno dell'angolo tra la direzione della PLS e gli assi delle coordinate. Perciò la PLS crea un iperpiano A-dimensionale nello spazio X che approssimi bene **X.** Allo stesso tempo le posizioni delle proiezioni di questi punti sul piano, descritte dagli scores t_{ia}, sono in relazione ai valori delle risposte Y_{im}. [7]

1.2 Nota Storica

L'approccio con PLS fu ideato nel 1975 da Herman Wold per la modellazione di data-set complessi in termini di catene di matrici, chiamate "path models". Questo approccio includeva un modo semplice ma efficace di stimare i parametri di questi modelli, denominato NIPALS(Non-linear Iterative PArtial Least Squares). Da qui nasce l'acronimo PLS (Partial Least Squares). Questo nome si riferisce alla parte centrale della stima, dove ogni parametro del modello è stimato iterativamente come la pendenza di una regressione bivariata(least squares) tra una matrice Y e un vettore di parametri X. Il "partial" in PLS indica che questa è una regressione parziale, poiché il vettore x è considerato fissato nella stima. Ciò mostra anche che si può considerare una qualsiasi moltiplicazione matrice-vettore come un set di regressioni bivariate semplici, ottenendo una maniera semplice di risolvere il problema di dati mancanti, oltre a fornire un'interessante relazione tra due operazioni centrali nell'algebra matriciale e nella statistica.

Nel 1980 il modello di PLS con 2 blocchi **X** e **Y** fu leggermente modificato da Syante Wold e Harald Martens per adattarsi meglio a dati scientifici e tecnologici, dimostrando la sua utilità nell'analizzare data set complessi, dove era difficile o impossibile applicare la regressione ordinaria. Per dare alla PLS un significato più descrittivo, si è cominciato a interpretare PLS come Projection to Latent Structures.

Nel 1993 fu proposto da De Jong e al. un nuovo algoritmo, di nome SIMPLS, in grado di migliorare l'originale algoritmo NIPALS, grazie principalmente alla possibilità di calcolare tutte le variabili latenti come diretta combinazione lineare delle variabili originarie, come sarà spiegato in seguito. Questo algoritmo è tutt'ora uno dei più utilizzati per l'implementazione della PLS nei software di analisi statistica, come quello usato per il presente lavoro, di nome Eigenvector.

2. Assunzioni nella PLSR

2.1 Variabili latenti

Nella modellazione con PSLR si assume che i sistemi o i processi investigati siano influenzati solo da un numero ristretto di variabili nascoste, le variabili latenti(LV).

Il numero di queste LV non è solitamente conosciuto, inoltre gli X-scores della PLSR non sono spesso stima diretta delle LV, ma si trovano nel loro stesso spazio. Gli scores (denominati **T**) sono quindi in relazione alle LV(denominate **V**) tramite una matrice di rotazione **R** di solito sconosciuta, con la proprietà **R'R=1:**

V=TR'

Sia le variabili X che quelle Y sono modellate come combinazioni delle LV nascoste, e non vengono considerate quindi come indipendenti. Queste assunzioni sulle LV corrispondono a concetti già applicati alle molecole e alle reazioni in chimica e in biologia molecolare, rendendo quindi la PLSR adatta per la modellazione di dati chimici e biologici [7]. Per esempio, in spettroscopia è noto che lo spettro di un campione è dato dalla somma degli spettri dei suoi costituenti moltiplicati per la loro concentrazione nel campione. Identificando quest'ultima come **t** e gli spettri come **p** otteniamo lo stesso modello delle variabili latenti:

X=t1p1'+t2p2'+...=TP'

Gli scores **T** possono essere visti in un ulteriore modo, ovvero composti da derivate di una funzione non conosciuta sottostante al modello preso in esame.

La scelta dell'interpretazione dipende dalla conoscenza che si ha del sistema: maggiore è, più facile è dare un'interpretazione degli X-scores o della loro rotazione basata sulle variabili latenti.

Se il numero di LV eguaglia quello di variabili X allora queste sono indipendenti, e la PLSR da gli stessi risultati di una MLR. Possiamo quindi vedere la PLSR come una generalizzazione della MLR, contenendo quest'ultima come caso speciale quando il numero di variabili X e Y è relativamente piccolo rispetto al numero di osservazioni N. In molti dei casi pratici tuttavia le variabili X non sono indipendenti: in questi casi la PLSR fornisce una soluzione statisticamente più robusta della MLR.

La PLSR crea un modello di X in termini di proiezione bilineare, più i residui. Perciò la PLSR assume che ci possano essere parti di X non correlate con Y, le quali potrebbero contenere rumore o altri disturbi. La PLSR tollera perciò la presenza di rumore in X.

2.2 Derivazioni alternative

Il secondo fondamento teorico del modello con LV è quello dell'espansione di Taylor. Assumiamo che i dati **X** e **Y** siano generati da una funzione multidimensionale F(**u**,**v**), dove il vettore di variabili **u** descrive i cambiamenti tra le variabili. Creando un'espansione di Taylor nella direzione **u** e discretizzando per ‘i’ osservazioni e ‘k’ variabili, otteniamo il modello. Più piccolo è l'intervallo **u** che viene modellato meno termini si devono usare nell'espansione di Taylor, e meno componenti ci servono nel modello.

Possiamo quindi interpretare la PLS come un modello di similarità: i dati(variabili) misurati in un set di osservazioni simili(campioni, casi...) possono sempre essere modellati con la PLS.

Più simili sono fra loro queste osservazioni, meno componenti sono necessarie

nel modello.

Si hanno in conclusione due interpretazioni del modello con LV: i dati reali possono essere spiegati come combinazione lineare di “fattori” o come misure fatte su un set di osservazioni simili [7].

2.3 Omogeneità

Qualsiasi analisi di dati si basa su un'assunzione di omogeneità. Ciò implica che il sistema o il processo studiato deve essere in uno stato simile per tutte le analisi eseguite, e il meccanismo di influenza di **X** su **Y** deve sempre essere lo stesso. Ciò corrisponde nello stesso tempo ad avere una certa limitazione alla variabilità e diversità di **X** e **Y.**

E' perciò importante che l'analisi preveda una diagnostica su quanto queste assunzioni siano soddisfatte. La PLSR fornisce diagnostiche addizionali oltre a quelle classiche dei modelli di regressione, specialmente basate sulla matrice X(i plot di scores e loadings e i residui di X). Questi dati ci permettono di diagnosticare più facilmente e velocemente delle disomogeneità, cosa che nei modelli tradizionali era possibile solo con grandi residui di Y.

3. PLSR

3.1 I dati - X e Y

Il modello PLSR è sviluppato da un training set di N osservazioni con K variabili X indicate da x_k(k=1,…,K) e M variabili Y indicate da y_m(m=1,…,M). Questi dati formano le matrici **X** e **Y**, rispettivamente di dimensioni N*K e N*M.

Le stime del modello vengono fatte basandosi sui dati X, tramite X-scores(chiamati anche t-values), X-residuals, la deviazione standard dei residui, y-values, intervalli di confidenza e altri parametri.

.

3.2 Trasformazione, scaling e centering

Prima dell'analisi le variabili X e Y sono spesso trasformate, in modo da rendere la loro distribuzione il più possibile simmetrica. Per questo motivo variabili con range di più di una magnitudine di 10 sono spesso trasformate logaritmicamente. Se in una variabile è presente il valore zero, la radice quarta è una buona alternativa al logaritmo.

I risultati della PLSR, come quelli di tutti i modelli di proiezione, dipendono dallo scaling dei dati: con uno scaling appropriato si può guidare il modello su variabili Y più importanti, oppure aumentare il peso di variabili X con più informazioni.

Per migliorare la facilità di interpretazione e la stabilità numerica è raccomandabile centrare i dati prima dell'analisi. Questo si può ottenere sottraendo la media da tutte le variabili presenti in X e Y, prima o dopo lo scaling. Altri valori possono essere sottratti al posto della media, come i valori in punti conosciuti: in questo caso l'analisi è fatta riferendosi alla deviazione rispetto a questi punti. Il centering non modifica l'interpretazione dei dati.

In assenza di conoscenze sull'importanza relativa delle variabili, l'approccio standard è quello di scalare ogni variabile a una varianza unitaria dividendola per la sua SD e centrarla sottraendo la media, eseguendo quindi quello che viene definito un auto-scaling. Ciò corrisponde a dare a ogni variabile lo stesso peso e la stessa importanza nell'analisi. In alcuni casi tuttavia, come nell'ambito della chimica analitica, all'auto-scaling sono preferibili i dati X non scalati o a un livello medio di scaling [7].

E' possibile normalizzare anche le osservazioni Y, per esempio nell'analisi di profili spettrali o di cromatografi. In questi casi la normalizzazione è spesso eseguita ponendo la somma di tutti i picchi di un profilo a 100 o 1000. Ciò rimuove la grandezza delle osservazioni, che può essere desiderabile in caso essa sia irrilevante.

3.3 Il modello della PLSR

Il modello della PLSR lineare trova "nuove" variabili, stimate dalle Latent Variables(LV) o dalle loro rotazioni. Queste nuove variabili sono chiamate X-scores e indicate come t_a (a=1,2,…,A). Gli X-scores sono predittori di Y e modellano X. Sono "pochi"(A in numero) e ortogonali. Vengono stimati come una combinazione lineare delle variabili originali x_k con coefficienti w_{ka}, definiti "pesi".

$$t_{ia} = \sum_k W_{ka} X_{ik} \quad (T = XW)$$

Gli X-scores(t_a) hanno le seguenti proprietà:

1) Moltiplicati con i loadings p_{ak} sono buoni predittori di X, in modo tale che gli X-residuals e_{ik} siano piccoli.

$$X = \sum_a t_{ia} p_{ak} + e_{ik} \quad (X = TP' + E)$$

2) Con una Y multivariata (M>1) i corrispondenti Y-scores moltiplicati per i pesi c_{am} sono buoni predittori di Y, in modo tale che i residui g_{im} siano piccoli.

$$y_{im} = \sum_a u_{ia}\, c_{am} + g_{im} \quad (Y = UC' + G)$$

3) Gli X-scores sono buoni predittori di Y.

$$y_{im} = \sum_a c_{ma} t_{ia} + f_{im}$$

Gli Y-residuals f_{im} esprimono la deviazione tra la risposta osservata e quella modellata e formano la matrice F.

Quest'ultima equazione può essere riscritta in maniera da essere simile a un modello di regressione multiplo:

$$y_{im} \sum_a c_{ma} \sum_k w_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im}$$

$$(Y = XWC' + F = XB + F)$$

I coefficienti di regressione b_{mk} possono essere scritti come:

$$b_{mk} = \sum_{a} c_{ma} w_{ka} \quad (B = WC')$$

E' importante notare che i coefficienti b non sono indipendenti, a meno che A (il numero di componenti della PLSR) eguagli K(il numero di variabili X).
Un interessante caso si ha quando esiste un'unica variabile y e X'X è diagonale. In questo caso non c'è correlazione in X, e la PLSR arriva ad una soluzione con un solo componente, come la MLR.
Dopo aver calcolato un componente t_a, la matrice X è "ridotta" sottraendo $t_{ia}p_{ka}$ da x_{ik}. Questo fa in modo che il modello della PLSR, espresso tramite i pesi w_a, sia riferito ai residui della dimensione precedente E_{a-1}, invece che alle variabili X stesse. Possiamo quindi riformulare l'espressione della t_{ia} come:

$$t_{ia} = \sum_{k} w_{ka} e_{ik,a-1} \quad (t_a = E_{a-1} W_a)$$

dove l'errore $e_{ik,a-1}$ è dato da:

$$e_{ik,a-1} = e_{ik,a-2} - t_{i,a-1} p_{a-1,k} \quad \left(E_{a-1} = E_{a-2} - t_{a-1} p'_{a-1}\right)$$

$$e_{ik,0} = X_{ik} \quad (E_0 = X)$$

I pesi w possono comunque essere trasformati in modo da essere direttamente relazionati alla matrice X. La relazione è data da:

$$W = W(P'W)^{-1}$$

Anche la matrice Y può essere "ridotta" sottraendo $t_a c_a'$, ma non è necessario, in quanto i risultati non subirebbero modifiche.

3.4 Interpretazione del modello PLSR

Una possibile interpretazione della PLSR è che essa formi nuove variabili $x(t_a)$ come combinazioni lineari delle vecchie x, e che successivamente usi queste t_a come predittrici di **Y**. Da qui è possibile vedere come la PLSR sia basata su un modello lineare.

Tutti i parametri (**t**,**u**,**v**,**p**,**c)** sono calcolati dall'algoritmo(che sarà descritto in seguito). Per l'interpretazione del modello, i vettori score (**t** e **u**) contengono le informazioni sui dati e la loro similarità rispetto al modello e al problema dato, mentre i pesi $\mathbf{w_a}$ e $\mathbf{c_a}$ danno informazioni su come le variabili si combinano per formare la relazione quantitativa tra **X** e **Y**, fornendo quindi un'interpretazione degli score $\mathbf{t_a}$ e $\mathbf{u_a}$. Questi pesi sono utili per comprendere quali variabili X sono importanti (valori alti di w_a) e quali variabili ci danno lo stesso grado di informazione (valori simili di w_a).

I pesi w_a esprimono sia la correlazione positiva tra X e Y che la "compensazione della correlazione" necessaria a predire Y da X, eliminando la varianza secondaria da X.

Quest'ultima è formata da tutto ciò che varia in X che non è primariamente legato a Y. Ciò rende w_a difficile da interpretare direttamente, specialmente nelle componenti con a>1. Si può eventualmente usare un'espansione ortogonale dei parametri X (O-PLS) per ottenere la parte di w_a direttamente correlata a Y, rendendo quindi l'interpretazione più chiara.

La parte dei dati non spiegata dal modello, i residui, sono di grande interesse diagnostico.

Residui di Y elevati indicano un modello malfunzionante, e un plot della probabilità normale di un singolo residuo è utile per identificare outliers nella relazione tra T e Y. Nella PLSR abbiamo anche i residui di X: la parte dei dati non usata nella modellazione di Y. Questi residui di X sono utili per identificare outliers nello spazio delle X o processi che deviano dallo standard[7].

3.5 Algoritmo della PLSR: NIPALS

Esistono diverse varianti dell'algoritmo, da usare a seconda del tipo e della forma dei dati, e molti di questi sopportano un moderato numero di dati mancanti.
La maggior parte degli algoritmi lavorano con le matrici originali **X** e **Y**(scalate e centrate). Alternativamente i cosiddetti "kernel algorithms" lavorano con le matrici di covarianza **X'X, Y'Y** e **X'Y.** Essi sono più vantaggiosi quando il numero di osservazioni N differisce molto da quello delle variabili(K e M).
Il primo algoritmo implementato per la PLS è chiamato NIPALS (Wold e al., 2001). Esso parte con dati X e Y opportunamente trasformati, scalati e centrati, e prosegue come segue(notare come con una singola y il processo non sia iterativo).

1) Si sceglie un vettore di partenza **u**, solitamente una delle colonne di **Y**. Con una singola y, u=y.

2) Si calcolano i pesi di X, **w:**

$$w = \frac{Xu'}{u'u}$$

Si normalizza **w** in modo tale che ||**w**||=1.0

3) Si calcolano gli scores di X:

$$t = Xw$$

4) Si calcolano i pesi di Y:

$$c = \frac{Y't}{t't}$$

5) Si trova un nuovo set di scores di Y:

$$u = \frac{Yc}{c'c}$$

6) La convergenza è testata sui cambiamenti di **t:**

$$\frac{|t_{old} - t_{new}|}{|t_{old}|} < epsilon$$

dove epsilon è compresa tra 10^{-6} e 10^{-8}.

Se la convergenza non viene raggiunta si ritorna al 2), in caso contrario si continua con 7) e si torna poi a 1). Se è presente una sola variabile y, la procedura converge in una singola interazione e si procede direttamente a 7).

7) Si rimuovono i componenti già calcolati dalle matrici **X** e **Y,** per poi usare queste matrici ridotte come basi per le iterazioni successive. La riduzione di Y è in realtà opzionale: i risultati non cambiano a seconda che questa matrice sia ridotta o meno.

$$p = \frac{X't}{t't}$$

$$X = X - tp'$$

$$Y = Y - tp'$$

8) Si continua con la nuova componente fino a che la cross-validazione indica che non è più presente informazione significativa in X riguardo a Y.

E' interessante notare come il primo vettore dei pesi (w_1) è il primo autovettore della matrice combinata di varianza e autovarianza **X'YY'X** e i successivi vettori dei pesi sono gli autovettori della versione ridotta della matrice $\mathbf{Z_a'YY'Z_a}$, dove $\mathbf{Z_a}=\mathbf{Z_{a-1}}-\mathbf{T_{a-1}P_{a-1}'}$. Similmente, il primo vettore degli score(t_1) è un autovettore di **XX'YY'**, e i successivi vettori degli score autovettori di $\mathbf{Z_aZ_a'YY'}$.

Queste relazioni tra autovettori mostrano anche che i vettori $\mathbf{w_a}$ formano un set ortonormale e che i vettori $\mathbf{t_a}$ sono ortogonali tra di loro. I vettori di loading($\mathbf{p_a}$) non sono ortogonali, come non lo sono gli score di Y, $\mathbf{u_a}$. I vettori u e p sono ortogonali rispettivamente ai vettori t e w per tutte le componenti precedenti: $\mathbf{u_b't_a}=0$ e $\mathbf{p_b'w_a}=0$, se b>a. Inoltre, $\mathbf{w_a'p_a}=1$. [7]

Nel corso degli anni l'algoritmo NIPALS originario è stato più volte modificato ed evoluto. Uno dei suoi principali problemi è dato dal fatto che tutti i pesi $\mathbf{w_a}$ per a=2,3,...,A sono applicati a una matrice di residui X_{a-1} e non alla matrice originaria dei dati X_0. Questo rende notevolmente più difficile l'interpretazione degli scores **t**, poiché non si hanno informazioni sulle matrici ridotte X_s per a>1. La relazione tra fattori e variabili è mostrata in modo migliore dai loadings p_a, facendo quindi in modo che i pesi w abbiano poca importanza nell'interpretazione del modello di regressione.

E' quindi vantaggioso cercare di esprimere i fattori t_a in termini della matrice originaria dei dati X_0:

$$t_a = X_0 r_a \quad (T = X_0 R)$$

E' stato dimostrato che è possibile ricavare i vettori che compongono R a partire dalla matrice indentità I [9]:

$$r_a = G_a w_a \quad a = 1, \dots, A$$

$$G_{a+1} = G_a - r_a p_a' \quad a = 1, \dots, A-1$$

3.6 Algoritmo della PLS: SIMPLS

Il software Eigenvector, usato per la creazione dei modelli presenti in questo lavoro, sfrutta un diverso algoritmo, di nome SIMPLS, sviluppato nel 1993 da De Jong e al. La principale modifica proposta rispetto al NIPALS è la diretta computazione dei pesi R, in modo da evitare la costruzione delle matrici ridotte $X_{1,\dots,}X_A$ e $Y_{1,}\dots,Y_{A,}$ insieme al calcolo della matrice dei pesi W. [9]

La matrice R così definita è simile, ma non uguale, alla matrice R 'standard' descritta precedentemente.

Uno degli aspetti centrali dell'algoritmo è il calcolo dei vettori dei pesi r_a e q_a in maniera tale da poterli applicare direttamente ai dati centrati:

$$t_a = X_0 r_a \qquad a = 1, \dots, A$$

$$u_a = Y_0 q_a \qquad a = 1, \dots, A$$

Nello sviluppo dell'algoritmo sono state applicate quattro condizioni:

1) massimizzazione della covarianza: $u_a't_a = qa'(Y_0'X_0)r_a = max!$
2) Normalizzazione dei pesi r_a: $r_a'r_a=1$
3) Normalizzazione dei pesi q_a: $q_a'q_a=1$
4) Ortogonalità degli scores t: $t_b't_a= 0$ per a>b

Senza quest'ultima condizione la soluzione sarebbe solo una: r_1 e q_1 sono rispettivamente il primo vettore destro e sinistro della matrice $S_0=X_0Y_0$. Per ottenere più di una soluzione e generare un set di fattori ortogonali è stata inserita la quarta condizione, che può anche essere espressa come:

$$t_b' t_a = t_b' X_0 r_a = (t_b' t_b) p_b' r_a = 0 \qquad per\ a > b$$

In questa formula p_b è il vettore dei loadings che esprime la relazione tra la le variabili X originali e il fattore b-iesimo della PLS. La formula afferma che qualsiasi nuovo vettore r_a (a>1) deve essere ortogonale a tutti i precedenti vettori di loading, quindi alle colonne di $P_{a-1}=[p_1,p_2,...,p_{a-1}]$. Definendo il versore di ortogonalità come:

$$P_{a-1}^{\perp} = I_p - P_{a-1}(P_{a-1}' P_{a-1})^{-1} P_{a-1}'$$

è possibile ricavare che

$$r_a = P_{a-1}^{\perp} r_a, \qquad per\ a > 1$$

La soluzione per r_a e q_a è data quindi dai primi 2 vettori della decomposizione ai valori singolari(SVD) di S_0 proiettata su un sottospazio ortogonale a P_{a-1}.

$$S_a \equiv P_a^{\perp}(X_0'Y_0) = P_a^{\perp}S_0$$

E' possibile calcolare S_a a partire da S_{a-1}, sfruttando una base ortonormale di P_a, definita come $V_a=[v_1,v_2,...,v_a]$.

$$S_a = S_{a-1} - v_a(v_a'S_{a-1}) \qquad per\ a > 1$$

Da questa formula è possibile notare la principale differenza dell'algoritmo SIMPLS rispetto a quello classico: il processo di riduzione è applicato alla matrice S_0 e non alle più larghe matrici dei dati X_0 e Y_0.
I valori stimati dei campioni usati in calibrazione sono calcolati grazie alla formula:

$$\widehat{Y_0} = TT'Y_0 = X_0RR'X_0'Y_0 = X_0RR'S_0$$

da cui possiamo definire la matrice B_{PLS} dei coefficienti di regressione

$$B_{PLS} = R(R'S_0) = R(T'Y_0) = R(T'Y) = RQ'$$

dove Q sono i vettori di loading della matrice Y.

3.7 Sviluppo del modello

Per lo sviluppo di un modello corretto è di fondamentale importanza avere sin da subito una buona conoscenza del sistema, in particolar modo è importante avere chiaro quali sono le proprietà Y che si vuole analizzare e quali predittori X devono essere variati e misurati. E' inoltre necessario partire da buoni dati, sia per le Y che per le X.

Le Y multivariate danno maggiore informazione, in quanto possono essere prima analizzate tramite PCA. Ciò fornisce una buona idea sul livello di variazione in Y o quali variabili dovrebbero essere analizzate insieme, ad esempio. La prima informazione che si può estrarre dal modello è il numero A di componenti principali, ovvero la complessità del modello e quindi del sistema, dato essenziale per qualsiasi modello empirico.

Con numerose variabili X correlate fra loro c'è un forte rischio di ottenere un over-fitting: un modello che fitta bene i dati ma che ha poco o nessun potere predittivo.

E' perciò necessario un test della significatività predittiva di ogni componente principale, in modo da fermarsi quando le componenti cominciano ad essere non significative.

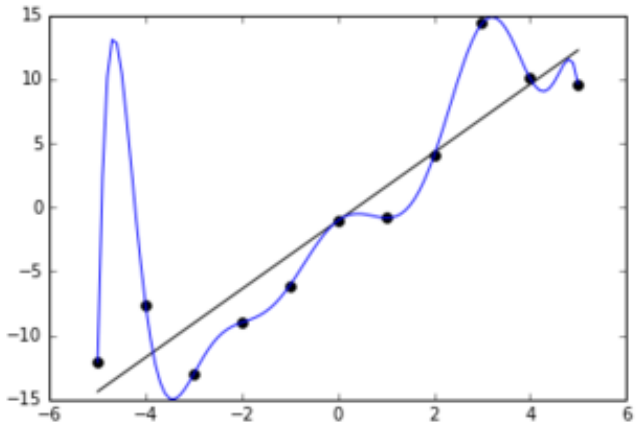

Fig. 2.3 Overfitting dei dati sperimentali

La cross-validazione(CV) è un metodo pratico e affidabile per testare questa significatività predittiva, tanto da diventare lo standard nell'analisi PLSR. La CV è effettuata dividendo i dati in un numero di gruppi G e successivamente sviluppando un insieme di modelli paralleli a partire dai dati ridotti grazie alla cancellazione di un insieme G alla volta. E' importante sottolineare come l'utilizzo di G=N ("leave one out") non sia raccomandato.

Dopo aver sviluppato un modello vengono calcolate le differenze tra i valori Y veri e quelli predetti per i dati eliminati.

La somma dei quadrati di queste differenze è calcolata e unita a quella degli altri modelli, per ottenere il parametro PRESS (Predictive Residuals Sum of Squares), che stima la capacità predittiva del modello.

Quando la CV è usata nella modalità "sequenziale" viene eseguita su un componente dopo l'altro, ma il "peeling-off" è eseguito solo una volta sulla matrice dei dati, e le conseguenti matrici dei residui **E** e **F** sono divise in gruppi per la CV delle componenti successive. Il rapporto $PRESS_a/SS_{a-1}$ è calcolato dopo ogni componente, che viene giudicato significativo se questo rapporto è minore di 0.9 per almeno una variabile Y. SS_{a-1} indica la somma dei quadrati dei residui prima della componente corrente *a*. Il calcolo continua fino a che una componente non è più significativa.

Alternativamente può essere usato un approccio "totale": si dividono inizialmente i dati in gruppi, per poi calcolare il PRESS per ogni componente fino a un numero a scelta(10 o 15 per esempio), eseguendo separati "peeling" della matrice dei dati di ogni gruppo. Il modello per cui si ottiene il valore minore di PRESS/(N-A-1) viene infine usato.

Con entrambi i modelli viene calcolato un PRESS del modello finale con il numero stimato di componenti significative. Questo viene spesso espresso come Q^2, calcolato con (1-PRESS/SS), dove SS è la somma dei quadrati di **Y** corretti con la media.

Questo può essere comparato con R^2=(1-RSS/SS), dove RSS è la somma dei quadrati dei residui. In modelli con diverse Y, si ottengono un R_m^2 e un Q_m^2 per ogni variabile y_m. Queste misure possono essere rispettivamente espresse come errore quadratico medio ottenuto durante la fase di calibrazione (Root Mean Square Error of Calibration: RMSEC) e durante la fase di validazione (Root Mean Square Error of Cross Validation: RMSECV).

Nella figura è rappresentato un possibile andamento in funzione del numero di variabili latenti del modello sia dell'errore di calibrazione (RMSEC) sia l'errore di validazione (RMSECV).

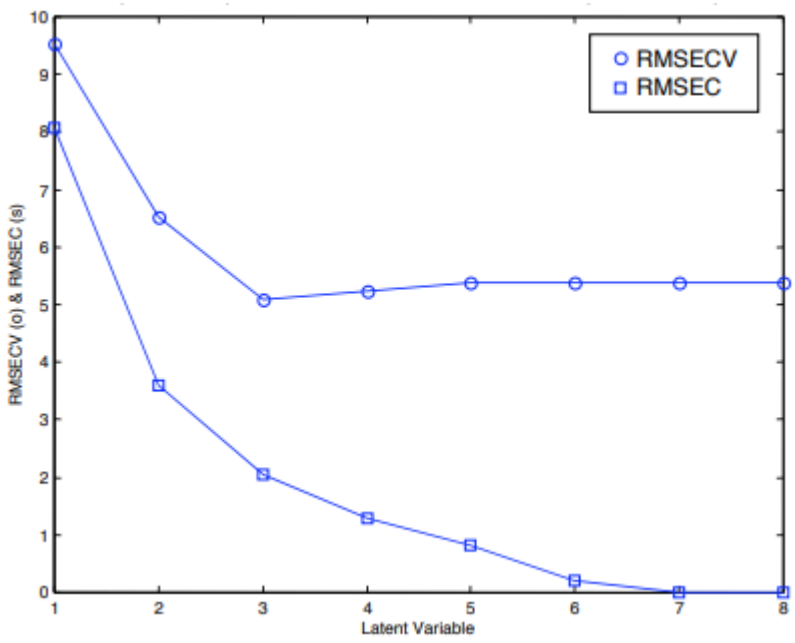

Fig. 2.3 Grafico dell'andamento di RMSEC e RMSECV.[8]

Si può osservare come l'errore di calibrazione diminuisca in maniera monotona all'aumentare del numero di variabili latenti, mentre l'errore di validazione presenta un valore minimo. L'andamento monotono di RMSEC indica come incrementando il numero di variabili latenti il modello tende a inglobare anche il rumore dei dati. RMSECV d'altro canto fornisce l'andamento dell'errore di predizione, ovviamente in fase di predizione la quantità di rumore dei dati sarà

sicuramente differente quindi la parte di modello che si adatta sul rumore dei dati di calibrazione provoca un errore quando viene utilizzata con dati con rumore differente. [8]

Il valore minimo di RMSECV indica quindi il numero di variabili latenti per il quale è massima la descrizione della parte deterministica dei dati ed oltre il quale il modello inizia a rappresentare il rumore dei dati di calibrazione. Il numero di variabili latenti in corrispondenza dei quali si ottiene il minimo di RMSECV indica le variabili latenti ottimali da considerare per il modello.

Un ulteriore miglioramento del modello si ottiene grazie agli score plots (u,t) delle prime due/tre dimensioni del modello, con i quali è possibile investigare la presenza di outliers, curvature o gruppi nei dati, caratteristiche che possono creare problemi al modello, insieme a valori bassi di R^2 e Q^2.

Per i singoli outliers deve essere investigata la bontà dei dati, e se questo non aiuta, possono essere eliminati dal modello solo se poco importanti. Le curvature nel plot (u,t) possono essere migliorate trasformando una parte dei dati (per esempio con una funzione log), oppure introducendo termini quadratici o cubici nel modello.

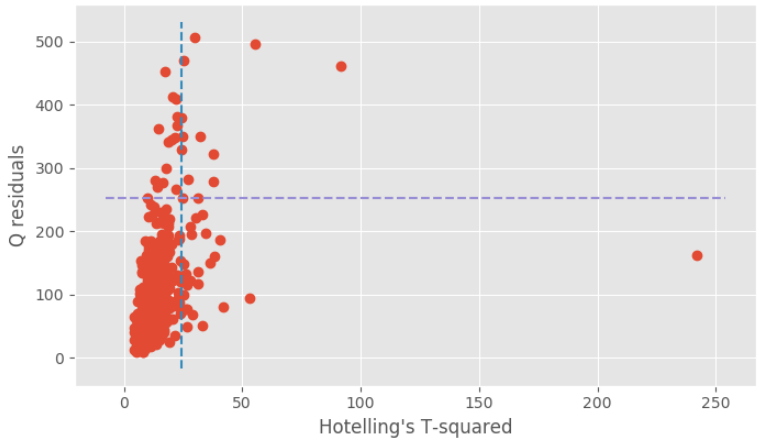

Fig 2.5 - Esempio di outliers nell'analisi PLSR.

Una volta eliminati i precedenti problemi si può cercare di ridurre il modello eliminando i dati poco importanti, ovvero quelli con un basso coefficiente di regressione o un basso valore del parametro VIP (Variable Importance in Projection).
Successivamente a ciò il modello è sviluppato, e può essere validato e interpretato.

3.8 Matrici X e Y incomplete (dati mancanti)

I metodi di proiezione come la PLSR tollerano numeri moderati di dati mancanti sia in X che in Y. Per avere dati mancanti in Y questa deve essere multivariata, ovvero deve avere almeno 2 colonne. Maggiori elementi hanno X e Y, maggiore è la percentuale di dati mancanti tollerata: per data set piccoli con circa 20 osservazioni e 20 variabili, circa il 10-20% di dati mancanti può essere ammesso, a meno che le mancanze non seguano un pattern preciso. [7]
L'algoritmo NIPALS della PLS si occupa automaticamente dei dati mancanti, sostituendoli iterativamente con predizioni del modello. Ciò corrisponde a fornire per ogni componente il valore del dato che ha zero residui e dunque nessuna influenza sui parametri t_a e p_a. Sono stati sviluppati anche altri approcci con algoritmi EM, migliori in casi di mancanza di una percentuale alta di dati. E' comunque importante ricordare che in questi casi qualsiasi predizione è incerta.

3.9 Errore di stima e intervalli di confidenza

Sono stati fatti numerosi sforzi per derivare in maniera teorica gli intervalli di confidenza dei parametri della PLSR. La maggior parte di questi sono però basati su assunzioni di regressione, dove si considera la PLSR come un modello parziale di regressione.

Solo con il lavoro di Burnham e al. (1999) si è cominciato a investigare questo problema ponendo la PLSR come un modello di regressione con variabili latenti.

Un modo di stimare l'errore standard e gli intervalli di confidenza direttamente dai dati è definito “jack-knifing”, raccomandato da Wold nel suo lavoro originale sulla PLS e successivamente rivisitato da Martens e Martens e al. (2000). L'idea è semplice: la variazione nei parametri dei vari sotto-modelli ottenuti durante la cross-validazione è usata per determinare la loro deviazione standard (chiamata errore standard), successivamente grazie alla distribuzione di t si determinano gli intervalli di confidenza. Poichè tutti i parametri della PLSR (scores, loadings, ecc...) sono combinazioni lineari dei dati originali, questi parametri sono vicini ad avere una distribuzione normale, facendo in modo che il jack-knifing funzioni correttamente.

3.10 Validazione del modello

Ogni modello prima di essere utilizzato deve essere validato. La migliore validazione di un modello è una buona capacità di stimare in modo costante e preciso i valori di Y da un nuovo set di dati X, un set di validazione, che è tuttavia difficile da trovare rappresentativo e indipendente. In assenza di esso due ragionevoli modi di eseguire la validazione sono la cross-validazione, che simula come il modello predica nuovi dati, e la stima del modello dopo una randomizzazione dei dati, trovando quindi la probabilità di ottenere un buon fitting con dati random.

Capitolo 4

Descrizione del sistema

1. Sensore

Il sistema utilizzato è un dispositivo elettronico a batteria, creato per eseguire misure sull'umidità del terreno. Esso è diviso in 2 parti: una sezione metallica rettangolare e aperta che funge da guida d'onda e un box contenete i circuiti elettronici. Il box presenta inoltre un connettore USB, per permettere la connessione a PC, un indicatore della carica della batteria, un pulsante di accensione e un connettore jack per caricare la batteria interna. La guida d'onda contiene al suo interno una coppia di antenne, trasmittente e ricevente. L'intero sensore ha la caratteristica di essere completamente impermeabile, essendo stato progettato per resistere agli agenti atmosferici.

Fig. 4.1 - Sensore per l'umidità

Il sensore sfrutta i principi della spettroscopia d'impedenza descritti in precedenza per predire il valore di umidità del terreno, basandosi su una serie di acquisizioni dei valori di guadagno e fase dell'impedenza di esso. Ciò è reso possibile dalla trasmissione in direzione ortogonale al terreno di un segnale a radiofrequenze nella banda 1.5-2.7 Ghz.

Un'onda elettromagnetica che incide su di un'interfaccia tra due mezzi diversi può subire un cambiamento di direzione (rifrazione) o una riflessione nelle rispettive componenti di ampiezza e fase in funzione delle caratteristiche dei due mezzi e dell'onda stessa. Utilizzando un'antenna per trasmettere al terreno un'onda elettromagnetica ed un'antenna in grado di ricevere la parte di onda riflessa si riesce a misurare le variazioni di ampiezza e fase tra le due onde, in modo da risalire alle caratteristiche dielettriche del terreno stesso e quindi al suo livello di

umidità.

Il range di frequenze è stato scelto in quanto l'informazione di umidità misurata dal sensore è contenuta in diversi sotto-insiemi di questa banda, a seconda del tipo di campione analizzato. [10]

Il circuito elettronico genera quindi un segnale alla frequenza minima di 1.5 GHz, che viene trasmesso verso il terreno attraverso la guida d'onda, e attraverso l'antenna ricevente rileva la parte dell'onda riflessa dal terreno verso il circuito. Infine i 2 segnali vengono processati, ricavando i valori di guadagno e fase.

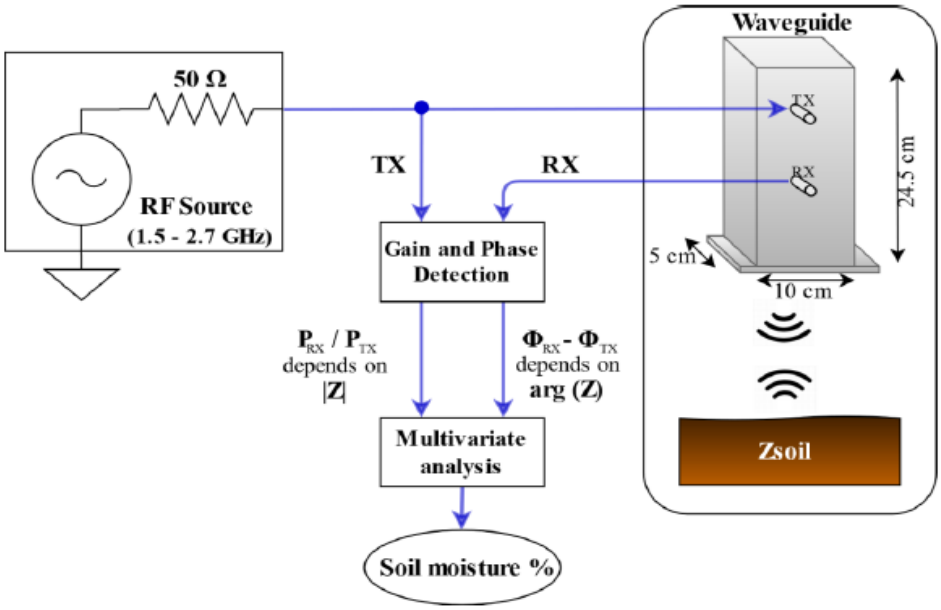

Fig 4.2 - Principio di funzionamento[10]

L'intera procedura di misura descritta è ripetuta per tutte le frequenze nel range 1.5-2.7 GHz, ottenendo un array di valori per il guadagno e uno per la fase. Questa coppia di array viene elaborata con un algoritmo di analisi multivariata, che permette di predire il valore di umidità del terreno.

1.1 Circuito elettronico

L'architettura del circuito può essere divisa in 4 sezioni principali:

1. *Control, Elaboration and Memorization;*
2. *RF Wave Generation;*
3. *Gain & Phase Measurement;*
4. *Power.*

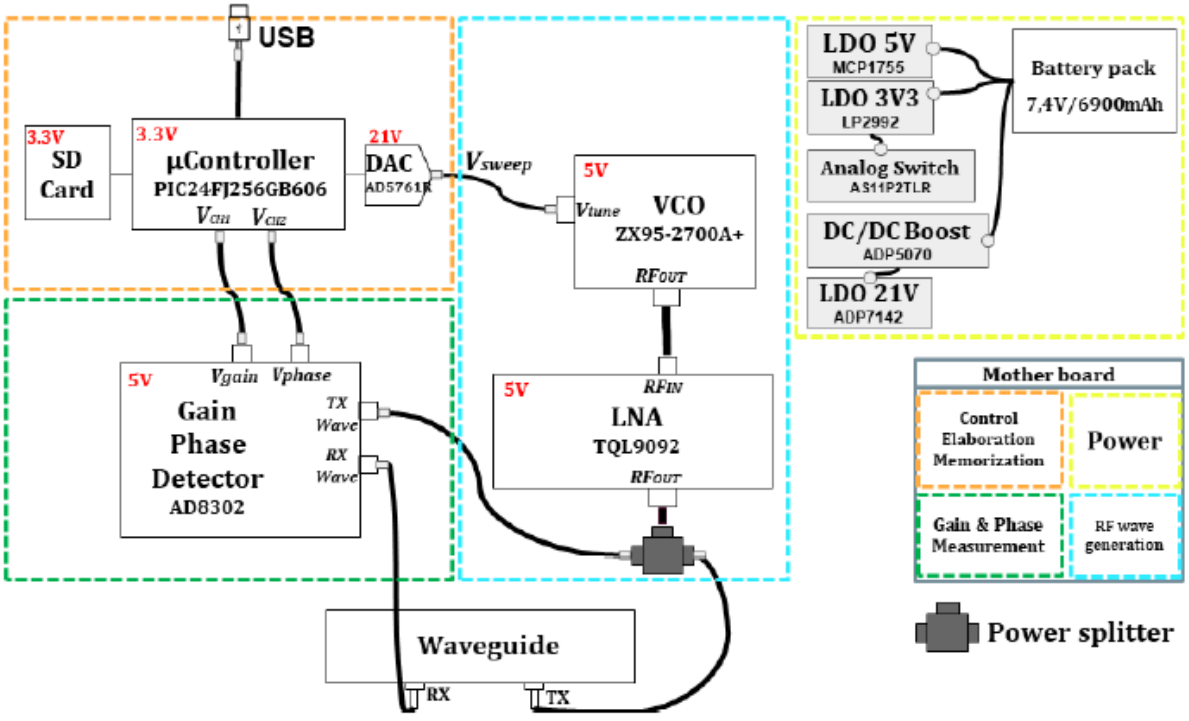

Fig.4.3 - Architettura del circuito elettronico [11]

1.1.1 Controllo, elaborazione e memorizzazione

Questa sezione è il “cervello” del circuito e svolge diverse funzioni. Consiste in un microcontrollore, un convertitore D/A, una SD memory card e un convertitore seriale-USB.

Microcontrollore

Il microcontrollore, o MCU (MicroController Unit), è un sistema elettronico integrato su di un singolo chip. Contiene una o più CPUs (Central Processing Units), una memoria e una serie di periferiche programmabili. Viene generalmente utilizzato in ogni sorta di sistema integrato come unità logica di controllo in quanto, grazie alla relativa semplicità di utilizzo e alla notevole versatilità, permette l’interfacciamento ad un numero considerevole di componenti elettronici.

In questa applicazione il microcontrollore gestisce il processo di misurazione, la comunicazione con l’interfaccia seriale e il salvataggio di dati sulla SD card. Per eseguire questi compiti è stato scelto di utilizzare un MICROCHIP PIC24FJ256GB606.

Convertitore D/A

Il convertitore D/A è un Analog Devices AD5761R a 16 bit. Il suo compito è quello di trasformare i valori digitali dati dal microcontrollore in un voltaggio analogico. Il convertitore ha una risoluzione di 16 bit, un range di tensione in uscita che può andare da 0 a 20 V(unipolare o bipolare), un rumore in output di 35 , un massimo INL di ±2 LSB e un tempo di setting massimo di 12.5 µs(con un gradino di 20 V). [11]

Il convertitore presenta un doppio buffer in ingresso composto dall'Input register e dal DAC register, che permette l'aggiornamento asincrono dell'uscita. Per il suo funzionamento esso utilizza un circuito interno che produce una tensione di riferimento pari a 2.5 V e che può essere fornita su di un pin d'uscita per un'eventuale riferimento esterno. E' presente infine un'interfaccia SPI 4-wire per la comunicazione con le periferiche esterne, che può raggiungere la frequenza di clock di 50 MHz, controllata da quattro pin chiamati rispettivamente *SYNC*, *SCLK*, SDI e SDO. Il pin di *SYNC* funge da Slave Select e deve essere mantenuto basso durante la fase di comunicazione col microcontrollore per un numero corretto di cicli di clock.

SD Memory Card

L'applicazione richiede il salvataggio delle misure compiute dal sensore su di un supporto fisico di memorizzazione non volatile e che possa essere trasferito semplicemente su PC. Inoltre, vista la presenza del microcontrollore, si richiede che la memorizzazione dei dati avvenga attraverso un protocollo gestibile dal microcontrollore stesso (SPI, USB, I^2C, ecc.). La scelta si è quindi indirizzata fin da subito sulla scheda di memoria SD, largamente utilizzata nei dispositivi elettronici presenti nel mercato odierno grazie alla possibilità di memorizzazione di notevoli quantità di dati in un supporto di dimensioni ridotte, e dotata di interfaccia di comunicazione SPI Standard a 3,3 V. La scheda di memoria utilizzata nel progetto ha capacità pari a 256 MB: considerando che ogni campione occupa quattro byte (due per il guadagno e due per la fase) ed il numero di campioni è pari a 7400, essa è in grado di memorizzare più di 9000 misurazioni. Presenta velocità di trasferimento massime pari a 23 MB/s in scrittura e 17 MB/s in lettura ed il suo consumo di corrente è pari a 105 μA in *Idle Mode* e 50 mA massimi in *Active Mode.*

1.1.2 Generazione di onde RF

Per la generazione del segnale a radiofrequenze è stato realizzato un circuito dedicato che, sulla base dei dati forniti dal circuito di controllo, genera un'onda sinusoidale a frequenza fissa. Quest'ultima viene fornita alla guida d'onda per compiere l'analisi spettroscopica.
Questo circuito è suddiviso in due sotto-circuiti:

- VCO (*Voltage Controlled Oscillator*): realizza l'onda sinusoidale alla frequenza definita dal valore di tensione in ingresso, proveniente dal convertitore DAC.

- LNA (*Low Noise Amplifier*): amplifica il segnale d'uscita del VCO al fine di raggiungere il livello di potenza dell'onda trasmessa dato dalle specifiche.

Controller Oscillator (VCO)

Questa sezione consiste in una serie di oscillatori MiniCircuits ZX95-2700A+, i quali trasformano il voltaggio in output del DAC in un'onda sinusoidale ideale con frequenza dipendente dal voltaggio in input.
Le caratteristiche più interessanti di questo modello sono la frequenza dell'onda generata tra 1.3 GHz e 2.7 GHz (compatibile quindi con le specifiche del progetto), la potenza in uscita di 3.3 dBm, il basso rumore di fase, la tensione di alimentazione di 5V e la corrente massima di alimentazione di 35 mA. [11]
I componenti scelti sono contenuti in un case con tre connettori per l'alimentazione e il tuning della tensione e un connettore SMA per l'output.

Dal grafico della forma d'onda in uscita del VCO per diverse frequenze si può notare come il trend tra la frequenza di output e la tensione di tuning non è strettamente lineare, e che la potenza di output sia massima per una tensione di circa 8 V.

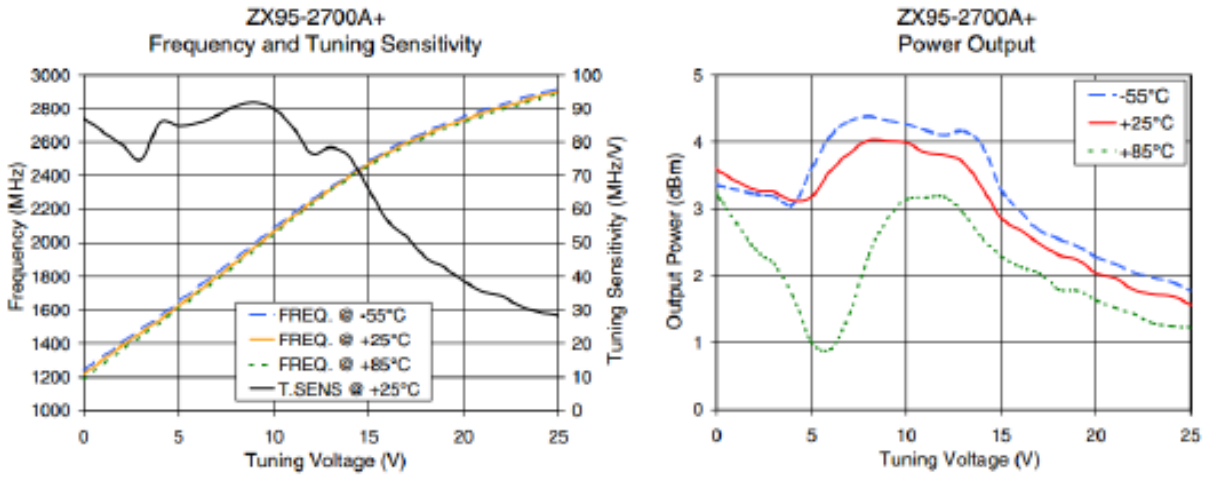

Fig. 4.4- frequenza e potenza di output dell'onda Rf rispetto alla tensione di tuning [10]

Low Noise Amplifier

Il segnale in uscita dal VCO presenta una potenza troppo bassa rispetto alle specifiche di progetto. Per questo motivo è necessario compierne un'amplificazione in modo da raggiungere la potenza trasmessa minima attraverso un circuito amplificatore a RF. Ciò è reso possibile dal QORVO TQL9092, un amplificatore ultra low noise con banda operativa da 0,6 a 4,2 GHz, in accordo quindi con le specifiche di frequenza del segnale a RF generato. Una specifica importante è il guadagno maggiore di 20 dB nella banda 1,5 – 3,8 GHz, ed è alimentato con una tensione di 5 V, conforme alle tensioni di alimentazione dei restanti componenti della parte a RF del circuito. La corrente massima di alimentazione del circuito in ON-STATE è di 85 mA.

1.1.3 Misura di guadagno e fase

Il segnale amplificato dall'LNA viene fornito alla guida rettangolare, che genera l'onda che raggiunge il terreno e riceve quella che quest'ultimo riflette, portando con sé l'informazione sulla misura di umidità. Il gain phase detector ha il compito di confrontare l'onda trasmessa e riflessa e fornire le informazioni misurate al microcontrollore. Esso include due porte d'ingresso *INPA* e *INPB* con connettori SMA che corrispondono rispettivamente all'onda riflessa e quella trasmessa, due porte d'uscita *VMAG e VPHS*, anch'esse con connettori SMA, che rappresentano il guadagno e la fase misurata dallo strumento ed un connettore per l'alimentazione. E' possibile grazie a questo strumento misurare variazioni del gain da -30 dB a + 30 dB e variazioni di fase tra 0º e 180º, mentre le tensioni d'uscita di guadagno e fase variano in un range da 0 V a 1,8 V. Inoltre, il componente rende disponibile una tensione di riferimento di 1,8V che funge da fondo scala per le tensioni d'uscita. [11]

1.1.4 Potenza

Le varie sezioni del circuito sono alimentate da diverse tensioni: è stato quindi necessario utilizzare un circuito di alimentazione che fornisse le giuste tensioni ai vari blocchi circuitali. Ciò è stato ottenuto inserendo una serie di regolatori LDO (MCP1755 - LP2992 - ADP7142), un regolatore DC/DC boost ADP5070 e uno switch analogico per dare energia al DAC e alla SD card. Il circuito include anche una batteria al litio da 7.4 V/6900 mAh.

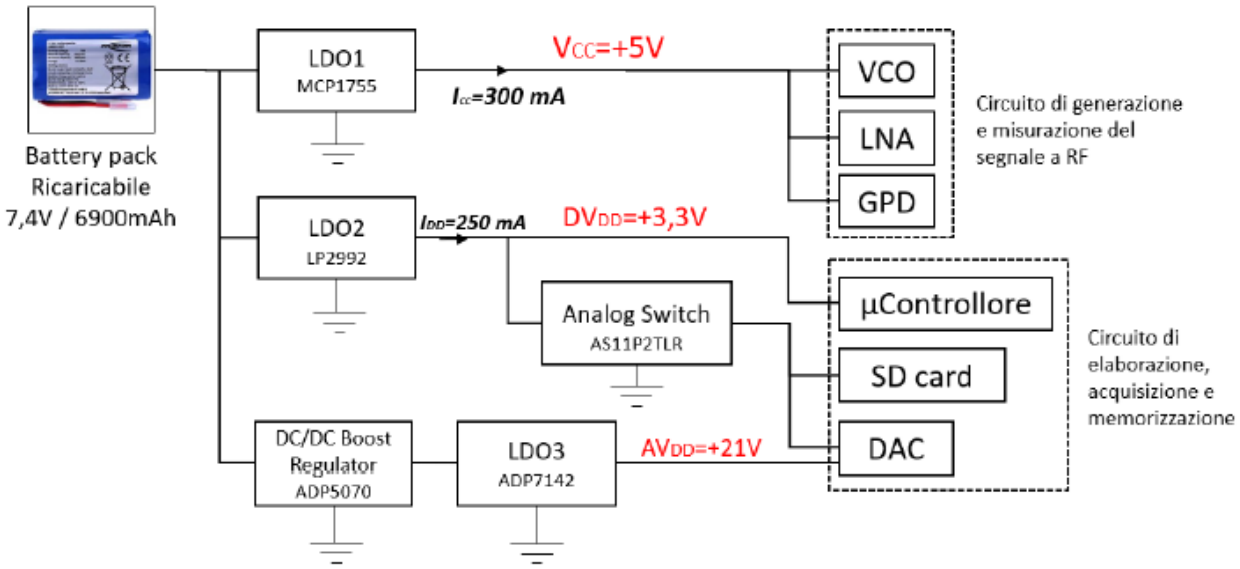

Fig. 4.5: Schema circuito di alimentazione [11]

1.2 Interfacce esterne

Nel box contenete il circuito elettronico sono state inserite alcune interfacce hardware esterne , che permettono un più semplice e appropriato funzionamento del sistema.

Lo switch ON/OFF è l'interruttore generale che connette la batteria al circuito e che permette di dargli potenza. Quando si trova nella posizione ON, si accende anche l'indicatore di carica della batteria, usato per monitorare il livello di carica della batteria . Quest'ultima ha una tensione nominale di 7.4 V e una tensione di scarica di 5 V.

Il tempo che la batteria impiega a scaricarsi è una funzione del voltaggio minimo a cui il circuito sta operando: nelle applicazioni pratiche si è messo il sensore in carica quando la tensione raggiunge circa 6 V. Per caricare la batteria è stato inserito un connettore jack DC nella parte inferiore destra del box, in cui può essere inserito il caricatore dedicato.

E' infine presente una porta USB, per connettere il sensore al PC grazie a un semplice cavo USB, e che presenta un rivestimento avvitabile, per renderla impermeabile e resistente alla polvere.

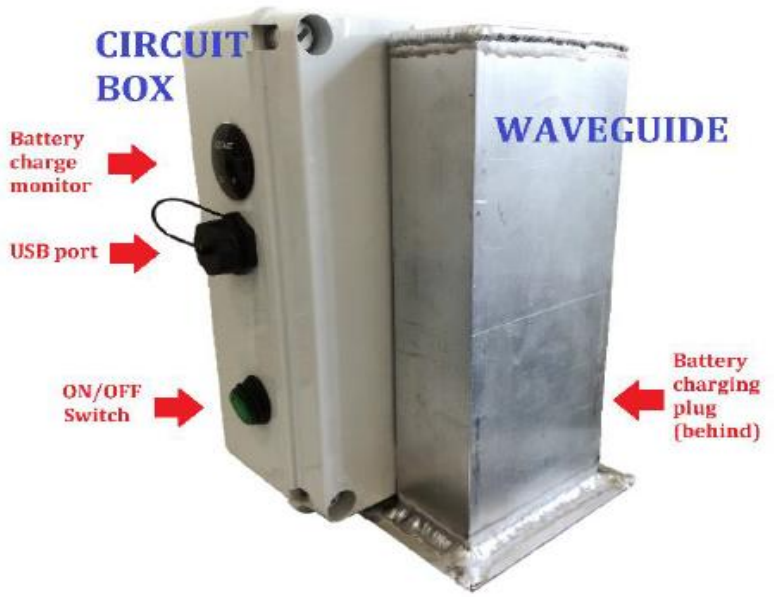

Fig. 4.6 - Interfacce hardware esterne del sensore [11]

2. Firmware

Il microcontrollore è stato programmato per mezzo dell'utilizzo del *software* di programmazione proprietario *Microchip Technology* MPLAB X IDE, dedicato alla famiglia di microcontrollori PIC e dsPIC a 8-bit, 16-bit e 32-bit. La scelta di un microcontrollore a 16-bit ha portato all'utilizzo del compilatore XC16 in linguaggio C. Per il trasferimento del codice sviluppato e per il relativo debug è stato utilizzato il programmatore PICkit3.

2.1 Sequenza di operazioni eseguite dal microcontrollore

La sezione digitale del sistema di misurazione viene coordinata dal microcontrollore, che comunica direttamente con il convertitore DAC e permette l'abilitazione delle alimentazioni fornite dai regolatori.
Il firmware originario, successivamente modificato per calcolare il valore di umidità, è stato creato per far eseguire al microcontrollore una serie di azioni, qui brevemente elencate:

1) Il microcontrollore inizializza le periferiche interne;

2) Si pone in un loop in attesa di un comando esterno, proveniente in questo caso dall'interfaccia in Matlab

3) Trasmette un nuovo valore di tensione al DAC;

4) Attende che il segnale trasmesso percorra l'intera catena di generazione RF e misurazione e che siano disponibili alle porte d'ingresso del modulo ADC i valori di guadagno e fase misurati.

5) Campiona e converte i suddetti valori, eseguendo una media tra 32 campioni misurati alla stessa frequenza. Successivamente salva i valori finali in una struttura dati, per poi inviare il valore della misurazione precedente alla seriale in modo che sia letto dal computer.

6) Qualora il valore trasmesso al DAC non coincidesse con l'ultimo valore (per il quale il VCO genererebbe la frequenza massima) si ritorna al punto 5. In caso contrario si prosegue.

7) Gli ultimi 2 valori finali di gain e phase vengono salvati nella struttura e scritti su seriale. Il sistema torna infine al loop del punto 2.

Nel seguito verranno analizzate in dettaglio le funzioni principali del *firmware*, seguendo quello che è il percorso descritto.

Inizializzazione delle periferiche

Come per ogni famiglia di microcontrollori PIC, il primo passo nella creazione del codice riguarda la sua configurazione iniziale; questa prevede la definizione di quelli che sono i "*Configuration bits*", che determinano il settaggio dei componenti di sistema all'accensione del dispositivo quali l'oscillatore primario, quello secondario, il *watchdog*, le porte di comunicazione con il PICkit3, ecc.
Il secondo passo della configurazione iniziale riguarda l'inizializzazione dei pin del microcontrollore, settando i loro valori di tensione iniziale, definendo quali fossero input o output e dichiarando i pin con uscita analogica.
In seguito il microcontrollore deve inizializzare le periferiche interne ed esterne

utilizzate. A questo scopo è presente una funzione chiamata ***SYSTEM_Initialize()*** che comprende diverse funzioni di inizializzazione, tra le quali le più importanti sono:

- ***OSCILLATOR_initialize()***: funzione che inizializza l'oscillatore primario e abilita quello secondario

- ***INTERRUPT_Initialize()***: funzione utilizzata per settare la priorità degli interrupt relativi alle periferiche. Prima di settare la priorità, la funzione attiva il *nesting* degli interrupt, con il quale ogni routine di interrupt a priorità maggiore può bloccare tutte le routine di interrupt a minore priorità. La priorità di un interrupt va dal valore di 1 al valore di 7; di norma tutti gli interrupt sono inizializzati ad una priorità pari a 4.

- ***TMR3_Initialize()***: funzione di inizializzazione del Timer 3, utilizzato durante la misura. Il periodo del timer è stato inizializzato sulla base del ritardo impiegato dal segnale a compiere il percorso μC-DAC-VCO-LNA-guida d'onda-GPD- μC. Questo ritardo è stato calcolato sulla base dei tempi che impiega il segnale ad attraversare i componenti della catena.

- ***ADC1_Initialize()***: funzione di inizializzazione del modulo A/D del microcontrollore. Quest'ultimo è stato inizializzato per compiere il campionamento continuo degli ingressi per mezzo della funzione *Auto-Sample* con la quale l'ADC campiona il nuovo dato una volta che ha completato la conversione di quello precedente in maniera del tutto automatica, memorizzando il campione nel registro a 16 bit ADC1BUFx.

Attesa del comando

Una volta completata l'inizializzazione il sistema si pone in uno stato di attesa, attraverso il main loop:

```
while(1){}

    return 1;
```

Il sistema attende un segnale esterno per cominciare la misurazione, segnale che il sistema legge nella seriale attraverso la funzione **U1RXInterrupt().**

Questa legge la seriale ed agisce in diversi modi a seconda del comando inviato da Matlab. Nel caso in esame il comando inviato per far partire la misura, in Test Mode, è il carattere "M", letto in ASCII come 0x4D. Una volta ricevuto il segnale, il sistema pone la variabile MODE a 0, in modo che si attivino tutte le funzionalità della Test Mode, e lancia la funzione **MeasureTask()**, che esegue la misura di gain e phase. All'interno di questa funzione è presente un ciclo for che esegue le misurazioni per 3700 valori di frequenza comprese tra Vstart e Vfinal:

for(V_DAC = Vstart;V_DAC<=Vfinal;V_DAC=V_DAC+Vstep)

Scrittura del valore di tensione al DAC

Come prima cosa si memorizza in una variabile *struct*, che contiene i dati della misurazione, il *timestamp* della misura, ossia l'istante di tempo nella quale avviene la misura, attraverso la funzione ***RTCC_TimeGet().*** Successivamente viene inviato al DAC il nuovo valore di tensione alla quale deve porsi la sua uscita per mezzo della funzione ***SPI1_Write2DACReg()***.

Avvio Timer e attesa

Una volta conclusa la fase di scrittura del dato al DAC il microcontrollore attende il tempo stimato di 250 μs prima di compiere la conversione A/D dei campioni. Questa operazione è effettuata grazie alla funzione ***TMR3_Start()***. In questo lasso di tempo il microcontrollore non esegue alcuna operazione.

Campionamento e conversione

Allo scadere del timer viene generato un interrupt grazie al quale il microcontrollore è pronto per campionare e convertire i canali di guadagno e fase. Nella funzione ***TMR3_CallBack()*** avviene la conversione e la memorizzazione dei campioni. I campioni di guadagno e fase vengono campionati otto alla volta e memorizzati nei relativi buffer in formato intero a 16 bit senza segno. Alla completa memorizzazione viene compiuta la media aritmetica degli otto valori di guadagno e degli otto valori di fase. Questo processo viene ripetuto una seconda volta e viene eseguita una media tra i 2 risultati finali in modo da mediare in totale tra 32 campioni. Dopo questa operazione il sistema torna alla funzione **MeasureTask()** che, prima di memorizzare i risultati finali nella struct apposita, scrive su seriale i valori calcolati nella misurazione precedente grazie alla funzione **UART1_writeSample().**

Fine della misurazione

Una volta completata la misurazione il firmware resetta il valore del DAC a 0, scrive gli ultimi valori di gain e phase su seriale e si riporta nel main loop, in attesa di eseguire una nuova misurazione.

3. Codice Matlab e GUI

Il processo di misurazione, sia nel Test Mode che nel Logger Mode(non modificato per questo lavoro di tesi), è controllato attraverso un interfaccia grafica(GUI) di Matlab, chiamata TerraGUI, che permette di eseguire misure, salvare i dati in formato .txt e programmare il sensore in modo da funzionare in maniera autonoma.

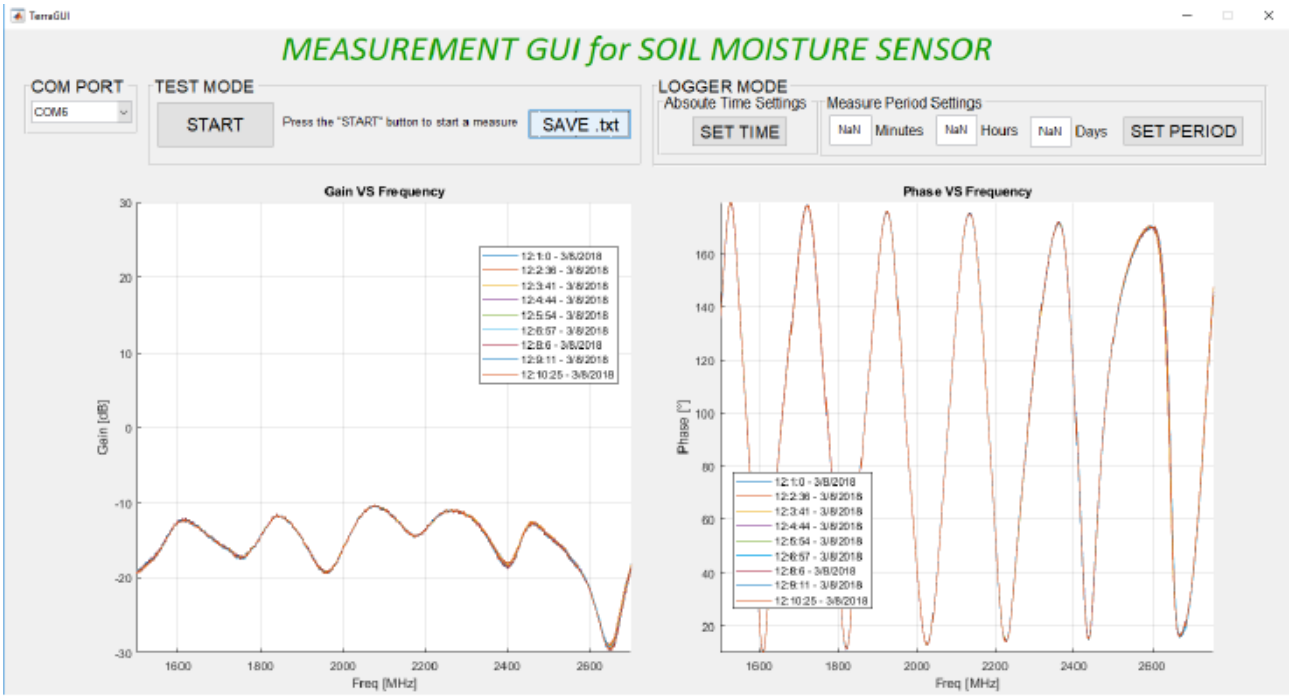

Fig. 4.7 – Interfaccia grafica in Matlab [10]

Nella parte in alto a sinistra è presente un pannello con un menù a tendina, da cui è possibile scegliere la porta seriale con la quale il sensore è connesso al PC tramite cavo USB.

La parte superiore comprende altri due pannelli: “TEST MODE” permette di eseguire una misura e salvare i risultati, “LOGGER MODE” di programmare il sensore per un funzionamento autonomo.

3.1 Test Mode

Il primo pannello presenta due pulsanti. Il primo, START, viene premuto per iniziare una misura, comportando una sua disattivazione fino al completamento di questa.

Quando il pulsante START viene premuto il programma salva il numero della porta seriale a cui è connesso il sensore ed esegue il programma “Terra.m”. Questo programma come prima cosa inizializza alcune variabili di interesse:

- Il numero di misurazioni, 3700
- L’array di frequenze, comprendente 3700 frequenze comprese tra 1498.779 MHz e 2757.009 MHz (trovate sperimentalmente)
- Il voltaggio Vref, 1.8 V
- Il numero di bit dell’ADC del microcontrollore, 12
- La sensitività di guadagno, 30 mV/dB, e di fase, 10 mV/°

Successivamente a questa parte di inizializzazione viene richiamato un altro programma, di nome “TerraSerialRead.m”, che ha lo scopo di leggere sulla seriale i dati inviati dal sensore durante la misura, ovvero il timestap e gli array di guadagno e fase. Inizialmente lo script chiude ogni connessione con altri dispositivi, mantenendo attiva solo la connessione con la porta seriale scelta in precedenza.

Vengono poi settati alcuni parametri della porta seriale, come il baud rate (9600 bps) e un buffer size di 512 bytes. Lo script inizializza gli array di gain e phase a 0 e scrive sulla porta seriale il carattere ASCII per “M”, in modo che il sensore cominci la misura, per poi leggere dalla seriale il timestap e i valori di gain e phase scritti dal sensore.

Dopo di ciò lo script “Terra.m” elabora i valori di guadagno e fase letti da “TerraSerialRead.m”, applicando le seguenti formule:

$$V_{gain} = \frac{gainSamples \cdot V_{ref}}{N_{levels}}$$

$$V_{phase} = \frac{phaseSamples \cdot V_{ref}}{N_{levels}}$$

$$gain = \frac{V_{gain}}{gainSensitivity} - 30$$

$$phase = -\frac{V_{phase}}{phaseSensitivity} + 180$$

I valori così calcolati sono infine usati da “TerraGUI.m” per creare i due grafici presenti nella parte centrale dell’interfaccia, che riportano i valori di guadagno e fase rispetto alle frequenze usate in misurazione. Quando la misura è finita il bottone START torna riutilizzabile e la GUI ritorna nella sua configurazione iniziale.

Il secondo pulsante presente nella “TEST MODE” viene utilizzato per salvare le misure in un file txt. I dati vengono salvati su tre colonne: la prima è per le frequenze, la seconda per il guadagno e la terza per la fase.

3.2 Logger Mode

Il pannello della LOGGER Mode comprende al suo interno due sezioni. La prima è chiamata “Absolute Time Settings” che permette, grazie al pulsante “SET TIME”, di settare il tempo assoluto del sensore. E’ stata implementata questa funzione poiché quando il sensore viene spento o quando la batteria si scarica si ha un reset del microcontrollore al tempo di inizializzazione.

Il secondo pannello, “Measure Period Settings” è stato creato per programmare il sensore in modo che esegua misure autonomamente, con un periodo settato dall’utente . Il pannello tre box di testo in cui è possibile inserire il valore di giorni,ore e minuti. Una volta inseriti è possibile programmare il sensore grazie al pulsante “SET PERIOD”.

Capitolo 5

Materiali e metodi (sviluppo del modello di calibrazione)

In questo capitolo vengono presentati i procedimenti messi in atto per lo sviluppo del modello di calibrazione del sensore, partendo dalla raccolta sperimentale dei dati di umidità e procedendo con l'analisi multivariata PLS.

1. Raccolta dei dati spettrali

Il primo step necessario per la creazione del modello di calibrazione è stato quello della raccolta di dati, ottenute eseguendo misurazioni con il sensore sul terreno. Sono state scelte cinque zone di test (nella sede della facoltà di agraria) in base all'assenza in superficie di erba, sassi o detriti organici che potessero falsare la misurazione. Nell'arco di due mesi (ottobre-novembre) sono state eseguite un totale di 315 misurazioni: per ogni punto scelto del terreno sono state fatte tre misurazioni, ognuna eseguita con il sensore ruotato con un' angolazione diversa, in modo da avere un numero maggiore di dati e rendere indipendenti i dati dalla posizione del sensore rispetto al suo asse centrale.

Fig 5.1 – Misurazione con sensore sul terreno

Dopo ogni tre misurazioni è stato eseguito un carotaggio del terreno, ottenendo un campione cilindrico di circa 10 cm di lunghezza. Questo è stato poi diviso a metà, in modo da ottenere due campioni, da cui verranno in seguito ricavati i corrispettivi valori di umidità per avere una stima sia per la zona superficiale che per quella sottostante. In seguito è stata eseguita anche una media di questi due valori.

Fig 5.2 – Carotaggio del terreno

Le misure dell'umidità sono state eseguite con il metodo considerato il gold standard per questo procedimento, ovvero il metodo termo-gravitometrico descritto nel capitolo 1. E' stata quindi eseguita una misura della massa dei campioni e della tara, ponendo successivamente i campioni in una stufa a 105 °C per circa 24 ore. Il giorno successivo è stata misurata nuovamente la massa dei campioni, e calcolata infine l'umidità percentuale.

In conclusione i dati usati per la creazione del modello di calibrazione sono composti da 345 misurazioni, per ognuna delle quali si hanno 3700 valori di guadagno e di fase rilevati dal sensore; e 115 valori di umidità superiore, inferiore e media: ad ogni gruppo di tre misure eseguite sullo stesso campione vengono associati valori di umidità uguali, come se fossero tre misure diverse su un terreno con l'umidità costante.

2. Il software Eigenvector

Per lo sviluppo del modello di calibrazione tramite PLS è stato usato un tool in ambiente Matlab di nome Eigenvector, sviluppato dalla Eigenvector Research Inc.
Grazie a questo software è possibile applicare ai dati un'analisi di regressione PLS in maniera ottimale, grazie a un'interfaccia user-friendly. Dispone inoltre di diversi strumenti e grafici con cui eseguire diagnostiche del modello creato per poterlo migliorare e validare.
Nella figura 5.3 si può osservare l'interfaccia grafica presente in Eigenvector per la PLS, dove è possibile caricare i dati, creare un modello, studiarlo e validarlo.

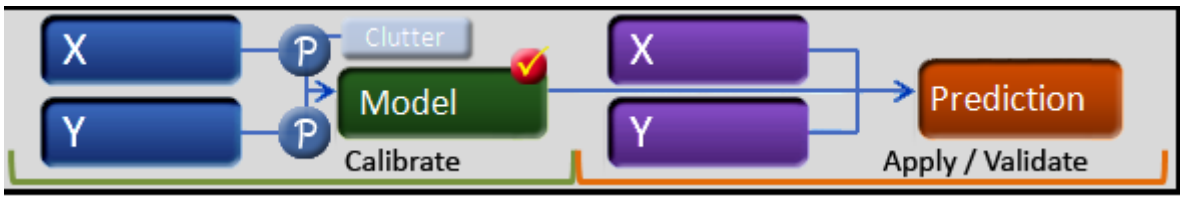

Fig 5.3 – Interfaccia del toolbox PLS

2.1 Dati sperimentali di input

Il primo passo è l'inserimento delle variabili dipendenti X e indipendenti Y. Sono state provate diverse combinazioni di dati, in modo da stabilire quali dessero i risultati migliori: per la Y sono stati testati i valori di umidità superiore, inferiore o media; mentre per le X sono stati state esplorate negli andamenti del gain, della phase, e di due parametri ottenuti come combinazione di questi:

- Modulo:

$$modulo = \sqrt[2]{(gain^2 + phase^2)}$$

- Angolo

$$angolo = arctg\left(\frac{gain}{phase}\right)$$

E’ importante specificare che sia i dati X di guadagno e fase che i dati Y sono stati filtrati attraverso un filtro gaussiano del nono ordine, in modo da eliminare una componente di rumore data dal circuito elettronico del sensore e da errori di misurazione. E’ stata a questo scopo creata una funzione di Matlab dove il filtro viene implementato grazie al comando gausswin(9) e applicato con filtfilt(): in questo modo si raddoppia l'ordine del filtro effettuando un doppio passaggio ed evitando il ritardo introdotto da un filtro classico.
Prima della creazione del modello i dati vengono inoltre preprocessati: per questi modelli è stato scelto di applicare solo l’autoscale, ovvero centrare i dati e ridurre la loro varianza a 1.

2.2 Cross-validation

La cross-validation è un tool fondamentale per l'analisi statistica, applicato durante la creazione del modello, che svolge due funzioni diverse:

1) Permette una valutazione della complessità ottimale del modello, nel caso della PSL quindi del numero necessario di LV

2) Permette una prima stima della validità del modello quando è applicato a dati sconosciuti

Per un data-set di partenza, la cross-validation si basa su una serie di esperimenti, chiamati esperimenti di sub-validazione, ognuno dei quali prevede la rimozione di un sottoinsieme di oggetti dal data-set, la creazione di un modello usando gli oggetti rimanenti nel test set, e una successiva applicazione del modello ottenuto agli oggetti rimossi inizialmente. In questo modo ogni esperimento di sub-validazione implica il testing di un modello con dati che non sono stati usati per crearlo.

Una tipica procedura di cross-validation comprende solitamente più di un esperimento di sub-validazione, ognuno dei quali richiede la selezione di diversi sottoinsiemi per la creazione del modello e il testing.
Esistono diversi metodi di cross-validation, che si differenziano tra di loro in base a in base a come i differenti sottoinsiemi sono creati. [12]
Per la seguente descrizione dei metodi si consideri n come numero totale di oggetti nel data-set e s come il numero di divisioni dei dati indicati nella procedura di cross-validazione, che deve sempre essere minore di n/2.

I principali metodi sono:

- *Venetian Blinds*: Ogni test-set è determinato selezionando ogni s-esimo oggetto nel data set, partendo dall'oggetto 1 fino all's.

- *Contiguous Blocks*: Ogni test-set è determinato selezionando blocchi contigui di n/s blocchi dal data-set, partendo dall'oggetto 1.

- *Random subset:* s differenti test-set sono determinati attraverso la scelta random di n/s oggetti del data-set. Questa procedura è ripetuta r volte, con r = numero di iterazioni. Il risultato finale è dato poi dalla media di queste iterazioni.
- *Leave-One-Out*: Ogni singolo oggetto del data set è usato come test set.

Una volta eseguita la cross-validazione viene calcolato il Root Mean Square Error of Cross-Validation(RMSECV), definito come:

$$RMSECV = \sqrt[2]{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

dove $\hat{y}$ contiene i valori della variabile Y stimati con la cross-validation e y i valori conosciuti di Y.

E' utile poi graficare il valore RMSECV in funzione del numero di variabili latenti utilizzati nel modello: questo grafico è molto utile per determinare il numero ottimale di variabili latenti da usare nel modello.

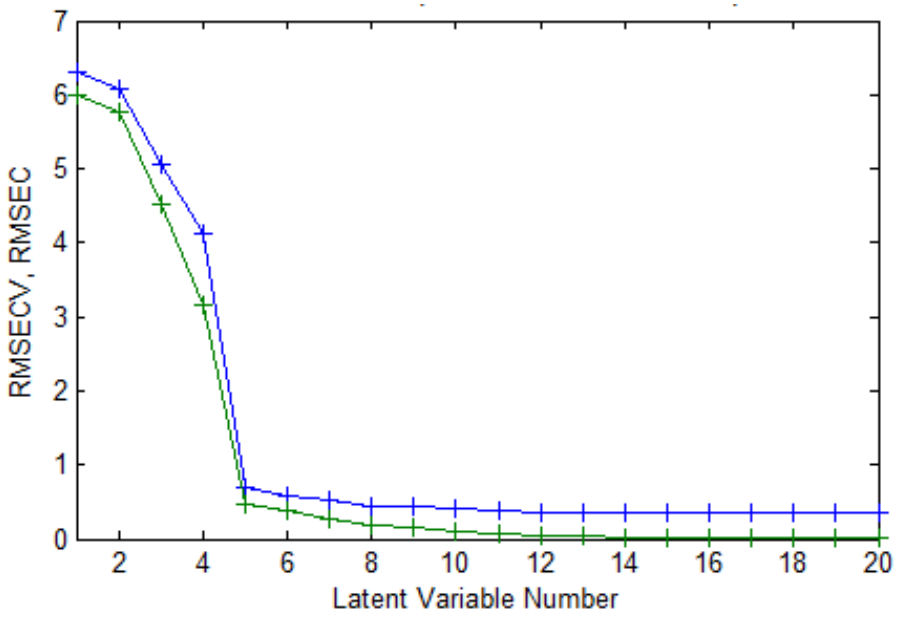

Fig 5.4 – Grafico RMSEC/RMSECV[8]

Il valore di RMSEC deve sempre diminuire al crescere del numero di LV, ma il valore di RMSECV non è detto che abbia lo stesso comportamento, poiché è calcolato da un esperimento in cui i valori di test non sono stati usati per la creazione del modello: è possibile quindi che esso aumenti nel caso in cui troppe LV siano aggiunte al modello. Il numero ottimale di variabili latenti viene tipicamente scelto osservando quando l'aggiunta di un'ulteriore LV non migliora di molto la performance del modello. Nell'esempio in figura 5.4 il corretto numero di LV potrebbe per esempio essere 5 o 8.

Nel caso che le conseguenze dell'overfitting siano particolarmente gravi è possibile eseguire più procedure di cross-validazione usando diversi metodi e paramentri, in modo da poter avere più informazioni con cui prendere una decisione.

2.3 Variabili latenti

Una volta inseriti i dati e creato il modello di partenza è possibile esplorare il risultato ottenuto attraverso vari indici e grafici dati dal modello. . In figura 5.5 sono riportati i parametri da prendere in considerazione per la scelta del numero delle variabili latenti, i valori di R^2 in calibrazione e in cross-validation, la varianza spiegata, e l'andamento dell'RMSECV all'aumentare del numero di LV.

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	38.47	38.47	84.17	84.17	1.3998	
2	49.40	87.86	1.97	86.14	1.3143	
3	6.27	94.13	2.00	88.14	1.2084	
4	0.77	94.90	2.83	90.98	1.0759	
5	0.81	95.71	1.90	92.87	0.97014	
6	0.46	96.18	1.79	94.67	0.85909	current*
7	0.84	97.02	0.65	95.32	0.83568	

Fig 5.5 – Tabella della variabili latenti

Per ogni variabile latente viene indicata la percentuale di varianza contenuta in questa nuova variabile, sia per i dati X che per quelli Y. Da definizione ogni variabile latente deve contenere più varianza di quella successiva, essendo la prima quella con il valore maggiore e così via. Si può anche notare come il sistema rilevi in maniera automatica il numero A di componenti principali necessario affinchè il sistema riesca ad avere una buona capacità predittiva ma non abbia problemi di overfitting

2.4 Outliers

Un'altra fonte di errore nella costruzione di un modello è data dalla presenza di outliers, ovvero un'osservazione distante da tutte le altre, che può essere dovuta alla varianza dei dati, ma anche ad errori sperimentali.

Il software Eigenvector permette di riconoscere e agire sugli outliers grazie ad alcuni grafici, mostrati nella figura 5.6.

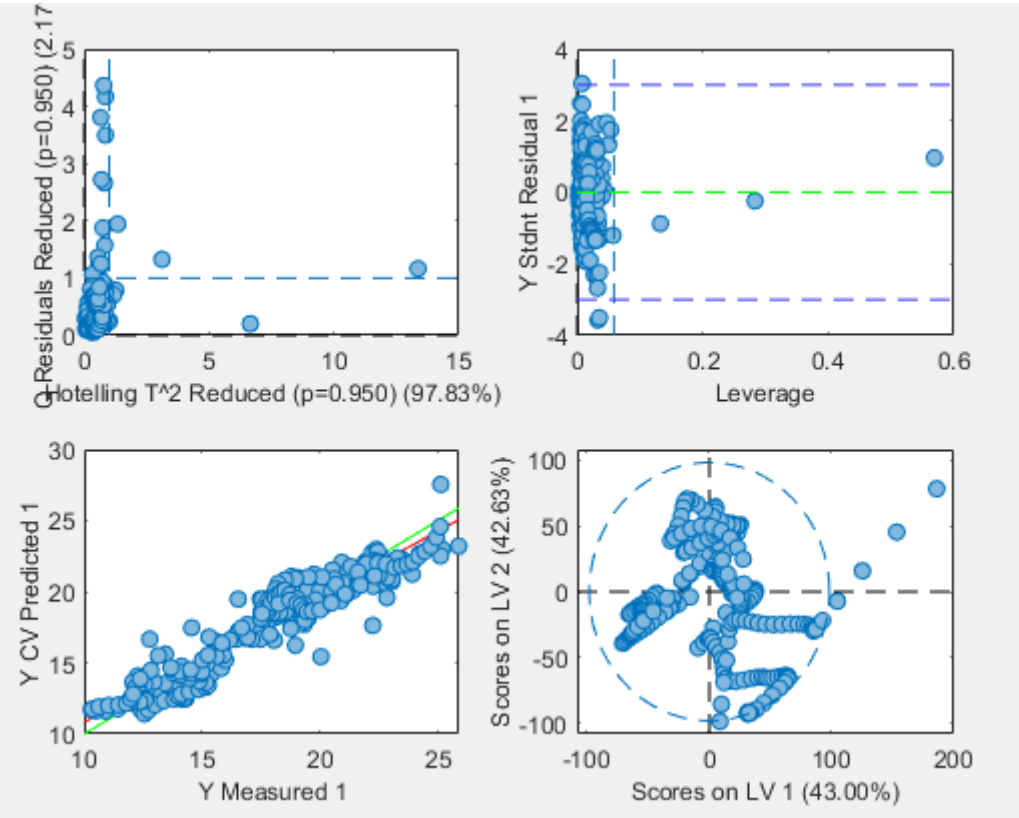

Fig 5.6 – Score plots

Quello di più immediata comprensione è il terzo (in basso a sinistra): lungo l'asse delle ascisse sono riportati i valori di umidità dati in ingresso al sistema (valori osservati), ovvero quelli calcolati con il gold standard; mentre sull'asse delle ordinate sono riportati i valori di umidità stimati dal modello creato per lo stesso campione. Si sarà quindi creato un buon modello quando i due valori saranno uguali o molto simili: si può osservare il comportamento lineare dei dati misurati o predetti grazie alle due

rette (rossa e verde) presenti nel grafico..
Il grafico accanto (in basso a destra) riporta il contributo che fornisce ogni campione alla varianza della prima componente principale (ascisse) e della seconda (ordinate) (scores plot).
Il sofware crea automaticamente un'ellissi contenente la maggior parte dei dati: quelli al di fuori di essa possono essere considerati outliers e investigati.
Il primo grafico (in alto a sinistra) si basa su due parametri statistici, chiamati Q residuals e Hotelling's T-squared, che aiutano a comprendere quanto un sistema descriva bene i campioni. Entrambi sono calcolati a partire dalla formula

$$X = TP^T + E$$

con T=scores, P=loadings e E=residuals.

Q residuals

Per ogni campione, i Q residual sono calcolati come la somma quadratica di ogni riga della matrice E Per esempio, per l'i-esimo campione di X (x_i), avremo:

$$Q_i = e_i e_i^T = x_i(I - P_k P_k^T)x_k^T$$

dove e_i = i-esima riga di E, P_k è la matrice dei k vettori di loadings calcolati dal modello e I è la matrice identità di appropriate proporzioni. Il parametro indica quanto ogni campione è conforme al modello: è una misura della differenza, o residuo, tra un campione e la sua proiezione nei k fattori del modello, permettendoci di capire se il non fitting del campione è dovuto a una deviazione sistematica o a una varianza random. [13]

Hotelling's T Squared

Mentre i Q residuals sono indicatori della magnitudine della variazione rimanente in ogni campione dopo la proiezione attraverso il modello, i valori dell'Hotelling T^2 rappresentano una misura della variazione all'interno del modello: indicano quanto ogni campione è distante dal centro del modello.
Questo parametro viene calcolato tramite la formula:

$$T_i^2 = t_i \lambda^{-1} t_i^T = x_i P_k \lambda^{-1} P_k^T x_i^T$$

dove t_i = i-esima riga di T_k, la matrice $m \times k$ degli scores, λ = matrice diagonale contente gli autovettori corrispondenti alle k variabili latenti. [13]
E' possibile calcolare questo parametro per ogni variabile, in modo da capire come contribuisce al valore di Hotelling's T^2 per ogni campione.
In questo caso prende il nome di T^2 contributions, e viene calcolato come:

$$t_{con,i} = t_i \lambda^{-1/2} P_k^T = x_i P_k \lambda^{-1/2} P_k^T$$

Questo termine può essere considerato una versione scalata dei dati del modello, in modo da rendere uguale la varianza catturata da ogni fattore. E' possibile ricavare facilmente T_i^2 da $t_{con,i}$, grazie alla formula:

$$T_i^2 = t_{con,i} t_{con,i}^T$$

Campioni con valori molto alti di Q residuals e Hotelling's T squared possono dare problemi alla corretta creazione del modello, e grazie al grafico corrispondente possono essere facilmente individuati ed eventualmente esclusi dai valori X.
La soglia di accettabilità per entrambi i parametri è posta a 0.950.

L’ultimo grafico (in alto a destra) infine presenta sull’asse delle ascisse il valore di leverage e sull’asse delle ordinate quello del residuo standard di Y.
Il parametro leverage indica quanto ogni osservazione contribuisce alla predizione del modello. I punti con alto leverage sono quindi quei campioni con valori delle variabili indipendenti molto alti, facendo in modo che il modello di regressione fittato passi molto vicino a questi punti.
E’ possibile calcolare il valore h di leverage grazie alla formula:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Campioni con valori alti di leverage vengono considerati estremi, e possono essere eliminati dal modello. La soglia viene posta a $\frac{2m}{n}$, dove m è il numero di componenti e n il numero di osservazioni.
Il residuo Y standard misura invece la differenza tra il valore di umidità misurato e quello predetto: i campioni con un alto valore in modulo di questo residuo sono gli stessi che venivano identificati come outliers dal primo grafico descritto. Per questo parametro la soglia di accettabilità è posta a ± 3.

2.5 Selezione delle variabili

Dopo aver esaminato ed eliminato gli outliers, è possibile sfruttare alcuni algoritmi presenti nel software Eigenvector che eseguono una selezione delle variabili di maggiore interesse per la costruzione del modello, in questo caso delle frequenze a cui sono state eseguite le misure dal sensore. Questo porta a un miglioramento generale del modello, in particolare per quanto riguarda la varianza dei dati spiegata dalle prime LV, aumentandone il valore. Sono presenti diversi algoritmi da poter

utilizzare per svolgere questa funzione: per la creazione del modello è stato sempre utilizzato l'automatico, implementato in Matlab grazie alla funzione selectvars (X,Y, maxlv). E' stata fatta questa scelta in quanto il modello ha molti campioni e forniva buoni risultati anche prima di applicare la selezione delle variabili, rendendo quindi superfluo applicare algoritmi più potenti ma lenti come il Genetic Algorithm o altri.

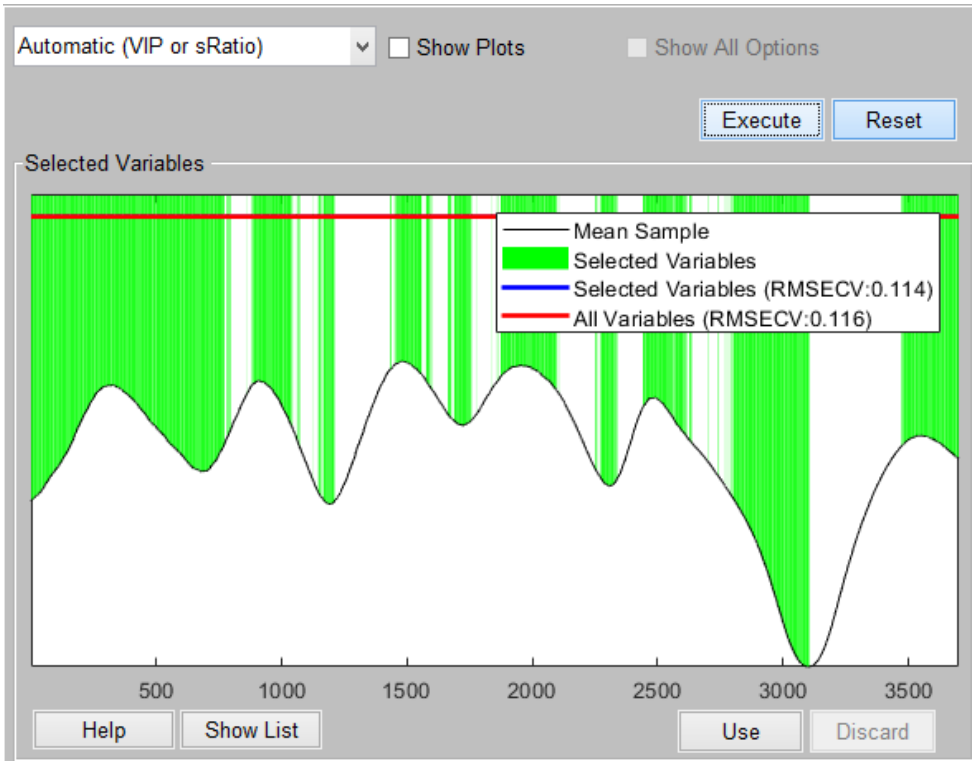

Fig 5.7 – Selezione delle variabili

La funzione selectvars richiede in ingresso i dati X e Y e il numero massimo di variabili latenti da utilizzare (maxlv). L'algoritmo cerca il miglior sottoinsieme di variabili usando i parametri VIP (Variable Importance in Projection) e SR (Selectivity Ratio) e presenta come risultato la selezione con il migliore valore di RMSECV. Come prima cosa le variabili con i valori percentuali minori di VIP e SR sono eliminati. Se il modello migliora questo processo è ripetuto iterativamente fino a che il modello smette di migliorare.

Per alcuni tipi di dati è più adatto rimuovere una grande parte di dati per ogni iterazione, per altri è meglio rimuoverne una frazione minore. Per valutare quale sia la frazione più adatta si eseguono rimozioni di diverse percentuali di dati: quelle di default sono [2 5 8 10 15 20 25 30 35 40 45]/100. Il processo iterativo è ripetuto per ognuna di esse, e solo il risultato con il miglior RMSECV è utilizzato. Per quanto riguarda il numero di iterazioni l'algoritmo ha come valore massimo di default 20. [14]

2.6 **Influenza della temperatura**

Uno dei parametri ambientali da tenere maggiormente in considerazione durante lo sviluppo del modello e l'utilizzo del sensore è sicuramente la temperatura, sia del suolo che dell'aria. La temperatura del suolo influenza la permittività dielettrica dell'acqua, come già spiegato nel capitolo 2, e quindi le misurazioni dell'impedenza; mentre la temperatura dell'aria può agire sul sistema elettronico, inficiandone la precisione. La figura 5.8, ottenuta graficando insieme misurazione fatte sul terreno a diverse temperature, mostra bene questo fenomeno

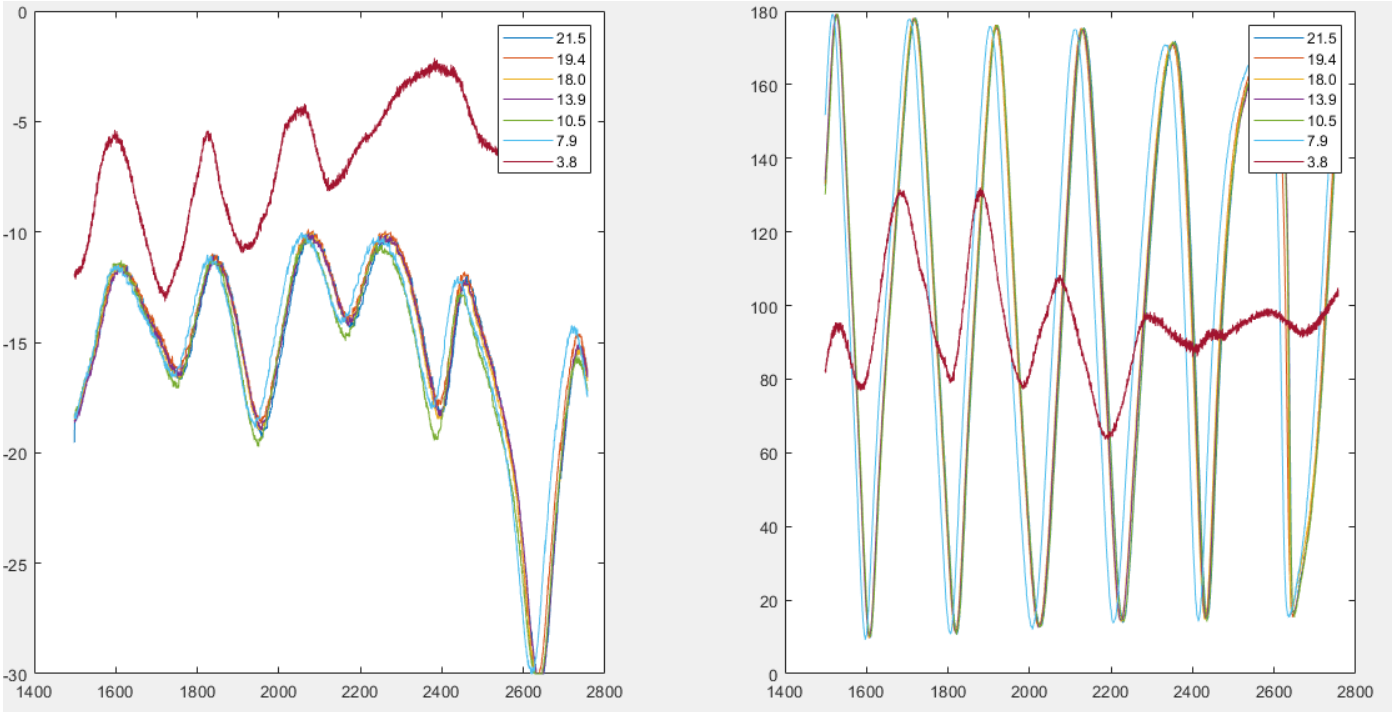

Fig 5.8– Grafici di misurazioni a diversa temperatura

Come è possibile osservare nel range 21.5 – 7.9 °C le misurazioni hanno un andamento molto simile, andamento che subisce sia un offset che una modifica dell'ampiezza (quest'ultimo solo nel guadagno). Per temperature più vicine allo zero invece l'andamento è profondamente diverso, sia come modulo che come forma: è stato quindi deciso di non inserire tra i dati X le misurazioni eseguite a temperature molto basse. Per eseguire misure accurate a queste temperature sarà necessaria la creazione di un modello ad hoc, raccogliendo molte più misure con queste caratteristiche termiche.

Una volta eliminate le misure eseguite con un clima più freddo il modello fornisce risultati molto buoni anche senza l'applicazione di preprocessing mirati per attenuare questa influenza, come potrebbe essere un Multiplicative Scatter Plot. E' possibile intuire il motivo di questo comportamento osservando uno degli score plot descritti in precedenza, dove è riportata lungo gli assi l'influenza di ogni campione sulla varianza spiegata dalle prime 2 componenti principali.

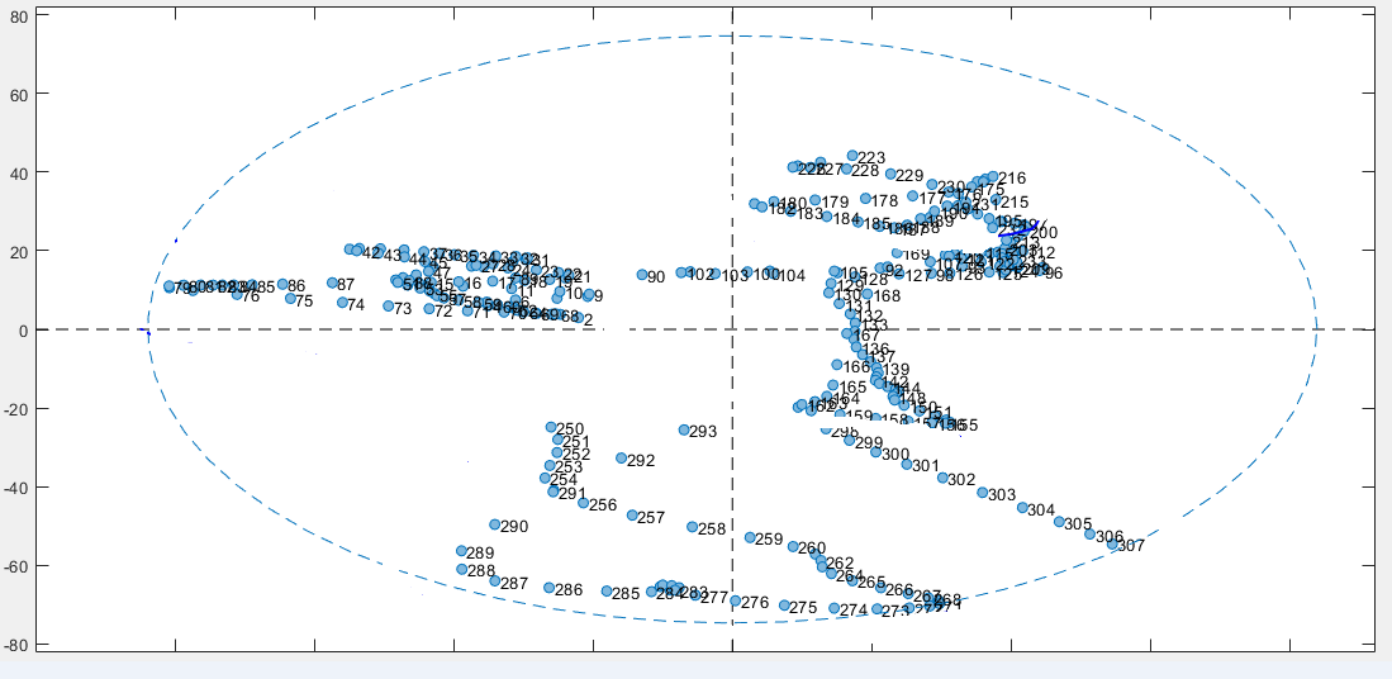

Fig 5.9–Score plot con numero dei campioni

Tramite questo grafico è possibile capire quali sono le prime variabili latenti, andando a ricercare andamenti nei dati lungo i due assi. Come è facile immaginare, nello score plot è possibile trovare un andamento lungo la prima componente principale riguardante l'umidità: all'estrema sinistra sono presenti i primi campioni raccolti, per i quali il valore di umidità è minore (circa 13%), mentre all'estrema destra abbiamo campioni con la maggiore percentuale di umidità (circa il 23%). Questo indica quindi che la prima variabile latente, ovvero quella che spiega la quantità maggiore di varianza nei dati, è legata all'informazione di umidità.

Per quanto riguarda la seconda LV, è possibile notare che nella parte inferiore del grafico sono presenti i campioni 250-307, raccolti alle temperature minori(circa 8 °C), mentre nella parte superiore i primi 90 campioni e i numeri 180-230, raccolti con una temperatura di circa 18 °C. La seconda variabile latente sembra quindi spiegare la variabilità spettrale dovuta alla variazione di temperatura: il modello tiene conto della varianza conseguente alla differenza di temperatura, senza considerare le variazioni dovute ad essa come un diverso valore di umidità

2.7 Validazione

Dopo aver creato un modello soddisfacente, è necessario eseguire una validazione, ovvero testare il modello con alcuni dati non usati nella sua costruzione. Questo viene eseguito caricando nell'interfaccia grafica un set di dati X e Y, circa il 20% dei dati totali, rappresentativi di tutto il range di umidità presente nei dati originari del modello. Il software userà poi il modello creato per stimare il valore di umidità partendo dai dati X forniti in validazione, e confronterà il risultato ottenuto con il valore delle Y fornite. Le misurazioni da fornire in validazione sono sempre state scelte in blocchi da tre, in modo da fornire tutte le misure per lo stesso campione, ed evitare di averne alcune in calibrazione ed alcune in validazione, falsando quindi i risultati.

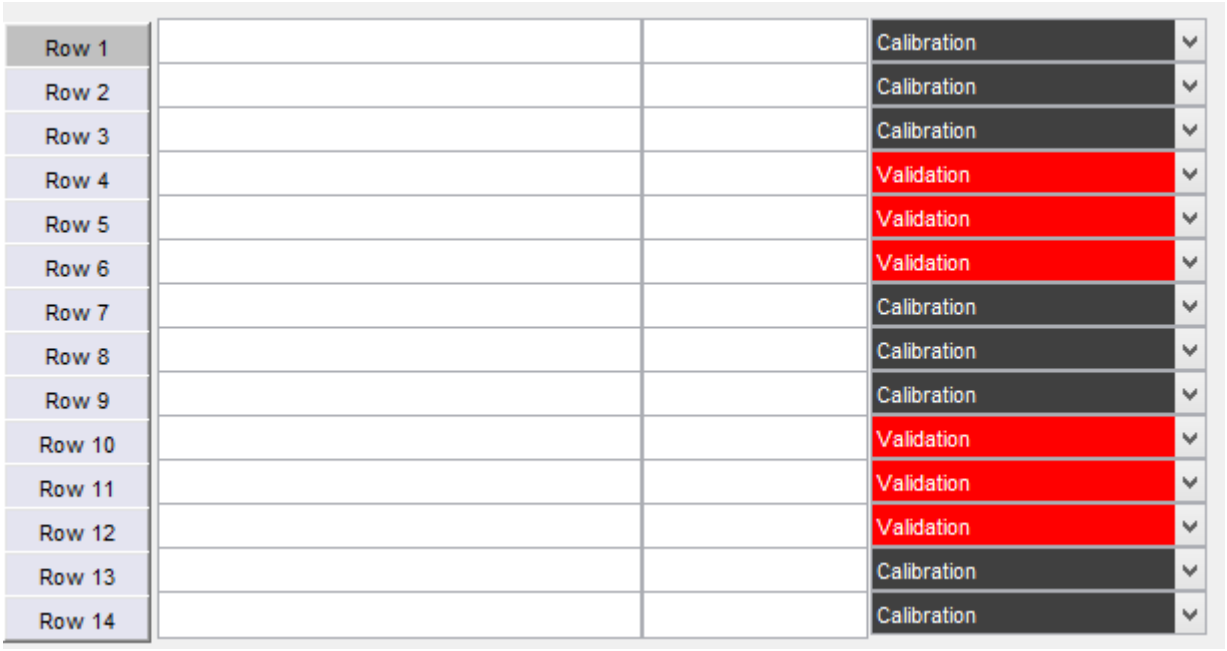

Fig 5.10 – Interfaccia di scelta calibrazione/validazione

Una volta eseguita la validazione è possibile ottenere i parametri RMSEP (Root Mean Square Error in Prediction) e R^2 in stima, che danno informazioni sulla bontà dei valori di umidità calcolati tramite al modello, insieme ai grafici degli score plot relativi ai dati inseriti in validazione.

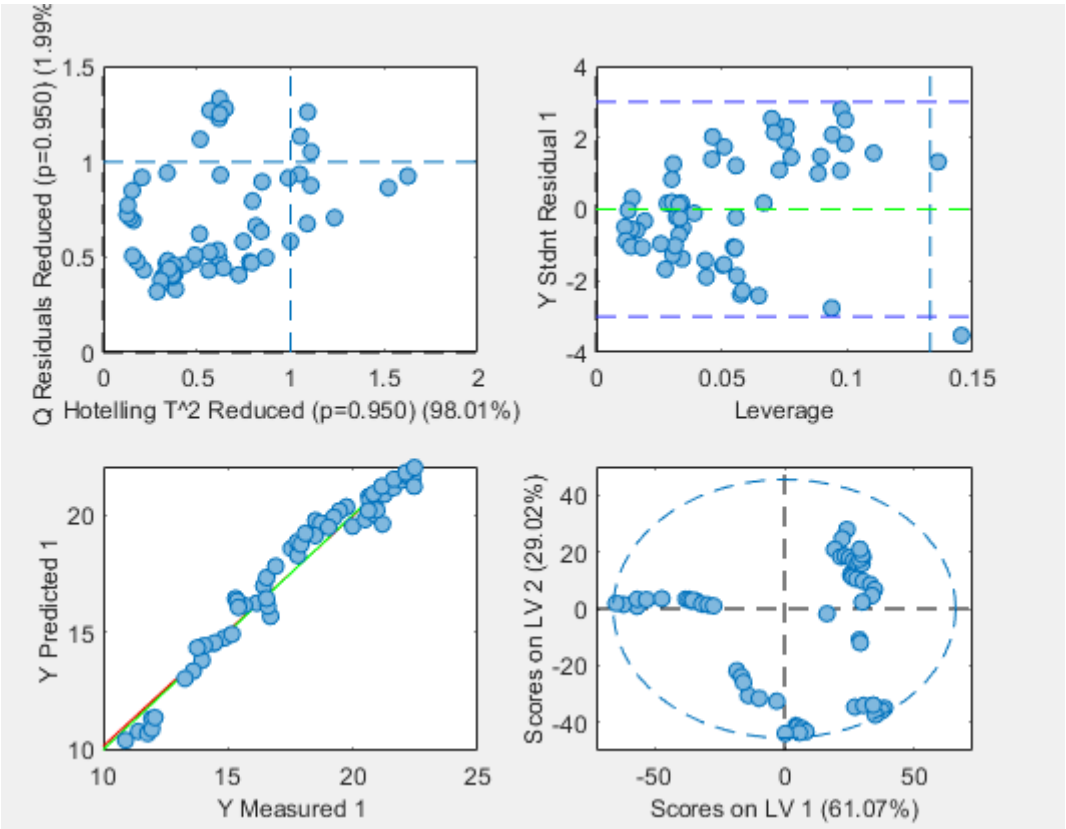

Fig 5.11– Score plot dei dati in validazione

3. Scelta del modello

Tutti i procedimenti descritti sono stati applicati ai modelli creati usando i diversi dati X citati all'inizio del capitolo, in modo da trovare il modello migliore da implementare nel sensore.

Per quanto riguarda i dati Y è stato deciso in partenza di usare i valori medi di umidità, in quanto maggiormente rappresentativi per questo parametro.

Il modello costruito con i valori di phase è stato scartato come primo in quanto, nonostante l'eliminazione degli outliers e la selezione delle variabili, la percentuale di varianza spiegata dalla prima LV è molto minore di quella legata alla seconda LV, andando quindi contro la definizione stessa di variabile latente.

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	24.13	24.13	82.16	82.16	1.482	
2	68.17	92.30	0.89	83.05	1.4442	
3	3.68	95.99	2.96	86.01	1.3184	
4	1.50	97.48	2.71	88.72	1.1911	
5	0.77	98.26	2.62	91.34	1.056	current*

Fig 5.12 – LV del modello con fase

Anche il modello calcolato con il modulo è stato scartato, a causa del basso numero di LV utilizzate, solo 2: è infatti necessario un buon compromesso di LV nella costruzione del modello. Se il numero è troppo alto si ha il rischio di over-fitting, se troppo basso non si avranno delle misurazioni corrette. Il modello del modulo aveva infatti i valori di R^2 e R^2 in validazione più bassi, rispettivamente 0.851 e 0.849.

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	63.96	63.96	71.07	71.07	1.806	
2	22.26	86.23	14.00	85.07	1.3028	current*
3	5.09	91.32	1.11	86.18	1.26	
4	2.73	94.04	1.33	87.51	1.2083	

Fig 5.13 – LV del modello con modulo

Questo stesso problema è stato anche rilevato nel modello con l'angolo che, sebbene desse risultati migliori, aveva solo 3 LV. Inoltre l'utilizzo dell'algoritmo di selezione delle variabili, necessario per avere corretti valori di varianza percentuale in ogni LV, selezionava un numero troppo basso di frequenze, 435.

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	70.98	70.98	84.67	84.67	1.3078	
2	17.71	88.68	2.08	86.75	1.2194	
3	4.59	93.27	2.28	89.03	1.1128	current*
4	4.50	97.78	0.72	89.75	1.0782	
5	0.38	98.16	1.97	91.73	0.99404	

Fig 5.14 – LV del modello con angolo

E' stato quindi deciso di utilizzare il modello creato a partire dai valori di gain. Dopo aver eseguito l'eliminazione degli outliers e la selezione delle variabili, il modello finale è stato calcolato a partire da 203 misurazioni, di cui vengono considerate 1158 frequenze.

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	61.07	61.07	85.26	85.26	1.3129	
2	29.02	90.09	2.60	87.86	1.1999	
3	2.04	92.13	1.87	89.73	1.1308	
4	2.66	94.79	1.36	91.09	1.0466	
5	1.08	95.87	2.57	93.66	0.90971	
6	0.77	96.64	1.46	95.12	0.8156	
7	0.77	97.41	1.18	96.30	0.72689	
8	0.35	97.77	1.33	97.64	0.6171	
9	0.24	98.01	0.77	98.40	0.52635	current*

Fig 5.15 – LV del modello con guadagno

Come è possibile osservare dalla figura 5.15 la varianza spiegata dalle 9 LV segue un andamento corretto.

E' possibile capire la scelta del numero di LV osservando il grafico di RMSEC e RMSECV in funzione del numero delle variabili, presentato in figura 5.16

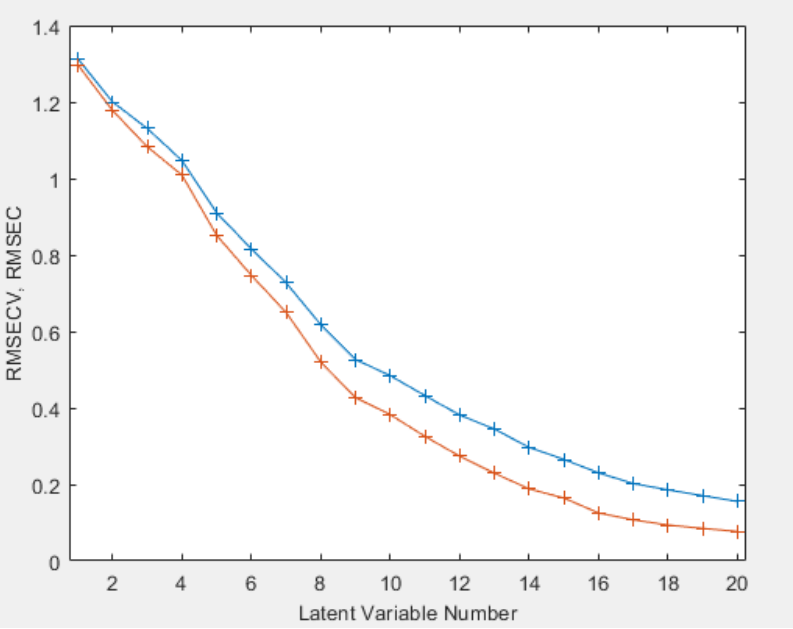

Fig 5.16 – RMSEC e RMSECV per il guadagno

9 LV è il valore da cui la diminuzione dell'errore quadratico comincia a diminuire, ed è perciò una buona scelta per quanto riguarda il numero di variabili latenti.

Il modello presenta i valori più alti di R^2 e R^2 in validazione tra tutti i modelli proposti, rispettivamente 0.984 e 0.976; mentre dagli score plot è si nota come sia il modello con il comportamento lineare più spiccato.

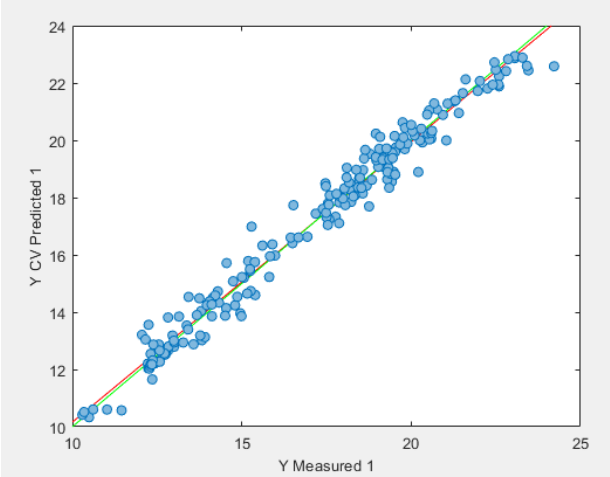

Fig 5.17 – grafico Y misurate/predette per il guadagno

Anche la validazione dà ottimi risultati, con un R^2 di predizione di 0.958 e un comportamento fortemente lineare presente anche nei campioni predetti per la validazione.

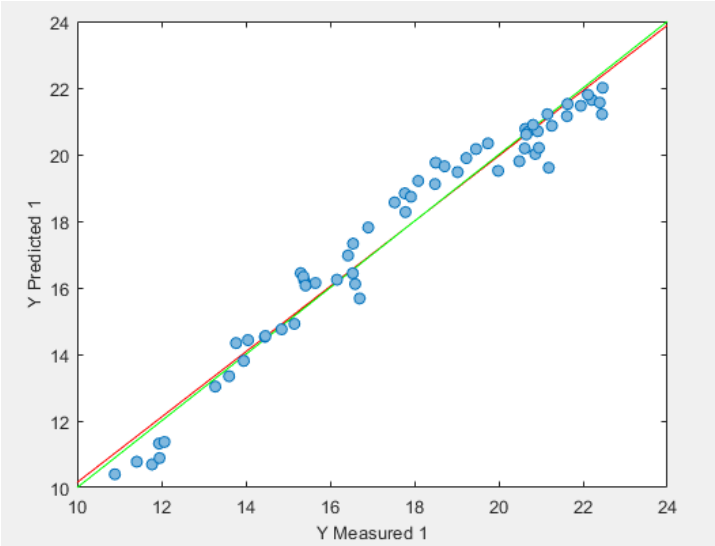

Fig 5.18 – grafico Y misurate/predette per il guadagno in validazione

3.1 La funzione regcon

L'ultimo passo prima dell'implementazione del modello nel MCU del sensore è l'utilizzo della funzione regcon, presente in Matlab. Questa funzione converte un modello generico creato tramite PLS o altri algoritmi in una forma espressa dall'equazione lineare y=ax+b.
La funzione si presenta nella forma

[a,b] = regcon(mod)

dove l'input mod indica un modello creato grazie al software Eigenvector e salvato come file mat, mentre gli output a e b indicano rispettivamente i coefficienti di regressione e l'intercetta, entrambi necessari per il calcolo del modello.
La funzione fornisce in uscita il vettore a di lunghezza 3700, in quanto riempie in maniera automatica i valori corrispondenti alle frequenze non usate dal modello con uno 0.
Salvando quindi questi valori all'interno della memoria del MCU è possibile calcolare l'umidità a partire dai valori di guadagno misurati dal sensore, grazie alla formula y=ax+b.

4. Modifiche del firmware

Le modifiche apportate al firmware hanno riguardato principalmente la funzione **MeasureTask()**, presente nel file "mcc.c", dove avviene il salvataggio dei valori finali di guadagno e fase e l'invio di questi alla porta seriale.
Nella parte iniziale del file sono state create nuove variabili, necessarie alla pre-elaborazione dei dati e al calcolo del valore di umidità.

```
unsigned int arrayindex;
int V_DAC;
unsigned int i;
float Vgain;
float gain;
float gainreg;
float moisture;
float moisture_b;
int moistureint;
const float Vref=1.8;
const float Nlevels=4095;
const float gainsensitivity=0.03;
const float b=197.285939536;
const float reg[3700]={0.287963754,0.204038267,-0.327234297,0
                       -0.219562168,0.184459623,0.046493377,-
                       0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
                       -0.35262728,0.056679571,0.318742654,0.
```

Fig 5.19 – variabili inizializzate in mcc.c

Tutte le variabili sono state create come float, in maniera da ottenere una maggiore precisione di calcolo, dopo essersi accertati che l'utilizzo di questa tipologia di dato non aumentasse eccessivamente il tempo di misurazione. Solo la variabile *moistureint* è stata creata come int, in modo da scrivere un dato intero su seriale, semplificando sia la fase di scrittura che quella di lettura in Matlab.

Le nuove variabili introdotte comprendono le costanti *Vref*, *Nlevels* e *gainsensitivity*, usate per eseguire l'elaborazione del dato di guadagno descritta nel capitolo 4, eseguita precedentemente su Matlab.

E' stato deciso di implementare questa elaborazione direttamente all'interno del microcontrollore poiché il modello è stato creato con i dati in uscita dallo script di Matlab: è quindi necessario applicare i coefficienti ai valori già elaborati. In questo modo inoltre si ha in uscita dal sensore il valore finale di umidità percentuale, senza dover eseguire ulteriori calcoli su pc.

Le variabili *reg* e *b* corrispondono ai dati a e b calcolati grazie alla funzione regcon. Come si può osservare nella figura 5.19, *reg* è un array di 3700 elementi, di cui tutti quelli corrispondenti a frequenze non utilizzate per la creazione del modello sono state riempite con 0; mentre b è un singolo valore.

Le variabili *gain*, *gainreg*,*moisture*, *moisture_b* e *moistureint* servono infine a eseguire i vari calcoli per il calcolo finale dell'umidità.

E' importante fare un osservazione su come sono state definite queste variabili: sfruttando il comando const per le variabili che non devono essere modificate è possibile sfruttare in maniera più efficace le memorie presenti nel PIC24.

Il MCU comprende infatti due memorie: la Program memory, in cui viene salvato il codice eseguibile, e la Data memory, contenente le variabili esterne, il system stack e i file register. Sebbene le due memorie siano in regioni distintamente separate, la famiglia di processori PIC24 contiene un supporto hardware comunemente chiamato Program Space Visibilityb(PSV), che permette il mappaggio di pagine di 32 Kb di program memory nella data memory. Il compilatore supporta di default l'accesso a una sola pagina, più che sufficiente per contenere tutti i valori costanti usati nel codice. [15]

In questo modo non viene ulteriormente appesantita la data memory, molto più piccola della program memory: quest'ultima è passata da un riempimento del 7% a uno del 16%, come possiamo vedere dalla figura 5.20:

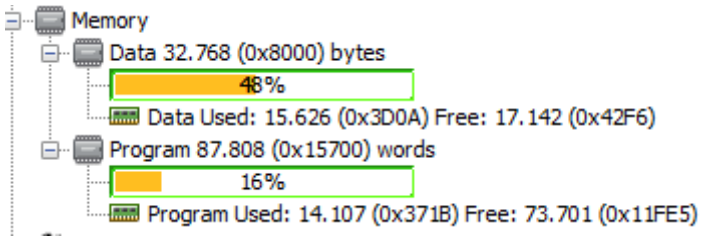

Fig 5.20 – memorie del PIC24

E' stato inoltre aggiunto un campo alla DataStruct già presente, in maniera da poter salvare all'interno di essa anche il valore finale di umidità.

```
typedef struct {
    unsigned int GainData[GainPhaseSamples];
    unsigned int PhaseData[GainPhaseSamples];
    int MoistureData;
```

Fig 5.21 – creazione della DataStruct

La parte principale del calcolo del valore di umidità è stata inserita all'interno del ciclo for presente nella funzione **MeasureTask(),** successivamente al salvataggio nella DataStruct dei valori ch0valuef (gain) e ch1valuef (phase), calcolati nella funzione **TMR3_Callback()**.

```
dataStruct.RData.GainData[arrayindex] = ch0valuef;
dataStruct.RData.PhaseData[arrayindex] = ch1valuef;
Vgain=(ch0valuef*Vref)/Nlevels;
gain=(Vgain/gainsensitivity)-30;
gainreg=gain*reg[arrayindex];
moisture=moisture+gainreg;
```

Fig 5.22 – codice all'interno del ciclo for

Viene come prima cosa applicato il preprocessing ad ogni dato del guadagno che viene misurato, salvando il risultato nella variabile *gain*. Questa viene successivamente moltiplicata per il coefficiente della rispettiva frequenza salvato in *reg*, per poi sommare i risultati tra di loro. In questo modo, alla fine del ciclo for, all'interno della variabile moisture sarà contenuto il risultato dell'operazione $a \cdot x$, con a il vettore reg e x i dati misurati di guadagno.

```
moisture_b=moisture+b;
moistureint=moisture_b;
dataStruct.RData.MoistureData= moistureint;

if (MODE == 0) { //TESTMODE
    // Write last samples to serial
    UART1_writeSample(dataStruct.RData.GainData[arrayindex - 1]);
    UART1_writeSample(dataStruct.RData.PhaseData[arrayindex - 1]);
    UART1_writeIntSample(dataStruct.RData.MoistureData);

}
```

Fig 5.23 – codice successivo al for

Dopo il ciclo for viene sommata a moisture la variabile b, in modo da completare il calcolo dell'equazione lineare y=ax+b. Il risultato viene poi salvato nella variabile moistureint, trasformandolo in un intero, e salvato nella DataStruct. Infine viene scritto su seriale successivamente ai due ultimi valori di gain e phase, grazie alla funzione UART1_writeIntSample. Questa funzione permette di inviare un numero intero alla seriale, dopo averlo trasformato in char.

5. Modifiche del codice Matlab

Le modifiche apportate al codice Matlab sono state poche, essendo il valore di umidità in uscita dal sensore già quello da poter mostrare nell'interfaccia grafica.
La funzione "TerraSerialRead.m" è stata modificata aggiungendo un comando con il quale leggere da seriale il valore di umidità.

```
% Receive gain samples
for i=1:Nsamples
    gainSamples(i,1)=fscanf(s,'%u');
    phaseSamples(i,1)=fscanf(s,'%u');


    disp(i)
end

%Receive moisture sample
moisture=fscanf(s,'%d');
```

Fig 5.24 – codice per la lettura da seriale

Come è possibile osservare dall'immagine 5.24, il comando è stato inserito successivamente alla lettura dei valori di gain e phase, in modo da rispettare l'ordine di scrittura su seriale del microchip. Per la lettura viene usato il comando fscaf, essendo i dati inviati dal sensore come char, con in ingresso il parametro s, in cui è stato salvato il numero della porta seriale appropriata.
Infine è stato aggiunto un riquadro all'interfaccia grafica, in cui viene mostrato il valore di umidità al completamento della misura.

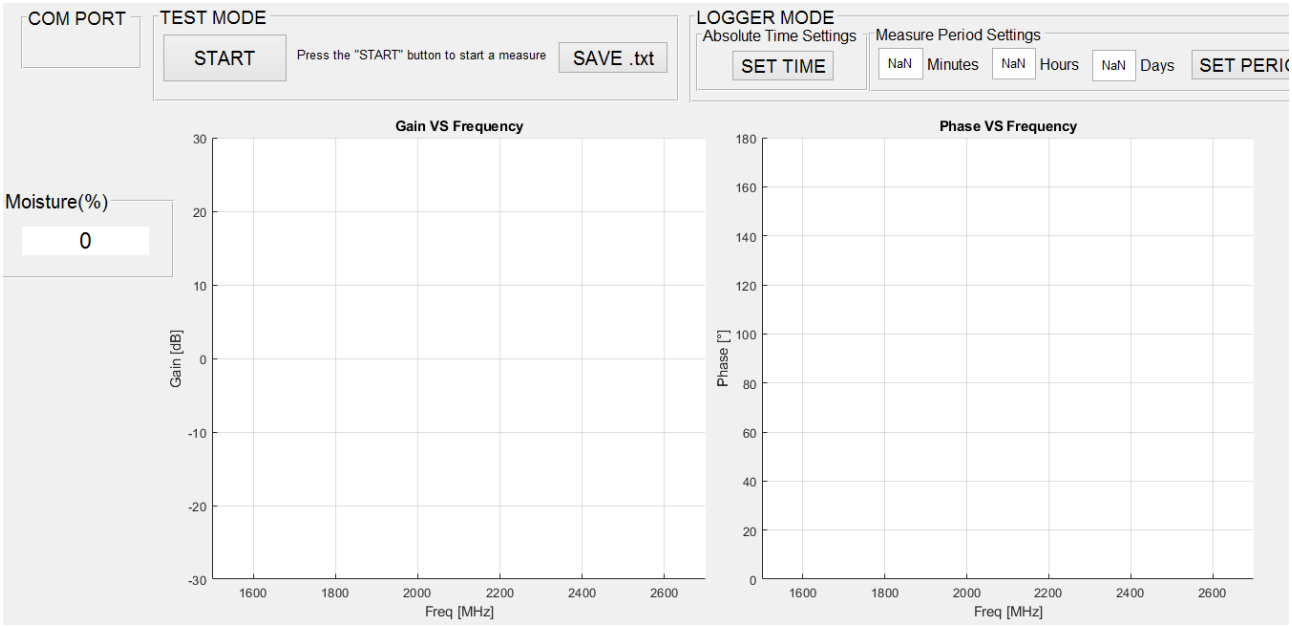

Fig 5.25 – GUI con pannello per l'umidità

Conclusioni

L'obbiettivo della tesi è stato quello di elaborare un modello di calibrazione per il sensore di umidità e di implementarlo all'interno del firmware, in modo da riuscire a ottenere in uscita dal sistema la percentuale di umidità del terreno.

Il sensore era stato testato e validato solamente con campioni creati appositamente in laboratorio. Per la creazione del modello si è partiti dunque dalla raccolta di dati sperimentali sul campo, in modo da creare un modello di calibrazione adatto all'utilizzo esterno del sensore. Si è quindi misurato il livello di umidità dei campioni grazie alla tecnica gold-standard termo-gravitometrica, ottenendo un totale di 315 misurazioni, usate per la creazione del modello.

Si è quindi sfruttato il software Eigenvector per creare e studiare diversi modelli, usando come dati in ingresso i valori di guadagno, di fase, di modulo e di angolo, insieme all'umidità media dei campioni. Tra questi è stato scelto il modello migliore grazie all'osservazione di numerosi parametri e grafici forniti dal software, come il numero di LV, RMSEC, RMSECV, R^2, ecc..
Il modello con la maggiore capacità predittiva si è rivelato essere quello creato con i dati di guadagno, e ad esso è stata applicata la funzione regcon(), in modo da renderlo implementabile in forma lineare.

Una volta ricavati i coefficienti di regressione è stato modificato il firmware del MCU presente nel sensore, in modo da eseguire durante la misura anche il calcolo dell'umidità percentuale. Infine è stato modificato anche lo script di Matlab, in modo da leggere questo ulteriore dato e mostrarlo nell'interfaccia grafica dedicata.

Si è quindi ora in grado di eseguire una misura dell'umidità in maniera veloce, poco invasiva e facilmente ripetibile.

Gli sviluppi futuri legati a questa tecnologia si dividono in due strade principali.
La prima, che sta già venendo investigata, è legata alla ricerca del modo migliore con cui inviare e salvare i dati misurati dal sensore senza doverlo connettere via USB a un pc, permettendo quindi di inserirlo in un sistema di rilevazione più complesso. Riguardo a ciò sta dando buoni risultati l'utilizzo della tecnologia LoRa, consistente in un antenna che permette di inviare i valori di umidità fino a 5 km di distanza dal sensore. Il dato viene inviato a un dispositivo definito gateway, che può essere sia un computer che una pagina di archiviazione online.
La seconda strada è invece legata alla creazione di modelli di calibrazioni che permettano di eseguire misurazioni corrette anche in condizioni diverse da quelle studiate in questa tesi. Le diversità possono essere sia legate al tipo di terreno, che potrebbe dare risposte diverse in caso di forte presenza di argilla o alte condizioni particolari; sia legate alla temperatura del terreno e dell'aria. Specialmente in caso di basse temperature (nell'ordine dei 0-5 °C) il circuito elettronico risponde in maniera diversa: nel caso si volesse usare il sensore a queste temperature sarà quindi necessario raccogliere ulteriori campioni, per poi creare un modello di calibrazione dedicato.

Bibliografia

[1] Pariva Dobriyal e al., «A review of the methods available for estimating soil moisture and its implications for water management», *Journal of Hydrology,* 2012.

[2] Nunzio Romano, «Soil moisture at local scale: Measurements and simulations », *Journal of Hydrology,* 2014.

[3] E.S. Mohamed, «Application of near-infrared reflectance for quantitative assessment of soil properties », *The Egyptian Journal of Remote Sensing and Space Sciences,* 2017.

[4] C. Ruva, *Spettroscopia di impedenza per la misura dell'umidità in applicazioni "green-bio",* 2017.

[5] K. Lomakin e al,, «Transmission Line Model for Rectangular Waveguides accurately incorporating Loss Effects», *IEEE 21st Workshop on Signal and Power Integrity (SPI),* 2017.

[6] A. Baldazzi, *Determinazione non distruttiva della qualità dei kiwi mediante microonde, 2019.*

[7] Svante Wold e al., «PLS-regression: a basic tool of chemometrics», *Chemometrics and Intelligent Laboratory Systems 58,* 2001.

[8] *Introduzione alla Analisi dei Dati Sperimentali,* 2005.

[9] Sijmen tie Jong, «SIMPLS: an alternative approach to partial least squares regression», *Chemometrics and Intelligent Laboratory System*,1993.

[10] M. Siboni, *Sensing dell'umidità del terreno con spettroscopia a microonde*, 2018.

[11] M. Siboni, *Instruction manual of the contactless microwave moisture sensor based on multivariate analysis*, 2018.

[12] Eigenvector Documentation Wiki, *Using Cross – Validation*, 2016.

[13] Eigenvector Documentation Wiki, *T – Squared Q Residuals and contributions*, 2012.

[14] Eigenvector Documentation Wiki, *Selectvars*, 2018.

[15] Microchip, *MPLAB® XC16 C Compiler User's Guide*, 2018

[16] A. Berardinelli, G. Luciani, M. Crescentini, A. Romani, M. Tartagni, and L. Ragni, «Application of non-linear statistical tools to a novel microwave dipole antenna moisture soil sensor, », *Sensors Actuators, A Phys.*, vol. 282, pp. 1–8, 2018.

[17] G. Luciani, A. Berardinelli, M. Crescentini, A. Romani, M. Tartagni, and L. Ragni, «Non-invasive soil moisture sensing based on open-ended waveguide and multivariate analysis», *Sensors Actuators, A Phys.*, vol. 265, pp. 236–245, 2017.