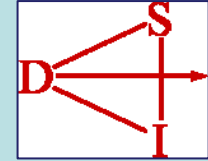




Università degli Studi di Firenze
**Corso di Laurea in Ingegneria per l'Ambiente e il
Territorio**



Tesi di Laurea

**Analisi della stratificazione lacustre
nell'invaso di Bilancino mediante
alberi delle decisioni**

Relatori

Prof. Stefano Marsili-Libelli

Dott. Elisabetta Pezzatini

Candidato

Alice Balducci

A.A. 2008/2009

Sommario

1. **Scopo della tesi**
2. **La stratificazione lacustre**
3. **Analisi dei dati**
4. **Scelta del modello**
5. **Pre-elaborazione dei dati: costruzione indicatori**
6. **L'ambiente Weka**
7. **Costruzione dell'albero delle decisioni**
8. **Uso previsionale dell'albero**
9. **Conclusioni**

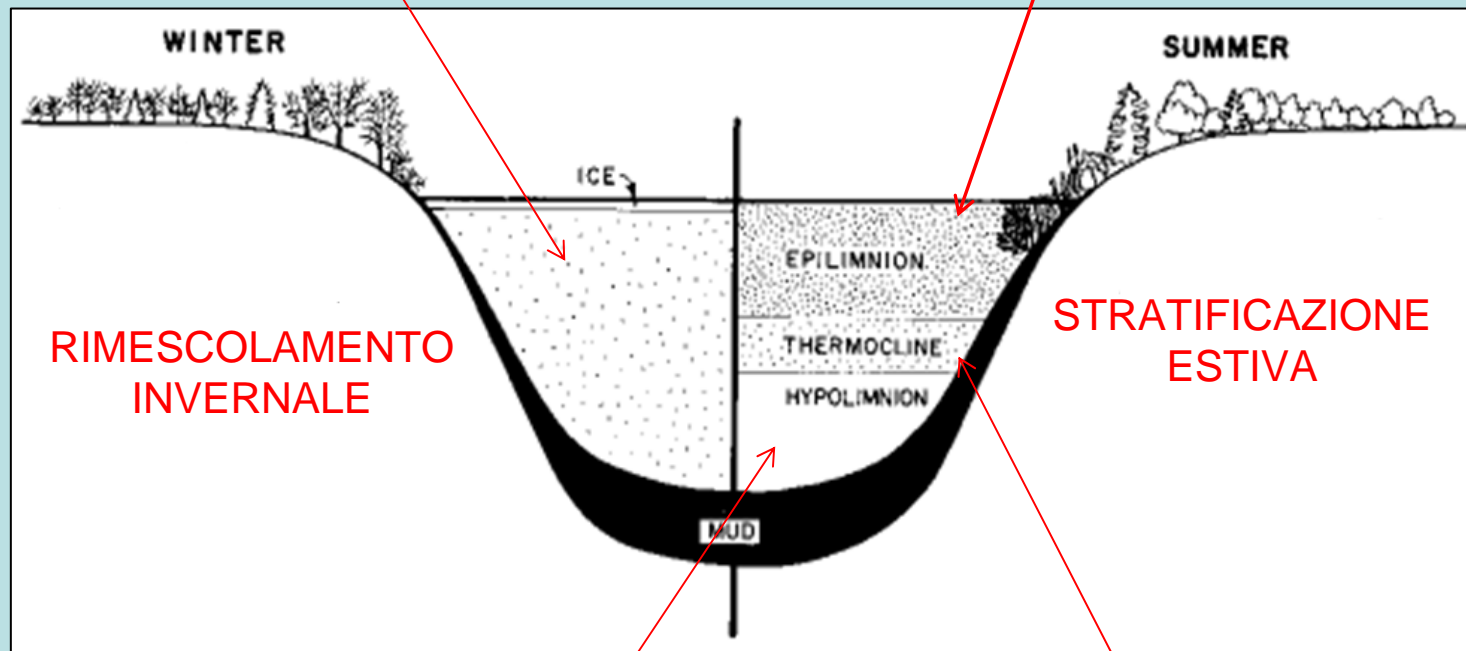
Scopo della tesi

- Costruire un modello previsionale delle condizioni dell'invaso, utilizzando i dati forniti da ARPAT
- Durante il tirocinio si sono analizzati i dati dell'invaso
- La realtà sottostante è apparsa troppo complessa per essere modellata in modo deterministico
- Si è scelto un approccio basato sui dati

La stratificazione lacustre

Completo rimescolamento verticale
Temperatura uniforme

Epilimnio: zona superficiale più calda, soggetta ad intensa circolazione

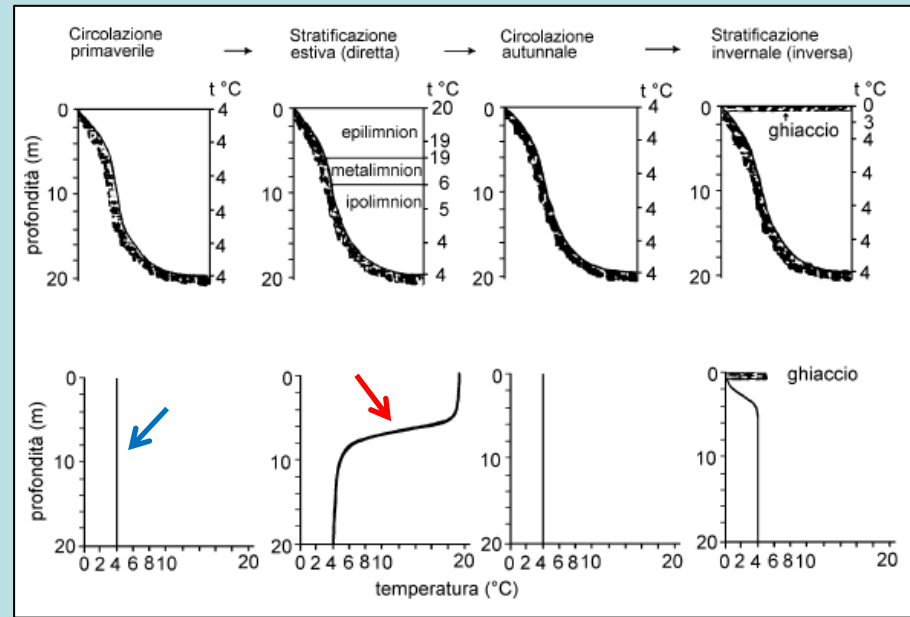


Ipolimnio: zona profonda e più fredda, isolata dall'epilimnio

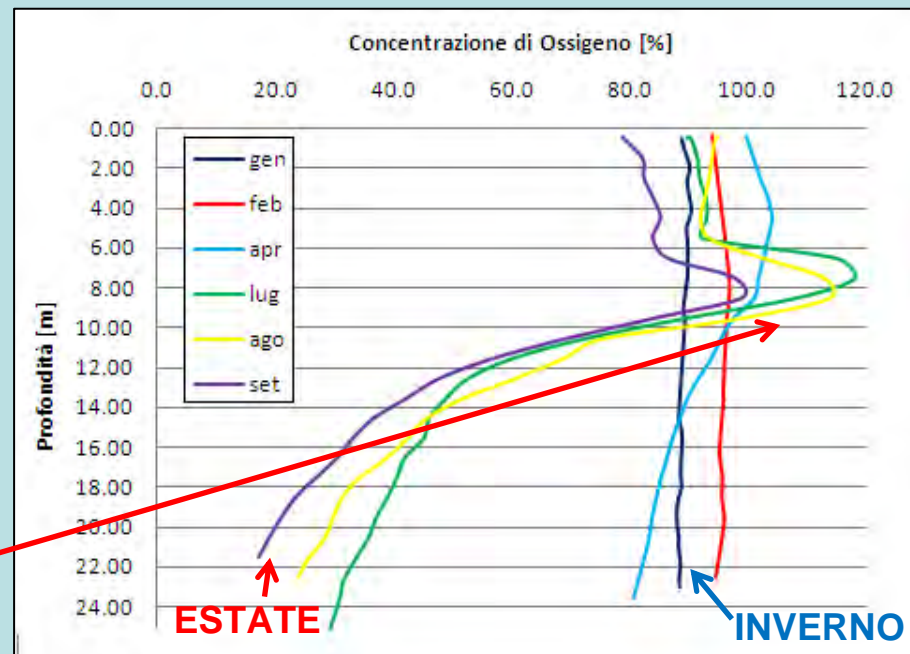
Termocline: zona in cui la variazione di temperatura è molto marcata che impedisce il mescolamento dei due strati

La stratificazione lacustre

Andamento della temperatura in funzione della profondità durante le stagioni dell'anno



La distribuzione dell'ossigeno disciolto all'interno dell'invaso è strettamente connessa all'andamento della temperatura



Incremento dovuto alla produzione da parte del fitoplancton

Analisi dei dati

➤ I dati sono stati forniti da ARPAT, dipartimento regionale di Firenze.

➤ Acquisizioni giornaliere ogni 6 ore nel periodo 2003-2005

Parametri rilevati:

- Temperatura
- Conducibilità
- Ossigeno disciolto
- Saturazione Ossigeno disciolto
- pH
- potenziale redox



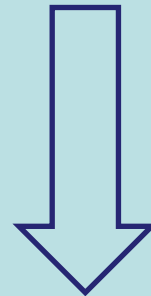
➤ Periodo 2006-2007: Dati mancanti

➤ Periodo 2008-2009 ⇒

Acquisizione con frequenza di
1-2 volte al mese
Stessi parametri rilevati

Scelta del modello

- Per definire un modello previsionale esistono due diversi approcci:
 - Modelli deterministici
 - Modelli basati sui dati



Per la complessità della struttura rispetto alla disponibilità di dati è scartata l'ipotesi di un modello deterministico

MODELLO BASATO SUI DATI ⇒ TECNICHE DI DATA MINING



Alberi delle decisioni

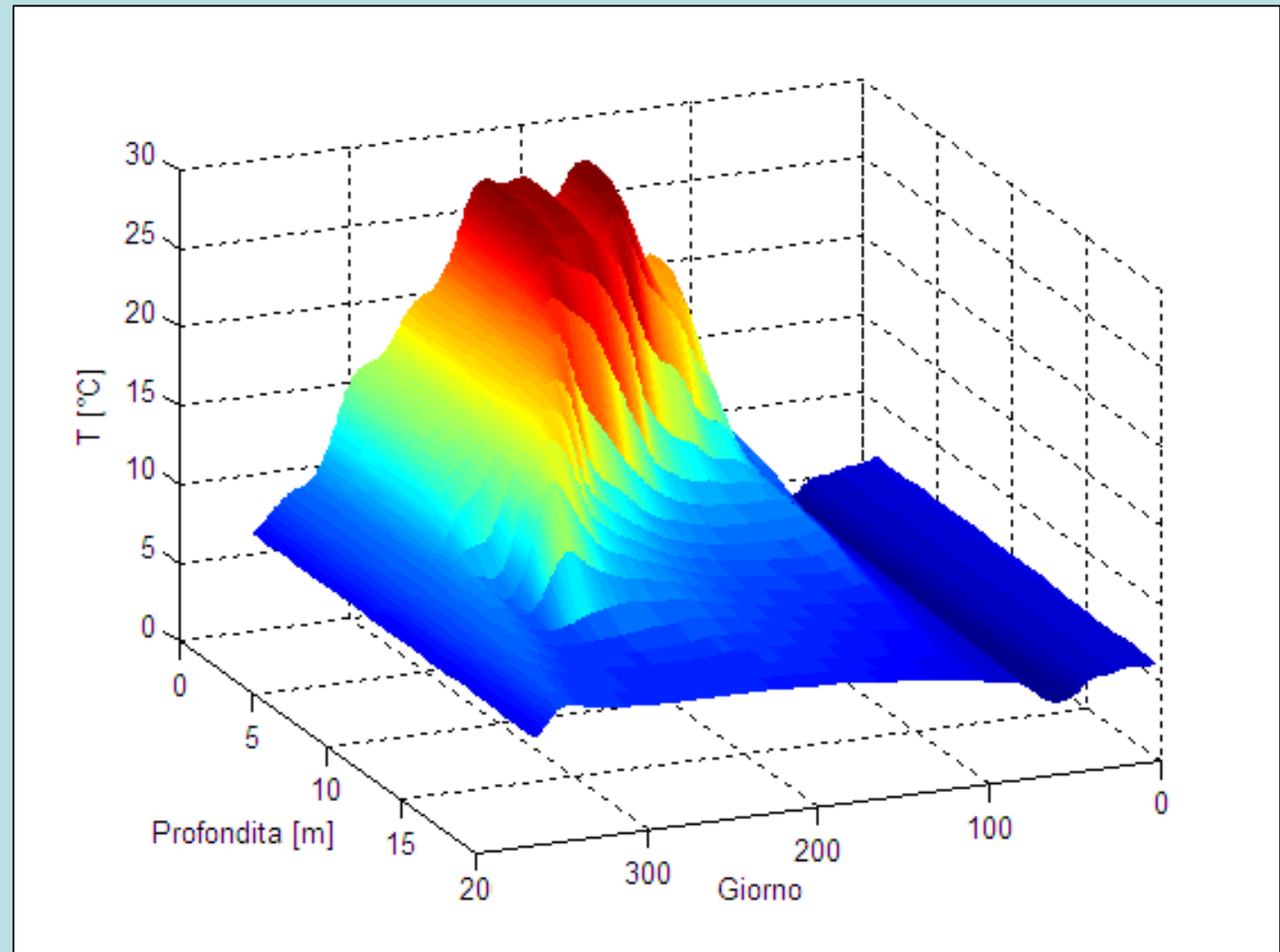
- Utilizziamo i dati solo di
 - Temperatura
 - Ossigeno disciolto
 - Profondità della termoclina



Parametri più rappresentativi per descrivere la dinamica della stratificazione

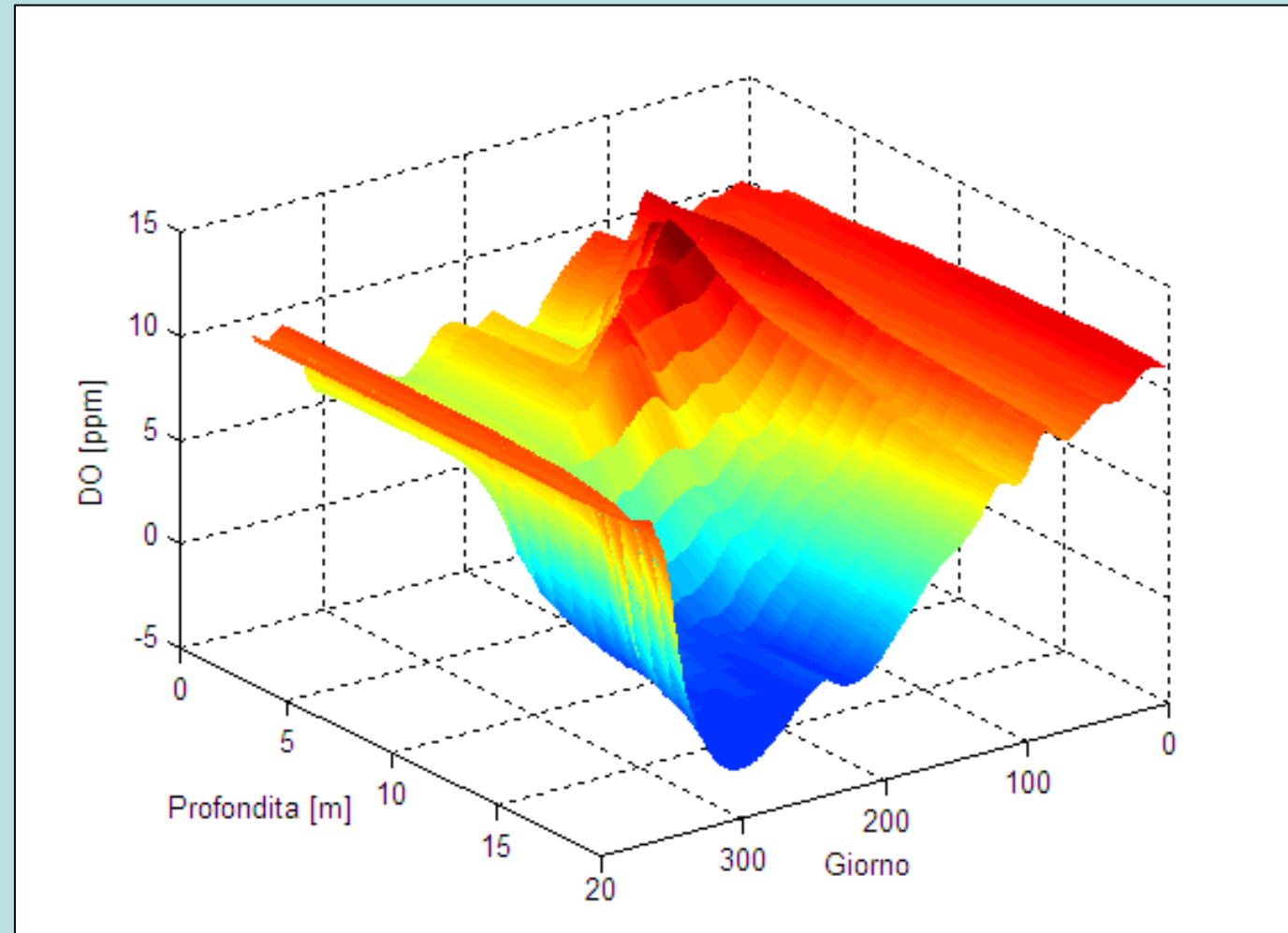
Pre-elaborazione dei dati: costruzione indicatori

Andamento della
temperatura in
funzione della
profondità nell'arco
dell'anno
[2003]



Pre-elaborazione dei dati: costruzione indicatori

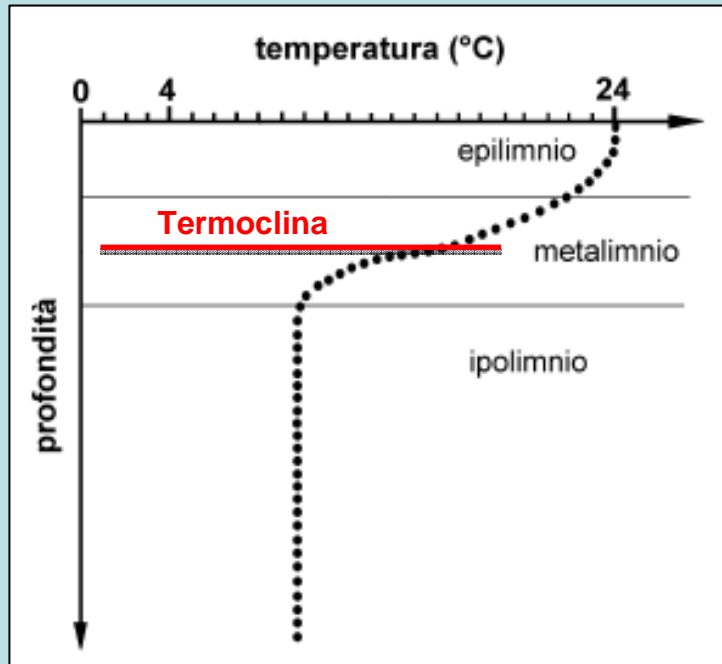
Andamento di
Ossigeno Disciolto
in funzione della
profondità nell'arco
dell'anno
[2003]



Pre-elaborazione dei dati: costruzione indicatori

➤ Dalla definizione di Termoclina

➔ Piano orizzontale che passa per il punto di flesso della curva termica

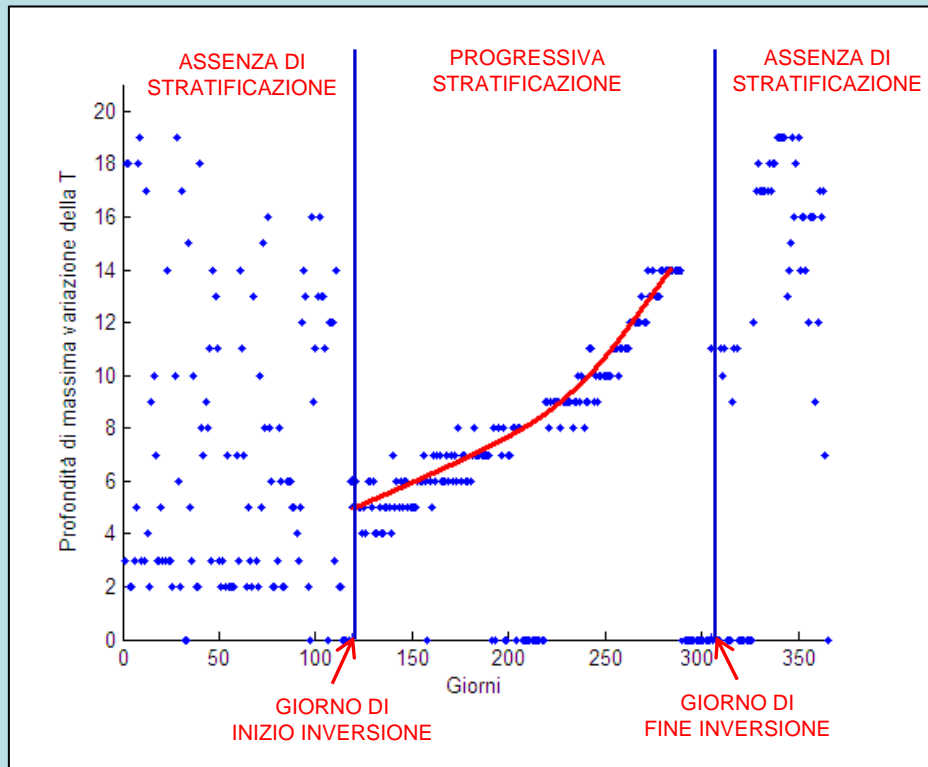


➔ Si utilizza un'approssimazione numerica della derivata seconda

$$\delta^2 T(z, k) = \frac{T(z + dz, k) - 2T(z, k) + T(z - dz, k)}{dz^2}$$

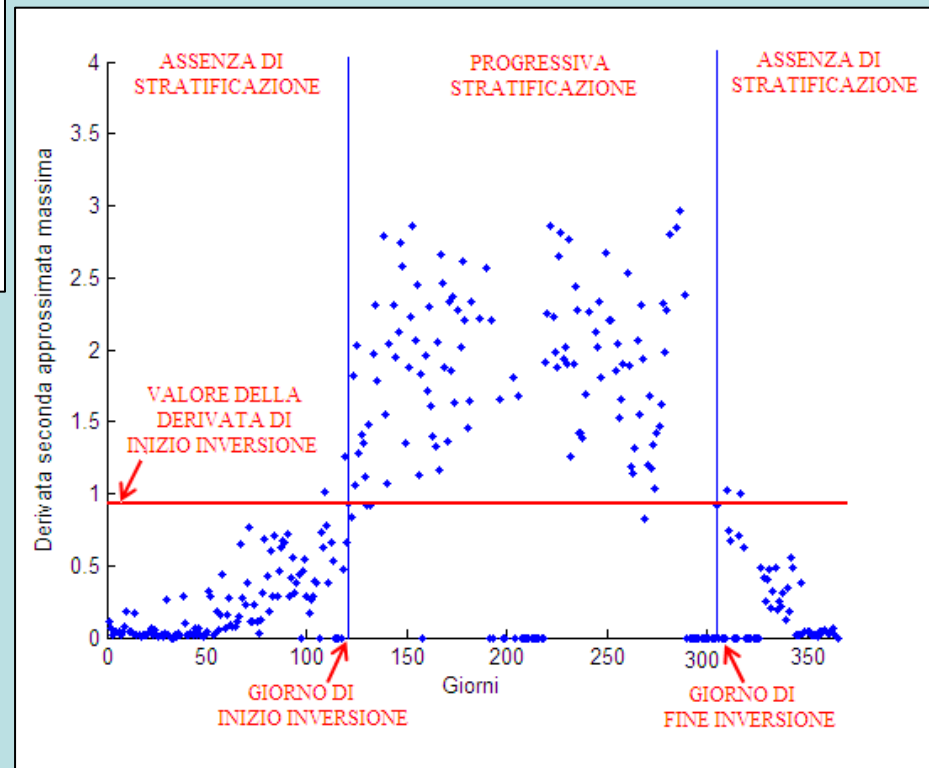
z = profondità
k = giorno

Pre-elaborazione dei dati: costruzione indicatori

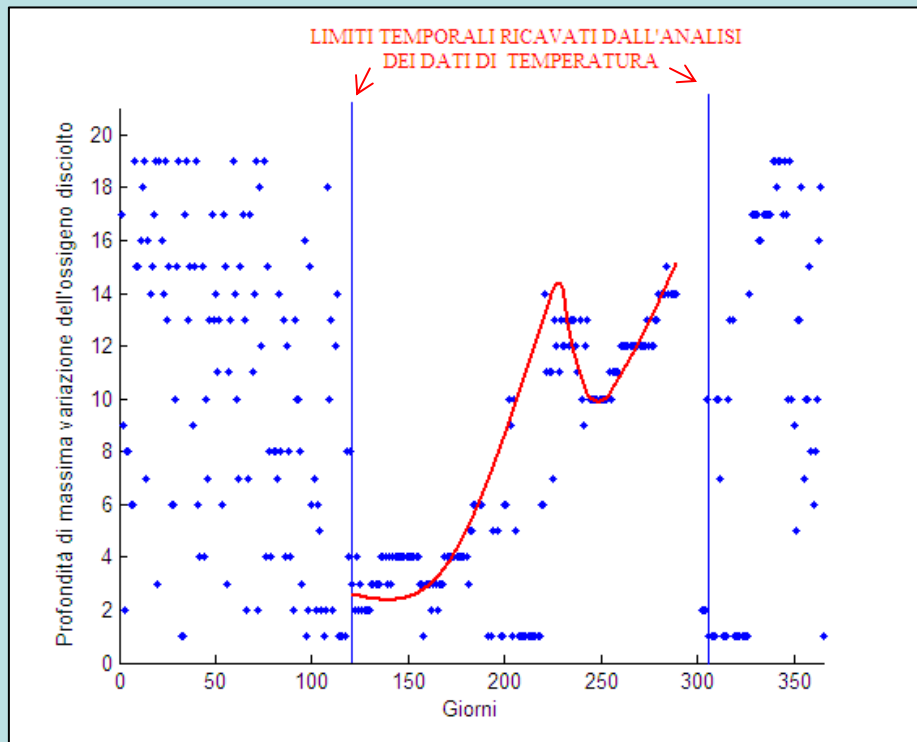


Le condizioni di stratificazione sono state ricavate sui dati relativi alla temperatura

COMPORTAMENTO A SOGLIA
sopra un certo valore si ha la presenza di stratificazione

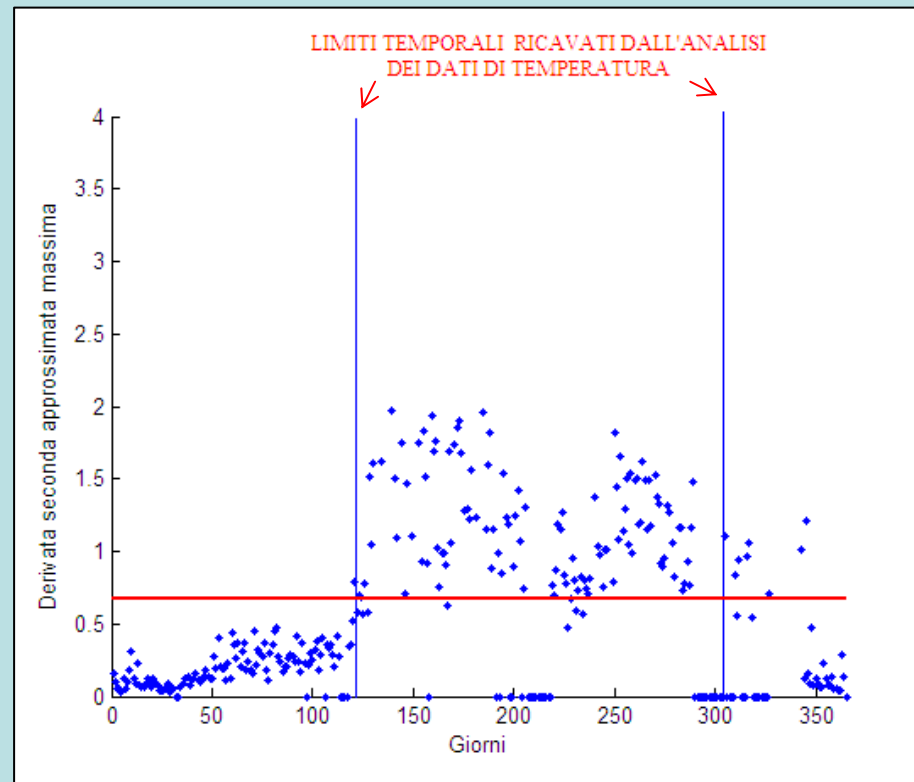


Pre-elaborazione dei dati: costruzione indicatori



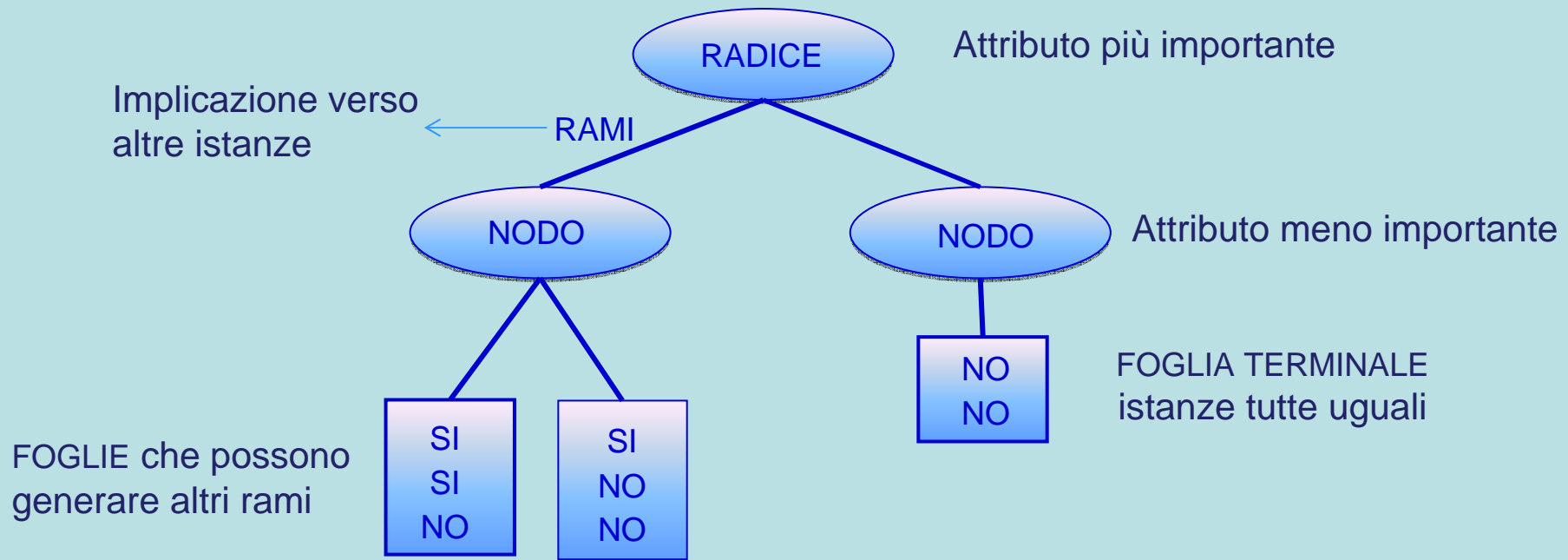
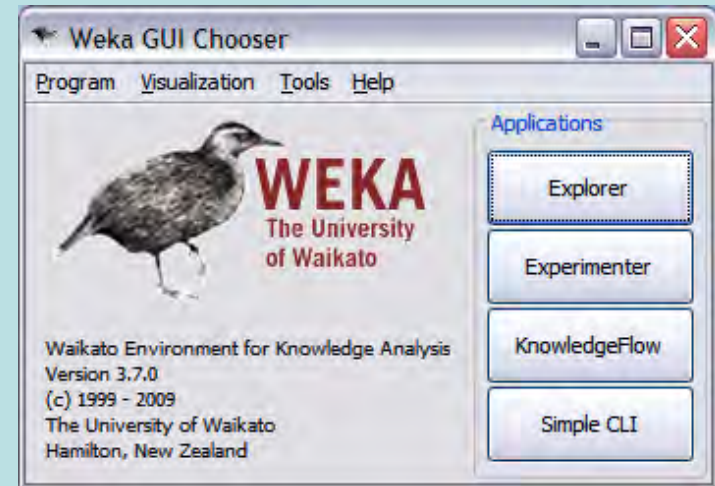
Si verifica se i risultati ottenuti sono congruenti con i dati relativi all'ossigeno disciolto

- le tre zone si delineano senza eccessivi errori
- La derivata seconda massima mantiene un comportamento a soglia



L'ambiente Weka

- Weka è un software che applica metodi di apprendimento automatici ad un set di dati estraendo l'eventuale informazione contenuta in essi.
- **Albero delle decisioni:** Struttura gerarchica che stabilisce delle priorità fra le proprietà delle variabili del modello



Costruzione dell'albero delle decisioni

➤ Attributi:

- Derivata seconda massima giornaliera della temperatura nell'intervallo di profondità
- Profondità della termoclina
- Derivata seconda massima giornaliera dell'ossigeno disciolto nell'intervallo di profondità
- Profondità relativa al flesso della curva batimetrica dell'Ossigeno Disciolto
- Verificarsi della stratificazione (Si/No) (albero binario)

➤ L'albero è stato allenato con i dati relativi agli anni **2003** e **2005**

➤ L'albero delle decisioni è stato costruito in maniera ricorsiva sfruttando l'entropia d'informazione e il guadagno d'informazione.

Costruzione dell'albero dell'albero delle decisioni

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M
Relation: 2003
Instances: 239
Attributes: 5
Temperatura
Profondità T
Ossigeno Disciolto
Profondità DO
Stratificazione
Test mode: 10-fold cross-validation

RIASSUNTO DEI DATI IN
INGRESSO E DEL TIPO DI
ANALI EFFETTUATA

=== Classifier model (full training set) ===

J48 pruned tree

ALGORITMO DI
CLASSIFICAZIONE

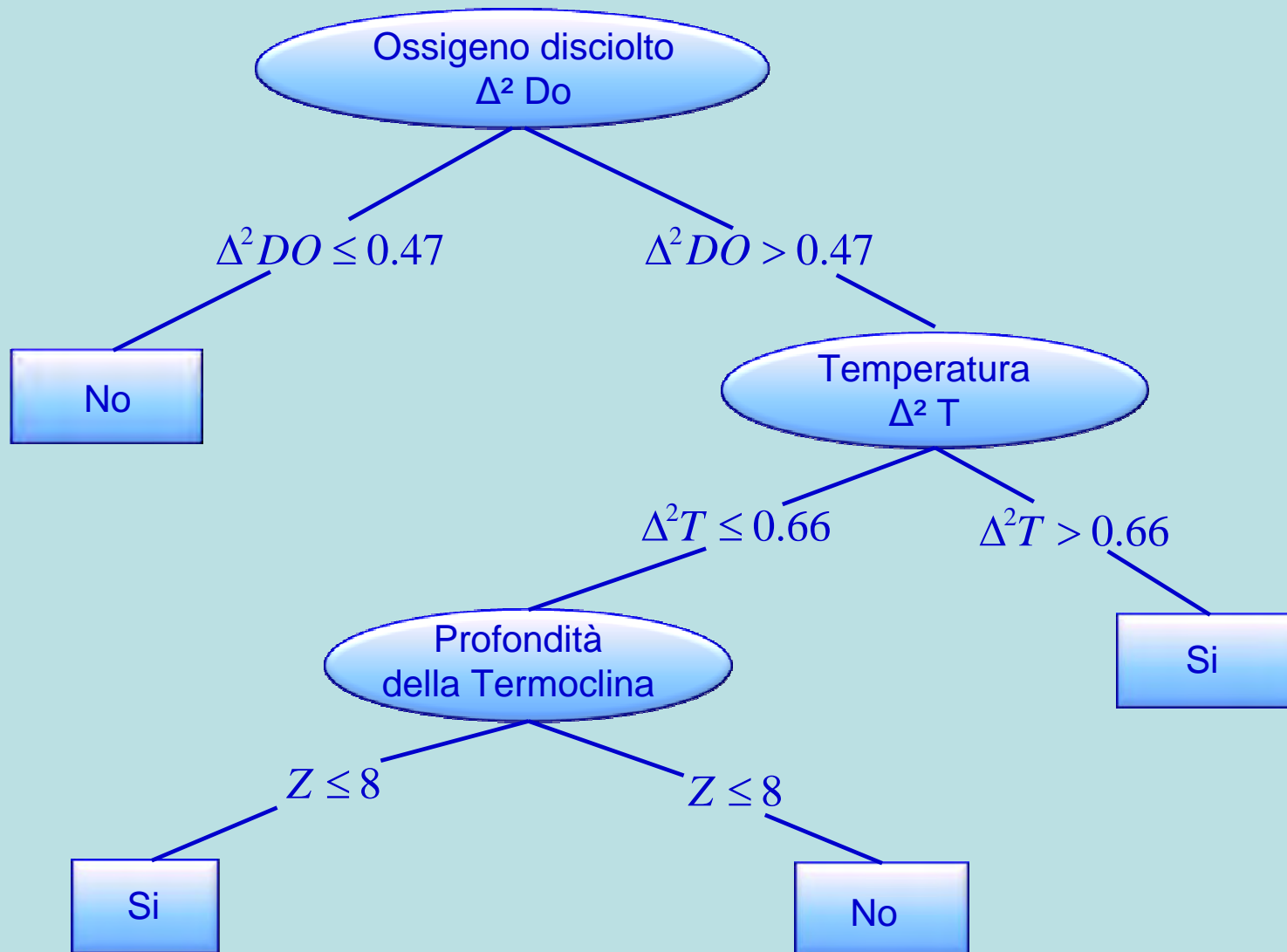
Ossigeno Disciolto <= 0.47: No (132.0/1.0)
Ossigeno Disciolto > 0.47
| Temperatura <= 0.66
| | Profondità T <= 8: Si (2.0)
| | Profondità T > 8: No (4.0)
| Temperatura > 0.66: Si (101.0)

STRUTTURA
DELL'ALBERO IN FORMA
DI TESTO

Number of Leaves : 4

Size of the tree : 7

Costruzione dell'albero dell'albero delle decisioni



Uso previsionale dell'albero

TRAINING SET



Il classificatore opera sulla classe di istanze con la quale è stato allenato

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	238	99.5816 %
Incorrectly Classified Instances	1	0.4184 %
Kappa statistic	0.9915	
Mean absolute error	0.0083	
Root mean squared error	0.0644	
Relative absolute error	1.6891 %	
Root relative squared error	12.9976 %	
Total Number of Instances	239	

QUASI LA TOTALITA' DELLE
ISTANZE E' CLASSIFICATA
CORRETTAMENTE

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.01	0.993	1	0.996	0.995	No
	0.99	0	1	0.99	0.995	0.995	Si
Weighted Avg.	0.996	0.005	0.996	0.996	0.996	0.995	

=== Confusion Matrix ===

a	b	<-- classified as
135	0	a = No
1	103	b = Si



	CLASSE PREDETTA	
	Si	No
CLASSE REALE	Si Veri Positivi	No Falsi Negativi
	No Falsi Positivi	Veri Negativi

Nel caso perfetto essa è diagonale, altrimenti ha fuori diagonale il numero di istanze si classificate come no e viceversa.

Uso previsionale dell'albero

CROSS-VALIDATION



Il classificatore opera una validazione incrociata



Raggruppa in maniera casuale l'insieme di dati specificato nei due insiemi di allenamento e di validazione

```
=== Stratified cross-validation ===
```

```
=== Summary ===
```

Correctly Classified Instances	233	97.4895 %
Incorrectly Classified Instances	6	2.5105 %
Kappa statistic	0.9489	
Mean absolute error	0.0303	
Root mean squared error	0.1591	
Relative absolute error	6.1578 %	
Root relative squared error	32.0892 %	
Total Number of Instances	239	

LA PERCENTUALE DI ISTANZE
CORRETTAMENTE
CLASSIFICATE SI MANTIENE
ELEVATA

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.978	0.029	0.978	0.978	0.978	0.972	No
	0.971	0.022	0.971	0.971	0.971	0.972	Si
Weighted Avg.	0.975	0.026	0.975	0.975	0.975	0.972	

```
=== Confusion Matrix ===
```

```
a  b  <-- classified as
132  3 |  a = No
  3 101 |  b = Si
```

Uso previsionale dell'albero

SUPPLIED TEST SET



Il classificatore opera su un nuovo insieme di dati

Giorno	Mese	Stratificazione	Classificazione
9	Gennaio	No	No
12	Gennaio	No	No
26	Gennaio	No	No
16	Febbraio	No	No
12	Marzo	No	No
16	Marzo	No	No
18	Marzo	No	No
7	Aprile	No	No
15	Aprile	Si	No
11	Maggio	Si	No
18	Maggio	Si	Si
22	Maggio	Si	Si
3	Giugno	Si	Si
5	Giugno	Si	Si
9	Giugno	Si	Si
9	Luglio	Si	Si
14	Luglio	Si	Si
6	Agosto	Si	Si
18	Agosto	Si	Si
22	Agosto	Si	Si
5	Settembre	Si	Si
10	Settembre	Si	Si
15	Settembre	Si	Si
6	Ottobre	Si	Si
10	Ottobre	Si	Si
20	Ottobre	Si	Si
5	Novembre	Si	Si
18	Novembre	No	No
4	Dicembre	No	No

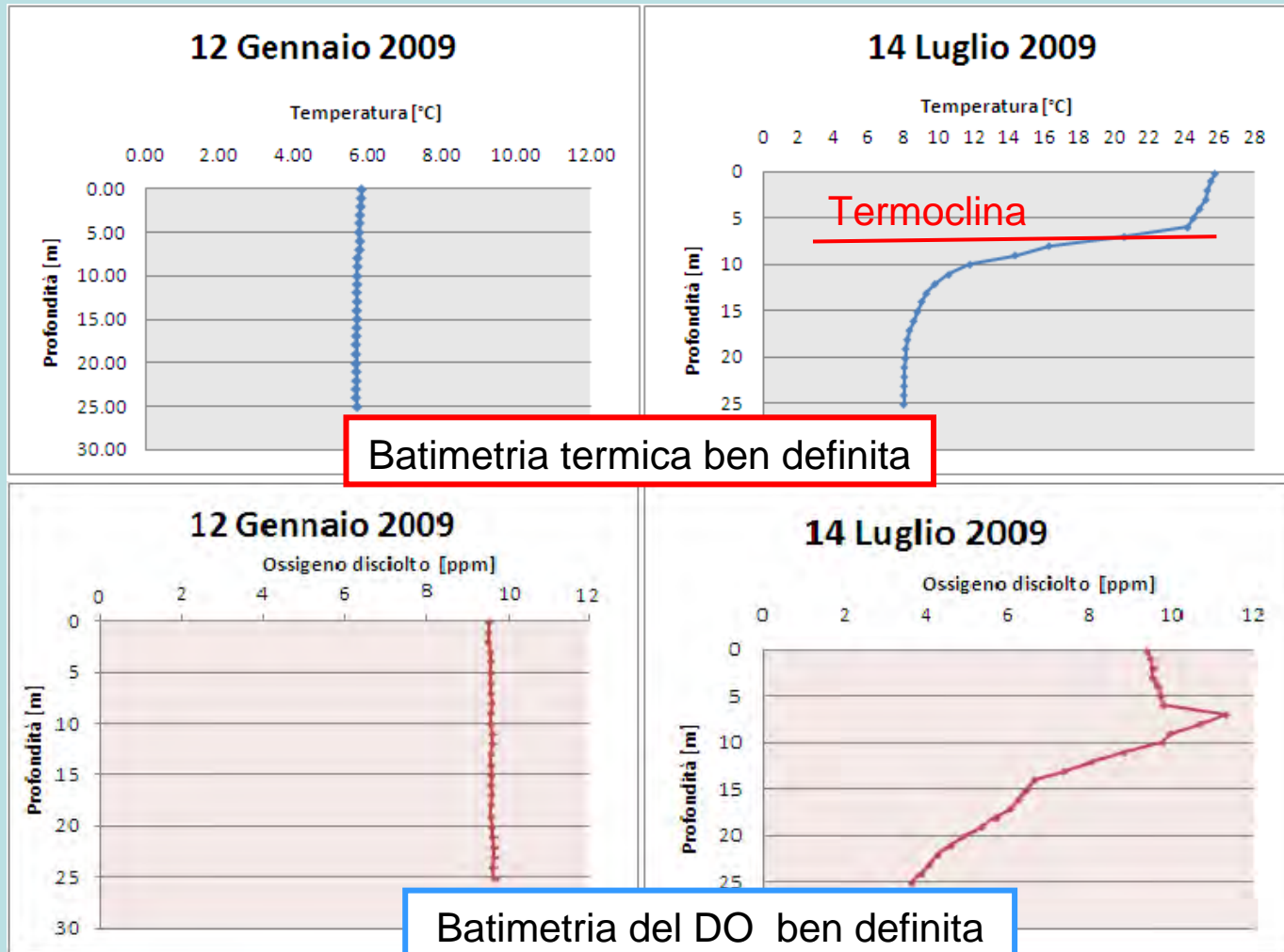
Il giorni usati corrispondono agli anni 2008 e 2009

Dei 29 giorni utilizzati per la validazione si sono avuti 2 classificazioni errate

La causa degli errori sta nella forma indefinita di uno dei due indicatori

Uso previsionale dell'albero

Comportamento osservato nei giorni correttamente classificati

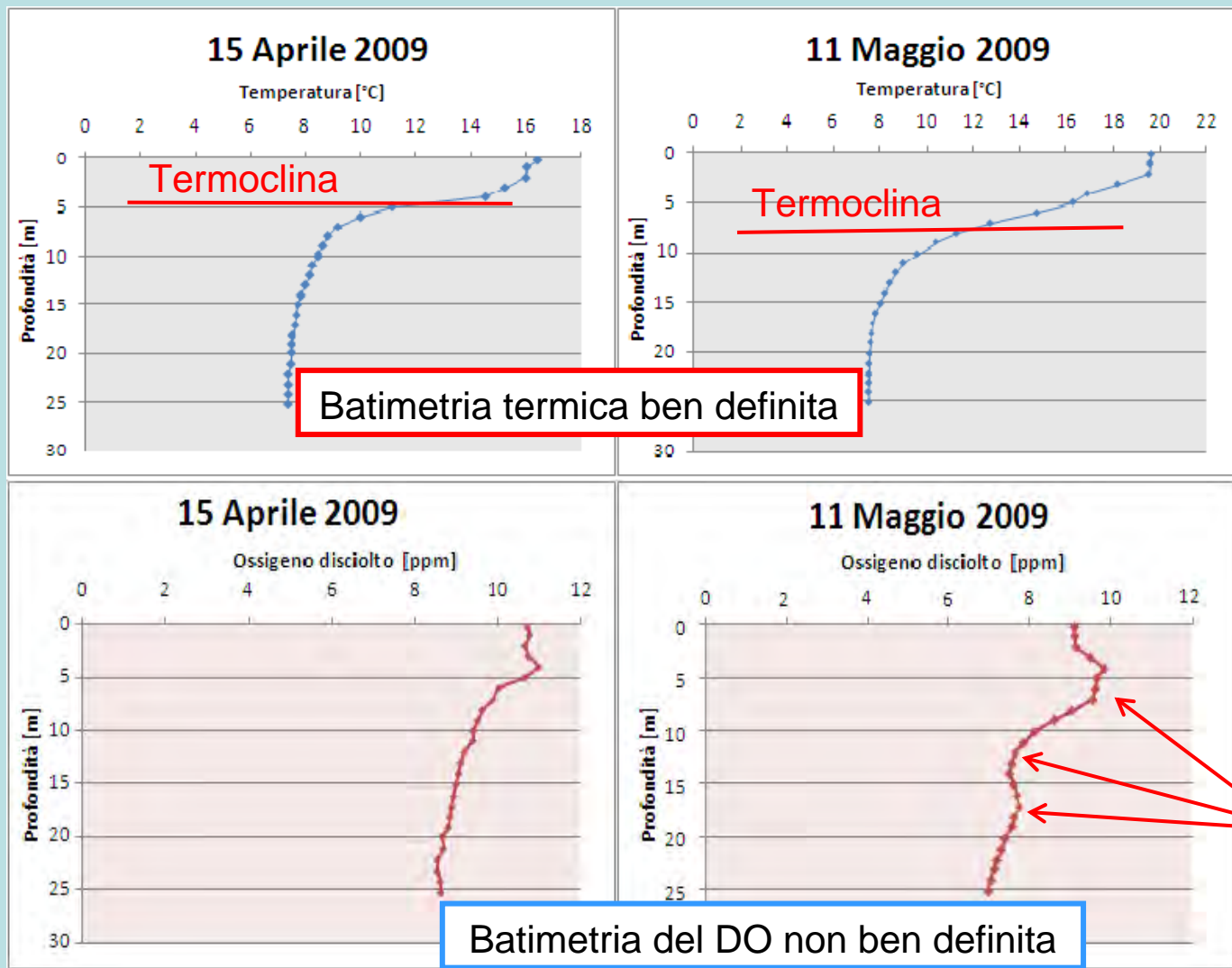


Sia l'andamento della temperatura che dell'ossigeno disciolto sono regolari

Si identificano nitidamente i profili caratteristici della **stratificazione estiva** e del **rimescolamento invernale**

Uso previsionale dell'albero

Comportamento osservato per i due giorni erroneamente classificati



L'andamento della temperatura è abbastanza regolare

L'errore dovuto alla forma indefinita del DO rispetto ai profili che il classificatore conosce (incertezza nella localizzazione del flesso)

Conclusioni

- **E' stato sviluppato un modello di stratificazione basato sui dati**
- **Dalla fase di pre-elaborazione dei dati è stato ricavato che:**
 - La derivata seconda massima giornaliera della temperatura e dell'ossigeno disciolto sono buoni indicatori della stratificazione
 - Tali indicatori presentano un comportamento a soglia
- **Il modello è costituito da un albero delle decisioni basato su Temperatura, Ossigeno Disciolto e Profondità della termoclina**
- **La classificazione è corretta nel:**
 - 99.5816% dei casi utilizzando il training set
 - 97.4895% dei casi nella cross-validation
 - 93.1034% dei casi nella validazione esterna utilizzando dati degli 2008 e 2009