Identificazione di Regimi Energetici negli ioni pesanti del Vento Solare tramite Hidden Markov Models

Alberto Prino (348174)

A.S. 2025/2026

Abstract

Lo scopo di questo lavoro è di trovare modelli statistici predittivi per un dataset contenente dati sul vento solare nell'arco di tempo 2005-2011. L'obiettivo è identificare e prevedere regimi temporali a diversa energia media del plasma. Si sono usate due tecniche: un Hidden Markov Model e un'analisi ARIMA. L'HMM viene allenato su un sottoinsieme temporale dei dati e testato in regime out-of-sample. I risultati mostrano una chiara separazione tra stati a bassa e alta energia, una dinamica persistente dei regimi e una buona coerenza statistica con l'ipotesi markoviana, come verificato tramite test di geometricità delle durate. In seguito, viene implementata una versione rolling del modello ARIMA che prevede l'andamento dell'energia per pochi intervalli di tempo, poi aggiorna la finestra coi nuovi dati osservati e ripete la previsione. Si segnala l'uso di chatbot ai al fine di migliorare la produzione di grafici e tabelle.

1 Introduzione e visualizzazione del dataset

Tutti i file del dataset e i codici possono essere trovati al seguente link: `https://github.com/s348174/Tesina_Statistica_Computazionale`.

Il vento solare è un plasma (ovvero ioni ad alta temperatura) che fluisce continuamente dalla corona solare nello spazio interplanetario. Le sue proprietà energetiche e composizionali mostrano variazioni temporali significative, spesso associate a diversi regimi fisici caratterizzati da diversi livelli di attività della corona solare.

L'identificazione automatica di tali regimi è un problema di particolare interesse per la progettazione di schermature attive (ovvero mediante campi elettromagnetici) dalle radiazioni solari. Poter prevedere i quando il plasma avrà alta o bassa energia, permette di programmare in modo efficiente la potenza necessaria per alimentare uno scudo magnetico e di consegeunza il suo consumo energetico nelle prossime ore, dato che le risorse su un velivolo spaziale sono limitate.

1.1 Dataset SWICS/ACE

Il dataset utilizzato in questo lavoro è derivato dalle osservazioni dello strumento *Solar Wind Ion Composition Spectrometer* (SWICS), a bordo della missione *Advanced Composition Explorer* (ACE). La sonda ACE opera in prossimità del punto di Lagrange L1 del sistema Sole–Terra (ovvero il punto di equilibrio tra la gravità terrestre e quella solare) e fornisce misure del vento solare prima che esso venga modificato dall'interazione con la magnetosfera terrestre.

I dati sono campionati con una cadenza temporale di due ore e descrivono le proprietà cinematiche, termiche e composizionali di diverse specie ioniche del vento solare. In particolare, il dataset include informazioni su ioni elio, carbonio, ossigeno e ferro, elementi chiave per lo studio dei processi coronali e della dinamica del plasma solare. Il dataset contiene 24924 entrate e 13 features.

1.2 Descrizione delle variabili

Le variabili presenti nel dataset sono elencate di seguito:

Table 1: Variabili del dataset SWICS/ACE e relative unità fisiche.

Variabile	**Descrizione fisica**	**Unità**
`time`	Tempo di osservazione (centro dell'intervallo)	–
`nHe2`	Densità numerica degli ioni He^{2+}	cm^{-3}
`vHe2`	Velocità di flusso degli ioni He^{2+}	$km\,s^{-1}$
`vthHe2`	Velocità termica degli ioni He^{2+}	$km\,s^{-1}$
`vC5`	Velocità di flusso degli ioni C^{5+}	$km\,s^{-1}$
`vthC5`	Velocità termica degli ioni C^{5+}	$km\,s^{-1}$
`vO6`	Velocità di flusso degli ioni O^{6+}	$km\,s^{-1}$
`vthO6`	Velocità termica degli ioni O^{6+}	$km\,s^{-1}$
`vFe10`	Velocità di flusso degli ioni Fe^{10+}	$km\,s^{-1}$
`vthFe10`	Velocità termica degli ioni Fe^{10+}	$km\,s^{-1}$
`HetoO`	Rapporto di abbondanza He/O	adimensionale
`CtoO`	Rapporto di abbondanza C/O	adimensionale
`FetoO`	Rapporto di abbondanza Fe/O	adimensionale

Le immagini successive mostrano l'andamento temporale di alcune features.

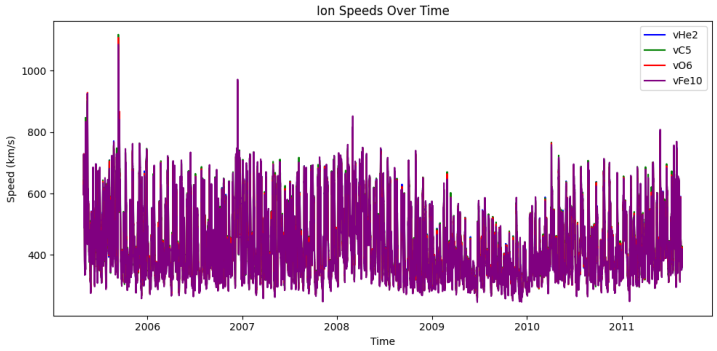

Figure 1: Andamento delle velocità dei 4 tipi di ioni nel tempo.

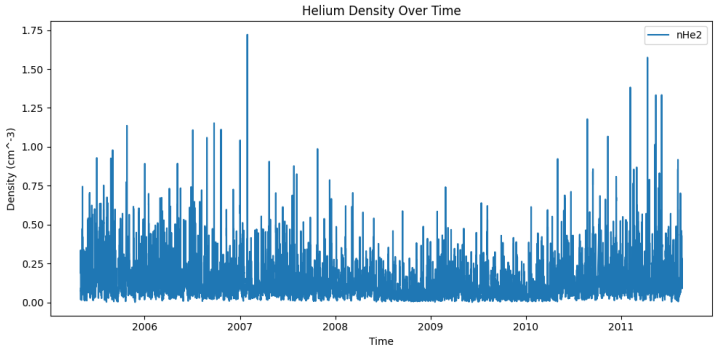

Figure 2: Andamento della densità degli ioni He^{2+} nel tempo.

1.3 Analisi bayesiana della composizione

L'analisi statistica della composizione del plasma è stata condotta a partire dalle densità ioniche e dai rapporti di abbondanza misurati dallo strumento SWICS/ACE. In primo luogo, la densità

assoluta degli ioni di ossigeno è stata ricostruita utilizzando il rapporto He/O e la densità misurata degli ioni He^{2+}. A partire da tale quantità, sono state stimate le abbondanze assolute di carbonio e ferro tramite i rispettivi rapporti C/O e Fe/O. Le abbondanze assolute sono state infine sommate sull'intero dataset temporale, ottenendo una stima globale dei conteggi relativi delle specie He, C, O e Fe.

Per modellare in modo probabilistico la composizione del plasma, è stato adottato un approccio bayesiano basato sulla distribuzione di Dirichlet. Assumendo un prior non informativo uniforme sulle abbondanze relative delle quattro specie ioniche considerate, la distribuzione a posteriori è stata costruita combinando il prior con i conteggi osservati e sfruttando il coniugio della Dirichlet con la multinomiale. Sono infine stati generati campioni della distribuzione a posteriori, consentendo di stimare le abbondanze relative medie e la loro dispersione statistica.

Successivamente, è stata testata l'ipotesi che la composizione del plasma segua una legge funzionale dipendente dalla massa atomica degli ioni, del tipo

$$p_i \propto \exp(-\lambda m_i),$$

dove m_i rappresenta la massa atomica della specie ionica e λ è un parametro libero. Il valore ottimale di λ è stato stimato minimizzando la distanza quadratica tra le abbondanze medie a posteriori e il modello teorico. Infine, la compatibilità statistica tra il modello e la distribuzione a posteriori è stata valutata tramite un test basato sulla distanza euclidea nello spazio delle composizioni, permettendo di stimare un valore di *p-value* associato all'ipotesi di dipendenza dalla massa.

L'algoritmo adottato per verificare l'ipotesi di una dipendenza esponenziale della composizione del plasma dalla massa atomica delle specie ioniche può essere riassunto nei seguenti passaggi:

1. Costruzione della distribuzione a posteriori delle abbondanze relative $\mathbf{p}$ mediante un modello bayesiano Dirichlet, a partire dai conteggi osservati sommati sull'intero dataset.

2. Calcolo del valore medio a posteriori $\langle \mathbf{p} \rangle$ delle abbondanze relative.

3. Definizione di un modello parametrico di composizione dipendente dalla massa atomica m_i,
$$p_i^{\text{model}}(\lambda) = \frac{\exp(-\lambda m_i)}{\sum_j \exp(-\lambda m_j)}.$$

4. Stima del parametro λ tramite minimizzazione numerica della funzione di loss quadratica
$$\mathcal{L}(\lambda) = \sum_i \left(p_i^{\text{model}}(\lambda) - \langle p_i \rangle \right)^2.$$

5. Valutazione della bontà del modello confrontando la distanza euclidea tra il modello ottimale e i campioni della distribuzione a posteriori, stimando un *p-value* come frazione dei campioni che risultano meno compatibili del valore medio a posteriori.

I risultati si possono osservare nella seguente immagine:

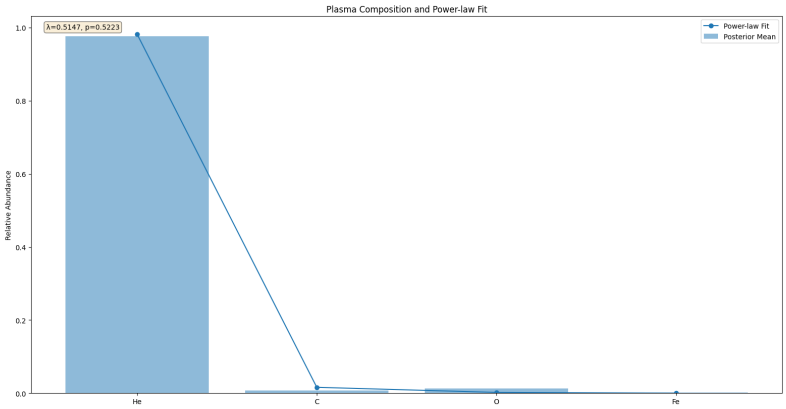

Figure 3: Adattamento esponenziale della composizione in base alla massa.

2 Definizione dell'energia media

I dati utilizzati provengono dallo strumento SWICS a bordo della sonda ACE e comprendono velocità ioniche per diverse specie (He, C, O, Fe). Per ogni specie i viene calcolata l'energia cinetica media:

$$E_i(t) = \frac{1}{2} m_i v_i^2(t), \tag{1}$$

espressa in elettronvolt.

L'energia media del plasma è definita come media pesata sulle abbondanze relative:

$$E_{\text{mean}}(t) = \sum_i w_i(t) E_i(t), \tag{2}$$

dove i pesi w_i sono determinati dai rapporti di abbondanza ionica misurati. La serie temporale risultante costituisce l'osservabile unidimensionale utilizzata dall'HMM.

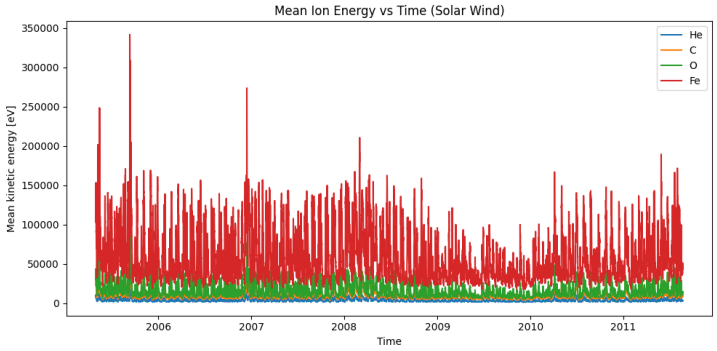

Figure 4: Andamento dell'energia dei vari tipi di ioni nel tempo.

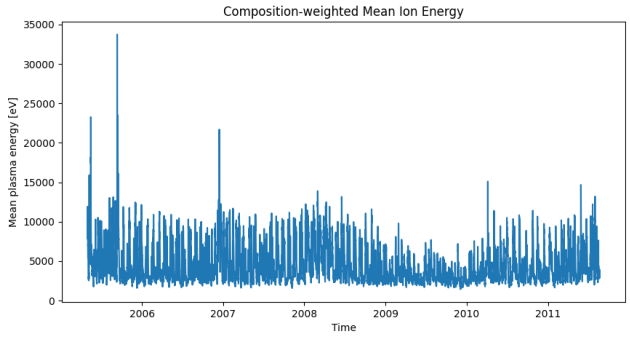

Figure 5: Andamento dell'energia media pesata nel tempo.

3 Teoria degli Hidden Markov Models

Osservando i dati, si notano massimi e minimi molto variabili nell'energia. Possiamo quindi chiederci se essi siano riconducibili a variabili nascoste di regime di attività solare (regime ad alta energia e regime a bassa energia) e se sia possibile identificare e prevedere tali regimi. Lo strumento statistico adatto sono gli Hidden Markov Models (HMM). Un HMM è definito da:

- una sequenza di stati latenti discreti $z_t \in \{1, \dots, K\}$,
- una matrice di transizione $\mathbf{P}$, con

$$\pi_{ij} = p(z_{t+1} = j \mid z_t = i),$$

- una distribuzione di emissione $p(x_t \mid z_t)$.

Nel caso considerato, si assume:

$$p(x_t \mid z_t = k) = \mathcal{N}(x_t \mid \mu_k, \sigma_k),$$

ossia un HMM gaussiano. L'insieme dei parametri per un modello a due stadi è:

$$\theta = (\mathbf{P}, \mu_1, \mu_2, \sigma_1, \sigma_2).$$

3.1 Inferenza negli Hidden Markov Models

Vediamo come è costruito il modello usato nella libreria *hmmlearn.GaussianHMM* in python. Come prior sulle probabilità di transizione si assume una Dirichlet:

$$\pi_{z_0,k} \sim Dir(\alpha_0),$$

con $\alpha_0 = (1, 1)$, ovvero la probabilità dello stato iniziale z_0 è uniforme. La stessa scelta di prior viene fatta sulle righe della matrice di transizione:

$$\pi_k \sim Dir(\alpha),\ \alpha = (1, 1).$$

La prior sulla media delle distribuzioni di emissione è normale: $\mu_k = \mathcal{N}(0, \frac{1}{w}\sigma_k^2)$, mentre sulle varianze è una gamma inversa: $\sigma_k^2 = IG(1, 0.01)$. Poiché il parametro default di w è 0, la prior non ha alcuna influenza e la media viene stimata dal modello basandosi sui dati osservati. La media viene attraverso il metodo MAP (maximum a posteriori) con la seguente formula:

$$\mu_k = \frac{w \cdot \mu_{prior} + \sum_t \gamma_k(t) x_t}{w + \sum_t \gamma_k(t)} = \frac{\sum_t \gamma_k(t) x_t}{\sum_t \gamma_t(k)}.$$

Poiché $w = 0$, dipende solo dal contributo dei dati moltiplicati per i pesi $\gamma_t(k)$.

Durante la fase di addestramento del modello, i parametri dell'HMM (matrice di transizione e parametri delle distribuzioni di emissione) vengono stimati massimizzando la log-verosimiglianza dei dati osservati tramite un algoritmo di Expectation–Maximization (EM), noto come algoritmo di Baum-Welch. L'algoritmo si articola in due step.

Passo E (Expectation) In questa fase si usa l'algoritmo forward-backward per valutare i dati, senza modificare i parametri. Si vuole ottenere la quantità:

$$\gamma_t(k) = P(z_t = k \mid \mathbf{x}_{1:T}, \theta^{(n)}),$$

ovvero la probabilità di essere nello stato k al tempo t. Lo si fa in tre step.

- **Passo forward:** sia $\alpha_t(i) = p(x_1, ..., x_t, z_t = k|\theta)$, la probablità di osservare $x_1, ..., x_t$ ed essere allo stato k al tempo t. Si calcola ricorsivamente.

  Inizializzazione:
  $\alpha_1(i) = \pi_i \cdot p(x_1|z_1 = i) = \pi_i \cdot \mathcal{N}(x_1; \mu_i, \Sigma_i), \quad 1 \leq i \leq N$

  Ricorsione:
  $$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\pi_{ij}\right] \cdot \mathcal{N}(x_{t+1}; \mu_j, \Sigma_j), \quad 1 \leq t \leq T-1,\ 1 \leq j \leq N$$

  Probabilità totale:
  $$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- **Passo backward:** sia $\beta_t(i) = p(x_{t+1}, ..., x_T|z_t = k, \theta)$ la probablità di osservare $x_{t+1}, ..., x_T$ sapendo che al tempo t il processo è allo stato k. Si calcola ricorsivamente.

  Inizializzazione:
  $\beta_T(i) = 1, \quad 1 \leq i \leq N$

  Ricorsione:
  $$\beta_t(i) = \sum_{j=1}^{N} \pi_{ij} \cdot \mathcal{N}(x_{t+1}; \mu_j, \Sigma_j) \cdot \beta_{t+1}(j), \quad t = T-1, T-2, \ldots, 1,\ 1 \leq i \leq N$$

- **Calcolo delle probabilità derivate:**

  Probabilità dello stato:
  $$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

  Probabilità di transizione:
  $$\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid \mathbf{x}_{1:T}, \theta^{(n)}) = \frac{\alpha_t(i)\pi_{ij}\mathcal{N}(x_{t+1}; \mu_j, \sigma_j)\beta_{t+1}(j)}{\sum_{k=1}^{N}\sum_{m=1}^{N} \alpha_t(k)\pi_{km}\mathcal{N}(x_{t+1}; \mu_m, \Sigma_m)\beta_{t+1}(m)}$$

Queste quantità rappresentano rispettivamente la probabilità a posteriori di occupazione di ciascuno stato e il numero atteso di transizioni tra stati consecutivi.

Passo M (Maximization) I parametri del modello vengono aggiornati per massimizzare la verosimiglianza:

$$\pi_{z_0,k}^{(n+1)} = \gamma_1(k), \tag{3}$$

$$\pi_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{4}$$

$$\mu_k^{(n+1)} = \frac{\sum_{t=1}^{T} \gamma_t(k)\, x_t}{\sum_{t=1}^{T} \gamma_t(k)}, \tag{5}$$

$$\sigma_k^{(n+1)} = \frac{\sum_{t=1}^{T} \gamma_t(k)\, (x_t - \mu_k)(x_t - \mu_k)^\top}{\sum_{t=1}^{T} \gamma_t(k)}. \tag{6}$$

Criterio di arresto I passi E e M vengono iterati fino al raggiungimento di uno dei seguenti criteri:

- convergenza della log-verosimiglianza,
- raggiungimento del numero massimo di iterazioni predefinito.

L'algoritmo EM garantisce che la log-verosimiglianza dei dati osservati non diminuisca a ogni iterazione, assicurando una convergenza monotona verso un massimo locale.

Una volta stimato il modello, la segmentazione della serie temporale in regimi discreti viene ottenuta tramite l'algoritmo di Viterbi, che calcola la singola sequenza più probabile di stati latenti nel senso della massima probabilità a posteriori (MAP):

$$\hat{z}_{1:T} = \arg\max_{z_{1:T}} P(z_{1:T} \mid x_{1:T}). \tag{7}$$

Questa sequenza fornisce una classificazione dei regimi energetici ed è utilizzata per l'analisi delle durate e per la visualizzazione dei cambi di regime nel tempo.

4 Applicazione del HMM

4.1 Addestramento e separazione dei regimi

L'applicazione degli *Hidden Markov Models* (HMM) al dataset ci permette di fare inferenza sui regimi di cui osserviamo solo l'effetto indiretto sull'energia del plasma. Lo scopo è allenare e testare un modello HMM su uno split 80/20 dei dati. La media e la varianza fornite, permetteranno quindi di calcolare a ogni istante di tempo la probabilità di un cambio di regime. Il numero di stati è fissato a $K = 2$, coerentemente con l'ipotesi di due principali regimi energetici.

Il modello è stato prima addestrato su tutto il dataset per osservare le energie medie associate agli stati:

Stato	μ [eV]
0	2918.8
1	6780.2

Questi valori indicano una chiara separazione tra uno stato a *bassa energia* e uno a *alta energia*, interpretabili rispettivamente come vento solare lento e veloce. La matrice di transizione stimata è:

$$\mathbf{P} = \begin{pmatrix} 0.983 & 0.017 \\ 0.023 & 0.977 \end{pmatrix}.$$

Le elevate probabilità diagonali indicano una forte persistenza temporale dei regimi. Nei successivi grafici si può osservare la classificazione fatta dal modello e dove il modello pensa che sia probabile avere alta attività solare.

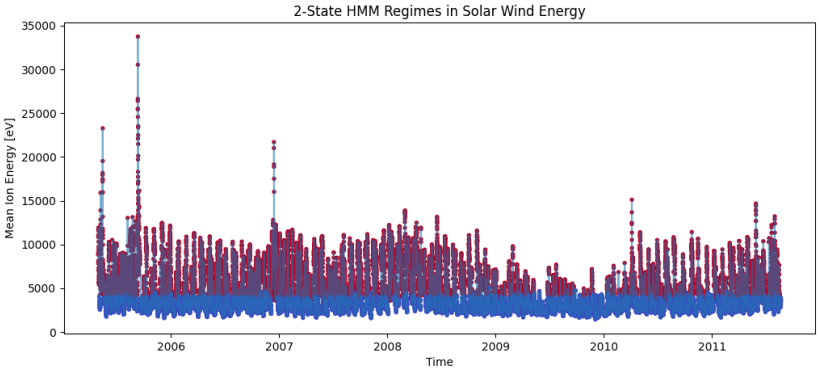

Figure 6: Classificazione dei punti in due stati.

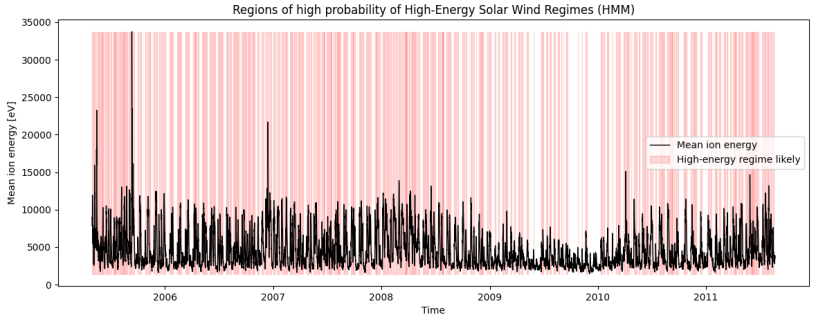

Figure 7: Regioni ad alta probabilità di intensa attività solare.

4.2 Valutazione predittiva

In seguito si è allenato il modello sul primo 80% dei dati e lo si è testato sul restante 20%. Le durate medie teoriche, date da:

$$\mathbb{E}[D_k] = \frac{1}{1 - \pi_{kk}},$$

risultano:

$$\langle D_0 \rangle \approx 59, \quad \langle D_1 \rangle \approx 44.$$

Le prestazioni del modello sono valutate tramite diverse metriche:

- Log-verosimiglianza:

$$\log \mathcal{L}_{\text{train}} = -1.68 \times 10^5, \quad \log \mathcal{L}_{\text{test}} = -4.18 \times 10^4,$$

  con log-likelihood per punto nel test pari a -8.39.

- Predictive log-score medio: -9.65.

- Errore assoluto medio (MAE):

$$\text{MAE}_{\text{baseline}} = 214 \text{ eV}, \quad \text{MAE}_{\text{HMM}} = 997 \text{ eV}.$$

Il peggioramento del MAE rispetto al baseline (ovvero il random walk naive) è atteso: l'HMM non è progettato per minimizzare errori puntuali, ma per catturare regimi latenti persistenti. Il valore predittivo del modello risiede quindi nella classificazione probabilistica dei regimi, non nella previsione punto-a-punto, come si può osservare dal grafico delle predizioni:

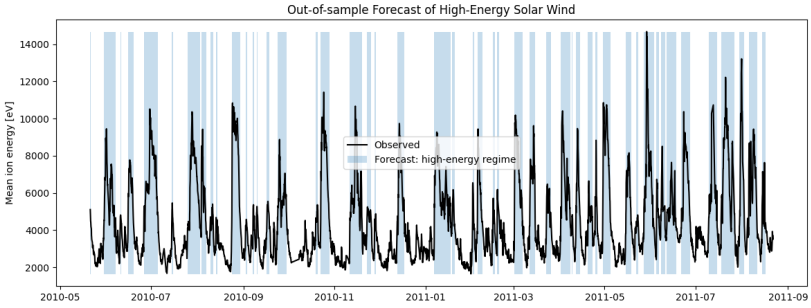

Figure 8: Previsione di regime sul test split del dataset.

4.3 Verifica dell'ipotesi HMM

Una proprietà fondamentale degli HMM è che la durata di permanenza in uno stato segue una distribuzione geometrica:

$$P(D = d) = (1 - p_{\text{stay}})p_{\text{stay}}^{d-1}, \tag{8}$$

dove $p_{\text{stay}} = \mathbf{P}_{kk}$ è la probabilità di rimanere nello stato k.

Si può osservare che le distribuzioni delle durate dei regimi assumono le seguenti forme:

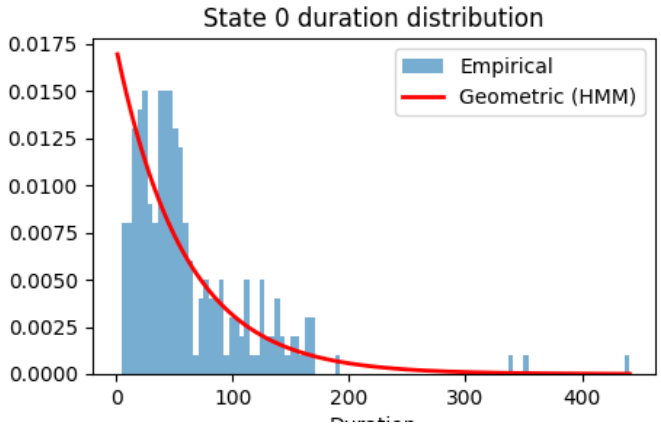

Figure 9: Distribuzione durata stato a bassa energia.

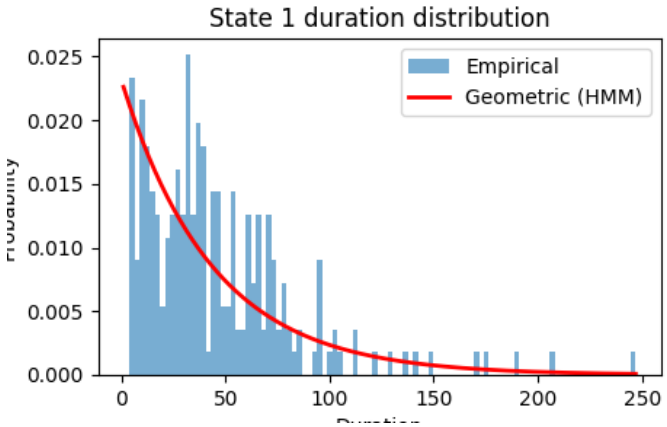

Figure 10: Distribuzione durata stato ad alta energia.

Per validare l'assunzione markoviana, si è testata la geometricità delle distribuzioni di durata tramite un test di Kolmogorov–Smirnov. Il test KS è un test non parametrico utilizzato per valutare la compatibilità tra una distribuzione empirica e una distribuzione teorica assegnata. Nel contesto di questo lavoro, esso viene impiegato per verificare se le durate dei regimi latenti siano coerenti con una distribuzione geometrica, come previsto dalla teoria degli HMM.

Dato un campione di osservazioni indipendenti $\{x_1, \ldots, x_n\}$, si definisce la funzione di distribuzione empirica (ECDF) come:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq x}, \tag{9}$$

dove $\mathbf{1}$ indica la funzione indicatrice.

Sia $F(x)$ la funzione di distribuzione cumulativa teorica sotto l'ipotesi nulla H_0. La statistica del test di Kolmogorov–Smirnov è definita come:

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{10}$$

ovvero la massima distanza tra la distribuzione empirica e quella teorica.

Valori elevati di D_n indicano una discrepanza significativa tra i dati osservati e il modello teorico. Il p-value del test è calcolato a partire dalla distribuzione asintotica di D_n sotto l'ipotesi H_0. I risultati sono:

Stato	$\hat{p}_{\text{stay}}$	KS	p-value
0	0.983	0.017	0.902
1	0.978	0.022	0.845

Gli elevati p-value indicano che non vi è evidenza statistica forte contro l'ipotesi di distribuzione geometrica, che giustifica la scelta del modello HMM.

5 Conclusioni su HMM

Questa analisi mostra che un HMM gaussiano su questo dataset a due stati è in grado di:

- identificare regimi energetici fisicamente interpretabili nel vento solare,
- descrivere correttamente la persistenza temporale dei regimi,
- fornire una previsione probabilistica dei periodi ad alta energia.

Sebbene non competitivo in termini di errore puntuale, il modello risulta adeguato come strumento di *regime detection* e analisi statistica dei processi dinamici del plasma solare. Il risultato del test KS, sebbene non squalifichi del tutto il modello HMM, suggerisce la possibilità di sperimentare con un modello HSMM, dato che in un processo fisico di questo tipo, si suppone che la probabilità di cambiare stato dipenda da quanto tempo si è trascorso in quello stato.

6 Teoria del modello ARIMA

Si vuole ora modellizzare non solo il regime energetico generale, ma i valori probabili che l'energia assumerà nel futuro. Un'analisi di questo tipo è una serie storica. Osservando i dati in figura 5, si osserva che, volendo modellizzare l'energia con un processo gaussiano, ovvero del tipo:

$$x_t = x_{t-1} + \omega_t,$$

il rumore ω_t ha probabilmente varianza non costante, dato che la serie è più ampia o meno a seconda dei cicli di attività solare. Quindi la serie non è stazionaria, di conseguenza si sceglie il modello ARIMA, che sfrutta il metodo delle differenze per renderla tale.

6.1 Definizione generale

Un modello ARIMA(p, d, q) è una generalizzazione dei modelli autoregressivi (AR) e a media mobile (MA), applicata a una serie temporale resa stazionaria mediante differenziazione. Indicando con x_t il processo gaussiano che rappresenta la serie temporale, il modello può essere espresso come:

$$\phi(B)(1-B)^d x_t = \theta(B)\omega_t, \tag{11}$$

dove:

- B è l'operatore di ritardo ($Bx_t = x_{t-1}$);
- p è l'ordine autoregressivo ($AR(p)$);
- d è il grado di differenziazione ($I(d)$);
- q è l'ordine della media mobile ($MA(q)$);
- ω_t è un rumore bianco con media nulla e varianza costante, non necessariamente gaussiano;
- $\phi(B)$ e $\theta(B)$ sono polinomi in B.

Il forma esplicita:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{12}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \tag{13}$$

L'equazione $\phi(B)X_t = \theta(B)\omega_t$ segue un modello $ARMA(p, q)$.

Il fattore $(1-B)^d$ rappresenta la serie integrata di ordine d, ovvero una serie tale che le differenziazioni di ordine d producono un rumore bianco:

$$(1-B)^d = (1-B) \cdot ... \cdot (1-B)x_t = \omega_t.$$

Per $d = 1$ (il valore usato nel codice) significa che:

$$x_t^\star = x_t - x_{t-1} = \omega_t.$$

6.2 Ipotesi e limiti

Il modello ARIMA si basa su alcune ipotesi fondamentali:

- stazionarietà (dopo differenziazione);
- linearità delle dipendenze temporali;

Nel contesto dei dati SWICS/ACE, tali ipotesi risultano solo parzialmente soddisfatte. L'energia del plasma è influenzata da processi fisici altamente non lineari (shock interplanetari, espulsioni di massa coronale, variazioni del vento solare), che rendono difficile una modellizzazione globale con parametri fissi nel tempo.

7 Fallimento dell'approccio ARIMA classico

L'applicazione di un ARIMA classico prevede la stima dei parametri (p, d, q) su un intervallo temporale fisso di addestramento e il loro utilizzo per l'intera fase di previsione. Nei dati analizzati questo approccio ha mostrato scarso successo, in quanto:

- i parametri ottimali variano nel tempo;
- la dinamica del plasma cambia a seconda del regime fisico osservato;

- gli errori di previsione crescono rapidamente allontanandosi dal periodo di training.

Il modello, quando applicato ha mostrato tendenza ad attestarsi sulla media generale, non descrivendo le variazioni di energia, come si osserva dal grafico.

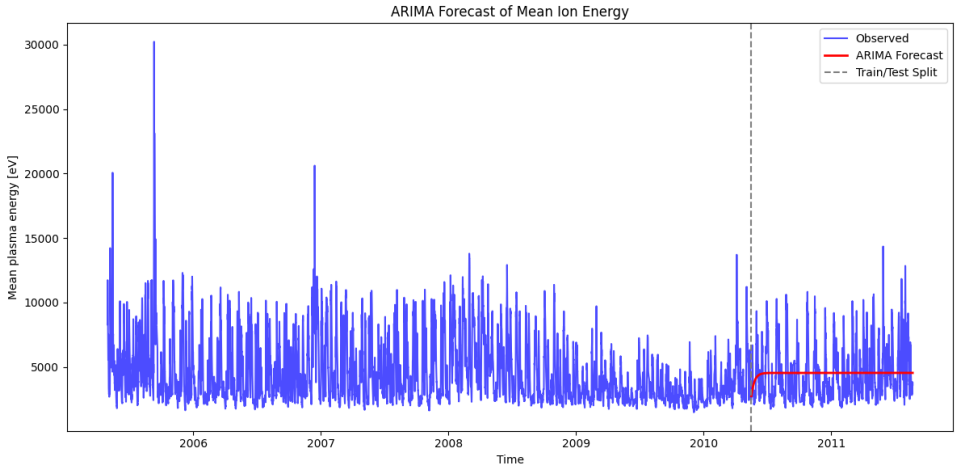

Figure 11: ARIMA con parametri $(2, 0, 2)$

Questi aspetti motivano l'adozione di un approccio più flessibile.

8 Approccio Rolling ARIMA

8.1 Descrizione del metodo

Nel modello Rolling ARIMA, il modello ARIMA viene riaddestrato iterativamente su una finestra temporale mobile (*rolling window*) di ampiezza fissata. Per ogni passo temporale:

1. si seleziona una finestra di dati recenti;
2. si stima un nuovo modello ARIMA sulla finestra;
3. si effettua la previsione a breve termine;
4. la finestra viene fatta scorrere in avanti nel tempo, aggiornandola con i nuovi dati osservati.

Questo approccio consente al modello di adattarsi gradualmente ai cambiamenti strutturali della serie temporale, risultando particolarmente adatto a sistemi fisici non stazionari come il vento solare.

8.2 Metriche di valutazione

Le prestazioni del modello Rolling ARIMA sono state valutate mediante le seguenti metriche:

- **RMSE** (Root Mean Square Error):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}i)^2}; \tag{14}$$

- **MAE** (Mean Absolute Error):

$$\text{MAE} = \frac{1}{N}\sum i = 1^{N}|y_i - \hat{y}_i|; \tag{15}$$

- **Errore relativo percentuale**:

$$\mathrm{RE} = \frac{\mathrm{MAE}}{\langle y \rangle} \times 100. \tag{16}$$

8.3 Ricerca degli iperparametri

Per trovare il valore di d che renda la serie stazionaria, si sono usati i test ADF (Augmented Dickey–Fuller) e KPSS (due test di ipotesi sulla stazionarietà di una serie), che hanno fornito i seguenti risultati:

Table 2: Risultati dei test di stazionarietà ADF e KPSS

Serie	Test	Statistica	p-value	Conclusione
Serie originale	ADF	-16.6304	$< 1e-04$	Rifiuto H_0 (stazionaria)
	KPSS	1.9896	0.0100†	Rifiuto H_0 (non stazionaria)
Serie differenziata ($d = 1$)	ADF	-29.1063	$< 1e-04$	Rifiuto H_0 (stazionaria)
	KPSS	0.0016	0.1000‡	Non rifiuto H_0 (stazionaria)

† Il warning di python indica che il p-value reale è inferiore a 0.01.
‡ Il warning di python indica che il p-value reale è superiore a 0.10.

Di conseguenza, imponiamo $d = 1$ negli iperparametri della gridsearch.

Per restringere il campo di ricerca di p e q, si è valutato il plot di ACF e PACF per la serie differenziata.

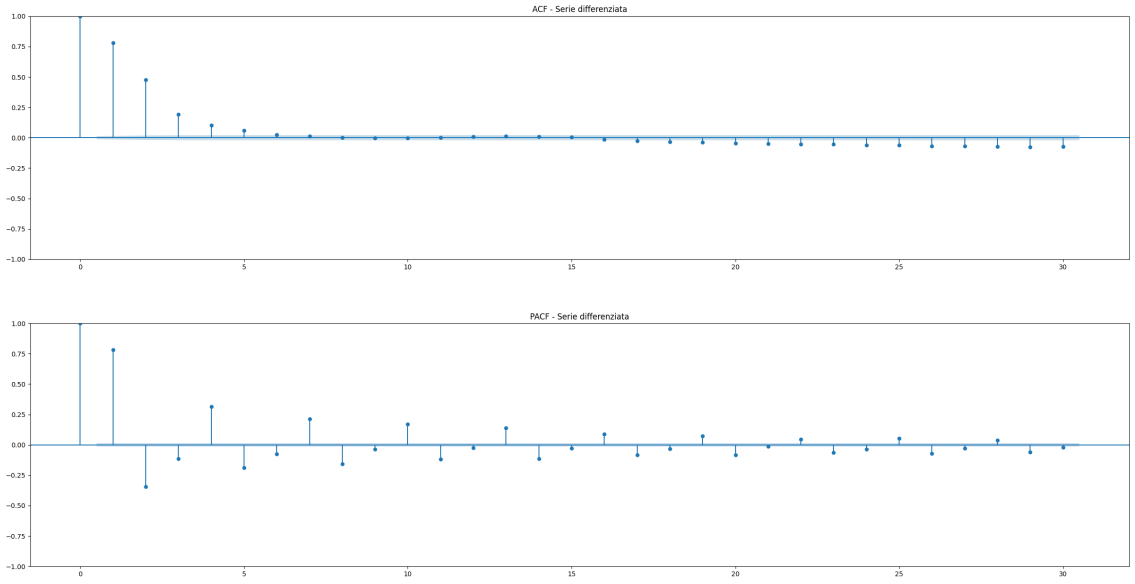

Figure 12: ACF e PACF della serie differenziata.

Il PACF suggerisce 1 o 2 come valori ottimali di p, mentre non da un chiaro responso per q, quindi nella gridsearch si cerca nell'intervallo $[0, 3]$. Qualora il valore ottimale risultante di q fosse 3, si aggiornerà la gridsearch per coprire anche il valore 4.

La gridsearch viene applicata solo sul primo 30% del dataset per ridurre il costo computazionale. La metrica da minimizzare è l'AIC (Akaike Information Criterion):

$$AIC = -2\log(L) + 2k,$$

Dove L è la massima likelyhood del modello e k è il numero di parametri stimati. La tupla con il valore AIC minimo tra quelle indagate è:

$$(p, d, q) = (1, 1, 2).$$

9 Conclusioni su ARIMA

Sono state provate tre combinazioni per (p, d, q): $(1, 0, 1)$ e $(2, 0, 1)$, che hanno fornito i seguenti risultati:

Table 3: Risultati del modello ARIMA.

Parametri	**RMSE (eV)**	**MAE (eV)**	**RE (%)**
$(1, 0, 1)$	508.10	293.99	6.07
$(2, 1, 2)$	423.29	227.10	4.66
$(1, 1, 2)$	423.25	224.31	4.60

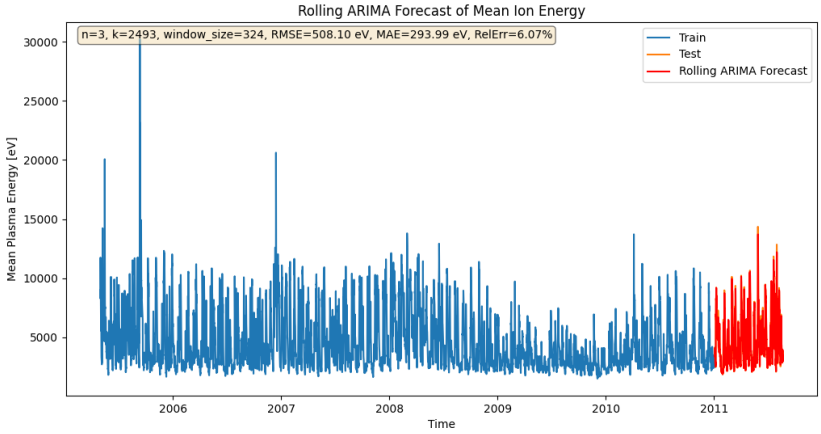

Figure 13: Rolling ARIMA con parametri $(1, 0, 1)$

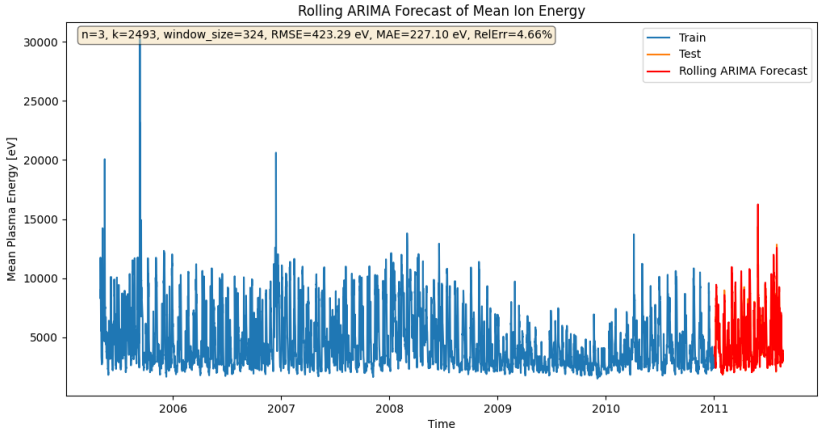

Figure 14: Rolling ARIMA con parametri $(2, 1, 2)$

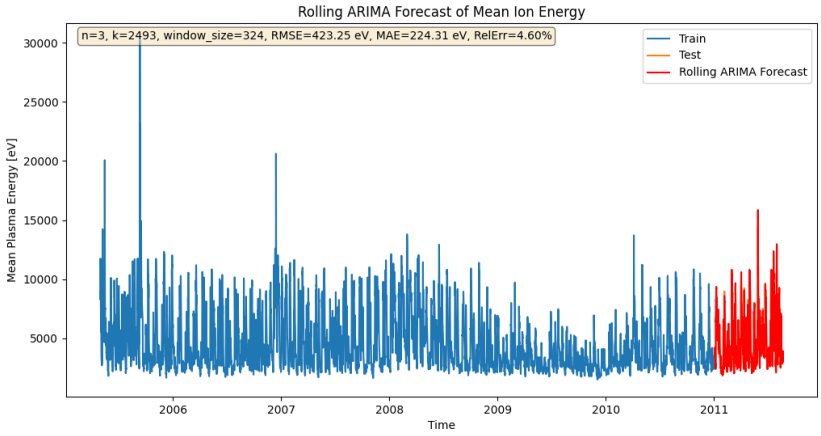

Figure 15: Rolling ARIMA con parametri $(1, 1, 2)$

Il valore di RMSE, superiore al MAE, indica la presenza di errori di previsione occasionalmente elevati, probabilmente associati a eventi fisici improvvisi come shock o transizioni di regime del vento solare. Tuttavia, il MAE e l'errore relativo mostrano che, in media, la previsione a breve termine rimane entro pochi punti percentuali rispetto all'energia tipica del plasma osservato.

9.1 Discussione fisica

Un errore relativo del $\sim 4.6\%$ è un ottimo risultato considerando la complessità del sistema fisico studiato. Il miglioramento rispetto all'ARIMA classico evidenzia come l'adattamento locale dei parametri sia cruciale per catturare la variabilità temporale dell'energia del plasma.

Tuttavia, il modello rimane limitato dalla sua natura lineare. Fenomeni fortemente non lineari non possono essere completamente descritti da ARIMA, suggerendo possibili estensioni future mediante modelli ibridi o approcci basati su reti neurali.

References

[1] CDAWeb. AC H_3 SWICS H^+ Measurements. URL: `https://cdaweb.gsfc.nasa.gov/cgi-bin/eval2.cgi?dataset=AC_H3_SWI&index=sp_phys`.

[2] hmmlearn Documentation. hmmlearn.hmm.gaussianhmm api documentation. URL: `https://hmmlearn.readthedocs.io/en/latest/api.html#hmmlearn.hmm.GaussianHMM`.

[3] Wikipedia contributors. Autoregressive integrated moving average. URL: `https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average#Choosing_the_order`.

[4] Wikipedia contributors. Baum–Welch algorithm. URL: `https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm`.

[5] Wikipedia contributors. Hidden Markov model. URL: `https://en.wikipedia.org/wiki/Hidden_Markov_model`.

[6] Wikipedia contributors. Kolmogorov–Smirnov test. URL: `https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test`.

[7] Wikipedia contributors. Test ADF. URL: `https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test`.

[8] Wikipedia contributors. Test KPSS. URL: https://it.wikipedia.org/wiki/Test_KPSS.