Analisi temporale del vento solare

applicazione di modelli statistici (HMM e ARIMA) alla previsione del vento solare

Alberto Prino

13 Febbraio 2026

Politecnico
di Torino

Presentation Overview

Introduzione al dataset

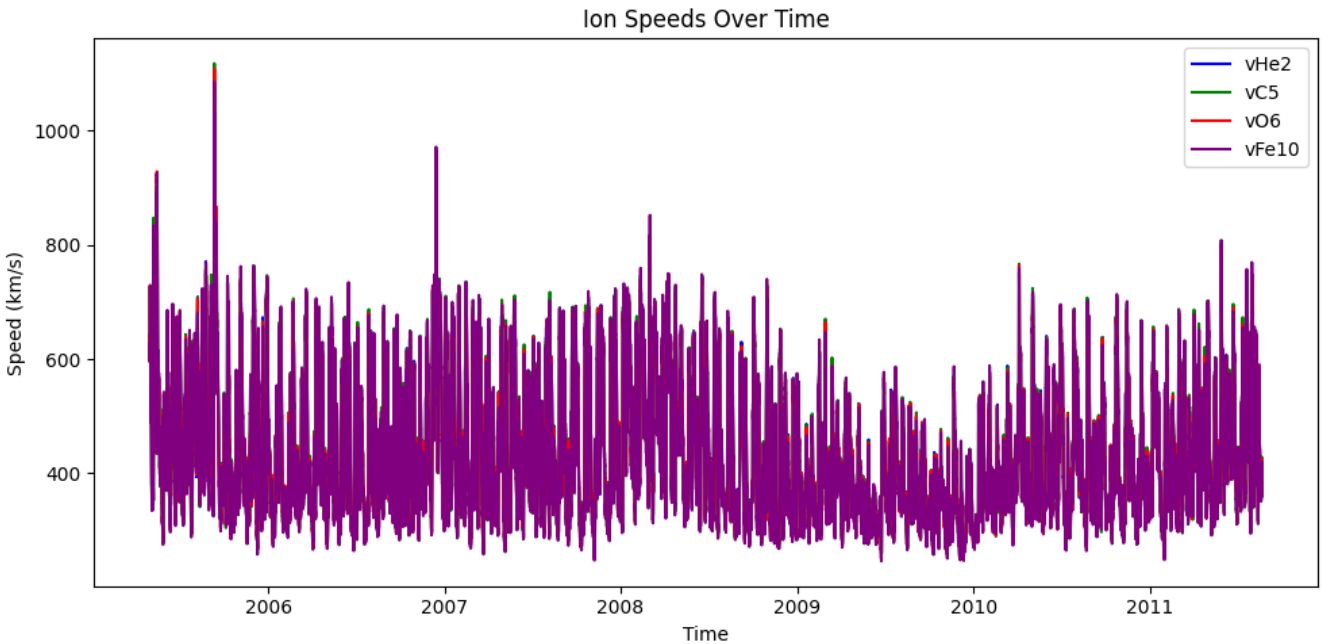
Il dataset utilizzato in questo lavoro è derivato dalle osservazioni dello strumento *Solar Wind Ion Composition Spectrometer* (SWICS), a bordo della missione *Advanced Composition Explorer* (ACE) della NASA, che si occupa di misurare dati riguardanti il plasma del vento solare.
I dati sono campionati con una cadenza temporale di due ore e descrivono le proprietà di diverse specie ioniche: elio, carbonio, ossigeno e ferro.
Il dataset contiene 24924 entrate e 13 features.

Features del dataset

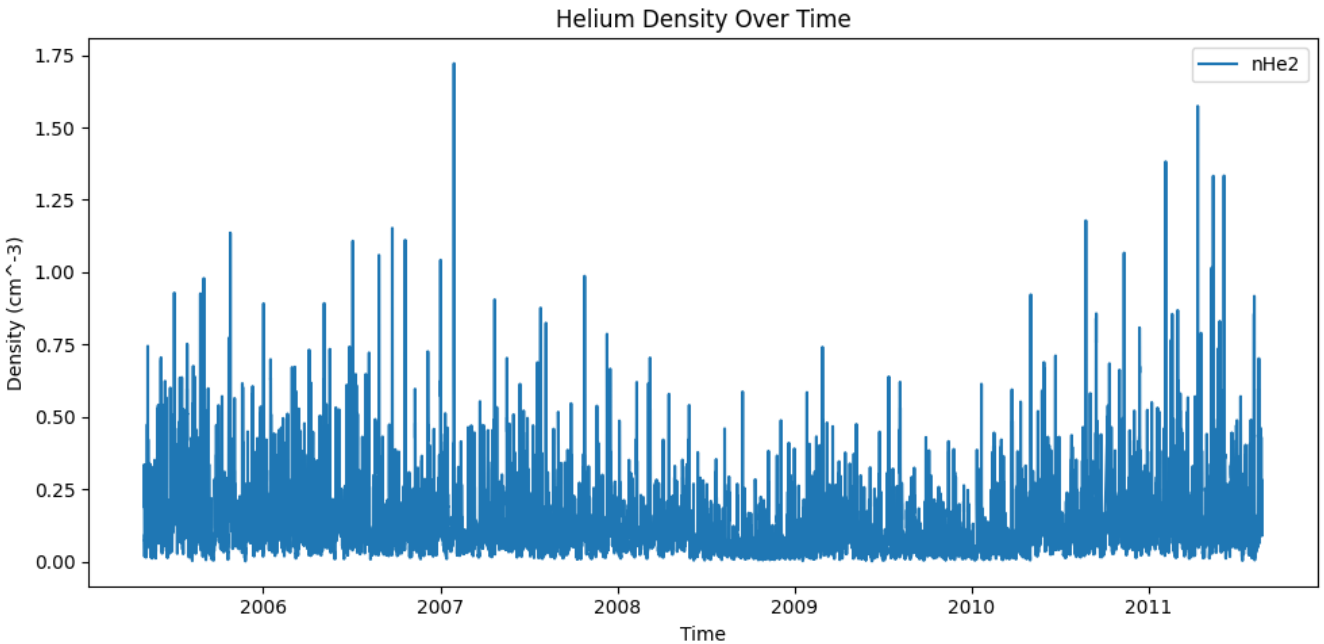
Tabella: Variabili del dataset SWICS/ACE e relative unità fisiche.

Variabile	**Descrizione fisica**	**Unità**
`time`	Tempo di osservazione	–
`nHe2`	Densità numerica degli ioni He^{2+}	cm^{-3}
`vHe2`	Velocità di flusso degli ioni He^{2+}	$\mathrm{km\,s}^{-1}$
`vC5`	Velocità di flusso degli ioni C^{5+}	$\mathrm{km\,s}^{-1}$
`vO6`	Velocità di flusso degli ioni O^{6+}	$\mathrm{km\,s}^{-1}$
`vFe10`	Velocità di flusso degli ioni Fe^{10+}	$\mathrm{km\,s}^{-1}$
`HetoO`	Rapporto di abbondanza He/O	–
`CtoO`	Rapporto di abbondanza C/O	–
`FetoO`	Rapporto di abbondanza Fe/O	–

Andamento delle velocità nel tempo



Andamento della densità di elio nel tempo



Analisi della composizione

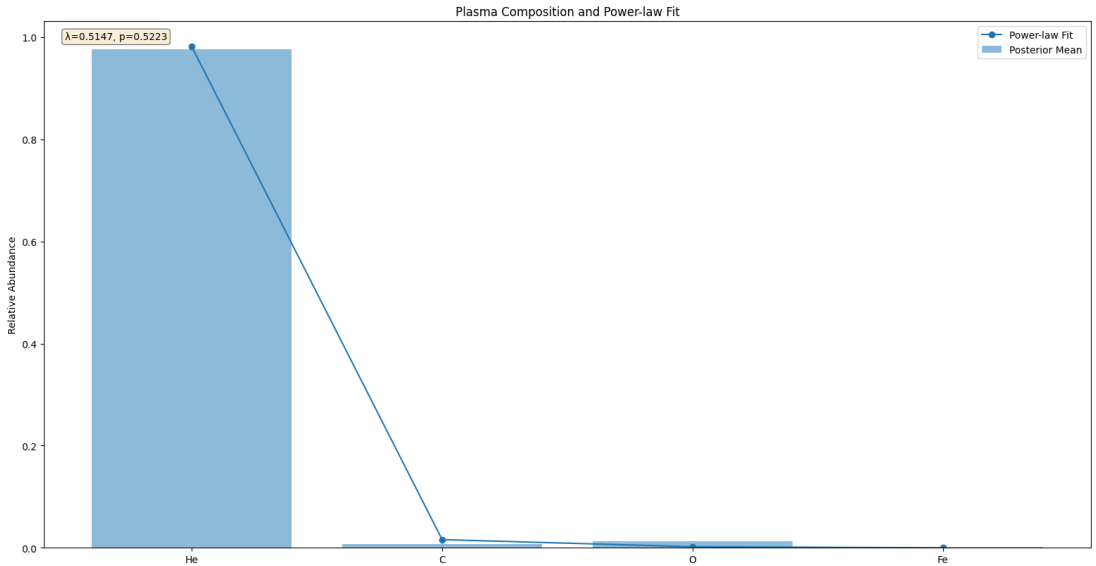


Figura: Adattamento esponenziale della composizione in base alla massa.

Test goodness-of-fit

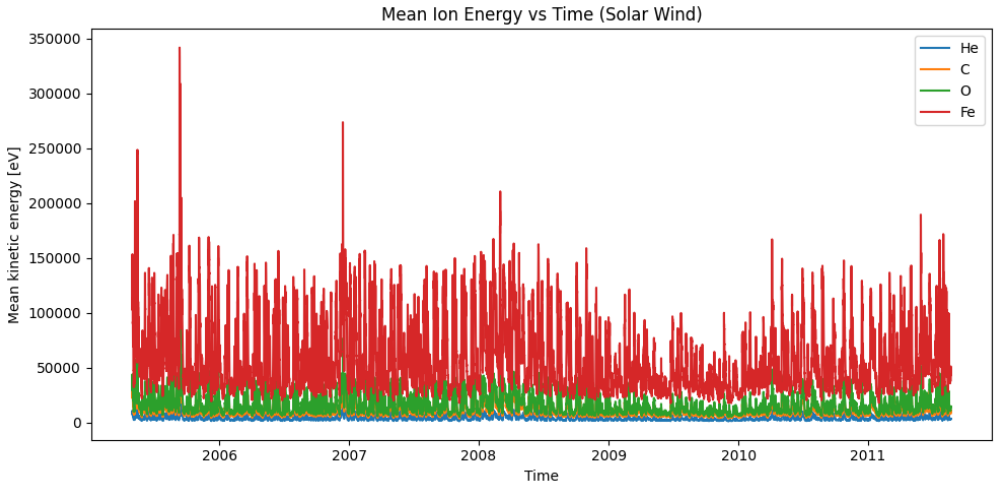
L'algoritmo adottato per verificare l'ipotesi di una dipendenza esponenziale della composizione del plasma dalla massa atomica delle specie ioniche può essere riassunto nei seguenti passaggi:

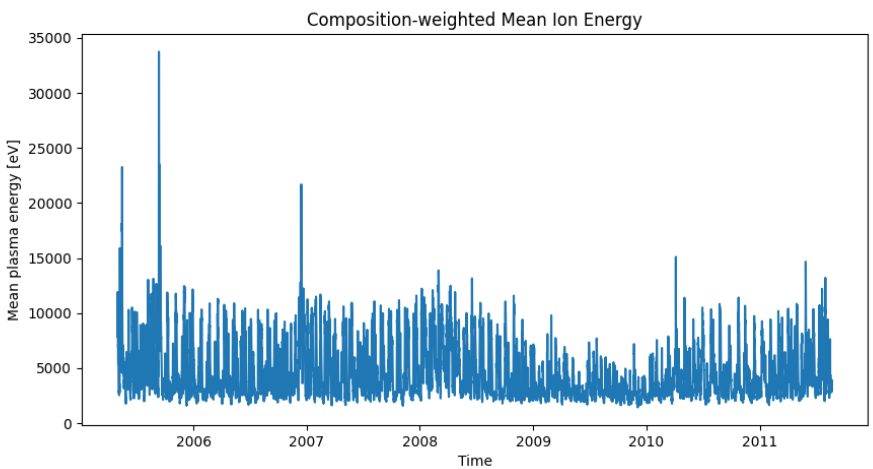
1. Costruzione della distribuzione a posteriori delle abbondanze relative $\mathbf{p}$ mediante un modello bayesiano Dirichlet, a partire dai conteggi osservati sommati sull'intero dataset.
2. Calcolo del valore medio a posteriori $\langle \mathbf{p} \rangle$ delle abbondanze relative.
3. Definizione 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

Andamento energia media nel tempo





Teoria degli HMM

Un HMM è definito da:

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

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

- una distribuzione di emissione
  $p(x_t \mid z_t = k) = \mathcal{N}(x_t \mid \mu_k, \sigma_k)$ (HMM gaussiano).

Algoritmo Baum-Welch

- Come prior sulle probabilità di transizione si assume una Dirichlet: $\pi_{z_0,k} \sim Dir(\alpha_0)$, con $\alpha_0 = (1,1)$.
- Prior sulle righe della matrice di transizione: $\pi_k \sim Dir(\alpha),\ \alpha = (1,1)$.
- Prior normale sulla media delle distribuzioni di emissione: $\mu_k = \mathcal{N}(0, \frac{1}{w}\sigma_k^2)$.
- Prior sulle varianze: $\sigma_k^2 = IG(1, 0.01)$ (Gamma inversa).

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:

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

Passo E

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:

$$\alpha_1(i) = \pi_i \cdot p(x_1|z_1 = i) = \pi_i \cdot \mathcal{N}(x_1; \mu_i, \Sigma_i).$$

Ricorsione:

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

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:

$\beta_T(i) = 1.$

Ricorsione:

$$\beta_t(i) = \sum_{j=1}^{N} \pi_{ij} \cdot \mathcal{N}(x_{t+1}; \mu_j, \Sigma_j) \cdot \beta_{t+1}(j).$$

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:

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

Passo M

I parametri del modello vengono aggiornati per massimizzare la likelyhood:

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

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

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

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

Algoritmo Viterbi

La previsione sugli stati z_t viene ottenuta con l'algoritmo di Viterbi, che calcola la singola sequenza più probabile di stati latenti nel senso della massima probabilità a posteriori (MAP):

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

Applicazione al dataset

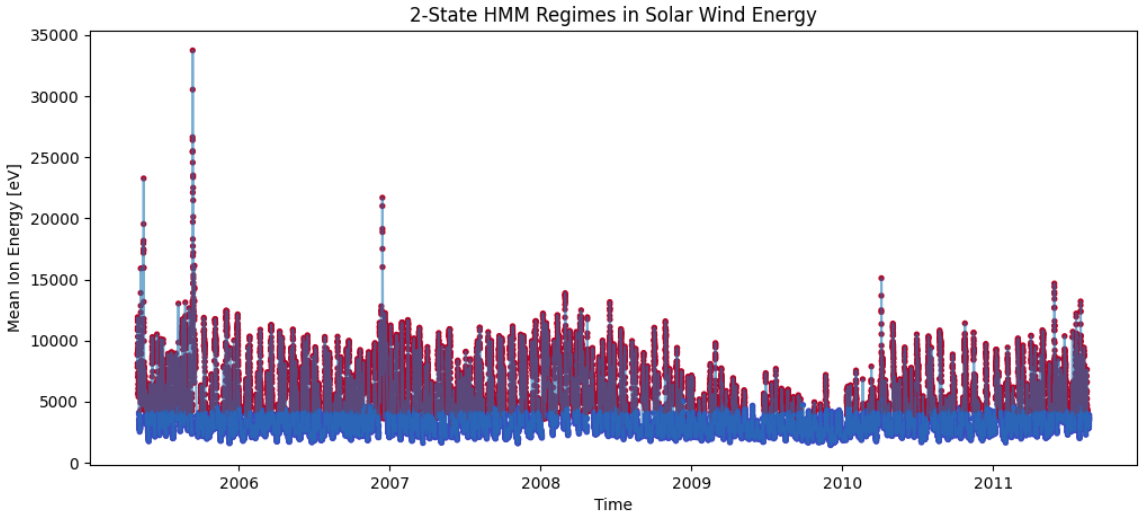
Il modello è stato prima addestrato su tutto il dataset per osservare le energie medie associate agli stati, con $K = 2$.

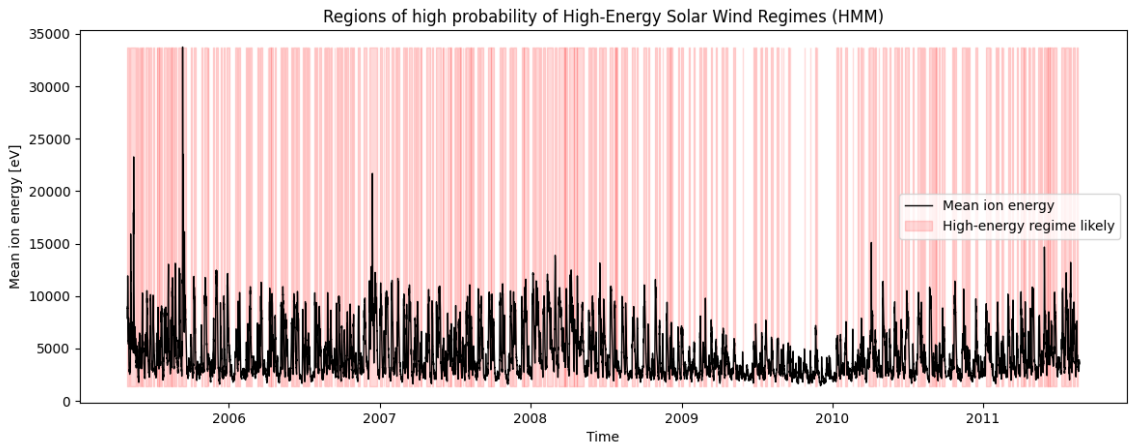
Stato	μ [eV]
0	2918.8
1	6780.2

Matrice di transizione stimata

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

Visualizzazione della classificazione





Testing HMM

Il modello è stato allenato sul primo 80% dei dati e testato sul restante 20%, coi seguenti risultati:

- Durate medie:

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

- Predictive log-score medio: -9.65.
- Errore assoluto medio (MAE):

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

Visualizzazione delle predizioni

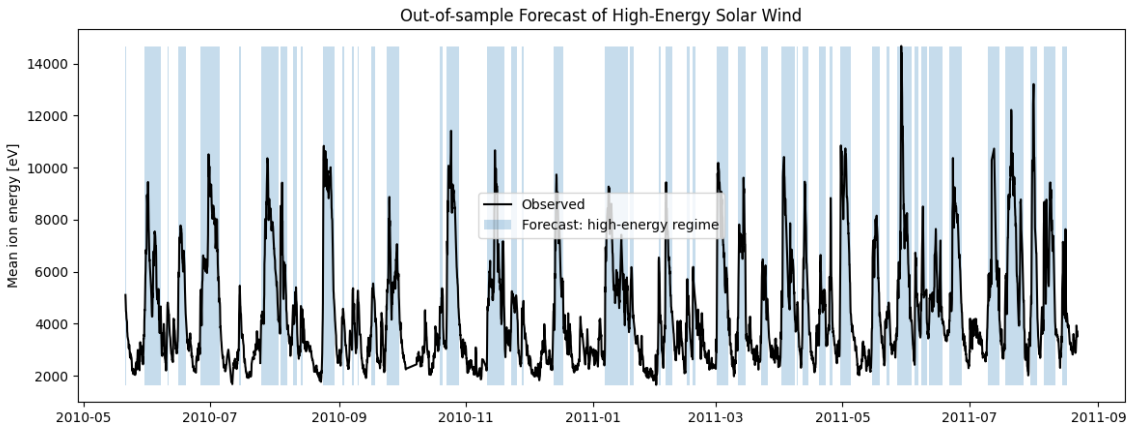


Figura: Previsione di regime sul test split del dataset.

Definizione del modello

Un modello ARIMA(p, d, q) applicato al processo gaussiano x_t può essere espresso come:

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

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.

Fallimento approccio ARIMA classico

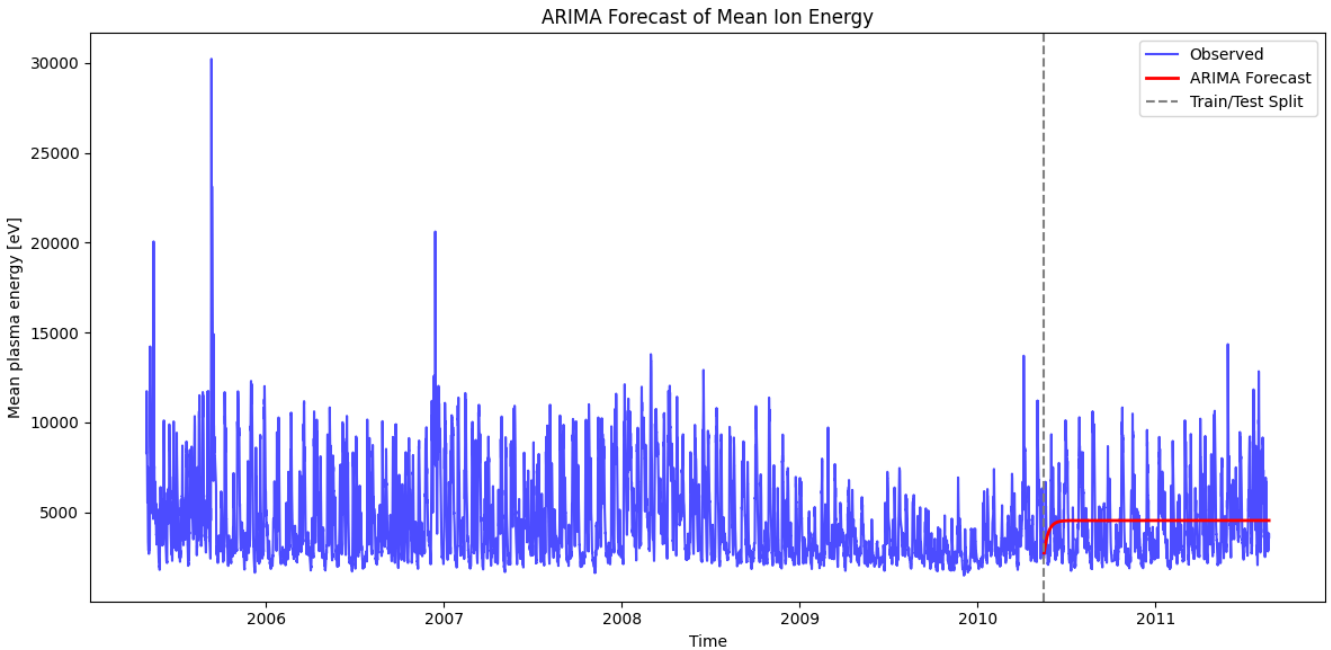


Figura: ARIMA con parametri $(2, 0, 2)$

Ricerca di d

Viene fatta attraverso i test ADF (H_0: serie non stazionaria) e KPSS (H_0: serie stazionaria).

Tabella: Risultati dei test di stazionarietà ADF e KPSS

Serie	Test	p-value	Conclusione
$d = 0$	ADF	$< 1e-04$	Rifiuto H_0
	KPSS	0.0100	Rifiuto H_0
$d = 1$	ADF	$< 1e-04$	Rifiuto H_0
	KPSS	0.1000	Non rifiuto H_0

Ricerca di p e q

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

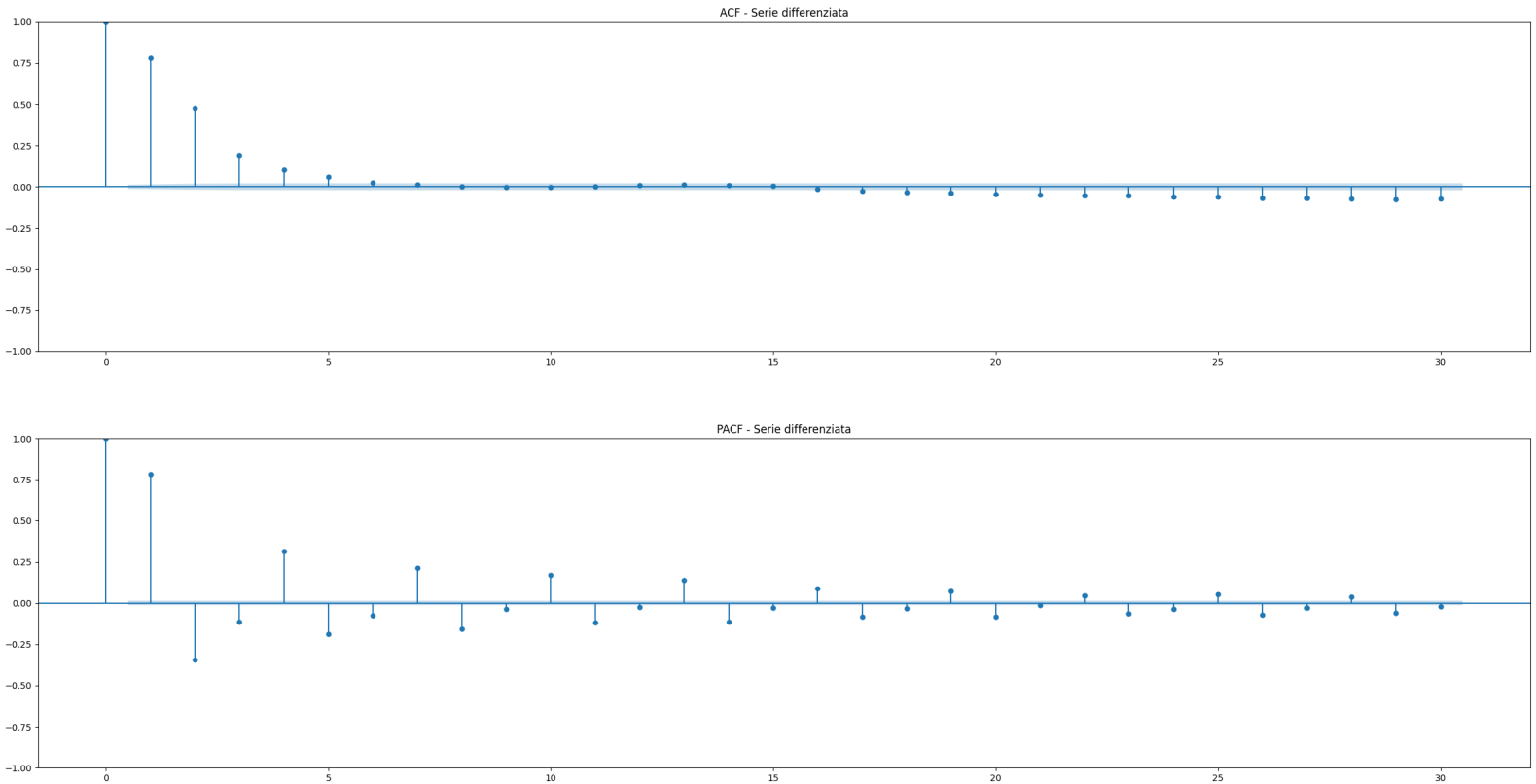


Figura: ACF e PACF della serie differenziata.

Forecasting

Dalla gridsearch emerge che la combinazione migliore è $(1,1,2)$, che fornisce i seguenti risultati:

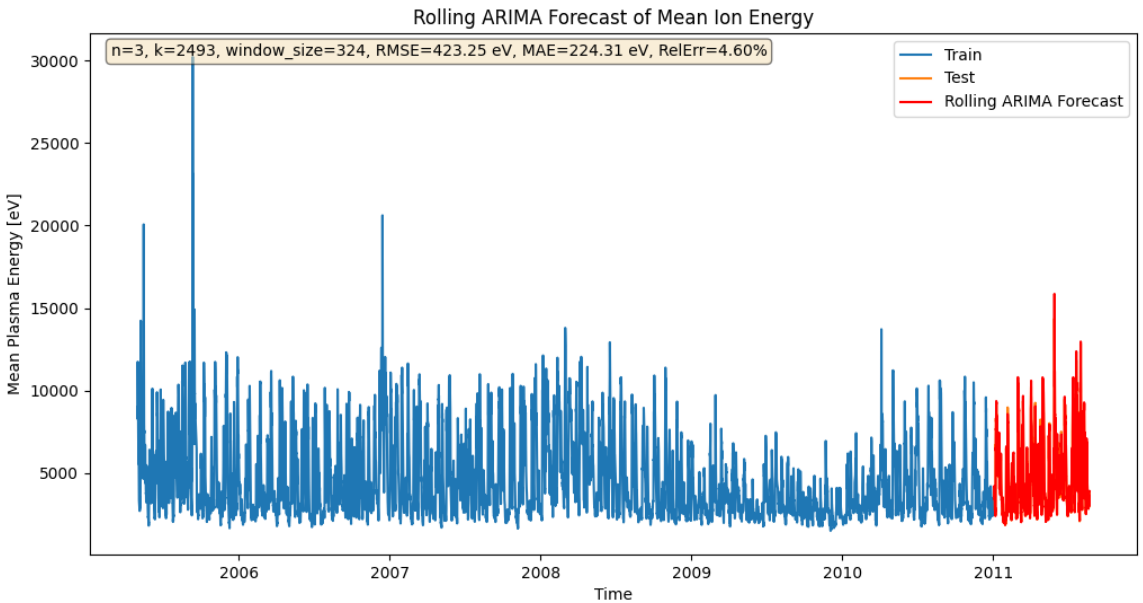


Figura: Rolling ARIMA con parametri $(1,1,2)$

Fine