ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΣΤΟΧΑΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Εργαστηριακή Αναφορά

Μάρκου Δήμητρα Π2019170

Κατά τη διάρκεια των εργαστηριακών μαθημάτων είδαμε τους παρακάτω όρους καθώς και την εφαρμογή τους σε διαφορετικά σύνολα δεδομένων.

Μέση Τιμή - Mean

Η μέση τιμή, είναι ένα μέτρο κεντρικής τάσης και αντιπροσωπεύει το κεντρικό σημείο των δεδομένων.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Διακύμανση - Variance

Η διακύμανση μετρά τη διασπορά ενός συνόλου σημείων δεδομένων γύρω από τη μέση τιμή τους. Η υψηλή διακύμανση υποδηλώνει ότι τα σημεία δεδομένων απέχουν πολύ από το μέσο όρο, ενώ η χαμηλή διακύμανση υποδηλώνει ότι βρίσκονται κοντά στο μέσο όρο.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Τυπική απόκλιση - Standard Deviation

Η τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης και παρέχει ένα μέτρο της μέσης απόστασης κάθε σημείου δεδομένων από τον μέσο όρο. Εκφράζεται στις ίδιες μονάδες με τα δεδομένα, καθιστώντας την πιο ερμηνεύσιμη από τη διακύμανση.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Συνδιακύμανση - Covariance

Η συνδιακύμανση δείχνει την κατεύθυνση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Εάν η συνδιακύμανση είναι θετική, καθώς η μία μεταβλητή αυξάνεται, η άλλη τείνει να αυξάνεται επίσης. Εάν είναι αρνητική, η μία μεταβλητή τείνει να μειώνεται καθώς η άλλη αυξάνεται. Το μέγεθος της συνδιακύμανσης εξαρτάται από τις μονάδες των μεταβλητών, γεγονός που καθιστά δύσκολη την άμεση ερμηνεία της.

$$\operatorname{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Συσχέτιση - Correlation

Η συσχέτιση είναι ένα τυποποιημένο μέτρο της ισχύος και της κατεύθυνσης της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Κυμαίνεται από -1 έως 1, όπου το 1 υποδηλώνει τέλεια θετική γραμμική σχέση, το -1 υποδηλώνει τέλεια αρνητική γραμμική σχέση και το 0 υποδηλώνει μηδενική γραμμική σχέση. Σε αντίθεση με τη συνδιακύμανση, η συσχέτιση είναι χωρίς διαστάσεις.

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Συνάρτηση διακύμανσης - Variance Function

Μια συνάρτηση διακύμανσης περιγράφει πώς μεταβάλλεται η διακύμανση μιας μεταβλητής ή μιας διαδικασίας με την πάροδο του χρόνου ή σε σχέση με μια άλλη μεταβλητή.

$$\gamma(T) = \frac{\sigma_T^2}{\sigma_0^2}$$

Κινητός μέσος όρος - Moving Average

Ο κινητός μέσος όρος χρησιμοποιείται για την εξομάλυνση των βραχυπρόθεσμων διακυμάνσεων και την ανάδειξη των μακροπρόθεσμων τάσεων σε δεδομένα χρονοσειρών.

$$MA_k = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}$$

Αυτοσυνδιακύμανση - Autocovariance

Η αυτοσυνδιακύμανση μετρά τη συνδιακύμανση μιας χρονοσειράς με μια μεταγενέστερη έκδοση του εαυτού της. Βοηθά στον εντοπισμό της γραμμικής εξάρτησης των τρεχουσών τιμών από τις παρελθοντικές τιμές. Η υψηλή αυτοσυνδιακύμανση υποδηλώνει ισχυρή ομοιότητα μεταξύ της χρονοσειράς και των μεταγενέστερων τιμών της.

$$\operatorname{cov}[X_{t_1}, X_{t_2}] = \mathbb{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})] = \mathbb{E}[X_{t_1} X_{t_2}] - \mu_{t_1}\mu_{t_2}$$

Αυτοσυσχέτιση - Autocorrelation

Η αυτοσυσχέτιση μετρά τη συσχέτιση μιας χρονοσειράς με τις δικές της παρελθοντικές και μελλοντικές τιμές. Βοηθά στον εντοπισμό επαναλαμβανόμενων μοτίβων ή τάσεων σε δεδομένα χρονοσειρών. Η αυτοσυσχέτιση κυμαίνεται από -1 έως 1, παρόμοια με τη συσχέτιση.

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

Συνέλιξη - Convolution

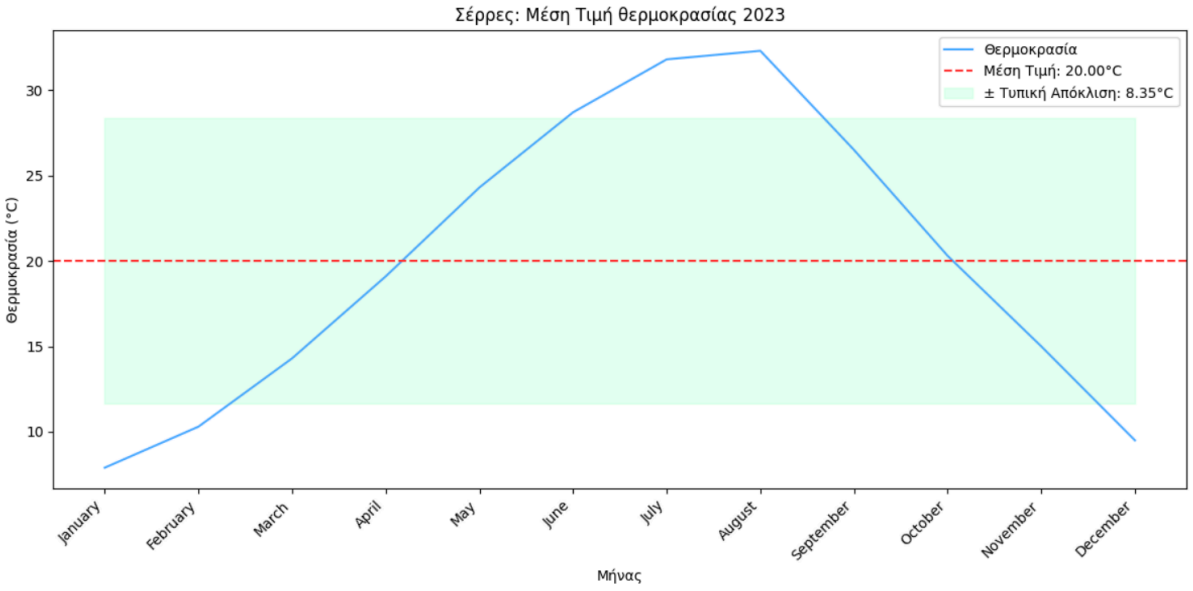
Η συνέλιξη χρησιμοποιείται για την ανάλυση της επίδρασης μιας συνάρτησης σε μια άλλη, η οποία εφαρμόζεται συχνά στην ανάλυση χρονοσειρών και στη θεωρία πιθανοτήτων. Συνδυάζει δύο συναρτήσεις για την παραγωγή μιας τρίτης συνάρτησης που αναπαριστά τον τρόπο με τον οποίο η μορφή της μιας τροποποιείται από την άλλη.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau)\, d\tau$$

Εφαρμογή σε Σύνολο Δεδομένων: serres-temp.csv

Για το σύνολο δεδομένων serres-temp.csv, το οποίο περιέχει τη μέση τιμή θερμοκρασίας για κάθε μήνα του έτους 2023, της πόλης των Σερρών, υπολόγισα τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση.

Μέση Τιμή (Mean): 20.0
Διακύμανση (Variance): 69.675
Τυπική Απόκλιση (Standard Deviation): 8.34715520402011



Ο παρακάτω κώδικας με τη χρήση των βιβλιοθηκών, numpy, pandas και matplotlib της γλώσσας προγραμματισμού Python, υπολογίζει τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση και δημιουργεί γράφημα το οποίο τα αντικατοπτρίζει. Στο γράφημα η μπλε συνεχόμενη γραμμή δείχνει τις θερμοκρασίες όλου του συνόλου δεδομένων, η κόκκινη διακεκομμένη γραμμή δείχνει τη μέση τιμή και το αχνό πράσινο πλαίσιο δείχνει τη ± τυπική απόκλιση γύρω από τη μέση τιμή.

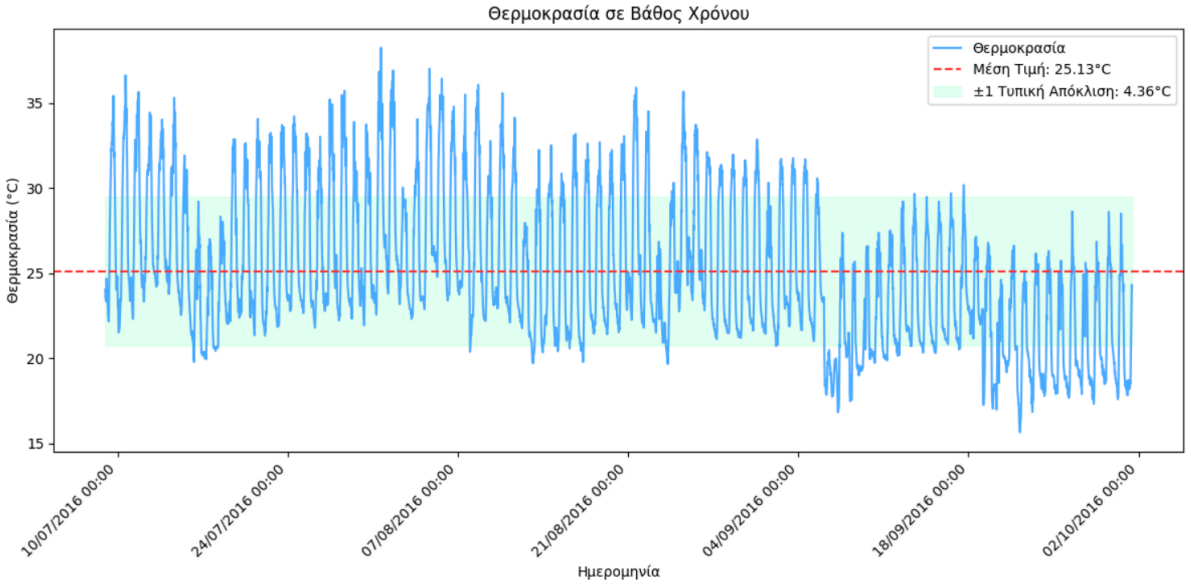
Κώδικας

```
# Σύνολο Δεδομένων - serres-temp.csv

file_path = '/kaggle/input/serres-temperatures-2023/serres-temp.csv'
serresTemps = pd.read_csv(file_path)
temperatures = serresTemps['Temperature (°C)']

# Υπολογισμός Μέσης τιμής, Διακύμανσης και Τυπικής απόκλισης
mean_temp = np.mean(temperatures)
variance_temp = np.var(temperatures)
std_dev_temp = np.std(temperatures)

print("Δεδομένα μέσης τιμής θερμοκρασίας για κάθε μήνα του έτους 2023 της πόλης των Σερρών,")
print("Μέση Τιμή (Mean):", mean_temp)
print("Διακύμανση (Variance):", variance_temp)
print("Τυπική Απόκλιση (Standard Deviation):", std_dev_temp)

# Γράφημα
fig, ax = plt.subplots(figsize=(12, 6))

ax.plot(serresTemps['Month'], temperatures, label='Θερμοκρασία',
color='#4AABFF')
ax.axhline(mean_temp, color='#FF3333', linestyle='--', label=f'Μέση Τιμή:
{mean_temp:.2f}°C')
ax.fill_between(serresTemps['Month'], mean_temp - std_dev_temp, mean_temp +
std_dev_temp, color='#00FF80', alpha=0.1, label=f'± Τυπική Απόκλιση:
{std_dev_temp:.2f}°C')
ax.set_title('Σέρρες: Μέση Τιμή θερμοκρασίας 2023')
ax.set_xlabel('Μήνας')
plt.setp(ax.xaxis.get_majorticklabels(), rotation=45, ha='right')
ax.set_ylabel('Θερμοκρασία (°C)')
ax.legend()

plt.tight_layout()
plt.show()
```

Εφαρμογή σε Σύνολο Δεδομένων: temp-dataset.csv

Για το σύνολο δεδομένων temp-dataset.csv, το οποίο περιέχει δεδομένα από την περιοχή του Αγίου Μάρκου Κέρκυρας, συλλεγμένα σε περίοδο περίπου 3 μηνών, με συχνότητα δειγματοληψίας κάθε 15 λεπτά, υπολόγισα τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση.

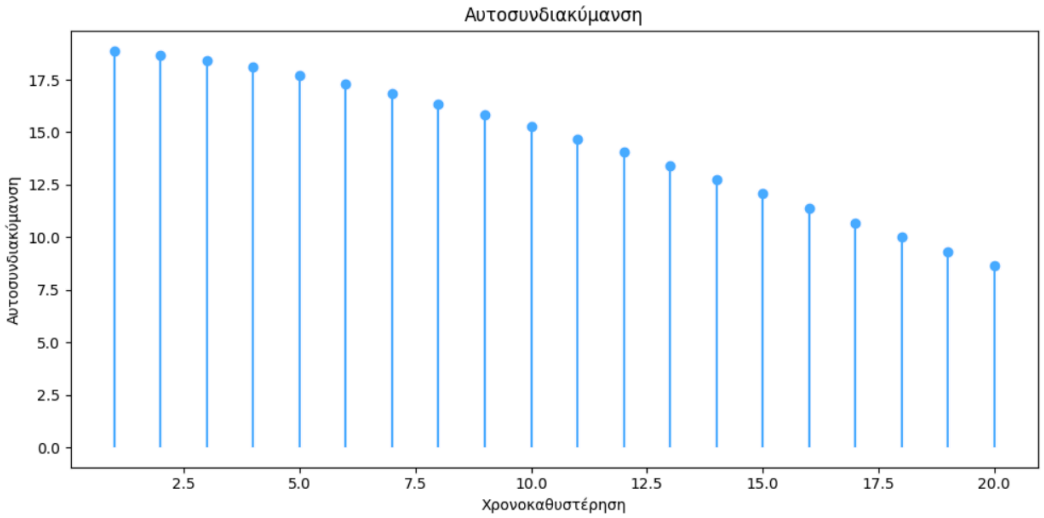
Μέση Τιμή (Mean): 25.131947342458997
Διακύμανση (Variance): 18.993359515850425
Τυπική Απόκλιση (Standard Deviation): 4.358137161202069



Η μέση θερμοκρασία κατά τη διάρκεια αυτής της περιόδου ήταν περίπου 25,13°C. Η υπολογιζόμενη διακύμανση των 18,99°C² και η τυπική απόκλιση των 4,36°C αντικατοπτρίζουν τη μεταβλητότητα των δεδομένων θερμοκρασίας. Στο γράφημα η μπλε συνεχόμενη γραμμή δείχνει τις θερμοκρασίες όλου του συνόλου δεδομένων, η κόκκινη διακεκομμένη γραμμή δείχνει τη μέση τιμή και το αχνό πράσινο πλαίσιο δείχνει τη ± τυπική απόκλιση γύρω από τη μέση τιμή.

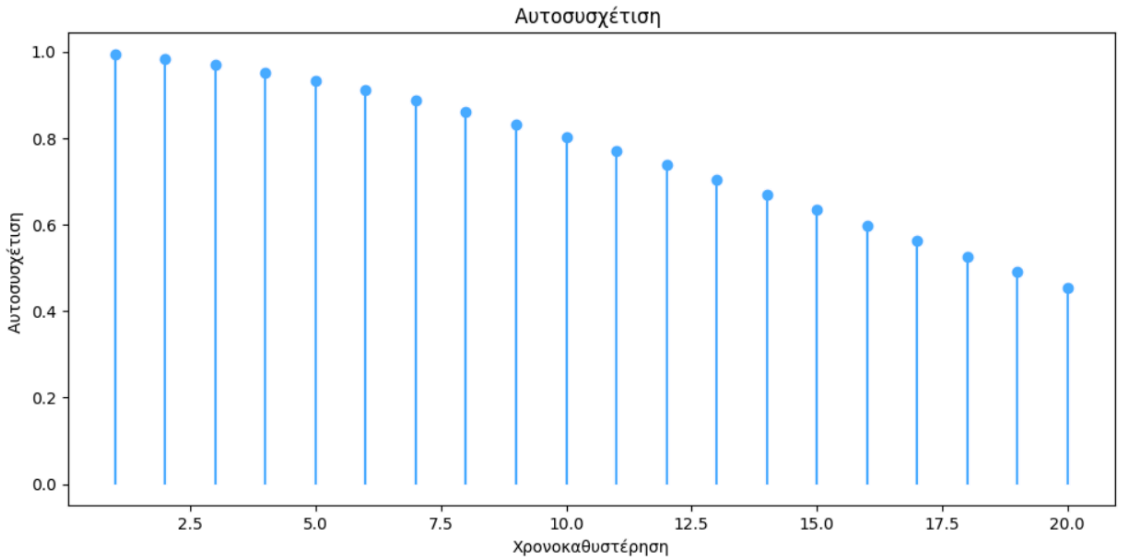
Η συνδιακύμανση μετρά τη γραμμική σχέση μεταξύ δύο διαφορετικών μεταβλητών. Σε αυτή την περίπτωση, έχουμε μόνο μία μεταβλητή ενδιαφέροντος: τη θερμοκρασία. Η αυτοσυνδιακύμανση μετρά τη συνδιακύμανση μιας χρονοσειράς με μια μεταγενέστερη έκδοση του εαυτού της. Για την ανάλυση αυτού του συνόλου δεδομένων, η αυτοσυνδιακύμανση είναι πιο σημαντική από τη συνδιακύμανση.

Αυτοσυνδιακύμανση



Το διάγραμμα αυτοσυνδιακύμανσης απεικονίζει τον βαθμό στον οποίο οι τιμές της θερμοκρασίας σχετίζονται γραμμικά με τις αντίστοιχες τιμές με χρονοκαθυστέρηση σε διάφορα χρονικά διαστήματα, από 1 έως 20. Οι υψηλές τιμές αυτοσυνδιακύμανσης σε μικρότερες χρονοκαθυστερήσεις υποδηλώνουν ισχυρή χρονική εξάρτηση στα δεδομένα θερμοκρασίας, με περιοδικές αιχμές που υποδηλώνουν εποχικό ή επαναλαμβανόμενο μοτίβο.

Αυτοσυσχέτιση



Το διάγραμμα αυτοσυσχέτισης απεικονίζει τη συσχέτιση μεταξύ των τιμών θερμοκρασίας και των μεταγενεστερων εκδόσεών τους, παρέχοντας ένα κανονικοποιημένο μέτρο της γραμμικής σχέσης. Οι χρονοκαθηστερυσεις κυμαίνονται από 1 έως 20 και τιμές αυτοσυσχέτισης κυμαίνονται μεταξύ -1 και 1. Οι τιμές κοντά στο 1 υποδηλώνουν ισχυρή θετική συσχέτιση, ενώ τιμές κοντά στο -1 υποδηλώνουν ισχυρή αρνητική συσχέτιση. Το γράφημα μας

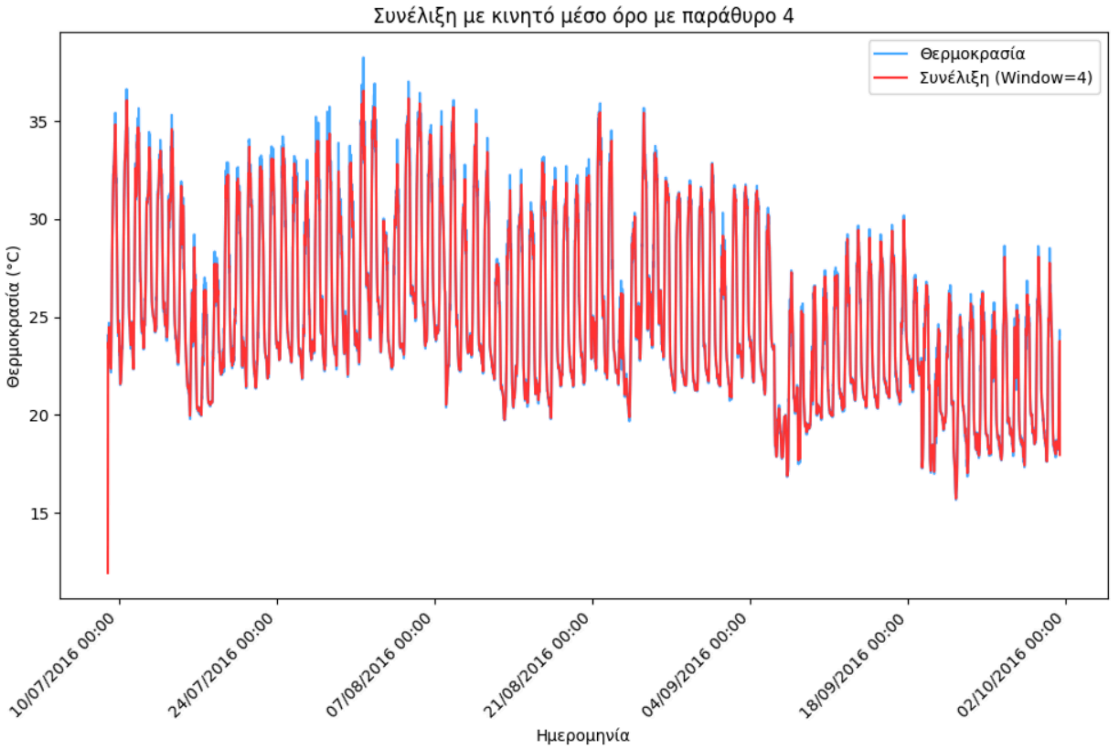
αποκαλύπτει ότι τα δεδομένα θερμοκρασίας έχουν υψηλή αυτοσυσχέτιση σε μικρές υστερήσεις, επιβεβαιώνοντας μια ισχυρή χρονική σχέση.

Συνέλιξη Με Κινητό Μέσο Όρο

Για τον υπολογισμό του κινητού μέσου όρου έχω επιλέξει 3 διαφορετικά μεγέθη παραθύρου. Το σύνολο δεδομένων που έχουμε η δειγματοληψία του είναι κάθε 15 λεπτά, επομένως το παράθυρο 4 υποδηλώνει μια ώρα, το παράθυρο 96 μια ημέρα και το 672 μια εβδομάδα.

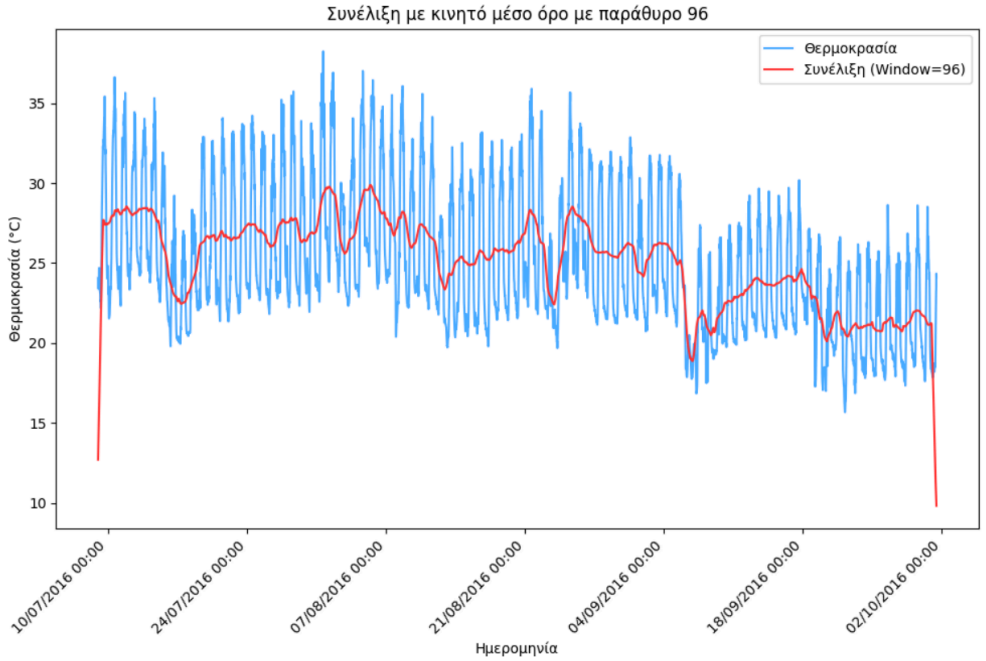
Τα γραφήματα συνέλιξης των δεδομένων θερμοκρασίας με πυρήνα κινητού μέσου όρου διαφορετικών μεγεθών παραθύρου: 4, 96 και 672, απεικονίζουν την επίδραση εξομάλυνσης των διαφορετικών μεγεθών παραθύρου στα δεδομένα θερμοκρασίας.

Συνέλιξη με μέγεθος παραθύρου 4 (1 ώρα)



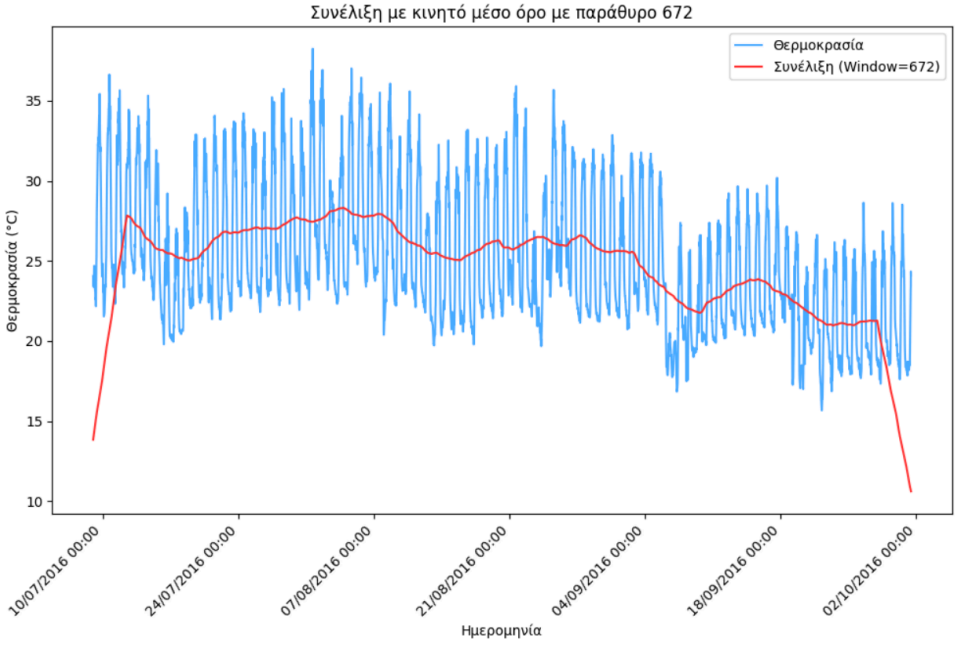
Αυτό το διάγραμμα εξομαλύνει τα δεδομένα θερμοκρασίας σε ένα σύντομο παράθυρο 1 ώρας, αποτυπώνοντας τις γρήγορες αλλαγές και μειώνοντας τον θόρυβο. Η προκύπτουσα γραμμή είναι ελαφρώς εξομαλυμένη, αλλά εξακολουθεί να δείχνει βραχυπρόθεσμες διακυμάνσεις.

Συνέλιξη με μέγεθος παραθύρου 96 (1 ημέρα)



Η χρήση ενός μεγέθους παραθύρου 1 ημέρας παρέχει μια πιο ομαλή καμπύλη που αναδεικνύει τις ημερήσιες τάσεις και μειώνει περαιτέρω τις βραχυπρόθεσμες διακυμάνσεις.

Συνέλιξη με μέγεθος παραθύρου 672 (1 εβδομάδα)



Με μέγεθος παραθύρου 1 εβδομάδας, η γραφική παράσταση δίνει έμφαση στις μακροπρόθεσμες τάσεις, εξομαλύνοντας τα εβδομαδιαία μοτίβα και αποτυπώνοντας με μεγαλύτερη σαφήνεια τη συνολική τάση της θερμοκρασίας.

Κώδικας

```
# Βιβλιοθήκες
import numpy as np
import pandas as pd
from scipy.signal import convolve
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

# Σύνολο Δεδομένων - temp-dataset.csv

file_path = '/kaggle/input/temperatures/temp-dataset-st-analysis.csv'
tempDataset = pd.read_csv(file_path)
tempDataset['Datetime'] = pd.to_datetime(tempDataset['Datetime'],
format='%d/%m/%Y %H:%M', errors='coerce')
tempDataset['Temperature'] = tempDataset['Temperature'].str.replace(',',
'.').astype(float)

# Υπολογισμός Μέσης τιμής, Διακύμανσης και Τυπικής απόκλισης
mean = tempDataset['Temperature'].mean()
variance = tempDataset['Temperature'].var()
standardDeviation = tempDataset['Temperature'].std()

print("Δεδομένα από την περιοχή του Αγίου Μάρκου Κέρκυρας, συλλεγμένα σε περίοδο
περίπου 3 μηνών, με συχνότητα δειγματοληψίας 'τιμή θερμοκρασίας/15 λεπτά'.")
print(f"Μέση Τιμή (Mean): {mean}")
print(f"Διακύμανση (Variance): {variance}")
print(f"Τυπική Απόκλιση (Standard Deviation): {standardDeviation}")

# Γράφημα
fig, ax = plt.subplots(figsize=(12, 6))

ax.plot(tempDataset['Datetime'], tempDataset['Temperature'],
label='Θερμοκρασία',color='#4AABFF')
ax.axhline(mean, color='#FF3333', linestyle='--', label=f'Μέση Τιμή:
{mean:.2f}°C')
ax.fill_between(tempDataset['Datetime'], mean - standardDeviation, mean +
standardDeviation, color='#00FF80', alpha=0.1, label=f'±1 Τυπική Απόκλιση:
{standardDeviation:.2f}°C')
ax.set_title('Θερμοκρασία σε Βάθος Χρόνου')
ax.set_xlabel('Ημερομηνία')
ax.xaxis.set_major_locator(mdates.DayLocator(interval=14))  # Two-week interval
ax.xaxis.set_major_formatter(mdates.DateFormatter('%d/%m/%Y %H:%M'))
plt.setp(ax.xaxis.get_majorticklabels(), rotation=45, ha='right')
ax.set_ylabel('Θερμοκρασία (°C)')
ax.legend()

plt.tight_layout()
plt.show()
```

```
# Αυτοσυνδιακύμανση και Αυτοσυσχέτιση
def autocovariance(series, lag):
       return series.cov(series.shift(lag))

lags = np.arange(1, 21)

autocovariances = [autocovariance(tempDataset['Temperature'], lag) for lag in
lags]
autocorrelations = [tempDataset['Temperature'].autocorr(lag) for lag in lags]

# Γράφημα
plt.figure(figsize=(10, 10))

# Αυτοσυνδιακύμανση
plt.subplot(2, 1, 1)
plt.stem(lags, autocovariances, linefmt='#4AABFF', markerfmt='o', basefmt=' ')
plt.title('Αυτοσυνδιακύμανση')
plt.xlabel('Χρονοκαθυστέρηση')
plt.ylabel('Αυτοσυνδιακύμανση')

# Αυτοσυσχέτιση
plt.subplot(2, 1, 2)
plt.stem(lags, autocorrelations, linefmt='#4AABFF', markerfmt='o', basefmt=' ')
plt.title('Αυτοσυσχέτιση ')
plt.xlabel('Χρονοκαθυστέρηση')
plt.ylabel('Αυτοσυσχέτιση ')

plt.tight_layout()
plt.show()
```

```
# Κινητός Μέσος Όρος με Διαφορετικά Παράθυρα και Συνέλιξη

window_sizes = [4, 96, 672] # 4= 1 ώρα, 96= 1 ημέρα, 672= 1 εβδομάδα

# Γράφημα Συνέλιξης με Κινητό Μέσο Όρο
fig, ax = plt.subplots(3, 1, figsize=(10, 20))

for i, window_size in enumerate(window_sizes):
      kernel = np.ones(window_size) / window_size
      moving_average_convolution = np.convolve(tempDataset['Temperature'],
kernel, mode='same')

      ax[i].plot(tempDataset['Datetime'], tempDataset['Temperature'],
label='Θερμοκρασία',color='#4AABFF')
      ax[i].plot(tempDataset['Datetime'], moving_average_convolution,
label=f'Συνέλιξη (Window={window_size})', color='#FF3333')
      ax[i].set_title(f'Συνέλιξη με κινητό μέσο όρο με παράθυρο {window_size}')
      ax[i].set_xlabel('Ημερομηνία')
      ax[i].xaxis.set_major_locator(mdates.DayLocator(interval=14))
      ax[i].xaxis.set_major_formatter(mdates.DateFormatter('%d/%m/%Y %H:%M'))
      plt.setp(ax[i].xaxis.get_majorticklabels(), rotation=45, ha='right')
      ax[i].set_ylabel('Θερμοκρασία (°C)')
      ax[i].legend()

plt.tight_layout()
plt.show()
```