Time Series

Introduction and Stationarity

Echcharif EL JAZOULI
Yakine TAHTAH

Sia Partners

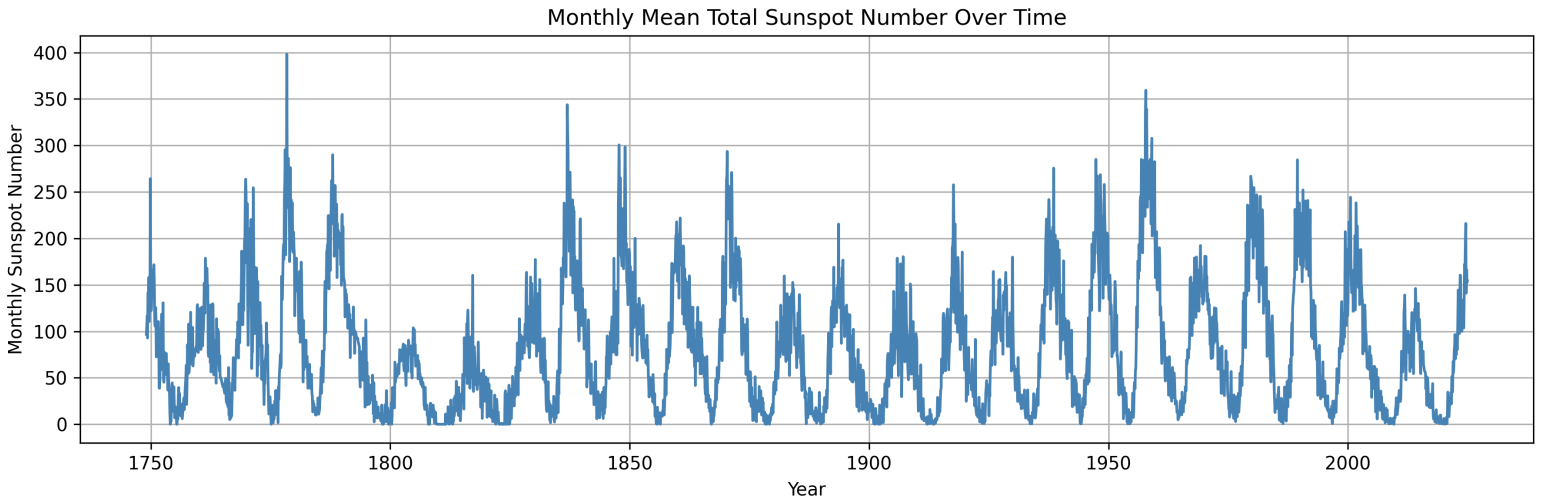
27 janvier 2025

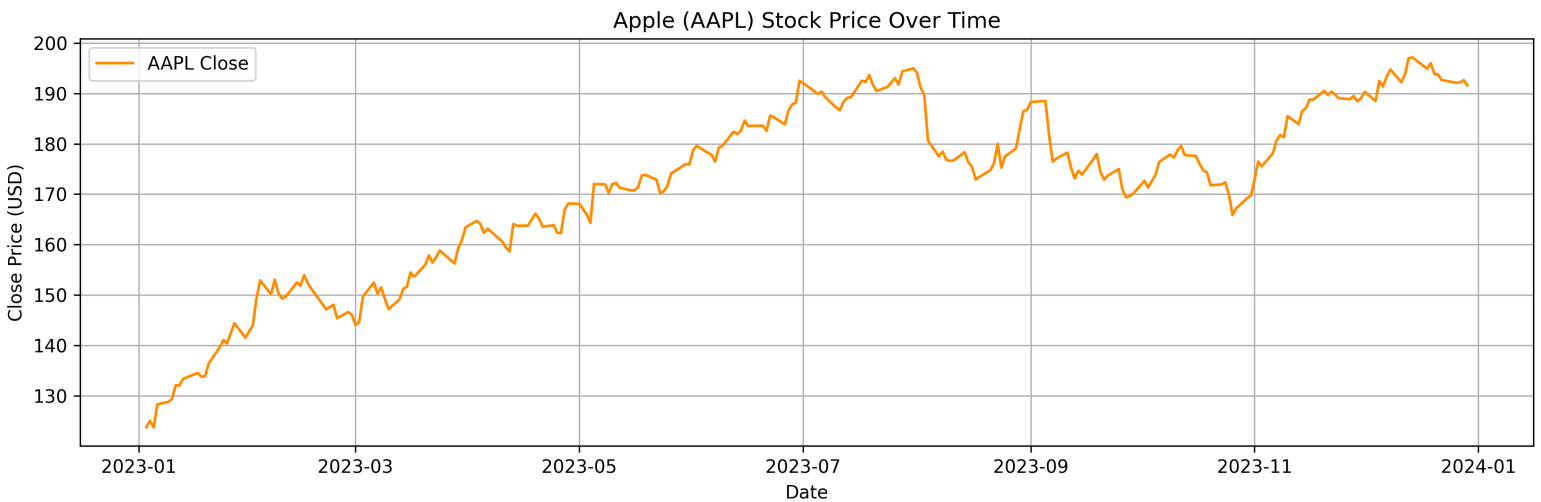
Table of Contents

Why Study Time Series?

- Observations taken over time often exhibit dependence.
- Examples in engineering:
  - Sensor readings (vibrations, temperatures, pressure).
  - Real-time data streams (network traffic, signals in control processes, etc.).
- Goal: **understand** the underlying correlation structure, **model** it, and **forecast** or **infer** future/hidden behavior.

Examples of Time Series Data





Stochastic Processes and Time Series

Definition 1 (Time Series) A *stochastic process* $\{X(t); t \in T\}$ is a collection of random variables defined on the same sample space, indexed by t. When T represents *time*, the stochastic process is called a *time series*.

Stochastic Processes and Time Series

A *stochastic process* is a very general concept: a family of random variables $\{X(\alpha) : \alpha \in A\}$ indexed by some set A. A *time series* is the same concept, but the index set T is specifically *time*, often discrete, e.g. $t = 1, 2, 3, \ldots$.

A **stochastic process** can be indexed by almost anything (space, temperature, location, etc.).

A **time series** is just a *stochastic process indexed by time*. In discrete form, $t \in \mathbb{Z}$ (or $\mathbb{N}$), and in continuous time, $t \in \mathbb{R}$.

Contextually, time series methods emphasize *temporal correlation*, *forecasting*, and *causality*, while more general stochastic processes might focus on broader or different types of dependencies.

Time Series vs. Markov Process

- A **Markov process** is a specific kind of stochastic process with the *Markov property*: future states depend on the current state but *not* directly on past states.
- By contrast, a general **time series** can exhibit *any* form of temporal dependence—ARMA models, long-memory processes, seasonal components, etc.—not necessarily limited to Markovian one-step-ahead dependence.
- In short, every Markov chain *is* a time series (with a particular, restricted dependence structure), but many time series do *not* meet the Markov property.

Is Natural Language a Time Series?

- At a high level, text (or speech) can be viewed as a **sequence** of linguistic tokens (words, phonemes, etc.) over *time* or *position*, so it is often treated *like* a time series in NLP.
- However, **stationarity** rarely holds. Language statistics (vocabulary, style, semantics) can shift significantly over the course of a document or conversation.
- *Most* real-world language data are *nonstationary*—topics change, new vocabulary appears, writing style evolves, etc.

Transformers and Time Series?

If transformers work so well on natural language, could they work on other time series data? (assuming we have large scale data from the same domain or same type as the time series of interest)

Realizations of a Time Series

Definition 2 (Realization) A *realization* (or sample path) of a time series $X(t)$ is the set of real values $\{X(t, \omega);\ t \in T\}$ obtained by fixing an elementary event ω.

In practice, we often see *one* realization (one path) of length n (in the discrete case):

$$\{x_1, x_2, \ldots, x_n\}.$$

Our challenge: infer properties of the *stochastic process* from just this single sequence!

Ensembles, Realizations and Ensemble Mean

1. A *stochastic process* $\{X(t) : t \in T\}$ is defined on a probability space $(\Omega, \mathcal{F}, P)$. Each point $\omega \in \Omega$ corresponds to an outcome.
2. A *realization* (or *sample path*) is what you get if you fix a particular ω. Then $t \mapsto X(t, \omega)$ is a single function of time.
3. The *ensemble* is the collection of *all* such realizations for *all* $\omega \in \Omega$.
4. The *ensemble mean* of $X(t)$ is called that because for each fixed time t, you average over the random variable values across Ω. Formally,
$$\mu_t = \mathrm{E}[X(t)] = \int_{\Omega} X(t, \omega)\, dP(\omega).$$
5. Hence, "ensemble mean" emphasizes that we are integrating across *the entire set of possible realizations* at a fixed instant in time, rather than averaging across different time points.

Strict vs. Covariance Stationarity

Definition 3 (Strict Stationarity)
A process $\{X(t)\}$ is *strictly stationary* if the joint distribution of

$$\{X(t_1), X(t_2), \ldots, X(t_k)\}$$

is identical to that of

$$\{X(t_1 + h), X(t_2 + h), \ldots, X(t_k + h)\}$$

for all $t_1, \ldots, t_k$ and for any h.

Definition 4 (Covariance Stationarity)
$\{X(t)\}$ is *covariance stationary* if:

1. $E[X(t)] = \mu$ (constant in time),
2. $\mathrm{Var}[X(t)] = \sigma^2 < \infty$ (constant in time),
3. $\gamma(t_1, t_2)$ depends only on $t_2 - t_1$ (i.e., depends on the lag only).

Covariance Stationary but not Strictly Stationary

Think of an example.

Autocovariance and Autocorrelation

Autocovariance and Autocorrelation for Stationary Time Series

If a time series $\{X_t\}$ is covariance stationary, the autocovariance function

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h})$$

depends only on the lag h. The autocorrelation function is given by

$$\rho(h) = \frac{\gamma(h)}{\sigma^2}.$$

Properties of $\gamma(h)$ and $\rho(h)$ for a stationary time series:

- $\gamma(0) = \sigma^2$, $\rho(0) = 1$.
- $|\gamma(h)| \leq \gamma(0)$ and $|\rho(h)| \leq 1$ for all h.
- $\gamma(h) = \gamma(-h)$, similarly, $\rho(h) = \rho(-h)$.
- $\gamma(h)$ and $\rho(h)$ are *positive semidefinite*.

Purely Random Process (White Noise)

Definition 5 (Discrete White Noise) A time series $\{X_t\}$ is called *discrete white noise* if:

1. X_t are identically distributed (often taken with mean 0) and variance $\gamma(t,t) = \sigma^2$
2. $\gamma(t_1, t_2) = 0$ when $t_1 \neq t_2$

Interpretation: No correlation at all across time.

Sample Mean Estimation

Observing one realization $\{x_1, \ldots, x_n\}$:

$$\overline{X} = \frac{1}{n} \sum_{t=1}^{n} x_t$$

is the natural estimator of the constant mean μ (if the process is stationary).

Theorem 1 (Ergodicity Condition) $\overline{X}$ is ergodic for μ if and only if

$$\lim_{k \to \infty} \frac{1}{k} \sum_{j=0}^{k-1} \gamma_j = 0.$$

A simpler sufficient condition is that $\lim_{k \to \infty} \gamma_k = 0$ (or equivalently $\rho_k \to 0$).

Stationarity, Ergodicity, and "Statistical Equilibrium"

Key Questions to Explore:

1. What does it mean for a time series to be "stationary"?
2. Does stationarity automatically imply "ergodicity"?
3. What is "statistical equilibrium"? Why do we care?

Informal Answers:

- **Stationarity:** The distributional properties (means, variances, correlations) do not change with shifts in time.
- **Ergodicity:** Long-run time averages from a single (sufficiently long) realization match the ensemble averages.
- **Statistical equilibrium:** Another way to say "the process's key statistical measures stay constant in time," which is essentially stationarity. Ergodicity further ensures we can reliably estimate those measures from one observed path.

Why Stationarity Is Not Always Enough

Stationary **doesn't imply** Ergodic

- Even if a process is stationary, a single observed realization might *not* reveal the full distributional properties.
- **Ergodicity** adds a condition that *time averages* converge to *ensemble (true) averages*.
- Example "mixture" scenario: We can randomly pick one of two stationary processes at the start and then remain stuck with it forever. Unconditionally, the mixture is still stationary, but from one path, you never see the other component.

Practical Implication:

- We often **assume** both stationarity and ergodicity in classic time-series analysis so that one observed time series is enough to estimate global properties like mean or autocorrelation.

The Coin-Flip Mixture

Goal: Construct a stationary process that is not ergodic.

The Coin-Flip Mixture

- Take two strictly *stationary* processes, $\{X_t^H\}$ and $\{X_t^T\}$, each with possibly different means and autocovariances.
- Flip a coin once at $t = 0$. If it's heads, use X_t^H for all t. If tails, use X_t^T for all t. Denote this event $Z = H$ or $Z = T$.
- Define the new process:

$$X_t = \begin{cases} X_t^H, & \text{if } Z = H, \\ X_t^T, & \text{if } Z = T. \end{cases}$$

The Coin-Flip Mixture

Why is X_t stationary?

- Before you look at any specific realization, the "unconditional" distribution is a blend $(pH + (1-p)T)$ of two stationary processes. This overall mixture does not change with time.

Why the Mixture Is Not Ergodic

Once the coin flip occurs, each *particular* realization stays locked into either the X_t^H branch or the X_t^T branch. Therefore:

- A single observed series comes from *only one* of the two processes, so the time average typically converges to μ_H or μ_T, not the mixture mean $p\,\mu_H + (1-p)\,\mu_T$.
- This violates the definition of ergodicity, which demands that the time average (of one realization) equals the ensemble average (over all possible outcomes).

Consequences:

- We have an example that is **stationary overall** (same distribution at all times) yet **fails ergodicity** (a single long path does not reveal the mixture distribution).

Step 1: Mean of the Mixture

The mean of the mixture equals

$$\mu = \mathbb{E}[X_t] = p\,\mu_H + (1-p)\,\mu_T.$$

Proof:

1. By definition, the mixture process is entirely one of the two stationary processes $\{X_t^H\}$ or $\{X_t^T\}$, depending on the outcome of a one-time "coin flip" at $t = 0$.
2. Unconditionally, we compute expectation by weighting with the probabilities p and $1-p$:
$$\mathbb{E}[X_t] = p\,\mathbb{E}[X_t^H] + (1-p)\,\mathbb{E}[X_t^T].$$
3. Each original process has its own constant mean (μ_H for the H-branch and μ_T for the T-branch) because they are stationary. Hence
$$\mathbb{E}[X_t^H] = \mu_H, \quad \mathbb{E}[X_t^T] = \mu_T.$$

Step 2: Autocovariance of the Mixture

For $k \geq 0$,

$$\gamma(k) \;=\; \mathrm{Cov}\big(X_t, X_{t+k}\big)$$

$$=\; p\left[\gamma^H(k) + (\mu_H - \mu)^2\right] \;+\; (1-p)\left[\gamma^T(k) + (\mu_T - \mu)^2\right].$$

Proof:

1. Recall that $\gamma(k) = \mathbb{E}[(X_t - \mu)(X_{t+k} - \mu)]$.
2. Condition on whether the entire sequence is in branch H (with probability p) or branch T (with probability $1-p$). Thus

$$\gamma(k) \;=\; p\,\mathbb{E}[(X_t^H - \mu)(X_{t+k}^H - \mu)] + (1-p)\,\mathbb{E}[(X_t^T - \mu)(X_{t+k}^T - \mu)].$$

3. Expand $X_t^H - \mu$ as $(X_t^H - \mu_H) + (\mu_H - \mu)$. Since $\mathbb{E}[X_t^H - \mu_H] = 0$ for the stationary process $\{X_t^H\}$, cross-terms with $\alpha^H := (\mu_H - \mu)$ lead to

$$\mathbb{E}[(X_t^H - \mu)(X_{t+k}^H - \mu)] \;=\; \gamma^H(k) + (\mu_H - \mu)^2.$$

A similar identity holds for the T-branch.

Step 3: $\gamma^H(k) \to 0$ and $\gamma^T(k) \to 0$

Assumption: Each original process $\{X_t^H\}$ and $\{X_t^T\}$ is a "nice" stationary process (e.g. an ARMA-type) whose autocovariance "dies out" as $k \to \infty$. Formally,

$$\lim_{k\to\infty} \gamma^H(k) = 0 \quad \text{and} \quad \lim_{k\to\infty} \gamma^T(k) = 0.$$

- Most linear stationary processes (e.g. AR(1), ARMA(p, q), or processes with sufficient mixing conditions) have autocovariances $\gamma(k)$ that converge to 0 as $k \to \infty$.
- Concretely, an AR(p) or ARMA(p, q) process often satisfies absolute summability of its impulse response and thus $\gamma(k)$ must tend to 0.
- Under this assumption, analyzing large-lag behavior in the mixture is straightforward.

Step 4: Violation of the Ergodic Criterion

Key Consequence: Since

$$\gamma(k) = p\left[\gamma^H(k) + (\mu_H - \mu)^2\right] + (1-p)\left[\gamma^T(k) + (\mu_T - \mu)^2\right],$$

we examine its limit as $k \to \infty$:

$$\lim_{k\to\infty} \gamma(k) \;=\; p\,(\mu_H - \mu)^2 \;+\; (1-p)\,(\mu_T - \mu)^2$$

Since $\mu = p\,\mu_H + (1-p)\,\mu_T$, a simple algebraic check shows $\lim_{k\to\infty} \gamma(k) \neq 0$ whenever $\mu_H \neq \mu_T$.

Hence, ergodic criterion fails: We see that because $\gamma(k)$ tends to a nonzero constant, its average remains nonzero. Thus:

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} \gamma(j) \;\neq\; 0,$$

violating the necessary and sufficient condition for ergodicity.

Conclusion

From a single realization, the time average of X_t locks onto either μ_H or μ_T rather than the mixture mean $p\,\mu_H + (1-p)\,\mu_T$. Thus, we cannot recover the full ensemble mean from a single path—ergodicity fails.

Stationarity vs. Ergodicity: Practical Takeaways

1. **Stationarity** says "the process's statistical properties don't change over time."
2. **Ergodicity** says "one long realization is representative of the entire distribution."
3. Many familiar stationary models (e.g. ARMA with typical mixing conditions) *are* ergodic. But the coin-flip mixture illustrates that "stationary" alone does *not* guarantee it.
4. In time-series analysis, assuming both properties often underlies standard estimation techniques—without ergodicity, even having a stable distribution may not allow accurate inference from a single path.

Stationarity vs. Ergodicity: Conclusion

Stationarity is about *time-invariance of distributions*, and ergodicity is about *time sampling sufficiency*. You can have stationarity without ergodicity if your single realization cannot traverse all possible states of the system.

Autocovariance Estimation

For a discrete parameter time series:

- Two estimators for autocovariance:

$$\tilde{\gamma}_k = \frac{1}{n - |k|} \sum_{t=1}^{n-|k|} (X_t - \bar{X})(X_{t+|k|} - \bar{X})$$

and:

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-|k|} (X_t - \bar{X})(X_{t+|k|} - \bar{X})$$

- Though $\tilde{\gamma}_k$ is less biased, $\hat{\gamma}_k$, also called sample autocovariance, has a lower **mean squared error** in most cases.

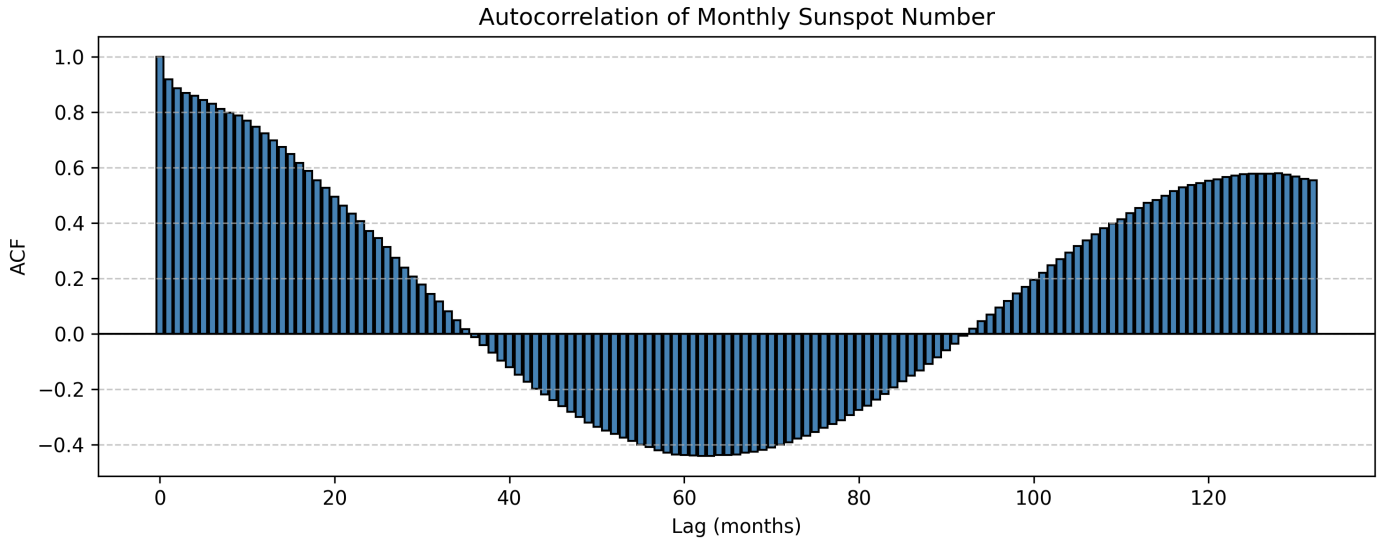
Autocorrelation Estimation

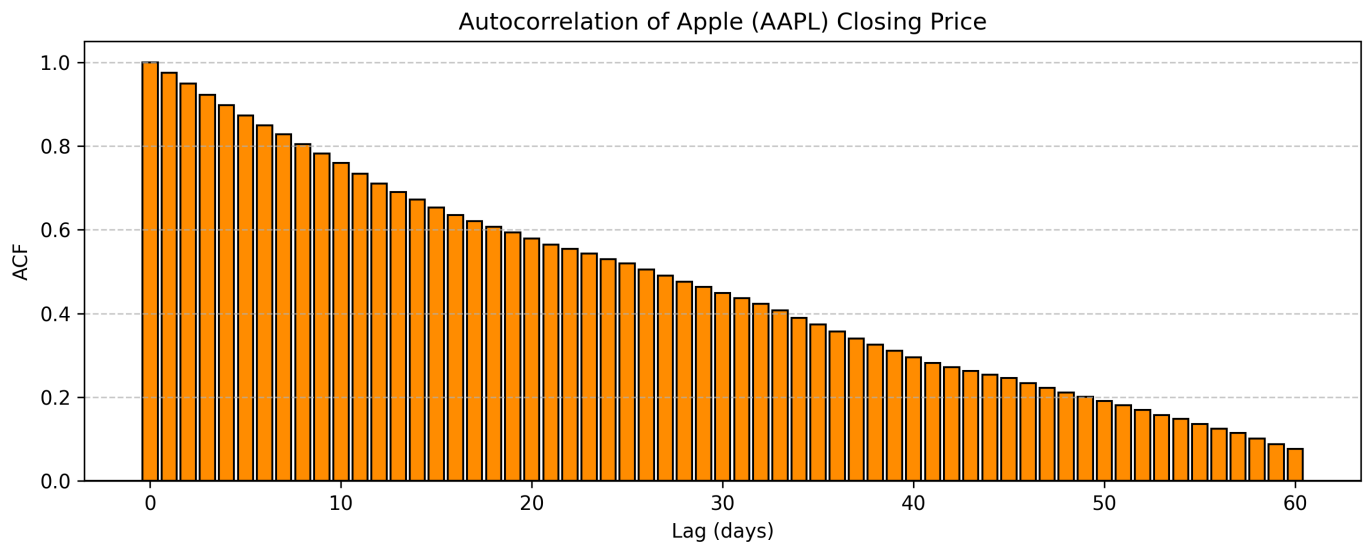
For a discrete parameter time series, the sample autocorrelation is

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

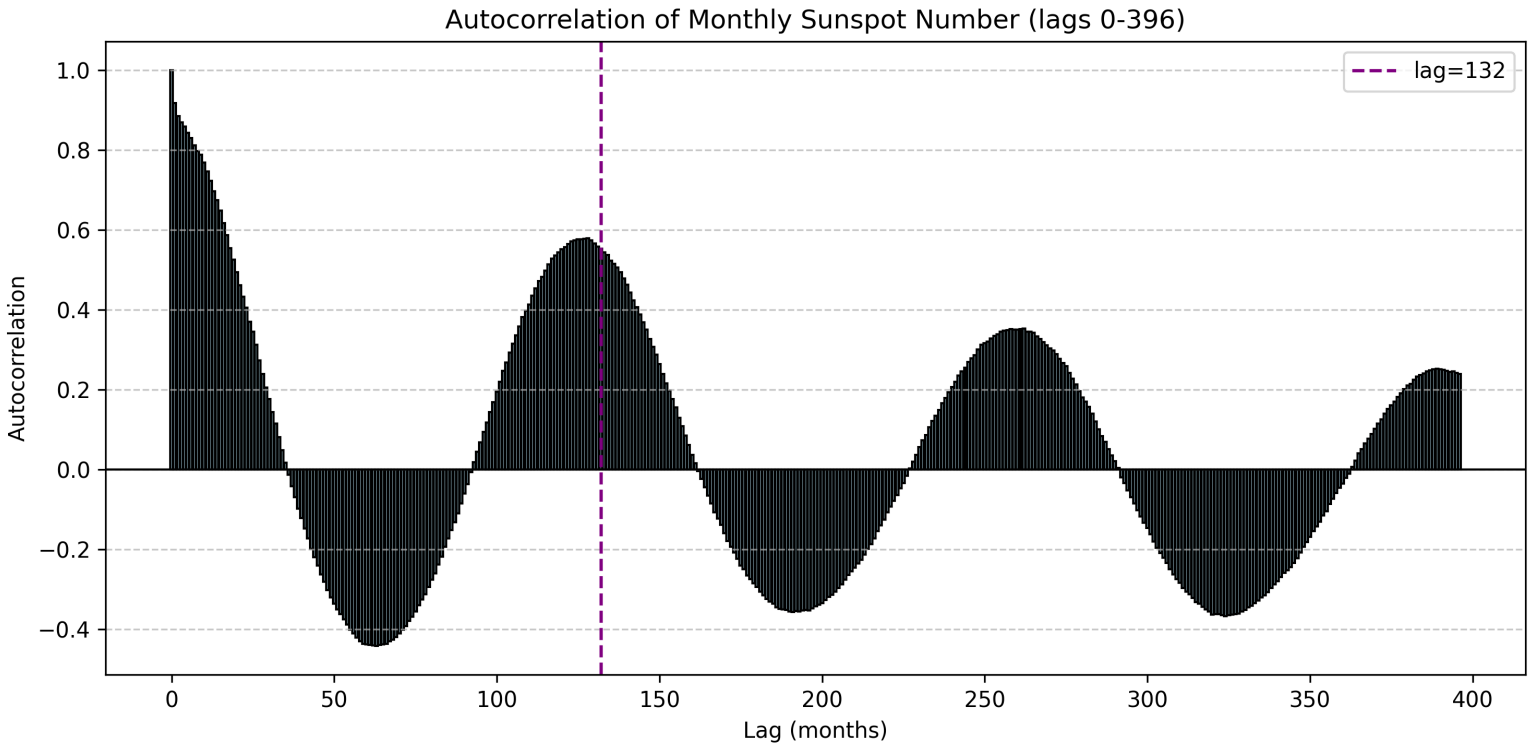
- $|\hat{\rho}_k| \leq 1$ (which is not guaranteed by the alternative estimate).
- There is correlation among $\hat{\rho}_k$ values.
- Plots of $\hat{\rho}_k$ vs. lag k (ACF plots) help identify correlation patterns.

ACF Examples





ACF Examples: Sunspot



Time vs. Frequency Domain

- **Time-domain**: focuses on autocorrelation, behavior of the process across time.
- **Frequency-domain**: focuses on identifying *cyclic* behavior by studying the *spectrum*.
- Autocovariance $\longleftrightarrow$ Power spectrum each contains the **same** information but in different representations.

Time-domain vs. Frequency-domain

- You do *not* need an obviously cyclical series to benefit from frequency-domain approaches. Frequency-domain analysis can be useful whenever you suspect (or want to test for) "hidden" repetitive behavior at various frequencies.
- **Frequency-domain analysis**, or spectral analysis, helps reveal how different frequencies contribute to the overall signal. Even if a time series does *not* exhibit a clean cycle, a frequency-based approach can uncover important components that repeat in a quasi-regular manner.
- The terms *seasonal*, *cyclical*, and *periodic* differ mostly by how predictable and strictly repeated the patterns are, and whether the period is known/fixed or not.

Time-domain vs. Frequency-domain

1. **Seasonal patterns** usually have a *known*, *fixed*, and *externally imposed* period. For instance, monthly data with a period of 12 (annual seasonality) or daily data with a period of 7 (weekly seasonality).
2. **Cyclical patterns** refer to *recurring* but *not strictly periodic* variations. They are often associated with economic or business cycles, where the length of each "cycle" can vary from occurrence to occurrence.
3. **Periodic** means *strict mathematical periodicity*: a function $x(t)$ is periodic with period T if $x(t+T) = x(t)$ for all t.

Hence, *seasonal* = a special case of a *cyclical* pattern with known, strictly repeated cycles imposed by an external calendar/factor,

Power Spectrum and Spectral Density

(Continuous-parameter version): For a stationary time series:

$$P_X(f) = \int_{-\infty}^{\infty} e^{-2\pi i f h}\, \gamma(h)\, dh, \quad S_X(f) = \frac{P_X(f)}{\sigma_X^2}.$$

In practice, for *discrete* time series sampled at unit time steps, we will use sums instead of integrals. For a **discrete time** stationary process defined on integers,

$$P_X(f) = \sum_{k=-\infty}^{\infty} \gamma_k\, e^{-2\pi i k f}, \quad S_X(f) = \frac{P_X(f)}{\sigma_X^2}, \quad \text{for } |f| \leq 0.5.$$

Relationship: $\gamma(k)$ and $P_X(f)$

If $\sum_{k=-\infty}^{\infty} |\gamma_k| < \infty$, then:

$$\gamma_k = \int_{-0.5}^{0.5} e^{2\pi i f k} P_X(f)\, df,$$

Key Point: The autocovariance function $\{\gamma_k\}$ and the power spectrum $P_X(f)$ are *Fourier transform pairs* and hence contain **equivalent** mathematical information about the process.

We can also show that:

$$P_X(f) = \sigma_X^2 + 2\sum_{k=1}^{\infty} \gamma_k \cos(2\pi f k), \quad S_X(f) = 1 + 2\sum_{k=1}^{\infty} \rho_k \cos(2\pi f k).$$

Nyquist Frequency and Aliasing

Definition 6 (Nyquist Frequency) If data are sampled at increments Δ, then the Nyquist frequency is $1/(2\Delta)$, and the shortest observable period is 2Δ.

- For discrete time data sampled at unit increments, the shortest period observable is 2 (in time units).
- Thus, the *highest* observable frequency is $f = 0.5$ cycles/sample, which is the Nyquist frequency.

Aliasing:

- Frequencies above 0.5 (in unit sampling) "fold" back into $[-0.5, 0.5]$.
- Example: $\cos(2\pi \cdot 0.3\, t)$ and $\cos(2\pi \cdot 1.3\, t)$ coincide on integer t, so $f = 1.3$ looks identical to $f = 0.3$ when sampled.

Sample Spectrum and Periodogram

Definition 7 (Sample Spectrum and Periodogram)
Let $\{X_t\}$ be discrete and stationary. Based on a realization $\{x_1, \ldots, x_n\}$:

$$\hat{P}_X(f) = \hat{\sigma}_X^2 + 2\sum_{k=1}^{n-1} \hat{\gamma}_k \cos(2\pi f k), \quad |f| \leq 0.5,$$

$$\hat{S}_X(f) = 1 + 2\sum_{k=1}^{n-1} \hat{\rho}_k \cos(2\pi f k).$$

The *periodogram* is simply the sample spectrum evaluated at the "harmonics" $f_j = j/n, j = 1, \ldots, \lfloor n/2 \rfloor$:

$$I(f_j) = \hat{\sigma}_X^2 + 2\sum_{k=1}^{n-1} \hat{\gamma}_k \cos(2\pi f_j k).$$

Properties of the Sample Spectrum

Theorem 2 Let X_t be a discrete stationary time series on the integers, and let $\hat{P}_X(f)$ be the sample spectrum. Then:

1. $\hat{P}_X(f)$ is *asymptotically unbiased*:

$$\lim_{n\to\infty} E[\hat{P}_X(f)] = P_X(f).$$

2. $\hat{P}_X(f)$ is **not consistent**: its variance *does not* go to zero as $n \to \infty$.

Often, we mitigate this by *smoothing* the periodogram (using window functions) to reduce volatility.

Smoothing the Periodogram

- A *smoothed spectral estimator* has the form

$$\hat{P}_S(f) \;=\; \lambda_0\,\hat{\sigma}_X^2 \;+\; 2\sum_{k=1}^{M}\lambda_k\,\hat{\gamma}_k\cos(2\pi f\,k),$$

  where $M \ll n$ and $\{\lambda_k\}$ is a chosen *window function*.
- Common windows: Bartlett (triangular), Tukey, Parzen, etc.
-
- Tradeoff: **smaller** $M \implies$ more smoothing (less variance) but more bias; **larger** $M \implies$ less smoothing but higher variance.
- A typical rule of thumb (especially for Parzen's window) is $M = 2\sqrt{n}$, balancing lower variance (by truncating at smaller M) with not overly smearing genuine peaks.

Conclusion and Further Reading

1. **Stationarity** underpins most classical time series approaches.
2. **Ergodicity** ensures single-run estimations can be trusted to reflect ensemble properties.
3. Modern practice often merges time-domain, frequency-domain, and nonstationary methods for real-world complexities.

Thank you!

References

- Brockwell & Davis : *Time Series: Theory and Methods.*
- Shumway & Stoffer : *Time Series Analysis and Its Applications.*
- Box & Jenkins : *Time Series Analysis: Forecasting and Control.*