# PGE 383

## Time Series Analysis

- Time Series Data
- Time Series Analysis
- Time Series Model
- Time Series Hands-on

Michael Pyrcz, The University of Texas at Austin

# PGE 383

## Time Series Analysis

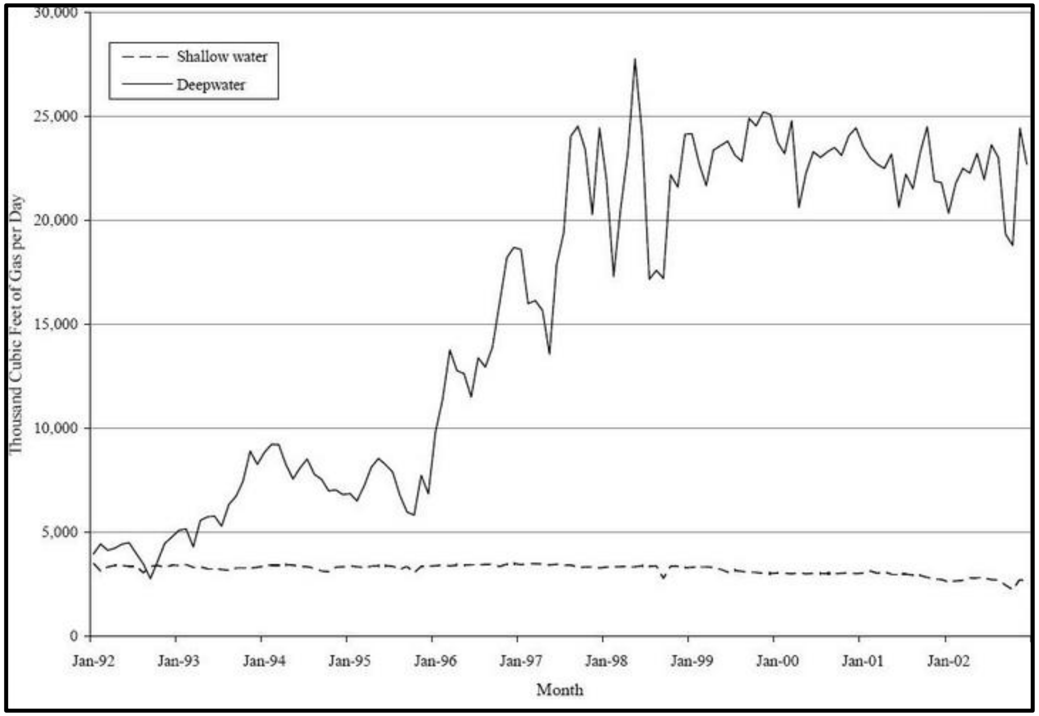
- **Time Series Data**

**Michael Pyrcz, The University of Texas at Austin**

# Time Series Data
# Temporal Data

Time series / temporal data

- measurements made over time
- e.g., production rates and composition

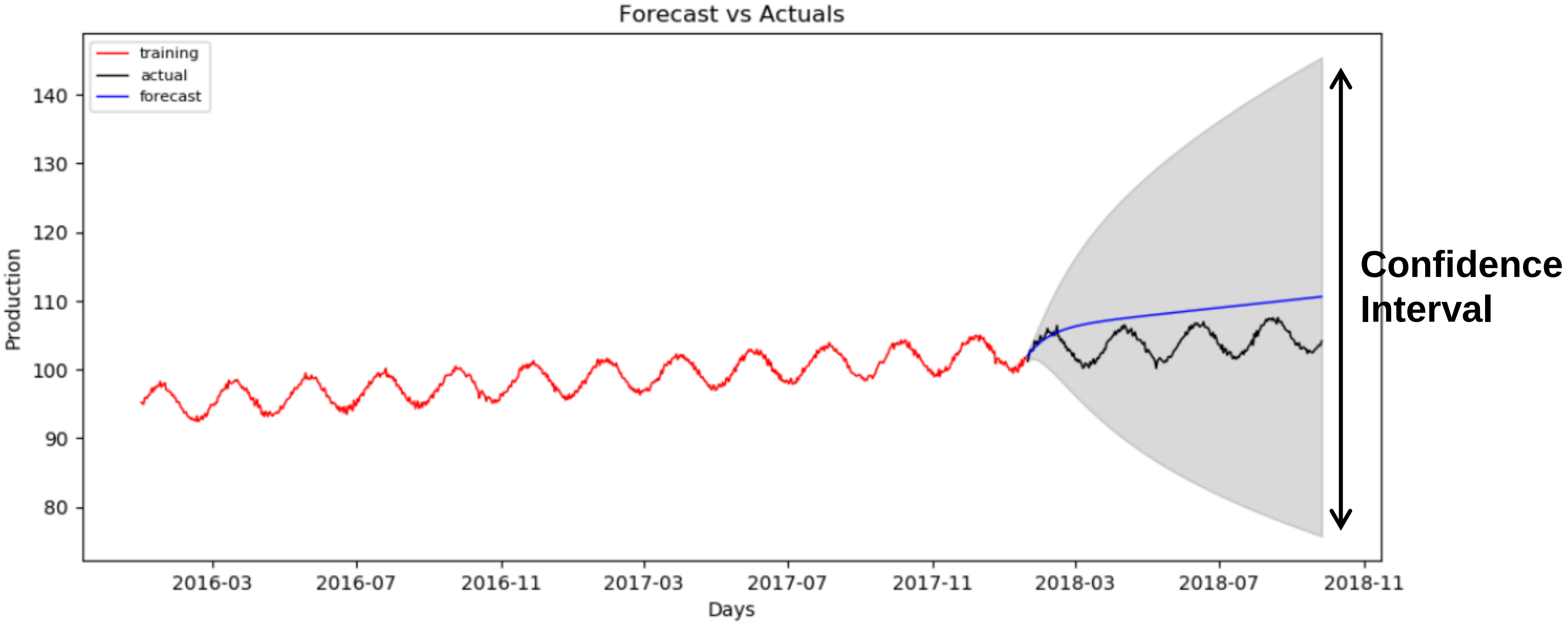


Average production rates for shallow-water and deepwater gas well completions.
https://geographic.org/deepwater_gulf_of_mexico/production_rates.html

# Modeling with Time Series Data vs. Spatial Data

Comparison to spatial data

1. one dimensional
2. observations tend to be clustered or exhaustive over a time interval
3. prediction is often future forecasting or forward extrapolation



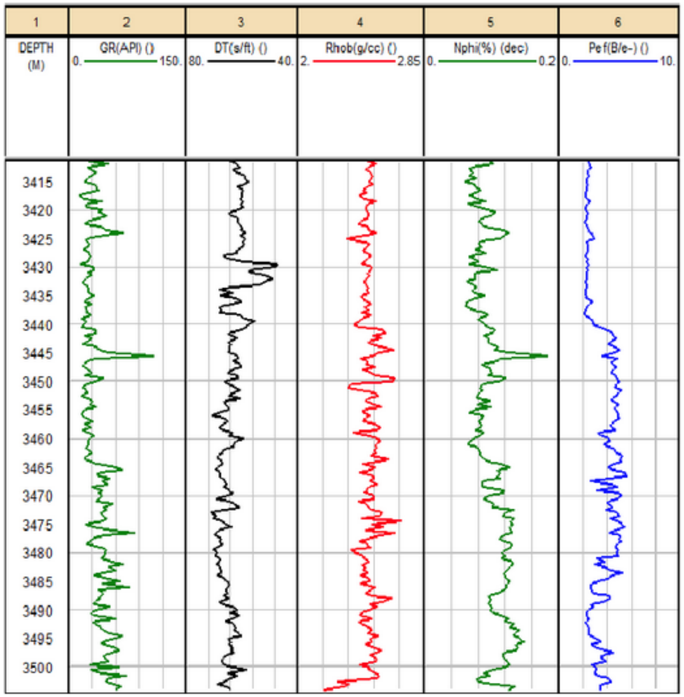
Training, testing and forecast for well production.

# Series Analysis

The general concepts of time series analysis may be applied on other 1D data sets

- Well logs

- Outcrop measured sections



Well logs from Ouadfeul and Alioune, 2013, Lithofacies prediction from well log data using a multilayer perceptron (MLP) and Kohonen's self-organizing map (SOM) - a case study from the Algerian Sahara.

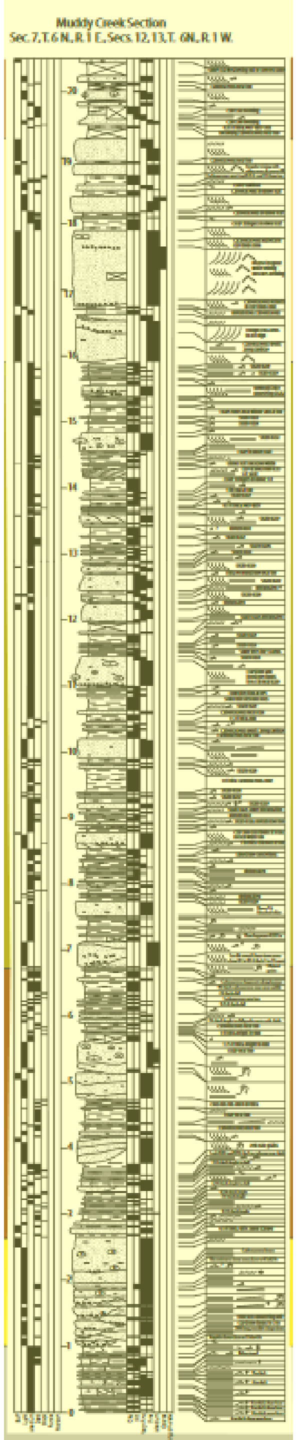
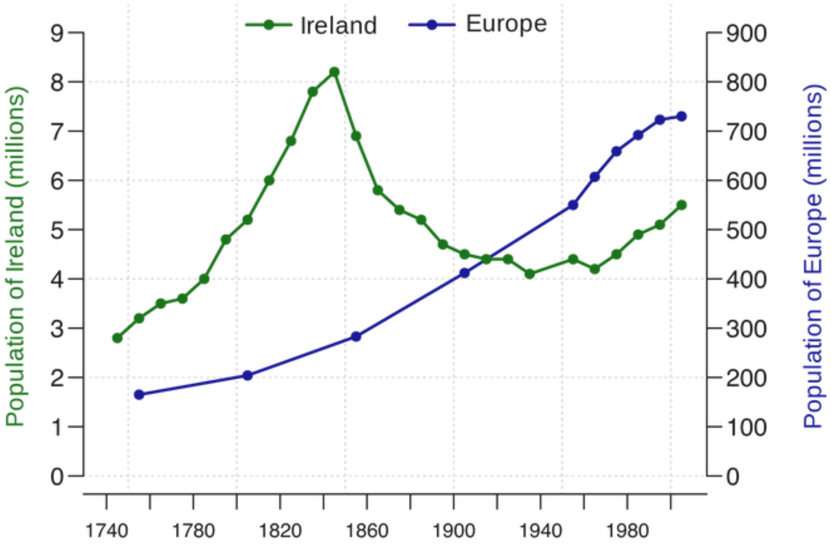


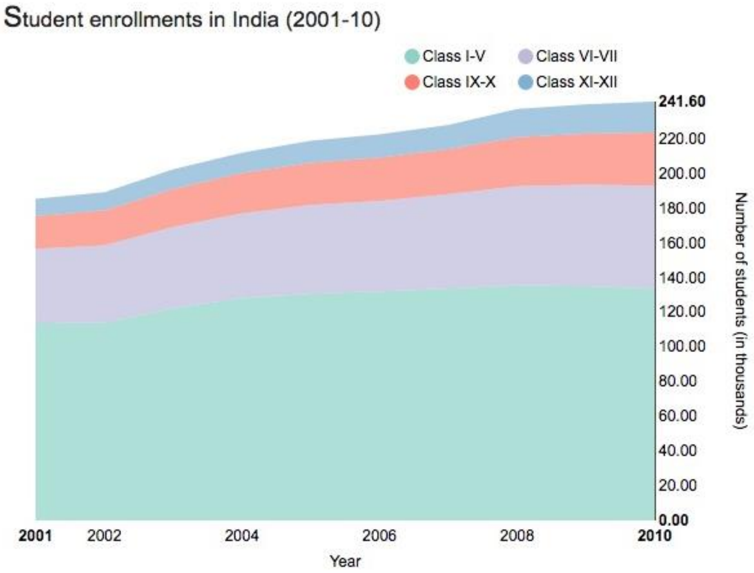
Image from Chapter 10 Detailed Measured Sections, Cross Sections, and Paleogeographic Reconstructions of the Upper Cretaceous and Lower Tertiary Nonmarine interval, Wind River Basin, Wyoming by Ronald C. JohnsonU.S. Geological Survey Digital Data Series DDS–69–J

# Time Series Data
# Temporal Data

## Visualizing Temporal Data

line graph





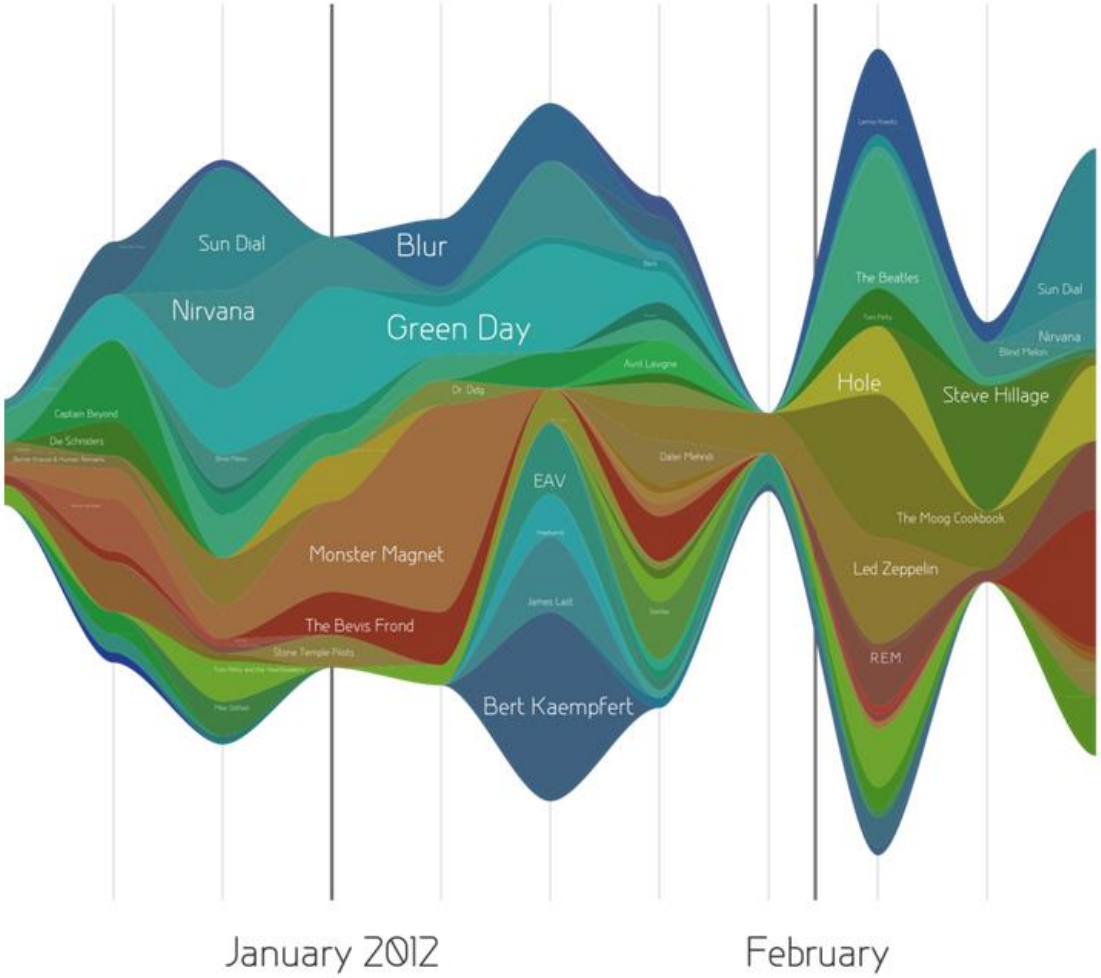
stacked area chart

images by Vasavi Ayalasomayajulafrom https://humansofdata.atlan.com/2016/11/visualizing-time-series-data/

# Time Series Data
# Temporal Data

Visualizing Temporal Data

steam graph



images by Vasavi Ayalasomayajulafrom https://humansofdata.atlan.com/2016/11/visualizing-time-series-data/

# PGE 383

## Time Series Analysis

- Time Series Analysis

# Variogram, Covariance Function and Correlogram

- **Variogram** – a measure of dissimilarity vs. distance. Calculated as ½ the average squared difference of values separated by a lag vector.

$$\boldsymbol{\gamma}(\mathbf{h}) = \frac{1}{2\boldsymbol{N}(\mathbf{h})} \sum_{\boldsymbol{\alpha}=1}^{\boldsymbol{N}(\mathbf{h})} \left(\boldsymbol{z}(\mathbf{u}_{\boldsymbol{\alpha}}) - \boldsymbol{z}(\mathbf{u}_{\boldsymbol{\alpha}} + \mathbf{h})\right)^{2}$$

  - The precise term is semivariogram (variogram if you remove the 1/2 ), but in practice the term variogram is used.
  - The ½ is used so that the covariance function and variogram may be related directly:

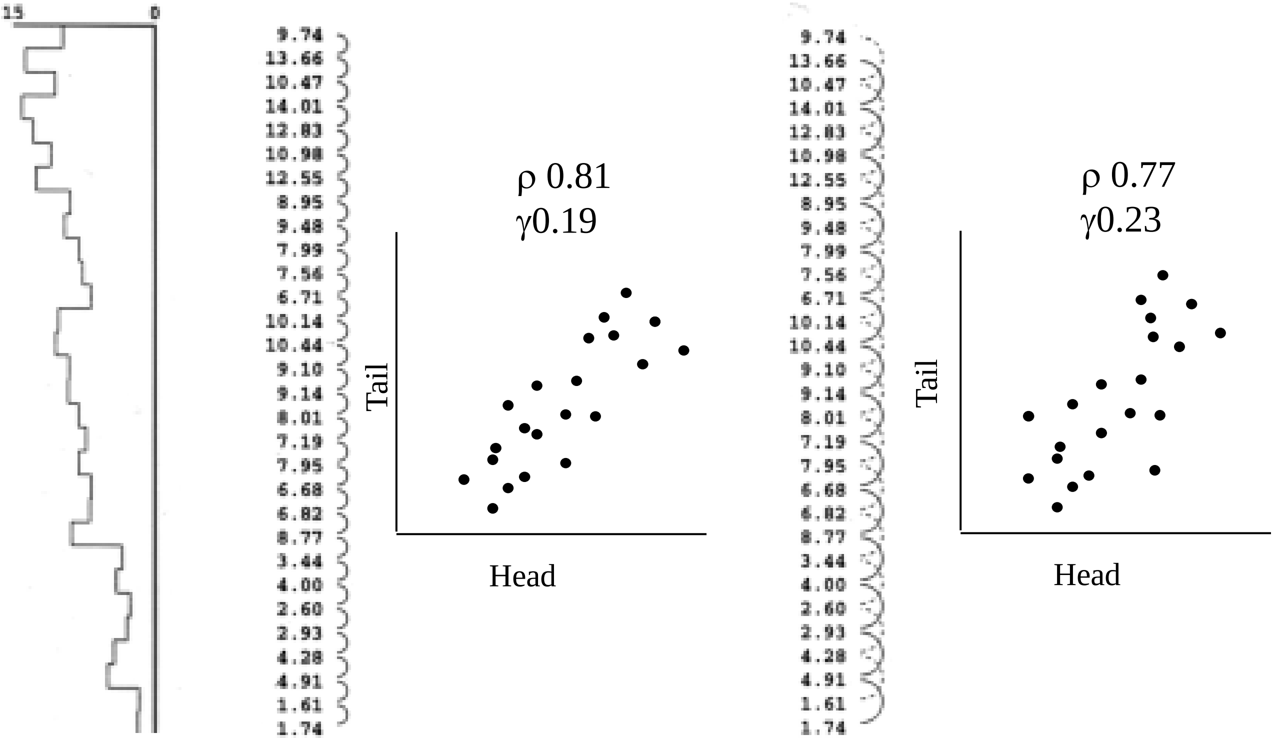
$$\boldsymbol{C}_{\boldsymbol{x}}(\boldsymbol{h}) = \boldsymbol{\sigma}_{\boldsymbol{x}}^{2} - \boldsymbol{\gamma}_{\boldsymbol{x}}(\mathbf{h})$$

  - Note the correlogram is related to the covariance function as:

$$\boldsymbol{\rho}_{\boldsymbol{x}}(\boldsymbol{h}) = \frac{\boldsymbol{C}_{\boldsymbol{x}}(\boldsymbol{h})}{\boldsymbol{\sigma}_{\boldsymbol{x}}^{2}}$$ , h-scatter plot correlation vs. lag distance

# Variogram Calculation

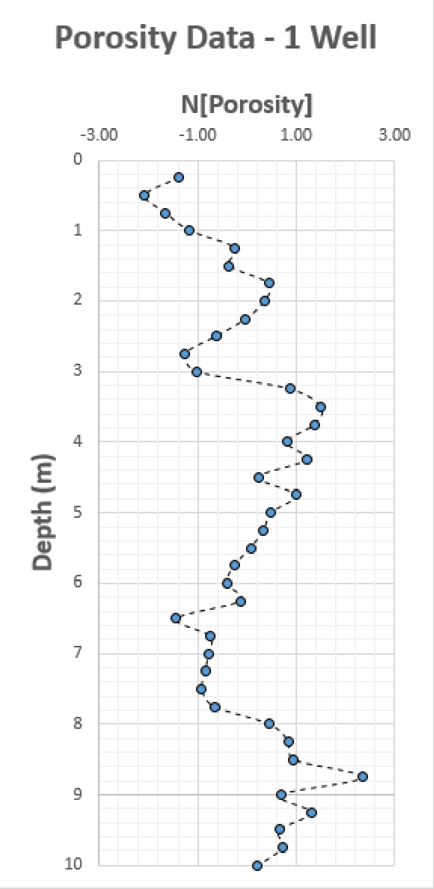
- Consider data values separated by *lag* vectors (the h values)
- Here are two examples of a lag vector equal to the data spacing and then twice the data spacing:



# Variogram Calculation Example

- Pick a lag distance and calculate the variogram for that one lag distance.

- Dataset is at GitHub/GeostatsGuy
- GeoDataSets/1D_Porosity.xlsx

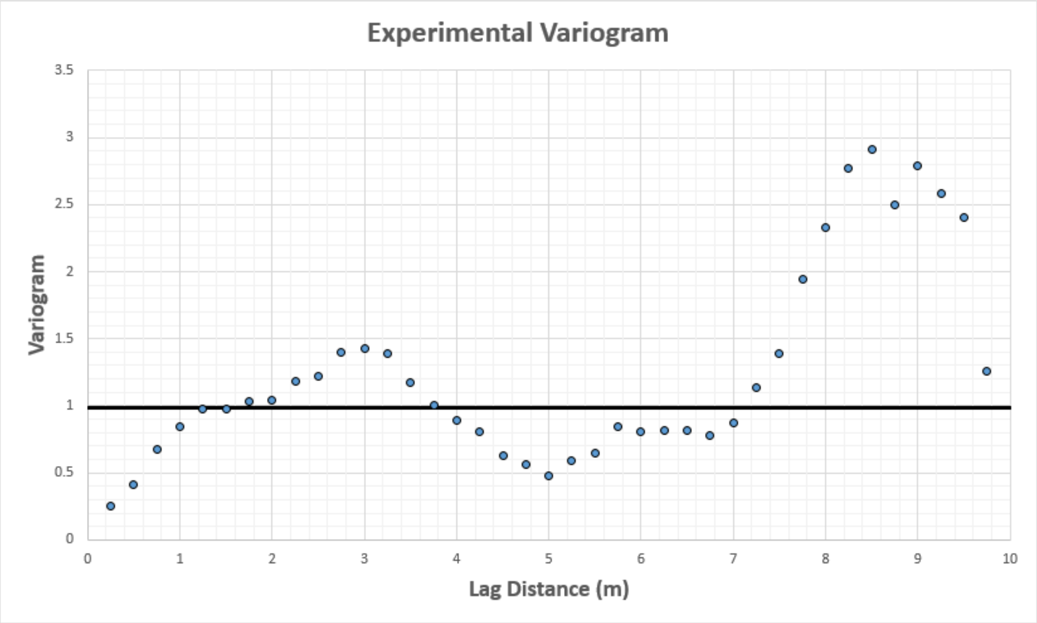


| Depth | N[Porosity] |
|---|---|
| 0.25 | -1.37 |
| 0.5 | -2.08 |
| 0.75 | -1.67 |
| 1 | -1.16 |
| 1.25 | -0.24 |
| 1.5 | -0.36 |
| 1.75 | 0.44 |
| 2 | 0.36 |
| 2.25 | -0.02 |
| 2.5 | -0.63 |
| 2.75 | -1.26 |
| 3 | -1.03 |
| 3.25 | 0.88 |
| 3.5 | 1.51 |
| 3.75 | 1.37 |
| 4 | 0.81 |
| 4.25 | 1.21 |
| 4.5 | 0.24 |
| 4.75 | 0.99 |
| 5 | 0.49 |
| 5.25 | 0.34 |
| 5.5 | 0.07 |
| 5.75 | -0.26 |
| 6 | -0.41 |
| 6.25 | -0.14 |
| 6.5 | -1.44 |
| 6.75 | -0.75 |
| 7 | -0.78 |
| 7.25 | -0.85 |
| 7.5 | -0.92 |
| 7.75 | -0.66 |
| 8 | 0.47 |
| 8.25 | 0.85 |
| 8.5 | 0.95 |
| 8.75 | 2.35 |
| 9 | 0.69 |
| 9.25 | 1.31 |
| 9.5 | 0.66 |
| 9.75 | 0.72 |
| 10 | 0.21 |

# Variogram Calculation Example

- Pick a lag distance and calculate the variogram for that one lag distance.

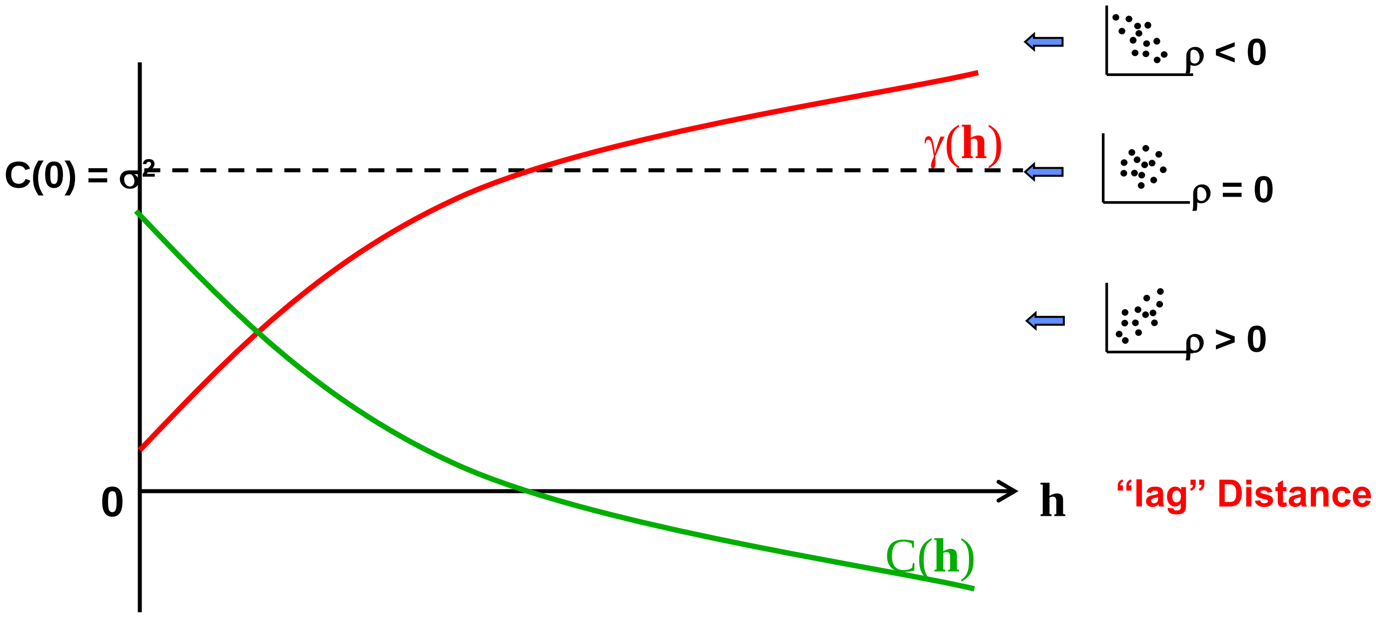
- Here’s all of them:



| Depth | N[Porosity] |
|---|---|
| 0.25 | -1.37 |
| 0.5 | -2.08 |
| 0.75 | -1.67 |
| 1 | -1.16 |
| 1.25 | -0.24 |
| 1.5 | -0.36 |
| 1.75 | 0.44 |
| 2 | 0.36 |
| 2.25 | -0.02 |
| 2.5 | -0.63 |
| 2.75 | -1.26 |
| 3 | -1.03 |
| 3.25 | 0.88 |
| 3.5 | 1.51 |
| 3.75 | 1.37 |
| 4 | 0.81 |
| 4.25 | 1.21 |
| 4.5 | 0.24 |
| 4.75 | 0.99 |
| 5 | 0.49 |
| 5.25 | 0.34 |
| 5.5 | 0.07 |
| 5.75 | -0.26 |
| 6 | -0.41 |
| 6.25 | -0.14 |
| 6.5 | -1.44 |
| 6.75 | -0.75 |
| 7 | -0.78 |
| 7.25 | -0.85 |
| 7.5 | -0.92 |
| 7.75 | -0.66 |
| 8 | 0.47 |
| 8.25 | 0.85 |
| 8.5 | 0.95 |
| 8.75 | 2.35 |
| 9 | 0.69 |
| 9.25 | 1.31 |
| 9.5 | 0.66 |
| 9.75 | 0.72 |
| 10 | 0.21 |

# Variogram, Covariance Function and Correlogram

- **Must plot variance, sill to interpret variogram:**
  - Positive correlation when semivariogram less than variance
  - No correlation when the semivariogram is equal to the variance
  - Negative correlation when the semivariogram points above variance



# Variogram, Covariance Function and Correlogram

- **Covariance Function** – a measure of similarity vs. distance. Calculated as the average product of values separated by a lag vector centered by the square of the mean.

$$\boldsymbol{C_x(\mathbf{h}) = \frac{\sum_{\alpha=1}^{n} x(\mathbf{u}_\alpha)\cdot x(\mathbf{u}_\alpha+\mathbf{h})}{n} - (\overline{x})^2}, \text{ if stationary mean}$$

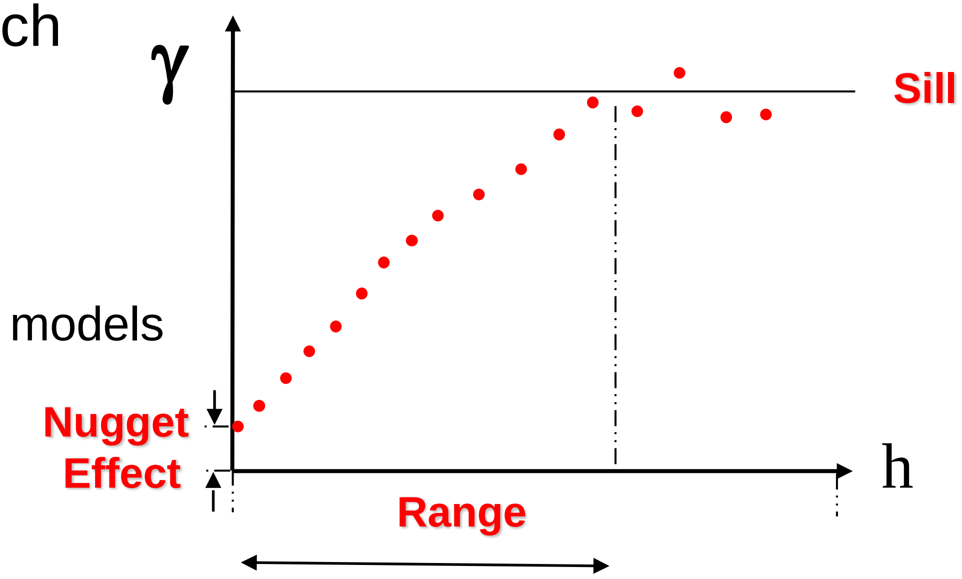
  - The covariance function is the variogram upside down. $\boldsymbol{\gamma_x(\mathbf{h}) = \sigma_x^2 - C_x(h)}$
  - We model variograms, but inside the kriging and simulation methods they are converted to covariance values for numerical convenience.

- **Autocorrelation or Correlogram** – a standardized covarianced function. They are the same when the sill, variance is one, $\boldsymbol{\sigma_x^2} = 1$.

$$\boldsymbol{\rho_x(h) = \frac{C_x(h)}{\sigma_x^2}}$$

# Variogram Components Definition

- **Nugget Effect** – discontinuity in the variogram at distances less than the minimum data spacing
  - As a ratio of nugget / sill, is known as relative nugget effect (%)
  - Measurement error, mixing populations cause apparent nugget effect
- **Sill** – the sample variance
  - Interpret spatial correlation relative to the sill, level of no correlation
- **Range** – lag distance to reach the sill
  - Up to that distance you have information
  - parameterization of variogram models



# Autocorrelation

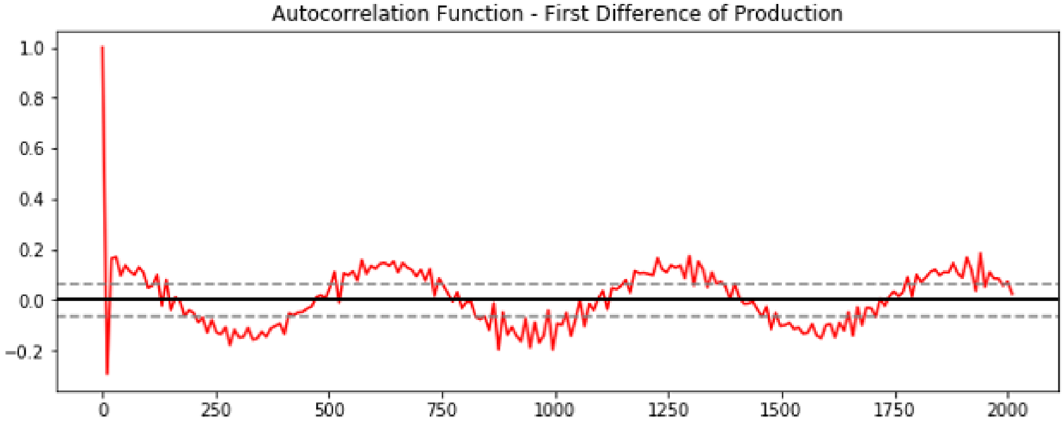
Autocorrelation:

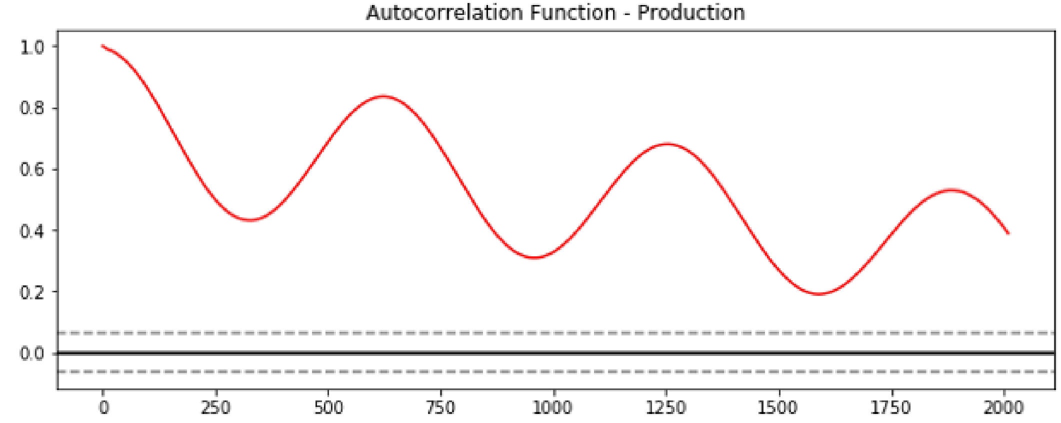
$$r_k = \frac{\sum_{\alpha=1}^{N-k}(y_i - \overline{y}) \cdot (y_{i+k} - \overline{y})}{\sum_{\alpha=1}^{N}(y_i - \overline{y})^2}$$

- where:
  - $\overline{y}$ is the global mean
  - $\sum_{\alpha=1}^{N}(y_i - \overline{y})^2$ is the variance
  - $r_k$ is the autocorrelation for lag $k$

- Note: this is the same as the correlogram:

$$\rho_x(h) = \frac{C_x(h)}{\sigma_x^2}$$





# Hurst Exponent

Hurst exponent to quantify the long-term memory of a time series.

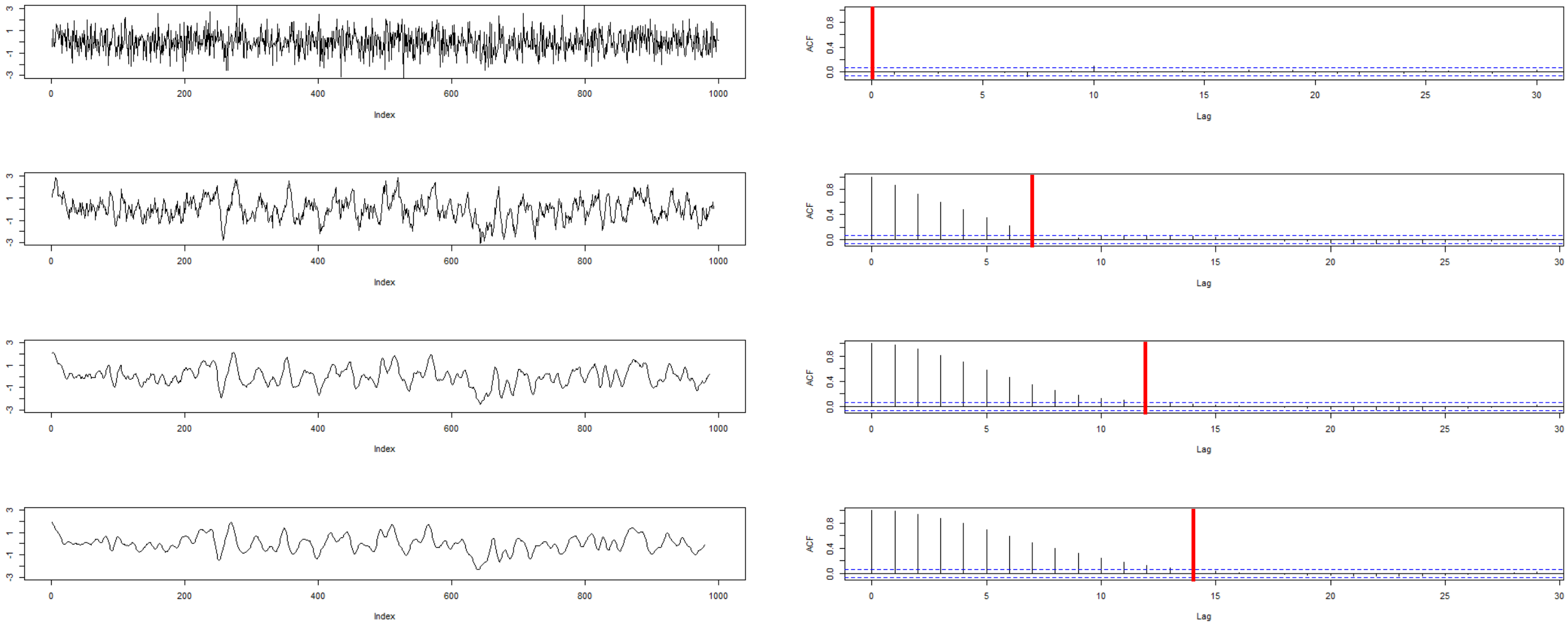
- Commonly used for modeling financial markets.
- The Hurst exponent is related to both autocorrelation and fractal dimension.
- Compare the results to the autocorrelation with the R Stats “zoo” package by Zeileis and others.

## Power Law Model

$$\boldsymbol{E\left[\frac{R(n)}{S(n)}\right] = Cn^{H} \textbf{ as } n \to \infty}$$

$R(n)$ is range and $S(n)$ is standard deviation, $\boldsymbol{C}$ is a constant, $\boldsymbol{n}$ is the number of samples in segment, and $\boldsymbol{E}$ is expected value over all partial time series lengths, $\boldsymbol{n}$.

**Time Series** **Autocorrelation** **Hurst (H)**



**0.54**

**0.69**

**0.71**

**0.73**

**0.0 < Hurst < 0.5 – long-term switching between high and low, 0.5 < Hurst < 1.0 – long-term correlation.**

# Time Series Stationarity

## Stationarity Definition:

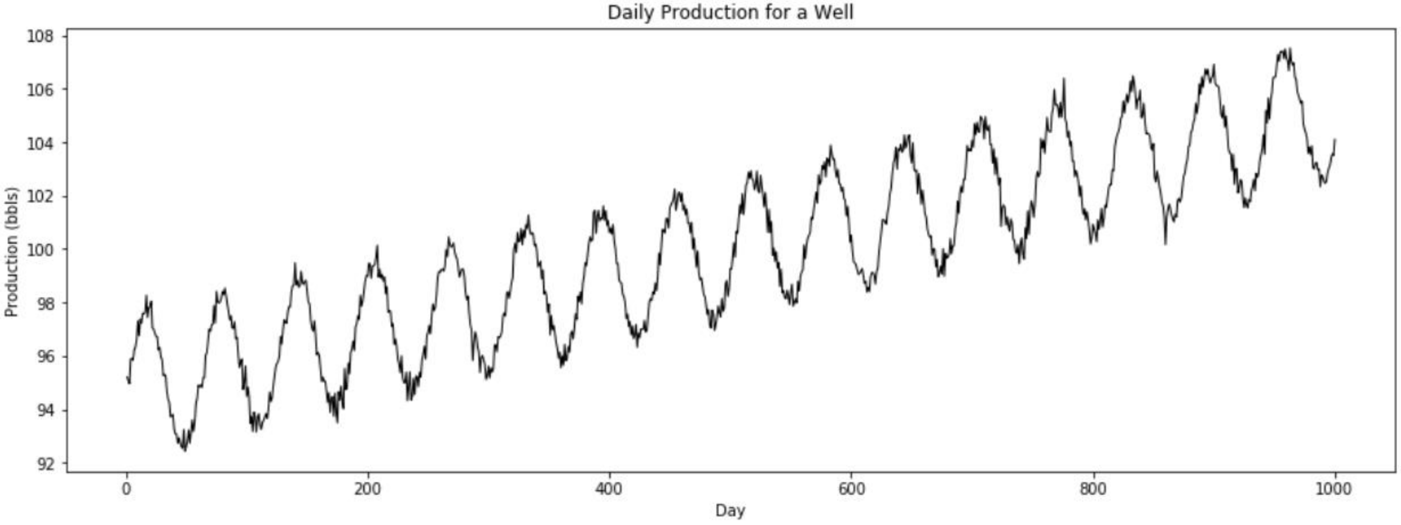
- The metric of interest is invariant under translation

- Stationarity in the mean, variance and entire CDF.
  - stationary mean, $m_x(\mathbf{u}) = m_x$
  - stationary variance, $\sigma_x^2(\mathbf{u}) = \sigma_x^2$
  - stationary CDF, $F_x(x; \mathbf{u}) = F_x(x)$
  - etc.

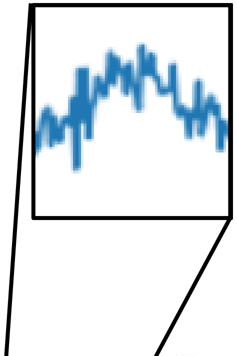
# Time Series Stationarity

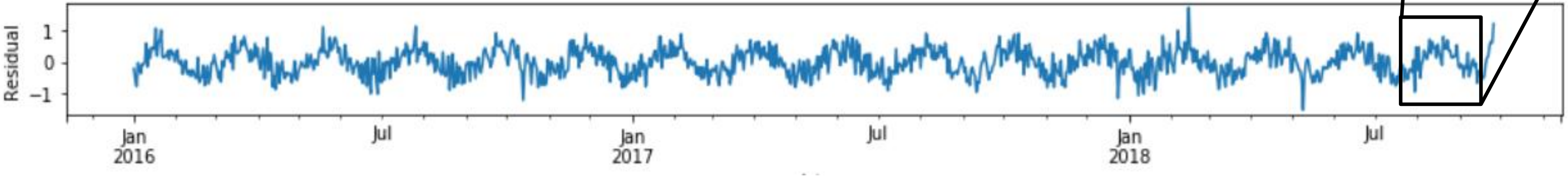
## Detecting Stationarity:

- Ocular inspection



- Dependent on the scale of observation

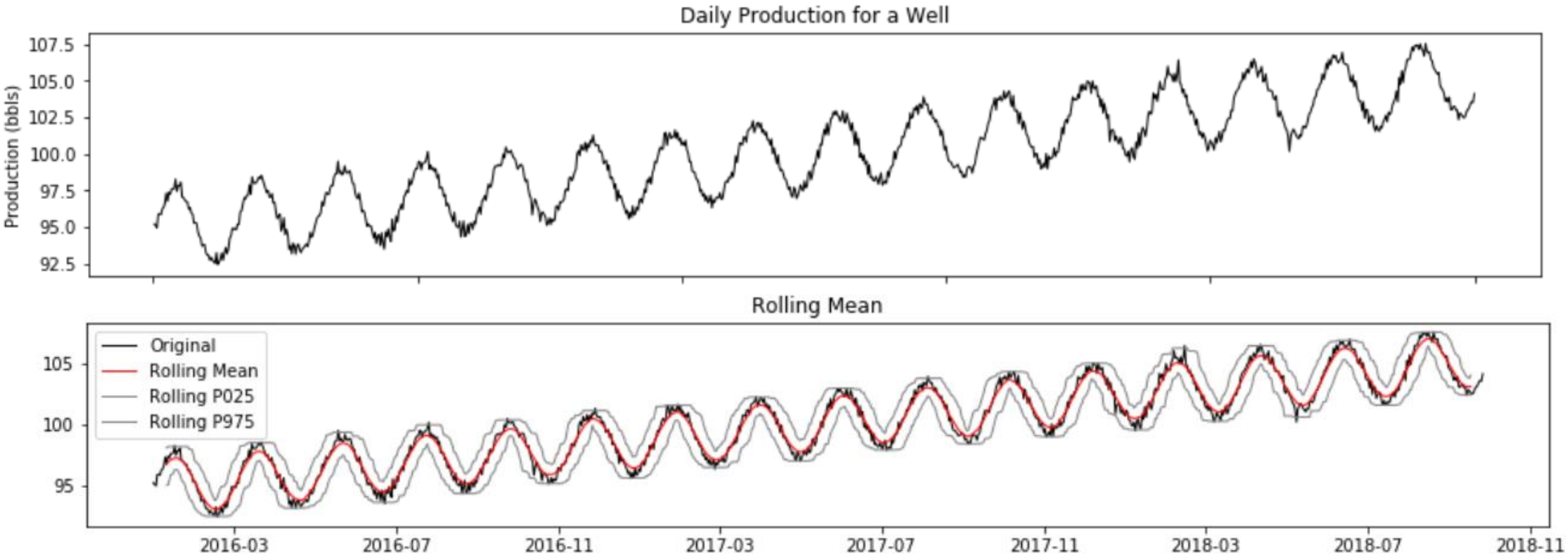




# Time Series Stationarity

## Detecting Stationarity:

- Local Statistics / Moving Window / Time Rolling Statistics
  - original dataset – well production



Rolling Mean, P025, P975 over 20-day window.

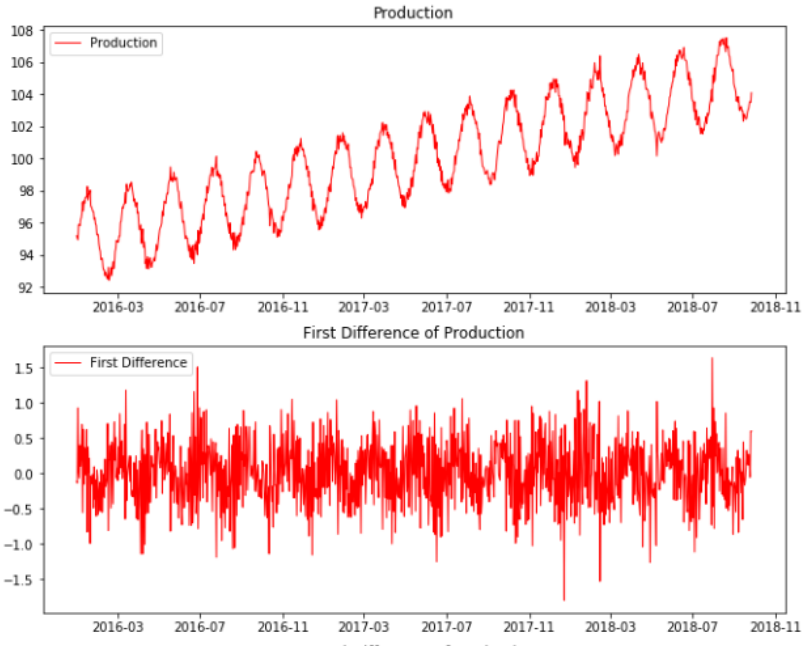
  - observe the local statistics calculated over a moving window

• $H_1$=H1= time series is stationary

# Time Series Stationarity

## Detecting Stationarity:

- Dickey-Fuller Hypothesis Test for a constant mean
  - $H_0$ = time series is non-stationary
  - $H_1$ = time series is stationary
  - if test statistic is less than the critical value at the alpha level (1 – significance level) or p-value less than alpha, then reject the null hypothesis and adopt the hypothesis that the time series is stationary



```
Dickey-Fuller Test Results:
Test Statistic                 -1.225206
p-value                         0.662660
#Lags Used                     22.000000
Number of Observations Used   977.000000
Critical Value (1%)            -3.437061
Critical Value (5%)            -2.864503
Critical Value (10%)           -2.568348
```

$H_0$ = time series is non-stationary

```
Dickey-Fuller Test Results - Second Differenced Production:
Test Statistic                -6.114406e+00
p-value                        9.168133e-08
#Lags Used                     2.200000e+01
Number of Observations Used    9.770000e+02
Critical Value (1%)           -3.437061e+00
Critical Value (5%)           -2.864503e+00
Critical Value (10%)          -2.568348e+00
```

$H_1$ = time series is stationary

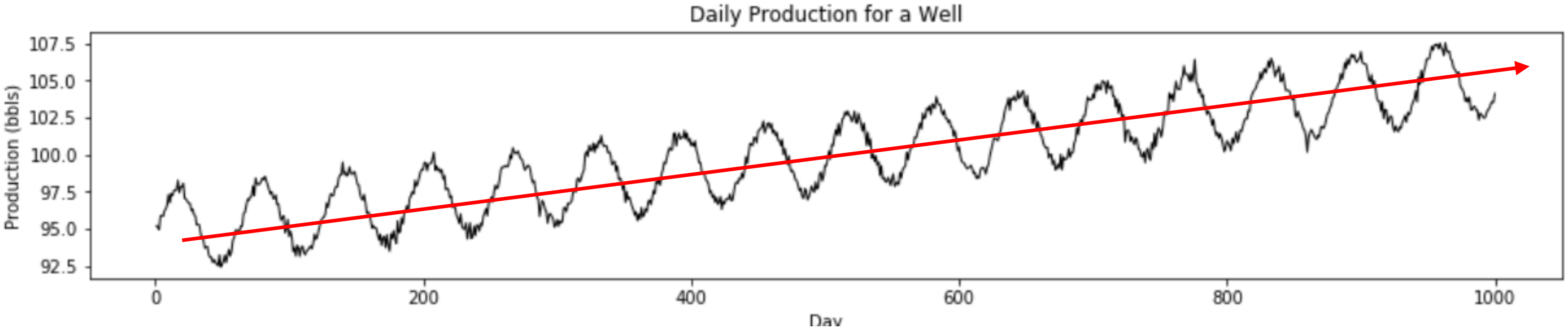
Production and first difference with Dickey-Fuller test.

- $H_1$=H1= time series is stationary

# Time Series Stationarity

## Dealing with Nonstationarity:

- For inference and prediction, we must deal with nonstationarity
- May include important information
- Could be modeled deterministically to improve prediction accuracy

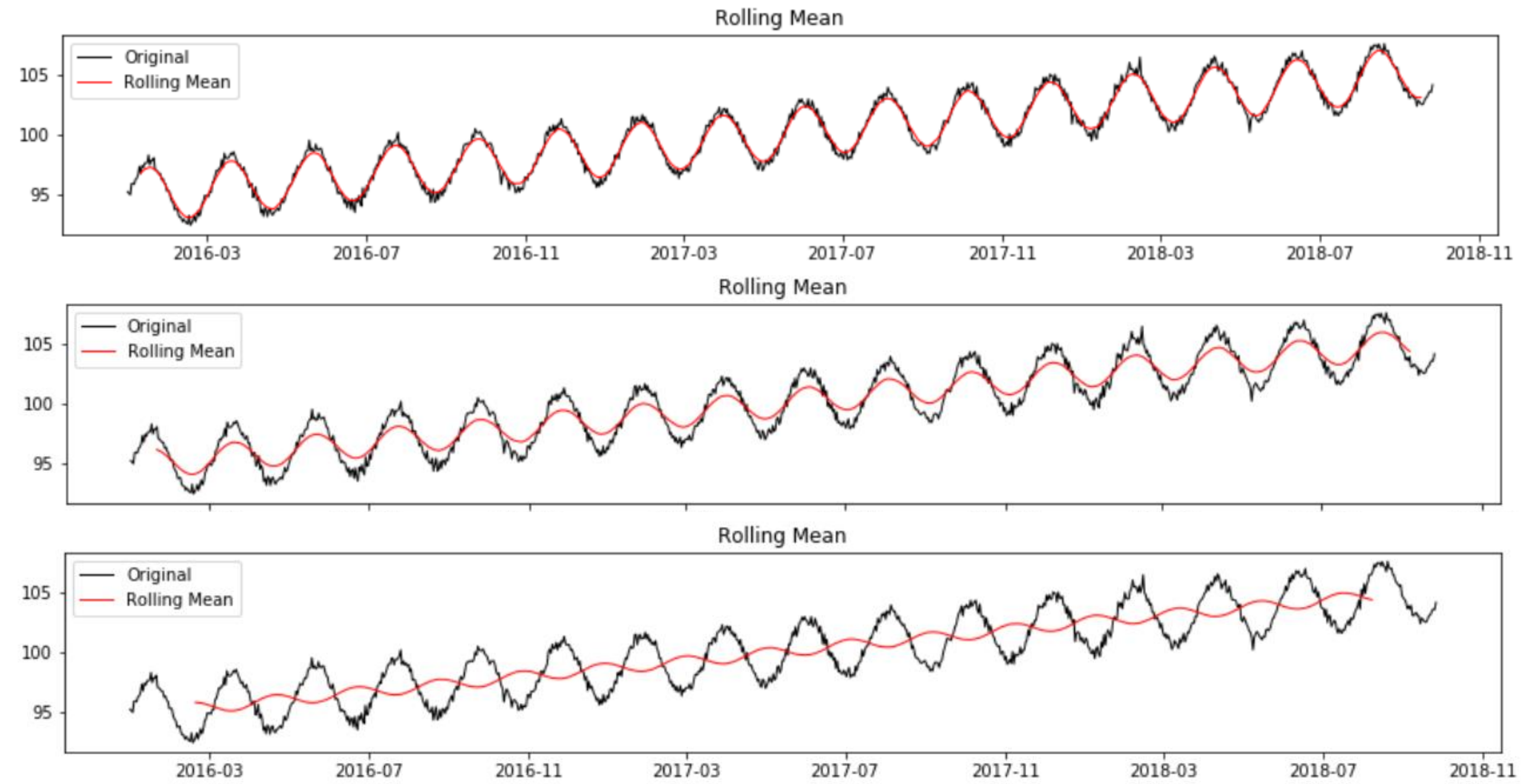


•$H_1$=H1= time series is stationary

# Time Series Stationarity

## Trend Models:

- A variety of methods to fit trends to characterize nonstationary behavior



20 day window

40 day window

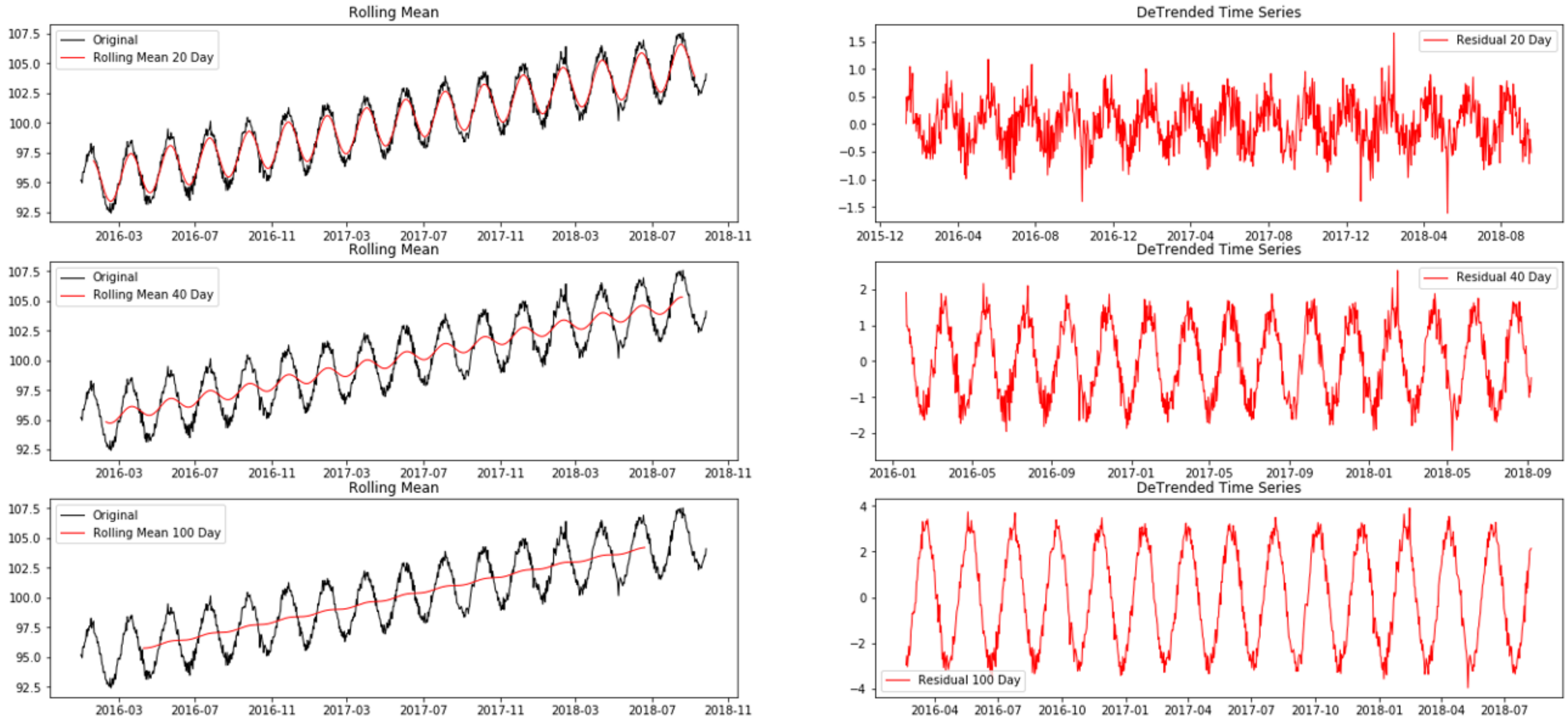
100 day window

•$H_1$=H1= time series is stationary

# Time Series Stationarity

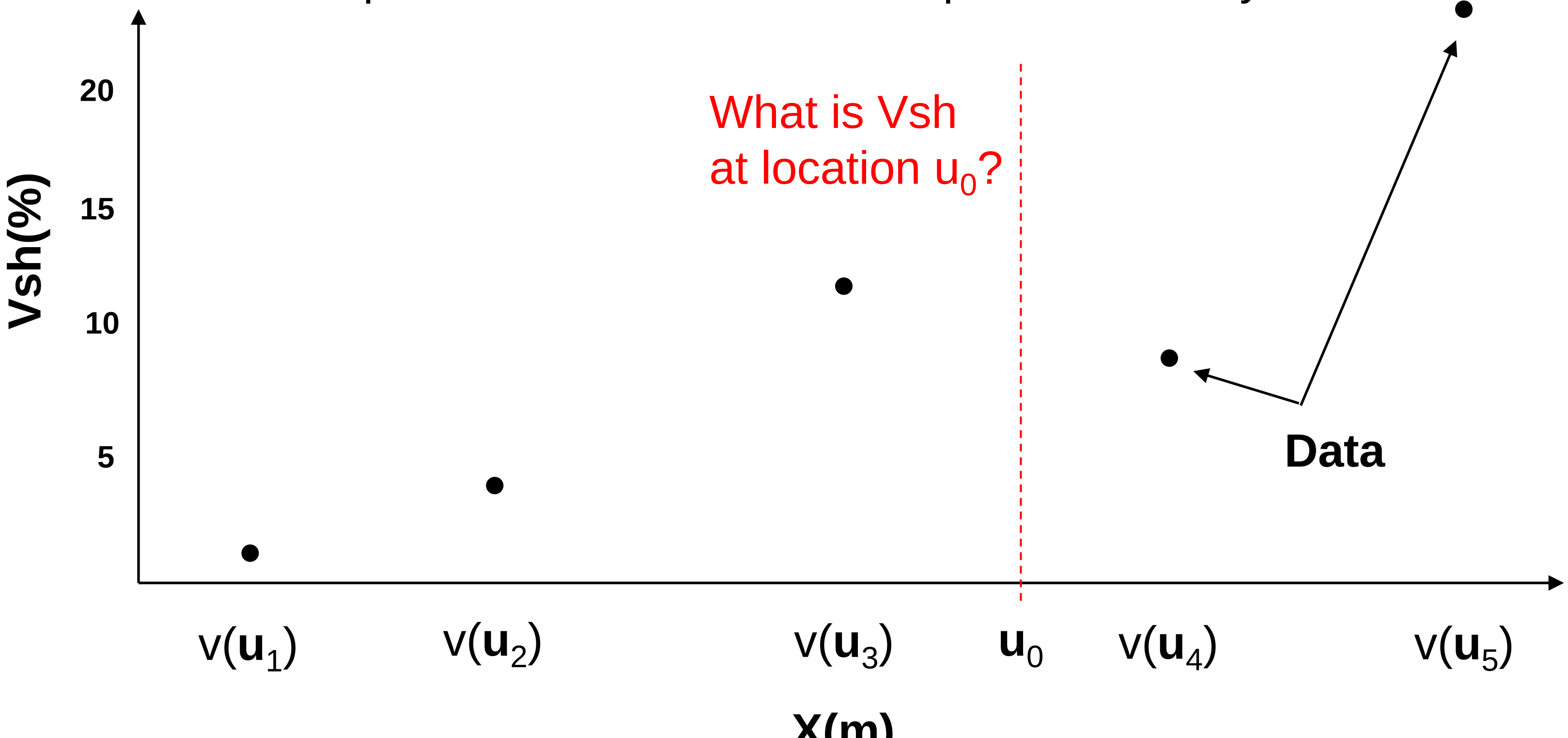
## Trend Models:

- Model trend and subtract, the remainder is the residual



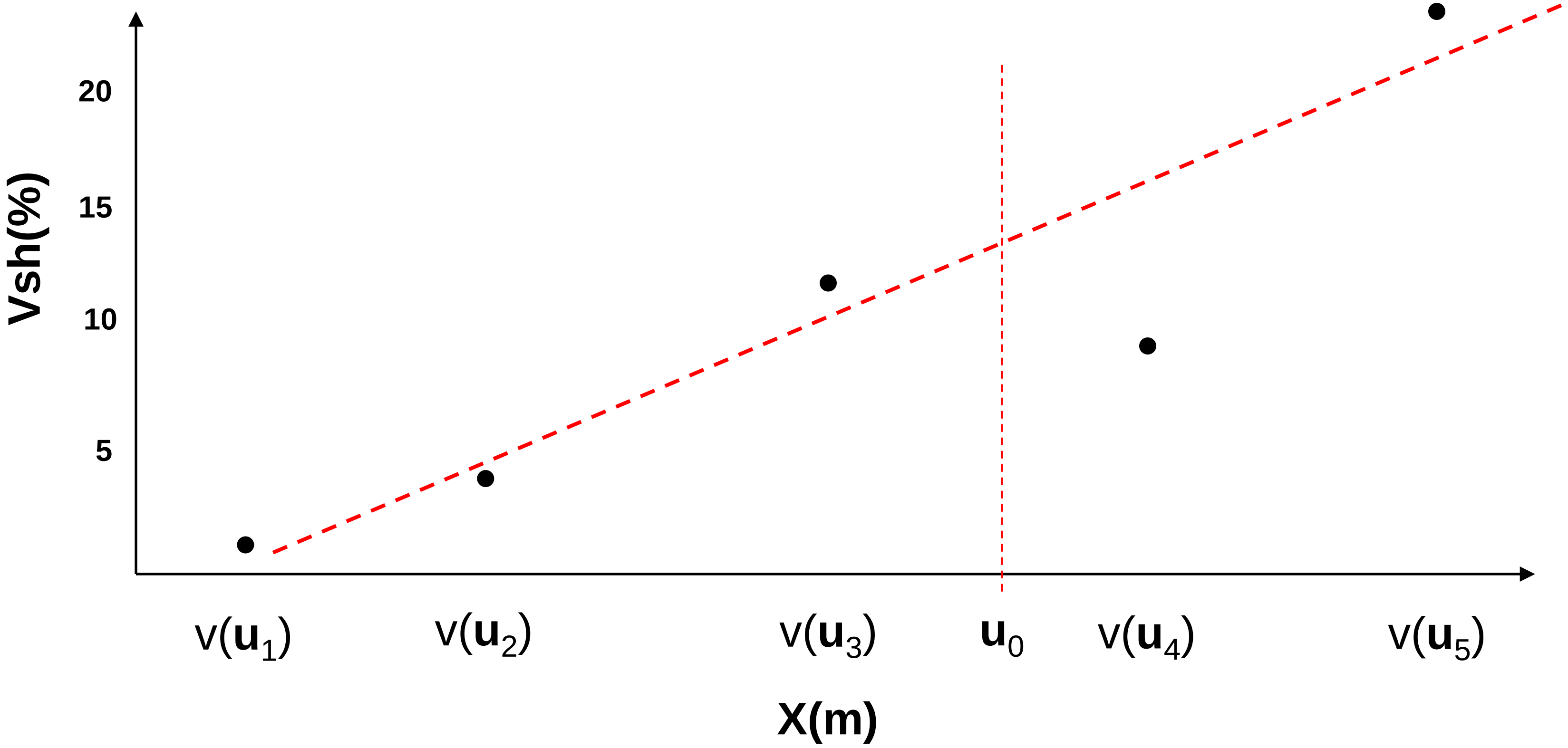
# Trend and Residual Method

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - In the presence of significant nonstationarity we would not rely 100% for spatial estimation on data + spatial continuity model



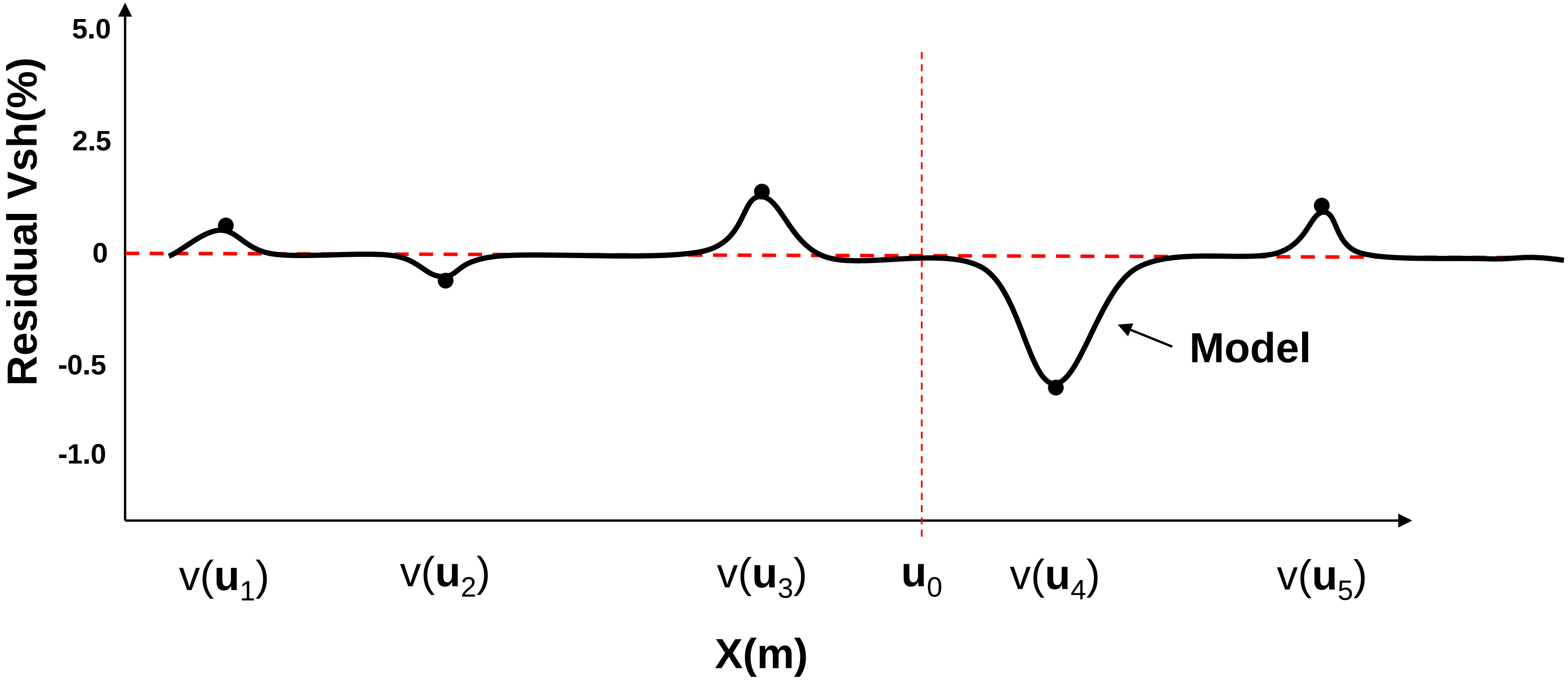
# Trend and Residual Method

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - – If we observe a trend, we should model the trend.



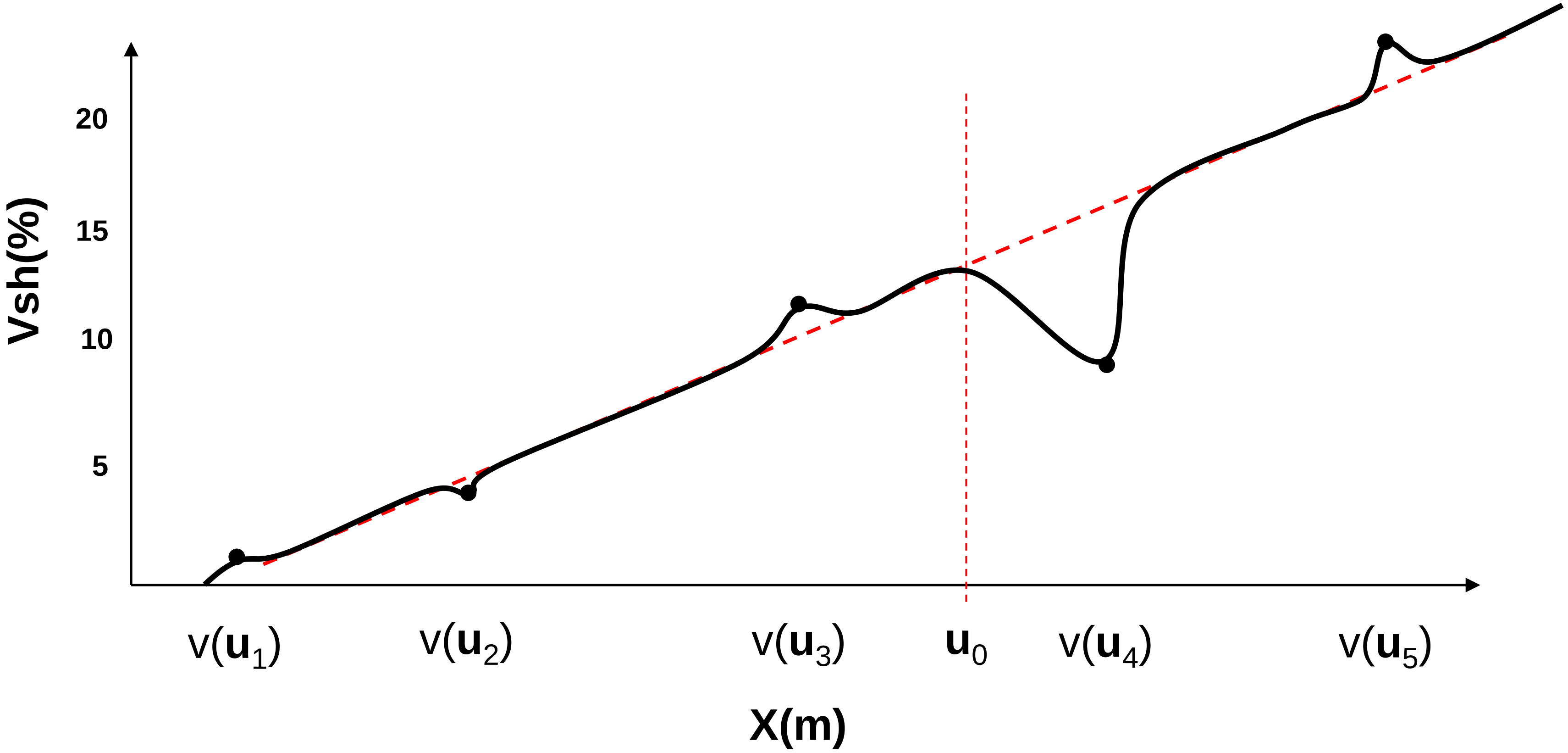
# Trend and Residual Method

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - – Then model the residuals.



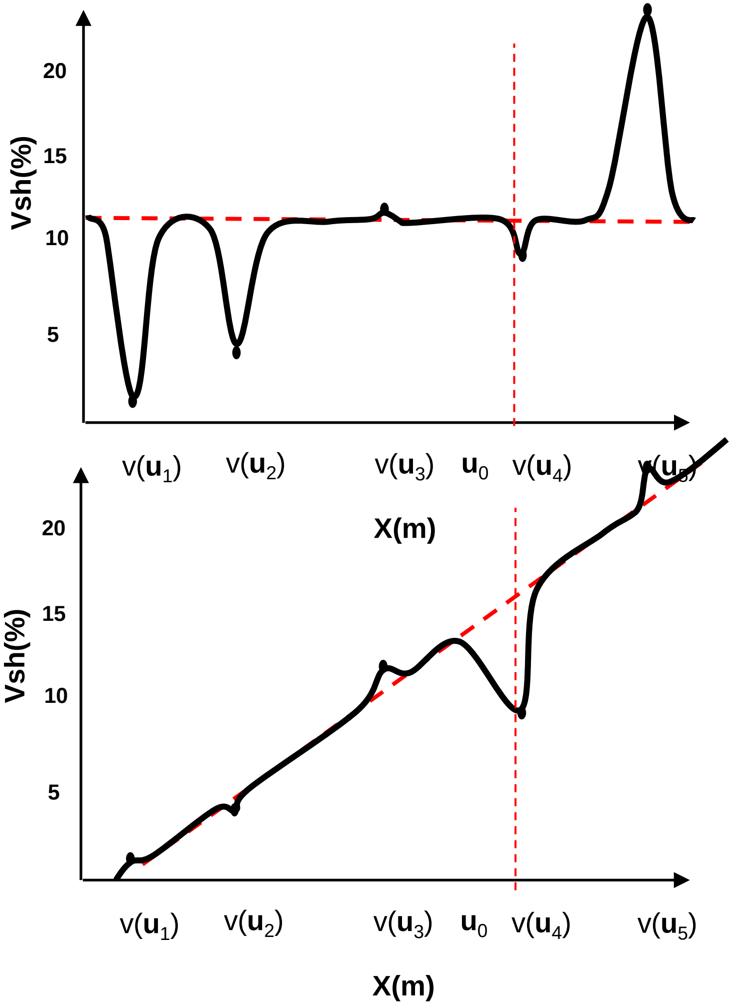
# Trend and Residual Method

- Geostatistical spatial estimation methods will make an assumption concerning stationarity
  - – After modeling, add the trend back to the modelled residuals



# Trend and Residual Method

- How bad could it be if we did not model a trend?
- Geostatistical estimation would assume stationarity* and away from data we would estimate with the global mean (simple kriging)!



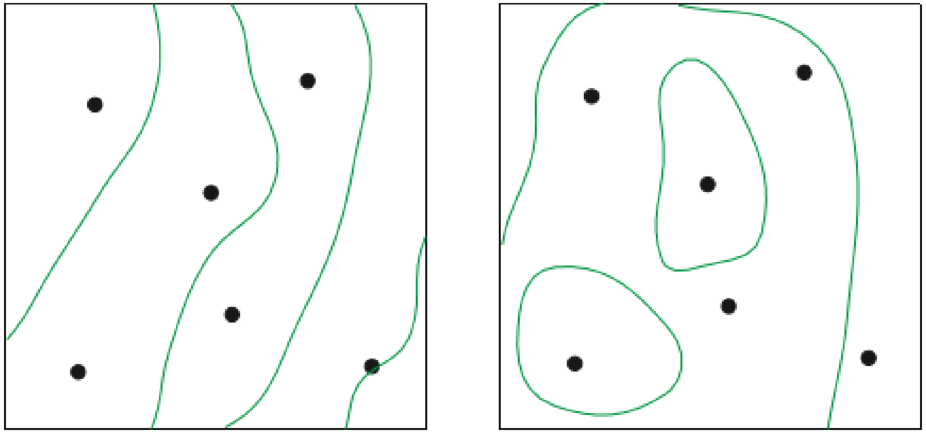
**Model with stationary mean + data.**

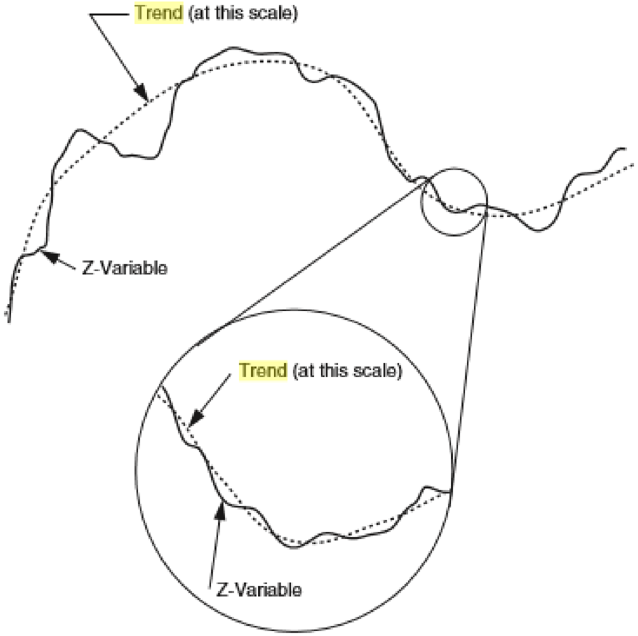
**Model with mean trend model and residual + data.**

***stationarity decision depends on type of method.**

# Trend and Residual Method

- Trend Modeling
  - We must identify and model trends / nonstationarities





- While we discuss data-driven trend modeling here any **trend modeling should include data integration** over the entire asset team
  - Geology
  - Geophysics
  - Petrophysics
  - Reservoir Engineering

Images from Pyrcz and Deutsch (2014)

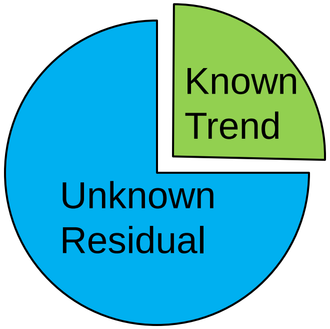
# Trend and Residual Method

- Any variance in the trend is removed from the residual:

$$\sigma_X^2 = \sigma_{X_t}^2 + \sigma_{X_r}^2 + 2C_{X_t,X_r}$$

  - if the $X_t \perp\!\!\!\perp X_r$, $C_{X_t,X_r} = 0$

$$\sigma_{X_r}^2 = \sigma_X^2 - \sigma_{X_t}^2$$



  - So if $\sigma_X^2$ is the total variance (variability), and $\sigma_{X_t}^2$ is the variability that is deterministically modelled, treated as known, and $\sigma_{X_r}^2$ is the component of the variability that is treated as unknown.
  - Result: the more variability explained by the trend the less variability that remains as uncertain.

# Additivity of Variance for Decomposing Trend and Residual

Can we partition variance of random variable Z between trend (X) and residual (Y)?

- Start with the variance of Z: $\sigma^2_Z = E(Z^2) - [E(Z)]^{\,2}$

- Substitute: $Z = \mathrm{X} + \mathrm{Y}$

$$\sigma^2_{X+Y} = E\left((X+Y)^2\right) - [E(X+Y)]^{\,2}$$

$$\sigma^2_{X+Y} = E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^{\,2}$$

$$\sigma^2_{X+Y} = E(X^2) + 2E(XY) + E(Y^2) - \left(E(X)^2 + 2E(X)E(Y) + E(Y)^2\right)$$

$$\sigma^2_{X+Y} = \underset{\sigma^2_X}{\boxed{E(X^2) - E(X)^2}} + \underset{\sigma^2_Y}{\boxed{E(Y^2) - E(Y)^2}} + 2\underset{C_{XY}(0)}{\boxed{\left(E(XY) - E(X)E(Y)\right)}}$$

- Note covariance: $C_{XY} = E(XY) - E(X)E(Y)$

$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2C_{XY}(0)$ ◁ **Additivity of variance**

- If the $X\_\_Y$, $C_{XY}(0)$ = 0, then $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ ◁ **In practice**

# Definition
# Deterministic Model

- Model that assumes perfect knowledge, without uncertainty

- Based on knowledge of the phenomenon or trend fitting to data
- Most subsurface models have a deterministic component (trend) to capture expert knowledge and to provide a stationary residual for geostatistical modeling.

- $H_1$=H1= time series is stationary

# Time Series Stationarity

## Time Series Differencing:

- A method to remove nonstationarity

- The differencing calculation proceeds as follows:

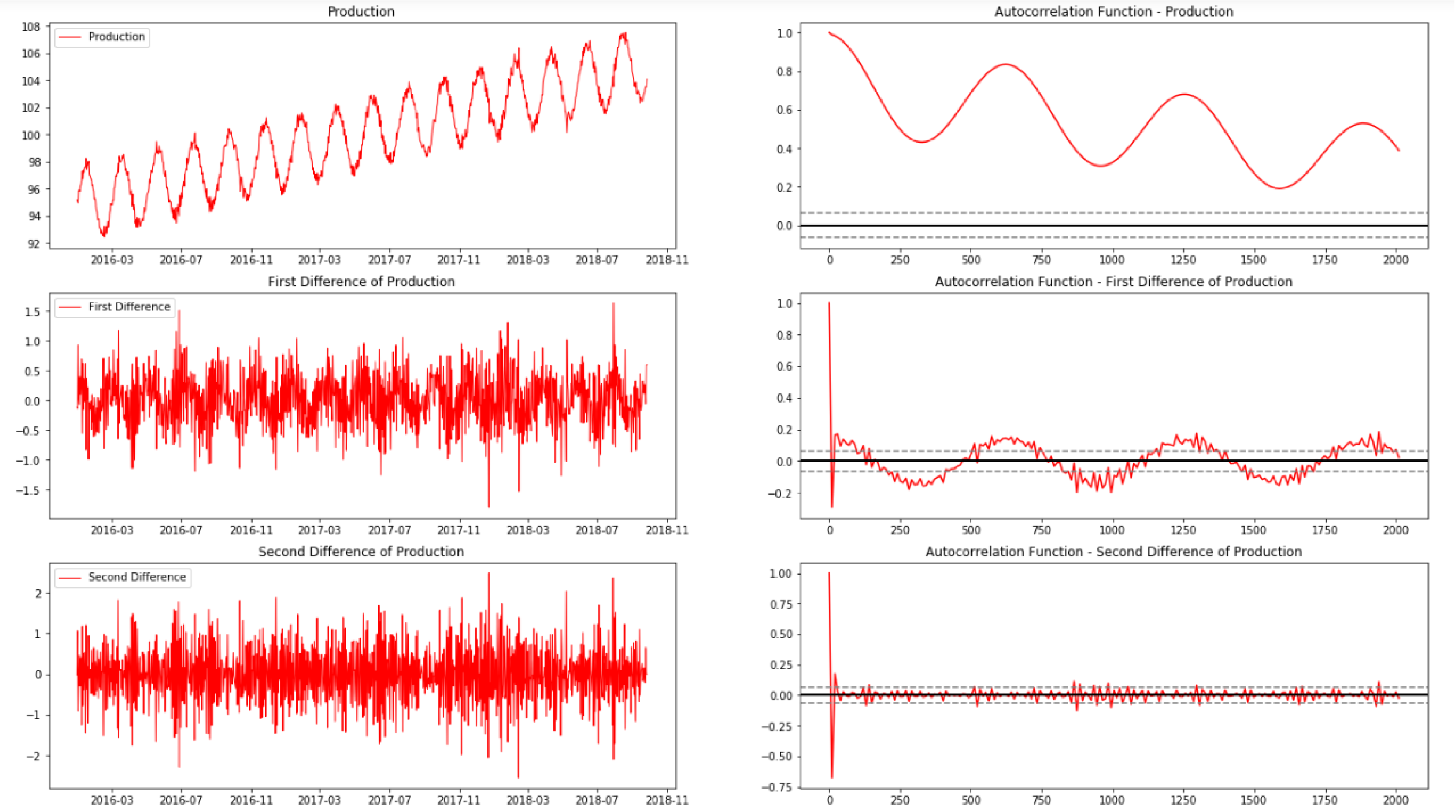
$$\text{First Order: } y^1(u_i) = z(u_{i+1}) - z(u_i)$$

- Differencing may be applied multiple times, until the result is stationary.

$$\text{Second Order: } y^2\ (u_i) = y^1(u_{i+1}) - y^1(u_i)$$

# Time Series Stationarity

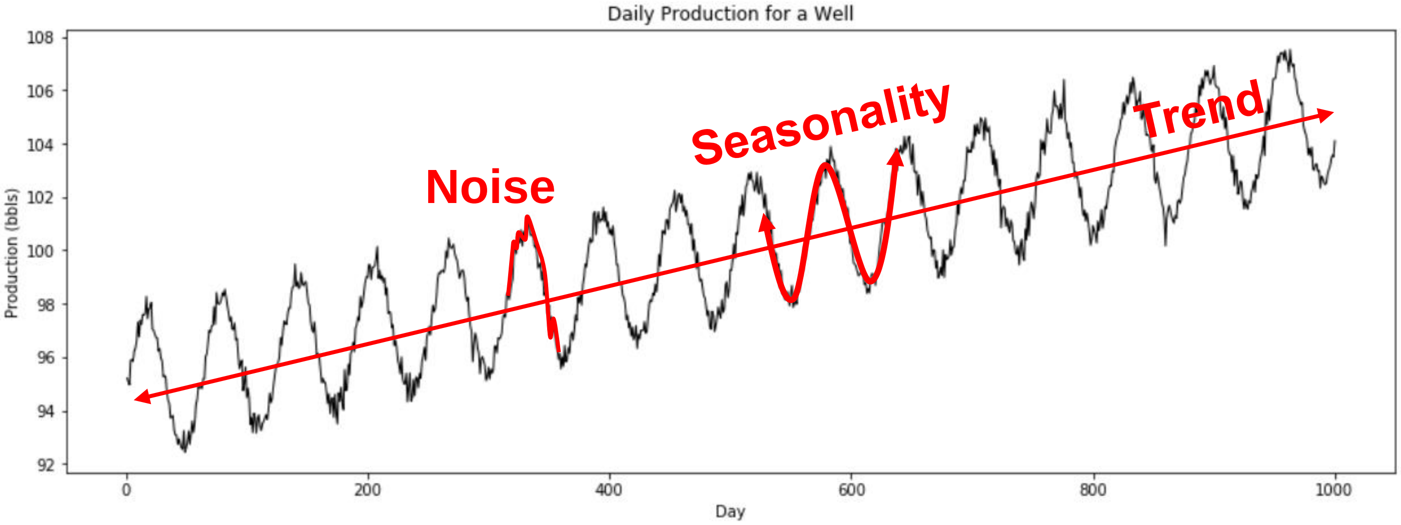
Time series differencing example with autocorrelation:



# Time Series Decomposition

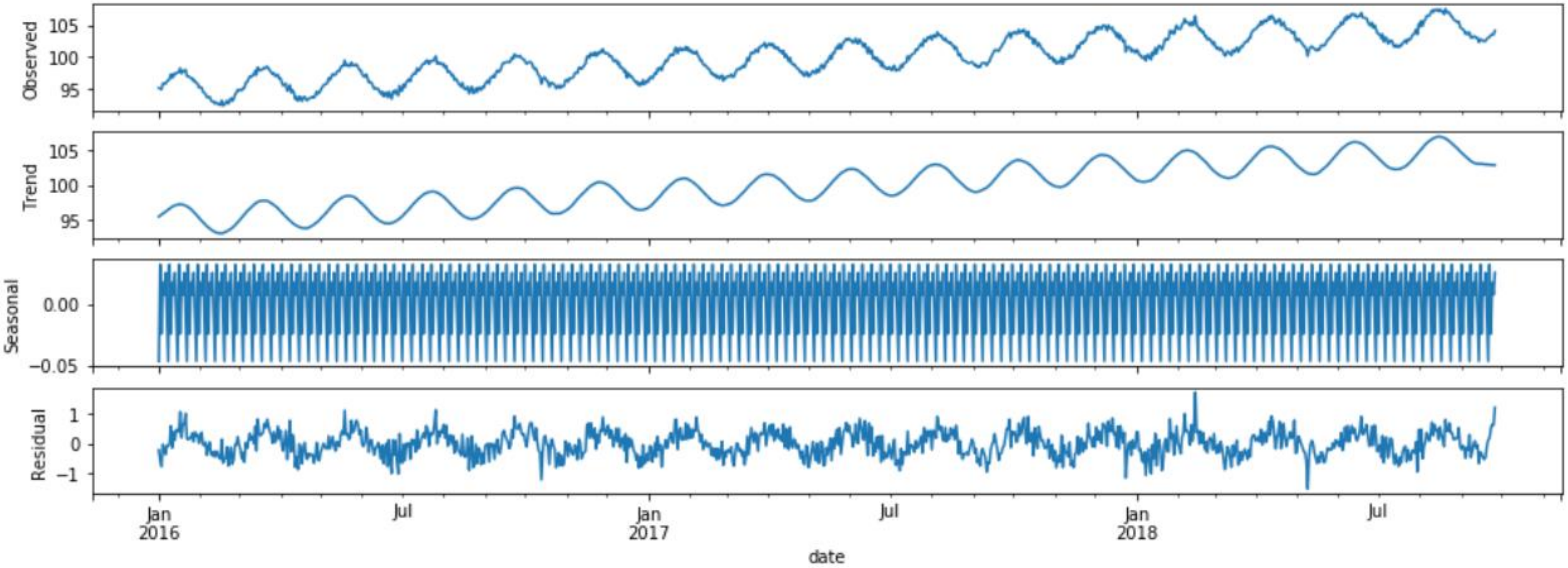
Method to decompose a time series into multiple additive components.

- The following components are common:
  1. trend – mid to long range autocorrelation structures
  2. seasonality – regular cycles
  3. noise / residual – the remainder after removal of the above



# Time Series Decomposition

Python StatsModel has a function to automatically decompose and model trend, seasonality and residual.



- failed to capture the cycles, included in the trend component

# PGE 383

## Time Series Analysis

- **Time Series Model**

# Time Series Modeling

ARIMA (Auto-regressive Integrated Moving Average) Model for Time Series

- **inference** – learning about the time series
- **prediction** - forecasting

- number of auto-regressive terms (p) is the number of lags for autocorrelation - we find this from the number of significant lags in the partial autocorrelation. There was only one lag (see above).

- number of moving average terms (q) is the number of differencing required remove the trend. From above we demonstrated that we had a stationary dataset after the first difference, but we improved the removal of the cycles after the second difference, let's use second difference.

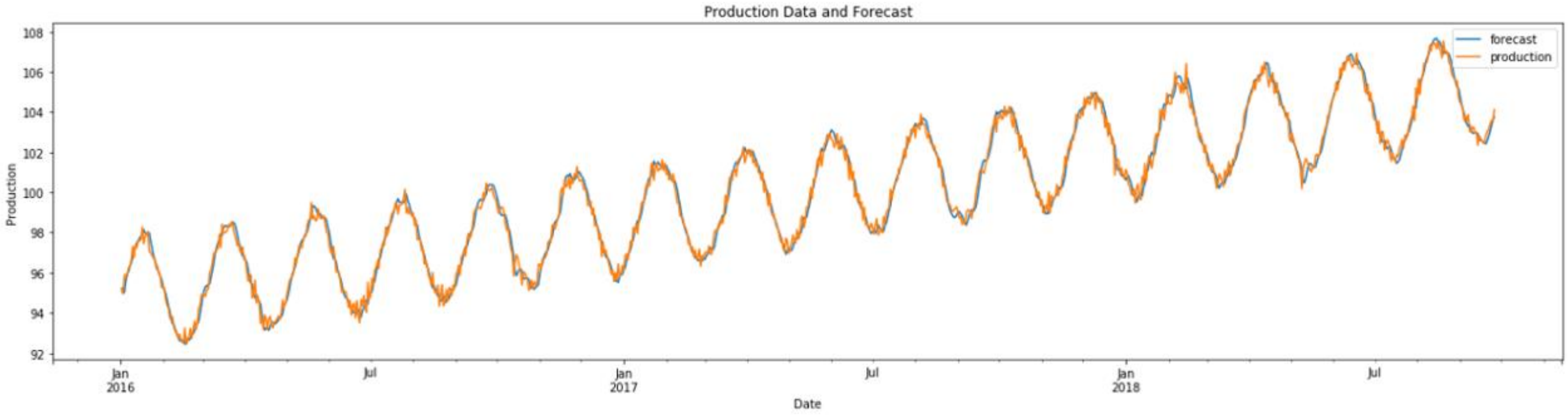
- number of nonseasonal differences (d) is the seasonal effect, our dataset did not show any significant seasonal cyclicity in the decomposition above, so we will just set it to one.

# Time Series Modeling

ARIMA (Auto-regressive Integrated Moving Average) Model for Time Series Example

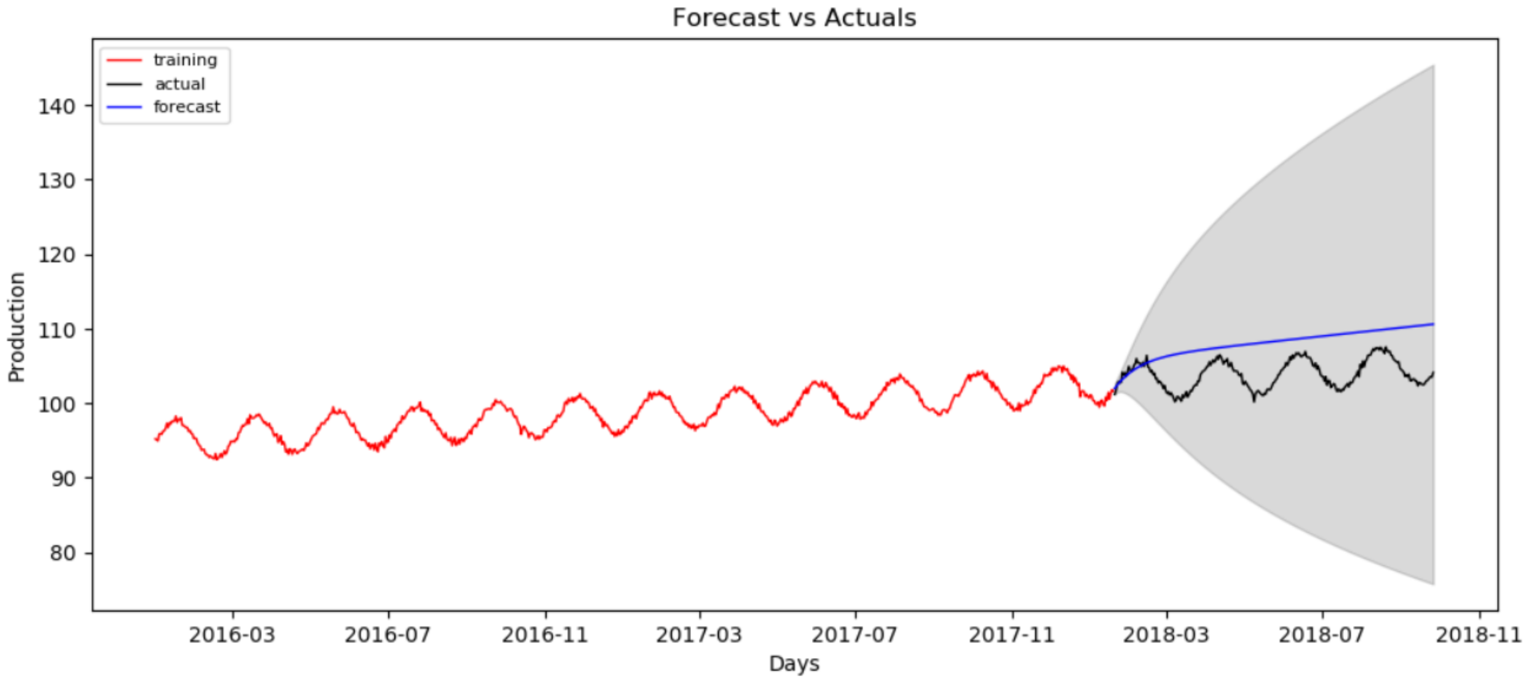
- This is the lag 1 forward forecast model

$$z(u_{i+1}) = f(z(u_1), \ldots, z(u_i))$$



# Time Series Modeling

ARIMA (Auto-regressive Integrated Moving Average) Model for Time Series Example



- Forecasting over 250 days with 750 days of training.
  - captured the long range trend and the short range autocorrelation
  - failed to capture the seasonality / cyclicity

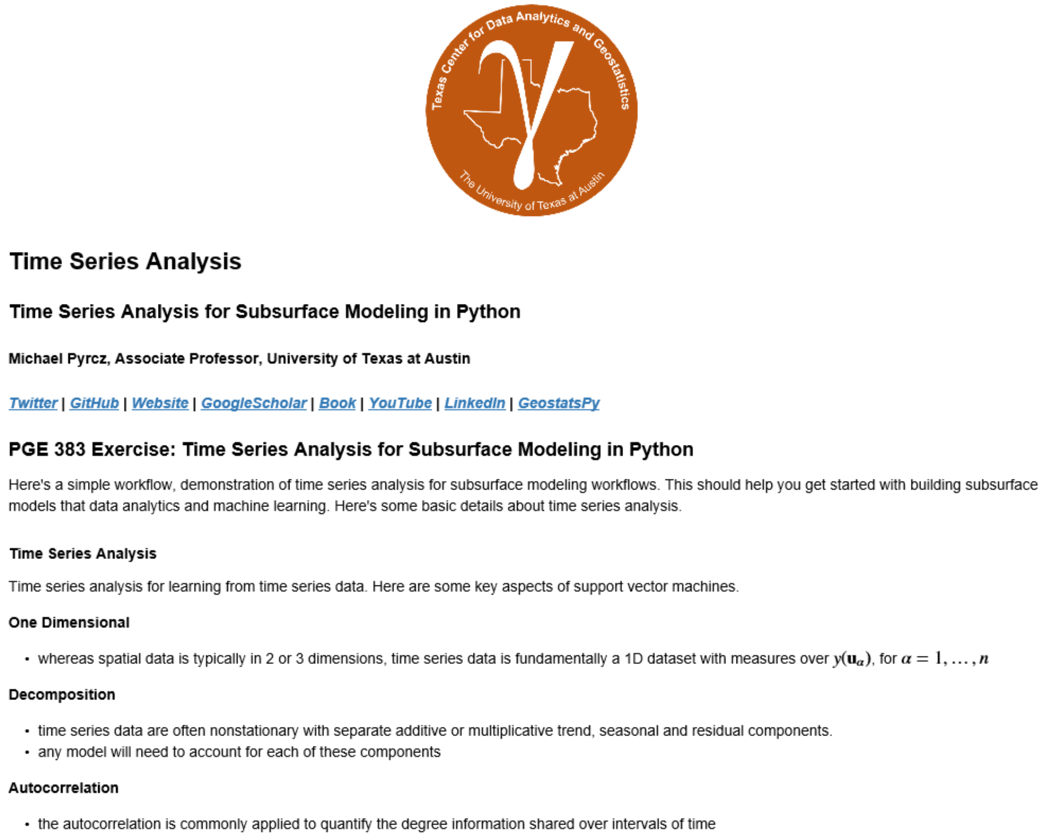
# PGE 383
## Time Series Analysis

- Time Series Hands-on

Michael Pyrcz, The University of Texas at Austin

# Time Series Analysis Demonstration

Demonstration workflow with time series analysis.



## Time Series Analysis

### Time Series Analysis for Subsurface Modeling in Python

**Michael Pyrcz, Associate Professor, University of Texas at Austin**

*Twitter* | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn* | *GeostatsPy*

### PGE 383 Exercise: Time Series Analysis for Subsurface Modeling in Python

Here's a simple workflow, demonstration of time series analysis for subsurface modeling workflows. This should help you get started with building subsurface models that data analytics and machine learning. Here's some basic details about time series analysis.

**Time Series Analysis**

Time series analysis for learning from time series data. Here are some key aspects of support vector machines.

**One Dimensional**

- whereas spatial data is typically in 2 or 3 dimensions, time series data is fundamentally a 1D dataset with measures over $y(\mathbf{u}_\alpha)$, for $\alpha = 1, \dots, n$

**Decomposition**

- time series data are often nonstationary with separate additive or multiplicative trend, seasonal and residual components.
- any model will need to account for each of these components

**Autocorrelation**

- the autocorrelation is commonly applied to quantify the degree information shared over intervals of time

File SubsurfaceDataAnalytics_TimeSeries.ipynb at https://git.io/fj6mZ.

# PGE 383

## Time Series Analysis

- Time Series Data
- Time Series Analysis
- Time Series Model
- Time Series Hands-on

Michael Pyrcz, The University of Texas at Austin