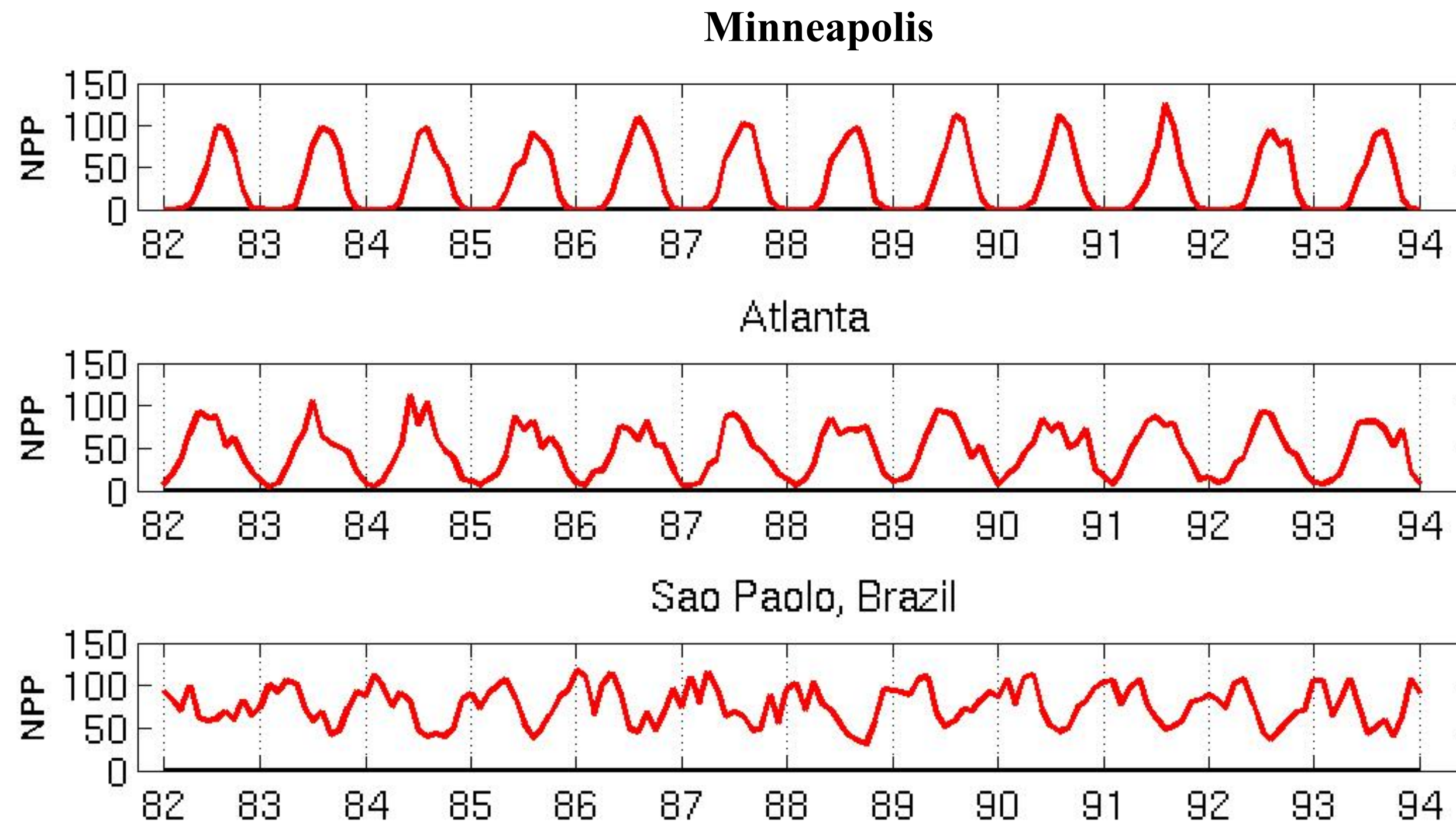


# Attribute Transformation

---

- An **attribute transform** is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
  - **Normalization**
    - ◆ Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - ◆ Take out unwanted, common signal, e.g., seasonality
  - In statistics, **standardization** refers to subtracting off the means and dividing by the standard deviation

# Example: Sample Time Series of Plant Growth



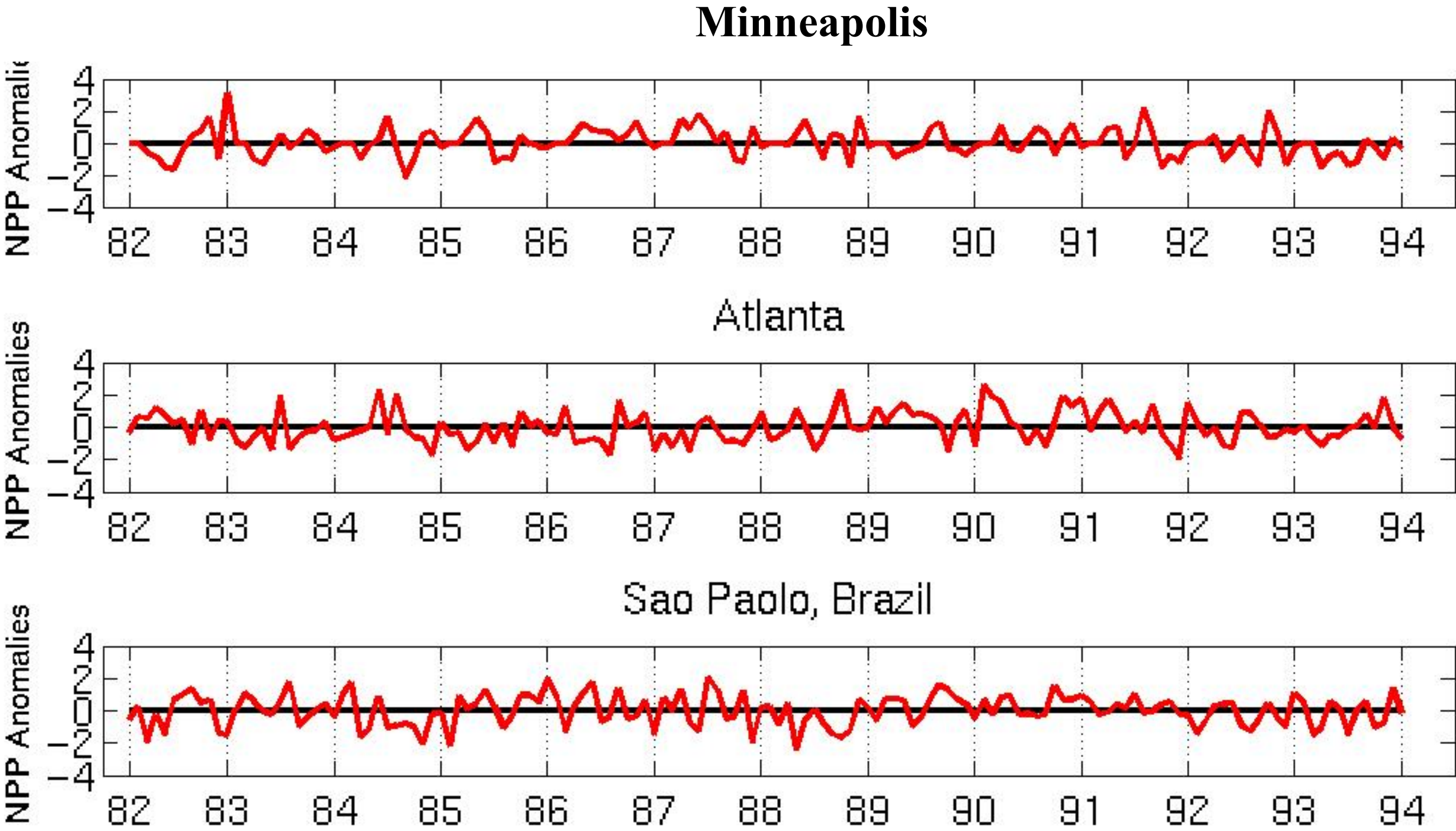
**Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.**

## Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000



# Seasonality Accounts for Much Correlation



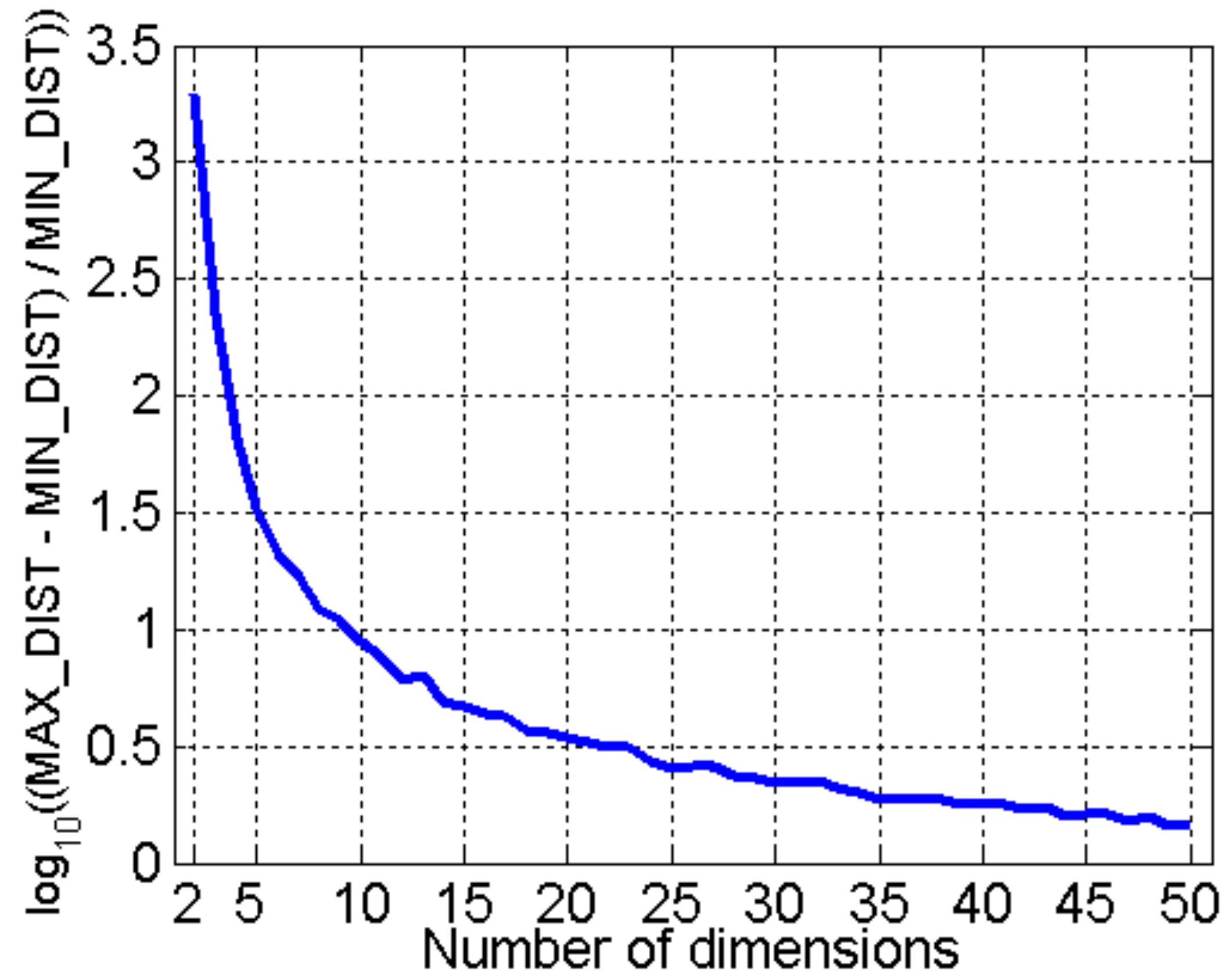
Normalized using  
monthly Z Score:  
Subtract off monthly  
mean and divide by  
monthly standard  
deviation

## Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	1.0000

# Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

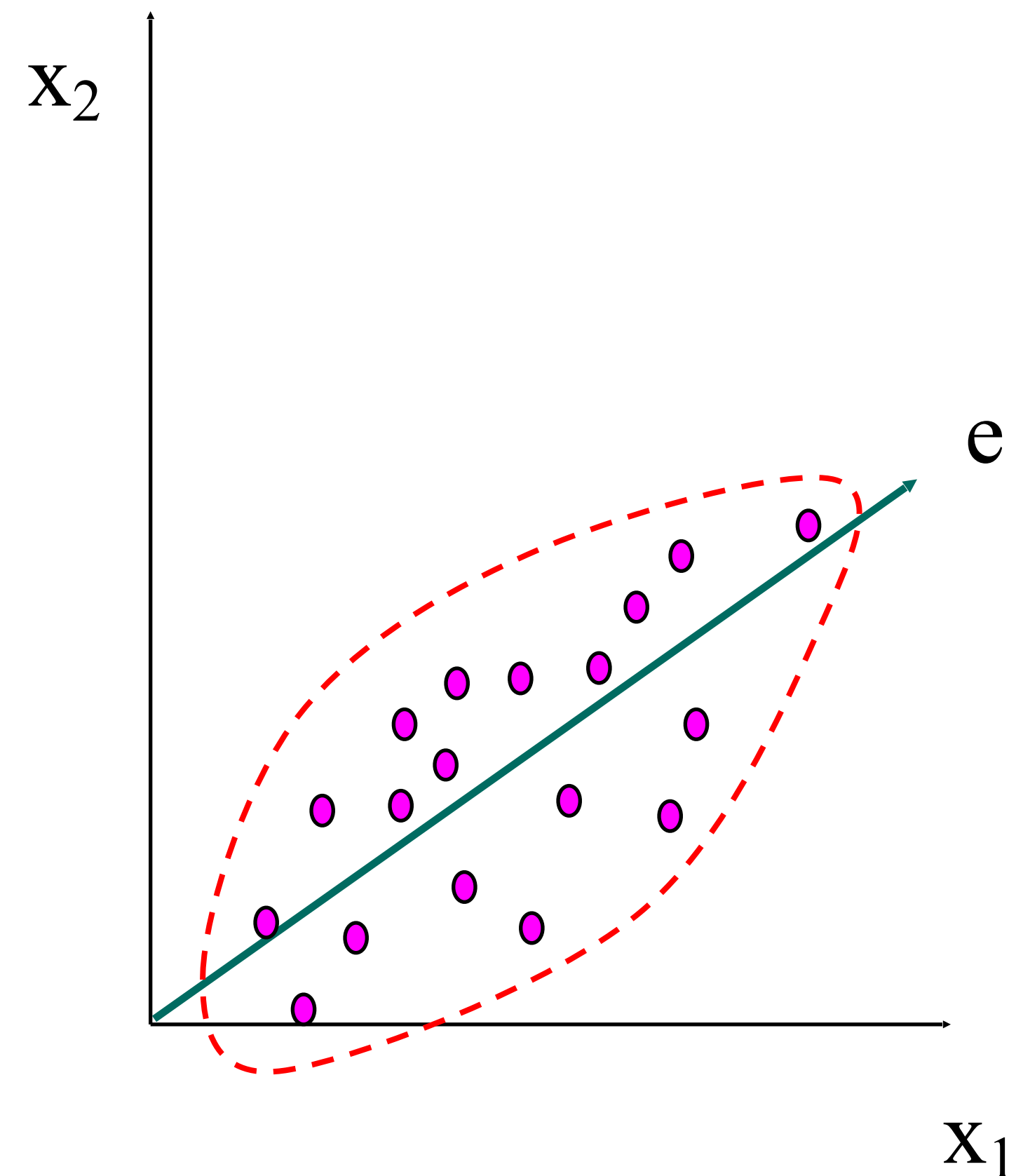
# Dimensionality Reduction

---

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques (e.g. t-SNE)

# Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data





# Dimensionality Reduction: PCA

---

256



# Feature Subset Selection

---

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification



# Feature Creation

---

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis