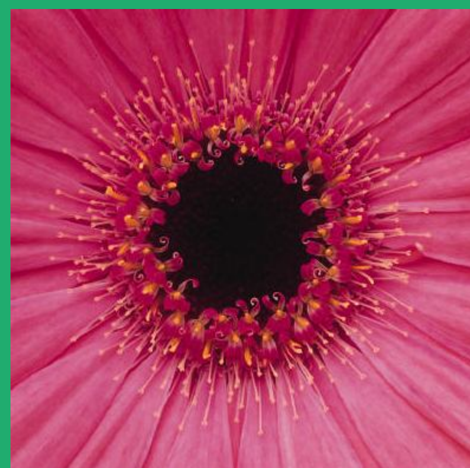
Unit-V

Mining time-series data



Dr. K.RAGHAVA RAO

Professor of CSE

Dept. of MCA

KL University

krraocse@gmail.com

http://datamining.blog.com

Mining Time-Series Data

- A **time series** is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals
- **Time series analysis:** A subfield of statistics, comprises methods that attempt to understand such time series, often either to understand the underlying context of the data points or to make forecasts (or predictions)
- Applications
  - Financial: stock price, inflation
  - Industry: power consumption
  - Scientific: experiment results
  - Meteorological: precipitation

Categories of Time-Series Movements

- Categories of Time-Series Movements
  - Long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time
  - Cyclic movements or cycle variations: long term oscillations about a trend line or curve
    - e.g., business cycles, may or may not be periodic
  - Seasonal movements or seasonal variations
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  - Irregular or random movements
- Time series analysis: decomposition of a time series into these four basic movements
  - Additive Modal: TS = T + C + S + I
  - Multiplicative Modal: TS = T × C × S × I

Estimation of Trend Curve

- The freehand method
  - Fit the curve by looking at the graph
  - Costly and barely reliable for large-scaled data mining
- The least-square method
  - Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points
- The moving-average method

Moving Average

- Moving average of order n

$$\frac{y_1 + y_2 + \cdots + y_n}{n}, \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \cdots$$

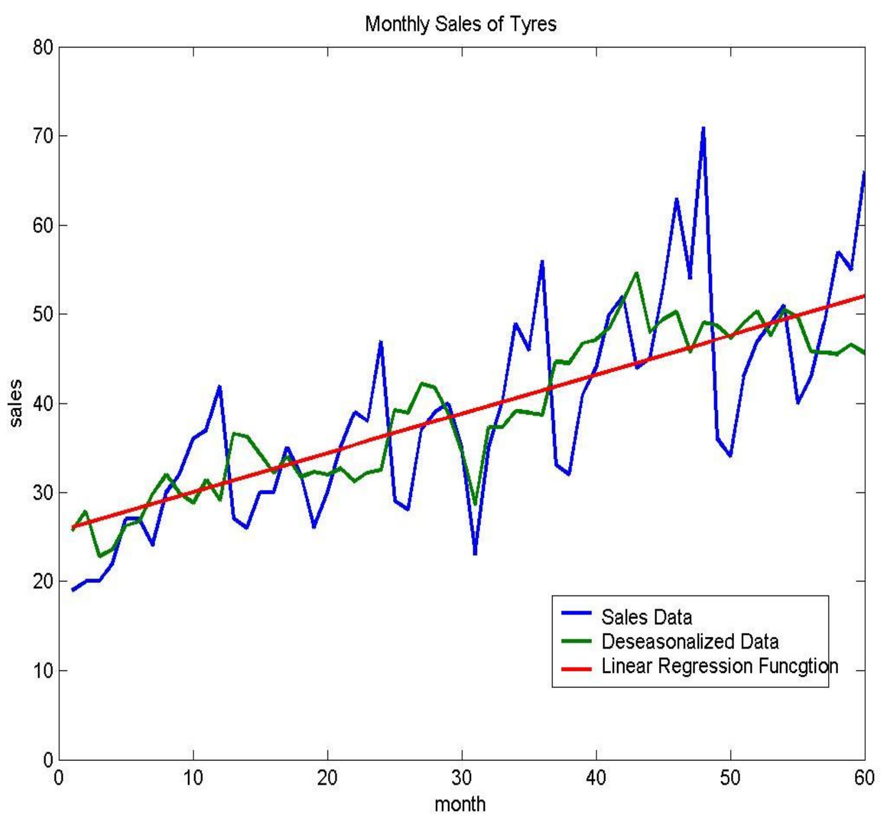
  - Smoothes the data
  - Eliminates cyclic, seasonal and irregular movements
  - Loses the data at the beginning or end of a series
  - Sensitive to outliers (can be reduced by weighted moving average)

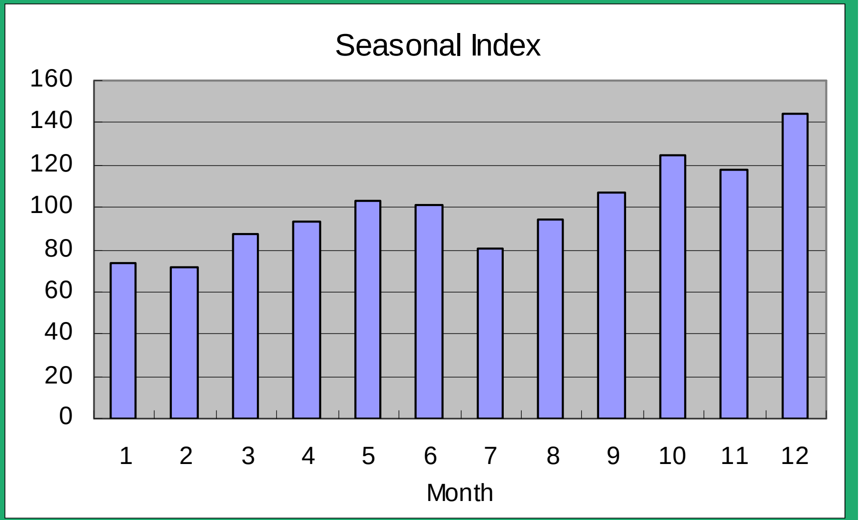
Trend Discovery in Time-Series (1): Estimation of Seasonal Variations

- Seasonal index
  - Set of numbers showing the relative values of a variable during the months of the year
  - E.g., if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months
- Deseasonalized data
  - Data adjusted for seasonal variations for better trend and cyclic analysis
  - Divide the original monthly data by the seasonal index numbers for the corresponding months

Seasonal Index

Fig-.Raw data from http://www.bbk.ac.uk/manop/man/docs/QII_2_2003%20Time%20series.pdf





Trend Discovery in Time-Series (2)

- Estimation of cyclic variations
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes
- Estimation of irregular variations
  - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality

Similarity Search in Time-Series Analysis

- Normal database query finds exact match
- Similarity search finds data sequences that differ only slightly from the given query sequence
- Two categories of similarity queries
  - Whole matching: find a sequence that is similar to the query sequence
  - Subsequence matching: find all pairs of similar sequences

Data reduction and Data Transformation

- Many techniques for signal analysis require the data to be in the frequency domain
- Usually data-independent transformations are used
  - The transformation matrix is determined a priori
    - discrete Fourier transform (DFT)
    - discrete wavelet transform (DWT)
- The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain

Multidimensional Indexing in Time-Series

- Multidimensional index construction
  - Constructed for efficient accessing using the first few Fourier coefficients
- Similarity search
  - Use the index to retrieve the sequences that are at most a certain small distance away from the query sequence
  - Perform post-processing by computing the actual distance between sequences in the time domain and discard any false matches

Subsequence Matching

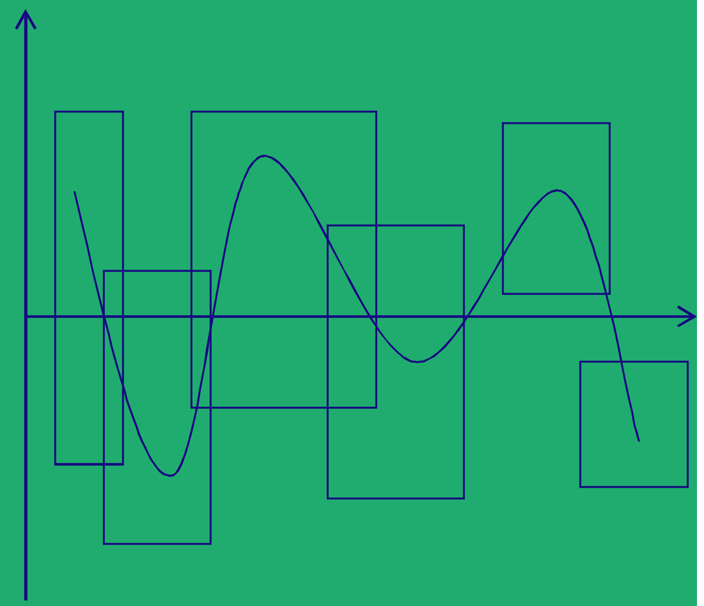
- Break each sequence into a set of pieces of window with length *w*

- Extract the features of the subsequence inside the window

- Map each sequence to a “trail” in the feature space

- Divide the trail of each sequence into “subtrails” and represent each of them with minimum bounding rectangle

- Use a multi-piece assembly algorithm to search for longer sequence matches



Analysis of Similar Time Series Methods

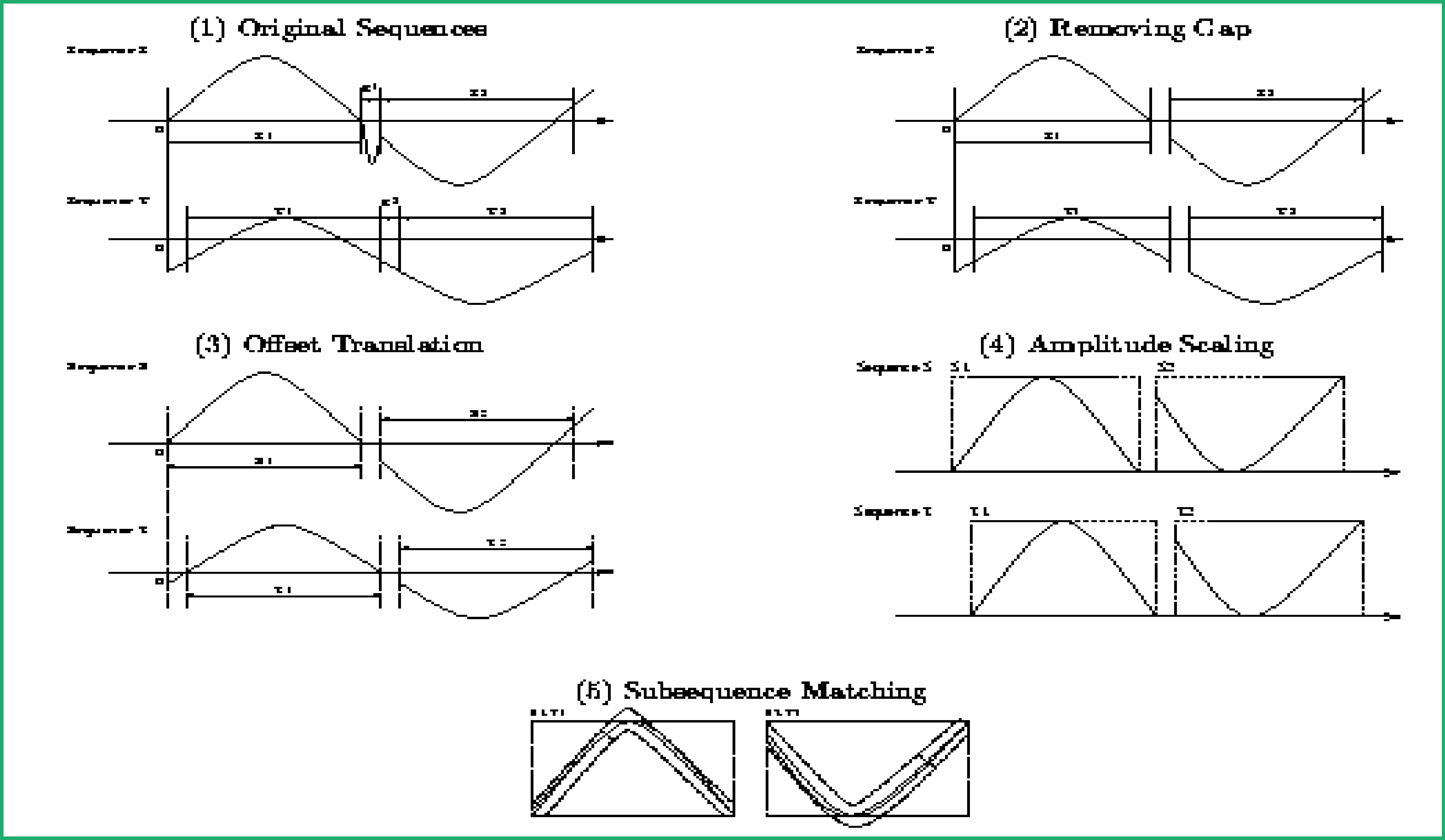


Fig.- Sub sequence matching in time-series data

Enhanced Similarity Search Methods

- Allow for gaps within a sequence or differences in offsets or amplitudes
- Normalize sequences with amplitude scaling and offset translation
- Two subsequences are considered similar if one lies within an envelope of ε width around the other, ignoring outliers
- Two sequences are said to be similar if they have enough non-overlapping time-ordered pairs of similar subsequences
- Parameters specified by a user or expert: sliding window size, width of an envelope for similarity, maximum gap, and matching fraction

Steps for Performing a Similarity Search

- Atomic matching
  - Find all pairs of gap-free windows of a small length that are similar
- Window stitching
  - Stitch similar windows to form pairs of large similar subsequences allowing gaps between atomic matches
- Subsequence Ordering
  - Linearly order the subsequence matches to determine whether enough similar pieces exist

Query Languages for Time Sequences

- Time-sequence query language
  - Should be able to specify sophisticated queries like

  Find all of the sequences that are similar to some sequence in class *A*, but not similar to any sequence in class *B*

  - Should be able to support various kinds of queries: range queries, all-pair queries, and nearest neighbor queries
- Shape definition language
  - Allows users to define and query the overall shape of time sequences
  - Uses human readable series of sequence transitions or macros
  - Ignores the specific details
    - E.g., the pattern up, Up, UP can be used to describe increasing degrees of rising slopes
    - Macros: spike, valley, etc.