

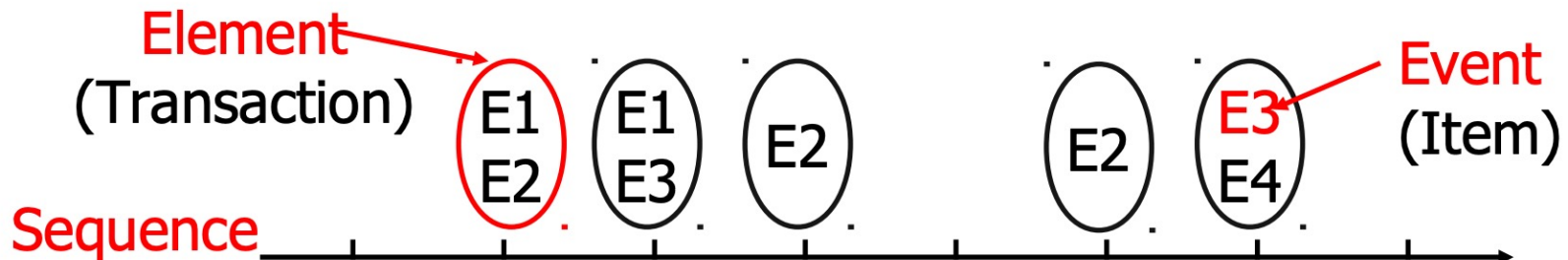
Time Series

Exercise 3



Sequential Pattern Mining

Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer Data	Purchase history of a given customer	A set of items bought by a customer at time t	Books, dairy products, CDs, etc
Web Server Logs	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Chord Progressions	Chords played in a song	Individual notes hit at a time	Notes (C, C#, D, ...)



Sequences

- A **sequence** is an ordered list of elements (transactions)

$$s = \langle e_1 e_2 e_3 \dots \rangle$$

- Each element contains a collection of items (events)

$$e_i = \{i_1, i_2, \dots, i_k\}$$

- Each element is attributed to a specific time
- **Length of a sequence** $|s|$ is given by the number of elements of the sequence.
- A **k-sequence** is a sequence that contains k events (items).

Subsequences

- A sequence $\langle a_1 a_2 \dots a_n \rangle$ is contained in another sequence $\langle b_1 b_2 \dots b_m \rangle$ ($m \geq n$) if there exist integers $i_1 < i_2 < \dots < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ..., $a_n \subseteq b_{i_n}$

Data sequence $\langle b \rangle$	Subsequence $\langle a \rangle$	Contain?
$\langle \{2,4\} \{3,5,6\} \{8\} \rangle$	$\langle \{2\} \{3,5\} \rangle$	
$\langle \{1,2\} \{3,4\} \rangle$	$\langle \{1\} \{2\} \rangle$	
$\langle \{2,4\} \{2,4\} \{2,5\} \rangle$	$\langle \{2\} \{4\} \rangle$	

- The **support** of a subsequence w is defined as the fraction of data sequences that contain w
- A **sequential pattern** is a frequent subsequence (i.e., a subsequence whose support is $\geq \text{minsup}$)

Generalized Sequential Pattern Algorithm (GSP)

- Step 1:
 - Make the first pass over the sequence database D to yield all the 1-element frequent subsequences
- Step 2: Repeat until no new frequent subsequences are found
 1. Candidate Generation:
 - Merge pairs of frequent subsequences found in the $(k-1)th$ pass to generate candidate sequences that contain k items
 2. Candidate Pruning:
 - Prune candidate k-sequences that contain infrequent $(k-1)$ -subsequences
(Apriori principle)
 3. Support Counting:
 - Make a new pass over the sequence database D to find the support for these candidate sequences
 4. Candidate Elimination:
 - Eliminate candidate k-sequences whose actual support is less than *minsup*

Applying GSP

Data sequence

< {2,4} {3,5,6} {8} >

< {1,2} {3,4} >

< {2,4} {2,4} {2,5} >

- **minsup: 50%**
- Frequent 1-sequences?
- Frequent 2-sequences?
- Frequent 3-sequences?

Task: Study Histories

Student	Semester 1	Semester 2	Semester 3	Semester 4
1235894	CS101, CS103, CS104, MA101	CS105, MA102, MA103	MA104	MA201, CS106
1237843	CS101, MA101, MA102	CS103, MA103	CS105	CS107
1238843	CS106	MA101, MA102	CS101, CS102, CS103	CS104, MA201
1240834	MA101, MA102	CS101, CS102, CS103	CS104, CS105	MA201, CS106
1243984	CS101, CS102, CS103	MA101	MA104, CS107	MA201, CS106
1245543	MA101, CS101, CS102, CS103	CS107	CS106	MA102
1247509	CS101, CS103, MA101	MA103	CS105	CS106
1256832	CS101, MA101	MA102, MA103	CS103, CS104	MA201
1256934	CS101, CS102, CS103	CS104, MA101	MA102, CS105	MA201, CS106
1257905	MA101	CS104, MA104	CS102	CS101, CS106, CS103

TODO

Find all sequences
of courses with
minsup = 0.75

Full list of courses:

CS101
CS102
CS103
CS104
CS105
CS106
CS107
MA101
MA102
MA103
MA104
MA201

Time Series Prediction

Component Models

A **time series** can consist of four components:

- Long - term trend (T_t)
- Cyclical effect (C_t)
- Seasonal effect (S_t)
- Random variation (R_t)

this is what we
want to find

we need to
eliminate those

Additive Model:

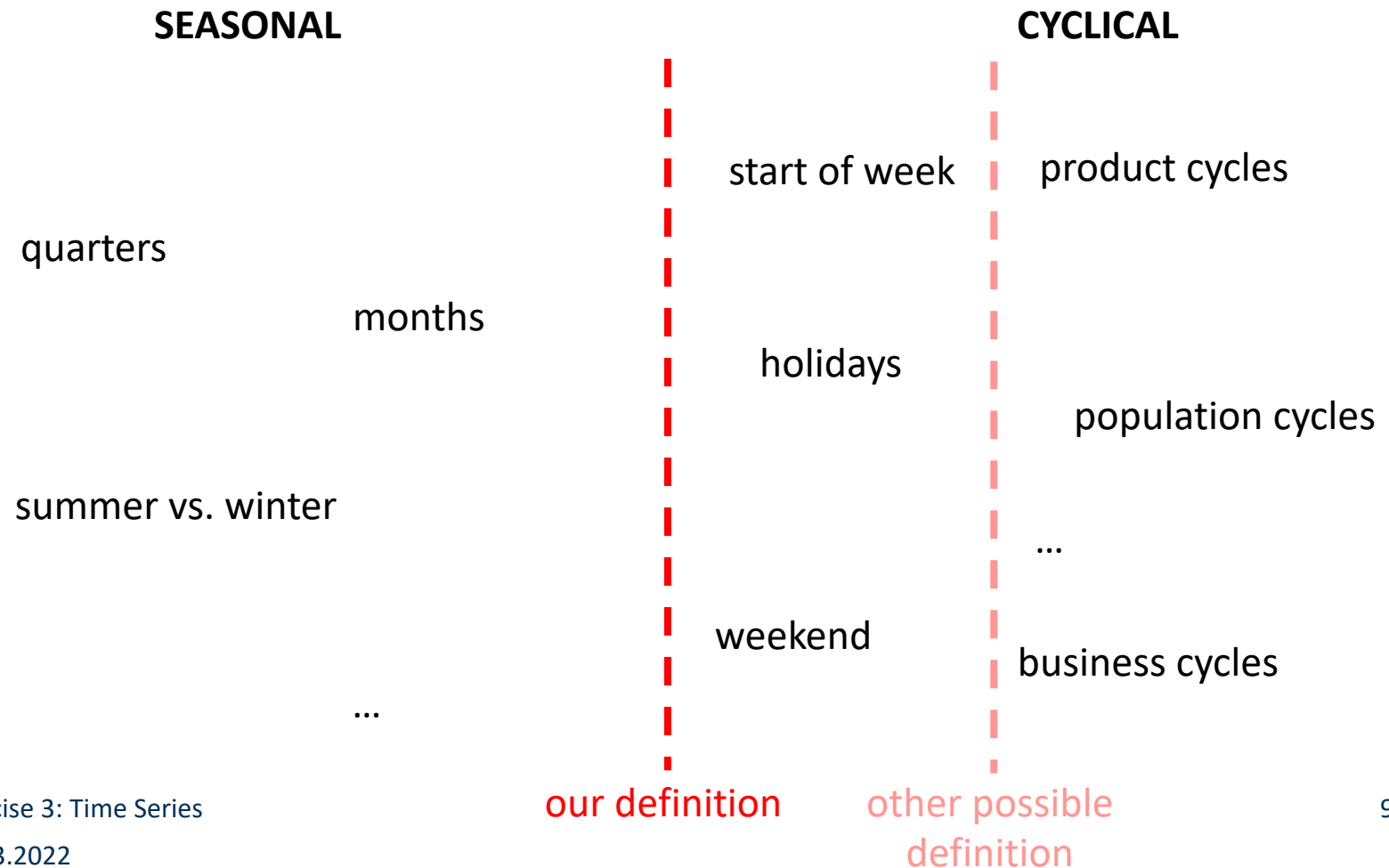
- $\text{Series} = T_t + C_t + S_t + R_t$

Multiplicative Model:

- $\text{Series} = T_t \times C_t \times S_t \times R_t$

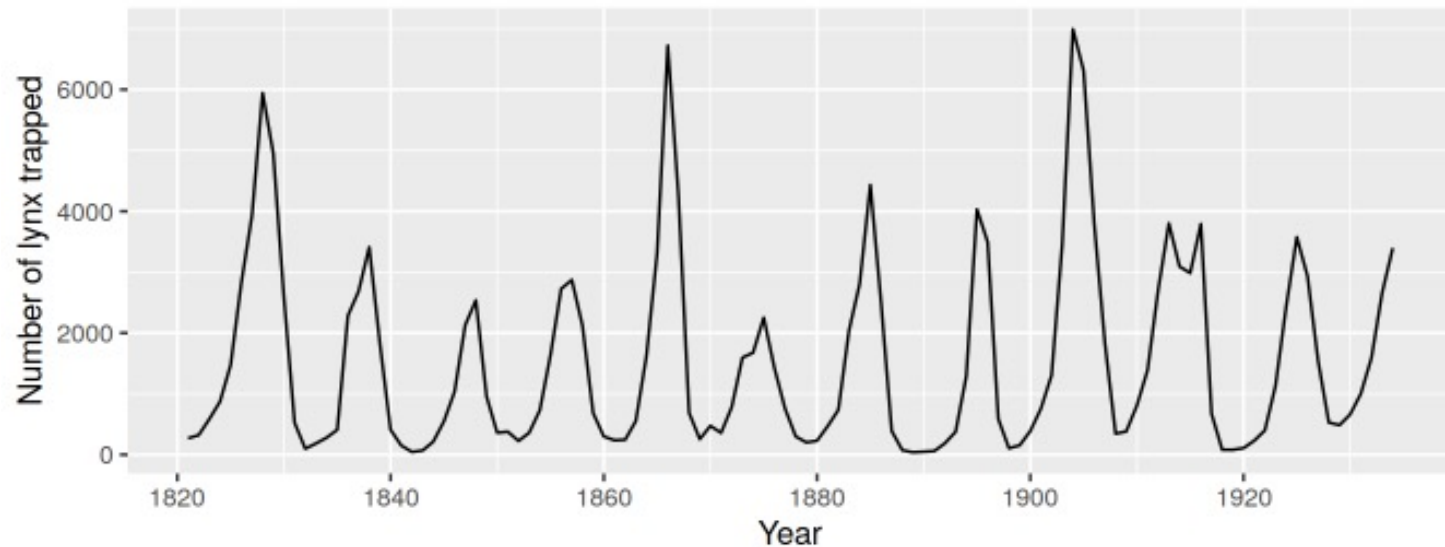
Time Series Prediction

Seasonal vs. Cyclical Effects



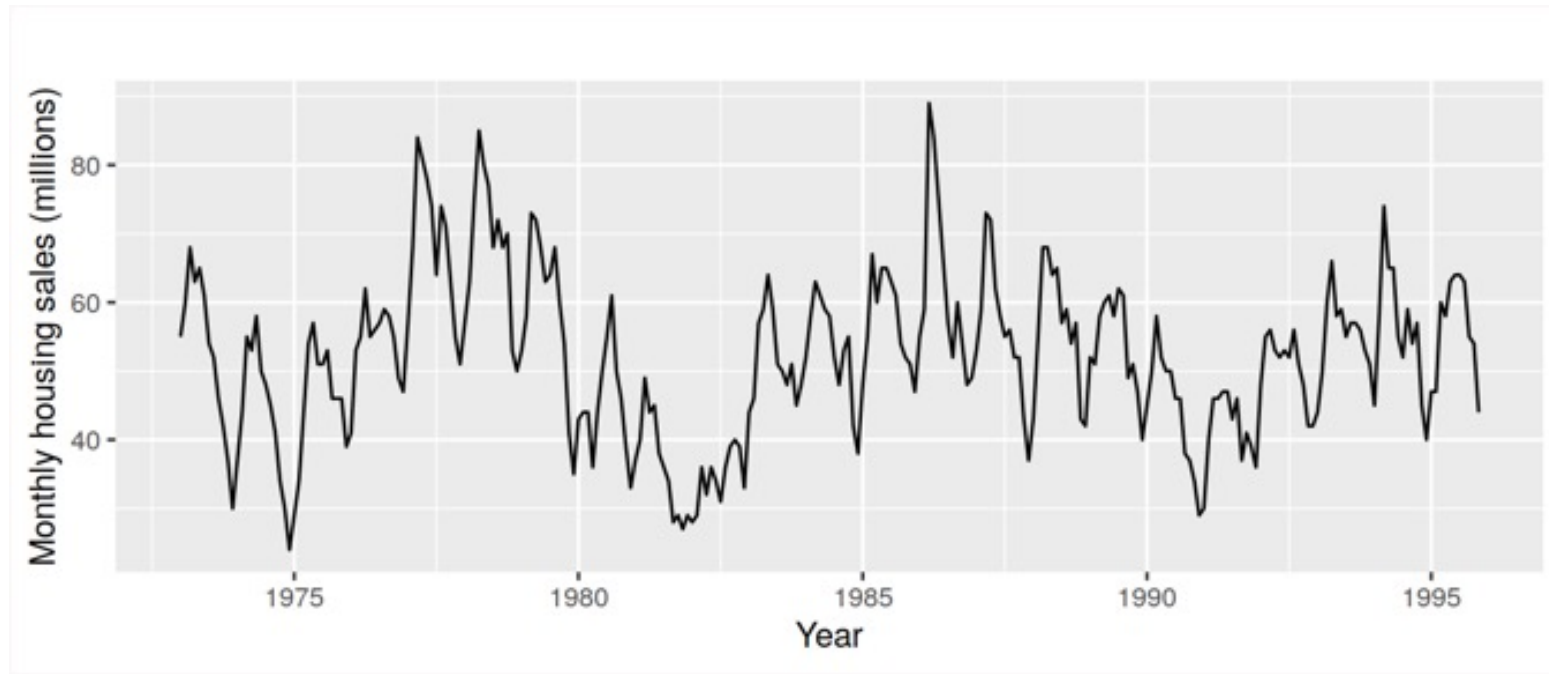
Time Series Prediction

Seasonal vs. Cyclical Effects



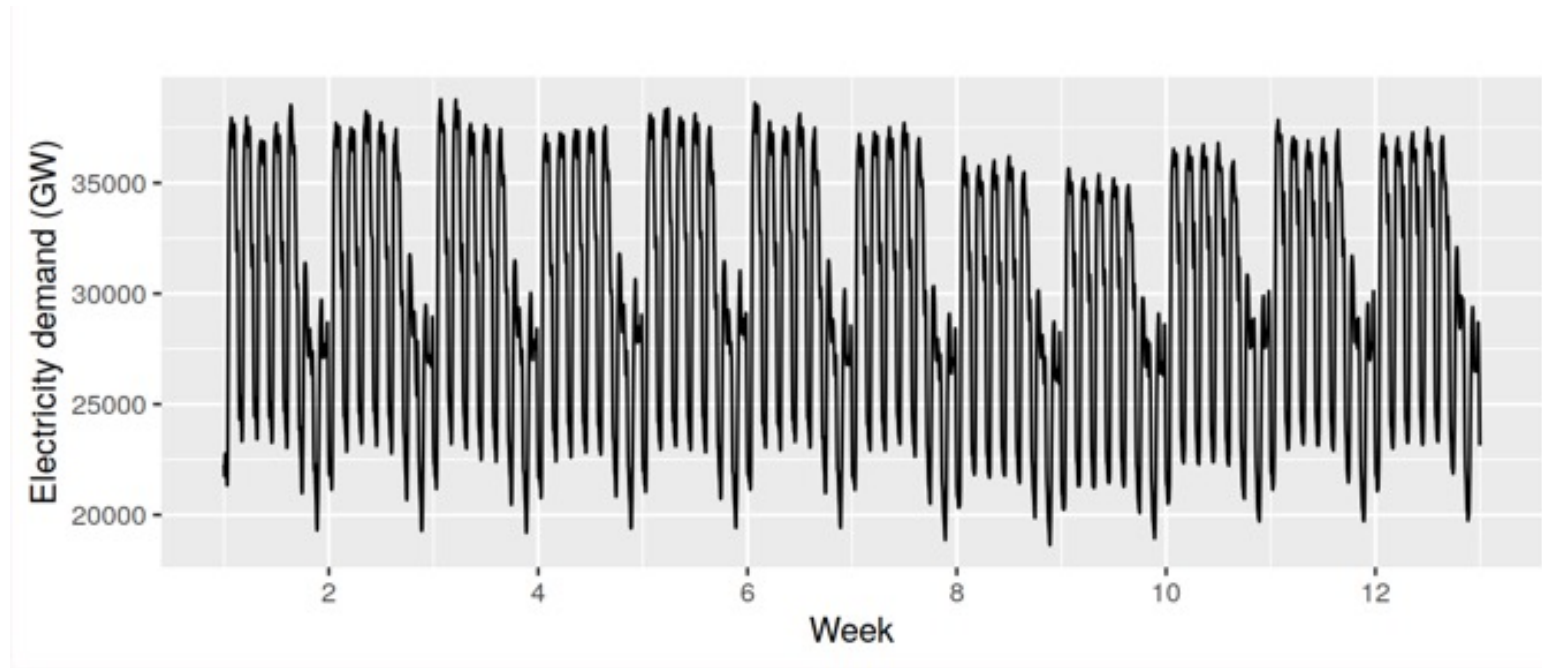
Time Series Prediction

Seasonal vs. Cyclical Effects



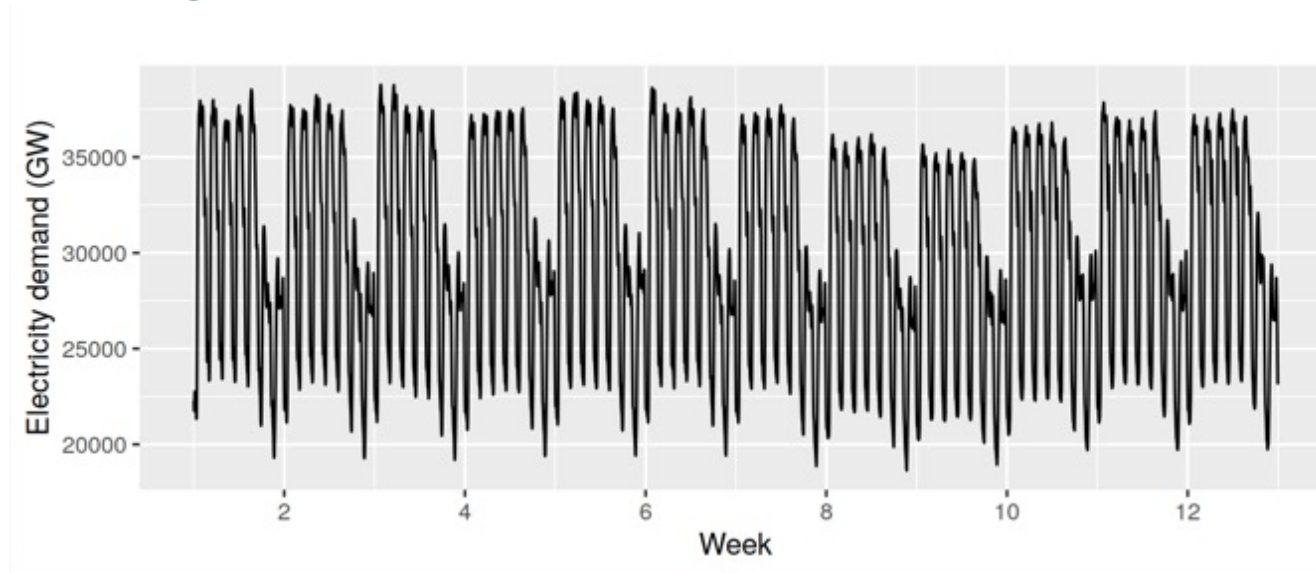
Time Series Prediction

Seasonal vs. Cyclical Effects



Time Series Prediction

Smoothing



From the given dataset, you get one datapoint per day which is the peak electricity demand. You want to create a predictor for the peak electricity demand at a given day t .

TASK:

- (1) If using simple exponential smoothing, will the peak demand for a Saturday likely be overpredicted or underpredicted? Why?
- (2) Would the Holt-Winters method solve this problem?