Stat 215a D. R. Brillinger 11/22/04

Data mining for time series

Case study Liu et al. (20??) Comp Stat + Data Analysis

Traditional t.s.a. - nonautomatic

sales and transaction data

goals - improve operations management
(labor scheduling, inventory replenishment,
product prep scheduling, strategic planning,
trends, brand loyalty, automatic forecasts, ...)

Fast-food restaurant franchises

30,000 stores worldwide

stores have PC-based servers

POS (point-of-sale) systems

orders keyed in, time stamped, stored, menu item, ingredients

Several hundred time-ordered series

corporate office higher level data

for each restaurant

data warehoused

<u>Store manager's</u> ability to anticipate:

(1) Labor requirements
(2) inventory levels
(3) food preparation scheduling

labor costs > 30% of $'s collected

high turnover rate of managers

Managers need to react on day-to-day basis (forecasts needed)

<u>Data mining</u> provides:

(1) automated processing of large amounts of data

(2) data → information → inventory planning, labor scheduling, food prep planning

(3) forecasting inventory and product depletion rates

E.g. <u>Daily time series</u>

$\Delta t = 1$ day (<u>granularity</u>)

day of week patterns

impact of: holidays, local sports events, outliers

data cleaning

Interest: forecast daily demand of particular perishable ingredient

identifying unusual data

for each shop

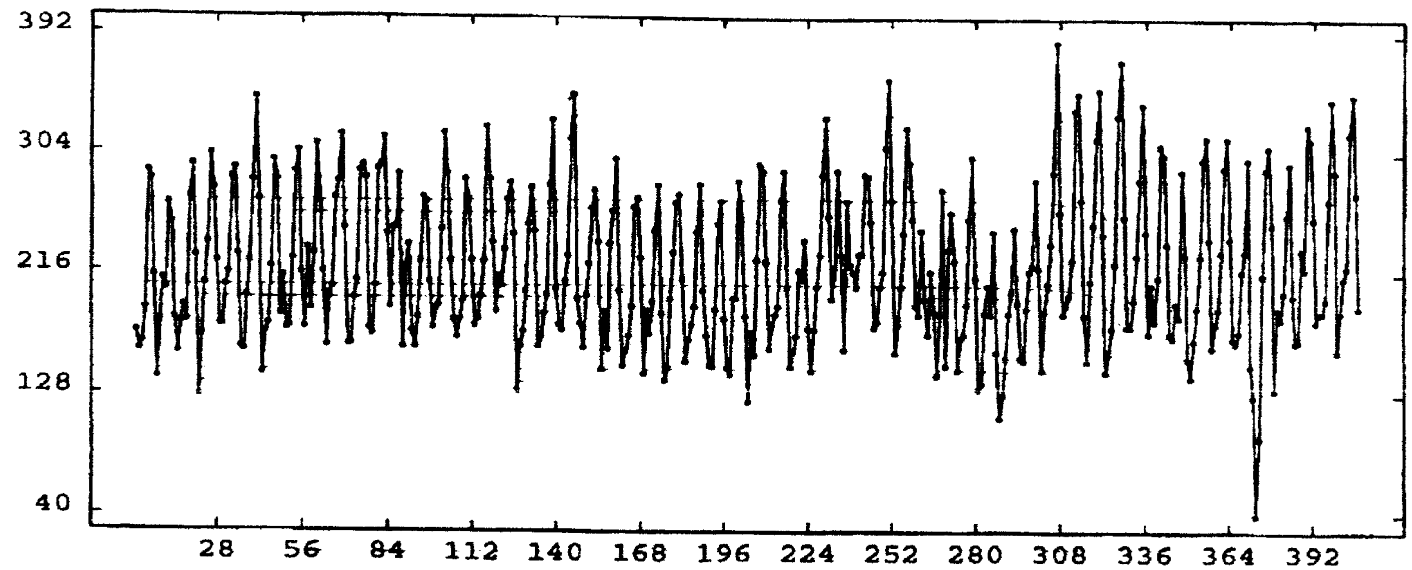

Fig. 1. Daily demand of a perishable ingredient for a fast-food restaurant (4/7/97 ~ 5/18/98).

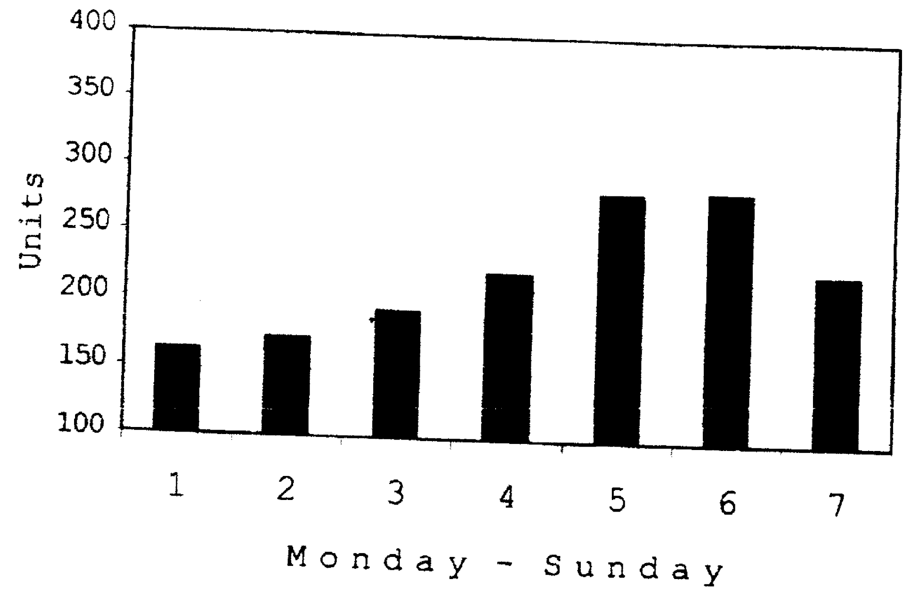

Fig. 2. Median daily demand (Monday through Sunday) of a perishable ingredient.

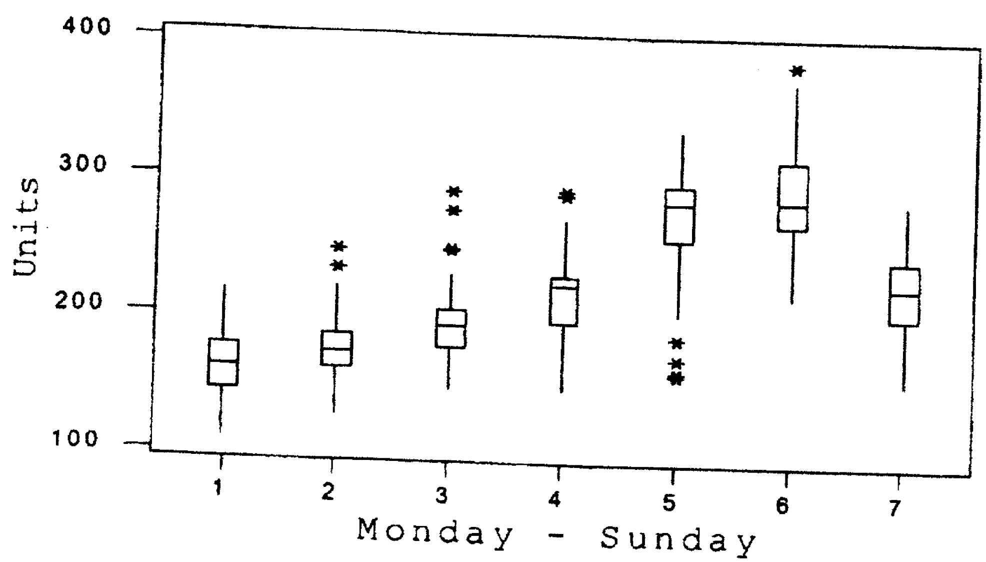

Fig. 3. Box-and-whisker plot (Monday through Sunday) for the daily demand of a perishable ingredient.

Some t.s. forecasts.

AR(1). $Y_t \sim \varphi Y_{t-1}$

$$\hat{Y}_{n+1} = \hat{\varphi} Y_n$$

residuals $\quad Y_t - \hat{\varphi} Y_{t-1}$

estimation $\quad$ OLS

ARIMA(0,1,1).

$$Y_t \sim (1-\theta)[Y_{t-1} + \theta Y_{t-2} + \theta^2 Y_{t-3} + \dots] = Y_t(\theta)$$

$$\hat{Y}_{n+1} = (1-\hat{\theta})[Y_n + \hat{\theta} Y_{n-1} + \hat{\theta}^2 Y_{n-2} + \dots]$$

residuals

$$SS(\theta) = \sum_t [Y_t - Y_t(\theta)]^2$$

SARIMA(0,1,1), period 7.

$$Y_t \sim (1-\Theta)\left[Y_{t-7} + \Theta Y_{t-14} + \Theta^2 Y_{t-21} + \dots\right] = Y_t(\Theta)$$

$$\hat{Y}_{n+1} = (1-\hat{\Theta})\left[Y_{n-6} + \hat{\Theta} Y_{n-13} + \dots\right]$$

$$SS(\Theta) = \sum_t \left[Y_t - Y_t(\Theta)\right]^2$$

Expression employed.

$$Y_t \sim (1-\Theta)\left[Y_{t-7} + \Theta Y_{t-14} + \Theta^2 Y_{t-21} + \dots\right]$$

$$+ \Phi Y_{t-1} - \Phi(1-\Theta)\left[Y_{t-8} + \Theta Y_{t-15} + \Theta^2 Y_{t-22} + \dots\right]$$

$$= Y_t(\Theta, \Phi)$$

$$SS(\Theta, \Phi) = \sum_t \left[Y_t - Y_t(\Theta, \Phi)\right]^2$$

Validation (diagnostic checking)

Sample ACF of residuals

Missing values

Take rough value

Use robust/resistant method

Data warehousing

Aggregate into hourly and daily intervals

Record special events (promotions, ...)

Forecast performance

365 days for analysis

42 for evaluation

Outliers.

Examine $Y_t - Y_t(\hat{\Theta}, \hat{\Phi})$

busload of kids

Robust/resistant version.

Weight terms so large residual terms have less effect

"outlier adjustment"

Corporate level

Warehousing formidable

Need flexibility

- new series

Rapid evaluation of promotional campaigns

- pool across stores
- intervention analysis

Examine best and worse performing stores

What makes this data mining?

Large amount of data

Automatic analysis

Business application

"Knowledge discovery"