

ネットワーク時代の ビジネスモデル構築



消費者データをもとにした 売上予測モデルと 購買構造の探索

SPSS Data Mining Day 2000
2000-5-15. 於・全日空ホテル
日経リサーチ・鈴木 督久
<http://www.nikkei-r.co.jp/>

1

平成11年度 データベース構築・技術開発促進 事業委託課題

- (財)データベース振興センター
- (株)日経リサーチ(責任者・鈴木督久)
- 課題「POSデータに対するデータマイニング手法群の比較研究と実用化技術の開発」
- 報告書「POSデータに対するデータマイニング事例集」

2

実施方法

- ユーザーである企業にとっての具体的な課題・問題を掲げて解決を試みる
- 現実にある大規模POSデータを使用する(NEEDS-SCANおよびSCAN/PANEL)
- 市販のデータマイニング・ツールを複数使用する
- 美しい成功事例だけでなく、失敗事例も

3

SPSS流 Concept: 5As

- Assess - Determine what data you need to collect
- Access - Get the data you need it
- Analyze - Use a variety of analytical techniques
- Act - Turn your information into action
- Automate - Build custom interfaces to deploy your models

4

Fayyad et al. (1996)

- Selection -
- Preprocessing -
- Transformation -
- Data Mining -
- Evaluation -

5

研究委員会

- 江原淳(専修大学)
 - 片桐嘉信(日本経済新聞社)
 - 清水順也(ハウス食品)
 - 中林三平(金融エンジニアリンググループ)
 - 茂木宏之(キッコーマン)
- 鈴木督久(日経リサーチ)事務局
- 日本経済新聞社&日経リサーチ:作業グループ

6

事例集

1. 新製品の立ち上がり売れ行き予測
2. 新製品をヒットに導く要因は何か
3. 焼肉たれの売れ行き要因分析
4. パネルデータによる購買行動の分析
5. パネルデータによるバスケット分析
6. 牛乳販売量の予測モデル

7

NEEDS—SCAN

- スーパー・コンビニ, 46チェーン・235店舗・1日80万人の来客
- 1日800万商品購入, 年間6000億円
- 日本チェーンストア協会加盟7376店舗の食・雑・医・化=10.5兆の5.7%に相当
- 流通システム開発センターのRDS/POSも導入(中堅420店舗:年商3000億円)してデータベース化.

8

NEEDS—SCAN／PANEL

- 1987年から13年間
- 関東地区2店舗, 2500世帯
- IDカードによる **ストアスキャン**
- 消費者の属性調査
- パネルに対するサーベイリサーチ

9

日経商品マスター

- 日経商品マスター:280万商品
- 新製品は1日200~250アイテム登場
- 年間で9万種類の新製品を登録
- 調査員による店頭調査で実施
- 日経独自の商品分類, 容量, 商品名

10

では, 事例の概要に~

11

1. 新製品の立ち上がり予測

- 新製品発売直後に将来の売行判断
 - スナック菓子(1997年10~1999年11月)
 - 1693商品→621→328(ローカル, 期間限定, 短期除去)
 - 地域(10分類), 日付, カテゴリ, 金額, 個数, 平均単価, カバー率(販売実績店比率).
 - 個数全国シェア基準(優位:0.2%, 中位0.05%, 劣位0.05%)
 - χ^2 で説明変数候補→C4.5と同様のアルゴリズム
- 1: 発売2Wまでに全国シェア0.1%以上拡大
2: かつ, 全国でのカバー率10%以上
→優位商品に成長可能性は90%程度(劣位と比較して)

12

2. 新製品をヒットに導く要因

- ポテトチップス←価格・カバー率・シェア
- 基準変数: 12週目シェア (1大ヒット, 2失敗でない)
- 説明変数: 地域別価格, カバー率, シェア
- **CHAID**. 発売2Wまでに北海道カバー率46.2%超→大ヒット(0.5%以上)65.8%. 下回るとヒットせず96.4%.
- 失敗しない(0.2%以上)→2W首都圏カバー率22.1%超
- **C5.0**. 2Wの全国シェア0.27%超→ヒット(操作できず)

13

CHAIDとC5.0

- 大ヒット要因(全国シェア0.5%超)
 - CHAID
 - 第1分岐: 第2週北海道カバー率
 - 第2分岐: 第2週全国シェア
 - C5.0
 - 第1分岐: 第2週全国シェア

14

CHAIDとC5.0

- 失敗しないための要因(全国シェア0%超)
 - CHAID
 - 第1分岐: 第2週首都圏カバー率
 - 第2分岐: 第2週全国カバー率
 - C5.0
 - 第1分岐: 第2週全国カバー率
 - 第2分岐: 第2週全国カバー率

15

3. 焼肉たれ売れ行き要因分析

- 首都圏3店舗. 1998.4~1999.3. 64商品
- 3店共通250日以上販売実績は5商品
- 目的変数: 定番3品の販売個数
 - 「エバラ黄金の味」「エバラおろしのたれ焼肉のたれ」「キッコーマン新撰焼肉赤と黒」
- 説明変数: 他商品値引率, 来店客数, 休日
- Clementine NN, CHAID

16

3. 焼肉たれ売れ行き要因分析

- A店では特異日除いてよく予測. 重要度は自身価格→休日→客数
- 店舗によって予測の良さが異なる
- NNとCHAIDで要因の結論が違う場合あり
- NNで0-1に事前変換した場合としない場合では結果が異なる(特に重要度指標)

17

4. パネルによる購買行動分析

- カレールー(1998.6~1999.5)30商品
- 年間10個以上購入パネル146世帯
- パネル別の購入履歴(購入の有無がY)
- 説明変数: メーカー, 辛さ, クラス, 容量, 値引率
- パネルごとに購入した日だけデータ生成(分散分析や2類のような形式のデータ)
- パネルごとに異なる購買要因を発見

18

4. パネルによる購買行動分析

- パネルごとにC5.0で分析→第1基準は何
- パネル属性と第1基準とのクロス分析
- 第1基準でパネルを4グループに分割してグループごとにC5.0で分析して構造を見る
- C5.0, CART, CHAIDでツリー比較
- 分散分析データと形式は同じだが, 購買データは実験計画的でないので整形をする

19

説明変数

メーカー	辛さ	クラス	容量	値引率
メーカーA	甘口	定番・一般	200g以上	注)
メーカーB	中辛	高級・本格	200g未満	
メーカーC	辛口	中級・後発		
PB				

注) 値引率 = (販売価格 - 通常価格) / 通常価格 *

20

分析データの形式

日付	購入	メーカー	辛さ	クラス	容量	値引き率
1998/7/18	買わない	メーカーA	中辛	定番・一般	200g未満	-16.5
	買わない	メーカーA	中辛	高級・本格	200g以上	0
	買わない	メーカーA	中辛	定番・一般	200g以上	0
	買った	メーカーC	中辛	定番・一般	200g未満	-41.1
	買わない	メーカーC	中辛	中級・後発	200g以上	0
	買った	メーカーC	中辛	中級・後発	200g未満	-24.7
1998/9/9	買わない	PB	中辛	定番・一般	200g未満	0
	買った	メーカーA	中辛	定番・一般	200g未満	-16.5
	買わない	メーカーA	中辛	高級・本格	200g以上	0
	買わない	メーカーA	中辛	定番・一般	200g以上	0
	買わない	メーカーC	中辛	定番・一般	200g未満	-33.6
	買わない	メーカーC	中辛	中級・後発	200g以上	0
1998/11/25	買わない	メーカーC	中辛	中級・後発	200g未満	-10.1
	買わない	PB	中辛	定番・一般	200g未満	0
	買った	メーカーA	中辛	定番・一般	200g未満	-29.1
	買わない	メーカーA	中辛	高級・本格	200g以上	-26.1

21

C5.0の出力結果

```

値引き率    =< -5.1
メーカー   その他 (0.0, 1.0) -> 買わない
メーカー   PB    (1.0, 1.0) -> 買わない
メーカー   メーカーC
容量       200g以上 (1.0, 1.0)   -> 買う
容量       200g未満 (11.0, 0.818) -> 買わない
メーカー   メーカーA
値引き率    =< -29.1
値引き率    =< -31.6
値引き率    =< -32.9 (1.0, 1.0) -> 買う
値引き率    > -32.9 (1.0, 1.0) -> 買わない
値引き率    > -31.6 (3.0, 1.0) -> 買う
値引き率    > -29.1 (5.0, 0.6) -> 買わない
    
```


22

5. パネルによるバスケット分析

- 1998.11~1999.10, データ数135万9254件から「14万レシート×12万商品」の形式に.
- 小分類単位にまとめ, 年間12件以下は除外して
 - 年間: 139, 010レコード(940分類)
 - 夏季: 34, 866レコード(736分類)

23

同時購買「信頼度」

 確信度, 信頼性....		商品 A	
		非購入	購入
商品 B	非購入	a	c
	購入	b	d

24

信頼度

		商品 A	
		非購入	購入
商品 B	非購入	a	c
	購入	b	d

Aを購買 → Bも買う : $\frac{d}{c+d}$

Bを購買 → Aも買う : $\frac{d}{b+d}$

信頼度(例)

		福神漬	
		非購入	購入
カレー	非購入	500	150
	購入	300	100

カレー → 福神漬 : $\frac{100}{300+100} = .250$

福神漬 → カレー : $\frac{100}{150+100} = .400$

Aprioriの出力結果 (夏季3ヶ月データ)

生中華そば<=中華そば用つゆ ②	(105:0.3%, 0.673)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮農産物 & 鶏卵	(105:0.3%, 0.626)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮農産物 & ハードヨーグルト	(70:0.2%, 0.6)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & はんぺん	(105:0.3%, 0.596)
菓子パン <= 蒸しパン & さつまいも揚げ・揚げかまぼこ	(70:0.2%, 0.595)
絹ごし豆腐・ソフト豆腐 <= 油揚げ & キムチ・朝鮮漬	(70:0.2%, 0.588)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮水産物 & 油揚げ	(139:0.4%, 0.587)
紙パック入り果汁入り清涼飲料 <= 紙パック入り無果汁清涼飲料 ③	(139:0.4%, 0.585)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 普通食パン & 鶏卵	(105:0.3%, 0.574)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生ちくわ・焼きちくわ	(209:0.6%, 0.568)
菓子パン <= デニッシュ・ペストリー & フルーツ入りヨーグルト	(174:0.5%, 0.554)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮農産物 & 即席カップ中華そば	(70:0.2%, 0.553)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 一般牛乳 & 油揚げ	(105:0.3%, 0.551)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮水産物 & ウィンナソーセージ	(105:0.3%, 0.549)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 冷凍お好み焼き・たこ焼き	(70:0.2%, 0.543)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮水産物 & 即席カップ中華そば	(70:0.2%, 0.542)
絹ごし豆腐・ソフト豆腐 <= 油揚げ & 鶏卵	(139:0.4%, 0.542)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮水産物 & その他漬物	(70:0.2%, 0.541)
粒納豆 <=油揚げ & キムチ・朝鮮漬	(70:0.2%, 0.541)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 釜めし・炊き込みごはんの素	(70:0.2%, 0.541)
菓子パン <= デニッシュ・ペストリー & 紙パック入り果汁混合野菜ジュース	(70:0.2%, 0.539)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮農産物 & 油揚げ	(139:0.4%, 0.538)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 生鮮水産物 & 鶏卵	(105:0.3%, 0.537)
絹ごし豆腐・ソフト豆腐 <= ロールパン・コッパン & 油揚げ	(174:0.5%, 0.536)
菓子パン <= 普通食パン & 生鮮農産物 & デニッシュ・ペストリー	(70:0.2%, 0.535)
絹ごし豆腐・ソフト豆腐 <= 粒納豆 & 普通食パン & 油揚げ	(139:0.4%, 0.532)
絹ごし豆腐・ソフト豆腐 <= 菓子パン & 粒納豆 & ハードヨーグルト	(70:0.2%, 0.532)

商品A <= 併買B (100, 17.0%, 0.25) →

B出現数

全レシート中のB出現率

信頼度

信頼度は正しい, 出現率は1位まるめ, 頻度はその乗算

全体的な併買パターン

- 「中華そば用つゆ」を買うと「生中華そば」を買う[信頼度67.3%]
- 「納豆」「精選農産物」「鶏卵」を買ったら「絹ごし豆腐・ソフト豆腐」を買う[62.6%]
- 「紙パック入り果汁入り清涼飲料」を買うと「紙パック入り無果汁清涼飲料」[58.5%]
- 全体として出現頻度高い日配品が上位に

29

日配品7分類を除去・再分析

- (絹ごし豆腐, 納豆, 一般牛乳, 生鮮農産物, 生鮮水産物, 普通食パン, 菓子パン)
- 生中華そば～中華そば用つゆ～めん用の具, メンマ
- カレールー～福神漬
- 缶入りコーラ飲料～缶入り無果汁炭酸飲料

30

コーザルデータを追加・分析

- 8月の火曜日に「粒納豆」と「絹ごし豆腐」→火曜日が特売日
- 7月の火曜日に「即席カップ焼きそば」を買った人の62.5%が「即席カップ中華そば」
- 「ポテトチップ」と「コーン系スナック」「小麦スナック」は休日に購入される
- 「カレールー」と「福神漬」は曜日に関係なく

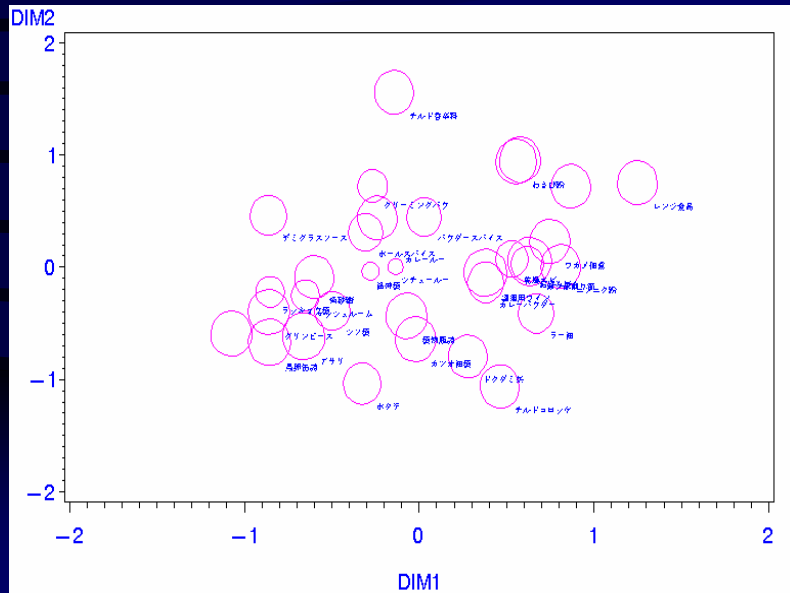
31

バーチャル・アイテムを活用

- 「ホームタイプマルチアイス」と「福神漬」を買ったら「カレールー」. 信頼度41%
- ブランド, 容量, 辛さ, クラスの4アイテム
- 「バーモントカレー」と「ビスケットクッキー」
- 「甘口」「やや甘」では「冷凍ポリドリンク」
- 「辛口」「やや辛」には「カレー粉」

32

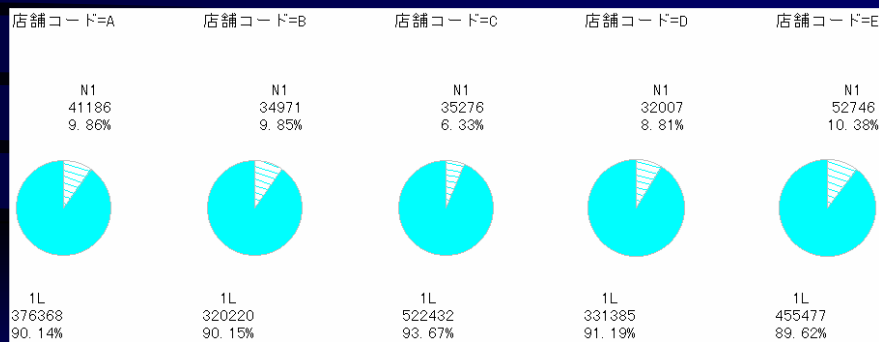
非対称MDS(岡太・今泉)



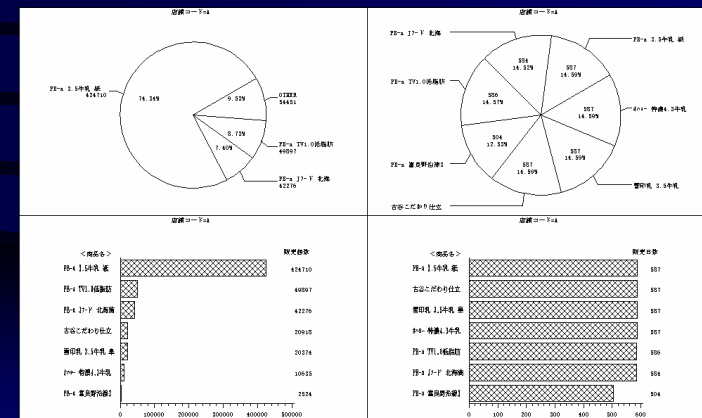
明日の牛乳の販売量予測

- 首都圏スーパーマーケット5店(ABCDE)
- 1998年4月~1999年11月(20か月)
- POSデータ, 気象データ
- データ吟味のうえで1リットル牛乳のみ分析
- 20ヶ月すべて販売実績のある商品のみ

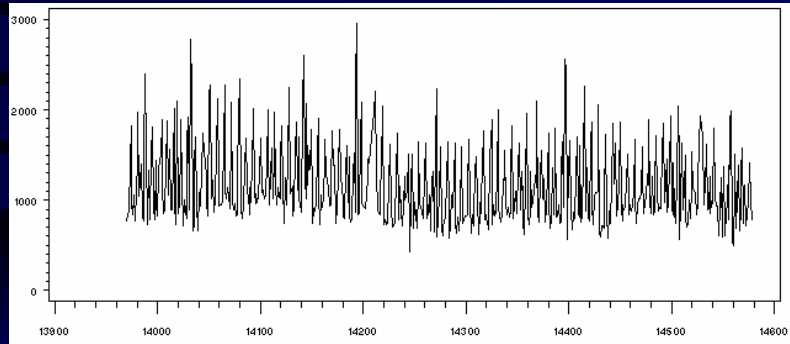
5店とも1リットル牛乳で9割以上



20ヶ月販売商品のシェアと日数

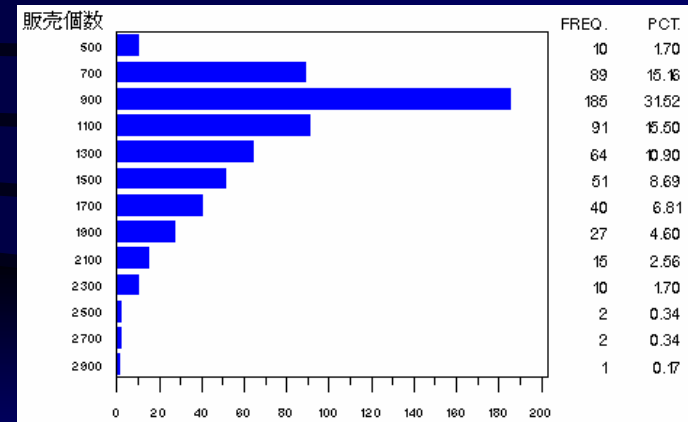


販売個数の日次変動(A店)



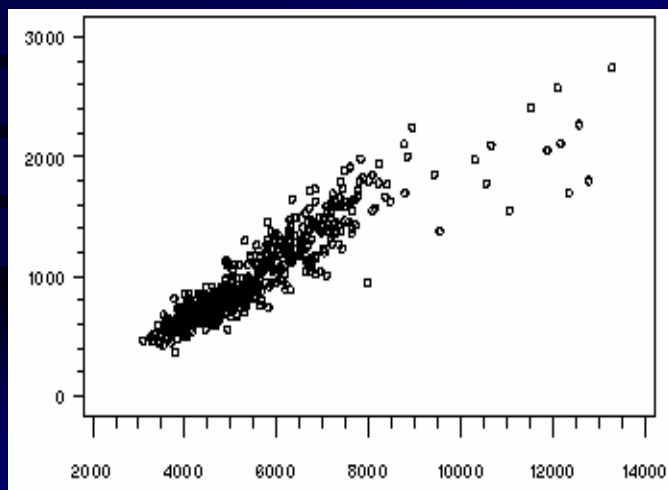
41

販売個数の日次変動(A店)



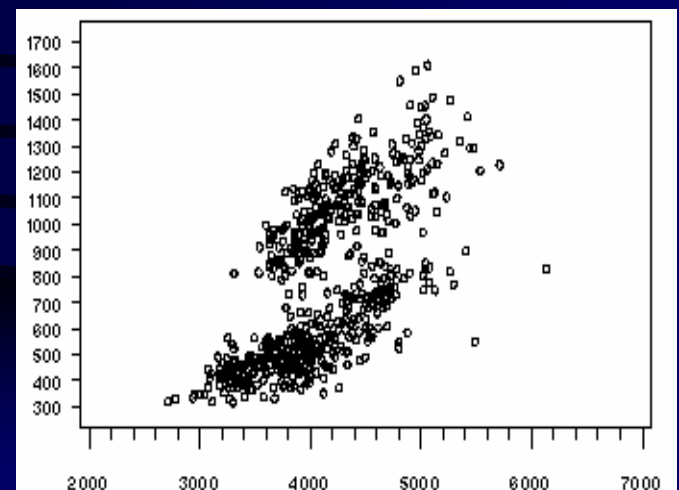
42

来店客数と販売個数(A店)



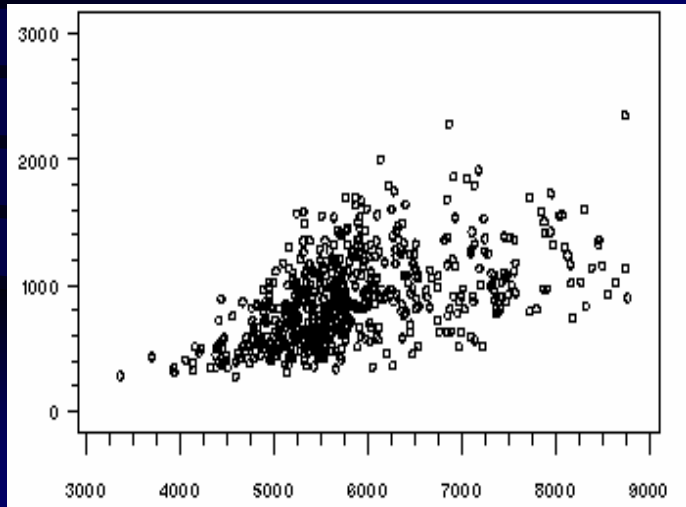
43

来店客数と販売個数(B店)



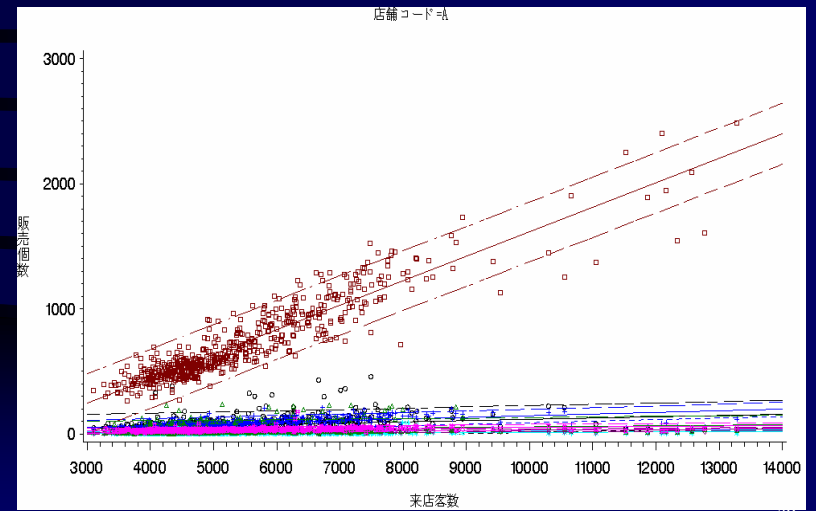
44

来店客数と販売個数(C店)



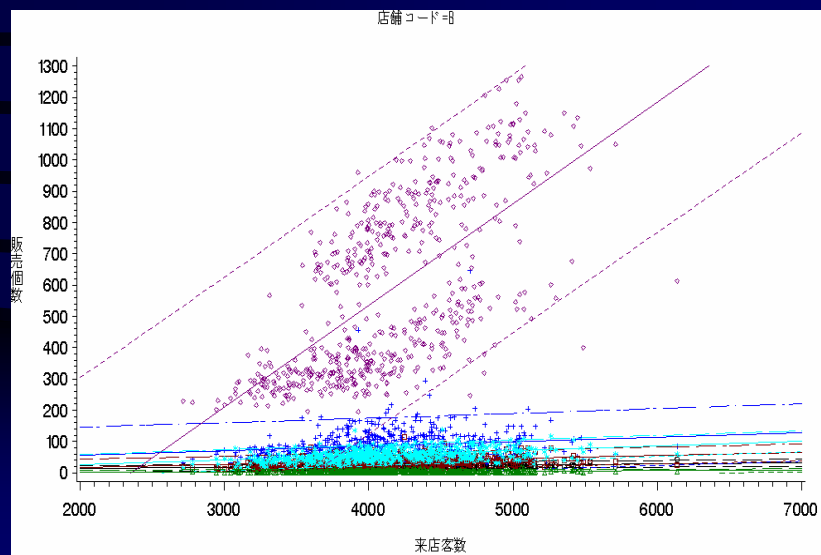
45

来店客数と販売個数(層別)A

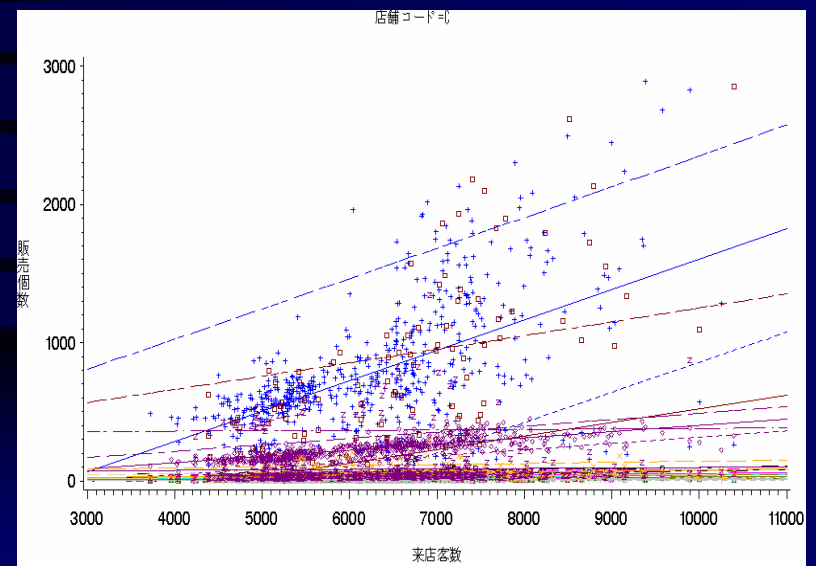


46

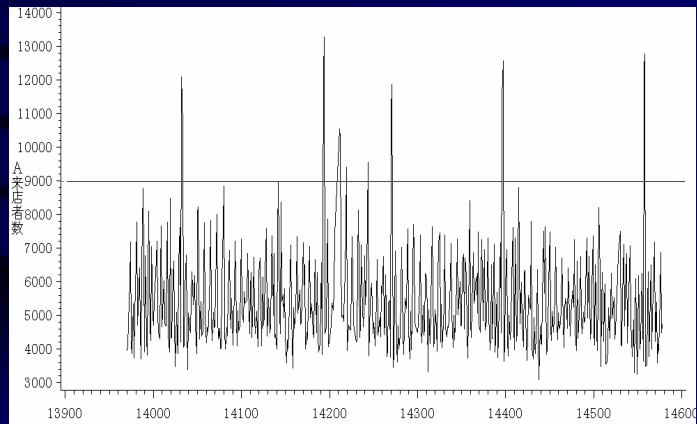
来店客数と販売個数(層別)B



来店客数と販売個数(層別)C



来店客数の飛び外れ値(A店)



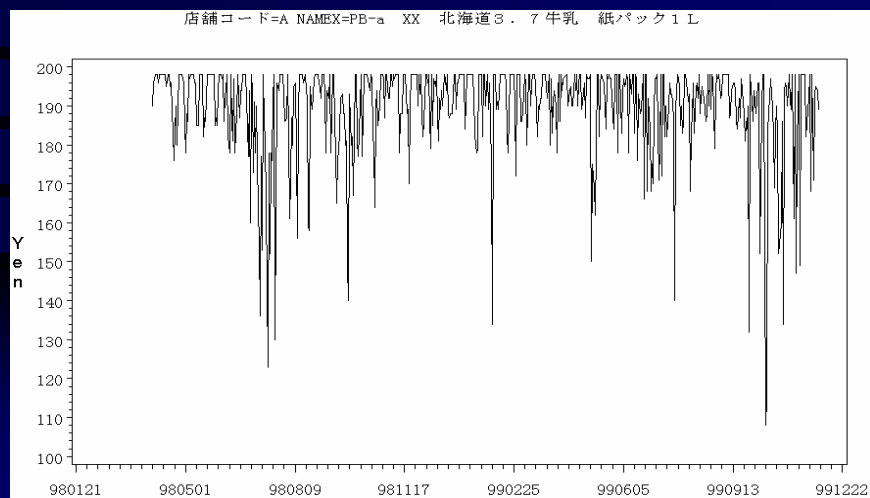
49

外れ値(特異日)は特殊な日

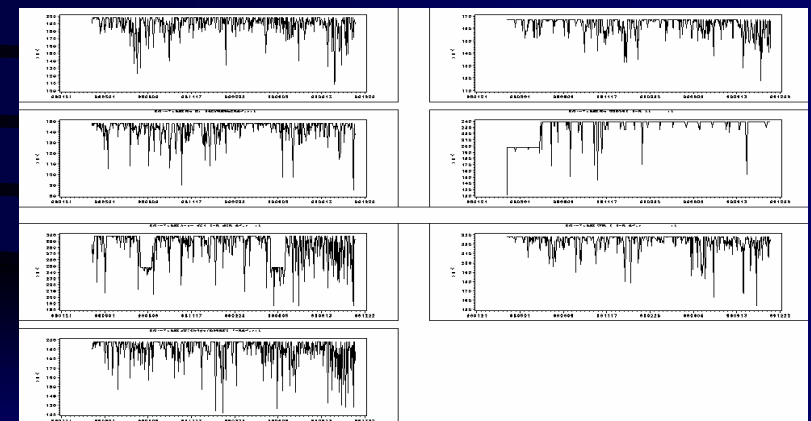
OBS	DATE	WEEK	WEEKDAY	N_STORE	TENKI	OUTLIER
1	1998-06-02	火	3	12095	くもり	*
2	1998-06-03	水	4	10661	雨	*
3	1998-11-10	火	3	12165	くもり	*
4	1998-11-11	水	4	13281	晴れ	*
5	1998-11-28	土	7	10554	晴れ	*
6	1998-11-29	日	1	10302	晴れ	*
7	1998-12-06	日	1	9420	晴れ	*
8	1998-12-31	木	5	9539	晴れ	*
9	1999-01-26	火	3	11051	晴れ	*
10	1999-01-27	水	4	11873	くもり	*
11	1999-06-01	tue	3	11524	晴れ	*
12	1999-06-02	wed	4	12574	晴れ	*
13	1999-11-09	tue	3	12338	晴れ	*
14	1999-11-10	wed	4	12774	晴れ	*

50

価格の日次変動(A店A商品)



価格の日次変動(A店. 各商品)



52

候補の予測変数

- 気象関係の変数
- 特異日のダミー変数
- 休店前後を示すダミー変数
- 商品別の価格
- 月、曜日を示す変数

53

変数選択の結果

- 天気, 不快指数
- 最高気温, 最高気温との差
- 月, 曜日
- 特異日
- 商品4の価格
- 商品5の価格
- 商品7の価格

54

NNとREGのモデル比較

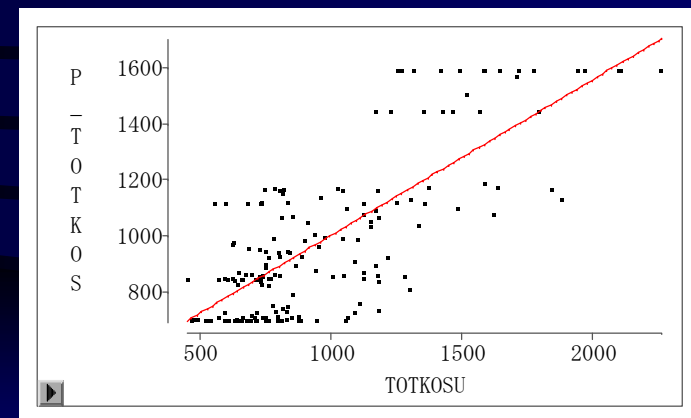
	R ²	絶対差	s
NN	0.581	190本	151
REG	0.687	163本	132

修正

	AIC	SBC	RMSE	RMSE(V)	RMSE(T)
NN	2651.9	2842	255.1	286.2	242.3
REG	2532	2591	211.1	258.2	209.9

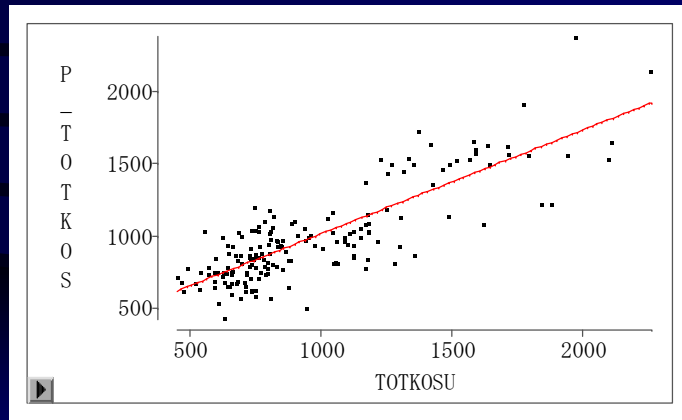
55

NNの予測値プロット



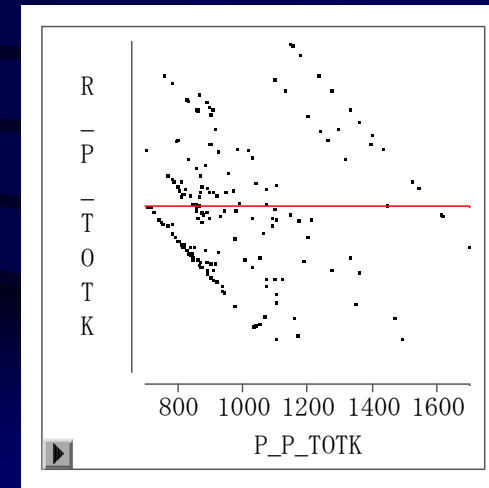
56

REGの予測値プロット



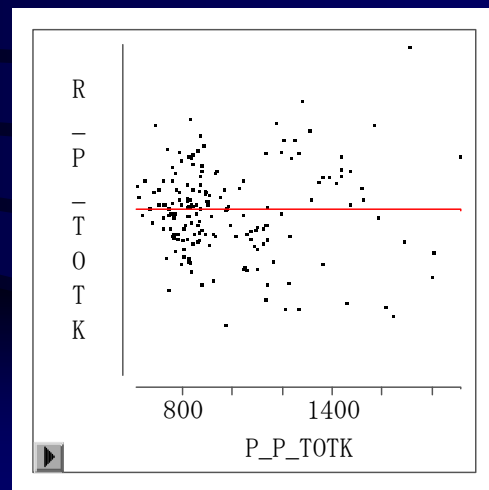
57

NNの残差プロット



58

REGの残差プロット



59

お知らせ

- 事例集は日経リサーチよりご提供できます
- 日経リサーチがDTMサービス業務実施
- 日経のPOSデータのマイニング研究など
 - 大滝厚他「ハイブリッドモデリングの手順提案とそのPOSデータへの適用」品質 (投稿中)
 - 「データマイニングとマーケティング」日科技連・多変量解析研究会・講演予定 (7月19日)
- 守口剛 (立教大学), 豊田秀樹 (早稲田大学)

60

牛乳(POSデータ)の予測

- 高橋幸一(森永乳業)
 - マクロ的な予測
- 近藤文代(総研大)
 - 値下げ効果, 曜日変動
 - 状態空間モデル(時系列解析)

61

デシジョン・ツリー分析

- **AID** 2分割 F(分散比)
- **CHAID** 多分割 χ^2 (乖離度)
- **CART** 2分割 不純度(改善度, Gini
多様性指標, エントロピー), S_w
- **C5.0** 多分割 利得比 gain ratio
- 介護

62