

Automated Trade Area Analysis. Case Study of G5 MWM Software Application



Alex.Zolotovitski@Generation5.net

Oct 29, 2004

Project Team

Statisticians and programmers

- **Dr. Alex Zolotovitski**
- **Dr. Ben Schwartz**
- **Don Rihm**
- **Dr. George Mersov**
- **Dr. Harold Reynolds**
- **Helen Li**
- **Ilya Millman**
- **Meiling Liu**
- **Dr. Victor Zurkowski**

+ Managers + Support

Project Objectives

- To create an automated tool for estimating
 $\$(\text{hh}, \text{Prd}, \text{Store})$
 - the dollar amount $\$$ that each US household **hh** (at the Zip+4 level, ~ 35 million zip+4's) spends for a specified group of Products **Prd** (60 groups) in each US grocery **Store** (~ 50,000)
- To run the tool to get the estimation
- To aggregate the results to estimate
 $\$(\text{Prd}, \text{Store})$
 - sales $\$$ for each specified group of products **Prd** in each US grocery **Store**.

Data sources

1. Panel data **\$(hhID, Prd, Bann)**
 - consumption of
37,000 Households **hh** x 290 brands of products x 717 Banners
2. Store data **(StoreID, Lon Lat, StoreName, Features)**

~50,000 stores with 20,243 unique store names x 15 Features and
3. G5 zip+4 level data **(zip9, Lon Lat, vars)**

~35M zip9 → ~12M gzip9 x ~ 10,000 demographic, expenditure, and behavioral variables

+

G5 estimated consumption **\$(zip9, Prd) .**

Challenges

- § A jumbo Problem

$$\begin{aligned} & \sim 35,000,000 \text{ zip+4} \times \sim 60 \text{ categories} \times \sim 100 \text{ stores/zip+4} \\ & = 210,000,000,000 \text{ numbers} \end{aligned}$$

- § No Consumption Data at the Store level
 - § Cannot simply apply MWM to historical data
 - § Need to match Stores to Banners

Clue Points - Methods

- StoreName – Banner Matching
- Non-parametric local + global regression k-NN (in G5 MWM) for Banner level prediction **\$(zip9, Prd, Bann)**

Parameters: Granulation (Grid) =200 m ; $d_{max} = 16 \text{ km}$

- Generalized Huff Model with local parameters for Store level prediction **\$(zip9, Prd, St)**

StoreName - Banner Matching. Grouping of Banners .

Store name		Banner	\$	RelPrice
Brookshire Bros Market		BROOKSHIRE BROTHERS	1,248,673	1.01
		BROOKSHIRE	1,704,430	1.04
Food Giant Supermarket		GIANT FOOD MARKETS	352,663	1.05
Giant Food Market		FOOD GIANT	359,839	1.06
		GIANT FOOD	1,342,456	1.15
Wyndalls Foodland		WYNDALLS FINER FOODS	107,333	1.02
		FOODLAND	1,399,029	1.05

Alternatives to make unique matching:

- 1) delete one or both links or
- 2) keep both links and create Group of Banners.

StoreName - Banner Matching.

Grouping of Banners .

Alternatives to make unique matching:

- 1) delete one or both links or
- 2) keep both links and create Group of Banners.

Criteria

1. Names similarity.
2. Responders \$ score. Relative price.
3. Max IHHI (Inverse Herfindahl-Hirschman Index)

$$\text{IHHI} = (\sum \text{SEL})^2 / \sum \text{SEL}^2 \equiv$$

number of big Groups of Banners

Distribution to Banner. Choosing model

- How to distribute \$\$: Zip9 -> Store or Zip9 -> Banner -> Store ??
- Some VIVAs, characterizing Banner-Store environment of Zip9
(we chose 52 variables from ~800 variables):

$$mSEL_gn = \text{lod}(bSELg / vSELg),$$

where $\text{lod}(x) = \ln(x/(1-x))$ (Johnson's transformation)

$$bSELg = \sum_{St \in \text{banner}} \text{SellSpace} / (1+d)^{1.273}$$

$$vSELg = \sum_{St \in \text{vicinity}} \text{SellSpace} / (1+d)^{1.273} ;$$

oCos = cos(angle between the Store(Banner) and Center of Stores) ;

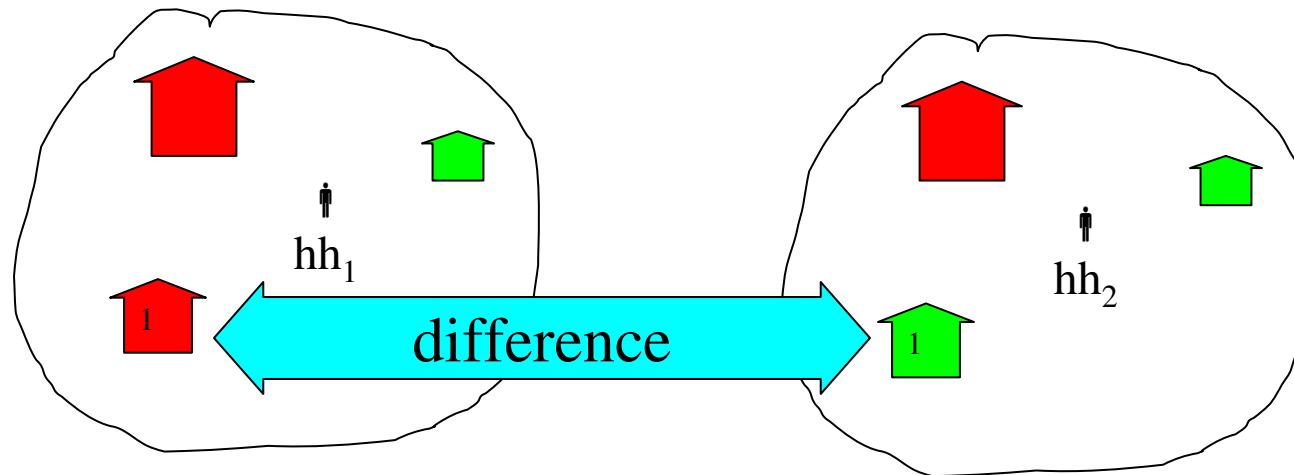
oALL_0 - AllFeatures of the nearest stores;

vHD4 = $\log_2(vSEL_8 / vSEL_4)$ – Hausdorff dimension of the Store Space.

- Criterion to choose model –

min GCV ~ min Mallow's C_p

Distribution of \$\$ from Group of Banners to Stores



$$\$ (hh_1, St_1) = \$ (hh_2, St_1) = \$ (hh_1, \text{Bann}_R) - \$ (hh_2, \text{Bann}_R) =$$

$$= -\$ (hh_1, \text{Bann}_G) + \$ (hh_2, \text{Bann}_G) \approx$$

$$\approx \sum \partial \$ (hh, \text{Bann}) / \partial \text{Feature} \times \text{Feature}(St_1)$$

Local Regr Coeff

Distribution \$\$ Banners \rightarrow Stores .

Methods

- Generalized Huff Model with local parameters for Store level prediction $\$(\text{zip9}, \text{Prd}, \text{St})$:

$$\$(\text{zip9}, \text{Prd}, \text{St}) = \$(\text{zip9}, \text{Prd}, \text{Bann}) * P(\text{zip9}, \text{Prd}, \text{St}) * I(\text{St} \in \text{Bann})$$

$$P(\text{zip9}, \text{Prd}, \text{St}) = \text{Norm} * U(\text{zip9}, \text{Prd}, \text{St}) \equiv U / \sum U$$

$$U(\text{zip9}, \text{Prd}, \text{St}) = \sum_{Ft} v_{Ft} * r_{Ft}(\text{zip9}, \text{Prd}, \text{Bann}) \quad (1)$$

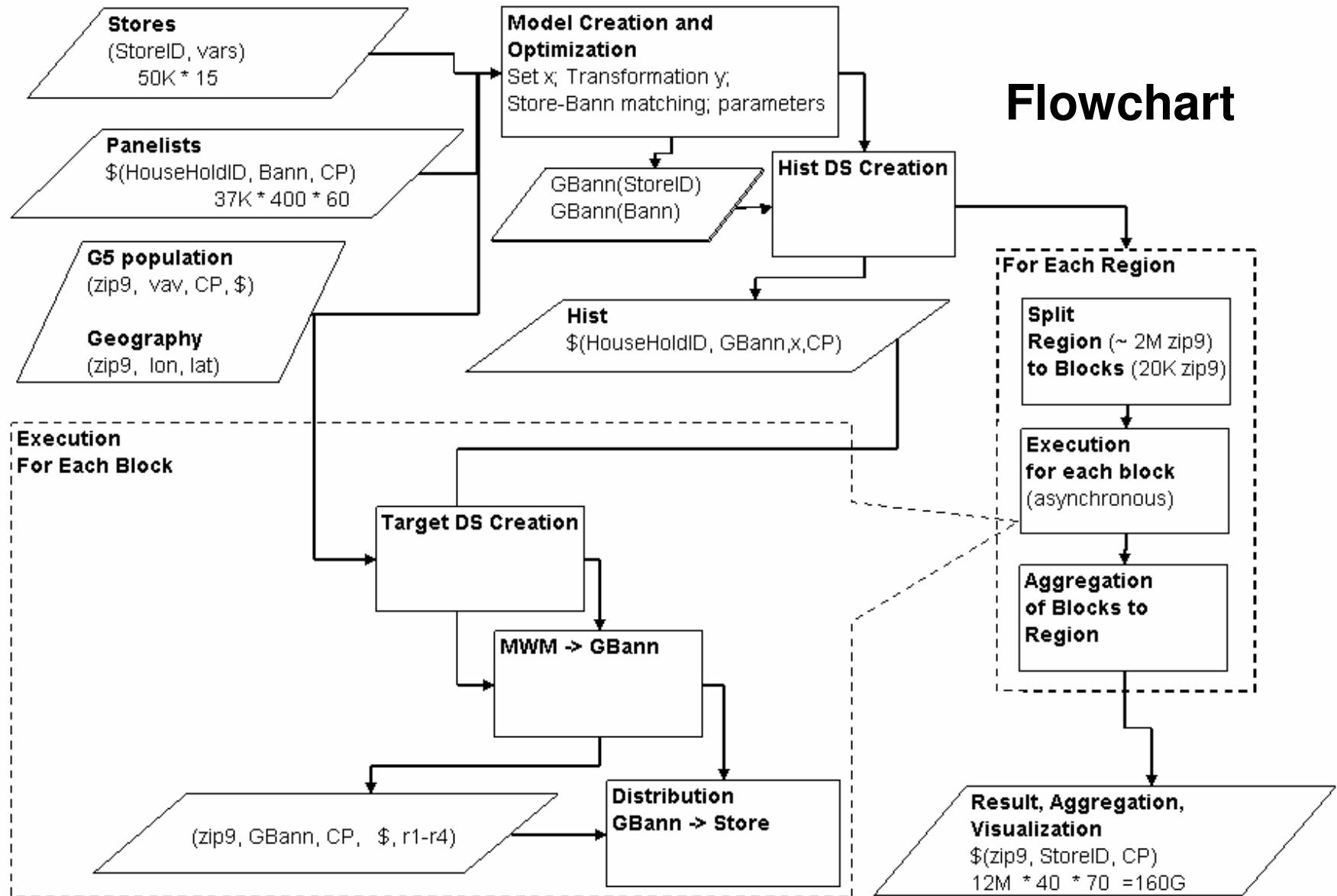
generalization of the Huff Model:

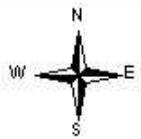
$$U(\text{zip9}, \text{Prd}, \text{St}) = S^\alpha / d^\beta \quad - \quad 1\text{-variable, global parameters}$$

Criterion to choose model (1):

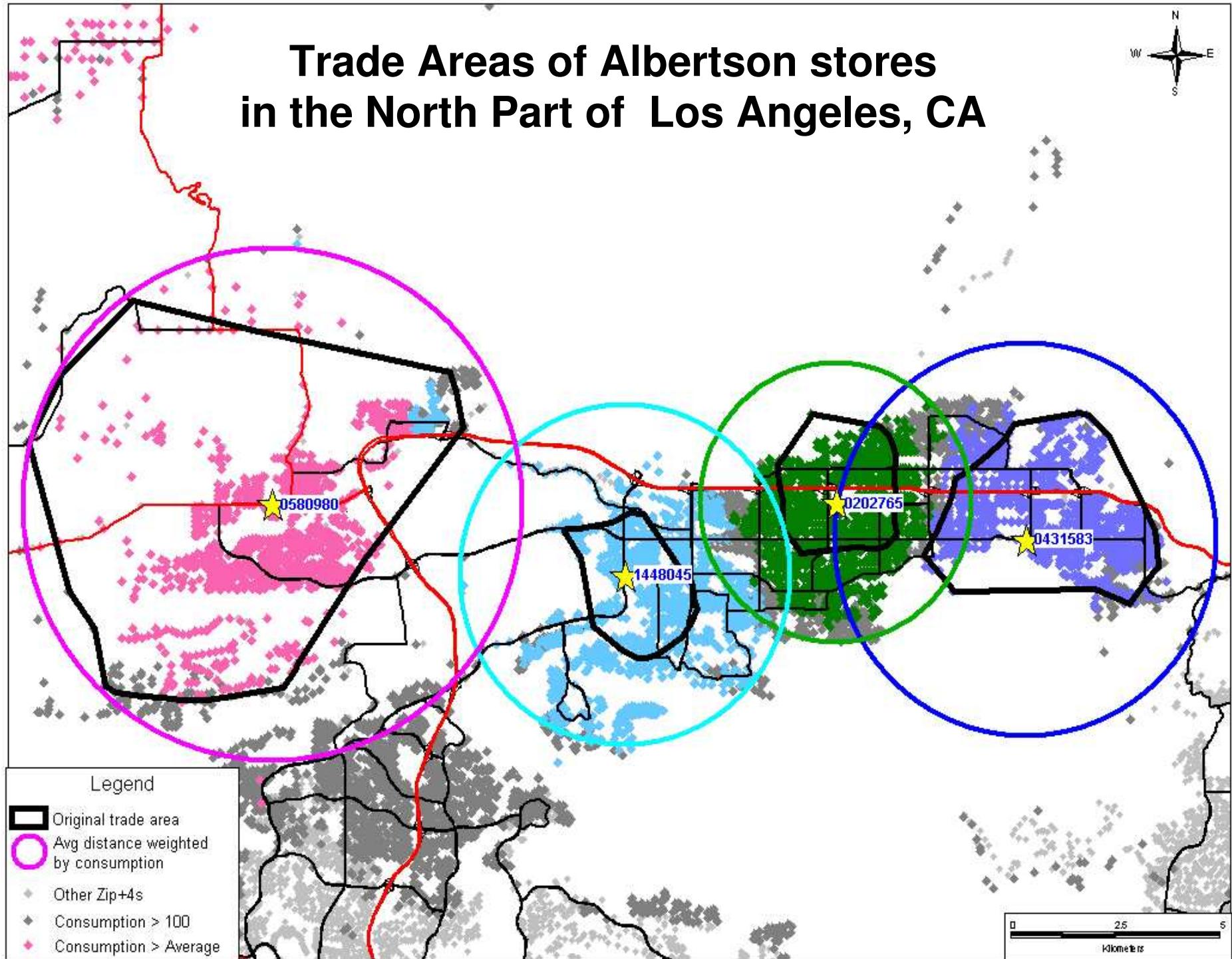
Correlation with stationary point y_∞ :

$$y_\infty = \lim_{n \rightarrow \infty} M^n(y), \quad \text{where} \quad \hat{y} = M(y) \equiv \text{Norm} \bullet \text{MWM}(y | \text{Hist} \rightarrow \text{Hist})$$





Trade Areas of Albertson stores in the North Part of Los Angeles, CA



G5 CRM Store Segmentation

Generation5
Mathematical Technologies Inc.

Parent Name

Albertsons Inc/HQ

Retailer Division

Albertsons Inc/Rocky Mtn Div

Category

Cheese

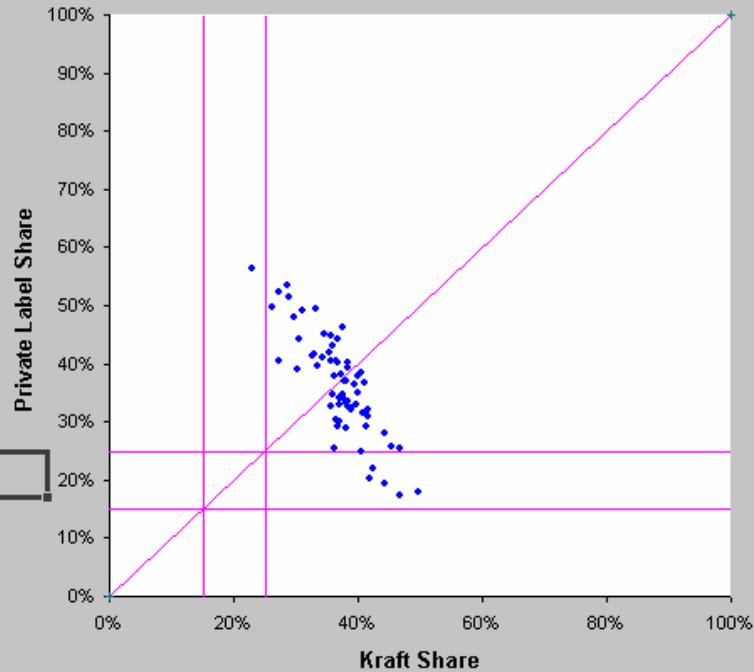
Share Ranges

Low ≤ < Medium ≤ ≤ High

Private Label	Kraft		
	Low	Med	High
High	* 0	* 1	* 59
Stores, %	0%	2%	91%
Med	* 0	* 0	* 5
Stores, %	0%	0%	8%
Low	* 0	* 0	* 0
Stores, %	0%	0%	0%

**** Total Stores : 65**

Albertsons Inc/Rocky Mtn Div Average Kraft = 37.%
 Cheese Average PrLb = 36.%



* Click to see detailed store address listing by segment.

Share is calculated by dividing estimated sales of Kraft or PL, by total sales (i.e.Kraft+PL+AO)

**Click to see complete store listing.

TDLinx Source: Retailer Division Name and All Commodity Volume

Results (Aggregated to Store)

? Questions ?

Automated Trade Area Analysis.
Case Study of G5 MWM Software Application



Alex.Zolotovitski@Generation5.net

Oct 29, 2004

Automated Trade Area Analysis. Case Study of G5 MWM Software Application



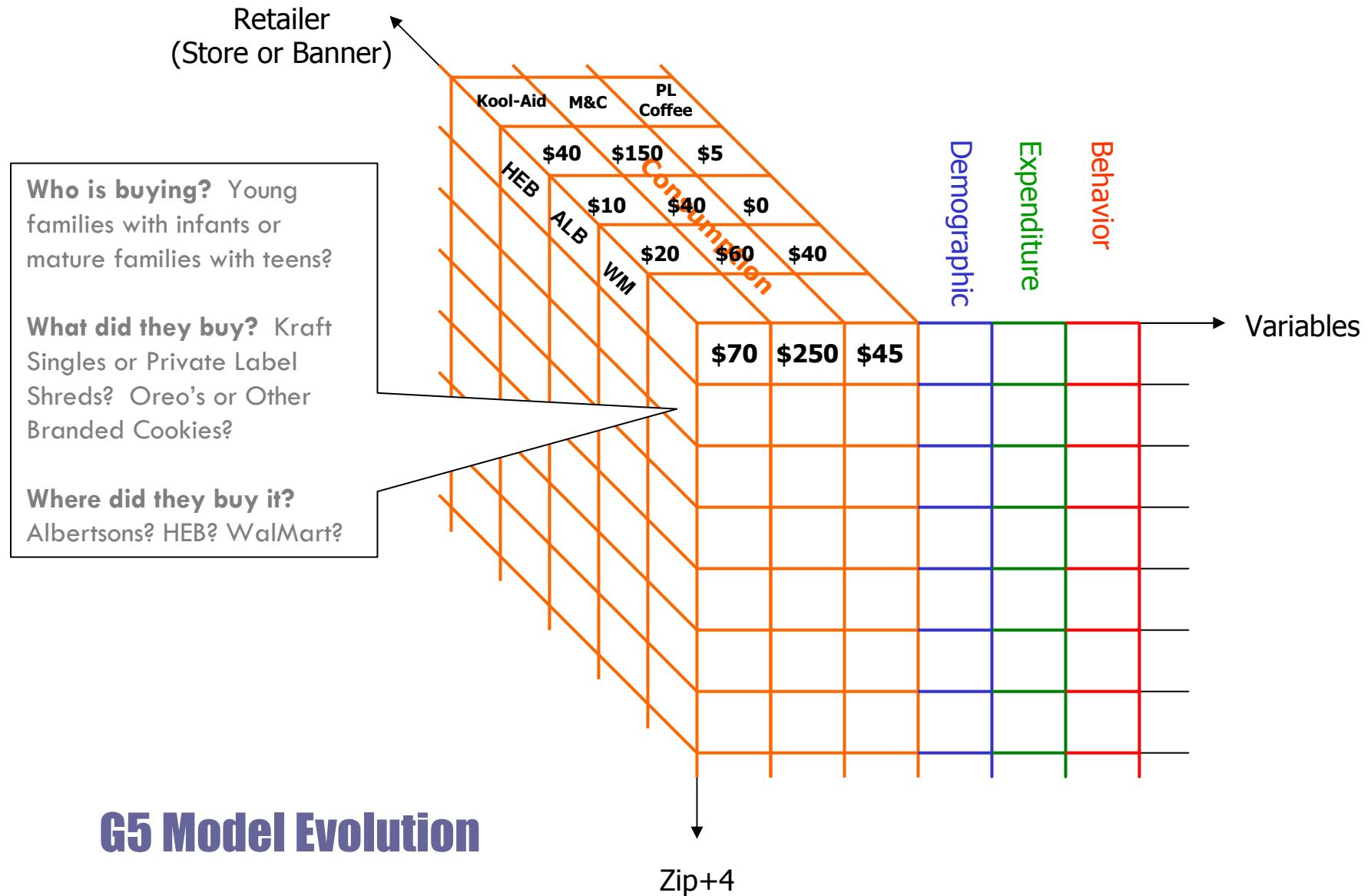
Alex.Zolotovitski@Generation5.net Oct 29, 2004

Generation5

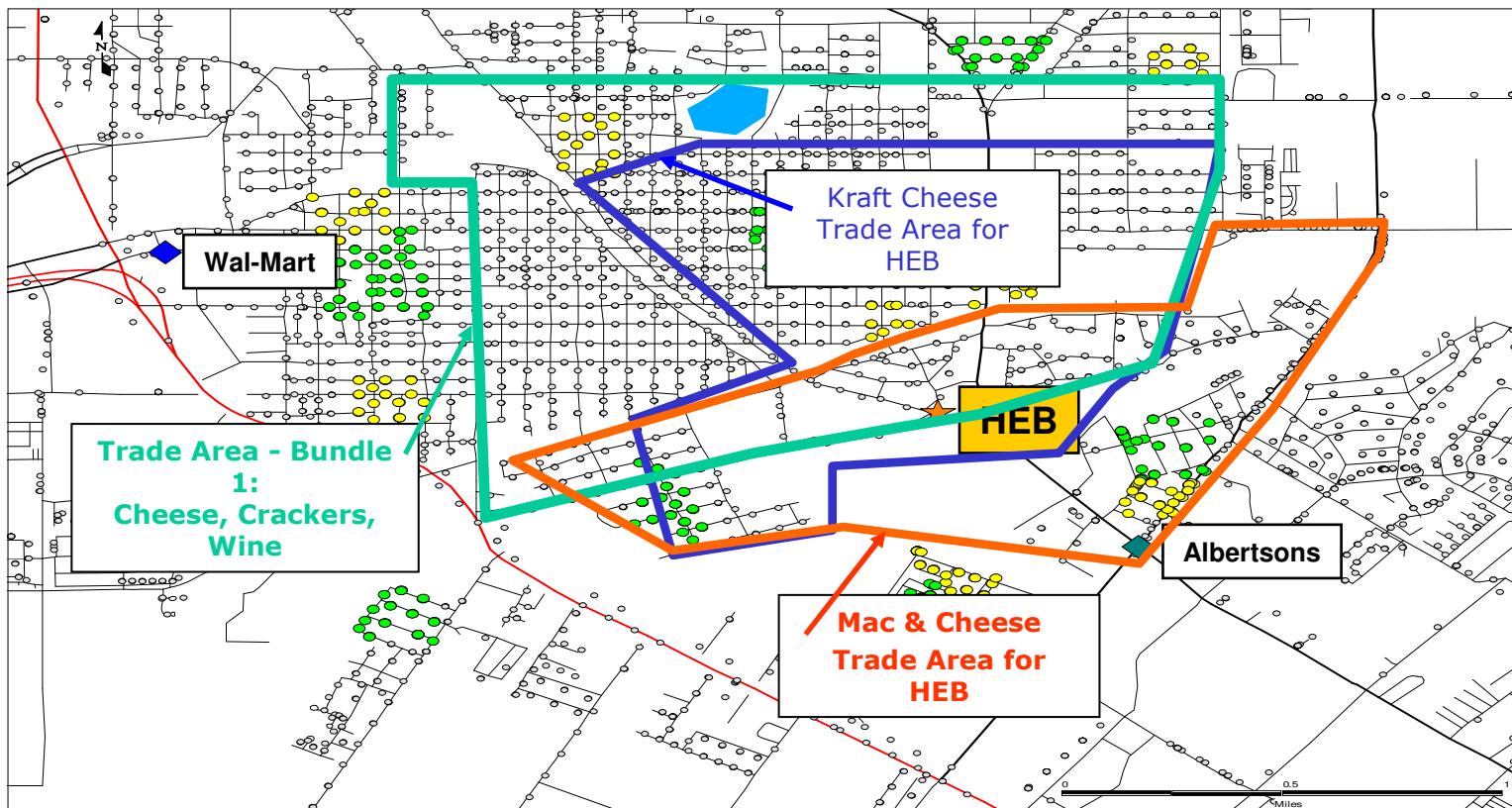
The pieces of the puzzle

1. Business Objectives and Applications
2. Data Sources
3. Challenges
4. Solution Approach & Automation
5. Deployment

Business Objectives and Applications



Business Objectives and Applications



Trade Area - Breaking point method for

Stages of the Trade Area Automation Project.

1. Data Analysis and preparation.

1. Matching ACN and TDLinx Data.(Olt-StoreName Matching).
Grouping Outlets (Olt) -> Groups of Outlets (GOlt), Criteria for the Grouping.

2. Modeling. Creating Outlet Group (GOlt) Level Historical Dataset.

2. Variable creation and selection.
3. Independent Variables.
4. Dependent Variables (**Non-linear Transformation**).
5. **Criterion for model optimization. Procedure of Express Cross-Validation.**
6. Model Selection (**MWM Distance, Granulation**, Sampling). **Refinement** of chosen model using Global Regression. Choosing local regression Coefficients to distribute \$ from GOlt to Store level.
7. Creating Store-Level Historical Dataset for validation of model and the next versions.
8. Validation of distribution to Store Level in two local regression models and 8 MWM iterations from them.

Stages of the Trade Area Automation Project.

3. Creating Outlet Group (GOlt) Level Target Dataset.

9. Joining Harold's (Zip9, Lon, Lat) and Jeff's (Zip9, G5, Prd\$) data.
10. Splitting areas (regions) to chunks (blocks).
- 11. Monitoring block execution.**
12. Creating GOlt Level Target Dataset in each block.

4. Deployment of model.

- 13. VB code for asynchronous execution of MWM in all blocks with a number of servers.**
- 14. Regression Refinement.**
- 15. Distribution \$ from GOlt to Store level.**
16. Aggregation of blocks.
17. Analysis of results.

Kraft US Trade Area Automation Project

Three Challenges --- Three Solutions

in

1. Data preparation.
2. Modeling.
3. Deployment.

Choosing model

Developed simple and accurate method of express cross-validation:

```
Data Target;    set Hist;   TargetInd = 1; y= . ;
Data CrossVal; set Hist Target; run;

/* proc MWM data = CrossVal ; */

proc SQL;
select sum((h.y-t.y)**2) into :Criterion
from Hist h,  CrossVal(where=(TargetInd=1)) t
where h.ID=t.ID;
```

Criterion to choose model: $\text{Min } \sum (y_{\text{fact}} - y_{\text{pred}})^2$



**Vladimir B. sends
Alex Z. and Helen
to TrArA.**