A Bayesian Approach to Data Base Integration

NYC INFORMS

March 17, 2004

James N. Arvesen, Ph.D.

The Basic Problem

- Have a data series-spatial/temporal-that is either a census or a statistical random sample.
- Have a "correlated" data series that is a statistical convenience sample.
- e.g. The random sample might be quarterly household demographics at the zip code level. The statistical convenience sample might be the number of credit cards by issuer at zip from several processors
- e.g. In the pharma industry, the random sample might be monthly drug Rx's (at the product level) by zip code. The convenience sample might be patients purchasing the particular drug each month for each zip code

Variety of data sources

The Pharmaceutical industry uses a variety of data sources to analyze markets, establish promotional strategies, deploy field organizations, measure representatives' performance, etc.

Some widely used data types include:

- Sales volume data (shipments, internal sales)
- Prescription delivery data
- Call reporting data (internal reports, 3rd party audits)
- "Patient level" data

Benefits

Each data source fulfills different needs and presents a different advantage/disadvantage profile

- Some sources enjoy a high level of coverage, but provide little insight in drug usage (shipment, call reporting)
- Some sources provide better tracking of drug usage, but lack in coverage
- Individually they don't provide a complete picture of the drug market
- Collectively, they suffer from disparate levels of coverage (by geography), making it difficult to compare and combine them

Other fields face incomplete data

Baseball illustration: 1970 Batting Averages for 18 MLB Players and Estimates

Player	Batting average first 45 at bats	Batting avg remainder of season	Bayesian Estimate
1. Clemente (Pitts, NL)	.400	.346	.351
2. F. Robinson (Balt, AL)	.378	.298	.329
3. F. Howard (Wash, AL)	.356	.276	.308
4. Johnstone (Cal, AL)	.333	.222	.287
5. Berry (Chi, AL)	.311	.273	.273
6. Spencer (Cal, AL)	.311	.270	.273
7. Kessinger (Chi, NL)	.289	.263	.268
8. L. Alvarado (Bos, AL)	.267	.210	.264
9. Santo (Chi, NL)	.244	.269	.259
10. Swoboda (NY, NL)	.244	.230	.259
11. Unser (Wash, AL)	.222	.264	.254
12. Williams (Chi, AL)	.222	.256	.254
13. Scott (Bos, AL)	.222	.303	.254
14. Petrocelli (Bos, AL)	.222	.264	.254
15. E. Rodriguez (KC, AL)	.222	.226	.254
16. Campaneris (Oak, AL)	.200	.285	.242
17. Munson (NY, AL)	.178	.316	.218
18. Alvis (Mil, NL)	.156	.200	.194

Searching for improved estimates

$$X_i \mid \theta_i \sim N(\theta_i, 1), i = 1, \ldots, k(=18)$$

after standard arcsine transformation.

IMPROVED ESTIMATOR

$$\overline{X} + (1 - (k-3)/V)(X_i - X)$$

$$V = \sum (X_i - X)^2$$

This is a standard shrinkage estimator.

See B. Efron and C. Morris, “Data analysis using Stein’s estimator and its generalizations”, *JASA,* 70, 311-319 (1975)

Example: Patient level data

Newer type of data, linking prescriptions to patients (unidentified) offers opportunities for improved understanding of market dynamics

- Compliance tracking
- New patient measurement
- Brand switching

Example: Patient level data

Data presented by suppliers poses several challenges to analysts, limiting its value

- Overall capture rate of samples may be low
- Capture rate of samples may vary by region
- Capture rate of samples may vary within a wide range at lower level of geography (5-digit zip codes)
- Capture rate of samples may vary over time
- Data often supplied with no projection at low geographic level

Situation is comparable to early days of prescription data

Projecting patient level data

Applying “simple” projection methodology to patient-level data results in reliable estimates at a higher geographic level (nation, region), but estimates at lower geographic levels (5-digit zip code) are less robust.

- Large samples, covering a high percentage of actual transactions, lower level of uncertainty in estimates
- Small samples, covering a small percentage of actual transactions, lead to projections with larger range of uncertainty

Symptoms of data issues

City	State	Zip Code	Total Population	Pop over 60	Product Shipment	Rx Capture	Simple Projection
BROOKLYN	NY	11225	114373	9527	1711235	46%	2767
CHICAGO	IL	60622	108608	9726	1996801	52%	2627
LOS ANGELES	CA	90095	94319	6505	1493657	39%	2281
SAN JOSE	CA	95111	44444	4922	757016	36%	716
GREENSBORO	NC	27405	44396	6783	793268	14%	0
COLUMBUS	OH	43228	44380	5359	876083	44%	1431
W. TOWNSEND	MA	01474	7301	1124	121153	41%	235
ATHENS	TN	37371	2436	397	45846	11%	0
SHADY DALE	GA	31085	2432	322	80876	40%	39
ROSE HILL	MS	39356	2432	447	133518	8%	0
FLORENCE	MA	01062	208	1	0	9%	6
CONCORD	MA	01742	416	61	6275	35%	6
PETALUMA	CA	94954	464	49	14369	20%	7
NOVATO	CA	94949	510	66	7564	18%	16
GILROY	CA	95020	1022	140	47475	17%	16
ABILENE	TX	79697	0	0	0	0%	0

Similar population and shipment, very different projections

Population and shipment values, no projection

Range in capture rates may explain “strange” projection results

Different population, shipment, same projection

PMSA
Pharmaceutical Management Science Association

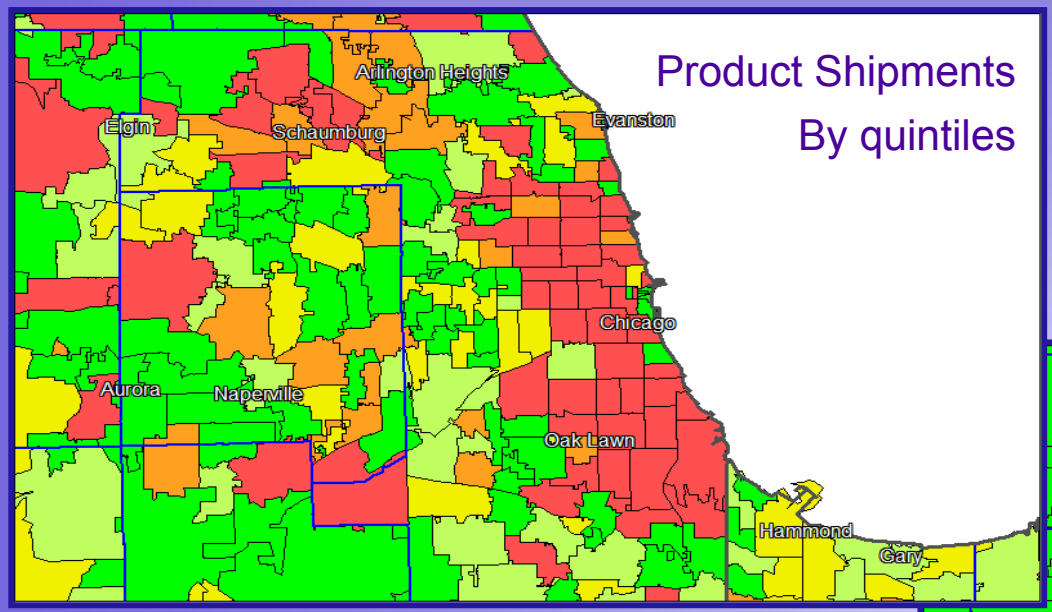
Symptoms of data issues

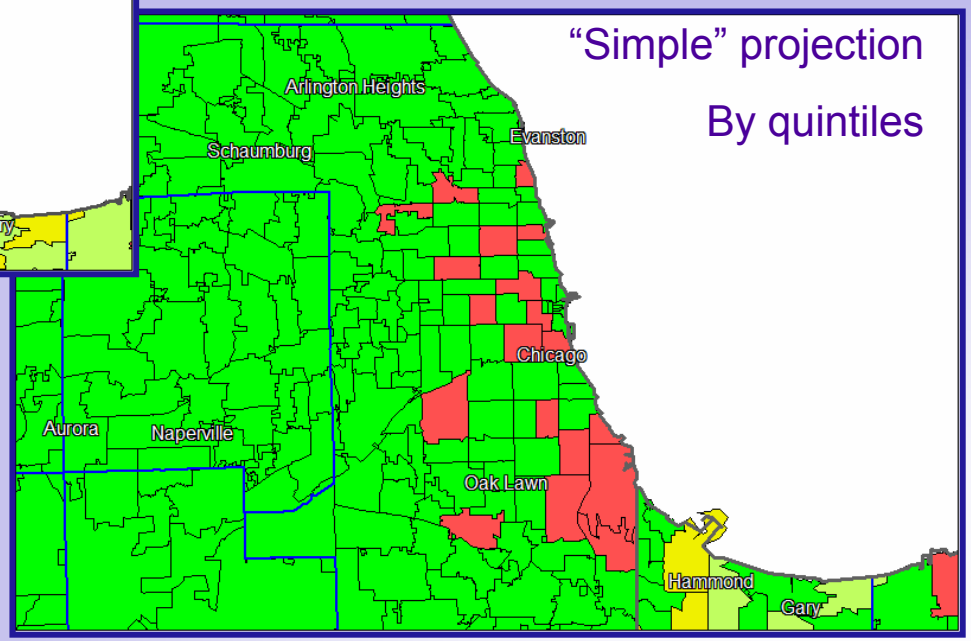
City	ST	Total Pop	Pop > 60	Product Ship $	Simple Projection	Rank by Pop	Rank by Pop > 60	Rank by Prod Ship	Rank by Proj
BROOKLYN	NY	114373	9527	1711235	2767	1	234	5	4
CHICAGO	IL	108608	9726	1996801	2627	2	213	1	9
LOS ANGELES	CA	94319	6505	1493657	2281	8	1008	14	23
SAN JOSE	CA	44444	4922	757016	716	560	2065	576	1786
GREENSBORO	NC	44396	6783	793268	0	561	876	471	20674
COLUMBUS	OH	44380	5359	876083	1431	562	1713	305	388
W. TOWNSEND	MA	7301	1124	121153	235	9197	9533	11159	5314
ATHENS	TN	2436	397	45846	0	15963	16446	15738	26278
SHADY DALE	GA	2432	322	80876	39	15979	18348	13204	11617
ROSE HILL	MS	2432	447	133518	0	15978	15538	10382	23823
FLORENCE	MA	208	1	0	6	33672	35494	28546	19078
CONCORD	MA	416	61	6275	6	31389	32001	23535	18975
PETALUMA	CA	464	49	14369	7	30867	32722	20741	18631
NOVATO	CA	510	66	7564	16	30329	31717	23111	16022
GILROY	CA	1022	140	47475	16	24404	26917	15603	15887
ABILENE	TX	0	0	0	0	35666	35540	24944	30658

Simple projection method leads to estimates that are inconsistent with other data profiles

PMSA
Pharmaceutical Management Science Association

“Simple” projection





Quintile 5
Quintile 4
Quintile 3
Quintile 2
Quintile 1

PMSA
Pharmaceutical Management Science Association

Goals of advanced projection

Advanced data projection methodology is required to improve the value of patient level data for analysis that is geographically driven

- Reduce the range of uncertainty around projections based on low capture rates of samples
- Make various data sources more “geographically comparable”
- Overcome sample variations over time

Bayesian estimation

Use Bayesian approach to improve estimates

- Bayesian smoothing estimation technique (borrowed strength) applied to improve estimates at zip code level when low capture samples are available
- Bayesian approach is superior to ratio estimation, especially when samples represent diverse percentages of transactions
- Bayesian approach can be used over time to further improve estimates

Bayesian estimation

Methodology concepts

- Bayesian Smoothing allows one to use estimates from regions that are similar in geography, demographics, and other important variables to improve estimates in regions where data is scarce or missing.
- Applied to patient level data, the approach consists in using other, more robust data variables to provide better estimates of patient data and reduce the range of uncertainty in areas with low capture rates

  Example: zip codes with similar demographic profile, shipments, or in similar geographic areas

Bayesian estimation

This methodology is used in a variety of applications, for example:

- Improve estimates of disease occurrence
  - L. A. Waller, B. P. Carlin, H, Xia, and A. E. Gelfand, “Hierarchical spatio-temporal mapping of disease rates” *JASA* **92** 607 – 617 (1997)
  - D. G. Clayton and L. Bernardinelli “Bayesian methods for mapping disease risk” in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies,* eds P. Elliot, J. Cuzick, D. English, and R. Stern. Oxford University Press, (1992)
- Design spam filters in email systems

Apply Bayesian Smoothing model to geographic estimates

Let $Y_i, i = 1,....,36875$

Be the true number of patients in zip code i.

Covariate data in zip code i:

- Total population
- Population over 60
- Product Shipments ($)

$\theta_i, i = 1,....,36875$, as in the baseball example

$$\varphi_i \mid \lambda \sim CAR(\lambda), i = 1,....,36875,$$

$$\varphi_i \mid \varphi_{j \neq i} \sim N(\overline{\varphi_i}, 1/(\lambda n_i)),$$

$$\overline{\varphi_i} = n_i^{-1} \sum_{j \in \delta_i} \varphi_j,$$

Where the summation is over the set of neighbors of region i.

Then the model is: $Y_i = X_i \beta + \theta_i + \varphi_i, i = 1,....,36875$

Bayesian estimates added to data

City	State	Zip Code	Total Population	Population over 60	Product Ship $	Simple Projection	Bayesian Estimate
BROOKLYN	NY	11225	114373	9527	1711235	2767	2338
CHICAGO	IL	60622	108608	9726	1996801	2627	2220
LOS ANGELES	CA	90095	94319	6505	1493657	2281	1927
SAN JOSE	CA	95111	44444	4922	757016	716	605
GREENSBORO	NC	27405	44396	6783	793268	0	438
COLUMBUS	OH	43228	44380	5359	876083	1431	1209
WEST TOWNSEND	MA	01474	7301	1124	121153	235	198
ATHENS	TN	37371	2436	397	45846	0	22
SHADY DALE	GA	31085	2432	322	80876	39	32
ROSE HILL	MS	39356	2432	447	133518	0	25
FLORENCE	MA	01062	208	1	0	6	5
CONCORD	MA	01742	416	61	6275	6	5
PETALUMA	CA	94954	464	49	14369	7	5
NOVATO	CA	94949	510	66	7564	16	13
GILROY	CA	95020	1022	140	47475	16	13
ABILENE	TX	79697	0	0	0	0	

Estimates improved in areas where simple projections appeared faulty

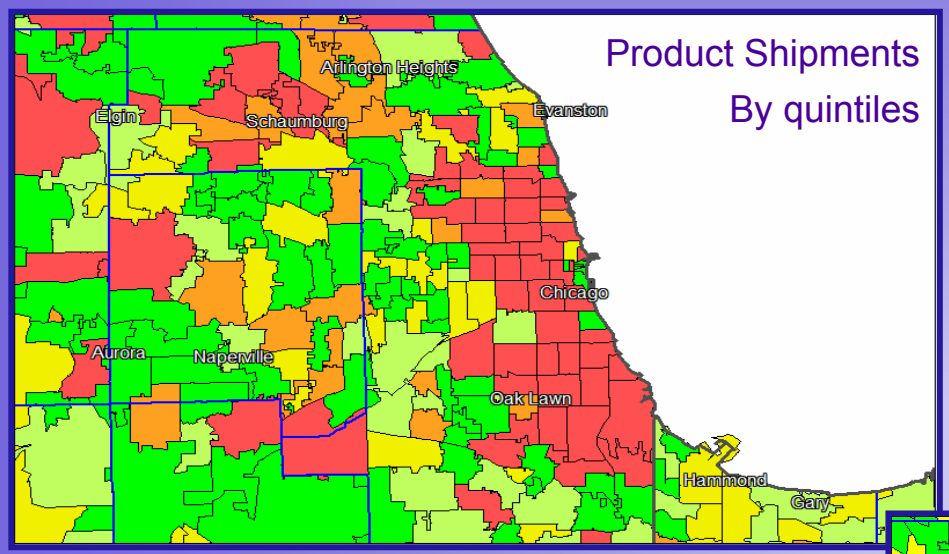
PMSA
Pharmaceutical Management Science Association

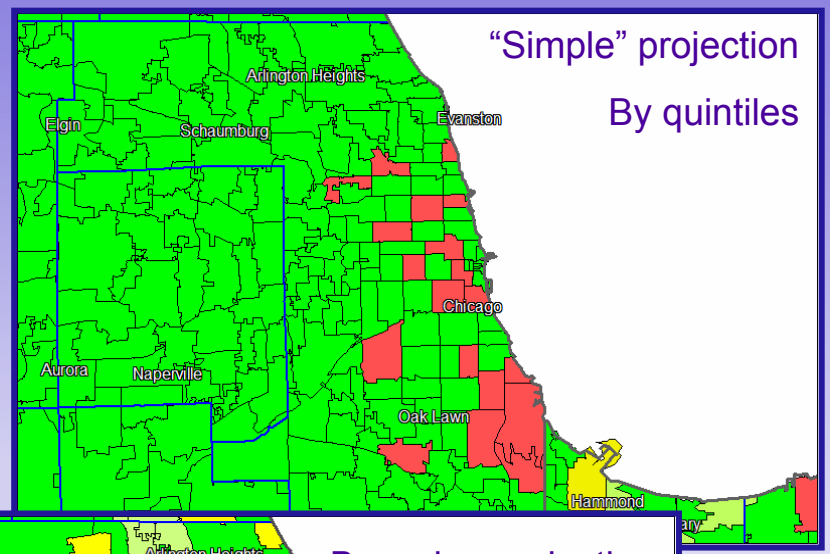
Bayesian improvements

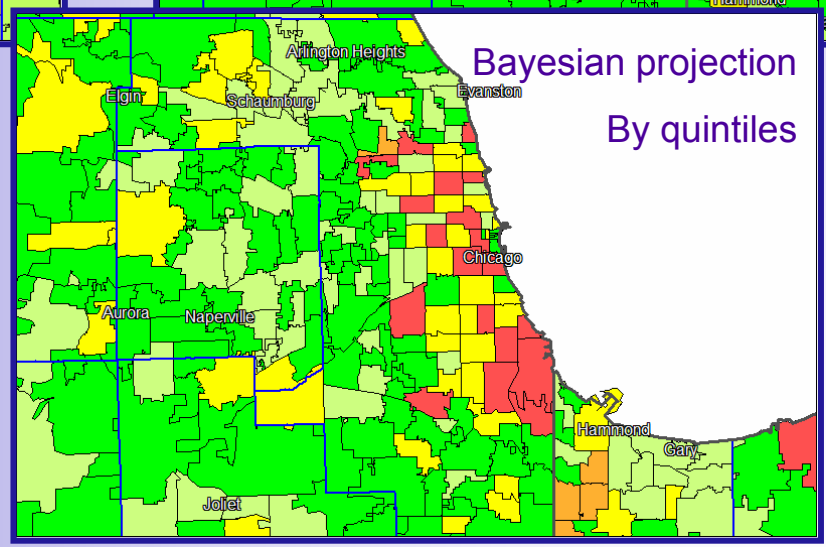
City	Total Population	Product Ship $	Simple Proj	Bayesian Estimate	Rank by Pop	Rank by Prod Ship	Rank by Proj	Rank by Bayesian
BROOKLYN	114373	1711235	2767	2338	1	5	4	4
CHICAGO	108608	1996801	2627	2220	2	1	9	9
LOS ANGELES	94319	1493657	2281	1927	8	14	23	23
SAN JOSE	44444	757016	716	605	560	576	1786	1787
GREENSBORO	44396	793268	0	438	561	471	20674	2858
COLUMBUS	44380	876083	1431	1209	562	305	388	388
W. TOWNSEND	7301	121153	235	198	9197	11159	5314	6073
ATHENS	2436	45486	0	22	15963	15738	26278	19565
SHADY DALE	2432	80876	39	32	15979	13204	11617	16166
ROSE HILL	2432	133518	0	25	15978	10382	23823	18333
FLORENCE	208	0	6	5	33672	28546	19078	31892
CONCORD	416	6275	6	5	31389	23535	18975	31789
PETALUMA	464	14369	7	5	30867	20741	18631	31445
NOVATO	510	7564	16	13	30329	23111	16022	24101
GILROY	1022	47475	16	13	24404	15603	15887	23966
ABILENE	0	0	0		35666	24944	30658	35666

Bayesian smoothing reduces the gap between estimates for records that are related

Improved projections



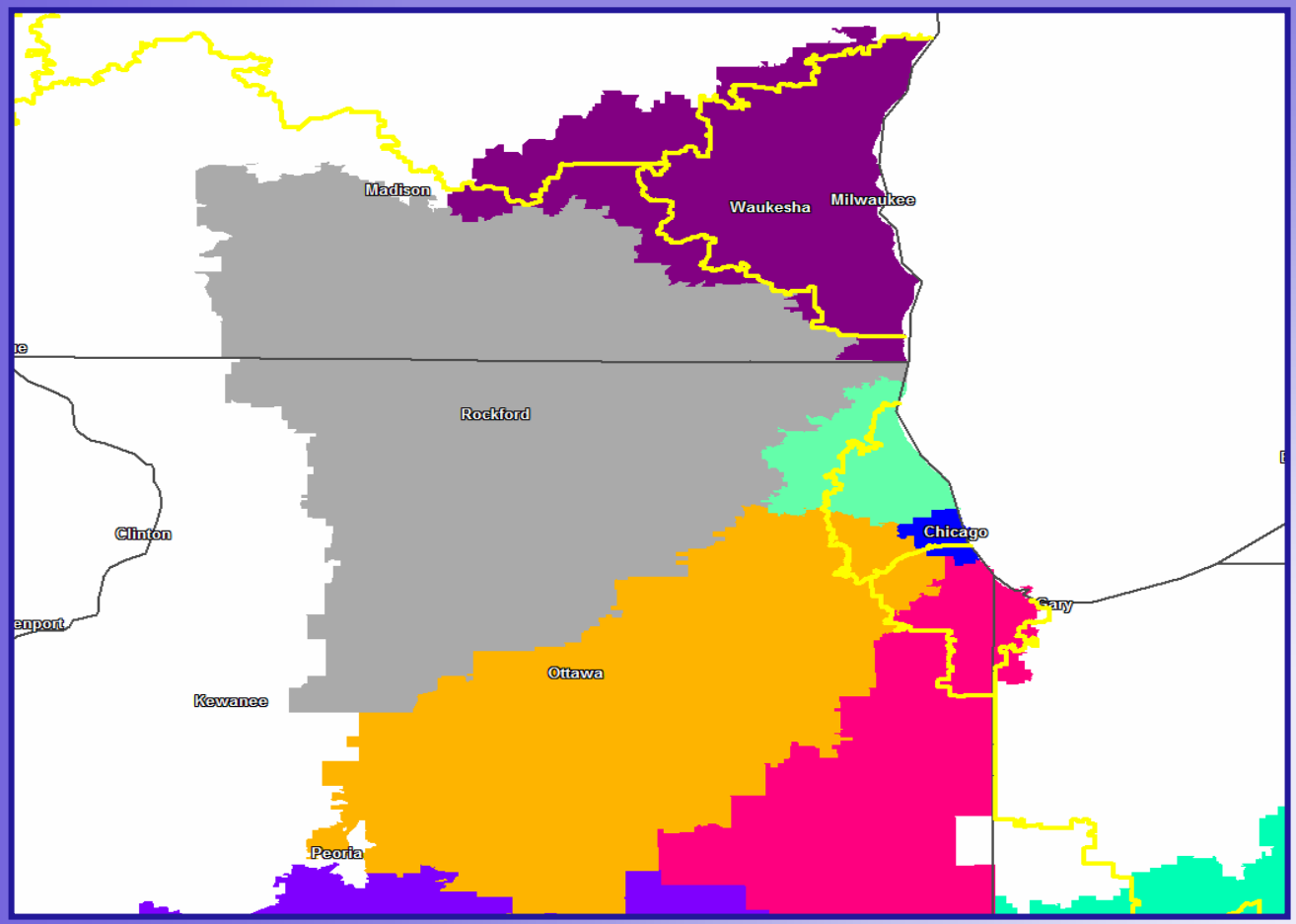




Quintile 5
Quintile 4
Quintile 3
Quintile 2
Quintile 1

PMSA
Pharmaceutical Management Science Association

Impact on territory alignment



Shaded territories: based on Bayesian estimates

Yellow boundary territories: based on “Simple” projections

Alignment based on Bayesian estimates recommends deploying 6 territories in SE Wisconsin and NE Illinois, when “simple” estimates suggest 4 territories (including 1 large territory extending across northern Illinois).

PMSA
Pharmaceutical Management Science Association

Conclusions

Bayesian smoothing approach helps improve projections at a finer geographic level

- Local estimates are more reliable
- Estimates from various regions are more comparable
- Estimates improve over time

Improved estimates provide stronger basis for geography sensitive tasks

- Market analysis
- Targeting
- Sales force deployment