

## Locating Prospective Performing Arts Patrons: A New Use of a Statistical Model

### **Carol Yamamoto**

Professor, Department of Arts, Entertainment & Media Management  
Columbia College Chicago  
Principal  
Collins Yamamoto Consulting, Chicago, Illinois

### **James Dauer, Ph. D.**

Graduate Program Coordinator  
Department of Computer Science & Information Systems  
Elmhurst College, Elmhurst, Illinois  
Associate, Collins Yamamoto Consulting, Chicago, Illinois

### **Maury Collins**

Principal  
Collins Yamamoto Consulting, Chicago, Illinois

### **Jeff Andreasen**

Senior Research Associate, Richard Day Research, Inc., Evanston, Illinois  
Associate, Collins Yamamoto Consulting, Chicago, Illinois

*Carol Yamamoto has taught graduate and undergraduate strategic planning and arts marketing courses since 1984, during which time she has also provided arts management counsel to clients including the National Endowment for the Arts.*

*James Dauer has held the Willet Chair for Distinguished Research in Business and Economics and has received the President's Award for Excellence in Teaching at Elmhurst College. He has authored over two dozen articles on topics ranging from artificial intelligence to statistical analysis.*

*Maury Collins co-founded Collins Yamamoto Consulting in 1991 after working for more than 20 years managing commercial and nonprofit theatres and working as a management consultant in Washington providing event management, lobbying and program development services.*

*Jeff Andreasen has provided research services to Collins Yamamoto Consulting for many arts marketing and feasibility studies. He has over twenty years of primary and secondary research experience, including analysis of large databases.*

### **Abstract**

A more efficient, faster way to identify and locate prospective ticket buyers of nonprofit performing arts enables an organization to target more effectively its most likely audience, achieve greater market penetration and increase earned income. This two-tiered approach uses statistical analysis of an arts organization's existing database of subscription and single ticket customers to determine key demographic characteristics. The subsequent deployment of a non-linear multiple regression model geographically identifies "pockets" of residents most likely to generate future sales, based on US Census tract demographics. The model calculates the current level of market penetration per census tract, suggests areas of potential market saturation, and identifies new markets outside of the organization's

traditional focus. The index of tract identification numbers gives the organization's staff uncomplicated means in future marketing campaigns to pursue tracts with the highest ticket sales potential.

### **Keywords**

Nonprofit Performing Arts Marketing, Audience Research, Correlation Analysis, Multiple Regression

## **Introduction**

Market research, as typically conducted in the nonprofit performing arts, usually outlines an audience's general demographic profile, but does not always reveal where potential ticket buyers can be found. For companies with very large databases, (>50,000 names), commercial mailing houses can match existing customers to potential customers in the so-called "lifestyle" databases, but the yield of names is typically a very small percentage of the available population, and the cost *per name* developed can run into the hundreds of dollars. Even with the identification of market prospects through these kinds of resources, performing arts organizations gain little new insight into their customers. Without additional effort, the organization often finds it difficult to focus its relatively limited marketing resources directly on the target market. With the intent of exploring whether a more efficient, faster way to obtain better audience information could be found, Chicago Opera Theater (COT) agreed to let Collins Yamamoto Consulting conduct a pilot project involving an innovative market analysis of COT's database of subscription and single ticket customers. COT is an opera company with an annual budget of approximately US\$3,000,000 that produces a season of three opera productions per year with 5 performances of each opera.

Collins Yamamoto Consulting (CYC) has worked for several years using statistical analysis and mathematical modeling to develop forecasts about various aspects of the performing arts and other business operations. One of the CYC associates has successfully performed analyses of this type in multi-billion dollar food and service industries with stand-alone locations, using non-linear multiple regression analysis for predicting potential revenue for proposed sites, as well as for estimating the impact a new site would have on existing stores' sales. Such studies recently have expanded into the international business arena, a further indication of their utility as a strategic planning tool. Combined with correlation analysis, regression models have also proven useful in identifying customer profiles, a prerequisite to the development of useful marketing strategies regardless of the industry involved. CYC developed this pilot study to see if these marketing decision tools could be as valuable to arts organizations as they have proven to be for for-profit businesses. As far as CYC has been able to determine, this type of tool has never been applied to a nonprofit performing arts organization before this study.

### **Problem Statement**

Chicago Opera Theater is a primary constituent of the new Joan W. and Irving B. Harris Theater for Music and Dance (the Harris) and was relocating its performances in the 2003-2004 season from a neighborhood-based venue to Chicago's central downtown Harris Theater. The premise of the Harris is that a consistent, central location will enhance marketing opportunities for the constituent groups. COT did not intend to decrease its number of performances, although, this move of about five miles (8 kilometers) represented an increase in seating capacity of more than 50%.

The goal of this pilot study was to test the usefulness of these statistical tools as applied to the performing arts. If proven useful, the analysis would enable COT to identify by census tract

location where its greatest potential for additional sales could be found. COT would have the opportunity to refine its marketing strategy so that it could more effectively target its most likely audience, achieve greater market penetration and consequently increase earned income.

## Methodology

### Overview

To customize this analysis to COT's specific audience, CYC applied multivariate statistical tools to the COT database of subscription and single ticket customers from the years 2000-2002. The analysis combined Year 2000 US Census information at the census tract and zip code level for the Chicago metropolitan area with the current COT audience ticket purchase data, which consisted of name, address, zip code and total sales. Neither demographic data nor fund raising information about these customers were available.

As defined by the United States Census Bureau, "US Census tracts are small, relatively permanent statistical subdivisions of a county that usually have between 2,500 and 8,000 persons and, when first delineated, are designed to be homogeneous with respect to population characteristics, economic status, and living conditions."<sup>1</sup> Census tracts were used because they tend to be smaller in area and contain more homogeneous demographic variables than 5-digit postal zip codes.

CYC developed a mathematical model to predict new sales potential as a function of tract demographics and distance from both the current and new venues. For the model, over a thousand variables<sup>2</sup> were considered, including a number that were created to adjust for the differences in sizes of the census tracts. For both venue locations, CYC identified a set of approximately 150 variables that correlated with ticket sales over the past three years. Of the 150 independent variables, 11 were the most statistically significant, and were selected to best describe the COT audience demographics.

These independent variables were checked for cross-correlation<sup>3</sup> and inherent redundancies were eliminated. Using the variables, CYC then developed the model to estimate potential ticket sales by census tract, based on tract demographics **and** distance from the venue. The model identified specific demographic characteristics of the COT audience and matched them with potential audience members in the metropolitan area at the census tract level.

Census tracts were then ranked in terms of their potential for generating additional ticket sales. Estimates of COT's market penetration identified for the company the most promising opportunities to further develop its audience.

### Procedure for Building the Model

COT provided its subscription and single ticket database, consisting of approximately 5,700 ticket buyer names, addresses and ticket purchase totals, for the three years. CYC analyzed the data, matched it to zip code and census tract data and developed the model to predict which census tracts held the greatest potential for additional sales.

Working in market research with performing arts companies over the last fifteen years, CYC repeatedly has observed that "nothing predicts behavior like behavior." In this study, the only behavioral information available was the total value of ticket purchases. In effect, a kind of two-

sided coin was created, with the COT data providing the behavioral information on one side and census data providing demographic information on the other side. Combining these two sides of information yielded a more complete identification of ticket buyers' characteristics. That information, in turn, was matched at the census tract level with other groups of people with similar characteristics.<sup>4</sup> Analysis at the five-digit zip code level is too big, and provides too much diversity; at the block (9-digit zip code) level, it is too small, too detailed and too much alike. However, census tracts do not line up by zip codes, and vary with population density, size, and geographic location.<sup>5</sup>

The model incorporated a multiple regression analysis, an inferential procedure used to assess the contribution of one or more **independent** variables (demographics of the tracts where COT patrons live) to a **dependent** variable (expressed as a three year average of sales). In other words, who are the types of people who buy tickets to COT? The regression analysis mathematically identified their characteristics and gave those characteristics "weights" of importance in relationship to the dependent variable – sales – relative to each other.

### **Predictive (Independent) Variables**

The chart in Exhibit 1 shows the independent variables that had the highest degree of relationship to ticket sales. They are ranked in a Relative Index that is a measure of the magnitude of each variable's influence versus the most highly correlated factor, in this case, College Degree. These variables were used in the regression analysis.

**Exhibit 1**

<u>Set of Predictive Variables</u>	<u>Relative Index</u>	<u>+ or – Influence</u>
• College Degree	1.000	+
• Total Population	0.679	+
• Age 55-59	0.519	+
• Age 45-54	0.450	+
• % HH Income US\$35-50K <sup>6</sup>	0.398	+
• % HH Income US\$75-100K	0.349	+
• % HH Income US\$15-25K	0.309	–
• % HH Income US\$100-150K	0.307	+
• % HH Income US\$25-35K	0.280	–
• % HH Income US\$150-250K	0.217	+
• Average HH Net Worth	0.211	+

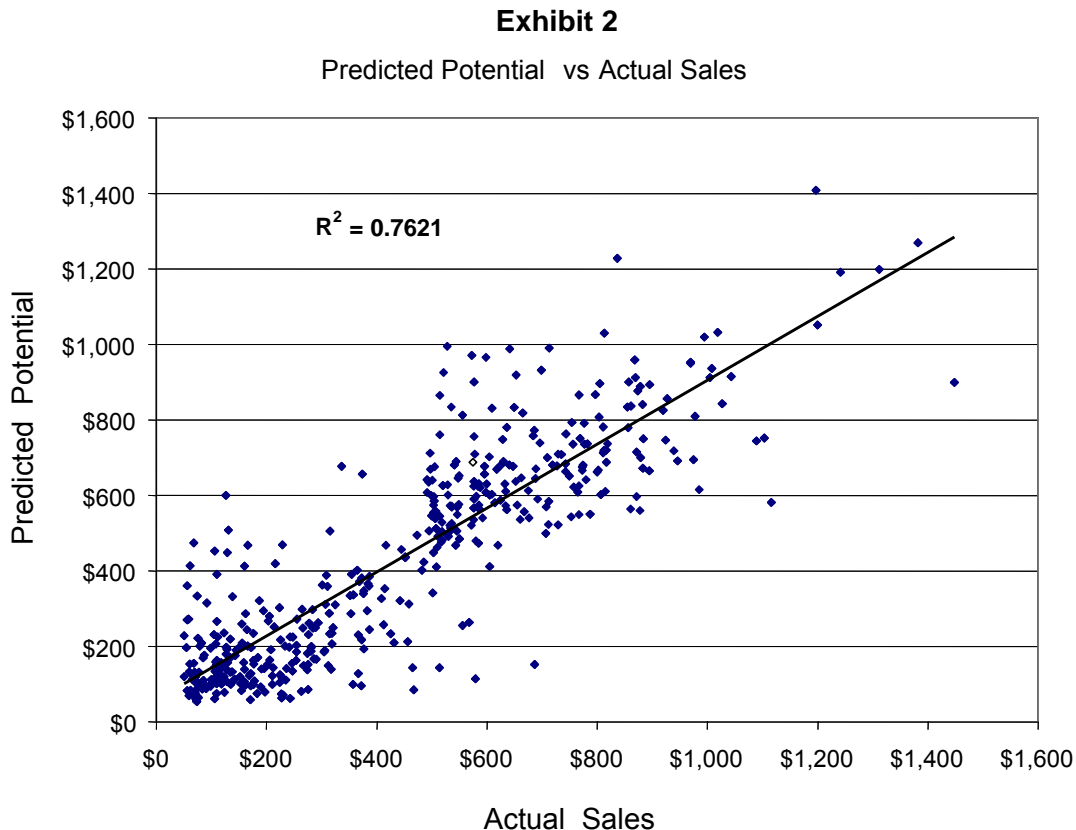
Separate and distinct from less or more advanced education levels, a four-year college degree was the strongest predictor of the more than 1,000 demographic variables considered. In other words, of all the things this audience may have in common, possessing a college degree was the most prevalent. The dominant position of college degree may also explain why often-related variables like higher income, family status or race/ethnicity did not appear to be stronger predictors. Because the analysis sought to minimize overlap among the variables, college degree may act as a surrogate for some others.

Total Population was also a strong predictor; i.e., the denser the population, the greater number of people within the census tract who will fit the COT audience profile, even though they may be

a small proportion of the total. As population grows, the number of “COT audience type” people will grow as well. Conversely, as population declines, other variables will decline as well.

Two of the variables had a negative influence or an inverse relationship; i.e., the more households in a tract with Household (HH) Income of US\$15-25,000 or US\$25-35,000, the less likely the tract’s households would buy tickets.

A random sample of 500 census tracts originating from approximately 5,700 COT current ticket buyers is shown in Exhibit 2. The model’s ability to correlate the 11 independent variables with sales in a tract has an overall accuracy of .76 or 76%, as measured by  $R^2$ .



It should be noted that this  $R^2$  score indicates a robust model and is significant because “generally, correlations tend to be higher in the physical sciences, where relationships between variables often obey uniform laws, and lower in the social sciences, where relationships may be harder to predict. A correlation of .4 between a pair of sociological variables may be more meaningful than a correlation of .7 between two variables in physics.”<sup>7</sup> Thus, while each variable by itself is not a strong predictor, together the 11 have an overall “accuracy” of 76%.

The model identified the census tracts that had the highest potential for additional sales, based on the degree of demographic characteristics that matched up to the model’s predictions. In order to realistically identify those potential audience census tracts, mathematical adjustments were made for distance, and minimums were set for ticket sales and tract population.

Sales potential needed to be “decayed” by distance in order to account for the customers’ willingness to drive so far, and no farther, to see the attraction. Distance was implicitly included

as a latent variable by its presence in the calculation for the decay of ticket sales as an inverse function of venue distance. A plot of the 3 year average ticket sales as a function of distance indicated that COT's ticket sales decrease non-linearly as distance from the venue increases. Non-linearly means that some audience members travel greater distances than others; CYC believes the difference in the distances traveled is attributable to their convenient access to public transportation and expressways. In order to realistically model this effect, average ticket sales were "decayed" using a non-linear exponential function. The resultant effect was to reduce sales very little for census tracts close to the venue (<1-3%) and up to 27% for tracts approaching or equal to 35 miles (56 kilometers) away. If sales had not been decayed, census tracts having similar demographics would yield comparable estimates of potential, regardless of distance from venue.

Average ticket sales and tract population filters eliminated the very small, scarcely populated areas. If any census tract in the sample failed one or more of the filters, the tract was excluded. The data filters were: average ticket sales  $\geq$  \$50 at tract level; distance from venue  $\leq$  35 miles; and census tract population  $\geq$  100 people.

By ranking the census tracts based on sales potential, the top 25% (approximately 500 tracts) were identified within the six county Metropolitan Statistical Area<sup>8</sup> of Chicago. These areas represent the greatest potential for marketing and outreach efforts because the residents of these census tracts most closely match the demographics of the opera company's existing customers.

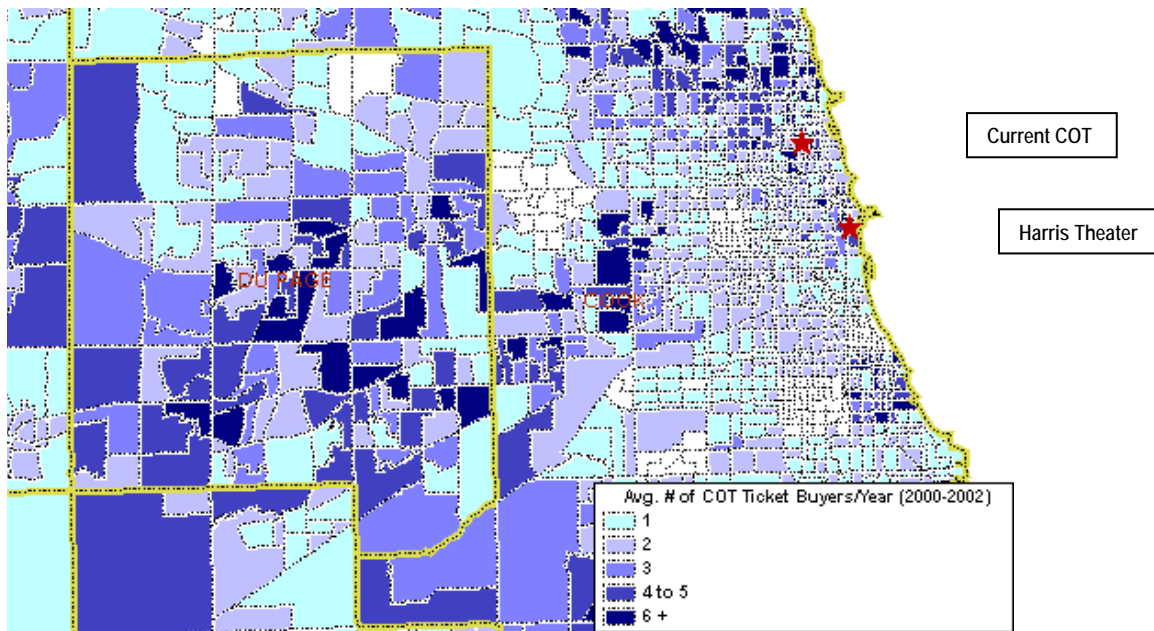
## Findings

### Overview

The model compared the three-year COT data with the year 2000 census tract data, at the neighborhood theater and Harris locations. It found the current market penetration of COT's audience by tract, as shown on the map<sup>9</sup> in Exhibit 3. Note that this map displays only a sample portion of the six county metropolitan area actually covered in the analysis.

This portion of the map shows the census tracts for a major portion of the city of Chicago, its suburbs, and the county due west of it, with heaviest market penetration in the darkest shade and decreasing numbers per tract in lighter shades. It also shows the locations of the neighborhood venue as "Current COT" and the Harris Theater. The darkest penetration number, six plus, actually covers a range of 6 (average number of ticket buyers per year) to as many as 300+ and in some cases, COT has >300 per census tract. For the range of census tracts sizes, 2,500 to 8,000 people, this would indicate penetration of as much as 15% to 3.75% of the population in the tract.

**Exhibit 3**  
**COT's existing Neighborhood Theater Audience – Location by Census Tract.**



### Tract Rankings

Estimates for additional ticket sales potential were found for each tract and the results sorted. The “Super tracts” (those with the greatest potential for additional sales) were grouped together, and the remainder was divided into Quartiles 1 through 4, or respectively, high, medium, low and poor potential for additional sales.

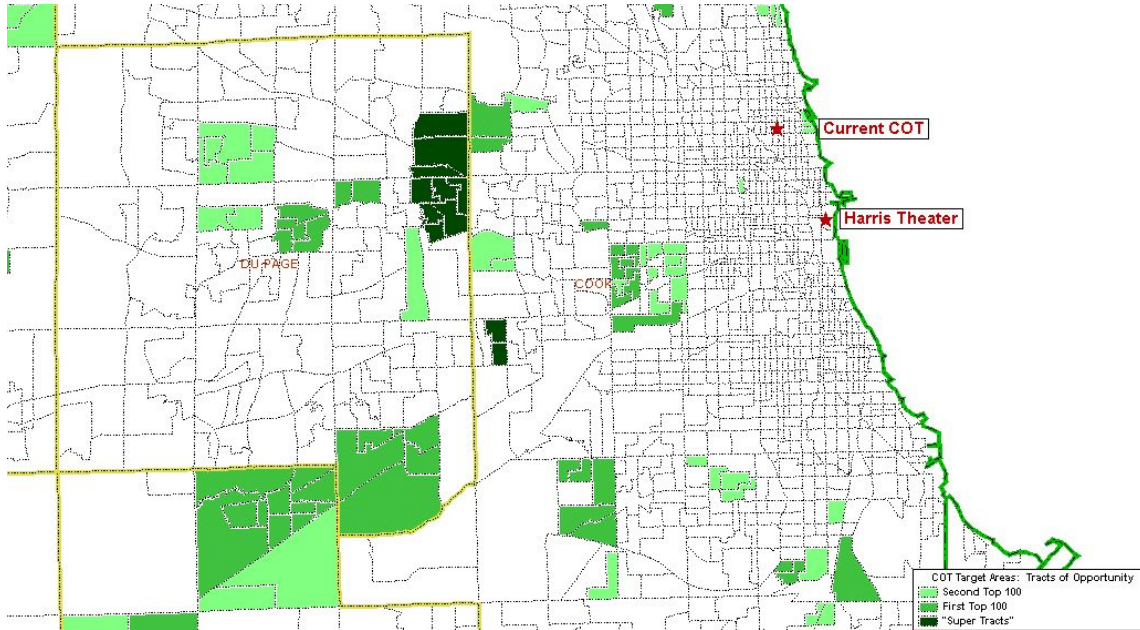
Estimates within each Quartile were indexed relative to the highest value within the group and then relative to the overall, regardless of Quartile. This index becomes a reliable locator (or where to look for) more customers like those COT has already attracted, taking into account the new location at the Harris.

### Sales Potential

Several maps were created to predict the potential ticket sales. Exhibit 4 shows the same portion of the area analyzed in Exhibit 3. This map shows estimates for additional ticket sales potential for each tract. The darker the shade, the more **potential** ticket buyers per census tract.



**Exhibit 4:  
COT's Potential Harris Audience – Location Identified By The Model And Shaded To Show “Super” Tracts, The First One Hundred Tracts Of Quartile 1, And The Second Hundred Tracts Of Quartile 1.**

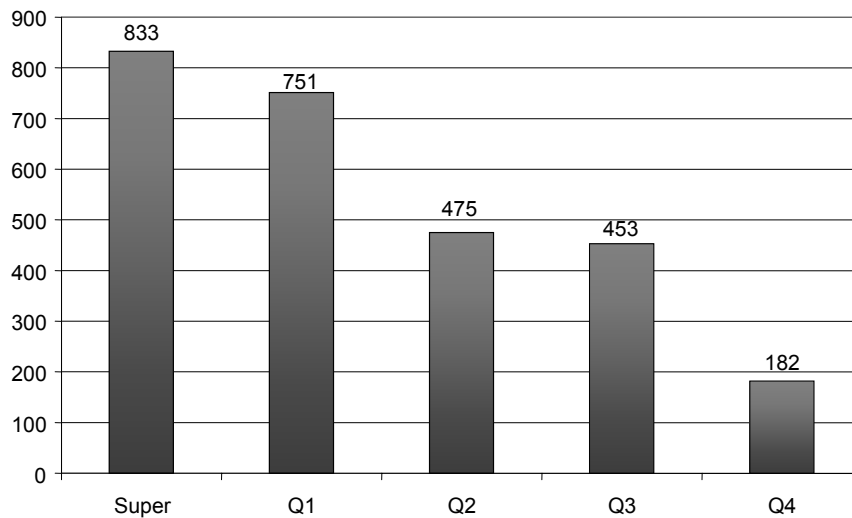


*Quartile Comparisons*

The graphs in Exhibits 5 - 9 show that the top two groups in potential sales, the “Super” group and Quartile 1, differed significantly in their demographic profiles from the remaining three groups.

**Exhibit 5**

**Average Number of College Graduates per Tract**

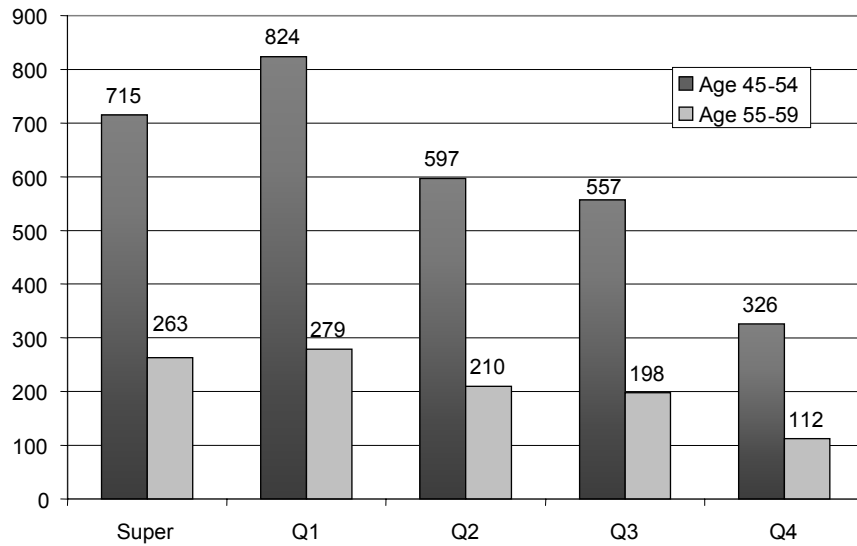




The “Super Tracts” group exceeds Quartile 2 by 75% in the average number of College Graduates.

### Exhibit 6

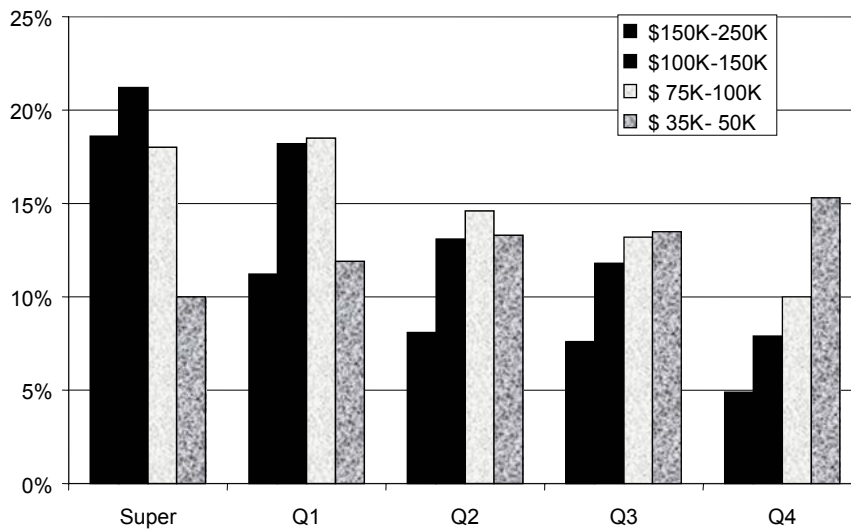
Average Persons per Age Group



“Super” and Q1 have significantly more of the key age group 45-54 than other Quartiles.

### Exhibit 7

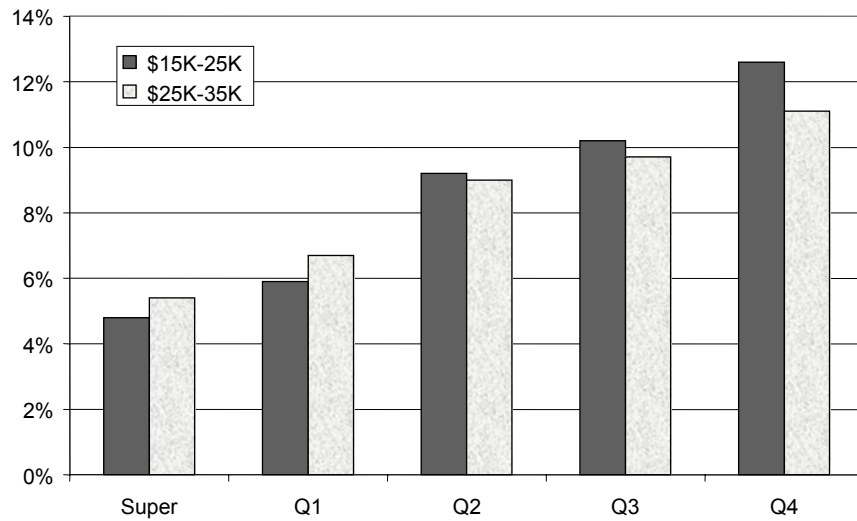
Income Levels with Positive Influence



Approximately 66% of the households in the “Super” tracts have income levels that have positive influence, compared to approximately 38% of the households in Q4.

### Exhibit 8

#### Income Levels with Negative Influence



In contrast, an aggregate 10% of the “Super” have income levels with negative influence, compared to an aggregate 25% of Q4.

### Exhibit 9

#### Average Household Net Worth

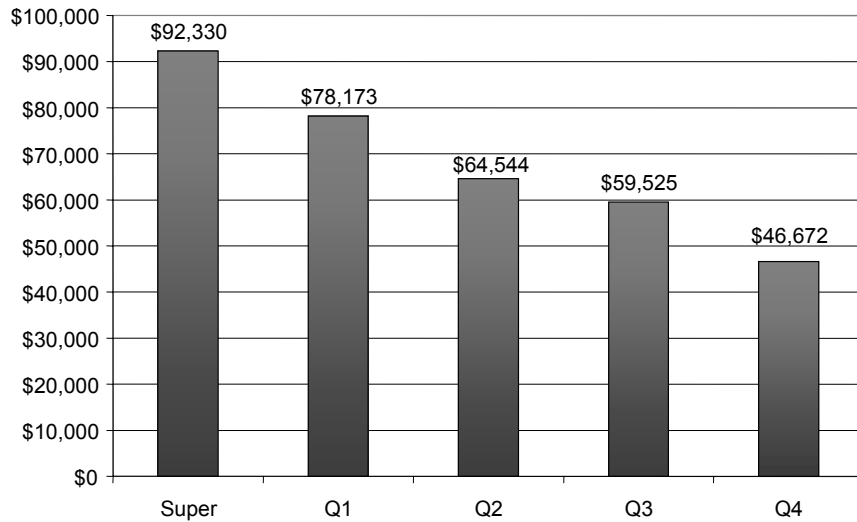


Exhibit 9 shows the “Supers” are 43% higher in household net worth than Q2 and 100% higher than Q4.

## Uses of These Findings

COT was advised to use the index and maps developed through this process to test these census tracts to see how much additional sales revenue could be generated by direct marketing to the potential customers who live in these identified areas. Tract identification numbers provided to COT could be used to purchase lists of residents and offers could be developed to appeal specifically to these potential customers. For example, there are 15,574 households in the nine “Super” tracts. A typical Chicago mailing house might charge approximately US\$0.05 per name or \$48 per thousand, resulting in a cost of approximately \$750 for the mailing list. With the right offer and using a conservative 5% response rate, these households might result in 750 new orders. COT could also decide to purchase lists for any of the top 100 tracts in Quartile 1; these total 211,843 households. If the budget precludes making this approximately \$10,000 investment, the tracts are ranked and COT can add portions of the Quartile incrementally. The second 100 tracts yield another 221,815 households.

Several observations were apparent about these census tracts. First, the *tracts with highest potential* were not found where most of the existing COT audience lived or within the usual focus of most Chicago performing arts marketing efforts. While the model seems to suggest that COT might be reaching saturation with their current direct marketing, the lack of demographic data on the customers makes it impossible to conclude that with certainty. Although the model decayed sales potential by distance, it is striking that the “Super” tracts all lay outside the Chicago city limits.

Next, the Super and Q1 tracts are located with convenient access to public transportation and the highway system. Given that travel is a factor in the ticket purchase decision, COT has an opportunity to develop promotions with the local public transit system to encourage audience members to come to the new, central downtown location by train. At the same time, free or discounted parking could be offered with subscriptions to induce those who drive.

Finally, the highest potential tracts are located in and around Chicago suburbs that are also home to ten suburban performing arts centers with 19 different venues ranging in size from 65 to 1,485 seats and averaging 510 seats per venue. COT has the chance to explore cross marketing opportunities with those organizations for run-outs, abbreviated productions, and outreach programs.

## Evaluation

Since this was a pilot study to test this analytical tool, follow-up is necessary to evaluate the model. In addition to the evaluation, more market research will be helpful in further refining the model. Questions that will form the basis of the evaluation are:

1. Did COT implement more effective targeted marketing to the identified potential market?
2. Did COT consequently increase earned income?

For each of the three years following the initial period analyzed (2000-2002), COT’s audience data will be collected and compared to that from the initial period. By running the evaluation over this period of time, any potential attendance spike resulting from moving to the new venue should have dissipated.

In 2004, the season after the study was concluded, COT began performing in the new Harris Theatre location. An initial follow up interview was conducted with the Marketing Director for the company in July 2004. She reported that the company did not implement targeted marketing to the “Super” tracts as advised in the Uses of These Findings section above. However, COT did expand its mass media buys to include the “Super” tract areas through purchases of zoned editions of the *Chicago Tribune*, a major daily newspaper, as well as through use of suburban newspapers. Exhibit 10 shows a comparison of COT’s last season in the neighborhood venue and its first season in the Harris.

**Exhibit 10**

	<u>2003 season - neighborhood</u>	<u>2004 season - Harris</u>
Single Tickets Sold	6,156	8,283
Subscription Tickets Sold	<u>4,166</u>	<u>7,335</u>
Total Tickets Sold	10,322	15,618

While the overall number of tickets sold increased by 51.3%, the number of subscription tickets sold increased by 76.1%. This is consistent with COT’s objective of increasing subscribers.

Additionally, as of July 15, 2004, and prior to its active subscription campaign for 2005, 57% of the 2004 subscribers had renewed their subscriptions and 270 new subscribers had purchased tickets for next season. (Subscriptions for the 2005 season are priced at US\$90, 150, 195, 270, and 315, respectively.)

As of the date of the writing of this paper, CYC is in the process of analyzing the 2003 through 2005 audience data. The 2003 data will be compared with the initial period to see if it is consistent with, and conforms to, the findings of the study detailed above. The 2004 and 2005 data will be analyzed using the model and the results compared with the original findings. The company indicates it has continued to use mass media rather than direct marketing to reach the “Super Tracts”. The evaluation will tell whether the new customers from 2004 and 2005 have come from the “Super Tracts” or from other parts of the market area. CYC expects that analysis to be completed and ready to report at the AIMAC conference in Montreal in July 2005.

It should be noted that while the evaluation will inform other potential users of these analytical tools and their efficacy, it is clear that no inference from this particular analysis should be made for other organizations’ audiences. The mathematical model yielded results specific to COT. Separate analyses for other companies will allow them to better understand their unique market presence and opportunities, and more effectively target their marketing efforts to their potential audiences.

**Conclusion**

This analysis matched COT’s database to the 2000 Census data because detailed demographic information about the COT audience was unavailable. More detailed demographic data would have made the process more efficient; for example, having such data would have made it possible to indicate where COT has reached saturation in its current marketing efforts, thereby enabling the company to make better choices about its marketing expenditures.

Efficiencies would also be gained in terms of time, analysis required, and most importantly, level of correlation. For example, in studies of for-profit retail companies, a correlation of .93 (as measured by  $R^2$ ) has been obtained with specific company-focused demographic data. The difference between an  $R^2$  of .76 and .93 in a business with the potential to do \$1 million in sales means that the projected sales could be off by almost a quarter million dollars. This result justifies the investment in additional market research to develop more detailed data.

However because resources are often limited in nonprofit performing arts organizations, if a choice between using this model versus traditional audience surveys must be made, this method yields the targeted locations of additional customers. Combining those addresses with the right offer should increase earned income compared to results obtained through traditional research.

## Acknowledgements

The authors are grateful to The Chicago Community Trust for underwriting the pilot project, especially Sarah Solotaroff for her important support and suggestions. Thanks are due also to Chicago Opera Theater: Brian Dickie, General Director; Joan Fox, Director of Finance and Administration; Monica Otting, Director of Development; Leslie Rodriguez, Director of Marketing; Alexis Klossner, Box Office & Customer Relations Manager, who cooperated with the study and made available the organization's audience database.

## Notes

- (1) US Census Bureau, Geography Division. Last revised online November 14, 2000. *Census Tracts and Block Numbering Areas*. Available from: [http://www.census.gov/geo/www/cen\\_tract.html](http://www.census.gov/geo/www/cen_tract.html) (cited June 14, 2004).
- (2) Claritas Inc. 2002. *Consumer Buying Power Variables- Entertainment*. San Diego: Claritas.
- (3) The set of independent variables was selected to minimize any intra-set correlation. This was validated using a standard statistical test. None of the predictor variables exhibited a cross-correlation greater than 0.23.
- (4) Claritas Inc. 2003. *Solution Series Analysis-Retail: Chicago, Illinois MSA by Zip Code and Census Tract*. San Diego: Claritas.
- (5) Claritas Inc. 2003. *Census Tract to Zip Code Cross Reference File*. San Diego: Claritas.
- (6) This refers to Census tracts with greater percentages of Households (HH) with total income in the range indicated.
- (7) Voelker, David H., Orton, Peter Z. and Adams, Scott V. 2001. *Statistics*. New York: HungryMinds Inc., p. 101.
- (8) US Census Bureau, Population Division. Last revised online June 14, 2004. *About Metropolitan and Micropolitan Statistical Areas*. Available from: <http://www.census.gov/population/www/estimates/aboutmetro.html> (cited July 8, 2004.) A Metropolitan Statistical Area, as designated by the United States Office of Management and Budget, includes "a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core."
- (9) Caliper Corporation. 2003. *Maptitude GIS Software License Version 4.6*. Newton, Massachusetts: Caliper Corporation.