Using ZIP Code Business Patterns Data to Measure Alcohol Outlet Density

Stephen A. Matthews, Ph.D.[*], **John D. McCarthy, Ph.D.**, and **Patrick S. Rafail, MA.**
Department of Sociology, Penn State University, 211 Oswald Tower, University Park, PA 16802

Abstract

Some states maintain high-quality alcohol outlet databases but quality varies by state, making comprehensive comparative analysis across US communities difficult. This study assesses the adequacy of using ZIP Code Business Patterns (ZIP-BP) data on establishments as estimates of the number of alcohol outlets by ZIP code. Specifically we compare ZIP-BP alcohol outlet counts with high-quality data from state and local records surrounding 44 college campus communities across 10 states plus the District of Columbia. Results show that a composite measure is strongly correlated (R=0.89) with counts of alcohol outlets generated from official state records. Analyses based on Generalized Estimation Equation models show that community and contextual factors have little impact on the concordance between the two data sources. There are also minimal inter-state differences in the level of agreement. To validate the use of a convenient secondary data set (ZIP-BP) it is important to have a high correlation with the more complex, high quality and more costly data product (i.e., datasets based on the acquisition and geocoding of state and local records) and then to clearly demonstrate that the discrepancy between the two to be unrelated to relevant explanatory variables. Thus our overall findings support the adequacy of using a conveniently available data set (ZIP-BP data) to estimate alcohol outlet densities in ZIP code areas in future research.

Keywords

Alcohol Outlets Density; College Campuses; ZIP Code Business Patterns data; Generalized Estimation Equation Models

1. Introduction

There is broad consensus within the research community recognizing the importance of the role of alcohol outlet density in understanding the role of alcohol availability in contributing to rates of alcohol consumption and how those rates are related a wide range of other

[*]**Corresponding author:** Stephen A. Matthews, Ph.D., Department of Sociology, The Pennsylvania State University, 211 Oswald Tower, University Park, PA 16802-6211. sxm27@psu.edu, Tel: (814) 863-8721, Fax: (814) 863-7216.

Contributors
Matthews and McCarthy designed the study. Matthews coordinated the data preparation. Rafail conducted the statistical analysis. All authors contributed to all sections of the manuscript.

Conflict of Interest
All authors declare that they have no conflicts of interest.

problems in local communities (Gruenewald, 2008). To measure alcohol outlet density in any local community the research must first specify the geographic boundaries and then identify all alcohol outlet establishments within those boundaries. Two measurement strategies have been employed to establish the number of alcohol outlets: *direct observation* and use of official records of state or local alcohol regulatory/licensing bodies. Direct observation is labor intensive rendering its use across widely dispersed communities prohibitive in cost, and thus quite rare, although it has been employed in studies of a single or a small number of dispersed communities (Kuo, et al., 2003; Laranjeira and Hinkly, 2002). The use of official data is far less costly to implement if comprehensive, timely and reliable records of alcohol outlets are available. Moreover, no universal protocols exist for how alcohol outlet data are classified, maintained or disseminated. Multi-site studies are liable to face challenges of data comparability because the temporal and spatial coverage of the data are less likely to conform to uniform standards of collection and categorization and these can vary between and within states.

The present study was motivated by a desire to find efficiencies in building comparable alcohol outlet databases for a larger study that included more than 270 college campus communities located in diverse geographical contexts. Without the resources to undertake detailed fieldwork of all alcohol outlets surrounding each college campus we needed to be able to utilize easily available secondary data to make informed estimates of the number of alcohol outlets.

2. Materials and Methods

Our goal was specifically to assess the adequacy of the U.S. Census Bureau's ZIP Code Business Patterns (hereafter ZIP-BP) estimates of counts of alcohol outlets compared with estimates generated from local and state listings. We did not expect ZIP-BP to provide accurate spatial and temporal estimates of outlet density compared against intensive local-area studies using high-quality local data, although we hoped that it might provide us a highly correlated proxy for alcohol outlets. To validate the use of a convenient data set (i.e., ZIP-BP) it is important to have a high correlation with the other data set and then to clearly demonstrate the discrepancy between the two to be unrelated to other explanatory variables.

We started by constructing campus specific alcohol outlet databases based on state and local licensing agencies. For comparability these campus-specific alcohol outlet databases needed to include street address data, provide complete spatial coverage for the areas surrounding the campuses, ideally be drawn from one state/local licensing authority, and be available for the same year (2006). Due to heterogeneity in alcohol availability it was important that for each state we could secure data on liquor stores, other outlets, drinking places and other broad parameters such as whether the state (or parts of states) alcohol outlets were controlled. We identified ten states and D.C. as having the most comprehensive data in terms of geographic and temporal coverage, the provision of regular updates, the inclusion of complete address information, and the available and ease of use of alcohol outlet listing data. The ten states (plus D.C.*) include: California*, Colorado*, Connecticut*, Florida*, Iowa, Michigan, Missouri*, North Carolina, Ohio and Virginia (* = license states). License states (n=32) issue licenses to private sellers allowing states to indirectly control the sale of alcohol while control states (n=18) directly regulate alcohol sales by controlling its retail and/or wholesale distribution.

The larger study, of which the current validation effort is a part, uses the largest 274 four-year college/university campuses in the U.S. by student enrollments in 2006 (IPEDS, 2006). Our validation study is based on 44 campuses and their surrounding communities and represents a diverse set of campus communities as measured by total population, total

student population, population density as well as social and economic characteristics. For comparative purposes we needed to define the campus community based on an aggregation of ZIP codes. Campus boundaries (polygons) were determined from consultation with multiple sources (campus maps, institution's website, local sources and on-line atlases/databases). We defined each campus community using a 2-mile buffer around the campus boundary, selecting ZIP codes if any part fell within the buffer. Our analysis below is based on 266 ZIP codes and 44 campuses.

To generate a count of alcohol outlets for each 266 ZIP codes we extracted alcohol outlet address from state and local listings, and then address matched each using ESRI/StreetMapPro.

The ZIP-BP includes data on establishment counts by industrial sector. We focused on North American Industry Classification System (NAICS) codes covering outlets where alcohol could be purchased and/or consumed. Specifically we used codes 445110 (Supermarkets and grocery stores except convenience), 445120 (Convenience stores), 445310 (Beer, wine and liquor stores), 447110 (Gasoline stations with convenience stores), 722110 (Restaurants – full-service), and 722410 (Drinking places – alcoholic beverages). Data drawn from the ZIP-BP database for 2006 was then assembled for the same 266 ZIP codes.

3. Results

3.1. Correspondence between State records and ZIP-BP data

Despite the presence of a few outlying data points there is overlap in the distribution for the ZIP-BP and state reports: correlation of 0.89 (R^2=0.79). A decomposition of the individual NAICS indicators used to construct our composite measure is provided in Table 1. The correlations for each NAICS category and the total state count are significant; most are in the range 0.4-0.6, though the correlation between full-service restaurants and the state totals is higher. There are also strong positive correlations for each individual NAICS item and the total count, meriting the inclusion of each in our composite measure. Within each of the NAICS classifications, there are uniformly positive and significant inter-correlations. However, none of the correlations between the fine-grained NAICS classifications are above 0.6. An item analysis suggests that any single NAICS classification will provide an insufficient picture of the total number of alcohol outlets relative to state counts.

3.2. Are discrepancies systematically linked to local demographic characteristics or state?

It is well established that community characteristics—particularly racial composition and socio-economic status—are related to alcohol outlet density (Bluthenthal, et al., 2008). One way to assess the adequacy of the ZIP-BP estimates, compared to the official counts, is to systematically examine whether differences are the result of the same local demographic or community contexts that are known to influence alcohol outlet density. To do so we analytically distinguish between the contextual features that predict the *density* of alcohol outlets and the factors that predict the *discrepancies* between the official and the ZIP-BP counts. If we find few differences between the contextual and community-level processes that predict discrepancies between ZIP-BP and state estimates—even if such factors are related to density rates—that helps to confirm our contention that the ZIP-BP database counts are roughly comparable to the more laboriously generated state count ZIP code estimates.

Our *response variable* is a continuous measure of the difference score between the number of alcohol outlets in a ZIP code as indicated by the official records minus the number of

outlets reported in ZIP-BP. Since our analytical units are ZIP codes around 44 college campuses these cannot be considered to be independent and identically distributed as is required for standard linear regression modeling. We use generalized estimating equations or GEE (Liang and Zeger, 1986) to account for dependencies in the data. We use an identity link function and an exchangeable correlation structure. For each of the 266 ZIP codes, we use 2000 Census data on the number of households, population density, median income, unemployment rates, percent of families in poverty, percent African-American, percent Hispanic, percent Asian, and the percent of all other non-white races. We also use indicators for each of the states.

Model 1 focuses on demographic variables to predict differences between the ZIP-BP and official counts of alcohol outlets. None of these variables are systematically related to differences in the number of alcohol outlets using traditional levels of significance. Indeed, even the model Chi-squared statistic is only 18.21 (*df*=9; p<0.05), suggesting that these demographic differences only provide a small improvement over predicting the mean difference for each case. When indicators for state are included (Model 2) the demographic variables still do not approach statistical significance. These results point towards lower differences in counts for Michigan relative to Virginia (the reference category), though other states are largely indistinguishable from Virginia. It is noteworthy that including the indicators for states yields a notable improvement in model fit based on the Chi-squared (134.91, *df*=19, p<0.001). We tested a model including a dichotomous indicator for license states and because of the potential influence of outliers on the difference score we also re-ran each model to ensure results were not biased by influential data.

Finally we estimated GEE models to explore whether the contextual-level variables predict alcohol outlet density treating the total reported counts from the ZIP-BP and the state counts as following a negative binomial distribution. Using the ZIP-BP data we find significant effects for number of households (+), poverty (+), and African American (−) with a model Chi-squared of 342.04 (*df* =18; p<0.0001). For the state counts, there are significant effects for number of households (+) and poverty (+), and the estimate for percentage African American is negative and nearly significant (Z=−1.83); model Chi-square 289.56 (*df* =18; p<0.0001). These results confirm that predictors in our model of count difference between the two data sources of alcohol outlets are indeed associated with alcohol outlet density.

To summarize, given the lack of statistically significant variables in our regression models coupled the high bivariate correlation between the ZIP-BP and state estimates, we suggest that ZIP-BP estimates provide an adequate representation of an alcohol environment.

4. Discussion

Knowing there is a high correlation (r = .89) between state/local data sources with the more readily available ZIP-BP NAICS data is important for alcohol researchers. Furthermore, the discrepancies between data sources are not the result of demographic characteristics of the campus contexts. Our conclusions only tentatively apply to Michigan. Michigan is a control state that does not control beer or wine but has a monopoly for distilled spirits. We have run models including and excluding Michigan. In Wayne County (Detroit) although convenience stores can sell alcohol those stores part of gas station cannot. A GEE analysis excluding Wayne County does not change our findings.

There are some limitations to this study. First, use of business patterns datasets limits analysis to certain geographic units and for many types of research ZIP codes do not represent a perfect analytical unit (Grubesic, 2008). ZIP codes are not constructed to facilitate the identification of homogeneous social neighborhoods – indeed ZIP code

boundaries change and do not neatly align with other administrative boundaries – and thus for researchers comfortable using other definitions of community then ZIP-BP does not provide an alternative to direct observation or procurement and geocoding of local/state lists. Second, the ZIP-BP has some limitations and there is likely to be an undercount (U.S. Census Bureau, 2009) due to disclosure policies, the existence of multi-unit companies, and unclassified and missing data. Finally, ZIP-BP databases cover approximately 40,000 5-digit ZIP codes in the U.S. but because of the fluctuation in the number of ZIP codes data on specific areas in any year may not be available.

The results we present should be further extended to communities more diverse than only those surrounding college campuses in order to increase our confidence that the patterns hold across diverse settings. We have no reason to believe, however, that the patterns we report are restricted to campus communities. If we are correct, and researchers adopt ZIP-BP alcohol outlet estimates as adequate proxies for community alcohol outlet densities, the number of studies will increase, as these data reduce the costs of expanding knowledge of the causes and consequences of alcohol outlet density on communities.

Acknowledgments

References

Bluthenthal RN, Cohen DA, Farley TA, Scribner R, Beighley C, Schonlau M, et al. Alcohol availability and neighborhood characteristics in Los Angeles, California and southern Louisiana. Journal of Urban Health. 2008; 85(2):191–205. [PubMed: 18228148]

Grubesic TH. Zip codes and spatial analysis: Problems and prospects. Socio-Economic Planning Sciences. 2008; 42(2):129–149.

Gruenewald P. Why do alcohol outlets matter anyway? A look into the future. Addiction. 2008; 103(10):1585–1587. [PubMed: 18821868]

IPEDS - Integrated Post-Secondary Educational Data System. National Center for Education Statistics; Washington D.C.: Dec 21. 2006 2010 Retrieved from http://nces.ed.gov/IPEDS/

Kuo M, Wechsler H, Greenberg P, Lee H. The marketing of alcohol to college students: the role of low prices and special promotions. American Journal of Preventive Medicine. 2003; 25(3):204–211. [PubMed: 14507526]

Laranjeira R, Hinkly D. Evaluation of alcohol outlet density and its relation with violence. Revista de Saude Publica. 2002; 36(4):455–461. [PubMed: 12364919]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22.

U.S. Census Bureau. County Business Patterns: Coverage and Methodology. Dec 21. 2009 2010 Retrieved from http://www.census.gov/econ/cbp/methodology.htm

Research Highlights

Variability in availability and quality of licensed alcohol outlet data raises problems.

We validate a composite measure of alcohol outlets based on secondary data.

The measure has a high correlation with more complex, high quality local data.

The discrepancies between measures are unrelated to relevant predictor variables.

Results support the adequacy of the measure to estimate alcohol outlet densities.

Table 1

Correlation matrix of the total state count and ZIP Code Business Patterns sub-classifications

	State Count	ZIP-BP Count	Supermkt Groceries	Conv Stores	Beer, Wine, Liquor Stores	Gas Stations	Full-Service Rest.	Drinking Places
State Count	1							
ZIP-BP Count	0.89***	1						
Supermkts/ Groceries	0.57***	0.64***	1					
Convenience Stores	0.42***	0.46***	0.55***	1				
Beer, Wine, Liquor Store	0.38***	0.48***	0.50***	0.30***	1			
Gas Stations	0.51***	0.57***	0.37***	0.22***	0.18**	1		
Full-Service Restaurant	0.81***	0.91***	0.40***	0.22***	0.31***	0.36***	1	
Drinking Places	0.58***	0.60***	0.28***	0.29***	0.16*	0.25***	0.45***	1

Notes: *n*=266 ZIP codes and 44 schools.

The sample breakdown for the different types of establishments across the study sites is: Supermarkets/Grocery: 13.15%; Convenience Stores: 6.38%; Beer/Wine/Liquor Stores: 6.48%; Gas Stations/Convenience Stores: 13.71%; Full-Service Restaurants: 50.17%; and Drinking Places: 10.09%.

*
p<0.05;

**
p<0.01;

***
p<0.001 (two-tailed tests)

Table 2

Estimates of the difference between state and ZIP Code Business Patterns counts of alcohol outlets from a Generalized Estimating Equation with an identity link function and exchangeable correlation structure.

	Model 1		Model 2	
# of Households (10,000)	3.95	(2.88)	4.85	(2.70)
Pop. Density (10,000)	2.35	(1.64)	1.85	(1.51)
Median Income (10,000)	−0.11	(0.66)	−1.02	(0.64)
Unemployed (1,000)	1.89	(2.16)	2.25	(2.01)
Families in Poverty (%)	15.44	(18.08)	12.60	(16.81)
African-American (%)	1.53	(7.29)	−0.77	(6.44)
Hispanic (%)	16.20	(12.46)	−4.04	(10.37)
Asian (%)	12.55	(31.46)	−28.73	(30.42)
Other Races (%)	−155.75	(119.65)	−139.56	(113.42)
California			11.83	(8.03)
Colorado			−11.72	(9.52)
Connecticut			−11.78	(8.74)
DC			−0.99	(8.79)
Florida			14.51	(8.55)
Iowa			−15.83	(9.98)
Michigan			−20.06[*]	(8.52)
Missouri			−5.10	(9.10)
North Carolina			−6.15	(8.33)
Ohio			−3.33	(8.81)
Constant	5.64	(5.77)	16.17	(10.77)
Model Chi-Squared (*df*)	18.21[*]	(9)	134.91[***]	(19)

Notes: Standard errors are in parentheses. *n*=266 ZIP codes and 44 schools. Virginia is the reference category.

[*] p<0.05;

[**] p<0.01;

[***] p<0.001