Applied Geography 30 (2010) 203–209

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

An analytical perspective on sporting events attendance: The 2007–2008 US NCAA college bowl games

Daniel A. Griffith*

Geospatial Information Sciences, School of Economic, Political and Policy Sciences, University of Texas at Dallas, 800 W. Campbell Road, GR31, Richardson, TX 75080-03021, USA

Keywords:
Football bowl game
Attendance
Poisson
NCAA

A B S T R A C T

Modern statistical and spatial statistical methodology—based upon generalized linear, mixed and semivariogram/autoregressive modeling theory—offers modeling techniques that are particularly suitable for analyzing sporting event attendances. Counts of attendees must be non-negative, and should follow a Poisson frequency distribution. The rapid expansion in recent years of US NCAA college football bowl games raises the practical research question of whether or not bowl game attendance (i.e., counts) can be predicted from readily available simple measures, which include: a team's win–loss record, distance separating a team's school from the city hosting its bowl game, and the payout of a bowl game. These three sets of variables are employed as covariates for predicting the geographic variation from city-to-city of college football bowl game attendance for the 2007–2008 season. A principal finding of this study is that bowl game attendance appears to be predictable with a contemporary spatial statistical model that is a special case of a Poisson probability model, whose mean is a linear combination of payoff levels and distance of the closer team to a stadium, two factors over which individual bowl organizing officials have some control. This analysis supplements and extends research findings pertaining to the mapping of intercollegiate sports and the geography of visitor attendance at college football games, and offers insight into factors over which cities hosting bowl games have some control.

1. Introduction

In 1989, Bale furnished an overview of the then emerging subdiscipline that has become known as the geography of sports; his 2003 edition of his original book updates this seminal piece of work. More general contemporary interest in this subdiscipline is attested to, in part, by publications penned by Mason and Robins (1991), Chase and Healey (1995), Pan and Baker, 1999, and Leonard (2005). But the annals of sports geography (e.g., *Sport Place*, before its demise) are relatively void of analytical treatments of sporting events attendance that exploit the power of modern spatial statistics and geographic information systems (GIS). The purpose of this paper is to furnish an example of this type of treatment. The thematic topic is attendance at the 2007–2008 US NCAA college football bowl games. As an aside, the methodology presented here can be adopted to provide interesting exercises in, for example, undergraduate quantitative GIS courses.

The research question being addressed here asks whether or not bowl game attendance can be predicted from readily available simple measures. The 2007–2008 bowl game season involved 32 football games whose attendance is characterized

* Tel.: +1 972 883 4950; fax: +1 972 883 4967.
E-mail address: dagriffith@utdallas.edu

0143-6228/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeog.2009.01.005

by (see http://www.ncaafootball.com): 28 sellouts, of which eight were sold out within 24 h of their matchup announcements, and 11 record attendances.

2. A brief history of college bowls[1]

The first major college bowl game, the initial bowl that eventually became known as the Rose Bowl (Pasadena, CA), took place at the end of the 1901 college football season. This remained the only annual bowl through the 1933 season. Three additional bowls [Sugar (New Orleans, LA), Orange (Miami, FL), and Sun (El Paso, TX)] were added for the 1934 post-season, with this set constituting the only bowls until the end of the 1957 college football season. After 57 years, the number of bowls began to increase, with both successful and unsuccessful additions to the original set of four bowls. The trend over the past half-century has been upwards (Fig. 1), with more stability in more recent years. Not counting the all-star bowls, over the years, a total of 47 bowls entered and then exited the set, many being short-lived, and only one successfully relocating.

The original bowl games were held in warm climates and linked to festivals, and consequently were played on New Year's Day. Expansion in the number of bowls as well as technology (e.g., dome stadia) have extended their geographic distribution, as well as the bowl season in order to avoid programming conflicts. One of the 2009 bowls is located in Washington, DC; the International Bowl, established in 2007, is located in Toronto. Meanwhile, games now are played from mid-December through mid-January. Today most bowls are funded through corporate sponsorships (which began in 1988 with the Sugar Bowl). A principal motivation for establishing and hosting these bowls is the promotion of tourism and business for a local community.

3. The NCAA college bowl dataset

As December is entered each year, many US newspapers begin presenting NCAA college football bowl matchup charts, completing these tabulations on a day-to-day basis after each game is finished. Available from *USA Today* is an archive of these annual sporting events dating back to 1998 (http://www.usatoday.com/sports/college/football/bowls.htm), often including attendance at each game. Bowl stadium locations and school team season records beginning with the 2000–2001 season can be retrieved from, for example, http://www.collegefootballresource.com. Web sites such as http://www.football-bowl.com/2007-bowl-games.html also list the payout for each bowl game. Consequently, a dataset can be compiled comprising participating school locations, team win percentages (note: the relatively recent BCS ratings are not available for all schools), payout value of a bowl game, and attendance figures, which can be extracted from such web sites as http://www.ncaafootball.com/index.php?s=&url_channel_id=34&url_article_id=11994&change_well_id=2.

Given that their names are known, locational coordinates of schools and bowl stadia can be gleaned from Google Earth. For example, the Pennsylvania State University (PSU) played Texas A&M University in the 2008 Alamo Bowl. Retrieved coordinates for this matchup are as follows (Fig. 2):

Location	Latitude (N)			Longitude (W)		
PSU	40°	48′	42.73″	77°	51′	22.21″
Texas A&M	30°	37′	7.16″	96°	20′	8.86″
Alamodome	29°	25′	6.26″	98°	28′	47.42″

Problems with Google Earth location search failures can be resolved by including zip code information.

These quantitative data can be organized into a dataset that includes attendance (ATT) as a dependent (i.e., response) variable, and the following as independent (i.e., covariate) variables: payout (in $ millions; PAY), latitude and longitude in decimal degrees for each of three locations, two season win/loss ratios, and the time sequence (SEQ) for when a bowl game was played. The two win/loss ratios can be organized such that the first variable records the team with the better record (WIN#1), whereas the second variable records the team with the poorer record (WIN#2). Next, the decimal degrees can be used to compute great circle distances (say, in 1000's of miles), with the first distance variable (DIST#1) recording the closer school to a stadium, and the second distance variable (DIST#2) recording the other school.

4. Data preprocessing

Attendance figures are counts, with a minimum value of 0, and a fuzzy upper limit defined by the size of a stadium. This context suggests that attendance should be treated as a Poisson variable. Because these counts are very sizeable, their associated frequency distribution tends to mimic a bell-shaped curve (Fig. 3a). Because a Poisson random variable converges upon a bell-shaped curve as its mean increases beyond, say, several thousand, the logarithm of these counts (LNATT) also has a frequency distribution that tends to mimic a bell-shaped curve (Fig. 3b); this latter relationship actually is an approximation.

[1] This section was gleaned from more than 100 Wikipedia annual college football season entries (e.g., http://en.wikipedia.org/wiki/1933_college_football_season).

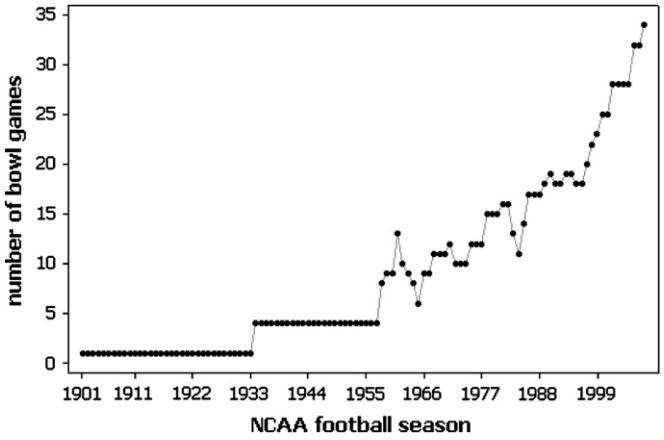

Fig. 1. Annual time series plot of number of post-season college football bowls, 1901–2008 football seasons.

This situation allows a reasonable comparison to be made between these two probability model specifications, allowing one to determine whether or not an analysis based upon the much simpler bell-shaped curve gives acceptably good results.

The simplest relationships involve trend lines that are straight (i.e., linear), rather than curved (i.e., non-linear). Bivariate scatterplots uncover non-linear relationships between both the original and log-transformed ATT and PAY (Fig. 4a), and both the original and log-transformed ATT and WIN#2 (Fig. 4b). In order to straighten these curved trend lines, the following Box–Tidwell transformations (Montgomery, Peck, & Vining, 2001) were identified for these two cases (using SAS PROC NLIN):

$$\text{TRPAY} = -1/(\text{PAY} + 0.73), \text{ and}$$
$$\text{TRWIN\#2} = -1/(\text{WIN\#2} + 0.57).$$

In each of these cases, the inverse function converts a positive relationship into a negative relationship. Multiplying by a -1 restores the nature of the original relationship, while the transformation itself improves upon the original degree of the straight-line relationship. The accompanying improvements are conspicuous in Fig. 4. Meanwhile, bivariate scatterplots fail to uncover apparent relationships between ATT and WIN#1. Of note is that, because of its remote location, the University of Hawaii is an outlier for both distance variables.

5. A spatial analysis of attendance

Conceptually based expectations imply the following hypotheses: attendance increases as

(1) payout increases,

(2) the win/loss ratio increases, and

(3) distance to a school decreases.

Given television coverage of college football games, as well as the uniqueness and esteem of each bowl game, the order in which games are played during the "bowl season" (i.e., the temporal sequencing) is not expected to be related to attendance. Rather, the unmeasured effects of competing sporting events may well play a more important role here.

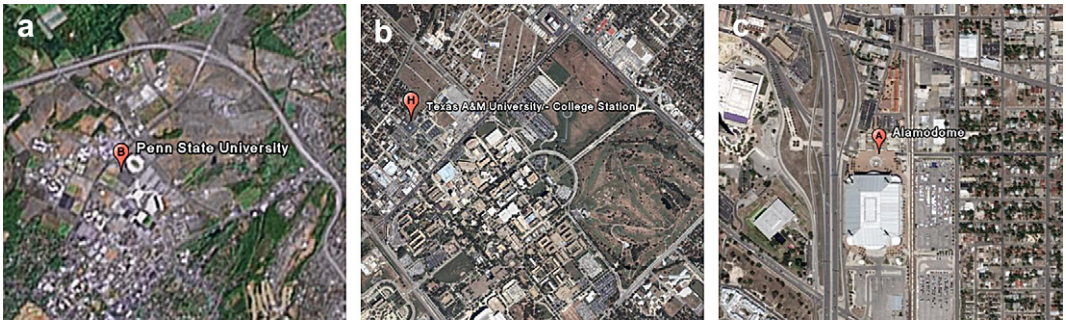

Fig. 2. Left (a): The Pennsylvania State University. Center (b): Texas A&M University. Right (c): Alamodome, stadium for the 2007 Alamo Bowl.

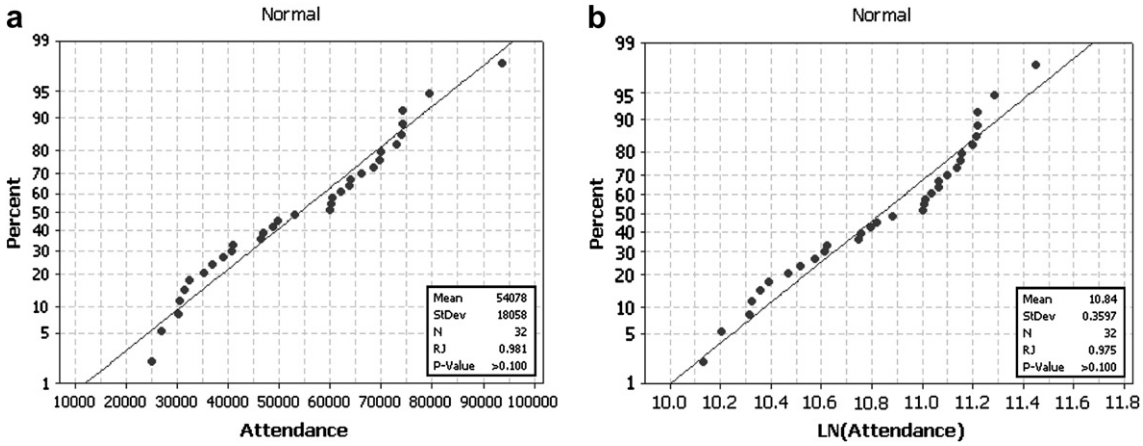

Fig. 3. Normal quantile plots. Left (a): attendance counts. Right (b): natural logarithm of attendance counts.

The initial model specification included all six independent variables, and then the least important one, according to selected statistical criteria, was eliminated, sequentially, until only statistically important independent variables remained in the model specification. This stepwise backward elimination regression procedures screened the set of predictor variables TR(PAY), WIN#1, TR(WIN#2), DIST#1, DIST#2, and SEQ. Both the linear (for a bell-shaped probability model) and the generalized linear (for a Poisson probability model) stepwise procedure (implemented both in SAS and STATA) identify as most important TR(PAY), DIST#1, and TR(WIN#2), in that order. The threshold significance level probability used for removal was 0.10 (i.e., at most a probability of 0.1 for a null hypothesis rejection).

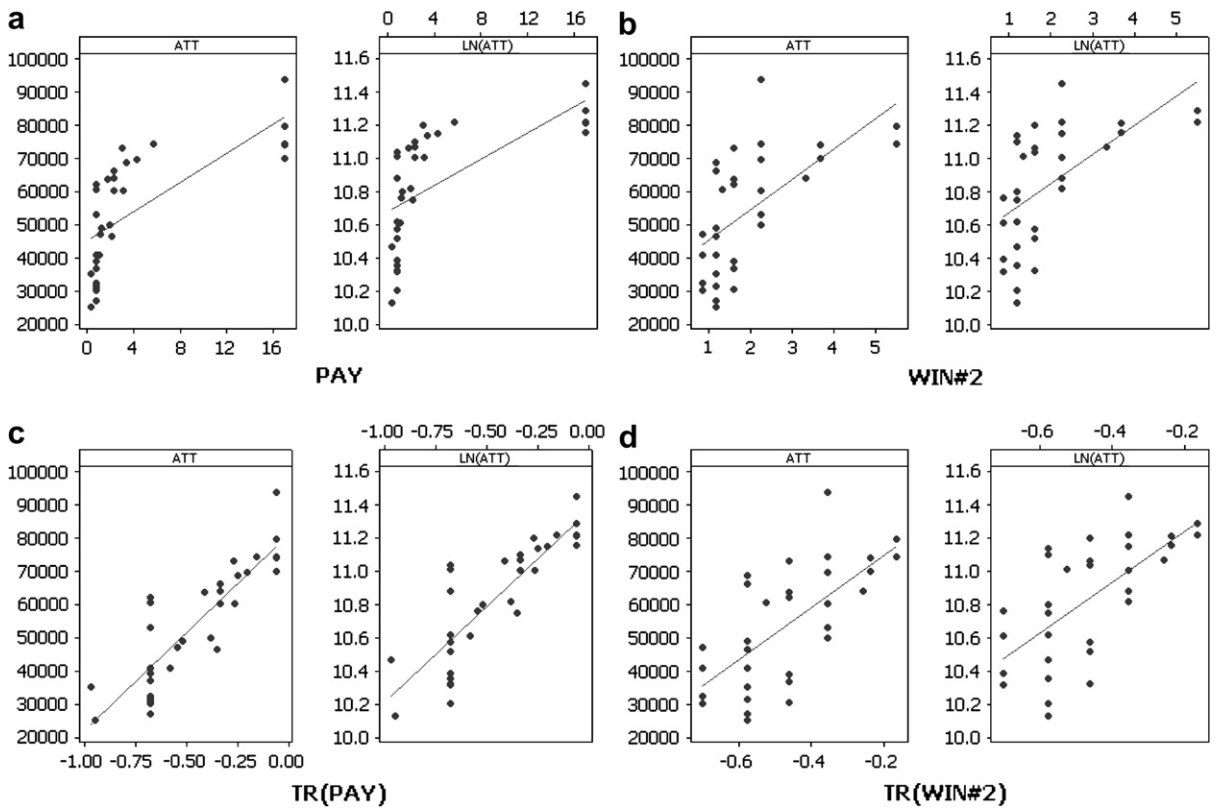

Fig. 4. Scatterplots with trend lines. Top left (a): the raw payout figures versus attendance. Top right (b): the lower win/loss ratio versus attendance. Bottom left (c): the transformed payout figures versus attendance. Bottom right (d): the transformed lower win/loss ratio versus attendance.

Diagnostics assessing the appropriateness of selected probability models also merit scrutiny. The normal (i.e., Gaussian) approximation conforms closely to a bell-shaped curve for both raw and log-transformed attendance (Fig. 3). The respective probabilities (Pr) for the Shapiro–Wilk normality diagnostic statistic, assuming a true null hypothesis of normality, are 0.2055 and 0.0825, both of which imply conformity with a bell-shaped curve at a 5% level of significance. These probabilities suggest that raw rather than log-transformed attendance should be modeled. But the linear regression residuals imply exactly the opposite, with the respective probabilities changing to 0.2188 and 0.5533. In addition, the log-transformed attendance figures are in keeping with a Poisson probability model specification, making them preferable on the basis of conceptual considerations. Meanwhile, the test for constant variance (i.e., heteroscedasticity) renders respective chi-square probabilities of 0.3484 and 0.4015, again supporting the use of log-transformed attendance figures coupled with a bell-shaped curve in an analysis.

Diagnostics for a Poisson probability specification include evaluation of over-dispersion [i.e., extra-Poisson variation such that the variance (σ^2) exceeds rather than equals the mean (μ), or $\sigma^2 > \mu$]. This outcome results from a myriad of different types of attendees mixing together to form attendance figures. Consequently, a non-constant (i.e., varying) mean is assumed here for attendance—specifically, a gamma-distributed mean—converting the Poisson into a negative binomial model specification. In other words, attendance is still correctly conceptualized as counts, but the same single Poisson mean does not apply to all attendees or stadia. This specification allows over-dispersion to be accounted for by including a dispersion parameter (v) that inflates the variance: $\sigma^2 = \mu(1 + v \times \mu)$.

Regression model estimation results are reported in Table 1. The variables are listed in this table in the order in which they were selected by a stepwise (linear/negative binomial) regression procedure. Of note is that the pseudo-R^2 (i.e., squared approximate multiple correlation coefficient) values increase with entry of the first two variables, and then slightly decrease with entry of the third variable. But this third variable plays an important role in controlling for heteroscedasticity/overdispersion (i.e., the extent to which the variance exceeds the mean): In other words, TR(WIN#2) helps to stabilize the log-attendance residual variable in the bell-shaped curve approximation, and further reduces over-dispersion in the negative binomial specification. It also accounts for roughly 2.5% of the geographic variation in the log-attendance bell-shaped curve approximation linear regression model specification. Consequently, the statistical prediction equation performs well across all 32 of the bowl games. Scatterplots with the generated predicted values appear in Fig. 5; results for the two probability models essentially are indistinguishable, suggesting that results from the simpler one based upon the bell-shaped curve are adequate.

Variable entered	Pr (chi-square)	Dispersion parameter
TR(PAY)	0.0089	0.0359
DIST#1	0.0139	0.0270
TR(WIN#2)	0.4015	0.0244

Interpretation of the statistical inference results includes that bowl game attendance does not covary with: (1) distance of the team whose school is further from a bowl stadium; (2) the better win/loss ratio; and (3) the sequencing of bowl games. This particular finding suggests the hypothesis that fans of some schools attend bowl games regardless of distance; this hypothesis should be evaluated in subsequent research. In contrast, bowl game attendance covaries with: (1) payout; (2) distance of the team whose school is closer to a bowl stadium; and (3) the poorer win/loss ratio. Not surprisingly, payout accounts for most of the geographic variation across bowl game stadium attendance. Its positive sign is consistent with the expectation of a tendency for attendance to increase with payout. The sign for distance is as expected: as distance of the closer school playing in a stadium increases, bowl game attendance tends to decrease. And, the sign for WIN#2 is sensible: as the win/loss ratio of the team with the poorer record increases, attendance tends to increase. This outcome may well relate to fans of the weaker of two teams in a bowl game matchup preferring to stay home, rather than attend a game in person, when they perceive a lower chance that their team will win the game. The accompanying slight drop in the pseudo-R^2 value with selection of TR(WIN#2) is possible because non-linear regression is being employed; it suggests a slight correction for prediction overfitting that is attributable to the presence of non-constant variance (e.g., extra-Poisson variation).

Statistically speaking, the bell-shaped curve approximation and negative binomial specifications render almost identical results. Changing from the approximate to the more appropriate counts random variable conceptualization results in only a slight decrease in the parameter estimate standard errors, as expected. Both specifications account for roughly 82% of the

Table 1
Summary regression estimation results for attendance prediction model specifications.

Variable	Log-Gaussian approximation			Negative binomial		
	Coefficient	Standard error	Change in pseudo-R^2[a]	Coefficient	Standard error	Change in pseudo-R^2
Intercept	11.5646	0.1014	***	11.5671	0.0960	***
TR(PAY)	0.9285	0.1577	0.7536	0.8993	0.1436	0.7536
DIST#1	−0.1836	0.0572	0.0737	−0.1916	0.0508	0.0761
TR(WIN#2)	0.4701	0.2693	−0.0058	0.4698	0.2499	−0.0060

[a] Multiple correlation based upon back-transformed predicted values.

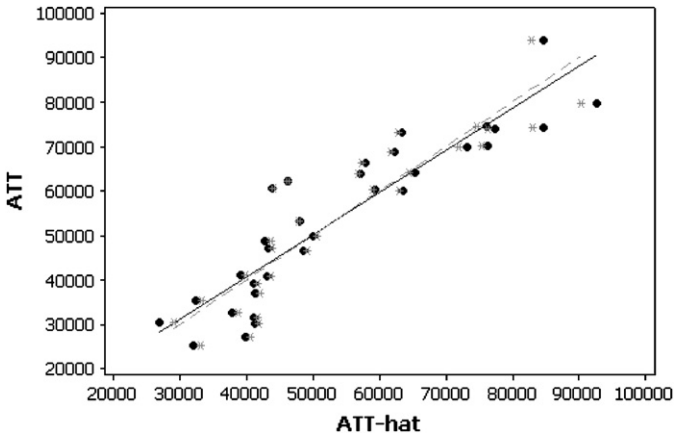

Fig. 5. Scatterplots of observed versus predicted college football bowl game attendance, 2007–2008: negative binomial predictions (black), and Gaussian approximations (gray).

geographic variation in 2007–2008 college football bowl stadium attendance. And, diagnostic statistics suggest that the statistical inferences drawn from these two models are reasonably sound.

6. Discussion

The geographic variation from city-to-city of college football bowl game attendance in 2007–2008 primarily is accounted for by the geographic variation in concomitant payoffs. This description is augmented slightly when the geographic variation in distance of the closer team to a bowl stadium is taken into account. Statistical model variance assumptions improve if the geographic variation in the lower team win/loss ratio also is taken into account, bolstering the inferential basis of the attendance prediction statistical models.

Because the lower team win/loss ratio variable accounts for virtually no variation, while helping to stabilize the variance, as well as the data being for a non-random sample (i.e., teams are not selected at random for matchups, and audience members do not attend a given bowl game at random) and attendance being composed of a large mixture of individuals, inclusion of a random effects term—which accounts for unknown independent variables missing from the model specification, and inter-attendee correlations such as family ties and friendship networks—is merited. Furthermore, the analysis thus far overlooks the issue of spatial autocorrelation in bowl stadium attendance. Accordingly, because the cities involved are spaced quite far apart geographically, presumably the only source of spatial autocorrelation is from three games being played in New Orleans, two in Orlando, and two in San Diego. Combining these two data features argues for the inclusion of a spatially structured random effects term. SAS PROC GLIMMIX implements such a model specification, supporting the use of an exponential, Gaussian, power and spherical semivariogram error model specification (see Griffith & Layne, 1999). Estimation including a spatially structured random effects term, geographically structured with an exponential semivariogram model, to account for residual spatial autocorrelation results in an extremely small geographic correlation field range, in keeping with the presence of the aforementioned three sets of replicates. This spatial structuring increases the pseudo-R^2 by about 1%, and has statistically non-significant exponential semivariogram model parameter estimates. Removing the spatial structuring and including an unstructured random effects term preserves the 1% increase in variance accounted for, but results in the lower team win/loss ratio variable becoming non-significant. This latter finding is consistent with TR(WIN#2) functioning mostly as a variance stabilizing covariate.

Variable	Coefficient	Standard error	Change in pseudo-R^2
Intercept	11.4307	0.0536	***
TR(PAY)	1.0619	0.1509	0.7463
DIST#1	−0.2041	0.0631	0.0658

Therefore, the final mixed model estimation results are as follows: Results for DIST#1 display a conspicuous change, whereas the intercept and TR(PAY) are reasonably similar to their fixed effects model counterparts. In other words, neither spatial autocorrelation nor the previously mentioned unmeasured effects appear to play an important role here. Consequently, college football bowl game attendance appears to be predictable with a negative binomial model specification including payoff levels and distance of the closer team to a stadium as covariates, two factors over which individual bowl organizing officials have some control.

In order to bolster findings reported here, this analysis needs to be replicated for other football seasons, a topic of future research.

7. Some policy implications

As mentioned earlier, communities establish college football bowls to promote tourism and local business, or more precisely, to increase community spirit and pride, attract tourists to a city, and stimulate local economic development. These events often result in windfalls for their host cities. These economic impacts arise from what is called event tourism (i.e., a bowl game is a single, although annual, event; see Getz, 2008), with local booster groups forecasting local sales, sales tax revenues, and job creation attributable to bowl games (Coates & Depken, 2006). Mondello and Rishe (2004) report that the proximity of bowl matchup schools to a host city contributes to the resulting impact, in terms of out-of-town visitor and organization spending, corroborating one of the findings summarized in this paper. These researchers also report that the number of non-local visitors contributes to this impact, a factor that is magnified by the package of bowl activities constituting a multiple-day event. This variable relates to attendance, which is the response variable analyzed in this paper, and is a function of stadium size.

To date, the literature contains little discussion about payout, which is found to be the principal explanatory factor in this study. This feature of bowl games is more broadly competitive, with cities seeking sponsorship from a large but limited number of national corporations. The 34 bowls for the 2008 football season, which involved 68 of a possible 119 teams (of which 72 were eligible by having at least a 50% win record), may well represent the maximum or near-maximum number of possible bowls (a 35th bowl, the Rocky Mountain Bowl in Salt Lake City, was not licensed for the 2008/2009 bowl season). Diminishing returns to the licensing of additional bowl games are materializing.

One policy implication implied by this paper is that cities hosting NCAA football bowl games should concentrate on increasing sponsorship revenues for payout. Bowls with bigger payouts can attract teams with better records, currently subject to the top five bowl games being governed by the Bowl Championship Series. The present structure also supports the implementation of a national championship post-season. In 2008/2009, the 34 bowls were distributed across 30 cities. Adding another two cities (e.g., Salt Lake City) would allow a play-off that involves 64 teams and five rounds. This structure not only would allow two additional bowl games to be licensed by the NCAA, but it also would increase the number of post-season games played by 28 (and would continue to allow Miami, New Orleans, Orlando, and San Diego each to host multiple games). Bowls in more cities could be licensed, and/or more cities could hold multiple licenses. And, the bowl season would not be much longer. The net effect would include increased payouts, albeit over a series of games rather than for a single game.

8. Summary and conclusions

This paper presents an analytical treatment of 2007–2008 NCAA college football bowl games attendance that exploits the power of modern spatial statistics and geographic information systems (GIS). Its principal finding is that bowl game attendance is predictable with a contemporary spatial statistical model, and that important predictive factors include payoff levels and distance of the closer team to a stadium. Proximity is an important local economic impact factor already identified in the literature, and confirmed here. The analytical finding for payoff levels is new, although conceptually not surprising. In addition, these findings can be used as input to policy debates in cities that host or contemplate hosting college football bowl games by offering sound inferential evidence about these factors.

Future research needs to replicate this study for other years. It also should include analyzing the space-time series of those bowls that have existed for many years. Finally, the methodology presented here furnishes an appealing prototype exercise for quantitative GIS courses.

References

Bale, J. (2003). *Sports geography* (2nd ed.). NY: Routledge.
Chase, J., & Healey, M. (1995). The spatial externality effects of football matches and rock concerts: the case of Portman Road Stadium, Ipswich, Suffolk. *Applied Geography, 15*, 18–34.
Coates, D., & Depken, C. (2006). Mega-Events: Is the Texas-Baylor game to Waco what the Super Bowl is to Houston? Working Paper Series, Paper No. 06-06. Limoges, France: International Association of Sports Economists.
Getz, D. (2008). Event tourism: definition, evolution, and research. *Tourism Management, 29*, 403–428.
Griffith, D., & Layne, L. (1999). *A casebook for spatial statistical data analysis: a compilation of analyses of different thematic datasets*. NY: Oxford University Press.
Leonard, J. (2005). The geography of visitor attendance at college football games. *Journal of Sport Behavior, 28*, 231–252.
Mason, C., & Robins, R. (1991). The spatial externality fields of football stadiums: the effects of football and non-football uses at Kenilworth Road, Luton. *Applied Geography, 11*, 251–266.
Mondello, M., & Rishe, P. (2004). Comparative economic impact analyses: differences across cities, events, and demographics. *Economic Development Quarterly, 18*, 331–342.
Montgomery, D., Peck, E., & Vining, G. (2001). *Introduction to linear regression analysis* (3rd ed.). NY: Wiley.
Pan, D., & Baker, J. (1999). Mapping of intercollegiate sports relative to selected attributes as determined by a product differentiation strategy. *Journal of Sport Behavior, 22*, 69–82.